Robust Uniform Recovery of Structured Signals from Nonlinear Observations*

Pedro Abdalla† Radu Balan‡ Junren Chen§

April 19, 2026

Abstract

Uniform signal recovery from a fixed ensemble is an important desideratum in mathematical signal processing. While it is well known that the restricted isometry property (RIP) guarantees uniform sparse recovery from noisy linear measurements, uniform recovery of structured signals from nonlinear observations remains much less understood. This paper shows that the restricted approximate invertibility condition (RAIC) provides a unified approach to this end. Particularly, uniform recovery is achieved by projected gradient descent (PGD) with gradients obeying RAIC for all signals. As an application, under a large class of piecewise Lipschitz link functions (possibly discontinuous), we develop a uniform recovery theory for Gaussian single-index model by establishing the uniform RAIC for the gradient of the (scaled) ℓ_2 loss via a covering argument. The theory generalizes the nonuniform recovery guarantees due to Plan and Vershynin (2016); Oymak and Soltanolkotabi (2017) and exhibits additional error terms that can be interpreted as the cost of uniform recovery. Intriguingly, in the three canonical settings of (a) sparse recovery via PGD with ℓ_0 projection (i.e., iterative hard thresholding (IHT)), (b) sparse recovery via PGD with ℓ_1 projection, and (c) recovering approximately sparse signals via PGD with ℓ_1 projection, the additional error terms are negligible and in turn our uniform recovery error rates are at the same order of existing nonuniform ones, up to log factors. Our results hence improve on Genzel and Stollenwerk (2023). Under the specific nonlinearity of 1-bit quantization, we use a VC dimension argument to show that the uniform recovery error of IHT is at the same order of the nonuniform recovery error, with no loss of log factor. In addition, we show that the robustness of PGD to noise and corruption can be incorporated elegantly by bounding a single additional random process that captures the gradient mismatch.

1 Introduction

We consider the recovery of $\boldsymbol{x} \in \mathcal{X}$ from the observations

$$y_i = f_i(\boldsymbol{a}_i^T \boldsymbol{x}), \quad i = 1, ..., m \tag{1}$$

where $\boldsymbol{a}_i$ are known sensing vectors, $\mathcal{X}$ is a set of structured signals, and f_i denote potential nonlinear transforms. Canonical examples include 1-bit compressed sensing (Boufounos and Baraniuk, 2008), sparse phase retrieval (Candes et al., 2015), generalized linear models (McCullagh and Nelder, 2019), and so on.

*The authors are listed in alphabetical order. Corresponding author: Junren Chen (jchen58@umd.edu)
†University of California, Irvine.
‡University of Maryland, College Park.
§University of Maryland, College Park.

It is standard to consider random design $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_m]^T$, under which an important feature of mathematical recovery guarantee is the (non)uniformity. A nonuniform guarantee for an estimator $\hat{\boldsymbol{x}}$ ensures the accurate recovery of a fixed $\boldsymbol{x}$ *oblivious* to $\boldsymbol{A}$, taking the form

$$\mathbb{P}(\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 < \eta_1) \geq 1 - \eta_2$$

for some small $\eta_1, \eta_2 > 0$. The probability is taken over $\boldsymbol{A}$ and other randomness with the model. In contrast, a uniform guarantee is stronger and states that $\hat{\boldsymbol{x}}$ well approximates all $\boldsymbol{x}$ in $\mathcal{X}$:

$$\mathbb{P}\big(\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 < \eta_1,\ \forall \boldsymbol{x} \in \mathcal{X}\big) \geq 1 - \eta_2.$$

Put differently, nonuniform and uniform guarantees characterize respectively the average-case and worst-case performance of an estimator.

Uniformity is an important desideratum because in real-world applications the sensing ensemble is fixed when designed and is expected to be able to work with all possible signals. Uniform guarantee is also more "robust" in that it allows for an adversarial generation of $\boldsymbol{x} \in \mathcal{X}$ that can be based on knowledge of $\boldsymbol{A}$. From this perspective, uniformity may find interests beyond signal processing, for example, in statistical learning.

For concreteness, we now consider the sparse recovery problem where $\mathcal{X} \subset \Sigma_k^n := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_0 \leq k\}$. In the linear setting with $f_i = \mathrm{Id}$, that is the reconstruction of $\boldsymbol{x} \in \Sigma_k^n$ from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, there exists a number of efficient algorithms that uniformly recover all $\boldsymbol{x} \in \Sigma_k^n$ as long as the matrix $\boldsymbol{A}$ satisfies restricted isometry property (RIP). See, e.g., Foucart and Rauhut (2013). However, if f_i is a nonlinear transform such as $\mathrm{sign}(\cdot)$ and $|\cdot|$, the problem of establishing uniform guarantee becomes much more entangled. In the literature, diverse algorithms are proposed for different nonlinear observations, and the majority of their theoretical guarantees is nonuniform.

The recent work Genzel and Stollenwerk (2023) aims to bridge this gap by developing a uniform recovery theory for the single index model (1) with (sub)Gaussian covariates.[1] The authors revisited the generalized Lasso approach of Plan and Vershynin (2016) which works under a fairly large class of nonlinear link functions f_i without the need of precise knowledge of f_i. It was observed that the major technical challenge arises in bounding a multiplier process taking the form

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{v} \in \mathcal{V}} \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{x} - \frac{f_i(\boldsymbol{a}_i^T \boldsymbol{x})}{\mu} \right) \boldsymbol{a}_i^T \boldsymbol{v}, \tag{2}$$

especially when f_i is discontinuous. Noticing that for Lipschitz continuous f_i the concentration inequalities due to Mendelson (2016) directly apply, Genzel and Stollenwerk proposed a unified scheme based on Lipschitz approximation of discontinuous f_i to bound (2). However, for discontinuous f_i, this approach yields a bound on (2) decaying in m no faster than $O(m^{-1/4})$, and therefore the uniform recovery error of Generalized Lasso for (1) with discontinuous f_i is no faster than $O(m^{-1/4})$. This is inferior to the nonuniform recovery error in Plan and Vershynin (2016) and the uniform recovery error under Lipschitz continuous f_i (Genzel and Stollenwerk, 2023, Theorem 1), both of which read $O(m^{-1/2})$. Based on the upper bound $O(m^{-1/4})$, the authors concluded "*the transition to uniform recovery with nonlinear output functions may result in a worse oversampling rate*" (Genzel and Stollenwerk, 2023, Page 916). Does this gap between the uniform recovery error rates under Lipsthitz link functions and under discontinuous link functions truly exist?

As another restriction of Genzel and Stollenwerk (2023), most developments therein are tailored to the uniform recovery of Generalized Lasso. This, however, may not apply to some nonlinear

[1] The model in Genzel and Stollenwerk (2023) is slightly more general but we sacrifice some of the generality to ease the presentation.

observations (e.g., it does not directly apply to phase retrieval), or may not be a premier solver for specific problem (e.g., for generalized linear models, maximum likelihood estimation is typically a preferred option). It is unclear how to adapt Genzel and Stollenwerk (2023) to general nonlinear observations with a solver based on generic loss functions.

We note in passing that Chen et al. (2023) used the Lipschitz approximation approach from Genzel and Stollenwerk (2023) to establish $O(m^{-1/2})$ uniform recovery error rates for recovering all signals with a generative prior from (1). Yet, Chen et al. (2023) is not directly comparable to the recovery of structured signals (like sparse vectors) treated in Genzel and Stollenwerk (2023) and the present paper, since the recovery program for recovering generative vectors is in general not computational tractable.

This paper develops a different approach to uniform recovery of structured signals from nonlinear observations. Our work is built upon a line of recent works (Friedlander et al., 2021; Matsumoto and Mazumdar, 2024a,b, 2025; Chen and Yuan, 2024a,b; Chen et al., 2025) that used a structured condition, referred to as the restricted approximate invertibility condition (RAIC), to analyze nonlinear observation models. A historical review of these works can be found in Section 5. While a difficulty of the problem is that the RIP is no longer effective for nonlinear observations (Genzel and Stollenwerk, 2023), the RAIC precisely serves as an analog of RIP in various nonlinear models (Chen et al., 2025).

In particular, RAIC states that some gradient operator, when restricted to the structured signals, well approximates the "ideal (invertibility) step" under a dual norm related to the signal structure; it is useful because projected gradient descent (PGD) with a gradient obeying RAIC linearly converges to the true signal. Therefore, PGD with gradient obeying RAIC for all $\boldsymbol{x} \in \mathcal{X}$, referred to as a uniform RAIC, achieves uniform recovery for all $\boldsymbol{x} \in \mathcal{X}$. While this perspective was (implicitly) used to establish uniform recovery guarantees for several problems (see Section 5), our work carefully elucidates on this approach and offers some extensions, such as structured signals living in a convex set with canonical example being the recovery of approximately sparse vectors, as well as the robustness to noise and corruption which follows from slightly more work. Our work is therefore of some pedagogical value.

As an application of this approach, our second main contribution is to develop a uniform recovery theory for solving (1) with Gaussian covariate via PGD with respect to a properly scaled ℓ_2 loss (similarly to Oymak and Soltanolkotabi (2017)), which can be viewed as a computational procedure for generalized Lasso (Plan and Vershynin, 2016; Genzel and Stollenwerk, 2023). The main bulk of technical work lies in establishing the uniform RAIC for all $\boldsymbol{x} \in \mathcal{X}$, where we encounter exactly the same multiplier process (2) as with Genzel and Stollenwerk (2023). However, unlike Genzel and Stollenwerk (2023), we restrict our attention to a large class of f_i that are piecewise Lipschitz continuous (see Assumption 4.2) and control the process by a delicate covering argument. In turn, we establish uniform recovery guarantees different from the ones of Genzel and Stollenwerk (2023). By analyzing several canonical settings, we show that our uniform guarantees are at most log factors worse than the nonuniform guarantees in Plan and Vershynin (2016); Oymak and Soltanolkotabi (2017).

For concreteness, we consider sparse recovery as an example. If $\mathcal{X} \subset \Sigma_k^n$ and the discontinuous f_i's satisfy Assumption 4.3, then the nonuniform recovery error rate (for a fixed $\boldsymbol{x}$) reads

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 = O\left(\sqrt{\frac{k \log(en/k)}{m}}\right);$$

our result shows that, if f_i satisfies some additional regularity conditions (see Assumption 4.2),

then the uniform recovery error rate (for all $\boldsymbol{x} \in \mathcal{X}$) reads

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 = \tilde{O}\left(\sqrt{\frac{k\log(en/k)}{m}}\right),$$

where $\tilde{O}(\cdot)$ hides log factor in (m, n, k). This degrades from the nonuniform error bound only by log factors and substantially improves on the uniform rate $\tilde{O}((\frac{k\log(en/k)}{m})^{1/4})$ from Genzel and Stollenwerk (2023). For the specific 1-bit observations $\{y_i = \mathrm{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\}_{i=1}^m$, we provide a sharper analysis based on VC dimension and establish

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 = O\left(\sqrt{\frac{k\log(en/k)}{m}}\right),$$

meaning that the gap between the uniform rate and nonuniform rate is at most a universal multiplicative constant. In a nutshell, we show that the gap in (Genzel and Stollenwerk, 2023, Page 916) does not exist for a large class of discontinuous f_i. In other words, *discontinuous link functions that are regular enough cannot lead to an essential gap between the uniform rate and nonuniform rate*. See Figure 1 for an intuitive explanation of our contribution.

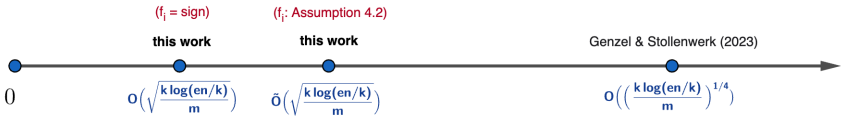

Figure 1: Existing uniform sparse recovery error rates for solving generalized Lasso. Note that existing nonuniform recovery error rates for generalized Lasso (Plan and Vershynin, 2016) and a corresponding PGD procedure (Oymak and Soltanolkotabi, 2017) are $O(\sqrt{k\log(en/k)/m})$ under possibly discontinuous f_i (satisfying a mild condition).

On a technical aspect, our covering argument for bounding (2) is analogous to Xu and Jacques (2020); Dirksen and Mendelson (2021); Chen et al. (2024, 2026) but require careful extensions of some steps to general and potentially discontinuous f_i (note that most previous works are specialized to quantization). Also, we provide a more refined argument for the specific 1-bit observations by VC-dimension theory with a suitable bound on multiplier processes; see Section 4.5 for further details.

The remainder of this paper is structured as follows. In Section 2, we introduce the PGD procedure and the RAIC for a cone or a convex set. Section 3 further unveils that RAIC implies the convergence of PGD to some statistical error and proposes our unified approach to uniform recovery of structured signals from nonlinear observations. In Section 4, the approach is applied to establishing uniform guarantees for single-index model along with careful comparisons with Plan and Vershynin (2016); Genzel and Stollenwerk (2023); Oymak and Soltanolkotabi (2017) (Sections 4.1–4.3); we also provide an instantiation of modulo measurements in Section 4.4 and a refined analysis of 1-bit measurements with no loss of log factors in Section 4.5. Section 5 reviews recent works on nonlinear observations through the lens of RAIC, and, building on these examples, Section 6 extends the robustness guarantees to noise and corruption. The missing proofs in Section 4 are provided in Appendices A–D.

2 PGD and RAIC

Our algorithm relies on a gradient operator $\boldsymbol{h}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$, which depends on $\boldsymbol{x}$ through the observations $y_i = f_i(\boldsymbol{a}_i^T \boldsymbol{x})$ only and is typically chosen as the (sub)gradient of some loss function. Specifically, $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ serves as the gradient at $\boldsymbol{u}$ in this work.

For the reconstruction of $\boldsymbol{x} \in \mathcal{X}$, we consider projected gradient descent (PGD) with step size η

$$\boldsymbol{x}_{t+1} = P_{\mathcal{K}}(\boldsymbol{x}_t - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t)), \quad t = 0, 1, \cdots, \tag{PGD}$$

where $P_{\mathcal{K}}$ denotes the projection onto a suitably chosen closed set $\mathcal{K}$, i.e.,

$$P_{\mathcal{K}}(\boldsymbol{u}) = \arg\min_{\boldsymbol{w} \in \mathcal{K}} \|\boldsymbol{w} - \boldsymbol{u}\|_2.$$

Note that (PGD) is computationally tractable as long as $P_{\mathcal{K}}$ is tractable. For instance, in sparse recovery, two standard choices of the projection are $\mathcal{K} = \Sigma_s^n$ (the ℓ_0 projection) and $\mathbb{B}_1^n(r)$ (the ℓ_1 projection), both of which can be implemented efficiently and aim to promote the sparsity. In particular, since $P_{\Sigma_s^n}$ reduces to the hard thresholding operator that only retains the s largest entries of the iterate, PGD with ℓ_0 projection is also referred to as a procedure of iterative hard thresholding (IHT) (Blumensath and Davies, 2009; Jacques et al., 2013). In general, we treat the two cases of (a) $\mathcal{K}$ is a cone and (b) $\mathcal{K}$ is a convex set. Also, $\mathcal{K}$ should be chosen such that $\mathcal{X} \subset \mathcal{K}$, that is, the desired signals live in $\mathcal{K}$.

For $\mathcal{U} \subset \mathbb{R}^n$ and any $\boldsymbol{v} \in \mathbb{R}^n$, the dual norm of $\boldsymbol{v}$ with respect to $\mathcal{U}$ is

$$\|\boldsymbol{v}\|_{\mathcal{U}^\circ} = \sup_{\boldsymbol{u} \in \mathcal{U}} \boldsymbol{u}^T \boldsymbol{v}.$$

Note that $\|\cdot\|_{\mathcal{U}^\circ}$ is a seminorm if $\mathcal{U}$ is symmetric, i.e., $\mathcal{U} = -\mathcal{U}$. It also satisfies $\|\boldsymbol{v}\|_{\mathcal{U}^\circ} \le \|\boldsymbol{v}\|_{\tilde{\mathcal{U}}^\circ}$ for any $\tilde{\mathcal{U}} \supset \mathcal{U}$. For $\phi > 0$, we let

$$\mathcal{K}_\phi = (\mathcal{K} - \mathcal{K}) \cap \phi B_2^n \qquad \text{and} \qquad \mathcal{K}_{\boldsymbol{x},\phi} = (\mathcal{K} - \boldsymbol{x}) \cap \phi B_2^n.$$

Our key technical component, the restricted approximate invertibility condition (RAIC), slightly differs for cone $\mathcal{K}$ and convex set $\mathcal{K}$. We shall start with the RAIC for a cone.

Definition 2.1 (RAIC for a cone (Chen et al., 2025))**.** Let $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{h}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathcal{K}$ be a cone. We say $\boldsymbol{h}_{\boldsymbol{x}}$ satisfies RAIC with respect to the cone $\mathcal{K}$, a constraint set $\mathcal{U}$, step size η, and an approximation error function $R_{\boldsymbol{x}} : \mathcal{U} \to \mathbb{R}$, if

$$\|\boldsymbol{u} - \boldsymbol{x} - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\|_{\mathcal{K}_1^\circ} \le R_{\boldsymbol{x}}(\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

We denote this by

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}(\mathcal{K}; \mathcal{U}, R_{\boldsymbol{x}}(\boldsymbol{u}), \eta).$$

For convex set, RAIC involves an additional scaling parameter $\phi > 0$.

Definition 2.2 (RAIC for a convex set)**.** Let $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{h}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$, $\mathcal{K}$ be a convex set and $\phi > 0$. We say $\boldsymbol{h}_{\boldsymbol{x}}$ satisfies RAIC at scale ϕ with respect to the convex set $\mathcal{K}$, a constraint set $\mathcal{U}$, step size η, and an approximation error function $R_{\boldsymbol{x},\phi} : \mathcal{U} \to \mathbb{R}$, if

$$\frac{1}{\phi}\|\boldsymbol{u} - \boldsymbol{x} - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\|_{\mathcal{K}_{\boldsymbol{x},\phi}^\circ} \le R_{\boldsymbol{x},\phi}(\boldsymbol{u}), \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

We denote this by

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}_\phi(\mathcal{K}; \mathcal{U}, R_{\boldsymbol{x},\phi}(\boldsymbol{u}), \eta).$$

Remark 2.1. Let $\phi' \geq \phi > 0$. Since $\mathcal{K} - \boldsymbol{x}$ is a convex set containing 0, $\frac{\mathcal{K}-\boldsymbol{x}}{\phi'} \subset \frac{\mathcal{K}-\boldsymbol{x}}{\phi}$ holds, and hence for any $\boldsymbol{w} \in \mathbb{R}^n$

$$\frac{1}{\phi}\|\boldsymbol{w}\|_{\mathcal{K}^\circ_{\boldsymbol{x},\phi}} = \|\boldsymbol{w}\|_{(\frac{\mathcal{K}-\boldsymbol{x}}{\phi}\cap B_2^n)^\circ} \geq \|\boldsymbol{w}\|_{(\frac{\mathcal{K}-\boldsymbol{x}}{\phi'}\cap B_2^n)^\circ} = \frac{1}{\phi'}\|\boldsymbol{w}\|_{\mathcal{K}^\circ_{\boldsymbol{x},\phi'}}.$$

This means that RAIC with smaller ϕ is harder to establish. In other words, suppose that

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}_\phi(\mathcal{K};\mathcal{U}, R_{\boldsymbol{x},\phi}(\boldsymbol{u}), \eta)$$

for any $\phi > 0$ with a set of approximation error functions $\{R_{\boldsymbol{x},\phi}(\boldsymbol{u})\}_{\phi>0}$, then one can assume $R_{\boldsymbol{x},\phi}(\boldsymbol{u}) \geq R_{\boldsymbol{x},\phi'}(\boldsymbol{u})$ $(\forall \boldsymbol{u} \in \mathcal{U})$ without loss of generality. ◊

At a current iterate is $\boldsymbol{u}$, $\eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ is the actual descent step, while $\boldsymbol{u}-\boldsymbol{x}$ is the ideal descent step (in light of $\boldsymbol{u} - (\boldsymbol{u} - \boldsymbol{x}) = \boldsymbol{x}$, which is the desired signal). Therefore, the RAIC essentially requires that the actual step approximates the ideal step under a dual norm adaptive to the low-dimensional signal structure. This is the key to establishing RAIC in high dimensions, as dual norm is in general much smaller than the ℓ_2 norm.

3 Unified Approach

This section proposes a (deterministic) unified approach to uniform recovery from nonlinear observations. It is built upon the linear convergence of PGD implied by RAIC.

In the case when $\mathcal{K}$ is a cone, the following statement shows that RAIC with approximation error $\mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2$ implies the convergence of PGD to $\boldsymbol{x}$ with per-iteration contraction rate $2\mu_1$ and error $\frac{2\mu_2}{1-2\mu_1}$. Let $B_2^n(\boldsymbol{x}; d) = \boldsymbol{x} + dB_2^n$.

Theorem 3.1. *(Chen et al., 2025, Theorem 3.1) Assume that $\mathcal{K}$ is a cone that contains $\boldsymbol{x}$. If*

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}(\mathcal{K};\mathcal{U}, R_{\boldsymbol{x}}(\boldsymbol{u}), \eta)$$

where $\mathcal{U} \supset \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$ for some $0 < d \leq \infty$ and $R_{\boldsymbol{x}}(\boldsymbol{u}) = \mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2$ with $\mu_1 < \frac{1}{2}$ and $d > \frac{2\mu_2}{1-2\mu_1}$, then $\{\boldsymbol{x}_t\}_{t\geq 0}$ generated by (PGD) with $\boldsymbol{x}_0 \in \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$ satisfies

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq (2\mu_1)^t\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + \frac{2\mu_2}{1-2\mu_1}, \quad \forall t \geq 0.$$

Remark 3.1. $\mathcal{U} \supset \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$ means that $\boldsymbol{x}$ is an interior point of $\mathcal{U}$ relative to $\mathcal{K}$. If RAIC only holds over $\mathcal{U}$ contained in the sphere $\|\boldsymbol{x}\|_2 S^{n-1}$, then under $\mathcal{U} \supset \|\boldsymbol{x}\|_2 S^{n-1} \cap \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$ one can instead use the normalized PGD

$$\boldsymbol{x}_{t+1} = \frac{\|\boldsymbol{x}\|_2 \cdot P_{\mathcal{K}}(\boldsymbol{x}_t - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t))}{\|P_{\mathcal{K}}(\boldsymbol{x}_t - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t))\|_2}, \quad t = 0, 1, 2, \cdots$$

with $x_0 \in \mathcal{U} \supset \|\boldsymbol{x}\|_2 S^{n-1} \cap \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$. See Case (ii) of Theorem 3.1 in Chen et al. (2025). ◊

In the case when $\mathcal{K}$ is a convex set, we have the following.

Theorem 3.2. *Assume that $\mathcal{K}$ is a convex set containing $\boldsymbol{x}$, and $\phi > 0$. If*

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}_\phi(\mathcal{K};\mathcal{U}, R_{\boldsymbol{x},\phi}(\boldsymbol{u}), \eta)$$

where $\mathcal{U} \supset \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$ for some $0 < d \leq \infty$ and $R_{\boldsymbol{x},\phi}(\boldsymbol{u}) = \mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2$ with $\mu_1 < \frac{1}{2}$ and $d > \frac{2\mu_2+\phi}{1-2\mu_1}$, then $\{\boldsymbol{x}_t\}_{t\geq 0}$ generated by (PGD) with $\boldsymbol{x}_0 \in \mathcal{K} \cap B_2^n(\boldsymbol{x}; d)$ satisfies

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq (2\mu_1)^t\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + \frac{2\mu_2 + \phi}{1-2\mu_1}, \quad \forall t \geq 0.$$

A key component to the proof is the following lemma which slightly tightens Corollary 8.3 in Plan et al. (2017) for the projection onto a convex set.

Lemma 3.1. *Let $\mathcal{K}$ be a convex set, $\boldsymbol{z} \in \mathcal{K}$ and $\boldsymbol{w} \in \mathbb{R}^n$. Then for every $t > 0$ we have*

$$\|P_{\mathcal{K}}(\boldsymbol{w}) - \boldsymbol{z}\|_2 \leq \max\left\{t, \frac{2}{t}\|\boldsymbol{w} - \boldsymbol{z}\|_{\mathcal{K}^\circ_{\boldsymbol{z},t}}\right\}.$$

Proof. This proof is omitted since it can be easily adapted from the argument in Plan et al. (2017) by noticing that $\mathcal{K} - \boldsymbol{x}$ is a star-shaped set. □

We now give the proof of Theorem 3.2.

Proof of Theorem 3.2. We define the sequence $\{f_t\}_{t\geq 0}$ by $f_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2$ and

$$f_{t+1} = 2\mu_1 f_t + 2\mu_2 + \phi, \quad t \geq 0.$$

It has a closed form expression

$$f_t = (2\mu_1)^t\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + (2\mu_2 + \phi)\frac{1 - (2\mu_1)^t}{1 - 2\mu_1} \leq (2\mu_1)^t\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + \frac{2\mu_2 + \phi}{1 - 2\mu_1}, \tag{3}$$

and thus it remains to prove $\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq f_t$ for all $t \geq 0$. We use induction to achieve this. The base case holds trivially. Suppose that $\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq f_t$, we seek to show $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|_2 \leq f_{t+1}$. From (3) and $d > \frac{2\mu_2+\phi}{1-2\mu_1}$, we reach $f_t \leq d$ for all $t \geq 0$, and thus $\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq d$. In view of (PGD) and $\boldsymbol{x}_0 \in \mathcal{K}$, we have $\boldsymbol{x}_t \in \mathcal{K}$. Taken collectively, $\boldsymbol{x}_t \in \mathcal{K} \cap B_2^n(\boldsymbol{x}; d) \subset \mathcal{U}$, and therefore we can use Lemma 3.1 and the RAIC to achieve

$$\begin{aligned}
&\|\boldsymbol{x}_{t+1} - \boldsymbol{x}\|_2 = \|P_{\mathcal{K}}(\boldsymbol{x}_t - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t)) - \boldsymbol{x}\|_2 \\
&\leq \max\left\{\phi, \frac{2}{\phi}\|\boldsymbol{x}_t - \boldsymbol{x} - \eta \cdot \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t)\|_{\mathcal{K}^\circ_{\boldsymbol{z},\phi}}\right\} \\
&\leq \max\{\phi, 2\mu_1\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 + 2\mu_2\} \\
&\leq 2\mu_1 f_t + 2\mu_2 + \phi = f_{t+1},
\end{aligned}$$

completing the proof. □

Remark 3.2. Theorem 3.2 shows the convergence of PGD to an error of $O(\mu_2 + \phi)$, with μ_2 and ϕ in equal footing. By Remark 2.1, one can think of $\mu_2 = \mu_2(\phi)$ as a non-decreasing function of ϕ;[2] therefore, one needs to deal with a tradeoff between μ_2 and ϕ in order to derive the minimal error. Specifically, to attain the best possible recovery error rate, the principle is to choose ϕ such that $\phi \asymp \mu_2$. ◊

Unified approach to uniform recovery. We are now ready to propose a unified approach to uniform signal recovery from nonlinear observations. The idea is simple: given Theorems 3.1 and 3.2 which show that RAIC of $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ implies the nonuniform recovery of the fixed $\boldsymbol{x}$ via PGD, one only needs to prove the RAIC of $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ for all $\boldsymbol{x} \in \mathcal{X}$ — referred to as a uniform RAIC — to establish the uniform recovery of $\boldsymbol{x} \in \mathcal{X}$ via PGD.

Note that our approach indeed goes beyond the specific observations in (1). To the end of uniform recovery of $\boldsymbol{x} \in \mathcal{X}$ from $\{D_{\boldsymbol{x}}\}_{\boldsymbol{x}\in\mathcal{X}}$, where $D_{\boldsymbol{x}}$ denotes what we can access for the estimation of $\boldsymbol{x}$, our approach consists of only two steps:

[2]Strictly, under $R_{\boldsymbol{x},\phi}(u) = \mu_1(\phi)\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2(\phi)$ and $\phi' \geq \phi$, $R_{\boldsymbol{x},\phi}(x) \geq R_{\boldsymbol{x},\phi'}(x)$ from Remark 2.1 yields $\mu_2(\phi) \geq \mu_2(\phi')$.

1. Construct the gradients $\boldsymbol{h}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$ from $D_{\boldsymbol{x}}$, for all $\boldsymbol{x} \in \mathcal{X}$.

2. Establish the (uniform) RAIC of $\boldsymbol{h}_{\boldsymbol{x}}$, for all $\boldsymbol{x} \in \mathcal{X}$.

The first step is typically model-specific, and a common practice is to choose $\boldsymbol{h}_{\boldsymbol{x}}$ as the (sub)gradient of some loss function. The second step is the main challenge and reduces to bounding an empirical process under random data. We have the following two theorems, which are proved by applying Theorems 3.1 and 3.2 to every $\boldsymbol{x} \in \mathcal{X}$, respectively. We omit the detailed proofs.

Theorem 3.3 ($\mathcal{K}$ is a cone)**.** *Suppose that we recover $\boldsymbol{x} \in \mathcal{X}$ by running (PGD) with some cone $\mathcal{K}$ satisfying $\mathcal{K} \supset \mathcal{X}$, which produces $\{\boldsymbol{x}_t\}_{t\geq 0}$. If there exist positive numbers $\mu_1, \mu_2, d_{\boldsymbol{x}}$ and sets $\mathcal{U}_{\boldsymbol{x}}$ such that*

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\big(\mathcal{K}; \mathcal{U}_{\boldsymbol{x}}, \mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2, \eta\big), \tag{4}$$

$$\mu_1 < \frac{1}{2}, \tag{5}$$

$$\mathcal{U}_{\boldsymbol{x}} \supset \mathcal{K} \cap B_2^n(\boldsymbol{x}; d_{\boldsymbol{x}}) \quad \text{for some } d_{\boldsymbol{x}} \in \Big(\frac{2\mu_2}{1-2\mu_1}, \infty\Big], \tag{6}$$

$$\boldsymbol{x}_0 \in \mathcal{K} \cap B_2^n(\boldsymbol{x}; d_{\boldsymbol{x}}) \tag{7}$$

hold for all $\boldsymbol{x} \in \mathcal{X}$, then

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq (2\mu_1)^t \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + \frac{2\mu_2}{1-2\mu_1}, \quad \forall t \geq 0, \quad \boldsymbol{x} \in \mathcal{X}. \tag{8}$$

Theorem 3.4 ($\mathcal{K}$ is a convex set)**.** *Suppose that we recover $\boldsymbol{x} \in \mathcal{X}$ by running (PGD) with some convex set $\mathcal{K}$ satisfying $\mathcal{K} \supset \mathcal{X}$, which produces $\{\boldsymbol{x}_t\}_{t\geq 0}$. If there exist positive numbers $\phi, \mu_1, \mu_2, d_{\boldsymbol{x}}$ and sets $\mathcal{U}_{\boldsymbol{x}}$ such that*

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}_{\phi}(\mathcal{K}; \mathcal{U}, \mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2, \eta), \tag{9}$$

$$\mu_1 < \frac{1}{2}, \tag{10}$$

$$\mathcal{U}_{\boldsymbol{x}} \supset \mathcal{K} \cap B_2^n(\boldsymbol{x}; d_{\boldsymbol{x}}) \quad \text{for some } d_{\boldsymbol{x}} \in \Big(\frac{2\mu_2+\phi}{1-2\mu_1}, \infty\Big], \tag{11}$$

$$\boldsymbol{x}_0 \in \mathcal{K} \cap B_2^n(\boldsymbol{x}; d_{\boldsymbol{x}}) \tag{12}$$

hold for all $\boldsymbol{x} \in \mathcal{X}$, then

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq (2\mu_1)^t \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + \frac{2\mu_2+\phi}{1-2\mu_1}, \quad \forall t \geq 0, \ \boldsymbol{x} \in \mathcal{X}. \tag{13}$$

4 Single-Index Model

As an application, we consider the uniform recovery of $\boldsymbol{x} \in \mathcal{X}$ from $\{y_i = f_i(\boldsymbol{a}_i^T\boldsymbol{x})\}_{i=1}^m$ as in (1). We will treat a fairly large class of f_i, which can be unknown, discontinuous or random, by an approach similar to Plan and Vershynin (2016); Genzel and Stollenwerk (2023); Oymak and Soltanolkotabi (2017). We shall proceed with the unified approach consisting of two steps.

Step 1: Choose the gradient. For $\boldsymbol{x} \in \mathcal{X}$, we adopt the ℓ_2 loss

$$L_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \mu \boldsymbol{a}_i^T \boldsymbol{u})^2 \tag{14}$$

with a re-scaling parameter $\mu > 0$. Here, μ is a tuning parameter; its role is clear from an interesting perspective that *a (Gaussian) nonlinear measurement can be viewed as a noisy linear measurement on a rescaled parameter*; see, e.g., Plan and Vershynin (2016); Thrampoulidis et al. (2015); Plan et al. (2017). Therefore, we take the gradient

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) = \nabla L_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^{m}(\mu \boldsymbol{a}_i^T \boldsymbol{u} - y_i)\mu \boldsymbol{a}_i. \tag{15}$$

We emphasize that $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ depends on $\boldsymbol{x}$ through $y_i = f_i(\boldsymbol{a}_i^T \boldsymbol{x})$.

Step 2: Prove the uniform RAIC. We shall discuss two cases:

- If $\mathcal{K}$ is a cone, then we seek to prove

  $$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\big(\mathcal{K}; \mathcal{U} := \mathcal{K}, \mu_1\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2, \eta := \mu^{-2}\big), \quad \forall \boldsymbol{x} \in \mathcal{X},$$

  namely

  $$\left\|\boldsymbol{u} - \boldsymbol{x} - \frac{1}{m}\sum_{i=1}^{m}\Big(\boldsymbol{a}_i^T \boldsymbol{u} - \frac{f_i(\boldsymbol{a}_i^T \boldsymbol{x})}{\mu}\Big)\boldsymbol{a}_i\right\|_{\mathcal{K}_1^\circ} \le \mu_1\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2, \quad \forall \boldsymbol{u} \in \mathcal{K}, \boldsymbol{x} \in \mathcal{X} \tag{16}$$

  with $\mu_1 < \frac{1}{2}$. Notice that (6)–(7) hold with $d_{\boldsymbol{x}} = \infty$ and $\boldsymbol{x}_0 = 0$. Therefore, by Theorem 3.3, (PGD) with $\boldsymbol{x}_0 = 0$ satisfies the uniform guarantee in (8).

- If instead $\mathcal{K}$ is a convex set, then for some $\phi > 0$, our goal is to prove

  $$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}_\phi\big(\mathcal{K}; \mathcal{U} := \mathcal{K}, \mu_1\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2, \eta := \mu^{-2}\big), \quad \forall \boldsymbol{x} \in \mathcal{X},$$

  that is,

  $$\frac{1}{\phi}\left\|\boldsymbol{u} - \boldsymbol{x} - \frac{1}{m}\sum_{i=1}^{m}\Big(\boldsymbol{a}_i^T \boldsymbol{u} - \frac{f_i(\boldsymbol{a}_i^T \boldsymbol{x})}{\mu}\Big)a_i\right\|_{\mathcal{K}_{\boldsymbol{x},\phi}^\circ} \le \mu_1\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2, \quad \forall \boldsymbol{u} \in \mathcal{K}, \boldsymbol{x} \in \mathcal{X}. \tag{17}$$

  with some $\mu_1 < \frac{1}{2}$. Since (11)–(12) hold with $d_{\boldsymbol{x}} = \infty$ and an arbitrary $\boldsymbol{x}_0$ in $\mathcal{K}$, Theorem 3.4 guarantees that (PGD) with any $\boldsymbol{x}_0 \in \mathcal{K}$ enjoys the uniform guarantee in (13).

At present, the conditions are deterministic. To proceed, a number of assumptions are needed to establish (16) and (17).

4.1 Assumptions

We work with Gaussian design and f_i that are either deterministic or independent across $i \in [m]$.

Assumption 4.1. The sensing vectors $\boldsymbol{a}_i$ are i.i.d. $N(0, \boldsymbol{I}_d)$ vectors. The nonlinear transforms f_i are either deterministic or independent across $i \in [m]$ (and also independent of $\boldsymbol{a}_i$).

Remark 4.1. Under some f_i, such as $f_i = \mathrm{Id} + \epsilon_i$ (i.e., noisy linear measurements (Raskutti et al., 2011)) and $f_i = \mathrm{sign}(\mathrm{Id} + \tau_i)$, $\tau_i \sim \mathrm{Uniform}[-\lambda, \lambda]$ (i.e., uniformly dithered 1-bit measurements (Dirksen and Mendelson, 2021)), $\boldsymbol{a}_i$ can be general isotropic sub-Gaussian vectors. The generality of sub-Gaussian $\boldsymbol{a}_i$, however, is not of interest to this work: we simply treat Gaussian $\boldsymbol{a}_i$. $\lozenge$

Similarly to Definition 1 of Genzel and Stollenwerk (2023), we introduce a model mismatch term

$$\rho(\boldsymbol{x}) := \left| \mathbb{E}_{g\sim N(0,1)} \left[\frac{f_i(\|\boldsymbol{x}\|_2 g) g}{\mu} \right] - \|\boldsymbol{x}\|_2 \right|, \quad \boldsymbol{x} \in \mathcal{X}. \tag{18}$$

Remark 4.2. In general, $\rho(\boldsymbol{x})$ is a contributor to the final error (since it appears in the approximation error of the RAIC in Theorems 4.1 and 4.2). As such, it is implicitly required that

$$\mathbb{E}[f_i(\|\boldsymbol{x}\|_2 g) g] \neq 0; \tag{19}$$

otherwise, $\rho(\boldsymbol{x}) \geq \|\boldsymbol{x}\|_2$ and error below $O(\|\boldsymbol{x}\|_2)$ cannot be achieved. As a consequence, the development in this section excludes even link functions f_i that satisfy $\mathbb{E}[f_i(\|\boldsymbol{x}\|_2 g) g] = 0$, thus excluding phase retrieval with $f_i = |\mathrm{Id}|$ or $|\mathrm{Id}|^2$. However, this does not mean that our unified approach does not apply to phase retrieval — indeed, it still works, see Section 5 and Chen et al. (2025). Indeed, this simply means that the gradient chosen in (15) is incompatible with phase retrieval.

On the other hand, once (19) holds, it is obvious that one can choose $\mu = \frac{\mathbb{E}[f_i(\|\boldsymbol{x}\|_2 g) g]}{\|\boldsymbol{x}\|_2}$ to render $\rho(\boldsymbol{x}) = 0$. Thus, if $\mathcal{X}$ is a subset of a sphere, say $\mathcal{X} \subset \lambda_0 S^{n-1}$ for some $\lambda_0 > 0$, then the universal choice $\mu = \frac{\mathbb{E}[f_i(\lambda_0 g) g]}{\lambda_0}$ zeroes $\rho(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. To fix idea, we shall restrict our attention to this case via Assumption 4.3 later on. $\lozenge$

To preclude pathological nonlinearities, we impose the following regularity conditions. We define

$$\tilde{f}_i := \mathrm{Id} - \frac{f_i}{\mu}.$$

We also let $\mathcal{D}_{f_i}$ be the points of discontinuity of f_i.

Assumption 4.2. For some constants $\varphi_1, \varphi_2, \varphi_3, \varphi_4, \varphi_5 > 0$, the following conditions are satisfied:

(C1) (*Sub-Gaussianity*) Let $g \sim N(0,1)$,

$$\|\tilde{f}_i(\|\boldsymbol{x}\|_2 g)\|_{\psi_2} \leq \varphi_1, \quad \forall \boldsymbol{x} \in \mathcal{X};$$

(C2) (*Separation of discontinuities*) If $|\mathcal{D}_{f_i}| > 1$, there exists some $\varphi_2 > 0$ such that

$$|\xi_1 - \xi_2| \geq \varphi_2, \quad \forall\, \xi_1, \xi_2 \in \mathcal{D}_{f_i} \ \text{ obeying } \ \xi_1 \neq \xi_2.$$

If $|\mathcal{D}_{f_i}| \leq 1$, we let $\varphi_2 := \infty$ and follow the convention $\frac{a}{\infty} = 0 \ \ (\forall a \in \mathbb{R})$.

(C3) (*Discontinuity with bounded jump height*) The discontinuity of f_i, if exists, is either removable discontinuity or jump discontinuity with bounded jump height:

$$\left| \lim_{a \to \xi^+} f_i(a) - \lim_{a \to \xi^-} f_i(a) \right| \leq \varphi_3, \quad \forall\, \xi \in \mathcal{D}_{f_i}.$$

We also assume $f_i(a) = \lim_{a \to \xi^-} f_i(a)$ with no loss of generality.

(C4) (*Piecewise Lipschitzness*) $\tilde{f}_i$ is φ_4-Lipschitz continuous over (b_1, b_2) for any $-\infty \le b_1 < b_2 \le \infty$ such that $(b_1, b_2) \cap \mathcal{D}_{f_i} = \varnothing$.

(C5) (*Small-ball probability*) Let $\mathrm{dist}(a, \mathcal{D}) := \inf_{\xi \in \mathcal{D}} |a - \xi|$. It holds for any $\boldsymbol{x} \in \mathcal{X}$ and $t \in (0, \frac{\varphi_2}{4})$ that

$$\mathbb{P}\big(\mathrm{dist}(\boldsymbol{a}_i^T \boldsymbol{x}, \mathcal{D}_{f_i}) \le t\big) \le \varphi_5 t.$$

Remark 4.3. In view of $\tilde{f}_i = \mathrm{Id} - \frac{f_i}{\mu}$, the above (C1) and (C4) imply the sub-Gaussianity of $f_i(\|\boldsymbol{x}\|_2 g)$ and piecewise Lipschitzness of f_i, vice versa. Since the two conditions enter our analysis in bounding the empirical process $\sup_{\boldsymbol{x},\boldsymbol{q}} \frac{1}{m} \sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T \boldsymbol{x}) \boldsymbol{a}_i^T \boldsymbol{q}$, they are enforced on $\tilde{f}_i$ for technical convenience. ♢

Remark 4.4. We note that (C2)–(C4) jointly imply

$$|\tilde{f}_i(b_1) - \tilde{f}_i(b_2)| \le \varphi_4 |b_1 - b_2| + \left(\frac{|b_1 - b_2|}{\varphi_2} + 1 \right) \frac{\varphi_3}{\mu}, \quad \forall\, b_1, b_2 \in \mathbb{R}, \tag{20}$$

since the number of discontinuities in $[\min\{b_1, b_2\}, \max\{b_1, b_2\}]$ is bounded by $\lfloor \frac{|b_1 - b_2|}{\varphi_2} \rfloor + 1$, and the jump heights of $\tilde{f}_i$ are bounded by $\frac{\varphi_3}{\mu}$. ♢

4.2 Uniform RAIC

We now establish the RAIC in (16) and (17). We first decompose the left-hand side into two terms and then bound them separately. For instance, in the conic case,

$$\begin{aligned} &\left\| \boldsymbol{u} - \boldsymbol{x} - \frac{1}{m} \sum_{i=1}^m \Big(\boldsymbol{a}_i^T \boldsymbol{u} - \frac{f_i(\boldsymbol{a}_i^T \boldsymbol{x})}{\mu} \Big) \boldsymbol{a}_i \right\|_{\mathcal{K}_1^\circ} \\ &\le \left\| \boldsymbol{u} - \boldsymbol{x} - \frac{1}{m} \sum_{i=1}^m \boldsymbol{a}_i \boldsymbol{a}_i^T (\boldsymbol{u} - \boldsymbol{x}) \right\|_{\mathcal{K}_1^\circ} + \left\| \frac{1}{m} \sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T \boldsymbol{x}) \boldsymbol{a}_i \right\|_{\mathcal{K}_1^\circ}. \end{aligned} \tag{21}$$

We then seek to bound the two terms in (21) uniformly for all $(\boldsymbol{u}, \boldsymbol{x}) \in \mathcal{K} \times \mathcal{X}$.

The first term is easy to control and in fact amounts to showing a RIP. It is much harder to uniformly control the second term due to the nonlinearity $f_i(\boldsymbol{a}_i^T \boldsymbol{x})$. To this end, we rely on a technically involved covering argument, whose ideas are, however, largely adapted from existing works of binary embedding (e.g., Plan and Vershynin (2014); Oymak and Recht (2015); Dirksen and Mendelson (2021)) and nonlinear compressed sensing (e.g., Xu and Jacques (2020); Chen et al. (2024, 2026)).

Our results of the uniform RAIC are given in the following, with proofs postponed to Section A. The statements involve two complexity measures for a general set $\mathcal{U} \subset \mathbb{R}^n$ — the Gaussian width

$$\omega(\mathcal{U}) = \mathbb{E} \sup_{\boldsymbol{u} \in \mathcal{U}} \langle \boldsymbol{g}, \boldsymbol{u} \rangle, \quad \text{where } \ \boldsymbol{g} \sim N(0, \boldsymbol{I}_n),$$

and the metric entropy

$$\mathscr{H}(\mathcal{U}, \varepsilon) = \log \mathscr{N}(\mathcal{U}, \varepsilon)$$

where $\mathscr{N}(\mathcal{U}, \varepsilon)$ is the covering number under radius ε, i.e., the minimal number of radius-ε ℓ_2 balls needed to cover $\mathcal{U}$. See Vershynin (2018) for a fuller account.

Theorem 4.1 (Uniform RAIC for a cone)**.** *Assume that Assumptions 4.1, 4.2 hold, and that $\mathcal{K}$ is a cone containing $\mathcal{X}$. For any sufficiently small $\varepsilon, \zeta > 0$ such that $\varphi_5 \varepsilon \sqrt{\log(1/\zeta)}$ is small enough, if*

$$m \gtrsim \mathscr{H}(\mathcal{X}, \varepsilon) + \Big(1 + \frac{\varphi_5^2 \varepsilon^2}{\zeta}\Big)\omega^2(\mathcal{K}_1) \tag{22}$$

then with probability at least $1 - \exp(-c_1\omega^2(\mathcal{K}_1)) - \exp(-c_2\mathscr{H}(\mathcal{X}, \varepsilon))$, $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ in (15) satisfies

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\bigg(\mathcal{K}; \mathcal{K}, \frac{C\omega(\mathcal{K}_1)}{\sqrt{m}}\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2 + O\Big(\sup_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x})\Big), \mu^{-2}\bigg), \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where

$$\mu_2 \lesssim \varphi_1 \sqrt{\frac{\mathscr{H}(\mathcal{X}, \varepsilon) + \omega^2(\mathcal{K}_1)}{m}} + \varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\Xi.$$

Here, Ξ is defined by

$$\Xi := \sqrt{\overline{\Xi} + \zeta}\bigg(\frac{\omega(\mathcal{K}_1)}{\sqrt{m}} + \sqrt{\overline{\Xi}\log\Big(\frac{1}{\overline{\Xi}}\Big)} + \sqrt{\zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg) \tag{23}$$

$$+ \frac{\varepsilon}{\varphi_2}\bigg(\frac{\omega^2(\mathcal{K}_1)}{m} + \overline{\Xi}\log\Big(\frac{1}{\overline{\Xi}}\Big) + \zeta\log\Big(\frac{1}{\zeta}\Big)\bigg), \tag{24}$$

$$\text{where} \quad \overline{\Xi} := \frac{\mathscr{H}(\mathcal{X}, \varepsilon)}{m} + \frac{\varphi_5\varepsilon\omega(\mathcal{K}_1)}{\sqrt{\zeta m}} + \varphi_5\varepsilon\sqrt{\log(e/\zeta)} \tag{25}$$

Theorem 4.2 (Uniform RAIC for a convex set)**.** *Assume that Assumptions 4.1, 4.2 hold, and that $\mathcal{K}$ is a convex set containing $\mathcal{X}$. For any small enough $\phi, \varepsilon, \zeta > 0$, if*

$$m \gtrsim \frac{\omega^2(\mathcal{K}_{\mathcal{X},\phi})}{\phi^2} + \Big(\frac{1}{\varepsilon^2} + \frac{\varphi_5^2}{\zeta}\Big)\omega^2(\mathcal{X}_\varepsilon) + \mathscr{H}(\mathcal{X}, \varepsilon), \tag{26}$$

then with probability at least $1 - \exp(-c_1\phi^{-2}\omega^2(\mathcal{K}_{\mathcal{X},\phi}))$, $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$ in (15) satisfies

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}_\phi\bigg(\mathcal{K}; \mathcal{K}, \frac{C\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}}\|\boldsymbol{u} - \boldsymbol{x}\|_2 + \mu_2 + O\Big(\sup_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x})\Big), \mu^{-2}\bigg), \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where

$$\mu_2 \lesssim \varphi_1 \sqrt{\frac{\mathscr{H}(\mathcal{X}, \varepsilon)}{m}} + \Big(1 + \frac{\varphi_1}{\phi}\Big)\frac{\omega(\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \varepsilon\varphi_4 + \sup_{\boldsymbol{x} \in \mathcal{X}} \rho(\boldsymbol{x}) + \frac{\varphi_3}{\mu}\Upsilon.$$

Here, Υ is defined by

$$\Upsilon := \frac{\varepsilon}{\varphi_2}\bigg(\frac{\omega(\varepsilon^{-1}\mathcal{X}_\varepsilon)}{\sqrt{m}} + \sqrt{\overline{\Upsilon}\log\Big(\frac{1}{\overline{\Upsilon}}\Big) + \zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg)\bigg(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \sqrt{\overline{\Upsilon}\log\Big(\frac{1}{\overline{\Upsilon}}\Big) + \zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg) \tag{27}$$

$$+ \sqrt{\overline{\Upsilon} + \zeta}\bigg(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \sqrt{\overline{\Upsilon}\log\Big(\frac{1}{\overline{\Upsilon}}\Big) + \zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg) \tag{28}$$

$$\text{where} \quad \overline{\Upsilon} := \frac{\mathscr{H}(\mathcal{X}, \varepsilon)}{m} + \frac{\varphi_5\omega(\mathcal{X}_\varepsilon)}{\sqrt{\zeta m}} + \varphi_5\varepsilon\sqrt{\log(e/\zeta)}. \tag{29}$$

4.3 General Uniform Recovery Guarantees

Under Assumptions 4.1 and 4.2, our goal is to uniformly recovery all $\boldsymbol{x} \in \mathcal{X}$ from $\{y_i = f_i(\boldsymbol{a}_i^T \boldsymbol{x})\}_{i=1}^m$ using (PGD). With the gradient chosen as in (15), the step size $\eta = \mu^{-2}$, and $\boldsymbol{x}_0$ be a point in $\mathcal{K}$, the algorithm reads

$$\boldsymbol{x}_{t+1} = P_{\mathcal{K}}\bigg(\boldsymbol{x}_t - \frac{1}{m}\sum_{i=1}^m \big(\boldsymbol{a}_i^T \boldsymbol{x}_t - \frac{y_i}{\mu}\big)\boldsymbol{a}_i\bigg), \quad t = 0, 1, \cdots. \tag{30}$$

We further make the following assumption.

Assumption 4.3. $\mathcal{X} \subset \mathbb{S}^{n-1}$ and

$$\mathbb{E}_{g\sim N(0,1)}[f_i(g)g] \neq 0.$$

Under this assumption, we shall set

$$\mu := \mathbb{E}_{g\sim N(0,1)}[f_i(g)g] \tag{31}$$

so that $\rho(\boldsymbol{x}) = 0$ holds for all $\boldsymbol{x} \in \mathcal{X}$.

We now present the uniform recovery guarantees which, under the unified framework introduced in Section 3, are direct outcomes of the uniform RAIC in Theorems 4.1 and 4.2.

Theorem 4.3 (PGD with cone $\mathcal{K}$)**.** *Assume that Assumptions 4.1, 4.2, 4.3 hold, and that $\mathcal{K}$ is a cone containing $\mathcal{X}$. For any sufficiently small $\varepsilon, \zeta > 0$, if (22) holds, then with probability at least $1 - \exp(-c_1\omega^2(\mathcal{K}_1)) - \exp(-c_2\mathscr{H}(\mathcal{X}, \varepsilon))$, for all $\boldsymbol{x} \in \mathcal{X}$, the sequence $\{\boldsymbol{x}_t\}_{t\geq 0}$ generated by (30) with some $\boldsymbol{x}_0 = 0$ satisfies*

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq \left(\frac{C_1\omega(\mathcal{K}_1)}{\sqrt{m}}\right)^t \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + C_2\varphi_1\sqrt{\frac{\mathscr{H}(\mathcal{X}, \varepsilon) + \omega^2(\mathcal{K}_1)}{m}} + C_3\varepsilon\varphi_4 + \frac{C_5\varphi_3}{\mu}\Xi$$

for any integer $t \geq 0$, where Ξ is defined in (23)–(25).

Proof. The result follows from invoking Theorem 4.1 to establish the uniform RAIC (for all $\boldsymbol{x} \in \mathcal{X}$) and then applying Theorem 3.3. □

Theorem 4.4 (PGD with convex set $\mathcal{K}$)**.** *Assume that Assumptions 4.1, 4.2, 4.3 hold, and that $\mathcal{K}$ is a convex set containing $\mathcal{X}$. For any sufficiently small $\phi, \varepsilon, \zeta > 0$, if (26) holds, then with probability at least $1 - \exp(-c_1\phi^{-2}\omega^2(\mathcal{K}_{\mathcal{X},\phi}))$, for all $\boldsymbol{x} \in \mathcal{X}$, the sequence $\{\boldsymbol{x}_t\}_{t\geq 0}$ generated by (30) with some $x_0 \in \mathcal{K}$ satisfies*

$$\begin{aligned}\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq \left(\frac{C_1\omega(\mathcal{K}_{\mathcal{X},\phi})}{\phi\sqrt{m}}\right)^t \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + C_2\varphi_1\sqrt{\frac{\mathscr{H}(\mathcal{X}, \varepsilon)}{m}} \\ + \frac{C_3\varphi_1\omega(\mathcal{K}_{\mathcal{X},\phi})}{\phi\sqrt{m}} + 2\phi + C_4\varepsilon\varphi_4 + \frac{C_5\varphi_3}{\mu}\Upsilon, \qquad t = 0, 1, 2, ...,\end{aligned} \tag{32}$$

where Υ is defined in (27)–(29).

Proof. The result follows from invoking Theorem 4.2 to establish the uniform RAIC (for all $\boldsymbol{x} \in \mathcal{X}$) and then applying Theorem 3.4. □

To put our results in perspective, we shall provide detailed comparisons to prior works.

Remark 4.5 (Comparison to nonuniform guarantees of Plan and Vershynin (2016); Oymak and Soltanolkotabi (2017); Cost of uniformity; and Cost of discontinuity). The nonuniform recovery guarantee in Theorem 1.9 of Plan and Vershynin (2016), when adapted to our setting with $\mu \asymp 1$, states the following: under Assumptions 4.1, 4.3 and $\|g - \frac{f_i(g)}{\mu}\|_{\psi_2} \leq \varphi_1$ with $g \sim N(0,1)$,[3] for a fixed $\boldsymbol{x} \in \mathcal{K} \cap \mathbb{S}^{n-1}$ and any small enough $\phi > 0$, if

$$m \gtrsim \frac{\omega^2(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi^2}, \tag{33}$$

then the convex program *generalized Lasso*[4]

$$\hat{\boldsymbol{x}}_{GLasso} = \arg\min_{u \in \mathcal{K}} \frac{1}{2m} \sum_{i=1}^{m} (y_i - \mu \boldsymbol{a}_i^T \boldsymbol{u})^2 \tag{34}$$

satisfies the high-probability error bound

$$\|\hat{\boldsymbol{x}}_{GLasso} - \boldsymbol{x}\|_2 \lesssim \frac{\varphi_1 \omega(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi\sqrt{m}} + \phi. \tag{35}$$

This result can be closely compared to our Theorem 4.4 concerning with *PGD in (30) with convex set* $\mathcal{K}$, which, indeed, can be viewed as a computational procedure for solving (34). The main difference is that our guarantee is uniform for all $\boldsymbol{x} \in \mathcal{X}$, under the additional Assumption 4.2.

Toward a comparison, we observe that our Theorem 4.4 with $\mathcal{X} = \{\boldsymbol{x}\}$ for a fixed $\boldsymbol{x} \in \mathcal{K} \cap \mathbb{S}^{n-1}$ — which concerns the nonuniform recovery of $\boldsymbol{x}$ — is consistent with the result in Plan and Vershynin (2016): since $\mathcal{X}$ is now a singleton, we have $\mathscr{H}(\mathcal{X}, \varepsilon) = 0$ $(\forall \varepsilon > 0)$ and $\omega(\mathcal{X}_\varepsilon) = 0$, and thus the sample complexity (26) reduces to (33); by further working with $\varepsilon \to 0$, Υ in (27)–(29) simplifies to

$$\Upsilon = \frac{\sqrt{\zeta} \cdot \omega(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi\sqrt{m}} + \zeta\sqrt{\log(1/\zeta)}$$

and thus also vanishes by taking $\zeta \to 0$; in turn, when t is sufficiently large, the guarantee in (32) reduces to

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \lesssim \frac{\varphi_1 \omega(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi\sqrt{m}} + \phi,$$

which is exactly identical to (35). As such, our Theorem 4.4 can be viewed as an extension of Plan and Vershynin (2016), recovering their result when setting $\mathcal{X}$ as a singleton.

When $\mathcal{X}$ is not a singleton, under sufficiently large t, Theorem 4.4 gives the uniform error bound

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \lesssim \frac{\varphi_1 \omega(\mathcal{K}_{\mathcal{X},\phi})}{\phi\sqrt{m}} + \phi + \varphi_1 \sqrt{\frac{\mathscr{H}(\mathcal{X}, \varepsilon)}{m}} + \varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\Upsilon.$$

[3] This is exactly (C1) in our Assumption 4.2 when $\mathcal{X} \subset \mathbb{S}^{n-1}$. However, this condition is not needed in Plan and Vershynin (2016); instead, their treatment is more general and relies on two parameters — $\hat{\sigma}^2 := \mathbb{E}(f_i(g) - \mu g)^2$ and $\hat{\eta}^2 := \mathbb{E}[(f(g) - \mu g)^2 g^2]$ — to achieve concentration; see Equation (I.5) therein. In the comparison, we stick with $\|g - \frac{f_i(g)}{\mu}\|_{\psi_2}$ and $\mu \asymp 1$, under which the two parameters are bounded by $\hat{\sigma}, \hat{\eta} \lesssim \varphi_1$.

[4] In Plan and Vershynin (2016), the authors assumed $\mu\boldsymbol{x} \in \mathcal{K}$ and used the regular ℓ_2-loss $\frac{1}{2m}\sum_{i=1}^m (y_i - \boldsymbol{a}_i^T \boldsymbol{u})^2$ in (34), which is slightly different from but technically equivalent to $\boldsymbol{x} \in \mathcal{K}$ and the loss (14) adopted here.

Therefore, a useful perspective is to interpret the additional terms

$$\varphi_1\sqrt{\frac{\mathscr{H}(\mathcal{X},\varepsilon)}{m}}+\varepsilon\varphi_4+\frac{\varphi_3}{\mu}\Upsilon$$

and the increased term $\frac{\varphi_1\omega(\mathcal{K}_{\mathcal{X},\phi})}{\phi\sqrt{m}}$ (compared to $\frac{\varphi_1\omega(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi\sqrt{m}}$ in (35)) as *the cost of the uniformity*.

Similarly, one can compare the nonuniform recovery guarantee due to Oymak and Soltanolkotabi (2017), who also considered PGD, with our Theorem 4.3. In this setting, the cost of uniformity is mainly captured by the additional terms [5]

$$\varphi_1\sqrt{\frac{\mathscr{H}(\mathcal{X},\varepsilon)}{m}}+\varepsilon\varphi_4+\frac{\varphi_3}{\mu}\Xi.$$

Moreover, we notice that f_i is Lipschitz continuous if and only if $\varphi_3=0$; see (C3) in Assumption 4.2. Therefore, the terms $\frac{\varphi_3}{\mu}\Xi$ in Theorem 4.3 and $\frac{\varphi_3}{\mu}\Upsilon$ in Theorem 4.4 can be further interpreted as *the cost of discontinuity* in uniform recovery.

Finally, the discontinuous f_i has only one discontinuity if and only if $\varphi_2=\infty$; see (C2) in Assumption 4.2. Therefore, the contributors to Ξ in Equation (24), namely

$$\frac{\varepsilon}{\varphi_2}\left(\frac{\omega^2(\mathcal{K}_1)}{m}+\overline{\Xi}\log\left(\frac{1}{\overline{\Xi}}\right)+\zeta\log\left(\frac{1}{\zeta}\right)\right),$$

and the contributors to Υ in Equation (27), that is

$$\frac{\varepsilon}{\varphi_2}\left(\frac{\omega(\varepsilon^{-1}\mathcal{X}_\varepsilon)}{\sqrt{m}}+\sqrt{\overline{\Upsilon}\log\left(\frac{1}{\overline{\Upsilon}}\right)+\zeta\log\left(\frac{1}{\zeta}\right)}\right)\left(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}}+\sqrt{\overline{\Upsilon}\log\left(\frac{1}{\overline{\Upsilon}}\right)+\zeta\log\left(\frac{1}{\zeta}\right)}\right),$$

capture *the cost of multiple points of discontinuity* in uniform recovery. ◊

In the following, we discuss the explicit error decay rates in some canonical examples. Our goal is to identify how much the uniform recovery error rate is worse than the nonuniform recovery error rate. We shall pause to distinguish two types of $\mathcal{X}$ based on the dependence of $\mathscr{H}(\mathcal{X},\varepsilon)$ on ε:

1. *$\mathcal{X}$ is a structured set.* The first type of $\mathcal{X}$ is the so-called structured set, whose defining feature is that $\mathscr{H}(\mathcal{X},\varepsilon)$ only logarithmically depends on ε. Canonical examples include $\Sigma_k^{n,*}$ and $M_r^{n_1,n_2,*}$, in light of their well-known covering number bounds (see, e.g., Plan and Vershynin (2012); Candes and Plan (2011)).

   Definition 4.1 (e.g., Xu and Jacques (2020); Oymak and Recht (2015); Jacques (2017); Chen and Ng (2023); Chen et al. (2024))**.** We say $\mathcal{X}$ is a structured set if

   $$\mathscr{H}(\mathcal{X},\varepsilon)\le C\omega^2(\mathcal{X})\log(1+\varepsilon^{-1}),\quad\forall\varepsilon>0 \tag{36}$$

   holds for some absolute constant C.

[5] As a passing note, we mention that Theorem 4.3 does not have a counterpart for generalized Lasso (34) and is unique to PGD: while projection onto cone $\mathcal{K}$ is computationally tractable for important instances such as $\mathcal{K}=\Sigma_k^n$, $M_r^{n_1,n_2}$, (34) with a low-complexity cone $\mathcal{K}$ is in general computationally intractable. This computational advantage of projected gradient descent (over the corresponding convex program) has been pointed out in Soltanolkotabi (2019); Oymak and Soltanolkotabi (2017).

Remark 4.6. For structured set $\mathcal{X}$ and any $\varepsilon \in (0,1)$ and $\eta > 0$, the definition of covering number yields

$$\mathscr{N}\Big(\frac{\mathcal{X}_\varepsilon}{\varepsilon}, \eta\Big) \le \mathscr{N}\Big(\frac{\mathcal{X}-\mathcal{X}}{\varepsilon}, \eta\Big) = \mathscr{N}(\mathcal{X}-\mathcal{X}, \varepsilon\eta) \le \mathscr{N}\Big(\mathcal{X}, \frac{\varepsilon\eta}{2}\Big)^2$$

and hence

$$\mathscr{H}\Big(\frac{\mathcal{X}_\varepsilon}{\varepsilon}, \eta\Big) \le 2\mathscr{H}\Big(\mathcal{X}, \frac{\varepsilon\eta}{2}\Big) \lesssim \omega^2(\mathcal{X}) \log\Big(1 + \frac{1}{\varepsilon\eta}\Big).$$

Then, Dudley's inequality gives

$$\omega(\varepsilon^{-1}\mathcal{X}_\varepsilon) \lesssim \int_0^2 \sqrt{\omega^2(\mathcal{X}) \log\Big(1 + \frac{1}{\varepsilon\eta}\Big)}\, d\eta \lesssim \omega(\mathcal{X}) \log(1+\varepsilon^{-1}). \tag{37}$$

$\Diamond$

2. $\mathcal{X}$ *is a general set.* The second is the general setting where $\mathscr{H}(\mathcal{X}, r)$ can only be bounded via Sudakov's inequality

$$\mathscr{H}(\mathcal{X}, \varepsilon) \le \frac{C\omega^2(\mathcal{X})}{\varepsilon^2} \tag{38}$$

where C is an absolute constant, and no more information is available on $\mathscr{H}(\mathcal{X}, r)$. Here, the quadratic dependence is essentially worse then the logarithmic one in (36). Nonetheless, note that (38) turns out to be nearly tight for some examples of interest, such as the set of approximately k-sparse vectors $\sqrt{k}\mathbb{B}_1^n \cap \mathbb{S}^{n-1}$ (Plan et al., 2017).

We are now ready to examine several concrete settings and determine the uniform recovery error rates.

Concrete setting (a): $\mathcal{X}$ is a structured set and $\mathcal{K}$ is a cone. The canonical examples include

$$(\mathcal{X}, \mathcal{K}) = (\Sigma_k^{n,*}, \Sigma_k^n) \quad \text{and} \quad (\mathcal{X}, \mathcal{K}) = (M_r^{n_1,n_2,*}, M_r^{n_1,n_2}).$$

Remark 4.7 (Uniform recovery error rate in setting (a)). The nonuniform error rate for recovering a fixed x via PGD, which was established in Oymak and Soltanolkotabi (2017), reads

$$\|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 \lesssim \varphi_1 \sqrt{\frac{\omega^2(\mathcal{K}_1)}{m}}. \tag{39}$$

By Theorem 4.3, (36) and $\mathcal{X} \subset \mathcal{K}_1$, PGD achieves the uniform recovery error rate

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 \lesssim \varphi_1 \sqrt{\frac{\omega^2(\mathcal{K}_1)}{m}} + \varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\Xi \tag{40}$$

up to a factor of $\sqrt{\log(1+\varepsilon^{-1})}$, where, by substituting (36) into (23)–(25) and *ignoring log factors*,

$$\Xi := \Big(1 + \frac{\varepsilon}{\varphi_2}\Big)\Big(\frac{\omega^2(\mathcal{K}_1)}{m} + \frac{\varphi_5\varepsilon\omega(\mathcal{K}_1)}{\sqrt{\zeta m}} + \varphi_5\varepsilon + \zeta\Big). \tag{41}$$

Suppose that

$$\varphi_1 = \Theta(1),\ \varphi_2 = \Omega(1) \text{ and } \varphi_3, \varphi_4, \varphi_5 = O(1).$$

We then choose $\varepsilon = \zeta$ to be at a sufficiently small order to guarantee that the terms with ε, ζ or $\frac{\varepsilon}{\sqrt{\zeta}} = \sqrt{\varepsilon}$ as a leading factor are negligible. For instance, we may choose

$$\varepsilon = \zeta = \Big(\frac{\omega^2(\mathcal{K}_1)}{m}\Big)^{10};$$

here, then $\varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\Xi = \tilde{O}(\frac{\omega^2(\mathcal{K}_1)}{m})$ when hiding log factors, and therefore the uniform recovery error from (40) reads

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 = \tilde{O}\bigg(\varphi_1\sqrt{\frac{\omega^2(\mathcal{K}_1)}{m}}\bigg).$$

This is identical to the nonuniform one (39) up to log factors, indicating that the uniformity costs very little in this setting. $\lozenge$

For any $\mathcal{U}$ we let $\mathrm{cone}(\mathcal{U})$ be the minimal cone containing $\mathcal{U}$. We say that $\mathcal{K}_{\mathcal{X}}$ has a descent cone structure if $\mathrm{cone}(\mathcal{K}_{\mathcal{X}})$ remains low-complexity in terms of Gaussian width. Two canonical examples are the Lasso-type convex relaxations in sparse or low-rank recovery such that the true signals lie in the boundary of $\mathcal{K}$ (we shall let $c_* \in (0, 1]$ in the following):

1. (sparse recovery) For

$$\mathcal{X} = \Sigma_k^{n,*} \cap \{\boldsymbol{u} : \|\boldsymbol{u}\|_1 = c_*\sqrt{k}\} \quad \text{and} \quad \mathcal{K} = \mathbb{B}_1^n(c_*\sqrt{k}), \tag{42}$$

   we have that $\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2) \lesssim k\log\Big(\frac{en}{k}\Big)$ is at the same order as $\omega^2(\Sigma_k^{n,*})$.

2. (low-rank recovery) For

$$\mathcal{X} = M_r^{n_1,n_2,*} \cap \{\boldsymbol{X} : \|\boldsymbol{X}\|_{nu} = c_*\sqrt{r}\} \quad \text{and} \quad \mathcal{K} = \mathbb{B}_{nu}(c_*\sqrt{r}), \tag{43}$$

   we have that $\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_F) \lesssim r(n_1 + n_2)$ is at the same order as $\omega^2(M_r^{n_1,n_2,*})$.

See, e.g., Chandrasekaran et al. (2012); Tropp (2015).

We now proceed to the second setting.

Concrete setting (b): $\mathcal{X}$ is a structured set, $\mathcal{K}$ is a convex set, $\mathcal{K}_{\mathcal{X}}$ exhibits a descent cone structure. Canonical examples include the above (42) and (43). We point out that, by using signal-dependent $\mathcal{K}$, $\mathcal{X}$ can be enlarged to all the sparse vectors or low-rank matrices — for instance, (42) can be adapted to $\mathcal{X} = \Sigma_k^{n,*}$ by using $\mathcal{K} = \mathbb{B}_1^n(\|\boldsymbol{x}\|_1)$ for recovering $\boldsymbol{x}$.

Remark 4.8 (Uniform recovery error rate in setting (b)). For recovering a fixed $\boldsymbol{x} \in \mathcal{X}$, the nonuniform recovery error rate of (34) due to Plan and Vershynin (2016), as reviewed in Remark 4.5, is

$$\|\hat{\boldsymbol{x}}_{GLasso} - \boldsymbol{x}\|_2 \lesssim \inf_{\phi\in(0,1/2)} \frac{\varphi_1\omega(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi\sqrt{m}} + \phi \lesssim \frac{\varphi_1\omega(\mathrm{cone}(\mathcal{K}_{\boldsymbol{x}}) \cap \mathbb{B}_2^n)}{\sqrt{m}}, \tag{44}$$

where in the second inequality we choose $\phi = \varphi_1\frac{\omega(\mathrm{cone}(\mathcal{K}_{\boldsymbol{x}})\cap\mathbb{B}_2^n)}{\sqrt{m}}$ and notice

$$\omega\Big(\frac{\mathcal{K}_{\boldsymbol{x},\phi}}{\phi}\Big) \le \omega(\mathrm{cone}(\mathcal{K} - \boldsymbol{x}) \cap \mathbb{B}_2^n), \quad \forall\phi > 0.$$

For simplicity, we enforce a very mild assumption $\omega(\mathcal{X}) \lesssim \omega(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n)$, which holds, e.g., when $0 \in \mathcal{K}$. By Theorem 4.4, (36), and $\omega(\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}) \le \omega(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n)$,[6] we reach the following: PGD in (30) achieves the uniform recovery error

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 \lesssim \frac{\varphi_1 \omega(\mathrm{cone}(\mathcal{K}_{\mathcal{X}} \cap \mathbb{B}_2^n))}{\sqrt{m}} + \phi + \varphi_4 \varepsilon + \frac{\varphi_3}{\mu} \Upsilon$$

up to a factor of $\sqrt{\log(1+\varepsilon^{-1})}$; moreover, by substituting (36), (37), and $\omega(\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}) \le \omega(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n)$ into (27)–(29) and ignoring log factors, the "cost of discontinuity" Υ is bounded by

$$\Upsilon \lesssim \Big(1 + \frac{\varepsilon}{\varphi_2}\Big)\bigg(\frac{\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n)}{m} + \frac{\varphi_5 \varepsilon \omega(\mathcal{X})}{\sqrt{\zeta m}} + \zeta + \varphi_5 \varepsilon\bigg).$$

Suppose that $\varphi_1 \asymp 1$, $\varphi_2 = \Omega(1)$ and $\varphi_3, \varphi_4, \varphi_5 = O(1)$. Then we can set $\phi = \varepsilon = \zeta$ at a sufficiently small scaling, say,

$$\Big(\frac{\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n)}{m}\Big)^{10},$$

to guarantee that $\phi + \varphi_4 \varepsilon + \frac{\varphi_3}{\mu}\Upsilon$ is dominated by $\frac{\varphi_1 \omega(\mathrm{cone}(\mathcal{K}_{\mathcal{X}} \cap \mathbb{B}_2^n))}{\sqrt{m}}$. Therefore, we obtain a uniform recovery error rate

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 = \tilde{O}\bigg(\varphi_1 \sqrt{\frac{\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n)}{m}}\bigg)$$

that is identical to (44) up to log factors as long as

$$\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n) = \tilde{O}(\omega^2(\mathrm{cone}(\mathcal{K}_{\boldsymbol{x}}) \cap \mathbb{B}_2^n)),$$

which is satisfied by (42) and (43). ◊

Concrete setting (c): General $\mathcal{X}$, convex set $\mathcal{K}$, and general $\mathcal{K}_{\mathcal{X}}$. We now drop the structured set assumption on $\mathcal{X}$ and the descent cone structure assumption on $\mathcal{K}_{\mathcal{X}}$. This setting is of interest because the "non-structured" $\mathcal{X}$ can be a much larger set — such as the set of approximately k-sparse vectors $\mathcal{X} = \mathbb{B}_1(\sqrt{k}) \cap \mathbb{S}^{n-1}$ and the set of approximately rank-r matrices $\mathcal{X} = \mathbb{B}_{nu}(\sqrt{r}) \cap \mathbb{S}_F$ — for which we shall choose $\mathcal{K} = \mathbb{B}_1(\sqrt{k})$ and $\mathcal{K} = \mathbb{B}_{nu}(\sqrt{r})$, respectively. In this setting, we only assume that $\mathcal{K}$ is convex so that (34) can be solved in polynomial time.

Remark 4.9 (Uniform recovery error rate in setting (c)). For recovering a fixed $\boldsymbol{x} \in \mathcal{X}$, the nonuniform recovery error (Plan and Vershynin, 2016) is no worse than (see Equation (35))

$$\|\hat{\boldsymbol{x}}_{GLasso} - \boldsymbol{x}\|_2 \lesssim \inf_{\phi \in (0,1)} \frac{\varphi_1 \omega(\mathcal{K}_{\boldsymbol{x},\phi})}{\phi\sqrt{m}} + \phi \le \inf_{\phi \in (0,1)} \frac{\varphi_1 \omega(\mathcal{K}_{\boldsymbol{x},1})}{\phi\sqrt{m}} + \phi \le 2\sqrt{\varphi_1}\bigg(\frac{\omega^2(\mathcal{K}_{\boldsymbol{x},1})}{m}\bigg)^{1/4}. \quad (45)$$

This worst-case error rate turns out to be tight in some setting, e.g., the recovery of approximately sparse vectors; see, e.g., Raskutti et al. (2011); Plan et al. (2017).

On the other hand, by (38), $\omega(\mathcal{X}_\varepsilon) \le \omega(\mathcal{X} - \mathcal{X}) \le 2\omega(\mathcal{X})$, $\omega(\mathcal{K}_{\mathcal{X},\phi}) \le \omega(\mathcal{K}_{\mathcal{X},1})$ (let $\varepsilon, \phi < 1$), along with a very mild assumption $\omega(\mathcal{X}) \lesssim \omega(\mathcal{K}_{\mathcal{X},1})$, our Theorem 4.4 yields the uniform recovery error

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 \lesssim \varphi_1 \frac{\omega(\mathcal{K}_{\mathcal{X},1})}{\min\{\phi, \varepsilon\}\sqrt{m}} + \phi + \varphi_4 \varepsilon + \frac{\varphi_3}{\mu}\Upsilon;$$

[6] This relaxation and $\omega(\frac{\mathcal{K}_{\boldsymbol{x},\phi}}{\phi}) \le \omega(\mathrm{cone}(\mathcal{K} - \boldsymbol{x}) \cap \mathbb{B}_2^n)$ used in (44) rely on the descent cone structure of $\mathcal{K}_{\mathcal{X}}$, which ensures that $\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n) \ll n$. In general (e.g., when $\boldsymbol{x}$ is an interior of $\mathcal{K}$), one may have $\omega^2(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2^n) \asymp n$.

moreover, by using $\omega(\mathcal{X}_\varepsilon) \le 2\omega(\mathcal{X})$, $\omega(\mathcal{K}_{\mathcal{X},\phi}) \le \omega(\mathcal{K}_{\mathcal{X},1})$, the mild assumption $\omega(\mathcal{X}) \lesssim \omega(\mathcal{K}_{\mathcal{X},1})$ and ignoring log factors in (27)–(29),

$$\Upsilon \lesssim \Big(1+\frac{\varepsilon}{\varphi_2}\Big)\Big(\frac{\omega^2(\mathcal{K}_{\mathcal{X},1})}{\min\{\varepsilon^2,\phi^2\}m}+\frac{\varphi_5\omega(\mathcal{X})}{\sqrt{\zeta m}}+\varphi_5\varepsilon+\zeta\Big).$$

Suppose $\varphi_1=\Theta(1)$, $\varphi_2=\Omega(1)$ and $\varphi_3,\varphi_4,\varphi_5=O(1)$. We set

$$\zeta=\Big(\frac{\varphi_5^2\omega^2(\mathcal{X})}{m}\Big)^{1/3} \quad \text{and} \quad \varepsilon=\phi=\sqrt{\varphi_1}\Big(\frac{\omega^2(\mathcal{K}_{\mathcal{X},1})}{m}\Big)^{1/4}$$

to reach

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{\boldsymbol{x}}_{pgd}-\boldsymbol{x}\|_2=\tilde{O}\bigg(\sqrt{\varphi_1}\Big(\frac{\omega^2(\mathcal{K}_{\mathcal{X},1})}{m}\Big)^{1/4}\bigg),$$

which is identical to (45) up to log factors. ◊

Remark 4.10 (Improvement on existing uniform guarantees in Genzel and Stollenwerk (2023)). The recent work of Genzel and Stollenwerk provides the first uniform recovery theory for single-index model via generalized Lasso (34). For Lipschitz continuous link functions (which render $(\varphi_2,\varphi_3,\varphi_5)=(\infty,0,0)$ in our Assumption 4.2), Genzel and Stollenwerk (2023) yielded a uniform recovery error rate comparable to the non-uniform one Plan and Vershynin (2016): for instance, if

$$\mathcal{X}=\Sigma_k^{n,*}\cap\{\boldsymbol{x}:\|\boldsymbol{x}\|_1=c_*\sqrt{k}\}, \quad \mathcal{K}=\mathbb{B}_1^n(c_*\sqrt{k}) \tag{46}$$

for $c_*\in(0,1)$, under $\mu=\Theta(1)$ and (C1) with $\varphi_1\asymp 1$, (C4) with $\varphi_4\asymp 1$ in our Assumption 4.2, Theorem 1 therein gives the high-probability uniform rate

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{\boldsymbol{x}}_{Glasso}-\boldsymbol{x}\|_2\lesssim\sqrt{\frac{k\log(en/k)}{m}}$$

which is identical to the nonuniform rate in Plan and Vershynin (2016). However, for discontinuous link functions, the unified approach of Genzel and Stollenwerk (2023) cannot derive a uniform recovery error decaying faster than $m^{-1/4}$: in the same example as per (46), in general, the uniform recovery error derived by Theorem 2 of Genzel and Stollenwerk (2023) is lower bounded by

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{\boldsymbol{x}}_{Glasso}-\boldsymbol{x}\|_2\gtrsim\Big(\frac{k\log(en/k)}{m}\Big)^{1/4},$$

due to the terms $t^{-2}(L_t^2+\hat{L}_2^2)\cdot\omega^2(T\mathcal{X})$ in Equation (2.3) therein; see also the discussion in (Genzel and Stollenwerk, 2023, Page 913).

Since (under mild assumption) the nonuniform error rate of Plan and Vershynin (2016) remains at

$$\|\hat{\boldsymbol{x}}_{Glasso}-\boldsymbol{x}\|_2\lesssim\sqrt{\frac{k}{m}\log\Big(\frac{en}{k}\Big)},$$

Genzel and Stollenwerk drew a conclusion that "*the transition to uniform recovery with (discontinuous) nonlinear output functions may result in a worse oversampling rate*" (Genzel and Stollenwerk, 2023, Page 916). Our work shows that this phenomenon does not occur for a large class of discontinuous link functions (obeying Assumption 4.2); instead, there exists no essential (non-log) gap between the uniform recovery errors and nonuniform recovery errors. See the three canonical settings analyzed in Remarks 4.7, 4.8, 4.9. ◊

Remark 4.11 (Key to the improvement and extension to Generalized Lasso). Although our work builds on the RAIC framework, we note that the transition from generalized Lasso to PGD (via RAIC) is not essential to our improvement over Genzel and Stollenwerk (2023). Rather, the analysis of uniform recovery via generalized Lasso ends up with bounding the same multiplier process (see Section 2.2 of Genzel and Stollenwerk (2023)), and indeed our improvement is due to establishing sharper bound on the multiplier process via a different argument. As such, the uniform error rate in our Theorem 4.4 carries over to the generalized Lasso. On the other hand, our RAIC treatment for PGD offers much more flexibility: (i) it encompasses PGD with projection onto some highly nonconvex sets such as Σ_k^n and $M_r^{n_1,n_2}$ that is beyond the scope of convex program (see Footnote 5); (ii) perhaps more importantly, it readily works many other nonlinear problems such as phase retrieval, generalized linear models, and ReLU regression (see Section 5), while generalized Lasso is tailored for Gaussian single-index model and an extension of Genzel and Stollenwerk (2023) on this regard is highly entangled (see Remark 3(2) therein). ♢

4.4 Uniform Sparse Recovery from Modulo Measurements

To further illustrate the abstract analysis in previous sections, we provide a concrete example of modulo measurements (Bhandari et al., 2020). Consider the uniform recovery of $\boldsymbol{x} \in \mathcal{X}$ from

$$y_i = m_\lambda(\boldsymbol{a}_i^T \boldsymbol{x}), \quad i = 1, 2, ..., m, \tag{47}$$

where m_λ is the modulo function given by $m_\lambda(v) = v - 2\lambda \left\lfloor \frac{v+\lambda}{2\lambda} \right\rfloor$ for some $\lambda \geq \frac{1}{4}$ (here, $\frac{1}{4}$ can be replaced by any positive constant); see Figure 2 for instance. We consider the following two sparse

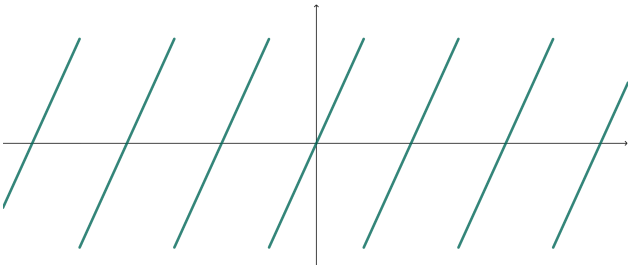

Figure 2: The graph of $m_{1/2}(v) = v - \lfloor v + \frac{1}{2} \rfloor$.

recovery settings that correspond to Remark 4.7 and Remark 4.8:

$$(\mathcal{X}, \mathcal{K}) = (\Sigma_k^{n,*}, \Sigma_k^n) \tag{48}$$

and

$$(\mathcal{X}, \mathcal{K}) = (\Sigma_k^{n,*} \cap \{\boldsymbol{u} : \|\boldsymbol{u}\|_1 = c_* \sqrt{k}\}, \mathbb{B}_1^n(c_* \sqrt{k})), \quad c_* \in (0, 1). \tag{49}$$

In the two settings, our unified framework yields the following statement that improves on Corollary 5 of Genzel and Stollenwerk (2023), where a uniform recovery error over $\boldsymbol{x} \in \mathcal{X}$ no faster than $(\frac{k \log(en/k)}{m})^{1/4}$ is achieved.

Theorem 4.5 (Uniform sparse recovery from modulo measurements)**.** *Consider (47) with Gaussian $\boldsymbol{a}_i$, suppose $\lambda \geq \frac{1}{4}$ and let $\mu = 1$. In either (48) or (49), if $m = \tilde{\Omega}(k \log \frac{en}{k})$, then with probability*

at least $1 - C\exp(-ck\log\frac{en}{k})$*, for all* $\boldsymbol{x} \in \mathcal{X}$*, the sequence* $\{\boldsymbol{x}_t\}_{t\geq 0}$ *generated by (30) with* $\boldsymbol{x}_0 = 0$ *satisfies*

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq \left(\frac{Ck\log(en/k)}{m}\right)^{t/2}\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + \tilde{O}\left(\sqrt{\frac{k\log(en/k)}{m}} + \frac{\lambda k\log(en/k)}{m}\right)$$

for any $t \geq 0$.

4.5 No Loss of Log Factors under 1-Bit Measurements

In the previous developments, we establish the uniform recovery error rates in Gaussian single-index model and show in canonical settings (a), (b), and (c) that the uniform rates are identical to the nonuniform rates, up to log factors. The log factors arise from the covering arguments in the proof of the uniform RAIC. A natural question is whether there exists a gap of log factors between the uniform and nonuniform recovery errors, or the log factors are simply proof artifacts.

In this subsection, we show for the specific 1-bit measurements and iterative hard thresholding algorithm (corresponding to setting (a)) that the log factors in Theorem 4.3 can be removed; we conjecture that, under more general discontinuous f_i, there is not a gap of log factors between the uniform and nonuniform recovery error rates. We shall discuss this further in Remark 4.12.

We consider the problem of recovery of $\boldsymbol{x} \in \Sigma_k^{n,*}$ from

$$y_i = \text{sign}(\boldsymbol{a}_i^T\boldsymbol{x}), \quad i = 1, ..., m$$

under Gaussian $\boldsymbol{a}_i$. This is the standard 1-bit compressed sensing problem which has been extensively studied (e.g., Boufounos and Baraniuk (2008); Jacques et al. (2013)). For the recovery of a fixed $\boldsymbol{x}$, Plan and Vershynin (2016) showed that the generalized Lasso

$$\hat{\boldsymbol{x}}_{Glasso} = \arg\min_{\boldsymbol{u}:\|\boldsymbol{u}\|_1 \leq \|\boldsymbol{x}\|_1} \frac{1}{2m}\sum_{i=1}^m \left(y_i - \sqrt{\frac{2}{\pi}}\boldsymbol{a}_i^T\boldsymbol{u}\right)^2$$

attains, with high probability, the non-uniform recovery error rate

$$\|\hat{\boldsymbol{x}}_{Glasso} - \boldsymbol{x}\|_2 \lesssim \sqrt{\frac{k\log(en/k)}{m}}.$$

We shall consider PGD in (30) with $\mathcal{K} = \Sigma_k^n$, that is a procedure of IHT:

$$\boldsymbol{x}_{t+1} = P_{\Sigma_k^n}\left(\boldsymbol{x}_t - \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^T\boldsymbol{x}_t - \sqrt{\frac{\pi}{2}}y_i\right)\boldsymbol{a}_i\right), \quad t = 0, 1, 2, \cdots. \tag{50}$$

Also note that the main result of Oymak and Soltanolkotabi (2017) yields a nonuniform error rate $\|\hat{\boldsymbol{x}}_{iht} - \boldsymbol{x}\|_2 \lesssim \sqrt{k\log(en/k)/m}$ identical to generalized Lasso. While an application of Theorem 4.3 yields the uniform rate (see Remark 4.7)

$$\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \|\hat{\boldsymbol{x}}_{iht} - \boldsymbol{x}\|_2 = \tilde{O}\left(\sqrt{\frac{k\log(en/k)}{m}}\right)$$

the following theorem shows that, indeed, the uniform recovery error rate is of order

$$O\left(\sqrt{\frac{k\log(en/k)}{m}}\right).$$

Hence, there is no loss of log factors in the recovery errors when shifting from nonuniform recovery to uniform recovery (rather, the loss is at most an absolute constant).

Theorem 4.6 (Uniform 1-bit compressed sensing with no loss of log factor)**.** *We solve the 1-bit compressed sensing problem by running PGD in (50) starting from an arbitrary* $\boldsymbol{x}_0 \in \Sigma_k^{n,*}$ *for all* $\boldsymbol{x} \in \Sigma_k^{n,*}$*. If* $m \gtrsim k \log(\frac{en}{k})$*, then*

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \le \left(\frac{Ck\log(en/k)}{m}\right)^{t/2} + \sqrt{\frac{C_1 k \log(en/k)}{m}}, \quad \forall t \ge 0, \quad \forall \boldsymbol{x} \in \Sigma_k^{n,*}$$

holds with probability at least $1 - C'\exp(-c'k\log\frac{en}{k})$*.*

The main technical ingredient to prove Theorem 4.6 is to show

$$\sup_{\boldsymbol{x},\boldsymbol{q}\in\Sigma_{2k}^{n,*}} \left| \sqrt{\frac{\pi}{2}} \frac{1}{m} \sum_{i=1}^{m} \operatorname{sign}(\boldsymbol{a}_i^T \boldsymbol{x}) \boldsymbol{a}_i^T \boldsymbol{q} - \boldsymbol{x}^T \boldsymbol{q} \right| \lesssim \sqrt{\frac{k\log(en/k)}{m}}\,.$$

Our key insight is to work conditionally on the sign functions and find a suitable metric that is essentially an interpolation between the standard ℓ_2 metric induced by the linear component $\boldsymbol{a}_i^T\boldsymbol{q}$ and the Lipschitz metric with respect to the empirical measure given by $\sum_{i=1}^m \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')$; importantly, covering number estimates are available through invoking VC theory. Finally, once the covering number under this interpolation metric is controlled, we can apply Dudley's inequality to control the supremum of the process of interest.

Remark 4.12 (Potential extension of Theorem 4.6). Compared to Theorems 4.3 and 4.4, the above result is sharper but nonetheless has two restrictions: (i) for IHT only, a representative example of setting (a) and Theorem 4.3; (ii) for $f_i = \operatorname{sign}$ only, corresponding to the 1-bit compressed sensing problem. To relax the restriction (ii), one may seek such an extension: in fact, once the conditional sub-Gaussianity in Lemma D.2 extends to a conditioning on $\{\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}-w), \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}'-w)\}$ for constant w, our VC dimension argument yields $\sup_{\boldsymbol{x},\boldsymbol{q}\in\Sigma_{2k}^{n,*}} |\frac{1}{m}\sum_{i=1}^m \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}-w)\boldsymbol{a}_i^T\boldsymbol{q} - \mathbb{E}[\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}-w)\boldsymbol{a}_i^T\boldsymbol{q}]| = O_w(\sqrt{k\log(en/k)/m})$, and hence IHT under $f_i(\cdot) = \operatorname{sign}(\cdot - w)$ achieves a sharp uniform recovery error rate of $O_w(\sqrt{k\log(en/k)/m})$; then immediately, the uniform rate $O(\sqrt{k\log(en/k)/m})$ extends to f_i that is a linear combination of a Lipschitz continuous function and functions in $\{\operatorname{sign}(\cdot - w) : w\}$. To keep the paper at a reasonable length, we leave this investigation for future work. ◊

5 Nonlinear Observations via RAIC

We provide a review of the recent line of works that analyze recovery or regression problems with nonlinear observations by establishing the RAIC.

Quantized measurements. To study 1-bit compressed sensing, RAIC was first (formally) introduced by Friedlander, Jeong, Plan and Yilmaz in Definition 8 of Friedlander et al. (2021) — in a multiscale manner — to show that the normalized binary iterative hard thresholding (NBIHT) algorithm achieves the rate $\tilde{O}(\frac{k^{3.5}}{m})$, although we mention in passing that Oymak and Soltanolkotabi (2017); Soltanolkotabi (2019, 2017) essentially established the RAIC of the gradient for some nonlinear regression problems. While Friedlander et al. (2021) is the first work to achieve $O(m^{-1})$ decay rate for an efficient 1-bit compressed sensing algorithm, the dependence on the sparsity k — i.e., $k^{3.5}$ up to log factors — is sub-optimal, in light of the information-theoretic optimal rate $\tilde{\Theta}(\frac{k}{m})$ (Jacques et al., 2013). To close this gap, the work of Matsumoto and Mazumdar (2024a) refined the definition of RAIC (see Definition 3.1 therein) and introduced a number of new ideas to

the analysis, showing that NBIHT is indeed nearly information-theoretic optimal and attains the uniform recovery error

$$\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \|\hat{\boldsymbol{x}}_{nbiht} - \boldsymbol{x}\|_2 = \tilde{O}\Big(\frac{k}{m}\Big).$$

In a follow-up work of the same authors, Matsumoto and Mazumdar (2024b) showed that NBIHT is robust to adversarial bit flips; in Section 6, we shall see that this robustness property of NBIHT follows from slightly more work (see Example 6.3 therein). Another follow-up work Chen and Yuan (2024b) provides two-fold extensions of Matsumoto and Mazumdar (2024a) toward more general quantization models and signal structures.

Other nonlinear observations. Let us now consider the statistical learning problem of sparse logistic regression with a temperature parameter $\beta > 0$, in which the goal is to learn the true parameter $\boldsymbol{\theta}^* \in \Sigma_k^{n,*}$ from known Gaussian covariates $\{\boldsymbol{x}_i\}_{i=1}^m$ and the Bernoulli responses

$$y_i = \text{Bernoulli}\bigg(\frac{1}{1+e^{-\beta \boldsymbol{x}_i^T\boldsymbol{\theta}^*}}\bigg), \quad i = 1, ..., m.$$

This problem is more general and hence strictly harder than 1-bit compressed sensing, since it reduces to 1-bit compressed sensing when $\beta \to \infty$. While the case of $\beta = 1$ is well understood (e.g., Negahban et al. (2012); Plan et al. (2017)), the information-theoretic limit for all $\beta > 0$ is a recent result due to Hsu and Mazumdar (2024), which exhibits a transition from $\tilde{O}(\sqrt{k/m})$ to $\tilde{O}(k/m)$ as β increases from a 0 to ∞. More recently, by establishing an RAIC (see Theorem 9 therein), Matsumoto and Mazumdar (2025) showed that an algorithm analogous to NBIHT is near-optimal and attains the information-theoretic limit in Hsu and Mazumdar (2024).

The problem of 1-bit phase retrieval is another nonlinear model that is more intricate than 1-bit compressed sensing. It concerns the reconstruction of $\boldsymbol{x}$ using only 1-bit information of each phaseless observation and has attracted some research attention (Mroueh and Rosasco, 2013; Kishore and Seelamantula, 2020; Domel-White and Bodmann, 2022; Eamaz et al., 2022), among which Domel-White and Bodmann (2022) is the only work providing non-asymptotic error rates — particularly, the authors analyzed a spectral method and derived a nonuniform recovery error rate $\text{dist}(\hat{\boldsymbol{x}}_{spectral}, \boldsymbol{x}) = \tilde{O}(\sqrt{n/m})$ and a substantially degraded uniform recovery error rate $\sup_{\boldsymbol{x}\in S^{n-1}} \text{dist}(\hat{\boldsymbol{x}}_{spectral}, \boldsymbol{x}) = \tilde{O}(\sqrt{n^2/m})$, where $\text{dist}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \min\{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2, \|\hat{\boldsymbol{x}} + \boldsymbol{x}\|_2\}$ is the phaseless ℓ_2 distance. The recent work of Chen and Yuan (2024a) analyzed the recovery of $\boldsymbol{x} \in \mathbb{A}_\alpha^\beta = \{\boldsymbol{u} : \alpha \le \|\boldsymbol{u}\|_2 \le \beta\}$ from $\boldsymbol{y} = \text{sign}(|\boldsymbol{A}\boldsymbol{x}| - \tau)$ with an $m \times n$ Gaussian matrix $\boldsymbol{A}$ and a fixed $\tau > 0$. Built upon Matsumoto and Mazumdar (2024a), the authors introduced a number of new ideas to prove a phaseless local RAIC (see Definition 4.1 therein), which then implies that the procedure of a spectral method followed by gradient descent attains the near-optimal uniform rate

$$\sup_{\boldsymbol{x}\in\mathbb{A}_\alpha^\beta} \text{dist}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \tilde{O}_{\alpha,\beta,\tau}\Big(\frac{n}{m}\Big).$$

The problem of 1-bit sparse phase retrieval was also analyzed in Chen and Yuan (2024a).

Unified approach. While the above-reviewed works provide sharp analysis to a specific nonlinear problem through establishing the RAIC, Chen et al. (2025) showed that RAIC is a unified approach to estimation problems with nonlinear observations, consisting of the two steps of choosing a gradient operator and establishing the RAIC. One central perspective of Chen et al. (2025)

is that RAIC serves as an analog of RIP for nonlinear models; to support this, the authors established three general convergence guarantees under RAIC — the convergence of PGD, the local convergence of Riemannian gradient descent for tensor regression, and the local convergence of gradient descent based on matrix factorization — all of which are implications of RIP under linear observations (e.g., Luo and Zhang (2024); Tu et al. (2016)). We note that our definitions of RAIC are taken from Chen et al. (2025).

The present paper is a follow-up work of Chen et al. (2025) and shows that RAIC provides a unified approach to uniform recovery of structured signals from nonlinear observations. This is reminiscent of the implication of RIP on uniform sparse recovery from (noisy) linear observations (Foucart and Rauhut, 2013, Section 6). As already noted, this contribution is mainly of pedagogical value, since indeed prior works already implicitly rely on this perspective to obtain uniform recovery guarantees from quantized measurements (Matsumoto and Mazumdar, 2024a; Chen and Yuan, 2024b,a) and logistic regression (Matsumoto and Mazumdar, 2025). Instead, our technical contributions mainly lie in the application of this approach to the Gaussian single-index model where we improve on Genzel and Stollenwerk (2023). Focusing on sparse recovery via IHT, a companion paper (Chen and Maleki, 2026) develops a unified approach to instance optimal sparse recovery.

6 Robustness of PGD

Robustness to noise or corruption is another desideratum in signal recovery and statistical estimation. The aim of this section is to show that the robustness of PGD can be incorporated into the RAIC approach in an elegant way — by *bounding one additional random process*. Taken collectively with the previous developments, our paper shows that RAIC is a unified approach to robust uniform recovery of structured signals from nonlinear observations, as per the title of the present paper.

We shall begin with an intuition. Suppose that in the noisy setting we can only access $\{(\boldsymbol{a}_i, \tilde{y}_i)\}_{i=1}^m$ for the estimation of $\boldsymbol{x} \in \mathcal{X}$, where $\{\tilde{y}_i\}_{i=1}^m$ are the noisy versions of the noiseless observations $\{y_i\}_{i=1}^m$. (Note that $\tilde{y}_i = y_i$ for $i \in [m]$ returns the noiseless case.) Our RAIC-based approach remains effective as the only difference is that the "noisy gradient" is now constructed from $\{(\boldsymbol{a}_i, \tilde{y}_i)\}_{i=1}^m$ and could differ from the "noiseless gradient" constructed from $\{(\boldsymbol{a}_i, y_i)\}_{i=1}^m$; all we need is to control an additional term capturing such difference.

To formalize the idea, for any $\boldsymbol{x} \in \mathcal{X}$ we shall use $\tilde{\boldsymbol{h}}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$ to denote the noisy gradient, which is a perturbed version of the noiseless gradient $\boldsymbol{h}_{\boldsymbol{x}} : \mathbb{R}^n \to \mathbb{R}^n$. Note that $\tilde{\boldsymbol{h}}_{\boldsymbol{x}} = \boldsymbol{h}_{\boldsymbol{x}}$ when $\tilde{y}_i = y_i$ for $i \in [m]$. It remains to establish the RAIC of $\tilde{\boldsymbol{h}}_{\boldsymbol{x}}$ and then invoke Theorems 3.1 or 3.2 to yield the recovery guarantee of PGD, and note that, if the RAIC is uniform over $\boldsymbol{x} \in \mathcal{X}_* \subset \mathcal{X}$, then the guarantee is uniform over $\mathcal{X}_*$. Built upon the RAIC of $\boldsymbol{h}_{\boldsymbol{x}}$, the RAIC of $\tilde{\boldsymbol{h}}_{\boldsymbol{x}}$ follows from a bound on

$$\sup_{\boldsymbol{x} \in \mathcal{X}_*} \sup_{\boldsymbol{u} \in \mathcal{U}_{\boldsymbol{x}}} \|\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) - \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\|_{\mathcal{K}_1^\circ}, \tag{51}$$

where $\mathcal{X}_*$ denotes the set of signals we want to uniformly recover (note that we recover nonuniform recovery when $\mathcal{X}_* = \{\boldsymbol{x}\}$ for a fixed $\boldsymbol{x} \in \mathcal{X}$). Intuitively, (51) captures the mismatch between the noisy gradient and the noiseless gradient under dual norm.

In the following, we provide a formal statement. We focus on signals living in a cone $\mathcal{K}$ for brevity while note that the case of $\mathcal{K}$ being a convex set is parallel.

Theorem 6.1 (Robust uniform recovery of signals in a cone)**.** *Suppose that $\mathcal{K}$ is a cone such that $\mathcal{X}_* \subset \mathcal{K}$. If*

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\big(\mathcal{K}; \mathcal{U}_{\boldsymbol{x}}, \mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2, \eta\big), \quad \forall \boldsymbol{x} \in \mathcal{X}_*, \tag{52}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}_*} \sup_{\boldsymbol{u}\in\mathcal{U}_{\boldsymbol{x}}} \eta\big\|\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) - \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\big\|_{\mathcal{K}_1^\circ} \le \tilde{\mu}, \tag{53}$$

where $\mu_1 < \frac{1}{2}$ and

$$\mathcal{U}_{\boldsymbol{x}} \supset \mathcal{K} \cap B_2^n(\boldsymbol{x}; d_{\boldsymbol{x}}) \ \text{ for some } \ \frac{2(\mu_2+\tilde{\mu})}{1-2\mu_1} < d_{\boldsymbol{x}} \le \infty, \quad \forall \boldsymbol{x} \in \mathcal{X}_*, \tag{54}$$

then $\{\boldsymbol{x}_t\}_{t\ge 0}$ generated by $\boldsymbol{x}_{t+1} = P_{\mathcal{K}}\big(\boldsymbol{x}_t - \eta\cdot\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{x}_t)\big)$ $(t=0,1,\cdots)$ and with initialization

$$\boldsymbol{x}_0 \in \mathcal{K} \cap B_2^n(\boldsymbol{x}; d_{\boldsymbol{x}}) \tag{55}$$

satisfies

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \le (2\mu_1)^t\|\boldsymbol{x}_0-\boldsymbol{x}\|_2 + \frac{2(\mu_2+\tilde{\mu})}{1-2\mu_1}, \quad \forall t\ge 0,\ \boldsymbol{x}\in\mathcal{X}_*.$$

Proof. By definition, (52)–(53) imply

$$\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\big(\mathcal{K}; \mathcal{U}_{\boldsymbol{x}}, \mu_1\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \mu_2 + \tilde{\mu}, \eta\big), \quad \forall \boldsymbol{x} \in \mathcal{X}_*,$$

and $\mu_1 < \frac{1}{2}$, (54) and (55) ensure (5)–(7). The result then follows by applying Theorem 3.1 to every $\boldsymbol{x}\in\mathcal{X}_*$. □

Remark 6.1. It is clear that $\tilde{\mu}$, as an upper bound on the random process $\sup_{\boldsymbol{x}\in\mathcal{X}_*}\sup_{\boldsymbol{u}\in\mathcal{U}_{\boldsymbol{x}}}\eta\big\|\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) - \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\big\|_{\mathcal{K}_1^\circ}$, captures the effect of noise or corruption. ♢

Let us proceed to a number of concrete examples to put the above approach to robustness in perspective. We treat the recovery of $\boldsymbol{x}\in\Sigma_k^n$ via iterative hard thresholding, i.e., (PGD) with $\mathcal{K}=\Sigma_k^n$. We also consider i.i.d. $\boldsymbol{a}_i \sim N(0,\boldsymbol{I}_n)$.

Example 6.1 (Noisy compressed sensing). As a warm-up example, we consider the recovery of $\boldsymbol{x}\in\Sigma_k^n$ from noisy linear observations $y_i = \boldsymbol{a}_i^T\boldsymbol{x} + \varepsilon_i$, $i\in[m]$ and adopt the ℓ_2 loss $L_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{2m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{u}-y_i)^2$. Hence, the "noisy" gradient is $\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{u}-y_i)\boldsymbol{a}_i = \frac{1}{m}\sum_{i=1}^m\big(\boldsymbol{a}_i^T\boldsymbol{u} - \boldsymbol{a}_i^T\boldsymbol{x}-\varepsilon_i\big)\boldsymbol{a}_i$. When $\varepsilon_i=0$ for $i\in[m]$, it reduces to the "noiseless" gradient $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^m\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{u}-\boldsymbol{x})$. It is not hard to show that [7]

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\bigg(\Sigma_k^n; \Sigma_k^n, \frac{1}{3}\|\boldsymbol{u}-\boldsymbol{x}\|_2, 1\bigg), \quad \forall \boldsymbol{x}\in\Sigma_k^n$$

holds with high probability. In light of Theorem 6.1 (with $d_{\boldsymbol{x}}=\infty$ for all $\boldsymbol{x}\in\Sigma_k^n$),

$$\boldsymbol{x}_{t+1} = P_{\Sigma_k^n}\bigg(\boldsymbol{x}_t - \frac{1}{m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{x}_t - y_i)\boldsymbol{a}_i\bigg), \quad t=0,1,2,\cdots$$

[7] This is indeed an implication of the RIP of $\boldsymbol{A}/\sqrt{m}$ over sparse vectors (e.g., (Chen et al., 2025, Section F.1)) and therefore holds over a large class of matrices $\boldsymbol{A}$ beyond Gaussian matrix.

with sufficiently many iterations uniformly recovers all $\boldsymbol{x} \in \Sigma_k^n$ to error of the order

$$\sup_{\boldsymbol{x}\in\Sigma_k^n} \sup_{\boldsymbol{u}\in\Sigma_k^n} \left\|\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) - \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\right\|_{(\Sigma_{2k}^{n,*})^\circ} = \sup_{\boldsymbol{x}\in\Sigma_k^n} \left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i\right\|_{(\Sigma_{2k}^{n,*})^\circ}.$$

This is solely the effect of noise (as exact recovery is achieved in noiseless case). Therefore,

$$\sup_{\boldsymbol{x}\in\Sigma_k^n} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 \lesssim \sup_{\boldsymbol{x}\in\Sigma_k^n} \left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i\right\|_{(\Sigma_{2k}^{n,*})^\circ}.$$

We consider two types of $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_m)^T$ and derive explicit bounds. First, suppose that $\boldsymbol{\varepsilon}$ is arbitrary (adversarial) corruption possibly depending on $(\boldsymbol{A}, \boldsymbol{x})$, then by Cauchy-Schwarz and the high-probability bound $\sup_{\boldsymbol{u}\in\Sigma_{2k}^{n,*}} \|\boldsymbol{A}\boldsymbol{u}\|_2 = O(\sqrt{m})$ (see, e.g., Theorem 9.1.1 of Vershynin (2018)),

$$\sup_{\boldsymbol{x}\in\Sigma_k^n} \left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i\right\|_{(\Sigma_{2k}^{n,*})^\circ} = \sup_{\boldsymbol{x}\in\Sigma_k^n} \sup_{\boldsymbol{u}\in\Sigma_{2k}^{n,*}} \frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i^T\boldsymbol{u} \le \sup_{\boldsymbol{u}\in\Sigma_{2k}^{n,*}} \frac{\|\boldsymbol{\varepsilon}\|_2\|\boldsymbol{A}\boldsymbol{u}\|_2}{m} \lesssim \frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{m}}. \tag{56}$$

The result is consistent with Theorem 5 and Equation (17) of Blumensath and Davies (2009). Next, we study a statistical learning setting where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ are oblivious to $\boldsymbol{x}$ and independent of $\boldsymbol{A}$. In light of $\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i\|_{\psi_2} \lesssim \frac{\sigma}{\sqrt{m}}$ (see, e.g., Section 2 of Vershynin (2018)), we have that

$$\sup_{\boldsymbol{x}\in\Sigma_k^n} \left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i\right\|_{(\Sigma_{2k}^{n,*})^\circ} = \sup_{\boldsymbol{u}\in\Sigma_{2k}^{n,*}} \frac{1}{m}\sum_{i=1}^m \varepsilon_i \boldsymbol{a}_i^T\boldsymbol{u} \lesssim \frac{\sigma\cdot\omega(\Sigma_{2k}^{n,*})}{\sqrt{m}} \lesssim \sigma\sqrt{\frac{k\log(en/k)}{m}}$$

holds with high probability (e.g., Section 8.6 of Vershynin (2018)). This recovers the minimax optimal rate in noisy sparse linear regression (Raskutti et al., 2011). △

Example 6.2 (Noisy sparse phase retrieval). In this problem, we seek to recover $\boldsymbol{x} \in \Sigma_k^n \setminus \{0\}$ from $y_i = |\boldsymbol{a}_i^T\boldsymbol{x}| + \varepsilon_i$, $i \in [m]$, or more compactly, $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}| + \boldsymbol{\varepsilon}$. To avoid the subtlety of the signal-to-noise ratio, let us focus on $\boldsymbol{x} \in \Sigma_k^{n,*}$. As with the reshaped Wirtinger flow (Zhang et al., 2017) and truncated amplitude flow (Wang et al., 2017a,b), we shall adopt the amplitude-based ℓ_2 loss $L_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{2m}\||\boldsymbol{A}\boldsymbol{u}| - \boldsymbol{y}\|_2^2 = \frac{1}{2m}\sum_{i=1}^m (|\boldsymbol{a}_i^T\boldsymbol{u}| - y_i)^2$, whose subgradient is given by

$$\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(u) = \frac{1}{m}\sum_{i=1}^m \left(|\boldsymbol{a}_i^T\boldsymbol{u}| - y_i\right)\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u})\boldsymbol{a}_i = \frac{1}{m}\sum_{i=1}^m \left(|\boldsymbol{a}_i^T\boldsymbol{u}| - |\boldsymbol{a}_i^T\boldsymbol{x}| - \varepsilon_i\right)\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u})\boldsymbol{a}_i,$$

and reduces to the noiseless gradient $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^m \left(|\boldsymbol{a}_i^T\boldsymbol{u}| - |\boldsymbol{a}_i^T\boldsymbol{x}|\right)\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u})\boldsymbol{a}_i$ when $\varepsilon = 0$. In Theorem 5.7 of Chen et al. (2025), the authors established the following uniform RAIC for $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$: if $m \gtrsim k\log\frac{en}{k}$, then for some absolute constant $c_* > 0$, with high probability,

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\left(\Sigma_k^n; \Sigma_k^n \cap B_2^n(\boldsymbol{x}; c_*), \frac{1}{4}\|\boldsymbol{u} - \boldsymbol{x}\|_2, 1\right), \quad \forall \boldsymbol{x} \in \Sigma_k^{n,*}.$$

This implies the uniform exact recovery of all $\boldsymbol{x} \in \Sigma_k^{n,*}$ by PGD starting from some $\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 \le c_*/2$. Such an initialization can be found by spectral method as long as the noise level is small enough and $m = \tilde{\Omega}(k^2)$ (Jagatap and Hegde, 2019), and we do not discuss this subtle issue. Rather, here we are interested in characterizing the impacts of noise. For iid Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent of $\boldsymbol{A}$, Theorem 5.7 of Chen et al. (2025) shows the uniform recovery error

$$\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 = \sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \sup_{\boldsymbol{u}\in B_2^n(\boldsymbol{x};c_*)} \left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u})\boldsymbol{a}_i\right\|_{(\Sigma_{2k}^{n,*})^\circ} = \tilde{O}\left(\sigma\sqrt{\frac{k\log(en/k)}{m}}\right),$$

which is minimax optimal up to log factor (e.g., Theorem 5.9 of Chen et al. (2025)). For arbitrary noise $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_m)^T$ obeying $\|\boldsymbol{\varepsilon}\|_2 \le c\sqrt{m}$ (for some small enough c) possibly depending on $(\boldsymbol{A}, \boldsymbol{x})$,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \|\hat{\boldsymbol{x}}_{pgd} - \boldsymbol{x}\|_2 &= \sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \sup_{\boldsymbol{u}\in B_2^n(\boldsymbol{x};c_*)} \left\| \frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u})\boldsymbol{a}_i \right\|_{(\Sigma_{2k}^{n,*})^\circ} \\ &\le \sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \sup_{\boldsymbol{u}\in B_2^n(\boldsymbol{x};c_*)} \sup_{\boldsymbol{v}\in\Sigma_{2k}^{n,*}} \frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u})\boldsymbol{a}_i^T\boldsymbol{v} \\ &\le \frac{\|\boldsymbol{\varepsilon}\|_2}{m} \sup_{\boldsymbol{v}\in\Sigma_{2k}^{n,*}} \|\boldsymbol{A}\boldsymbol{v}\|_2 = O\bigg(\frac{\|\boldsymbol{\varepsilon}\|_2}{\sqrt{m}}\bigg)\end{aligned}$$

by a reasoning parallel to (56). Note that this error term is consistent with Theorem 3 of Zhang et al. (2017). △

Example 6.3 (1-bit compressed sensing). In the noiseless case, this concerns the recovery of $\boldsymbol{x} \in \Sigma_k^{n,*}$ from $\boldsymbol{y} = \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})$ with $m \times n$ Gaussian matrix $\boldsymbol{A}$. While it can be encompassed into the single-index model, the solver in Section 4 only achieves $O(\sqrt{k\log(en/k)/m})$ (see Theorem 4.6) and falls short of the information-theoretic rate $\tilde{O}(k/m)$ (Jacques et al., 2013). The only known and provably optimal efficient algorithm is NBIHT (Matsumoto and Mazumdar, 2024a). NBIHT is in essence a projected (sub)gradient descent algorithm with regard to the ReLU loss $L_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^m \max\{0, -y_i\boldsymbol{a}_i^T\boldsymbol{u}\} = \frac{1}{2m}\sum_{i=1}^m \big[|\boldsymbol{a}_i^T\boldsymbol{u}| - y_i\boldsymbol{a}_i^T\boldsymbol{u}\big]$, which is a convex relaxation of the hamming distance loss and possesses the (noiseless) subgradient $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{2m}\sum_{i=1}^m \big(\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u}) - \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\big)\boldsymbol{a}_i$. With step size $\eta = \sqrt{2\pi}$ and an additional normalization step, NBIHT proceeds as

$$\boldsymbol{x}_{t+1} = \frac{P_{\Sigma_k^n}\big(\boldsymbol{x}_t - \sqrt{2\pi}\cdot\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t)\big)}{\|P_{\Sigma_k^n}\big(\boldsymbol{x}_t - \sqrt{2\pi}\cdot\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{x}_t)\big)\|_2}, \quad t = 0, 1, 2, ... \tag{57}$$

The optimality of NBIHT follows from the following uniform RAIC: if $m = \tilde{\Omega}(k\log\frac{en}{k})$, then with high probability, [8]

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\bigg(\Sigma_k^n; \Sigma_k^{n,*}, \frac{1}{3}\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \tilde{O}\Big(\frac{k}{m}\Big), \sqrt{2\pi}\bigg), \quad \forall \boldsymbol{x} \in \Sigma_k^{n,*};$$

see (Matsumoto and Mazumdar, 2024a, Theorem 3.3) or (Chen and Yuan, 2024b, Theorem 5). Matsumoto and Mazumdar (2024b) showed the robustness of NBIHT to η-fraction of adversarial bit flips: suppose that our observation is $\tilde{\boldsymbol{y}} \in \{-1, 1\}^m$ obeying $d_H(\tilde{\boldsymbol{y}}, \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})) := \sum_{i=1}^m \mathbb{1}\big(\tilde{y}_i \ne \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\big) \le \eta m$ for some small enough $\eta \in (0, 1)$, then NBIHT (with $\tilde{\boldsymbol{y}}$ as input) satisfies

$$\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \|\hat{\boldsymbol{x}}_{nbiht} - \boldsymbol{x}\|_2 = \tilde{O}\Big(\frac{k}{m}\Big) + O(\eta\sqrt{\log(1/\eta)}),$$

where the term $O(\eta\sqrt{\log(1/\eta)})$ captures the effect of adversarial bit flips.

The main aim of this example is to show that this robustness result can be derived from our framework. Let $\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{2m}\sum_{i=1}^m (\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{u}) - \tilde{y}_i)\boldsymbol{a}_i$ be the noisy gradient, then the effect of the

[8] Although $\Sigma_k^{n,*} \not\supseteq \Sigma_k^n \cap B_2^n(\boldsymbol{x}; d)$ for any $d > 0$ and hence Theorem 3.1 does not apply, (57) is capable of yielding the convergence of NBIHT to $\tilde{O}(k/m)$ ℓ_2 error due to the normalization step; see Remark 3.1.

adversarial corruption is captured by the term in (53). With high probability, it can be bounded by

$$
\begin{aligned}
&\sup_{\boldsymbol{x},\boldsymbol{u}\in\Sigma_k^{n,*}} \sqrt{2\pi}\|\tilde{\boldsymbol{h}}_{\boldsymbol{x}}(\boldsymbol{u}) - \boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})\|_{(\Sigma_{2k}^{n,*})^\circ} \\
&= \sup_{\boldsymbol{x},\boldsymbol{u}\in\Sigma_k^{n,*}} \sqrt{2\pi} \sup_{\boldsymbol{v}\in\Sigma_{2k}^{n,*}} \frac{1}{2m}\sum_{i=1}^m (y_i - \tilde{y}_i)\boldsymbol{a}_i^T\boldsymbol{v} \\
&\le \sqrt{2\pi} \sup_{\boldsymbol{v}\in\Sigma_{2k}^{n,*}} \max_{\substack{I\subset[m]\\|I|\le\eta m}} \frac{1}{m}\sum_{i\in I} |\boldsymbol{a}_i^T\boldsymbol{v}| \\
&\qquad \blacktriangleright \text{ by triangle inequality and } d_H(\tilde{\boldsymbol{y}}, \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})) \le \eta m \\
&\le \sqrt{2\pi\eta} \sup_{\boldsymbol{v}\in\Sigma_{2k}^{n,*}} \max_{\substack{I\subset[m]\\|I|\le\eta m}} \frac{1}{\sqrt{m}} \bigg(\sum_{i\in I} |\boldsymbol{a}_i^T\boldsymbol{v}|^2\bigg)^{1/2} \\
&\qquad \blacktriangleright \text{ by Cauchy-Schwarz inequality} \\
&\lesssim \eta\sqrt{\log(\eta^{-1})}, \\
&\qquad \blacktriangleright \text{ by a concentration bound in Lemma A.2}
\end{aligned}
$$

as desired. △

Remark 6.2. Note that the error term $O(\eta\sqrt{\log(\eta^{-1})})$ in Example 6.3 is near optimal up to log factor, since there exists some $\boldsymbol{x}' \in \Sigma_k^{n,*}$ obeying $\|\boldsymbol{x}' - \boldsymbol{x}\|_2 \asymp \eta$ such that $\boldsymbol{x}'$ and $\boldsymbol{x}$ are not distinguishable under η-fraction of adversarial bit flips: let $P_{\boldsymbol{x},\boldsymbol{x}'} = \mathbb{P}(\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')) \le \frac{\eta}{2}$ (e.g., see Plan and Vershynin (2014)), then in light of $d_H(\operatorname{sign}(\boldsymbol{A}\boldsymbol{x}'), \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})) = \sum_{i=1}^m \mathbb{1}\big(\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}') \neq \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\big) \sim \mathrm{Bin}(m, P_{\boldsymbol{x},\boldsymbol{x}'})$, Chernoff bound yields $d_H(\operatorname{sign}(\boldsymbol{A}\boldsymbol{x}'), \operatorname{sign}(\boldsymbol{A}\boldsymbol{x})) \le \eta m$ with high probability. ◊

References

A. Bhandari, F. Krahmer, and R. Raskar. On unlimited sampling and reconstruction. *IEEE Transactions on Signal Processing*, 69:3827–3839, 2020.

T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.

E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

J. Chen and A. Maleki. Iterative hard thresholding is instance optimal for sparse recovery from phaseless, one-bit, relu, and phase-only measurements. *Preprint, available upon request*, 2026.

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*, 69(10):6739–6764, 2023.

J. Chen and M. Yuan. One-bit phase retrieval: Optimal rates and efficient algorithms. *arXiv preprint arXiv:2405.04733*, 2024a.

J. Chen and M. Yuan. Optimal quantized compressed sensing via projected gradient descent. *arXiv preprint arXiv:2407.04951*, 2024b.

J. Chen, J. Scarlett, M. K. Ng, and Z. Liu. A unified framework for uniform signal recovery in nonlinear generative compressed sensing. *Advances in Neural Information Processing Systems*, 2023.

J. Chen, Z. Liu, M. Ding, and M. K. Ng. Uniform recovery guarantees for quantized corrupted sensing using structured or generative priors. *SIAM Journal on Imaging Sciences*, 17(3):1909–1977, 2024.

J. Chen, L. Ding, D. Xia, and M. Yuan. A unified approach to statistical estimation under nonlinear observations: tensor estimation and low-rank factorization. *arXiv preprint arXiv:2510.16965*, 2025.

J. Chen, M. K. Ng, and J. Scarlett. Robust instance optimal phase-only compressed sensing. *Information and Inference: A Journal of the IMA (to appear)*, 2026.

J. Depersin. Robust subgaussian estimation with vc-dimension. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 60, pages 971–989. Institut Henri Poincaré, 2024.

S. Dirksen and S. Mendelson. Non-gaussian hyperplane tessellations and robust one-bit compressed sensing. *Journal of the European Mathematical Society*, 23(9):2913–2947, 2021.

D. Domel-White and B. G. Bodmann. Phase retrieval by binary questions: Which complementary subspace is closer? *Constructive Approximation*, 56(1):1–33, 2022.

A. Eamaz, F. Yeganegi, and M. Soltanalian. One-bit phase retrieval: More samples means less complexity? *IEEE Transactions on Signal Processing*, 70:4618–4632, 2022.

S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. In *Applied and Numerical Harmonic Analysis*, 2013.

M. P. Friedlander, H. Jeong, Y. Plan, and Ö. Yılmaz. Nbiht: An efficient algorithm for 1-bit compressed sensing with optimal error decay rate. *IEEE Transactions on Information Theory*, 68(2):1157–1177, 2021.

M. Genzel and A. Stollenwerk. A unified approach to uniform signal recovery from nonlinear observations. *Foundations of Computational Mathematics*, 23(3):899–972, 2023.

D. Hsu and A. Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2418–2437. PMLR, 2024.

L. Jacques. Small width, low distortions: quantized random embeddings of low-complexity sets. *IEEE Transactions on Information Theory*, 63(9):5477–5495, 2017.

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*, 59 (4):2082–2102, 2013.

G. Jagatap and C. Hegde. Sample-efficient algorithms for recovering structured signals from magnitude-only measurements. *IEEE Transactions on Information Theory*, 65(7):4434–4456, 2019.

V. Kishore and C. S. Seelamantula. Wirtinger flow algorithms for phase retrieval from binary measurements. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5750–5754. IEEE, 2020.

Y. Luo and A. R. Zhang. Tensor-on-tensor regression: Riemannian optimization, over-parameterization, statistical-computational gap and their interplay. *The Annals of Statistics*, 52(6):2583–2612, 2024.

N. Matsumoto and A. Mazumdar. Binary iterative hard thresholding converges with optimal number of measurements for 1-bit compressed sensing. *Journal of the ACM*, 71(5):1–64, 2024a.

N. Matsumoto and A. Mazumdar. Robust 1-bit compressed sensing with iterative hard thresholding. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2941–2979. SIAM, 2024b.

N. Matsumoto and A. Mazumdar. Learning sparse generalized linear models with binary outcomes via iterative hard thresholding. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3933–4032. PMLR, 30 Jun–04 Jul 2025.

P. McCullagh and J. A. Nelder. *Generalized linear models*. Routledge, 2019.

S. Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680, 2016.

Y. Mroueh and L. Rosasco. Quantization and greed are good: One bit phase retrieval, robustness and greedy refinements. *arXiv preprint arXiv:1312.1830*, 2013.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4): 538–557, 2012.

S. Oymak and B. Recht. Near-optimal bounds for binary embeddings of arbitrary sets. *arXiv preprint arXiv:1512.04433*, 2015.

S. Oymak and M. Soltanolkotabi. Fast and reliable parameter estimation from nonlinear observations. *SIAM Journal on Optimization*, 27(4):2276–2300, 2017.

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

Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

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

Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, 2016.

Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

M. Soltanolkotabi. Learning relus via gradient descent. *Advances in neural information processing systems*, 30, 2017.

M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *IEEE Transactions on Information Theory*, 65 (4):2374–2400, 2019.

C. Thrampoulidis, E. Abbasi, and B. Hassibi. Lasso with non-linear measurements is equivalent to one with linear measurements. *Advances in Neural Information Processing Systems*, 28, 2015.

J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. *Sampling theory, a renaissance*, pages 67–101, 2015.

S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017a.

G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2017b.

C. Xu and L. Jacques. Quantized compressive sensing with rip matrices: The benefit of dithering. *Information and Inference: A Journal of the IMA*, 9(3):543–586, 2020.

H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18, 2017.

A Proof of Theorem 4.1 (Uniform RAIC for a cone)

Our goal is to prove (16), and we only need to bound the two terms in (21). Let us introduce the shorthands $I_1(\boldsymbol{u},\boldsymbol{x}) := \|\boldsymbol{u}-\boldsymbol{x}-\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{u}-\boldsymbol{x})\|_{\mathcal{K}_1^\circ}$ and $I_2(\boldsymbol{x}) := \|\frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i\|_{\mathcal{K}_1^\circ}$. We seek to bound $I_1(\boldsymbol{u},\boldsymbol{x})$ and $I_2(\boldsymbol{x})$ uniformly for all $(\boldsymbol{u},\boldsymbol{x}) \in \mathcal{K}\times\mathcal{X}$ (also recall $\mathcal{X}\subset\mathcal{K}$).

A.1 Controlling $I_1(\mathbf{u}, \mathbf{x})$

For any $\boldsymbol{u} \in \mathcal{K}$ and $\boldsymbol{x} \in \mathcal{X} \subset \mathcal{K}$ such that $\boldsymbol{u} \neq \boldsymbol{x}$, we have $\frac{\boldsymbol{u}-\boldsymbol{x}}{\|\boldsymbol{u}-\boldsymbol{x}\|_2} \in \mathcal{K}_1$ and hence

$$\begin{aligned} I_1(\boldsymbol{u}, \boldsymbol{x}) &= \|\boldsymbol{u}-\boldsymbol{x}\|_2 \cdot \sup_{\boldsymbol{q}\in\mathcal{K}_1} \frac{1}{m}\sum_{i=1}^m (\frac{\boldsymbol{u}-\boldsymbol{x}}{\|\boldsymbol{u}-\boldsymbol{x}\|_2})^T \boldsymbol{a}_i\boldsymbol{a}_i^T\boldsymbol{q} - (\frac{\boldsymbol{u}-\boldsymbol{x}}{\|\boldsymbol{u}-\boldsymbol{x}\|_2})^T\boldsymbol{q} \\ &\le \|\boldsymbol{u}-\boldsymbol{x}\|_2 \cdot \sup_{\boldsymbol{p},\boldsymbol{q}\in\mathcal{K}_1} \Big|\frac{1}{m}\sum_{i=1}^m \boldsymbol{p}^T\boldsymbol{a}_i\boldsymbol{a}_i^T\boldsymbol{q} - \boldsymbol{p}^T\boldsymbol{q}\Big|. \end{aligned} \tag{58}$$

We pause to introduce a concentration inequality for product process.

Lemma A.1 (Outcome of Theorem 1.13 in Mendelson (2016))**.** *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$ be i.i.d. copies of a random vector $\boldsymbol{a}$, and let $\{g_{\boldsymbol{u}}(\boldsymbol{a})\}_{\boldsymbol{u}\in\mathcal{U}}$ and $\{h_{\boldsymbol{v}}(\boldsymbol{a})\}_{\boldsymbol{v}\in\mathcal{V}}$ be real-valued stochastic processes indexed by $\mathcal{U}, \mathcal{V} \subset \mathbb{R}^n$, respectively. Assume that there exist $K_1, K_2, r_1, r_2 \ge 0$ such that*

$$\|g_{\boldsymbol{u}}(\boldsymbol{a}) - g_{\boldsymbol{u}'}(\boldsymbol{a})\|_{\psi_2} \le K_1\|\boldsymbol{u}-\boldsymbol{u}'\|_2, \quad \|g_{\boldsymbol{u}}(\boldsymbol{a})\|_{\psi_2} \le r_1, \quad \forall\, \boldsymbol{u}, \boldsymbol{u}' \in \mathcal{U},$$

and

$$\|h_{\boldsymbol{v}}(\boldsymbol{a}) - h_{\boldsymbol{v}'}(\boldsymbol{a})\|_{\psi_2} \le K_2\|\boldsymbol{v}-\boldsymbol{v}'\|_2, \quad \|h_{\boldsymbol{v}}(\boldsymbol{a})\|_{\psi_2} \le r_2, \quad \forall\, \boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}.$$

Then for every $t \ge 1$, with probability at least $1 - 2\exp(-ct^2)$, it holds that

$$\begin{aligned} &\sup_{\boldsymbol{u}\in\mathcal{U}}\sup_{\boldsymbol{v}\in\mathcal{V}} \Big|\frac{1}{m}\sum_{i=1}^m g_{\boldsymbol{u}}(\boldsymbol{a}_i)h_{\boldsymbol{v}}(\boldsymbol{a}_i) - \mathbb{E}\big[g_{\boldsymbol{u}}(\boldsymbol{a})h_{\boldsymbol{v}}(\boldsymbol{a})\big]\Big| \\ &\le C\bigg(\frac{(K_1\omega(\mathcal{U}) + tr_1)\cdot(K_2\omega(\mathcal{V}) + tr_2)}{m} + \frac{r_1K_2\omega(\mathcal{V}) + r_2K_1\omega(\mathcal{U}) + tr_1r_2}{\sqrt{m}}\bigg). \end{aligned}$$

Continuing from (58), a straightforward application of Lemma A.1 yields

$$\mathbb{P}\bigg(I_1(\boldsymbol{u}, \boldsymbol{x}) \lesssim \sqrt{\frac{\omega^2(\mathcal{K}_1)}{m}}\|\boldsymbol{u}-\boldsymbol{x}\|_2,\, \forall \boldsymbol{u} \in \mathcal{K},\, \boldsymbol{x} \in \mathcal{X}\bigg) \ge 1 - 2\exp(-\omega^2(\mathcal{K}_1)). \tag{59}$$

A.2 Controlling $I_2(\mathbf{x})$

We seek to bound $\sup_{\boldsymbol{x}\in\mathcal{X}} I_2(\boldsymbol{x}) = \sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} \frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}$. We start with

$$\begin{aligned} \Big|\mathbb{E}\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}\Big| &= \Big|\mathbb{E}\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\langle \boldsymbol{q}, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\rangle\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\Big| \\ &\qquad \blacktriangleright \text{ by rotational invariance of } \boldsymbol{a}_i \\ &= \Big|\langle \boldsymbol{q}, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\rangle\Big|\rho(\boldsymbol{x}) \le \rho(\boldsymbol{x}) \\ &\qquad \blacktriangleright \text{ recall } \rho(\boldsymbol{x}) \text{ defined in (18)} \end{aligned} \tag{60}$$

for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{q} \in \mathcal{K}_1$. Therefore,

$$\begin{aligned} \sup_{\boldsymbol{x}\in\mathcal{X}} I_2(\boldsymbol{x}) &= \sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} \frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q} \\ &\le \sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} \bigg\{\frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q} - \mathbb{E}\Big[\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}\Big]\bigg\} + \sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} \mathbb{E}\Big[\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}\Big] \end{aligned}$$

$$\leq \sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\Bigg\{\underbrace{\frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})\boldsymbol{a}_i^T\boldsymbol{q}-\mathbb{E}\Big[\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})\boldsymbol{a}_i^T\boldsymbol{q}\Big]}_{:=J_{\boldsymbol{p},\boldsymbol{q}}}\Bigg\}+\sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x}) \tag{61}$$

We invoke a covering argument to achieve uniform bound on $\boldsymbol{p}\in\mathcal{X}$. For some $\varepsilon>0$ to be chosen, we let $\mathcal{N}_\varepsilon$ be a minimal ε-net of $\mathcal{X}$ and hence $\log|\mathcal{N}_\varepsilon|=\mathscr{H}(\mathcal{X},\varepsilon)$. By (C1) in Assumption 4.2, we use Lemma A.1 to reach

$$\mathbb{P}\bigg(\sup_{\boldsymbol{q}\in\mathcal{K}_1} J_{\boldsymbol{p},\boldsymbol{q}}\lesssim\frac{\varphi_1(t+\omega(\mathcal{K}_1))}{\sqrt{m}}\bigg)\geq 1-2\exp(-t^2),\quad\forall\boldsymbol{p}\in\mathcal{X},\ 0<t\leq\sqrt{m}. \tag{62}$$

We then take a union bound over $p\in\mathcal{N}_\varepsilon$ and set $t=\sqrt{2\mathscr{H}(\mathcal{X},\varepsilon)}+\omega(\mathcal{K}_1)$ to achieve

$$\mathbb{P}\bigg(\sup_{\boldsymbol{p}\in\mathcal{N}_\varepsilon}\sup_{\boldsymbol{q}\in\mathcal{K}_1} J_{\boldsymbol{p},\boldsymbol{q}}\lesssim\frac{\varphi_1(\sqrt{\mathscr{H}(\mathcal{X},\varepsilon)}+\omega(\mathcal{K}_1))}{\sqrt{m}}\bigg)\geq 1-2\exp(-\mathscr{H}(\mathcal{X},\varepsilon)-\omega^2(\mathcal{K}_1)). \tag{63}$$

For any $\boldsymbol{p}\in\mathcal{X}$, we let

$$\hat{\boldsymbol{p}}=\arg\min_{\boldsymbol{w}\in\mathcal{N}_\varepsilon}\|\boldsymbol{w}-\boldsymbol{p}\|_2, \tag{64}$$

which depends on $\boldsymbol{p}$ but such dependence is dropped in notation. By triangle inequality and substituting $J_{\boldsymbol{p},\boldsymbol{q}}$,

$$\begin{aligned}&\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} J_{\boldsymbol{p},\boldsymbol{q}}-\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} J_{\hat{\boldsymbol{p}},\boldsymbol{q}}\leq\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}(J_{\boldsymbol{p},\boldsymbol{q}}-J_{\hat{\boldsymbol{p}},\boldsymbol{q}})\\ &\overset{(60)}{\leq}\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\frac{1}{m}\sum_{i=1}^m\big(\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})-\tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big)\boldsymbol{a}_i^T\boldsymbol{q}+2\sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x})\end{aligned} \tag{65}$$

All that remains is to bound

$$G:=\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\frac{1}{m}\sum_{i=1}^m\big(\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})-\tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big)\boldsymbol{a}_i^T\boldsymbol{q}.$$

For some $\eta\in(0,\frac{\varphi_2}{2})$ and any $\boldsymbol{x}\in\mathcal{X}$, we define

$$\mathcal{J}_{\boldsymbol{x},\eta}:=\{i\in[m]:\operatorname{dist}(\boldsymbol{a}_i^T\boldsymbol{x},\mathcal{D}_{f_i})\leq\eta\}. \tag{66}$$

By (C5) in Assumption 4.2 and a Chernoff bound for binomial variable,

$$\mathbb{P}\big(|\mathcal{J}_{\boldsymbol{x},\eta}|\leq 2\varphi_5\eta m\big)\geq 1-\exp(-c'\varphi_5\eta m). \tag{67}$$

Under the condition

$$\varphi_5\eta m\gtrsim\mathscr{H}(\mathcal{X},\varepsilon), \tag{68}$$

a union bound over $\boldsymbol{x}\in\mathcal{N}_\varepsilon$ yields

$$\mathbb{P}\bigg(\sup_{\boldsymbol{x}\in\mathcal{N}_\varepsilon}|\mathcal{J}_{\boldsymbol{x},\eta}|\leq 2\varphi_5\eta m\bigg)\geq 1-\exp(-\frac{c'\varphi_5\eta m}{2}). \tag{69}$$

For $\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}$, we also define

$$\mathcal{G}_{\boldsymbol{x},\boldsymbol{x}',\eta}:=\Big\{i\in[m]:|\boldsymbol{a}_i^T(\boldsymbol{x}-\boldsymbol{x}')|\geq\frac{\eta}{2}\Big\}. \tag{70}$$

We pause to introduce a useful concentration bound.

Lemma A.2 (Theorem 2.10 in Dirksen and Mendelson (2021)). *Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m$ be i.i.d. copies of $N(0, I_n)$, and let $\mathcal{U} \subset \mathbb{R}^n$. For any $\zeta \in \{\frac{1}{m}, \frac{2}{m}, \cdots, \frac{m-1}{m}, 1\}$, the event*

$$\sup_{\boldsymbol{u}\in\mathcal{U}} \max_{\substack{\mathcal{G}\subset[m]\\|\mathcal{G}|=\zeta m}} \left(\frac{1}{\eta m}\sum_{i\in I}|\boldsymbol{a}_i^T\boldsymbol{u}|^2\right)^{1/2} \lesssim \frac{\omega(\mathcal{U})}{\sqrt{\zeta m}} + \Big(\sup_{\boldsymbol{u}\in\mathcal{U}}\|\boldsymbol{u}\|_2\Big)\sqrt{\log\frac{e}{\zeta}}$$

holds with probability at least $1 - 2\exp(-c\zeta m \log(\frac{e}{\zeta}))$.

For any $\zeta \in \{\frac{i}{m} : i \in [m]\}$, under Assumption 4.1, Lemma A.2 yields

$$\sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon} \max_{\substack{\mathcal{I}\subset[m]\\|\mathcal{I}|\le\zeta m}} \left(\frac{1}{\zeta m}\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{w}|^2\right)^{1/2} \lesssim \frac{\omega(\mathcal{X}_\varepsilon)}{\sqrt{\zeta m}} + \varepsilon\sqrt{\log(e/\zeta)}, \quad \text{w.p. } \ge 1 - 2\exp(-c''\zeta m\log(e/\zeta))$$

We enforce the condition

$$\frac{\omega(\mathcal{X}_\varepsilon)}{\sqrt{\zeta m}} + \varepsilon\sqrt{\log(e/\zeta)} \lesssim \eta \ \text{ with small enough hidden constant} \tag{71}$$

so that

$$\mathbb{P}\bigg(\sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon} \max_{\substack{\mathcal{I}\subset[m]\\|\mathcal{I}|\le\zeta m}} \Big(\frac{1}{\zeta m}\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{w}|^2\Big)^{1/2} \le \frac{\eta}{4}\bigg) \ge 1 - 2\exp(-c''\zeta m\log(e/\zeta)). \tag{72}$$

By $\mathcal{X} \subset \mathcal{K}$ for a cone $\mathcal{K}$, $\mathcal{X}_\varepsilon \subset \varepsilon\mathcal{K}_1$, (71) can be ensured by

$$\frac{\varepsilon\cdot\omega(\mathcal{K}_1)}{\sqrt{\zeta m}} + \varepsilon\sqrt{\log(e/\zeta)} \lesssim \eta \ \text{ with small enough hidden constant.} \tag{73}$$

Since for any $\boldsymbol{p} \in \mathcal{X}$ we have $\boldsymbol{p} - \hat{\boldsymbol{p}} \in \mathcal{X}_\varepsilon$ by (64), the event in (72) implies

$$|\mathcal{G}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}| \le \zeta m\,, \quad \forall\, \boldsymbol{p}\in\mathcal{X}; \tag{74}$$

otherwise, if $|\mathcal{G}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}| > \zeta m$ for some $\boldsymbol{p} \in \mathcal{X}$, then we can find $\hat{\mathcal{G}} \subset \mathcal{G}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}$ such that $|\hat{\mathcal{G}}| = \zeta m$, then by the definition of $\mathcal{G}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}$,

$$\left(\frac{1}{\zeta m}\sum_{i\in\hat{\mathcal{G}}}|\boldsymbol{a}_i^T(\boldsymbol{p}-\hat{\boldsymbol{p}})|^2\right)^{1/2} \ge \frac{\eta}{2},$$

which contradicts (72).

In summary, we can assume that the events (69) and (74) hold with probability at least

$$1 - \exp(-c'\varphi_5\eta m) - 2\exp(-c'\zeta m\log(e/\zeta)). \tag{75}$$

We now let

$$\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta} := \mathcal{J}_{\hat{\boldsymbol{p}},\eta} \cup \mathcal{G}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}\,, \quad \boldsymbol{p}\in\mathcal{X} \tag{76}$$

which, in view of $\hat{\boldsymbol{p}} \in \mathcal{N}_\varepsilon$, has small cardinality

$$|\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}| \le |\mathcal{J}_{\hat{\boldsymbol{p}},\eta}| + |\mathcal{G}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}| \le (2\varphi_5\eta + \zeta)m\,, \quad \forall \boldsymbol{p}\in\mathcal{X}. \tag{77}$$

To bound G, we separately treat the measurements in $\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}$ and $\bar{\mathcal{I}}^c_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta} := [m] \setminus \bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}$:

$$G \le \underbrace{\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\frac{1}{m}\sum_{i\in\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}}\big(\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p}) - \tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big)\boldsymbol{a}_i^T\boldsymbol{q}}_{:=G_1} + \underbrace{\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\frac{1}{m}\sum_{i\in\bar{\mathcal{I}}^c_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}}\big(\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p}) - \tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big)\boldsymbol{a}_i^T\boldsymbol{q}}_{:=G_2}\,. \tag{78}$$

It remains to bound G_1 and G_2 separately.

A.2.1 Bounding G_1

We shall start with

$$
\begin{aligned}
G_1 &\le \sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\frac{1}{m}\sum_{i\in\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}}\big|\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})-\tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big||\boldsymbol{a}_i^T\boldsymbol{q}| \\
&\quad \blacktriangleright \text{ by triangle inequality} \\
&\le \sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\frac{1}{m}\sum_{i\in\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}}\Big((\varphi_4+\frac{\varphi_3}{\mu\varphi_2})|\boldsymbol{a}_i^T(\boldsymbol{p}-\hat{\boldsymbol{p}})|+\frac{\varphi_3}{\mu}\Big)|\boldsymbol{a}_i^T\boldsymbol{q}| \\
&\quad \blacktriangleright \text{ by Equation (20)} \\
&\le (\varphi_4+\frac{\varphi_3}{\mu\varphi_2})\sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\max_{\substack{\mathcal{I}\subset[m]\\|\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}}\frac{1}{m}\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{w}||\boldsymbol{a}_i^T\boldsymbol{q}|+\sup_{\boldsymbol{q}\in\mathcal{K}_1}\max_{\substack{\mathcal{I}\subset[m]\\|\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}}\frac{\varphi_3}{\mu}\frac{1}{m}\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{q}| \\
&\quad \blacktriangleright \text{ by } \boldsymbol{p}-\hat{\boldsymbol{p}}\in\mathcal{X}_\varepsilon \text{ and Equation (77)} \\
&:= G_{11}+G_{12}.
\end{aligned}
\tag{79}
$$

To bound G_{11},

$$
\begin{aligned}
G_{11} &\le \Big(\varphi_4+\frac{\varphi_3}{\mu\varphi_2}\Big)\bigg(\sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon}\max_{|\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}\frac{1}{m}\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{w}|^2\bigg)^{1/2}\bigg(\sup_{\boldsymbol{q}\in\mathcal{K}_1}\max_{|\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}\frac{1}{m}\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{q}|^2\bigg)^{1/2} \\
&\quad \blacktriangleright \text{ by Cauchy-Schwarz inequality} \\
&\lesssim \Big(\varphi_4+\frac{\varphi_3}{\mu\varphi_2}\Big)\bigg(\frac{\omega(\mathcal{X}_\varepsilon)}{\sqrt{m}}+\varepsilon\sqrt{\varphi_5\eta\log\Big(\frac{1}{\varphi_5\eta}\Big)+\zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg)\bigg(\frac{\omega(\mathcal{K}_1)}{\sqrt{m}}+\sqrt{\varphi_5\eta\log\Big(\frac{1}{\varphi_5\eta}\Big)+\zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg) \\
&\quad \blacktriangleright \text{ by Lemma A.2, this holds w.p. } \ge 1-4\exp(-c'\zeta m\log(e/\zeta)) \\
&\lesssim \varepsilon\Big(\varphi_4+\frac{\varphi_3}{\mu\varphi_2}\Big)\bigg(\frac{\omega^2(\mathcal{K}_1)}{m}+\varphi_5\eta\log\Big(\frac{1}{\varphi_5\eta}\Big)+\zeta\log\Big(\frac{1}{\zeta}\Big)\bigg). \\
&\quad \blacktriangleright \text{ by } \mathcal{X}\subset\mathcal{K} \text{ and hence } \mathcal{X}_\varepsilon\subset\mathcal{K}_\varepsilon=\varepsilon\mathcal{K}_1
\end{aligned}
\tag{80}
$$

Similarly,

$$
\begin{aligned}
G_{12} &\le \frac{\varphi_3}{\mu}\sqrt{\frac{2\varphi_5\eta+\zeta}{m}}\sup_{\boldsymbol{q}\in\mathcal{K}_1}\max_{|\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}\bigg(\sum_{i\in\mathcal{I}}|\boldsymbol{a}_i^T\boldsymbol{q}|^2\bigg)^{1/2} \\
&\lesssim \frac{\varphi_3\sqrt{\varphi_5\eta+\zeta}}{\mu}\bigg(\frac{\omega(\mathcal{K}_1)}{\sqrt{m}}+\sqrt{\varphi_5\eta\log\Big(\frac{1}{\varphi_5\eta}\Big)}+\sqrt{\zeta\log\Big(\frac{1}{\zeta}\Big)}\bigg)
\end{aligned}
\tag{81}
$$

with probability at least $1-2\exp(-c'\zeta m\log(e/\zeta))$.

A.2.2 Bounding G_2

We now bound G_2. For $i\in\bar{\mathcal{I}}^c_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}$, we have $\mathrm{dist}(\boldsymbol{a}_i^T\hat{\boldsymbol{p}},f_i)\ge\eta$ and $|\boldsymbol{a}_i^T\boldsymbol{p}-\boldsymbol{a}_i^T\hat{\boldsymbol{p}}|<\frac{\eta}{2}$, which implies

$$
(\min\{\boldsymbol{a}_i^T\boldsymbol{p},\boldsymbol{a}_i^T\boldsymbol{q}\},\max\{\boldsymbol{a}_i^T\boldsymbol{p},\boldsymbol{a}_i^T\boldsymbol{q}\})\cap\mathcal{D}_{f_i}=\varnothing,
$$

and therefore (C4) in Assumption 4.2 yields

$$
|\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})-\tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})|\le\varphi_4|\boldsymbol{a}_i^T(\boldsymbol{p}-\hat{\boldsymbol{p}})|. \tag{82}
$$

Thus,

$$
\begin{aligned}
G_2 &\le \sup_{\boldsymbol{q}\in\mathcal{K}_1} \frac{1}{m} \sum_{i\in\bar{\mathcal{I}}^c_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}} \varphi_4 |\boldsymbol{a}_i^T(\boldsymbol{p}-\hat{\boldsymbol{p}})||\boldsymbol{a}_i^T\boldsymbol{q}| \\
&\quad \blacktriangleright \text{ by Equation (82)} \\
&\le \sup_{\boldsymbol{q}\in\mathcal{K}_1} \sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon} \frac{1}{m} \sum_{i=1}^m \varphi_4 |\boldsymbol{a}_i^T\boldsymbol{w}\boldsymbol{a}_i^T\boldsymbol{q}| \\
&\quad \blacktriangleright \text{ by } \boldsymbol{p}-\hat{\boldsymbol{p}} \in \mathcal{X}_\varepsilon \\
&\le \varphi_4 \sup_{\boldsymbol{q}\in\mathcal{K}_1} \left(\frac{1}{m}\sum_{i=1}^m |\boldsymbol{a}_i^T\boldsymbol{q}|^2\right)^{1/2} \sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon} \left(\frac{1}{m}\sum_{i=1}^m |\boldsymbol{a}_i^T\boldsymbol{w}|^2\right)^{1/2} \\
&\quad \blacktriangleright \text{ by Cauchy-Schwarz inequality} \\
&\lesssim \varphi_4 \left(\frac{\omega(\mathcal{K}_1)}{\sqrt{m}}+1\right)\left(\frac{\omega(\mathcal{X}_\varepsilon)}{\sqrt{m}}+\varepsilon\right) \\
&\quad \blacktriangleright \text{ by Lemma A.2, w.p. } \ge 1-2\exp(-c'm) \\
&\lesssim \varepsilon\varphi_4. \qquad (83) \\
&\quad \blacktriangleright \text{ by } \mathcal{X}_\varepsilon \subset \mathcal{K}_\varepsilon = \varepsilon\mathcal{K}_1 \text{ and } m \gtrsim \omega^2(\mathcal{K}_1)
\end{aligned}
$$

A.3 Putting pieces together

We choose η such that

$$
\varphi_5\eta \asymp \frac{\mathscr{H}(\mathcal{X},\varepsilon)}{m} + \frac{\varphi_5\varepsilon\omega(\mathcal{K}_1)}{\sqrt{\zeta m}} + \varphi_5\varepsilon\sqrt{\log(e/\zeta)} := \overline{\Xi}
$$

to fulfill (68) and (73). Under small enough $\varphi_5\varepsilon\sqrt{\log(1/\zeta)}$ and (22), $\overline{\Xi}$ is small enough. Recall that we have established yield

$$
G_{11} \lesssim \varepsilon\Big(\varphi_4 + \frac{\varphi_3}{\mu\varphi_2}\Big)\left(\frac{\omega^2(\mathcal{K}_1)}{m} + \overline{\Xi}\log\left(\frac{1}{\overline{\Xi}}\right) + \zeta\log\left(\frac{1}{\zeta}\right)\right), \tag{84}
$$

$$
G_{12} \lesssim \frac{\varphi_3\sqrt{\overline{\Xi}+\zeta}}{\mu}\left(\frac{\omega(\mathcal{K}_1)}{\sqrt{m}} + \sqrt{\overline{\Xi}\log\left(\frac{1}{\overline{\Xi}}\right)} + \sqrt{\zeta\log\left(\frac{1}{\zeta}\right)}\right) \tag{85}
$$

We now substitute (84) and (85) into the decomposition $G_1 \le G_{11}+G_{12}$ to yield a bound on G_1. Combining with $G_2 \lesssim \varepsilon\varphi_4$ from (83), $G \le G_1+G_2$, and the fact that $\frac{\omega^2(\mathcal{K}_1)}{m} + \overline{\Xi}\log(\frac{1}{\overline{\Xi}}) + \zeta\log(\frac{1}{\zeta})$ is small enough, we obtain

$$
G \lesssim \varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\cdot\Xi
$$

where, as per (23)–(25),

$$
\begin{aligned}
&\Xi := \sqrt{\overline{\Xi}+\zeta}\Big(\frac{\omega(\mathcal{K}_1)}{\sqrt{m}} + \sqrt{\overline{\Xi}\log(\frac{1}{\overline{\Xi}})} + \sqrt{\zeta\log(\frac{1}{\zeta})}\Big) + \frac{\varepsilon}{\varphi_2}\Big(\frac{\omega^2(\mathcal{K}_1)}{m} + \overline{\Xi}\log(\frac{1}{\overline{\Xi}}) + \zeta\log(\frac{1}{\zeta})\Big), \\
&\text{where} \quad \overline{\Xi} := \frac{\mathscr{H}(\mathcal{X},\varepsilon)}{m} + \frac{\varphi_5\varepsilon\omega(\mathcal{K}_1)}{\sqrt{\zeta m}} + \varphi_5\varepsilon\sqrt{\log(e/\zeta)}
\end{aligned}
$$

In light of (65), we have

$$\begin{aligned}\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} J_{\boldsymbol{p},\boldsymbol{q}} &\le \sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\mathcal{K}_1} J_{\hat{\boldsymbol{p}},\boldsymbol{q}} + 2\sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x}) + O\Big(\varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\Xi\Big)\\ &\lesssim \sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x}) + \varphi_1\sqrt{\frac{\omega^2(\mathcal{K}_1)+\mathscr{H}(\mathcal{X},\varepsilon)}{m}} + \varepsilon\varphi_4 + \frac{\varphi_3}{\mu}\Xi\\ &\quad \blacktriangleright \text{ by } \hat{\boldsymbol{p}}\in\mathcal{N}_\varepsilon \text{ and Equation (63)}\end{aligned}$$

Substituting this into (61) yields the same bound for I_2, then combining with (59) and (21) completes the proof.

B Proof of Theorem 4.2 (Uniform RAIC for a convex set)

The proof closely follows that of Theorem 4.1, except that some steps need adaptations since $\mathcal{K}$ is not a cone. We omit some of the details that are parallel to Theorem 4.1. For any $\boldsymbol{u}\in\mathcal{K}$ and $\boldsymbol{x}\in\mathcal{X}\subset\mathcal{K}$, we start with the decomposition

$$\begin{aligned}&\frac{1}{\phi}\Big\|\boldsymbol{u}-\boldsymbol{x}-\frac{1}{m}\sum_{i=1}^m\big(\boldsymbol{a}_i^T\boldsymbol{u}-\frac{f_i(\boldsymbol{a}_i^T\boldsymbol{x})}{\mu}\big)a_i\Big\|_{\mathcal{K}_{\boldsymbol{x},\phi}^\circ}\\ &\le \underbrace{\frac{1}{\phi}\Big\|\boldsymbol{u}-\boldsymbol{x}-\frac{1}{m}\sum_{i=1}^m\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{u}-\boldsymbol{x})\Big\|_{\mathcal{K}_{\boldsymbol{x},\phi}^\circ}}_{:=I_1(\boldsymbol{u},\boldsymbol{x})} + \underbrace{\frac{1}{\phi}\Big\|\frac{1}{m}\sum_{i=1}^m\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i\Big\|_{\mathcal{K}_{\boldsymbol{x},\phi}^\circ}}_{:=I_2(\boldsymbol{x})}.\end{aligned}\tag{86}$$

We seek to bound $I_1(\boldsymbol{u},\boldsymbol{x})$ and $I_2(\boldsymbol{x})$ uniformly for all $(\boldsymbol{u},\boldsymbol{x})\in\mathcal{K}\times\mathcal{X}$.

B.1 Controlling $I_1(\mathbf{u},\mathbf{x})$

For any $(\boldsymbol{u},\boldsymbol{x})\in\mathcal{K}\times\mathcal{X}$,

$$I_1(\boldsymbol{u},\boldsymbol{x}) = \sup_{\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\boldsymbol{x},\phi}}\Big|\frac{1}{m}\sum_{i=1}^m(\boldsymbol{u}-\boldsymbol{x})^T\boldsymbol{a}_i\boldsymbol{a}_i^T\boldsymbol{q}-(\boldsymbol{u}-\boldsymbol{x})^T\boldsymbol{q}\Big| \le I_1'\phi + I_1'\|\boldsymbol{u}-\boldsymbol{x}\|_2,$$

where

$$\begin{aligned}I_1' &:= \sup_{\boldsymbol{p},\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}}\Big|\frac{1}{m}\sum_{i=1}^m\boldsymbol{p}^T\boldsymbol{a}_i\boldsymbol{a}_i^T\boldsymbol{q}-\boldsymbol{p}^T\boldsymbol{q}\Big|,\\ \mathcal{K}_{\mathcal{X},\phi} &= \bigcup_{\boldsymbol{x}\in\mathcal{X}}\mathcal{K}_{\boldsymbol{x},\phi} = (\mathcal{K}-\mathcal{X})\cap\phi\mathbb{B}_2^n.\end{aligned}$$

This can be seen by noticing that $\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}$ always holds and discussing the following two cases:

- If $\|\boldsymbol{u}-\boldsymbol{x}\|_2\le\phi$, then we have $\boldsymbol{u}-\boldsymbol{x}\in\mathcal{K}_{\mathcal{X},\phi}$ and therefore $\frac{\boldsymbol{u}-\boldsymbol{x}}{\phi}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}$;
- If $\|\boldsymbol{u}-\boldsymbol{x}\|_2\ge\phi$, then due to being $\mathcal{K}-\mathcal{X}$ star-shaped,

$$\frac{\boldsymbol{u}-\boldsymbol{x}}{\|\boldsymbol{u}-\boldsymbol{x}\|_2}\in\frac{\mathcal{K}-\mathcal{X}}{\|\boldsymbol{u}-\boldsymbol{x}\|_2}\cap\mathbb{S}^{n-1}\subset\frac{\mathcal{K}-\mathcal{X}}{\phi}\cap\mathbb{S}^{n-1}\subset\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}.$$

Under $m \gtrsim \phi^{-2}\omega^2(\mathcal{K}_{\mathcal{X},\phi})$, Lemma A.1 yields

$$I_1' \lesssim \frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}}$$

with probability at least $1 - 2\exp(-c'\phi^{-2}\omega^2(\mathcal{K}_{\mathcal{X},\phi}))$, and therefore with the same probability

$$I_1(\boldsymbol{u},\boldsymbol{x}) \lesssim \frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}}\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \frac{\omega(\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}}, \quad \forall(\boldsymbol{u},\boldsymbol{x}) \in \mathcal{K}\times\mathcal{X}. \tag{87}$$

B.2 Controlling $I_2(\mathbf{x})$

To bound

$$I_2 = \sup_{\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\boldsymbol{x},\phi}} \frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q} \le \sup_{\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}} \frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}$$

for all $\boldsymbol{x}\in\mathcal{X}$, we use a covering argument parallel to the corresponding part in the proof of Theorem 4.1, with the following two differences: first, due to the difference between Definitions 16 and 17, the range of $\boldsymbol{q}$, which is $\boldsymbol{q}\in\mathcal{K}_1$ in the previous proof, should be replaced by $\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}$; second, some steps that rely on the conic structure of $\mathcal{K}$, such as (80) and (83), should be revised.

To start, we revisit the argument in (61) to obtain

$$\sup_{\boldsymbol{x}\in\mathcal{X}} I_2(\boldsymbol{x}) \le \sup_{\boldsymbol{p}\in\mathcal{X}} \sup_{\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}} J_{\boldsymbol{p},\boldsymbol{q}} + \sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x}) \tag{88}$$

where

$$J_{\boldsymbol{p},\boldsymbol{q}} := \frac{1}{m}\sum_{i=1}^m \tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})\boldsymbol{a}_i^T\boldsymbol{q} - \mathbb{E}[\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p})\boldsymbol{a}_i^T\boldsymbol{q}].$$

For some $\varepsilon\in(0,1)$, we let $\mathcal{N}_\varepsilon$ be the minimal ε-net of $\mathcal{X}$, and find $\hat{\boldsymbol{p}}$ by (64) for each $\boldsymbol{p}\in\mathcal{X}$. We now revisit the argument in (62)–(63) to establish

$$\mathbb{P}\Big(\sup_{\boldsymbol{p}\in\mathcal{N}_\varepsilon}\sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} J_{\boldsymbol{p},\boldsymbol{q}} \lesssim \frac{\varphi_1(\sqrt{\mathscr{H}(\mathcal{X},\varepsilon)}+\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}))}{\sqrt{m}}\Big) \ge 1 - 2\exp\Big(-\mathscr{H}(\mathcal{X},\varepsilon) - \omega^2(\frac{\omega(\mathcal{K}_{\mathcal{X},\phi})}{\phi})\Big). \tag{89}$$

Then, the argument in (65) with $\boldsymbol{q}\in\mathcal{K}_1$ being replaced with $\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\boldsymbol{x},\phi}$ then yields

$$\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} J_{\boldsymbol{p},\boldsymbol{q}} \le \sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} J_{\hat{\boldsymbol{p}},\boldsymbol{q}} + \underbrace{\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} \frac{1}{m}\sum_{i=1}^m \big(\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p}) - \tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big)\boldsymbol{a}_i^T\boldsymbol{q}}_{:=G} + 2\sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x}). \tag{90}$$

All that is left is to bound G.

We shall pause to make some preparations. For $\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}$ and small $\eta\in(0,\frac{\varphi_2}{2})$, we define the index sets $\mathcal{J}_{\boldsymbol{x},\eta}$ and $\mathcal{G}_{\boldsymbol{x},\boldsymbol{x}',\eta}$ as in (66) and (70), respectively, and further define $\bar{\mathcal{I}}_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}$ as in (76). By the arguments in Equations (67)–(77), under the conditions (68) and (71), the event (77) holds with probability in (75).

To bound G, we revisit the arguments in (78) and (79) to establish

$$
\begin{aligned}
G \le (\varphi_4 + \frac{\varphi_3}{\mu\varphi_2}) \sup_{\boldsymbol{w}\in\mathcal{X}_\varepsilon} \sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} \max_{\substack{\mathcal{I}\subset[m]\\ |\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}} \frac{1}{m}\sum_{i\in\mathcal{I}} |\boldsymbol{a}_i^T\boldsymbol{w}||\boldsymbol{a}_i^T\boldsymbol{q}| \\
\blacktriangleright \text{ corresponding to } G_{11} \text{ in (79)} \\
+ \sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} \max_{\substack{\mathcal{I}\subset[m]\\ |\mathcal{I}|\le(2\varphi_5\eta+\zeta)m}} \frac{\varphi_3}{\mu}\frac{1}{m}\sum_{i\in\mathcal{I}} |\boldsymbol{a}_i^T\boldsymbol{q}| \\
\blacktriangleright \text{ corresponding to } G_{12} \text{ in (79)} \\
+ \sup_{\boldsymbol{p}\in\mathcal{X}} \sup_{\boldsymbol{q}\in\frac{\mathcal{K}_{\mathcal{X},\phi}}{\phi}} \frac{1}{m}\sum_{i\in\bar{\mathcal{I}}^c_{\boldsymbol{p},\hat{\boldsymbol{p}},\eta}} \big(\tilde{f}_i(\boldsymbol{a}_i^T\boldsymbol{p}) - \tilde{f}_i(\boldsymbol{a}_i^T\hat{\boldsymbol{p}})\big)\boldsymbol{a}_i^T\boldsymbol{q} \\
\blacktriangleright \text{ corresponding to } G_2 \text{ in (78)} \\
:= G_{11} + G_{12} + G_2. \qquad (91)
\end{aligned}
$$

We shall bound these terms separately. To bound G_{11}, we revisit the argument in Equation (80) to obtain

$$
\begin{aligned}
G_{11} \lesssim (\varphi_4 + \frac{\varphi_3}{\mu\varphi_2})\bigg(\frac{\omega(\mathcal{X}_\varepsilon)}{\sqrt{m}} + \varepsilon\sqrt{\varphi_5\eta\log(\frac{1}{\varphi_5\eta}) + \zeta\log(\frac{1}{\zeta})}\bigg) \\
\cdot\bigg(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \sqrt{\varphi_5\eta\log(\frac{1}{\varphi_5\eta}) + \zeta\log(\frac{1}{\zeta})}\bigg) \qquad (92)
\end{aligned}
$$

with probability at least $1 - 4\exp(-c'\zeta m\log(e/\zeta))$.

Similarly, as with (81),

$$
G_{12} \lesssim \frac{\varphi_3\sqrt{\varphi_5\eta+\zeta}}{\mu}\bigg(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \sqrt{\varphi_5\eta\log(\frac{1}{\varphi_5\eta})} + \sqrt{\zeta\log(\frac{1}{\zeta})}\bigg) \qquad (93)
$$

with probability at least $1 - 2\exp(-c'\zeta m\log(e/\zeta))$.

To bound G_2, we revisit (83) without the final inequality therein that relies on $\mathcal{K}$ being a cone, yielding

$$
\mathbb{P}\bigg(G_2 \lesssim \varphi_4\big(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + 1\big)\big(\frac{\omega(\mathcal{X}_\varepsilon)}{\sqrt{m}} + \varepsilon\big)\,,\ \forall \boldsymbol{p}\in\mathcal{X}\bigg) \ge 1 - 2\exp(-c'm). \qquad (94)
$$

B.3 Putting pieces together

Suppose

$$
m \gtrsim \omega^2(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}) + \omega^2(\varepsilon^{-1}\mathcal{X}_\varepsilon),
$$

then (94) implies $G_2 \lesssim \varphi_4\varepsilon$ for all $\boldsymbol{p}\in\mathcal{X}$. We shall choose η before proceeding. Recall that conditions (68) and (71) are needed to support our analysis, so we set the minimal η such that

$$
\varphi_5\eta \asymp \frac{\mathscr{H}(\mathcal{X},\varepsilon)}{m} + \frac{\varphi_5\omega(\mathcal{X}_\varepsilon)}{\sqrt{\zeta m}} + \varphi_5\varepsilon\sqrt{\log(e/\zeta)}. \qquad (95)
$$

We suppose $\zeta m \gtrsim \varphi_5^2\omega^2(\mathcal{X}_\varepsilon)$ and $\varphi_5\varepsilon\sqrt{\log(e/\zeta)}$ is small enough to guarantee small enough $\varphi_5\eta$. Substituting the bounds in (92), (93) and (94) into Equations (91), along with using $\overline{\Upsilon}$ to replace the choice of $\varphi_5\eta$ in (95), establishes

$$
G \lesssim \varphi_4\varepsilon + \frac{\varphi_3}{\mu}\Upsilon, \qquad (96)
$$

where, as per Equations (27)–(29),

$$\Upsilon := \frac{\varepsilon}{\varphi_2}\Big(\frac{\omega(\varepsilon^{-1}\mathcal{X}_\varepsilon)}{\sqrt{m}} + \sqrt{\overline{\Upsilon}\log(\frac{1}{\overline{\Upsilon}}) + \zeta\log(\frac{1}{\zeta})}\Big)\Big(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \sqrt{\overline{\Upsilon}\log(\frac{1}{\overline{\Upsilon}}) + \zeta\log(\frac{1}{\zeta})}\Big)$$
$$+ \sqrt{\overline{\Upsilon}+\zeta}\Big(\frac{\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi})}{\sqrt{m}} + \sqrt{\overline{\Upsilon}\log(\frac{1}{\overline{\Upsilon}}) + \zeta\log(\frac{1}{\zeta})}\Big)$$
$$\text{where } \overline{\Upsilon} := \frac{\mathscr{H}(\mathcal{X},\varepsilon)}{m} + \frac{\varphi_5\omega(\mathcal{X}_\varepsilon)}{\sqrt{\zeta m}} + \varphi_5\varepsilon\sqrt{\log(e/\zeta)}.$$

Substituting (96) into (90) and using (89) to bound $\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}} J_{\hat{\boldsymbol{p}},\boldsymbol{q}}$ yields a bound on $\sup_{\boldsymbol{p}\in\mathcal{X}}\sup_{\boldsymbol{q}\in\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}} J_{\boldsymbol{p},\boldsymbol{q}}$. Further combining with (88), we obtain

$$\sup_{\boldsymbol{x}\in\mathcal{X}} I_2(\boldsymbol{x}) \lesssim \frac{\varphi_1(\sqrt{\mathscr{H}(\mathcal{X},\varepsilon)} + \omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}))}{\sqrt{m}} + \varphi_4\varepsilon + \frac{\varphi_3}{\mu}\Upsilon + \sup_{\boldsymbol{x}\in\mathcal{X}}\rho(\boldsymbol{x}).$$

Combining with (87) and (86) finishes the proof.

C Proof of Theorem 4.5 (Uniform sparse recovery from modulo measurements)

Note that Assumption 4.1 holds trivially. In light of

$$m_\lambda(v) = \sum_{k\in\mathbb{Z}}(v-2\lambda k)\mathbb{1}\big(\lambda(2k-1)\le v<\lambda(2k+1)\big) = v - 2\lambda\sum_{k\in\mathbb{Z}} k\cdot\mathbb{1}\big(\lambda(2k-1)\le v<\lambda(2k+1)\big), \tag{97}$$

we have

$$\begin{aligned}\mathbb{E}\Big[g m_\lambda(g)\Big] &= \mathbb{E}\Big[g^2 - 2\lambda\sum_{k\in\mathbb{Z}} kg\mathbb{1}\big(\lambda(2k-1)\le g<\lambda(2k+1)\big)\Big]\\ &= 1 - 2\lambda\sum_{k\in\mathbb{Z}}\mathbb{E}\Big[kg\mathbb{1}\big(\lambda(2k-1)\le g<\lambda(2k+1)\big)\Big] = 1.\end{aligned}$$

Thus, Assumption 4.3 holds and the choice in (31) is $\mu = 1$.

We now validate Assumption 4.2. By $|m_\lambda(v)|\le|v|$ and $\tilde{f}_i = \mathrm{Id} - m_\lambda$, for $g\sim N(0,1)$,

$$\|\tilde{f}_i(g)\|_{\psi_2} \le \|g\|_{\psi_2} + \|m_\lambda(g)\|_{\psi_2} \le 2\||g|\|_{\psi_2} = O(1).$$

Thus, (C1) of Assumption 4.2 holds for $\varphi_1 = O(1)$. Since the points of discontinuity of m_λ lie in $\{\lambda(2k-1) : k\in\mathbb{Z}\}$, so φ_2 in (C2) of Assumption 4.2 satisfies $\varphi_2\ge 2\lambda\ge\frac{1}{2}$. Moreover, (C3) of Assumption 4.2 holds with $\varphi_3 = 2\lambda$, and in light of (97), (C4) of Assumption 4.2 holds with $\varphi_4 = 0$. By $\boldsymbol{a}_i^T\boldsymbol{x}\sim N(0,1)$ and $\mathcal{D}_{f_i} = \{(2k-1)\lambda : k\in\mathbb{Z}\}$, for $g\sim N(0,1)$ and any $t\in(0,\frac{\lambda}{2})$,

$$\begin{aligned}\mathbb{P}\big(\mathrm{dist}(\boldsymbol{a}_i^T\boldsymbol{x},\mathcal{D}_{f_i})\le t\big) &= \sum_{k\in\mathbb{Z}}\mathbb{P}\Big(g\in[(2k-1)\lambda - t,(2k-1)\lambda+t]\Big)\\ &= 2\sum_{k=1}^{\infty}\int_{(2k-1)\lambda-t}^{(2k-1)\lambda+t}\frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2})\,dt\\ &\le \frac{4t}{\sqrt{2\pi}}\sum_{k=1}^{\infty}\exp\Big(-\frac{(2k-3/2)^2\lambda^2}{2}\Big)\end{aligned}$$

$$\leq \frac{4t}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \exp\Big(-\frac{k\lambda^2}{8}\Big) = \sqrt{\frac{8}{\pi}} \frac{\exp(-\lambda^2/8)t}{1-\exp(-\lambda^2/8)}$$

By $\lambda \geq \frac{1}{4}$, (C5) in Assumption 4.2 holds with $\varphi_5 = O(\exp(-\lambda^2/8))$. It remains to apply Theorems 4.3 and 4.4 to the two settings (48) and (49), separately. Before proceeding, we note the following: for $\mathcal{X} = \Sigma_k^{n,*}$ or $\Sigma_k^{n,*} \cap \{\boldsymbol{x} : \|\boldsymbol{x}\|_1 = c_*\sqrt{k}\}$ and any $\varepsilon > 0$,

$$\mathscr{H}(\mathcal{X}, \varepsilon) \leq k \log(\frac{Cn}{\varepsilon k}), \tag{98}$$

$$\omega(\varepsilon^{-1}\mathcal{X}_\varepsilon) \lesssim \sqrt{k \log(\frac{en}{k})}\,; \tag{99}$$

for $\mathcal{K} = \Sigma_k^n$,

$$\omega(\mathcal{K}_1) \lesssim \sqrt{k \log(\frac{en}{k})}; \tag{100}$$

for $\mathcal{X} = \Sigma_k^{n,*} \cap \{\boldsymbol{x} : \|\boldsymbol{x}\|_1 = c_*\sqrt{k}\}$ and $\mathcal{K} = \mathbb{B}_1^n(c_*\sqrt{k})$,

$$\omega(\phi^{-1}\mathcal{K}_{\mathcal{X},\phi}) \leq \omega(\mathrm{cone}(\mathcal{K}_{\mathcal{X}}) \cap \mathbb{B}_2) \lesssim \sqrt{k \log(\frac{en}{k})}\,, \quad \forall \phi > 0. \tag{101}$$

See Plan and Vershynin (2012, 2013); Chandrasekaran et al. (2012) for instance.

C.1 Applying Theorem 4.3 to (48)

By Theorem 4.3, (98) and (100), for any small $\varepsilon = \zeta$ such that $m \gtrsim k \log\left(\frac{en}{\varepsilon k}\right)$, then with probability at least $1 - \exp(-c'k\log(\frac{en}{k}))$, it holds for all $\boldsymbol{x} \in \mathcal{X}$ and any $t \geq 0$ that

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq \Big(\frac{Ck\log(en/k)}{m}\Big)^{t/2} \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + C_1\sqrt{\frac{k\log(n/\varepsilon k)}{m}} + C_2\lambda\Xi,$$

where

$$\Xi = \overline{\Xi}\log^{1/2}(\frac{1}{\overline{\Xi}}) + \varepsilon\log^{1/2}(\frac{1}{\varepsilon}) + \frac{\varepsilon}{\lambda}\big(\overline{\Xi}\log(\frac{1}{\overline{\Xi}}) + \varepsilon\log(\frac{1}{\varepsilon})\big)$$
$$\text{where } \overline{\Xi} := \frac{k\log(n/k\varepsilon)}{m} + \exp(-\frac{\lambda^2}{8})\left(\sqrt{\frac{\varepsilon k\log(en/k)}{m}} + \varepsilon\sqrt{\log(1/\varepsilon)}\right).$$

Setting $\varepsilon = (\frac{k\log(en/k)}{m})^{10}$ yields the claim.

C.2 Applying Theorem 4.4 to (49)

By Theorem 4.4 with sufficiently small $\varepsilon = \zeta$ and ϕ, along with (98), (99) and (101), we reach the following: if $m \gtrsim k\log(\frac{en}{k})$, then with probability at least $1 - \exp(-c'k\log\frac{en}{k})$, it holds for all $\boldsymbol{x} \in \mathcal{X}$ and any $t \geq 0$ that

$$\|\boldsymbol{x}_t - \boldsymbol{x}\|_2 \leq \Big(\frac{Ck\log(en/k)}{m}\Big)^{t/2} \|\boldsymbol{x}_0 - \boldsymbol{x}\|_2 + C_1\sqrt{\frac{k\log(\frac{n}{k\varepsilon})}{m}} + C_2\phi + C_3\lambda\Upsilon,$$

where

$$\Upsilon = \overline{\Upsilon}\log^{1/2}\big(\frac{1}{\overline{\Upsilon}}\big) + \zeta\log^{1/2}\big(\frac{1}{\zeta}\big) + \frac{\varepsilon}{\lambda}\bigg(\overline{\Upsilon}\log\big(\frac{1}{\overline{\Upsilon}}\big) + \zeta\log\big(\frac{1}{\zeta}\big)\bigg),$$
$$\text{where } \ \overline{\Upsilon} = \frac{k\log(\frac{n}{k\varepsilon})}{m} + \exp\big(-\frac{\lambda^2}{8}\big)\bigg[\sqrt{\frac{\varepsilon k\log(en/k)}{m}} + \varepsilon\log^{1/2}(\frac{1}{\varepsilon})\bigg].$$

As it turns out, the uniform recovery error rate here is identical to the one for the setting of (48). In particular, we set $\phi = \varepsilon = (\frac{k\log(en/k)}{m})^{10}$ yields the claim.

D Proof of Theorem 4.6 (Uniform 1-bit compressed sensing with no loss of log factor)

In this section, we establish a sharp uniform recovery error rate for sparse recovery from 1-bit measurements, which does not lose a log factor compared to the nonuniform error rates Plan and Vershynin (2016); Oymak and Soltanolkotabi (2017).

Under $\mu = \mathbb{E}_{g\sim N(0,1)}[g\,\mathrm{sign}(g)] = \sqrt{\frac{2}{\pi}}$ and $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^m(\sqrt{\frac{2}{\pi}}\boldsymbol{a}_i^T\boldsymbol{u} - y_i)\sqrt{\frac{2}{\pi}}\boldsymbol{a}_i$ from (15), the procedure of (50) is simply PGD with $\mathcal{K} = \Sigma_k^n$, gradient $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u})$, and step size $\eta = \mu^{-2} = \frac{\pi}{2}$. We now claim that, it suffices to prove the following uniform RAIC

$$\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{u}) \sim \mathrm{RAIC}\bigg(\Sigma_k^n; \Sigma_k^n, C_1\sqrt{\frac{k\log(en/k)}{m}}\|\boldsymbol{u}-\boldsymbol{x}\|_2 + C_2\sqrt{\frac{k\log(en/k)}{m}}, \frac{\pi}{2}\bigg), \quad \forall \boldsymbol{x}\in\Sigma_k^{n,*}$$

i.e.,

$$\begin{aligned}&\bigg\|\boldsymbol{u}-\boldsymbol{x} - \frac{1}{m}\sum_{i=1}^m\big(\boldsymbol{a}_i^T\boldsymbol{u} - \sqrt{\frac{\pi}{2}}\,\mathrm{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\big)\boldsymbol{a}_i\bigg\|_{(\Sigma_{2k}^{n,*})^\circ}\\ &\qquad\lesssim \sqrt{\frac{k\log(en/k)}{m}}\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \sqrt{\frac{k\log(en/k)}{m}}, \quad \forall(\boldsymbol{u},\boldsymbol{x})\in\Sigma_k^n\times\Sigma_k^{n,*},\end{aligned} \tag{102}$$

Again, we emphasize that Theorem 4.1 only yields a uniform RAIC with approximation error

$$C'\sqrt{\frac{k\log(en/k)}{m}}\|\boldsymbol{u}-\boldsymbol{x}\|_2 + \tilde{O}\bigg(\sqrt{\frac{k\log(en/k)}{m}}\bigg),$$

exhibiting extra log factors compared to the desired (102).

To start, we revisit the decomposition in (21) and the arguments in Appendix A.1 and find that it remains to establish

$$\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}}\sup_{\boldsymbol{q}\in\Sigma_{2k}^{n,*}}\frac{1}{m}\sum_{i=1}^m\bigg(\sqrt{\frac{\pi}{2}}\,\mathrm{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) - \boldsymbol{a}_i^T\boldsymbol{x}\bigg)\boldsymbol{a}_i^T\boldsymbol{q} \lesssim \sqrt{\frac{k\log(en/k)}{m}}$$

with the promised probability under $m \gtrsim k\log(en/k)$. To this end, the first step is to decouple the process into two centered processes as follows

$$\underbrace{\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}}\sup_{\boldsymbol{q}\in\Sigma_{2k}^{n,*}}\frac{1}{m}\sum_{i=1}^m\Big(\boldsymbol{x}^T\boldsymbol{q} - (\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{q})\Big)}_{:=I_1} + \underbrace{\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}}\sup_{\boldsymbol{q}\in\Sigma_{2k}^{n,*}}\frac{1}{m}\sum_{i=1}^m\bigg(\sqrt{\frac{\pi}{2}}\,\mathrm{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q} - \boldsymbol{x}^T\boldsymbol{q}\bigg)}_{:=I_2}.$$

We control each process separately as they require different techniques.

D.1 Controlling I_1

By Lemma A.1, we conclude that for any $m \geq w^2(\Sigma_k^{n,*})$, for $u = w(\Sigma_k^{n,*})$, we have that with probability at least $1 - 2^{-w^2(\Sigma_k^{n,*})}$

$$\sup_{\boldsymbol{x}\in\Sigma_k^{n,*}} \sup_{\boldsymbol{q}\in\Sigma_{2k}^{n,*}} \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{x}^T\boldsymbol{q} - (\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{q})\right) \lesssim \sqrt{\frac{w^2(\Sigma_k^{n,*})}{m}}.$$

Recalling the estimate $w^2(\Sigma_k^{n,*}) \asymp k\log(en/k)$, we conclude the estimate for I_1.

D.2 Controlling I_2

We would like to proceed similarly to control the process I_2. However, the sign function is not a class of function with subgaussian increments. To bypass the lack of regularity of the sign function, we rely on a covering argument exploiting the metric entropy estimates for Boolean classes of functions that exploit the VC dimension of the class. We refer the reader to Vershynin (2018) for a background on VC dimension, but it will not be strictly necessary here as the estimates based on VC-dimension used here are well-known facts in the literature.

To simplify the notation, we shall embed $\boldsymbol{x} \in \Sigma_k^{n,*} \subset \Sigma_{2k}^{n,*}$ and seek to bound

$$\sup_{\boldsymbol{x},\boldsymbol{q}\in\Sigma_{2k}^{n,*}} \left|\sqrt{\frac{\pi}{2}}\frac{1}{m}\sum_{i=1}^m \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q} - \boldsymbol{x}^T\boldsymbol{q}\right|.$$

To start, we reduce the control of I_2 to its symmetrized version. More accurately, we apply the Gine-Zinn symmetrization for tail estimates (see, e.g., (Mendelson, 2016, Theorem 1.14)) to obtain that for any $u \geq 1$

$$\begin{aligned}&\mathbb{P}\left(\sup_{\boldsymbol{x},\boldsymbol{q}\in\Sigma_{2k}^{n,*}} \left|\sqrt{\frac{\pi}{2}}\frac{1}{m}\sum_{i=1}^m \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q} - \boldsymbol{x}^T\boldsymbol{q}\right| \geq u\right)\\ &\leq 4\mathbb{P}\left(\sup_{\boldsymbol{x},\boldsymbol{q}\in\Sigma_{2k}^{n,*}} \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}\right| \geq \frac{u}{4}\sqrt{\frac{2}{\pi}}\right),\end{aligned} \tag{103}$$

where $\varepsilon_1, \cdots, \varepsilon_m$ are random signs independent from $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m$. Next, we focus on the process $Z_{\boldsymbol{x},\boldsymbol{q}}$ given by

$$Z_{\boldsymbol{x},\boldsymbol{q}} \coloneqq \frac{1}{\sqrt{m}}\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T\boldsymbol{q}, \quad (\boldsymbol{x},\boldsymbol{q}) \in \Sigma_{2k}^{n,*} \times \Sigma_{2k}^{n,*}.$$

We require two technical lemmas that brings together some technical results in the literature. To state them accurately, let $\mathcal{N}(T, d, \varepsilon)$ be covering number of T with respect to d at the scale of ε, that is, the smallest number of balls of radius ε with respect to the metric d whose centers lie in T.

Lemma D.1. *Let $\mathcal{X} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m\}$ be any collection of points in $\mathbb{R}^n$ and set $\|\cdot\|_{\mathcal{X}}$ to be the following metric in $\mathbb{R}^n$*

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathcal{X}} \coloneqq \left(\frac{1}{m}\sum_{i=1}^m \left|\operatorname{sign}(\boldsymbol{y}_i^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{y}_i^T\boldsymbol{x}')\right|^2\right)^{1/2}, \quad \boldsymbol{x},\boldsymbol{x}' \in \mathbb{R}^n.$$

Then, for any $\varepsilon \in (0,2)$,

$$\log\mathcal{N}(\Sigma_{2k}^{n,*}, \|\cdot\|_{\mathcal{X}}, \varepsilon) \lesssim k\log\left(\frac{en}{k}\right)\log\left(\frac{2}{\varepsilon}\right).$$

Proof. We first observe that

$$|\operatorname{sign}(\boldsymbol{y}^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{y}^T\boldsymbol{x}')| = 2|\mathbb{1}(\langle \boldsymbol{y}, \boldsymbol{x}\rangle \geq 0) - \mathbb{1}(\langle \boldsymbol{y}, \boldsymbol{x}'\rangle \geq 0)|$$

Consequently, up to a factor of 2, it suffices to estimate the metric entropy of the boolean class of function $\mathcal{F} = \{\mathbb{1}(\langle \cdot, \boldsymbol{x}\rangle \geq 0) : \boldsymbol{x} \in \Sigma_{2k}^{n,*}\}$ with respect to the L^2 metric endowed by the empirical measure of $\mathcal{X}$, that is,

$$\|f - f'\|_{L^2(\mathcal{X})} \coloneqq \left(\frac{1}{m}\sum_{i=1}^{m}\big(f - f'\big)(y_i)\big)^2\right)^{1/2}, \quad f, f' \in \mathcal{F}.$$

By the metric entropy estimate for VC class (see for (Vershynin, 2018, Theorem 8.3.13)), we obtain that

$$\log \mathcal{N}(\Sigma_{2k}^{n,*}, \|\cdot\|_{\mathcal{X}}, \varepsilon) = \log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^2(\mathcal{X})}, \varepsilon/2) \lesssim vc(\mathcal{F}) \log\left(\frac{2}{\varepsilon}\right),$$

where $vc(\mathcal{F})$ is the VC-dimension of the class $\mathcal{F}$. The proof now follows by combining the estimate above with $vc(\mathcal{F}) \lesssim k\log(en/k)$ (see (Depersin, 2024, Corollary 2)). □

Lemma D.2. *Let $\boldsymbol{a} \sim N(0, \boldsymbol{I}_n)$ be a standard gaussian random vector in $\mathbb{R}^n$. Given $\boldsymbol{x}, \boldsymbol{x}' \in S^{n-1}$ such that $\boldsymbol{x} \neq \pm\boldsymbol{x}'$, set $\|\cdot\|_{\psi_2}$ be the ψ_2 norm conditioned on $\operatorname{sign}\langle \boldsymbol{a}, \boldsymbol{x}\rangle$ and $\operatorname{sign}\langle \boldsymbol{a}, \boldsymbol{x}'\rangle$. Then for any vector $\boldsymbol{v} \in \mathbb{R}^n$, it the following holds that*

$$\|\boldsymbol{a}^T\boldsymbol{v}\|_{\psi_2} \lesssim \|\boldsymbol{v}\|_2.$$

Proof. By homogeneity of the ψ_2 norm, it suffices to prove the result when v is a unit norm vector. Next, we decompose the vector $\boldsymbol{v}$ along the directions $\boldsymbol{\beta}_1 \coloneqq (\boldsymbol{x} - \boldsymbol{x}')/\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ and $\boldsymbol{\beta}_2 \coloneqq (\boldsymbol{x} + \boldsymbol{x}')/\|\boldsymbol{x} + \boldsymbol{x}'\|_2$,

$$\boldsymbol{v} = (\boldsymbol{v}^T\boldsymbol{\beta}_1)\boldsymbol{\beta}_1 + (\boldsymbol{v}^T\boldsymbol{\beta}_2)\boldsymbol{\beta}_2 + \boldsymbol{v}^{\perp},$$

where $\boldsymbol{v}^{\perp}$ is orthogonal to both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. By triangle inequality and Cauchy-Schwarz,

$$\|\boldsymbol{a}^T\boldsymbol{v}\|_{\psi_2} \leq \|\boldsymbol{a}^T\boldsymbol{\beta}_1\|_{\psi_2} + \|\boldsymbol{a}^T\boldsymbol{\beta}_2\|_{\psi_2} + \|\boldsymbol{a}^T\boldsymbol{v}^{\perp}\|_{\psi_2}.$$

Since $\boldsymbol{a} \sim N(0, \boldsymbol{I}_n)$ and $\boldsymbol{v}^{\perp}$ is orthogonal to $\boldsymbol{x}$ and $\boldsymbol{x}'$, we conclude that $\boldsymbol{a}^T\boldsymbol{v}^{\perp}$ is distributed as a univariate centered Gaussian with variance $\|\boldsymbol{v}^{\perp}\|_2^2$, which implies that $\|\boldsymbol{a}^T\boldsymbol{v}^{\perp}\|_{\psi_2} \lesssim 1$. Next, by triangle inequality

$$\|\boldsymbol{a}^T\boldsymbol{\beta}_1\|_{\psi_2} \leq \|\boldsymbol{a}^T\boldsymbol{\beta}_1\mathbb{1}(\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}'))\|_{\psi_2} + \|\boldsymbol{a}^T\boldsymbol{\beta}_1\mathbb{1}(\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}'))\|_{\psi_2}.$$

We argue that each term is bounded by a constant. By (Matsumoto and Mazumdar, 2024a, Lemma B.1) we have that

$$\mathbb{P}\bigg(\bigg|\boldsymbol{a}^T\boldsymbol{\beta}_1 - \sqrt{\frac{\pi}{2}}\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{\arccos(\boldsymbol{x}^T\boldsymbol{x}')}\bigg| \geq t \big| \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}')\bigg) \lesssim e^{-t^2/2},$$

Consequently, by (Vershynin, 2018, Proposition 2.6.6)

$$\bigg\|\bigg(\boldsymbol{a}^T\boldsymbol{\beta}_1 - \sqrt{\frac{\pi}{2}}\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2}{\arccos(\boldsymbol{x}^T\boldsymbol{x}')}\bigg)\mathbb{1}(\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}'))\bigg\|_{\psi_2} \lesssim 1,$$

and by triangular inequality

$$\|\boldsymbol{a}^T\boldsymbol{\beta}_1\mathbb{1}(\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}'))\|_{\psi_2} \lesssim 1 + \sqrt{\frac{\pi}{2}}\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2}{\arccos(\boldsymbol{x}^T\boldsymbol{x}')} \lesssim 1.$$

Next, we argue that

$$\|\boldsymbol{a}^T\boldsymbol{\beta}_1\mathbb{1}(\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}'))\|_{\psi_2} \lesssim 1.$$

Indeed, by change of variables $\boldsymbol{y} = -\boldsymbol{x}'$, it is enough to argue that

$$\left\|\boldsymbol{a}^T\frac{\boldsymbol{x}+\boldsymbol{y}}{\|\boldsymbol{x}+\boldsymbol{y}\|_2}\mathbb{1}(\operatorname{sign}(\boldsymbol{a}^T\boldsymbol{x}) \neq \operatorname{sign}(\boldsymbol{a}^T\boldsymbol{y}))\right\|_{\psi_2} \lesssim 1,$$

which follows directly from (Matsumoto and Mazumdar, 2024a, Lemma B.2). By symmetry, the proof that $\|\boldsymbol{a}^T\boldsymbol{\beta}_2\|_{\psi_2} \lesssim 1$ follows an analogous argument. □

We now focus on obtaining the estimate for the supremum of the process $Z_{\boldsymbol{x},\boldsymbol{q}}$. Notice that the metric entropy estimates in Lemma D.1 holds for any collection of points $\mathcal{X}$, thus the law of $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m$ does not affect the estimates there. This is the typical advantage of an approach relying on covering number of boolean classes of function.

Thus, the first step is to exploit the behavior $\|Z_{\boldsymbol{x},\boldsymbol{q}} - Z_{\boldsymbol{x}',\boldsymbol{q}'}\|_{\psi_2}$, conditionally on the signs $\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})$ and $\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')$. In fact, by triangle inequality

$$\|Z_{\boldsymbol{x},\boldsymbol{q}} - Z_{\boldsymbol{x}',\boldsymbol{q}'}\|_{\psi_2} \leq \|Z_{\boldsymbol{x},\boldsymbol{q}} - Z_{\boldsymbol{x},\boldsymbol{q}'}\|_{\psi_2} + \|Z_{\boldsymbol{x},\boldsymbol{q}'} - Z_{\boldsymbol{x}',\boldsymbol{q}'}\|_{\psi_2}.$$

For the first term, notice that Lemma D.2 combined with (Vershynin, 2018, Proposition 2.7.1) implies that

$$\|Z_{\boldsymbol{x},\boldsymbol{q}} - Z_{\boldsymbol{x},\boldsymbol{q}'}\|_{\psi_2} = \frac{1}{\sqrt{m}}\left\|\sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i^T(\boldsymbol{q}-\boldsymbol{q}')\right\|_{\psi_2} \leq \left(\frac{1}{m}\sum_{i=1}^m \|\boldsymbol{a}_i^T(\boldsymbol{q}-\boldsymbol{q}')\|_{\psi_2}^2\right)^{1/2} \lesssim \|\boldsymbol{q}-\boldsymbol{q}'\|_2.$$

For the second term, by Lemma D.2 again

$$\begin{aligned}
\|Z_{\boldsymbol{x},\boldsymbol{q}'} - Z_{\boldsymbol{x}',\boldsymbol{q}'}\|_{\psi_2} &= \frac{1}{\sqrt{m}}\left\|\sum_{i=1}^m \varepsilon_i\big(\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')\big)\boldsymbol{a}_i^T\boldsymbol{q}'\right\|_{\psi_2} \\
&\leq \left(\frac{1}{m}\sum_{i=1}^m \|\big(\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')\big)\boldsymbol{a}_i^T\boldsymbol{q}'\|_{\psi_2}^2\right)^{1/2} \\
&\lesssim \left(\frac{1}{m}\sum_{i=1}^m \big|\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')\big|^2\right)^{1/2}.
\end{aligned}$$

We conclude that the process $Z_{\boldsymbol{x},\boldsymbol{q}}$ is dominated by the following metric

$$\|Z_{\boldsymbol{x},\boldsymbol{q}} - Z_{\boldsymbol{x}',\boldsymbol{q}'}\|_{\psi_2} \lesssim \underbrace{\|\boldsymbol{q}-\boldsymbol{q}'\|_2 + \left(\frac{1}{m}\sum_{i=1}^m |\operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}) - \operatorname{sign}(\boldsymbol{a}_i^T\boldsymbol{x}')|^2\right)^{1/2}}_{:=d\left((\boldsymbol{x},\boldsymbol{x}'),(\boldsymbol{q},\boldsymbol{q}')\right)}.$$

We now aim to apply Dudley's inequality (see for example Vershynin (2018)). To estimate the covering number of the set $\Sigma_{2k}^{n,*} \times \Sigma_{2k}^{n,*}$ with respect to d, we first observe that

$$\mathcal{N}(\Sigma_{2k}^{n,*} \times \Sigma_{2k}^{n,*}, d, \varepsilon) \leq \mathcal{N}(\Sigma_{2k}^{n,*}, \|\cdot\|_{\mathcal{X}}, \varepsilon/2) \cdot \mathcal{N}(\Sigma_{2k}^{n,*}, \|\cdot\|_2, \varepsilon/2),$$

where $\mathcal{X}$ is the collection of points $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$. By Lemma D.1 and the standard (Euclidean) covering number of $\Sigma_{2k}^{n,*}$, we have

$$\log \mathcal{N}(\Sigma_{2k}^{n,*} \times \Sigma_{2k}^{n,*}, d, \varepsilon) \lesssim k \log\left(\frac{en}{k}\right) \log\left(\frac{4}{\varepsilon}\right).$$

Since the diameter of $\Sigma_{2k}^{n,*} \times \Sigma_{2k}^{n,*}$ with respect to d is clearly bounded by 4, we apply the Dudley's tail bound: with probability at least $1 - 2e^{-u^2}$,

$$\begin{aligned}
\sup_{\boldsymbol{x},\boldsymbol{q} \in \Sigma_{2k}^{n,*}} Z_{\boldsymbol{x},\boldsymbol{q}} &\lesssim \int_0^4 \sqrt{\log \mathcal{N}(\Sigma_{2k}^{n,*} \times \Sigma_{2k}^{n,*}, d, \varepsilon)} d\varepsilon + u \\
&\lesssim \int_0^4 \sqrt{k \log\left(\frac{en}{k}\right) \log\left(\frac{4}{\varepsilon}\right)} d\varepsilon + u \\
&\lesssim \sqrt{k \log\left(\frac{en}{k}\right)} + u.
\end{aligned}$$

Diving both sides by $\sqrt{m}$ and choosing $u = \sqrt{k \log(en/k)}$, we obtain that with probability at least $1 - 2^{-k \log(en/k)}$,

$$\sup_{\boldsymbol{x},\boldsymbol{q} \in \Sigma_{2k}^{n,*}} \frac{1}{m} \sum_{i=1}^m \varepsilon_i \operatorname{sign}(\boldsymbol{a}_i^T \boldsymbol{x}) \boldsymbol{a}_i^T \boldsymbol{q} \lesssim \sqrt{\frac{k \log(en/k)}{m}},$$

which together with (103) concludes the estimate for I_2.