Composite optimization for robust blind deconvolution

Vasileios Charisopoulos[*]   Damek Davis[†]   Mateo Díaz[‡]
Dmitriy Drusvyatskiy[§]

Abstract

The blind deconvolution problem seeks to recover a pair of vectors from a set of rank one bilinear measurements. We consider a natural nonsmooth formulation of the problem and show that under standard statistical assumptions, its moduli of weak convexity, sharpness, and Lipschitz continuity are all dimension independent. This phenomenon persists even when up to half of the measurements are corrupted by noise. Consequently, standard algorithms, such as the subgradient and prox-linear methods, converge at a rapid dimension-independent rate when initialized within constant relative error of the solution. We then complete the paper with a new initialization strategy, complementing the local search algorithms. The initialization procedure is both provably efficient and robust to outlying measurements. Numerical experiments, on both simulated and real data, illustrate the developed theory and methods.

1 Introduction

A variety of tasks in data science amount to solving a nonlinear system $F(x) = 0$, where $F \colon \mathbb{R}^d \to \mathbb{R}^m$ is a highly structured smooth map. The setting when F is a quadratic map already subsumes important problems such as phase retrieval [12,37,47], blind deconvolution [4,33,36,49], matrix completion [13,18,48], and covariance matrix estimation [15,35], to name a few. Recent works have suggested a number of two-stage procedures for globally solving such problems. The first stage—*initialization*—yields a rough estimate x_0 of an optimal solution, often using spectral techniques. The second stage—*local refinement*—uses a local search algorithm that rapidly converges to an optimal solution, when initialized at x_0. For a detailed discussion, we refer the reader to the recent survey [16].

The typical starting point for local refinement is to form an optimization problem

$$\min_{x \in \mathcal{X}} \ f(x) := h(F(x)), \tag{1.1}$$

---

[*]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; `people.orie.cornell.edu/vc333/`

[†]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; `people.orie.cornell.edu/dsd95/`

[‡]Center for Applied Mathematics, Cornell University Ithaca, NY 14850, USA; `people.cam.cornell.edu/md825/`

[§]Department of Mathematics, U. Washington, Seattle, WA 98195; `www.math.washington.edu/~ddrusv`. Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 1740551 awards.

where $h(\cdot)$ is a carefully chosen penalty function and $\mathcal{X}$ is a constraint set. Most widely-used penalties are smooth and convex; e.g., the squared ℓ_2-norm $h(z) = \frac{1}{2}\|z\|_2^2$ is ubiquitous in this context. Equipped with such penalties, the problem (1.1) is smooth and therefore gradient-based methods become immediately applicable. The main analytic challenge is that the condition number $\frac{\lambda_{\max}(\nabla^2 f)}{\lambda_{\min}(\nabla^2 f)}$ of the problem (1.1) often grows with the dimension of the ambient space d. This is the case for example for phase retrieval, blind deconvolution, and matrix completion problems; see e.g. [16] and references therein. Consequently, generic nonlinear programming guarantees yield efficiency estimates that are far too pessimistic. Instead, a fruitful strategy is to recognize that the Hessian may be well-conditioned along the "relevant" set of directions, which suffice to guarantee rapid convergence. This is where new insight and analytic techniques for each particular problem come to bare (e.g. [37,39,49]).

Smoothness of the penalty function $h(\cdot)$ in (1.1) is crucially used by the aforemention techniques. A different recent line of work [6, 20, 21, 25] has instead suggested the use of nonsmooth convex penalties—most notably the ℓ_1-norm $h(z) = \|z\|_1$. Such a nonsmooth formulation will play a central role in our work. A number of algorithms are available for nonsmooth compositional problems (1.1), most notably the subgradient method

$$x_{t+1} = \mathrm{proj}_{\mathcal{X}}(x_t - \alpha_t v_t) \qquad \text{with} \qquad v_t \in \partial f(x_t),$$

and the prox-linear algorithm

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \; h\Big(F(x_t) + \nabla F(x_t)(x - x_t)\Big) + \frac{1}{2\alpha_t}\|x - x_t\|_2^2.$$

The local convergence guarantees of both methods can be succinctly described as follows. Set $\mathcal{X}^* := \operatorname{argmin}_{\mathcal{X}} f$ and suppose there exist constants $\rho, \mu, L > 0$ satisfying:

- **(approximation)** $\left| h(F(y)) - h\Big(F(x) + \nabla F(x)(y - x)\Big) \right| \leq \frac{\rho}{2}\|y - x\|_2^2$ for all $x \in \mathcal{X}$,

- **(sharpness)** $f(x) - \inf f \geq \mu \cdot \mathrm{dist}^2(x, \mathcal{X}^*)$ for all $x \in \mathcal{X}$,

- **(Lipschitz bound)** $\|v\|_2 \leq L$ for all $v \in \partial f(x)$ with $\mathrm{dist}(x, \mathcal{X}^*) \leq \frac{\rho}{\mu}$.

Then when equipped with an appropriate sequence α_t and initialized at a point x_0 satisfying $\mathrm{dist}(x_0, \mathcal{X}^*) \leq \frac{\rho}{\mu}$, both the subgradient and prox-linear iterates will converge to an optimal solution of the problem. The prox-linear algorithm converges quadratically, while the subgradient method converges at a linear rate governed by the ratio $\frac{\mu}{L} \in (0, 1)$.

A possible advantage of nonsmooth techniques can be gleaned from the phase retrieval problem. The papers [25, Corollary 3.1,3.2], [21, Corollary 3.8] recently, showed that for the phase retrieval problem, standard statistical assumptions imply that with high probability all the constants $\rho, \mu, L > 0$ are dimension independent. Consequently, completely generic guarantees outlined above, without any modification, imply that both methods converge at a dimension-independent rate, when initialized within constant relative error of the optimal solution. This is in sharp contrast to the smooth formulation of the problem, where a more nuanced analysis is required, based on restricted smoothness and convexity. Moreover, this approach is robust to outliers in the sense that analogous guarantees persist even when up to half of the measurements are corrupted by noise.

In light of the success of the nonsmooth penalty approach for phase retrieval, it is intriguing to determine if nonsmooth techniques can be fruitful for a wider class of large-scale problems. Our current work fits squarely within this research program. In this work, we analyze a nonsmooth penalty technique for the problem of *blind deconvolution.* Formally, we consider the task of robustly recovering a pair $(\bar{w}, \bar{x}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ from m bilinear measurements:

$$y_i = \langle \ell_i, \bar{w} \rangle \langle r_i, \bar{x} \rangle + \eta_i, \tag{1.2}$$

where η is an arbitrary noise corruption with frequency $p_{\text{fail}} := \frac{|\text{supp } \eta|}{m}$ that is at most one half, and $\ell_i \in \mathbb{R}^{d_1}$ and $r_i \in \mathbb{R}^{d_2}$ are known measurement vectors. Such bilinear systems and their complex analogues arise often in biological systems, control theory, coding theory, and image deblurring, among others. Most notably such problems appear when recovering a pair $(u, v) \in \mathbb{C}^m \times \mathbb{C}^m$ from the convolution measurements $y = (Lu) * (Rv) \in \mathbb{C}^m$. When passing to the Fourier domain this problem is equivalent to that of solving a complex bilinear system of equations; see the pioneering work [4]. All the arguments we present can be extended to the complex case. We focus on the real case for simplicity.

In this work we analyze the following nonsmooth formulation of the problem:

$$\min_{\|w\|_2, \|x\|_2 \leq \nu\sqrt{M}} f(w, x) := \frac{1}{m} \sum_{i=1}^{m} |\langle \ell_i, w \rangle \langle r_i, x \rangle - y_i|, \tag{1.3}$$

where $\nu \geq 1$ is a user-specified constant and $M = \|\bar{w}\bar{x}^{\top}\|_F$. Our contributions are two-fold:

1. (**Local refinement**) Suppose that the vectors ℓ_i and r_i are both i.i.d. Sub-Gaussian and satisfy a mild growth condition (which is automatically satisfied for Gaussian random vectors). We will show that as long as the number of measurements satisfies $m \gtrsim \frac{d_1 + d_2}{(1 - 2p_{\text{fail}})^2} \ln(\frac{1}{1 - 2p_{\text{fail}}})$, the formulation (1.3) admits dimension independent constants ρ, L, and μ with high probability. Consequently, subgradient and prox-linear methods rapidly converge to the optimal solution at a dimension-independent rate when initialized at a point x_0 with constant relative error $\frac{\|w_0 x_0^{\top} - \bar{w}\bar{x}^{\top}\|_F}{\|\bar{w}\bar{x}^T\|_F} \lesssim 1$. Analogous results also hold under more general incoherence assumptions.

2. (**Initialization**) Suppose now that ℓ_i and r_i are both i.i.d. Gaussian and are independent from the noise η. We develop an initialization procedure that in the regime $m \gtrsim d_1 + d_2$ and $p_{\text{fail}} \in [0, 1/10]$, will find a point x_0 satisfying $\frac{\|w_0 x_0^{\top} - \bar{w}\bar{x}^{\top}\|_F}{\|\bar{w}\bar{x}^T\|_F} \lesssim 1$, with high probability. To the best of our knowledge, this is the only available initialization procedure with provable guarantees in presence of gross outliers. We also develop complementary guarantees under the weaker assumption that the vectors (ℓ_i, r_i) corresponding to exact measurements are independent from the noise η_i in the outlying measurements. This noise model allows one to plant outlying measurements from a completely different pair of signals, and is therefore computationally more challenging.

The literature studying bilinear systems is rich. From the information-theoretic perspective [17, 29, 34], the optimal sample complexity in the noiseless regime is $m \gtrsim d_1 + d_2$ if no further assumptions (e.g. sparsity) are imposed on the signals. Therefore, from a sample complexity viewpoint, our guarantees are optimal. Incidentally, to our best knowledge, all

alternative approaches are either suboptimal by a polylogarithmic factor in d_1, d_2 or require knowing the sign pattern of one of the underlying signals [3,4].

Recent algorithmic advances for blind deconvolution can be classified into two main approaches: works based on convex relaxations and those employing gradient descent on a smooth nonconvex function. The influential convex techniques of [3,4] "lift" the objective to a higher dimension, thereby necessitating the resolution of a high-dimensional semidefinite program. The more recent work of [1,2] instead relaxes the feasible region in the natural parameter space, under the assumption that the coordinate signs of either $\bar{w}$ or $\bar{x}$ are known a priori. Finally, with the exception of [4], the aforementioned works do not provide guarantees in the noisy regime.

Nonconvex approaches for blind deconvolution typically apply gradient descent to a smooth formulation of the problem [27, 33, 37]. Since the condition number of the problem scales with dimension, as we mentioned previously, these works introduce a nuanced analysis that is specific to the gradient method. The authors of [33] propose applying gradient descent on a regularized objective function, and identify a "basin of attraction" around the solution. The paper [37] instead analyzes gradient descent on the unregularized objective. They use the leave-one-out technique and prove that the iterates remain within a region where the objective function satisfies restricted strong convexity and smoothness conditions. The sample complexities of the methods in [27, 33, 37, 37] are optimal up to polylog factors.

The nonconvex strategies mentioned above all use spectral methods for initialization. These methods are not robust to outliers, since they rely on the leading singular vectors/values of a potentially noisy measurement operator. Adapting the spectral initialization of [25] to bilinear inverse problems enables us to deal with gross outliers of arbitrary magnitude. Indeed, high variance noise makes it easier for our initialization to "reject" outlying measurements.

The outline of the paper is as follows. Section 2 records basic notation we will use throughout the paper. Section 3 reviews the impact of sharpness and weak convexity on the rapid convergence of numerical methods. Section 4 establishes estimates of weak convexity, sharpness, and Lipschitz moduli for the blind deconvolution problem under both deterministic and statistical assumptions on the data. Section 5 introduces the initialization procedure and proves its correctness even if a constant fraction of measurements is corrupted by gross outliers. The final Section 6 presents numerical experiments illustrating the theoretical results in the paper.

2 Notation

The section records basic notation that we will use throughout the paper. To this end, we always endow $\mathbb{R}^d$ with the dot product, $\langle x, y \rangle = x^\top y$, and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. The symbol $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$, while $\mathbb{B}$ denotes the open unit ball. When convenient, we will use the notation $\mathbb{B}^d$ to emphasize the dimension of the ambient space. More generally, $\mathbb{B}_r(x)$ will stand for the open ball around x of radius r. We define the distance and the nearest-point projection of a point x onto a closed set $Q \subseteq \mathbb{R}^d$ by

$$\text{dist}(x, Q) = \inf_{y \in Q} \|x - y\|_2 \quad \text{and} \quad \text{proj}_Q(x) = \operatorname*{argmin}_{y \in Q} \|x - y\|_2,$$

respectively. For any pair of real-valued functions $f, g \colon \mathbb{R}^d \to \mathbb{R}$, the notation $f \lesssim g$ means that there exists a positive constant C such that $f(x) \leq Cg(x)$ for all $x \in \mathbb{R}^d$. We write $f \asymp g$ if both $f \lesssim g$ and $g \lesssim f$.

We will always use the trace inner product $\langle X, Y \rangle = \mathrm{Tr}(X^T Y)$ on the space of matrices $\mathbb{R}^{d_1 \times d_2}$. The symbols $\|A\|_{\mathrm{op}}$ and $\|A\|_F$ will denote the operator and Frobenius norm of A, respectively. Assuming $d \leq m$, the map $\sigma \colon \mathbb{R}^{d \times m} \to \mathbb{R}_+^d$ returns the vector of ordered singular values $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_d(A)$. Note the equalities $\|A\|_F = \|\sigma(A)\|_2$ and $\|A\|_{\mathrm{op}} = \sigma_1(A)$.

Nonsmooth functions will appear throughout this work. Consequently will use some basic constructions of generalized differentiation, as set out for example in the monographs [8, 38, 42, 45]. Consider a function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and a point x, with $f(x)$ finite. Then the *Fréchet subdifferential* of f at $\bar{x}$, denoted by $\partial f(x)$, is the set of all vectors $v \in \mathbb{R}^d$ satisfying

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \qquad \text{as } y \to x. \tag{2.1}$$

Thus, a vector v lies in the subdifferential $\partial f(x)$ precisely when the function $y \mapsto f(x) + \langle v, y - x \rangle$ locally minorizes f up to first-order. We say that a point x is *stationary for* f whenever the inclusion, $0 \in \partial f(x)$, holds. Standard results show for convex functions f the subdifferential $\partial f(x)$ reduces to the subdifferential in the sense of convex analysis, while for differentiable functions f it consists only of the gradient $\partial f(x) = \{\nabla f(x)\}$.

Notice that in general, the little-o term in (2.1) may depend on the base-point x, and the estimate (2.1) therefore may be nonuniform. In this work, we will only encounter functions whose subgradients automatically satisfy a uniform type of lower-approximation property. We say that a function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *ρ-weakly convex*[1] if the perturbed function $x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2$ is convex. It is straightforward to see that for any ρ-weakly convex function f, subgradients automatically satisfy the uniform bound:

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathbb{R}^d, \forall v \in \partial f(x).$$

We will comment further on the class of weakly convex functions in Section 3.

We say a that a random vector X in $\mathbb{R}^d$ is *η-sub-gaussian* whenever $\mathbb{E} \exp\left(\frac{\langle u, X \rangle^2}{\eta^2}\right) \leq 2$ for all vectors $u \in \mathbb{S}^{d-1}$. The *sub-gaussian norm* of a real-valued random variable X is defined to be $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E} \exp\left(\frac{X^2}{t^2}\right) \leq 2\}$, while the *sub-exponential norm* is defined by $\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E} \exp\left(\frac{|X|}{t}\right) \leq 2\}$. Given a sample $y = (y_1, \ldots, y_n)$, we will write $\mathtt{med}(y)$ to denote its median.

3   Algorithms for sharp weakly convex problems

The central thrust of this work is that under reasonable statistical assumptions, the penalty formulation (1.3) satisfies two key properties: (1) the objective function is weakly convex and (2) grows at least linearly as one moves away from the solution set. In this section, we review

---

[1]Weakly convex functions also go by other names such as lower-C^2, uniformly prox-regularity, paraconvex, and semiconvex.

the consequences of these two properties for local rapid convergence of numerical methods. The discussion mostly follows the recent work [20], though elements of this viewpoint can already be seen in the two papers [21, 25] on robust phase retrieval.

Setting the stage, we introduce the following assumption.

Assumption A. Consider the optimization problem,

$$\min_{x \in \mathcal{X}} \ f(x). \tag{3.1}$$

Suppose that the following properties hold for some real $\mu, \rho > 0$.

1. **(Weak convexity)** The set $\mathcal{X}$ is closed and convex, while the function $f : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex.

2. **(Sharpness)** The set of minimizers $\mathcal{X}^* := \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x)$ is nonempty and the inequality

$$f(x) - \inf f \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*) \qquad \text{holds for all } x \in \mathcal{X}.$$

The class of weakly convex functions is broad and its importance in optimization is well documented [5, 40, 43, 44, 46]. It trivially includes all convex functions and all C^1-smooth functions with Lipschitz gradient. More broadly, it includes all compositions

$$f(x) = h(F(x)),$$

where $h(\cdot)$ is convex and L-Lipschitz, and $F(\cdot)$ is C^1-smooth with β-Lipschitz Jacobian. Indeed then the composite function $f = h \circ F$ is weakly convex with parameter $\rho = L\beta$; see e.g. [24, Lemma 4.2]. In particular, our target problem (1.3) is clearly weakly convex, being a composition of the ℓ_1 norm and a quadratic map. The estimate $\rho = L\beta$ on the weak convexity constant is often much too pessimistic, however. Indeed, under statistical assumptions, we will see that the target problem (1.3) has a much better weak convexity constant. The notion of sharpness, and the related error bound property, is now ubiquitous in nonlinear optimization. Indeed, sharpness underlies much of perturbation theory and rapid convergence guarantees of various numerical methods. For a systematic treatment of error bounds and their applications, we refer the reader to the monographs of Dontchev-Rockafellar [22] and Ioffe [28], and the article of Lewis-Pang [31].

Taken together, weak convexity and sharpness provide an appealing framework for deriving local rapid convergence guarantees for numerical methods. In this work, we specifically focus on two such procedures: the subgradient and prox-linear algorithms. To this end, we aim to estimate both the radius of rapid converge around the solution set and the rate of convergence. Our ultimate goal is to show that when specialized to our target problem (1.3), with high probability, both of these quantities are independent of the ambient dimensions d_1 and d_2 as soon as the number of measurements is sufficiently large.

Both the subgradient and prox-linear algorithms have the property that when initialized at a stationary point of the problem, they could stay there for all subsequent iterations. Since we are interested in finding global minima, and not just stationary points, we must therefore estimate the neighborhood of the solution set that has no extraneous stationary points. This is the content of the following simple lemma [20, Lemma 3.1].

Lemma 3.1. *Suppose that Assumption A holds. Then the problem* (3.1) *has no stationary points x satisfying*

$$0 < \text{dist}(x; \mathcal{X}^*) < \frac{2\mu}{\rho}.$$

Proof. Fix a critical point $x \in \mathcal{X} \notin \mathcal{X}^*$. Letting $x^* := \text{proj}_{\mathcal{X}^*}(x)$, we deduce $\mu \cdot \text{dist}(x, \mathcal{X}^*) \leq f(x) - f(x^*) \leq \frac{\rho}{2} \cdot \|x - x^*\|^2 = \frac{\rho}{2} \cdot \text{dist}^2(x, \mathcal{X}^*)$. Dividing by $\text{dist}(x, \mathcal{X}^*)$, the result follows. $\square$

The estimate $\frac{2\mu}{\rho}$ of the radius in Lemma 3.1 is tight. To see this, consider minimizing the univariate function $f(x) = |\lambda^2 x^2 - 1|$ on the real line $\mathcal{X} = \mathbb{R}$. Observe that the set of minimizers is $\mathcal{X}^* = \{\pm\frac{1}{\lambda}\}$, while $x = 0$ is always an extraneous stationary point. A quick computation shows that the smallest valid weak convexity is $\rho = 2\lambda^2$ while the largest valid sharpness constant is $\mu = \lambda$.

We therefore deduce $\text{dist}(0, \mathcal{X}^*) = \frac{1}{\lambda} = \frac{2\mu}{\rho}$. Hence the radius of the region $\frac{2\mu}{\rho}$ that is devoid of extraneous stationary points is tight.

In light of Lemma 3.1, let us define for any $\gamma > 0$ the tube

$$\mathcal{T}_\gamma := \left\{ z \in \mathbb{R}^d \colon \text{dist}(z, \mathcal{X}^*) \leq \gamma \cdot \frac{\mu}{\rho} \right\}. \tag{3.2}$$

Thus we would like to search for algorithms whose basin of attraction is a tube $\mathcal{T}_\gamma$ for some numerical constant $\gamma > 0$. Due to the above discussion, such a basin of attraction is in essence optimal.

We next discuss two rapidly converging algorithms. The first is the Polyak subgradient method, outlined in Algorithm 1. Notice that the only parameter that is needed to implement the procedure is the minimal value of the problem (3.1). This value is sometimes known; case in point, the minimal value of the penalty formulation (1.3) is zero when the bilinear measurements are exact.

Algorithm 1: Polyak Subgradient Method

Data: $x_0 \in \mathbb{R}^d$
Step k: $(k \geq 0)$
    Choose $\zeta_k \in \partial f(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.
    Set $x_{k+1} = \text{proj}_{\mathcal{X}} \left(x_k - \dfrac{f(x_k) - \min_{\mathcal{X}} f}{\|\zeta_k\|^2} \zeta_k \right)$.

The rate of convergence of the method relies on the Lipschitz constant and the condition measure:

$$L := \sup\{\|\zeta\| : \zeta \in \partial f(x), x \in \mathcal{T}_1\} \qquad \text{and} \qquad \tau := \frac{\mu}{L}.$$

A straightforward argument [20, Lemma 3.2] shows $\tau \in [0, 1]$. The following theorem appears as [20, Theorem 4.1], while its application to phase retrieval was investigated in [21].

Theorem 3.2 (Polyak subgradient method). *Suppose that Assumption A holds and fix a real $\gamma \in (0, 1)$. Then Algorithm 1 initialized at any point $x_0 \in \mathcal{T}_\gamma$ produces iterates that converge Q-linearly to $\mathcal{X}^*$, that is*

$$\text{dist}^2(x_{k+1}, \mathcal{X}^*) \leq \left(1 - (1 - \gamma)\tau^2\right) \text{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0. \tag{3.3}$$

7

When the minimal value of the problem (3.1) is unknown, there is a straightforward modification of the subgradient method that converges R-linearly. The idea is to choose a geometrically decaying control sequence for the stepsize. The disadvantage is that the convergence guarantees rely on being able to tune estimates of L, ρ, and μ.

Algorithm 2: Subgradient method with geometrically decreasing stepsize

---
Data: Real $\lambda > 0$ and $q \in (0, 1)$.

Step k: $(k \geq 0)$

    Choose $\zeta_k \in \partial g(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.

    Set stepsize $\alpha_k = \lambda \cdot q^k$.

    Update iterate $x_{k+1} = \text{proj}_{\mathcal{X}}\left(x_k - \alpha_k \frac{\zeta_k}{\|\zeta_k\|}\right)$.

The following theorem appears as [20, Theorem 6.1]. The convex version of the result dates back to Goffin [26].

Theorem 3.3 (Geometrically decaying subgradient method). *Suppose that Assumption A holds, fix a real $\gamma \in (0, 1)$, and suppose $\tau \leq \sqrt{\frac{1}{2-\gamma}}$. Set $\lambda := \frac{\gamma\mu^2}{\rho L}$ and $q := \sqrt{1 - (1 - \gamma)\tau^2}$. Then the iterates x_k generated by Algorithm 2, initialized at a point $x_0 \in \mathcal{T}_\gamma$, satisfy:*

$$\text{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{\rho^2}\left(1 - (1-\gamma)\tau^2\right)^k \qquad \forall k \geq 0. \tag{3.4}$$

Notice that both subgradient algorithms 1 and 2 are at best locally linearly convergent, with a relatively cheap per-iteration cost. As the last example we discuss an algorithm that is specifically designed for convex compositions, which is locally *quadratically convergent*. The caveat is that the method may have a high per-iteration cost, since in each iteration one must solve an auxiliary convex optimization problem.

Setting the stage, let us introduce the following assumption.

Assumption B. Consider the optimization problem,

$$\min_{x \in \mathcal{X}} f(x) := h(F(x)). \tag{3.5}$$

Suppose that the following properties holds for some real $\mu, \rho > 0$.

1. **(Convexity and smoothness)** The function $h(\cdot)$ and the set $\mathcal{X}$ are convex and $F(\cdot)$ is differentiable.

2. **(Approximation accuracy)** The convex models $f_x(y) := h(F(x) + \nabla F(x)(y - x))$ satisfy the estimate:

$$|f(y) - f_x(y)| \leq \frac{\rho}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathcal{X}.$$

3. **(Sharpness)** The set of minimizers $\mathcal{X}^* := \underset{x \in \mathcal{X}}{\text{argmin}}\, f(x)$ is nonempty and the inequality

$$f(x) - \inf f \geq \mu \cdot \text{dist}(x, \mathcal{X}^*) \qquad \text{holds for all } x \in \mathcal{X}.$$

8

It is straightforward to see that Assumption B implies that f is ρ-weakly convex; see e.g. [24, Lemma 7.3]. Therefore Assumption B implies Assumption A.

Algorithm 3 describes the prox-linear method—a close variant of Gauss-Newton. For a historical account of the prox-linear method, see e.g., [10, 24, 32] and the references therein.

Algorithm 3: Prox-linear algorithm

Data: Initial point $x_0 \in \mathbb{R}^d$, proximal parameter $\beta > 0$.

Step k: $(k \geq 0)$

$\quad$ Set $x_{k+1} \leftarrow \underset{x \in \mathcal{X}}{\mathrm{argmin}} \left\{ h\left(F(x_k) + \nabla F(x_k)(x - x_k)\right) + \frac{\beta}{2}\|x - x_k\|^2 \right\}.$

The following theorem proves that under Assumption B, the prox-linear method converges quadratically, when initialized sufficiently close to the solution set. Guarantees of this type have appeared, for example, in [11, 23, 23, 25]. For the sake of completeness, we provide a quick argument.

Theorem 3.4 (Prox-linear algorithm). *Suppose Assumption B holds. Choose any $\beta \geq \rho$ and set $\gamma := \rho/\beta$. Then Algorithm 3 initialized at any point $x_0 \in \mathcal{T}_\gamma$ converges quadratically:*

$$\mathrm{dist}(x_{k+1}, \mathcal{X}^*) \leq \tfrac{\beta}{\mu} \cdot \mathrm{dist}^2(x_k, \mathcal{X}^*) \qquad \forall k \geq 0.$$

Proof. Consider an iterate x_k and choose any $x^* \in \mathrm{proj}_{\mathcal{X}^*}(x_k)$. Taking into account that the function $x \mapsto f_{x_k}(x) + \frac{\beta}{2}\|x - x_k\|^2$ is strongly convex and x_{k+1} is its minimizer, we deduce

$$\left(f_{x_k}(x_{k+1}) + \frac{\beta}{2}\|x_{k+1} - x_k\|^2 \right) + \frac{\beta}{2}\|x_{k+1} - x^*\|^2 \leq f_{x_k}(x^*) + \frac{\beta}{2}\|x^* - x_k\|^2.$$

Using Assumption B.2, we therefore obtain

$$f(x_{k+1}) + \frac{\beta}{2}\|x_{k+1} - x^*\|^2 \leq f(x^*) + \beta\|x^* - x_k\|^2.$$

Rearranging and using sharpness (Assumption B.3), we conclude

$$\mu \cdot \mathrm{dist}(x_{k+1}, \mathcal{X}^*) \leq f(x_{k+1}) - f(x^*) \leq \beta \cdot \mathrm{dist}^2(x_k, \mathcal{X}^*),$$

as claimed. $\qquad\square$

4 Assumptions and Models

In this section, we aim to interpret the efficiency of the subgradient and prox-linear algorithms discussed in Section 3, when applied to our target problem (1.3). To this end, we must estimate the three parameters $\rho, \mu, L > 0$. These quantities control both the size of the attraction neighborhood around the optimal solution set and the rate of convergence within the neighborhood. In particular, we will show that these quantities are independent of the ambient dimension d_1, d_2 under natural assumptions on the data generating mechanism.

It will be convenient for the time being to abstract away from the formulation (1.3), and instead consider the function

$$g(w, x) := \frac{1}{m} \|\mathcal{A}(wx^\top) - y\|_1,$$

where $\mathcal{A} \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ is an arbitrary linear map and $y \in \mathbb{R}^m$ is an arbitrary vector. The formulation (1.3) corresponds to the particular linear map $\mathcal{A}(X) = (\ell_i^\top X r_i)_{i=1}^m$. Since we will be interested in the prox-linear method, let us define the convex model

$$g_{(w,x)}(\hat{w}, \hat{x}) := \frac{1}{m} \|\mathcal{A}(wx^\top + w(\hat{x} - x)^\top + (\hat{w} - w)x^\top) - y\|_1.$$

Our strategy is as follows. Section 4.1 identifies deterministic assumptions on the data, $\mathcal{A}$ and y, that yield favorable estimates of $\rho, \mu, L > 0$. Then Section 4.2 shows that these deterministic assumptions hold with high probability under natural statistical assumptions on the data generating mechanism.

4.1 Favorable Deterministic Properties

The following property, widely used in the literature, will play a central role in our analysis.

Assumption C (Restricted Isometry Property (RIP)). *There exist constants $c_1, c_2 > 0$ such that for all matrices $X \in \mathbb{R}^{d_1 \times d_2}$ of rank at most two the following bound holds:*

$$c_1 \|X\|_F \leq \frac{1}{m} \|\mathcal{A}(X)\|_1 \leq c_2 \|X\|_F.$$

The following proposition estimates the two constants ρ and L, governing the performance of the subgradient and prox-linear methods under Assumption C.

Proposition 4.1 (Approximation accuracy and Lipschitz continuity). *Suppose Assumption C holds and let $K > 0$ be arbitrary. Then the following estimates hold:*

$$|g(\hat{w}, \hat{x}) - g_{(w,x)}(\hat{w}, \hat{x})| \leq \frac{c_2}{2} \cdot \|(w, x) - (\hat{w}, \hat{x})\|_2^2 \qquad \forall x, \hat{x} \in \mathbb{R}^{d_1}, \forall w, \hat{w} \in \mathbb{R}^{d_2},$$

$$|g(w, x) - g(\hat{w}, \hat{x})| \leq \sqrt{2}c_2 K \cdot \|(w, x) - (\hat{w}, \hat{x})\|_2 \qquad \forall x, \hat{x} \in K\mathbb{B}, w, \hat{w} \in K\mathbb{B}.$$

Proof. To see the first estimate, observe

$$
\begin{aligned}
|g(\hat{w}, \hat{x}) - g_{(w,x)}(\hat{w}, \hat{x})| &= \left| \frac{1}{m} \|\mathcal{A}(\hat{w}\hat{x}^\top) - y\|_1 - \frac{1}{m} \|\mathcal{A}(wx^\top + w(\hat{x} - x)^\top + (\hat{w} - w)x^\top) - y\|_1 \right| \\
&\leq \frac{1}{m} \|\mathcal{A}(\hat{w}\hat{x}^\top - wx^\top - w(\hat{x} - x)^\top - (\hat{w} - w)x^\top)\|_1 \\
&= \frac{1}{m} \|\mathcal{A}((w - \hat{w})(x - \hat{x})^\top)\|_1 \\
&\leq c_2 \|(w - \hat{w})(x - \hat{x})^\top\|_F \\
&\leq \frac{c_2}{2} \left(\|w - \hat{w}\|_2^2 + \|x - \hat{x}\|_2^2 \right),
\end{aligned}
$$

10

where the last estimate follows from Young's inequality $2ab \leq a^2 + b^2$. Now suppose $w, \hat{w} \in K\mathbb{B}$ and $x, \hat{x} \in K\mathbb{B}$. We then successively compute:

$$
\begin{aligned}
|g(w,x) - g(\hat{w}, \hat{x})| \leq \frac{1}{m} \|\mathcal{A}(wx^\top - \hat{w}\hat{x}^\top)\|_1 &\leq c_2 \|wx^\top - \hat{w}\hat{x}^\top\|_F \\
&= c_2 \|(w - \hat{w})x^\top + \hat{w}(x - \hat{x})^\top\|_F \\
&\leq c_2 \|x\|_2 \|w - \hat{w}\|_2 + c_2 \|\hat{w}\|_2 \|x - \hat{x}\|_2 \\
&\leq \sqrt{2} c_2 K \cdot \|(w,x) - (\hat{w}, \hat{x})\|_2.
\end{aligned}
$$

The proof is complete. □

We next move on to estimates of the sharpness constant μ. To this end, consider two vectors $\bar{w} \in \mathbb{R}_1^d$ and $\bar{x} \in \mathbb{R}^{d_2}$, and set $M := \|\bar{x}\bar{w}^T\|_F = \|\bar{x}\|_2 \cdot \|\bar{w}^T\|_2$. Without loss of generality, henceforth, we suppose $\|\bar{w}\|_2 = \|\bar{x}\|_2$. Our estimates on the sharpness constant will be valid only on bounded sets. Consequently, define the two sets:

$$
\mathcal{S}_\nu := \nu\sqrt{M} \cdot (\mathbb{B}^{d_1} \times \mathbb{B}^{d_2}), \qquad \mathcal{S}_\nu^* := \{(\alpha\bar{w}, (1/\alpha)\bar{x}) : 1/\nu \leq |\alpha| \leq \nu\}.
$$

The set $\mathcal{S}_\nu$ simply encodes a bounded region, while $\mathcal{S}_\nu^*$ encodes all rank-1 factorizations of the matrix $\bar{w}\bar{x}^\top$ with bounded factors. We begin with the following proposition, which analyzes the sharpness properties of the idealized function

$$
(x, w) \mapsto \|wx^\top - \bar{w}\bar{x}^\top\|_F.
$$

The proof is quite long, and therefore we have placed it in Appendix A.1.

Proposition 4.2. *For any $\nu \geq 1$, we have the following bound*

$$
\|wx^\top - \bar{w}\bar{x}^\top\|_F \geq \frac{\sqrt{M}}{2\sqrt{2}(\nu + 1)} \mathrm{dist}\big((w, x), \mathcal{S}_\nu^*\big) \qquad \text{for all } (w, x) \in \mathcal{S}_\nu.
$$

Thus the function $(x, w) \mapsto \|wx^\top - \bar{w}\bar{x}^\top\|_F$ is sharp on the set $\mathcal{S}_\nu$ with coefficient $\frac{\sqrt{M}}{2\sqrt{2}(\nu+1)}$. We note in passing that the analogue of Proposition 4.2 for symmetric matrices was proved in [49, Lemma 5.4].

The sharpness of the loss $g(\cdot, \cdot)$ in the noiseless regime (i.e. when $y = \mathcal{A}(\bar{w}\bar{x}^\top)$) is now immediate.

Proposition 4.3 (Sharpness in the Noiseless Regime)**.** *Suppose that Assumption C holds and that equality, $y = \mathcal{A}(\bar{w}\bar{x}^\top)$, holds. Then for any $\nu \geq 1$, we have the following bound:*

$$
g(w, x) - g(\bar{w}, \bar{x}) \geq \frac{c_1\sqrt{M}}{2\sqrt{2}(\nu + 1)} \mathrm{dist}\big((w, x), \mathcal{S}_\nu^*\big) \qquad \text{for all } (w, x) \in \mathcal{S}_\nu.
$$

Proof. Using Assumption C and Proposition 4.2, we deduce

$$
g(w, x) - g(\bar{w}, \bar{x}) = \frac{1}{m} \|\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\|_1 \geq c_1 \|wx^\top - \bar{w}\bar{x}^\top\|_F \geq \frac{c_1\sqrt{M}}{2\sqrt{2}(\nu + 1)} \mathrm{dist}\big((w, x), \mathcal{S}_\nu^*\big),
$$

as claimed. □

11

Sharpness in the noisy case requires an additional assumption. We record it below. Henceforth, for any set $\mathcal{I}$, we define the restricted linear map $\mathcal{A}_{\mathcal{I}} \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{|\mathcal{I}|}$ by setting $\mathcal{A}_{\mathcal{I}}(X) := (\mathcal{A}(X))_{i \in \mathcal{I}}$.

Assumption D ($\mathcal{I}$-outliner bounds). *There exists a set $\mathcal{I} \subset \{1, \ldots, m\}$, vectors $\bar{w} \in \mathbb{R}^{d_1}$, $\bar{x} \in \mathbb{R}^{d_2}$, and a constant $c_3 > 0$ such that the following hold.*

(C1) *Equality $y_i = \mathcal{A}(\bar{w}\bar{x}^\top)_i$ holds for all $i \notin \mathcal{I}$.*

(C2) *For all matrices $X \in \mathbb{R}^{d_1 \times d_2}$ of rank at most two, we have*

$$c_3 \|X\|_F \leq \frac{1}{m} \|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \frac{1}{m} \|\mathcal{A}_{\mathcal{I}}(X)\|_1. \tag{4.1}$$

Combining Assumption D with Proposition 4.2 quickly yields sharpness of the objective even in the noisy setting.

Proposition 4.4 (Sharpness in the noisy regime). *Suppose that Assumption D holds. Then*

$$g(w, x) - g(\bar{w}, \bar{x}) \geq \frac{c_3 \sqrt{M}}{2\sqrt{2}(\nu + 1)} \text{dist}\big((w, x), \mathcal{S}_\nu^*\big) \qquad \textit{for all } (w, x) \in \mathcal{S}_\nu.$$

Proof. Defining $\eta = \mathcal{A}(\bar{w}\bar{x}^T) - y$, we have the following bound:

$$
\begin{aligned}
&g(w, x) - g(\bar{w}, \bar{x}) \\
&\quad = \frac{1}{m} \left(\|\mathcal{A}\left(wx^\top - \bar{w}\bar{x}^\top\right) + \eta\|_1 - \|\eta\|_1 \right) \\
&\quad = \frac{1}{m} \left(\|\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\|_1 + \sum_{i \in \mathcal{I}} \left(\left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i + \eta_i\right| - \left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| - |\eta_i| \right) \right) \\
&\quad \geq \frac{1}{m} \left(\|\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\|_1 - 2 \sum_{i \in \mathcal{I}} \left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| \right) \\
&\quad = \frac{1}{m} \sum_{i \in \mathcal{I}^c} \left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| - \frac{1}{m} \sum_{i \in \mathcal{I}} \left|\left(\mathcal{A}(wx^\top - \bar{w}\bar{x}^\top)\right)_i\right| \\
&\quad \geq c_3 \|wx^\top - \bar{w}\bar{x}^\top\|_F \geq \frac{c_3 \sqrt{M}}{2\sqrt{2}(\nu + 1)} \text{dist}\big((w, x), \mathcal{S}_\nu^*\big),
\end{aligned}
$$

where the first inequality follows by the reverse triangle inequality, the second inequality follows by Assumption (C2), and the final inequality follows from Proposition 4.2. The proof is complete. $\square$

To summarize, suppose Assumptions C and D are valid. Then in the notation of Section 3 we may set:

$$\boxed{\rho = c_2, \quad L = c_2 \nu \sqrt{2M}, \quad \mu = \frac{c_3 \sqrt{M}}{2\sqrt{2}(\nu + 1)}.}$$

Consequently, the tube radius of $\mathcal{T}_1$ is $\frac{2\mu}{\rho} = \frac{c_3}{c_2} \cdot \frac{\sqrt{M}}{\sqrt{2}(\nu+1)}$ and the the linear convergence rate of the subgradient method is governed by $\tau = \frac{\mu}{L} = \frac{c_3}{c_2} \cdot \frac{1}{4(\nu+1)^2}$. In particular, the local search algorithms must be initialized at a point (x, w), whose relative distance to the solution set $\frac{\text{dist}((x,w),\mathcal{S}_\nu^*)}{\sqrt{\|\bar{x}\bar{w}^\top\|_F}}$ is upper bounded by a constant. We record this conclusion below.

Corollary 4.5 (Convergence guarantees). *Suppose Assumptions C and D are valid, and consider the optimization problem*

$$\min_{(x,w)\in\mathcal{S}_\nu} \ g(w,x) = \frac{1}{m}\|\mathcal{A}(wx^\top) - y\|_1.$$

Choose any pair (x_0, y_0) satisfying

$$\frac{\text{dist}((w_0,x_0),\mathcal{S}_\nu^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq \frac{c_3}{4\sqrt{2}c_2(\nu+1)}.$$

Then the following are true.

1. **(Polyak subgradient)** *Algorithm 1 initialized (x_0, y_0) produces iterates that converge linearly to $\mathcal{S}_\nu^*$, that is*

$$\frac{\text{dist}^2((w_k,x_k),\mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leq \left(1 - \frac{c_3^2}{32c_2^2(\nu+1)^4}\right)^k \cdot \frac{c_3^2}{32c_2^2(\nu+1)^2} \qquad \forall k \geq 0.$$

2. **(geometric subgradient)** *Set $\lambda := \frac{c_3^2\sqrt{\|\bar{w}\bar{x}^\top\|_F}}{16\sqrt{2}c_2^2\nu(\nu+1)^2}$ and $q := \sqrt{1 - \frac{c_3^2}{32c_2^2(\nu+1)^4}}$. Then the iterates x_k generated by Algorithm 2, initialized at (w_0, x_0) converge linearly:*

$$\frac{\text{dist}^2((w_k,x_k),\mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leq \left(1 - \frac{c_3^2}{32c_2^2(\nu+1)^4}\right)^k \cdot \frac{c_3^2}{32c_2^2(\nu+1)^2} \qquad \forall k \geq 0.$$

3. **(prox-linear)** *Algorithm 3 with $\beta = \rho$ and initialized at (w_0, x_0) converges quadratically:*

$$\frac{\text{dist}((w_k,x_k),\mathcal{S}_\nu^*))}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq 2^{-2^k} \cdot \frac{c_3}{2\sqrt{2}c_2(\nu+1)} \qquad \forall k \geq 0.$$

4.2   Assumptions under generative models

In this section, we present natural generative models under which Assumptions C and D are guaranteed to hold. Recall that at the high level, we aim to recover the pair of signals $(\bar{w}, \bar{x})$ based on given corrupted bilinear measurements y. Formally, let us fix two disjoint sets $\mathcal{I}_{\text{in}} \subseteq [m]$ and $\mathcal{I}_{\text{out}} \subseteq [m]$, called the *inlier* and *outlier* sets. Intuitively, the index set $\mathcal{I}_{\text{in}}$ encodes exact measurements while $\mathcal{I}_{\text{out}}$ encodes measurements that have been replaced by gross outliers. Define the corruption frequency $p_{\text{fail}} := \frac{|\mathcal{I}_{\text{out}}|}{m}$; henceforth, we will suppose $p_{\text{fail}} \in [0, 1/2)$. Then for an arbitrary, potentially random sequence $\{\xi_i\}_{i=1}^m$, we consider the measurement model:

$$y_i := \begin{cases} \langle \ell_i, \bar{w}\rangle\langle r_i, \bar{x}\rangle & \text{if } i \in \mathcal{I}_{\text{in}}, \\ \xi_i & \text{if } i \in \mathcal{I}_{\text{out}}. \end{cases} \tag{4.2}$$

In accordance with the previous section, we define the linear map $\mathcal{A}\colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ by $\mathcal{A}(X) = (\ell_i^\top X r_i)_{i=1}^m$. To simplify notation, we let $L \in \mathbb{R}^{m \times d_1}$ denote the matrix whose rows, in column form, are ℓ_i and we let $R \in \mathbb{R}^{m \times d_2}$ denote the matrix whose rows are r_i. Note that we make no assumptions about the nature of ξ_i. In particular, ξ_i can even encode exact measurements for a different signal.

We focus on two measurement matrix models. The first model requires both matrices L and R to be random. For simplicity, the reader may assume both are Gaussian with i.i.d. entries, though the results of this paper extend beyond this case. The second model allows semi-deterministic matrices, namely deterministic L and Gaussian R with i.i.d. entries. In the later parts of the paper, we will put further incoherence assumptions on the deterministic matrix L.

Random matrix models.

M1 The vectors ℓ_i and r_i are i.i.d. realizations of η-sub-gaussian random vectors $\ell \in \mathbb{R}^{d_1}$ and $r \in \mathbb{R}^{d_2}$, respectively. Suppose moreover that ℓ and r are independent and satisfy the nondegeneracy condition,

$$\inf_{\substack{X:\ \operatorname{rank} X \leq 2 \\ \|X\|_F = 1}} \mathbb{P}(|\ell^\top X r| \geq \mu_0) \geq p_0, \tag{4.3}$$

for some real $\mu_0, p_0 > 0$.

M2 The matrix L is arbitrary and the matrix R is standard Gaussian.

Some comments are in order. The model **M1** is fully stochastic, in the sense that ℓ_i and r_i are generated by independent sub-gaussian random vectors. The nondegeneracy condition (4.3) essentially asserts that with positive probability, the products $\ell^\top X r$ are non-negligible, uniformly over all unit norm rank two matrices X. In particular, the following example shows that Gaussian matrices with i.i.d. entries are admissible under Model **M1**. In contrast, the model **M2** is semi-stochastic: it allows L to be deterministic, while making the stronger assumption that R is Gaussian.

Example 4.1 (Gaussian Matrices Satisfy Model **M1**)**.** Assume that ℓ and r are standard Gaussian random vectors in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. We claim this setting is admissible under **M1**. To see this, fix a rank 2 matrix X having unit Frobenius norm. Consider now a singular value decomposition $X = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top$, and note the equality, $\sigma_1^2 + \sigma_2^2 = 1$. For each index $i = 1, 2$ define $a_i := \langle \ell, u_i \rangle$ and $b_i := \langle v_i, r \rangle$. Then clearly a_1, a_2, b_1, b_2 are i.i.d. standard Gaussian; see e.g. [51, Exercise 3.3.6]. Thus, for any $c \geq 0$, we compute

$$\mathbb{P}(|\ell^\top X r| \geq c) = \mathbb{P}(|\sigma_1 a_1 b_1 + \sigma_2 a_2 b_2| \geq c) = \mathbb{E}\left(\mathbb{P}(|\sigma_1 a_1 b_1 + \sigma_2 a_2 b_2| \geq c \mid a_1, a_2)\right).$$

Notice that conditioned on a_1, a_2, we have $\sigma_1 a_1 b_1 + \sigma_2 a_2 b_2 \sim \mathsf{N}(0, (\sigma_1 a_1)^2 + (\sigma_2 a_2)^2)$. Thus letting z be a standard normal, we have

$$\mathbb{P}(|\ell^\top X r| \geq c) = \mathbb{E}\left(\mathbb{P}(\sqrt{(\sigma_1 a_1)^2 + (\sigma_2 a_2)^2}|z| \geq c \mid a_1, a_2)\right)$$

$$= \mathbb{P}(\sqrt{(\sigma_1 a_1)^2 + (\sigma_2 a_2)^2}|z| \geq c)$$

$$\geq \mathbb{P}(\sigma_1 |a_1 z| \geq c) \geq \mathbb{P}(|a_1 z| \geq \sqrt{2}c).$$

Therefore, we may simply set $\mu_0 = \operatorname{median}(|a_1 z|)/\sqrt{2}$ and $p_0 = \frac{1}{2}$.

4.2.1 Assumptions C and D under Model M1

In this section, we aim to prove the following theorem, which shows validity of Assumptions C and D under **M1**, with high probability.

Theorem 4.6 (Measurement Model **M1**). *Consider a set $\mathcal{I} \subseteq \{1, \ldots, m\}$ satisfying $|\mathcal{I}| < m/2$. Then there exist constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$ depending only on μ_0, p_0, η such that the following holds. As long as $m \geq \frac{c_1(d_1+d_2+1)}{(1-2|\mathcal{I}|/m)^2} \ln\left(c_2 + \frac{c_2}{1-2|\mathcal{I}|/m}\right)$, then with probability at least $1 - 4\exp\left(-c_3 \left(1 - 2|\mathcal{I}|/m\right)^2 m\right)$, every matrix $X \in \mathbb{R}^{d_1 \times d_2}$ of rank at most two satisfies*

$$c_4\|X\|_F \leq \frac{1}{m}\|\mathcal{A}(X)\|_1 \leq c_5\|X\|_F, \tag{4.4}$$

and

$$\frac{1}{m}\left[\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1\right] \geq c_6 \left(1 - \frac{2|\mathcal{I}|}{m}\right)\|X\|_F. \tag{4.5}$$

Due to scale invariance, in the proof we only concern ourselves with matrices X of rank at most two satisfying $\|X\|_F = 1$. Let us fix such a matrix X and an arbitrary index set $\mathcal{I} \subseteq \{1, \ldots, m\}$ with $|\mathcal{I}| < m/2$. We begin with the following lemma.

Lemma 4.7 (Pointwise concentration). *The random variable $|\ell^\top X r|$ is sub-exponential with parameter $\sqrt{2}\eta^2$. Consequently, the estimate holds:*

$$\mu_0 p_0 \leq \mathbb{E}|\ell^\top X r| \lesssim \eta^2. \tag{4.6}$$

Moreover, there exists a numerical constant $c > 0$ such that for any $t \in (0, \sqrt{2}\eta^2]$, we have with probability at least $1 - 2\exp(-\frac{ct^2}{\eta^4}m)$ the estimate:

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1\right]\right| \leq t. \tag{4.7}$$

Proof. Markov's inequality along with (4.3) implies

$$\mathbb{E}|\ell^\top X r| \geq \mu_0 \cdot \mathbb{P}(|\ell^\top X r| \geq \mu_0) \geq \mu_0 p_0,$$

which is the lower bound in (4.6). Now we address the upper bound. To that end, suppose that X has a singular value decomposition $X = \sigma_1 U_1 V_1^\top + \sigma_2 U_2 V_2^\top$. We then deduce

$$\begin{aligned}
\||\ell^\top X r|\|_{\psi_1} = \|\ell^\top \left(\sigma_1 U_1 V_1^\top + \sigma_2 U_2 V_2^\top\right) r\|_{\psi_1} &= \|\sigma_1 \langle \ell, U_1 \rangle \langle V_1, r \rangle + \sigma_2 \langle \ell, U_2 \rangle \langle V_2, r \rangle\|_{\psi_1} \\
&\leq \sigma_1 \|\langle \ell, U_1 \rangle \langle V_1, r \rangle\|_{\psi_1} + \sigma_2 \|\langle \ell, U_2 \rangle \langle V_2, r \rangle\|_{\psi_1} \\
&\leq \sigma_1 \|\langle \ell, U_1 \rangle\|_{\psi_2} \|\langle V_1, r \rangle\|_{\psi_2} + \sigma_2 \|\langle \ell, U_2 \rangle\|_{\psi_2} \|\langle V_2, r \rangle\|_{\psi_2} \\
&\leq (\sigma_1 + \sigma_2)\eta^2 \leq \sqrt{2}\eta^2,
\end{aligned}$$

where the second inequality follows since $\|\cdot\|_{\psi_1}$ is a norm and $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$ [51, Lemma 2.7.7]. This bound has two consequences: first $|\ell^\top X r|$ is a sub-exponential random variable with parameter $\sqrt{2}\eta^2$ and second $\mathbb{E}|\ell^\top X r| \leq \sqrt{2}\eta^2$, see [51, Exercise 2.7.2]. The first bound will be useful momentarily, while the second completes the proof of (4.6).

15

Next define the sub-exponential random variable

$$Y_i = \begin{cases} |\ell_i^\top X r_i| - \mathbb{E}|\ell_i^\top X r_i| & \text{if } i \notin \mathcal{I} \\ -(|\ell_i^\top X r_i| - \mathbb{E}|\ell_i^\top X r_i|) & \text{if } i \in \mathcal{I}. \end{cases}$$

Standard results (e.g. [51, Exercise 2.7.10]) imply $\|Y_i\|_{\psi_1} \lesssim \sqrt{2}\eta^2$ for all i. Using Bernstein inequality for sub-exponential random variables, Theorem C.6, to upper bound $\mathbb{P}\left(\frac{1}{m}|\sum_{i=1}^m Y_i| \geq t\right)$ completes the proof. $\qquad\square$

Proof of Theorem 4.6. Choose $\epsilon \in (0, \sqrt{2})$ and let $\mathcal{N}$ be the $(\epsilon/\sqrt{2})$-net guaranteed by Lemma C.2. Let $\mathcal{E}$ denote the event that the following two estimates hold for all matrices in $X \in \mathcal{N}$:

$$\frac{1}{m}\left| \|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1\right] \right| \leq t, \tag{4.8}$$

$$\frac{1}{m}\left| \|\mathcal{A}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}(X)\|_1\right] \right| \leq t. \tag{4.9}$$

Throughout the proof, we will assume that the event $\mathcal{E}$ holds. We will estimate the probability of $\mathcal{E}$ at the end of the proof. Meanwhile, seeking to establish RIP, define the quantity

$$c_2 := \sup_{X \in S_2} \frac{1}{m}\|\mathcal{A}(X)\|_1.$$

We aim first to provide a high probability bound on c_2.

Let $X \in S_2$ be arbitrary and let $X_\star$ be the closest point to X in $\mathcal{N}$. Then we have

$$\frac{1}{m}\|\mathcal{A}(X)\|_1 \leq \frac{1}{m}\|\mathcal{A}(X_\star)\|_1 + \frac{1}{m}\|\mathcal{A}(X - X_\star)\|_1$$

$$\leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(X_\star)\|_1 + t + \frac{1}{m}\|\mathcal{A}(X - X_\star)\|_1 \tag{4.10}$$

$$\leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(X)\|_1 + t + \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(X - X_\star)\|_1 + \|\mathcal{A}(X - X_\star)\|_1\right), \tag{4.11}$$

where (4.10) follows from (4.7) and (4.11) follows from the triangle inequality. To simplify the third term in (4.11), using SVD, we deduce that there exist two orthogonal matrices X_1, X_2 of rank at most two satisfying $X - X_\star = X_1 + X_2$. With this decomposition in hand, we compute

$$\frac{1}{m}\|\mathcal{A}(X - X_\star)\|_1 \leq \frac{1}{m}\|\mathcal{A}(X_1)\|_1 + \frac{1}{m}\|\mathcal{A}(X_2)\|_1$$

$$\leq c_2(\|X_1\|_F + \|X_2\|_F) \leq \sqrt{2}c_2\|X - X_\star\|_F \leq c_2\epsilon, \tag{4.12}$$

where the second inequality follows from the definition of c_2 and the estimate $\|X_1\|_F + \|X_2\|_F \leq \sqrt{2}\|(X_1, X_2)\|_F = \sqrt{2}\|X_1 + X_2\|_F$. Thus, we arrive at the bound

$$\frac{1}{m}\|\mathcal{A}(X)\|_1 \leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(X)\|_1 + t + 2c_2\epsilon. \tag{4.13}$$

16

As X was arbitrary, we may take the supremum of both sides of the inequality, yielding $c_2 \leq \frac{1}{m}\sup_{X \in S_2} \mathbb{E}\|\mathcal{A}(X)\|_1 + t + 2c_2\epsilon$. Rearranging yields the bound

$$c_2 \leq \frac{\frac{1}{m}\sup_{X \in S_2}\mathbb{E}\|\mathcal{A}(X)\|_1 + t}{1 - 2\epsilon}.$$

Assuming that $\epsilon \leq 1/4$, we further deduce that

$$c_2 \leq \bar{\sigma} := \frac{2}{m}\sup_{X \in S_2}\mathbb{E}\|\mathcal{A}(X)\|_1 + 2t, \tag{4.14}$$

establishing that the random variable c_2 is bounded by $\bar{\sigma}$ in the event $\mathcal{E}$.

Now let $\hat{\mathcal{I}}$ denote either $\hat{\mathcal{I}} = \emptyset$ or $\hat{\mathcal{I}} = \mathcal{I}$. We now provide a uniform lower bound on $\frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1$. Indeed,

$$\frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1$$

$$= \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X_\star) + \mathcal{A}_{\hat{\mathcal{I}}^c}(X - X_\star)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(X_\star) + \mathcal{A}_{\hat{\mathcal{I}}}(X - X_\star)\|_1$$

$$\geq \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X_\star)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(X_\star)\|_1 - \frac{1}{m}\|\mathcal{A}(X - X_\star)\|_1 \tag{4.15}$$

$$\geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X_\star)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X_\star)\|_1\right] - t - \frac{1}{m}\|\mathcal{A}(X - X_\star)\|_1 \tag{4.16}$$

$$\geq \frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1\right] - t - \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(X - X_\star)\|_1 + \|\mathcal{A}(X - X_\star)\|_1\right) \tag{4.17}$$

$$\geq \frac{1}{m}\mathbb{E}\left[\|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1\right] - t - 2\bar{\sigma}\epsilon, \tag{4.18}$$

where (4.15) uses the forward and reverse triangle inequalities, (4.16) follows from (4.8), the estimate (4.17) follows from the forward and reverse triangle inequalities, and (4.18) follows from (4.12) and (4.14). Switching the roles of $\mathcal{I}$ and $\mathcal{I}^c$ in the above sequence of inequalities, and choosing $\epsilon = t/4\bar{\sigma}$, we deduce

$$\frac{1}{m}\sup_{X \in S_2}\left|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1 - \mathbb{E}\left[\|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1\right]\right| \leq \frac{3t}{2}.$$

In particular, setting $\hat{\mathcal{I}} = \emptyset$, we deduce

$$\frac{1}{m}\sup_{X \in S_2}\left|\|\mathcal{A}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}(X)\|_1\right]\right| \leq \frac{3t}{2}$$

and therefore using (4.6), we conclude the RIP property

$$\mu_0 p_0 - \frac{3t}{2} \leq \frac{1}{m}\|\mathcal{A}(X)\|_1 \lesssim \eta^2 + \frac{3t}{2}, \qquad \forall X \in S_2. \tag{4.19}$$

Next, let $\hat{\mathcal{I}} = \mathcal{I}$ and note that

$$\frac{1}{m}\mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1\right] = \frac{|\mathcal{I}^c| - |\mathcal{I}|}{m} \cdot \mathbb{E}|\ell^\top X r| \geq \mu_0 p_0\left(1 - \frac{2|\mathcal{I}|}{m}\right).$$

Therefore every $X \in S_2$ satisfies

$$\frac{1}{m} \left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1 \right] \geq \mu_0 p_0 \left(1 - \frac{2|\mathcal{I}|}{m} \right) - \frac{3t}{2}. \tag{4.20}$$

Setting $t = \frac{2}{3} \min\{\mu_0 p_0/2, \mu_0 p_0 (1 - 2|\mathcal{I}|/m)/2\} = \frac{1}{3}\mu_0 p_0 (1 - 2|\mathcal{I}|/m)$ in (4.19) and (4.20), we deduce the claimed estimates (4.4) and (4.5). Finally, let us estimate the probability of $\mathcal{E}$. Using Lemma 4.7 and the union bound yields

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\leq \sum_{X \in \mathcal{N}} \mathbb{P}\big\{ (4.8) \text{ or } (4.9) \text{ fails at } X \big\} \\
&\leq 4|\mathcal{N}| \exp\left(-\frac{ct^2}{\eta^4} m \right) \\
&\leq 4 \left(\frac{9}{\epsilon} \right)^{2(d_1 + d_2 + 1)} \exp\left(-\frac{ct^2}{\eta^4} m \right) \\
&= 4 \exp\left(2(d_1 + d_2 + 1) \ln(9/\epsilon) - \frac{ct^2}{\eta^4} m \right)
\end{aligned}$$

where the second inequality follows from lemma C.2 and c is a constant.

Then we deduce since $1/\epsilon = 4\bar{\sigma}/t \lesssim 2 + \eta^2/(1 - 2|\mathcal{I}|/m)$.

$$\mathbb{P}(\mathcal{E}^c) \leq 4 \exp\left(c_1(d_1 + d_2 + 1) \ln\left(c_2 + \frac{c_2}{1 - 2|\mathcal{I}|/m} \right) - \frac{4c\mu_0^2 p_0^2 (1 - \frac{2|\mathcal{I}|}{m})^2}{9\eta^4} m \right).$$

Hence as long as $m \geq \frac{18\eta^4 c_1(d_1 + d_2 + 1) \ln\left(c_2 + \frac{c_2}{1 - 2|\mathcal{I}|/m} \right)}{4c\mu_0^2 p_0^2 (1 - \frac{2|\mathcal{I}|}{m})^2}$, we can be sure $\mathbb{P}(\mathcal{E}^c) \leq 4 \exp\left(-\frac{4c\mu_0^2 p_0^2 (1 - \frac{2|\mathcal{I}|}{m})^2}{18\eta^4} m \right)$.

The result follows immediately.

$\square$

Combining Theorem 4.6 with Corollary 4.5 we obtain the following guarantee.

Corollary 4.8 (Convergence guarantees). *Consider the measurement model (4.2) and suppose that model **M1** is valid. Consider the optimization problem*

$$\min_{(x,w) \in \mathcal{S}_\nu} f(w, x) = \frac{1}{m} \sum_{i=1}^m |\langle \ell_i, w \rangle \langle r_i, x \rangle - y_i|.$$

Then there exist constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$ depending only on μ_0, p_0, η such that as long as $m \geq \frac{c_1(d_1 + d_2 + 1)}{(1 - 2p_{\text{fail}})^2} \ln\left(c_2 + \frac{c_2}{1 - 2p_{\text{fail}}} \right)$ and you choose any pair (x_0, y_0) with relative error

$$\frac{\text{dist}((w_0, x_0), \mathcal{S}_\nu^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq \frac{c_6 (1 - 2p_{\text{fail}})}{4\sqrt{2}c_5(\nu + 1)}, \tag{4.21}$$

then with probability at least $1 - 4\exp\left(-c_3(1 - 2p_{\text{fail}})^2 m \right)$ the following are true.

18

1. **(Polyak subgradient)** *Algorithm 1 initialized (x_0, y_0) produces iterates that converge linearly to $\mathcal{S}_\nu^*$, that is*

$$\frac{\text{dist}^2((w_k, x_k), \mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \le \left(1 - \frac{c_6^2 (1 - 2p_{\text{fail}})^2}{32c_5^2(\nu + 1)^4}\right)^k \cdot \frac{c_6^2 (1 - 2p_{\text{fail}})^2}{32c_5^2(\nu + 1)^2} \qquad \forall k \ge 0.$$

2. **(geometric subgradient)** *Set* $\lambda := \frac{c_6^2(1 - 2p_{\text{fail}})^2 \sqrt{\|\bar{w}\bar{x}^\top\|_F}}{16\sqrt{2}c_5^2\nu(\nu + 1)^2}$ *and* $q := \sqrt{1 - \frac{c_6^2(1 - 2p_{\text{fail}})^2}{32c_5^2(\nu + 1)^4}}$.
Then the iterates x_k generated by Algorithm 2, initialized at (w_0, x_0) converge linearly:

$$\frac{\text{dist}^2((w_k, x_k), \mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \le \left(1 - \frac{c_6^2 (1 - 2p_{\text{fail}})^2}{32c_5^2(\nu + 1)^4}\right)^k \cdot \frac{c_6^2 (1 - 2p_{\text{fail}})^2}{32c_5^2(\nu + 1)^2} \qquad \forall k \ge 0.$$

3. **(prox-linear)** *Algorithm 3 with $\beta = \rho$ and initialized at (w_0, x_0) converges quadratically:*

$$\frac{\text{dist}((w_k, x_k), \mathcal{X}^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \le 2^{-2^k} \cdot \frac{c_6 (1 - 2p_{\text{fail}})}{2\sqrt{2}c_5(\nu + 1)} \qquad \forall k \ge 0.$$

Thus with high probability, if one initializes the subgradient and prox-linear methods at a pair (w_0, x_0) satisfying $\frac{\text{dist}((w_0, x_0), \mathcal{S}_\nu^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \le \frac{c_6(1 - 2p_{\text{fail}})}{4\sqrt{2}c_5(\nu + 1)}$, then the methods will converge to the optimal solution set at a dimension independent rate.

4.2.2 Assumptions C and D under Model M2

In this section, we verify Assumptions C and D under Model **M2** and an extra incoherence condition. Namely, we impose further conditions on ℓ_p/ℓ_2 singular values of L ($p \ge 1$)

$$\sigma_{p,\min}(L) = \inf_{w \in \mathbb{S}^{d-1}} \|Lw\|_p \qquad \text{and} \qquad \sigma_{p,\max}(L) := \sup_{w \in \mathbb{S}^{d-1}} \|Lw\|_p,$$

which intuitively guarantee that the entries of any vector in $\{\mathcal{A}(X) \mid \text{rank}(X) \le 2\}$ are "well-spread."

Proposition 4.9 (Measurement Model **M2**). *Assume Model **M2** and fix an arbitrary index $\mathcal{I} \subseteq \{1, \ldots, m\}$. Define the parameter*

$$\Delta := \frac{\sigma_{1,\min}(L)}{2\sqrt{\pi}m} - 2\sigma_{\infty,\max}(L)\sqrt{\frac{2}{\pi}\frac{|\mathcal{I}|}{m}},$$

and suppose $\Delta > 0$. Then there exist numerical constants $c_1, c_2, c_3 > 0$ such that with probability

$$1 - 4\exp\left(c_1(d_1 + d_2 + 1)\ln\left(c_2\left(1 + \frac{\sigma_{1,\max}(L)}{\sigma_{1,\min}(L)}\right)\right) - c_3 \cdot \frac{\sigma_{1,\min}^2(L)}{\sigma_{2,\max}^2(L)}\right),$$

every matrix $X \in \mathbb{R}^{d_1 \times d_2}$ of rank at most two satisfies

$$\frac{\sigma_{1,\min}(L)}{2\sqrt{\pi m}}\|X\|_F \leq \frac{1}{m}\|\mathcal{A}(X)\|_1 \leq \frac{(2^{5/2}+1)\sigma_{1,\max}(L)}{2\sqrt{\pi m}} \cdot \|X\|_F, \tag{4.22}$$

and

$$\frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{I}}(X)\|_1 \geq \Delta\|X\|_F. \tag{4.23}$$

Proof. The argument mirrors the proof of Proposition 4.6 and therefore we only provide a sketch. Fix a unit Frobenius norm matrix X of rank at most two. We aim to show that for any fixed $\hat{\mathcal{I}} \subseteq \{1, \ldots, m\}$, the following random variable is highly concentrated around its mean:

$$Z_{\hat{\mathcal{I}}} = \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \frac{1}{m}\|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1.$$

To that end, fix a singular value decomposition $X = s_1 u_1 v_1^\top + s_2 u_2 v_2^\top$. We then compute

$$(\mathcal{A}(X))_i = \ell_i^\top(s_1 u_1 v_1^\top + s_2 u_2 v_2^\top)r_i = s_1\langle\ell_i, u_1\rangle\widehat{r}_i^{(1)} + s_2\langle\ell_i, u_2\rangle\widehat{r}_i^{(2)},$$

where u_1 and u_2 are orthogonal, $s_1^2 + s_2^2 = 1$, and $\widehat{r}_i^{(1)}, \widehat{r}_i^{(2)}$ are i.i.d. standard normal random variables. This decomposition, together with the rotation invariance of the normal distribution, furnishes us with the following distributional equivalence:

$$(\mathcal{A}(X))_i \overset{(d)}{=} \sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2}\ \widehat{r}_i,$$

where $\widehat{r}_i$ is a standard normal random variable. Consequently, we have the following expression for the expectation:

$$\mathbb{E}[Z_{\hat{\mathcal{I}}}] = \sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i\in\hat{\mathcal{I}}^c}\sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2} - \sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i\in\hat{\mathcal{I}}}\sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2}.$$

We now upper/lower bound this expectation. The upper bound follows from the estimate

$$\mathbb{E}[Z_{\hat{\mathcal{I}}}] \leq \mathbb{E}\left[\frac{1}{m}\|\mathcal{A}(X)\|_1\right] \leq \sqrt{\frac{2}{\pi}}\frac{1}{m}(\|Lu_1\|_1 + \|Lu_2\|_1) = \frac{2^{3/2}\sigma_{1,\max}(L)}{\sqrt{\pi m}}.$$

The lower bound uses the following two dimensional inequality $\frac{\|z\|_1}{\sqrt{2}} \leq \|z\|_2 \leq \|z\|_1$, which holds for all $z \in \mathbb{R}^2$:

$$\mathbb{E}[Z_{\hat{\mathcal{I}}}] = \sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i\in\hat{\mathcal{I}}^c}\sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2} - \sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i\in\hat{\mathcal{I}}}\sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2}$$

$$= \sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i=1}^{m}\sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2} - 2\sqrt{\frac{2}{\pi}}\frac{1}{m}\sum_{i\in\hat{\mathcal{I}}}\sqrt{(s_1\langle\ell_i, u_1\rangle)^2 + (s_2\langle\ell_i, u_2\rangle)^2}$$

$$\geq \frac{1}{\sqrt{\pi m}}(|s_1|\|Lu_1\|_1 + |s_2|\|Lu_2\|_1) - 2\sqrt{\frac{2}{\pi}}\frac{|\hat{\mathcal{I}}|}{m}\max_{i=1,\ldots,m}\|\ell_i\|_2$$

$$\geq \frac{\sigma_{1,\min}(L)}{\sqrt{\pi m}} - 2\sigma_{\infty,\max}(L)\sqrt{\frac{2}{\pi}\frac{|\hat{\mathcal{I}}|}{m}}.$$

In particular, setting $\hat{\mathcal{I}} = \emptyset$, we deduce

$$\frac{\sigma_{1,\min}(L)}{\sqrt{\pi}m} \leq \frac{1}{m}\mathbb{E}\|\mathcal{A}(X)\|_1 \leq \frac{2^{3/2}\sigma_{1,\max}(L)}{\sqrt{\pi}m}. \tag{4.24}$$

To establish concentration of the random variable $Z_{\hat{\mathcal{I}}}$, we apply a standard result (Theorem C.5) on the concentration of weighted sums of mean zero independent sub-gaussian random variables. In particular, to apply Theorem C.5, we write $Y_i = |\hat{r}_i| - \mathbb{E}|\hat{r}_i|$, and define weights

$$a_i = \frac{1}{m} \cdot \begin{cases} \sqrt{(s_1\langle \ell_i, u_1\rangle)^2 + (s_2\langle \ell_i, u_2\rangle)^2} & \text{if } i \notin \hat{\mathcal{I}}, \\ -\sqrt{(s_1\langle \ell_i, u_1\rangle)^2 + (s_2\langle \ell_i, u_2\rangle)^2} & \text{if } i \in \hat{\mathcal{I}}. \end{cases}$$

Noticing that $\||\hat{r}_i| - \mathbb{E}|\hat{r}_i|\|_{\psi_2} \leq K$, where $K > 0$ is an absolute constant, and

$$\|a\|_2^2 = \frac{1}{m^2}\sum_{i=1}^m \left((s_1\langle \ell_i, u_1\rangle)^2 + (s_2\langle \ell_i, u_2\rangle)^2\right) \leq \frac{2\sigma_{2,\max}^2(L)}{m^2},$$

it follows that for any fixed unit Frobenius norm matrix X of rank at most two, with probability at least $1 - 2\exp\left(-\frac{ct^2m^2}{2K^2\sigma_{2,\max}^2(L)}\right)$, we have

$$\frac{1}{m}\left\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1\right] \leq t. \tag{4.25}$$

We have thus established concentration for any fixed X. We now proceed with a covering argument in the same way as in the proof of Theorem 4.6. To this end, choose $\epsilon \in (0, \sqrt{2})$ and let $\mathcal{N}$ be the $(\epsilon/\sqrt{2})$-net guaranteed by lemma C.2. Let $\mathcal{E}$ denote the event that the following two estimates hold for all matrices $X \in \mathcal{N}$:

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1\right]\right| \leq t,$$

$$\frac{1}{m}\left|\|\mathcal{A}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}(X)\|_1\right]\right| \leq t.$$

Throughout the proof, we will assume that the event $\mathcal{E}$ holds. By exactly the same covering argument as in Theorem 4.6, setting $\epsilon = t/4\bar{\sigma}$ with $\bar{\sigma} = \frac{2}{m}\sup_{X \in S_2}\mathbb{E}\|\mathcal{A}(X)\|_1 + 2t$, we deduce

$$\frac{1}{m}\sup_{X \in S_2}\left|\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\hat{\mathcal{I}}^c}(X)\|_1 - \|\mathcal{A}_{\hat{\mathcal{I}}}(X)\|_1\right]\right| \leq \frac{3t}{2},$$

where either $\hat{\mathcal{I}} = \emptyset$ or $\hat{\mathcal{I}} = \mathcal{I}$.

In particular, setting $\hat{\mathcal{I}} = \emptyset$ and using the bound (4.24), we deduce

$$\frac{\sigma_{1,\min}(L)}{\sqrt{\pi}m} - \frac{3t}{2} \leq \frac{1}{m}\|\mathcal{A}(X)\|_1 \leq \frac{2^{3/2}\sigma_{1,\max}(L)}{\sqrt{\pi}m} + \frac{3t}{2}$$

for all $X \in S_2$. In turn, setting $\hat{\mathcal{I}} = \mathcal{I}$ we deduce

$$\frac{1}{m}\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \frac{1}{m}\|\mathcal{A}_{\mathcal{I}}(X)\|_1 \geq \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1\right] - \frac{3t}{2}$$

$$\geq \frac{\sigma_{1,\min}(L)}{\sqrt{\pi}m} - 2\sigma_{\infty,\max}(L)\sqrt{\frac{2}{\pi}}\frac{|\mathcal{I}|}{m} - \frac{3t}{2}.$$

21

Setting $t := \frac{\sigma_{1,\min}(L)}{3\sqrt{\pi m}}$, the estimates (4.22) and (4.23) follow immediately. Finally, estimating the probability of $\mathcal{E}$ using the union bound quickly yields:

$$\mathbb{P}(\mathcal{E}^c) \leq 4 \exp\left(c_1(d_1 + d_2 + 1) \ln\left(c_2\left(1 + \frac{\sigma_{1,\max}(L)}{\sigma_{1,\min}(L)} \right) \right) - c_3 \cdot \frac{\sigma_{1,\min}^2(L)}{\sigma_{2,\max}^2(L)} \right).$$

The result follows. $\qquad\square$

5 Initialization

Previous sections have focused on local convergence guarantees under various statistical assumptions. In particular, under Assumptions C and D, one must initialize the local search procedures at a point (w, x), whose relative distance to the solution set $\frac{\text{dist}((x,w),\mathcal{S}_\nu^*)}{\sqrt{\|\bar{x}\bar{w}^\top\|_F}}$ is upper bounded by a constant. In this section, we present a new spectral initialization routine (Algorithm 4) that is able to efficiently find such point (w, x). The algorithm is inspired by [25, Section 4] and [52].

Before describing the intuition behind the procedure, let us formally introduce our assumptions. Throughout this section, we make the following assumption on the data generating mechanism, which is stronger than Model **M1**:

$\overline{\textbf{M1}}$ The entries of matrices L and R are i.i.d. Gaussian.

Our arguments rely heavily on properties of the Gaussian distribution. We note, however, that our experimental results suggest that Algorithm 4 provides high-quality initializations under weaker distributional assumptions.

Recall that in the previous sections, the noise ξ was arbitrary. In this section, however, we must assume more about the nature of the noise. We will consider two different settings.

N1 The measurement vectors $\{(\ell_i, r_i)\}_{i=1}^m$ and the noise sequence $\{\xi_i\}_{i=1}^m$ are independent.

N2 The inlying measurement vectors $\{(\ell_i, r_i)\}_{i \in \mathcal{I}_{\text{in}}}$ and the corrupted observations $\{\xi_i\}_{i \in \mathcal{I}_{\text{out}}}$ are independent.

The noise models **N1** and **N2** differ in how an adversary may choose to corrupt the measurements. Model **N1** allows an adversary to corrupt the signal, but does not allow observation of the measurement vectors $\{(\ell_i, r_i)\}_{i=1}^m$. On the other hand, Model **N2** allows an adversary to observe the outlying measurement vectors $\{(\ell_i, r_i)\}_{i \in \mathcal{I}_{\text{out}}}$ and arbitrarily corrupt those measurements. For example, the adversary may replace the outlying measurements with those taken from a completely different signal: $y_i = \left(\mathcal{A}(\widetilde{w}\widetilde{x}^\top) \right)_i$ for $i \in \mathcal{I}_{\text{out}}$.

We can now describe the intuition underlying Algorithm 4. Throughout we denote unit vectors parallel to $\bar{w}$ and $\bar{x}$ by $\bar{w}_\star$ and $\bar{x}_\star$, respectively. Algorithm 4 exploits the expected near orthogonality of the random vectors ℓ_i and r_i to the directions $\bar{w}_\star$ and $\bar{x}_\star$, respectively, in order to select a "good" set of measurement vectors. Namely, since $\mathbb{E}\left[\langle \ell_i, \bar{w}_\star \rangle\right] = \mathbb{E}\left[\langle r_i, \bar{x}_\star \rangle\right] = 0$, we expect minimal eigenvectors of L^{init} and R^{init} to be near $\bar{w}_\star$ and $\bar{x}_\star$, respectively. Since our measurements are bilinear, we cannot necessarily select vectors for which $|\langle \ell_i, \bar{w}_\star \rangle|$ and $|\langle r_i, \bar{x}_\star \rangle|$ are both small, rather, we may only select vectors

---
Algorithm 4: Initialization.

    Data: $y \in \mathbb{R}^m, L \in \mathbb{R}^{m \times d_1}, R \in \mathbb{R}^{m \times d_2}$

    $\mathcal{I}^{\mathrm{sel}} \leftarrow \{i \mid |y_i| \leq \mathtt{med}(|y|)\}$

    Form directional estimates:

$$L^{\mathrm{init}} \leftarrow \tfrac{1}{m} \sum_{i \in \mathcal{I}^{\mathrm{sel}}} \ell_i \ell_i^\top, \quad R^{\mathrm{init}} \leftarrow \tfrac{1}{m} \sum_{i \in \mathcal{I}^{\mathrm{sel}}} r_i r_i^\top$$

$$\widehat{w} \leftarrow \operatorname{argmin}_{p \in \mathbb{S}^{d_1-1}} p^\top L^{\mathrm{init}} p, \quad \text{and} \quad \widehat{x} \leftarrow \operatorname{argmin}_{q \in \mathbb{S}^{d_2-1}} q^\top R^{\mathrm{init}} q.$$

    Estimate the norm of the signal:

$$\widehat{M} \leftarrow \operatorname*{argmin}_{\beta \in \mathbb{R}} G(\beta) := \frac{1}{m} \sum_{i=1}^m |y_i - \beta \langle \ell_i, \widehat{w} \rangle \langle r_i, \widehat{x} \rangle|,$$

$$w_0 \leftarrow \operatorname{sign}(\widehat{M}) \left|\widehat{M}\right|^{1/2} \widehat{w}, \quad \text{and} \quad x_0 \leftarrow \left|\widehat{M}\right|^{1/2} \widehat{x}.$$

    return (w_0, x_0)

---

for which the product $|\langle \ell_i, \bar{w}_\star \rangle \langle r_i, \bar{x}_\star \rangle|$ is small, leading to subtle ambiguities not present in [25, Section 4] and [52]; see Figure 1. Corruptions add further ambiguities since the noise model **N2** allows a constant fraction of measurements to be adversarially modified.
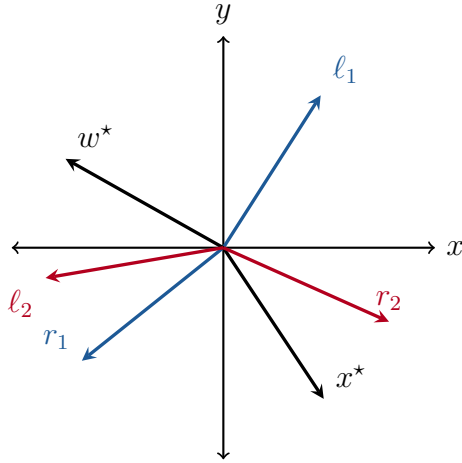

Figure 1: Intuition behind spectral initialization. The pair ℓ_1, r_1 will be included since both vectors are almost orthogonal to the true directions. ℓ_2, r_2 is unlikely to be included since r_2 is almost aligned with $x^\star$.

Formally, Algorithm 4 estimates an initial signal (w_0, x_0) in two stages: first it constructs a pair of directions $(\widehat{w}, \widehat{x})$ which estimate the true directions

$$\bar{w}_\star := \frac{1}{\|\bar{w}\|_2} \bar{w} \quad \text{and} \quad \bar{x}_\star := \frac{1}{\|\bar{x}\|_2} \bar{x}$$

23

(up to sign); then it constructs an estimate $\widehat{M}$ of the signed signal norm $\pm M$, which corrects for sign errors in the first stage. We now discuss both stages in more detail, starting with the direction estimate. Most proofs will be deferred to Appendix B. The general proof strategy we follow is analogous to [25, Section 4] for phase retrieval, with some subtle modifications due to asymmetry.

Direction Estimate. In the first stage of the algorithm, we estimate the directions $\bar{w}_\star$ and $\bar{x}_\star$, up to sign. Key to our argument is the following decomposition for model **N1** (which will be proved in Appendix B.1):

$$L^{\text{init}} = \frac{|\mathcal{I}^{\text{sel}}|}{m} \cdot I_{d_1} - \gamma_1 \bar{w}_\star \bar{w}_\star^\top + \Delta_L, \quad R^{\text{init}} = \frac{|\mathcal{I}^{\text{sel}}|}{m} \cdot I_{d_2} - \gamma_2 \bar{x}_\star \bar{x}_\star^\top + \Delta_R,$$

where $\gamma_1, \gamma_2 \gtrsim 1$ and the matrices Δ_L, Δ_R have small operator norm (decreasing with $(d_1 + d_2)/m$), with high probability. Using the Davis-Kahan $\sin\theta$ theorem [19], we can then show that the minimal eigenvectors of L^{init} and R^{init} are sufficiently close to $\{\pm\bar{w}_\star\}$ and $\{\pm\bar{x}_\star\}$, respectively.

Proposition 5.1 (Directional estimates). *There exist numerical constants $c_1, c_2, C > 0$, so that for any $p_{\text{fail}} \in [0, 1/10]$ and $t \in [0, 1]$, with probability at least $1 - c_1 \exp(-c_2 mt)$, the following hold:*

$$\min_{s \in \{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - sw^\star x^{\star\top} \right\|_F \leq \begin{cases} C \cdot \left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t \right) & \text{under Model } \textbf{N1}, \text{ and} \\ C \cdot \left(p_{\text{fail}} + \sqrt{\frac{\max\{d_1, d_2\}}{m}} + t \right) & \text{under Model } \textbf{N2}. \end{cases}$$

Norm estimate. In the second stage of the algorithm, we estimate M as well as correct the sign of the direction estimates from the previous stage. In particular, for any $(\widehat{w}, \widehat{x}) \in \mathbb{S}^{d_1-1} \times \mathbb{S}^{d_2-1}$ define the quantity

$$\delta := \left(1 + \frac{c_5}{c_6(1 - 2p_{\text{fail}})} \right) \min_{s \in \{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top \right\|_F, \tag{5.1}$$

where c_5 and c_6 are as in Theorem 4.6. Then we prove the following estimate (see Appendix B.2).

Proposition 5.2 (Norm Estimate). *Under either noise model, **N1** and **N2**, there exist numerical constants $c_1, \ldots, c_6 > 0$ so that if $m \geq \frac{c_1(d_1 + d_2 + 1)}{(1 - 2p_{\text{fail}})^2} \ln\left(c_2 + \frac{c_2}{1 - 2p_{\text{fail}}} \right)$, then with probability at least $1 - 4\exp(-c_3(1 - 2p_{\text{fail}})^2 m)$, we have that any minimizer $\widehat{M}$ of the function*

$$G(\beta) := \frac{1}{m} \sum_{i=1}^m |y_i - \beta\langle \ell_i, \widehat{w}\rangle\langle \widehat{x}, r_i\rangle|$$

satisfies $||\widehat{M}| - M| \leq \delta M$. Moreover, if in this event $\delta < 1$, then we have $\text{sign}(\widehat{M}) = \text{argmin}_{s \in \{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top \right\|_F$.

Thus, the preceding proposition shows that tighter estimates on the norm M result from better directional estimates in the first stage of Algorithm 4. In light of Proposition 5.2, we next estimate the probability of the event $\delta \leq 1/2$, which in particular implies with high probability $\text{sign}(\widehat{M}) = \text{argmin}_{s \in \{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top \right\|_F$.

Proposition 5.3 (Sign estimate). *Under either Model **N1** and **N2**, there exist numerical constants $c_0, c_1, c_2, c_3 > 0$ such that if $p_{\text{fail}} < c_0$ and $m \geq c_3(d_1 + d_2)$, then the estimate holds:[2]*

$$\mathbb{P}\left(\delta > 1/2\right) \leq c_1 \exp\left(-c_2 m\right).$$

Proof. Using Theorem 4.6 and Propositions 5.1, we deduce that for any $t \in [0, 1]$, with probability $1 - c_1 \exp\left(-c_2 mt\right)$ we have

$$\delta \leq \begin{cases} C \cdot \left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t\right) & \text{under Model } \mathbf{N1}, \text{ and} \\ C \cdot \left(p_{\text{fail}} + \sqrt{\frac{\max\{d_1, d_2\}}{m}} + t\right) & \text{under Model } \mathbf{N2}. \end{cases}$$

Thus under model **N1** it suffices to set $t = (2C)^{-2} - \frac{\max\{d_1, d_2\}}{m}$. Then the probability of the event $\delta \leq 1/2$ is at least $1 - c_1 \exp\left(-c_2((2C)^{-2}m - \max\{d_1, d_2\})\right)$. On the other hand, under model **N2**, it suffices to assume $2Cp_{\text{fail}} < 1$ and then we can set $t = (((2C)^{-1} - p_{\text{fail}})^2 - \frac{\max\{d_1, d_2\}}{m})$. The probability of the event $\delta \leq 1/2$ is then at least $1 - c_1(\exp\left(-c_2(m((2C)^{-1} - p_{\text{fail}})^2 - \max\{d_1, d_2\})\right))$. Finally using the bound $\max\{d_1, d_2\} \leq d_1 + d_2 \leq \frac{m}{c_3}$ yields the result. $\square$

Step 3: Final estimate. Putting the directional and norm estimates together, we arrive at the following theorem.

Theorem 5.4. *There exist numerical constants $c_0, c_1, c_2, c_3, C > 0$ such that if $p_{\text{fail}} \leq c_0$ and $m \geq c_4(d_1 + d_2)$, then for all $t \in [0, 1]$, with probability at least $1 - c_1 \exp\left(-c_3 mt\right)$, we have*

$$\frac{\left\| w_0 x_0^\top - \bar{w}\bar{x}^\top \right\|_F}{\left\| \bar{w}\bar{x}^\top \right\|_F} \leq \begin{cases} C \cdot \left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t\right) & \text{under Model } \mathbf{N1}, \text{ and} \\ C \cdot \left(p_{\text{fail}} + \sqrt{\frac{\max\{d_1, d_2\}}{m}} + t\right) & \text{under Model } \mathbf{N2}. \end{cases}$$

Proof. Suppose that we are in the events guaranteed by Propositions 5.1, 5.2, and 5.3. Then noting that

$$w_0 = \text{sign}(\widehat{M})|\widehat{M}|^{1/2}\widehat{w}, \ x_0 = |\widehat{M}|^{1/2}\widehat{x},$$

---

[2]In the case of model **N1**, one can set $c_0 = 1/10$.

we find that

$$
\begin{aligned}
\left\| w_0 x_0^\top - \bar{w}\bar{x}^\top \right\|_F &= \left\| \operatorname{sign}(\widehat{M})|\widehat{M}|\widehat{w}\widehat{x}^\top - M\bar{w}_\star \bar{x}_\star^\top \right\|_F \\
&= M \left\| \widehat{w}\widehat{x}^\top - \operatorname{sign}(\widehat{M})\bar{w}_\star \bar{x}_\star^\top + \frac{|\widehat{M}| - M}{M}\widehat{w}\widehat{x}^\top \right\|_F \\
&\leq M \left\| \widehat{w}\widehat{x}^\top - \operatorname{sign}(\widehat{M})w^\star x^{\star\top} \right\|_F + M\delta \\
&= M \cdot \left(2 + \frac{c_5}{c_6(1 - 2p_{\text{fail}})} \right) \min_{s \in \{\pm 1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star \bar{x}_\star^\top \right\|_F ,
\end{aligned}
$$

where c_5 and c_6 are defined in Theorem 4.6. Appealing to Proposition 5.1, the result follows.
$\qquad\square$

Combining Corollary 4.5 and Theorem 5.4, we arrive at the following guarantee for the stage procedure.

Corollary 5.5 (Efficiency estimates). *Suppose either of the models **N1** and **N2**. Let (w_0, x_0) be the output of the initialization Algorithm 4. Set $\widehat{M} = \|w_0 x_0^\top\|_F$ and consider the optimization problem*

$$
\min_{\|x\|_2, \|w\|_2 \leq \sqrt{2\widehat{M}}} g(w, x) = \frac{1}{m}\|\mathcal{A}(wx^\top) - y\|_1. \tag{5.2}
$$

Set $\nu := \sqrt{\frac{2\widehat{M}}{M}}$ and notice that the feasible region of (5.2) coincides with $\mathcal{S}_\nu$. Then there exist constants $c_0, c_1, c_2, c_3, c_5 > 0$ and $c_4 \in (0, 1)$ such that as long as $m \geq c_3(d_1 + d_2)$ and $p_{\text{fail}} \leq c_0$, the following properties hold with probability $1 - c_1 \exp(-c_2 m)$.[3]

1. *(**subgradient**) Both Algorithms 1 and 2 (with appropriate λ, q) initialized (x_0, y_0) produce iterates that converge linearly to $\mathcal{S}_\nu^*$, that is*

$$
\frac{\operatorname{dist}^2((w_k, x_k), \mathcal{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leq c_4 (1 - c_4)^k \qquad \forall k \geq 0.
$$

2. *(**prox-linear**) Algorithm 3 initialized at (w_0, x_0) (with appropriate $\beta > 0$) converges quadratically:*

$$
\frac{\operatorname{dist}((w_k, x_k), \mathcal{S}_\nu^*))}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leq c_5 \cdot 2^{-2^k} \qquad \forall k \geq 0.
$$

Proof. We provide the proof under model **N1**. The proof under model **N2** is completely analogous. Combining Proposition 5.2, Proposition 5.3, and Theorem 5.4, we deduce that there exist constants c_0, c_1, c_2, c_3, C such that as long as $m \geq c_3(d_1 + d_2)$ and $p_{\text{fail}} < c_0$, then for any $t \in [0, 1]$, with probability $1 - c_1 \exp(-c_2 mt)$, we have

$$
\left| \frac{\widehat{M}}{M} - 1 \right| \leq \delta \leq \frac{1}{2}, \tag{5.3}
$$

---

[3]In the case of model **N1**, one can set $c_0 = 1/10$.

and

$$\frac{\left\|w_0 x_0^\top - \bar{w}\bar{x}^\top\right\|_F}{M} \le C\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t.$$

In particular, notice from (5.3) that $1 \le \nu \le \sqrt{3}$ and therefore the feasible region $\mathcal{S}_\nu$ contains an optimal solution of the original problem (1.3). Using Proposition 4.2, we have

$$\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F \ge \frac{\sqrt{M}}{2\sqrt{2}(\nu+1)}\mathrm{dist}\big((w_0, x_0), \mathcal{S}_\nu^*\big).$$

Combining the estimates, we conclude

$$\frac{\mathrm{dist}((w_0, x_0), \mathcal{S}_\nu^*)}{\sqrt{M}} \le 2\sqrt{2}(\nu+1)\frac{\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F}{M} \le 2\sqrt{2}(\nu+1)C\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t.$$

Thus to ensure the relative error assumption (4.21), it suffices to ensure the inequality

$$2\sqrt{2}(\nu+1)C\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t \le \frac{c_6\,(1 - 2p_{\mathrm{fail}})}{4\sqrt{2}c_5(\nu+1)},$$

where c_5, c_6 are the constants from Corollary 4.8. Using the bound $\nu \le \sqrt{3}$, it suffices to set

$$t = \left(\frac{c_6(1 - 2p)}{16\sqrt{3}c_5 C}\right)^2 - \frac{\max\{d_1, d_2\}}{m}.$$

Thus the probability of the desired event becomes $1 - c_2(\exp\left(-c_3(c_4 m - \max\{d_1, d_2\})\right)$ for some constant c_4. Finally, using the bound $\max\{d_1, d_2\} \le d_1 + d_2 \le \frac{m}{c_3}$ and applying Corollary 4.8 completes the proof. $\qquad\square$

6 Numerical Experiments

In this section we demonstrate the performance and stability of the prox-linear and subgradient methods, and the initialization procedure, when applied to real and artificial instances of Problem (1.3). All experiments were performed using the `Julia` [7] programming language.

Subgradient method implementation. Implementation of the subgradient method for Problem (1.3) is simple, and has low per-iteration cost. Indeed, one may simply choose the subgradient

$$\frac{1}{m}\sum_{i=1}^m \mathrm{sign}(\langle \ell_i, w\rangle\langle x, r_i\rangle - y)\left(\langle x, r_i\rangle\begin{bmatrix}\ell_i \\ 0\end{bmatrix} + \langle \ell_i, w\rangle\begin{bmatrix}0 \\ r_i\end{bmatrix}\right) \in \partial f(w, x),$$

where $\mathrm{sign}(t)$ denotes the sign of t, with the convention $\mathrm{sign}(0) = 0$. The cost of computing this subgradient is on the order of four matrix multiplications. When applying Algorithm 2, choosing the correct parameters is important, since its convergence is especially sensitive to the value of the step-size decay q; the experiment described in Section 6.1.2, which aided us empirically in choosing q for the rest of the experiments, demonstrates this phenomenon. Setting $\lambda = 1.0$ seemed to suffice for all the experiments depicted hereafter.

Prox-linear method implementation. Recall that the convex models used by the prox-linear method take the form:

$$f_{(w_k,x_k)}(w,x) = \frac{1}{m}\|\mathcal{A}(w_k x_k^\top + w_k(x-x_k)^\top + (w-w_k)x_k^\top) - y\|_1 \tag{6.1}$$

Equivalently, one may rewrite this expression as a Least Absolute Deviation (LAD) objective:

$$f_{(w_k,x_k)}(w,x) = \frac{1}{m}\sum_{i=1}^{m}\left| \underbrace{\left(\langle x_k, r_i\rangle \ell_i^\top \mid \langle \ell_i, w_k\rangle r_i^\top \right)}_{=:A_i} \underbrace{\begin{pmatrix} w - w_k \\ x - x_k \end{pmatrix}}_{=:z} - \underbrace{(y_i - \langle \ell_i, w_k\rangle\langle x_k, r_i\rangle)}_{=:\tilde{y}_i} \right|$$

$$= \frac{1}{m}\|Az - \tilde{y}\|_1.$$

Thus, each iteration of Algorithm 3 requires solving a strongly convex optimization problem:

$$z_{k+1} = \operatorname*{argmin}_{z\in\mathcal{S}_\nu}\left\{ \frac{1}{m}\|Az - \tilde{y}\|_1 + \frac{1}{2\alpha}\|z\|_2^2 \right\}.$$

Motivated by the work of [25] on robust phase retrieval, we solve this subproblem with the *graph splitting* variant of the Alternating Direction Method of Multipliers, as described in [41]. This iterative method applies to problems of the form

$$\min_{z\in\mathcal{X}}\quad \frac{1}{m}\|t - \tilde{y}\|_1 + \frac{1}{2\alpha}\|z\|_2^2$$
$$\text{s.t.}\quad t = Az.$$

Yielding the following subproblems, which are repeatedly executed:

$$z' \leftarrow \operatorname*{argmin}_{z\in\mathcal{S}_\nu}\left\{ \frac{1}{2\alpha}\|z\|_2^2 + \frac{\rho}{2}\|z - (z_k - \lambda_k)\|_2^2 \right\}$$

$$t' \leftarrow \operatorname*{argmin}_{t}\left\{ \frac{1}{m}\|t - \tilde{y}\|_1 + \frac{\rho}{2}\|t - (t_k - \nu_k)\|_2^2 \right\}$$

$$\begin{pmatrix} z_+ \\ t_+ \end{pmatrix} \leftarrow \begin{bmatrix} I_{d_1+d_2} & A^\top \\ A & -I_m \end{bmatrix}^{-1}\begin{bmatrix} I_{d_1+d_2} & A^\top \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{pmatrix} z'+\lambda \\ t'+\nu \end{pmatrix}$$

$$\lambda_+ \leftarrow \lambda + (z' - z_+), \nu_+ \leftarrow \nu + (t' - t_+),$$

where $\lambda \in \mathbb{R}^{d_1+d_2}$ and $\nu \in \mathbb{R}^m$ are dual multipliers and $\rho > 0$ is a control parameter. Each above step may be computed analytically. We found in our experiments that choosing $\alpha = 1$ and $\rho \sim \frac{1}{m}$ yielded fast convergence. Our stopping criteria for this subproblem is considered met when the primal residual satisfies $\|(z_+, t_+) - (z, t)\| \leq \epsilon_k \cdot \left(\sqrt{d_1 + d_2} + \max\{\|z\|_2, \|t\|_2\}\right)$ and the dual residual satisfies $\|(\lambda_+, \nu_+) - (\lambda, \nu)\| \leq \epsilon_k \cdot \left(\sqrt{d_1 + d_2} + \max\{\|\lambda\|_2, \|\nu\|_2\}\right)$ with $\epsilon_k = 2^{-k}$.

6.1 Artificial Data

We first illustrate the performance of the prox-linear and subgradient methods under noise model **N1** with i.i.d. standard Gaussian noise ξ_i. Both methods are initialized with Algorithm 4. We experimented with Gaussian noise of varying variances, and observed that

higher levels did not adversely affect the performance of our algorithm. This is not surprising, since the theory suggests that both the objective and the initialization procedure are robust to gross outliers. We analyze the performance with problem dimensions $d_1 \in \{400, 1000\}$ and $d_2 = 500$ and with number of measurements $m = c \cdot (d_1 + d_2)$ with c varying from 1 to 8. In Fig. 2 and 3, we have depicted how the quantity

$$\frac{\left\| w_k x_k^\top - \bar{w}\bar{x}^\top \right\|_F}{\left\| \bar{w}\bar{x}^\top \right\|_F}$$

changes per iteration for the prox-linear and subgradient methods. We conducted tests in both the moderate corruption ($p_{\text{fail}} = .25$) and high corruption ($p_{\text{fail}} = .45$) regimes. For both methods, under moderate corruption ($p_{\text{fail}} = .25$) we see that exact recovery is possible as long as $c \geq 5$. Likewise, even in high corruption regime ($p_{\text{fail}} = .45$) exact recovery is still possible as long as $c \geq 8$. We also illustrate the performance of Algorithm 1 when there is no corruption at all in Fig. 2, which converges an order of magnitude faster than Algorithm 2.

In terms of algorithm performance, we see that the prox-linear method takes few outer iterations, approximately 15, to achieve very high accuracy, while the subgradient method requires a few hundred iterations. This behavior is expected as the prox-linear method converges quadratically and the subgradient method converges linearly. Although the number of iterations of the prox-linear method is small, we demonstrate in the sequel that its total run-time, including the cost of solving subproblems, can be higher than the subgradient method.

6.1.1 Number of matrix-vector multiplications

Each iteration of the prox-linear method requires the numerical resolution of a convex optimization problem. We solve this subproblem using the *graph splitting* ADMM algorithm, as described in [41], the cost of which is dominated by the number of matrix vector products required to reach the target accuracy. The number of "inner iterations" of the prox-linear method and thus the number of matrix vector products is not determined a priori. The cost of each iteration of the subgradient method, on the other hand, is on the order of 4 matrix vector products. In the subsequent plots, we solve a sequence of synthetic problems for $d_1 = d_2 = 100$ and keep track of the total number of matrix-vector multiplications performed. We run both methods until we obtain $\frac{\left\| wx^\top - \bar{w}\bar{x}^\top \right\|_F}{\left\| \bar{w}\bar{x}^\top \right\|_F} \leq 10^{-5}$. Additionally, we keep track of the same statistics for the subgradient method. We present the results in Fig. 4. We observe that the number of matrix-vector multiplications required by the prox-linear method can be much greater than those required by the subgradient method. Additionally, they seem to be much more sensitive to the ratio $\frac{m}{d_1 + d_2}$.

6.1.2 Choice of step size decay

Due to the sensitivity of Algorithm 2 to the step size decay q, we experiment with different choices of q in order to find an empirical range of values which yield acceptable performance. To that end, we generate synthetic problems of dimension 100×100 and choose $q \in \{0.90, 0.905, \ldots, 0.995\}$, and record the average error of the final iterate after 1000 iterations of the subgradient method for different choices of $m = c \cdot (d_1 + d_2)$. The average is taken
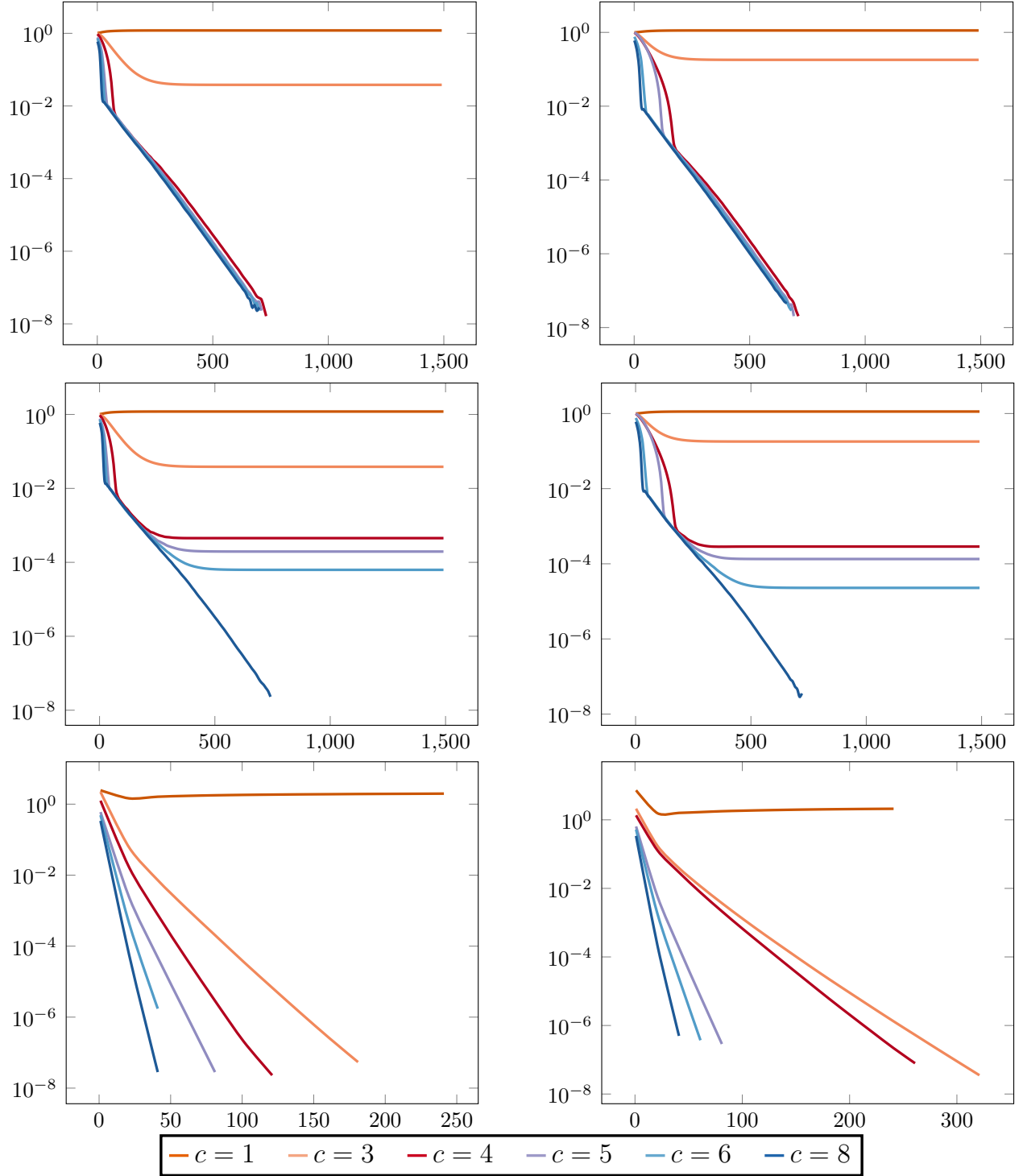
Figure 2: Dimensions are $(d_1, d_2) = (400, 500)$ in the first column and $(d_1, d_2) = (1000, 500)$ in the second column. We plot the error $\left\| w_k x_k^\top - \bar{w}\bar{x}^\top \right\|_F / \left\| \bar{w}\bar{x}^\top \right\|_F$ vs iteration count. Top row is using Algorithm 2 with $p_{\text{fail}} = 0.25$. Second row is using Algorithm 2 with $p_{\text{fail}} = 0.45$. Third row is using Algorithm 1 with $p_{\text{fail}} = 0$.
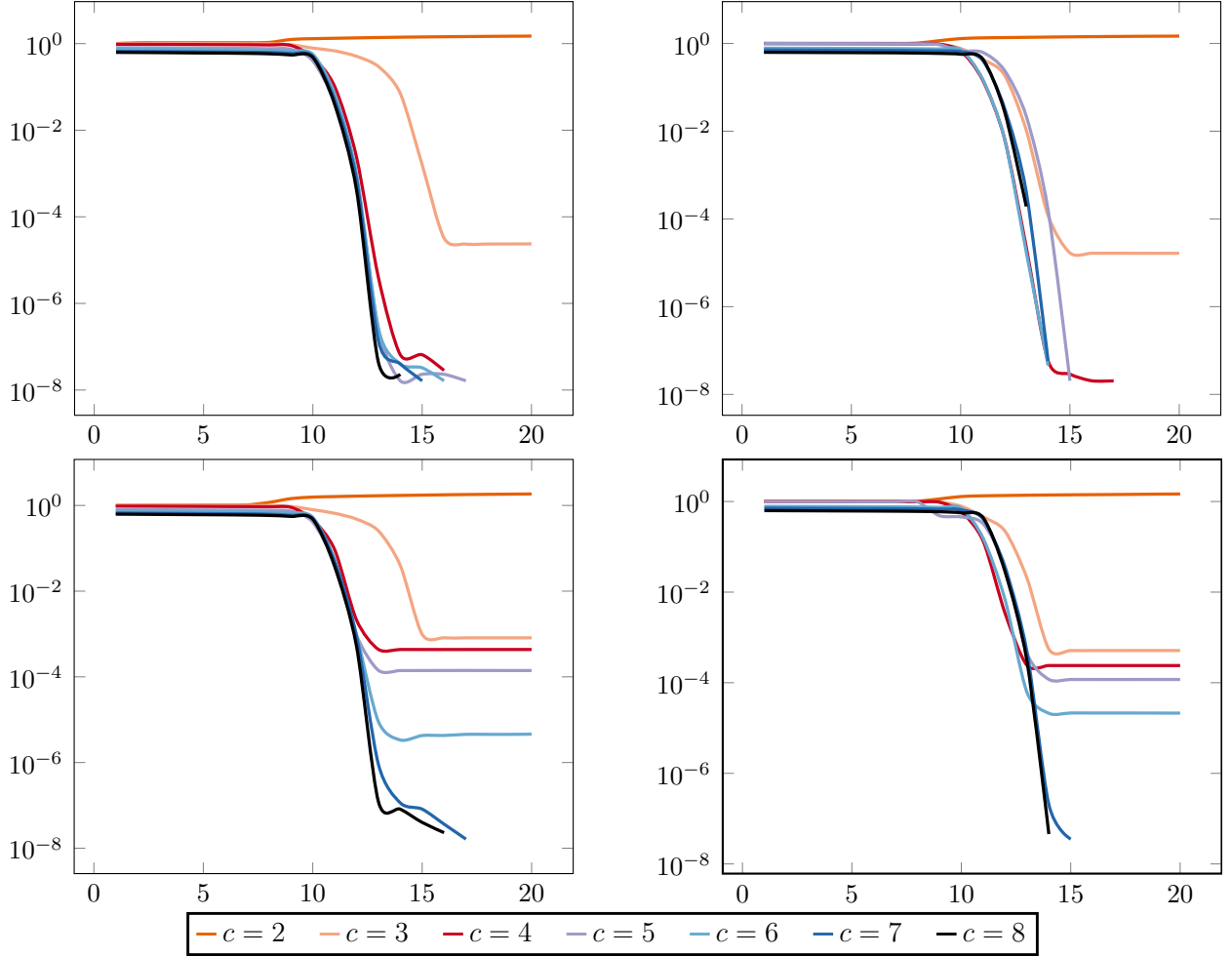
Figure 3: Dimensions are $(d_1, d_2) = (400, 500)$ in the first column and $(d_1, d_2) = (1000, 500)$ in the second column. We plot the error $\left\| w_k x_k^\top - \bar{w}\bar{x}^\top \right\|_F / \left\| \bar{w}\bar{x}^\top \right\|_F$ vs iteration count for an application of Algorithm 3 in the two settings: $p_{\text{fail}} = 0.25$ (top row) and $p_{\text{fail}} = 0.45$ (bottom row).

over 50 test runs with $\lambda = 1.0$. We test both noisy and noiseless instances to see if corruption of entries significantly changes the effective range of q. Results are shown in Fig. 5.

6.1.3 Robustness to noise

We now empirically validate the robustness of the prox-linear and subgradients algorithms to noise. In a setup familiar from other recent works [4, 25], we generate *phase transition plots*, where the x-axis varies with the level of corruption p_{fail}, the y-axis varies as the ratio $\frac{m}{d_1+d_2}$ changes, and the shade of each pixel represents the percentage of problem instances solved successfully. For every configuration $(p_{\text{fail}}, m/(d_1 + d_2))$, we run 100 experiments.

Noise model N1 - independent noise Initially, we experiment with Gaussian random matrices and $(d_1, d_2) \in \{(100, 100), (200, 200)\}$, the results for which can be found in Fig. 6.
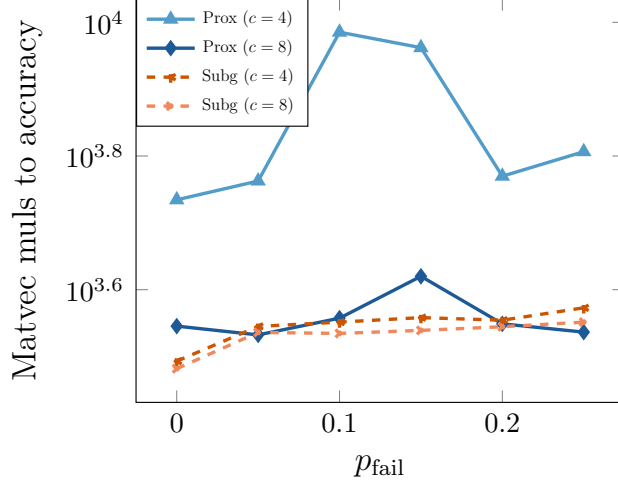
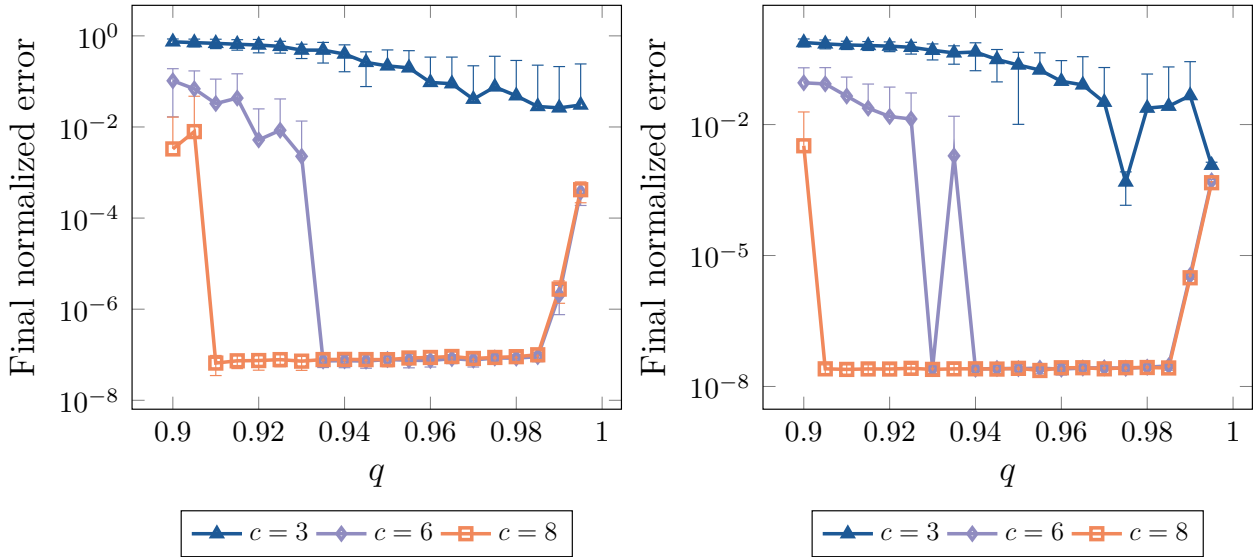Figure 4: Matrix-vector multiplications to reach rel. accuracy of 10^{-5}.

Figure 5: Final normalized error $\left\| w_k x_k^\top - \bar{w}\bar{x}^\top \right\|_F / \left\| \bar{w}\bar{x}^\top \right\|_F$ for Algorithm 2 with different choices of q, in the settings $p_{\text{fail}} = 0$ (left) and $p_{\text{fail}} = 0.25$ (right).

The phase transition plots are similar for both dimensionality choices, revealing that in the moderate independent noise regime ($p_{\text{fail}} \leq 25\%$), setting $m \geq 4(d_1 + d_2)$ suffices. On the other hand, for exact recovery in high noise regimes ($p_{\text{fail}} \simeq 45\%$), one may need to choose m as large as $8 \cdot (d_1 + d_2)$.

We repeat the same experiment in the setting where the matrix L is deterministic and has orthogonal columns of Euclidean norm $\sqrt{m}$, and R is a gaussian random matrix. Specifically, we take L to be a partial Hadamard matrix, from the first d_1 columns of an $m \times m$ Hadamard matrix. In that case, the operator $v \mapsto Lv$ can be computed efficiently in $\mathcal{O}(m \log m)$ time by 0-padding v to length m and computing its Fast Walsh-Hadamard Transform (FWHT). Additionally, the products $w \mapsto L^\top w$ can also be computed in $\mathcal{O}(m \log m)$ time by taking the FWHT of w and keeping the first d_1 coordinates of the result.

The phase transition plots can be found in Fig. 7. A comparison with the phase transition plot in Fig. 6 shows a different trend. In this case, exact recovery does not occur when the noise is above $p_{\text{fail}} \simeq 20\%$ and m is in the range $\{1, \ldots, 8\}$.

Noise model N2 - arbitrary noise   We now repeat the previous experiments, but switch to noise model **N2**. In particular, we now adversarially hide a different signal in a subset of measurements, i.e., we set

$$y_i = \begin{cases} \langle \ell_i, \bar{w} \rangle \langle \bar{x}, r_i \rangle, & i \notin \mathcal{I}_{\text{in}}, \\ \langle \ell_i, \bar{w}_{\text{imp}} \rangle \langle \bar{x}_{\text{imp}}, r_i \rangle & i \in \mathcal{I}_{\text{out}}, \end{cases}$$

where in the above $(\bar{w}_{\text{imp}}, \bar{x}_{\text{imp}}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ is an arbitrary pair of signals. Intuitively, this is a more challenging noise model than **N1**, since it allows an adversary try to trick the algorithm into recovering an entirely different signal. Our experiments confirm that this regime is indeed more difficult for the proposed algorithms, which is why we only depict the range $p_{\text{fail}} \in [0, 0.38]$ in Figs. 8 and 9 below.

6.2   Performance of initialization on real data

We now demonstrate the proposed initialization strategy on real world images. Specifically, we set $\bar{w}$ and $\bar{x}$ to be two random digits from the training subset of the MNIST dataset [30]. In this experiment, the measurement matrices $L, R \in \mathbb{R}^{(16 \cdot 784) \times 784}$ have i.i.d. Gaussian entries, and the noise follows Model **N1** with $p_{\text{fail}} = 0.45$. We apply the initialization method and plot the resulting images (initial estimates) in Fig. 10. Evidently, the initial estimates of the images are visually similar to the true digits, up to sign; in other examples, the foreground appears to be switched with the background, which corresponds to the natural sign ambiguity. Finally, we plot the normalized error for the two recovery methods (subgradient and prox-linear) in Fig. 11.

6.3   Experiments on Big Data

We apply the subgradient method for recovering large-scale real color images $W, X \in \mathbb{R}^{n \times n \times 3}$. In this setting, $p_{\text{fail}} = 0.0$ so using Algorithm 1 is applicable with $\min_{\mathcal{X}} f = 0$. We "flatten" the matrices W, X into $3n^2$ dimensional vectors w, x. In contrast to the previous experiments, our sensing matrices are of the following form:

$$L = \begin{bmatrix} HS_1 \\ \vdots \\ HS_k \end{bmatrix}, \; R = \begin{bmatrix} HS'_1 \\ \vdots \\ HS'_k \end{bmatrix},$$

where $H \in \{-1, 1\}^{d \times d} / \sqrt{d}$ is the $d \times d$ symmetric normalized Hadamard matrix and $S_i = \text{diag}(\xi_1, \ldots, \xi_d)$, $\xi \sim_{\text{i.i.d}} \{-1, 1\}$ is a diagonal random sign matrix. The same holds for S'_i. Notice that we can perform the operations $w \mapsto Lw$, $x \mapsto Rx$ in $\mathcal{O}(kd \log d)$ time: we first form the elementwise product between the signal and the random signs, and then take its Hadamard transform, which can be performed in $\mathcal{O}(d \log d)$ flops. We can efficiently compute
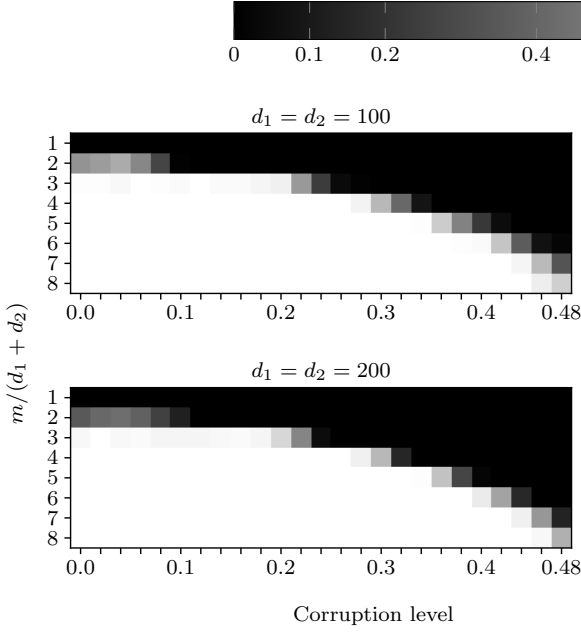
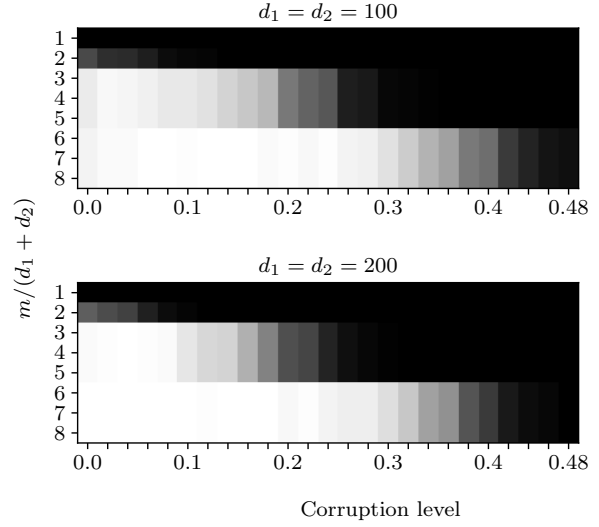Figure 6: Phase transition for **M1**, **N1**.

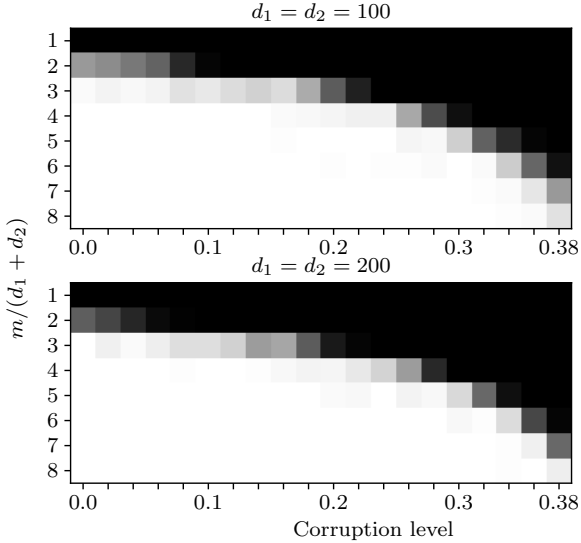Figure 7: Phase transition for **M2**, **N1**.

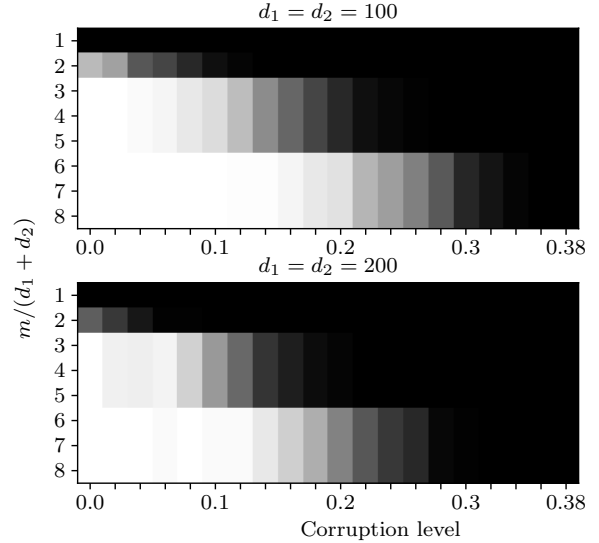Figure 8: Phase transition for **M1**, **N2**.

Figure 9: Phase transition for **M2**, **N2**.

$p \mapsto L^\top p$, $q \mapsto R^\top q$, required for the subgradient method, in a similar fashion. We recover each channel separately, which means we essentially have to solve three similar minimization problems. Notice that this results in dimensionality $d_1 = d_2 = n^2$, $m = kn^2$ for each channel.

We observed that our initialization procedure (Algorithm 4) is extremely accurate in this setting. Therefore to better illustrate the performance of the local search algorithms, we perform the following heuristic initialization. For each channel, we first sample $\widehat{w}, \widehat{x} \sim \mathbb{S}^{d-1}$, rescale by the true magnitude of the signal, and run Algorithm 1 for one step to obtain our initial estimates w_0, x_0.
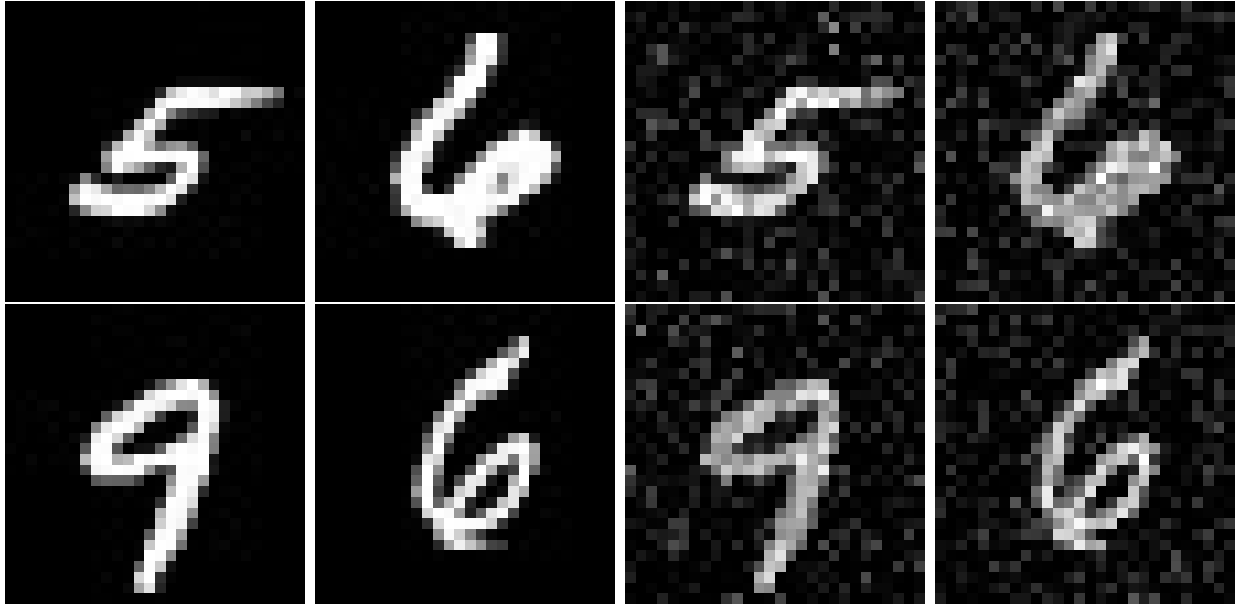
Figure 10: Digits $5, 6$ (top) and $9, 6$ (bottom). Original images are shown on the left, estimates on the right. Parameters: $p_{\text{fail}} = 0.45, m = 16 \cdot 784$.
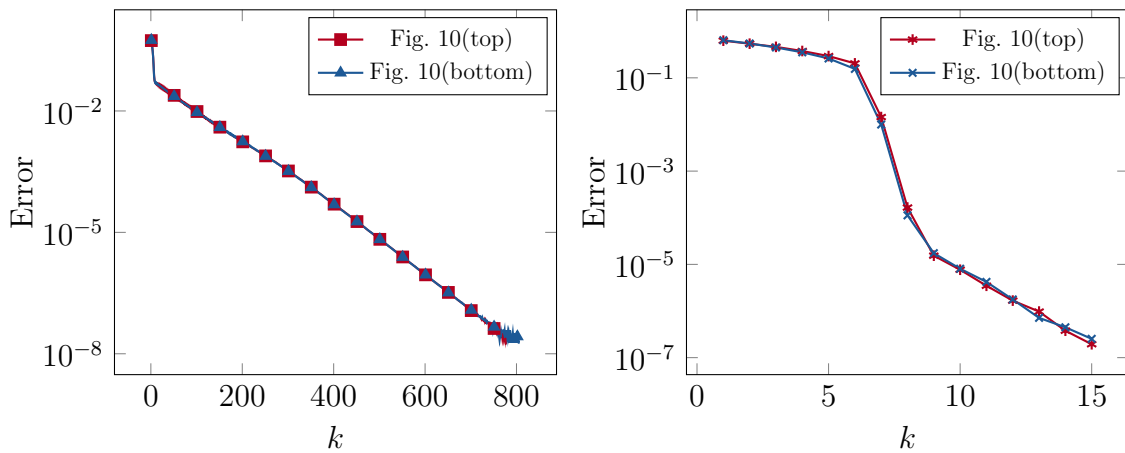

Figure 11: Relative error vs iteration count on `mnist` digits for subgradient method (left) and prox-linear method (right).

An example where we recover a pair of 512×512 color images using the Polyak subgradient method (Algorithm 1) is shown below; Fig. 12 shows the progression of the estimates w_k, up until the 90-th iteration, while Fig. 13 depicts the normalized error at each iteration for the different channels of the images.

References

[1] Alireza Aghasi, Ali Ahmed, and Paul Hand. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. *arXiv preprint arXiv:1702.04342*,
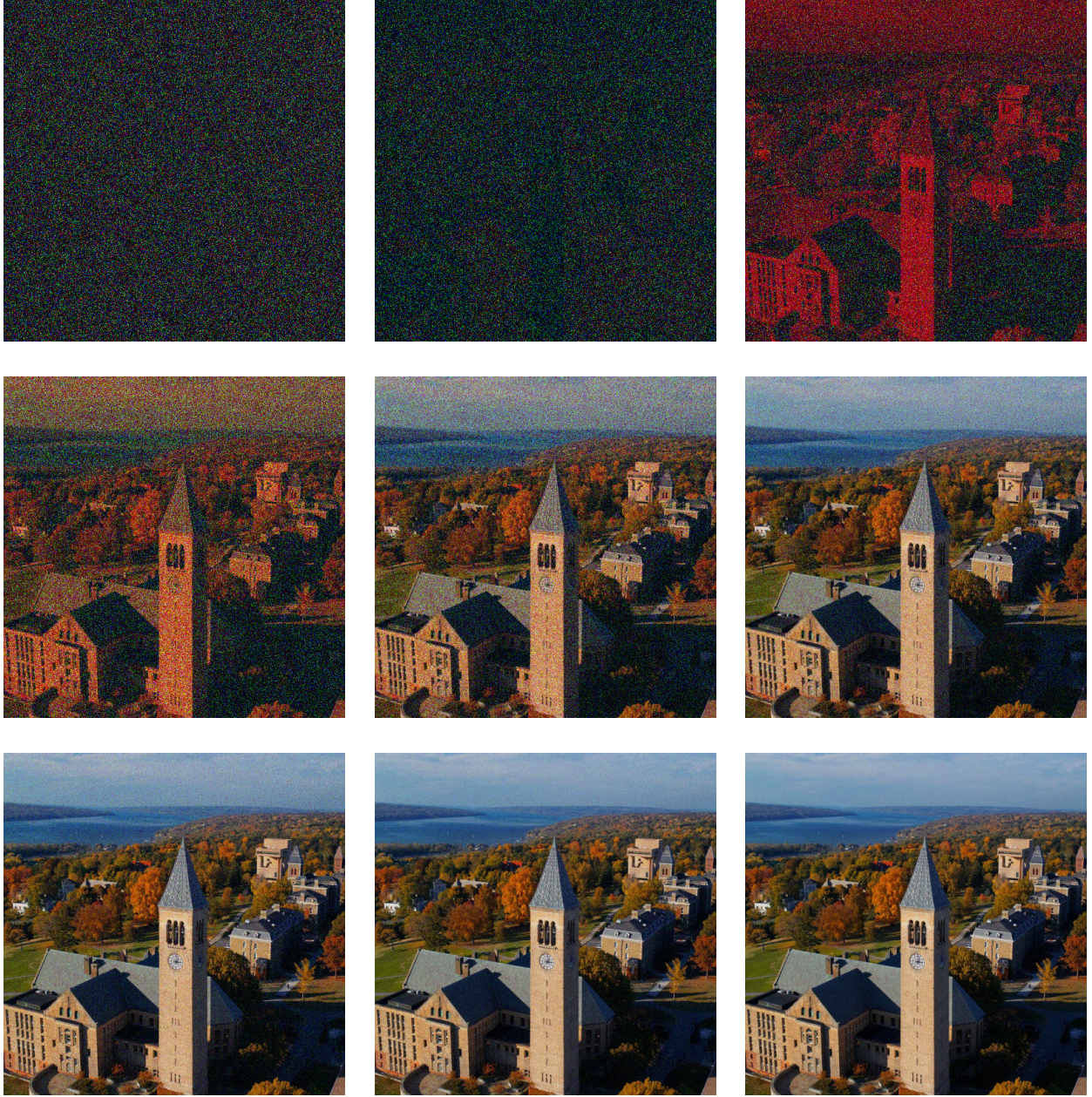
Figure 12: Iterates w_{10i}, $i = 1, \ldots, 9$. $(m, k, d, n) = (2^{22}, 16, 2^{18}, 512)$.

2017.

[2] Alireza Aghasi, Ali Ahmed, Paul Hand, and Babhru Joshi. A convex program for bilinear inversion of sparse vectors. *arXiv preprint arXiv:1809.08359*, 2018.

[3] Ali Ahmed, Alireza Aghasi, and Paul Hand. Blind deconvolutional phase retrieval via convex programming. *arXiv preprint arXiv:1806.08091*, 2018.

[4] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.
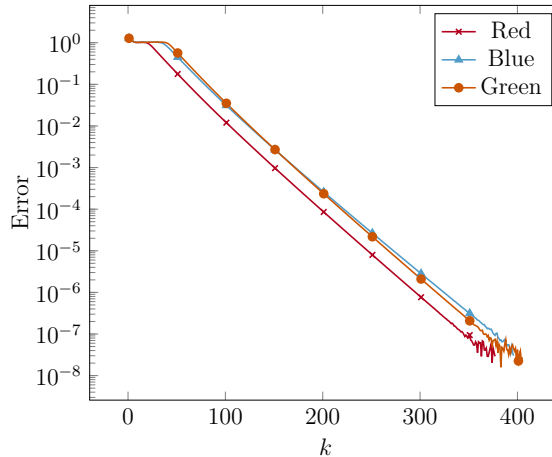
Figure 13: Normalized error for different channels in image recovery.

[5] Paolo Albano and Piermarco Cannarsa. Singularities of semiconcave functions in Banach spaces. In *Stochastic analysis, control, optimization and applications*, Systems Control Found. Appl., pages 171–190. Birkhäuser Boston, Boston, MA, 1999.

[6] Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. *arXiv preprint arXiv:1810.10702*, 2018.

[7] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[8] J.M. Borwein and Q.J. Zhu. *Techniques of Variational Analysis*. Springer Verlag, New York, 2005.

[9] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[10] J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Math. Programming*, 33(3):260–279, 1985.

[11] J.V. Burke and M.C. Ferris. A Gauss-Newton method for convex composite optimization. *Math. Programming*, 71(2, Ser. A):179–194, 1995.

[12] E.J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[13] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

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

[15] Yuxin Chen, Yuejie Chi, and Andrea J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inform. Theory*, 61(7):4034–4059, 2015.

[16] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

[17] Sunav Choudhary and Urbashi Mitra. Sparse blind deconvolution: What cannot be done. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3002–3006. IEEE, 2014.

[18] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *arXiv preprint arXiv:1601.06422*, 2016.

[19] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[20] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *arXiv preprint arXiv:1803.02461*, 2018.

[21] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[22] A.L. Dontchev and R.T. Rockafellar. *Implicit functions and solution mappings*. Monographs in Mathematics, Springer-Verlag, 2009.

[23] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Math. Oper. Res., arXiv:1602.06661*, 2016.

[24] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Preprint arXiv:1605.00125*, 2016.

[25] J.C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Preprint arXiv:1705.02356*, 2017.

[26] J. L. Goffin. On convergence rates of subgradient optimization methods. *Math. Programming*, 13(3):329–347, 1977.

[27] Wen Huang and Paul Hand. Blind deconvolution by steepest descent algorithm on a quotient manifold. *arXiv preprint arXiv:1710.03309v2*, 2018.

[28] Alexander D. Ioffe. *Variational analysis of regular mappings*. Springer Monographs in Mathematics. Springer, Cham, 2017. Theory and applications.

[29] Michael Kech and Felix Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

[30] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[31] Adrian S. Lewis and Jong-Shi Pang. Error bounds for convex inequality systems. In *Generalized convexity, generalized monotonicity: recent results (Luminy, 1996)*, volume 27 of *Nonconvex Optim. Appl.*, pages 75–110. Kluwer Acad. Publ., Dordrecht, 1998.

[32] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, pages 1–46, 2015.

[33] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

[34] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transactions on Information Theory*, 62(7):4266–4275, 2016.

[35] Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. *arXiv preprint arXiv:1802.06286*, 2018.

[36] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 31, 2015.

[37] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

[38] B.S. Mordukhovich. *Variational analysis and generalized differentiation. I*, volume 330 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2006. Basic theory.

[39] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, 27(4):538–557, 2012.

[40] E. A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[41] Neal Parikh and Stephen Boyd. Block splitting for distributed optimization. *Mathematical Programming Computation*, 6(1):77–102, 2014.

[42] J.-P. Penot. *Calculus without derivatives*, volume 266 of *Graduate Texts in Mathematics*. Springer, New York, 2013.

[43] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[44] R.T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg, 1982.

[45] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis.* Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[46] S. Rolewicz. On paraconvex multifunctions. In *Third Symposium on Operations Research (Univ. Mannheim, Mannheim, 1978), Section I*, volume 31 of *Operations Res. Verfahren*, pages 539–546. Hain, Königstein/Ts., 1979.

[47] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, May 2015.

[48] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inform. Theory*, 62(11):6535–6579, 2016.

[49] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 964–973. JMLR.org, 2016.

[50] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

[51] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press, 2018.

[52] G. Wang, G.B. Giannakis, and Y.C. Eldar. Solving systems of random quadratic equations via a truncated amplitude flow. *arXiv:1605.08285*, 2016.

Appendix A   Sharpness

A.1   Proof of Proposition 4.2

Without loss of generality, we assume that $M = 1$ (by rescaling) and that $\bar{w} = e_1 \in \mathbb{R}^{d_1}$ and $\bar{x} = e_1 \in \mathbb{R}^{d_2}$ (by rotation invariance). Recall that the distance to $\mathcal{S}_\nu^*$ may be written succinctly as

$$\text{dist}((w, x), \mathcal{S}_\nu^*) = \sqrt{\inf_{(1/\nu) \leq |\alpha| \leq \nu} \{\|w - \alpha\bar{w}\|_2^2 + \|x - (1/\alpha)\bar{x}\|_2^2\}}.$$

Before we establish the general result, we first consider the simpler case, $d_1 = d_2 = 1$.

Claim 1. The following bound holds:

$$|wx - 1| \geq \frac{1}{\sqrt{2}} \cdot \sqrt{\inf_{(1/\nu) \leq |\alpha| \leq \nu} \{|w - \alpha|^2 + |x - (1/\alpha)|^2\}},$$

for all $w, x \in [-\nu, \nu]$.

Proof of Claim. Consider a pair $(w, x) \in \mathbb{R}^2$ with $|w|, |x| \leq \nu$. It is easy to see that without loss of generality, we may assume $w \geq |x|$. We then separate the proof into two cases, which are graphically depicted in Figure 14.
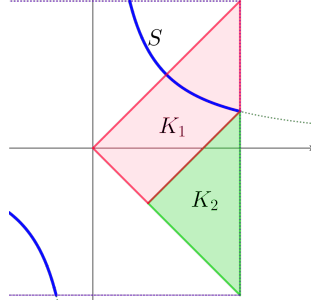
Figure 14: The regions K_1, K_2 correspond to cases 1 and 2 of the proof of Claim 1, respectively.

Case 1: $w - x \leq \frac{\nu^2 - 1}{\nu}$. In this case, we will traverse from (w, x) to the $\mathcal{S}_\nu^*$ in the direction $(1, 1)$. See Figure 14. First, consider the equation

$$wx - \sqrt{2}(w + x)t + t^2/2 = 1,$$

in the variable t and note the equality

$$wx - \sqrt{2}(w + x)t + t^2/2 = (w - t/\sqrt{2})(x - t/\sqrt{2}).$$

Using the quadratic formula to solve for t, we get

$$t = \sqrt{2}(w + x) - \sqrt{2(w + x)^2 - 2(wx - 1)}.$$

Note that the discriminant is nonnegative since $(w + x)^2 - (wx - 1) = w^2 + x^2 + xw + 1 \geq 1$.
Set $\alpha = (w - t/\sqrt{2})$ and note the identity $1/\alpha = (x - t/\sqrt{2})$. Therefore,

$$
\begin{aligned}
|wx - 1| &= |(1/\alpha)(w - \alpha) + \alpha(x - 1/\alpha) + (w - \alpha)(x - 1/\alpha)| \\
&= |(x - t/\sqrt{2})(t/\sqrt{2}) + (w - t/\sqrt{2})(t/\sqrt{2}) + t^2/2| \\
&= \frac{|t|}{\sqrt{2}}|(w + x) - t/\sqrt{2}| = \frac{|t|}{2}\sqrt{2(w + x)^2 - 2(wx - 1)} \geq \frac{|t|}{\sqrt{2}}.
\end{aligned}
$$

Observe now the equality

$$\frac{|t|}{\sqrt{2}} = \frac{1}{\sqrt{2}} \cdot (|w - \alpha|^2 + |x - 1/\alpha|^2)^{1/2}.$$

Hence it remains to bound α. First we note that $\alpha \geq 0, 1/\alpha \geq 0$, since

$$
\begin{aligned}
\alpha + 1/\alpha &= (w - t/\sqrt{2}) + (x - t/\sqrt{2}) \\
&= -(w + x) + 2\sqrt{(w + x)^2 - (wx - 1)} \geq 0.
\end{aligned}
$$

In addition, since $w \geq x$, we have $\alpha = w - t/\sqrt{2} \geq x - t/\sqrt{2} = 1/\alpha$. Since α and $1/\alpha$ are positive, we must therefore have $\alpha \geq 1 \geq 1/\nu$. Thus, it remains to verify the bound $\alpha \leq \nu$. To that end, notice that

$$1/\alpha = x - t/\sqrt{2} \geq w - t/\sqrt{2} - \frac{\nu^2 - 1}{\nu} = \alpha - \frac{\nu^2 - 1}{\nu}.$$

Therefore, $\frac{\nu^2 - 1}{\nu} \geq \frac{\alpha^2 - 1}{\alpha}$. Since the function $t \mapsto \frac{t^2 - 1}{t}$ is increasing, we deduce $\alpha \leq \nu$.

41

Case 2: $w - x \geq \frac{\nu^2 - 1}{\nu}$. In this case, we will simply set $\alpha = \nu$. Define

$$t = \left((w - \nu)^2 + (x - 1/\nu)^2\right)^{1/2}, \qquad a = \frac{w - \nu}{t}, \qquad \text{and} \qquad b = \frac{x - 1/\nu}{t}.$$

Notice that proving the desired bound amounts to showing $|wx - 1| \geq \frac{t}{\sqrt{2}}$. Observe the following estimates

$$a, b \leq 0, \qquad b \leq a, \qquad a^2 + b^2 = 1, \qquad \text{and} \qquad t \leq -\frac{1}{(a + b)}\left(\frac{\nu^2 + \nu}{\nu}\right),$$

where the the first inequality follows from the bounds $w \leq \nu$ and $\nu \geq w \geq x + \nu - 1/\nu$, second inequality follows from the bound $w - x \geq (\nu^2 - 1)/\nu$, the equality follows from algebraic manipulations, and the third inequality follows from the estimate $w + x \geq 0$. Observe

$$|wx - 1| = |(\nu + ta)(1/\nu + tb) - 1| = |t^2 ab + t\nu b + ta/\nu|.$$

Thus, by dividing through by t, we need only show that

$$|tab + \nu b + a/\nu| \geq \frac{1}{\sqrt{2}}. \tag{A.1}$$

To prove this bound, note that since $2b^2 \geq a^2 + b^2 = 1$, we have the $-\nu b - a/\nu \geq -\nu b \geq 1/\sqrt{2}$. Therefore, in the particular case when $ab = 0$ the estimate A.1 follows immediately. Define the linear function $p(s) := -(ab)s - \nu b - a/\nu$. Hence, assume $ab \neq 0$. Notice $p(0) \geq 1/\sqrt{2}$. Thus it suffices to show that the solution s^* of the equation $p(s) = 1/\sqrt{2}$ satisfies $s^* \geq t$. To see this, we compute:

$$\begin{aligned}
s^* &= -\frac{1}{ab}\left(\nu b + a/\nu + \frac{1}{\sqrt{2}}\right) \\
&= -\frac{1}{(a + b)}(a + b)\left(\frac{\nu}{a} + \frac{1}{b\nu} + \frac{1}{\sqrt{2}ab}\right) \\
&= -\frac{1}{(a + b)}\left(\nu\left(1 + \frac{b}{a}\right) + \frac{1}{\nu}\left(1 + \frac{a}{b}\right) + \frac{1}{\sqrt{2}}\left(\frac{1}{a} + \frac{1}{b}\right)\right) \\
&\geq -\frac{1}{(a + b)}\left(\nu + \frac{1}{\nu}\left(1 + \frac{a}{b} + \frac{b}{a} + \frac{1}{\sqrt{2}b} + \frac{1}{\sqrt{2}a}\right)\right) \\
&= -\frac{1}{(a + b)}\left(\nu + \frac{1}{\nu}\left(1 + \frac{\sqrt{2}(a^2 + b^2) - (|a| + |b|)}{\sqrt{2}ab}\right)\right) \\
&= -\frac{1}{(a + b)}\left(\nu + \frac{1}{\nu}\left(1 + \frac{\sqrt{2} - (|a| + |b|)}{\sqrt{2}ab}\right)\right) \\
&\geq -\frac{1}{(a + b)}\left(\nu + \frac{1}{\nu}\right) \geq t,
\end{aligned}$$

where the first inequality follows since $\nu \geq 1$ and the second inequality follows since $a^2 + b^2 = 1$ and $\sqrt{2}\|(a, b)\|_2 \geq \|(a, b)\|_1$, as desired. $\qquad\square$

Now we prove the general case. First suppose that $\|wx^\top - \bar{w}\bar{x}^\top\|_F \geq 1/2$. Since $\|w - \bar{w}\|_2 \leq (\nu + 1)$ and $\|x - \bar{x}\|_2 \leq (\nu + 1)$, we have

$$\text{dist}((w, x), \mathcal{S}_\nu^*) \leq \sqrt{2}(\nu + 1) \leq 2\sqrt{2}(\nu + 1)\|wx^\top - \bar{w}\bar{x}^\top\|_F,$$

which proves the desired bound.

On the other hand, suppose that $\|wx^\top - \bar{w}\bar{x}^\top\|_F < 1/2$. Define the two vectors:

$$\tilde{w} = (w_1, 0, \ldots, 0)^\top \in \mathbb{R}^{d_1} \qquad \text{and} \qquad \tilde{x} = (x_1, 0, \ldots, 0)^\top \in \mathbb{R}^{d_2}.$$

With this notation, we find that by Claim 1, there exists an α satisfying $(1/\nu) \leq |\alpha| \leq \nu$, such that the following holds:

$$\begin{aligned}
\|wx^\top - \bar{w}\bar{x}^\top\|_F^2 &= \|wx^\top - \tilde{w}\tilde{x}^\top + \tilde{w}\tilde{x}^\top - \bar{w}\bar{x}^\top\|_F^2 \\
&= \|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 + \|\tilde{w}\tilde{x}^\top - \bar{w}\bar{x}^\top\|_F^2 \\
&\geq \|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 + \frac{1}{2}\left(\|\tilde{w} - \alpha\bar{w}\|_F^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_F^2\right).
\end{aligned}$$

We now turn our attention to lower bounding the first term. Observe since $|w_1 x_1 - \bar{w}_1 \bar{x}_1| \leq \|wx^T - \bar{w}\bar{x}^T\|_F < 1/2$, we have

$$|w_1 x_1| \geq |\bar{w}_1 \bar{x}_1| - |w_1 x_1 - \bar{w}_1 \bar{x}_1| \geq (1/2)|\bar{w}_1 \bar{x}_1| = 1/2,$$

Moreover, note the estimates, $\nu|w_1| \geq |x_1||w_1| \geq 1/2$ and $\nu|x_1| \geq |x_1||w_1| \geq 1/2$, which imply that $|w_1| \geq 1/2\nu$ and $|x_1| \geq 1/2\nu$. Thus, we obtain the lower bound

$$\begin{aligned}
\|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 &= \|(w - \tilde{w})\tilde{x}^\top + \tilde{w}(x - \tilde{x})^\top + (w - \tilde{w})(x - \tilde{x})^\top\|_F^2 \\
&= |x_1|^2\|w - \tilde{w}\|^2 + |w_1|^2\|x - \tilde{x}\|_2^2 + \|(w - \tilde{w})(x - \tilde{x})^\top\|_F^2 \\
&\geq |x_1|^2\|w - \tilde{w}\|_2^2 + |w_1|^2\|x - \tilde{x}\|_2^2 \\
&\geq \left(\frac{1}{2\nu}\right)^2\left(\|w - \tilde{w}\|_2^2 + \|x - \tilde{x}\|_2^2\right).
\end{aligned}$$

Finally, we obtain the bound

$$\begin{aligned}
\|wx^\top - \bar{w}\bar{x}^\top\|_F^2 &\geq \|wx^\top - \tilde{w}\tilde{x}^\top\|_F^2 + \frac{1}{2}\left(\|\tilde{w} - \alpha\bar{w}\|_F^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_F^2\right) \\
&\geq \left(\frac{1}{2\nu}\right)^2\left(\|w - \tilde{w}\|_2^2 + \|x - \tilde{x}\|_2^2\right) + \frac{1}{2}\left(\|\tilde{w} - \alpha\bar{w}\|_2^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_2^2\right) \\
&\geq \min\left\{\frac{1}{2}, \left(\frac{1}{2\nu}\right)^2\right\}\left(\|w - \tilde{w}\|_2^2 + \|x - \tilde{x}\|_2^2 + \|\tilde{w} - \alpha\bar{w}\|_2^2 + \|\tilde{x} - (1/\alpha)\bar{x}\|_2^2\right) \\
&= \left(\frac{1}{2\nu}\right)^2 \cdot \text{dist}^2((w, x), \mathcal{S}_\nu^*).
\end{aligned}$$

By recalling that $1/2\nu \geq 1/2\sqrt{2}(\nu + 1)$, the proof is complete.

Appendix B   Initialization

B.1   Proof of Proposition 5.1

As stated in Section 5, we first verify that L^{init} and R^{init} are nearby matrices with minimal eigenvectors equal to $\bar{w}_\star$ and $\bar{x}_\star$. Then we apply the Davis-Kahan $\sin\theta$ theorem [19] to prove that the minimal eigenvectors of L^{init} and R^{init} must also be close to the optimal directions.

Throughout the rest of the proof, we define the sets of "selected" inliers and outliers:

$$\mathcal{I}_{\text{in}}^{\text{sel}} = \mathcal{I}_{\text{in}} \cap \mathcal{I}^{\text{sel}} \qquad \text{and} \qquad \mathcal{I}_{\text{out}}^{\text{sel}} = \mathcal{I}_{\text{out}} \cap \mathcal{I}^{\text{sel}}.$$

We record the relative size of these parameters as well, since they appear in the bounds that follow:

$$S_{\text{in}} := \frac{1}{m} \left| \mathcal{I}_{\text{in}}^{\text{sel}} \right| \qquad \text{and} \qquad S_{\text{out}} = \frac{1}{m} \left| \mathcal{I}_{\text{out}}^{\text{sel}} \right|.$$

Theorem B.1. *There exist numerical constants $c_1, c_2, c_3, c_4, c_5 > 0$, so that for any $p_{\text{fail}} \in [0, 1/10]$ and $t \in [0,1]$, with probability at least $1 - c_1(\exp(-c_2 mt))$ the following hold:*

1. *Under noise model **N1***

$$L^{\text{init}} = (S_{\text{in}} + S_{\text{out}})I_{d_1} - \gamma_1 \bar{w}_\star \bar{w}_\star^\top + \Delta_1, \quad R^{\text{init}} = (S_{\text{in}} + S_{\text{out}})I_{d_2} - \gamma_2 \bar{x}_\star \bar{x}_\star^\top + \Delta_2,$$

   where $\gamma_1 \geq c_3$ and $\gamma_2 \geq c_4$ and

$$\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\} \leq c_5 \left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t \right).$$

2. *Under noise model **N2***

$$L^{\text{init}} = S_{\text{in}} I_{d_1} - \gamma_1 \bar{w}_\star \bar{w}_\star^\top + \Delta_1, \quad R^{\text{init}} = S_{\text{in}} I_{d_2} - \gamma_2 \bar{x}_\star \bar{x}_\star^\top + \Delta_2,$$

   where $\gamma_1 \geq c_3$ and $\gamma_2 \geq c_4$ and

$$\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\} \leq p_{\text{fail}} + c_5 \left(\sqrt{\frac{\max\{d_1, d_2\}}{m}} + t \right).$$

Proof. Without loss of generality, we only prove the result for L^{init}; the result for R^{init} follows by a symmetric argument.

Define the projection operators $P_{\bar{w}_\star} := \bar{w}_\star \bar{w}_\star^\top$ and let $P_{\bar{w}_\star}^\perp := I - \bar{w}_\star \bar{w}_\star^\top$. Then decompose L^{init} into the sums of four matrices Y_0, Y_1, Y_2, Y_3, as follows:

$$L^{\text{init}} = \frac{1}{m} \Bigg(\underbrace{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}}_{m \cdot Y_0} + \underbrace{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \left(P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp + P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star} \right)}_{m \cdot Y_1} + \underbrace{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp}_{m \cdot Y_2} + \underbrace{\sum_{i \in \mathcal{I}_{\text{out}}^{\text{sel}}} \ell_i \ell_i^\top}_{m \cdot Y_3} \Bigg)..$$

$$\text{(B.1)}$$

We will now study the properties of these four matrices under both noise models.

First, note that in either case we may write $Y_0 = y_0 \bar{w}_\star \bar{w}_\star^\top$, where

$$y_0 := \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} (\ell_i^\top \bar{w}_\star)^2.$$

In addition, we will present a series of Lemmas showing the following high probability deviation bounds:

$$\gamma_1 := S_{\text{in}} - y_0 \gtrsim 1, \quad \|Y_1\|_{\text{op}} \lesssim \sqrt{\frac{d_1}{m}}, \quad \text{and} \quad \|Y_2 - S_{\text{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)\|_{\text{op}} \lesssim \sqrt{\frac{d_1}{m}}.$$

Finally, our bounds on the term Y_3 as well as the definition of Δ_1 depend on the noise model under consideration. Thus, we separate this bound into two cases:

Noise model N1. Under this noise model, we have

$$\|Y_3 - S_{\text{out}} I_{d_1}\|_{\text{op}} \lesssim \sqrt{\frac{d_1}{m}}.$$

Thus, we set $\Delta_1 = Y_1 + \left(Y_2 - S_{\text{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)\right) + \left(Y_3 - S_{\text{out}} I_{d_1}\right).$

Noise model N2. Under this noise model, we have

$$\|Y_3\|_{\text{op}} \lesssim p_{\text{fail}} + \sqrt{\frac{d_1}{m}}.$$

Thus, we set $\Delta_1 = Y_1 + (Y_2 - S_{\text{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)) + Y_3.$

Therefore, under either noise model, the result will follow immediately from the following four Lemmas. We defer the proofs for the moment.

Lemma B.2. *There exist constants $c, c_1, c_2 > 0$ such that for any $p_{\text{fail}} \in [0, 1/10]$ the following holds:*

$$\mathbb{P}\left(S_{\text{in}} - y_0 \geq c\right) \geq 1 - c_1 \exp\left(-c_2 m\right).$$

Lemma B.3. *For $t \geq 0$, we have*

$$\mathbb{P}\left(\|Y_1\|_{\text{op}} \geq 2\sqrt{\frac{d_1 - 1}{m}} + t\right) \leq \exp\left(-\frac{mt^2}{8}\right) + \exp\left(-\frac{m}{2}\right).$$

Lemma B.4. *There exist numerical constants $C, c > 0$ such that for any $t > 0$ we have*

$$\mathbb{P}\left(\left\|Y_2 - S_{\text{in}}(I_{d_1} - \bar{w}_\star \bar{w}_\star^\top)\right\|_{\text{op}} \geq C\sqrt{\frac{d_1}{m}} + t\right) \leq 2\exp(-cmt).$$

Lemma B.5. *There exist constants $C_1, C_2, c_1, c_2 > 0$ such that for any $t > 0$ the following hold. Under the noise Model **N1**, we have the estimate*

$$\mathbb{P}\left(\|Y_3 - S_{\text{out}}I_{d_1}\|_{\text{op}} \geq c_3\sqrt{\frac{d_1}{m}} + t\right) \leq 2\exp(-c_4 mt),$$

*while under the noise model **N2** we have*

$$\mathbb{P}\left(\|Y_3\|_{\text{op}} \geq p_{\text{fail}} + c_1\sqrt{\frac{d_1}{m}} + t\right) \leq 2\exp(-c_2 mt).$$

The proof of the the theorem is complete. $\qquad\square$

We now apply the Davis-Kahan $\sin\theta$ theorem [19] as stated in Lemma C.1. Throughout we assume that we are in the event described in Theorem B.1.

Proof of Proposition 5.1. We will use the notation from Theorem B.1. We only prove the result under **N1**, since the proof under **N2** is completely analogous. Define matrices $V_1 = \gamma_1 \bar{w}_\star \bar{w}_\star^\top - (S_{\text{in}} + S_{\text{out}})I_{d_1}$ and $V_2 = \gamma_2 \bar{x}_\star \bar{x}_\star^\top - (S_{\text{in}} + S_{\text{out}})I_{d_2}$. Matrix V_1 has spectral gap γ_1 and top eigenvector $\bar{w}_\star$, while matrix V_2 has spectral gap γ_2 and top eigenvector $\bar{x}_\star$. Therefore, since $-L^{\text{init}} = V_1 - \Delta_1$ and $-R^{\text{init}} = V_2 - \Delta_2$, Lemma C.1 implies that

$$\min_{s\in\{\pm 1\}} \|\widehat{w} - s\bar{w}_\star\|_2 \leq \frac{\sqrt{2}\|\Delta_1\|_{\text{op}}}{\gamma_1} \qquad \text{and} \qquad \min_{s\in\{\pm 1\}} \|\widehat{x} - s\bar{x}_\star\|_2 \leq \frac{\sqrt{2}\|\Delta_2\|_{\text{op}}}{\gamma_2}.$$

We will use these two inequalities to bound $\min_{s\in\{\pm 1\}} \|\widehat{w}\widehat{x}^\top - s\bar{w}_\star\bar{x}_\star^\top\|_F$. To do so, we need to analyze $s_1 = \text{argmin}_{s\in\{\pm 1\}} \|\widehat{w} - s\bar{w}_\star\|$ and $s_2 = \text{argmin}_{s\in\{\pm 1\}} \|\widehat{x} - s\bar{x}_\star\|$. We split the argument into two cases.

Suppose first $s_1 = s_2$. Then

$$\|\widehat{w}\widehat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\|_F = \|\widehat{w}(\widehat{x} - s_2\bar{x}_\star)^\top - (\bar{w}_\star - s_1\widehat{w})\bar{x}_\star^\top\|_F \leq \|\widehat{x} - s_2\bar{x}_\star\|_2 + \|\bar{w}_\star - s_1\widehat{w}\|_2$$

$$\leq \frac{2\sqrt{2}\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\}}{\min\{\gamma_1, \gamma_2\}},$$

as desired.

Suppose instead $s_1 = -s_2$. Then

$$\|\widehat{w}\widehat{x}^\top + \bar{w}_\star\bar{x}_\star^\top\|_F = \|\widehat{w}(\widehat{x} - s_2\bar{x}_\star)^\top + (\bar{w}_\star + s_2\widehat{w})\bar{x}_\star^\top\|_F \leq \|\widehat{x} - s_2\bar{x}_\star\|_2 + \|\bar{w}_\star - s_1\widehat{w}\|_2$$

$$\leq \frac{2\sqrt{2}\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\}}{\min\{\gamma_1, \gamma_2\}},$$

as desired. Bounding $\max\{\|\Delta_1\|_{\text{op}}, \|\Delta_2\|_{\text{op}}\}$ using Theorem B.1 completes the proof. $\qquad\square$

The next sections present the proof of Lemmas B.2-B.5. We next set up the notation. For any sequence of vectors $\{w_i\}_{i=1}^m$ in $\mathbb{R}^d$, we will use the symbol $w_{i,2:d}$ to denote the vector in $\mathbb{R}^{d-1}$ consisting of the last $d-1$ coordinates of w_i.

We will use the following two observations throughout. First, by rotation invariance we will assume, without loss of generality, that $\bar{w}_\star = e_1$ and $\bar{x}_\star = e_1$. Second, and crucially, this assumption implies that $\mathcal{I}_{\text{in}}^{\text{sel}}$ depends on $\{\ell_i\}_{i=1}^m$ only through the first component. In particular, we have that $\{\ell_{i,2:d_1}\}_{i=1}^m$ and $\mathcal{I}_{\text{in}}^{\text{sel}}$ are independent. Similarly, $\{r_{i,2:d_2}\}_{i=1}^m$ and $\mathcal{I}_{\text{in}}^{\text{sel}}$ are independent as well.

B.1.1 Proof of Lemma B.2

Our goal is to lower bound the quantity

$$S_{\text{in}} - y_0 = \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} (1 - \ell_{i,1}^2).$$

To prove a lower bound, we need to control the random variables $\ell_{i,1}^2$ on the set $\mathcal{I}_{\text{in}}^{\text{sel}}$.

Before proving the key claim, we first introduce some notation. First, define

$$q_{\text{fail}} := \frac{5 - 2p_{\text{fail}}}{8(1 - p_{\text{fail}})},$$

which is strictly less than one since $p_{\text{fail}} < 1/2$. Let $a, b \sim \mathsf{N}(0, 1)$ and define Q_{fail} to be the q_{fail}-quantile of the random variable $|ab|$. In particular, the following relationship holds

$$q_{\text{fail}} = \mathbb{P}\left(|ab| \leq Q_{\text{fail}}\right).$$

Additionally, define the conditional expected value

$$\omega_{\text{fail}} = \mathbb{E}\left[a^2 \mid |ab| \leq Q_{\text{fail}}\right].$$

Rather than analyzing $\mathcal{I}_{\text{in}}^{\text{sel}}$ directly, we introduce the following set $\mathcal{I}_{\text{in}}^Q$, which is simpler to analyze:

$$\mathcal{I}_{\text{in}}^Q := \left\{i \in \mathcal{I}_{\text{in}} \mid \left|\ell_i^\top \bar{w}_\star \bar{x}_\star^\top r_i\right| \leq Q_{\text{fail}}\right\}.$$

Then we prove the following claim.

Claim 2. There exist numerical constants $c, K > 0$ such that for all $t \geq 0$ the following inequalities hold true:

1. $\frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} \geq \frac{1 - 2p_{\text{fail}}}{2}$.

2. $\mathbb{P}\left(\mathcal{I}_{\text{in}}^Q \supseteq \mathcal{I}_{\text{in}}^{\text{sel}}\right) \geq 1 - \exp\left(-\frac{3(1 - 2p_{\text{fail}})}{160}m\right)$,

3. $\mathbb{P}\left(|\mathcal{I}_{\text{in}}^Q| \geq \frac{6251m}{10000}\right) \leq \exp\left(-\frac{m}{2 \cdot 10^8}\right)$,

4. $\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|} \sum_{i \in \mathcal{I}_{\text{in}}^Q} \ell_{i,1}^2 \geq \omega_{\text{fail}} + t\right) \leq \exp\left(-c \min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\} \frac{m(1 - 2p_{\text{fail}})}{2}\right) + \exp\left(-\frac{3(1 - 2p_{\text{fail}})}{160}m\right)$.

47

Before we prove the claim, we show it leads to the conclusion of the lemma. Assuming we are in the event

$$\mathcal{E} = \left\{ \mathcal{I}_{\text{in}}^Q \supseteq \mathcal{I}_{\text{in}}^{\text{sel}}, \ \ |\mathcal{I}_{\text{in}}^Q| < \frac{6251m}{10000}, \ \ \frac{1}{|\mathcal{I}_{\text{in}}^Q|} \sum_{i \in \mathcal{I}_{\text{in}}^Q} \ell_{i,1}^2 \le \frac{101}{100} \omega_{\text{fail}} \right\},$$

it follows that

$$S - y_0 = \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} (1 - \ell_{i,1}^2) \ge \frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} - \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^Q} \ell_{i,1}^2 \ge \frac{1 - 2p_{\text{fail}}}{2} - \frac{|\mathcal{I}_{\text{in}}^Q|}{m \left| \mathcal{I}_{\text{in}}^Q \right|} \sum_{i \in \mathcal{I}_{\text{in}}^Q} \ell_{i,1}^2$$

$$\ge \frac{1 - 2p_{\text{fail}}}{2} - \frac{631351}{1000000} \omega_{\text{fail}} \ge 0.04644344.$$

where the first three inequalities follow by the definition of the event $\mathcal{E}$. The fourth inequality follows by the definition of $\mathcal{E}$ and Lemma C.13, which implies $\omega_{\text{fail}} \le .56$ when $p_{\text{fail}} = .1$ and that the difference is minimized over $p_{\text{fail}} \in [0, .1]$ at the endpoint $p_{\text{fail}} = .1$. To get the claimed probabilities, we note that by Lemma C.13, we have $\omega_{\text{fail}} \ge .5$ for any setting of p_{fail}.

Now we prove the claim.

Proof of the Claim. We separate the proof into four parts.

Part 1. By definition, we have

$$\frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} = \frac{|\mathcal{I}_{\text{in}} \cap \mathcal{I}^{\text{sel}}|}{m} = \frac{|\mathcal{I}^{\text{sel}}| - |\mathcal{I}_{\text{out}} \cap \mathcal{I}^{\text{sel}}|}{m} \ge \frac{\frac{m}{2} - |\mathcal{I}_{\text{out}} \cap \mathcal{I}^{\text{sel}}|}{m} \ge \frac{\frac{m}{2} - mp_{\text{fail}}}{m} = \frac{1 - 2p_{\text{fail}}}{2}.$$

Part 2. By the definitions of $\mathcal{I}_{\text{in}}^{\text{sel}}$ and $\mathcal{I}_{\text{in}}^Q$, the result will follow once we show that

$$\mathbb{P}\left(\text{med}(\{|y_i|\}_i^m) \ge Q_{\text{fail}} M\right) \le \exp\left(-\frac{3(1 - 2p_{\text{fail}})}{160} m\right).$$

To that end, first note that

$$\text{med}(\{|y_i|\}_i^m) = \min\left\{ |y_j| : j \in [m], \sum_{i=1}^m \mathbf{1}\{|y_i| \le |y_j|\} \ge \frac{m}{2} \right\}$$

$$= \min\left\{ |y_j| : j \in [m], \sum_{i=1}^m \mathbf{1}\{|y_i| \le |y_j|\} \ge \frac{|\mathcal{I}_{\text{in}}|}{2(1 - p_{\text{fail}})} \right\}$$

$$\le \min\left\{ |y_j| : j \in \mathcal{I}_{\text{in}}, \sum_{i=1}^m \mathbf{1}\{|y_i| \le |y_j|\} \ge \frac{|\mathcal{I}_{\text{in}}|}{2(1 - p_{\text{fail}})} \right\}$$

$$\le \min\left\{ |y_j| : j \in \mathcal{I}_{\text{in}}, \sum_{i \in \mathcal{I}_{\text{in}}} \mathbf{1}\{|y_i| \le |y_j|\} \ge \frac{|\mathcal{I}_{\text{in}}|}{2(1 - p_{\text{fail}})} \right\}$$

$$= \text{quant}_{\frac{1}{2(1 - p_{\text{fail}})}}\left(\{|y_i|\}_{i \in \mathcal{I}_{\text{in}}}\right),$$

where the first equality follows since $\frac{|\mathcal{I}_{\text{in}}|}{2(1-p_{\text{fail}})} = \frac{(1-p_{\text{fail}})m}{2(1-p_{\text{fail}})} = m/2$, the first inequality follows since the minimum is taken over a smaller set, and the second inequality follows since the sum is taken over a smaller set of indices. Therefore, we find that

$$\mathbb{P}\left(\texttt{med}(\{|y_i|\}_i^m) \geq Q_{\text{fail}}M\right) \leq \mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|\}_{i\in\mathcal{I}_{\text{in}}}\right) \geq Q_{\text{fail}}M\right)$$

$$= \mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|/M\}_{i\in\mathcal{I}_{\text{in}}}\right) \geq Q_{\text{fail}}\right),$$

and our remaining task is to bound this probability.

To bound this probability, we apply Lemma C.9 to the i.i.d. sample $\{|y_i|/M : i \in \mathcal{I}_{\text{in}}\}$, which is sampled from the distribution of $\mathcal{D}$ of $|ab|$ where $a, b \sim \mathsf{N}(0,1)$ and a, b are independent. Therefore, using the identities (for $i \in \mathcal{I}_{\text{in}}$)

$$q = \mathbb{P}\left(|y_i|/M \leq Q_{\text{fail}}\right) = q_{\text{fail}} = \frac{5 - 2p_{\text{fail}}}{8(1 - p_{\text{fail}})}$$

and choosing $p := (2(1 - p_{\text{fail}}))^{-1} < q$, we find that

$$\mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|/M\}_{i\in\mathcal{I}_{\text{in}}}\right) \geq Q_{\text{fail}}\right) \leq \exp\left(\frac{m(q-p)^2}{2(q-p)/3 + 2q(1-q)}\right)$$

$$= \exp\left(\frac{m(q-p)}{2/3 + 6q}\right)$$

$$= \exp\left(-\frac{3(1-2p_{\text{fail}})m}{8(1-p_{\text{fail}})(2+18q)}\right)$$

$$\leq \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right),$$

where we have used the identity $q - p = \frac{1-2p_{\text{fail}}}{8(1-p_{\text{fail}})} = (1-q)/3$ in the first equality. This completes the bound and implies that $\mathcal{I}_{\text{in}}^Q \supseteq \mathcal{I}_{\text{in}}^{\text{sel}}$ with high probability, as desired.

Part 3. Since $\{|y_i|/M : i \in \mathcal{I}_{\text{in}}\}$ is an i.i.d. sample from the distribution of $|ab|$ where $a, b \sim \mathsf{N}(0,1)$ are independent, we have for each $i \in \mathcal{I}_{\text{in}}$, that

$$\mathbb{P}\left(i \in \mathcal{I}_{\text{in}}^Q\right) = \mathbb{P}\left(|y_i|/M \leq Q_{\text{fail}}\right) = \mathbb{P}\left(|ab| \leq Q_{\text{fail}}\right) = q_{\text{fail}}.$$

Therefore, $\mathbb{E}\left[|\mathcal{I}_{\text{in}}^Q|\right] = q_{\text{fail}}|\mathcal{I}_{\text{in}}| \leq \frac{5-2p_{\text{fail}}}{8(1-p_{\text{fail}})}(1 - p_{\text{fail}})m \leq \frac{5}{8}m$. Finally, we apply Hoeffding's inequality (Lemma C.3) to the i.i.d. Bernoulli random variables $\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^q\} - \mathbb{E}\left[\mathbf{1}\{i \in \mathcal{I}_{\text{in}}^q\}\right]$ ($i \in \mathcal{I}_{\text{in}}$) to deduce that

$$\mathbb{P}\left(\frac{6251m}{10000} \leq |\mathcal{I}_{\text{in}}^Q|\right) = \mathbb{P}\left(\frac{m}{10000} \leq |\mathcal{I}_{\text{in}}^Q| - \frac{5m}{8}\right) \leq \mathbb{P}\left(\frac{m}{10000} \leq |\mathcal{I}_{\text{in}}^Q| - \mathbb{E}|\mathcal{I}_{\text{in}}^Q|\right)$$

$$\leq \exp\left(-\frac{(1/10000)^2 m}{2(1-p_{\text{fail}})}\right) \leq \exp\left(-\frac{m}{2 \cdot 10^8}\right),$$

as desired.

Part 4. First write

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t\right)$$

$$= \mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t \text{ and } |\mathcal{I}_{\text{in}}^Q| \supseteq \mathcal{I}_{\text{in}}^{\text{sel}}\right) + \mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t \text{ and } \mathcal{I}_{\text{in}}^Q \not\supseteq \mathcal{I}_{\text{in}}^{\text{sel}}\right)$$

$$\leq \mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t \text{ and } |\mathcal{I}_{\text{in}}^Q| \geq \frac{m(1-2p_{\text{fail}})}{2}\right) + \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right),$$

where first inequality follows from Part 2 and the bound $\frac{|\mathcal{I}_{\text{in}}^{\text{sel}}|}{m} \geq \frac{1-2p_{\text{fail}}}{2}$. Thus, we focus on bounding the first term.

To that end, notice that

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t \text{ and } |\mathcal{I}_{\text{in}}^Q| \geq \frac{m(1-2p_{\text{fail}})}{2}\right)$$

$$= \mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t \,\middle|\, |\mathcal{I}_{\text{in}}^Q| \geq \frac{m(1-2p_{\text{fail}})}{2}\right) \mathbb{P}\left(|\mathcal{I}_{\text{in}}^Q| \geq \frac{m(1-2p_{\text{fail}})}{2}\right).$$

Observe that for any index $i \in \mathcal{I}_{\text{in}}$ and $t \geq 0$, we have $\mathbb{P}\left(\ell_{i,1}^2 \geq t \mid i \in \mathcal{I}_{\text{in}}^Q\right) = \mathbb{P}\left(a^2 \geq t \mid |ab| \leq Q_{\text{fail}}\right)$, where $a, b \sim \mathsf{N}(0,1)$ are independent. In addition, we have $q_{\text{fail}} = P(|ab| \leq Q_{\text{fail}}) = \frac{5-2p_{\text{fail}}}{8(1-p_{\text{fail}})} \geq 5/8 > 1/2$, where we have used the fact that q_{fail} is an increasing function of p_{fail}. Therefore, applying Lemma C.10, we have the following bound:

$$\mathbb{P}\left(\ell_{i,1}^2 \geq t \mid i \in \mathcal{I}_{\text{in}}^Q\right) \leq 2\exp(-t/2K_1) \qquad \text{for all } t \geq 0 \text{ and } i \in \mathcal{I}_{\text{in}},$$

where K_1 is a numerical constant. In particular, by Theorem C.7 and the identity $\omega_{\text{fail}} = \mathbb{E}\left[a^2 \geq t \mid |ab| \leq Q_{\text{fail}}\right]$, we have the following bound

$$\mathbb{P}\left(\frac{1}{|\mathcal{I}_{\text{in}}^Q|}\sum_{i\in\mathcal{I}_{\text{in}}^Q}\ell_{i,1}^2 \geq \omega_{\text{fail}} + t \,\middle|\, |\mathcal{I}_{\text{in}}^Q| > \frac{m(1-2p_{\text{fail}})}{2}\right) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\frac{m(1-2p_{\text{fail}})}{2}\right)$$

for numerical constants c and K, as desired.

$\square$

The proof is complete.

$\square$

B.1.2 Proof of Lemma B.3

Our goal is to bound the operator norm of the following matrix:

$$Y_1 = \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \left(P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp + P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star} \right) = \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1} \left(e_1 \ell_{i,2:d}^\top + \ell_{i,2:d} e_1^\top \right).$$

Simplifying, we find that

$$Y_1 = \begin{bmatrix} 0 & \lambda_{2:d_1}^\top \\ \lambda_{2:d_1} & 0 \end{bmatrix} \qquad \text{for} \qquad \lambda := \begin{bmatrix} 0 \\ \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1} \ell_{i,2:d_1} \end{bmatrix} \in \mathbb{R}^{d_1}.$$

Evidently, $\|Y_1\|_{\text{op}} \leq \|\lambda_{2:d_1}\|_2$, so our focus will be to bound this quantity. We will bound this quantity through the following claim, which is based on Gaussian concentration for Lipschitz functions.

Claim 3. Consider the (random) function $F : \mathbb{R}^{m \times (d_1 - 1)} \to \mathbb{R}$, given by

$$F(a_1, \ldots, a_m) = \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1} a_i \right\|_2.$$

Then F is $\widehat{\eta} = \frac{1}{m} \sqrt{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \ell_{i,1}^2}$ Lipschitz continuous and

$$\mathbb{P}\left(F(\ell_{1,2:d}, \ldots, \ell_{m,2:d}) \geq 2\sqrt{\frac{d_1 - 1}{m}} + t \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}} \right) \leq \exp\left(-\frac{mt^2}{8} \right).$$

Moreover, the following bound holds:

$$\mathbb{P}\left(\widehat{\eta} \geq \frac{2}{\sqrt{m}} \right) \leq \exp\left(-\frac{m}{2} \right).$$

Proof of Claim. For any $A = \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix} \in \mathbb{R}^{m \times (d_1 - 1)}$ and $B = \begin{bmatrix} b_1 & \cdots & b_m \end{bmatrix} \in \mathbb{R}^{m \times (d_1 - 1)}$, we have

$$|F(A) - F(B)| \leq \frac{1}{m} \|(A - B)(\ell_{i,1} \mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2 \leq \frac{1}{m} \|(A - B)\|_{\text{op}} \|(\ell_{i,1} \mathbf{1}\{i \in \mathcal{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2 \leq \widehat{\eta} \|A - B\|_F,$$

which proves that F is $\widehat{\eta}$-Lipschitz. Therefore, since for all i the variables $\ell_{i,1}$ and $\ell_{i,2:d_1}$ are independent, standard results on Gaussian concentration for Lipschitz functions (applied conditionally), Theorem C.8, imply that

$$\mathbb{P}\left(F(\ell_{1,2:d}, \ldots, \ell_{m,2:d}) - \mathbb{E}\left[F(\ell_{1,2:d}, \ldots, \ell_{m,2:d}) \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}} \right] \geq t \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathcal{I}_{\text{in}}^{\text{sel}} \right)$$
$$\leq \exp\left(-\frac{mt^2}{8} \right).$$

Thus, the first part of the claim is a consequence of the following bound:

$$\mathbb{E}\left[F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\;\middle|\;\widehat{\eta}<\frac{2}{\sqrt{m}},\{\ell_{1,i}\}_{i=1}^{m},\mathcal{I}_{\text{in}}^{\text{sel}}\right]\leq\sqrt{\mathbb{E}\left[F(\ell_{1,2:d},\ldots,\ell_{m,2:d})^{2}\;\middle|\;\widehat{\eta}<\frac{2}{\sqrt{m}},\{\ell_{1,i}\}_{i=1}^{m},\mathcal{I}_{\text{in}}^{\text{sel}}\right]}$$

$$=\sqrt{\frac{1}{m^{2}}\mathbb{E}\left[\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\ell_{i,1}^{2}(d_{1}-1)\;\middle|\;\hat{\eta}<\frac{2}{\sqrt{m}}\right]}\leq 2\sqrt{\frac{d_{1}-1}{m}}.$$

We now turn our attention to the high probability bound on $\widehat{\eta}$.

To that end, notice that the (random) function $E\colon\mathbb{R}^{m}\to\mathbb{R}$ given by

$$E(a)=\frac{1}{m}\sqrt{\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}a_{i}^{2}}=\frac{1}{m}\|(a_{i}\mathbf{1}\{i\in\mathcal{I}_{\text{in}}^{\text{sel}}\})_{i=1}^{m}\|_{2}.$$

is m^{-1}-Lipschitz continuous. Moreover, we have that $\mathbb{E}\left[E(\ell_{1,i},\ldots,\ell_{1,d})\right]\leq\frac{1}{m}\mathbb{E}\left[\|(\ell_{1,i})_{i=1}^{m}\|_{2}\right]\leq m^{-1/2}$. Therefore, by Gaussian concentration we have

$$\mathbb{P}\left(\widehat{\eta}\geq\frac{2}{\sqrt{m}}\right)\geq\mathbb{P}\left(E(\ell_{1,i},\ldots,\ell_{1,d})-\mathbb{E}\left[E(\ell_{1,i},\ldots,\ell_{1,d})\right]\geq\frac{1}{\sqrt{m}}\right)\leq\exp\left(-\frac{m}{2}\right),$$

as desired. $\qquad\square$

To complete the proof, observe that

$$\mathbb{P}\left(\|\lambda_{2:d_{1}}\|_{2}\geq 2\sqrt{\frac{d_{1}-1}{m}}+t\right)$$

$$=\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\ell_{i,1}\ell_{i,2:d_{1}}\right\|_{2}\geq 2\sqrt{\frac{d_{1}-1}{m}}+t\right)$$

$$\leq\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathcal{I}_{\text{in}}^{\text{sel}}}\ell_{i,1}\ell_{i,2:d_{1}}\right\|_{2}\geq 2\sqrt{\frac{d_{1}-1}{m}}+t\;\middle|\;\widehat{\eta}<\frac{2}{\sqrt{m}}\right)\mathbb{P}\left(\widehat{\eta}<\frac{2}{\sqrt{m}}\right)+\mathbb{P}\left(\widehat{\eta}\geq\frac{2}{\sqrt{m}}\right)$$

$$\leq\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\geq 2\sqrt{\frac{d_{1}-1}{m}}+t\;\middle|\;\widehat{\eta}<\frac{2}{\sqrt{m}}\right)+\exp\left(-\frac{m}{2}\right),$$

where the second inequality is due to Claim 3. Finally, by Claim 3, the conditional probability is bounded as follows

$$\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\geq 2\sqrt{\frac{d_{1}-1}{m}}+t\;\middle|\;\widehat{\eta}<\frac{2}{\sqrt{m}}\right)$$

$$=\mathbb{E}_{\mathcal{I}_{\text{in}}^{\text{sel}},\{\ell_{i,1}\}_{i=1}^{m}}\left[\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d})\geq 2\sqrt{\frac{d_{1}-1}{m}}+t\;\middle|\;\widehat{\eta}<\frac{2}{\sqrt{m}},\{\ell_{1,i}\}_{i=1}^{m},\mathcal{I}_{\text{in}}^{\text{sel}}\right)\right]$$

$$\leq\exp\left(-\frac{mt^{2}}{8}\right),$$

which completes the proof.

B.1.3 Proof of Lemma B.4

Observe the equality

$$Y_2 = \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \begin{bmatrix} 0 \\ \ell_{i,2:d_1} \end{bmatrix} \begin{bmatrix} 0 & \ell_{i,2:d_1}^\top \end{bmatrix}.$$

Therefore, we seek to bound the following operator norm:

$$\left\| Y_2 - S_{\text{in}} \left(I_{d_1} - e_1 e_1^\top \right) \right\|_{\text{op}} = \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \left(\ell_{i,2:d_1} \ell_{i,2:d_1}^\top - I_{d_1-1} \right) \right\|_{\text{op}}.$$

Using the tower rule for expectations and appealing to Corollary C.12, we therefore deduce

$$\mathbb{P} \left(\left\| Y_2 - S_{\text{in}} \left(I_{d_1} - e_1 e_1^\top \right) \right\|_{\text{op}} \geq C \sqrt{\frac{d_1}{m} + t} \right)$$

$$\leq \mathbb{E}_{\mathcal{I}_{\text{in}}^{\text{sel}}} \left[\mathbb{P} \left(\left\| Y_2 - S_{\text{in}} \left(I_{d_1} - e_1 e_1^\top \right) \right\|_{\text{op}} \geq C \sqrt{\frac{d_1}{m} + t} \;\middle|\; \mathcal{I}_{\text{in}}^{\text{sel}} = \mathcal{I} \right) \right] \leq 2 \exp(-cmt),$$

as desired.

$\square$

B.1.4 Proof of Lemma B.5

Noise model N1 Under this noise model, we write

$$\left\| Y_3 - S_{\text{out}} I_{d_1} \right\|_{\text{op}} = \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{out}}^{\text{sel}}} \ell_i \ell_i^\top - S_{\text{out}} I_{d_1} \right\|_{\text{op}}.$$

The proof follows by repeating the conditioning argument as in the proof of lemma B.4.

Noise model N2 Observe that

$$\left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{out}}^{\text{sel}}} \ell_i \ell_i^\top \right\|_{\text{op}} \leq \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{out}}} \ell_i \ell_i^\top \right\|_{\text{op}} \leq \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{out}}} (\ell_i \ell_i^\top - I_{d_1}) \right\|_{\text{op}} + \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{out}}} I_{d_1} \right\|_{\text{op}}$$

$$= \left\| \frac{1}{m} \sum_{i \in \mathcal{I}_{\text{out}}} (\ell_i \ell_i^\top - I_{d_1}) \right\|_{\text{op}} + p_{\text{fail}}.$$

Appealing to Corollary C.12, the result follows immediately.

$\square$

B.2 Proof of Proposition 5.2

We will assume that $\|\hat{w}\hat{x}^\top - \bar{w}_\star \bar{x}_\star^\top\|_F \leq \|\hat{w}\hat{x}^\top + \bar{w}_\star \bar{x}_\star^\top\|_F$. We will show that with high probability, $|\widehat{M} - M| \leq \delta M$, and moreover in this event if $\delta < 1$, we have $\widehat{M} > 0$. The other setting $\|\hat{w}\hat{x}^T - \bar{w}_\star \bar{x}_\star^\top\|_F \geq \|\hat{w}\hat{x}^T + \bar{w}_\star \bar{x}_\star^\top\|_F$ can treated similarly.

We will use the guarantees of Proposition 4.6. In particular, there exist numerical constants $c_1, \ldots, c_6 > 0$ so that as long as $m \geq \frac{c_1(d_1 + d_2 + 1)}{(1 - \frac{2|\mathcal{I}|}{m})^2} \ln\left(c_2 + \frac{1}{1 - 2|\mathcal{I}|/m}\right)$, then with probability at least $1 - 4\exp\left(-c_3(1 - \frac{2|\mathcal{I}|}{m})^2 m\right)$ we have

$$c_4\|X\|_F \leq \frac{1}{m}\|\mathcal{A}(X)\|_1 \leq c_5\|X\|_F \qquad \text{for all rank} \leq 2 \text{ matrices } X \in \mathbb{R}^{d_1 \times d_2},$$

and

$$c_6\left(1 - 2p_{\mathrm{fail}}\right)\|X\|_F \leq \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{in}}}|\ell_i^\top X r_i| - \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{out}}}|\ell_i^\top X r_i| \qquad \text{for all rank} \leq 2 \text{ matrices } X \in \mathbb{R}^{d_1 \times d_2}.$$

Throughout the remainder of the proof, suppose we are in this event. Define the two univariate functions

$$\hat{g}(a) := \frac{1}{m}\sum_{i=1}^m \left| y_i - (1+a)M\ell_i^\top \hat{w}\hat{x}^\top r_i \right|,$$

$$g(a) := \frac{1}{m}\sum_{i=1}^m \left| y_i - (1+a)M\ell_i^\top \bar{w}\bar{x}^\top r_i \right|$$

By construction, if $a^\star$ minimizes $\hat{g}(\cdot)$ then $(1+a^\star)M$ minimizes G. Thus, to prove the claim we need only show that any minimizer $a^\star$ of $\hat{g}$ satisfies $-\delta \leq a^\star \leq \delta$.

To that end, first note that $g(0)$ and $\hat{g}(0)$ are close:

$$|\hat{g}(0) - g(0)| \leq \frac{M}{m}\sum_{i=1}^m |\ell_i^\top \hat{w}\hat{x}^\top r_i - \ell_i^\top \bar{w}_\star \bar{x}_\star^\top r_i| \leq c_5 M \left\|\hat{w}\hat{x}^\top - \bar{w}_\star \bar{x}_\star^\top\right\|_F, \qquad (\mathrm{B.2})$$

Therefore, setting $\mu_3 = c_6\left(1 - 2p_{\mathrm{fail}}\right)$, we obtain

$$\hat{g}(a) = \frac{1}{m}\sum_{i=1}^m \left| y_i - (1+a)M\ell_i^\top \hat{w}\hat{x}^\top r_i \right|$$

$$= \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{in}}} \left| \ell_i^\top \bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top \hat{w}\hat{x}^\top r_i \right| + \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{out}}} \left| y_i - (1+a)M\ell_i^\top \hat{w}\hat{x}^\top r_i \right|$$

$$\geq \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{in}}} \left| \ell_i^\top \bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top \hat{w}\hat{x}^\top r_i \right| - \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{out}}} \left| \ell_i^\top \bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top \hat{w}\hat{x}^\top r_i \right|$$

$$\quad + \frac{1}{m}\sum_{i \in \mathcal{I}_{\mathrm{out}}} \left| y_i - \ell_i^\top \bar{w}\bar{x}^\top r_i \right|$$

$$\geq g(0) + \mu_3\|(1+a)M\hat{w}\hat{x}^\top - \bar{w}\bar{x}^\top\|_F$$

$$\geq \hat{g}(0) + \mu_3\|(1+a)M\hat{w}\hat{x}^\top - \bar{w}\bar{x}^\top\|_F - c_5 M \left\|\hat{w}\hat{x}^\top - \bar{w}_\star \bar{x}_\star^\top\right\|_F$$

$$\geq \hat{g}(0) + \mu_3|a|M - (\mu_3 M + c_5 M)\left\|\hat{w}\hat{x}^\top - \bar{w}_\star \bar{x}_\star^\top\right\|_F,$$

where the second inequality follows from Theorem 4.6, the third inequality follows from Equation (B.2), and the fourth follows from the reverse triangle inequality. Thus, any minimizer $a^\star$ of $\hat{g}$ must satisfy

$$|a^\star| \leq \left(1 + \frac{c_5}{\mu_3}\right) \left\| \widehat{w}\widehat{x}^\top - \bar{w}_\star \bar{x}_\star^\top \right\|_F = \delta,$$

as desired. Finally suppose $\delta < 1$. Then we deduce $\widehat{M} = (1 + |a^\star|)M \geq (1 - \delta)M > 0$. The proof is complete.

Appendix C Auxiliary Lemmas

C.1 Technical Results

This subsection presents technical lemmas we employed in our proofs. The first result we need is a special case of the celebrated Davis-Kahan $\sin\theta$ Theorem (see [19]). For any two unit vectors $u_1, v_1 \in \mathbb{S}^{d-1}$, define $\theta(u_1, v_1) = \cos^{-1}(|\langle u_1, v_1 \rangle|)$.

Lemma C.1. *Consider symmetric matrices $X, \Delta, Z \in \mathbb{R}^{n \times n}$, where $Z = X + \Delta$. Define δ to be the eigengap $\lambda_1(X) - \lambda_2(X)$, and denote the first eigenvectors of X, Z by u_1, v_1, respectively. Then*

$$\frac{1}{\sqrt{2}} \min\left\{ \|u - v\|_2, \|u + v\|_2 \right\} \leq \sqrt{1 - \langle u_1, v_1 \rangle^2} = |\sin\theta(u_1, v_1)| \leq \frac{\|\Delta\|_{\mathrm{op}}}{\delta}.$$

Additionally, we need the following fact about ϵ-nets over low-rank matrices, which we employ frequently to prove uniform concentration inequalities.

Lemma C.2 (Lemma 3.1 in [14]). *Let $S_r := \left\{ X \in \mathbb{R}^{d_1 \times d_2} \mid \mathrm{rank}(X) \leq r, \|X\|_F = 1 \right\}$. There exists an ϵ-net $\mathcal{N}$ (with respect to $\|\cdot\|_F$) of S_r obeying*

$$|\mathcal{N}| \leq \left(\frac{9}{\epsilon}\right)^{(d_1 + d_2 + 1)r}.$$

C.2 Concentration Inequalities

In this subsection, we first provide a few well-known concentration inequalities about sub-gaussian and sub-exponential random variables.

Theorem C.3 (Hoeffding's Inequality - Theorem 2.2.2 in [51]). *Let $X_1, \ldots, X_N$ be independent symmetric Bernoulli random variables. Then for any $t \geq 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^N X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2N}\right).$$

Theorem C.4 (Bernstein's Inequality - Theorem 2.8.4 in [51]). *Let $X_1, \ldots, X_N$ be independent mean-zero random variables, such that for $|X_i| \leq K$ for all i. Then for any $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(\sigma^2 + Kt/3\right)}\right)$$

here $\sigma^2 = \sum \mathbb{E}[X_i^2]$ is the variance of the sum.

Theorem C.5 (Sub-gaussian Concentration - Theorem 2.6.3 in [51]). *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables and $(a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Theorem C.6 (Sub-exponential Concentration - Theorem 2.8.2 in [51]). *Let $Z_1, \ldots, Z_m$ be an independent, mean zero, sub-exponential random variables and let $a \in \mathbb{R}^m$ be a fixed vector. Then, for any $t \geq 0$ we have that*

$$\mathbb{P}\left(\sum_{i=1}^{m} a_i Z_i \leq -t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right)$$

where $K := \max_i \|Z_i\|_{\psi_1}$ and $c > 0$ is a numerical constant.

Theorem C.7 (Corollary 2.8.3 in [51]). *Let $X_1, \ldots, X_m$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i\right| \geq t\right) \leq 2\exp\left[-cm\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right]$$

where $c > 0$ is a numerical constant and $K := \max_i \|X_i\|_{\psi_1}$.

Theorem C.8 (Theorem 5.6 in [9]). *Let $X = (X_1, \ldots, X_m)$ be a vector of n independent standard normal random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ denote an L-Lipschitz function. Then, for every $t \geq 0$, we have*

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

The following concentration inequalities deal with quantiles of distributions:

Lemma C.9. *Let $X_1, \ldots, X_m$ be an i.i.d. sample with distribution $\mathcal{D}$, choose Q_q to be the q population quantile of the distribution $\mathcal{D}$, that is $q = \mathbb{P}(X_1 \leq Q_q)$, and let $p \in (0,1)$ be any probability with $p < q$. Then,*

$$\mathbb{P}\left(\mathtt{quant}_p(\{X_i\}_{i=1}^{m}) \geq Q_q\right) \leq \exp\left(\frac{m(q-p)^2}{2(q-p)/3 + 2q(1-q)}\right),$$

where $\mathtt{quant}_p(\{X_i\}_{i=1}^{m})$ denotes the p-th quantile of the sample $\{X_i\}$.

Proof. It is easy to see that the following holds, $\mathtt{quant}_p(\{X_i\}_i^m) \geq Q_q$ if, and only if, $\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{X_i \leq Q_q\} \leq p$. Notice that $\mathbf{1}\{X_i \leq Q_q\} \sim \mathrm{B}(q)$ are i.i.d. Bernoulli random variables and thus $\mathrm{Var}(\mathbf{1}\{X_i \leq Q_q\}) = q(1-q)$. Then, the result follows by applying Bernstein's inequality (Theorem C.6) to $\frac{1}{m} \sum \mathbf{1}\{X_i \leq Q_q\} - q$. □

Lemma C.10. *Let a, b be i.i.d. sub-gaussian random variables. For any $Q > 0$ such that $q := \mathbb{P}(|ab| \leq Q) > 1/2$, consider the random variable c^2 defined as a^2 conditioned on the event $|ab| \leq Q$, namely for all t*

$$\mathbb{P}\left(c^2 \leq t\right) = \mathbb{P}\left(a^2 \leq t \mid |ab| \leq Q\right).$$

Then, c^2 is a sub-exponential random variable, in other words for all $t \geq 0$ we have that

$$\mathbb{P}\left(c^2 \geq t\right) \leq 2\exp(-t/2K)$$

where K is the minimum scalar such that $\mathbb{P}\left(a^2 \geq t\right) \leq 2\exp(-t/K)$.

Proof. Let us consider two cases. Suppose first $t \leq 2K\log 2$. Then we have that $1 \leq 2\exp(-t/2K)$ and therefore the stated inequality is trivial.

Suppose now $t \geq 2K\log 2$. Then we have that

$$\frac{t}{2K} \geq \log 2 \iff \exp(t/K - t/2K) \geq 2 \iff \exp(-t/2K) \geq 2\exp(-t/K).$$

With this we can bound the probability

$$\mathbb{P}\left(c^2 \geq t\right) = \frac{1}{q}\mathbb{P}\left(a^2 \mathbf{1}\{|ab| \leq Q\} \geq t\right) \leq \frac{1}{q}\mathbb{P}\left(a^2 \geq t\right) \leq \frac{2}{q}\exp(-t/K)$$
$$\leq 4\exp(-t/K) \leq 2\exp(-t/2K),$$

as claimed. □

The following Theorem from [50] is especially useful in bounding the operator norm of random matrices:

Theorem C.11 (Operator norm of random matrices)**.** *Consider an $m \times n$ matrix A whose rows A_i are independent, sub-gaussian, isotropic random vectors in $\mathbb{R}^n$. Then, for every $t \geq 0$, one has*

$$\mathbb{P}\left(\left\|\frac{1}{m}AA^\top - I_n\right\|_{\mathrm{op}} \leq C\sqrt{\frac{n}{m}} + t\right) \geq 1 - 2\exp\left(-cmt\right),$$

where C depends only on $K := \max_i \|A_i\|_{\psi_2}$.

Proof. The Theorem is a direct Corollary of [50, Theorem 5.39]. Specifically, the concavity of the square root gives us $\sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a+b}$, implying that

$$C\sqrt{\frac{n}{m}} + \sqrt{\frac{t}{m}} \leq C\sqrt{2}\sqrt{\frac{n}{m} + \frac{t}{m}}.$$

Additionally, [50, Theorem 5.39] gives us that

$$\mathbb{P}\left(\left\|\frac{1}{m}AA^\top - I_n\right\|_{\mathrm{op}} \le C\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right) \ge 1 - 2\exp\left(-ct^2\right).$$

Setting $t' = C\sqrt{m}t$ and a bit of relabeling, along with the square root inequality, gives us the desired inequality. $\qquad\square$

Let us record the following elementary consequence.

Corollary C.12. *Let $a_1, \ldots, a_m \in \mathbb{R}^d$ be independent, sub-gaussian, isotropic random vectors in $\mathbb{R}^n$ and let $\mathcal{I} \subset \{1, \ldots, m\}$ be an arbitrary set. Then, for every $t \ge 0$, one has*

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathcal{I}}(a_i a_i^\top - I_d)\right\|_{\mathrm{op}} \le C\sqrt{\frac{d}{m}} + t\right) \ge 1 - 2\exp\left(-cmt\right),$$

where C depends only on $K := \max_i \|A_i\|_{\psi_2}$.

Proof. Consider the matrix $A \in \mathbb{R}^{|\mathcal{I}|\times d}$ whose rows are the vectors a_i for $i \in \mathcal{I}$. Then we deduce

$$\left\|\frac{1}{m}\sum_{i\in\mathcal{I}}(a_i a_i^\top - I_d)\right\|_{\mathrm{op}} = \frac{|\mathcal{I}|}{m}\left\|\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}a_i a_i^\top - I_d\right\|_{\mathrm{op}} = \frac{|\mathcal{I}|}{m}\left\|\frac{1}{|\mathcal{I}|}AA^\top - I_d\right\|_{\mathrm{op}}.$$

Appealing to Theorem C.11, we therefore deduce for any $\gamma > 0$ the estimate

$$\left\|\frac{1}{m}\sum_{i\in\mathcal{I}}(a_i a_i^\top - I_d)\right\|_{\mathrm{op}} \le \frac{|\mathcal{I}|}{m}\sqrt{\frac{d}{|\mathcal{I}|}} + \gamma \le C\sqrt{\frac{d|\mathcal{I}|}{m^2} + \frac{\gamma|\mathcal{I}|^2}{m^2}},$$

holds with probability $1 - 2\exp(-c|\mathcal{I}|\gamma)$. Now for any $t > 0$, choose γ such that, $\frac{d|\mathcal{I}|}{m^2} + \frac{\gamma|\mathcal{I}|^2}{m^2} = \frac{d}{m} + t$, namely $\gamma = \frac{m^2}{|\mathcal{I}|^2}[\frac{d}{m}(1 - \frac{|\mathcal{I}|}{m}) + t]$. Noting

$$|\mathcal{I}|\gamma = m \cdot \frac{m}{|\mathcal{I}|}\left[\frac{d}{m}\left(1 - \frac{|\mathcal{I}|}{m}\right) + t\right] \ge mt,$$

completes the proof. $\qquad\square$

Recall that we defined the functions $q_{\mathrm{fail}}(p_{\mathrm{fail}}) = \frac{5 - 2p_{\mathrm{fail}}}{8(1 - p_{\mathrm{fail}})}$ and $Q_{\mathrm{fail}}(q_{\mathrm{fail}})$ given as the q_{fail}-quantile of $|ab|$ where a, b are i.i.d. standard normal. Furthermore we defined $\omega_{\mathrm{fail}} = \mathbb{E}[a^2 \mid |ab| \le Q_{\mathrm{fail}}]$.

Lemma C.13. *The function $\omega : [0, 1] \to \mathbb{R}_+$ given by*

$$p_{\mathrm{fail}} \mapsto \mathbb{E}[a^2 \mid |ab| \le Q_{\mathrm{fail}}]$$

is nondecreasing. In particular, there exist numerical constants $c_1, c_2 > 0$ such that for any $0 \le p_{\mathrm{fail}} \le 0.1$ we have

$$c_1 \le \omega_{\mathrm{fail}} \le c_2,$$

where the tightest constants are given by $c_1 = \omega(0) \ge 0.5$ and $c_2 = \omega(0.1) \le 0.56$.

Proof. The bulk of this result is contained in the following claim.

Claim 4. Let $0 \leq Q \leq Q'$ be arbitrary numbers, then

$$\mathbb{P}(a^2 \geq t \mid |ab| \leq Q) \leq \mathbb{P}(a^2 \geq t \mid |ab| \leq Q') \qquad \forall t \in \mathbb{R}.$$

We defer the proof of the claim and show how it implies the lemma. Observe that the functions $p_{\text{fail}} \mapsto q_{\text{fail}}$ and $q_{\text{fail}} \mapsto Q_{\text{fail}}$ are nondecreasing, thus it suffices to show that the function $Q \mapsto \mathbb{E}[a^2 \mid |ab| \leq Q]$ is nondecreasing. Let $0 \leq Q \leq Q'$

$$\mathbb{E}[a^2 \mid |ab| \leq Q] = \int_0^\infty \mathbb{P}(a^2 \geq t \mid |ab| \leq Q)dt \leq \int_0^\infty \mathbb{P}(a^2 \geq t \mid |ab| \leq Q')dt = \mathbb{E}[a^2 \mid |ab| \leq Q'],$$

where the inequality follows from the claim and the equalities follow from the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)dt$ for nonnegative random variables X. Hence ω is a nondecreasing function.

The above implies that for any $p_{\text{fail}} \in [0, 0.1]$ we have $\omega(0) \leq \omega_{\text{fail}} \leq \omega(0.1)$. Note that $\omega(0)$ is positive since it is defined by a positive integrand on a set of non-negligible measure. The bounds on $\omega(0)$ and $\omega(0.1)$ follow by a numerical computation. In particular we obtain that with $Q = 0.6$ the probability $\mathbb{P}(|ab| \leq Q) \geq 0.6679 \geq 2/3 = q_{\text{fail}}(0.1)$. Then computing numerically (with precision set to 32 digits) we obtain $\omega(0.1) \leq \mathbb{E}[a^2 \mid |ab| \leq Q] \leq 0.56$. Similarly we find that if we set $Q = 0.5$ we get $\mathbb{P}(|ab| \leq Q) \leq 0.5903 \leq 5/8 = q_{\text{fail}}(0)$. Then evaluating we find $\omega(0) \geq \mathbb{E}[a^2 \mid |ab| \leq Q] \geq 0.5$.

Proof of the claim. The statement of the claim is equivalent to having that for any $t \in \mathbb{R}_+$ the function $h_t : \mathbb{R}_+ \to \mathbb{R}$ given by

$$Q \mapsto \frac{\mathbb{P}(a^2 \leq t; |ab| \leq Q)}{\mathbb{P}(|ab| \leq Q)}$$

is nonincreasing. Our goal is to show that $h_t' \leq 0$. In order to prove this result we proceed as follows. Define

$$g(Q) := \frac{\pi}{2}\mathbb{P}(|ab| \leq Q) = \int_0^\infty \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx,$$

and

$$f_t(Q) := \frac{\pi}{2}\mathbb{P}(a^2 \leq t; |ab| \leq Q) = \int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx.$$

Observe $h_t = f_t/g$. Thus it suffices to show $f_t'g - f_tg' \leq 0$. Invoking Leibniz rule we get

$$\begin{aligned}
f_t'(Q) &= \frac{\partial}{\partial Q} \int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx \\
&= \int_0^{\sqrt{t}} \frac{\partial}{\partial Q} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx \\
&= \int_0^{\sqrt{t}} \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx.
\end{aligned}$$

Repeating the same procedure we get $g'(Q) = \int_0^\infty \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx$. Some algebra reveals we want to show

$$\xi(t) := \frac{\left(\int_0^{\sqrt{t}} \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx\right)}{\left(\int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx\right)} \leq \frac{\left(\int_0^\infty \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\right)}{\left(\int_0^\infty \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx\right)}.$$

It is enough to show that the function $\xi(t)$ is monotonically increasing. Define

$$\zeta_Q(t) = \int_0^{\sqrt{t}} \frac{1}{x} \exp(-(x^2 + Q^2/x^2)/2)\, dx \qquad \text{and} \qquad \psi_Q(t) = \int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx,$$

Thus we have

$$\zeta_Q'(t) = \frac{1}{2t} \exp(-(t + Q^2/t)/2) \qquad \text{and} \qquad \psi_Q'(t) = \frac{1}{2\sqrt{t}} \int_0^{Q/\sqrt{t}} \exp(-(t + y^2)/2)\, dy.$$

Again, $\xi(t) = \zeta_Q(t)/\psi_Q(t)$, hence we need to show $\zeta_Q' \psi_Q \geq \zeta_Q \psi_Q'$. After some algebra, this amounts to proving

$$\left(\int_0^{\sqrt{t}} \int_0^{Q/x} \exp(-(x^2 + y^2)/2)\, dy\, dx\right)$$

$$\geq \left(\int_0^{\sqrt{t}} \frac{\sqrt{t}}{x} \exp(-(Q^2/x^2 - Q^2/t)/2) \int_0^{Q/\sqrt{t}} \exp(-(x^2 + y^2)/2)\, dy\, dx\right).$$

The inequality is true if in particular the same holds for the integrands, i.e.

$$\int_0^{Q/x} \exp(-y^2/2)\, dy \geq \frac{\sqrt{t}}{x} \exp\left(-\left(\frac{Q^2}{x^2} - \frac{Q^2}{t}\right)/2\right) \int_0^{Q/\sqrt{t}} \exp(-y^2/2)\, dy.$$

Since $x \leq \sqrt{t}$, the previous inequality holds if

$$x \mapsto \frac{1}{x} \frac{\exp\left(-\frac{Q^2}{2x^2}\right)}{\int_0^{Q/x} \exp(-y^2/2) dy}$$

is increasing. By taking derivatives and reordering terms we see that this is equivalent to

$$\frac{Q - x^2}{Qx} \int_0^{Q/x} \exp(-y^2/2)dy + \exp(-Q^2/2x^2) \geq 0.$$

Since $\exp(-y^2/2)$ is decreasing, we have

$$\frac{Q - x^2}{qx} \int_0^{Q/x} \exp(-y^2/2)dy \geq \frac{Q - x^2}{Qx} \frac{Q}{x} \exp(-Q^2/2x^2) \geq -\exp(-Q^2/2x^2)$$

proving the claim. $\qquad\square$

Thus the proof is complete. $\qquad\square$