

Finding a sparse vector in a subspace: linear sparsity using alternating directions

Qing Qu, *Student Member, IEEE*, Ju Sun, *Student Member, IEEE*, and John Wright, *Member, IEEE*

Abstract—Is it possible to find the sparsest vector (direction) in a generic subspace $\mathcal{S} \subseteq \mathbb{R}^p$ with $\dim(\mathcal{S}) = n < p$? This problem can be considered a homogeneous variant of the sparse recovery problem, and finds connections to sparse dictionary learning, sparse PCA, and many other problems in signal processing and machine learning. In this paper, we focus on a *planted sparse model* for the subspace: the target sparse vector is embedded in an otherwise random subspace. Simple convex heuristics for this planted recovery problem provably break down when the fraction of nonzero entries in the target sparse vector substantially exceeds $O(1/\sqrt{n})$. In contrast, we exhibit a relatively simple nonconvex approach based on alternating directions, which provably succeeds even when the fraction of nonzero entries is $\Omega(1)$. To the best of our knowledge, this is the first practical algorithm to achieve linear scaling under the planted sparse model. Empirically, our proposed algorithm also succeeds in more challenging data models, e.g., sparse dictionary learning.

Index Terms—Sparse vector, Subspace modeling, Sparse recovery, Homogeneous recovery, Dictionary learning, Nonconvex optimization, Alternating direction method

I. INTRODUCTION

Suppose that a linear subspace \mathcal{S} embedded in \mathbb{R}^p contains a sparse vector $\mathbf{x}_0 \neq \mathbf{0}$. Given an arbitrary basis of \mathcal{S} , can we efficiently recover \mathbf{x}_0 (up to scaling)? Equivalently, provided a matrix $\mathbf{A} \in \mathbb{R}^{(p-n) \times p}$ with $\text{Null}(\mathbf{A}) = \mathcal{S}$,¹ can we efficiently find a nonzero sparse vector \mathbf{x} such that $\mathbf{A}\mathbf{x} = \mathbf{0}$? In the language of sparse recovery, can we solve

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{0}, \mathbf{x} \neq \mathbf{0} \quad ? \quad (\text{I.1})$$

In contrast to the standard sparse recovery problem ($\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{b} \neq \mathbf{0}$), for which convex relaxations perform nearly optimally for broad classes of designs \mathbf{A} [2, 3], the computational properties of problem (I.1) are not

This work was partially supported by grants ONR N00014-13-1-0492, NSF 1343282, NSF 1527809, and funding from the Moore and Sloan Foundations. Q. Qu, J. Sun and J. Wright are all with the Electrical Engineering Department, Columbia University, New York, NY, 10027, USA (e-mail: {qq2105, js4038, jw2966}@columbia.edu). This paper is an extension of our previous conference version [1].

¹ $\text{Null}(\mathbf{A}) \doteq \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{A}\mathbf{x} = \mathbf{0}\}$ denotes the null space of \mathbf{A} .

nearly as well understood. It has been known for several decades that the basic formulation

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S} \setminus \{\mathbf{0}\}, \quad (\text{I.2})$$

is NP-hard for an arbitrary subspace [4, 5]. In this paper, we assume a specific random *planted sparse model* for the subspace \mathcal{S} : a target sparse vector is embedded in an otherwise random subspace. We will show that under the specific random model, problem (I.2) is tractable by an efficient algorithm based on nonconvex optimization.

A. Motivation

The general version of Problem (I.2), in which \mathcal{S} can be an arbitrary subspace, takes several forms in numerical computation and computer science, and underlies several important problems in modern signal processing and machine learning. Below we provide a sample of these applications.

Sparse Null Space and Matrix Sparsification: The *sparse null space* problem is finding the sparsest matrix \mathbf{N} whose columns span the null space of a given matrix \mathbf{A} . The problem arises in the context of solving linear equality problems in constrained optimization [5], null space methods for quadratic programming [6], and solving underdetermined linear equations [7]. The *matrix sparsification* problem is of similar flavor, the task is finding the sparsest matrix \mathbf{B} which is equivalent to a given full rank matrix \mathbf{A} under elementary column operations. Sparsity helps simplify many fundamental matrix operations (see [8]), and the problem has applications in areas such as machine learning [9] and in discovering cycle bases of graphs [10]. [11] discusses connections between the two problems and also to other problems in complexity theory.

Sparse (Complete) Dictionary Learning: In dictionary learning, given a data matrix \mathbf{Y} , one seeks an approximation $\mathbf{Y} \approx \mathbf{A}\mathbf{X}$, such that \mathbf{A} is a representation dictionary with certain desired structure and \mathbf{X} collects the representation coefficients with maximal sparsity. Such compact representation naturally allows signal compression, and also facilitates efficient signal acquisition and classification (see relevant discussion in [12]). When

\mathbf{A} is required to be complete (i.e., square and invertible), by linear algebra, we have² $\text{row}(\mathbf{Y}) = \text{row}(\mathbf{X})$ [13]. Then the problem reduces to finding sparsest vectors (directions) in the known subspace $\text{row}(\mathbf{Y})$, i.e. (I.2). Insights into this problem have led to new theoretical developments on complete dictionary learning [13–15].

Sparse Principal Component Analysis (Sparse PCA):

In geometric terms, Sparse PCA (see, e.g., [16–18] for early developments and [19, 20] for discussion of recent results) concerns stable estimation of a linear subspace spanned by a sparse basis, in the data-poor regime, i.e., when the available data are not numerous enough to allow one to decouple the subspace estimation and sparsification tasks. Formally, given a data matrix $\mathbf{Z} = \mathbf{U}_0\mathbf{X}_0 + \mathbf{E}$,³ where $\mathbf{Z} \in \mathbb{R}^{p \times n}$ collects column-wise n data points, $\mathbf{U}_0 \in \mathbb{R}^{p \times r}$ is the sparse basis, and \mathbf{E} is a noise matrix, one is asked to estimate \mathbf{U}_0 (up to sign, scale, and permutation). Such a factorization finds applications in gene expression, financial data analysis and pattern recognition [24]. When the subspace is known (say by the PCA estimator with enough data samples), the problem again reduces to instances of (I.2) and is already nontrivial⁴. The full geometric sparse PCA can be treated as finding sparse vectors in a subspace that is subject to perturbation.

In addition, variants and generalizations of the problem (I.2) have also been studied in applications regarding control and optimization [25], nonrigid structure from motion [26], spectral estimation and Prony’s problem [27], outlier rejection in PCA [28], blind source separation [29], graphical model learning [30], and sparse coding on manifolds [31]; see also [32] and the references therein.

B. Prior Arts

Despite these potential applications of problem (I.2), it is only very recently that efficient computational surrogates with nontrivial recovery guarantees have been discovered for some cases of practical interest. In the context of sparse dictionary learning, Spielman et al. [13] introduced a convex relaxation which replaces the nonconvex problem (I.2) with a sequence of linear programs (ℓ^1/ℓ^∞ Relaxation):

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ s.t. } x(i) = 1, \mathbf{x} \in \mathcal{S}, 1 \leq i \leq p. \quad (\text{I.3})$$

They proved that when \mathcal{S} is generated as a span of n random sparse vectors, with high probability (w.h.p.), the

²Here, $\text{row}(\cdot)$ denotes the row space.

³Variants of multiple-component formulations often add an additional orthonormality constraint on \mathbf{U}_0 but involve a different notation of sparsity; see, e.g., [16, 21–23].

⁴[14] has also discussed this data-rich sparse PCA setting.

relaxation recovers these vectors, provided the probability of an entry being nonzero is at most $\theta \in O(1/\sqrt{n})$. In the *planted sparse model*, in which \mathcal{S} is formed as direct sum of a single sparse vector \mathbf{x}_0 and a “generic” subspace, Hand and Demanet proved that (I.3) also correctly recovers \mathbf{x}_0 , provided the fraction of nonzeros in \mathbf{x}_0 scales as $\theta \in O(1/\sqrt{n})$ [14]. One might imagine improving these results by tightening the analyses. Unfortunately, the results of [13, 14] are essentially sharp: *when θ substantially exceeds $\Omega(1/\sqrt{n})$, in both models the relaxation (I.3) provably breaks down.* Moreover, the most natural semidefinite programming (SDP) relaxation of (I.1),

$$\min_{\mathbf{X} \succeq 0} \|\mathbf{X}\|_1, \text{ s.t. } \langle \mathbf{A}^\top \mathbf{A}, \mathbf{X} \rangle = 0, \text{ trace}[\mathbf{X}] = 1 \quad (\text{I.4})$$

also breaks down at exactly the same threshold of $\theta \sim O(1/\sqrt{n})$.⁵

One might naturally conjecture that this $1/\sqrt{n}$ threshold is simply an intrinsic price we must pay for having an efficient algorithm, even in these random models. Some evidence towards this conjecture might be borrowed from the superficial similarity of (I.2)-(I.4) and *sparse PCA* [16]. In sparse PCA, there is a substantial gap between what can be achieved with currently available efficient algorithms and the information theoretic optimum [19, 33]. Is this also the case for recovering a sparse vector in a subspace? *Is $\theta \in O(1/\sqrt{n})$ simply the best we can do with efficient, guaranteed algorithms?*

Remarkably, this is not the case. Recently, Barak et al. introduced a new rounding technique for sum-of-squares relaxations, and showed that the sparse vector \mathbf{x}_0 in the planted sparse model can be recovered when $p \geq \Omega(n^2)$ and $\theta = \Omega(1)$ [34]. It is perhaps surprising that this is possible at all with a polynomial time algorithm. Unfortunately, the runtime of this approach is a high-degree polynomial in p (see Table I); for machine learning problems in which p is often either the feature dimension or the sample size, this algorithm is mostly of theoretical interest only. However, it raises an interesting algorithmic question: *Is there a practical algorithm that provably recovers a sparse vector with $\theta \gg 1/\sqrt{n}$ portion of nonzeros from a generic subspace \mathcal{S} ?*

C. Contributions and Recent Developments

In this paper, we address the above problem under the planted sparse model. We allow \mathbf{x}_0 to have up to $\theta_0 p$ nonzero entries, where $\theta_0 \in (0, 1)$ is a constant.

⁵This breakdown behavior is again in sharp contrast to the standard sparse approximation problem (with $\mathbf{b} \neq \mathbf{0}$), in which it is possible to handle very large fractions of nonzeros (say, $\theta = \Omega(1/\log n)$), or even $\theta = \Omega(1)$) using a very simple ℓ^1 relaxation [2, 3]

TABLE I
COMPARISON OF EXISTING METHODS FOR RECOVERING A PLANTED SPARSE VECTOR IN A SUBSPACE

Method	Recovery Condition	Time Complexity ⁶
ℓ^1/ℓ^∞ Relaxation [14]	$\theta \in O(1/\sqrt{n})$	$O(n^3 p \log(1/\epsilon))$
SDP Relaxation	$\theta \in O(1/\sqrt{n})$	$O(p^{3.5} \log(1/\epsilon))$
SOS Relaxation [34]	$p \geq \Omega(n^2), \theta \in O(1)$	$\sim O(p^7 \log(1/\epsilon))$ ⁷
Spectral Method [35]	$p \geq \Omega(n^2 \text{poly log}(n)), \theta \in O(1)$	$O(np \log(1/\epsilon))$
This work	$p \geq \Omega(n^4 \log n), \theta \in O(1)$	$O(n^5 p^2 \log n + n^3 p \log(1/\epsilon))$

We provide a relatively simple algorithm which, w.h.p., exactly recovers \mathbf{x}_0 , provided that $p \geq \Omega(n^4 \log n)$. A comparison of our results with existing methods is shown in Table I. After initial submission of our paper, Hopkins et al. [35] proposed a different simple algorithm based on the spectral method. This algorithm guarantees recovery of the planted sparse vector also up to linear sparsity, whenever $p \geq \Omega(n^2 \text{poly log}(n))$, and comes with better time complexity.⁸

Our algorithm is based on alternating directions, with two special twists. First, we introduce a special data driven initialization, which seems to be important for achieving $\theta = \Omega(1)$. Second, our theoretical results require a second, linear programming based rounding phase, which is similar to [13]. Our core algorithm has very simple iterations, of linear complexity in the size of the data, and hence should be scalable to moderate-to-large scale problems.

Besides enjoying the $\theta \sim \Omega(1)$ guarantee that is out of the reach of previous practical algorithms, our algorithm performs well in simulations – empirically succeeding with $p \geq \Omega(n \text{poly log}(n))$. It also performs well empirically on more challenging data models, such as the complete dictionary learning model, in which the subspace of interest contains not one, but n random target sparse vectors. This is encouraging, as breaking the $O(1/\sqrt{n})$ sparsity barrier with a practical algorithm and optimal guarantee is an important problem in theoretical dictionary learning [36–40]. In this regard, our recent work [15] presents an efficient algorithm based on Riemannian optimization that guarantees recovery up to linear sparsity. However, the result is based on different ideas: a different nonconvex formulation, optimization algorithm, and analysis methodology.

⁸Despite these improved guarantees in the planted sparse model, our method still produces more appealing results on real imagery data – see Section V-B for examples.

D. Paper Organization, Notations and Reproducible Research

The rest of the paper is organized as follows. In Section II, we provide a nonconvex formulation and show its capability of recovering the sparse vector. Section III introduces the alternating direction algorithm. In Section IV, we present our main results and sketch the proof ideas. Experimental evaluation of our method is provided in Section V. We conclude the paper by drawing connections to related work and discussing potential improvements in Section VI. Full proofs are all deferred to the appendix sections.

For a matrix \mathbf{X} , we use \mathbf{x}_i and \mathbf{x}^j to denote its i -th column and j -th row, respectively, all in column vector form. Moreover, we use $x^{(i)}$ to denote the i -th component of a vector \mathbf{x} . We use the compact notation $[k] \doteq \{1, \dots, k\}$ for any positive integer k , and use c or C , and their indexed versions to denote absolute numerical constants. The scope of these constants are always local, namely within a particular lemma, proposition, or proof, such that the apparently same constant in different contexts may carry different values. For probability events, sometimes we will just say the event holds “with high probability” (w.h.p.) if the probability of failure is dominated by $p^{-\kappa}$ for some $\kappa > 0$.

The codes to reproduce all the figures and experimental results can be found online at:

<https://github.com/sunju/psv>.

II. PROBLEM FORMULATION AND GLOBAL OPTIMALITY

We study the problem of recovering a sparse vector $\mathbf{x}_0 \neq \mathbf{0}$ (up to scale), which is an element of a known subspace $\mathcal{S} \subset \mathbb{R}^p$ of dimension n , provided an arbitrary orthonormal basis $\mathbf{Y} \in \mathbb{R}^{p \times n}$ for \mathcal{S} . Our starting point is the nonconvex formulation (I.2). Both the objective and the constraint set are nonconvex, and hence it is not easy to optimize over. We relax (I.2) by replacing the ℓ^0 norm with the ℓ^1 norm. For the constraint $\mathbf{x} \neq \mathbf{0}$, since in most applications we only care about the solution up to

scaling, it is natural to force \mathbf{x} to live on the unit sphere \mathbb{S}^{n-1} , giving

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S}, \|\mathbf{x}\|_2 = 1. \quad (\text{II.1})$$

This formulation is still nonconvex, and for general nonconvex problems it is known to be NP-hard to find even a local minimizer [41]. Nevertheless, the geometry of the sphere is benign enough, such that for well-structured inputs it actually *will* be possible to give algorithms that find the global optimizer.

The formulation (II.1) can be contrasted with (I.3), in which effectively we optimize the ℓ^1 norm subject to the constraint $\|\mathbf{x}\|_\infty = 1$: because the set $\{\mathbf{x} : \|\mathbf{x}\|_\infty = 1\}$ is polyhedral, the ℓ^∞ -constrained problem immediately yields a sequence of linear programs. This is very convenient for computation and analysis. However, it suffers from the aforementioned breakdown behavior around $\|\mathbf{x}_0\|_0 \sim p/\sqrt{n}$. In contrast, though the sphere $\|\mathbf{x}\|_2 = 1$ is a more complicated geometric constraint, it will allow much larger number of nonzeros in \mathbf{x}_0 . Indeed, if we consider the global optimizer of a reformulation of (II.1):

$$\min_{\mathbf{q} \in \mathbb{R}^n} \|\mathbf{Y}\mathbf{q}\|_1, \quad \text{s.t.} \quad \|\mathbf{q}\|_2 = 1, \quad (\text{II.2})$$

where \mathbf{Y} is any orthonormal basis for \mathcal{S} , the sufficient condition that guarantees exact recovery under the planted sparse model for the subspace is as follows:

Theorem II.1 (ℓ^1/ℓ^2 recovery, planted sparse model). *There exists a constant $\theta_0 > 0$, such that if the subspace \mathcal{S} follows the planted sparse model*

$$\mathcal{S} = \text{span}(\mathbf{x}_0, \mathbf{g}_1, \dots, \mathbf{g}_{n-1}) \subset \mathbb{R}^p,$$

where $\mathbf{g}_i \sim_{i.i.d.} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I})$, and $\mathbf{x}_0 \sim_{i.i.d.} \frac{1}{\sqrt{\theta p}} \text{Ber}(\theta)$ are all jointly independent and $1/\sqrt{n} < \theta < \theta_0$, then the unique (up to sign) optimizer \mathbf{q}^* to (II.2), for any orthonormal basis \mathbf{Y} of \mathcal{S} , produces $\mathbf{Y}\mathbf{q}^* = \xi\mathbf{x}_0$ for some $\xi \neq 0$ with probability at least $1 - cp^{-2}$, provided $p \geq Cn$. Here c and C are positive constants.

Hence, if we could find the global optimizer of (II.2), we would be able to recover \mathbf{x}_0 whose number of nonzero entries is quite large – even linear in the dimension p ($\theta = \Omega(1)$). On the other hand, it is not obvious that this should be possible: (II.2) is nonconvex. In the next section, we will describe a simple heuristic algorithm for approximately solving a relaxed version of the ℓ^1/ℓ^2 problem (II.2). More surprisingly, we will then prove that for a class of random problem instances, this algorithm, plus an auxiliary rounding technique, actually recovers the global optimizer – the target sparse vector \mathbf{x}_0 . The proof requires a detailed probabilistic analysis, which is

sketched in Section IV-B.

Before continuing, it is worth noting that the formulation (II.1) is in no way novel – see, e.g., the work of [29] in blind source separation for precedent. However, our algorithms and subsequent analysis are novel.

III. ALGORITHM BASED ON ALTERNATING DIRECTION METHOD (ADM)

To develop an algorithm for solving (II.2), it is useful to consider a slight relaxation of (II.2), in which we introduce an auxiliary variable $\mathbf{x} \approx \mathbf{Y}\mathbf{q}$:

$$\min_{\mathbf{q}, \mathbf{x}} f(\mathbf{q}, \mathbf{x}) \doteq \frac{1}{2} \|\mathbf{Y}\mathbf{q} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1, \quad \text{s.t.} \quad \|\mathbf{q}\|_2 = 1. \quad (\text{III.1})$$

Here, $\lambda > 0$ is a penalty parameter. It is not difficult to see that this problem is equivalent to minimizing the *Huber M-estimator* over $\mathbf{Y}\mathbf{q}$. This relaxation makes it possible to apply the alternating direction method to this problem. This method starts from some initial point $\mathbf{q}^{(0)}$, alternates between optimizing with respect to (w.r.t.) \mathbf{x} and optimizing w.r.t. \mathbf{q} :

$$\mathbf{x}^{(k+1)} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{Y}\mathbf{q}^{(k)} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1, \quad (\text{III.2})$$

$$\mathbf{q}^{(k+1)} = \arg \min_{\mathbf{q} \in \mathbb{S}^{n-1}} \frac{1}{2} \|\mathbf{Y}\mathbf{q} - \mathbf{x}^{(k+1)}\|_2^2, \quad (\text{III.3})$$

where $\mathbf{x}^{(k)}$ and $\mathbf{q}^{(k)}$ denote the values of \mathbf{x} and \mathbf{q} in the k -th iteration. Both (III.2) and (III.3) have simple closed form solutions:

$$\mathbf{x}^{(k+1)} = S_\lambda[\mathbf{Y}\mathbf{q}^{(k)}], \quad \mathbf{q}^{(k+1)} = \frac{\mathbf{Y}^\top \mathbf{x}^{(k+1)}}{\|\mathbf{Y}^\top \mathbf{x}^{(k+1)}\|_2}, \quad (\text{III.4})$$

where $S_\lambda[x] = \text{sign}(x) \max\{|x| - \lambda, 0\}$ is the soft-thresholding operator. The proposed ADM algorithm is summarized in Algorithm 1.

Algorithm 1 Nonconvex ADM Algorithm

Input: A matrix $\mathbf{Y} \in \mathbb{R}^{p \times n}$ with $\mathbf{Y}^\top \mathbf{Y} = \mathbf{I}$, initialization $\mathbf{q}^{(0)}$, threshold parameter $\lambda > 0$.

Output: The recovered sparse vector $\hat{\mathbf{x}}_0 = \mathbf{Y}\mathbf{q}^{(k)}$

1: **for** $k = 0, \dots, O(n^4 \log n)$ **do**

2: $\mathbf{x}^{(k+1)} = S_\lambda[\mathbf{Y}\mathbf{q}^{(k)}]$,

3: $\mathbf{q}^{(k+1)} = \frac{\mathbf{Y}^\top \mathbf{x}^{(k+1)}}{\|\mathbf{Y}^\top \mathbf{x}^{(k+1)}\|_2}$,

4: **end for**

The algorithm is simple to state and easy to implement. However, if our goal is to recover the *sparsest* vector \mathbf{x}_0 , some additional tricks are needed.

Initialization. Because the problem (II.2) is nonconvex, an arbitrary or random initialization may not produce a

global minimizer.⁹ In fact, good initializations are critical for the proposed ADM algorithm to succeed in the linear sparsity regime. For this purpose, we suggest using every normalized row of \mathbf{Y} as initializations for \mathbf{q} , and solving a sequence of p nonconvex programs (II.2) by the ADM algorithm.

To get an intuition of why our initialization works, recall the planted sparse model $\mathcal{S} = \text{span}(\mathbf{x}_0, \mathbf{g}_1, \dots, \mathbf{g}_{n-1})$ and suppose

$$\bar{\mathbf{Y}} = [\mathbf{x}_0 \mid \mathbf{g}_1 \mid \dots \mid \mathbf{g}_{n-1}] \in \mathbb{R}^{p \times n}. \quad (\text{III.5})$$

If we take a row $\bar{\mathbf{y}}^i$ of $\bar{\mathbf{Y}}$, in which $x_0(i)$ is nonzero, then $x_0(i) = \Theta(1/\sqrt{\theta p})$. Meanwhile, the entries of $\mathbf{g}_1(i), \dots, \mathbf{g}_{n-1}(i)$ are all $\mathcal{N}(0, 1/p)$, and so their magnitude have size about $1/\sqrt{p}$. Hence, when θ is not too large, $x_0(i)$ will be somewhat bigger than most of the other entries in $\bar{\mathbf{y}}^i$. Put another way, $\bar{\mathbf{y}}^i$ is biased towards the first standard basis vector \mathbf{e}_1 . Now, under our probabilistic model assumptions, $\bar{\mathbf{Y}}$ is very well conditioned: $\bar{\mathbf{Y}}^\top \bar{\mathbf{Y}} \approx \mathbf{I}$.¹⁰ Using the Gram-Schmidt process¹¹, we can find an orthonormal basis \mathbf{Y} for \mathcal{S} via:

$$\bar{\mathbf{Y}} = \mathbf{Y}\mathbf{R}, \quad (\text{III.6})$$

where \mathbf{R} is upper triangular, and \mathbf{R} is itself well-conditioned: $\mathbf{R} \approx \mathbf{I}$. Since the i -th row $\bar{\mathbf{y}}^i$ of $\bar{\mathbf{Y}}$ is biased in the direction of \mathbf{e}_1 and \mathbf{R} is well-conditioned, the i -th row \mathbf{y}^i of \mathbf{Y} is also biased in the direction of \mathbf{e}_1 . In other words, with this canonical orthobasis \mathbf{Y} for the subspace, the i -th row of \mathbf{Y} is biased in the direction of the global optimizer. The heuristic arguments are made rigorous in Appendix B and Appendix D.

What if we are handed some other basis $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$, where \mathbf{U} is an arbitrary orthogonal matrix? Suppose \mathbf{q}_* is a global optimizer to (II.2) with the input matrix \mathbf{Y} , then it is easy to check that, $\mathbf{U}^\top \mathbf{q}_*$ is a global optimizer to (II.2) with the input matrix $\hat{\mathbf{Y}}$. Because

$$\langle (\mathbf{Y}\mathbf{U})^\top \mathbf{e}_i, \mathbf{U}^\top \mathbf{q}_* \rangle = \langle \mathbf{Y}^\top \mathbf{e}_i, \mathbf{q}_* \rangle,$$

our initialization is *invariant* to any rotation of the orthobasis. Hence, *even if we are handed an arbitrary orthobasis for \mathcal{S} , the i -th row is still biased in the direction of the global optimizer.*

Rounding by linear programming (LP). Let $\bar{\mathbf{q}}$ denote the output of Algorithm 1. As illustrated in Fig. 1, we

will prove that with our particular initialization and an appropriate choice of λ , ADM algorithm uniformly moves towards the optimal over a large portion of the sphere, and its solution falls within a certain small radius of the globally optimal solution \mathbf{q}_* to (II.2). To exactly recover \mathbf{q}_* , or equivalently to recover the exact sparse vector $\mathbf{x}_0 = \gamma \mathbf{Y} \mathbf{q}_*$ for some $\gamma \neq 0$, we solve the linear program

$$\min_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1 \quad \text{s.t.} \quad \langle \mathbf{r}, \mathbf{q} \rangle = 1 \quad (\text{III.7})$$

with $\mathbf{r} = \bar{\mathbf{q}}$. Since the feasible set $\{\mathbf{q} \mid \langle \bar{\mathbf{q}}, \mathbf{q} \rangle = 1\}$ is essentially the tangent space of the sphere \mathbb{S}^{n-1} at $\bar{\mathbf{q}}$, whenever $\bar{\mathbf{q}}$ is close enough to \mathbf{q}_* , one should expect that the optimizer of (III.7) exactly recovers \mathbf{q}_* and hence \mathbf{x}_0 up to scale. We will prove that this is indeed true under appropriate conditions.

IV. MAIN RESULTS AND SKETCH OF ANALYSIS

A. Main Results

In this section, we describe our main theoretical result, which shows that w.h.p. the algorithm described in the previous section succeeds.

Theorem IV.1. *Suppose that \mathcal{S} obeys the planted sparse model, and let the columns of \mathbf{Y} form an arbitrary orthonormal basis for the subspace \mathcal{S} . Let $\mathbf{y}^1, \dots, \mathbf{y}^p \in \mathbb{R}^n$ denote the (transposes of) the rows of \mathbf{Y} . Apply Algorithm 1 with $\lambda = 1/\sqrt{p}$, using initializations $\mathbf{q}^{(0)} = \mathbf{y}^1 / \|\mathbf{y}^1\|_2, \dots, \mathbf{y}^p / \|\mathbf{y}^p\|_2$, to produce outputs $\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_p$. Solve the linear program (III.7) with $\mathbf{r} = \bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_p$, to produce $\hat{\mathbf{q}}_1, \dots, \hat{\mathbf{q}}_p$. Set $i^* \in \arg \min_i \|\mathbf{Y}\hat{\mathbf{q}}_i\|_1$. Then*

$$\mathbf{Y}\hat{\mathbf{q}}_{i^*} = \gamma \mathbf{x}_0, \quad (\text{IV.1})$$

for some $\gamma \neq 0$ with probability at least $1 - cp^{-2}$, provided

$$p \geq Cn^4 \log n, \quad \text{and} \quad \frac{1}{\sqrt{n}} \leq \theta \leq \theta_0. \quad (\text{IV.2})$$

Here C, c and θ_0 are positive constants.

Remark IV.2. *We can see that the result in Theorem IV.1 is suboptimal in sample complexity compared to the global optimality result in Theorem II.1 and Barak et al.'s result [34] (and the subsequent work [35]). For successful recovery, we require $p \geq \Omega(n^4 \log n)$, while the global optimality and Barak et al. demand $p \geq \Omega(n)$ and $p \geq \Omega(n^2)$, respectively. Aside from possible deficiencies in our current analysis, compared to Barak et al., we believe this is still the first practical and efficient method which is guaranteed to achieve $\theta \sim \Omega(1)$ rate. The lower bound on θ in Theorem IV.1 is mostly for convenience in*

⁹More precisely, in our models, random initialization *does* work, but only when the subspace dimension n is *extremely* low compared to the ambient dimension p .

¹⁰This is the common heuristic that ‘‘tall random matrices are well conditioned’’ [42].

¹¹...QR decomposition in general with restriction that $R_{11} = 1$.

the proof; in fact, the LP rounding stage of our algorithm already succeeds w.h.p. when $\theta \in O(1/\sqrt{n})$.

B. A Sketch of Analysis

In this section, we briefly sketch the main ideas of proving our main result in Theorem IV.1, to show that the “initialization + ADM + LP rounding” pipeline recovers \mathbf{x}_0 under the stated technical conditions, as illustrated in Fig. 1. The proof of our main result requires rather detailed technical analysis of the iteration-by-iteration properties of Algorithm 1, most of which is deferred to the appendices.

As noted in Section III, the ADM algorithm is invariant to change of basis. So w.l.o.g., let us assume $\bar{\mathbf{Y}} = [\mathbf{x}_0 \mid \mathbf{g}_1 \mid \cdots \mid \mathbf{g}_{n-1}]$ and let \mathbf{Y} to be its orthogonalization, i.e.,¹²

$$\mathbf{Y} = \left[\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \mid \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \left(\mathbf{G}^\top \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right)^{-1/2} \right]. \quad (\text{IV.3})$$

When p is large, $\bar{\mathbf{Y}}$ is nearly orthogonal, and hence $\bar{\mathbf{Y}}$ is very close to \mathbf{Y} . Thus, in our proofs, whenever convenient, we make the arguments on $\bar{\mathbf{Y}}$ first and then “propagate” the quantitative results onto \mathbf{Y} by perturbation arguments. With that noted, let $\mathbf{y}^1, \dots, \mathbf{y}^p$ be the transpose of the rows of \mathbf{Y} , and note that these are all independent random vectors. To prove the result of Theorem IV.1, we need the following results. First, given the specified \mathbf{Y} , we show that our initialization is biased towards the global optimum:

Proposition IV.3 (Good initialization). *Suppose $\theta > 1/\sqrt{n}$ and $p \geq Cn$. It holds with probability at least $1 - cp^{-2}$ that at least one of our p initialization vectors suggested in Section III, say $\mathbf{q}_i^{(0)} = \mathbf{y}^i / \|\mathbf{y}^i\|_2$, obeys*

$$\left| \left\langle \frac{\mathbf{y}^i}{\|\mathbf{y}^i\|_2}, \mathbf{e}_1 \right\rangle \right| \geq \frac{1}{10\sqrt{\theta n}}. \quad (\text{IV.4})$$

Here C, c are positive constants.

Proof. See Appendix D. \square

Second, we define a vector-valued random process $\mathbf{Q}(\mathbf{q})$ on $\mathbf{q} \in \mathbb{S}^{n-1}$, via

$$\mathbf{Q}(\mathbf{q}) = \frac{1}{p} \sum_{i=1}^p \mathbf{y}^i S_\lambda \left[\mathbf{q}^\top \mathbf{y}^i \right], \quad (\text{IV.5})$$

¹²Note that with probability one, the inverse matrix square-root in \mathbf{Y} is well defined. So \mathbf{Y} is well defined w.h.p. (i.e., except for $\mathbf{x}_0 = 0$). See more quantitative characterization of \mathbf{Y} in Appendix B.

so that based on (III.4), one step of the ADM algorithm takes the form:

$$\mathbf{q}^{(k+1)} = \frac{\mathbf{Q}(\mathbf{q}^{(k)})}{\|\mathbf{Q}(\mathbf{q}^{(k)})\|_2} \quad (\text{IV.6})$$

This is a very favorable form for analysis: the term in the numerator $\mathbf{Q}(\mathbf{q}^{(k)})$ is a sum of p independent random vectors with $\mathbf{q}^{(k)}$ viewed as fixed. We study the behavior of the iteration (IV.6) through the random process $\mathbf{Q}(\mathbf{q}^{(k)})$. We want to show that w.h.p. the ADM iterate sequence $\mathbf{q}^{(k)}$ converges to some small neighborhood of $\pm \mathbf{e}_1$, so that the ADM algorithm plus the LP rounding (described in Section III) successfully retrieves the sparse vector $\mathbf{x}_0 / \|\mathbf{x}_0\| = \mathbf{Y} \mathbf{e}_1$. Thus, we hope that in general, $\mathbf{Q}(\mathbf{q})$ is more concentrated on the first coordinate than $\mathbf{q} \in \mathbb{S}^{n-1}$. Let us partition the vector \mathbf{q} as $\mathbf{q} = [q_1; \mathbf{q}_2]$, with $q_1 \in \mathbb{R}$ and $\mathbf{q}_2 \in \mathbb{R}^{n-1}$; and correspondingly $\mathbf{Q}(\mathbf{q}) = [Q_1(\mathbf{q}); \mathbf{Q}_2(\mathbf{q})]$. The inner product of $\mathbf{Q}(\mathbf{q}) / \|\mathbf{Q}(\mathbf{q})\|_2$ and \mathbf{e}_1 is strictly larger than the inner product of \mathbf{q} and \mathbf{e}_1 if and only if

$$\frac{|Q_1(\mathbf{q})|}{|q_1|} > \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2}.$$

In the following proposition, we show that w.h.p., this inequality holds uniformly over a significant portion of the sphere

$$\Gamma \doteq \left\{ \mathbf{q} \in \mathbb{S}^{n-1} \mid \frac{1}{10\sqrt{n\theta}} \leq |q_1| \leq 3\sqrt{\theta}, \|\mathbf{q}_2\|_2 \geq \frac{1}{10} \right\}, \quad (\text{IV.7})$$

so the algorithm moves in the correct direction. Let us define the gap $G(\mathbf{q})$ between the two quantities $|Q_1(\mathbf{q})| / |q_1|$ and $\|\mathbf{Q}_2(\mathbf{q})\|_2 / \|\mathbf{q}_2\|_2$ as

$$G(\mathbf{q}) \doteq \frac{|Q_1(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2}, \quad (\text{IV.8})$$

and we show that the following result is true:

Proposition IV.4 (Uniform lower bound for finite sample gap). *There exists a constant $\theta_0 \in (0, 1)$, such that when $p \geq Cn^4 \log n$, the estimate*

$$\inf_{\mathbf{q} \in \Gamma} G(\mathbf{q}) \geq \frac{1}{10^4 \theta^2 n p}$$

holds with probability at least $1 - cp^{-2}$, provided $\theta \in (1/\sqrt{n}, \theta_0)$. Here C, c are positive constants.

Proof. See Appendix E. \square

Next, we show that whenever $|q_1| \geq 3\sqrt{\theta}$, w.h.p. the iterates stay in a “safe region” with $|q_1| \geq 2\sqrt{\theta}$ which is enough for LP rounding (III.7) to succeed.

Proposition IV.5 (Safe region for rounding). *There exists*

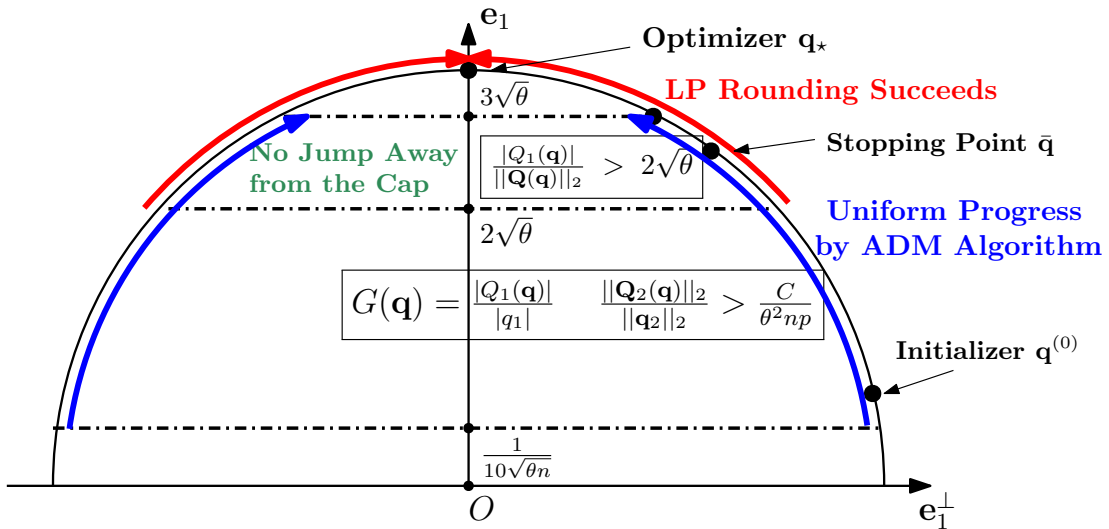


Fig. 1. An illustration of the proof sketch for our ADM algorithm.

a constant $\theta_0 \in (0, 1)$, such that when $p \geq Cn^4 \log n$, it holds with probability at least $1 - cp^{-2}$ that

$$\frac{|Q_1(\mathbf{q})|}{\|\mathbf{Q}(\mathbf{q})\|_2} \geq 2\sqrt{\theta}$$

for all $\mathbf{q} \in \mathbb{S}^{n-1}$ satisfying $|q_1| > 3\sqrt{\theta}$, provided $\theta \in (1/\sqrt{n}, \theta_0)$. Here C, c are positive constants.

Proof. See Appendix F. \square

In addition, the following result shows that the number of iterations for the ADM algorithm to reach the safe region can be bounded grossly by $O(n^4 \log n)$ w.h.p..

Proposition IV.6 (Iteration complexity of reaching the safe region). *There is a constant $\theta_0 \in (0, 1)$, such that when $p \geq Cn^4 \log n$, it holds with probability at least $1 - cp^{-2}$ that the ADM algorithm in Algorithm 1, with any initialization $\mathbf{q}^{(0)} \in \mathbb{S}^{n-1}$ satisfying $|q_1^{(0)}| \geq \frac{1}{10\sqrt{\theta n}}$, will produce some iterate $\bar{\mathbf{q}}$ with $|\bar{q}_1| > 3\sqrt{\theta}$ at least once in at most $O(n^4 \log n)$ iterations, provided $\theta \in (1/\sqrt{n}, \theta_0)$. Here C, c are positive constants.*

Proof. See Appendix G. \square

Moreover, we show that the LP rounding (III.7) with input $\mathbf{r} = \bar{\mathbf{q}}$ exactly recovers the optimal solution w.h.p., whenever the ADM algorithm returns a solution $\bar{\mathbf{q}}$ with first coordinate $|\bar{q}_1| > 2\sqrt{\theta}$.

Proposition IV.7 (Success of rounding). *There is a constant $\theta_0 \in (0, 1)$, such that when $p \geq Cn$, the following holds with probability at least $1 - cp^{-2}$ provided $\theta \in (1/\sqrt{n}, \theta_0)$: Suppose the input basis is \mathbf{Y} defined in (IV.3) and the ADM algorithm produces an output $\bar{\mathbf{q}} \in \mathbb{S}^{n-1}$ with $|\bar{q}_1| > 2\sqrt{\theta}$. Then the rounding procedure*

with $\mathbf{r} = \bar{\mathbf{q}}$ returns the desired solution $\pm \mathbf{e}_1$. Here C, c are positive constants.

Proof. See Appendix H. \square

Finally, given $p \geq Cn^4 \log n$ for a sufficiently large constant C , we combine all the results above to complete the proof of Theorem IV.1.

Proof of Theorem IV.1. W.l.o.g., let us again first consider $\bar{\mathbf{Y}}$ as defined in (III.5) and its orthogonalization \mathbf{Y} in a “natural/canonical” form (IV.3). We show that w.h.p. our algorithmic pipeline described in Section III exactly recovers the optimal solution up to scale, via the following argument:

- 1) **Good initializers.** Proposition IV.3 shows that w.h.p., at least one of the p initialization vectors, say $\mathbf{q}_i^{(0)} = \mathbf{y}^i / \|\mathbf{y}^i\|_2$, obeys

$$\left| \langle \mathbf{q}_i^{(0)}, \mathbf{e}_1 \rangle \right| \geq \frac{1}{10\sqrt{\theta n}},$$

which implies that $\mathbf{q}_i^{(0)}$ is biased towards the global optimal solution.

- 2) **Uniform progress away from the equator.** By Proposition IV.4, for any $\theta \in (1/\sqrt{n}, \theta_0)$ with a constant $\theta_0 \in (0, 1)$,

$$G(\mathbf{q}) = \frac{|Q_1(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}\|_2} \geq \frac{1}{10^4 \theta^2 n p} \quad (\text{IV.9})$$

holds uniformly for all $\mathbf{q} \in \mathbb{S}^{n-1}$ in the region $\frac{1}{10\sqrt{\theta n}} \leq |q_1| \leq 3\sqrt{\theta}$ w.h.p.. This implies that with an input $\mathbf{q}^{(0)}$ such that $|q_1^{(0)}| \geq \frac{1}{10\sqrt{\theta n}}$, the ADM algorithm will eventually obtain a point $\mathbf{q}^{(k)}$ for

which $|q^{(k)}| \geq 3\sqrt{\theta}$, if sufficiently many iterations are allowed.

- 3) **No jumps away from the caps.** Proposition IV.5 shows that for any $\theta \in (1/\sqrt{n}, \theta_0)$ with a constant $\theta_0 \in (0, 1)$, w.h.p.,

$$\frac{Q_1(\mathbf{q})}{\|\mathbf{Q}(\mathbf{q})\|_2} \geq 2\sqrt{\theta}$$

holds for all $\mathbf{q} \in \mathbb{S}^{n-1}$ with $|q_1| \geq 3\sqrt{\theta}$. This implies that once $|q_1^{(k)}| \geq 3\sqrt{\theta}$ for some iterate k , all the future iterates produced by the ADM algorithm stay in a ‘‘spherical cap’’ region around the optimum with $|q_1| \geq 2\sqrt{\theta}$.

- 4) **Location of stopping points.** As shown in Proposition IV.6, w.h.p., the strictly positive gap $G(\mathbf{q})$ in (IV.9) ensures that one needs to run at most $O(n^4 \log n)$ iterations to first encounter an iterate $\mathbf{q}^{(k)}$ such that $|q_1^{(k)}| \geq 3\sqrt{\theta}$. Hence, the steps above imply that, w.h.p., Algorithm 1 fed with the proposed initialization scheme successively produces iterates $\bar{\mathbf{q}} \in \mathbb{S}^{n-1}$ with its first coordinate $|\bar{q}_1| \geq 2\sqrt{\theta}$ after $O(n^4 \log n)$ steps.
- 5) **Rounding succeeds when $|r_1| \geq 2\sqrt{\theta}$.** Proposition IV.7 proves that w.h.p., the LP rounding (III.7) with an input $\mathbf{r} = \bar{\mathbf{q}}$ produces the solution $\pm \mathbf{x}_0$ up to scale.

Taken together, these claims imply that from at least one of the initializers $\mathbf{q}^{(0)}$, the ADM algorithm will produce an output $\bar{\mathbf{q}}$ which is accurate enough for LP rounding to exactly return $\mathbf{x}_0/\|\mathbf{x}_0\|_2$. On the other hand, our ℓ^1/ℓ^2 optimality theorem (Theorem II.1) implies that $\pm \mathbf{x}_0$ are the unique vectors with the smallest ℓ^1 norm among all unit vectors in the subspace. Since w.h.p. $\mathbf{x}_0/\|\mathbf{x}_0\|_2$ is among the p unit vectors $\hat{\mathbf{q}}_1, \dots, \hat{\mathbf{q}}_p$ our p row initializers finally produce, our minimal ℓ^1 norm selector will successfully locate $\mathbf{x}_0/\|\mathbf{x}_0\|_2$ vector.

For the general case when the input is an arbitrary orthonormal basis $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$ for some orthogonal matrix \mathbf{U} , the target solution is $\mathbf{U}^\top \mathbf{e}_1$. The following technical pieces are perfectly parallel to the argument above for \mathbf{Y} .

- 1) Discussion at the end of Appendix D implies that w.h.p., at least one row of $\hat{\mathbf{Y}}$ provides an initial point $\mathbf{q}^{(0)}$ such that $|\langle \mathbf{q}^{(0)}, \mathbf{U}^\top \mathbf{e}_1 \rangle| \geq \frac{1}{10\sqrt{\theta n}}$.
- 2) Discussion following Proposition IV.4 in Appendix E indicates that for all \mathbf{q} such that $\frac{1}{10\sqrt{\theta n}} \leq |\langle \mathbf{q}, \mathbf{U}^\top \mathbf{e}_1 \rangle| \leq 3\sqrt{\theta}$, there is a strictly positive gap, indicating steady progress towards a point $\mathbf{q}^{(k)}$ such that $|\langle \mathbf{q}^{(k)}, \mathbf{U}^\top \mathbf{e}_1 \rangle| \geq 3\sqrt{\theta}$.
- 3) Discussion at the end of Appendix F implies that once \mathbf{q} satisfies $|\langle \mathbf{q}, \mathbf{U}^\top \mathbf{e}_1 \rangle|$, the next iterate will

not move far away from the target:

$$\left| \left\langle \mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}) / \left\| \mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}) \right\|_2, \mathbf{U}^\top \mathbf{e}_1 \right\rangle \right| \geq 2\sqrt{\theta}.$$

- 4) Repeating the argument in Appendix G for general input $\hat{\mathbf{Y}}$ shows it is enough to run the ADM algorithm $O(n^4 \log n)$ iterations to cross the range $\frac{1}{10\sqrt{\theta n}} \leq |\langle \mathbf{q}, \mathbf{U}^\top \mathbf{e}_1 \rangle| \leq 3\sqrt{\theta}$. So the argument above together dictates that with the proposed initialization, w.h.p., the ADM algorithm produces an output $\bar{\mathbf{q}}$ that satisfies $|\langle \bar{\mathbf{q}}, \mathbf{U}^\top \mathbf{e}_1 \rangle| \geq 2\sqrt{\theta}$, if we run at least $O(n^4 \log n)$ iterations.
- 5) Since the ADM returns $\bar{\mathbf{q}}$ satisfying $|\langle \bar{\mathbf{q}}, \mathbf{R}^\top \mathbf{e}_1 \rangle| \geq 2\sqrt{\theta}$, discussion at the end of Appendix H implies that we will obtain a solution $\mathbf{q}_\star = \pm \mathbf{U}^\top \mathbf{e}_1$ up to scale as the optimizer of the rounding program, exactly the target solution.

Hence, we complete the proof. \square

Remark IV.8. *Under the planted sparse model, in practice the ADM algorithm with the proposed initialization converges to a global optimizer of (III.1) that correctly recovers \mathbf{x}_0 . In fact, simple calculation shows such desired point for successful recovery is indeed the only critical point of (III.1) near the pole in Fig. 1. Unfortunately, using the current analytical framework, we did not succeed in proving such convergence in theory. Proposition IV.5 and IV.6 imply that after $O(n^4 \log n)$ iterations, however, the ADM sequence will stay in a small neighborhood of the target. Hence, we proposed to stop after $O(n^4 \log n)$ steps, and then round the output using the LP that provable recover the target, as implied by Proposition IV.5 and IV.7. So the LP rounding procedure is for the purpose of completing the theory, and seems not necessary in practice. We suspect alternative analytical strategies, such as the geometrical analysis that we will discuss in Section VI, can likely get around the artifact.*

V. EXPERIMENTAL RESULTS

In this section, we show the performance of the proposed ADM algorithm on both synthetic and real datasets. On the synthetic dataset, we show the phase transition of our algorithm on both the planted sparse and the dictionary learning models; for the real dataset, we demonstrate how seeking sparse vectors can help discover interesting patterns on face images.

A. Phase Transition on Synthetic Data

For the planted sparse model, for each pair of (k, p) , we generate the n dimensional subspace $\mathcal{S} \subset \mathbb{R}^p$ by direct sum of \mathbf{x}_0 and \mathbf{G} : $\mathbf{x}_0 \in \mathbb{R}^p$ is a k -sparse vector with uniformly random support and all nonzero entries

equal to 1, and $\mathbf{G} \in \mathbb{R}^{p \times (n-1)}$ is an i.i.d. Gaussian matrix distributed by $\mathcal{N}(0, 1/p)$. So one basis \mathbf{Y} of the subspace \mathcal{S} can be constructed by $\mathbf{Y} = \text{GS}([\mathbf{x}_0, \mathbf{G}])\mathbf{U}$, where $\text{GS}(\cdot)$ denotes the Gram-Schmidt orthonormalization operator and $\mathbf{U} \in \mathbb{R}^{n \times n}$ is an arbitrary orthogonal matrix. For each p , we set the regularization parameter in (III.1) as $\lambda = 1/\sqrt{p}$, use all the normalized rows of \mathbf{Y} as initializations of \mathbf{q} for the proposed ADM algorithm, and run the alternating steps for 10^4 iterations. We determine the recovery to be successful whenever $\|\mathbf{x}_0/\|\mathbf{x}_0\|_2 - \mathbf{Y}\mathbf{q}\|_2 \leq 10^{-2}$ for at least one of the p trials (we set the tolerance relatively large as we have shown that LP rounding exactly recovers the solutions with approximate input). To determine the empirical recovery performance of our ADM algorithm, first we fix the relationship between n and p as $p = 5n \log n$, and plot out the phase transition between k and p . Next, we fix the sparsity level $\theta = 0.2$ (or $k = 0.2p$), and plot out the phase transition between p and n . For each pair of (p, k) or (n, p) , we repeat the simulation for 10 times. Fig. 2 shows both phase transition plots.

We also experiment with the complete dictionary learning model as in [13] (see also [15]). Specifically, the observation is assumed to be $\mathbf{Y} = \mathbf{A}_0\mathbf{X}_0$, where \mathbf{A}_0 is a square, invertible matrix, and \mathbf{X}_0 a $n \times p$ sparse matrix. Since \mathbf{A}_0 is invertible, the row space of \mathbf{Y} is the same as that of \mathbf{X}_0 . For each pair of (k, n) , we generate $\mathbf{X}_0 = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top$, where each vector $\mathbf{x}_i \in \mathbb{R}^p$ is k -sparse with every nonzero entry following i.i.d. Gaussian distribution, and construct the observation by $\mathbf{Y}^\top = \text{GS}(\mathbf{X}_0^\top)\mathbf{U}^\top$. We repeat the same experiment as for the planted sparse model described above. The only difference is that here we determine the recovery to be successful as long as one sparse row of \mathbf{X}_0 is recovered by one of those p programs. Fig. 3 shows both phase transition plots.

Fig. 2(a) and Fig. 3(a) suggest our ADM algorithm could work into the linear sparsity regime for both models, provided $p \geq \Omega(n \log n)$. Moreover, for both models, the $\log n$ factor seems necessary for working into the linear sparsity regime, as suggested by Fig. 2(b) and Fig. 3(b): there are clear nonlinear transition boundaries between success and failure regions. For both models, $O(n \log n)$ sample requirement is near optimal: for the planted sparse model, obviously $p \geq \Omega(n)$ is necessary; for the complete dictionary learning model, [13] proved that $p \geq \Omega(n \log n)$ is required for exact recovery. For the planted sparse model, our result $p \geq \Omega(n^4 \log n)$ is far from this much lower empirical requirement. Fig 2(b) further suggests that alternative reformulation and algorithm are needed to solve (II.1) so that the optimal recovery guarantee as depicted in Theorem II.1 can be

obtained.

B. Exploratory Experiments on Faces

It is well known in computer vision that the collection of images of a convex object only subject to illumination changes can be well approximated by a low-dimensional subspaces in raw-pixel space [43]. We will play with face subspaces here. First, we extract face images of one person (65 images) under different illumination conditions. Then we apply *robust principal component analysis* [44] to the data and get a low dimensional subspace of dimension 10, i.e., the basis $\mathbf{Y} \in \mathbb{R}^{32256 \times 10}$. We apply the ADM + LP algorithm to find the sparsest elements in such a subspace, by randomly selecting 10% rows of \mathbf{Y} as initializations for \mathbf{q} . We judge the sparsity in the ℓ^1/ℓ^2 sense, that is, the sparsest vector $\hat{\mathbf{x}}_0 = \mathbf{Y}\mathbf{q}^*$ should produce the smallest $\|\mathbf{Y}\mathbf{q}\|_1 / \|\mathbf{Y}\mathbf{q}\|_2$ among all results. Once some sparse vectors are found, we project the subspace onto orthogonal complement of the sparse vectors already found¹³, and continue the seeking process in the projected subspace. Fig. 4(Top) shows the first four sparse vectors we get from the data. We can see they correspond well to different extreme illumination conditions. We also implemented the spectral method (with the LP post-processing) proposed in [35] for comparison under the same protocol. The result is presented as Fig. 4(Bottom): the ratios $\|\cdot\|_{\ell^1} / \|\cdot\|_{\ell^2}$ are significantly higher, and the ratios $\|\cdot\|_{\ell^4} / \|\cdot\|_{\ell^2}$ (this is the metric to be maximized in [35] to promote sparsity) are significantly lower. By these two criteria the spectral method with LP rounding consistently produces vectors with higher sparsity levels under our evaluation protocol. Moreover, the resulting images are harder to interpret physically.

Second, we manually select ten different persons' faces under the normal lighting condition. Again, the dimension of the subspace is 10 and $\mathbf{Y} \in \mathbb{R}^{32256 \times 10}$. We repeat the same experiment as stated above. Fig. 5 shows four sparse vectors we get from the data. Interestingly, the sparse vectors roughly correspond to differences of face images concentrated around facial parts that different people tend to differ from each other, e.g., eye brows, forehead hair, nose, etc. By comparison, the vectors returned by the spectral method [35] are relatively denser and the sparsity patterns in the images are less structured physically.

In sum, our algorithm seems to find useful sparse vectors for potential applications, such as peculiarity discovery in first setting, and locating differences in second setting. Nevertheless, the main goal of this

¹³The idea is to build a sparse, orthonormal basis for the subspace in a greedy manner.

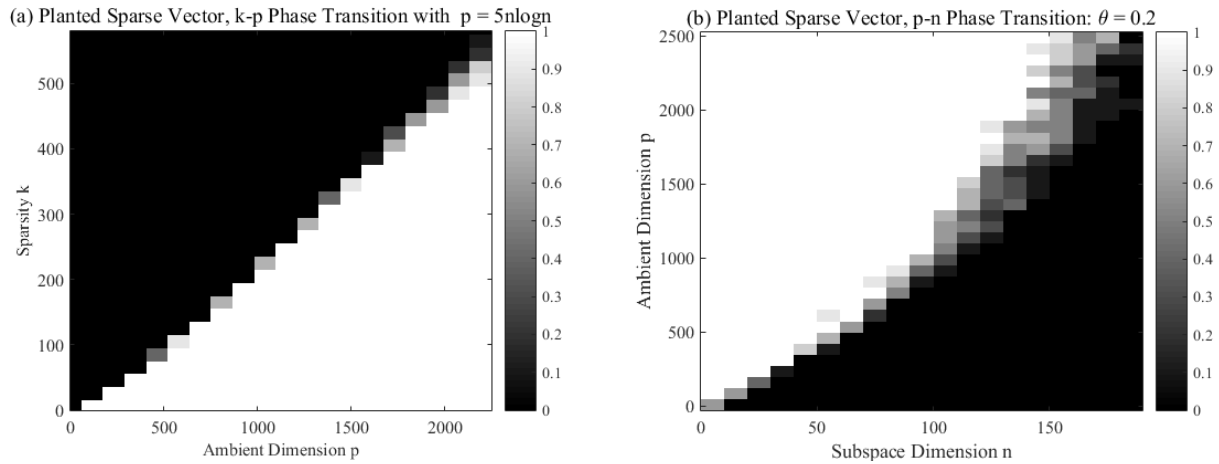


Fig. 2. Phase transition for the planted sparse model using the ADM algorithm: (a) with fixed relationship between p and n : $p = 5n \log n$; (b) with fixed relationship between p and k : $k = 0.2p$. White indicates success and black indicates failure.

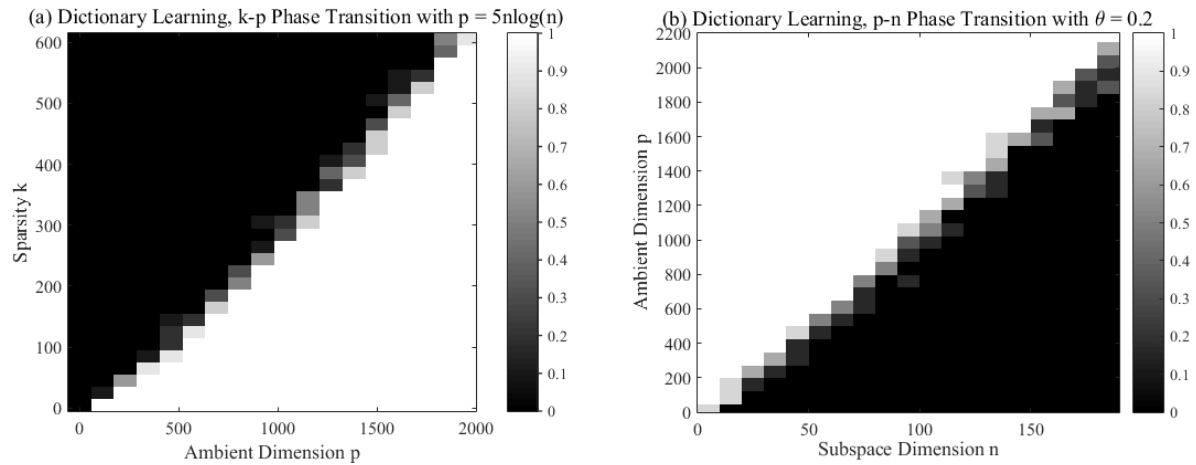


Fig. 3. Phase transition for the dictionary learning model using the ADM algorithm: (a) with fixed relationship between p and n : $p = 5n \log n$; (b) with fixed relationship between p and k : $k = 0.2p$. White indicates success and black indicates failure.

experiment is to invite readers to think about similar pattern discovery problems that might be cast as the problem of seeking sparse vectors in a subspace. The experiment also demonstrates in a concrete way the practicality of our algorithm, both in handling data sets of realistic size and in producing meaningful results even beyond the (idealized) planted sparse model that we adopted for analysis.

VI. CONNECTIONS AND DISCUSSION

For the planted sparse model, there is a substantial performance gap in terms of p - n relationship between the our optimality theorem (Theorem II.1), empirical simulations, and guarantees we have obtained via efficient algorithm (Theorem IV.1). More careful and tighter analysis based on decoupling [45] and chaining [46, 47] and geometrical analysis described below can probably help bridge the gap between our theoretical and empirical results. Matching

the theoretical limit depicted in Theorem II.1 seems to require novel algorithmic ideas. The random models we assume for the subspace can be extended to other random models, particularly for dictionary learning where all the bases are sparse (e.g., Bernoulli-Gaussian random model).

This work is part of a recent surge of research efforts on deriving provable and practical nonconvex algorithms to central problems in modern signal processing and machine learning. These problems include low-rank matrix recovery/completion [48–56], tensor recovery/decomposition [57–61], phase retrieval [62–65], dictionary learning [15, 36–40], and so on.¹⁴ Our approach, like the others, is to start with a carefully chosen, problem-specific initialization, and then perform a local analysis of the subsequent iterates to guarantee

¹⁴The webpage <http://sunju.org/research/nonconvex/> maintained by the second author contains pointers to the growing list of work in this direction.

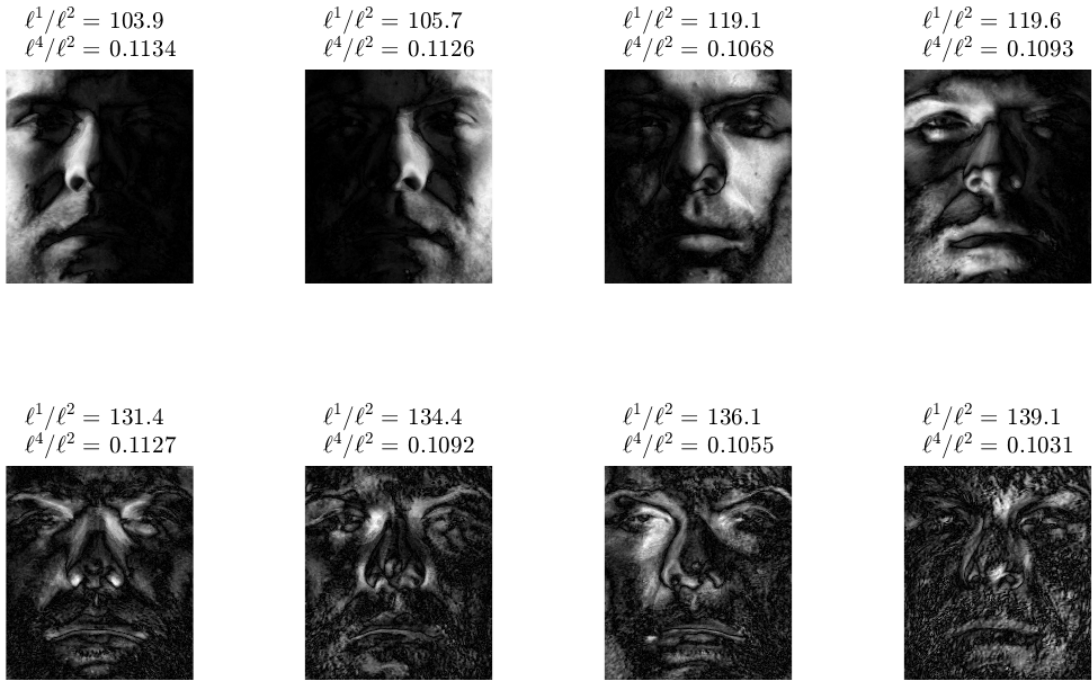


Fig. 4. The first four sparse vectors extracted for one person in the Yale B database under different illuminations. (Top) by our ADM algorithm; (Bottom) by the speeding-up SOS algorithm proposed in [35].

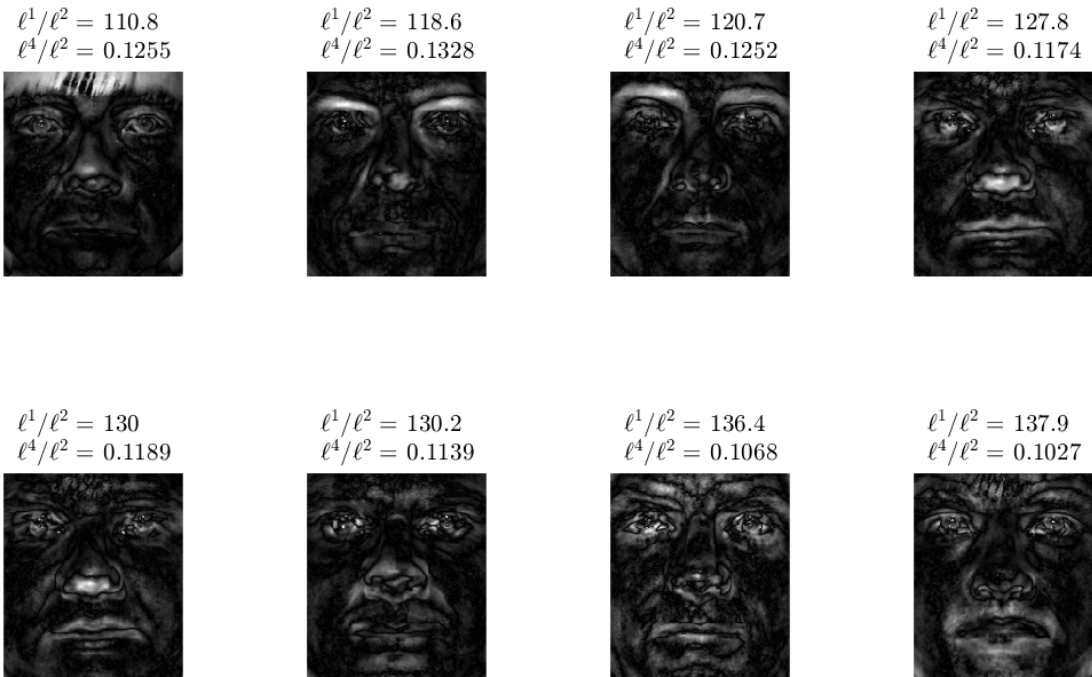


Fig. 5. The first four sparse vectors extracted for 10 persons in the Yale B database under normal illuminations. (Top) by our ADM algorithm; (Bottom) by the speeding-up SOS algorithm proposed in [35].

convergence to a good solution. In comparison, our subsequent work on complete dictionary learning [15] and generalized phase retrieval [65] has taken a geometrical approach by characterizing the function landscape and designing efficient algorithm accordingly. The geometric approach has allowed provable recovery via efficient algorithms, with an *arbitrary initialization*. The article [66] summarizes the geometric approach and its applicability to several other problems of interest.

A hybrid of the initialization and the geometric approach discussed above is likely to be a powerful computational framework. To see it in action for the current planted sparse vector problem, in Fig. 6 we provide the asymptotic function landscape (i.e., $p \rightarrow \infty$) of the Huber loss on the sphere \mathbb{S}^2 (aka the relaxed formulation we tried to solve (III.1)). It is clear that with an initialization that is biased towards either the north or the south pole, we are situated in a region where the gradients are always nonzero and points to the favorable directions such that many reasonable optimization algorithms can take the gradient information and make steady progress towards the target. This will probably ease the algorithm development and analysis, and help yield tight performance guarantees.

We provide a very efficient algorithm for finding a sparse vector in a subspace, with strong guarantee. Our algorithm is practical for handling large datasets—in the experiment on the face dataset, we successfully extracted some meaningful features from the human face images. However, the potential of seeking sparse/structured element in a subspace seems largely unexplored, despite the cases we mentioned at the start. We hope this work could inspire more application ideas.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their valuable comments and constructive criticism. JS thanks the Wei Family Private Foundation for their generous support. We thank Cun Mu, IEOR Department of Columbia University, for helpful discussion and input regarding this work. We thank the anonymous reviewers for their constructive comments that helped improve the manuscript. This work was partially supported by grants ONR N00014-13-1-0492, NSF 1343282, NSF 1527809, and funding from the Moore and Sloan Foundations.

APPENDIX A

TECHNICAL TOOLS AND PRELIMINARIES

In this appendix, we record several lemmas that are useful for our analysis.

Lemma A.1. *Let $\psi(x)$ and $\Psi(x)$ to denote the probability density function (pdf) and the cumulative distribution function (cdf) for the standard normal distribution:*

$$\begin{aligned} \text{(Standard Normal pdf)} \quad \psi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \\ \text{(Standard Normal cdf)} \quad \Psi(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{t^2}{2}\right\} dt, \end{aligned}$$

Suppose a random variable $X \sim \mathcal{N}(0, \sigma^2)$, with the pdf $f_\sigma(x) = \frac{1}{\sigma} \psi\left(\frac{x}{\sigma}\right)$, then for any $t_2 > t_1$ we have

$$\begin{aligned} \int_{t_1}^{t_2} f_\sigma(x) dx &= \Psi\left(\frac{t_2}{\sigma}\right) - \Psi\left(\frac{t_1}{\sigma}\right), \\ \int_{t_1}^{t_2} x f_\sigma(x) dx &= -\sigma \left[\psi\left(\frac{t_2}{\sigma}\right) - \psi\left(\frac{t_1}{\sigma}\right) \right], \\ \int_{t_1}^{t_2} x^2 f_\sigma(x) dx &= \sigma^2 \left[\Psi\left(\frac{t_2}{\sigma}\right) - \Psi\left(\frac{t_1}{\sigma}\right) \right] \\ &\quad - \sigma \left[t_2 \psi\left(\frac{t_2}{\sigma}\right) - t_1 \psi\left(\frac{t_1}{\sigma}\right) \right]. \end{aligned}$$

Lemma A.2 (Taylor Expansion of Standard Gaussian cdf and pdf). *Assume $\psi(x)$ and $\Psi(x)$ be defined as above. There exists some universal constant $C_\psi > 0$ such that for any $x_0, x \in \mathbb{R}$,*

$$\begin{aligned} |\psi(x) - [\psi(x_0) - x_0 \psi(x_0)(x - x_0)]| &\leq C_\psi (x - x_0)^2, \\ |\Psi(x) - [\Psi(x_0) + \psi(x_0)(x - x_0)]| &\leq C_\psi (x - x_0)^2. \end{aligned}$$

Lemma A.3 (Matrix Induced Norms). *For any matrix $\mathbf{A} \in \mathbb{R}^{p \times n}$, the induced matrix norm from $\ell^p \rightarrow \ell^q$ is defined as*

$$\|\mathbf{A}\|_{\ell^p \rightarrow \ell^q} \doteq \sup_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_q.$$

In particular, let $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_n] = [\mathbf{a}^1, \dots, \mathbf{a}^p]^\top$, we have

$$\begin{aligned} \|\mathbf{A}\|_{\ell^2 \rightarrow \ell^1} &= \sup_{\|\mathbf{x}\|_2=1} \sum_{k=1}^p |\mathbf{a}_k^\top \mathbf{x}|, \quad \|\mathbf{A}\|_{\ell^2 \rightarrow \ell^\infty} = \max_{1 \leq k \leq p} \|\mathbf{a}^k\|_2, \\ \|\mathbf{A}\mathbf{B}\|_{\ell^p \rightarrow \ell^r} &\leq \|\mathbf{A}\|_{\ell^q \rightarrow \ell^r} \|\mathbf{B}\|_{\ell^p \rightarrow \ell^q}, \end{aligned}$$

and \mathbf{B} is any matrix of size compatible with \mathbf{A} .

Lemma A.4 (Moments of the Gaussian Random Variable). *If $X \sim \mathcal{N}(0, \sigma_X^2)$, then it holds for all integer $m \geq 1$ that*

$$\begin{aligned} \mathbb{E}[|X|^m] &= \sigma_X^m (m-1)!! \left[\sqrt{\frac{2}{\pi}} \mathbb{1}_{m=2k+1} + \mathbb{1}_{m=2k} \right] \\ &\leq \sigma_X^m (m-1)!!, \quad k = \lfloor m/2 \rfloor. \end{aligned}$$

Lemma A.5 (Moments of the χ Random Variable). *If $X \sim \chi(n)$, i.e., $X = \|\mathbf{x}\|_2$ for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then it*

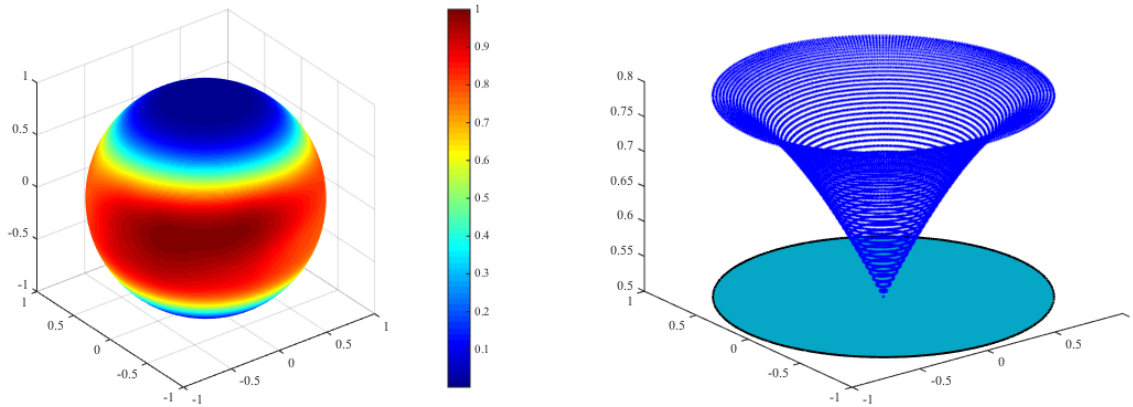


Fig. 6. Function landscape of $f(\mathbf{q})$ with $\theta = 0.4$ for $n = 3$. (Left) $f(\mathbf{q})$ over the sphere \mathbb{S}^2 . Note that near the spherical caps around the north and south poles, there are no critical points and the gradients are always nonzero; (Right) Projected function landscape by projecting the upper hemisphere onto the equatorial plane. Mathematically the function $g(\mathbf{w}) : \mathbf{e}_3^1 \mapsto \mathbb{R}$ obtained via the reparameterization $\mathbf{q}(\mathbf{w}) = [\mathbf{w}; \sqrt{1 - \|\mathbf{w}\|^2}]$. Corresponding to the left, there is no undesired critical point around $\mathbf{0}$ within a large radius.

holds for all integer $m \geq 1$ that

$$\mathbb{E}[X^m] = 2^{m/2} \frac{\Gamma(m/2 + n/2)}{\Gamma(n/2)} \leq m!! n^{m/2}.$$

Lemma A.6 (Moments of the χ^2 Random Variable). *If $X \sim \chi^2(n)$, i.e., $X = \|\mathbf{x}\|_2^2$ for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then it holds for all integer $m \geq 1$ that*

$$\begin{aligned} \mathbb{E}[X^m] &= 2^m \frac{\Gamma(m + n/2)}{\Gamma(n/2)} \\ &= \prod_{k=1}^m (n + 2k - 2) \leq \frac{m!}{2} (2n)^m. \end{aligned}$$

Lemma A.7 (Moment-Control Bernstein's Inequality for Random Variables [67]). *Let X_1, \dots, X_p be i.i.d. real-valued random variables. Suppose that there exist some positive numbers R and σ_X^2 such that*

$$\mathbb{E}[|X_k|^m] \leq \frac{m!}{2} \sigma_X^2 R^{m-2}, \quad \text{for all integers } m \geq 2.$$

Let $S \doteq \frac{1}{p} \sum_{k=1}^p X_k$, then for all $t > 0$, it holds that

$$\mathbb{P}[|S - \mathbb{E}[S]| \geq t] \leq 2 \exp\left(-\frac{pt^2}{2\sigma_X^2 + 2Rt}\right).$$

Lemma A.8 (Moment-Control Bernstein's Inequality for Random Vectors [15]). *Let $\mathbf{x}_1, \dots, \mathbf{x}_p \in \mathbb{R}^d$ be i.i.d. random vectors. Suppose there exist some positive number R and σ_X^2 such that*

$$\mathbb{E}[\|\mathbf{x}_k\|_2^m] \leq \frac{m!}{2} \sigma_X^2 R^{m-2}, \quad \text{for all integers } m \geq 2.$$

Let $\mathbf{s} = \frac{1}{p} \sum_{k=1}^p \mathbf{x}_k$, then for any $t > 0$, it holds that

$$\mathbb{P}[\|\mathbf{s} - \mathbb{E}[\mathbf{s}]\|_2 \geq t] \leq 2(d+1) \exp\left(-\frac{pt^2}{2\sigma_X^2 + 2Rt}\right).$$

Lemma A.9 (Gaussian Concentration Inequality). *Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be an L -Lipschitz function. Then we have for all $t > 0$ that*

$$\mathbb{P}[f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geq t] \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

Lemma A.10 (Bounding Maximum Norm of Gaussian Vector Sequence). *Let $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$ be a sequence of (not necessarily independent) standard Gaussian vectors in \mathbb{R}^{n_2} . It holds that*

$$\mathbb{P}\left[\max_{i \in [n_1]} \|\mathbf{x}_i\|_2 > \sqrt{n_2} + 2\sqrt{2 \log(2n_1)}\right] \leq (2n_1)^{-3}.$$

Proof. Since the function $\|\cdot\|_2$ is 1-Lipschitz, by Gaussian concentration inequality, for any $i \in [n_1]$, we have

$$\begin{aligned} \mathbb{P}\left[\|\mathbf{x}_i\|_2 - \sqrt{\mathbb{E}\|\mathbf{x}_i\|_2^2} > t\right] &\leq \mathbb{P}[\|\mathbf{x}_i\|_2 - \mathbb{E}\|\mathbf{x}_i\|_2 > t] \\ &\leq \exp\left(-\frac{t^2}{2}\right) \end{aligned}$$

for all $t > 0$. Since $\mathbb{E}\|\mathbf{x}_i\|_2^2 = n_2$, by a simple union bound, we obtain

$$\mathbb{P}\left[\max_{i \in [n_1]} \|\mathbf{x}_i\|_2 > \sqrt{n_2} + t\right] \leq \exp\left(-\frac{t^2}{2} + \log n_1\right)$$

for all $t > 0$. Taking $t = 2\sqrt{2 \log(2n_1)}$ gives the claimed result. \square

Corollary A.11. *Let $\Phi \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathcal{N}(0, 1)$. It*

holds that

$$\|\Phi \mathbf{x}\|_\infty \leq \left(\sqrt{n_2} + 2\sqrt{2 \log(2n_1)} \right) \|\mathbf{x}\|_2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^{n_2}, \quad \mathbb{P}[\mathcal{E}^c] \leq 2 \exp\left(-\frac{\delta^2 n_1}{\pi} + n_2 \log\left(1 + \frac{2}{\delta}\right)\right).$$

with probability at least $1 - (2n_1)^{-3}$.

Proof. Let $\Phi = [\phi^1, \dots, \phi^{n_1}]^\top$. Without loss of generality, let us only consider $\mathbf{x} \in \mathbb{S}^{n_2-1}$, we have

$$\|\Phi \mathbf{x}\|_\infty = \max_{i \in [n_1]} |\mathbf{x}^\top \phi^i| \leq \max_{i \in [n_1]} \|\phi^i\|_2. \quad (\text{A.1})$$

Invoking Lemma A.10 returns the claimed result. \square

Lemma A.12 (Covering Number of a Unit Sphere [42]). *Let $\mathbb{S}^{n-1} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1\}$ be the unit sphere. For any $\varepsilon \in (0, 1)$, there exists some ε cover of \mathbb{S}^{n-1} w.r.t. the ℓ^2 norm, denoted as \mathcal{N}_ε , such that*

$$|\mathcal{N}_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^n \leq \left(\frac{3}{\varepsilon}\right)^n.$$

Lemma A.13 (Spectrum of Gaussian Matrices, [42]). *Let $\Phi \in \mathbb{R}^{n_1 \times n_2}$ ($n_1 > n_2$) contain i.i.d. standard normal entries. Then for every $t \geq 0$, with probability at least $1 - 2 \exp(-t^2/2)$, one has*

$$\begin{aligned} \sqrt{n_1} - \sqrt{n_2} - t &\leq \sigma_{\min}(\Phi) \\ &\leq \sigma_{\max}(\Phi) \leq \sqrt{n_1} + \sqrt{n_2} + t. \end{aligned}$$

Lemma A.14. *For any $\varepsilon \in (0, 1)$, there exists a constant $C(\varepsilon) > 1$, such that provided $n_1 > C(\varepsilon)n_2$, the random matrix $\Phi \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathcal{N}(0, 1)$ obeys*

$$(1 - \varepsilon) \sqrt{\frac{2}{\pi}} n_1 \|\mathbf{x}\|_2 \leq \|\Phi \mathbf{x}\|_1 \leq (1 + \varepsilon) \sqrt{\frac{2}{\pi}} n_1 \|\mathbf{x}\|_2,$$

for all $\mathbf{x} \in \mathbb{R}^{n_2}$ with probability at least $1 - 2 \exp(-c(\varepsilon)n_1)$ for some $c(\varepsilon) > 0$.

Geometrically, this lemma roughly corresponds to the well known almost spherical section theorem [68, 69], see also [70]. A slight variant of this version has been proved in [3], borrowing ideas from [71].

Proof. By homogeneity, it is enough to show that the bounds hold for every \mathbf{x} of unit ℓ^2 norm. For a fixed \mathbf{x}_0 with $\|\mathbf{x}_0\|_2 = 1$, $\Phi \mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. So $\mathbb{E} \|\Phi \mathbf{x}\|_1 = \sqrt{\frac{2}{\pi}} n_1$. Note that $\|\cdot\|_1$ is $\sqrt{n_1}$ -Lipschitz, by concentration of measure for Gaussian vectors in Lemma A.9, we have

$$\mathbb{P}[\left| \|\Phi \mathbf{x}\|_1 - \mathbb{E}[\|\Phi \mathbf{x}\|_1] \right| > t] \leq 2 \exp\left(-\frac{t^2}{2n_1}\right)$$

for any $t > 0$. For a fixed $\delta \in (0, 1)$, \mathcal{S}^{n_2-1} can be covered by a δ -net N_δ with cardinality $\#N_\delta \leq (1 + 2/\delta)^{n_2}$. Now consider the event

$$\mathcal{E} \doteq \left\{ \mathbf{x} \in N_\delta \mid (1 - \delta) \sqrt{\frac{2}{\pi}} n_1 \leq \|\Phi \mathbf{x}\|_1 \leq (1 + \delta) \sqrt{\frac{2}{\pi}} n_1 \right\}. \quad \mathbf{g}_\ell \sim_{i.i.d.} \mathcal{N}\left(\mathbf{0}, \frac{1}{p} \mathbf{I}\right), \quad 1 \leq \ell \leq n - 1. \quad (\text{B.1})$$

A simple application of union bound yields

$$\mathbb{P}[\mathcal{E}^c] \leq 2 \exp\left(-\frac{\delta^2 n_1}{\pi} + n_2 \log\left(1 + \frac{2}{\delta}\right)\right).$$

Choosing δ small enough such that

$$(1 - 3\delta)(1 - \delta)^{-1} \geq 1 - \varepsilon, \quad (1 + \delta)(1 - \delta)^{-1} \leq 1 + \varepsilon,$$

then conditioned on \mathcal{E} , we can conclude that

$$(1 - \varepsilon) \sqrt{\frac{2}{\pi}} n_1 \leq \|\Phi \mathbf{x}\|_1 \leq (1 + \varepsilon) \sqrt{\frac{2}{\pi}} n_1 \quad \forall \mathbf{x} \in \mathbb{S}^{n_2-1}.$$

Indeed, suppose \mathcal{E} holds. Then it can easily be seen that any $\mathbf{z} \in \mathbb{S}^{n_2-1}$ can be written as

$$\mathbf{z} = \sum_{k=0}^{\infty} \lambda_k \mathbf{x}_k, \quad \text{with } |\lambda_k| \leq \delta^k, \mathbf{x}_k \in N_\delta \text{ for all } k.$$

Hence we have

$$\begin{aligned} \|\Phi \mathbf{z}\|_1 &= \left\| \Phi \sum_{k=0}^{\infty} \lambda_k \mathbf{x}_k \right\|_1 \leq \sum_{k=0}^{\infty} \delta^k \|\Phi \mathbf{x}_k\|_1 \\ &\leq (1 + \delta)(1 - \delta)^{-1} \sqrt{\frac{2}{\pi}} n_1. \end{aligned}$$

Similarly,

$$\begin{aligned} \|\Phi \mathbf{z}\|_1 &= \left\| \Phi \sum_{k=0}^{\infty} \lambda_k \mathbf{x}_k \right\|_1 \\ &\geq \left[1 - \delta - \delta(1 + \delta)(1 - \delta)^{-1} \right] \sqrt{\frac{2}{\pi}} n_1 \\ &= (1 - 3\delta)(1 - \delta)^{-1} \sqrt{\frac{2}{\pi}} n_1. \end{aligned}$$

Hence, the choice of δ above leads to the claimed result. Finally, given $n_1 > Cn_2$, to make the probability $\mathbb{P}[\mathcal{E}^c]$ decaying in n_1 , it is enough to set $C = \frac{2\pi}{\delta^2} \log\left(1 + \frac{2}{\delta}\right)$. This completes the proof. \square

APPENDIX B

THE RANDOM BASIS VS. ITS ORTHONORMALIZED VERSION

In this appendix, we consider the planted sparse model

$$\bar{\mathbf{Y}} = [\mathbf{x}_0 \mid \mathbf{g}_1 \mid \dots \mid \mathbf{g}_{n-1}] = [\mathbf{x}_0 \mid \mathbf{G}] \in \mathbb{R}^{p \times n}$$

as defined in (III.5), where

$$x_0(k) \sim_{i.i.d.} \frac{1}{\sqrt{\theta p}} \text{Ber}(\theta), \quad 1 \leq k \leq p, \quad (\text{B.1})$$

$$\mathbf{g}_\ell \sim_{i.i.d.} \mathcal{N}\left(\mathbf{0}, \frac{1}{p} \mathbf{I}\right), \quad 1 \leq \ell \leq n - 1.$$

Recall that one ‘‘natural/canonical’’ orthonormal basis for the subspace spanned by columns of $\bar{\mathbf{Y}}$ is

$$\mathbf{Y} = \left[\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \mid \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \left(\mathbf{G}^\top \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right)^{-1/2} \right],$$

which is well-defined with high probability as $\mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G}$ is well-conditioned (proved in Lemma B.2). We write

$$\mathbf{G}' \doteq \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \left(\mathbf{G}^\top \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right)^{-1/2} \quad (\text{B.2})$$

for convenience. When p is large, $\bar{\mathbf{Y}}$ has nearly orthonormal columns, and so we expect that \mathbf{Y} closely approximates $\bar{\mathbf{Y}}$. In this section, we make this intuition rigorous. We prove several results that are needed for the proof of Theorem II.1, and for translating results for $\bar{\mathbf{Y}}$ to results for \mathbf{Y} in Appendix E-D.

For any realization of \mathbf{x}_0 , let $\mathcal{I} = \text{supp}(\mathbf{x}_0) = \{i \mid \mathbf{x}_0(i) \neq 0\}$. By Bernstein’s inequality in Lemma A.7 with $\sigma_X^2 = 2\theta$ and $R = 1$, the event

$$\mathcal{E}_0 \doteq \left\{ \frac{1}{2}\theta p \leq |\mathcal{I}| \leq 2\theta p \right\} \quad (\text{B.3})$$

holds with probability at least $1 - 2\exp(-\theta p/16)$. Moreover, we show the following:

Lemma B.1. *When $p \geq Cn$ and $\theta > 1/\sqrt{n}$, the bound*

$$\left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| \leq \frac{4\sqrt{2}}{5} \sqrt{\frac{n \log p}{\theta^2 p}} \quad (\text{B.4})$$

holds with probability at least $1 - cp^{-2}$. Here C, c are positive constants.

Proof. Because $\mathbb{E} \left[\|\mathbf{x}_0\|_2^2 \right] = 1$, by Bernstein’s inequality in Lemma A.7 with $\sigma_X^2 = 2/(\theta p^2)$ and $R = 1/(\theta p)$, we have

$$\begin{aligned} \mathbb{P} \left[\left| \|\mathbf{x}_0\|_2^2 - \mathbb{E} \left[\|\mathbf{x}_0\|_2^2 \right] \right| > t \right] &= \mathbb{P} \left[\left| \|\mathbf{x}_0\|_2^2 - 1 \right| > t \right] \\ &\leq 2 \exp \left(-\frac{\theta p t^2}{4 + 2t} \right) \end{aligned}$$

for all $t > 0$, which implies

$$\begin{aligned} &\mathbb{P} \left[\left| \|\mathbf{x}_0\|_2 - 1 \right| > \frac{t}{\|\mathbf{x}_0\|_2 + 1} \right] \\ &= \mathbb{P} \left[\left| \|\mathbf{x}_0\|_2 - 1 \right| (\|\mathbf{x}_0\|_2 + 1) > t \right] \\ &\leq 2 \exp \left(-\frac{\theta p t^2}{4 + 2t} \right). \end{aligned}$$

On the intersection with \mathcal{E}_0 , $\|\mathbf{x}_0\|_2 + 1 \geq \frac{1}{2} + 1 \geq 5/4$ and setting $t = \sqrt{\frac{n \log p}{\theta^2 p}}$, we obtain

$$\mathbb{P} \left[\left| \|\mathbf{x}_0\|_2 - 1 \right| \geq \frac{4}{5} \sqrt{\frac{n \log p}{\theta^2 p}} \mid \mathcal{E}_0 \right]$$

$$\leq 2 \exp \left(-\sqrt{np \log p} \right).$$

Unconditionally, this implies that with probability at least $1 - 2 \exp(-p\theta/16) - 2 \exp(-\sqrt{np \log p})$, we have

$$\left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| = \frac{|1 - \|\mathbf{x}_0\|_2|}{\|\mathbf{x}_0\|_2} \leq \frac{4\sqrt{2}}{5} \sqrt{\frac{n \log p}{\theta^2 p}},$$

as desired. \square

Let $\mathbf{M} \doteq \left(\mathbf{G}^\top \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right)^{-1/2}$. Then $\mathbf{G}' = \mathbf{G} \mathbf{M} - \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M}$. We show the following results hold:

Lemma B.2. *Provided $p \geq Cn$, it holds that*

$$\|\mathbf{M}\| \leq 2, \quad \|\mathbf{M} - \mathbf{I}\| \leq 4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}}$$

with probability at least $1 - (2p)^{-2}$. Here C is a positive constant.

Proof. First observe that

$$\|\mathbf{M}\| = \left(\sigma_{\min} \left(\mathbf{G}^\top \mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right) \right)^{-1/2} = \sigma_{\min}^{-1} \left(\mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right).$$

Now suppose \mathbf{B} is an orthonormal basis spanning \mathbf{x}_0^\perp . Then it is not hard to see the spectrum of $\mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G}$ is the same as that of $\mathbf{B}^\top \mathbf{G} \in \mathbb{R}^{(p-1) \times (n-1)}$; in particular,

$$\sigma_{\min} \left(\mathcal{P}_{\mathbf{x}_0^\perp} \mathbf{G} \right) = \sigma_{\min} \left(\mathbf{B}^\top \mathbf{G} \right).$$

Since each entry of $\mathbf{G} \sim_{i.i.d.} \mathcal{N}\left(0, \frac{1}{p}\right)$, and \mathbf{B}^\top has orthonormal rows, $\mathbf{B}^\top \mathbf{G} \sim_{i.i.d.} \mathcal{N}\left(0, \frac{1}{p}\right)$, we can invoke the spectrum results for Gaussian matrices in Lemma A.13 and obtain that

$$\begin{aligned} \sqrt{\frac{p-1}{p}} - \sqrt{\frac{n-1}{p}} - 2\sqrt{\frac{\log(2p)}{p}} &\leq \sigma_{\min} \left(\mathbf{B}^\top \mathbf{G} \right) \\ &\leq \sigma_{\max} \left(\mathbf{B}^\top \mathbf{G} \right) \leq \sqrt{\frac{p-1}{p}} + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}} \end{aligned}$$

with probability at least $1 - (2p)^{-2}$. Thus, when $p \geq C_1 n$ for some sufficiently large constant C_1 , by using the results above we have

$$\begin{aligned} \|\mathbf{M}\| &= \sigma_{\min}^{-1} \left(\mathbf{B}^\top \mathbf{G} \right) \\ &= \left(\sqrt{\frac{p-1}{p}} - \sqrt{\frac{n-1}{p}} - 2\sqrt{\frac{\log(2p)}{p}} \right)^{-1} \leq 2, \end{aligned}$$

$$\begin{aligned} &\|\mathbf{I} - \mathbf{M}\| \\ &= \max \left(|\sigma_{\max}(\mathbf{M}) - 1|, |\sigma_{\min}(\mathbf{M}) - 1| \right) \\ &= \max \left(\left| \sigma_{\min}^{-1} \left(\mathbf{B}^\top \mathbf{G} \right) - 1 \right|, \left| \sigma_{\max}^{-1} \left(\mathbf{B}^\top \mathbf{G} \right) - 1 \right| \right) \end{aligned}$$

$$\begin{aligned}
 &\leq \max \left\{ \left(\sqrt{\frac{p-1}{p}} - \sqrt{\frac{n-1}{p}} - 2\sqrt{\frac{\log(2p)}{p}} \right)^{-1} - 1, \right. \\
 &\quad \left. 1 - \left(\sqrt{\frac{p-1}{p}} + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}} \right)^{-1} \right\} \\
 &= \max \left\{ \left(1 - \sqrt{\frac{p-1}{p}} + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}} \right) \right. \\
 &\quad \times \left(\sqrt{\frac{p-1}{p}} - \sqrt{\frac{n-1}{p}} - 2\sqrt{\frac{\log(2p)}{p}} \right)^{-1}, \\
 &\quad \left(\sqrt{\frac{p-1}{p}} - 1 + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}} \right) \\
 &\quad \times \left. \left(\sqrt{\frac{p-1}{p}} + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}} \right)^{-1} \right\} \\
 &\leq 2 \left(1 - \sqrt{\frac{p-1}{p}} + \sqrt{\frac{n-1}{p}} + 2\sqrt{\frac{\log(2p)}{p}} \right) \\
 &\leq 4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}},
 \end{aligned}$$

with probability at least $1 - (2p)^{-2}$. \square

Lemma B.3. *Let $\mathbf{Y}_{\mathcal{I}}$ be a submatrix of \mathbf{Y} whose rows are indexed by the set \mathcal{I} . There exists a constant $C > 0$, such that when $p \geq Cn$ and $1/2 > \theta > 1/\sqrt{n}$, the following*

$$\begin{aligned}
 \|\bar{\mathbf{Y}}\|_{\ell^2 \rightarrow \ell^1} &\leq 3\sqrt{p}, \\
 \|\mathbf{Y}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} &\leq 7\sqrt{2\theta p}, \\
 \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} &\leq 4\sqrt{n} + 7\sqrt{\log(2p)}, \\
 \|\bar{\mathbf{Y}}_{\mathcal{I}} - \mathbf{Y}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} &\leq 20\sqrt{\frac{n \log p}{\theta}}, \\
 \|\bar{\mathbf{Y}} - \mathbf{Y}\|_{\ell^2 \rightarrow \ell^1} &\leq 20\sqrt{\frac{n \log p}{\theta}}
 \end{aligned}$$

hold simultaneously with probability at least $1 - cp^{-2}$ for a positive constant c .

Proof. First of all, we have

$$\begin{aligned}
 \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^1} &\leq \frac{1}{\|\mathbf{x}_0\|_2^2} \|\mathbf{x}_0\|_{\ell^2 \rightarrow \ell^1} \left\| \mathbf{x}_0^\top \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^2} \\
 &= \frac{2}{\|\mathbf{x}_0\|_2^2} \|\mathbf{x}_0\|_1 \left\| \mathbf{x}_0^\top \mathbf{G} \right\|_2,
 \end{aligned}$$

where in the last inequality we have applied the fact $\|\mathbf{M}\| \leq 2$ from Lemma B.2. Now $\mathbf{x}_0^\top \mathbf{G}$ is an i.i.d. Gaussian vectors with each entry distributed as $\mathcal{N}\left(0, \frac{\|\mathbf{x}_0\|_2^2}{p}\right)$,

where $\|\mathbf{x}_0\|_2^2 = \frac{|\mathcal{I}|}{\theta p}$. So by Gaussian concentration inequality in Lemma A.9, we have

$$\left\| \mathbf{x}_0^\top \mathbf{G} \right\|_2 \leq 2 \|\mathbf{x}_0\|_2 \sqrt{\frac{\log(2p)}{p}}$$

with probability at least $1 - c_1 p^{-2}$. On the intersection with \mathcal{E}_0 , this implies

$$\left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^1} \leq 2\sqrt{2\theta \log(2p)},$$

with probability at least $1 - c_2 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Moreover, when intersected with \mathcal{E}_0 , Lemma A.14 implies that when $p \geq C_1 n$,

$$\|\mathbf{G}\|_{\ell^2 \rightarrow \ell^1} \leq \sqrt{p}, \quad \|\mathbf{G}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} \leq \sqrt{2\theta p}$$

with probability at least $1 - c_3 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Hence, by Lemma B.2, when $p > C_2 n$,

$$\begin{aligned}
 \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} &\leq \|\mathbf{G}\|_{\ell^2 \rightarrow \ell^1} \|\mathbf{I} - \mathbf{M}\| + \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^1} \\
 &\leq \sqrt{p} \left(4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}} \right) + 2\sqrt{2\theta \log(2p)} \\
 &\leq 4\sqrt{n} + 7\sqrt{\log(2p)}, \\
 \|\bar{\mathbf{Y}}\|_{\ell^2 \rightarrow \ell^1} &\leq \|\mathbf{x}_0\|_{\ell^2 \rightarrow \ell^1} + \|\mathbf{G}\|_{\ell^2 \rightarrow \ell^1} \\
 &\leq \|\mathbf{x}_0\|_1 + \sqrt{p} \leq 2\sqrt{\theta p} + \sqrt{p} \leq 3\sqrt{p}, \\
 \|\mathbf{G}'_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} &\leq \|\mathbf{G}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} \|\mathbf{M}\| + \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^1} \\
 &\leq 2\sqrt{2\theta p} + 2\sqrt{2\theta \log(2p)} \leq 4\sqrt{2\theta p}, \\
 \|\mathbf{G}_{\mathcal{I}} - \mathbf{G}'_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} &\leq \|\mathbf{G}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} \|\mathbf{I} - \mathbf{M}\| + \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^1} \\
 &\leq \sqrt{2\theta p} \left(4\sqrt{\frac{n}{p}} + 4\sqrt{\frac{\log(2p)}{p}} \right) \\
 &\quad + 2\sqrt{2\theta \log(2p)} \leq 4\sqrt{2\theta n} + 6\sqrt{2\theta \log(2p)}, \\
 \|\mathbf{Y}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} &\leq \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_{\ell^2 \rightarrow \ell^1} + \|\mathbf{G}'_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} \\
 &\leq \frac{\|\mathbf{x}_0\|_1}{\|\mathbf{x}_0\|_2} + 6\sqrt{2\theta p} \leq 7\sqrt{2\theta p}
 \end{aligned}$$

with probability at least $1 - c_4 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Finally, by Lemma B.1 and the results above, we obtain

$$\begin{aligned}
 \|\bar{\mathbf{Y}} - \mathbf{Y}\|_{\ell^2 \rightarrow \ell^1} &\leq \left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| \|\mathbf{x}_0\|_1 + \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} \\
 &\leq 20\sqrt{\frac{n \log p}{\theta}},
 \end{aligned}$$

$$\begin{aligned} \|\bar{\mathbf{Y}}_{\mathcal{I}} - \mathbf{Y}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} &\leq \left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| \|\mathbf{x}_0\|_1 + \|\mathbf{G}_{\mathcal{I}} - \mathbf{G}'_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} && \leq 2\sqrt{\frac{n}{p}} + 4\sqrt{\frac{2\log(2p)}{p}} + \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta p}} \\ &\leq 20\sqrt{\frac{n\log p}{\theta}}, && \leq 2\sqrt{\frac{n}{p}} + 8\sqrt{\frac{2\log(2p)}{p}}, \end{aligned}$$

holding with probability at least $1 - c_5 p^{-2}$. \square

Lemma B.4. *Provided $p \geq Cn$ and $\theta > 1/\sqrt{n}$, the following*

$$\begin{aligned} \|\mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty} &\leq 2\sqrt{\frac{n}{p}} + 8\sqrt{\frac{2\log(2p)}{p}}, \\ \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty} &\leq \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{21\sqrt{n\log(2p)}}{p} && \leq \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{(8\sqrt{2} + 4)\sqrt{n\log(2p)}}{p} + \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta p}} \\ & && \leq \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{21\sqrt{n\log(2p)}}{p}, \end{aligned}$$

hold simultaneously with probability at least $1 - cp^{-2}$ for some constant $c > 0$.

Proof. First of all, we have when $p \geq C_1 n$, it holds with probability at least $1 - c_2 p^{-2}$ that

$$\begin{aligned} \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^\infty} &\leq \frac{1}{\|\mathbf{x}_0\|_2^2} \|\mathbf{x}_0\|_{\ell^2 \rightarrow \ell^\infty} \left\| \mathbf{x}_0^\top \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^2} \\ &\leq \frac{2}{\|\mathbf{x}_0\|_2^2} \|\mathbf{x}_0\|_\infty \left\| \mathbf{x}_0^\top \mathbf{G} \right\|_2, \end{aligned}$$

where at the last inequality we have applied the fact $\|\mathbf{M}\| \leq 2$ from Lemma B.2. Moreover, from proof of Lemma B.3, we know that $\|\mathbf{x}_0^\top \mathbf{G}\|_2 \leq 2\sqrt{\log(2p)/p} \|\mathbf{x}_0\|_2$ with probability at least $1 - c_3 p^{-2}$ provided $p \geq C_4 n$. Therefore, conditioned on \mathcal{E}_0 , we obtain that

$$\begin{aligned} \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^\infty} &\leq \frac{4\|\mathbf{x}_0\|_\infty}{\|\mathbf{x}_0\|_2} \sqrt{\frac{\log(2p)}{p}} \\ &\leq \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta p}} \end{aligned}$$

holds with probability at least $1 - c_5 p^{-2}$ provided $\theta > 1/\sqrt{n}$. Now by Corollary A.11, we have that

$$\|\mathbf{G}\|_{\ell^2 \rightarrow \ell^\infty} \leq \sqrt{\frac{n}{p}} + 2\sqrt{\frac{2\log(2p)}{p}}$$

with probability at least $1 - c_6 p^{-2}$. Combining the above estimates and Lemma B.2, we have that with probability at least $1 - c_7 p^{-2}$

$$\begin{aligned} \|\mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty} &\leq \|\mathbf{G} \mathbf{M}\|_{\ell^2 \rightarrow \ell^\infty} + \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^\infty} \\ &\leq \|\mathbf{G}\|_{\ell^2 \rightarrow \ell^\infty} \|\mathbf{M}\| + \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^\infty} \end{aligned}$$

where the last simplification is provided that $\theta > 1/\sqrt{n}$ and $p \geq C_8 n$ for a sufficiently large C_8 . Similarly,

$$\begin{aligned} \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty} &\leq \|\mathbf{G}\|_{\ell^2 \rightarrow \ell^\infty} \|\mathbf{I} - \mathbf{M}\| + \left\| \frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2} \mathbf{G} \mathbf{M} \right\|_{\ell^2 \rightarrow \ell^\infty} \\ &\leq \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{(8\sqrt{2} + 4)\sqrt{n\log(2p)}}{p} + \frac{4\sqrt{2\log(2p)}}{\sqrt{\theta p}} \\ &\leq \frac{4n}{p} + \frac{8\sqrt{2}\log(2p)}{p} + \frac{21\sqrt{n\log(2p)}}{p}, \end{aligned}$$

completing the proof. \square

APPENDIX C PROOF OF ℓ^1/ℓ^2 GLOBAL OPTIMALITY

In this appendix, we prove the ℓ^1/ℓ^2 global optimality condition in Theorem II.1 of Section II.

Proof of Theorem II.1. We will first analyze a canonical version, in which the input orthonormal basis is \mathbf{Y} as defined in (III.6) of Section III:

$$\min_{\mathbf{q} \in \mathbb{R}^n} \|\mathbf{Y} \mathbf{q}\|_1, \quad \text{s.t. } \|\mathbf{q}\|_2 = 1.$$

Let $\mathbf{q} = \begin{bmatrix} q_1 \\ \mathbf{q}_2 \end{bmatrix}$ and let \mathcal{I} be the support set of \mathbf{x}_0 , we have

$$\begin{aligned} \|\mathbf{Y} \mathbf{q}\|_1 &= \|\mathbf{Y}_{\mathcal{I}} \mathbf{q}\|_1 + \|\mathbf{Y}_{\mathcal{I}^c} \mathbf{q}\|_1 \\ &\geq |q_1| \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 - \|\mathbf{G}'_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}'_{\mathcal{I}^c} \mathbf{q}_2\|_1 \\ &\geq |q_1| \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 - \|\mathbf{G}_{\mathcal{I}} \mathbf{q}_2\|_1 - \|(\mathbf{G}_{\mathcal{I}} - \mathbf{G}'_{\mathcal{I}}) \mathbf{q}_2\|_1 \\ &\quad + \|\mathbf{G}_{\mathcal{I}^c} \mathbf{q}_2\|_1 - \|(\mathbf{G}_{\mathcal{I}^c} - \mathbf{G}'_{\mathcal{I}^c}) \mathbf{q}_2\|_1 \\ &\geq |q_1| \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 - \|\mathbf{G}_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}_{\mathcal{I}^c} \mathbf{q}_2\|_1 \\ &\quad - \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} \|\mathbf{q}_2\|_2, \end{aligned}$$

where \mathbf{G} and \mathbf{G}' are defined in (B.1) and (B.2) of Appendix B. By Lemma A.14 and intersecting with \mathcal{E}_0 defined in (B.3), as long as $p \geq C_1 n$, we have

$$\begin{aligned} \|\mathbf{G}_{\mathcal{I}} \mathbf{q}_2\|_1 &\leq \frac{2\theta p}{\sqrt{p}} \|\mathbf{q}_2\|_2 = 2\theta \sqrt{p} \|\mathbf{q}_2\|_2, \\ \|\mathbf{G}_{\mathcal{I}^c} \mathbf{q}_2\|_1 &\geq \frac{1}{2} \frac{p - 2\theta p}{\sqrt{p}} \|\mathbf{q}_2\|_2 = \frac{1}{2} \sqrt{p} (1 - 2\theta) \|\mathbf{q}_2\|_2 \end{aligned}$$

for all $\mathbf{q}_2 \in \mathbb{R}^{n-1}$ with probability at least $1 - c_2p^{-2}$. Moreover, by Lemma B.3,

$$\|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} \leq 4\sqrt{n} + 7\sqrt{\log(2p)}$$

hold with probability at least $1 - c_3p^{-2}$ when $p \geq C_4n$ and $\theta > 1/\sqrt{n}$. So we obtain that

$$\begin{aligned} \|\mathbf{Y}\mathbf{q}\|_1 &\geq g(\mathbf{q}) \doteq |q_1| \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 \\ &+ \|\mathbf{q}_2\|_2 \left(\frac{1}{2}\sqrt{p}(1 - 2\theta) - 2\theta\sqrt{p} - 4\sqrt{n} - 7\sqrt{\log(2p)} \right) \end{aligned}$$

holds with probability at least $1 - c_5p^{-2}$. Assuming \mathcal{E}_0 , we observe

$$\left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 \leq \sqrt{|\mathcal{I}|} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_2 \leq \sqrt{2\theta p}.$$

Now $g(\mathbf{q})$ is a linear function in $|q_1|$ and $\|\mathbf{q}_2\|_2$. Thus, whenever θ is sufficiently small and $p \geq C_6n$ such that

$$\sqrt{2\theta p} < \frac{1}{2}\sqrt{p}(1 - 2\theta) - 2\theta\sqrt{p} - 4\sqrt{n} - 7\sqrt{\log(2p)},$$

$\pm \mathbf{e}_1$ are the unique minimizers of $g(\mathbf{q})$ under the constraint $q_1^2 + \|\mathbf{q}_2\|_2^2 = 1$. In this case, because $\|\mathbf{Y}(\pm \mathbf{e}_1)\|_1 = g(\pm \mathbf{e}_1)$, and we have

$$\|\mathbf{Y}\mathbf{q}\|_1 \geq g(\mathbf{q}) > g(\pm \mathbf{e}_1)$$

for all $\mathbf{q} \neq \pm \mathbf{e}_1$, $\pm \mathbf{e}_1$ are the unique minimizers of $\|\mathbf{Y}\mathbf{q}\|_1$ under the spherical constraint. Thus there exists a universal constant $\theta_0 > 0$, such that for all $1/\sqrt{n} \leq \theta \leq \theta_0$, $\pm \mathbf{e}_1$ are the only global minimizers of (II.2) if the input basis is \mathbf{Y} .

Any other input basis can be written as $\widehat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$, for some orthogonal matrix \mathbf{U} . The program now is written as

$$\min_{\mathbf{q} \in \mathbb{R}^n} \left\| \widehat{\mathbf{Y}}\mathbf{q} \right\|_1, \quad \text{s.t. } \|\mathbf{q}\|_2 = 1,$$

which is equivalent to

$$\min_{\mathbf{q} \in \mathbb{R}^n} \left\| \widehat{\mathbf{Y}}\mathbf{q} \right\|_1, \quad \text{s.t. } \|\mathbf{U}\mathbf{q}\|_2 = 1,$$

which is obviously equivalent to the canonical program we analyzed above by a simple change of variable, i.e., $\bar{\mathbf{q}} \doteq \mathbf{U}\mathbf{q}$, completing the proof. \square

APPENDIX D GOOD INITIALIZATION

In this appendix, we prove Proposition IV.3. We show that the initializations produced by the procedure described in Section III are biased towards the optimal.

Proof of Proposition IV.3. Our previous calculation has shown that $\theta p/2 \leq |\mathcal{I}| \leq 2\theta p$ with probability at least

$1 - c_1p^{-2}$ provided $p \geq C_2n$ and $\theta > 1/\sqrt{n}$. Let $\mathbf{Y} = [\mathbf{y}^1, \dots, \mathbf{y}^p]^\top$ as defined in (III.6). Consider any $i \in \mathcal{I}$. Then $x_0(i) = \frac{1}{\sqrt{\theta p}}$, and

$$\begin{aligned} \langle \mathbf{e}_1, \mathbf{y}^i / \|\mathbf{y}^i\|_2 \rangle &= \frac{1/\sqrt{\theta p}}{\|\mathbf{x}_0\|_2 \|\mathbf{y}^i\|_2} \\ &\geq \frac{1/\sqrt{\theta p}}{\|\mathbf{x}_0\|_2 (\|\mathbf{x}_0\|_\infty / \|\mathbf{x}_0\|_2 + \|(\mathbf{g}')^i\|_2)} \\ &\geq \frac{1/\sqrt{\theta p}}{\|\mathbf{x}_0\|_2 (\|\mathbf{x}_0\|_\infty / \|\mathbf{x}_0\|_2 + \|\mathbf{g}^i\|_2 + \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty})}, \end{aligned}$$

where \mathbf{g}^i and $(\mathbf{g}')^i$ are the i -th rows of \mathbf{G} and \mathbf{G}' , respectively. Since such \mathbf{g}^i 's are independent Gaussian vectors in \mathbb{R}^{n-1} distributed as $\mathcal{N}(\mathbf{0}, 1/p)$, by Gaussian concentration inequality and the fact that $|\mathcal{I}| \geq \theta p/2$ w.h.p.,

$$\mathbb{P} \left[\exists i \in \mathcal{I} : \|\mathbf{g}^i\|_2 \leq 2\sqrt{n/p} \right] \geq 1 - \exp(-c_3n\theta p) \leq c_4p^{-2},$$

provided $p \geq C_5n$ and $\theta > 1/\sqrt{n}$. Moreover,

$$\|\mathbf{x}_0\|_2 = \sqrt{|\mathcal{I}|} \times \frac{1}{\theta p} \leq \sqrt{2\theta p} \times \frac{1}{\theta p} = \sqrt{2}.$$

Combining the above estimates and result of Lemma B.4, we obtain that provided $p \geq C_6n$ and $\theta > 1/\sqrt{n}$, with probability at least $1 - c_7p^{-2}$, there exists an $i \in [p]$, such that if we set $\mathbf{q}^{(0)} = \mathbf{y}^i / \|\mathbf{y}^i\|_2$, it holds that

$$\begin{aligned} |q_1^{(0)}| &\geq \frac{1}{\sqrt{\theta p}} \left[\sqrt{2} \left(\frac{4n}{p} + 8\sqrt{2} \frac{\log(2p)}{p} + 21 \frac{\sqrt{n \log(2p)}}{p} \right) \right. \\ &\quad \left. + \frac{1}{\sqrt{\theta p}} + 2\sqrt{\frac{2n}{p}} \right]^{-1} \\ &\geq \frac{1/\sqrt{\theta p}}{1/\sqrt{\theta p} + 6\sqrt{2}\sqrt{n/p}} \quad (\text{using } p \geq C_6n) \\ &= \frac{1}{1 + 6\sqrt{2}\sqrt{\theta n}} \\ &\geq \frac{1}{(1 + 6\sqrt{2})\sqrt{\theta n}} \quad (\text{as } \theta > 1/\sqrt{n}) \\ &\geq \frac{1}{10\sqrt{\theta n}}, \end{aligned}$$

completing the proof. \square

We will next show that for an arbitrary orthonormal basis $\widehat{\mathbf{Y}} \doteq \mathbf{Y}\mathbf{U}$ the initialization still biases towards the target solution. To see this, suppose w.l.o.g. $(\mathbf{y}^i)^\top$ is a row of \mathbf{Y} with nonzero first coordinate. We have shown above that with high probability $\left| \left\langle \frac{\mathbf{y}^i}{\|\mathbf{y}^i\|_2}, \mathbf{e}_1 \right\rangle \right| \geq \frac{1}{10\sqrt{\theta n}}$ if \mathbf{Y} is the input orthonormal basis. For $\widehat{\mathbf{Y}}$, as $\mathbf{x}_0 = \mathbf{Y}\mathbf{e}_1 = \mathbf{Y}\mathbf{U}\mathbf{U}^\top \mathbf{e}_1$, we know $\mathbf{q}_* = \mathbf{U}^\top \mathbf{e}_1$ is the target

solution corresponding to $\hat{\mathbf{Y}}$. Observing that

$$\begin{aligned} \left| \left\langle \mathbf{U}^\top \mathbf{e}_1, \frac{\left(\mathbf{e}_i^\top \hat{\mathbf{Y}} \right)^\top}{\left\| \left(\mathbf{e}_i^\top \hat{\mathbf{Y}} \right)^\top \right\|_2} \right\rangle \right| &= \left| \left\langle \mathbf{U}^\top \mathbf{e}_1, \frac{\mathbf{U}^\top \mathbf{Y}^\top \mathbf{e}_i}{\left\| \mathbf{U}^\top \mathbf{Y}^\top \mathbf{e}_i \right\|_2} \right\rangle \right| \\ &= \left| \left\langle \mathbf{e}_1, \frac{\left(\mathbf{Y} \right)^\top \mathbf{e}_i}{\left\| \mathbf{Y}^\top \mathbf{e}_i \right\|_2} \right\rangle \right| \\ &= \left| \left\langle \mathbf{e}_1, \frac{\mathbf{y}^i}{\left\| \mathbf{y}^i \right\|_2} \right\rangle \right| \geq \frac{1}{10\sqrt{n\theta}}, \end{aligned}$$

corroborating our claim.

APPENDIX E

LOWER BOUNDING FINITE SAMPLE GAP $G(\mathbf{q})$

In this appendix, we prove Proposition IV.4. In particular, we show that the gap $G(\mathbf{q})$ defined in (IV.8) is strictly positive over a large portion of the sphere \mathbb{S}^{n-1} .

Proof of Proposition IV.4. Without loss of generality, we work with the ‘‘canonical’’ orthonormal basis \mathbf{Y} defined in (III.6). Recall that \mathbf{Y} is the orthogonalization of the planted sparse basis $\bar{\mathbf{Y}}$ as defined in (III.5). We define the processes $\bar{\mathbf{Q}}(\mathbf{q})$ and $\mathbf{Q}(\mathbf{q})$ on $\mathbf{q} \in \mathbb{S}^{n-1}$, via

$$\begin{aligned} \bar{\mathbf{Q}}(\mathbf{q}) &= \frac{1}{p} \sum_{i=1}^p \bar{\mathbf{y}}^i S_\lambda \left[\mathbf{q}^\top \bar{\mathbf{y}}^i \right], \\ \mathbf{Q}(\mathbf{q}) &= \frac{1}{p} \sum_{i=1}^p \mathbf{y}^i S_\lambda \left[\mathbf{q}^\top \mathbf{y}^i \right]. \end{aligned}$$

Thus, we can separate $\bar{\mathbf{Q}}(\mathbf{q})$ as $\bar{\mathbf{Q}}(\mathbf{q}) = \begin{bmatrix} \bar{Q}_1(\mathbf{q}) \\ \bar{\mathbf{Q}}_2(\mathbf{q}) \end{bmatrix}$, where

$$\begin{aligned} \bar{Q}_1(\mathbf{q}) &= \frac{1}{p} \sum_{i=1}^p x_{0i} S_\lambda \left[\mathbf{q}^\top \bar{\mathbf{y}}^i \right] \\ \bar{\mathbf{Q}}_2(\mathbf{q}) &= \frac{1}{p} \sum_{i=1}^p \mathbf{g}_i S_\lambda \left[\mathbf{q}^\top \bar{\mathbf{y}}^i \right], \end{aligned} \tag{E.1}$$

and separate $\mathbf{Q}(\mathbf{q})$ correspondingly. Our task is to lower bound the gap $G(\mathbf{q})$ for finite samples as defined in (IV.8). Since we can deterministically constrain $|q_1|$ and $\|\mathbf{q}_2\|_2$ over the set Γ as defined in (IV.7) (e.g., $\frac{1}{10\sqrt{n\theta}} \leq |q_1| \leq 3\sqrt{\theta}$ and $\|\mathbf{q}_2\|_2 \geq \frac{1}{10}$, where the choice of $\frac{1}{10}$ for \mathbf{q}_2 is arbitrary here, as we can always take a sufficiently small θ), the challenge lies in lower bounding $|Q_1(\mathbf{q})|$ and upper bounding $\|\mathbf{Q}_2(\mathbf{q})\|_2$, which depend on the orthonormal basis \mathbf{Y} . The unnormalized basis $\bar{\mathbf{Y}}$ is much

easier to work with than \mathbf{Y} . Our proof will follow the observation that

$$\begin{aligned} |Q_1(\mathbf{q})| &\geq |\mathbb{E}[\bar{Q}_1(\mathbf{q})]| - |\bar{Q}_1(\mathbf{q}) - \mathbb{E}[\bar{Q}_1(\mathbf{q})]| \\ &\quad - |Q_1(\mathbf{q}) - \bar{Q}_1(\mathbf{q})|, \\ \|\mathbf{Q}_2(\mathbf{q})\| &\leq \|\mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2 + \|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2 \\ &\quad + \|\mathbf{Q}_2(\mathbf{q}) - \bar{\mathbf{Q}}_2(\mathbf{q})\|_2. \end{aligned}$$

In particular, we show the following:

- Appendix E-A shows that the expected gap is lower bounded for all $\mathbf{q} \in \mathbb{S}^{n-1}$ with $|q_1| \leq 3\sqrt{\theta}$:

$$\bar{G}(\mathbf{q}) \doteq \frac{|\mathbb{E}[\bar{Q}_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \geq \frac{1}{50} \frac{q_1^2}{\theta p}.$$

As $|q_1| \geq \frac{1}{10\sqrt{n\theta}}$, we have

$$\inf_{\mathbf{q} \in \Gamma} \frac{|\mathbb{E}[\bar{Q}_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \geq \frac{1}{5000} \frac{1}{\theta^2 n p}.$$

- Appendix E-B, as summarized in Proposition E.8, shows that whenever $p \geq \Omega(n^4 \log n)$, it holds with high probability that

$$\begin{aligned} &\sup_{\mathbf{q} \in \Gamma} \frac{|\bar{Q}_1(\mathbf{q}) - \mathbb{E}[\bar{Q}_1(\mathbf{q})]|}{|q_1|} + \frac{\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \\ &\leq \frac{10\sqrt{\theta n}}{4 \times 10^5 \theta^{5/2} n^{3/2} p} + \frac{10}{4 \times 10^5 \theta^2 n p} = \frac{1}{2 \times 10^4 \theta^2 n p}. \end{aligned}$$

- Appendix E-D shows that whenever $p \geq \Omega(n^4 \log n)$, it holds with high probability that

$$\begin{aligned} &\sup_{\mathbf{q} \in \Gamma} \frac{|\bar{Q}_1(\mathbf{q}) - Q_1(\mathbf{q})|}{|q_1|} + \frac{\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2} \\ &\leq \frac{10\sqrt{\theta n}}{4 \times 10^5 \theta^{5/2} n^{3/2} p} + \frac{10}{4 \times 10^5 \theta^2 n p} = \frac{1}{2 \times 10^4 \theta^2 n p}. \end{aligned}$$

Observing that

$$\begin{aligned} &\inf_{\mathbf{q} \in \Gamma} G(\mathbf{q}) \\ &\geq \inf_{\mathbf{q} \in \Gamma} \left(\frac{|\mathbb{E}[\bar{Q}_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \right) \\ &\quad - \sup_{\mathbf{q} \in \Gamma} \left(\frac{|\bar{Q}_1(\mathbf{q}) - \mathbb{E}[\bar{Q}_1(\mathbf{q})]|}{|q_1|} + \frac{\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E}[\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \right) \\ &\quad - \sup_{\mathbf{q} \in \Gamma} \left(\frac{|\bar{Q}_1(\mathbf{q}) - Q_1(\mathbf{q})|}{|q_1|} + \frac{\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2} \right), \end{aligned}$$

we obtain the result as desired. \square

For the general case when the input orthonormal basis is $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$ with target solution $\mathbf{q}_\star = \mathbf{U}^\top \mathbf{e}_1$, a straightforward extension of the definition for the gap

would be:

$$G(\mathbf{q}; \hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}) \doteq \frac{|\langle \mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}), \mathbf{U}^\top \mathbf{e}_1 \rangle|}{|\langle \mathbf{q}, \mathbf{U}^\top \mathbf{e}_1 \rangle|} - \frac{\|(\mathbf{I} - \mathbf{U}^\top \mathbf{e}_1 \mathbf{e}_1^\top \mathbf{U}) \mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}})\|_2}{\|(\mathbf{I} - \mathbf{U}^\top \mathbf{e}_1 \mathbf{e}_1^\top \mathbf{U}) \mathbf{q}\|_2}.$$

Since $\mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}) = \frac{1}{p} \sum_{k=1}^p \mathbf{U}^\top \mathbf{y}^k S_\lambda(\mathbf{q}^\top \mathbf{U}^\top \mathbf{y}^k)$, we have

$$\begin{aligned} \mathbf{U}\mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}) &= \frac{1}{p} \sum_{k=1}^p \mathbf{U}\mathbf{U}^\top \mathbf{y}^k S_\lambda(\mathbf{q}^\top \mathbf{U}^\top \mathbf{y}^k) \\ &= \frac{1}{p} \sum_{k=1}^p \mathbf{y}^k S_\lambda[(\mathbf{U}\mathbf{q})^\top \mathbf{y}^k] \\ &= \mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y}). \end{aligned} \quad (\text{E.2})$$

Hence we have

$$G(\mathbf{q}; \hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}) = \frac{|\langle \mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y}), \mathbf{e}_1 \rangle|}{|\langle \mathbf{U}\mathbf{q}, \mathbf{e}_1 \rangle|} - \frac{\|(\mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^\top) \mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y})\|_2}{\|(\mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^\top) \mathbf{U}\mathbf{q}\|_2}.$$

Therefore, from Proposition IV.4 above, we conclude that under the same technical conditions as therein,

$$\inf_{\mathbf{q} \in \mathbb{S}^{n-1}: \frac{1}{10\sqrt{\theta n}} \leq |\langle \mathbf{U}\mathbf{q}, \mathbf{e}_1 \rangle| \leq 3\sqrt{\theta}} G(\mathbf{q}; \hat{\mathbf{Y}}) \geq \frac{1}{10^4 \theta^2 n p}$$

with high probability.

A. Lower Bounding the Expected Gap $\bar{G}(\mathbf{q})$

In this section, we provide a nontrivial lower bound for the gap

$$\bar{G}(\mathbf{q}) = \frac{|\mathbb{E}[\bar{Q}_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\bar{Q}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2}. \quad (\text{E.3})$$

More specifically, we show that:

Proposition E.1. *There exists some numerical constant $\theta_0 > 0$, such that for all $\theta \in (0, \theta_0)$, it holds that*

$$\bar{G}(\mathbf{q}) \geq \frac{1}{50} \frac{q_1^2}{\theta p} \quad (\text{E.4})$$

for all $\mathbf{q} \in \mathbb{S}^{n-1}$ with $|q_1| \leq 3\sqrt{\theta}$.

Estimating the gap $\bar{G}(\mathbf{q})$ requires delicate estimates for $\mathbb{E}[\bar{Q}_1(\mathbf{q})]$ and $\mathbb{E}[\bar{Q}_2(\mathbf{q})]$. We first outline the main proof in Appendix E-A1, and delay these detailed technical calculations to the subsequent subsections.

1) *Sketch of the Proof:* W.l.o.g., we only consider the situation that $q_1 > 0$, because the case of $q_1 < 0$ can be

similarly shown by symmetry. By (E.1), we have

$$\begin{aligned} \mathbb{E}[\bar{Q}_1(\mathbf{q})] &= \mathbb{E}\left[x_0 S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right], \\ \mathbb{E}[\bar{Q}_2(\mathbf{q})] &= \mathbb{E}\left[\mathbf{g} S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right], \end{aligned}$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I})$, and $x_0 \sim \frac{1}{\sqrt{\theta p}} \text{Ber}(\theta)$. Let us decompose

$$\mathbf{g} = \mathbf{g}_\parallel + \mathbf{g}_\perp,$$

with $\mathbf{g}_\parallel = \mathcal{P}_\parallel \mathbf{g} = \frac{\mathbf{q}_2 \mathbf{q}_2^\top}{\|\mathbf{q}_2\|_2^2} \mathbf{g}$, and $\mathbf{g}_\perp = (\mathbf{I} - \mathcal{P}_\parallel) \mathbf{g}$. In this notation, we have

$$\begin{aligned} \mathbb{E}[\bar{Q}_2(\mathbf{q})] &= \mathbb{E}\left[\mathbf{g}_\parallel S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}_\parallel\right]\right] \\ &\quad + \mathbb{E}\left[\mathbf{g}_\perp S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}_\perp\right]\right] \\ &= \mathbb{E}\left[\mathbf{g}_\parallel S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right] \\ &\quad + \mathbb{E}[\mathbf{g}_\perp] \mathbb{E}\left[S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right] \\ &= \frac{\mathbf{q}_2}{\|\mathbf{q}_2\|_2^2} \mathbb{E}\left[\mathbf{q}_2^\top \mathbf{g} S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right], \end{aligned}$$

where we used the facts that $\mathbf{q}_2^\top \mathbf{g} = \mathbf{q}_2^\top \mathbf{g}_\parallel$, \mathbf{g}_\perp and \mathbf{g}_\parallel are uncorrelated Gaussian vectors and therefore independent, and $\mathbb{E}[\mathbf{g}_\perp] = \mathbf{0}$. Let $Z \doteq \mathbf{g}^\top \mathbf{q}_2 \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \|\mathbf{q}_2\|_2^2/p$, by partial evaluation of the expectations with respect to x_0 , we get

$$\mathbb{E}[\bar{Q}_1(\mathbf{q})] = \sqrt{\frac{\theta}{p}} \mathbb{E}\left[S_\lambda\left[\frac{q_1}{\sqrt{\theta p}} + Z\right]\right], \quad (\text{E.5})$$

$$\begin{aligned} \mathbb{E}[\bar{Q}_2(\mathbf{q})] &= \frac{\theta \mathbf{q}_2}{\|\mathbf{q}_2\|_2^2} \mathbb{E}\left[Z S_\lambda\left[\frac{q_1}{\sqrt{\theta p}} + Z\right]\right] \\ &\quad + \frac{(1-\theta)\mathbf{q}_2}{\|\mathbf{q}_2\|_2^2} \mathbb{E}[Z S_\lambda[Z]]. \end{aligned} \quad (\text{E.6})$$

Straightforward integration based on Lemma A.1 gives an explicit form of the expectations as follows

$$\begin{aligned} \mathbb{E}[\bar{Q}_1(\mathbf{q})] &= \sqrt{\frac{\theta}{p}} \left\{ \left[\alpha \Psi\left(-\frac{\alpha}{\sigma}\right) + \beta \Psi\left(\frac{\beta}{\sigma}\right) \right] \right. \\ &\quad \left. + \sigma \left[\psi\left(-\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right) \right] \right\} \end{aligned} \quad (\text{E.7})$$

$$\begin{aligned} \mathbb{E}[\bar{Q}_2(\mathbf{q})] &= \mathbf{q}_2 \left\{ \frac{2(1-\theta)}{p} \Psi\left(-\frac{\lambda}{\sigma}\right) \right. \\ &\quad \left. + \frac{\theta}{p} \left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) \right] \right\} \end{aligned} \quad (\text{E.8})$$

where the scalars α and β are defined as

$$\alpha = \frac{q_1}{\sqrt{\theta p}} + \lambda, \quad \beta = \frac{q_1}{\sqrt{\theta p}} - \lambda,$$

and $\psi(t)$ and $\Psi(t)$ are *pdf* and *cdf* for standard normal distribution, respectively, as defined in Lemma A.1. Plugging (E.7) and (E.8) into (E.3), by some simplifications, we obtain

$$\begin{aligned} \bar{G}(\mathbf{q}) &= \frac{1}{q_1} \sqrt{\frac{\theta}{p}} \left[\alpha \Psi\left(-\frac{\alpha}{\sigma}\right) + \beta \Psi\left(\frac{\beta}{\sigma}\right) - \frac{2q_1}{\sqrt{\theta p}} \Psi\left(-\frac{\lambda}{\sigma}\right) \right] \\ &\quad - \frac{\theta}{p} \left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) - 2\Psi\left(-\frac{\lambda}{\sigma}\right) \right] \\ &\quad + \frac{\sigma}{q_1} \sqrt{\frac{\theta}{p}} \left[\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right) \right]. \end{aligned} \quad (\text{E.9})$$

With $\lambda = 1/\sqrt{p}$ and $\sigma^2 = \|\mathbf{q}_2\|_2^2/p = (1 - q_1^2)/p$, we have

$$-\frac{\alpha}{\sigma} = -\frac{\delta + 1}{\sqrt{1 - q_1^2}}, \quad \frac{\beta}{\sigma} = \frac{\delta - 1}{\sqrt{1 - q_1^2}}, \quad \frac{\lambda}{\sigma} = \frac{1}{\sqrt{1 - q_1^2}},$$

where $\delta = q_1/\sqrt{\theta}$ for $q_1 \leq 3\sqrt{\theta}$. To proceed, it is natural to consider estimating the gap $\bar{G}(\mathbf{q})$ by Taylor's expansion. More specifically, we approximate $\Psi(-\frac{\alpha}{\sigma})$ and $\psi(-\frac{\alpha}{\sigma})$ around $-1 - \delta$, and approximate $\Psi(\frac{\beta}{\sigma})$ and $\psi(\frac{\beta}{\sigma})$ around $-1 + \delta$. Applying the estimates for the relevant quantities established in Lemma E.2, we obtain

$$\begin{aligned} \bar{G}(\mathbf{q}) &\geq \frac{1 - \theta}{p} \Phi_1(\delta) - \frac{1}{\delta p} \Phi_2(\delta) + \frac{1 - \theta}{p} \psi(-1) q_1^2 \\ &\quad + \frac{1}{p} \left(\sigma\sqrt{p} + \frac{\theta}{2} - 1 \right) \eta_2(\delta) q_1^2 \\ &\quad + \frac{1}{2\delta p} [1 + \delta^2 - \theta\delta^2 - \sigma(1 + \delta^2)\sqrt{p}] q_1^2 \eta_1(\delta) \\ &\quad + \frac{\sigma}{\delta\sqrt{p}} \eta_1(\delta) - \frac{5C_T\sqrt{\theta}q_1^3}{p} (\delta + 1)^3, \end{aligned}$$

where we define

$$\begin{aligned} \Phi_1(\delta) &= \Psi(-1 - \delta) + \Psi(-1 + \delta) - 2\Psi(-1), \\ \Phi_2(\delta) &= \Psi(-1 + \delta) - \Psi(-1 - \delta), \\ \eta_1(\delta) &= \psi(-1 + \delta) - \psi(-1 - \delta), \\ \eta_2(\delta) &= \psi(-1 + \delta) + \psi(-1 - \delta), \end{aligned}$$

and C_T is as defined in Lemma E.2. Since $1 - \sigma\sqrt{p} \geq 0$, dropping those small positive terms $\frac{q_1^2}{p}(1 - \theta)\psi(-1)$, $\frac{\theta q_1^2}{2p}\eta_2(\delta)$, and $(1 + \delta^2)(1 - \sigma\sqrt{p})q_1^2\eta_1(\delta)/(2\delta p)$, and using the fact that $\delta = q_1/\sqrt{\theta}$, we obtain

$$\begin{aligned} \bar{G}(\mathbf{q}) &\geq \frac{1 - \theta}{p} \Phi_1(\delta) - \frac{1}{\delta p} [\Phi_2(\delta) - \sigma\sqrt{p}\eta_1(\delta)] \\ &\quad - \frac{q_1^2}{p} (1 - \sigma\sqrt{p}) \eta_2(\delta) - \frac{\sqrt{\theta}}{2p} q_1^3 \eta_1(\delta) \end{aligned}$$

$$\begin{aligned} &- \frac{C_1\sqrt{\theta}q_1^3}{p} \max\left(\frac{q_1^3}{\theta^{3/2}}, 1\right) \\ &\geq \frac{1 - \theta}{p} \Phi_1(\delta) - \frac{1}{\delta p} [\Phi_2(\delta) - \eta_1(\delta)] \\ &\quad - \frac{q_1^2}{p} \frac{\eta_1(\delta)}{\delta} - \frac{q_1^2}{\theta p} \left(\frac{2\theta}{\sqrt{2\pi}} + \frac{3\theta^2}{2\sqrt{2\pi}} + C_1\theta^2 \right), \end{aligned}$$

for some constant $C_1 > 0$, where we have used $q_1 \leq 3\sqrt{\theta}$ to simplify the bounds and the fact $\sigma\sqrt{p} = \sqrt{1 - q_1^2} \geq 1 - q_1^2$ to simplify the expression. Substituting the estimates in Lemma E.4 and use the fact $\delta \mapsto \eta_1(\delta)/\delta$ is bounded, we obtain

$$\begin{aligned} \bar{G}(p) &\geq \frac{1}{p} \left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta \right) \delta^2 - \frac{q_1^2}{\theta p} (c_1\theta + c_2\theta^2) \\ &\geq \frac{q_1^2}{\theta p} \left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta - c_1\theta - c_2\theta^2 \right) \end{aligned}$$

for some positive constants c_1 and c_2 . We obtain the claimed result once θ_0 is made sufficiently small.

2) Auxiliary Results Used in the Proof:

Lemma E.2. *Let $\delta \doteq q_1/\sqrt{\theta}$. There exists some universal constant $C_T > 0$ such that we have the follow polynomial approximations hold for all $q_1 \in (0, \frac{1}{2})$:*

$$\begin{aligned} \left| \psi\left(-\frac{\alpha}{\sigma}\right) - \left[1 - \frac{1}{2}(1 + \delta)^2 q_1^2\right] \psi(-1 - \delta) \right| &\leq C_T (1 + \delta)^2 q_1^4, \\ \left| \psi\left(\frac{\beta}{\sigma}\right) - \left[1 - \frac{1}{2}(\delta - 1)^2 q_1^2\right] \psi(\delta - 1) \right| &\leq C_T (\delta - 1)^2 q_1^4, \\ \left| \Psi\left(-\frac{\alpha}{\sigma}\right) - \left[\Psi(-1 - \delta) - \frac{1}{2}\psi(-1 - \delta)(1 + \delta)q_1^2 \right] \right| &\leq C_T (1 + \delta)^2 q_1^4, \\ \left| \Psi\left(\frac{\beta}{\sigma}\right) - \left[\Psi(\delta - 1) + \frac{1}{2}\psi(\delta - 1)(\delta - 1)q_1^2 \right] \right| &\leq C_T (\delta - 1)^2 q_1^4, \\ \left| \Psi\left(-\frac{\lambda}{\sigma}\right) - \left[\Psi(-1) - \frac{1}{2}\psi(-1)q_1^2 \right] \right| &\leq C_T q_1^4. \end{aligned}$$

Proof. First observe that for any $q_1 \in (0, \frac{1}{2})$ it holds that

$$0 \leq \frac{1}{\sqrt{1 - q_1^2}} - \left(1 + \frac{q_1^2}{2}\right) \leq q_1^4.$$

Thus, we have

$$-(1 + \delta) \left(1 + \frac{1}{2}q_1^2 + q_1^4\right) \leq -\frac{\alpha}{\sigma} \leq -(1 + \delta) \left(1 + \frac{1}{2}q_1^2\right),$$

when $\delta \geq 1$, we have

$$(\delta - 1) \left(1 + \frac{1}{2}q_1^2\right) \leq \frac{\beta}{\sigma} \leq (\delta - 1) \left(1 + \frac{1}{2}q_1^2 + q_1^4\right),$$

when $\delta \leq 1$, we have

$$(\delta - 1) \left(1 + \frac{1}{2}q_1^2 + q_1^4 \right) \leq \frac{\beta}{\sigma} \leq (\delta - 1) \left(1 + \frac{1}{2}q_1^2 \right).$$

So we obtain

$$\begin{aligned} \psi \left(-(1 + \delta) \left(1 + \frac{1}{2}q_1^2 + q_1^4 \right) \right) &\leq \psi \left(-\frac{\alpha}{\sigma} \right) \\ &\leq \psi \left(-(1 + \delta) \left(1 + \frac{1}{2}q_1^2 \right) \right). \end{aligned}$$

By Taylor expansion of the left and right sides of the above two-side inequality around $-1 - \delta$ using Lemma A.2, we obtain

$$\begin{aligned} \left| \psi \left(-\frac{\alpha}{\sigma} \right) - \psi(-1 - \delta) - \frac{1}{2}(1 + \delta)^2 q_1^2 \psi(-1 - \delta) \right| \\ \leq C_T (1 + \delta)^2 q_1^4, \end{aligned}$$

for some numerical constant $C_T > 0$ sufficiently large. In the same way, we can obtain other claimed results. \square

Lemma E.3. *For any $\delta \in [0, 3]$, it holds that*

$$\Phi_2(\delta) - \eta_1(\delta) \geq \frac{\eta_1(3)}{9} \delta^3 \geq \frac{1}{20} \delta^3. \quad (\text{E.10})$$

Proof. Let us define

$$h(\delta) = \Phi_2(\delta) - \eta_1(\delta) - C\delta^3$$

for some $C > 0$ to be determined later. Then it is obvious that $h(0) = 0$. Direct calculation shows that

$$\begin{aligned} \frac{d}{d\delta} \Phi_1(\delta) &= \eta_1(\delta), & \frac{d}{d\delta} \Phi_2(\delta) &= \eta_2(\delta), \\ \frac{d}{d\delta} \eta_1(\delta) &= \eta_2(\delta) - \delta\eta_1(\delta). \end{aligned} \quad (\text{E.11})$$

Thus, to show (E.10), it is sufficient to show that $h'(\delta) \geq 0$ for all $\delta \in [0, 3]$. By differentiating $h(\delta)$ with respect to δ and use the results in (E.11), it is sufficient to have

$$h'(\delta) = \delta\eta_1(\delta) - 3C\delta^2 \geq 0 \iff \eta_1(\delta) \geq 3C\delta$$

for all $\delta \in [0, 3]$. We obtain the claimed result by observing that $\delta \mapsto \eta_1(\delta)/3\delta$ is monotonically decreasing over $\delta \in [0, 3]$ as justified below.

Consider the function

$$p(\delta) \doteq \frac{\eta_1(\delta)}{3\delta} = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{\delta^2 + 1}{2}\right) \frac{e^\delta - e^{-\delta}}{\delta}.$$

To show it is monotonically decreasing, it is enough to show $p'(\delta)$ is always nonpositive for $\delta \in (0, 3)$, or equivalently

$$g(\delta) \doteq \left(e^\delta + e^{-\delta} \right) \delta - (\delta^2 + 1) \left(e^\delta - e^{-\delta} \right) \leq 0$$

for all $\delta \in (0, 3)$, which can be easily verified by noticing that $g(0) = 0$ and $g'(\delta) \leq 0$ for all $\delta \geq 0$. \square

Lemma E.4. *For any $\delta \in [0, 3]$, we have*

$$(1 - \theta)\Phi_1(\delta) - \frac{1}{\delta} [\Phi_2(\delta) - \eta_1(\delta)] \geq \left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta \right) \delta^2. \quad (\text{E.12})$$

Proof. Let us define

$$g(\delta) = (1 - \theta)\Phi_1(\delta) - \frac{1}{\delta} [\Phi_2(\delta) - \eta_1(\delta)] - c_0(\theta) \delta^2,$$

where $c_0(\theta) > 0$ is a function of θ . Thus, by the results in (E.11) and L'Hospital's rule, we have

$$\begin{aligned} \lim_{\delta \rightarrow 0} \frac{\Phi_2(\delta)}{\delta} &= \lim_{\delta \rightarrow 0} \eta_2(\delta) = 2\psi(-1), \\ \lim_{\delta \rightarrow 0} \frac{\eta_1(\delta)}{\delta} &= \lim_{\delta \rightarrow 0} [\eta_2(\delta) - \delta\eta_1(\delta)] = 2\psi(-1). \end{aligned}$$

Combined that with the fact that $\Phi_1(0) = 0$, we conclude $g(0) = 0$. Hence, to show (E.12), it is sufficient to show that $g'(\delta) \geq 0$ for all $\delta \in [0, 3]$. Direct calculation using the results in (E.11) shows that

$$g'(\delta) = \frac{1}{\delta^2} [\Phi_2(\delta) - \eta_1(\delta)] - \theta\eta_1(\delta) - 2c_0(\theta) \delta.$$

Since $\eta_1(\delta)/\delta$ is monotonically decreasing as shown in Lemma E.3, we have that for all $\delta \in (0, 3)$

$$\eta_1(\delta) \leq \delta \lim_{\delta \rightarrow 0} \frac{\eta_1(\delta)}{\delta} \leq \frac{2}{\sqrt{2\pi}} \delta.$$

Using the above bound and the main result from Lemma E.3 again, we obtain

$$g'(\delta) \geq \frac{1}{20} \delta - \frac{2}{\sqrt{2\pi}} \theta \delta - 2c_0 \delta.$$

Choosing $c_0(\theta) = \frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta$ completes the proof. \square

B. Finite Sample Concentration

In the following two subsections, we estimate the deviations around the expectations $\mathbb{E}[\bar{Q}_1(\mathbf{q})]$ and $\mathbb{E}[\mathbf{Q}_2(\mathbf{q})]$, i.e., $|\bar{Q}_1(\mathbf{q}) - \mathbb{E}[\bar{Q}_1(\mathbf{q})]|$ and $\|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2$, and show that the total deviations fit into the gap $\bar{G}(\mathbf{q})$ we derived in Appendix E-A. Our analysis is based on the scalar and vector Bernstein's inequalities with moment conditions. Finally, in Appendix E-C, we uniformize the bound by applying the classical discretization argument.

1) Concentration for $\bar{Q}_1(\mathbf{q})$:

Lemma E.5 (Bounding $|\bar{Q}_1(\mathbf{q}) - \mathbb{E}[\bar{Q}_1(\mathbf{q})]|$). *For each $\mathbf{q} \in \mathbb{S}^{n-1}$, it holds for all $t > 0$ that*

$$\mathbb{P} [|\bar{Q}_1(\mathbf{q}) - \mathbb{E}[\bar{Q}_1(\mathbf{q})]| \geq t] \leq 2 \exp \left(-\frac{\theta p^3 t^2}{8 + 4pt} \right).$$

Proof. By (E.1), we know that

$$\bar{Q}_1(\mathbf{q}) = \frac{1}{p} \sum_{k=1}^p X_k^1, \quad X_k^1 = x_0(k) \mathcal{S}_\lambda [x_0(k) q_1 + Z_k]$$

where $Z_k = \mathbf{q}_2^\top \mathbf{g}^k \sim \mathcal{N}\left(0, \frac{\|\mathbf{q}_2\|_2^2}{p}\right)$. Thus, for any $m \geq 2$, by Lemma A.4, we have

$$\begin{aligned} & \mathbb{E} [|X_k^1|^m] \\ & \leq \theta \left(\frac{1}{\sqrt{\theta p}} \right)^m \mathbb{E} \left[\left| \frac{q_1}{\sqrt{\theta p}} + Z_k \right|^m \right] \\ & = \theta \left(\frac{1}{\sqrt{\theta p}} \right)^m \sum_{l=0}^m \binom{m}{l} \left(\frac{q_1}{\sqrt{\theta p}} \right)^l \mathbb{E} [|Z_k|^{m-l}] \\ & = \theta \left(\frac{1}{\sqrt{\theta p}} \right)^m \sum_{l=0}^m \binom{m}{l} \left(\frac{q_1}{\sqrt{\theta p}} \right)^l (m-l-1)!! \left(\frac{\|\mathbf{q}_2\|_2}{\sqrt{p}} \right)^{m-l} \\ & \leq \frac{m!}{2} \theta \left(\frac{1}{\sqrt{\theta p}} \right)^m \left(\frac{q_1}{\sqrt{\theta p}} + \frac{\|\mathbf{q}_2\|_2}{\sqrt{p}} \right)^m \\ & \leq \frac{m!}{2} \theta \left(\frac{2}{\theta p} \right)^m = \frac{m!}{2} \frac{4}{\theta p^2} \left(\frac{2}{\theta p} \right)^{m-2} \end{aligned}$$

let $\sigma_X^2 = 4/(\theta p^2)$ and $R = 2/(\theta p)$, apply Lemma A.7, we get

$$\mathbb{P} [|\bar{Q}_1(\mathbf{q}) - \mathbb{E} [\bar{Q}_1(\mathbf{q})]| \geq t] \leq 2 \exp \left(-\frac{\theta p^3 t^2}{8 + 4pt} \right).$$

as desired. \square

2) Concentration for $\bar{Q}_2(\mathbf{q})$:

Lemma E.6 (Bounding $\|\bar{Q}_2(\mathbf{q}) - \mathbb{E} [\bar{Q}_2(\mathbf{q})]\|_2$). *For each $\mathbf{q} \in \mathbb{S}^{n-1}$, it holds for all $t > 0$ that*

$$\begin{aligned} & \mathbb{P} [\|\bar{Q}_2(\mathbf{q}) - \mathbb{E} [\bar{Q}_2(\mathbf{q})]\|_2 > t] \\ & \leq 2(n+1) \exp \left(-\frac{\theta p^3 t^2}{128n + 16\sqrt{\theta n p t}} \right). \end{aligned}$$

Before proving Lemma E.6, we record the following useful results.

Lemma E.7. *For any positive integer $s, l > 0$, we have*

$$\mathbb{E} \left[\|\mathbf{g}^k\|_2^s \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \right] \leq \frac{(l+s)!!}{2} \|\mathbf{q}_2\|_2^l \frac{(2\sqrt{n})^s}{(\sqrt{p})^{s+l}}.$$

In particular, when $s = l$, we have

$$\mathbb{E} \left[\|\mathbf{g}^k\|_2^l \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \right] \leq \frac{l!}{2} \|\mathbf{q}_2\|_2^l \left(\frac{4\sqrt{n}}{p} \right)^l$$

Proof. Let $\mathcal{P}_{\mathbf{q}_2} = \frac{\mathbf{q}_2 \mathbf{q}_2^\top}{\|\mathbf{q}_2\|_2^2}$ and $\mathcal{P}_{\mathbf{q}_2^\perp} = \left(\mathbf{I} - \frac{1}{\|\mathbf{q}_2\|_2^2} \mathbf{q}_2 \mathbf{q}_2^\top \right)$ denote the projection operators onto \mathbf{q}_2 and its orthogonal complement, respectively. By Lemma A.4, we have

$$\mathbb{E} \left[\|\mathbf{g}^k\|_2^s \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \right]$$

$$\begin{aligned} & \leq \mathbb{E} \left[\left(\|\mathcal{P}_{\mathbf{q}_2} \mathbf{g}^k\|_2 + \|\mathcal{P}_{\mathbf{q}_2^\perp} \mathbf{g}^k\|_2 \right)^s \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \right] \\ & = \sum_{i=0}^s \binom{s}{i} \mathbb{E} \left[\|\mathcal{P}_{\mathbf{q}_2} \mathbf{g}^k\|_2^i \right] \mathbb{E} \left[\left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \|\mathcal{P}_{\mathbf{q}_2^\perp} \mathbf{g}^k\|_2^{s-i} \right] \\ & = \sum_{i=0}^s \binom{s}{i} \mathbb{E} \left[\|\mathcal{P}_{\mathbf{q}_2} \mathbf{g}^k\|_2^i \right] \mathbb{E} \left[\left| \mathbf{q}_2^\top \mathbf{g}^k \right|^{l+s-i} \right] \frac{1}{\|\mathbf{q}_2\|_2^{s-i}} \\ & \leq \|\mathbf{q}_2\|_2^l \sum_{i=0}^s \binom{s}{i} \mathbb{E} \left[\|\mathcal{P}_{\mathbf{q}_2^\perp} \mathbf{g}^k\|_2^i \right] \left(\frac{1}{\sqrt{p}} \right)^{l+s-i} (l+s-i-1)!!. \end{aligned}$$

Using Lemma A.5 and the fact that $\|\mathcal{P}_{\mathbf{q}_2} \mathbf{g}^k\|_2 \leq \|\mathbf{g}^k\|_2$, we obtain

$$\begin{aligned} & m \mathbb{E} \left[\|\mathbf{g}^k\|_2^s \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \right] \\ & \leq \|\mathbf{q}_2\|_2^l \sum_{i=0}^s \binom{s}{i} \left(\frac{\sqrt{n}}{\sqrt{p}} \right)^i i! \left(\frac{1}{\sqrt{p}} \right)^{l+s-i} (l+s-i-1)!! \\ & \leq \|\mathbf{q}_2\|_2^l \left(\frac{1}{\sqrt{p}} \right)^l \frac{(l+s)!!}{2} \left(\frac{\sqrt{n}}{\sqrt{p}} + \frac{1}{\sqrt{p}} \right)^s \\ & \leq \frac{(l+s)!!}{2} \|\mathbf{q}_2\|_2^l \frac{(2\sqrt{n})^s}{(\sqrt{p})^{s+l}}. \end{aligned}$$

\square

Now, we are ready to prove Lemma E.6,

Proof. By (E.1), note that

$$\bar{Q}_2 = \frac{1}{p} \sum_{k=1}^p \mathbf{X}_k^2, \quad \mathbf{X}_k^2 = \mathbf{g}^k \mathcal{S}_\lambda [x_0(k) q_1 + Z_k]$$

where $Z_k = \mathbf{q}_2^\top \mathbf{g}^k$. Thus, for any $m \geq 2$, by Lemma E.7, we have

$$\begin{aligned} & \mathbb{E} [\|\mathbf{X}_k^2\|_2^m] \\ & \leq \theta \mathbb{E} \left[\|\mathbf{g}^k\|_2^m \left| \frac{q_1}{\sqrt{\theta p}} + \mathbf{q}_2^\top \mathbf{g}^k \right|^m \right] \\ & \quad + (1-\theta) \mathbb{E} \left[\|\mathbf{g}^k\|_2^m \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^m \right] \\ & \leq \theta \sum_{l=0}^m \binom{m}{l} \mathbb{E} \left[\left| \mathbf{q}_2^\top \mathbf{g}^k \right|^l \|\mathbf{g}^k\|_2^{m-l} \right] \left| \frac{q_1}{\sqrt{\theta p}} \right|^{m-l} \\ & \quad + (1-\theta) \mathbb{E} \left[\|\mathbf{g}^k\|_2^m \left| \mathbf{q}_2^\top \mathbf{g}^k \right|^m \right] \\ & \leq \theta \left(\frac{2\sqrt{n}}{\sqrt{p}} \right)^m \sum_{l=0}^m \binom{m}{l} \frac{(m+l)!!}{2} \left(\frac{\|\mathbf{q}_2\|_2}{\sqrt{p}} \right)^l \left| \frac{q_1}{\sqrt{\theta p}} \right|^{m-l} \\ & \quad + (1-\theta) \frac{m!}{2} \|\mathbf{q}_2\|_2^m \left(\frac{4\sqrt{n}}{p} \right)^m \\ & \leq \theta \frac{m!}{2} \left(\frac{4\sqrt{n}}{\sqrt{p}} \right)^m \left(\frac{\|\mathbf{q}_2\|_2}{\sqrt{p}} + \frac{q_1}{\sqrt{\theta p}} \right)^m \\ & \quad + (1-\theta) \frac{m!}{2} \|\mathbf{q}_2\|_2^m \left(\frac{4\sqrt{n}}{p} \right)^m \end{aligned}$$

$$\leq \frac{m!}{2} \left(\frac{8\sqrt{n}}{\sqrt{\theta p}} \right)^m.$$

Taking $\sigma_X^2 = 64n/(\theta p^2)$ and $R = 8\sqrt{n}/(\sqrt{\theta p})$ and using vector Bernstein's inequality in Lemma A.8, we obtain

$$\begin{aligned} \mathbb{P} \left[\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E} [\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2 \geq t \right] \\ \leq 2(n+1) \exp \left(-\frac{\theta p^3 t^2}{128n + 16\sqrt{\theta n p t}} \right), \end{aligned}$$

as desired. \square

C. Union Bound

Proposition E.8 (Uniformizing the Bounds). *Suppose that $\theta > 1/\sqrt{n}$. Given any $\xi > 0$, there exists some constant $C(\xi)$, such that whenever $p \geq C(\xi) n^4 \log n$, we have*

$$\begin{aligned} |\bar{Q}_1(\mathbf{q}) - \mathbb{E} [\bar{Q}_1(\mathbf{q})]| &\leq \frac{2\xi}{\theta^{5/2} n^{3/2} p}, \\ \|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E} [\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2 &\leq \frac{2\xi}{\theta^2 n p} \end{aligned}$$

hold uniformly for all $\mathbf{q} \in \mathbb{S}^{n-1}$, with probability at least $1 - c(\xi)p^{-2}$ for a positive constant $c(\xi)$.

Proof. We apply the standard covering argument. For any $\varepsilon \in (0, 1)$, by Lemma A.12, the unit hemisphere of interest can be covered by an ε -net \mathcal{N}_ε of cardinality at most $(3/\varepsilon)^n$. For any $\mathbf{q} \in \mathbb{S}^{n-1}$, it can be written as

$$\mathbf{q} = \mathbf{q}' + \mathbf{e}$$

where $\mathbf{q}' \in \mathcal{N}_\varepsilon$ and $\|\mathbf{e}\|_2 \leq \varepsilon$. Let a row of $\bar{\mathbf{Y}}$ be $\bar{\mathbf{y}}^k = [x_0(k), \mathbf{g}^k]^\top$, which is an independent copy of $\bar{\mathbf{y}} = [x_0, \mathbf{g}]^\top$. By (E.1), we have

$$\begin{aligned} &|\bar{Q}_1(\mathbf{q}) - \mathbb{E} [\bar{Q}_1(\mathbf{q})]| \\ &= \left| \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda \left[\langle \bar{\mathbf{y}}^k, \mathbf{q}' + \mathbf{e} \rangle \right] \right. \\ &\quad \left. - \frac{1}{p} \sum_{k=1}^p \mathbb{E} \left[x_0(k) \mathcal{S}_\lambda \left[\langle \bar{\mathbf{y}}^k, \mathbf{q}' + \mathbf{e} \rangle \right] \right] \right| \\ &\leq \left| \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda \left[\langle \bar{\mathbf{y}}^k, \mathbf{q}' + \mathbf{e} \rangle \right] \right. \\ &\quad \left. - \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda \left[\langle \bar{\mathbf{y}}^k, \mathbf{q}' \rangle \right] \right| \\ &\quad + \left| \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda \left[\langle \bar{\mathbf{y}}^k, \mathbf{q}' \rangle \right] - \mathbb{E} [x_0 \mathcal{S}_\lambda [\langle \bar{\mathbf{y}}, \mathbf{q}' \rangle]] \right| \\ &\quad + \left| \mathbb{E} [x_0 \mathcal{S}_\lambda [\langle \bar{\mathbf{y}}, \mathbf{q}' \rangle]] - \mathbb{E} [x_0 \mathcal{S}_\lambda [\langle \bar{\mathbf{y}}, \mathbf{q}' + \mathbf{e} \rangle]] \right|. \end{aligned}$$

Using Cauchy-Schwarz inequality and the fact that $\mathcal{S}_\lambda[\cdot]$ is a nonexpansive operator, we have

$$\begin{aligned} &|\bar{Q}_1(\mathbf{q}) - \mathbb{E} [\bar{Q}_1(\mathbf{q})]| \\ &\leq |\bar{Q}_1(\mathbf{q}') - \mathbb{E} [\bar{Q}_1(\mathbf{q}')]| \\ &\quad + \left(\frac{1}{p} \sum_{k=1}^p |x_0(k)| \|\bar{\mathbf{y}}^k\|_2 + \mathbb{E} [|x_0| \|\bar{\mathbf{y}}\|_2] \right) \|\mathbf{e}\|_2 \\ &\leq |\bar{Q}_1(\mathbf{q}') - \mathbb{E} [\bar{Q}_1(\mathbf{q}')]| \\ &\quad + \varepsilon \frac{1}{\sqrt{\theta p}} \left(\frac{2}{\sqrt{\theta p}} + \max_{k \in [p]} \|\mathbf{g}^k\|_2 + \mathbb{E} [\|\mathbf{g}\|_2] \right). \end{aligned}$$

By Lemma A.10, $\max_{k \in [p]} \|\mathbf{g}^k\|_2 \leq \sqrt{n/p} + 2\sqrt{2 \log(2p)/p}$ with probability at least $1 - c_1 p^{-3}$. Also $\mathbb{E} [\|\mathbf{g}\|_2] \leq \left(\mathbb{E} [\|\mathbf{g}\|_2^2] \right)^{1/2} \leq \sqrt{n/p}$. Taking $t = \xi \theta^{-5/2} n^{-3/2} p^{-1}$ in Lemma E.5 and applying a union bound with $\varepsilon = \xi \theta^{-2} n^{-2} (\log 2p)^{-1/2} / 7$, and combining with the above estimates, we obtain that

$$\begin{aligned} |\bar{Q}_1(\mathbf{q}) - \mathbb{E} [\bar{Q}_1(\mathbf{q})]| &\leq \frac{\xi}{\theta^{5/2} n^{3/2} p} \\ &+ \frac{\xi}{7\theta^{5/2} n^2 \sqrt{\log(2p)p}} \left(4\sqrt{n} + 2\sqrt{2 \log(2p)} \right) \leq \frac{2\xi}{\theta^{5/2} n^{3/2} p} \end{aligned}$$

holds for all $\mathbf{q} \in \mathbb{S}^{n-1}$, with probability at least

$$\begin{aligned} &1 - c_1 p^{-3} - 2 \exp(-c_3(\xi) p / (\theta^4 n^3)) \\ &\quad + c_4(\xi) n \log n + c_5(\xi) n \log \log(2p). \end{aligned}$$

Similarly, by (E.1), we have

$$\begin{aligned} &\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E} [\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2 \\ &= \left\| \frac{1}{p} \sum_{k=1}^p \left\{ \mathbf{g}^k \mathcal{S}_\lambda \left[\langle \bar{\mathbf{y}}^k, \mathbf{q}' + \mathbf{e} \rangle \right] - \mathbb{E} [\mathbf{g} \mathcal{S}_\lambda [\langle \bar{\mathbf{y}}, \mathbf{q}' + \mathbf{e} \rangle]] \right\} \right\|_2 \\ &\leq \|\bar{\mathbf{Q}}_2(\mathbf{q}') - \mathbb{E} [\bar{\mathbf{Q}}_2(\mathbf{q}')] \|_2 \\ &\quad + \left(\frac{1}{p} \sum_{k=1}^p \|\mathbf{g}^k\|_2 \|\bar{\mathbf{y}}^k\|_2 + \mathbb{E} [\|\mathbf{g}\|_2 \|\bar{\mathbf{y}}\|_2] \right) \|\mathbf{e}\|_2 \\ &\leq \|\bar{\mathbf{Q}}_2(\mathbf{q}') - \mathbb{E} [\bar{\mathbf{Q}}_2(\mathbf{q}')] \|_2 \\ &\quad + \varepsilon \left[\max_{k \in [p]} \|\mathbf{g}^k\|_2 \left(\frac{1}{\sqrt{\theta p}} + \max_{k \in [p]} \|\mathbf{g}^k\|_2 \right) + \frac{\sqrt{n}}{\sqrt{\theta p}} + \frac{n}{p} \right]. \end{aligned}$$

Applying the above estimates for $\max_{k \in [p]} \|\mathbf{g}^k\|_2$, and taking $t = \xi \theta^{-2} n^{-1} p^{-1}$ in Lemma E.6 and applying a union bound with $\varepsilon = \xi \theta^{-2} n^{-2} \log^{-1}(2p) / 30$, we obtain that

$$\begin{aligned} &\|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbb{E} [\bar{\mathbf{Q}}_2(\mathbf{q})]\|_2 \\ &\leq \frac{\xi}{\theta^2 n p} + \frac{\xi}{30\theta^2 n^2 \log(2p)} \left\{ 4 \left(\sqrt{\frac{n}{p}} + \sqrt{\frac{2 \log(2p)}{p}} \right)^2 + \frac{2n}{p} \right\} \end{aligned}$$

$$\begin{aligned} &\leq \frac{\xi}{\theta^2 np} + \frac{\xi}{30\theta^2 n^2 \log(2p)} \left\{ \frac{16 \log(2p)}{p} + \frac{10n}{p} \right\} \\ &\leq \frac{2\xi}{\theta^2 np} \end{aligned}$$

holds for all $\mathbf{q} \in \mathbb{S}^{n-1}$, with probability at least

$$1 - c_1 p^{-3} - \exp(-c_6(\xi)p/(\theta^3 n^3) + c_7(\xi)n \log n + c_8(\xi)n \log \log(2p)).$$

Taking $p \geq C_9(\xi)n^4 \log n$ and simplifying the probability terms complete the proof. \square

D. $\mathbf{Q}(\mathbf{q})$ approximates $\bar{\mathbf{Q}}(\mathbf{q})$

Proposition E.9. *Suppose $\theta > 1/\sqrt{n}$. For any $\xi > 0$, there exists some constant $C(\xi)$, such that whenever $p \geq C(\xi)n^4 \log n$, the following bounds*

$$\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1(\mathbf{q}) - \bar{Q}_1(\mathbf{q})| \leq \frac{\xi}{\theta^{5/2} n^{3/2} p} \quad (\text{E.13})$$

$$\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\mathbf{Q}_2(\mathbf{q}) - \bar{\mathbf{Q}}_2(\mathbf{q})\|_2 \leq \frac{\xi}{\theta^2 np}, \quad (\text{E.14})$$

hold with probability at least $1 - c(\xi)p^{-2}$ for a positive constant $c(\xi)$.

Proof. First, for any $\mathbf{q} \in \mathbb{S}^{n-1}$, from (E.1), we know that

$$\begin{aligned} &|\bar{Q}_1(\mathbf{q}) - Q_1(\mathbf{q})| \\ &= \left| \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] - \frac{1}{p} \sum_{k=1}^p \frac{x_0(k)}{\|\mathbf{x}_0\|_2} \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right| \\ &\leq \left| \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] - \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right| \\ &+ \left| \frac{1}{p} \sum_{k=1}^p x_0(k) \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] - \frac{1}{p} \sum_{k=1}^p \frac{x_0(k)}{\|\mathbf{x}_0\|_2} \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right| \\ &\leq \frac{1}{p} \sum_{k=1}^p |x_0(k)| \left| \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] - \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right| \\ &\quad + \frac{1}{p} \sum_{k=1}^p |x_0(k)| \left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| \left| \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right|. \end{aligned}$$

For any $\mathcal{I} = \text{supp}(\mathbf{x}_0)$, using the fact that $\mathcal{S}_\lambda[\cdot]$ is a nonexpansive operator, we have

$$\begin{aligned} &\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |\bar{Q}_1(\mathbf{q}) - Q_1(\mathbf{q})| \\ &\leq \frac{1}{p} \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \sum_{k \in \mathcal{I}} |x_0(k)| \left| \mathbf{q}^\top (\bar{\mathbf{y}}^k - \mathbf{y}^k) \right| \\ &\quad + \left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| \frac{1}{p} \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \sum_{k \in \mathcal{I}} |x_0(k)| \left| \mathbf{q}^\top \mathbf{y}^k \right| \\ &= \frac{1}{\sqrt{\theta} p^{3/2}} \left(\|\bar{\mathbf{Y}}_{\mathcal{I}} - \mathbf{Y}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} + \left| 1 - \frac{1}{\|\mathbf{x}_0\|_2} \right| \|\mathbf{Y}_{\mathcal{I}}\|_{\ell^2 \rightarrow \ell^1} \right) \end{aligned}$$

By Lemma B.1 and Lemma B.3 in Appendix B, we have the following holds

$$\begin{aligned} &\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |\bar{Q}_1(\mathbf{q}) - Q_1(\mathbf{q})| \\ &\leq \frac{1}{\sqrt{\theta} p^{3/2}} \left(20 \sqrt{\frac{n \log p}{\theta}} + \frac{4\sqrt{2}}{5} \sqrt{\frac{n \log p}{\theta^2 p}} \times 7 \sqrt{2\theta p} \right) \\ &\leq \frac{32}{\theta p^{3/2}} \sqrt{n \log p}, \end{aligned}$$

with probability at least $1 - c_1 p^{-2}$, provided $p \geq C_2 n$ and $\theta > 1/\sqrt{n}$. Simple calculation shows that it is enough to have $p \geq C_3(\xi)n^4 \log n$ for some sufficiently large $C_1(\xi)$ to obtain the claimed result in (E.13). Similarly, by Lemma B.3 and Lemma B.4 in Appendix B, we have

$$\begin{aligned} &\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\bar{\mathbf{Q}}_2(\mathbf{q}) - \mathbf{Q}_2(\mathbf{q})\|_2 \\ &= \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \left\| \frac{1}{p} \sum_{k=1}^p \mathbf{g}^k \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] - \frac{1}{p} \sum_{k=1}^p \mathbf{g}^k \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right\|_2 \\ &\leq \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \left\| \frac{1}{p} \sum_{k=1}^p \mathbf{g}^k \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] - \frac{1}{p} \sum_{k=1}^p \mathbf{g}^k \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] \right\|_2 \\ &\quad + \left\| \frac{1}{p} \sum_{k=1}^p \mathbf{g}^k \mathcal{S}_\lambda[\mathbf{q}^\top \bar{\mathbf{y}}^k] - \frac{1}{p} \sum_{k=1}^p \mathbf{g}^k \mathcal{S}_\lambda[\mathbf{q}^\top \mathbf{y}^k] \right\|_2 \\ &\leq \frac{1}{p} \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \sum_{k=1}^p \left\| \mathbf{g}^k - \mathbf{g}^k \right\|_2 \left| \mathbf{q}^\top \bar{\mathbf{y}}^k \right| \\ &\quad + \frac{1}{p} \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \sum_{k=1}^p \left\| \mathbf{g}^k \right\|_2 \left| \mathbf{q}^\top (\bar{\mathbf{y}}^k - \mathbf{y}^k) \right| \\ &\leq \frac{1}{p} (\|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty} \|\bar{\mathbf{Y}}\|_{\ell^2 \rightarrow \ell^1} + \|\mathbf{G}'\|_{\ell^2 \rightarrow \ell^\infty} \|\bar{\mathbf{Y}} - \mathbf{Y}\|_{\ell^2 \rightarrow \ell^1}) \\ &\leq \frac{1}{p} \left(\frac{120 \max(n, \log(2p))}{\sqrt{p}} \right. \\ &\quad \left. + \frac{300 \sqrt{n \log(2p)} \max(\sqrt{n}, \sqrt{\log(2p)})}{\sqrt{\theta} p} \right) \\ &\leq \frac{420 \sqrt{n \log(2p)} \max(\sqrt{n}, \sqrt{\log(2p)})}{\theta^{1/2} p^{3/2}} \end{aligned}$$

with probability at least $1 - c_4 p^{-2}$ provided $p \geq C_4 n$ and $\theta > 1/\sqrt{n}$. It is sufficient to have $p \geq C_5(\xi)n^4 \log n$ to obtain the claimed result (E.14). \square

APPENDIX F

LARGE $|q_1|$ ITERATES STAYING IN SAFE REGION FOR ROUNDING

In this appendix, we prove Proposition IV.5 in Section IV.

Proof of Proposition IV.5. For notational simplicity, w.l.o.g. we will proceed to prove assuming $q_1 > 0$. The

proof for $q_1 < 0$ is similar by symmetry. It is equivalent to show that

$$\frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{|Q_1(\mathbf{q})|} < \sqrt{\frac{1}{4\theta} - 1},$$

which is implied by

$$\begin{aligned} \mathcal{L}(\mathbf{q}) &\doteq \frac{\|\mathbb{E}[\overline{\mathbf{Q}}_2(\mathbf{q})]\|_2 + \|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}[\overline{\mathbf{Q}}_2(\mathbf{q})]\|_2}{\mathbb{E}[Q_1(\mathbf{q})] - |Q_1(\mathbf{q}) - \mathbb{E}[\overline{Q}_1(\mathbf{q})]|} \\ &< \sqrt{\frac{1}{4\theta} - 1} \end{aligned}$$

for any $\mathbf{q} \in \mathbb{S}^{n-1}$ satisfying $q_1 > 3\sqrt{\theta}$. Recall from (E.7) that

$$\begin{aligned} \mathbb{E}[\overline{Q}_1(\mathbf{q})] &= \sqrt{\frac{\theta}{p}} \left\{ \left[\alpha \Psi\left(-\frac{\alpha}{\sigma}\right) + \beta \Psi\left(\frac{\beta}{\sigma}\right) \right] \right. \\ &\quad \left. + \sigma \left[\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right) \right] \right\}, \end{aligned}$$

where

$$\alpha = \frac{1}{\sqrt{p}} \left(\frac{q_1}{\sqrt{\theta}} + 1 \right), \quad \beta = \frac{1}{\sqrt{p}} \left(\frac{q_1}{\sqrt{\theta}} - 1 \right), \quad \sigma = \frac{\|\mathbf{q}_2\|_2}{\sqrt{p}}.$$

Noticing the fact that

$$\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right) \geq 0,$$

and for $q_1 > 3\sqrt{\theta}$

$$\Psi\left(\frac{\beta}{\sigma}\right) = \Psi\left(\frac{1}{\sqrt{1-q_1^2}} \left(\frac{q_1}{\sqrt{\theta}} - 1 \right)\right) \geq \Psi(2) \geq \frac{19}{20},$$

we have

$$\begin{aligned} \mathbb{E}[\overline{Q}_1(\mathbf{q})] &\geq \frac{\sqrt{\theta}}{p} \left\{ \frac{q_1}{\sqrt{\theta}} \left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) \right] \right. \\ &\quad \left. + \Psi\left(-\frac{\alpha}{\sigma}\right) - \Psi\left(\frac{\beta}{\sigma}\right) \right\} \\ &\geq \frac{2\sqrt{\theta}}{p} \Psi\left(\frac{\beta}{\sigma}\right) \geq \frac{19\sqrt{\theta}}{10p}. \end{aligned}$$

Moreover, from (E.8), we have

$$\begin{aligned} \|\mathbb{E}[\overline{\mathbf{Q}}_2(\mathbf{q})]\|_2 &= \|\mathbf{q}_2\|_2 \left\{ \frac{2(1-\theta)}{p} \Psi\left(-\frac{\lambda}{\sigma}\right) \right. \\ &\quad \left. + \frac{\theta}{p} \left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) \right] \right\} \\ &\leq \frac{2(1-\theta)}{p} \Psi(-1) + \frac{\theta}{p} [\Psi(-1) + 1] \\ &\leq \frac{2}{p} \Psi(-1) + \frac{\theta}{p} \leq \frac{2}{5p} + \frac{\theta}{p}, \end{aligned}$$

where we have used the fact that $-\lambda/\sigma \leq -1$ and $-\alpha/\sigma \leq -1$. Moreover, from results in Proposition E.8

and Proposition E.9 in Appendix E, we know that

$$\begin{aligned} &\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1(\mathbf{q}) - \mathbb{E}[\overline{Q}_1(\mathbf{q})]| \\ &\leq \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1(\mathbf{q}) - \overline{Q}_1(\mathbf{q})| + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |\overline{Q}_1(\mathbf{q}) - \mathbb{E}[\overline{Q}_1(\mathbf{q})]| \\ &\leq \frac{1}{2 \times 10^5 \theta^{5/2} n^{3/2} p}, \end{aligned}$$

and

$$\begin{aligned} &\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\mathbf{Q}(\mathbf{q}) - \mathbb{E}[\overline{\mathbf{Q}}(\mathbf{q})]\|_2 \\ &\leq \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\mathbf{Q}(\mathbf{q}) - \overline{\mathbf{Q}}(\mathbf{q})\|_2 + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\overline{\mathbf{Q}}(\mathbf{q}) - \mathbb{E}[\overline{\mathbf{Q}}(\mathbf{q})]\|_2 \\ &\leq \frac{1}{2 \times 10^5 \theta^2 n p} \end{aligned}$$

hold with probability at least $1 - c_1 p^{-2}$ provided that $p \geq \Omega(n^4 \log n)$. Hence, with high probability, we have

$$\begin{aligned} \mathcal{L}(\mathbf{q}) &\leq \frac{2/(5p) + \theta/p + (2 \times 10^5 \theta^2 n p)^{-1}}{19\sqrt{\theta}/(10p) - (2 \times 10^5 \theta^{5/2} n^{3/2} p)^{-1}} \\ &\leq \frac{3/5}{18\sqrt{\theta}/10} \leq \frac{1}{3\sqrt{\theta}} < \sqrt{\frac{1}{4\theta} - 1}, \end{aligned}$$

whenever θ is sufficiently small. This completes the proof. \square

Now, keep the notation in Appendix E for general orthonormal basis $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$. For any current iterate $\mathbf{q} \in \mathbb{S}^{n-1}$ that is close enough to the target solution, i.e., $|\langle \mathbf{q}, \mathbf{U}^\top \mathbf{e}_1 \rangle| = |\langle \mathbf{U}\mathbf{q}, \mathbf{e}_1 \rangle| \geq 3\sqrt{\theta}$, we have

$$\begin{aligned} \frac{|\langle \mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}), \mathbf{U}^\top \mathbf{e}_1 \rangle|}{\|\mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}})\|_2} &= \frac{|\langle \mathbf{U}\mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}}), \mathbf{e}_1 \rangle|}{\|\mathbf{U}\mathbf{Q}(\mathbf{q}; \hat{\mathbf{Y}})\|_2} \\ &= \frac{|\langle \mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y}), \mathbf{e}_1 \rangle|}{\|\mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y})\|_2}, \end{aligned}$$

where we have applied the identity proved in (E.2). Taking $\mathbf{U}\mathbf{q} \in \mathbb{S}^{n-1}$ as the object of interest, by Proposition IV.5, we conclude that

$$\frac{|\langle \mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y}), \mathbf{e}_1 \rangle|}{\|\mathbf{Q}(\mathbf{U}\mathbf{q}; \mathbf{Y})\|_2} \geq 2\sqrt{\theta}$$

with high probability.

APPENDIX G BOUNDING ITERATION COMPLEXITY

In this appendix, we prove Proposition IV.6 in Section IV.

Proof of Proposition IV.6. Recall from Proposition IV.4 in Section IV, the gap

$$G(\mathbf{q}) = \frac{|Q_1(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}\|_2} \geq \frac{1}{10^4\theta^2np}$$

holds uniformly over $\mathbf{q} \in \mathbb{S}^{n-1}$ satisfying $\frac{1}{10\sqrt{\theta n}} \leq |q_1| \leq 3\sqrt{\theta}$, with probability at least $1 - c_1p^{-2}$, provided $p \geq C_2n^4 \log n$. The gap $G(\mathbf{q})$ implies that

$$\begin{aligned} \left| \tilde{Q}_1(\mathbf{q}) \right| &\doteq \frac{|Q_1(\mathbf{q})|}{\|\mathbf{Q}(\mathbf{q})\|_2} \geq \frac{|q_1| \|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}\|_2 \|\mathbf{Q}(\mathbf{q})\|_2} \\ &\quad + \frac{|q_1|}{10^4\theta^2np \|\mathbf{Q}(\mathbf{q})\|_2} \\ \iff \left| \tilde{Q}_1(\mathbf{q}) \right| &\geq \frac{|q_1|}{\|\mathbf{q}_2\|_2} \sqrt{1 - \left| \tilde{Q}_1(\mathbf{q}) \right|^2} \\ &\quad + \frac{|q_1|}{10^4\theta^2np \|\mathbf{Q}(\mathbf{q})\|_2} \\ \implies \left| \tilde{Q}_1(\mathbf{q}) \right|^2 &\geq |q_1|^2 \left(1 + \frac{\|\mathbf{q}_2\|_2^2}{10^8\theta^4n^2p^2 \|\mathbf{Q}(\mathbf{q})\|_2^2} \right). \end{aligned}$$

Given the set Γ defined in (IV.7), now we know that

$$\begin{aligned} &\sup_{\mathbf{q} \in \Gamma} \|\mathbf{Q}(\mathbf{q})\|_2 \\ &\leq \sup_{\mathbf{q} \in \Gamma} \left| \mathbb{E} [\bar{Q}_1(\mathbf{q})] \right| + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \left| \mathbb{E} [\bar{Q}_1(\mathbf{q})] - \bar{Q}_1(\mathbf{q}) \right| \\ &\quad + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1(\mathbf{q}) - \bar{Q}_1(\mathbf{q})| + \sup_{\mathbf{q} \in \Gamma} \left\| \mathbb{E} [\bar{Q}_2(\mathbf{q})] \right\|_2 \\ &\quad + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \left\| \mathbb{E} [\bar{Q}_2(\mathbf{q})] - \bar{Q}_2(\mathbf{q}) \right\|_2 \\ &\quad + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \left\| \mathbf{Q}_2(\mathbf{q}) - \bar{Q}_2(\mathbf{q}) \right\|_2 \\ &\leq \sup_{\mathbf{q} \in \Gamma} \left| \mathbb{E} [\bar{Q}_1(\mathbf{q})] \right| + \sup_{\mathbf{q} \in \Gamma} \left| \mathbb{E} [\bar{Q}_2(\mathbf{q})] \right| + \frac{1}{np} \end{aligned}$$

with probability at least $1 - c_3p^{-2}$ provided $p \geq C_4n^4 \log n$ and $\theta > 1/\sqrt{n}$. Here we have used Proposition E.8 and Proposition E.9 to bound the magnitudes of the four difference terms. To bound the magnitudes of the expectations, we have

$$\begin{aligned} &\left| \mathbb{E} [\bar{Q}_1(\mathbf{q})] \right| \\ &= \left| \mathbb{E} \left[\frac{1}{p} \sum_{k=1}^p x_0(k) S_\lambda \left[x_0(k)q_1 + \mathbf{q}_2^\top \mathbf{g}^k \right] \right] \right| \\ &\leq \frac{1}{\sqrt{\theta p}} \left(\frac{1}{\sqrt{\theta p}} + \mathbb{E} [\|\mathbf{g}\|_2] \right) \leq \frac{3\sqrt{n}}{\sqrt{\theta p}} \leq \frac{3n}{p}, \\ &\left\| \mathbb{E} [\bar{Q}_2(\mathbf{q})] \right\|_2 \\ &= \left\| \mathbb{E} \left[\frac{1}{p} \sum_{k=1}^p \mathbf{g}^k S_\lambda \left[x_0(k)q_1 + \mathbf{q}_2^\top \mathbf{g}^k \right] \right] \right\|_2 \end{aligned}$$

$$\leq \frac{1}{\sqrt{\theta p}} \mathbb{E} [\|\mathbf{g}\|_2] + \mathbb{E} [\|\mathbf{g}\|_2^2] \leq \frac{3n}{p}$$

hold uniformly for all $\mathbf{q} \in \Gamma$, provided $\theta > 1/\sqrt{n}$. Thus, we obtain that

$$\sup_{\mathbf{q} \in \Gamma} \|\mathbf{Q}(\mathbf{q})\|_2 \leq \frac{3n}{p} + \frac{3n}{p} + \frac{1}{np} \leq \frac{7n}{p}$$

with probability at least $1 - c_3p^{-2}$ provided $p \geq C_4n^4 \log n$ and $\theta > 1/\sqrt{n}$. So we conclude that

$$\left| \frac{\tilde{Q}_1(\mathbf{q})}{|q_1|} \right| \geq \sqrt{1 + \frac{1 - 9\theta}{10^8 \times 7^2 \times \theta^4 n^4}}.$$

Thus, starting with any $\mathbf{q} \in \mathbb{S}^{n-1}$ such that $|q_1| \geq \frac{1}{10\sqrt{\theta n}}$, we will need at most

$$\begin{aligned} T &= \frac{2 \log \left(3\sqrt{\theta} / \frac{1}{10\sqrt{\theta n}} \right)}{\log \left(1 + \frac{1-9\theta}{10^8 \times 7^2 \times \theta^4 n^4} \right)} = \frac{2 \log (30\theta\sqrt{n})}{\log \left(1 + \frac{1-9\theta}{10^8 \times 7^2 \times \theta^4 n^4} \right)} \\ &\leq \frac{2 \log (30\theta\sqrt{n})}{(\log 2) \frac{1-9\theta}{10^8 \times 7^2 \times \theta^4 n^4}} \leq C_5 n^4 \log n \end{aligned}$$

steps to arrive at a $\bar{\mathbf{q}} \in \mathbb{S}^{n-1}$ with $|\bar{q}_1| \geq 3\sqrt{\theta}$ for the first time. Here we have assumed $\theta_0 < 1/9$ and used the fact that $\log(1+x) \geq x \log 2$ for $x \in [0, 1]$ to simplify the final result. \square

APPENDIX H

ROUNDING TO THE DESIRED SOLUTION

In this appendix, we prove Proposition IV.7 in Section IV. For convenience, we will assume the notations we used in Appendix B. Then the rounding scheme can be written as

$$\min_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1, \quad \text{s.t. } \langle \bar{\mathbf{q}}, \mathbf{q} \rangle = 1. \quad (\text{H.1})$$

We will show the rounding procedure get us to the desired solution with high probability, regardless of the particular orthonormal basis used.

Proof of Proposition IV.7. The rounding program (H.1) can be written as

$$\inf_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1, \quad \text{s.t. } \bar{q}_1 q_1 + \langle \bar{\mathbf{q}}_2, \mathbf{q}_2 \rangle = 1. \quad (\text{H.2})$$

Consider its relaxation

$$\inf_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1, \quad \text{s.t. } \bar{q}_1 q_1 + \|\bar{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 \geq 1. \quad (\text{H.3})$$

It is obvious that the feasible set of (H.3) contains that of (H.2). So if \mathbf{e}_1/\bar{q}_1 is the unique optimal solution (UOS)

of (H.3), it is also the UOS of (H.2). Let $\mathcal{I} = \text{supp}(\mathbf{x}_0)$, and consider a modified problem

$$\begin{aligned} \inf_{\mathbf{q}} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| - \|\mathbf{G}'_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}'_{\mathcal{I}^c} \mathbf{q}_2\|_1, \quad (\text{H.4}) \\ \text{s.t. } \bar{q}_1 q_1 + \|\bar{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 \geq 1. \end{aligned}$$

The objective value of (H.4) lower bounds the objective value of (H.3), and are equal when $\mathbf{q} = \mathbf{e}_1/\bar{q}_1$. So if $\mathbf{q} = \mathbf{e}_1/\bar{q}_1$ is the UOS to (H.4), it is also UOS to (H.3), and hence UOS to (H.2) by the argument above. Now

$$\begin{aligned} & - \|\mathbf{G}'_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}'_{\mathcal{I}^c} \mathbf{q}_2\|_1 \\ \geq & - \|\mathbf{G}_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}_{\mathcal{I}^c} \mathbf{q}_2\|_1 - \|(\mathbf{G} - \mathbf{G}') \mathbf{q}_2\|_1 \\ \geq & - \|\mathbf{G}_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}_{\mathcal{I}^c} \mathbf{q}_2\|_1 - \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} \|\mathbf{q}_2\|_2. \end{aligned}$$

When $p \geq C_1 n$, by Lemma A.14 and Lemma B.3, we know that

$$\begin{aligned} & - \|\mathbf{G}_{\mathcal{I}} \mathbf{q}_2\|_1 + \|\mathbf{G}_{\mathcal{I}^c} \mathbf{q}_2\|_1 - \|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \rightarrow \ell^1} \|\mathbf{q}_2\|_2 \\ \geq & - \frac{6}{5} \sqrt{\frac{2}{\pi}} 2\theta \sqrt{p} \|\mathbf{q}_2\|_2 + \frac{24}{25} \sqrt{\frac{2}{\pi}} (1 - 2\theta) \sqrt{p} \|\mathbf{q}_2\|_2 \\ & - 4\sqrt{n} \|\mathbf{q}_2\|_2 - 7\sqrt{\log(2p)} \|\mathbf{q}_2\|_2 \doteq \zeta \|\mathbf{q}_2\|_2 \end{aligned}$$

holds with probability at least $1 - c_2 p^{-2}$. Thus, we make a further relaxation of problem (H.2) by

$$\begin{aligned} \inf_{\mathbf{q}} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| + \zeta \|\mathbf{q}_2\|_2, \quad (\text{H.5}) \\ \text{s.t. } \bar{q}_1 q_1 + \|\bar{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 \geq 1, \end{aligned}$$

whose objective value lower bounds that of (H.4). By similar arguments, if \mathbf{e}_1/\bar{q}_1 is UOS to (H.5), it is UOS to (H.2). At the optimal solution to (H.5), notice that it is necessary to have $\text{sign}(q_1) = \text{sign}(\bar{q}_1)$ and $\bar{q}_1 q_1 + \|\bar{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 = 1$. So (H.5) is equivalent to

$$\begin{aligned} \inf_{\mathbf{q}} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| + \zeta \|\mathbf{q}_2\|_2, \quad (\text{H.6}) \\ \text{s.t. } \bar{q}_1 q_1 + \|\bar{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 = 1. \end{aligned}$$

which is further equivalent to

$$\inf_{q_1} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| + \zeta \frac{1 - \bar{q}_1 |q_1|}{\|\bar{\mathbf{q}}_2\|_2}, \quad \text{s.t. } |q_1| \leq \frac{1}{\bar{q}_1}. \quad (\text{H.7})$$

Notice that the problem in (H.7) is linear in $|q_1|$ with a compact feasible set. Since the objective is also monotonic in $|q_1|$, it indicates that the optimal solution only occurs at the boundary points $|q_1| = 0$ or $|q_1| = 1/\bar{q}_1$. Therefore, $\mathbf{q} = \mathbf{e}_1/\bar{q}_1$ is the UOS of (H.7) if and only if

$$\frac{1}{\bar{q}_1} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 < \frac{\zeta}{\|\bar{\mathbf{q}}_2\|_2}.$$

Since $\left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 \leq \sqrt{2\theta p}$ conditioned on \mathcal{E}_0 , it is sufficient to have

$$\frac{\sqrt{2\theta p}}{2\sqrt{\theta}} \leq \zeta = \frac{24}{25} \sqrt{\frac{2}{\pi}} \sqrt{p} \left(1 - \frac{9}{2}\theta - \frac{25}{6} \sqrt{\frac{\pi}{2}} \sqrt{\frac{n}{p}} - \frac{175}{24} \sqrt{\frac{\pi}{2}} \sqrt{\frac{\log(2p)}{p}} \right).$$

Therefore there exists a constant $\theta_0 > 0$, such that whenever $\theta \leq \theta_0$ and $p \geq C_3(\theta_0)n$, the rounding returns \mathbf{e}_1/\bar{q}_1 . A bit of thought suggests one can take a universal C_3 for all possible choice of θ_0 , completing the proof. \square

When the input basis is $\hat{\mathbf{Y}} = \mathbf{Y}\mathbf{U}$ for some orthogonal matrix $\mathbf{U} \neq \mathbf{I}$, if the ADM algorithm produces some $\bar{\mathbf{q}} = \mathbf{U}^\top \mathbf{q}'$, such that $q'_1 > 2\sqrt{\theta}$. It is not hard to see that now the rounding (H.1) is equivalent to

$$\min_{\mathbf{q}} \|\mathbf{Y}\mathbf{U}\mathbf{q}\|_1, \quad \text{s.t. } \langle \mathbf{q}', \mathbf{U}\mathbf{q} \rangle = 1.$$

Renaming $\mathbf{U}\mathbf{q}$, it follows from the above argument that at optimum \mathbf{q}_* it holds that $\mathbf{U}\mathbf{q}_* = \gamma \mathbf{e}_1$ for some constant γ with high probability.

REFERENCES

- [1] Q. Qu, J. Sun, and J. Wright, "Finding a sparse vector in a subspace: Linear sparsity using alternating directions," in *Advances in Neural Information Processing Systems*, 2014.
- [2] E. J. Candès and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [3] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ^1 -norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [4] S. T. McCormick, "A combinatorial approach to some sparse matrix problems," tech. rep., DTIC Document, 1983.
- [5] T. F. Coleman and A. Pothén, "The null space problem i. complexity," *SIAM Journal on Algebraic Discrete Methods*, vol. 7, no. 4, pp. 527–537, 1986.
- [6] M. Berry, M. Heath, I. Kaneko, M. Lawo, R. Plemmons, and R. Ward, "An algorithm to compute a sparse basis of the null space," *Numerische Mathematik*, vol. 47, no. 4, pp. 483–504, 1985.
- [7] J. R. Gilbert and M. T. Heath, "Computing a sparse basis for the null space," *SIAM Journal on Algebraic Discrete Methods*, vol. 8, no. 3, pp. 446–459, 1987.
- [8] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*. New York, NY, USA: Oxford University Press, Inc., 1986.

- [9] A. J. Smola and B. Schlkopf, "Sparse greedy matrix approximation for machine learning," pp. 911–918, Morgan Kaufmann, 2000.
- [10] T. Kavitha, K. Mehlhorn, D. Michail, and K. Paluch, "A faster algorithm for minimum cycle basis of graphs," in *31st International Colloquium on Automata, Languages and Programming*, pp. 846–857, Springer, 2004.
- [11] L.-A. Gottlieb and T. Neylon, "Matrix sparsification and the sparse null space problem," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 205–218, Springer, 2010.
- [12] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *arXiv preprint arXiv:1411.3230*, 2014.
- [13] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.
- [14] P. Hand and L. Demanet, "Recovering the sparsest element in a subspace," *arXiv preprint arXiv:1310.1654*, 2013.
- [15] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," *arXiv preprint arXiv:1504.06785*, 2015.
- [16] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.
- [17] I. M. Johnstone and A. Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, 2009.
- [18] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," *SIAM review*, vol. 49, no. 3, pp. 434–448, 2007.
- [19] R. Krauthgamer, B. Nadler, D. Vilenchik, *et al.*, "Do semidefinite relaxations solve sparse PCA up to the information limit?," *The Annals of Statistics*, vol. 43, no. 3, pp. 1300–1322, 2015.
- [20] T. Ma and A. Wigderson, "Sum-of-squares lower bounds for sparse pca," *arXiv preprint arXiv:1507.06370*, 2015.
- [21] V. Q. Vu, J. Cho, J. Lei, and K. Rohe, "Fantope projection and selection: A near-optimal convex relaxation of sparse pca," in *Advances in Neural Information Processing Systems*, pp. 2670–2678, 2013.
- [22] J. Lei, V. Q. Vu, *et al.*, "Sparsistency and agnostic inference in sparse pca," *The Annals of Statistics*, vol. 43, no. 1, pp. 299–322, 2015.
- [23] Z. Wang, H. Lu, and H. Liu, "Nonconvex statistical optimization: Minimax-optimal sparse pca in polynomial time," *arXiv preprint arXiv:1408.5352*, 2014.
- [24] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet, "A direct formulation of sparse PCA using semidefinite programming," *SIAM Review*, vol. 49, no. 3, 2007.
- [25] Y.-B. Zhao and M. Fukushima, "Rank-one solutions for homogeneous linear matrix equations over the positive semidefinite cone," *Applied Mathematics and Computation*, vol. 219, no. 10, pp. 5569–5583, 2013.
- [26] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2018–2025, IEEE, 2012.
- [27] G. Beylkin and L. Monzón, "On approximation of functions by exponential sums," *Applied and Computational Harmonic Analysis*, vol. 19, no. 1, pp. 17–48, 2005.
- [28] C. T. Manolis and V. Rene, "Dual principal component pursuit," *arXiv preprint arXiv:1510.04390*, 2015.
- [29] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.
- [30] A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade, "When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity," in *Advances in Neural Information Processing Systems*, pp. 1986–1994, 2013.
- [31] J. Ho, Y. Xie, and B. Vemuri, "On a nonlinear generalization of sparse coding and dictionary learning," in *Proceedings of The 30th International Conference on Machine Learning*, pp. 1480–1488, 2013.
- [32] Y. Nakatsukasa, T. Soma, and A. Uschmajew, "Finding a low-rank basis in a matrix subspace," *CoRR*, vol. abs/1503.08601, 2015.
- [33] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Conference on Learning Theory*, pp. 1046–1066, 2013.
- [34] B. Barak, J. Kelner, and D. Steurer, "Rounding sum-of-squares relaxations," *arXiv preprint arXiv:1312.6652*, 2013.
- [35] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer, "Speeding up sum-of-squares for tensor decompo-

- sition and planted sparse vectors,” *arXiv preprint arXiv:1512.02337*, 2015.
- [36] S. Arora, R. Ge, and A. Moitra, “New algorithms for learning incoherent and overcomplete dictionaries,” *arXiv preprint arXiv:1308.6273*, 2013.
- [37] A. Agarwal, A. Anandkumar, and P. Netrapalli, “Exact recovery of sparsely used overcomplete dictionaries,” *arXiv preprint arXiv:1309.1952*, 2013.
- [38] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, “Learning sparsely used overcomplete dictionaries via alternating minimization,” *arXiv preprint arXiv:1310.7991*, 2013.
- [39] S. Arora, A. Bhaskara, R. Ge, and T. Ma, “More algorithms for provable dictionary learning,” *arXiv preprint arXiv:1401.0579*, 2014.
- [40] S. Arora, R. Ge, T. Ma, and A. Moitra, “Simple, efficient, and neural algorithms for sparse coding,” *arXiv preprint arXiv:1503.00778*, 2015.
- [41] K. G. Murty and S. N. Kabadi, “Some NP-complete problems in quadratic and nonlinear programming,” *Mathematical programming*, vol. 39, no. 2, pp. 117–129, 1987.
- [42] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*, 2010.
- [43] R. Basri and D. W. Jacobs, “Lambertian reflectance and linear subspaces,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 2, pp. 218–233, 2003.
- [44] E. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?,” *Journal of the ACM*, vol. 58, May 2011.
- [45] V. De la Pena and E. Giné, *Decoupling: from dependence to independence*. Springer, 1999.
- [46] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, vol. 60. Springer Science & Business Media, 2014.
- [47] K. Luh and V. Vu, “Dictionary learning with few samples and matrix concentration,” *arXiv preprint arXiv:1503.08854*, 2015.
- [48] P. Jain, P. Netrapalli, and S. Sanghavi, “Low-rank matrix completion using alternating minimization,” in *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pp. 665–674, ACM, 2013.
- [49] M. Hardt, “On the provable convergence of alternating minimization for matrix completion,” *arXiv preprint arXiv:1312.0925*, 2013.
- [50] M. Hardt and M. Wootters, “Fast matrix completion without the condition number,” in *Proceedings of The 27th Conference on Learning Theory*, pp. 638–678, 2014.
- [51] M. Hardt, “Understanding alternating minimization for matrix completion,” in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pp. 651–660, IEEE, 2014.
- [52] P. Jain and P. Netrapalli, “Fast exact matrix completion with finite samples,” *arXiv preprint arXiv:1411.1087*, 2014.
- [53] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, “Non-convex robust pca,” in *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.
- [54] Q. Zheng and J. Lafferty, “A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements,” *arXiv preprint arXiv:1506.06081*, 2015.
- [55] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht, “Low-rank solutions of linear matrix equations via procrustes flow,” *arXiv preprint arXiv:1507.03566*, 2015.
- [56] Y. Chen and M. J. Wainwright, “Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees,” *arXiv preprint arXiv:1509.03025*, 2015.
- [57] P. Jain and S. Oh, “Provable tensor factorization with missing data,” in *Advances in Neural Information Processing Systems*, pp. 1431–1439, 2014.
- [58] A. Anandkumar, R. Ge, and M. Janzamin, “Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates,” *arXiv preprint arXiv:1402.5180*, 2014.
- [59] A. Anandkumar, R. Ge, and M. Janzamin, “Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models,” *arXiv preprint arXiv:1411.1488*, 2014.
- [60] A. Anandkumar, P. Jain, Y. Shi, and U. Niranjan, “Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations,” *arXiv preprint arXiv:1510.04747*, 2015.
- [61] R. Ge, F. Huang, C. Jin, and Y. Yuan, “Escaping from saddle points—online stochastic gradient for tensor decomposition,” in *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.
- [62] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Advances in Neural Information Processing Systems*, pp. 2796–2804, 2013.
- [63] E. J. Candès, X. Li, and M. Soltanolkotabi, “Phase retrieval via wirtinger flow: Theory and algorithms,” *arXiv preprint arXiv:1407.1065*, 2014.
- [64] Y. Chen and E. J. Candès, “Solving random quadratic systems of equations is nearly as

- easy as solving linear systems,” *arXiv preprint arXiv:1505.05114*, 2015.
- [65] J. Sun, Q. Qu, and J. Wright, “A geometric analysis of phase retrieval,” *arXiv preprint arXiv:1602.06664*, 2016.
- [66] J. Sun, Q. Qu, and J. Wright, “When are nonconvex problems not scary?,” *arXiv preprint arXiv:1510.06096*, 2015.
- [67] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.
- [68] T. Figiel, J. Lindenstrauss, and V. D. Milman, “The dimension of almost spherical sections of convex bodies,” *Acta Mathematica*, vol. 139, no. 1, pp. 53–94, 1977.
- [69] A. Y. Garnaev and E. D. Gluskin, “The widths of a euclidean ball,” in *Dokl. Akad. Nauk SSSR*, vol. 277, pp. 1048–1052, 1984.
- [70] E. Gluskin and V. Milman, “Note on the geometric-arithmetic mean inequality,” in *Geometric aspects of Functional analysis*, pp. 131–135, Springer, 2003.
- [71] G. Pisier, *The volume of convex bodies and Banach space geometry*, vol. 94. Cambridge University Press, 1999.