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

Qing Qu, Ju Sun, and John Wright
{qq2105, js4038, jw2966}@columbia.edu

December 15, 2014

Abstract

We consider the problem of recovering the sparsest vector in a 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 applications in sparse dictionary learning, sparse PCA, and other problems in signal processing and machine learning. Simple convex heuristics for this problem provably break down when the fraction of nonzero entries in the target sparse vector substantially exceeds $1/\sqrt{n}$. 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 our knowledge, this is the first practical algorithm to achieve this linear scaling. This result assumes a planted sparse model, in which the target sparse vector is embedded in an otherwise random subspace. Empirically, our proposed algorithm also succeeds in more challenging data models arising, e.g., from sparse dictionary learning.

1 Introduction

Suppose we are given a linear subspace $\mathcal{S}$ of a high-dimensional space $\mathbb{R}^p$, which contains a sparse vector $\mathbf{x}_0 \neq \mathbf{0}$. Given arbitrary basis of $\mathcal{S}$, can we efficiently recover $\mathbf{x}_0$? Equivalently, provided a matrix $\mathbf{A} \in \mathbb{R}^{(p-n)\times p}$, can we efficiently find a nonzero sparse vector $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{0}$? In the language of sparse approximation, can we solve

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{0},\ \mathbf{x} \neq \mathbf{0} \qquad ? \tag{1.1}$$

Variants of this problem have been studied in the context of applications to numerical linear algebra [CP86], system control and optimizations [ZF13], nonrigid structure from motion [DLH12], spectral estimation and Prony's problem [BM05], sparse PCA [ZHT06], blind source separation [ZP01], dictionary learning [SWW12], graphical model learning [AHJK13], and sparse coding on manifolds [HXV13].

However, in contrast to the standard sparse regression problem ($\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{b} \neq \mathbf{0}$), for which convex relaxations perform nearly optimally for broad classes of designs $\mathbf{A}$ [CT05, Don06], the computational properties of problem (1.1) are not nearly as well understood. It has been known for several decades that the basic formulation

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S} \setminus \{\mathbf{0}\}, \tag{1.2}$$

is NP-hard [CP86]. However, it is only recently that efficient computational surrogates with nontrivial recovery guarantees have been discovered for some practical cases of interest. In the context of sparse dictionary learning, Spielman et al. [SWW12] introduced a relaxation which replaces the nonconvex problem (1.2) with a sequence of linear programs:

$$\ell^1/\ell^\infty \text{ Relaxation:} \qquad \min_{\mathbf{x}} \|\mathbf{x}\|_1, \quad \text{s.t.} \quad x_i = 1,\ \mathbf{x} \in \mathcal{S},\ 1 \leq i \leq p, \tag{1.3}$$

Table 1: Comparison of existing methods for recovering a planted sparse vector in a subspace

Method	Recovery Condition	Total Complexity
ℓ^1/ℓ^∞ Relaxation[HD13]	$\theta \in O(1/\sqrt{n})$	$O(np^3)$
SDP Relaxation	$\theta \in O(1/\sqrt{n})$	$O(p^{3.5})$
SOS Relaxation [BKS13]	$p \geq \Omega(n^2), \theta \in O(1)$	high order $poly(p)$
This work	$p \geq \Omega(n^4 \log n),\ \theta \in O(1)$	$O(n^5 p^2 \log n)$

and proved that when $\mathcal{S}$ is generated as a span of n random sparse vectors, with high probability the relaxation recovers these vectors, provided the probability of an entry being nonzero is at most $\theta \in O\left(1/\sqrt{n}\right)$. In a *planted sparse model*, in which $\mathcal{S}$ consists of a single sparse vector $\mathbf{x}_0$ embedded in a "generic" subspace, Hand and Demanent proved that (1.3) also correctly recovers $\mathbf{x}_0$, provided the fraction of nonzeros in $\mathbf{x}_0$ scales as $\theta \in O\left(1/\sqrt{n}\right)$ [HD13]. Unfortunately, the results of [SWW12, HD13] are essentially sharp: *when θ substantially exceeds $1/\sqrt{n}$, in both models the relaxation* (1.3) *provably breaks down.* Moreover, the most natural semidefinite programming (SDP) relaxation of (1.1),

$$\min_{\mathbf{X}} \|\mathbf{X}\|_1, \quad \text{s.t.} \quad \left\langle \mathbf{A}^\top \mathbf{A}, \mathbf{X} \right\rangle = 0,\ \text{trace}[\mathbf{X}] = 1,\ \mathbf{X} \succeq \mathbf{0}. \tag{1.4}$$

also breaks down at exactly the same threshold of $\theta \sim 1/\sqrt{n}$.[1]

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 (1.2)-(1.4) and *sparse PCA* [ZHT06]. In sparse PCA, there is a substantial gap between what can be achieved with efficient algorithms and the information theoretic optimum [BR13]. Is this also the case for recovering a sparse vector in a subspace? *Is $\theta \in O\left(1/\sqrt{n}\right)$ 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 $p \geq \Omega\left(n^2\right)$ and $\theta = \Omega(1)$ [BKS13]. 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, and so for machine learning problems in which p is either a feature dimension or sample size, this algorithm is 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 $\mathcal{S}$?*

In this paper, we address this problem, under the following hypotheses: we assume the planted sparse model, in which a target sparse vector $\mathbf{x}_0$ is embedded in an otherwise random n-dimensional subspace of $\mathbb{R}^p$. We allow $\mathbf{x}_0$ to have up to $\theta_0 p$ nonzero entries, where $\theta_0 \in (0, 1)$ is a constant. We provide a relatively simple algorithm which, with very high probability, exactly recovers $\mathbf{x}_0$, provided that $p \geq \Omega\left(n^4 \log n\right)$, where the comparison with existing results is shown in Table 1.

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 [SWW12]. 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.

In addition to enjoying theoretical guarantees in a regime ($\theta = \Omega(1)$) that is out of the reach of previous practical algorithms, it performs well in simulations – succeeding with $p \geq \Omega\left(n \log n\right)$. It also performs well empirically on more challenging data models, such as the dictionary learning model, in which the subspace of interest contains not one, but n target sparse vectors. Breaking the $O(1/\sqrt{n})$ sparsity barrier with a practical algorithm is an important open problem in the nascent literature on algorithmic guarantees for dictionary learning [AGM13, ABGM14, AAN13, AAJ+13]. We are optimistic that the techniques introduced here will be applicable in this direction.[2]

[1] This 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 [CT05, Don06]

[2] In work currently in preparation [SQW14], we show that in the dictionary learning problem, efficient algorithms based on nonconvex

2 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 (1.2). Both the objective and constraint are nonconvex, and hence not easy to optimize over. We relax (1.2) by replacing the ℓ^0 norm with the ℓ^1 norm. For the constraint $\mathbf{x} \neq \mathbf{0}$, which is necessary to avoid a trivial solution, we force $\mathbf{x}$ to live on the unit sphere $\|\mathbf{x}\|_2 = 1$, giving

$$\min_{\mathbf{x}} \ \|\mathbf{x}\|_1 , \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{S},\ \|\mathbf{x}\|_2 = 1. \tag{2.1}$$

This formulation is still nonconvex, and so we should not expect to obtain an efficient algorithm that can solve it globally for general inputs $\mathcal{S}$. Nevertheless, the geometry of the sphere is benign enough that for well-structured inputs it actually *will* be possible to give algorithms that find the global optimum.

The formulation (2.1) can be contrasted with (1.3), in which effectively we optimize the ℓ^1 norm subject to the constraint $\|\mathbf{x}\|_\infty = 1$. Because $\|\cdot\|_\infty$ is polyhedral, that formulation immediately yields a sequence of linear programs. This is very convenient for computation and analysis, but suffers from the aforementioned breakdown behavior around $\|\mathbf{x}_0\|_0 \sim p/\sqrt{n}$. In contrast, the sphere $\|\mathbf{x}\|_2 = 1$ is a more complicated geometric constraint, but will allow much larger numbers of nonzeros in $\mathbf{x}_0$. Indeed, if we consider the global optimizer of a reformulation of (2.1):

$$\min_{\mathbf{q}\in\mathbb{R}^n} \ \|\mathbf{Y}\mathbf{q}\|_1 , \quad \text{s.t.} \quad \|\mathbf{q}\|_2 = 1, \tag{2.2}$$

where $\mathbf{Y}$ is any orthonormal basis for $\mathcal{S}$, we have strong recovery guarantees as follows:

Theorem 2.1 (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*

$$\mathcal{S} = \operatorname{span}(\mathbf{x}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{n-1}) \ \subset \ \mathbb{R}^p, \tag{2.3}$$

where $\mathbf{g}_i \sim_{\text{i.i.d.}} \mathcal{N}(\mathbf{0}, \frac{1}{p}\mathbf{I})$, and $\mathbf{x}_0 \sim_{\text{i.i.d.}} \frac{1}{\sqrt{\theta p}}\mathrm{Ber}(\theta)$ are all mutually independent and $1/\sqrt{n} < \theta < \theta_0$, then optimum $\mathbf{q}^\star$ to (2.2), for any orthonormal basis $\mathbf{Y}$ of $\mathcal{S}$, produces $\mathbf{Y}\mathbf{q}^\star = \xi\mathbf{x}_0$ for some $\xi \neq 0$ with high probability, provided $p \geq \Omega(n \log n)$. [3]

Hence, *if* we could find the global optimizer of (2.2), 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: (2.2) is nonconvex. In the next section, we will describe a simple heuristic algorithm for (a near approximation of) the ℓ^1/ℓ^2 problem (2.2), which guarantees to find a stationary point. 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 optimum – the target sparse vector $\mathbf{x}_0$. The proof requires a detailed probabilistic analysis, which is sketched in Section 4.2.

Before continuing, it is worth noting that the formulation (2.1) is in no way novel – see, e.g., the work of [ZP01] in blind source separation for precedent. However, our algorithms and subsequent analysis are novel.

3 Algorithm based on Alternating Direction Method (ADM)

To develop an algorithm for solving (2.2), it is useful to consider a slight relaxation of (2.2), in which we introduce an auxiliary variable $\mathbf{x} \approx \mathbf{Y}\mathbf{q}$:

$$\min_{\mathbf{q},\mathbf{x}} \frac{1}{2}\|\mathbf{Y}\mathbf{q} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 , \quad \text{s.t.} \quad \|\mathbf{q}\|_2 = 1, \tag{3.1}$$

optimization also produce global solutions, even when $\theta = \Omega(1)$.

[3] Note that this version is much stronger and more practical than that appearing in the conference version [QSW14].

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 $\mathbf{Yq}$. 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 $\mathbf{x}$ and optimizing with respect to $\mathbf{q}$:

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2} \left\| \mathbf{Y}\mathbf{q}^{(k)} - \mathbf{x} \right\|_2^2 + \lambda \left\| \mathbf{x} \right\|_1, \tag{3.2}$$

$$\mathbf{q}^{(k+1)} = \arg\min_{\mathbf{q}} \frac{1}{2} \left\| \mathbf{Y}\mathbf{q} - \mathbf{x}^{(k+1)} \right\|_2^2 \text{ s.t. } \left\| \mathbf{q} \right\|_2 = 1. \tag{3.3}$$

Both (3.2) and (3.3) have simple closed form solutions:

$$\mathbf{x}^{(k+1)} = S_\lambda[\mathbf{Y}\mathbf{q}^{(k)}], \qquad \mathbf{q}^{(k+1)} = \frac{\mathbf{Y}^\top \mathbf{x}^{(k+1)}}{\left\| \mathbf{Y}^\top \mathbf{x}^{(k+1)} \right\|_2}, \tag{3.4}$$

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

Algorithm 1 Nonconvex ADM for solving (3.1)

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** $k = 0, \ldots, O\left(n^4 \log n\right)$ **do**
2: $\quad \mathbf{x}^{(k+1)} = S_\lambda[\mathbf{Y}\mathbf{q}^{(k)}]$,
3: $\quad \mathbf{q}^{(k+1)} = \frac{\mathbf{Y}^\top \mathbf{x}^{(k+1)}}{\left\| \mathbf{Y}^\top \mathbf{x}^{(k+1)} \right\|_2}$,
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 (2.2) is nonconvex, an arbitrary or random initialization is unlikely to produce a global minimizer.[4] Therefore, good initializations are critical for the proposed ADM algorithm to succeed. For this purpose, we suggest to use every normalized row of $\mathbf{Y}$ as initializations for $\mathbf{q}$, and solve a sequence of p nonconvex programs (2.2) by the ADM algorithm.

To get an intuition of why our initialization works, recall the planted sparse model: $\mathcal{S} = \mathrm{span}(\mathbf{x}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{n-1})$. Write $\mathbf{Z} = [\mathbf{x}_0 \mid \mathbf{g}_1 \mid \cdots \mid \mathbf{g}_{n-1}] \in \mathbb{R}^{p \times n}$. Suppose we take a row $\mathbf{z}^i$ of $\mathbf{Z}$, in which $\mathbf{x}_0(i)$ is nonzero, then $\mathbf{x}_0(i) = \Theta\left(1/\sqrt{\theta p}\right)$. Meanwhile, the entries of $\mathbf{g}_1(i), \ldots \mathbf{g}_{n-1}(i)$ are all $\mathcal{N}(0, 1/p)$, and so have size about $1/\sqrt{p}$. Hence, when θ is not too large, $\mathbf{x}_0(i)$ will be somewhat bigger than most of the other entries in $\mathbf{z}_i$. Put another way, $\mathbf{z}_i$ *is biased towards the first standard basis vector* $\mathbf{e}_1$.

Now, under our probabilistic assumptions, $\mathbf{Z}$ is very well conditioned: $\mathbf{Z}^\top \mathbf{Z} \approx \mathbf{I}$.[5] Using, e.g., Gram-Schmidt, we can find a basis $\overline{\mathbf{Y}}$ for $\mathcal{S}$ of the form

$$\overline{\mathbf{Y}} = \mathbf{ZR}, \tag{3.5}$$

where $\mathbf{R}$ is upper triangular, and $\mathbf{R}$ is itself well-conditioned: $\mathbf{R} \approx \mathbf{I}$. Since the i-th row of $\mathbf{Z}$ is biased in the direction of $\mathbf{e}_1$ and $\mathbf{R}$ is well-conditioned, the i-th row $\overline{\mathbf{y}}^i$ is also biased in the direction of $\mathbf{e}_1$. Moreover, we know that the global optimizer $\mathbf{q}_\star$ should satisfy $\overline{\mathbf{Y}}\mathbf{q}_\star = \mathbf{x}_0$. Since $\mathbf{Z}\mathbf{e}_1 = \mathbf{x}_0$, we have $\mathbf{q}_\star = \mathbf{R}^{-1}\mathbf{e}_1 \approx \mathbf{e}_1$. Here, the approximation comes from $\mathbf{R} \approx \mathbf{I}$. Hence, for this particular choice of $\mathbf{Y}$, described in (3.5), *the i-th row is biased in the direction of the global optimizer*.

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

[5] This is the common heuristic that "tall random matrices are well conditioned" [Ver10].

What if we are handed some other basis $\mathbf{Y} = \overline{\mathbf{Y}}\mathbf{U}$, where $\mathbf{U}$ is an orthogonal matrix? Suppose $\mathbf{q}_\star$ is a global optimizer to (2.2) with input matrix $\overline{\mathbf{Y}}$, then it is easy to check that, with input matrix $\mathbf{Y}$, $\mathbf{U}^\top \mathbf{q}_\star$ is also a global optimizer to (2.2), which implies that our initialization is *invariant* to any rotation of the basis. Hence, *even if we are handed an arbitrary basis for $\mathcal{S}$, the i-th row is still biased in the direction of the global optimizer*.

Rounding. Let $\overline{\mathbf{q}}$ denote the output of Algorithm 1. We will prove that with our particular initialization and an appropriate choice of λ, the solution of our ADM algorithm falls within a certain radius of the globally optimal solution $\mathbf{q}_\star$ to (2.2). To recover $\mathbf{q}_\star$, or equivalently to recover the sparse vector $\mathbf{x}_0 = \xi \mathbf{Y}\mathbf{q}_\star$ for some $\xi \neq 0$, we solve the linear program

$$\min_{\mathbf{q}} \|\mathbf{Y}\mathbf{q}\|_1 \quad \text{s.t.} \quad \langle \mathbf{r}, \mathbf{q} \rangle = 1 \tag{3.6}$$

with $\mathbf{r} = \overline{\mathbf{q}}$. We will prove that if $\overline{\mathbf{q}}$ is close enough to $\mathbf{q}^\star$, then (3.6) exactly recovers $\mathbf{q}^\star$, and hence $\mathbf{x}_0$.

4 Analysis

4.1 Main Results

In this section, we describe our main theoretical result, which shows that with high probability, the algorithm described in the previous section succeeds.

Theorem 4.1. *Suppose that $\mathcal{S}$ satisfies the planted sparse model, and let the columns of $\mathbf{Y}$ be an arbitrary orthonormal basis for the subspace $\mathcal{S}$. Let $\mathbf{y}^1, \ldots, \mathbf{y}^p \in \mathbb{R}^n$ denote the (transposes of) the rows of $\mathbf{Y}$. Apply Algorithm 1 with $\lambda = 1/\sqrt{p}$, using initializations $\mathbf{q}^{(0)} = \mathbf{y}^1, \ldots, \mathbf{y}^p$, to produce outputs $\overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_p$. Solve the linear program* (3.6) *with $\mathbf{r} = \overline{\mathbf{q}}_1, \ldots, \overline{\mathbf{q}}_p$, to produce $\widehat{\mathbf{q}}_1, \ldots, \widehat{\mathbf{q}}_p$. Set $i^\star \in \arg\min_i \|\mathbf{Y}\widehat{\mathbf{q}}_i\|_0$. Then*

$$\mathbf{Y}\widehat{\mathbf{q}}_{i^\star} = \gamma \mathbf{x}_0, \tag{4.1}$$

for some $\gamma \neq 0$ with overwhelming probability, provided

$$\exp(n/2)/2 \geq p \geq Cn^4 \log n, \qquad \textit{and} \qquad \frac{1}{\sqrt{n}} \leq \theta \leq \theta_0. \tag{4.2}$$

Here, C and $\theta_0 > 0$ are universal constants.

We can see that the result in Theorem 4.1 is suboptimal compared to the global optimality result in Theorem 2.1 and Barak et al.'s result [BKS13] in sampling complexity. For successful recovery, we require $p \geq \Omega(n^4 \log n)$, while the global optimality and Barak et al. demand $p \geq \Omega(n \log n)$ 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 $\theta \sim O(1)$ rate. The lower bound on θ in Theorem 4.1 is mostly for convenience in the proof; in fact, the LP rounding stage of our algorithm already succeeds with high probability when $\theta \in O(1/\sqrt{n})$.

4.2 A Sketch of Analysis

The proof of our main result requires rather detailed technical analysis of the iteration-by-iteration properties of Algorithm 1. In this section, as illustrated in Fig. 1, we briefly sketch the main ideas. Detailed proofs are deferred to the appendices.

As noted in Section 3, the ADM algorithm is invariant to change of basis. So, we can assume without loss of generality that we are working with the particular basis $\overline{\mathbf{Y}} = \mathbf{ZR}$ defined in that section. In order to further streamline the presentation, we are going to sketch the proof under the assumption that

$$\mathbf{Y} = [\mathbf{x}_0 \mid \mathbf{g}_1 \mid \cdots \mid \mathbf{g}_{n-1}], \tag{4.3}$$

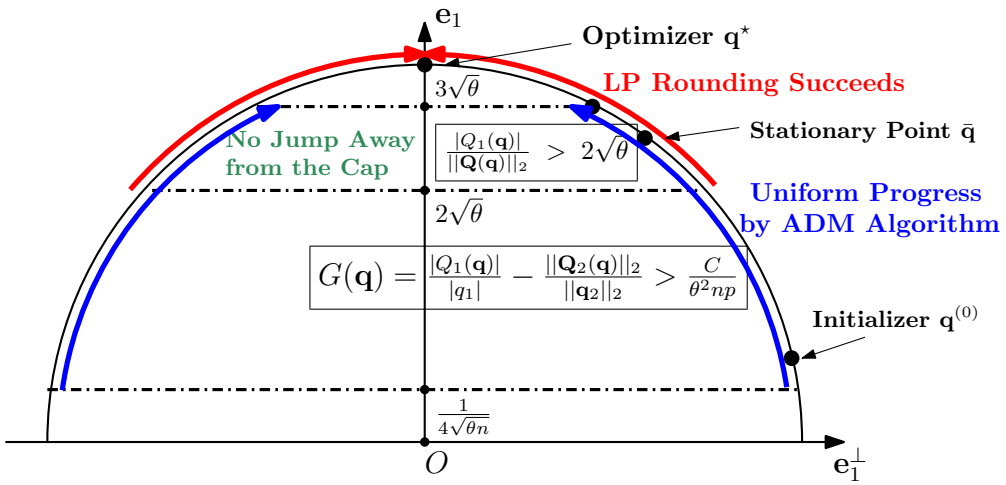

Figure 1: An illustration of the proof sketch for our ADM algorithm.

rather than the orthogonalized version $\overline{\mathbf{Y}}$. When p is large $\mathbf{Y}$ is already nearly orthogonal, and hence $\mathbf{Y}$ is very close to $\overline{\mathbf{Y}}$. In fact, in our proof, we simply carry through the argument for $\mathbf{Y}$, and then note that $\mathbf{Y}$ and $\overline{\mathbf{Y}}$ are close enough that all steps of the proof still hold with $\mathbf{Y}$ replaced by $\overline{\mathbf{Y}}$. With that noted, let $\mathbf{y}^1, \ldots, \mathbf{y}^p \in \mathbb{R}^n$ denote the transposes of the rows of $\mathbf{Y}$, and note that these are independent random vectors. From (3.4), we can see one step of the ADM algorithm takes the form:

$$\mathbf{q}^{(k+1)} = \frac{\frac{1}{p}\sum_{i=1}^p \mathbf{y}^i S_\lambda\left[\left(\mathbf{q}^{(k)}\right)^\top \mathbf{y}^i\right]}{\left\|\frac{1}{p}\sum_{i=1}^p \mathbf{y}^i S_\lambda\left[\left(\mathbf{q}^{(k)}\right)^\top \mathbf{y}^i\right]\right\|_2}. \tag{4.4}$$

This is a very favorable form for analysis: if $\mathbf{q}^{(k)}$ is viewed as fixed, the term in the numerator is a sum of p independent random vectors. To this end, we define a vector valued random process $\mathbf{Q}(\mathbf{q})$ on $\mathbf{q} \in \mathbb{S}^{n-1}$, via

$$\mathbf{Q}(\mathbf{q}) = \frac{1}{p}\sum_{i=1}^p \mathbf{y}^i S_\lambda[\mathbf{q}^\top \mathbf{y}^i]. \tag{4.5}$$

We study the behavior of the iteration (4.4) through the random process $\mathbf{Q}(\mathbf{q})$. We wish to show that with overwhelming probability in our choice of $\mathbf{Y}$, $\mathbf{q}^{(k)}$ converges to some small neighborhood of $\pm\mathbf{e}_1$, so that the ADM algorithm plus the LP rounding (described in Section 3) successfully retrieves the sparse vector $\mathbf{x}_0 = \mathbf{Y}\mathbf{e}_1$. Thus, we hope that in general, $\mathbf{Q}(\mathbf{q})$ is more concentrated on the first coordinate than $\mathbf{q}$. Let us partition the vector $\mathbf{q}$ as $\mathbf{q} = \begin{bmatrix} q_1 \\ \mathbf{q}_2 \end{bmatrix}$, with $q_1 \in \mathbb{R}$ and $\mathbf{q}_2 \in \mathbb{R}^{n-1}$, and correspondingly partition $\mathbf{Q}(\mathbf{q}) = \begin{bmatrix} Q_1(\mathbf{q}) \\ \mathbf{Q}_2(\mathbf{q}) \end{bmatrix}$, where

$$Q_1(\mathbf{q}) = \frac{1}{p}\sum_{i=1}^p x_{0i} S_\lambda\left[\mathbf{q}^\top \mathbf{y}^i\right] \qquad \text{and} \qquad \mathbf{Q}_2(\mathbf{q}) = \frac{1}{p}\sum_{i=1}^p \mathbf{g}_i S_\lambda\left[\mathbf{q}^\top \mathbf{y}^i\right]. \tag{4.6}$$

The inner product of $\mathbf{Q}(\mathbf{q})/\left\|\mathbf{Q}(\mathbf{q})\right\|_2$ and $\mathbf{e}_1$ is strictly larger than the inner product of $\mathbf{q}$ and $\mathbf{e}_1$ if and only if

$$\frac{|Q_1(\mathbf{q})|}{|q_1|} > \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2}. \tag{4.7}$$

In the appendix, we show that with overwhelming probability, this inequality holds uniformly over a significant portion of the sphere, so the algorithm moves in the correct direction. To complete the proof of Theorem 4.1, we combine the following observations, provided $\exp(n/2)/2 \geq p \geq \Omega\left(n^4 \log n\right)$:

1. **Good initializers**. With overwhelming probability, at least one of the initializers $\mathbf{q}^{(0)}$ satisfies $|q_1^{(0)}| > \frac{1}{4\sqrt{\theta n}}$.
2. **Uniform progress away from the equator**. With overwhelming probability, for every $\mathbf{q} \in \mathbb{S}^{n-1}$ such that $\frac{1}{4\sqrt{\theta n}} \leq |q_1| \leq 3\sqrt{\theta}$, the bound

$$\frac{|Q_1(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2(\mathbf{q})\|_2}{\|\mathbf{q}\|_2} > \frac{C}{\theta^2 np} \tag{4.8}$$

   holds, for some numerical constant $C > 0$. This implies that if at any iteration k of the algorithm, $|q_1^{(k)}| > \frac{1}{4\sqrt{\theta n}}$, the algorithm will eventually obtain a point $\mathbf{q}^{(k')}$, $k' > k$, for which $|q_1^{(k')}| > 3\sqrt{\theta}$, if sufficiently many iterations are allowed.
3. **No jumps away from the caps**. With overwhelming probability,

$$\frac{|Q_1(\mathbf{q})|}{\sqrt{|Q_1(\mathbf{q})^2| + \|\mathbf{Q}_2(\mathbf{q})\|_2^2}} \geq 2\sqrt{\theta} \tag{4.9}$$

   for all $\mathbf{q} \in \mathbb{S}^{n-1}$ with $|q_1| > 3\sqrt{\theta}$.
4. **Location of stationary points**. The above steps imply that, with overwhelming probability, Algorithm 1 fed with the proposed initialization scheme produces at least one stopping point $\overline{\mathbf{q}} \in \mathbb{S}^{n-1}$ satisfying $|\overline{q}_1| \geq 2\sqrt{\theta}$.
5. **Rounding succeeds when** $|r_1| > 2\sqrt{\theta}$. With overwhelming probability, the linear programming based rounding (3.6) will produce $\pm\mathbf{x}_0$, up to scale, whenever it is provided with an input $\mathbf{r}$ whose first coordinate has magnitude at least $2\sqrt{\theta}$.

Taken together, these claims imply that from at least one of the initializers $\mathbf{q}^{(0)}$, the ADM algorithm will produce an output $\overline{\mathbf{q}}$ which is accurate enough for LP rounding to exactly return $\mathbf{x}_0$, up to scale. As $\mathbf{x}_0$ is the sparsest nonzero vector in the subspace $\mathcal{S}$ with overwhelming probability, it will be selected as $\mathbf{Y}\mathbf{q}_{i^*}$, and hence produced by the algorithm.

5 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 vector and dictionary learning models; for the real dataset, we demonstrate how seeking sparse vectors can help discover interesting patterns.

5.1 Phase Transition on Synthetic Data

For the planted sparse model, for each pair of (k, p), we generate the n dimensional subspace $\mathcal{S} \in \mathbb{R}^p$ by a k sparse vector $\mathbf{x}_0$ with nonzero entries equal to 1 and a random Gaussian matrix $\mathbf{G} \in \mathbb{R}^{p \times (n-1)}$ with $G_{ij} \sim_{i.i.d.} \mathcal{N}(0, 1/p)$, so that one basis $\mathbf{Y}$ of the subspace $\mathcal{S}$ can be constructed by $\mathbf{Y} = \mathtt{GS}\left([\mathbf{x}_0, \mathbf{G}]\right)\mathbf{U}$, where $\mathtt{GS}(\cdot)$ denotes the Gram-Schmidt orthonormalization operator and $\mathbf{U} \in \mathbb{R}^{n \times n}$ is an arbitrary orthogonal matrix. We fix the relationship between n and p as $p = 5n \log n$, and set the regularization parameter in (3.1) as $\lambda = 1/\sqrt{p}$. We use all the normalized rows of $\mathbf{Y}$ as initializations of $\mathbf{q}$ for the proposed ADM

algorithm, and run every program for 5000 iterations. We determine the recovery to be successful whenever $\|\mathbf{x}_0/\|\mathbf{x}_0\|_2 - \mathbf{Y}\mathbf{q}\|_2 \leq \varepsilon$ for at least one of the p programs, where $\varepsilon = 10^{-3}$. For each pair of (k, p), we repeat the simulation for five times.

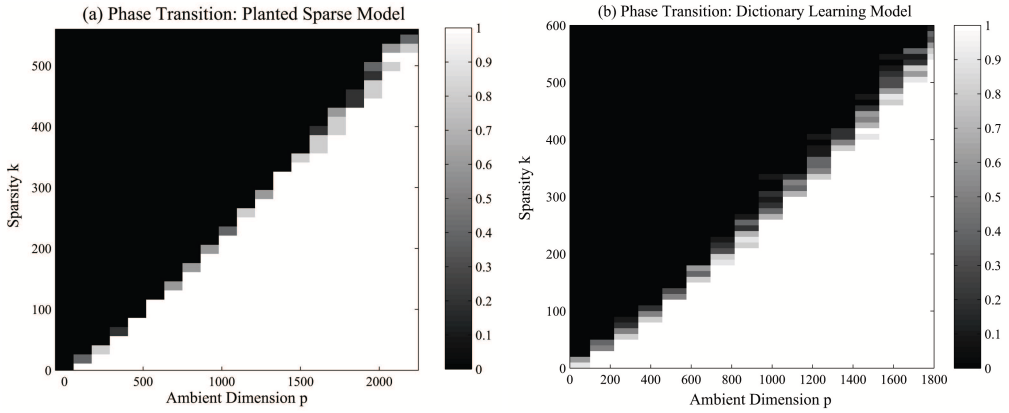

Figure 2: Phase transition for the planted sparse model (left) and dictionary learning model (right) using the ADM algorithm, with fixed relationship between p and n: $p = 5n \log n$. White indicates success and black indicates failure.

Second, we consider the same dictionary learning model as in [SWW12]. 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 $\mathbf{X}_0 = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^\top$, 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 $\mathbf{Y}^\top = \mathtt{GS}\left(\mathbf{X}_0^\top\right)\mathbf{U}^\top$. We repeat the same experiment as for the planted sparse model presented 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.

Figure 2 shows the phase transition between the sparsity level k and p for both models. It seems clear for both problems our algorithm can work well into (even beyond) the linear sparsity regime whenever $p \sim n \log n$. Hence for the planted sparse model, to close the gap between our algorithm and practice is one future direction. Also, how to extend our analysis for dictionary learning is another interesting direction.

5.2 Exploratory Experiments on Faces

It is well known in computer vision that appearance of convex objects only subject to illumination changes leads to image collection that can be well approximated by low-dimensional space in raw-pixel space [BJ03]. 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* [CLMW11] 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 algorithm to find the sparsest element in such a subspace, by randomly selecting 10% rows as initializations for $\mathbf{q}$. We judge the sparsity in a ℓ^1/ℓ^2 sense, that is, the sparsest vector $\widehat{\mathbf{x}}_0 = \mathbf{Y}\mathbf{q}^\star$ 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 found, and continue the seeking process in the projected subspace. Figure 3 shows the first four sparse vectors we get from the data. We can see they correspond well to different extreme illumination conditions.

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. Figure 4 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 bows, forehead hair, nose, etc.

Figure 3: Four sparse vectors extracted by the ADM algorithm for one person in the Yale B database under different illuminations.

Figure 4: Four sparse vectors extracted by the ADM algorithm for 10 persons in the Yale B database under normal illuminations.

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

6 Discussion

The random models we assume for the subspace can be easily extended to other random models, particularly for dictionary learning. Moreover we believe the algorithm paradigm works far beyond the idealized models, as our preliminary experiments on face data have clearly shown. For the particular planted sparse model, the performance gap in terms of (p, n, θ) between the empirical simulation and our result is likely due to analysis itself. Advanced techniques to bound the empirical process, such as decoupling [DlPG99] techniques, can be deployed in place of our crude union bound to cover all iterates. On the application side, 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 can invite more application ideas.

This paper is part of a recent surge of research into provable and practical nonconvex approaches to estimating various types of low-dimensional structures, often in large-scale settings [CLS14, JNS13, Har13, NJS13, YCS13]. The dominant approach is to start with a clever, problem-specific initialization, and then perform a local analysis of the subsequent iterates. Our forthcoming work [SQW14] on dictionary learning takes a more geometric approach, and proves global recovery via efficient algorithms, with arbitrary initialization. The approach developed there may be applicable to the planted sparse model studied here, as well as to many other interesting nonconvex problems.

Acknowledgement

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. This work was partially supported by grants ONR N00014-13-1-0492, NSF 1343282, and funding from the Moore and Sloan Foundations.

References

[AAJ+13] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *arXiv preprint arXiv:1310.7991*, 2013.

[AAN13] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013.

[ABGM14] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, 2014.

[AGM13] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *arXiv preprint arXiv:1308.6273*, 2013.

[AHJK13] Anima Anandkumar, Daniel Hsu, Majid Janzamin, and Sham M Kakade. When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity. In *Advances in Neural Information Processing Systems*, pages 1986–1994, 2013.

[BJ03] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):218–233, 2003.

[BKS13] Boaz Barak, Jonathan Kelner, and David Steurer. Rounding sum-of-squares relaxations. *arXiv preprint arXiv:1312.6652*, 2013.

[BM05] Gregory Beylkin and Lucas Monzón. On approximation of functions by exponential sums. *Applied and Computational Harmonic Analysis*, 19(1):17–48, 2005.

[BR13] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[CLMW11] Emmanuel Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), May 2011.

[CLS14] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *arXiv preprint arXiv:1407.1065*, 2014.

[CP86] Thomas F Coleman and Alex Pothen. The null space problem i. complexity. *SIAM Journal on Algebraic Discrete Methods*, 7(4):527–537, 1986.

[CT05] Emmanuel J Candès and Terence Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[DLH12] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2018–2025. IEEE, 2012.

[DlPG99] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer, 1999.

[Don06] David 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*, 59(6):797–829, 2006.

[FLM77] Tadeusz Figiel, Joram Lindenstrauss, and Vitali D Milman. The dimension of almost spherical sections of convex bodies. *Acta Mathematica*, 139(1):53–94, 1977.

[GG84] Andrej Y Garnaev and Efim D Gluskin. The widths of a euclidean ball. In *Dokl. Akad. Nauk SSSR*, volume 277, pages 1048–1052, 1984.

[GM03] E Gluskin and V Milman. Note on the geometric-arithmetic mean inequality. In *Geometric aspects of Functional analysis*, pages 131–135. Springer, 2003.

[Har13] Moritz Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv preprint arXiv:1312.0925*, 2013.

[HD13] Paul Hand and Laurent Demanet. Recovering the sparsest element in a subspace. *arXiv preprint arXiv:1310.1654*, 2013.

[HXV13] Jeffrey Ho, Yuchen Xie, and Baba Vemuri. On a nonlinear generalization of sparse coding and dictionary learning. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1480–1488, 2013.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pages 665–674. ACM, 2013.

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[Pis99] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

[QSW14] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In *Advances in Neural Information Processing Systems*, 2014.

[SQW14] Ju Sun, Qing Qu, and John Wright. Complete dictionary learning over the sphere. *In preparation*, 2014.

[SWW12] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[YCS13] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. *arXiv preprint arXiv:1310.3745*, 2013.

[ZF13] Yun-Bin Zhao and Masao Fukushima. Rank-one solutions for homogeneous linear matrix equations over the positive semidefinite cone. *Applied Mathematics and Computation*, 219(10):5569–5583, 2013.

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[ZP01] Michael Zibulevsky and Barak A Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.

Appendices

Notes on notations. For a matrix $\mathbf{X}$, $\mathbf{x}_i$ denotes its i-th column, and $\mathbf{x}^j$ denotes its j-th row, all in column vector form. So $\left(\mathbf{x}^j\right)^\top$ is the j-th row in row vector form. We will use the compact notation $[k] \doteq \{1, \ldots, k\}$ for any positive integer k. We will use C or c, and their indexed versions to denote 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 probable events, sometimes we will just say the event holds with "high probability" if the probability of failure is dominated by some polynomial $\mathrm{poly}\,(n, p)$ which diminished to zero whenever n or p is large, with "overwhelming probability" if the failure probability is dominated some exponential function $\exp\left(\mathrm{poly}\,(n, p)\right)$ which diminishes to zero whenever n or p is large.

A Technical Tools and Preliminaries

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:*

$$(\textit{Standard Normal pdf}) \quad \psi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \tag{A.1}$$

$$(\textit{Standard Normal cdf}) \quad \Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{t^2}{2}\right\} dt, \tag{A.2}$$

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

$$\int_{t_1}^{t_2} f_\sigma(x)dx = \Psi\left(\frac{t_2}{\sigma}\right) - \Psi\left(\frac{t_1}{\sigma}\right), \tag{A.3}$$

$$\int_{t_1}^{t_2} x f_\sigma(x)dx = -\sigma\left[\psi\left(\frac{t_2}{\sigma}\right) - \psi\left(\frac{t_1}{\sigma}\right)\right], \tag{A.4}$$

$$\int_{t_1}^{t_2} x^2 f_\sigma(x)dx = \sigma^2\left[\Psi\left(\frac{t_2}{\sigma}\right) - \Psi\left(\frac{t_1}{\sigma}\right)\right] - \sigma\left[t_2\psi\left(\frac{t_2}{\sigma}\right) - t_1\psi\left(\frac{t_1}{\sigma}\right)\right]. \tag{A.5}$$

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*

$$|\psi(x) - [\psi(x_0) - x_0\psi(x_0)(x - x_0)]| \le C_\psi(x - x_0)^2, \tag{A.6}$$

$$|\Psi(x) - [\Psi(x_0) + \psi(x_0)(x - x_0)]| \le C_\psi(x - x_0)^2. \tag{A.7}$$

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

$$\|\mathbf{A}\|_{\ell^p\to\ell^q} \doteq \sup_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\mathbf{x}\|_q. \tag{A.8}$$

In particular, we have

$$\|\mathbf{A}\|_{\ell^2\to\ell^1} = \sup_{\|\mathbf{x}\|_2=1} \sum_{k=1}^{p} \left|\mathbf{a}_k^\top \mathbf{x}\right|, \quad \|\mathbf{A}\|_{\ell^2\to\ell^\infty} = \max_{1\le k\le p} \left\|\mathbf{a}^k\right\|_2, \tag{A.9}$$

$$\|\mathbf{A}\mathbf{B}\|_{\ell^p\to\ell^r} \le \|\mathbf{A}\|_{\ell^q\to\ell^r} \|\mathbf{B}\|_{\ell^p\to\ell^q}, \tag{A.10}$$

and $\mathbf{A}$ and $\mathbf{B}$ are any matrices of compatible size.

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

$$\mathbb{E}\left[|X|^m\right] = \sigma_X^m \left(m-1\right)!! \left[\sqrt{\frac{2}{\pi}} \mathbb{1}_{m=2k+1} + \mathbb{1}_{m=2k}\right] \leq \sigma_X^m \left(m-1\right)!!, \; k = \lfloor m/2 \rfloor. \tag{A.11}$$

Lemma A.5 (Moments of the χ Random Variable). *If $X \sim \chi\left(n\right)$, i.e., $X \equiv_d \left\|\mathbf{x}\right\|_2$[6] for $\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$. Then it holds for all integer $m \geq 1$ that*

$$\mathbb{E}\left[X^m\right] = 2^{m/2} \frac{\Gamma\left(m/2 + n/2\right)}{\Gamma\left(n/2\right)} \leq m! n^{m/2} \tag{A.12}$$

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

$$\mathbb{E}\left[X^m\right] = 2^m \frac{\Gamma\left(m + n/2\right)}{\Gamma\left(n/2\right)} = \prod_{k=1}^{m} (n + 2k - 2) \leq \frac{m!}{2} (2n)^m. \tag{A.13}$$

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

$$\mathbb{E}\left[|X_k|^m\right] \leq \frac{m!}{2} \sigma_X^2 R^{m-2}, \;\; \textit{for all integers } m \geq 2. \tag{A.14}$$

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

$$\mathbb{P}\left[\left|S - \mathbb{E}\left[S\right]\right| \geq t\right] \leq 2 \exp\left(-\frac{pt^2}{2\sigma_X^2 + 2Rt}\right). \tag{A.15}$$

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

$$\mathbb{E}\left[\left\|\mathbf{x}_k\right\|_2^m\right] \;\leq\; \frac{m!}{2} \sigma_X^2 R^{m-2}, \quad \textit{for all integers } m \geq 2. \tag{A.16}$$

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

$$\mathbb{P}\left[\left\|\mathbf{s} - \mathbb{E}\left[\mathbf{s}\right]\right\|_2 \geq t\right] \;\leq\; 2(d+1) \exp\left(-\frac{pt^2}{2\sigma_X^2 + 2Rt}\right). \tag{A.17}$$

Lemma A.9 (Hoeffding's Inequality). *Let $X_1, \cdots, X_p$ be independent random variables such that X_k takes its values in $[a_k, b_k]$ almost surely for all $1 \leq k \leq p$. Let $S = \sum_{k=1}^{p} \left(X_k - \mathbb{E}X_k\right)$, then for every $t > 0$,*

$$\mathbb{P}\left[S \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{p} (b_k - a_k)^2}\right). \tag{A.18}$$

Lemma A.10 (Gaussian Concentration Inequality). *Let $\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right)$. Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ be an L-Lipschitz function. Then we have for all $t > 0$ that*

$$\mathbb{P}\left[f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geq t\right] \leq \exp\left(-\frac{t^2}{2L^2}\right). \tag{A.19}$$

[6]The notation $\equiv_d$ means equivalent in distribution.

Lemma A.11 (Bounding Maximum Norm of Gaussian Vector Sequence)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be a sequence of (not necessarily independent) standard Gaussian vectors in $\mathbb{R}^n$. Then, it holds that*

$$\mathbb{P}\left[\max_{i\in[p]} \|\mathbf{x}_i\|_2 > \sqrt{2\log p} + 2\sqrt{n}\right] \leq \exp\left(-\frac{1}{2}n\right). \tag{A.20}$$

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

$$\mathbb{P}\left[\|\mathbf{x}_i\|_2 - \sqrt{\mathbb{E}\|\mathbf{x}_i\|_2^2} > t\right] \leq \mathbb{P}\left[\|\mathbf{x}_i\|_2 - \mathbb{E}\|\mathbf{x}_i\|_2 > t\right] \leq \exp\left(-\frac{t^2}{2}\right) \tag{A.21}$$

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

$$\mathbb{P}\left[\max_{i\in[p]} \|\mathbf{x}_i\| > \sqrt{n} + t\right] \leq \exp\left(-\frac{t^2}{2} + \log p\right) \tag{A.22}$$

for all $t > 0$. Taking $t = \sqrt{2\log p} + \sqrt{n}$ and simplifying the terms gives the claimed result. □

Lemma A.12 (Covering Number of a Unit Ball)**.** *Let $\mathcal{B} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq 1\}$ be a unit ball. For any $\varepsilon \in (0,1)$, there exists some ε cover of $\mathcal{B}$ w.r.t. the normal $\mathbb{R}^n$ metric, denoted as $\mathcal{N}_\varepsilon$, such that*

$$|\mathcal{N}_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^n \leq \left(\frac{3}{\varepsilon}\right)^n. \tag{A.23}$$

Lemma A.13 (Spectrum of Gaussian Matrices, [Ver10])**.** *Let $\mathbf{A} \in \mathbb{R}^{p\times n}$ $(p > n)$ contain i.i.d. standard normal entries. Then for every $t \geq 0$, with probability at least $1 - 2\exp\left(-t^2/2\right)$, one has*

$$\sqrt{p} - \sqrt{n} - t \leq \sigma_{\min}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{A}) \leq \sqrt{p} + \sqrt{n} + t. \tag{A.24}$$

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 $\mathbf{\Phi} \in \mathbb{R}^{n_1\times n_2} \sim_{i.i.d.} \mathcal{N}(0,1)$ obeys*

$$(1-\varepsilon)\sqrt{\frac{2}{\pi}} n_1 \|\mathbf{x}\|_2 \leq \|\mathbf{\Phi}\mathbf{x}\|_1 \leq (1+\varepsilon)\sqrt{\frac{2}{\pi}} n_1 \|\mathbf{x}\|_2 \quad \textit{for all } \mathbf{x} \in \mathbb{R}^{n_2}, \tag{A.25}$$

with probability at least $1 - 2\exp(-c(\varepsilon)\, n_2)$ for some $c(\varepsilon) > 0$.

Geometrically, this lemma roughly corresponds to the well known almost spherical section theorem [FLM77, GG84], see also [GM03]. A slight variant of this version has been proved in [Don06], borrowing ideas from [Pis99].

Proof. By homogeneity, it is enough to consider all $\mathbf{x}$ with unit norms. For a fixed $\mathbf{x}_0$ with $\|\mathbf{x}_0\|_2 = 1$, $\mathbf{\Phi}\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. So $\mathbb{E}\|\mathbf{\Phi}\mathbf{x}\|_1 = \sqrt{\frac{2}{\pi}} n_1$. By concentration of measure for Gaussian vectors,

$$\mathbb{P}\left[\left|\|\mathbf{\Phi}\mathbf{x}\|_1 - \mathbb{E}\left[\|\mathbf{\Phi}\mathbf{x}\|_1\right]\right| > t\right] \leq 2\exp\left(-\frac{t^2}{2n_1}\right) \tag{A.26}$$

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

$$\mathcal{E} \doteq \left\{(1-\delta)\sqrt{\frac{2}{\pi}} n_1 \leq \|\mathbf{\Phi}\mathbf{x}\|_1 \leq (1+\delta)\sqrt{\frac{2}{\pi}} n_1 \;\forall\, \mathbf{x} \in N_1\right\}. \tag{A.27}$$

A simple application of union bound yields

$$\mathbb{P}\left[\mathcal{E}^c\right] \leq 2\exp\left(-\frac{\delta^2 n_1}{\pi} + n_2 \log\left(1+\frac{2}{\delta}\right)\right). \tag{A.28}$$

Choosing δ small enough such that

$$(1-3\delta)(1-\delta)^{-1} \geq 1-\varepsilon \text{ and } (1+\delta)(1-\delta)^{-1} \leq 1+\varepsilon, \tag{A.29}$$

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

$$(1-\varepsilon)\sqrt{\frac{2}{\pi}}n_1 \leq \|\mathbf{\Phi x}\|_1 \leq (1+\varepsilon)\sqrt{\frac{2}{\pi}}n_1 \;\forall\, \mathbf{x} \in \mathbb{S}^{n_2-1}. \tag{A.30}$$

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

$$\mathbf{z} = \sum_{k=0}^{\infty} \lambda_k \mathbf{x}_k, \qquad \text{with } |\lambda_k| \leq \delta^k, \mathbf{x}_k \in N_1 \text{ for all } k. \tag{A.31}$$

Hence we have

$$\|\mathbf{\Phi z}\|_1 = \left\|\mathbf{\Phi}\sum_{k=0}^{\infty}\lambda_k \mathbf{x}_k\right\|_1 \leq \sum_{k=0}^{\infty}\delta^k \|\mathbf{\Phi x}_k\|_1 \leq (1+\delta)(1-\delta)^{-1}\sqrt{\frac{2}{\pi}}n_1. \tag{A.32}$$

Similarly,

$$\|\mathbf{\Phi z}\|_1 = \left\|\mathbf{\Phi}\sum_{k=0}^{\infty}\lambda_k \mathbf{x}_k\right\|_1 \geq \left[1-\delta-\delta(1+\delta)(1-\delta)^{-1}\right]\sqrt{\frac{2}{\pi}}n_1 = (1-3\delta)(1-\delta)^{-1}\sqrt{\frac{2}{\pi}}n_1. \tag{A.33}$$

Hence, the choice of δ above leads to the claimed result. To make $\mathbb{P}\left[\mathcal{E}^c\right]$ small, it is enough to choose C such that

$$C\delta^2/\pi > \log\left(1+\frac{2}{\delta}\right). \tag{A.34}$$

Setting $C = 2\log\left(1+\frac{2}{\delta}\right)\pi/\delta^2$ completes the proof. □

Lemma A.15. *Suppose $n_1 \leq \frac{1}{2}\exp(n_2/2)$. Fix $\varepsilon \in (0,1)$. Then for any ξ such that $\xi^2 > 2\log(1+2\varepsilon)$. The random matrix $\mathbf{\Phi} \in \mathbb{R}^{n_1 \times n_2} \sim_{i.i.d.} \mathcal{N}(0,1)$ obeys*

$$\|\mathbf{\Phi x}\|_\infty \leq \frac{1+\xi}{1-\varepsilon}\sqrt{n_2}\,\|\mathbf{x}\|_2 \quad \textit{for all } \mathbf{x} \in \mathbb{R}^{n_2}, \tag{A.35}$$

with probability at least $1-\exp\left(-n_2\left(\xi^2/2-\log(1+2\varepsilon)\right)\right)$.

Proof. Again for a fixed $\mathbf{x}_0 \in \mathbb{S}^{n_2-1}$, $\mathbf{\Phi x}_0 \equiv_d \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For any fixed $\beta > 0$ to be decided later,

$$\mathbb{E}\left[\beta\|\mathbf{v}\|_\infty\right] = \mathbb{E}\left[\beta \max_{i\in[n_1]}|v_i|\right] = \mathbb{E}\left[\log \max_{i\in[n_1]}\exp(\beta|v_1|)\right] \leq \log\mathbb{E}\left[\max_{i\in[n_1]}\exp(\beta|v_1|)\right] \tag{A.36}$$

$$\leq \log\mathbb{E}\left[\sum_{i=1}^{n_1}\exp(\beta|v_1|)\right] = \log n_1\mathbb{E}\left[\exp(\beta|v_1|)\right] \leq \log 2n_1\exp\left(\beta^2/2\right). \tag{A.37}$$

Hence

$$\mathbb{E}\left[\|\mathbf{v}\|_\infty\right] \leq \frac{\log 2n_1\exp\left(\beta^2/2\right)}{\beta}. \tag{A.38}$$

Taking $\beta = \sqrt{2\log(2n_1)}$, we obtain

$$\mathbb{E}\left[\|\mathbf{v}\|_\infty\right] \le \sqrt{2\log(2n_1)}. \tag{A.39}$$

Because the mapping $\mathbf{v} \mapsto \|\mathbf{v}\|_\infty$ is 1-Lipschitz, by concentration of measure for Gaussian vectors, we obtain

$$\mathbb{P}\left[\|\mathbf{\Phi x}\|_\infty - \mathbb{E}\left[\|\mathbf{\Phi x}\|_\infty\right] > t\right] \le \exp\left(-\frac{t^2}{2}\right). \tag{A.40}$$

Taking $t = \xi\sqrt{n_2}$, and consider an ε-net $\mathcal{N}_\varepsilon$ that covers $\mathcal{S}^{n_2-1}$ with cardinality $|\mathcal{N}_\varepsilon| \le (1+2/\varepsilon)^{n_2}$, we have the event

$$\mathcal{E} \doteq \{\|\mathbf{\Phi x}\|_\infty \le (1+\xi)\sqrt{n_2}\ \forall\ \mathbf{x} \in N_\varepsilon\} \tag{A.41}$$

holds with probability at least $1 - \exp\left(-\xi^2 n_2/2 + n_2\log(1+2\varepsilon)\right)$. Conditioned on $\mathcal{E}$, we have

$$\sup_{\|\mathbf{z}\|_2=1}\|\mathbf{\Phi z}\|_\infty \le \sup_{\mathbf{z}'\in\mathcal{N}_\varepsilon}\|\mathbf{\Phi z}'\|_\infty + \sup_{\|\mathbf{e}\|_2\le\varepsilon}\|\mathbf{\Phi e}\|_\infty = \sup_{\mathbf{z}'\in\mathcal{N}_\varepsilon}\|\mathbf{\Phi z}'\|_\infty + \varepsilon\sup_{\|\mathbf{e}\|_2=1}\|\mathbf{\Phi e}\|_\infty. \tag{A.42}$$

Hence we have

$$\sup_{\|\mathbf{z}\|_2=1}\|\mathbf{\Phi z}\|_\infty \le \frac{1}{1-\varepsilon}\sup_{\mathbf{z}'\in\mathcal{N}_\varepsilon}\|\mathbf{\Phi z}'\|_\infty = \frac{1+\xi}{1-\varepsilon}\sqrt{n_2}, \tag{A.43}$$

completing the proof. □

B The Random Basis vs. Its Orthonormalized Version

We consider $\mathbf{Y}$ obeying the planted sparse model:

$$\mathbf{Y} = [\mathbf{x}_0 \mid \mathbf{G}] \in \mathbb{R}^{p\times n} \tag{B.1}$$

with

$$\mathbf{x}_0 \sim_{i.i.d.} \frac{1}{\sqrt{\theta p}}\mathrm{Ber}(\theta), \mathbf{G} \sim_{i.i.d.} \mathcal{N}\left(0, \frac{1}{p}\right). \tag{B.2}$$

One "natural/canonical" orthonormal basis for the subspace spanned by columns of $\mathbf{Y}$ is

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

We also write $\mathbf{G}' \doteq \mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\left(\mathbf{G}^\top\mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\right)^{-1/2}$ for convenience. In this section, we want to show that the intuition $\mathbf{Y}'$ well approximating $\mathbf{Y}$[7] can be made rigorous. These results are needed when we prove Theorem 2.1 for the global optimality of the natural ℓ^2 constrained formulation (2.1), as well as when we translate the results for $\mathbf{Y}$ to quantitative statements about $\mathbf{Y}'$ in Appendix F.4.

For any realization of $\mathbf{x}_0$, let the support (index set of nonzero elements) of $\mathbf{x}_0$ be $\mathcal{I}$. By Hoeffding's inequality in Lemma A.9, we have the event

$$\mathcal{E}_0 \doteq \left\{\frac{1}{2}\theta p \le |\mathcal{I}| \le 2\theta p\right\} \tag{B.4}$$

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

[7] When n and p are large, $\mathbf{Y}$ has nearly orthonormal columns.

Lemma B.1. *The bound*

$$\left|1-\frac{1}{\|\mathbf{x}_0\|_2}\right| \;\le\; \frac{2\sqrt{2}}{5}\sqrt{\frac{n\log p}{\theta^3 p}} \tag{B.5}$$

holds with probability at least $1-2\exp\left(-p\theta^2/2\right)-2\exp\left(-2n\log p\right)$.

Proof. Because $\mathbb{E}\left[\|\mathbf{x}_0\|_2^2\right]=1$, by Hoeffding's inequality in Lemma A.9, we have

$$\mathbb{P}\left[\left|\|\mathbf{x}_0\|_2^2-\mathbb{E}\left[\|\mathbf{x}_0\|_2^2\right]>t\right|\right]=\mathbb{P}\left[\left|\|\mathbf{x}_0\|_2^2-1\right|>t\right]\le 2\exp\left(-2\theta^2pt^2\right) \tag{B.6}$$

for all $t>0$, which implies

$$\mathbb{P}\left[\left|\|\mathbf{x}_0\|_2-1\right|\left(\|\mathbf{x}_0\|_2+1\right)>t\right]\le 2\exp\left(-2\theta^2pt^2\right). \tag{B.7}$$

On the intersection with $\mathcal{E}_0$, $\|\mathbf{x}_0\|_2+1\le\sqrt{2}+1\le 5/2$, and setting $t=\sqrt{\frac{n\log p}{\theta^3 p}}$, we obtain

$$\mathbb{P}\left[\left|\|\mathbf{x}_0\|_2-1\right|>\frac{2}{5}\sqrt{\frac{n\log p}{\theta^3 p}}\right]\le 2\exp\left(-2n\log p\right). \tag{B.8}$$

So we obtain that with probability at least $1-2\exp\left(-p\theta^2/2\right)-2\exp\left(-2n\log p\right)$,

$$\left|1-\frac{1}{\|\mathbf{x}_0\|_2}\right|=\frac{\left|1-\|\mathbf{x}_0\|_2\right|}{\|\mathbf{x}_0\|_2}\le\frac{2\sqrt{2}}{5}\sqrt{\frac{n\log p}{\theta^3 p}}, \tag{B.9}$$

as desired. □

Next, 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* $p\ge Cn\ge 2$ *for some large enough constant* C*, it holds that*

$$\|\mathbf{M}\|\le 2,\quad \|\mathbf{M}-\mathbf{I}\|\le 4\sqrt{\frac{n}{p}} \tag{B.10}$$

with probability at least $1-c'\exp\left(-c''n\right)$ *for some positive constants* c' *and* c''.

Proof. First observe that

$$\|\mathbf{M}\|=\left(\sigma_{\min}\left(\mathbf{G}^\top\mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\right)\right)^{-1/2}=\left(\sigma_{\min}\left(\mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\right)\right)^{-1}. \tag{B.11}$$

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,

$$\sigma_{\min}\left(\mathcal{P}_{\mathbf{x}_0^\perp}\mathbf{G}\right)=\sigma_{\min}\left(\mathbf{B}^\top\mathbf{G}\right). \tag{B.12}$$

Since $\mathbf{G}\sim_{i.i.d.}\mathcal{N}\left(0,\frac{1}{p}\right)$, 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

$$\sqrt{\frac{p-1}{p}}-2\sqrt{\frac{n-1}{p-1}}\le\sigma_{\min}\left(\mathbf{B}^\top\mathbf{G}\right)\le\sigma_{\max}\left(\mathbf{B}^\top\mathbf{G}\right)\le\sqrt{\frac{p-1}{p}}+2\sqrt{\frac{n-1}{p-1}} \tag{B.13}$$

with probability at least $1 - c_1 \exp(-c_2 n)$ for some $c_1, c_2 > 0$. Thus, when $p \geq C_1 n$ for some large constant C_1, we have

$$\|\mathbf{M}\| = \left(\sqrt{\frac{p-1}{p}} - 2\sqrt{\frac{n-1}{p-1}}\right)^{-1} \leq 2, \tag{B.14}$$

$$\|\mathbf{I} - \mathbf{M}\| = \max\left(|\sigma_{\max}(\mathbf{M}) - 1|, |\sigma_{\min}(\mathbf{M}) - 1|\right) \leq 2\sqrt{\frac{n-1}{p-1}}\left(\sqrt{\frac{p-1}{p}} - 2\sqrt{\frac{n-1}{p-1}}\right)^{-1} \leq 4\sqrt{\frac{n}{p}}, \tag{B.15}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. □

Lemma B.3. *There exists a constant $C > 0$, such that when $p \geq Cn$, the following*

$$\|\mathbf{Y}\|_{\ell^2 \to \ell^1} \leq 3\sqrt{p}, \tag{B.16}$$

$$\|\mathbf{Y}'_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \leq 7\sqrt{2\theta p}, \tag{B.17}$$

$$\|\mathbf{Y}_{\mathcal{I}} - \mathbf{Y}'_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \leq \frac{10}{\theta}\sqrt{n \log p}, \tag{B.18}$$

$$\|\mathbf{G} - \mathbf{G}'\|_{\ell^2 \to \ell^1} \leq 8\sqrt{n}, \tag{B.19}$$

$$\|\mathbf{Y} - \mathbf{Y}'\|_{\ell^2 \to \ell^1} \leq \frac{10}{\theta}\sqrt{n \log p} \tag{B.20}$$

hold simultaneously with probability at least $1 - c' \exp(-c'' n) - 2\exp\left(-p\theta^2/2\right)$ for some positive constants c' and c''.

Proof. First of all, we have

$$\left\|\frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2 \to \ell^1} \leq \frac{1}{\|\mathbf{x}_0\|_2^2}\|\mathbf{x}_0\|_{\ell^2 \to \ell^1}\left\|\mathbf{x}_0^\top \mathbf{G}\mathbf{M}\right\|_{\ell^2 \to \ell^2} = \frac{2}{\|\mathbf{x}_0\|_2^2}\|\mathbf{x}_0\|_1\left\|\mathbf{x}_0^\top \mathbf{G}\right\|_2, \tag{B.21}$$

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 measure concentration inequality for Gaussian vectors, we have

$$\left\|\mathbf{x}_0^\top \mathbf{G}\right\|_2 \leq 2\|\mathbf{x}_0\|_2\sqrt{\frac{n}{p}} \tag{B.22}$$

with probability at least $1 - \exp(-n/2)$. On the intersection with $\mathcal{E}_0$, this implies

$$\left\|\frac{\mathbf{x}_0 \mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2 \to \ell^1} \leq 4\sqrt{|\mathcal{I}|}\sqrt{\frac{n}{p}} \leq 4\sqrt{2\theta n}, \tag{B.23}$$

with probability at least $1 - \exp(-n/2) - 2\exp\left(-p\theta^2/2\right)$. Moreover, when intersected with $\mathcal{E}_0$, Lemma A.14 implies that when $p \geq \Omega(n)$,

$$\|\mathbf{G}\|_{\ell^2 \to \ell^1} \leq \sqrt{p}, \quad \|\mathbf{G}_{\mathcal{I}}\|_{\ell^2 \to \ell^1} \leq \sqrt{2\theta p} \tag{B.24}$$

with probability at least $1 - c_1 \exp(-c_2 n) - 2\exp\left(-p\theta^2/2\right)$, for some positive constants c_1, c_2. So when

$p \geq \Omega(n)$,

$$\|\mathbf{G}-\mathbf{G}'\|_{\ell^2\to\ell^1} \leq \|\mathbf{G}-\mathbf{G}\mathbf{M}\|_{\ell^2\to\ell^1} + \left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^1} \leq \|\mathbf{G}\|_{\ell^2\to\ell^1}\|\mathbf{I}-\mathbf{M}\| + 4\sqrt{2\theta n} \leq 8\sqrt{n}, \quad \text{(B.25)}$$

$$\|\mathbf{Y}\|_{\ell^2\to\ell^1} \leq \|\mathbf{x}_0\|_{\ell^2\to\ell^1} + \|\mathbf{G}\|_{\ell^2\to\ell^1} \leq \|\mathbf{x}_0\|_1 + \sqrt{p} \leq 2\sqrt{\theta p} + \sqrt{p} \leq 3\sqrt{p}, \quad \text{(B.26)}$$

$$\|\mathbf{G}_{\mathcal{I}}'\|_{\ell^2\to\ell^1} \leq \|\mathbf{G}_{\mathcal{I}}\mathbf{M}\|_{\ell^2\to\ell^1} + \left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^1} \leq \|\mathbf{G}_{\mathcal{I}}\|_{\ell^2\to\ell^1}\|\mathbf{M}\| + 4\sqrt{2\theta n} \leq 6\sqrt{2\theta p}, \quad \text{(B.27)}$$

$$\|\mathbf{G}_{\mathcal{I}}-\mathbf{G}_{\mathcal{I}}'\|_{\ell^2\to\ell^1} \leq \|\mathbf{G}_{\mathcal{I}}(\mathbf{I}-\mathbf{M})\|_{\ell^2\to\ell^1} + \left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^1} \leq \|\mathbf{G}_{\mathcal{I}}\|_{\ell^2\to\ell^1}\|\mathbf{I}-\mathbf{M}\| + 4\sqrt{2\theta n} \leq 8\sqrt{2\theta n}, \quad \text{(B.28)}$$

$$\|\mathbf{Y}_{\mathcal{I}}'\|_{\ell^2\to\ell^1} \leq \left\|\frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2}\right\|_{\ell^2\to\ell^1} + \|\mathbf{G}_{\mathcal{I}}'\|_{\ell^2\to\ell^1} \leq \frac{\|\mathbf{x}_0\|_1}{\|\mathbf{x}_0\|_2} + 6\sqrt{2\theta p} \leq 7\sqrt{2\theta p} \quad \text{(B.29)}$$

with probability at least $1-c_3\exp(-c_4 n)-2\exp\left(-p\theta^2/2\right)$ for some positive constants c_3, c_4, where we have used the above estimates and the results in Lemma B.2. Finally, by Lemma B.1, we obtain

$$\|\mathbf{Y}-\mathbf{Y}'\|_{\ell^2\to\ell^1} \leq \left|1-\frac{1}{\|\mathbf{x}_0\|_2}\right|\|\mathbf{x}_0\|_1 + \|\mathbf{G}-\mathbf{G}'\|_{\ell^2\to\ell^1} \leq \frac{10}{\theta}\sqrt{n\log p}, \quad \text{(B.30)}$$

$$\|\mathbf{Y}_{\mathcal{I}}-\mathbf{Y}_{\mathcal{I}}'\|_{\ell^2\to\ell^1} \leq \left|1-\frac{1}{\|\mathbf{x}_0\|_2}\right|\|\mathbf{x}_0\|_1 + \|\mathbf{G}_{\mathcal{I}}-\mathbf{G}_{\mathcal{I}}'\|_{\ell^2\to\ell^1} \leq \frac{10}{\theta}\sqrt{n\log p}, \quad \text{(B.31)}$$

holding with probability at least $1-c_5\exp(-c_6 n)-2\exp\left(-p\theta^2/2\right)$ for some positive constants c_5, c_6. □

Lemma B.4. *Provided $Cn \leq p \leq \exp(n/2)/2$ for some constant $C>0$, the following*

$$\|\mathbf{G}'\|_{\ell^2\to\ell^\infty} \leq 16\sqrt{\frac{n}{\theta p}}, \quad \text{(B.32)}$$

$$\|\mathbf{G}-\mathbf{G}'\|_{\ell^2\to\ell^\infty} \leq \frac{32n}{\sqrt{\theta}p} \quad \text{(B.33)}$$

hold simultaneously with probability at least $1-c'\exp(-c''n)-2\exp\left(-p\theta^2/2\right)$ for some positive constants c' and c''.

Proof. First of all, we have

$$\left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^\infty} \leq \frac{1}{\|\mathbf{x}_0\|_2^2}\|\mathbf{x}_0\|_{\ell^2\to\ell^\infty}\left\|\mathbf{x}_0^\top\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^2} = \frac{2}{\|\mathbf{x}_0\|_2^2}\|\mathbf{x}_0\|_\infty\left\|\mathbf{x}_0^\top\mathbf{G}\right\|_2, \quad \text{(B.34)}$$

where at the last inequality we have applied the fact $\|\mathbf{M}\| \leq 2$ from Lemma B.2. Similar to the proof to Lemma B.3, we have that $\left\|\mathbf{x}_0^\top\mathbf{G}\right\|_2 \leq 2\|\mathbf{x}_0\|_2\sqrt{n/p}$ with probability at least $1-\exp(-n/2)$. So on the intersection with $\mathcal{E}_0$, we obtain that

$$\left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\|\mathbf{x}_0\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^\infty} \leq \frac{4\|\mathbf{x}_0\|_\infty}{\|\mathbf{x}_0\|_2}\sqrt{\frac{n}{p}} \leq \frac{4\sqrt{2n}}{\sqrt{\theta}p} \quad \text{(B.35)}$$

holds with probability at least $1-\exp(-n/2)-2\exp\left(-p\theta^2/2\right)$. Now taking $\xi=2$ and $\varepsilon=1/2$ in Lemma A.15, we have that

$$\|\mathbf{G}\|_{\ell^2\to\ell^\infty} \leq 6\sqrt{\frac{n}{p}} \quad \text{(B.36)}$$

with probability at least $1-\exp\left(-n\left(2-\log 2\right)\right)$. Combining with results in Lemma B.2, we obtain

$$\begin{aligned}\left\|\mathbf{G}'\right\|_{\ell^2\to\ell^\infty} &\le \left\|\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^\infty} + \left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\left\|\mathbf{x}_0\right\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^\infty} \\ &\le \left\|\mathbf{G}\right\|_{\ell^2\to\ell^\infty}\left\|\mathbf{M}\right\| + \frac{4\sqrt{2n}}{\sqrt{\theta}p} \le 12\sqrt{\frac{n}{p}} + \frac{4\sqrt{2n}}{\sqrt{\theta}p} \le 16\sqrt{\frac{n}{\theta p}}, \end{aligned} \tag{B.37}$$

$$\left\|\mathbf{G}-\mathbf{G}'\right\|_{\ell^2\to\ell^\infty} \le \left\|\mathbf{G}\right\|_{\ell^2\to\ell^\infty}\left\|\mathbf{I}-\mathbf{M}\right\| + \left\|\frac{\mathbf{x}_0\mathbf{x}_0^\top}{\left\|\mathbf{x}_0\right\|_2^2}\mathbf{G}\mathbf{M}\right\|_{\ell^2\to\ell^\infty} \le \frac{24n}{p} + \frac{4\sqrt{2n}}{\sqrt{\theta}p} \le \frac{32n}{\sqrt{\theta}p} \tag{B.38}$$

with probability at least $1-c_7\exp\left(-c_8 n\right)-2\exp\left(-p\theta^2/2\right)$ for some positive constants c_7, c_8.

□

C Proof of ℓ^1/ℓ^2 Global Optimality

Proof. We will first analyze a canonical version, in which the input basis is $\mathbf{Y}'$:

$$\min_{\mathbf{q}\in\mathbb{R}^n}\left\|\mathbf{Y}'\mathbf{q}\right\|_1, \qquad \text{s.t. } \left\|\mathbf{q}\right\|_2 = 1. \tag{C.1}$$

Let $\mathbf{q}=[q_1;\mathbf{q}_2]$. For any fixed support $\mathcal{I}$ of $\mathbf{x}_0$, we have

$$\begin{aligned}\left\|\mathbf{Y}'\mathbf{q}\right\|_1 &= \left\|\mathbf{Y}'_{\mathcal{I}}\mathbf{q}\right\|_1 + \left\|\mathbf{Y}'_{\mathcal{I}^c}\mathbf{q}\right\|_1 \\ &\ge \left|q_1\right|\left\|\frac{\mathbf{x}_0}{\left\|\mathbf{x}_0\right\|_2}\right\|_1 - \left\|\mathbf{G}'_{\mathcal{I}}\mathbf{q}_2\right\|_1 + \left\|\mathbf{G}'_{\mathcal{I}^c}\mathbf{q}_2\right\|_1 \\ &\ge \left|q_1\right|\left\|\frac{\mathbf{x}_0}{\left\|\mathbf{x}_0\right\|_2}\right\|_1 - \left\|\mathbf{G}_{\mathcal{I}}\mathbf{q}_2\right\|_1 - \left\|\left(\mathbf{G}_{\mathcal{I}}-\mathbf{G}'_{\mathcal{I}}\right)\mathbf{q}_2\right\|_1 + \left\|\mathbf{G}_{\mathcal{I}^c}\mathbf{q}_2\right\|_1 - \left\|\left(\mathbf{G}_{\mathcal{I}^c}-\mathbf{G}'_{\mathcal{I}^c}\right)\mathbf{q}_2\right\|_1 \\ &\ge \left|q_1\right|\left\|\frac{\mathbf{x}_0}{\left\|\mathbf{x}_0\right\|_2}\right\|_1 - \left\|\mathbf{G}_{\mathcal{I}}\mathbf{q}_2\right\|_1 + \left\|\mathbf{G}_{\mathcal{I}^c}\mathbf{q}_2\right\|_1 - \left\|\mathbf{G}-\mathbf{G}'\right\|_{\ell^2\to\ell^1}\left\|\mathbf{q}_2\right\|_2. \end{aligned} \tag{C.2}$$

By Lemma A.14 and intersecting with $\mathcal{E}_0$, we have that as long as $p\ge\Omega\left(n\right)$, there exists constant $c_1>0$ such that

$$\left\|\mathbf{G}_{\mathcal{I}}\mathbf{q}_2\right\|_1 \le \frac{2\theta p}{\sqrt{p}}\left\|\mathbf{q}_2\right\|_2 = 2\theta\sqrt{p}\left\|\mathbf{q}_2\right\|_2 \text{ for all } \mathbf{q}_2\in\mathbb{R}^{n-1}, \tag{C.3}$$

$$\left\|\mathbf{G}_{\mathcal{I}^c}\mathbf{q}_2\right\|_1 \ge \frac{1}{2}\frac{p-2\theta p}{\sqrt{p}}\left\|\mathbf{q}_2\right\|_2 = \frac{1}{2}\sqrt{p}\left(1-2\theta\right)\left\|\mathbf{q}_2\right\|_2 \text{ for all } \mathbf{q}_2\in\mathbb{R}^{n-1}, \tag{C.4}$$

hold with probability at least $1-2\exp\left(-c_1 n_2\right)-2\exp\left(-p\theta^2/2\right)$. Moreover, by Lemma B.3,

$$\left\|\mathbf{G}-\mathbf{G}'\right\|_{\ell^2\to\ell^1} \le 8\sqrt{n} \tag{C.5}$$

holds with probability at least $1-c_2\exp\left(-c_3 n\right)-2\exp\left(-p\theta^2/2\right)$ when $p\ge\Omega\left(n\right)$. So we obtain that

$$\left\|\mathbf{Y}'\mathbf{q}\right\|_1 \ge \left|q_1\right|\left\|\frac{\mathbf{x}_0}{\left\|\mathbf{x}_0\right\|_2}\right\|_1 + \left\|\mathbf{q}_2\right\|_2\left(\frac{1}{2}\sqrt{p}\left(1-2\theta\right) - 2\theta\sqrt{p} - 8\sqrt{n}\right) \tag{C.6}$$

holds with probability at least $1-c_4\exp\left(-c_5 n\right)-2\exp\left(-p\theta^2/2\right)$ for some positive c_4 and c_5. Assuming $\mathcal{E}_0$, we observe

$$\left\|\frac{\mathbf{x}_0}{\left\|\mathbf{x}_0\right\|_2}\right\|_1 \le \sqrt{\left|\mathcal{I}\right|}\left\|\frac{\mathbf{x}_0}{\left\|\mathbf{x}_0\right\|_2}\right\|_2 \le \sqrt{2\theta p}. \tag{C.7}$$

So in order to minimize the objective $\|\mathbf{Y}'\mathbf{q}\|_1$ at $\mathbf{e}_1$ or $-\mathbf{e}_1$, i.e., $q_1 = 1$, subject to the constraint $q_1^2 + \|\mathbf{q}_2\|_2^2 = 1$, it suffices to have

$$\sqrt{2\theta p} < \frac{1}{2}\sqrt{p}\,(1-2\theta) - 2\theta\sqrt{p} - 8\sqrt{n}, \tag{C.8}$$

which is satisfied when θ is sufficiently small. Thus there exists a universal constant $\theta_0 > 0$, such that for all $1/\sqrt{n} \le \theta \le \theta_0$, when $p \ge \Omega(n)$, $\pm\mathbf{e}_1$ are the global minimizers of (2.2) with probability at least $1 - c_4\exp(-c_5 n) - 2\exp\left(-p\theta^2/2\right)$, if the input basis is $\mathbf{Y}'$. As $\theta > 1/\sqrt{n}$ by assumption, from (B.4), to make the above probability high, it is enough to make $p \ge \Omega(n\log n)$.

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

$$\min_{\mathbf{q}\in\mathbb{R}^n} \|\mathbf{Y}'\mathbf{R}\mathbf{q}\|_1, \qquad \text{s.t. } \|\mathbf{q}\|_2 = 1, \tag{C.9}$$

which is equivalent to

$$\min_{\mathbf{q}\in\mathbb{R}^n} \|\mathbf{Y}'\mathbf{R}\mathbf{q}\|_1, \qquad \text{s.t. } \|\mathbf{R}\mathbf{q}\|_2 = 1, \tag{C.10}$$

which is obviously equivalent to the canonical program we analyze above by a simple change of variable, i.e., $\overline{\mathbf{q}} \doteq \mathbf{R}\mathbf{q}$, completing the proof. □

D Proof of Main Result

In this appendix, we prove our main result in Theorem 4.1. In particular, we will first show that when the $\mathbf{Y}'$ defined in (B.3) is the input orthonormal basis, the "initialization + ADM + LP rounding" pipeline recovers $\mathbf{x}_0$ under the stated technical conditions. Then we will upgrade the recovery result to all orthonormal basis by observing that all three stages are "invariant" to the input orthonormal basis $\mathbf{Y}$.

Keep the notation in Section 3, let $\mathbf{y}^1, \cdots, \mathbf{y}^p$ be the transpose of the rows of $\mathbf{Y}$, and let $\mathbf{y}'^1, \cdots, \mathbf{y}'^p$ be the transpose of the rows of $\mathbf{Y}'$. For $\mathbf{q} \in \mathbb{S}^{n-1}$, set

$$\mathbf{Q}(\mathbf{q}) = \frac{1}{p}\sum_{k=1}^{p} \mathbf{y}^k S_\lambda\left[\mathbf{q}^\top \mathbf{y}^k\right], \tag{D.1}$$

$$\mathbf{Q}'(\mathbf{q}) = \frac{1}{p}\sum_{k=1}^{p} \mathbf{y}'^k S_\lambda\left[\mathbf{q}^\top \mathbf{y}'^k\right]. \tag{D.2}$$

Further, we write $\mathbf{Q}(\mathbf{q}) = [Q_1(\mathbf{q}); \mathbf{Q}_2(\mathbf{q})]$, where $Q_1(\mathbf{q})$ is the first coordinate, and define similar notations for $\mathbf{Q}'(\mathbf{q})$. In addition, for any $k = 1, \cdots, p$, set

$$X_k^1(Z_k) = x_{0k} S_\lambda\left[\mathbf{q}^\top \mathbf{y}^k\right] = x_{0k} S_\lambda\left[x_{0k}q_1 + Z_k\right], \tag{D.3}$$

$$\mathbf{X}_k^2(Z_k) = \mathbf{g}^k S_\lambda\left[\mathbf{q}^\top \mathbf{y}^k\right] = \mathbf{g}^k S_\lambda\left[x_{0k}q_1 + Z_k\right], \tag{D.4}$$

where $Z_k = \mathbf{q}_2^\top \mathbf{g}^k \sim \mathcal{N}(0, \sigma^2)$ for $\sigma = \|\mathbf{q}_2\|_2/\sqrt{p}$, and x_{0k} denotes the k-th coordinate of $\mathbf{x}_0$. Hence we obviously have

$$Q_1 = \frac{1}{p}\sum_{k=1}^{p} X_k^1, \quad \mathbf{Q}_2 = \frac{1}{p}\sum_{k=1}^{p} \mathbf{X}_k^2, . \tag{D.5}$$

Next we sketch the main technical pieces for establishing the recovery results for $\mathbf{Y}'$ first. All detailed proofs are deferred to later sections of the appendix. We will assume $\exp(n/2)/2 \ge p \ge Cn^4\log n$ for some large constant C for all the subsequent claims.

1. **Good initialization**. Proposition E.1 in Appendix E shows that with overwhelming probability, at least one of our p initialization vectors suggested in Section 3, say $\mathbf{q}_i^{(0)} = \mathbf{y}'^i$, obeys that

$$\left| \left\langle \frac{\mathbf{y}'^i}{\|\mathbf{y}'^i\|_2}, \mathbf{e}_1 \right\rangle \right| \geq \frac{1}{4\sqrt{\theta n}}. \tag{D.6}$$

2. **Uniform progress away from the equator**. By Proposition F.1 in Appendix F, there exists some constant $\theta_0 > 0$, such that for any $\theta \in \left(\frac{1}{\sqrt{n}}, \theta_0\right)$,

$$G'(\mathbf{q}) = \frac{|Q_1'(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2'(\mathbf{q})\|_2}{\|\mathbf{q}\|_2} \ \geq \ \frac{1}{4000\theta^2 np} \tag{D.7}$$

holds uniformly for all $\mathbf{q} \in \mathbb{S}^{n-1}$ in the region $\frac{1}{4\sqrt{\theta n}} \leq |q_1| \leq 3\sqrt{\theta}$ with overwhelming probability.

3. **No jumps away from the cap**. Proposition G.1 in Appendix G shows that for any $\theta \in \left(\frac{1}{\sqrt{n}}, \theta_0\right)$, with overwhelming probability,

$$\frac{Q_1'(\mathbf{q})}{\|\mathbf{Q}'(\mathbf{q})\|_2} \ \geq \ 2\sqrt{\theta} \tag{D.8}$$

holds for all $\mathbf{q}$ with $|q_1| \geq 3\sqrt{\theta}$.

4. **Location of the stationary/stopping point**. The first point above ensures that with overwhelming probability at least one starting point $\mathbf{q}^{(0)}$ will satisfy $\left|q_1^{(0)}\right| \geq \frac{1}{4\sqrt{\theta n}}$. As shown in Appendix H, the strictly positive gap of the second point ensures that one needs to run at most $O\left(n^4 \log n\right)$ iterations to first encounter an iterate $\mathbf{q}^{(k)}$ such that $\left|q_1^{(k)}\right| \geq 3\sqrt{\theta}$. The third point suggests extra iterations will not move away from the cap area, and hence the stationary point $\overline{\mathbf{q}}$ of the ADM algorithm will satisfy $|\overline{q}_1| \geq 2\sqrt{\theta}$. If one enforces a hard stop after $O\left(n^4 \log n\right)$ iterations, the stopping point will similarly stay in the region $|q_1| \geq 2\sqrt{\theta}$.

5. **LP Rounding succeeds**. We know that in the LP rounding stage, described in Section 3, will receive a vector $\mathbf{r} = \bar{\mathbf{q}}$ with its first coordinate $|r_1| \geq 2\sqrt{\theta}$. Proposition I.1 in Appendix I proves that with overwhelming probability, the LP rounding (3.6) (operated on $\mathbf{Y}'$) will output a solution $\mathbf{q}^\star = \mathbf{e}_1$.

In summary, our ADM algorithm in Algorithm 1 using a smart initialization, plus an LP rounding stage (3.6), will output $\mathbf{q}^\star = \pm\mathbf{e}_1$ with overwhelming probability, or $\mathbf{Y}'\mathbf{q}^\star$ as a nontrivial scaled version of $\mathbf{x}_0$.

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

- Discussion at the end of Appendix E suggests with overwhelming probability, at least one row of $\widehat{\mathbf{Y}}$ provides an initial point $\mathbf{q}^{(0)}$ such that $\left|\left\langle \mathbf{q}^{(0)}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right| \geq \frac{1}{4\sqrt{\theta n}}$.
- Discussion following Proposition F.1 in Appendix F suggests that for all $\mathbf{q}$ such that $\frac{1}{4\sqrt{\theta n}} \leq \left|\left\langle \mathbf{q}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right| \leq 3\sqrt{\theta}$, there is a strictly positive gap, indicating steady progress towards a point $\mathbf{q}^{(k)}$ such that $\left|\left\langle \mathbf{q}^{(k)}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right| \geq 3\sqrt{\theta}$.
- Discussion at the end of Appendix G indicates once $\mathbf{q}$ satisfying $\left|\left\langle \mathbf{q}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right|$, the next iterate will not move far away from the target:

$$\frac{\left\langle \mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right), \mathbf{R}^\top \mathbf{e}_1 \right\rangle}{\left\| \mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right) \right\|_2} \geq 2\sqrt{\theta}. \tag{D.9}$$

- Repeating the argument in Appendix H for general input $\widehat{\mathbf{Y}}$ shows it is enough to run the ADM algorithm $O\left(n^4 \log n\right)$ iterations to cross the range $\frac{1}{4\sqrt{\theta n}} \leq \left|\left\langle \mathbf{q}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right| \leq 3\sqrt{\theta}$. So the above three points together dictates that with the proposed initialization, with overwhelming probability, we finally obtain a point $\overline{\mathbf{q}}$ that satisfies $\left|\left\langle \overline{\mathbf{q}}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right| \geq 2\sqrt{\theta}$, if we run at least $O\left(n^4 \log n\right)$ iterations.
- Since the ADM returns $\overline{\mathbf{q}}$ satisfying $\left|\left\langle \overline{\mathbf{q}}, \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right| \geq 2\sqrt{\theta}$, discussion at the end of Appendix I dictates that we will obtain $\mathbf{q}^\star = \mathbf{R}^\top \mathbf{e}_1$ as the optimizer of the rounding program, exactly the target solution.

We complete the proof.

E Good Initialization

Proposition E.1. *Let $\mathbf{y}'^k$ for $k = 1, \ldots, p$ be the transpose of the rows of the orthonormal bases $\mathbf{Y}'$ defined in* (B.3)*. If $\theta > 1/\sqrt{n}$ and $\exp(n/2)/2 \geq p \geq Cn^2$ for some constant $C > 0$, it holds that at least one of our p initialization vectors suggested in Section 3, say $\mathbf{q}_i^{(0)} = \mathbf{y}'^i$, obeys*

$$\left|\left\langle \frac{\mathbf{y}'^i}{\left\|\mathbf{y}'^i\right\|_2}, \mathbf{e}_1 \right\rangle\right| \geq \frac{1}{4\sqrt{\theta n}}, \tag{E.1}$$

with probability at least $1 - c' \exp(-c''n)$ for some positive constants c' and c''.

Proof. Since $\mathbf{x}_0$ is i.i.d. Bernoulli, with probability at least $1-(1-\theta)^p \geq 1-\exp(-\theta p)$, at least one component of $\mathbf{x}_0$ is nonzero. Without loss of generality (w.l.o.g.), assume the k-th component of $\mathbf{x}_0$ is nonzero. Then $x_{0k} = \frac{1}{\sqrt{\theta p}}$, and

$$|q_{i1}| \;=\; \frac{\frac{1}{\sqrt{\theta p}}}{\left\|\mathbf{x}_0\right\|_2 \left\|\mathbf{y}'^i\right\|_2} \geq \frac{\frac{1}{\sqrt{\theta p}}}{\left\|\mathbf{x}_0\right\|_2 \left(\left\|\mathbf{x}_0/\left\|\mathbf{x}_0\right\|_2\right\|_{\ell^2\to\ell^\infty} + \left\|\mathbf{G}'\right\|_{\ell^2\to\ell^\infty}\right)} = \frac{1}{\sqrt{\theta p}\left\|\mathbf{x}_0\right\|_\infty + \left\|\mathbf{x}_0\right\|_2 \left\|\mathbf{G}'\right\|_{\ell^2\to\ell^\infty}}. \tag{E.2}$$

We know that with probability at least $1 - \exp(-p\theta^2/2)$, it holds that

$$\left\|\mathbf{x}_0\right\|_2 = \sqrt{|\mathcal{I}| \times \frac{1}{\theta p}} \leq \sqrt{2\theta p \times \frac{1}{\theta p}} = \sqrt{2}. \tag{E.3}$$

Moreover, using Lemma B.4 , and Lemma A.15 with $\varepsilon = 1/16$ and $\xi = 1/2$, we know when $p \geq C_1 n$ for some large $C_1 > 0$, it holds that (note that $\|\mathbf{M}\|$ can be arbitrarily close to 1 for large C_1 in Lemma B.2)

$$\left\|\mathbf{G}'\right\|_{\ell^2\to\ell^\infty} \leq \left\|\mathbf{G}\right\|_{\ell^2\to\ell^\infty} \|\mathbf{M}\| + \frac{4\sqrt{2n}}{\sqrt{\theta}p} \leq \frac{9}{5}\sqrt{\frac{n}{p}} + \frac{4\sqrt{2n}}{\sqrt{\theta}p} \leq 2\sqrt{\frac{n}{p}} \tag{E.4}$$

with probability at least $1 - c_1 \exp(-c_2 n)$ for some positive constants c_1 and c_2. Therefore with probability at least $1 - \exp(-\theta p) - \exp\left(-\theta^2 p\right) - c_1 \exp(-c_2 n)$, it holds that

$$|q_{i1}| \geq \frac{1}{1 + \sqrt{2\theta p} \times 2\sqrt{\frac{n}{p}}} = \frac{1}{1 + 2\sqrt{2}\sqrt{\theta n}}. \tag{E.5}$$

Using the fact that $\theta \geq 1/\sqrt{n}$, we obtain $|q_{i1}| \geq \frac{1}{\left(1+2\sqrt{2}\right)\sqrt{\theta n}}$. It is sufficient to set $p \geq C_2 n^2$ for some large enough $C_2 > 0$ to make the probability overwhelming, as desired. □

We will next show that for an arbitrary orthonormal basis $\widehat{\mathbf{Y}} \doteq \mathbf{Y}'\mathbf{R}$ the initialization still biases towards the target solution. To see this, suppose w.l.o.g. $\left(\mathbf{y}'^i\right)^\top$ is a row of $\mathbf{Y}'$ with nonzero first coordinate. We have

shown above that with high probability $\left|\left\langle \frac{\mathbf{y}'^i}{\|\mathbf{y}'^i\|_2}, \mathbf{e}_1 \right\rangle\right| \geq \frac{1}{4\sqrt{\theta n}}$ if $\mathbf{Y}'$ is the input orthonormal basis. For $\mathbf{Y}'$, as $\mathbf{x}_0 = \mathbf{Y}'\mathbf{e}_1 = \mathbf{Y}'\mathbf{R}\mathbf{R}^\top\mathbf{e}_1$, we know $\mathbf{q}^\star = \mathbf{R}^\top\mathbf{e}_1$ is the target solution corresponding to $\widehat{\mathbf{Y}}$. Observing that

$$\left|\left\langle \mathbf{R}^\top\mathbf{e}_1, \frac{\left(\mathbf{e}_i^\top\widehat{\mathbf{Y}}\right)^\top}{\left\|\left(\mathbf{e}_i^\top\widehat{\mathbf{Y}}\right)^\top\right\|_2} \right\rangle\right| = \left|\left\langle \mathbf{R}^\top\mathbf{e}_1, \frac{\mathbf{R}^\top\left(\mathbf{Y}'\right)^\top\mathbf{e}_i}{\left\|\mathbf{R}^\top\left(\mathbf{Y}'\right)^\top\mathbf{e}_i\right\|_2} \right\rangle\right| = \left|\left\langle \mathbf{e}_1, \frac{\left(\mathbf{Y}'\right)^\top\mathbf{e}_i}{\left\|\left(\mathbf{Y}'\right)^\top\mathbf{e}_i\right\|_2} \right\rangle\right| = \left|\left\langle \mathbf{e}_1, \frac{\mathbf{y}'^i}{\|\mathbf{y}'^i\|_2} \right\rangle\right| \geq \frac{1}{4\sqrt{n\theta}}, \tag{E.6}$$

corroborating our claim.

F Lower Bounding Finite Sample Gap $G'(\mathbf{q})$

We will first work with the "canonical" orthonormal basis $\mathbf{Y}'$. The task is to lower bound the gap for finite samples

$$G'(\mathbf{q}) = \frac{|Q_1'(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2'(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2}. \tag{F.1}$$

Since we can deterministically constrain $|q_1|$ and $\|\mathbf{q}_2\|_2$ (e.g., $\frac{1}{4\sqrt{n\theta}} \leq |q_1| \leq 3\sqrt{\theta}$ and $\|\mathbf{q}_2\|_2 \geq \frac{1}{4}$, where the choice of $\frac{1}{4}$ 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 depends on the orthonormal basis $\mathbf{Y}'$. It turns out to cook up a typical expectation-concentration style argument, the unnormalized basis $\mathbf{Y}$ is much easier to work with than $\mathbf{Y}'$. Hence our proof will follow the observation that

$$|Q_1'(\mathbf{q})| \geq |\mathbb{E}[Q_1(\mathbf{q})]| - |Q_1(\mathbf{q}) - \mathbb{E}[Q_1(\mathbf{q})]| - |Q_1'(\mathbf{q}) - Q_1(\mathbf{q})|, \tag{F.2}$$

$$\|\mathbf{Q}_2'(\mathbf{q})\| \leq \|\mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2 + \|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2 + \|\mathbf{Q}_2'(\mathbf{q}) - \mathbf{Q}_2(\mathbf{q})\|_2. \tag{F.3}$$

In particular, define the set $\Gamma = \left\{\mathbf{q} \in \mathbb{S}^{n-1} \,:\, \frac{1}{4\sqrt{n\theta}} \leq |q_1| \leq 3\sqrt{\theta}, \|\mathbf{q}_2\|_2 \geq \frac{1}{4}\right\}$:

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

$$G(\mathbf{q}) \doteq \frac{|\mathbb{E}[Q_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \geq \frac{1}{50}\frac{q_1^2}{\theta p}. \tag{F.4}$$

  As $|q_1| \geq \frac{1}{4\sqrt{n\theta}}$, we have

$$\inf_{\mathbf{q}\in\Gamma} \frac{|\mathbb{E}[Q_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \geq \frac{1}{800}\frac{1}{\theta^2 np}. \tag{F.5}$$

- Appendix F.2, as summarized in Proposition F.9, shows that whenever $\exp(n) \geq p \geq \Omega\left(n^4 \log n\right)$, it holds with overwhelmingly probability that

$$\begin{aligned} &\sup_{\mathbf{q}\in\Gamma} \frac{|Q_1(\mathbf{q}) - \mathbb{E}[Q_1(\mathbf{q})]|}{|q_1|} + \frac{\|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \\ \leq\ &\frac{4\sqrt{\theta n}}{16000\theta^{5/2}n^{3/2}p} + \frac{4}{16000\theta^2 np} = \frac{1}{2000\theta^2 np}. \end{aligned} \tag{F.6}$$

- Appendix F.4 shows that whenever $\exp(n/2)/2 \geq p \geq \Omega\left(n^4 \log n\right)$, it holds with overwhelmingly probability that

$$\begin{aligned} &\sup_{\mathbf{q}\in\Gamma} \frac{|Q_1(\mathbf{q}) - Q_1'(\mathbf{q})|}{|q_1|} + \frac{\|\mathbf{Q}_2(\mathbf{q}) - \mathbf{Q}_2'(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2} \\ &\leq \frac{4\sqrt{\theta n}}{16000\theta^{5/2}n^{3/2}p} + \frac{4}{16000\theta^2 np} = \frac{1}{2000\theta^2 np}. \end{aligned} \tag{F.7}$$

Observing that

$$\begin{aligned} \inf_{\mathbf{q}\in\Gamma} G'(\mathbf{q}) \geq \inf_{\mathbf{q}\in\Gamma} &\left(\frac{|\mathbb{E}[Q_1(\mathbf{q})]|}{|q_1|} - \frac{\|\mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2}{\|\boldsymbol{q}_2\|_2} \right) - \sup_{\mathbf{q}\in\Gamma} \left(\frac{|Q_1(\mathbf{q}) - \mathbb{E}[Q_1(\mathbf{q})]|}{|q_1|} + \frac{\|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2}{\|\mathbf{q}_2\|_2} \right) \\ &- \sup_{\mathbf{q}\in\Gamma} \left(\frac{|Q_1(\mathbf{q}) - Q_1'(\mathbf{q})|}{|q_1|} + \frac{\|\mathbf{Q}_2(\mathbf{q}) - \mathbf{Q}_2'(\mathbf{q})\|_2}{\|\mathbf{q}_2\|_2} \right), \end{aligned} \tag{F.8}$$

we obtain the following:

Proposition F.1. *There exists some constant $\theta_0 > 0$ such that, for all $\theta \in \left(\frac{1}{\sqrt{n}}, \theta_0\right)$, when $\exp(n/2)/2 \geq p \geq Cn^4 \log n$ for some large constant $C > 0$, we have*

$$\inf_{\mathbf{q}\in\Gamma} G'(\mathbf{q}) \;\geq\; \frac{1}{4000\theta^2 np},, \tag{F.9}$$

with probability at least $1 - c'\exp(-c''n)$ for some positive constants c' and c''.

For the general case when the input orthonormal basis is $\widehat{\mathbf{Y}} = \mathbf{Y}'\mathbf{R}$ with target solution $\mathbf{q}^\star = \mathbf{R}^\top \mathbf{e}_1$, a straightforward extension of the definition for the gap would be:

$$G'\left(\mathbf{q}; \widehat{\mathbf{Y}} = \mathbf{Y}'\mathbf{R}\right) \doteq \frac{\left|\left\langle \mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right), \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right|}{|\langle \mathbf{q}, \mathbf{R}^\top \mathbf{e}_1\rangle|} - \frac{\left\|\left(\mathbf{I} - \mathbf{R}^\top \mathbf{e}_1 \mathbf{e}_1^\top \mathbf{R}\right) \mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right)\right\|_2}{\left\|\left(\mathbf{I} - \mathbf{R}^\top \mathbf{e}_1 \mathbf{e}_1^\top \mathbf{R}\right) \mathbf{q}\right\|_2}. \tag{F.10}$$

Since $\mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right) = \frac{1}{p}\sum_{k=1}^p \mathbf{R}^\top \mathbf{y}^k S_\lambda\left(\mathbf{q}^\top \mathbf{R}^\top \mathbf{y}^k\right)$, we have

$$\mathbf{R}\mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right) = \frac{1}{p}\sum_{k=1}^p \mathbf{R}\mathbf{R}^\top \mathbf{y}'^k S_\lambda\left(\mathbf{q}^\top \mathbf{R}^\top \mathbf{y}'^k\right) = \frac{1}{p}\sum_{k=1}^p \mathbf{y}'^k S_\lambda\left[(\mathbf{R}\mathbf{q})^\top \mathbf{y}'^k\right] = \mathbf{Q}'(\mathbf{R}\mathbf{q}; \mathbf{Y}'). \tag{F.11}$$

Hence we have

$$G'\left(\mathbf{q}; \widehat{\mathbf{Y}} = \mathbf{Y}'\mathbf{R}\right) = \frac{|\langle \mathbf{Q}'(\mathbf{R}\mathbf{q}; \mathbf{Y}'), \mathbf{e}_1\rangle|}{|\langle \mathbf{R}\mathbf{q}, \mathbf{e}_1\rangle|} - \frac{\left\|\left(\mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^\top\right) \mathbf{Q}'(\mathbf{R}\mathbf{q}; \mathbf{Y}')\right\|_2}{\left\|\left(\mathbf{I} - \mathbf{e}_1 \mathbf{e}_1^\top\right) \mathbf{R}\mathbf{q}\right\|_2}. \tag{F.12}$$

Therefore, from Proposition F.1 above, we conclude that under the same technical conditions as therein,

$$\inf_{\mathbf{q}\in\mathbb{S}^{n-1}: \frac{1}{4\sqrt{\theta n}} \leq |\langle \mathbf{R}\mathbf{q}, \mathbf{e}_1\rangle| \leq 3\sqrt{\theta}} G'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right) \geq \frac{1}{4000\theta^2 np} \tag{F.13}$$

with overwhelmingly probability.

F.1 Lower Bounding the Expected Gap $G(\mathbf{q})$

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

$$G(\mathbf{q}) = \frac{|\mathbb{E}\left[Q_1(\mathbf{q})\right]|}{|q_1|} - \frac{\|\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right]\|_2}{\|\mathbf{q}_2\|_2}. \tag{F.14}$$

More specifically, we show that:

Proposition F.2. *There exists some numerical constant $\theta_0 > 0$, such that for all $\theta \in (0, \theta_0)$, it holds that*

$$G(\mathbf{q}) \geq \frac{1}{50}\frac{q_1^2}{\theta p} \tag{F.15}$$

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

Because the estimation for the gap $G(\mathbf{q})$ involves dedicated estimation for $\mathbb{E}\left[Q_1(\mathbf{q})\right]$ and $\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right]$, we sketch the main proof in Appendix F.1.1, and leave those detailed technical calculations in the subsequent subsections.

F.1.1 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 (D.5), (D.3) and (D.4), we have

$$\mathbb{E}\left[Q_1(\mathbf{q})\right] = \mathbb{E}\left[x_0 S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right], \tag{F.16}$$

$$\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right] = \mathbb{E}\left[\mathbf{g} S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right], \tag{F.17}$$

where $\mathbf{q} = \left[q_1, \mathbf{q}_2^\top\right]^\top$, $\mathbf{g} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{p}\mathbf{I}\right)$, and $x_0 \sim \frac{1}{\sqrt{\theta p}}\mathrm{Ber}(\theta)$. Let us decompose

$$\mathbf{g} = \mathbf{g}_\parallel + \mathbf{g}_\perp, \tag{F.18}$$

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 $\mathbf{g}_\perp = (\mathbf{I} - \mathcal{P}_\parallel)\mathbf{g}$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right] &= \mathbb{E}\left[\mathbf{g}_\parallel S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}_\parallel\right]\right] + \mathbb{E}\left[\mathbf{g}_\perp S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}_\parallel\right]\right] \\
&= \mathbb{E}\left[\mathbf{g}_\parallel S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right] + \mathbb{E}\left[\mathbf{g}_\perp\right]\mathbb{E}\left[S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right] \\
&= \frac{\mathbf{q}_2}{\|\mathbf{q}_2\|_2^2}\mathbb{E}\left[\mathbf{q}_2^\top \mathbf{g} S_\lambda\left[x_0 q_1 + \mathbf{q}_2^\top \mathbf{g}\right]\right],
\end{aligned} \tag{F.19}$$

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 $\mathbb{E}\left[\mathbf{g}_\perp\right] = \mathbf{0}$. 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

$$\mathbb{E}\left[Q_1(\mathbf{q})\right] = \sqrt{\frac{\theta}{p}}\mathbb{E}\left[S_\lambda\left[\frac{q_1}{\sqrt{\theta p}} + Z\right]\right], \tag{F.20}$$

$$\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right] = \frac{\theta \mathbf{q}_2}{\|\mathbf{q}_2\|_2^2}\mathbb{E}\left[Z S_\lambda\left[\frac{q_1}{\sqrt{\theta p}} + Z\right]\right] + \frac{(1-\theta)\mathbf{q}_2}{\|\mathbf{q}_2\|_2^2}\mathbb{E}\left[Z S_\lambda\left[Z\right]\right]. \tag{F.21}$$

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

$$\mathbb{E}\left[Q_1(\mathbf{q})\right] = \sqrt{\frac{\theta}{p}}\left\{\left[\alpha\Psi\left(-\frac{\alpha}{\sigma}\right) + \beta\Psi\left(\frac{\beta}{\sigma}\right)\right] + \sigma\left[\psi\left(-\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right)\right]\right\}, \tag{F.22}$$

$$\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right] = \left\{\frac{2\left(1-\theta\right)}{p}\Psi\left(-\frac{\lambda}{\sigma}\right) + \frac{\theta}{p}\left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right)\right]\right\}\mathbf{q}_2, \tag{F.23}$$

where the scalars α and β are defined as

$$\alpha = \frac{q_1}{\sqrt{\theta p}} + \lambda, \qquad \beta = \frac{q_1}{\sqrt{\theta p}} - \lambda, \tag{F.24}$$

and $\psi(t)$ and $\Psi(t)$ are *pdf* and *cdf* for standard normal distribution, respectively, as defined in Lemma A.1. Plugging (F.22) and (F.23) into (F.14), by some simplifications, we obtain

$$\begin{aligned} G(\mathbf{q}) \ =& \frac{1}{q_1}\sqrt{\frac{\theta}{p}}\left[\alpha\Psi\left(-\frac{\alpha}{\sigma}\right) + \beta\Psi\left(\frac{\beta}{\sigma}\right) - \frac{2q_1}{\sqrt{\theta p}}\Psi\left(-\frac{\lambda}{\sigma}\right)\right] - \frac{\theta}{p}\left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) - 2\Psi\left(-\frac{\lambda}{\sigma}\right)\right] \\ &+ \frac{\sigma}{q_1}\sqrt{\frac{\theta}{p}}\left[\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right)\right]. \end{aligned} \tag{F.25}$$

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

$$-\frac{\alpha}{\sigma} \ = \ -\frac{\delta+1}{\sqrt{1-q_1^2}}, \qquad \frac{\beta}{\sigma} \ = \ \frac{\delta-1}{\sqrt{1-q_1^2}}, \qquad \frac{\lambda}{\sigma} \ = \ \frac{1}{\sqrt{1-q_1^2}}, \tag{F.26}$$

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

$$\begin{aligned} G(\mathbf{q}) \ \ge& \ \frac{1-\theta}{p}\Phi_1(\delta) - \frac{1}{\delta p}\Phi_2(\delta) + \frac{1-\theta}{p}\psi(-1)q_1^2 + \frac{1}{p}\left(\sigma\sqrt{p} + \frac{\theta}{2} - 1\right)\eta_2(\delta)q_1^2 \\ &+ \frac{1}{2\delta p}\left[1+\delta^2-\theta\delta^2-\sigma\left(1+\delta^2\right)\sqrt{p}\right]q_1^2\eta_1(\delta) + \frac{\sigma}{\delta\sqrt{p}}\eta_1(\delta) - \frac{5C_T\sqrt{\theta}q_1^3}{p}(\delta+1)^3, \end{aligned} \tag{F.27}$$

where we define

$$\Phi_1(\delta) \ = \ \Psi(-1-\delta) + \Psi(-1+\delta) - 2\Psi(-1), \qquad \Phi_2(\delta) \ = \ \Psi(-1+\delta) - \Psi(-1-\delta), \tag{F.28}$$

$$\eta_1(\delta) \ = \ \psi(-1+\delta) - \psi(-1-\delta), \qquad \eta_2(\delta) \ = \ \psi(-1+\delta) + \psi(-1-\delta), \tag{F.29}$$

and C_T is as defined in Lemma F.3. Since $1-\sigma\sqrt{p} \ge 0$, dropping those small positive terms $\frac{q_1^2}{p}(1-\theta)\psi(-1)$, $\frac{\theta q_1^2}{2p}\eta_2(\delta)$, and $\left(1+\delta^2\right)\left(1-\sigma\sqrt{p}\right)q_1^2\eta_1(\delta)/(2\delta p)$, and using the fact that $\delta = q_1/\sqrt{\theta}$, we obtain

$$\begin{aligned} G(\mathbf{q}) &\ge \frac{1-\theta}{p}\Phi_1(\delta) - \frac{1}{\delta p}\left[\Phi_2(\delta) - \sigma\sqrt{p}\eta_1(\delta)\right] - \frac{q_1^2}{p}\left(1-\sigma\sqrt{p}\right)\eta_2(\delta) - \frac{\sqrt{\theta}}{2p}q_1^3\eta_1(\delta) - \frac{C_1\sqrt{\theta}q_1^3}{p}\max\left(\frac{q_1^3}{\theta^{3/2}}, 1\right) \\ &\ge \frac{1-\theta}{p}\Phi_1(\delta) - \frac{1}{\delta p}\left[\Phi_2(\delta) - \eta_1(\delta)\right] - \frac{q_1^2}{p}\frac{\eta_1(\delta)}{\delta} - \frac{q_1^2}{p\theta}\frac{2}{\sqrt{2\pi}}\theta - \frac{q_1^2}{\theta p}\frac{3\theta^2}{2\sqrt{2\pi}} - \frac{q_1^2}{\theta p}\left(C_1\theta^2\right), \end{aligned} \tag{F.30}$$

for some constant $C_1 > 0$, where we have used $q_1 \le 3\sqrt{\theta}$ to simplify the bounds and the fact $\sigma\sqrt{p} = \sqrt{1-q_1^2} \ge 1-q_1^2$ to simplify the expression. Substituting the estimates in Lemma F.5 and use the fact $\delta \mapsto \eta_1(\delta)/\delta$ is bounded, we obtain

$$G(p) \ge \frac{1}{p}\left(\frac{1}{40} - \frac{1}{2\sqrt{2\pi}}\theta\right)\delta^2 - \frac{q_1^2}{\theta p}\left(c_1\theta + c_2\theta^2\right) \tag{F.31}$$

$$\ge \frac{q_1^2}{\theta p}\left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta - c_1\theta - c_2\theta^2\right) \tag{F.32}$$

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

F.1.2 Auxiliary Results Used in the Proof

Lemma F.3. *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 $q_1 \in \left(0, \frac{1}{2}\right)$:*

$$\left|\psi\left(-\frac{\alpha}{\sigma}\right) - \left[1 - \frac{1}{2}(1+\delta)^2 q_1^2\right]\psi(-1-\delta)\right| \;\le\; C_T\,(1+\delta)^2\, q_1^4, \tag{F.33}$$

$$\left|\psi\left(\frac{\beta}{\sigma}\right) - \left[1 - \frac{1}{2}(\delta-1)^2 q_1^2\right]\psi(\delta-1)\right| \;\le\; C_T\,(\delta-1)^2\, q_1^4, \tag{F.34}$$

$$\left|\Psi\left(-\frac{\alpha}{\sigma}\right) - \left[\Psi(-1-\delta) - \frac{1}{2}\psi(-1-\delta)(1+\delta)q_1^2\right]\right| \;\le\; C_T\,(1+\delta)^2\, q_1^4, \tag{F.35}$$

$$\left|\Psi\left(\frac{\beta}{\sigma}\right) - \left[\Psi(\delta-1) + \frac{1}{2}\psi(\delta-1)(\delta-1)q_1^2\right]\right| \;\le\; C_T\,(\delta-1)^2\, q_1^4, \tag{F.36}$$

$$\left|\Psi\left(-\frac{\lambda}{\sigma}\right) - \left[\Psi(-1) - \frac{1}{2}\psi(-1)q_1^2\right]\right| \;\le\; C_T q_1^4. \tag{F.37}$$

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

$$0 \le \frac{1}{\sqrt{1-q_1^2}} - \left(1 + \frac{q_1^2}{2}\right) \le q_1^4. \tag{F.38}$$

Hence we have

$$-(1+\delta)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right) \le -\frac{\alpha}{\sigma} \le -(1+\delta)\left(1 + \frac{1}{2}q_1^2\right), \tag{F.39}$$

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

$$(\delta-1)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right) \le \frac{\beta}{\sigma} \le (\delta-1)\left(1 + \frac{1}{2}q_1^2\right), \text{ when } \delta \le 1. \tag{F.40}$$

So we have

$$\psi\left(-(1+\delta)\left(1 + \frac{1}{2}q_1^2 + q_1^4\right)\right) \le \psi\left(-\frac{\alpha}{\sigma}\right) \le \psi\left(-(1+\delta)\left(1 + \frac{1}{2}q_1^2\right)\right). \tag{F.41}$$

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

$$\left|\psi\left(-\frac{\alpha}{\sigma}\right) - \psi(-1-\delta) - \frac{1}{2}(1+\delta)^2 q_1^2 \psi(-1-\delta)\right| \;\le\; C_T\,(1+\delta)^2\, q_1^4, \tag{F.42}$$

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

Lemma F.4. *For any $\delta \in [0,3]$, it holds that*

$$\Phi_2(\delta) - \eta_1(\delta) \;\ge\; \frac{\eta_1(3)}{9}\delta^3 \ge \frac{1}{20}\delta^3. \tag{F.43}$$

Proof. Let us define

$$h(\delta) \;=\; \Phi_2(\delta) - \eta_1(\delta) - C\delta^3 \tag{F.44}$$

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

$$\frac{d}{d\delta}\Phi_1(\delta) = \eta_1(\delta), \quad \frac{d}{d\delta}\Phi_2(\delta) = \eta_2(\delta), \quad \frac{d}{d\delta}\eta_1(\delta) = \eta_2(\delta) - \delta\eta_1(\delta). \tag{F.45}$$

Thus, to show (F.43), 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 (F.45), it is sufficient to have

$$h'(\delta) = \delta\eta_1(\delta) - 3C\delta^2 \geq 0 \Longleftrightarrow \eta_1(\delta) \geq 3C\delta \tag{F.46}$$

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

$$p(\delta) \doteq \frac{\eta_1(\delta)}{3\delta} = \frac{1}{3\sqrt{2\pi}} \exp\left(-\frac{\delta^2+1}{2}\right) \frac{e^{\delta} - e^{-\delta}}{\delta}. \tag{F.47}$$

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

$$g(\delta) \doteq \left(e^{\delta} + e^{-\delta}\right)\delta - \left(\delta^2 + 1\right)\left(e^{\delta} - e^{-\delta}\right) \leq 0 \tag{F.48}$$

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

Lemma F.5. *For any $\delta \in [0, 3]$, we have*

$$(1-\theta)\Phi_1(\delta) - \frac{1}{\delta}\left[\Phi_2(\delta) - \eta_1(\delta)\right] \geq \left(\frac{1}{40} - \frac{1}{\sqrt{2\pi}}\theta\right)\delta^2. \tag{F.49}$$

Proof. Let us define

$$g(\delta) \;=\; (1-\theta)\Phi_1(\delta) - \frac{1}{\delta}\left[\Phi_2(\delta) - \eta_1(\delta)\right] - c_0(\theta)\,\delta^2, \tag{F.50}$$

where $c_0(\theta) > 0$ is a function of θ. Thus, by the results in (F.45) and L'Hospital's rule, we have

$$\lim_{\delta\to 0}\frac{\Phi_2(\delta)}{\delta} = \lim_{\delta\to 0}\eta_2(\delta) = 2\psi(-1), \quad \lim_{\delta\to 0}\frac{\eta_1(\delta)}{\delta} = \lim_{\delta\to 0}\left[\eta_2(\delta) - \delta\eta_1(\delta)\right] = 2\psi(-1). \tag{F.51}$$

Combined that with the fact that $\Phi_1(0) = 0$, we conclude $g(0) = 0$. Hence, to show (F.49), it is sufficient to show that $g'(\delta) \geq 0$ for all $\delta \in [0, 3]$. Direct calculation using the results in (F.45) shows that

$$g'(\delta) = \frac{1}{\delta^2}\left[\Phi_2(\delta) - \eta_1(\delta)\right] - \theta\eta_1(\delta) - 2c_0(\theta)\,\delta. \tag{F.52}$$

Since $\eta_1(\delta)/\delta$ is monotonically decreasing as shown in Lemma F.4, we have that for all $\delta \in (0, 3)$

$$\eta_1(\delta) \leq \delta \lim_{\delta\to 0}\frac{\eta(\delta)}{\delta} \leq \frac{2}{\sqrt{2\pi}}\delta. \tag{F.53}$$

Using the above bound and the main result from Lemma F.4 again, we obtain

$$g'(\delta) \;\geq\; \frac{1}{20}\delta - \frac{2}{\sqrt{2\pi}}\theta\delta - 2c_0\delta. \tag{F.54}$$

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

F.2 Finite Sample Concentration

In the following two subsections, we estimate the deviations around the expectations $\mathbb{E}Q_1$ and $\mathbb{E}\mathbf{Q}_2$, i.e., $|Q_1 - \mathbb{E}Q_1|$ and $\|\mathbf{Q}_2 - \mathbb{E}\mathbf{Q}_2\|_2$, and show that the total deviations fit into the gap $G(\mathbf{q})$ we derived in Appendix F.1. Our analysis is based on the scalar and vector Bernstein's inequalities with moment conditions. Finally, in Appendix F.3, we uniform the bound by applying the classical discretization argument.

F.2.1 Concentration for Q_1

Lemma F.6 (Bounding $|Q_1 - \mathbb{E}[Q_1(\mathbf{q})]|$)**.** *For each $\mathbf{q} \in \mathbb{S}^{n-1}$, it holds for all $t > 0$ that*

$$\mathbb{P}\left[|Q_1(\mathbf{q}) - \mathbb{E}[Q_1(\mathbf{q})]| \geq t\right] \leq 2\exp\left(-\frac{\theta p^3 t^2}{8 + 4pt}\right). \tag{F.55}$$

Proof. By (D.3) and (D.5), we know that

$$Q_1(\mathbf{q}) = \frac{1}{p}\sum_{k=1}^{p} X_k^1, \quad X_k^1 = x_{0k}\mathcal{S}_\lambda\left[x_{0k}q_1 + Z_k\right] \tag{F.56}$$

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

$$\begin{aligned}
\mathbb{E}\left[\left|X_k^1\right|^m\right] &\leq \theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \mathbb{E}\left[\left|\frac{q_1}{\sqrt{\theta p}} + Z_k\right|^m\right] \\
&= \theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \sum_{l=0}^{m}\binom{m}{l}\left(\frac{q_1}{\sqrt{\theta p}}\right)^l \mathbb{E}\left[|Z_k|^{m-l}\right] \\
&= \theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \sum_{l=0}^{m}\binom{m}{l}\left(\frac{q_1}{\sqrt{\theta p}}\right)^l (m-l-1)!!\left(\frac{\|\mathbf{q}_2\|_2}{\sqrt{p}}\right)^{m-l} \\
&\leq \frac{m!}{2}\theta\left(\frac{1}{\sqrt{\theta p}}\right)^m \left(\frac{q_1}{\sqrt{\theta p}} + \frac{\|\mathbf{q}_2\|_2}{\sqrt{p}}\right)^m \\
&\leq \frac{m!}{2}\theta\left(\frac{2}{\theta p}\right)^m = \frac{m!}{2}\frac{4}{\theta p^2}\left(\frac{2}{\theta p}\right)^{m-2}
\end{aligned} \tag{F.57}$$

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

$$\mathbb{P}\left[|Q_1(\mathbf{q}) - \mathbb{E}[Q_1(\mathbf{q})]| \geq t\right] \leq 2\exp\left(-\frac{\theta p^3 t^2}{8 + 4pt}\right). \tag{F.58}$$

as desired. □

F.2.2 Concentration for $\mathbf{Q}_2$

Lemma F.7 (Bounding $\|\mathbf{Q}_2 - \mathbb{E}[\mathbf{Q}_2]\|_2$)**.** *For each $\mathbf{q} \in \mathbb{S}^{n-1}$, it holds for all $t > 0$ that*

$$\mathbb{P}\left[\|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}[\mathbf{Q}_2(\mathbf{q})]\|_2 > t\right] \leq 2(n+1)\exp\left(-\frac{\theta p^3 t^2}{128n + 16\sqrt{\theta n}pt}\right). \tag{F.59}$$

Before proving Lemma F.7, we record the following useful results.

Lemma F.8. *For any positive integer $s, l > 0$, we have*

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^s \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l\right] \leq \frac{(l+s)!}{2}\|\mathbf{q}_2\|_2^l \frac{\left(2\sqrt{n}\right)^s}{\left(\sqrt{p}\right)^{s+l}} \tag{F.60}$$

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

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^l \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l\right] \leq \frac{l!}{2} \left\|\mathbf{q}_2\right\|_2^l \left(\frac{4\sqrt{n}}{p}\right)^l \tag{F.61}$$

Proof. Let $\mathcal{P}_{\mathbf{q}_2^\parallel} = \frac{\mathbf{q}_2\mathbf{q}_2^\top}{\|\mathbf{q}_2\|_2^2}$ 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

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^s \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l\right] &\leq \mathbb{E}\left[\left(\left\|\mathcal{P}_{\mathbf{q}_2^\parallel}\mathbf{g}^k\right\|_2 + \left\|\mathcal{P}_{\mathbf{q}_2^\perp}\mathbf{g}^k\right\|_2\right)^s \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l\right] \\
&= \sum_{i=0}^{s} \binom{s}{i} \mathbb{E}\left[\left\|\mathcal{P}_{\mathbf{q}_2^\perp}\mathbf{g}^k\right\|_2^i\right] \mathbb{E}\left[\left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l \left\|\mathcal{P}_{\mathbf{q}_2^\parallel}\mathbf{g}^k\right\|_2^{s-i}\right] \\
&= \sum_{i=0}^{s} \binom{s}{i} \mathbb{E}\left[\left\|\mathcal{P}_{\mathbf{q}_2^\perp}\mathbf{g}^k\right\|_2^i\right] \mathbb{E}\left[\left|\mathbf{q}_2^\top \mathbf{g}^k\right|^{l+s-i}\right] \frac{1}{\|\mathbf{q}_2\|_2^{s-i}} \\
&\leq \|\mathbf{q}_2\|_2^l \sum_{i=0}^{s} \binom{s}{i} \mathbb{E}\left[\left\|\mathcal{P}_{\mathbf{q}_2^\perp}\mathbf{g}^k\right\|_2^i\right] \left(\frac{1}{\sqrt{p}}\right)^{l+s-i} (l+s-i-1)!!.
\end{aligned} \tag{F.62}$$

Using Lemma A.5 and the fact that $\left\|\mathcal{P}_{\mathbf{q}_2^\perp}\mathbf{g}^k\right\|_2^2 \leq \left\|\mathbf{g}^k\right\|_2^2$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^s \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l\right] &\leq \|\mathbf{q}_2\|_2^l \sum_{i=0}^{s} \binom{s}{i} \left(\frac{\sqrt{n}}{\sqrt{p}}\right)^i i! \left(\frac{1}{\sqrt{p}}\right)^{l+s-i} (l+s-i-1)!! \\
&\leq \|\mathbf{q}_2\|_2^l \left(\frac{1}{\sqrt{p}}\right)^l \frac{(l+s)!}{2} \left(\frac{\sqrt{n}}{\sqrt{p}} + \frac{1}{\sqrt{p}}\right)^s \\
&\leq \frac{(l+s)!}{2} \|\mathbf{q}_2\|_2^l \frac{(2\sqrt{n})^s}{\left(\sqrt{p}\right)^{s+l}}.
\end{aligned} \tag{F.63}$$

□

Now, we are ready to prove Lemma F.7,

Proof. By (D.5) and (D.4), note that

$$\mathbf{Q}_2 = \frac{1}{p}\sum_{k=1}^{p} \mathbf{X}_k^2, \quad \mathbf{X}_k^2 = \mathbf{g}^k \mathcal{S}_\lambda\left[x_{0k}q_1 + Z_k\right] \tag{F.64}$$

where $Z_k = \mathbf{q}_2^\top \mathbf{g}^k$. Thus, for any $m \geq 2$, by Lemma F.8, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{X}_k^2\right\|_2^m\right] &\leq \theta\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^m \left|\frac{q_1}{\sqrt{\theta p}} + \mathbf{q}_2^\top \mathbf{g}^k\right|^m\right] + (1-\theta)\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^m \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^m\right] \\
&\leq \theta\sum_{l=0}^{m} \binom{m}{l} \mathbb{E}\left[\left|\mathbf{q}_2^\top \mathbf{g}^k\right|^l \left\|\mathbf{g}^k\right\|_2^m\right] \left|\frac{q_1}{\sqrt{\theta p}}\right|^{m-l} + (1-\theta)\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^m \left|\mathbf{q}_2^\top \mathbf{g}^k\right|^m\right] \\
&\leq \theta\left(\frac{2\sqrt{n}}{\sqrt{p}}\right)^m \sum_{l=0}^{m} \binom{m}{l} \frac{(m+l)!}{2} \left(\frac{\|\mathbf{q}_2\|_2}{\sqrt{p}}\right)^l \left|\frac{q_1}{\sqrt{\theta p}}\right|^{m-l} + (1-\theta)\frac{m!}{2}\|\mathbf{q}_2\|_2^m \left(\frac{4\sqrt{n}}{p}\right)^m \\
&\leq \theta\frac{m!}{2}\left(\frac{4\sqrt{n}}{\sqrt{p}}\right)^m \left(\frac{\|\mathbf{q}_2\|_2}{\sqrt{p}} + \frac{q_1}{\sqrt{\theta p}}\right)^m + (1-\theta)\frac{m!}{2}\|\mathbf{q}_2\|_2^m \left(\frac{4\sqrt{n}}{p}\right)^m \qquad \text{(F.65)} \\
&\leq \frac{m!}{2}\left(\frac{8\sqrt{n}}{\sqrt{\theta}p}\right)^m. \qquad \text{(F.66)}
\end{aligned}$$

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

$$\mathbb{P}\left[\left\|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right]\right\|_2 \geq t\right] \;\leq\; 2(n+1)\exp\left(-\frac{\theta p^3 t^2}{128n + 16\sqrt{\theta n}pt}\right), \tag{F.67}$$

as desired. □

F.3 Union Bound

Proposition F.9 (Uniformizing the Bounds). *Suppose that $\theta > \frac{1}{\sqrt{n}}$. Given any $\xi > 0$, there exists some constant $C(\xi)$, such that whenever $\exp(n) \geq p \geq C(\xi) n^4 \log n$, we have*

$$\left|Q_1(\mathbf{q}) - \mathbb{E}\left[Q_1(\mathbf{q})\right]\right| \;\leq\; \frac{2\xi}{\theta^{5/2}n^{3/2}p}, \tag{F.68}$$

$$\left\|\mathbf{Q}_2(\mathbf{q}) - \mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right]\right\|_2 \;\leq\; \frac{2\xi}{\theta^2 np} \tag{F.69}$$

hold uniformly for all $\mathbf{q} \in \mathbb{S}^{n-1}$, with probability at least $1 - c'\exp(-c''n)$ for some positive constants c' and c''.

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 $\mathcal{N}_\varepsilon$ of cardinality at most $(3/\varepsilon)^n$. For any $\mathbf{q} \in \mathbb{S}^{n-1}$, it can be written as

$$\mathbf{q} = \mathbf{q}' + \mathbf{e} \tag{F.70}$$

where $\mathbf{q}' \in \mathcal{N}_\varepsilon$ and $\|\mathbf{e}\|_2 \leq \varepsilon$. Let $\mathbf{y}^k = \left[x_{0k}, \mathbf{g}^k\right]^\top$ be a row of $\mathbf{Y}$, by (D.3) and (D.5), we have

$$\begin{aligned}
&\left|Q_1(\mathbf{q}) - \mathbb{E}\left[Q_1(\mathbf{q})\right]\right| \\
&= \left|\frac{1}{p}\sum_{k=1}^{p}\left\{x_{0k}\mathcal{S}_\lambda\left[\left\langle \mathbf{y}^k, \mathbf{q}' + \mathbf{e}\right\rangle\right] - \mathbb{E}\left[x_{0k}\mathcal{S}_\lambda\left[\left\langle \mathbf{y}^k, \mathbf{q}' + \mathbf{e}\right\rangle\right]\right]\right\}\right| \\
&\leq \left|\frac{1}{p}\sum_{k=1}^{p} x_{0k}\mathcal{S}_\lambda\left[\left\langle \mathbf{y}^k, \mathbf{q}' + \mathbf{e}\right\rangle\right] - \frac{1}{p}\sum_{k=1}^{p} x_{0k}\mathcal{S}_\lambda\left[\left\langle \mathbf{y}^k, \mathbf{q}'\right\rangle\right]\right| + \left|\frac{1}{p}\sum_{k=1}^{p} x_{0k}\mathcal{S}_\lambda\left[\left\langle \mathbf{y}^k, \mathbf{q}'\right\rangle\right] - \mathbb{E}\left[x_0\mathcal{S}_\lambda\left[\left\langle \mathbf{y}, \mathbf{q}'\right\rangle\right]\right]\right| \\
&+ \left|\mathbb{E}\left[x_0\mathcal{S}_\lambda\left[\left\langle \mathbf{y}, \mathbf{q}'\right\rangle\right]\right] - \mathbb{E}\left[x_0\mathcal{S}_\lambda\left[\left\langle \mathbf{y}, \mathbf{q}' + \mathbf{e}\right\rangle\right]\right]\right|.
\end{aligned} \tag{F.71}$$

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

$$\left|Q_1(\mathbf{q}) - \mathbb{E}\left[Q_1(\mathbf{q})\right]\right| \leq \left|Q_1(\mathbf{q}') - \mathbb{E}\left[Q_1(\mathbf{q}')\right]\right| + \left(\frac{1}{p}\sum_{k=1}^{p}\left|x_{0k}\right|\left\|\mathbf{y}^k\right\|_2 + \mathbb{E}\left[\left|x_0\right|\left\|\mathbf{y}\right\|_2\right]\right)\left\|\mathbf{e}\right\|_2 \tag{F.72}$$

$$\leq \left|Q_1(\mathbf{q}') - \mathbb{E}\left[Q_1(\mathbf{q}')\right]\right| + 2\varepsilon\frac{1}{\sqrt{\theta p}}\left(\frac{1}{\sqrt{\theta p}} + \max_{k\in[p]}\left\|\mathbf{g}^k\right\|_2\right). \tag{F.73}$$

By Lemma A.11 and the assumption that $p \leq \exp(n)$, we have that $\max_{k\in[p]}\left\|\mathbf{g}^k\right\|_2 \leq 4\sqrt{n/p}$ with probability at least $1 - \exp(-n/2)$. Taking $t = \xi\theta^{-5/2}n^{-3/2}p^{-1}$ in Lemma F.6 and applying a union bound, setting $\varepsilon = \xi\theta^{-2}n^{-2}/10$ and combining with the above estimate, we obtain that

$$\left|Q_1(\mathbf{q}) - \mathbb{E}\left[Q_1(\mathbf{q})\right]\right| \leq \frac{\xi}{\theta^{5/2}n^{3/2}p} + \frac{\xi}{5}\frac{1}{\theta^2 n^2}\frac{5\sqrt{n}}{\sqrt{\theta}p} \leq \frac{2\xi}{\theta^{5/2}n^{3/2}p} \tag{F.74}$$

holds for all $\mathbf{q} \in \mathbb{S}^{n-1}$, with probability at least $1 - \exp(-n/2) - \exp\left(-\frac{c_1(\xi)p}{\theta^4 n^3} + c_2(\xi)\, n\log n\right)$ for some numerical constants $c_1(\xi)$ and $c_2(\xi)$.

Similarly, by (D.3) and (D.5), we have

$$\begin{aligned}\left\|\mathbf{Q}_2(\mathbf{q})-\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right]\right\|_2 &= \left\|\frac{1}{p}\sum_{k=1}^{p}\left\{\mathbf{g}^k\mathcal{S}_\lambda\left[\left\langle\mathbf{y}^k,\mathbf{q}'+\mathbf{e}\right\rangle\right]-\mathbb{E}\left[\mathbf{g}^k\mathcal{S}_\lambda\left[\left\langle\mathbf{y}^k,\mathbf{q}'+\mathbf{e}\right\rangle\right]\right]\right\}\right\|_2\\ &\le \left\|\mathbf{Q}_2(\mathbf{q}')-\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q}')\right]\right\|_2+\left(\frac{1}{p}\sum_{k=1}^{p}\left\|\mathbf{g}^k\right\|_2\left\|\mathbf{y}^k\right\|_2+\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2\left\|\mathbf{y}^k\right\|_2\right]\right)\|\mathbf{e}\|_2\\ &\le \left\|\mathbf{Q}_2(\mathbf{q}')-\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q}')\right]\right\|_2+2\varepsilon\max_{k\in[p]}\left\|\mathbf{g}^k\right\|_2\left(\frac{1}{\sqrt{\theta p}}+\max_{k\in[p]}\left\|\mathbf{g}^k\right\|_2\right). \end{aligned}\tag{F.75}$$

Applying the above estimates for $\max_{k\in[p]}\left\|\mathbf{g}^k\right\|_2$, and taking $t=\xi\theta^{-2}n^{-1}p^{-1}$ in Lemma F.7 and applying a union bound, then setting $\varepsilon=\xi\theta^{-2}n^{-2}/40$, we obtain that

$$\left\|\mathbf{Q}_2(\mathbf{q})-\mathbb{E}\left[\mathbf{Q}_2(\mathbf{q})\right]\right\|_2\le\frac{\xi}{\theta^2np}+\frac{\xi}{20\theta^2n^2}4\sqrt{\frac{n}{p}}\left(\frac{1}{\sqrt{\theta p}}+4\sqrt{\frac{n}{p}}\right)\le\frac{2\xi}{\theta^2np}\tag{F.76}$$

holds for all $\mathbf{q}\in\mathbb{S}^{n-1}$, with probability at least $1-\exp\left(-n/2\right)-\exp\left(-\frac{c_3(\xi)p}{\theta^3n^3}+c_4\left(\xi\right)n\log n\right)$.

Overall, it is enough to take $p\ge Cn^4\log n$ for some large C to make the above events to hold with overwhelming probability, as desired. □

F.4 $\mathbf{Q}'(\mathbf{q})$ approximates $\mathbf{Q}(\mathbf{q})$

Proposition F.10. *Suppose $\theta>\frac{1}{\sqrt{n}}$. For any $\xi>0$, there exists some constant $C\left(\xi\right)$, such that whenever $\exp\left(n/2\right)/2\ge p\ge C\left(\xi\right)n^4\log n$, the following bounds*

$$\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\left|Q_1'(\mathbf{q})-Q_1(\mathbf{q})\right| \le \frac{\xi}{\theta^{5/2}n^{3/2}p}\tag{F.77}$$

$$\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\left\|\mathbf{Q}_2'(\mathbf{q})-\mathbf{Q}_2(\mathbf{q})\right\|_2 \le \frac{\xi}{\theta^2np},\tag{F.78}$$

hold for all $\mathbf{q}\in\mathbb{S}^{n-1}$, with probability at least $1-c'\exp\left(-c''n\right)$ for some positive constants c' and c''.

Proof. First, for any $\mathbf{q}\in\mathbb{S}^{n-1}$, from D.5, we know that

$$\begin{aligned}&\left|Q_1(\mathbf{q})-Q_1'(\mathbf{q})\right|\\ =&\left|\frac{1}{p}\sum_{k=1}^{p}x_{0k}\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]-\frac{1}{p}\sum_{k=1}^{p}\frac{x_{0k}}{\left\|\mathbf{x}_0\right\|_2}\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right|\\ \le&\left|\frac{1}{p}\sum_{k=1}^{p}x_{0k}\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]-\frac{1}{p}\sum_{k=1}^{p}x_{0k}\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right|+\left|\frac{1}{p}\sum_{k=1}^{p}x_{0k}\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]-\frac{1}{p}\sum_{k=1}^{p}\frac{x_{0k}}{\left\|\mathbf{x}_0\right\|_2}\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right|\\ \le&\frac{1}{p}\sum_{k=1}^{p}\left|x_{0k}\right|\left|\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]-\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right|+\frac{1}{p}\sum_{k=1}^{p}\left|x_{0k}\right|\left|1-\frac{1}{\left\|\mathbf{x}_0\right\|_2}\right|\left|\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right|. \end{aligned}\tag{F.79}$$

Let $\mathcal{I}=\mathrm{supp}(\mathbf{x}_0)$. Conditioned on the support, using the facts that $\mathcal{S}_\lambda[\cdot]$ is a nonexpansive operator, we obtain

$$\begin{aligned}\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\left|Q_1(\mathbf{q})-Q_1'(\mathbf{q})\right|&\le\frac{1}{p}\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\sum_{k\in\mathcal{I}}\left|x_{0k}\right|\left\|\mathbf{q}^\top\left(\mathbf{y}^k-\mathbf{y}'^k\right)\right\|_2+\left|1-\frac{1}{\left\|\mathbf{x}_0\right\|_2}\right|\frac{1}{p}\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\sum_{k\in\mathcal{I}}\left|x_{0k}\right|\left|\mathbf{q}^\top\mathbf{y}'^k\right|\\ &=\frac{1}{\sqrt{\theta}p^{3/2}}\left(\left\|\mathbf{Y}_{\mathcal{I}}-\mathbf{Y}_{\mathcal{I}}'\right\|_{\ell^2\to\ell^1}+\left|1-\frac{1}{\left\|\mathbf{x}_0\right\|_2}\right|\left\|\mathbf{Y}_{\mathcal{I}}'\right\|_{\ell^2\to\ell^1}\right). \end{aligned}\tag{F.80}$$

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

$$\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}|Q_1(\mathbf{q})-Q_1'(\mathbf{q})| \;\le\; \frac{1}{\sqrt{\theta}p^{3/2}}\left(\frac{10}{\theta}\sqrt{n\log p}+\frac{2\sqrt{2}}{5}\sqrt{\frac{n\log p}{\theta^3 p}}\times 7\sqrt{2\theta p}\right)\le \frac{16}{\theta^{3/2}p^{3/2}}\sqrt{n\log p}, \quad \text{(F.81)}$$

with probability at least $1-c_1\exp\left(-c_2 n\right)$ for some positive constants c_1 and c_2. Since the above holds uniformly for any support pattern $\mathcal{I}$, we conclude that

$$\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}|Q_1(\mathbf{q})-Q_1'(\mathbf{q})|\le \frac{16}{\theta^{3/2}p^{3/2}}\sqrt{n\log p} \quad \text{(F.82)}$$

with probability at least $1-c_1\exp\left(-c_2 n\right)$. Now it is sufficient to let $p\ge C\left(\xi\right)n^4\log n$ for some $C\left(\xi\right)>0$ to obtain the claimed result.

Similarly, by Lemma B.3 and Lemma B.4 in Appendix B, we have

$$\begin{aligned}
&\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\left\|\mathbf{Q}_2(\mathbf{q})-\mathbf{Q}_2'(\mathbf{q})\right\|_2\\
=&\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\left\|\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}^k\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]-\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}'^k\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right\|_2\\
\le&\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\left\|\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}^k\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]-\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}'^k\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]\right\|_2+\left\|\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}'^k\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}^k\right]-\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}'^k\mathcal{S}_\lambda\left[\mathbf{q}^\top\mathbf{y}'^k\right]\right\|_2\\
\le&\frac{1}{p}\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\sum_{k=1}^{p}\left\|\mathbf{g}^k-\mathbf{g}'^k\right\|_2\left|\mathbf{q}^\top\mathbf{y}^k\right|+\frac{1}{p}\sup_{\mathbf{q}\in\mathbb{S}^{n-1}}\sum_{k=1}^{p}\left\|\mathbf{g}'^k\right\|_2\left|\mathbf{q}^\top\left(\mathbf{y}^k-\mathbf{y}'^k\right)\right|\\
\le&\frac{1}{p}\left(\left\|\mathbf{G}-\mathbf{G}'\right\|_{\ell^2\to\ell^\infty}\left\|\mathbf{Y}\right\|_{\ell^2\to\ell^1}+\left\|\mathbf{G}'\right\|_{\ell^2\to\ell^\infty}\left\|\mathbf{Y}-\mathbf{Y}'\right\|_{\ell^2\to\ell^1}\right)\\
\le&\frac{1}{p}\left(\frac{32n}{\sqrt{\theta}p}\times 3\sqrt{p}+16\sqrt{\frac{n}{\theta p}}\times\frac{10}{\theta}\sqrt{n\log p}\right)\;\le\;\frac{192n\sqrt{\log p}}{\theta^{3/2}p^{3/2}}
\end{aligned} \quad \text{(F.83)}$$

holds conditioned on any support pattern $\mathcal{I}$, with probability at least $1-c_3\exp\left(-c_4 n\right)$ for some positive constants c_3 and c_4, which similarly implies the bound holds uniformly, regardless of the support, with the same probability. It is sufficiently to have $\exp\left(n/2\right)/2\ge p\ge C_2\left(\xi\right)n^4\log n$ to obtain the claimed result. □

G Large $|q_1|$ Iterates Staying in Safe Region for Rounding

Proposition G.1. *There exists a constant $\theta_0>0$, such that for any $\theta\in\left(\frac{1}{\sqrt{n}},\theta_0\right)$, whenever $\exp\left(n\right)\ge p\ge Cn^4\log n$ for some large constant $C>0$, we have*

$$\frac{|Q_1'(\mathbf{q})|}{\left\|\mathbf{Q}'(\mathbf{q})\right\|_2}\;\ge\;2\sqrt{\theta}, \quad \text{(G.1)}$$

for all $\mathbf{q}\in\mathbb{S}^{n-1}$ satisfying $|q_1|>3\sqrt{\theta}$, with probability at least $1-c'\exp\left(-c''n\right)$ for some positive constants c' and c''.

Proof. 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

$$\frac{\left\|\mathbf{Q}_2'\left(\mathbf{q}\right)\right\|_2}{\left|Q_1'\left(\mathbf{q}\right)\right|}<\sqrt{\frac{1}{4\theta}-1}, \quad \text{(G.2)}$$

which is implied by

$$\mathcal{L}\left(\mathbf{q}\right) \doteq \frac{\|\mathbb{E}\mathbf{Q}_2(\mathbf{q})\|_2 + \|\mathbf{Q}_2'(\mathbf{q}) - \mathbb{E}\mathbf{Q}_2(\mathbf{q})\|_2}{\mathbb{E}Q_1\left(\mathbf{q}\right) - \left|Q_1'\left(\mathbf{q}\right) - \mathbb{E}Q_1\left(\mathbf{q}\right)\right|} < \sqrt{\frac{1}{4\theta} - 1} \tag{G.3}$$

for any $\mathbf{q} \in \mathbb{S}^{n-1}$ satisfying $q_1 > 3\sqrt{\theta}$. Recall from (F.22) that

$$\mathbb{E}Q_1(\mathbf{q}) = \sqrt{\frac{\theta}{p}} \left\{ \left[\alpha\Psi\left(-\frac{\alpha}{\sigma}\right) + \beta\Psi\left(\frac{\beta}{\sigma}\right) \right] + \sigma \left[\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right) \right] \right\}, \tag{G.4}$$

where

$$\alpha = \frac{1}{\sqrt{p}} \left(\frac{q_1}{\sqrt{\theta}} + 1 \right), \quad \beta = \frac{1}{\sqrt{p}} \left(\frac{q_1}{\sqrt{\theta}} - 1 \right), \quad \sigma = \|\mathbf{q}_2\|_2 / \sqrt{p}. \tag{G.5}$$

Noticing the fact that

$$\psi\left(\frac{\beta}{\sigma}\right) - \psi\left(-\frac{\alpha}{\sigma}\right) \geq 0, \tag{G.6}$$

$$\Psi\left(\frac{\beta}{\sigma}\right) = \Psi\left(\frac{1}{\sqrt{1-q_1^2}} \left(\frac{q_1}{\sqrt{\theta}} - 1 \right) \right) \geq \Psi\left(2\right) \geq \frac{19}{20} \quad \text{for } q_1 > 3\sqrt{\theta}, \tag{G.7}$$

we have

$$\mathbb{E}Q_1\left(\mathbf{q}\right) \geq \frac{\sqrt{\theta}}{p} \left\{ \frac{q_1}{\sqrt{\theta}} \left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) \right] + \Psi\left(-\frac{\alpha}{\sigma}\right) - \Psi\left(\frac{\beta}{\sigma}\right) \right\} \geq \frac{2\sqrt{\theta}}{p} \Psi\left(\frac{\beta}{\sigma}\right) \geq \frac{19}{10} \frac{\sqrt{\theta}}{p}. \tag{G.8}$$

Moreover, from (F.23), we have

$$\|\mathbb{E}\mathbf{Q}_2\left(\mathbf{q}\right)\|_2 = \|\mathbf{q}_2\|_2 \left\{ \frac{2\left(1-\theta\right)}{p} \Psi\left(-\frac{\lambda}{\sigma}\right) + \frac{\theta}{p} \left[\Psi\left(-\frac{\alpha}{\sigma}\right) + \Psi\left(\frac{\beta}{\sigma}\right) \right] \right\} \tag{G.9}$$

$$\leq \frac{2\left(1-\theta\right)}{p} \Psi\left(-1\right) + \frac{\theta}{p} \left[\Psi\left(-1\right) + 1 \right] \leq \frac{2}{p} \Psi\left(-1\right) + \frac{\theta}{p} \leq \frac{2}{5p} + \frac{\theta}{p}, \tag{G.10}$$

where we have used the fact that $-\lambda/\sigma \leq -1$ and $-\alpha/\sigma \leq -1$. Moreover, from results in Proposition F.9 and Proposition F.10 in Appendix F , we know that

$$\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1'(\mathbf{q}) - \mathbb{E}Q_1(\mathbf{q})| \leq \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1'(\mathbf{q}) - Q_1(\mathbf{q})| + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} |Q_1(\mathbf{q}) - \mathbb{E}Q_1(\mathbf{q})| \leq \frac{1}{8000\theta^{5/2} n^{3/2} p}, \tag{G.11}$$

$$\sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\mathbf{Q}'(\mathbf{q}) - \mathbb{E}\mathbf{Q}(\mathbf{q})\|_2 \leq \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\mathbf{Q}'(\mathbf{q}) - \mathbf{Q}(\mathbf{q})\|_2 + \sup_{\mathbf{q} \in \mathbb{S}^{n-1}} \|\mathbf{Q}(\mathbf{q}) - \mathbb{E}\mathbf{Q}(\mathbf{q})\|_2 \leq \frac{1}{8000\theta^2 np} \tag{G.12}$$

hold with probability at least $1 - c' \exp\left(-c'' n\right)$ for some positive constants c' and c'' when $p \geq \Omega\left(n^4 \log n\right)$. Hence, with overwhelming probability, we have

$$\mathcal{L}\left(\mathbf{q}\right) \leq \frac{\frac{2}{5p} + \frac{\theta}{p} + \frac{1}{8000\theta^2 np}}{\frac{19}{10}\frac{\sqrt{\theta}}{p} - \frac{1}{8000\theta^{5/2} n^{3/2} p}} \leq \frac{\frac{3}{5}}{\frac{18\sqrt{\theta}}{10}} \leq \frac{1}{3\sqrt{\theta}} < \sqrt{\frac{1}{4\theta} - 1}, \tag{G.13}$$

whenever θ is sufficiently small. This completes the proof. □

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

$$\frac{\left|\left\langle \mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right), \mathbf{R}^\top \mathbf{e}_1 \right\rangle\right|}{\left\|\mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right)\right\|_2} = \frac{\left|\left\langle \mathbf{R}\mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right), \mathbf{e}_1 \right\rangle\right|}{\left\|\mathbf{R}\mathbf{Q}'\left(\mathbf{q}; \widehat{\mathbf{Y}}\right)\right\|_2} = \frac{\left|\left\langle \mathbf{Q}'\left(\mathbf{R}\mathbf{q}; \mathbf{Y}'\right), \mathbf{e}_1 \right\rangle\right|}{\left\|\mathbf{Q}'\left(\mathbf{R}\mathbf{q}; \mathbf{Y}'\right)\right\|_2}, \tag{G.14}$$

where we have applied the identity proved in (F.11). Taking $\mathbf{Rq} \in \mathbb{S}^{n-1}$ as the object of interest, by Proposition G.1, we conclude that

$$\frac{\left|\left\langle \mathbf{Q}'\left(\mathbf{Rq};\mathbf{Y}'\right),\mathbf{e}_1\right\rangle\right|}{\left\|\mathbf{Q}'\left(\mathbf{Rq};\mathbf{Y}'\right)\right\|_2} \geq 2\sqrt{\theta} \tag{G.15}$$

with overwhelming probability.

H Bounding Iteration Complexity

Proposition H.1. *There is a constant $\theta_0 > 0$, such that for any $\theta \in \left(\frac{1}{\sqrt{n}}, \theta_0\right)$, with probability at least $1 - c' \exp\left(-c''n\right)$ (c' and c'' are positive constants), the ADM algorithm in Algorithm 1, with any initialization $\mathbf{q}^{(0)} \in \mathbb{S}^{n-1}$ satisfying $\left|q_1^{(0)}\right| \geq \frac{1}{4\sqrt{\theta n}}$, will produce some iterate $\overline{\mathbf{q}}$ with $|\bar{q}_1| > 3\sqrt{\theta}$ at least once in at most $O(n^4 \log n)$ iterations, provided $\exp(n) \geq p \geq Cn^4 \log n$ for some large constant C.*

Proof. Recall from Proposition F.1 in Appendix F, the gap

$$G'(\mathbf{q}) = \frac{|Q_1'(\mathbf{q})|}{|q_1|} - \frac{\|\mathbf{Q}_2'(\mathbf{q})\|_2}{\|\mathbf{q}\|_2} \;\geq\; \frac{1}{4000\theta^2 np} \tag{H.1}$$

holds uniformly over $\mathbf{q} \in \mathbb{S}^{n-1}$ satisfying $\frac{1}{4\sqrt{\theta p}} \leq |q_1| \leq 3\sqrt{\theta}$ with probability at least $1 - c_1 \exp\left(-c_2 n\right)$ for positive constants c_1 and c_2, provided $p \geq \Omega\left(n^4 \log n\right)$. The gap $G'(\mathbf{q})$ implies that

$$\left|\widetilde{Q}_1'\left(\mathbf{q}\right)\right| \doteq \frac{|Q_1'(\mathbf{q})|}{\|\mathbf{Q}'\left(\mathbf{q}_2\right)\|_2} \geq \frac{|q_1|\,\|\mathbf{Q}_2'(\mathbf{q})\|_2}{\|\mathbf{q}\|_2\,\|\mathbf{Q}'\left(\mathbf{q}\right)\|_2} + \frac{|q_1|}{4000\theta^2 np\,\|\mathbf{Q}'\left(\mathbf{q}\right)\|_2} \tag{H.2}$$

$$\Longleftrightarrow \left|\widetilde{Q}_1'\left(\mathbf{q}\right)\right| \geq \frac{|q_1|}{\|\mathbf{q}_2\|_2}\sqrt{1 - \left|\widetilde{Q}_1'\left(\mathbf{q}\right)\right|^2} + \frac{|q_1|}{4000\theta^2 np\,\|\mathbf{Q}'\left(\mathbf{q}\right)\|_2} \tag{H.3}$$

$$\Longrightarrow \left|\widetilde{Q}_1'\left(\mathbf{q}\right)\right|^2 \geq |q_1|^2\left(1 + \frac{\|\mathbf{q}_2\|_2^2}{4000^2\theta^4 n^2 p^2\,\|\mathbf{Q}'\left(\mathbf{q}\right)\|_2^2}\right). \tag{H.4}$$

Now we know that

$$\sup_{\mathbf{q}\in\Gamma}\|\mathbf{Q}'\left(\mathbf{q}\right)\|_2 \leq \sup_{\mathbf{q}\in\Gamma}|\mathbf{Q}_1'\left(\mathbf{q}\right)| + \sup_{\mathbf{q}\in\Gamma}\|\mathbf{Q}_2'\left(\mathbf{q}\right)\|_2 \tag{H.5}$$

$$= \sup_{\mathbf{q}\in\Gamma}\left|\frac{1}{p}\sum_{k=1}^{p} x_{0k} S_\lambda\left[x_{0k}q_1 + \mathbf{q}_2^\top \mathbf{g}^k\right]\right| + \sup_{\mathbf{q}\in\Gamma}\left\|\frac{1}{p}\sum_{k=1}^{p}\mathbf{g}^k S_\lambda\left[x_{0k}q_1 + \mathbf{q}_2^\top\mathbf{g}^k\right]\right\|_2 \tag{H.6}$$

$$\leq \frac{1}{p}\left(\sup_{\mathbf{q}\in\Gamma}\sum_{k=1}^{p}|x_{0k}|\left|x_{0k}q_1 + \mathbf{q}_2^\top\mathbf{g}^k\right| + \sup_{\mathbf{q}\in\Gamma}\sum_{k=1}^{p}\left\|\mathbf{g}^k\right\|_2\left|x_{0k}q_1 + \mathbf{q}_2^\top\mathbf{g}^k\right|\right) \tag{H.7}$$

$$\leq 2\left(\frac{1}{\sqrt{\theta p}} + \sup_{k\in[p]}\left\|\mathbf{g}^k\right\|_2\right)\sup_{\mathbf{q}\in\Gamma}\left(\frac{|q_1|}{\sqrt{\theta p}} + \|\mathbf{q}_2\|_2\sup_{k\in[p]}\left\|\mathbf{g}^k\right\|_2\right) \tag{H.8}$$

$$\leq 2\left(\frac{1}{\sqrt{\theta p}} + \sup_{k\in[p]}\left\|\mathbf{g}^k\right\|_2\right)^2. \tag{H.9}$$

From Lemma A.11, we know that $\sup_{k\in[p]}\left\|\mathbf{g}^k\right\|_2 \leq \sqrt{2\log p}/\sqrt{p} + 2\sqrt{n}/\sqrt{p}$ with probability at least $1 - \exp\left(-n/2\right)$. Then provided $p \leq \exp\left(n\right)$ and $\theta \geq 1/\sqrt{n}$, we obtain

$$\sup_{\mathbf{q}\in\Gamma}\|\mathbf{Q}'\left(\mathbf{q}\right)\|_2 \leq \frac{50n}{p}. \tag{H.10}$$

So we conclude that

$$\frac{\left|\widetilde{Q}_1'(\mathbf{q})\right|}{|q_1|} \geq \sqrt{1+\frac{1-9\theta}{4000^2\times 50^2\times\theta^4 n^4}}. \tag{H.11}$$

Therefore, starting with any $\mathbf{q}\in\mathbb{S}^{n-1}$ such that $|q_1|\geq\frac{1}{4\sqrt{\theta n}}$, we will need at most

$$T=\frac{2\log\left(3\sqrt{\theta}/\frac{1}{4\sqrt{\theta n}}\right)}{\log\left(1+\frac{1-9\theta}{4000^2\times 50^2\times\theta^4 n^4}\right)}=\frac{2\log\left(12\theta\sqrt{n}\right)}{\log\left(1+\frac{1-9\theta}{4000^2\times 50^2\times\theta^4 n^4}\right)}\leq\frac{2\log\left(12\theta\sqrt{n}\right)}{(\log 2)\,\frac{1-9\theta}{4000^2\times 50^2\times\theta^4 n^4}}\leq Cn^4\log n \tag{H.12}$$

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

I Rounding to the Desired Solution

For convenience, we will assume the notations we used in Appendix B. Then the rounding scheme can be written as

$$\min_{\mathbf{q}}\ \left\|\mathbf{Y}'\mathbf{R}\mathbf{q}\right\|_1,\quad \text{s.t. } \langle\overline{\mathbf{q}},\mathbf{q}\rangle=1, \tag{I.1}$$

for some orthogonal matrix $\mathbf{R}$. We will show the rounding procedure get us to the desired solution with overwhelming probability, regardless of the particular orthonormal basis used.

Proposition I.1. *Suppose the input basis is $\mathbf{Y}'$ defined in* (B.3) *and the ADM algorithm produces $\bar{\mathbf{q}}\in\mathbb{S}^{n-1}$ with $\bar{q}_1>2\sqrt{\theta}$. Then there exists some constants $C,\theta_0>0$, such that when $p\geq Cn^2$ and $\theta\in\left(\frac{1}{\sqrt{n}},\theta_0\right)$, the rounding procedure with $\mathbf{r}=\overline{\mathbf{q}}$ returns the desired solution $\mathbf{e}_1$ with probability at least $1-c\exp(-c'n)$ for some numerical constants $c,c'>0$.*

Proof. The rounding program (I.1) can be written as

$$\inf_{\mathbf{q}}\ \left\|\mathbf{Y}'\mathbf{q}\right\|_1,\quad \text{s.t. } \bar{q}_1 q_1+\langle\overline{\mathbf{q}}_2,\mathbf{q}_2\rangle=1. \tag{I.2}$$

Consider its relaxation

$$\inf_{\mathbf{q}}\ \left\|\mathbf{Y}'\mathbf{q}\right\|_1,\quad \text{s.t. } \bar{q}_1 q_1+\left\|\overline{\mathbf{q}}_2\right\|_2\left\|\mathbf{q}_2\right\|_2\geq 1. \tag{I.3}$$

It is obvious that the feasible set of (I.3) contains that of (I.2). So if $\mathbf{e}_1$ is the unique optimal solution (UOS) of (I.3), it is also the UOS of (I.2). Let $\mathcal{I}=\mathrm{supp}(\mathbf{x}_0)$, and consider a modified problem

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

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

$$\begin{aligned}-\left\|\mathbf{G}'_{\mathcal{I}}\mathbf{q}_2\right\|_1+\left\|\mathbf{G}'_{\mathcal{I}^c}\mathbf{q}_2\right\|_1 &\geq -\left\|\mathbf{G}_{\mathcal{I}}\mathbf{q}_2\right\|_1+\left\|\mathbf{G}_{\mathcal{I}^c}\mathbf{q}_2\right\|_1-\left\|(\mathbf{G}-\mathbf{G}')\,\mathbf{q}_2\right\|_1 &\text{(I.5)}\\ &\geq -\left\|\mathbf{G}_{\mathcal{I}}\mathbf{q}_2\right\|_1+\left\|\mathbf{G}_{\mathcal{I}^c}\mathbf{q}_2\right\|_1-\left\|\mathbf{G}-\mathbf{G}'\right\|_{\ell^2\to\ell^1}\left\|\mathbf{q}_2\right\|_2. &\text{(I.6)}\end{aligned}$$

When $p\geq\Omega\left(n^2\right)$, by Lemma A.14 and Lemma B.3, we know that

$$\begin{aligned}&-\left\|\mathbf{G}_{\mathcal{I}}\mathbf{q}_2\right\|_1+\left\|\mathbf{G}_{\mathcal{I}^c}\mathbf{q}_2\right\|_1-\left\|\mathbf{G}-\mathbf{G}'\right\|_{\ell^2\to\ell^1}\left\|\mathbf{q}_2\right\|_2\\ &\qquad\geq -\frac{6}{5}\sqrt{\frac{2}{\pi}}2\theta\sqrt{p}\left\|\mathbf{q}_2\right\|_2+\frac{24}{25}\sqrt{\frac{2}{\pi}}(1-2\theta)\sqrt{p}\left\|\mathbf{q}_2\right\|_2-8\sqrt{n}\left\|\mathbf{q}_2\right\|_2\doteq\zeta\left\|\mathbf{q}_2\right\|_2\end{aligned} \tag{I.7}$$

holds with probability at least $1 - c_1 \exp(-c_2 n)$ for some positive constants c_1 and c_2. Thus, we make a further relaxation of problem (I.2) by

$$\inf_{\mathbf{q}} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| + \zeta \|\mathbf{q}_2\|_2, \quad \text{s.t. } \overline{q}_1 q_1 + \|\overline{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 \geq 1, \tag{I.8}$$

whose objective value lower bounds that of (I.4). By similar arguments, if $\mathbf{e}_1$ is UOS to (I.8), it is UOS to (I.2). At the optimal solution to (I.8), notice that it is necessary to have $\text{sign}(q_1) = \text{sign}(\overline{q}_1)$ and $\overline{q}_1 q_1 + \|\overline{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 = 1$. So (I.8) is equivalent to

$$\inf_{\mathbf{q}} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| + \zeta \|\mathbf{q}_2\|_2, \quad \text{s.t. } \overline{q}_1 q_1 + \|\overline{\mathbf{q}}_2\|_2 \|\mathbf{q}_2\|_2 = 1. \tag{I.9}$$

which is further equivalent to

$$\inf_{q_1} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 |q_1| + \zeta \frac{1 - |\overline{q}_1|\,|q_1|}{\|\overline{\mathbf{q}}_2\|_2}, \quad \text{s.t. } |q_1| \leq \frac{1}{|\overline{q}_1|}. \tag{I.10}$$

Notice that the problem in (I.10) is linear in $|q_1|$ with a compact feasible set, which indicates that the optimal solution only occur at the boundary points $|q_1| = 0$ and $|q_1| = 1/|\overline{q}_1|$. Therefore, $\mathbf{q} = \mathbf{e}_1$ is the UOS of (I.10) if and only if

$$\frac{1}{|\overline{q}_1|} \left\| \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2} \right\|_1 < \frac{\zeta}{\|\overline{\mathbf{q}}_2\|_2}. \tag{I.11}$$

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

$$\frac{\sqrt{2\theta p}}{2\sqrt{\theta}} \leq \zeta = \frac{24}{25}\sqrt{\frac{2}{\pi}}\sqrt{p}\left(1 - \frac{9}{2}\theta - \frac{25}{3}\sqrt{\frac{n}{p}}\right). \tag{I.12}$$

Therefore there exists a constant $\theta_0 > 0$, such that whenever $\theta \leq \theta_0$, the rounding returns $\mathbf{e}_1$, completing the proof. □

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

$$\min_{\mathbf{q}} \|\mathbf{Y}'\mathbf{R}\mathbf{q}\|_1, \quad \text{s.t. } \langle \mathbf{q}', \mathbf{R}\mathbf{q} \rangle = 1. \tag{I.13}$$

Renaming $\mathbf{R}\mathbf{q}$, it follows from the above argument that at optimum $\mathbf{q}^\star$ it holds that $\mathbf{R}\mathbf{q}^\star = \mathbf{e}_1$ with overwhelming probability.