A Unified Framework for Uniform Signal Recovery in Nonlinear Generative Compressed Sensing

Junren Chen*
University of Hong Kong
chenjr58@connect.hku.hk

Jonathan Scarlett
National University of Singapore
scarlett@comp.nus.edu.sg

Michael K. Ng
Hong Kong Baptist University
michael-ng@hkbu.edu.hk

Zhaoqiang Liu*
UESTC
zqliu12@gmail.com

Abstract

In generative compressed sensing (GCS), we want to recover a signal $\boldsymbol{x}^* \in \mathbb{R}^n$ from m measurements ($m \ll n$) using a generative prior $\boldsymbol{x}^* \in G(\mathbb{B}_2^k(r))$, where G is typically an L-Lipschitz continuous generative model and $\mathbb{B}_2^k(r)$ represents the radius-r ℓ_2-ball in $\mathbb{R}^k$. Under nonlinear measurements, most prior results are non-uniform, i.e., they hold with high probability for a fixed $\boldsymbol{x}^*$ rather than for all $\boldsymbol{x}^*$ simultaneously. In this paper, we build a unified framework to derive uniform recovery guarantees for nonlinear GCS where the observation model is nonlinear and possibly discontinuous or unknown. Our framework accommodates GCS with 1-bit/uniformly quantized observations and single index models as canonical examples. Specifically, using a single realization of the sensing ensemble and generalized Lasso, *all* $\boldsymbol{x}^* \in G(\mathbb{B}_2^k(r))$ can be recovered up to an ℓ_2-error at most ϵ using roughly $\tilde{O}(k/\epsilon^2)$ samples, with omitted logarithmic factors typically being dominated by $\log L$. Notably, this almost coincides with existing non-uniform guarantees up to logarithmic factors, hence the uniformity costs very little. As part of our technical contributions, we introduce the Lipschitz approximation to handle discontinuous observation models. We also develop a concentration inequality that produces tighter bounds for product processes whose index sets have low metric entropy. Experimental results are presented to corroborate our theory.

1 Introduction

In compressed sensing (CS) that concerns the reconstruction of low-complexity signals (typically sparse signals) [5, 6, 15], it is standard to employ a random measurement ensemble, i.e., a random sensing matrix and other randomness that produces the observations. Thus, a recovery guarantee involving a single draw of the measurement ensemble could be *non-uniform* or *uniform* — the non-uniform one ensures the accurate recovery of any fixed signal with high probability, while the uniform one states that one realization of the measurements works simultaneously for all structured signals of interest. Uniformity is a highly desired property in CS, since in applications the measurement ensemble is typically fixed and should work for all signals [17]. Besides, the derivation of a uniform guarantee is often significantly harder than a non-uniform one, making uniformity an interesting theoretical problem in its own right.

*Corresponding authors.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Inspired by the tremendous success of deep generative models in different applications, it was recently proposed to use a generative prior to replace the commonly used sparse prior in CS [2], which led to numerical success such as a significant reduction of the measurement number. This new perspective for CS, which we call generative compressed sensing (GCS), has attracted a large volume of research interest, e.g., nonlinear GCS [29, 33, 45], MRI applications [24, 46], and information-theoretic bounds [27, 34], among others. This paper focuses on the uniform recovery problem for nonlinear GCS, which is formally stated below. Our main goal is to build a unified framework that can produce uniform recovery guarantees for various nonlinear measurement models.

Problem: Let $B_2^k(r)$ be the ℓ_2-ball with radius r in $\mathbb{R}^k$. Suppose that $G : B_2^k(r) \to \mathbb{R}^n$ is an L-Lipschitz continuous generative model, $a_1, \ldots, a_m \in \mathbb{R}^n$ are the sensing vectors, $x^\star \in \mathcal{K} := G(B_2^k(r))$ is the underlying signal, and we have the observations $y_i = f_i(a_i^\top x^\star),\ i = 1, \ldots, m$, where $f_1(\cdot), \ldots, f_m(\cdot)$ are possibly unknown,[2] possibly random non-linearities. Given a single realization of $\{a_i, f_i\}_{i=1}^m$, under what conditions we can uniformly recover all $x^\star \in \mathcal{K}$ from the corresponding $\{a_i, y_i\}_{i=1}^m$ up to an ℓ_2-norm error of ϵ?

1.1 Related Work

We divide the related works into nonlinear CS (based on traditional structures like sparsity) and nonlinear GCS.

Nonlinear CS: Beyond the standard linear CS model where one observes $y_i = a_i^\top x^\star$, recent years have witnessed rapidly increasing literature on nonlinear CS. An important nonlinear CS model is 1-bit CS that only retains the sign $y_i = \mathrm{sign}(a_i^\top x^\star)$ [3, 22, 41, 42]. Subsequent works also considered 1-bit CS with dithering $y_i = \mathrm{sign}(a_i^\top x^\star + \tau_i)$ to achieve norm reconstruction under sub-Gaussian sensing vectors [9, 14, 48]. Besides, the benefit of using dithering was found in uniformly quantized CS with observation $y_i = \mathcal{Q}_\delta(a_i^\top x^\star + \tau_i)$, where $\mathcal{Q}_\delta(\cdot) = \delta\lfloor \frac{\cdot}{\delta} \rfloor + \frac{\delta}{2}$ is the uniform quantizer with resolution δ [8, 48, 52]. Moreover, the authors of [16, 43, 44] studied the more general single index model (SIM) where the observation $y_i = f_i(a_i^\top x^\star)$ involves (possibly) unknown nonlinearity f_i.

While the restricted isometry property (RIP) of the sensing matrix $A = [a_1, \ldots, a_m]^\top$ leads to uniform recovery in linear CS [4, 15, 49], this is not true in nonlinear CS. In fact, many existing results are non-uniform [9, 16, 21, 41, 43, 44, 48], and some uniform guarantees can be found in [7, 8, 14, 17, 41, 42, 52]. Most of these uniform guarantees suffer from a slower error rate.

The most relevant work to this paper is the recent work [17] that described a unified approach to uniform signal recovery for nonlinear CS. The authors of [17] showed that in the aforementioned models with k-sparse $x^\star$, a uniform ℓ_2-norm recovery error of ϵ could be achieved via generalized Lasso using roughly k/ϵ^4 measurements [17, Section 4]. In this work, we build a unified framework for uniform signal recovery in nonlinear GCS. To achieve a uniform ℓ_2-norm error of ϵ in the above models with the generative prior $x^\star \in G(B_2^k(r))$, our framework only requires a number of samples proportional to k/ϵ^2. Unlike [17] that used the technical results [36] to bound the product process, we develop a concentration inequality that produces a tighter bound in the setting of generative prior, thus allowing us to derive a sharper uniform error rate.

Nonlinear GCS: Building on the seminal work by Bora *et al.* [2], numerous works have investigated linear or nonlinear GCS [1, 11, 12, 19, 20, 23, 25, 30, 39, 40, 51], with a recent survey [47] providing a comprehensive overview. Particularly for nonlinear GCS, 1-bit CS with generative models has been studied in [26, 31, 45], and generative priors have been used for SIM in [29, 32, 33]. In addition, score-based generative models have been applied to nonlinear CS in [10, 38].

The majority of research for nonlinear GCS focuses on non-uniform recovery, with only a few exceptions [33, 45]. Specifically, under a generative prior, [33, Section 5] presented uniform recovery guarantees for SIM where $y_i = f_i(a_i^\top x^\star)$ with deterministic Lipschitz f_i or $f_i(x) = \mathrm{sign}(x)$. Their proof technique is based on the local embedding property developed in [31], which is a geometric property that is often problem-dependent and currently only known for 1-bit measurements and deterministic Lipschitz link functions. In contrast, our proof technique does not rely on such

[2]In order to establish a unified framework, our recovery method (2.1) involves a parameter T that should be chosen according to f_i. For the specific single index model with possibly unknown f_i, we can follow prior works [33, 43] to assume that $Tx^\star \in \mathcal{K}$, and recover $Tx^\star$ without using T. See Remark 5 for more details.

geometric properties and yields a unified framework with more generality. Furthermore, [33] did not consider dithering, which limits their ability to estimate the norm of the signal.

The authors of [45] derived a uniform guarantee from dithered 1-bit measurements under bias-free ReLU neural network generative models, while we obtain a uniform guarantee with the comparable rate for more general Lipschitz generative models. Additionally, their recovery program differs from the generalized Lasso approach (Section 2.1) used in our work. Specifically, they minimize an ℓ_2 loss with $\|\mathbf{x}\|_2^2$ as the quadratic term, while generalized Lasso uses $\|\mathbf{A}\mathbf{x}\|_2^2$ that depends on the sensing vector. As a result, our approach can be readily generalized to sensing vectors with an unknown covariance matrix [33, Section 4.2], unlike [45] that is restricted to isotropic sensing vectors. Under random dithering, while [45] only considered 1-bit measurements, we also present new results for uniformly quantized measurements (also referred to as multi-bit quantizer in some works [13]).

1.2 Contributions

In this paper, we build a unified framework for uniform signal recovery in nonlinear GCS. We summarize the paper structure and our main contributions as follows:

- We present Theorem 1 as our main result in Section 2. Under rather general observation models that can be discontinuous or unknown, Theorem 1 states that the uniform recovery of all $\mathbf{x}^* \in G(\mathbb{B}_2^k(r))$ up to an ℓ_2-norm error of ϵ can be achieved using roughly $O\big(\frac{k\log L}{\epsilon^2}\big)$ samples. Specifically, we obtain uniform recovery guarantees for 1-bit GCS, 1-bit GCS with dithering, Lipschitz-continuous SIM, and uniformly quantized GCS with dithering.
- We provide a proof sketch in Section 3. Without using the embedding property as in [33], we handle the discontinuous observation model by constructing a Lipschitz approximation. Compared to [17], we develop a new concentration inequality (Theorem 2) to derive tighter bounds for the product processes arising in the proof.

We also perform proof-of-concept experiments on the MNIST [28] and CelebA [35] datasets for various nonlinear models to demonstrate that by using a single realization of $\{\mathbf{a}_i, f_i\}_{i=1}^m$, we can obtain reasonably accurate reconstruction for multiple signals. Due to the page limit, the experimental results and detailed proofs are provided in the supplementary material.

1.3 Notation

We use boldface letters to denote vectors and matrices, while regular letters are used for scalars. For a vector $\mathbf{x}$, we let $\|\mathbf{x}\|_q$ ($1 \le q \le \infty$) denote its ℓ_q-norm. We use $\mathbb{B}_q^n(r) := \{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{z}\|_q \le r\}$ to denote the ℓ_q ball in $\mathbb{R}^n$, and $(\mathbb{B}_q^n(r))^c$ represents its complement. The unit Euclidean sphere is denoted by $\mathbb{S}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$. We use C, C_i, c_i, c to denote absolute constants whose values may differ from line to line. We write $A = O(B)$ or $A \lesssim B$ (resp. $A = \Omega(B)$ or $A \gtrsim B$) if $A \le CB$ for some C (resp. $A \ge cB$ for some c). We write $A \asymp B$ if $A = O(B)$ and $A = \Omega(B)$ simultaneously hold. We sometimes use $\tilde{O}(\cdot)$ to further hide logarithmic factors, where the hidden factors are typically dominated by $\log L$ in GCS, or $\log n$ in CS. We let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Given $\mathcal{K}_1, \mathcal{K}_2 \subset \mathbb{R}^n$, $\mathbf{a} \in \mathbb{R}^n$ and some $a \in \mathbb{R}$, we define $\mathcal{K}_1 - \mathcal{K}_2 := \{\mathbf{x}_1 - \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{K}_1, \mathbf{x}_2 \in \mathcal{K}_2\}$, $\mathbf{a} + \mathcal{K}_1 := \{\mathbf{a}\} + \mathcal{K}_1$, and $a\mathcal{K}_1 := \{a\mathbf{x} : \mathbf{x} \in \mathcal{K}_1\}$. We also adopt the conventions of $a \wedge b = \min\{a, b\}$, and $a \vee b = \max\{a, b\}$.

2 Main Results

We first give some preliminaries.

Definition 1. For a random variable X, we define the sub-Gaussian norm $\|X\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\}$ and the sub-exponential norm $\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \le 2\}$. X is sub-Gaussian (resp. sub-exponential) if $\|X\|_{\psi_2} < \infty$ (resp. $\|X\|_{\psi_1} < \infty$). For a random vector $\mathbf{x} \in \mathbb{R}^n$, we let $\|\mathbf{x}\|_{\psi_2} := \sup_{\mathbf{v} \in \mathbb{S}^{n-1}} \|\mathbf{v}^\top \mathbf{x}\|_{\psi_2}$.

Definition 2. Let $\mathcal{S}$ be a subset of $\mathbb{R}^n$. We say that a subset $\mathcal{S}_0 \subset \mathcal{S}$ is an η-net of $\mathcal{S}$ if every point in $\mathcal{S}$ is at most η distance away from some point in $\mathcal{S}_0$, i.e., $\mathcal{S} \subset \mathcal{S}_0 + \mathbb{B}_2^n(\eta)$. Given a radius η, we define the covering number $\mathcal{N}(S;\epsilon)$ as the minimal cardinality of an ϵ-net of S. The metric entropy of S with respect to radius ϵ is defined as $\mathscr{H}(S;\epsilon) = \log \mathcal{N}(S;\epsilon)$.

2.1 Problem Setup

We make the following assumptions on the observation model.

Assumption 1. Let $a \sim \mathcal{N}(0, I_n)$ and let f be a possibly unknown, possibly random non-linearity that is independent of a. Let $(a_i, f_i)_{i=1}^m$ be i.i.d. copies of (a, f). With a single draw of $(a_i, f_i)_{i=1}^m$, for $x^\star \in K = G(B_2^k(r))$, where $G : B_2^k(r) \to \mathbb{R}^n$ is an L-Lipschitz generative model, we observe $\{y_i := f_i(a_i^\top x^\star)\}_{i=1}^m$. We can express the model more compactly as $y = f(Ax^\star)$, where $A = [a_1, \dots, a_m]^\top \in \mathbb{R}^{m\times n}$, $f = (f_1, \dots, f_m)^\top$ and $y = (y_1, \dots, y_m)^\top \in \mathbb{R}^m$.

In this work, we consider the generalized Lasso as the recovery method [16, 33, 43], whose core idea is to ignore the non-linearity and minimize the regular ℓ_2 loss. In addition, we need to specify a constraint that reflects the low-complexity nature of $x^\star$, and specifically, we introduce a problem-dependent scaling factor $T \in \mathbb{R}$ and use the constraint "$x \in TK$". Note that this is necessary even if the problem is linear; for example, with observations $y = 2Ax^\star$, one needs to minimize the ℓ_2 loss over "$x \in 2K$". Also, when the generative prior is given by $Tx^\star \in K = G(B_2^k(r))$, we should simply use "$x \in K$" as constraint; this is technically equivalent to the treatment adopted in [33] (see more discussions in Remark 5 below). Taken collectively, we consider

$$\hat{x} = \arg\min_{x \in TK} \|y - Ax\|_2. \tag{2.1}$$

Importantly, we want to achieve uniform recovery of all $x^\star \in K$ with a single realization of (A, f).

2.2 Assumptions

Let f be the function that characterizes our nonlinear measurements. We introduce several assumptions on f here, and then verify them for specific models in Section 2.3. We define the set of discontinuities as

$$D_f = \{a \in \mathbb{R} : f \text{ is discontinuous at } a\}.$$

We define the notion of jump discontinuity as follows.

Definition 3. (Jump discontinuity). A function $f : \mathbb{R} \to \mathbb{R}$ has a jump discontinuity at x_0 if both $L^- := \lim_{x\to x_0^-} f(x)$ and $L^+ := \lim_{x\to x_0^+} f(x)$ exist but $L^- \neq L^+$. We simply call the oscillation at x_0, i.e., $|L^+ - L^-|$, the jump.

Roughly put, our framework applies to piece-wise Lipschitz continuous f_i with (at most) countably infinite jump discontinuities, which have bounded jumps and are well separated. The precise statement is given below.

Assumption 2. For some (B_0, L_0, β_0), the following statement unconditionally holds true for any realization of f (specifically, $f_1, \dots, f_m$ in our observations):

- D_f is one of the following: $\varnothing$, a finite set, or a countably infinite set;
- All discontinuities of f (if any) are jump discontinuities with the jump bounded by B_0;
- f is L_0-Lipschitz on any interval (a, b) satisfying $(a, b) \cap D_f = \varnothing$.
- $|a - b| \geq \beta_0$ holds for any $a, b \in D_f$, $a \neq b$ (we set $\beta_0 = \infty$ if $|D_f| \leq 1$).

For simplicity, we assume $f(x_0) = \lim_{x\to x_0^+} f(x)$ for $x_0 \in D_f$.[3]

We note that Assumption 2 is satisfied by L-Lipschitz f with $(B_0, L_0, \beta_0) = (0, L, \infty)$, 1-bit quantized observation $f(\cdot) = \mathrm{sign}(\cdot + \tau)$ (τ is the potential dither, similarly below) with $(B_0, L_0, \beta_0) = (2, 0, \infty)$, and uniformly quantized observation $f(\cdot) = \delta\left(\left\lfloor \frac{\cdot + \tau}{\delta} \right\rfloor + \frac{1}{2}\right)$ with $(B_0, L_0, \beta_0) = (\delta, 0, \delta)$.

[3] This is very mild because the observations are $f_i(a_i^\top x)$, while $\mathbb{P}(a^\top x \in D_{f_i}) = 0$ (as D_{f_i} is at most countably infinite and $a \sim \mathcal{N}(0, I_n)$).

Under Asssumption 2, for any $\beta \in [0, \frac{\beta_0}{2})$ we construct $f_{i,\beta}$ as the Lipschitz approximation of f_i to deal with the potential discontinuity of f_i (i.e., $D_{f_i} \neq \varnothing$). Specifically, $f_{i,\beta}$ modifies f_i in $D_{f_i} + [-\frac{\beta}{2}, \frac{\beta}{2}]$ to be piece-wise linear and Lipschitz continuous; see its precise definition in (3.4).

We develop Theorem 2 to bound certain product processes appearing in the analysis, which produces bounds tighter than [36] when the index sets have low metric entropy. To make Theorem 2 applicable, we further make the following Assumption 3, which can be checked case-by-case by estimating the sub-Gaussian norm and probability tail. Also, $U_g^{(1)}$ and $U_g^{(2)}$ can even be a bit crude because the measurement number in Theorem 1 depends on them in a logarithmic manner.

Assumption 3. Let $a \sim \mathcal{N}(0, I_n)$, under Assumptions 1-2, we define the Lipschitz approximation $f_{i,\beta}$ as in (3.4). We let

$$\xi_{i,\beta}(a) := f_{i,\beta}(a) - Ta; \quad \varepsilon_{i,\beta}(a) := f_{i,\beta}(a) - f_i(a). \tag{2.2}$$

For all $\beta \in (0, \frac{\beta_0}{2})$, we assume the following holds with some parameters $(A_g^{(1)}, U_g^{(1)}, P_0^{(1)})$ and $(A_g^{(2)}, U_g^{(2)}, P_0^{(2)})$:

- $\sup_{x \in \mathcal{K}} \|\xi_{i,\beta}(a^\top x)\|_{\psi_2} \le A_g^{(1)}$, $\mathbb{P}\big(\sup_{x \in \mathcal{K}} |\xi_{i,\beta}(a^\top x)| \le U_g^{(1)}\big) \ge 1 - P_0^{(1)}$;
- $\sup_{x \in \mathcal{K}} \|\varepsilon_{i,\beta}(a^\top x)\|_{\psi_2} \le A_g^{(2)}$, $\mathbb{P}\big(\sup_{x \in \mathcal{K}} |\varepsilon_{i,\beta}(a^\top x)| \le U_g^{(2)}\big) \ge 1 - P_0^{(2)}$.

To build a more complete theory we further introduce two useful quantities. For some $x \in \mathcal{K}$, we define the target mismatch $\rho(x)$ as in [17, Definition 1]:

$$\rho(x) = \big\| \mathbb{E}\big[f_i(a_i^\top x)a_i\big] - Tx \big\|_2. \tag{2.3}$$

It is easy to see that $\mathbb{E}\big[f_i(a_i^\top x)a_i\big]$ minimizes the expected ℓ_2 loss $\mathbb{E}\big[\|y - Ax\|_2^2\big]$, thus one can roughly understand $\mathbb{E}\big[f_i(a_i^\top x)a_i\big]$ as the expectation of $\hat{x}$. Since Tx is the desired ground truth, a small $\rho(x)$ is intuitively an important ingredient for generalized Lasso to succeed. Fortunately, in many models, $\rho(x)$ with a suitably chosen T will vanish (e.g., linear model [2], single index model [33], 1-bit model [31]) or at least be sufficiently small (e.g., 1-bit model with dithering [45]).

As mentioned before, our method to deal with discontinuity of f_i is to introduce its approximation $f_{i,\beta}$, which differs from f_i only in $D_{f_i} + [-\frac{\beta}{2}, \frac{\beta}{2}]$. This will produce some bias because the actual observation is $f_i(a_i^\top x)$ rather than $f_{i,\beta}(a_i^\top x)$. Hence, for some $x \in \mathcal{K}$ we define the following quantity to measure the bias induced by $f_{i,\beta}$:

$$\mu_\beta(x) = \mathbb{P}\Big(a^\top x \in D_{f_i} + \Big[-\frac{\beta}{2}, \frac{\beta}{2}\Big]\Big), \quad a \sim \mathcal{N}(0, I_n). \tag{2.4}$$

The following assumption can often be satisfied by choosing suitable T and sufficiently small β_1.

Assumption 4. Suppose Assumptions 1-3 hold true with parameters $B_0, L_0, \beta_0, A_g^{(1)}, A_g^{(2)}$. For the T used in (2.1), $\rho(x)$ defined in (2.3) satisfies

$$\sup_{x \in \mathcal{K}} \rho(x) \lesssim (A_g^{(1)} \vee A_g^{(2)}) \sqrt{\frac{k}{m}}. \tag{2.5}$$

Moreover, there exists some $0 < \beta_1 < \frac{\beta_0}{2}$ such that

$$(L_0\beta_1 + B_0) \sup_{x \in \mathcal{K}} \sqrt{\mu_{\beta_1}(x)} \lesssim (A_g^{(1)} \vee A_g^{(2)}) \sqrt{\frac{k}{m}}. \tag{2.6}$$

In the proof, the estimation error $\|\hat{x} - Tx\|$ is contributed by a concentration term of scaling $\tilde{O}\big((A_g^{(1)} \vee A_g^{(2)})\sqrt{k/m}\big)$ and some bias terms. The main aim of Assumption 4 is to pull down the bias terms so that the concentration term is dominant.

2.3 Main Theorem and its Implications

We now present our general theorem and apply it to some specific models.

Theorem 1. Under Assumptions 1-4, given any recovery accuracy $\epsilon \in (0,1)$, if it holds that $m \gtrsim (A_g^{(1)} \vee A_g^{(2)})^2 \frac{k\mathscr{L}}{\epsilon^2}$, then with probability at least $1 - m(P_0^{(1)} + P_0^{(2)}) - m\exp(-\Omega(n)) - C\exp(-\Omega(k))$ on a single realization of $(\mathbf{A}, \mathbf{f}) := (\mathbf{a}_i, f_i)_{i=1}^m$, we have the uniform signal recovery guarantee $\|\hat{\mathbf{x}} - T\mathbf{x}\|_2 \le \epsilon$ for all $\mathbf{x} \in \mathcal{K}$, where $\hat{\mathbf{x}}$ is the solution to (2.1) with $\mathbf{y} = \mathbf{f}(\mathbf{A}\mathbf{x})$, and $\mathscr{L} = \log \tilde{P}$ is a logarithmic factor with $\tilde{P}$ being polynomial in (L, n) and other parameters that typically scale as $O(L + n)$. See (C.11) for the precise expression of $\mathscr{L}$.

To illustrate the power of Theorem 1, we specialize it to several models to obtain concrete uniform signal recovery results. Starting with Theorem 1, the remaining work is to select parameters that justify Assumptions 2-4. We summarize the strategy as follows: (i) Determine the parameters in Assumption 2 by the measurement model; (ii) Set T that verifies (2.5) (see Lemmas 8-11 for the following models); (iii) Set the parameters in Assumption 3, for which bounding the norm of Gaussian vector is useful; (iv) Set β_1 to guarantee (2.6) based on some standard probability argument. We only provide suitable parameters for the following concrete models due to space limit, while leaving more details to Appendix E.

(A) 1-bit GCS. Assume that we have the 1-bit observations $y_i = \mathrm{sign}(\mathbf{a}_i^\top \mathbf{x})$; then $f_i(\cdot) = f(\cdot) = \mathrm{sign}(\cdot)$ satisfies Assumption 2 with $(B_0, L_0, \beta_0) = (2, 0, 1)$. In this model, it is hopeless to recover the norm of $\|\mathbf{x}\|_2$; as done in previous work, we assume $\mathbf{x} \in \mathcal{K} \subset \mathbb{S}^{n-1}$ [31, Remark 1]. We set $T = \sqrt{2/\pi}$ and take the parameters in Assumption 3 as $A_g^{(1)} \asymp 1, U_g^{(1)} \asymp \sqrt{n}, P_0^{(1)} \asymp \exp(-\Omega(n)); A_g^{(2)} \asymp 1, U_g^{(2)} \asymp 1, P_0^{(2)} = 0$. We take $\delta = \delta_1 \asymp \frac{k}{m}$ to guarantee (2.6). With these choices, Theorem 1 specializes to the following:

Corollary 1. Consider Assumption 1 with $f_i(\cdot) = \mathrm{sign}(\cdot)$ and $\mathcal{K} \subset \mathbb{S}^{n-1}$, let $\epsilon \in (0,1)$ be any given recovery accuracy. If $m \gtrsim \frac{k}{\epsilon^2} \log \left(\frac{Lr\sqrt{mn}}{\epsilon \wedge (k/m)}\right)$,[4] then with probability at least $1 - 2m\exp(-cn) - m\exp(-\Omega(k))$ on a single draw of $(\mathbf{a}_i)_{i=1}^m$, we have the uniform signal recovery guarantee $\left\|\hat{\mathbf{x}} - \sqrt{\frac{2}{\pi}}\mathbf{x}\right\|_2 \le \epsilon$ for all $\mathbf{x} \in \mathcal{K}$, where $\hat{\mathbf{x}}$ is the solution to (2.1) with $\mathbf{y} = \mathrm{sign}(\mathbf{A}\mathbf{x})$ and $T = \sqrt{\frac{2}{\pi}}$.

Remark 1. A uniform recovery guarantee for generalized Lasso in 1-bit GCS was obtained in [33, Section 5]. Their proof relies on the local embedding property in [31]. Note that such geometric property is often problem-dependent and highly nontrivial. By contrast, our argument is free of geometric properties of this kind.

Remark 2. For traditional 1-bit CS, [17, Corollary 2] requires $m \gtrsim \tilde{O}(k/\epsilon^4)$ to achieve uniform ℓ_2-accuracy of ϵ for all k-sparse signals, which is inferior to our $\tilde{O}(k/\epsilon^2)$. This is true for all remaining examples. To obtain such a sharper rate, the key technique is to use our Theorem 2 (rather than [36]) to obtain tighter bound for the product processes, as will be discussed in Remark 8.

(B) 1-bit GCS with dithering. Assume that the $\mathbf{a}_i^\top \mathbf{x}$ is quantized to 1-bit with dither[5] $\tau_i \overset{iid}{\sim} \mathcal{U}[-\lambda, \lambda]$ for some λ to be chosen, i.e., we observe $y_i = \mathrm{sign}(\mathbf{a}_i^\top \mathbf{x} + \tau_i)$. Following [45] we assume $\mathcal{K} \subset \mathbb{B}_2^n(R)$ for some $R > 0$. Here, using dithering allows the recovery of signal norm $\|\mathbf{x}\|_2$, so we do not need to assume $\mathcal{K} \subset \mathbb{S}^{n-1}$ as in Corollary 1. We set $\lambda = CR\sqrt{\log m}$ with sufficiently large C, and $T = \lambda^{-1}$. In Assumption 3, we take $A_g^{(1)} \asymp 1$, $U_g^{(1)} \asymp \sqrt{n}$, $P_0^{(1)} \asymp \exp(-\Omega(n))$; $A_g^{(2)} \asymp 1$, $U_g^{(2)} \asymp 1$, and $P_0^{(2)} = 0$. Moreover, we take $\delta = \delta_1 = \frac{k}{m}$ to guarantee (2.6). Now we can invoke Theorem 1 to get the following.

Corollary 2. Consider Assumption 1 with $f_i(\cdot) = \mathrm{sign}(\cdot + \tau_i)$, $\tau_i \sim \mathcal{U}[-\lambda, \lambda]$ and $\mathcal{K} \subset \mathbb{B}_2^n(R)$, and $\lambda = CR\sqrt{\log m}$ with sufficiently large C. Let $\epsilon \in (0,1)$ be any given recovery accuracy. If $m \gtrsim \frac{k}{\epsilon^2} \log \left(\frac{Lr\sqrt{mn}}{\lambda(\epsilon \wedge (k/m))}\right)$, then with probability at least $1 - 2m\exp(-cn) - m\exp(-\Omega(k))$ on a single draw of $(\mathbf{a}_i, \tau_i)_{i=1}^m$, we have the uniform signal recovery guarantee $\|\hat{\mathbf{x}} - \lambda^{-1}\mathbf{x}\|_2 \le \epsilon$ for all $\mathbf{x} \in \mathcal{K}$, where $\hat{\mathbf{x}}$ is the solution to (2.1) with $\mathbf{y} = \mathrm{sign}(\mathbf{A}\mathbf{x} + \boldsymbol{\tau})$ (here, $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_m]^\top$) and $T = \lambda^{-1}$.

Remark 3. To our knowledge, the only related prior result is in [45, Theorem 3.2]. However, their result is restricted to ReLU networks. By contrast, we deal with the more general Lipschitz generative models; by specializing our result to the ReLU network that is typically $(n^{\Theta(d)})$-Lipschitz [2] (d is

---

[4] Here and in other similar statements, we implicitly assume a large enough implied constant.

[5] Throughout this work, the random dither is independent of the $\{\mathbf{a}_i\}_{i=1}^m$.

the number of layers), our error rate coincides with theirs up to a logarithmic factor. Additionally, as already mentioned in the Introduction Section, our result can be generalized to a sensing vector with an unknown covariance matrix, unlike theirs which is restricted to isotropic sensing vectors. The advantage of their result is in allowing sub-exponential sensing vectors.

(C) Lipschitz-continuous SIM with generative prior. Assume that any realization of f is unconditionally $\hat{L}$-Lipschitz, which implies Assumption 2 with $(B_0; L_0; \ _0) = (0; \hat{L}; 1\ )$. We further assume $P(f(0) \quad \hat{B}) \quad 1 \quad P_0^0$ for some $(\hat{B}; P_0^0)$. Because the norm of x is absorbed into the unknown $f(\ )$, we assume $K \quad S^{n\ 1}$. We set $\ = 0$ so that $f_{i;} \ = f_i$. We introduce the quantities $\ = E[f(g)g]; \ = kf(g)k_{\ 2};$ where $g \quad N(0; 1)$. We choose $T = \ $ and set parameters in Assumption 3 as $A_g^{(1)} \quad + \ ; U_g^{(1)} \quad (\hat{L} + \ )^{p}\overline{n} + \hat{B}; P_0^{(1)} \quad P_0^0 + \exp(\ \ (\ n)); A_g^{(2)} \ + \ ; U_g^{(2)} = 0; P_0^{(2)} = 0$. Now we are ready to apply Theorem 1 to this model. We obtain:

Corollary 3. Consider Assumption 1 with $\hat{L}$-Lipschitz f, suppose that $P(f(0) \quad \hat{B}) \quad 1 \quad P_0^0$, and define the parameters $\ = E[f(g)g], \ = kf(g)k_{\ 2}$ with $g \quad N(0; 1)$. Let $\ 2 (0; 1)$ be any given recovery accuracy. If $m \gtrsim \frac{(\ +\ )k}{\ ^2} \log \frac{Lr^{p}\overline{m}[n(\ +\ )(\hat{L}+\ )+^{p}\overline{n}\ \hat{B}+\ ]}{(\ +\ )}$, then with probability at least $1 \quad 2m\exp(\ cn) \quad mP_0^0 \quad c_1\exp(\ \ (\ k))$ on a single draw of $(a_i; f_i)_{i=1}^m$, we have the uniform signal recovery guarantee $k\hat{x} \quad x\ k_2 \quad$ for all $x \quad 2 K$, where $\hat{x}$ is the solution to (2.1) with $y = f(Ax\ )$ and $T = \ $.

Remark 4. While the main result of [33] is non-uniform, it was noted in [33, Section 5] that a similar uniform error rate can be established for any deterministic 1-Lipschitz f. Our result here is more general in that the $\hat{L}$-Lipschitz f is possibly random. Note that randomness on f is significant because it provides much more flexibility (e.g., additive random noise).

Remark 5. For SIM with unknown f_i it may seem impractical to use (2.1) as it requires $\ = E[f(g)g]$ where $g \quad N(0; 1)$. However, by assuming $x \quad 2 K = G(B_2^k(r))$ as in [33], which is natural for sufficiently expressive $G(\ )$, we can simply use $x \quad 2 K$ as constraint in (2.1). Our Corollary 3 remains valid in this case under some inessential changes of $\log$ factors in the sample complexity.

(D) Uniformly quantized GCS with dithering. The uniform quantizer with resolution $\ > 0$ is defined as $Q\ (a) = \ b\frac{a}{\ }c + \frac{1}{2}$ for $a\ 2 R$. Using dithering $\ _i \quad U[\ \frac{\ }{2}; \frac{\ }{2}]$, we suppose that the observations are $y_i = Q\ (a_i^> x\ + \ _i)$. This satisfies Assumption 2 with $(B_0; L_0; \ _0) = (\ ; 0; \ )$. We set $T = 1$ and take parameters for Assumption 3 as follows: $A_g^{(1)}; U_g^{(1)}; A_g^{(2)}; U_g^{(2)}$, and $P_0^{(1)} = P_0^{(2)} = 0$: We take $\ = \ _1 \ \frac{k}{m}$ to confirm (2.6). With these parameters, we obtain the following from Theorem 1.

Corollary 4. Consider Assumption 1 with $f(\ ) = Q\ (\ +\ ),\ \quad U[\ \frac{\ }{2}; \frac{\ }{2}]$ for some quantization resolution $\ > 0$. Let $\ > 0$ be any given recovery accuracy. If $m \gtrsim \frac{\ ^2 k}{\ ^2} \log \frac{Lr^{p}\overline{mn}}{\ \wedge[k = (m^{p}\overline{n})]}$, then with probability at least $1 \quad 2m\exp(\ cn) \quad c_1\exp(\ \ (\ k))$ on a single draw of $(a_i; \ _i)_{i=1}^m$, we have the uniform recovery guarantee $k\hat{x} \quad x\ k_2 \quad$ for all $x \quad 2 K$, where $\hat{x}$ is the solution to (2.1) with $y = Q\ (Ax\ + \ )$ and $T = 1$ (here, $\ = [\ _1; \ldots; \ _m]^>$).

Remark 6. While this dithered uniform quantized model has been widely studied in traditional CS (e.g., non-uniform recovery [8, 48], uniform recovery [17, 52]), it has not been investigated in GCS even for non-uniform recovery. Thus, this is new to the best of our knowledge.

A simple extension to the noisy model $y = f(Ax\ ) + \ $ where $\ 2 R^m$ has i.i.d. sub-Gaussian entries can be obtained by a fairly straightforward extension of our analysis; see Appendix F.

3 Proof Sketch

To provide a sketch of our proof, we begin with the optimality condition $ky \quad A\hat{x}k_2^2 \quad ky \quad A(Tx\ )k_2^2$. We expand the square and plug in $y = f(Ax\ )$ to obtain

$$\left\| \frac{A}{\sqrt{m}}(\hat{x} \quad Tx\ ) \right\|_2^2 \quad \frac{2}{m}\ f(Ax\ ) \quad TAx\ ; A(\hat{x} \quad Tx\ )\ : \qquad (3.1)$$

For the final goal $k\hat{x} \quad Tx\ k_2 \quad$, up to rescaling, it is enough to prove $k\hat{x} \quad Tx\ k_2 \quad 3\ $. We assume for convenience that $k\hat{x} \quad Tx\ k_2 > 2\ $, without loss of generality. Combined with

$\hat{x}, Tx \in TK$, we know $\hat{x} - Tx \in K^{\epsilon}_-$, where $K^{\epsilon}_- := (TK_-) \cap \left(\mathbb{B}_2^n(2\epsilon)\right)^c$; $K_- = K - K$: We further define

$$(K^{\epsilon}_-)^* := \left\{ z/\|z\|_2 : z \in K^{\epsilon}_- \right\} \tag{3.2}$$

where the normalized error lives, i.e. $\frac{\hat{x} - Tx}{\|\hat{x} - Tx\|_2} \in (K^{\epsilon}_-)^*$. Our strategy is to establish a uniform lower bound (resp., upper bound) for the left-hand side (resp., the right-hand side) of (3.1). We emphasize that these bounds must hold uniformly for all $x \in K$.

It is relatively easy to use set-restricted eigenvalue condition (S-REC) [2] to establish a uniform lower bound for the left-hand side of (3.1), see Appendix B.1 for more details. It is significantly more challenging to derive an upper bound for the right-hand side of (3.1). As the upper bound must hold uniformly for all x, we first take the supremum over x and $\hat{x}$ and consider bounding the following:

$$\begin{aligned} R &:= \frac{1}{m}\left\langle f(Ax) - TAx, A(\hat{x} - Tx)\right\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(f_i(a_i^\top x) - Ta_i^\top x\right)\left(a_i^\top[\hat{x} - Tx]\right) \\ &\le \|\hat{x} - Tx\|_2 \cdot \sup_{x \in K}\sup_{v \in (K^{\epsilon}_-)^*} \frac{1}{m}\sum_{i=1}^{m}\left(f_i(a_i^\top x) - Ta_i^\top x\right)\left(a_i^\top v\right) := \|\hat{x} - Tx\|_2 \cdot R_u; \end{aligned} \tag{3.3}$$

where $(K^{\epsilon}_-)^*$ is defined in (3.2). Clearly, R_u is the supremum of a product process, whose factors are indexed by $x \in K$ and $v \in (K^{\epsilon}_-)^*$. It is, in general, challenging to control a product process, and existing results often require both factors to satisfy a certain "sub-Gaussian increments" condition (e.g., [36, 37]). However, the first factor of R_u (i.e., $f_i(a_i^\top x) - Ta_i^\top x$) does not admit such a condition when f_i is not continuous (e.g., the 1-bit model $f_i = \text{sign}(\cdot)$). We will construct the Lipschitz approximation of f_i to overcome this difficulty shortly in Section 3.1.

Remark 7. We note that these challenges stem from our pursuit of uniform recovery. In fact, a non-uniform guarantee for SIM was presented in [33, Theorem 1]. In its proof, the key ingredient is [33, Lemma 3] that bounds R_u without the supremum on x. This can be done as long as $f_i(a_i^\top x)$ is sub-Gaussian, while the potential discontinuity of f_i is totally unproblematic.

3.1 Lipschitz Approximation

For any $x_0 \in \mathscr{D}_{f_i}$ we define the one-sided limits as $f_i^-(x_0) = \lim_{x \to x_0^-} f_i(x)$ and $f_i^+(x_0) = \lim_{x \to x_0^+} f_i(x)$, and write their average as $f_i^a(x_0) = \frac{1}{2}(f_i^-(x_0) + f_i^+(x_0))$. Given any approximation accuracy $\beta \in (0, \frac{\beta_0}{2})$, we construct the Lipschitz continuous function $f_{i,\beta}$ as:

$$f_{i,\beta}(x) = \begin{cases} f_i(x), & \text{if } x \notin \mathscr{D}_{f_i} + [-\frac{\beta}{2}, \frac{\beta}{2}] \\ f_i^a(x_0) - \frac{2[f_i^a(x_0) - f_i(x_0 - \frac{\beta}{2})](x_0 - x)}{\beta}, & \text{if } \exists x_0 \in \mathscr{D}_{f_i} \text{ s.t. } x \in [x_0 - \frac{\beta}{2}, x_0] \\ f_i^a(x_0) + \frac{2[f_i(x_0 + \frac{\beta}{2}) - f_i^a(x_0)](x - x_0)}{\beta}, & \text{if } \exists x_0 \in \mathscr{D}_{f_i}, \text{ s.t. } x \in [x_0, x_0 + \frac{\beta}{2}] \end{cases} \tag{3.4}$$

We have defined the approximation error $\varepsilon_{i,\beta}(\cdot) = f_{i,\beta}(\cdot) - f_i(\cdot)$ in Assumption 3. An important observation is that both $f_{i,\beta}$ and $|\varepsilon_{i,\beta}|$ are Lipschitz continuous (see Lemma 1 below). Here, it is crucial to consider $|\varepsilon_{i,\beta}|$ rather than $\varepsilon_{i,\beta}$ as the latter is not continuous; see Figure 1 for an intuitive graphical illustration and more explanations in Appendix B.2.

Lemma 1. With B_0, L_0, β_0 given in Assumption 2, for any $\beta \in (0, \frac{\beta_0}{2})$, $f_{i,\beta}$ is $\left(L_0 + \frac{B_0}{\beta}\right)$-Lipschitz over $\mathbb{R}$, and $|\varepsilon_{i,\beta}|$ is $\left(2L_0 + \frac{B_0}{\beta}\right)$-Lipschitz over $\mathbb{R}$.

3.2 Bounding the product process

We now present our technique to bound R_u. Recall that $\xi_{i,\beta}(a)$ and $\varepsilon_{i,\beta}(a)$ were defined in (2.2). By Lemma 1, $\xi_{i,\beta}$ is $\left(L_0 + T + \frac{B_0}{\beta}\right)$-Lipschitz. Now we use $f_i(a) - Ta = \xi_{i,\beta}(a) - \varepsilon_{i,\beta}$ to decompose R_u (in the following, we sometimes shorten "$\sup_{x \in K}\sup_{v \in (K^{\epsilon}_-)^*}$" as "$\sup_{x,v}$"):

$$R_u \le \underbrace{\sup_{x,v}\frac{1}{m}\sum_{i=1}^{m}\xi_{i,\beta}(a_i^\top x)\left(a_i^\top v\right)}_{R_{u1}} + \underbrace{\sup_{x,v}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_{i,\beta}(a_i^\top x)\left(a_i^\top v\right)}_{R_{u2}}. \tag{3.5}$$

Figure 1: (Left): f_i and its approximation $f_{i,0.5}$; (Right): approximation error $\varepsilon_{i,0.5}$, $|\varepsilon_{i,0.5}|$.

It remains to control R_{u1} and R_{u2}. By the Lipschitz continuity of $f_{i,\beta}$ and $|\varepsilon_{i,\beta}|$, the factors of R_{u1} and R_{u2} admit sub-Gaussian increments, so it is natural to first center them and then invoke the concentration inequality for product process due to Mendelson [36, Theorem 1.13], which we restate in Lemma 5 (Appendix A). However, this does not produce a tight bound and would eventually require $\tilde{O}(k/\epsilon^4)$ to achieve a uniform ℓ_2-error of ϵ, as is the case in [17, Section 4].

In fact, Lemma 5 is based on Gaussian width and hence blind to the fact that $\mathcal{K}, (\mathcal{K}^-_\epsilon)^*$ here have low metric entropy (Lemma 6). By characterizing the low intrinsic dimension of index sets via metric entropy, we develop the following concentration inequality that can produce tighter bound for R_{u1} and R_{u2}. This also allows us to derive uniform error rates sharper than those in [17, Section 4].

Theorem 2. Let $g_x = g_x(a)$ and $h_v = h_v(a)$ be stochastic processes indexed by $x \in \mathcal{X} \subset \mathbb{R}^{p_1}, v \in \mathcal{V} \subset \mathbb{R}^{p_2}$, both defined with respect to a common random variable a. Assume that:

- (A1.) $g_x(a), h_v(a)$ are sub-Gaussian for some (A_g, A_h) and admit sub-Gaussian increments regarding ℓ_2 distance for some (M_g, M_h):

$$\begin{aligned} \|g_x(a) - g_{x'}(a)\|_{\psi_2} &\le M_g\|x - x'\|_2,\ \|g_x(a)\|_{\psi_2} \le A_g,\ \forall\, x, x' \in \mathcal{X}; \\ \|h_v(a) - h_{v'}(a)\|_{\psi_2} &\le M_h\|v - v'\|_2,\ \|h_v(a)\|_{\psi_2} \le A_h,\ \forall\, v, v' \in \mathcal{V}. \end{aligned} \tag{3.6}$$

- (A2.) On a single draw of a, for some (L_g, U_g, L_h, U_h) the following events simultaneously hold with probability at least $1 - P_0$:

$$\begin{aligned} |g_x(a) - g_{x'}(a)| &\le L_g\|x - x'\|_2,\ |g_x(a)| \le U_g,\ \forall\, x, x' \in \mathcal{X}; \\ |h_v(a) - h_{v'}(a)| &\le L_h\|v - v'\|_2,\ |h_v(a)| \le U_h,\ \forall\, v, v' \in \mathcal{V}. \end{aligned} \tag{3.7}$$

Let $a_1, \dots, a_m$ be i.i.d. copies of a, and introduce the shorthand $S_{g,h} = L_gU_h + M_gA_h$ and $T_{g,h} = L_hU_g + M_hA_g$. If $m \gtrsim \mathscr{H}\big(\mathcal{X}, \frac{A_gA_h}{\sqrt{m}S_{g,h}}\big) + \mathscr{H}\big(\mathcal{V}, \frac{A_gA_h}{\sqrt{m}T_{g,h}}\big)$, where $\mathscr{H}(\cdot,\cdot)$ is the metric entropy defined in Definition 2, then with probability at least $1 - mP_0 - 2\exp\big[-\Omega\big(\mathscr{H}(\mathcal{X}, \frac{A_gA_h}{\sqrt{m}S_{g,h}}) + \mathscr{H}(\mathcal{V}, \frac{A_gA_h}{\sqrt{m}T_{g,h}})\big)\big]$ we have $I \lesssim \frac{A_gA_h}{\sqrt{m}}\sqrt{\mathscr{H}(\mathcal{X}, \frac{A_gA_h}{\sqrt{m}S_{g,h}}) + \mathscr{H}(\mathcal{V}, \frac{A_gA_h}{\sqrt{m}T_{g,h}})}$, where $I := \sup_{x\in\mathcal{X}}\sup_{v\in\mathcal{V}}\Big|\frac{1}{m}\sum_{i=1}^{m}\big(g_x(a_i)h_v(a_i) - \mathbb{E}[g_x(a_i)h_v(a_i)]\big)\Big|$ is the supremum of a product process.

Remark 8. We use R_{u2} as an example to illustrate the advantage of Theorem 2 over Lemma 5. The key step is on bounding the centered process

$$R_{u2,c} := \sup_{x\in\mathcal{K}}\ \sup_{v\in(\mathcal{K}^-_\epsilon)^*}\Big(|\varepsilon_{i,\beta}(a_i^\top x)||a_i^\top v| - \mathbb{E}[|\varepsilon_{i,\beta}(a_i^\top x)||a_i^\top v|]\Big).$$

Let $g_x(a_i) = |\varepsilon_{i,\beta}(a_i^\top x)|$ and $h_v(a_i) = |a_i^\top v|$, then one can use Theorem 2 or Lemma 5 to bound $R_{u2,c}$. Note that $\|a_i^\top v\|_{\psi_2} = O(1)$ justifies the choice $A_h = O(1)$, and both $\mathscr{H}(\mathcal{K}, \cdot)$ and $\mathscr{H}((\mathcal{K}^-_\epsilon)^*, \cdot)$ depend linearly on k but only logarithmically on $\cdot$ (Lemma 6), so Theorem 2 could bound $R_{u2,c}$ by $\tilde{O}\big(A_g\sqrt{k/m}\big)$ that depends on M_g in a logarithmic manner. However, the bound produced by Lemma 5 depends linearly on M_g; see term $\frac{M_gA_h\omega(\mathcal{K})}{\sqrt{m}}$ in (A.1). From (3.6), M_g should be proportional to the Lipschitz constant of $|\varepsilon_{i,\beta}|$, which scales as $\frac{1}{\beta}$ (Lemma 1). The issue is that in many cases we need to take extremely small β to guarantee that (2.6) holds true (e.g., we take $\beta \asymp k/m$ in 1-bit GCS). Thus, Lemma 5 produces a worse bound compared to our Theorem 2.

4 Conclusion

In this work, we built a unified framework for uniform signal recovery in nonlinear generative compressed sensing. We showed that using generalized Lasso, a sample size of $\tilde{O}(k/\epsilon^2)$ suffices to uniformly recover all $x \in G(\mathbb{B}_2^k(r))$ up to an ℓ_2-error of ϵ. We specialized our main theorem to 1-bit GCS with/without dithering, single index model, and uniformly quantized GCS, deriving uniform guarantees that are new or exhibit some advantages over existing ones. Unlike [33], our proof is free of any non-trivial embedding property. As part of our technical contributions, we constructed the Lipschitz approximation to handle potential discontinuity in the observation model, and also developed a concentration inequality to derive tighter bound for the product processes arising in the proof, allowing us to obtain a uniform error rate faster than [17]. Possible future directions include extending our framework to handle the adversarial noise and representation error.

Acknowledgment. J. Chen was supported by a Hong Kong PhD Fellowship from the Hong Kong Research Grants Council (RGC). J. Scarlett was supported by the Singapore National Research Foundation (NRF) under grant A-0008064-00-00. M. K. Ng was partially supported by the HKRGC GRF 17201020, 17300021, CRF C7004-21GF and Joint NSFC-RGC N-HKU76921.

References

[1] M. Asim, M. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: Mitigating representation error and dataset bias," in *International Conference on Machine Learning*. PMLR, 2020, pp. 399–409.

[2] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International Conference on Machine Learning*. PMLR, 2017, pp. 537–546.

[3] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Annual Conference on Information Sciences and Systems*. IEEE, 2008, pp. 16–21.

[4] T. T. Cai and A. Zhang, "Sparse representation of a polytope and recovery of sparse signals and low-rank matrices," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 122–132, 2013.

[5] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[6] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[7] J. Chen and M. K. Ng, "Uniform exact reconstruction of sparse signals and low-rank matrices from phase-only measurements," *IEEE Transactions on Information Theory*, vol. 69, no. 10, pp. 6739–6764, 2023.

[8] J. Chen, M. K. Ng, and D. Wang, "Quantizing heavy-tailed data in statistical estimation: (Near) Minimax rates, covariate quantization, and uniform recovery," *arXiv preprint arXiv:2212.14562*, 2022.

[9] J. Chen, C.-L. Wang, M. K. Ng, and D. Wang, "High dimensional statistical estimation under uniformly dithered one-bit quantization," *IEEE Transactions on Information Theory*, vol. 69, no. 8, pp. 5151–5187, 2023.

[10] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in *International Conference on Learning Representations*, 2022.

[11] G. Daras, J. Dean, A. Jalal, and A. Dimakis, "Intermediate layer optimization for inverse problems using deep generative models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2421–2432.

[12] M. Dhar, A. Grover, and S. Ermon, "Modeling sparse deviations for compressed sensing using generative models," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1214–1223.

[13] S. Dirksen, “Quantized compressed sensing: a survey,” in *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*. Springer, 2019, pp. 67–95.

[14] S. Dirksen and S. Mendelson, “Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing,” *Journal of the European Mathematical Society*, vol. 23, no. 9, pp. 2913–2947, 2021.

[15] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[16] M. Genzel, “High-dimensional estimation of structured signals from non-linear observations with general convex loss functions,” *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1601–1619, 2016.

[17] M. Genzel and A. Stollenwerk, “A unified approach to uniform signal recovery from nonlinear observations,” *Foundations of Computational Mathematics*, pp. 1–74, 2022.

[18] R. M. Gray and T. G. Stockham, “Dithered quantizers,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[19] P. Hand, O. Leong, and V. Voroninski, “Phase retrieval under a generative prior,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[20] P. Hand and V. Voroninski, “Global guarantees for enforcing deep generative priors by empirical risk,” in *Conference On Learning Theory*. PMLR, 2018, pp. 970–978.

[21] L. Jacques and T. Feuillen, “The importance of phase in complex compressive sensing,” *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 4150–4161, 2021.

[22] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, “Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors,” *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[23] A. Jalal, S. Karmalkar, A. Dimakis, and E. Price, “Instance-optimal compressed sensing via posterior sampling,” in *International Conference on Machine Learning (ICML)*, 2021.

[24] A. Jalal, M. Arvinte, G. Daras, E. Price, A. G. Dimakis, and J. Tamir, “Robust compressed sensing MRI with deep generative priors,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 938–14 954, 2021.

[25] A. Jalal, L. Liu, A. G. Dimakis, and C. Caramanis, “Robust compressed sensing using generative models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 713–727, 2020.

[26] Y. Jiao, D. Li, M. Liu, X. Lu, and Y. Yang, “Just least squares: Binary compressive sampling with low generative intrinsic dimension,” *Journal of Scientific Computing*, vol. 95, no. 1, p. 28, 2023.

[27] A. Kamath, E. Price, and S. Karmalkar, “On the power of compressed sensing with generative models,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 5101–5109.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] J. Liu and Z. Liu, “Non-iterative recovery from nonlinear observations using generative models,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 233–243.

[30] Z. Liu, S. Ghosh, and J. Scarlett, “Towards sample-optimal compressive phase retrieval with sparse and generative priors,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 656–17 668, 2021.

[31] Z. Liu, S. Gomes, A. Tiwari, and J. Scarlett, “Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 6216–6225.

[32] Z. Liu and J. Han, “Projected gradient descent algorithms for solving nonlinear inverse problems with generative priors,” *arXiv preprint arXiv:2209.10093*, 2022.

[33] Z. Liu and J. Scarlett, “The generalized Lasso with nonlinear observations and generative priors,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 125–19 136, 2020.

[34] Z. Liu and J. Scarlett, “Information-theoretic lower bounds for compressive sensing with generative models,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 292–303, 2020.

[35] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

[36] S. Mendelson, “Upper bounds on product and multiplier empirical processes,” Stochastic Processes and their Applications, vol. 126, no. 12, pp. 3652–3680, 2016.

[37] S. Mendelson, “On multiplier processes under weak moment assumptions,” in Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016. Springer, 2017, pp. 301–318.

[38] X. Meng and Y. Kabashima, “Quantized compressed sensing with score-based generative models,” in International Conference on Learning Representations, 2023.

[39] A. Naderi and Y. Plan, “Sparsity-free compressed sensing with applications to generative priors,” IEEE Journal on Selected Areas in Information Theory, 2022.

[40] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, “Deep learning techniques for inverse problems in imaging,” IEEE Journal on Selected Areas in Information Theory, vol. 1, no. 1, pp. 39–56, 2020.

[41] Y. Plan and R. Vershynin, “Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach,” IEEE Transactions on Information Theory, vol. 59, no. 1, pp. 482–494, 2012.

[42] Y. Plan and R. Vershynin, “One-bit compressed sensing by linear programming,” Communications on Pure and Applied Mathematics, vol. 66, no. 8, pp. 1275–1297, 2013.

[43] Y. Plan and R. Vershynin, “The generalized lasso with non-linear observations,” IEEE Transactions on information theory, vol. 62, no. 3, pp. 1528–1537, 2016.

[44] Y. Plan, R. Vershynin, and E. Yudovina, “High-dimensional estimation with geometric constraints,” Information and Inference: A Journal of the IMA, vol. 6, no. 1, pp. 1–40, 2017.

[45] S. Qiu, X. Wei, and Z. Yang, “Robust one-bit recovery via ReLU generative networks: Near-optimal statistical rate and global landscape analysis,” in International Conference on Machine Learning. PMLR, 2020, pp. 7857–7866.

[46] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, “Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss,” IEEE Transactions on Medical Imaging, vol. 37, no. 6, pp. 1488–1497, 2018.

[47] J. Scarlett, R. Heckel, M. R. Rodrigues, P. Hand, and Y. C. Eldar, “Theoretical perspectives on deep learning methods in inverse problems,” IEEE Journal on Selected Areas in Information Theory, 2022.

[48] C. Thrampoulidis and A. S. Rawat, “The generalized lasso for sub-gaussian measurements with dithered quantization,” IEEE Transactions on Information Theory, vol. 66, no. 4, pp. 2487–2500, 2020.

[49] Y. Traonmilin and R. Gribonval, “Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all,” Applied and Computational Harmonic Analysis, vol. 45, no. 1, pp. 170–205, 2018.

[50] R. Vershynin, High-dimensional probability: An introduction with applications in data science. Cambridge University Press, 2018, vol. 47.

[51] J. Whang, E. Lindgren, and A. Dimakis, “Composing normalizing ows for inverse problems,” in International Conference on Machine Learning, 2021, pp. 11 158–11 169.

[52] C. Xu and L. Jacques, “Quantized compressive sensing with RIP matrices: The bene t of dithering,” Information and Inference: A Journal of the IMA, vol. 9, no. 3, pp. 543–586, 2020.

Supplementary Material

A Unified Framework for Uniform Signal Recovery in Nonlinear Generative Compressed Sensing

(NeurIPS 2023)

A Technical Lemmas

Lemma 2. (Lemma 2.7.7, [50]) Let X, Y be sub-Gaussian, then XY is sub-exponential with $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$.

Lemma 3. (Centering, [50, Exercise 2.7.10]) For some absolute constant C, $\|X - \mathbb{E}X\|_{\psi_1} \le C\|X\|_{\psi_1}$.

Lemma 4. (Bernstein's inequality, [50, Theorem 2.8.1]) Let $X_1, \ldots, X_N$ be independent, zero-mean, sub-exponential random variables. Then for every $t \ge 0$, for some absolute constant c we have

$$\mathbb{P}\left(\left|\sum_{i=1}^N X_i\right| \ge t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{1\le i\le N}\|X_i\|_{\psi_1}}\right\}\right).$$

Lemma 5. ([36], statement adapted from [16, Theorem 8]) Let $g_x = g_x(a)$ and $h_v = h_v(a)$ be stochastic processes indexed by $x \in \mathcal{X} \subset \mathbb{R}^{p_1}$, $v \in \mathcal{V} \subset \mathbb{R}^{p_2}$, both defined on some common random variable a. Assume that (A1.) in Theorem 2 holds, and let $a_1, \ldots, a_m$ be i.i.d. copies of a. Then for any $u \ge 1$, with probability at least $1 - 2\exp(-cu^2)$ we have the bound

$$\begin{aligned}\sup_{\substack{x\in\mathcal{X}\\ v\in\mathcal{V}}}\left|\frac{1}{m}\sum_{i=1}^m g_x(a_i)h_v(a_i) - \mathbb{E}[g_x(a_i)h_v(a_i)]\right| \le\ & C\left(\frac{(M_g\cdot\omega(\mathcal{X}) + u\cdot A_g)\cdot(M_h\cdot\omega(\mathcal{V}) + u\cdot A_h)}{m}\right.\\ &\left. + \frac{A_g\cdot M_h\cdot\omega(\mathcal{V}) + A_h\cdot M_g\cdot\omega(\mathcal{X}) + u\cdot A_gA_h}{\sqrt{m}}\right),\end{aligned} \tag{A.1}$$

where $\omega(\cdot)$ is the Gaussian width defined as $\omega(\mathcal{X}) = \mathbb{E}\sup_{x\in\mathcal{X}} g^\top x$ where $g \sim \mathcal{N}(0, I_{p_1})$.

The proofs of the remaining lemmas will be provided in Appendix D. (Some simple facts such as Lemma 8 were already used in prior works; while we provide the proofs for completeness.)

Lemma 6. (Metric entropy of some constraint sets) Assume $\mathcal{K} = G(\mathbb{B}_2^k(r))$ for some L-Lipschitz generative model G. Let $\mathcal{K}^- = \mathcal{K} - \mathcal{K}$, for some $T > 0$, $\epsilon \in (0,1)$ let $\mathcal{K}^-_\epsilon := (T\mathcal{K}^-) \cap (\mathbb{B}_2^n(2\epsilon))^c$, and further define $(\mathcal{K}^-_\epsilon)^* = \{\frac{z}{\|z\|_2} : z \in \mathcal{K}^-_\epsilon\}$. Then for any $\eta \in (0, Lr)$, we have

$$\mathscr{H}(\mathcal{K},\eta) \le k\log\frac{3Lr}{\eta};\quad \mathscr{H}(\mathcal{K}^-,\eta) \le 2k\log\frac{6Lr}{\eta};$$

$$\mathscr{H}(\mathcal{K}^-_\epsilon,\eta) \le 2k\log\frac{12TLr}{\eta};\quad \mathscr{H}((\mathcal{K}^-_\epsilon)^*,\eta) \le 2k\log\frac{12TLr}{\epsilon\eta};$$

where $\mathscr{H}(\cdot,\cdot)$ is the metric entropy defined in Definition 2.

Lemma 7. (Bound the ℓ_2-norm of Gaussian vector) If $a \sim \mathcal{N}(0, I_n)$, then $\mathbb{P}(|\|a\|_2 - \sqrt{n}| \ge t) \le 2\exp(-Ct^2)$. In particular, setting $t \asymp \sqrt{n}$ yields $\mathbb{P}(\|a\|_2 \gtrsim \sqrt{n}) \le 2\exp(-\Omega(n))$.

In the following, Lemmas 8-11 indicate suitable choices of T in the concrete models we consider. These choices can make $\rho(x)$ in (2.3) sufficiently small or even zero.

Lemma 8. (Choice of T in 1-bit GCS) If $a \sim \mathcal{N}(0, I_n)$, then for any $x \in \mathbb{S}^{n-1}$ it holds that $\mathbb{E}[\mathrm{sign}(a^\top x)a] = \sqrt{\frac{2}{\pi}}x$.

Lemma 9. (Choice of T in 1-bit GCS with dithering) If $a \sim \mathcal{N}(0, I_n)$ and $\tau \sim \mathscr{U}[-\lambda, \lambda]$ are independent, and $\lambda = CR\sqrt{\log m}$ with sufficiently large C, then for any $x \in \mathbb{B}_2^n(R)$ it holds that $\|\mathbb{E}[\mathrm{sign}(a^\top x + \tau)a] - \frac{x}{\lambda}\|_2 = O(m^{-9})$.

Lemma 10. (Choice of T in SIM). If $a \sim \mathcal{N}(0, I_n)$, for some function f and any $x \in S^{n-1}$ it holds that $\mathbb{E}[f(a^\top x)a] = \mu x$ for $\mu = \mathbb{E}[f(g)g]$ with $g \sim \mathcal{N}(0,1)$.

Lemma 11. (Choice of T in uniformly quantized GCS with dithering) Given any $\delta > 0$, let $\tau \sim \mathcal{U}[-\frac{\delta}{2}, \frac{\delta}{2}]$ and $\mathcal{Q}_\delta(\cdot) = \delta\lfloor\frac{\cdot}{\delta}\rfloor + \frac{\delta}{2}$. Then, for any $a \in \mathbb{R}$, it holds that $\mathbb{E}[\mathcal{Q}_\delta(a+\tau)] = a$. In particular, let $a \in \mathbb{R}^n$ be a random vector satisfying $\mathbb{E}(aa^\top) = I_n$, and $\tau \sim \mathcal{U}[-\frac{\delta}{2}, \frac{\delta}{2}]$ be independent of a, then we have $\mathbb{E}[\mathcal{Q}_\delta(a^\top x + \tau)a] = x$.

Lemma 12 facilitates our analysis of the uniform quantizer.

Lemma 12. Let $f_i(\cdot) = \delta\lfloor\frac{\cdot+\tau_i}{\delta}\rfloor + \frac{\delta}{2}$ for $\tau_i \sim \mathcal{U}[-\frac{\delta}{2}, \frac{\delta}{2}]$, and $f_{i,\beta}(\cdot)$ be defined in (3.4) for some $0 < \beta < \frac{\delta}{2}$. Moreover, let $\xi_{i,\beta}(a) = f_{i,\beta}(a) - \mu a$, $\varepsilon_{i,\beta}(a) = f_{i,\beta}(a) - f_i(a)$, then for any $a \in \mathbb{R}$, $|\xi_{i,\beta}(a)| \le 2\delta$, $|\varepsilon_{i,\beta}(a)| \le \delta$ holds deterministically.

More generally, the approximation error $|\varepsilon_{i,\beta}(a)|$ can always be bounded as follows.

Lemma 13. Suppose that f_i satisfies Assumption 2, and for any $\beta \in [0, \frac{\beta_0}{2}]$ we construct $f_{i,\beta}$ as in (3.4). Then, for any $a \in \mathbb{R}$, we have $|\varepsilon_{i,\beta}(a)| \le \left(\frac{3L_0\beta}{2} + B_0\right)\mathbb{1}(a \in \mathcal{D}_{f_i} + [-\frac{\beta}{2}, \frac{\beta}{2}])$.

B More Details of the Proof Sketch

B.1 Set-Restricted Eigenvalue Condition

Definition 4. Let $S \subset \mathbb{R}^n$. For parameters $\gamma, \delta > 0$, a matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy S-REC$(S; \gamma, \delta)$ if the following holds:

$$\|A(x_1 - x_2)\|_2 \ge \gamma\|x_1 - x_2\|_2 - \delta, \ \forall x_1, x_2 \in S.$$

It was proved in [2] that $\frac{1}{\sqrt{m}}A$ satisfies the S-REC with high probability if the entries of A are i.i.d. standard Gaussian.

Lemma 14. (Lemma 4.1 in [2]) Let $G : \mathbb{B}_2^k(r) \to \mathbb{R}^n$ be L-Lipschitz for some $r, L > 0$, and define $\mathcal{K} = G(\mathbb{B}_2^k(r))$. For any $\alpha \in (0,1)$, if $A \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and $m = \Omega\left(\frac{k}{\alpha^2}\log\frac{Lr}{\delta}\right)$, then $\frac{1}{\sqrt{m}}A$ satisfies S-REC$(\mathcal{K}, 1-\alpha, \delta)$ with probability at least $1 - \exp(-\Omega(\alpha^2 m))$.

B.2 Lipschitz Approximation

The approximation error $\varepsilon_{i,\beta}(\cdot)$ can be expanded as:

$$\varepsilon_{i,\beta}(x) = \begin{cases} 0, & \text{if } x \notin \mathcal{D}_{f_i} + [-\frac{\beta}{2}, \frac{\beta}{2}] \\ f_i^a(x_0) - f_i(x) - \frac{2[f_i^a(x_0) - f_i(x_0 - \frac{\beta}{2})](x_0 - x)}{\beta}, & \text{if } x \in [x_0 - \frac{\beta}{2}, x_0],\ (\exists x_0 \in \mathcal{D}_{f_i}) \\ f_i^a(x_0) - f_i(x) + \frac{2[f_i(x_0 + \frac{\beta}{2}) - f_i^a(x_0)](x - x_0)}{\beta}, & \text{if } x \in [x_0, x_0 + \frac{\beta}{2}],\ (\exists x_0 \in \mathcal{D}_{f_i}) \end{cases}.$$

Although $|\varepsilon_{i,\beta}|$ is Lipschitz continuous, $\varepsilon_{i,\beta}$ is not. In particular, given $x_0 \in \mathcal{D}_{f_i}$ we note that

$$\varepsilon_{i,\beta}^-(x_0) = \lim_{x \to x_0^-}\varepsilon_{i,\beta}(x) = f_i^a(x_0) - f_i^-(x_0) = \frac{1}{2}\left(f_i^+(x_0) - f_i^-(x_0)\right);$$

$$\varepsilon_{i,\beta}^+(x_0) = \lim_{x \to x_0^+}\varepsilon_{i,\beta}(x) = f_i^a(x_0) - f_i^+(x_0) = \frac{1}{2}\left(f_i^-(x_0) - f_i^+(x_0)\right).$$

Thus, it is crucial to include the absolute value for rendering the continuity.

C Proofs of Main Results

C.1 Proof of Theorem 1

Proof. Up to rescaling, we only need to prove that $\|\hat{x} - Tx\|_2 \le 3\epsilon$ holds uniformly for all $x \in \mathcal{K}$. We can assume $\|\hat{x} - Tx\|_2 \ge 2\epsilon$; otherwise, the desired bound is immediate.

(1) Lower bounding the left-hand side of (3.1).

We use S-REC to find a lower bound for $\left\|\frac{A}{\sqrt{m}}(\hat{x} - \mu T x^\star)\right\|_2^2$. Specifically, we invoke Lemma 14 with $\alpha = \frac{1}{2}$ and $\delta = \frac{\epsilon}{2T}$, which gives that under $m = \Omega\left(k \log \frac{LTr}{\epsilon}\right)$, with probability at least $1 - \exp(-cm)$, the following holds:

$$\left\|\frac{A}{\sqrt{m}}(x_1 - x_2)\right\|_2 \geq \frac{1}{2}\|x_1 - x_2\|_2 - \frac{\epsilon}{2T};\ \forall\, x_1; x_2 \in K: \tag{C.1}$$

Recall that we assume $\|\hat{x} - \mu T x^\star\|_2 \geq 2\epsilon$, $T^{-1}\hat{x} \in K$, and $x^\star \in K$, so we set $x_1 = \frac{\hat{x}}{T}; x_2 = \mu x^\star$ in (C.1) to obtain

$$\left\|\frac{A}{\sqrt{m}}\left(\frac{\hat{x}}{T} - \mu x^\star\right)\right\|_2 \geq \frac{1}{2}\left\|\frac{\hat{x}}{T} - \mu x^\star\right\|_2 - \frac{\epsilon}{2T} \geq \frac{1}{4T}\left\|\hat{x} - \mu T x^\star\right\|_2:$$

Thus, the left-hand side of (3.1) can be lower bounded by $\Omega\left(\|\hat{x} - \mu T x^\star\|_2^2\right)$:

(2) Upper bounding the right-hand side of (3.1).

As analysed in (3.3) and (3.5), the right-hand side of (3.1) is bounded by $2\|\hat{x} - \mu T x^\star\|_2 \cdot \left(R_{u1} + R_{u2}\right)$, so all that remains is to bound $R_{u1}; R_{u2}$. In the rest of the proof, we simply write $\sup_{x;v} :=$ $\sup_{x \in K; v \in (K_{-\epsilon})}$ and recall the shorthand $\tilde{f}_{i;\tau}(a) = f_{i;\tau}(a) - Ta$. Thus, the first factor of R_{u1} is given by $\tilde{f}_{i;\tau}(a_i^\top x)$. By centering, we have

$$R_{u1} \leq \underbrace{\sup_{x;v} \frac{1}{m}\sum_{i=1}^m \left\{[\tilde{f}_{i;\tau}(a_i^\top x)](a_i^\top v) - \mathbb{E}\,[\tilde{f}_{i;\tau}(a_i^\top x)](a_i^\top v)\right\}}_{R_{u1;c}} + \underbrace{\sup_{x;v} \mathbb{E}\,[\tilde{f}_{i;\tau}(a_i^\top x)](a_i^\top v)}_{R_{u1;e}}; \tag{C.2}$$

and

$$R_{u2} \leq \underbrace{\sup_{x;v} \left\{\frac{1}{m}\sum_{i=1}^m \xi_{i;\tau}(a_i^\top x)\,a_i^\top v - \mathbb{E}\,\xi_{i;\tau}(a_i^\top x)\,a_i^\top v\right\}}_{R_{u2;c}} + \underbrace{\sup_{x;v} \mathbb{E}\,\xi_{i;\tau}(a_i^\top x)\,a_i^\top v}_{R_{u2;e}}: \tag{C.3}$$

We will invoke Theorem 2 multiple times to derive the required bounds.

(2.1) Bounding the centered product process $R_{u1;c}$.

We let $g_x(a_i) = \tilde{f}_{i;\tau}(a_i^\top x)$ and $h_v(a_i) = a_i^\top v$, and write

$$R_{u1;c} = \sup_{x;v} \frac{1}{m}\sum_{i=1}^m \left\{g_x(a_i)h_v(a_i) - \mathbb{E}[g_x(a_i)h_v(a_i)]\right\}:$$

We verify conditions in Theorem 2 as follows:

- For any $x; x' \in K$, because $\tilde{f}_{i;\tau}$ is $\left(L_0 + \frac{B_0}{\tau} + T\right)$-Lipschitz continuous (Lemma 1), we have

$$\begin{aligned}\|g_x(a_i) - g_{x'}(a_i)\|_{\psi_2} &\leq (L_0 + T + \frac{B_0}{\tau})\|a_i^\top x - a_i^\top x'\|_{\psi_2} \\ &= O\left(L_0 + T + \frac{B_0}{\tau}\right)\|x - x'\|_2:\end{aligned}$$

- Since $a_i \sim \mathcal{N}(0; I_n)$, by Lemma 7, with probability $1 - 2\exp(-\Omega(n))$ we have $\|a_i\|_2 = O(\sqrt{n})$. On this event, we have

$$\begin{aligned}|g_x(a_i) - g_{x'}(a_i)| &\leq (L_0 + T + \frac{B_0}{\tau})|a_i^\top x - a_i^\top x'| \\ &\leq (L_0 + T + \frac{B_0}{\tau})\|a_i\|_2\|x - x'\|_2 \\ &= O\left(\sqrt{n}\left[L_0 + T + \frac{B_0}{\tau}\right]\right)\|x - x'\|_2:\end{aligned}$$

- Recall $\mathcal{N}(\mathcal{K})$ in (3.2). Since $\mathcal{N}(\mathcal{K}) \subseteq \mathbb{S}^{n-1}$, for any $v, v' \in \mathcal{N}(\mathcal{K})$, we have $\|a_i^\top v - a_i^\top v'\|_{\psi_2} = O(1)\|v - v'\|_2$, and when $\|a_i\| = O(\sqrt{n})$ we have $|a_i^\top v - a_i^\top v'| \le \|a_i\|_2\|v - v'\|_2 = O(\sqrt{n})\|v - v'\|_2$. Moreover, because $\mathcal{N}(\mathcal{K}) \subseteq \mathbb{B}_2^n$, for any $v \in \mathcal{N}(\mathcal{K})$ we have $\|a_i^\top v\|_{\psi_2} = O(1)$, $|a_i^\top v| \le \|a_i\|_2 = O(\sqrt{n})$.

Combined with Assumption 3 and its parameters $(A_g^{(1)}, U_g^{(1)}, P_0^{(1)})$ and $(A_g^{(2)}, U_g^{(2)}, P_0^{(2)})$, $R_{u1;c}$ satisfies the conditions of Theorem 2 with the following parameters

$$M_g \asymp L_0 + T + \frac{B_0}{\zeta};\ A_g = A_g^{(1)};\ M_h \asymp 1;\ A_h \asymp 1$$

$$L_g \asymp \sqrt{n}\left(L_0 + T + \frac{B_0}{\zeta}\right);\ U_g = U_g^{(1)};\ L_h \asymp \sqrt{n};\ U_h \asymp \sqrt{n}$$

and $P_0 = P_0^{(1)} + 2\exp(-\Omega(n))$. Now suppose that we have

$$m \gtrsim \mathscr{H}\left(\mathcal{K}, \frac{A_g^{(1)}}{\sqrt{m}n[L_0 + T + \frac{B_0}{\zeta}]}\right) + \mathscr{H}\left(\mathcal{N}(\mathcal{K}), \frac{A_g^{(1)}}{\sqrt{m}(\sqrt{n}U_g^{(1)} + A_g^{(1)})}\right); \tag{C.4}$$

and note that by using Lemma 6, (C.4) can be guaranteed by

$$m \gtrsim k\log\left(\frac{Lr\sqrt{m}}{A_g^{(1)}}\left[n\left(L_0 + T + \frac{B_0}{\zeta}\right) + \frac{T(\sqrt{n}U_g^{(1)} + A_g^{(1)})}{\zeta}\right]\right). \tag{C.5}$$

Then Theorem 2 yields that the following bound holds with probability at least $1 - mP_0^{(1)} - C\exp(-\Omega(k)) - m\exp(-\Omega(n))$:

$$\begin{aligned} |R_{u1;c}| &\lesssim \frac{A_g^{1}}{\sqrt{m}}\sqrt{\mathscr{H}\left(\mathcal{K}, \frac{A_g^{(1)}}{\sqrt{m}n[L_0 + T + \frac{B_0}{\zeta}]}\right) + \mathscr{H}\left(\mathcal{N}(\mathcal{K}), \frac{A_g^{(1)}}{\sqrt{m}(\sqrt{n}U_g^{(1)} + A_g^{(1)})}\right)} \\ &\lesssim A_g^{(1)}\sqrt{\frac{k}{m}\log\left(\frac{Lr\sqrt{m}}{A_g^{(1)}}\left[n\left(L_0 + T + \frac{B_0}{\zeta}\right) + \frac{T(\sqrt{n}U_g^{(1)} + A_g^{(1)})}{\zeta}\right]\right)}. \end{aligned} \tag{C.6}$$

(2.2) Bounding the centered product process $R_{u2;c}$.

We let $g_x(a_i) = |\bar\varepsilon_{i,\zeta}(a_i^\top x)|$ and $h_v(a_i) = |a_i^\top v|$, and write

$$R_{u2;c} = \sup_{x,v} \frac{1}{m}\sum_{i=1}^{m}\left\{g_x(a_i)h_v(a_i) - \mathbb{E}[g_x(a_i)h_v(a_i)]\right\}.$$

We verify the conditions in Theorem 2 as follows:

- For any $x, x' \in \mathcal{K}$, because $|\bar\varepsilon_{i,\zeta}|$ is $(2L_0 + \frac{B_0}{\zeta})$-Lipschitz continuous (Lemma 1), we have

$$\begin{aligned} \|g_x(a_i) - g_{x'}(a_i)\|_{\psi_2} &\le \left(2L_0 + \frac{B_0}{\zeta}\right)\|a_i^\top x - a_i^\top x'\|_{\psi_2} \\ &= O\left(L_0 + \frac{B_0}{\zeta}\right)\|x - x'\|_2. \end{aligned}$$

- By Lemma 7, with probability at least $1 - 2\exp(-\Omega(n))$ we have $\|a_i\|_2 = O(\sqrt{n})$. On this event, we have

$$\begin{aligned} |g_x(a_i) - g_{x'}(a_i)| &\le \left(2L_0 + \frac{B_0}{\zeta}\right)|a_i^\top x - a_i^\top x'| \\ &\le \left(2L_0 + \frac{B_0}{\zeta}\right)\|a_i\|_2\|x - x'\|_2 \\ &= O\left(\sqrt{n}\left(L_0 + \frac{B_0}{\zeta}\right)\right)\|x - x'\|_2. \end{aligned}$$

- For any $v, v' \in \mathcal{B}(K, \delta)$ we have $\| |a_i^\top v| - |a_i^\top v'| \|_{\psi_2} \le \|a_i^\top (v - v')\|_{\psi_2} = O(1)\|v - v'\|_2$. Similarly as before, we assume $\|a_i\|_2 = O(\sqrt{n})$, which gives $|a_i^\top v - a_i^\top v'| \le \|a_i\|_2 \|v - v'\|_2 = O(\sqrt{n})\|v - v'\|_2$. Moreover, $\mathcal{B}(K, \delta) \subset \mathbb{B}_2^n$ implies $\| |a_i^\top v| \|_{\psi_2} = O(1)$ and $|a_i^\top v| \le \|a_i\|_2 \|v\|_2 = O(\sqrt{n})$ holds for all $v \in \mathcal{B}(K, \delta)$.

Combined with Assumption 3, $R_{u2;c}$ satisfies the conditions of Theorem 2 with

$$
\begin{aligned}
&M_g \lesssim L_0 + \frac{B_0}{\delta};\ A_g = A_g^{(2)};\ M_h \lesssim 1;\ A_h \asymp O(1) \\
&L_g \lesssim \sqrt{n}\left(L_0 + \frac{B_0}{\delta}\right);\ U_g = U_g^{(2)};\ L_h \lesssim \sqrt{n};\ U_h \lesssim \sqrt{n}
\end{aligned}
$$

and $P_0 = P_0^{(2)} + 2\exp(-\Omega(n))$. Suppose we have

$$
m \gtrsim \mathscr{H}\left(K, \frac{A_g^{(2)}}{\sqrt{m}n[L_0 + \frac{B_0}{\delta}]}\right) + \mathscr{H}\left(\mathcal{B}(K,\delta), \frac{A_g^{(2)}}{\sqrt{m}(A_g^{(2)} + \sqrt{n}U_g^{(2)})}\right),
$$

which can be guaranteed (from Lemma 6) by

$$
m \gtrsim k \log\left(\frac{Lr\sqrt{m}}{A_g^{(2)}}\left[n\left(L_0 + \frac{B_0}{\delta}\right) + \frac{T(A_g^{(2)} + \sqrt{n}U_g^{(2)})}{\delta}\right]\right). \tag{C.7}
$$

Then, we can invoke Theorem 2 to obtain that the following bound holds with probability at least $1 - mP_0^{(2)} - 2m\exp(-\Omega(n)) - C\exp(-\Omega(k))$:

$$
\begin{aligned}
|R_{u2;c}| &\lesssim \frac{A_g^{(2)}}{\sqrt{m}}\sqrt{\mathscr{H}\left(K, \frac{A_g^{(2)}}{\sqrt{m}n[L_0 + \frac{B_0}{\delta}]}\right) + \mathscr{H}\left(\mathcal{B}(K,\delta), \frac{A_g^{(2)}}{\sqrt{m}(A_g^{(2)} + \sqrt{n}U_g^{(2)})}\right)} \\
&\lesssim A_g^{(2)}\sqrt{\frac{k}{m}\log\left(\frac{Lr\sqrt{m}}{A_g^{(2)}}\left[n\left(L_0 + \frac{B_0}{\delta}\right) + \frac{T(A_g^{(2)} + \sqrt{n}U_g^{(2)})}{\delta}\right]\right)}.
\end{aligned} \tag{C.8}
$$

(2.3) Bounding the expectation terms $R_{u1;e}, R_{u2;e}$.

Recall that $\xi_{i;\delta}(a) = f_{i;\delta}(a) - Ta$ and $\varepsilon_{i;\delta}(a) = f_{i;\delta}(a) - f_i(a)$, and so $\xi_{i;\delta}(a) = \varepsilon_{i;\delta}(a) + f_i(a) - Ta$. Hence, by using $\mathbb{E}[a_i a_i^\top] = I_n$ and $\|v\|_2 = 1$, we have

$$
\begin{aligned}
R_{u1;e} &\le \sup_{x,v} \mathbb{E}\left[\left(f_i(a_i^\top x) - Ta_i^\top x\right)(a_i^\top v)\right] + \sup_{x,v} \mathbb{E}\left[\varepsilon_{i;\delta}(a_i^\top x) a_i^\top v\right] \\
&\le \sup_{x \in X} \left\|\mathbb{E}[f_i(a_i^\top x)a_i] - Tx\right\|_2 + \sup_{x,v} \mathbb{E}\,|\varepsilon_{i;\delta}(a_i^\top x)||a_i^\top v| \\
&\le \sup_{x \in X} \rho(x) + R_{u2;e};
\end{aligned} \tag{C.9}
$$

where $\rho(x)$ is the model mismatch defined in (2.3), and $R_{u2;e}$ is defined in (C.3). It remains to bound $R_{u2;e}$, for which we first apply Cauchy-Schwarz (with $\|v\|_2 \le 1$) and then use Lemma 13 to obtain

$$
\begin{aligned}
R_{u2;e} &= \sup_{x,v} \mathbb{E}\,|\varepsilon_{i;\delta}(a_i^\top x)||a_i^\top v| \\
&\le \sup_{x,v} \sqrt{\mathbb{E}[|\varepsilon_{i;\delta}(a_i^\top x)|^2]}\sqrt{\mathbb{E}[|a_i^\top v|^2]} \\
&\le \sup_{x \in K} \sqrt{\mathbb{E}\left[|\varepsilon_{i;\delta}(a_i^\top x)|^2 \mathbb{1}\left(a_i^\top x \in \mathscr{D}_{f_i} + \left[-\frac{\delta}{2}, \frac{\delta}{2}\right]\right)\right]} \\
&\le \left(\frac{3L_0\delta}{2} + B_0\right) \sup_{x \in K} \sqrt{\mathbb{P}\left(a_i^\top x \in \mathscr{D}_{f_i} + \left[-\frac{\delta}{2}, \frac{\delta}{2}\right]\right)} \\
&\le \left(\frac{3L_0\delta}{2} + B_0\right) \sup_{x \in K} \sqrt{\mu_\delta(x)};
\end{aligned} \tag{C.10}
$$

where we use Lemma 13 in the third and fourth line, and $\psi(x)$ is defined in (2.4).

(3) Combining everything to conclude the proof.

Recall that in Assumption 4, we assume that

$$\sup_{x\in\mathcal{X}} \psi(x) \lesssim (A_g^{(1)} \vee A_g^{(2)})\sqrt{\frac{k}{m}};$$

and we take sufficiently small δ_1 such that

$$(L_0\delta_1 + B_0)\sup_{x\in\mathcal{K}} \sqrt{\psi_{\delta_1}(x)} \lesssim (A_g^{(1)} \vee A_g^{(2)})\sqrt{\frac{k}{m}};$$

then by setting $\delta = \delta_1$, the derived bound of $R_{u1;c} + R_{u2;c}$ (see (C.6) and (C.8)) dominates that of $R_{u1;e} + R_{u2;e}$ (see (C.9) and (C.10)), and so $R_{u1} + R_{u2} \lesssim R_{u1;c} + R_{u2;c}$.

Recall that (C.6) and (C.8) are guaranteed by the sample size of (C.5) and (C.7), while (C.5) and (C.7) hold as long as

$$\begin{aligned} m &\gtrsim k\log\left(\frac{Lr\sqrt{m}}{A_g^{(1)}\wedge A_g^{(2)}}\left[n\left(L_0+T+\frac{B_0}{\delta_1}\right)+\frac{T(\sqrt{n}(U_g^{(1)}\vee U_g^{(2)})+(A_g^{(1)}\vee A_g^{(2)}))}{}\right]\right)\\ &:= kL \qquad \text{(here we use } L \text{ to abbreviate the log factors)} \end{aligned} \tag{C.11}$$

with probability at least $1 - m(P_0^{(1)} + P_0^{(2)}) - m\exp(-\Omega(n)) - C\exp(-\Omega(k))$ we have

$$R_{u1;c} + R_{u2;c} \lesssim (A_g^{(1)} \vee A_g^{(2)})\sqrt{\frac{kL}{m}}.$$

Therefore, the right-hand side of (3.1) can be uniformly bounded by

$$O\left(\|\hat{x} - Tx\|_2\,(A_g^{(1)} \vee A_g^{(2)})\sqrt{\frac{kL}{m}}\right). \tag{C.12}$$

Combining with the uniform lower bound for the left-hand side of (3.1), i.e., $\Omega(\|\hat{x} - Tx\|_2^2)$, we obtain the following bound uniformly for all $x \in \mathcal{X}$:

$$\|\hat{x} - Tx\|_2 \lesssim (A_g^{(1)} \vee A_g^{(2)})\sqrt{\frac{kL}{m}}.$$

Hence, as long as

$$m \gtrsim \left(A_g^{(1)} \vee A_g^{(2)}\right)^2 \frac{kL}{\epsilon^2};$$

we again obtain $\|\hat{x} - Tx\|_2 \le 3\epsilon$, which completes the proof. □

C.2 Proof of Theorem 2

Proof. Step 1. Control the process over finite nets.

Recall that $\mathcal{X}$ and $\mathcal{V}$ are the index sets of x and v, as stated in Theorem 2. We first establish the desired concentration for a fixed pair $(x, v) \in \mathcal{X} \times \mathcal{V}$. By Lemma 2, $\|g_x(a_i)h_v(a_i)\|_{\psi_1} \le \|g_x(a_i)\|_{\psi_2}\|h_v(a_i)\|_{\psi_2} \le A_gA_h$. Furthermore, centering (Lemma 3) gives

$$\|g_x(a_i)h_v(a_i) - \mathbb{E}[g_x(a_i)h_v(a_i)]\|_{\psi_1} = O(A_gA_h).$$

Thus, for fixed $(x, v) \in \mathcal{X} \times \mathcal{V}$ we define

$$I_{x,v} = \frac{1}{m}\sum_{i=1}^{m} g_x(a_i)h_v(a_i) - \mathbb{E}[g_x(a_i)h_v(a_i)].$$

Then, we can invoke Bernstein's inequality (Lemma 4) to obtain for any $t \ge 0$ that

$$\mathbb{P}\left(|I_{x,v}| \ge t\right) \le 2\exp\left(-cm\min\left\{\left(\frac{t}{A_gA_h}\right)^2, \frac{t}{A_gA_h}\right\}\right). \tag{C.13}$$

We construct G_1 as an ϵ_1-net of X, and G_2 as an ϵ_2-net of V, with both nets being minimal in that $\log|G_1| = H(X;\epsilon_1)$, $\log|G_2| = H(V;\epsilon_2)$, and where ϵ_1, ϵ_2 are to be chosen later. Then, we take a union bound of (C.13) over $(x,v)\in G_1\times G_2$ to obtain

$$P\left(\sup_{\substack{x\in G_1\\ v\in G_2}} |I_{x,v}| \ge t\right) \le 2\exp\left(H(X;\epsilon_1)+H(V;\epsilon_2) - cm\min\left(\left(\frac{t}{A_gA_h}\right)^2, \frac{t}{A_gA_h}\right)\right). \tag{C.14}$$

Now we set $t \asymp A_gA_h\sqrt{\frac{H(X;\epsilon_1)+H(V;\epsilon_2)}{m}}$ for a sufficiently large hidden constant. Then, if $m \ge C(H(X;\epsilon_1)+H(V;\epsilon_2))$ for large enough C so that $\frac{t}{A_gA_h}\le 1$ (we assume this now and will confirm it in (C.21) after specifying ϵ_1, ϵ_2), (C.14) gives

$$P\left(\sup_{\substack{x\in G_1\\ v\in G_2}} |I_{x,v}| \gtrsim A_gA_h\sqrt{\frac{H(X;\epsilon_1)+H(V;\epsilon_2)}{m}}\right) \le 2\exp\left(-\left(H(X;\epsilon_1)+H(V;\epsilon_2)\right)\right)$$

Hence, from now on we proceed with the proof on the event

$$\sup_{\substack{x\in G_1\\ v\in G_2}} |I_{x,v}| \lesssim A_gA_h\sqrt{\frac{H(X;\epsilon_1)+H(V;\epsilon_2)}{m}}, \tag{C.15}$$

which holds within the promised probability.

Step 2. Control the approximation error of the nets.

We have derived a bound for $\sup_{x\in G_1, v\in G_2}|I_{x,v}|$, while we want to control $I = \sup_{x\in X, v\in V}|I_{x,v}|$, so we further investigate how close these two quantities are. We define the event as

$$E_1 = \{\text{the events in (3.7) hold for all } a_i, i\in[m]\},$$

then by assumption (A2.) in the theorem statement, a union bound gives $P(E_1)\ge 1 - mP_0$. In the following, we proceed with the analysis of the event E_1. Combining with (C.15) we now bound $|I_{x,v}|$ for any given $x\in X, v\in V$. Specifically, we pick $x'\in G_1, v'\in G_2$ such that $\|x'-x\|_2\le \epsilon_1, \|v'-v\|_2\le\epsilon_2$, and thus we have

$$|I_{x,v}| \le |I_{x',v'}| + |I_{x,v} - I_{x',v'}| \le O\left(A_gA_h\sqrt{\frac{H(X;\epsilon_1)+H(V;\epsilon_2)}{m}}\right) + |I_{x,v} - I_{x',v'}|. \tag{C.16}$$

Moreover, we have

$$\begin{aligned} &|I_{x,v} - I_{x',v'}| \\ &= \left|\frac{1}{m}\sum_{i=1}^m \left(g_x(a_i)h_v(a_i) - g_{x'}(a_i)h_{v'}(a_i)\right) - \mathbb{E}\left(g_x(a_i)h_v(a_i) - g_{x'}(a_i)h_{v'}(a_i)\right)\right| \\ &\le \underbrace{\frac{1}{m}\sum_{i=1}^m |g_x(a_i)h_v(a_i) - g_{x'}(a_i)h_{v'}(a_i)|}_{\mathrm{err}_1} + \underbrace{\mathbb{E}|g_x(a_i)h_v(a_i) - g_{x'}(a_i)h_{v'}(a_i)|}_{\mathrm{err}_2} \end{aligned} \tag{C.17}$$

We bound err_1 using the event E_1 as follows:

$$\begin{aligned} \mathrm{err}_1 &\le \frac{1}{m}\sum_{i=1}^m \left(|g_x(a_i) - g_{x'}(a_i)||h_v(a_i)| + |h_v(a_i) - h_{v'}(a_i)||g_{x'}(a_i)|\right) \\ &\le \frac{1}{m}\sum_{i=1}^m \left(L_g\|x - x'\|_2 U_h + L_h\|v - v'\|_2 U_g\right) \\ &\le \frac{1}{m}\sum_{i=1}^m \left(L_gU_h\epsilon_1 + L_hU_g\epsilon_2\right) = L_gU_h\epsilon_1 + L_hU_g\epsilon_2. \end{aligned} \tag{C.18}$$

On the other hand, we bound err_2 using assumption (A1.) in the theorem statement. Noting that $\mathbb{E}|X| = O(\|X\|_{\psi_1})$ [50, Proposition 2.7.1(b)], and further applying Lemma 2 we obtain

$$
\begin{aligned}
\mathrm{err}_2 &\lesssim \|g_x(a_i)h_v(a_i) - g_{x'}(a_i)h_{v'}(a_i)\|_{\psi_1} \\
&\le \|(g_x(a_i) - g_{x'}(a_i))h_v(a_i)\|_{\psi_1} + \|g_{x'}(a_i)(h_v(a_i) - h_{v'}(a_i))\|_{\psi_1} \\
&\le \|g_x(a_i) - g_{x'}(a_i)\|_{\psi_2}\|h_v(a_i)\|_{\psi_2} + \|g_{x'}(a_i)\|_{\psi_2}\|h_v(a_i) - h_{v'}(a_i)\|_{\psi_2} \\
&\le M_g\|x - x'\|_2 \cdot A_h + A_g \cdot M_h\|v - v'\|_2 \le M_gA_h\epsilon_1 + A_gM_h\epsilon_2.
\end{aligned}
\tag{C.19}
$$

Note that the bounds (C.18) and (C.19) hold uniformly for all $(x, v) \in \mathcal{X} \times \mathcal{V}$, and hence, we can substitute them into (C.16) and (C.17) to obtain

$$
\sup_{\substack{x \in \mathcal{X} \\ v \in \mathcal{V}}} |I_{x,v}| \le O\left(A_gA_h\sqrt{\frac{H(\mathcal{X};\epsilon_1) + H(\mathcal{V};\epsilon_2)}{m}} + (L_gU_h + M_gA_h)\epsilon_1 + (L_hU_g + M_hA_g)\epsilon_2\right). \tag{C.20}
$$

Recall that we use the shorthand $S_{g,h} := L_gU_h + M_gA_h$ and $T_{g,h} := L_hU_g + M_hA_g$. We set $\epsilon_1 \asymp \frac{A_gA_h}{\sqrt{m}S_{g,h}}$, $\epsilon_2 \asymp \frac{A_gA_h}{\sqrt{m}T_{g,h}}$ so that the right-hand side of (C.20) is dominated by the first term. Overall, with a sample size satisfying

$$
m = \Omega\left(H\left(\mathcal{X}; \frac{A_gA_h}{\sqrt{m}S_{g,h}}\right) + H\left(\mathcal{V}; \frac{A_gA_h}{\sqrt{m}T_{g,h}}\right)\right), \tag{C.21}
$$

we can bound $I = \sup_{x\in\mathcal{X}}\sup_{v\in\mathcal{V}} |I_{x,v}|$ (defined in the theorem statement) as

$$
I \lesssim A_gA_h\sqrt{\frac{H(\mathcal{X}; A_gA_hm^{-1/2}S_{g,h}^{-1}) + H(\mathcal{V}; A_gA_hm^{-1/2}T_{g,h}^{-1})}{m}} \tag{C.22}
$$

with probability at least

$$
1 - mP_0 - 2\exp\left(-\Omega\left(H\left(\mathcal{X}; \frac{A_gA_h}{\sqrt{m}S_{g,h}}\right) + H\left(\mathcal{V}; \frac{A_gA_h}{\sqrt{m}T_{g,h}}\right)\right)\right).
$$

This completes the proof. □

D Other Omitted Proofs

D.1 Proof of Lemma 1 (Lipschitz continuity of $f_{i,\delta}$ and $\varepsilon_{i,\delta}$).

Proof. It is straightforward to check that $f_{i,\delta}$ and $|\varepsilon_{i,\delta}|$ are piece-wise continuous functions; hence, it suffices to prove that they are Lipschitz with the claimed Lipschitz constant over each piece. In any interval contained in the part of $x \notin D_{f_i} + [-\frac{\delta}{2}, \frac{\delta}{2}]$, $f_{i,\delta} = f_i$ and $|\varepsilon_{i,\delta}| = 0$ trivially satisfy the claim. In any interval contained in $[x_0 - \frac{\delta}{2}, x_0]$ for some $x_0 \in D_{f_i}$, $f_{i,\delta}$ is linear with slope $\frac{2}{\delta}\left(f_i^a(x_0) - f_i(x_0 - \frac{\delta}{2})\right)$, combined with the bound $|f_i^a(x_0) - f_i(x_0 - \frac{\delta}{2})| \le |f_i^a(x_0) - f_i^-(x_0)| + |f_i^-(x_0) - f_i(x_0 - \frac{\delta}{2})| \le \frac{|f_i^+(x_0) - f_i^-(x_0)|}{2} + \frac{L_0\delta}{2} \le \frac{1}{2}(B_0 + L_0\delta)$, we know that $f_{i,\delta}$ is $\left(L_0 + \frac{B_0}{\delta}\right)$-Lipschitz. Further $|\varepsilon_{i,\delta}| = |f_{i,\delta} - f_i|$, and f_i is L_0-Lipschitz over this interval, so $|\varepsilon_{i,\delta}|$ is $\left(2L_0 + \frac{B_0}{\delta}\right)$-Lipschitz continuous. A similar argument applies to an interval contained in $[x_0, x_0 + \frac{\delta}{2}]$. □

D.2 Proof of Lemma 6 (Metric entropy of constraint sets).

Proof. Bounding $H(\mathcal{K}; \epsilon)$.

By [50, Corollary 4.2.13], there exists an $\frac{\epsilon}{Lr}$-net G_1 of $\mathbb{B}_2^k$ such that

$$
\log|G_1| \le k\log\left(\frac{2Lr}{\epsilon} + 1\right) \le k\log\frac{3Lr}{\epsilon},
$$

where we use $\epsilon \le Lr$. Note that rG_1 is an $\frac{\epsilon}{L}$-net of $\mathbb{B}_2^k(r)$, and because G is L-Lipschitz, $G(rG_1)$ is an ϵ-net of $\mathcal{K}$, thus yielding $H(\mathcal{K}; \epsilon) \le k\log\frac{3Lr}{\epsilon}$.

Bounding $\mathcal{H}(K^{-};\epsilon)$ and $\mathcal{H}(K^{*};\epsilon)$.

We construct G_2 as an $\frac{\epsilon}{2}$-net of K satisfying $\log|G_2| \le k\log\frac{6Lr}{\epsilon}$. Then, it is easy to see that $G_2 - G_2$ is an ϵ-net of $K^{-} = K - K$, showing that

$$\mathcal{H}(K^{-};\epsilon) \le \log|G_2|^2 \le 2k\log\frac{6Lr}{\epsilon}.$$

For a given $T>0$, this directly implies $\mathcal{H}(TK^{-};\epsilon) \le 2k\log\frac{6TLr}{\epsilon}$. Moreover, because $K^{*} \subset TK^{-}$, by [50, Exercise 4.2.10] (which states that $\mathcal{H}(K_1;r) \le \mathcal{H}(K_2;\frac{r}{2})$ holds for any $r>0$ if $K_1 \subset K_2$) we obtain

$$\mathcal{H}(K^{*};\epsilon) \le \mathcal{H}\left(TK^{-};\frac{\epsilon}{2}\right) \le 2k\log\frac{12TLr}{\epsilon}.$$

Bounding $\mathcal{H}((K^{*})^{*};\epsilon)$.

We construct G_3 as an $\frac{\epsilon}{2}$-net of K^{*} satisfying $\log|G_3| \le 2k\log\frac{12TLr}{\epsilon}$, then we consider $(G_3)^{*} := \{\frac{z}{\|z\|_2} : z \in G_3\}$. We aim to prove that $(G_3)^{*}$ is an ϵ-net of $(K^{*})^{*}$. Note that any $x_1 \in (K^{*})^{*}$ can be written as $\frac{z_1}{\|z_1\|_2}$ for some $z_1 \in K^{*}$ and recall that $\|z_1\|_2 \ge 2\tau$. Moreover, by construction, there exists some $z_2 \in G_3$ such that $\|z_1 - z_2\|_2 \le \frac{\epsilon}{2}$. Note that $\frac{z_2}{\|z_2\|_2} \in (G_3)^{*}$, and moreover we have

$$\begin{aligned}\left\|\frac{z_1}{\|z_1\|_2}-\frac{z_2}{\|z_2\|_2}\right\|_2 &\le \left\|\frac{z_1}{\|z_1\|_2}-\frac{z_2}{\|z_1\|_2}\right\|_2+\left\|\frac{z_2}{\|z_1\|_2}-\frac{z_2}{\|z_2\|_2}\right\|_2\\ &= \frac{\|z_1-z_2\|_2}{\|z_1\|_2}+\frac{\big|\|z_2\|_2-\|z_1\|_2\big|}{\|z_1\|_2}\\ &\le \frac{2\|z_1-z_2\|_2}{\|z_1\|_2} \le \frac{2\epsilon}{2}=\epsilon.\end{aligned}$$

Hence, we obtain

$$\mathcal{H}((K^{*})^{*};\epsilon) \le \log|(G_3)^{*}| \le \log|G_3| \le 2k\log\frac{12TLr}{\epsilon},$$

which completes the proof. □

D.3 Proof of Lemma 8 (Choice of T in 1-bit GCS).

Proof. Since $x \in \mathbb{S}^{n-1}$, for some orthogonal matrix P we have $Px = e_1$ (the first column of I_n). Since $\tilde{a} := Pa = [\tilde{a}_i]$ has the same distribution as a, we have

$$\begin{aligned}\mathbb{E}[\mathrm{sign}(a^\top x)a] &= \mathbb{E}[\mathrm{sign}(\tilde{a}^\top e_1)P^\top\tilde{a}] = P^\top\mathbb{E}[\mathrm{sign}(\tilde{a}_1)\tilde{a}]\\ &= P^\top\sqrt{\frac{2}{\pi}}e_1 = \sqrt{\frac{2}{\pi}}x.\end{aligned}$$

□

D.4 Proof of Lemma 9 (Choice of T in 1-bit GCS with dithering).

Proof. We first note that

$$\left\|\mathbb{E}[\mathrm{sign}(a^\top x+\tau)a]-\frac{x}{\delta}\right\|_2=\frac{1}{\delta}\sup_{v\in\mathbb{S}^{n-1}}\left(\mathbb{E}\left[\delta\,\mathrm{sign}(a^\top x+\tau)a^\top v\right]-x^\top v\right). \qquad \text{(D.1)}$$

We first fix a and expect over $\tau \sim \mathcal{U}[-\delta,\delta]$ to obtain

$$\begin{aligned}&\mathbb{E}_\tau\left[\delta\,\mathrm{sign}(a^\top x+\tau)a^\top v\right]\\ &=(\delta a^\top v)\left[\mathbb{1}(|a^\top x|>\delta)\,\mathrm{sign}(a^\top x)+\mathbb{1}(|a^\top x|\le\delta)\left(\frac{\delta+a^\top x}{2\delta}-\frac{\delta-a^\top x}{2\delta}\right)\right]\\ &=(a^\top x)(a^\top v)\mathbb{1}(|a^\top x|\le\delta)+(\delta a^\top v)\,\mathrm{sign}(a^\top x)\mathbb{1}(|a^\top x|>\delta).\end{aligned}$$

We plug this into (D.1), and note that $x^{>} v = E[(a^{>} x)(a^{>} v)]$, which gives

$$
\begin{aligned}
& E[\mathrm{sign}(a^{>} x + \ )a] \quad \frac{x}{\ \ }_{2} \\
= & \frac{1}{\ } \sup_{v \, 2 \, S^{n \ 1}} E \ \Big[(\ a^{>} v) \, \mathrm{sign}(a^{>} x) \quad (a^{>} x)(a^{>} v) \ 1(j a^{>} x j > \ ) \Big] \qquad \text{(D.2)} \\
& \frac{1}{\ } \sup_{v \, 2 \, S^{n \ 1}} E \ [\ j a^{>} v j + j a^{>} x j j a^{>} v j] 1(j a^{>} x j > \ )
\end{aligned}
$$

For any $x \ 2 \ B_2^n(R)$ and $v \ 2 \ S^{n \ 1}$, we have $k a^{>} x k_{\ 2} = O(R)$ and $k a^{>} v k_{\ 2} = O(1)$. Applying the Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
& E \ [\ j a^{>} v j + j a^{>} x j j a^{>} v j] 1(j a^{>} x j > \ ) \\
& \sqrt{E[(\ j a^{>} v j + j x^{>} a a^{>} v j)^2]} \ \sqrt{P(j a^{>} x j > \ )} \\
& \sqrt{2 \ \ ^2 E[j a^{>} v j^2] + E[(a^{>} x)^2 (a^{>} v)^2]} \ \sqrt{2 \exp(\ c \ ^2 = R^2)} \\
. \ \ & \exp \ \frac{c \ ^2}{R^2} \ . \ \frac{\ }{m^9}
\end{aligned}
$$

Note that in the third line, we use the probability tail bound of the sub-Gaussian $j a^{>} x j$, and in the last line, we use $= C R \sqrt{\log m}$ with some suf ciently large C. The proof is completed by substituting this into (D.2). □

D.5 Proof of Lemma 10 (Choice of T in SIM).

Proof. This lemma slightly generalizes that of Lemma 8. We again choose an orthogonal matrix P such that $P x = e_1$, where e_1 represents the rst column of I_n. Since a and $P a$ have the same distribution, we have

$$
\begin{aligned}
E[f(a^{>} x)a] &= P^{>} E[f((P a)^{>} e_1) P a] \\
&= P^{>} E[f(a^{>} e_1)a] = P^{>} (\ e_1) = \ x:
\end{aligned}
$$

□

D.6 Proof of Lemma 11 (Choice of T in uniformly quantized GCS with dithering).

Proof. In the theorem, the statement before "In particular" can be found in [18, Theorem 1]. Based on this, we have $E[Q \ (a^{>} x + \ )a] = E_a E \ [Q \ (a^{>} x + \ )a] = E_a (a a^{>} x) = x:$ □

D.7 Proof of Lemma 12. (Bounds on $j \ _{i;} \ j$ and $j "_{i;} \ j$ for the uniform quantizer).

Proof. By the de nition of $f_{i;}$ in (3.4), we have $j "_{i;} \ (a) j = j f_{i;} \ (a) \quad f_i(a) j \quad$. It follows that $f_i(\ ) = Q \ (\ + \ )$ with $Q \ (a) = \ b \frac{a}{\ } c + \frac{1}{2}$, and $j Q \ (a) \quad a j \quad \frac{\ }{2}$ holds for any $a \ 2 \ R$. Hence, we have $j f_i(a) \quad a j = j Q \ (a + \ ) \quad (a + \ ) + \ j \ \ j Q \ (a + \ ) \quad (a + \ ) j + j \ j \ \ \frac{\ }{2} + \frac{\ }{2} = \ $. To complete the proof, we use the inequalities $j \ _{i;} \ (a) j \ \ j \ f_{i;} \ (a) \quad f_i(a) j + j f_i(a) \quad a j \quad \ + \ = 2 \ .$ □

D.8 Proof of Lemma 13. (Bound on the approximation error $j "_{i;} \ j$)

Proof. For any $a \ 2 = D_{f_i} + [\ \frac{\ }{2}; \frac{\ }{2}]$, by the de nition in (3.4) we have $"_{i;} \ (a) = 0$. If $a \ 2 \ [x_0 \ \ \frac{\ }{2}; x_0]$ for some $x_0 \ 2 \ D_{f_i}$, then we have

$$
\begin{aligned}
j "_{i;} \ (a) j = & j f_{i;} \ (a) \quad f_i(a) j \\
& j \ f_{i;} \ (a) \quad f_{i;} \ (x_0) j + j f_{i;} \ (x_0) \quad f_i \ (x_0) j + j f_i \ (x_0) \quad f_i(a) j \\
& 2 L_0 + \frac{B_0}{\ } \ j a \quad x_0 j + j f_i^{a}(x_0) \quad f_i \ (x_0) j + L_0 j x_0 \quad a j \\
& 3 L_0 + \frac{B_0}{\ } \quad \frac{\ }{2} + \frac{1}{2} j f_i^{+}(x_0) \quad f_i \ (x_0) j \quad \frac{3 L_0}{2} + B_0;
\end{aligned}
$$

where we use Lemma 1 and Assumption 2 in the third line, and use $j a \quad x_0 j \quad \frac{\ }{2}$ and $f_i^{a}(x_0) = \frac{1}{2} \ f_i \ (x_0) + f_i^{+}(x_0)$ in the fourth line. □

E Parameter Selection for Specific Models

E.1 1-bit GCS

To specialize Theorem 1 to this model, we select the parameters as follows:

- Assumption 2. Under the 1-bit observation model $y_i = \mathrm{sign}(a_i^\top x^\star)$, the function $f_i(\cdot) = f(\cdot) = \mathrm{sign}(\cdot)$ satisfies Assumption 2 with $(B_0, L_0, \beta_0) = (2, 0, \infty)$.
- (2.5) in Assumption 4. Recall that $\mathcal{K} \subset \mathbb{S}^{n-1}$. Under the assumption $\|x\|_2 = 1$, we set $T = \sqrt{2/\pi}$ so that $\rho(x) = 0$ holds for all $x \in \mathcal{X}$ (Lemma 8), which provides (2.5).
- Assumption 3. By Lemma 7, we have $\mathbb{P}(\|a\|_2 = O(\sqrt{n})) \ge 1 - 2\exp(-\Omega(n))$, and we suppose that this high-probability event holds. Also note that $|f_{i,\beta}| \le 1$. Hence, we have

$$\|\xi_{i,\beta}(a^\top x)\|_{\psi_2} \le \|f_{i,\beta}(a^\top x)\|_{\psi_2} + \|Ta^\top x\|_{\psi_2} = O(1);$$
$$\sup_{x \in \mathcal{K}} |\xi_{i,\beta}(a^\top x)| \le |f_{i,\beta}(a^\top x)| + |Ta^\top x| \le 1 + T\|a\|_2 = O(\sqrt{n}).$$

  Because $\varepsilon_{i,\beta} = f_{i,\beta} - f_i$, we have $\|\varepsilon_{i,\beta}(a^\top x)\|_{\psi_2} = O(1)$, and $|\varepsilon_{i,\beta}(a^\top x)| \le 2$ holds deterministically. Hence, regarding the parameters in Assumption 3, we can take

$$A_g^{(1)} \asymp 1; U_g^{(1)} \asymp \sqrt{n}; P_0^{(1)} \asymp \exp(-\Omega(n)); A_g^{(2)} \asymp 1; U_g^{(2)} \asymp 1; P_0^{(2)} = 0.$$

- (2.6) in Assumption 4. It remains to pick β_1 that satisfies (2.6). Note that $\mathscr{D}_{f_i} = \{0\}$, and for any $x \in \mathcal{K}$, $a^\top x \sim \mathcal{N}(0, 1)$, so we have

$$\mu_\beta(x) = \mathbb{P}\left(a^\top x \in \left[-\frac{\beta}{2}, \frac{\beta}{2}\right]\right) = O(\beta).$$

  Thus, we take $\beta = \beta_1 \asymp \frac{k}{m}$ to guarantee (2.6).

E.2 1-bit GCS with dithering

To specialize Theorem 1 to this model, we select the parameters as follows:

- Assumption 2. The observation function can be written as $f_i(x) = \mathrm{sign}(x + \tau_i)$ with $\tau_i \sim \mathcal{U}[-\lambda, \lambda]$, which satisfies Assumption 2 with $(B_0, L_0, \beta_0) = (2, 0, \infty)$.
- (2.5) in Assumption 4. We set $\lambda = CR\sqrt{\log m}$ with C large enough, so that Lemma 9 justifies (2.5).
- Assumption 3. By Lemma 7, we have $\mathbb{P}(\|a\|_2 = O(\sqrt{n})) \ge 1 - 2\exp(-\Omega(n))$. Assume this event holds, and note that $f_{i,\beta}$ is still bounded by 1, we have

$$\|\xi_{i,\beta}(a^\top x)\|_{\psi_2} \le \|f_{i,\beta}(a^\top x)\|_{\psi_2} + \|\lambda^{-1} a^\top x\|_{\psi_2} = O(R/\lambda) + O(1) = O(1);$$
$$\sup_{x \in \mathcal{K}} |\xi_{i,\beta}(a^\top x)| \le 1 + \sup_{x \in \mathcal{K}} |\lambda^{-1} a^\top x| \le 1 + \sup_{x \in \mathcal{K}} \lambda^{-1}\|a\|_2\|x\|_2 = O(\sqrt{n}).$$

  Moreover, because $\varepsilon_{i,\beta} = f_{i,\beta} - f_i$, the following hold deterministically: $\|\varepsilon_{i,\beta}(a^\top x)\|_{\psi_2} = O(1)$, $\sup_{x \in \mathcal{K}} |\varepsilon_{i,\beta}(a^\top x)| = O(1)$. Thus, regarding the parameters in Assumption 3 we can take

$$A_g^{(1)} \asymp 1;\ U_g^{(1)} \asymp \sqrt{n};\ P_0^{(1)} \asymp \exp(-\Omega(n));\ A_g^{(2)} \asymp 1;\ U_g^{(2)} \asymp 1;\ P_0^{(2)} = 0.$$

- (2.6) in Assumption 4. It remains to confirm (2.6) for suitable β_1. For any β, note that $\mathscr{D}_{f_i} + [-\frac{\beta}{2}, \frac{\beta}{2}] = [-\tau_i - \frac{\beta}{2}, -\tau_i + \frac{\beta}{2}]$, and hence for any $x \in \mathcal{K} \subset \mathbb{B}_2^n(R)$ we have

$$\mu_\beta(x) = \mathbb{P}\left(a^\top x \in \left[-\tau_i - \frac{\beta}{2}, -\tau_i + \frac{\beta}{2}\right]\right) = \mathbb{P}\left(a^\top x + \tau_i \in \left[-\frac{\beta}{2}, \frac{\beta}{2}\right]\right) \le \frac{\beta}{\lambda},$$

  which can be seen by conditioning on a. Hence, we can take $\beta = \lambda\beta_1 = \frac{k}{m}$ to guarantee (2.6).

E.3 Lipschitz-continuous SIM with generative prior

To specialize Theorem 1 to this model, we select the parameters as follows:

- Assumption 2. Since f is $\hat{L}$-Lipschitz by assumption, it satisfies Assumption 2 with $(B_0; L_0; \ _0) = (0; \hat{L}; 1)$.
- (2.5) in Assumption 4. Recall that we have defined the quantities $= E[f(g)g];$ $= \|f(g)\|_{ 2};$ where g $N(0;1)$. Then, we choose $T =$ so that $(x) = 0$ holds for any x (Lemma 10), thus justifying (2.5).
- Assumption 3. Because f_i is $\hat{L}$-Lipschitz and does not contain any discontinuity, there is no need to construct the Lipschitz approximation $f_{i;}$ for some > 0, while we simply use $= 0$, which implies $f_{i;} = f_i$ and $"_{i;} = 0$. Note that $_{i;}(a) = f_i(a)$ a, and so we have

$$\|f_i(a^\top x) \quad a^\top x\|_{ 2} \quad \| f_i(a^\top x)\|_{ 2} + \| a^\top x\|_{ 2} = O(\ + \ ):$$

  We suppose $\|a\|_2 = O(\sqrt{n})$, which holds with probability at least $1 \ 2\exp(\ (n))$ (Lemma 7); we also suppose $f_i(0)$ $\hat{B}$, which holds with probability at least $1 \ P_0^0$ by assumption. On these two events, we have

$$|f_i(a^\top x) \quad a^\top x| \quad |f_i(a^\top x) \quad f_i(0)| + |f_i(0)| + \ \|a\|_2\|x\|_2$$
$$\hat{L}\|a\|_2 + \hat{B} + \ \|a\|_2 \lesssim (\hat{L} + \ )\sqrt{n} + \hat{B}:$$

  Combined with $"_{i;} = 0$, we can set the parameters in Assumption 3 as follows:

$$A_g^{(1)} \quad + \ ; \ U_g^{(1)} \quad (\hat{L} + \ )\sqrt{n} + \hat{B}; \ P_0^{(1)} \quad P_0^0 + \exp(\ (n));$$
$$A_g^{(2)} \quad + \ ; \ U_g^{(2)} = 0; \ P_0^{(2)} = 0:$$

- (2.6) in Assumption 4. Because $= 0$ and $D_{f_i} = ?$, (2.6) is trivially satisfied.

E.4 Uniformly quantized GCS with dithering

To specialize Theorem 1 to this model, we select the parameters as follows:

- Assumption 2. The uniform quantizer with resolution > 0 is defined as $Q(a) = \ b\frac{a}{}c + \frac{1}{2}$ for $a\ 2\ R$. We consider this quantizer with dithering $U[\frac{}{2}; \frac{}{2}]$. Specifically, we observe $y_i = Q(a_i^\top x + \ _i)$, so the observation function is $f_i(\ ) = Q(\ + \ )$ with $U[\frac{}{2}; \frac{}{2}]$. Hence, Assumption 2 is satisfied with $(B_0; L_0; \ _0) = (\ ; 0; \ )$.
- (2.5) in Assumption 4. The benefit of dithering is to whiten the quantization noise. With $T = 1$, for any $x\ 2\ K$, Lemma 11 implies $(x) = \|E[Q(a_i^\top x + \ _i)a_i] \quad x\|_2 = 0$, thus justifying (2.5).
- Assumption 3. Note that for any $2\ (0; \frac{}{2})$, by Lemma 12, we can take the parameters for Assumption 3 as follows:

$$A_g^{(1)}; U_g^{(1)}; A_g^{(2)}; U_g^{(2)} \quad ; \ P_0^{(1)} = P_0^{(2)} = 0:$$

- (2.6) in Assumption 4. All that remains is to pick $= \ _1$ that satisfies (2.6). Because $D_{f_i} = \ _i + \ Z$, hence for any $x\ 2\ K$ we have

$$(x) = P \ a^\top x\ 2 \quad + \ Z + \left[\frac{}{2}; \frac{}{2}\right] = P \ a^\top x + \ 2\ Z + \left[\frac{}{2}; \frac{}{2}\right] = O \ \frac{}{} \ ;$$

  which can be seen by using the randomness of $U[\frac{}{2}; \frac{}{2}]$ conditionally on x. Hence, we take $= \ _1 \ \frac{k}{m}$, which provides (2.6).

F Handling Sub-Gaussian Additive Noise

In this appendix, we describe how our results can be extended to the noisy model $y = f(Ax) + \ $, where $2\ R^m$ is the noise vector that is independent of $(A; f)$ and has i.i.d. sub-Gaussian entries $_i$ satisfying $\| _i\|_{ 2} = O(1)$. Along similar lines as in (3.1)-(3.3), we find that gives rise to an additional term $\frac{2}{m}h \ ; A(\hat{x} \quad Tx)i$ to the right-hand side of (3.1), which is bounded by

$2\|\hat{x} - Tx\|_2 \sup_{v\in\Sigma(K_\epsilon)} \frac{1}{m}\langle \epsilon, Av\rangle$, with the constraint set $\Sigma(K_\epsilon)$ defined in (3.2). Thus, in (3.3), in addition to R_u, in the noisy setting we need to bound the additional term

$$R_u^0 := \sup_{v\in\Sigma(K_\epsilon)} \frac{1}{m}\langle \epsilon, Av\rangle = \sup_{v\in\Sigma(K_\epsilon)} \frac{1}{m}\sum_{i=1}^m \epsilon_i a_i^\top v.$$

This can be done by the following lemma, which indicates that the sharp (uniform) rate in Theorem 1 can be retained in the presence of noise

Lemma 15. *(Bounding the additional term R_u^0). In the noisy setting described above, with probability at least $1 - C_1\exp(-\Omega(k\log\frac{TLr}{\epsilon})) - C_2\exp(-\Omega(m))$, we have $R_u^0 \lesssim \sqrt{\frac{k\log\frac{TLr}{\epsilon}}{m}}$.*

Proof. Conditioning on ϵ, the randomness of a_i's gives $\frac{1}{m}\sum_{i=1}^m \epsilon_i a_i \sim \mathcal{N}(0; \frac{\|\epsilon\|_2^2}{m^2} I_n)$, and so $\|(\frac{1}{m}\sum_{i=1}^m \epsilon_i a_i)^\top v_1 - (\frac{1}{m}\sum_{i=1}^m \epsilon_i a_i)^\top v_2\|_{\psi_2} \le \frac{C_0\|\epsilon\|_2\|v_1 - v_2\|_2}{m}$ holds for any $v_1, v_2 \in \mathbb{R}^n$. Let $\omega(\cdot)$ be the Gaussian width as defined in Lemma 5. Then, using the randomness of a_i's, Talagrand's comparison inequality [50, Exercise 8.6.5] yields that for any $t \ge 0$, we have

$$\mathbb{P}\left(R_u^0 \le \frac{C_1\|\epsilon\|_2\,[\omega(\Sigma(K_\epsilon)) + t]}{m}\right) \ge 1 - 2\exp(-t^2). \qquad \text{(F.1)}$$

Next, we bound the Gaussian width $\omega(\Sigma(K_\epsilon))$. Recall that $\Sigma(K_\epsilon)$ is defined in (3.2), and Lemma 6 bounds its metric entropy as $\mathscr{H}(\Sigma(K_\epsilon), \eta) \le 2k\log\frac{12TLr}{\epsilon\eta}$. Thus, we can invoke Dudley's integral inequality [50, Theorem 8.1.3] to obtain

$$\omega(\Sigma(K_\epsilon)) \le C_2\int_0^2 \sqrt{2k\log\frac{12TLr}{\epsilon\eta}}\,d\eta \lesssim \sqrt{k\log\frac{TLr}{\epsilon}}.$$

Now, we further let $t = \sqrt{k\log\frac{TLr}{\epsilon}}$ in (F.1) to obtain that $R_u^0 \lesssim \frac{\|\epsilon\|_2\sqrt{k\log\frac{TLr}{\epsilon}}}{m}$ holds with probability at least $1 - 2\exp(-k\log\frac{TLr}{\epsilon})$. It remains to deal with the randomness of ϵ and bound $\|\epsilon\|_2$. Because ϵ has i.i.d. entries with $\|\epsilon_i\|_{\psi_2} = O(1)$, by [50, Theorem 3.1.1] we can obtain that $\|\epsilon\|_2 \le C_3\sqrt{m}$ with probability at least $1 - 2\exp(-c_3 m)$. Substituting this bound into $R_u^0 \lesssim \frac{\|\epsilon\|_2\sqrt{k\log\frac{TLr}{\epsilon}}}{m}$, the result follows. □

To close this appendix, we briefly state how to adapt the proof of Theorem 1 to explicitly include the additive noise ϵ. Specifically, the left-hand side of (3.1) and its uniform lower bound $\Omega(\|\hat{x} - Tx\|_2^2)$ remain unchanged, while the right-hand side of (3.1) is now bounded by $2\|\hat{x} - Tx\|_2(R_{u1} + R_{u2} + R_u^0)$ (with $2\|\hat{x} - Tx\|_2 R_u^0$ being the additional term); thus, combining the bound (C.12) on $2\|\hat{x} - Tx\|_2(R_{u1} + R_{u2})$ and Lemma 15, we establish a uniform upper bound

$$O\left(\|\hat{x} - Tx\|_2\left[(A_g^{(1)} \vee A_g^{(2)})\sqrt{\frac{kL}{m}} + \sqrt{\frac{k\log\frac{TLr}{\epsilon}}{m}}\right]\right)$$

for the right-hand side of (3.1). Therefore, to ensure uniform recovery up to the ℓ_2-norm accuracy of ϵ under the sub-Gaussian noise, it suffices to have a sample complexity

$$m \gtrsim (A_g^{(1)} \vee A_g^{(2)})^2\frac{kL}{\epsilon^2} + \frac{k\log\frac{TLr}{\epsilon}}{\epsilon^2}. \qquad \text{(F.2)}$$

Since the logarithmic factors L in (C.11) dominates $\log\frac{TLr}{\epsilon}$, (F.2) indeed coincides with the sample complexity $m \gtrsim (A_g^{(1)} \vee A_g^{(2)})^2\frac{kL}{\epsilon^2}$ in Theorem 1 under the mild condition of $A_g^{(1)} \vee A_g^{(2)} = \Omega(1)$.

G Experimental Results for the MNIST dataset

G.1 Details of the Settings

In this section, we conduct experiments on the MNIST dataset [28] to support our theoretical framework. We use various nonlinear measurement models, including 1-bit, dithered 1-bit, ReLU,

Figure 2: Reconstructed images of the MNIST dataset for the noiseless 1-bit measurements with $m = 150$.

and uniformly quantized CS with dithering (UQD). We select 30 images from the MNIST testing set, ensuring that there are three images from each of the 10 classes for maximum variability. A single measurement matrix A is generated and used for all 30 test images. All the experiments are repeated for 10 random trials. All the experiments are run using Python 3.10.6 and PyTorch 2.0.0, with an NVIDIA RTX 3060 Laptop 6GB GPU.

We train a variational autoencoder (VAE) on the training set of the MNIST dataset, which has 60,000 images, each of size 784. The decoder of the VAE is a fully connected neural network with ReLU activations, with input dimension $k = 20$ and output dimension $n = 784$, and two hidden layers with 500 neurons each. We train the VAE using the Adam optimizer with a mini-batch size of 100 and a learning rate of 0.001.

Since our contributions are primarily theoretical, we only provide simple proof-of-concept experimental results. In particular, since (2.1) is intractable to solve exactly, to estimate the underlying signal, we choose to use the algorithm proposed in [2] (referred to as CSGM) to approximate it. CSGM performs a gradient descent algorithm in the latent space in $\mathbb{R}^k$ with random restarts. In addition, we compare with the Lasso program that is solved by the iterative shrinkage thresholding algorithm.

For CSGM, we follow the setting in [2] and perform 10 random restarts with 1000 gradient descent steps per restart and pick the reconstruction with the best measurement error.

G.2 Experimental Results for Noiseless 1-bit Measurements and Uniformly Quantized Measurements with Dithering

In this subsection, we present the numerical results for 1-bit measurements and uniformly quantized measurements with dithering, while the results for dithered 1-bit measurements and the Lipschitz SIM where the nonlinear link function is ReLU are similarly provided in Appendix G.3. For 1-bit measurements, since the underlying signal is assumed to be a unit vector and we aim to recover the direction of the signal, we use cosine similarity that is calculated as $\hat{x}^T x/(\|\hat{x}\|_2 \cdot \|x\|_2)$ with $\hat{x}$ being the estimated vector to measure the reconstruction performance. For uniformly quantized measurements with dithering, we use the relative ℓ_2-norm distance between the underlying signal and the estimated vector, i.e., $\|\hat{x} - x\|/\|x\|_2$, to measure the reconstruction performance.

Since this paper is concerned with uniform recovery performance, in each trial, we record the worst-case reconstruction performance (i.e., the smallest cosine similarity or the largest relative error) over the 30 test images, and the worst-case cosine similarity or relative error is averaged over 10 trials.

Figures 2, 3, and 4 show that for noiseless 1-bit measurements and uniformly quantized measurements with dithering with $\delta = 3$, the CSGM approach can produce reasonably accurate reconstruction for all the test images when the number of measurements m is as small as 150 and 100 respectively.

G.3 Experimental Results for ReLU and Dithered 1-bit Measurements

We present the experimental results for the ReLU link function and dithered 1-bit measurements in Figures 5, 6, and 7. For dithered 1-bit measurements, we set $R = \sqrt{\log m}$ with $R > 0$ being a tuning parameter. For the case of using the ReLU link function, similarly to noiseless 1-bit measurements, we calculate the cosine similarity to measure the reconstruction performance. For dithered 1-bit measurements, similarly to uniformly quantized measurements with dithering, we calculate the relative ℓ_2-norm distance. We observe that for these two nonlinear measurement models

Figure 3: Reconstructed images of the MNIST dataset for UQD with $m = 100$ and $= 3$.

(a) 1-bit with varying m (b) UQD with xed $= 3$ (c) UQD with xed $m = 200$

Figure 4: Quantitative results of the performance of CSGM for 1-bit and UQD measurements on the MNIST dataset.

with a single realization of the random measurement ensemble, CSGM can also lead to reasonably good reconstruction for all the test images when the number of measurements is small compared to the ambient dimension.

H Experimental Results for the CelebA dataset

In this section, we present numerical results for the CelebA dataset [35], which contains more than 200,000 face images for celebrities with an ambient dimension of $n = 12288$. We train a deep convolutional generative adversarial network (DCGAN) following the settings in https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html. The latent dimension of the generator is $k = 100$ and the number of epochs for training is 20. Since the experiments for CelebA are more time-consuming than those of MNIST, we select 20 images from the test set of CelebA and perform 5 random trials. Other settings are the same as those for the MNIST dataset.

Since we have observed from the numerical results for MNIST that the experiments for the ReLU link function and dithered 1-bit measurements are similar, we only present the results for noiseless 1-bit measurements and uniformly quantized observations with dithering.

From Figures 8 and 10, we observe that for noiseless 1-bit measurements with 1500 samples, a single measurement matrix A can lead to reasonably accurate reconstruction for all the 20 test images. In addition, from Figures 9 and 10, we observe that for uniformly quantized measurements with

Figure 5: Reconstructed images of the MNIST dataset for the ReLU link function with $m = 150$ and $= 0.2$.

Figure 6: Examples of reconstructed images of the MNIST dataset for dithered 1-bit measurements with $m = 250$ and $R = 5$.

(a) ReLU with fixed $= 0.1$ (b) ReLU with fixed $m = 150$ (c) Dithered 1-bit with fixed $R = 5$

Figure 7: Quantitative results of the performance of CSGM for the ReLU link function and dithered 1-bit measurements on the MNIST dataset.

dithering, a single realization of the measurement matrix and random dither is sufficient for the reasonably accurate recovery of the 20 test images when $m = 1000$ and $= 20$.

Figure 8: Reconstructed images of the CelebA dataset for the noiseless 1-bit measurements with $m = 1500$.

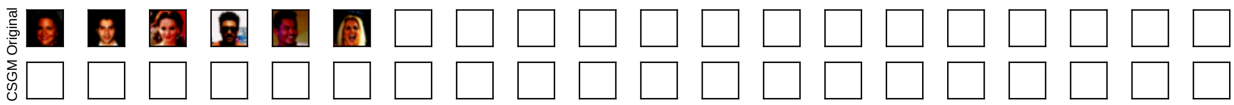
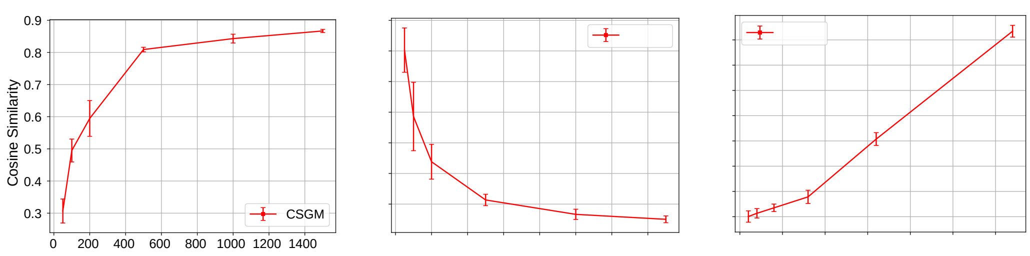


Figure 9: Reconstructed images of the CelebA dataset for UQD with $m = 1000$ and $\delta = 20$.



(a) 1-bit with varying m (b) UQD with fixed = 20 (c) UQD with fixed $m = 500$

Figure 10: Quantitative results of the performance of CSGM for 1-bit and UQD measurements on the CelebA dataset.