Information and Inference: A Journal of the IMA (2020) **00**, 1–64
doi:10.1093/imaiai/iaaa027

Composite optimization for robust rank one bilinear sensing

VASILEIOS CHARISOPOULOS AND DAMEK DAVIS
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA

MATEO DÍAZ
Center for Applied Mathematics, Cornell University, Ithaca, NY 14850, USA

AND

DMITRIY DRUSVYATSKIY[†]
Department of Mathematics, University of Washington, Seattle, WA 98195, USA
[†]Corresponding author. Email: ddrusv@uw.edu

[Received on 18 January 2019; revised on 3 July 2020; accepted on 30 July 2020]

We consider the task of recovering a pair of vectors from a set of rank one bilinear measurements, possibly corrupted by noise. Most notably, the problem of robust blind deconvolution can be modeled in this way. We consider a natural nonsmooth formulation of the rank one bilinear sensing problem and show that its moduli of weak convexity, sharpness and Lipschitz continuity are all dimension independent, under favorable statistical assumptions. 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 a constant relative error of the solution. We 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.

Keywords: blind deconvolution; Gauss–Newton; subgradient method; weak convexity; composite optimization; spectral.

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 [**14**,**45**,**56**], blind deconvolution [**5**,**41**,**44**,**58**], matrix completion [**15**,**22**,**57**] and covariance matrix estimation [**18**,**43**], 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 [**19**].

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

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

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. [19] 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 bear (e.g. [45,47,58]).

Smoothness of the penalty function $h(\cdot)$ in (1.1) is crucially used by the aforementioned techniques. A different recent line of work [7,24,25,30] 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 $\mathcal{X}^* := \mathrm{argmin}_{\mathcal{X}} f$ and suppose there exist constants $\rho, \mu, \mathsf{L}_f > 0$ satisfying:

- **(approximation)** $\big|h(F(y)) - h\big(F(x) + \nabla F(x)(y - x)\big)\big| \leqslant \frac{\rho}{2}\|y - x\|_2^2$ for all $x \in \mathcal{X}$,
- **(sharpness)** $f(x) - \inf f \geqslant \mu \cdot \mathrm{d}ist(x, \mathcal{X}^*)$ for all $x \in \mathcal{X}$,
- **(Lipschitz bound)** $\|v\|_2 \leqslant \mathsf{L}_f$ for all $v \in \partial f(x)$ with $\mathrm{dist}(x, \mathcal{X}^*) \leqslant \frac{\rho}{\mu}$.

Then when equipped with an appropriate sequence α_t and initialized at x_0 satisfying $\mathrm{dist}(x_0, \mathcal{X}^*) \leqslant \frac{\rho}{\mu}$, 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 $\frac{\mu}{\mathsf{L}_f} \in (0, 1)$.

A possible advantage of nonsmooth techniques can be gleaned from the phase retrieval problem. The papers [30, Corollary 3.1,3.2], [25, Corollary 3.8] recently, showed that for the phase retrieval problem, standard statistical assumptions imply that with high probability all the constants $\rho, \mu, \mathsf{L}_f > 0$ 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 rank-1 bilinear sensing. 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 [5] on blind deconvolution. 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:

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

where $\nu \geqslant 1$ 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 (automatic for Gaussian random vectors). We will show that as long as the number of measurements satisfies $m \gtrsim \frac{d_1+d_2}{(1-2p_{\text{fail}})^2} \ln\left(C + \frac{1}{1-2p_{\text{fail}}}\right)$, where C is a small dimension-independent constant, the formulation (1.3) admits dimension-independent constants ρ, L_f 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 $\frac{\|w_0 x_0^\top - \bar{w}\bar{x}^\top\|_F}{\|\bar{w}\bar{x}^T\|_F} \lesssim 1$.
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. The proposed procedure is motivated by the analogous initialization algorithm for robust phase retrieval [30,61]. To the best of our knowledge, this is the only available initialization procedure for rank-1 bilinear sensing 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. It is well known [20,34,42] that 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 [4,5].

As we alluded to above, our main motivation for studying rank-1 bilinear sensing is the blind deconvolution problem. 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 [4,5] '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 [5], 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 [32,41,45]. 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 [41] propose applying gradient descent on a regularized objective function and identify a 'basin of attraction' around the solution. The paper [45] 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 [32,41,45] are optimal up to polylog factors.

The popular formulation for the blind deconvolution problem [5] necessitates one of the sets of measurement vectors r_i or ℓ_i to be deterministic. Indeed, they are built from the columns of a discrete Fourier transform (DFT) matrix. Consequently, our assumptions that both r_i and ℓ_i are random is an oversimplification. Similar assumptions are made in [1–3] for example. Nonetheless, extensive experiments in Section 6.4 show that even in this semi-stochastic setting the proposed algorithms work remarkably well. In particular, for difficult instances with a large incoherence parameter, we observe that the proposed algorithms perform on par and often better than gradient descent on the smooth formulations of the problem.

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 [30] 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 recent work [26] studied the landscape of an unconstrained version of (1.3) and showed that the spurious critical points concentrate close to a co-dimension two subspace. Thus, suggesting there might be a large region with friendly geometry.

A related line of work [36,37,62] considers the original blind deconvolution problem of recovering a pair (u, v) from their convolution $u * v$ when u is low-dimensional and v is a sparse vector. These works are based on a very different approach to modeling the problem than the one we consider here. It would be interesting to see if similar ideas can be extended to this setting.

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 rank-1 bilinear sensing problem under 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

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

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 $f(x) \leqslant Cg(x)$ 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 $\langle X, Y\rangle = \mathrm{Tr}(X^\top Y)$ 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 $d \leqslant m$, the map $\sigma : \mathbb{R}^{d\times m} \to \mathbb{R}^d_+$ returns the vector of ordered singular values $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \cdots \geqslant \sigma_d(A)$. 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 [**9**,**46**,**50**,**54**]. 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

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

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:

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

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

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

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

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 the consequences of these two properties for local rapid convergence of numerical methods. The discussion mostly follows the recent work [**24**], though elements of this viewpoint can already be seen in the two papers [**25**,**30**] on robust phase retrieval.

Setting the stage, we introduce the following assumption.

ASSUMPTION A Consider the optimization problem,

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

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

1. **(Weak convexity)** The set $\mathscr{X}$ is closed and convex, while the function $f : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex.
2. **(Sharpness)** The set of minimizers $\mathscr{X}^* := \operatorname{argmin}_{x\in\mathscr{X}} f(x)$ is nonempty and the inequality

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

The class of weakly convex functions is broad and its importance in optimization is well documented [**6**,**48**,**51**,**53**,**55**]. 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_h-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 $\rho = \mathsf{L}_h\beta$, see e.g. [**29**, 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 $\rho = \mathsf{L}_h\beta$ 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 [**27**] and Ioffe [**33**] and the article of Lewis–Pang [**40**].

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 the 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 [24, Lemma 3.1].

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

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

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

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 $\mathscr{X} = \mathbb{R}$. Observe that the set of minimizers is $\mathscr{X}^* = \left\{\pm\frac{1}{\lambda}\right\}$, while $x = 0$ is always an extraneous stationary point. A quick computation shows that the smallest valid weak convexity constant is $\rho = 2\lambda^2$, while the largest valid sharpness constant is $\mu = \lambda$. We therefore deduce $\operatorname{dist}(0, \mathscr{X}^*) = \frac{1}{\lambda} = \frac{2\mu}{\rho}$. 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

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

Thus, we would like to search for algorithms whose basin of attraction is a tube $\mathscr{T}_\gamma$ 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**:** $(k \geqslant 0)$
Choose $\zeta_k \in \partial f(x_k)$. **If** $\zeta_k = 0$, then exit algorithm.
Set $x_{k+1} = \operatorname{proj}_{\mathscr{X}}\left(x_k - \frac{f(x_k) - \min_{\mathscr{X}} f}{\|\zeta_k\|^2} \zeta_k\right)$.

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

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

A straightforward argument [24, Lemma 3.2] shows $\tau \in [0, 1]$. The following theorem appears as [24, Theorem 4.1], while its application to phase retrieval was investigated in [25].

THEOREM 3.1 (Polyak subgradient method). Suppose that Assumption A holds and fix a real $\gamma \in (0,1)$. Then Algorithm 1 initialized at any $x_0 \in \mathcal{T}_\gamma$ produces iterates that converge Q-linearly to $\mathcal{X}^*$, that is

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

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 L_f, ρ and μ, thus a successful implementation may need to have access to reliable estimates of those quantities.

Algorithm 2 Subgradient method withgeometrically decreasing stepsize

Data: Real $\lambda > 0$ and $q \in (0,1)$.
Step k**:** $(k \geqslant 0)$
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 $x_{k+1} = \operatorname{proj}_{\mathcal{X}}\left(x_k - \alpha_k \frac{\zeta_k}{\|\zeta_k\|}\right)$.

The following theorem appears as [24, Theorem 6.1]. The convex version of the result dates back to Goffin [31].

THEOREM 3.2 (Geometrically decaying subgradient method). Suppose that Assumption A holds, fix a real $\gamma \in (0,1)$ and suppose $\tau \leqslant \sqrt{\frac{1}{2-\gamma}}$. Set $\lambda := \frac{\gamma\mu^2}{\rho \mathsf{L}_f}$ 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:

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

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 higher per-iteration cost, since in each iteration one must solve an auxiliary convex 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:

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

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

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

It is straightforward to see that Assumption B implies that f is ρ-weakly convex, see e.g. [29, 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. [12,29,39] and the references therein.

Algorithm 3 Prox-linear algorithm

Data: Initial point $x_0 \in \mathbb{R}^d$, proximal parameter $\beta > 0$.
Step k**:** $(k \geqslant 0)$

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

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 [11,28,30]. For the sake of completeness, we provide a quick argument.

THEOREM 3.3 (Prox-linear algorithm). Suppose Assumption B holds. Choose any $\beta \geqslant \rho$ and set $\gamma := \rho/\beta$. Then Algorithm 3 initialized at any point $x_0 \in \mathcal{T}_\gamma$ converges quadratically:

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

Proof. Consider an iterate x_k and choose any $x^* \in \operatorname{proj}_{\mathcal{X}^*}(x_k)$. 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

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

Using Assumption B.2, we therefore obtain

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

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

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

as claimed. □

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 $\rho, \mu, \mathsf{L}_f > 0$. 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

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

where $\mathscr{A}\colon \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ 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 $\mathscr{A}(X) = (\ell_i^\top X r_i)_{i=1}^m$. Since we will be interested in the prox-linear method, let us define the convex model

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

Our strategy is as follows. Section 4.1 identifies deterministic assumptions on the data, $\mathscr{A}$ and y, that yield favorable estimates of $\rho, \mu, \mathsf{L}_f > 0$. 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:

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

The classical RIP condition [**17**,**52**] used the ℓ_2-norm instead of the ℓ_1-norm. The ℓ_1 version that we use here has appeared in previous works [**16**,**18**] and also goes by the name of restricted uniform boundedness [**13**].

The following proposition estimates the two constants ρ and L_f, governing the performance of the subgradient and prox-linear methods under Assumption C. See Appendix A.1 for a proof.

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

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

$$|g(w,x) - g(\hat{w},\hat{x})| \leqslant \sqrt{2}c_2K\cdot\|(w,x)-(\hat{w},\hat{x})\|_2 \qquad \forall x,\hat{x}\in K\mathbb{B}, w,\hat{w}\in K\mathbb{B}.$$

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 $M := \|\bar{x}\bar{w}^\top\|_F = \|\bar{x}\|_2 \cdot \|\bar{w}^\top\|_2$. 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:

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

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 theorem, which analyzes the sharpness properties of the idealized function

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

This key geometrical property will allow us to establish sharpness for any measurement map $\mathcal{A}$ satisfying RIP. The proof is quite tedious, and therefore we have placed it in Appendix A.2.

THEOREM 4.2 For any $\nu \geqslant 1$, we have the following bound:

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

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 Theorem 4.2 for symmetric matrices was proved in [**58**, 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 $\nu \geqslant 1$, we have the following bound:

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

Proof. Using Assumption C and Theorem 4.2, we deduce

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

as claimed. □

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 ($\mathcal{I}$-outlier 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

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

Combining Assumption D with Theorem 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

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

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

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

where the first inequality follows by the reverse triangle inequality, the second inequality follows by Assumption (C2) and the final inequality follows from Theorem 4.2. The proof is complete. □

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

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

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

COROLLARY 4.1 (Convergence guarantees). Suppose Assumptions C and D are valid, and consider the optimization problem

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

Choose any pair (w_0, x_0) satisfying

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

Then the following are true.

1. **(Polyak subgradient)** Algorithm 1 initialized (w_0, x_0) produces iterates that converge linearly to $\mathscr{S}_\nu^*$, that is

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

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:

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

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

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

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 $\mathscr{I}_{\text{in}} \subseteq [m]$ and $\mathscr{I}_{\text{out}} \subseteq [m]$, called the *inlier* and *outlier* sets. Intuitively, the index set $\mathscr{I}_{\text{in}}$ encodes exact measurements while $\mathscr{I}_{\text{out}}$ encodes measurements that have been replaced by gross outliers. Define the corruption frequency $p_{\text{fail}} := \frac{|\mathscr{I}_{\text{out}}|}{m}$; 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 the following fully stochastic matrix model. For simplicity, the reader may assume both L and R are Gaussian with i.i.d. entries, though the results of this paper extend beyond this case. We should note that in more realistic circumstances, such as the problem of blind deconvolution, it is more appropriate for one of the matrices L or R to be deterministic. Though the theoretical guarantees we present only hold in the fully stochastic setting, numerical experiments in Section 6.4 indicate that the proposed methods are effective even when one of the matrices is deterministic.

Random matrix model

M 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,

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

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

Thus, the model M asserts 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 M.

Example 4.5 (Gaussian matrices satisfy model M). 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 M. 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. [**60**, Exercise 3.3.6]. Thus, for any $c \geqslant 0$, we compute

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

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

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

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

4.2.1 *Assumptions C and D under model M* In this section, we record the following theorem, which shows validity of Assumptions C and D under M, with high probability. We defer the proof of the next result to Appendix B.1.

THEOREM 4.6 (Measurement model M). 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 $m \geqslant \frac{c_1(d_1+d_2+1)}{(1-2|\mathcal{I}|/m)^2} \ln\left(c_2 + \frac{c_2}{1-2|\mathcal{I}|/m}\right)$, 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

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

and

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

Combining Theorem 4.6 with Corollary 4.1, we obtain the following guarantee.

COROLLARY 4.2 (Convergence guarantees). Consider the measurement model (4.2) and suppose that model M 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 $m \geqslant \frac{c_1(d_1+d_2+1)}{(1-2p_{\text{fail}})^2} \ln\left(c_2 + \frac{c_2}{1-2p_{\text{fail}}}\right)$ and you choose any pair (w_0, x_0) with relative error

$$\frac{\text{dist}((w_0, x_0), \mathcal{S}_\nu^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leqslant \frac{c_6\left(1 - 2p_{\text{fail}}\right)}{4\sqrt{2}c_5(\nu+1)}, \tag{4.6}$$

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

1. **(Polyak subgradient)** Algorithm 1 initialized (w_0, x_0) produces iterates that converge linearly to $\mathscr{S}_\nu^*$, that is

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

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

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

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

$$\frac{\operatorname{dist}((w_k, x_k), \mathscr{X}^*)}{\sqrt{\|\bar{w}\bar{x}^\top\|_F}} \leqslant 2^{-2^k} \cdot \frac{c_6\left(1-2p_{\mathrm{fail}}\right)}{2\sqrt{2}c_5(\nu+1)} \qquad \forall k \geqslant 0.$$

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

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 $\frac{\operatorname{dist}((x,w),\mathscr{S}_\nu^*)}{\sqrt{\|\bar{x}\bar{w}^\top\|_F}}$ 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 [30, Section 4] and [61].

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 **M**:

$\overline{\mathbf{M}}$ 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 $i \in \mathcal{I}_{\text{out}}$.

Algorithm 4 Initialization.

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

$\mathcal{I}^{\text{sel}} \leftarrow \{i \mid |y_i| \leqslant \texttt{med}(|y|)\}$

Form directional estimates:

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

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

Estimate the norm of the signal:

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

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

return (w_0, x_0)

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 for which the product $\left|\langle \ell_i, \bar{w}_\star\rangle\langle r_i, \bar{x}_\star\rangle\right|$ is small, leading to subtle ambiguities not present in [30, Section 4] and [61], see Fig. 1. Corruptions add further ambiguities since the noise model **N2** allows a constant fraction of measurements to be adversarially modified.

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

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

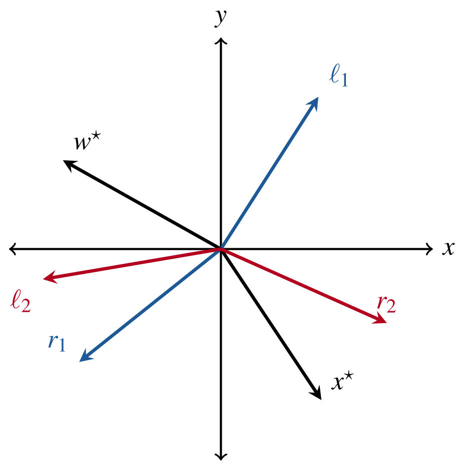

FIG. 1. Intuition behind spectral initialization. The pair ℓ_1, r_1 will be included since both vectors are almost orthogonal to the true directions. ℓ_2, r_2 is unlikely to be included since r_2 is almost aligned with $x^\star$.

(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 C. The general proof strategy we follow is analogous to [30, 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 C.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 [23], 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 $1 - c_1 \exp\left(-c_2 mt\right)$, the following hold:

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

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

$$\delta := \left(1 + \frac{c_5}{c_6(1-2p_{\mathrm{fail}})}\right) \min_{s\in\{\pm1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star\bar{x}_\star^\top \right\|_F, \tag{5.1}$$

where c_5 and c_6 are as in Theorem 4.6. Then we prove the following estimate (see Appendix C.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 $m \geqslant \frac{c_1(d_1+d_2+1)}{(1-2p_{\mathrm{fail}})^2} \ln\left(c_2 + \frac{c_2}{1-2p_{\mathrm{fail}}}\right)$, then with probability at least $1 - 4\exp\left(-c_3(1-2p_{\mathrm{fail}})^2 m\right)$, 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 $||\widehat{M}| - M| \leqslant \delta M$. Moreover, if in this event $\delta < 1$, then we have $\mathrm{sign}(\widehat{M}) = \mathrm{argmin}_{s\in\{\pm1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star\bar{x}_\star^\top \right\|_F$.

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 $\delta \leqslant 1/2$, which in particular implies with high probability $\mathrm{sign}(\widehat{M}) = \mathrm{argmin}_{s\in\{\pm1\}} \left\| \widehat{w}\widehat{x}^\top - s\bar{w}_\star\bar{x}_\star^\top \right\|_F$.

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 $p_{\mathrm{fail}} < c_0$ and $m \geqslant c_3(d_1 + d_2)$, then the estimate holds:[2]

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

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

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

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

STEP 3: FINAL ESTIMATE. Putting the directional and norm estimates together, we arrive at the following theorem.

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

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

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

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

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

we find that

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

where c_5 and c_6 are defined in Theorem 4.6. Appealing to Proposition 5.1, the result follows. □

Combining Corollary 4.1 and Theorem 5.4, we arrive at the following guarantee for the stage procedure.

COROLLARY 5.1 (Efficiency estimates). Suppose either of the models **N1** and **N2**. Let (w_0, x_0) be the output of the initialization Algorithm4. Set $\widehat{M} = \|w_0 x_0^\top\|_F$ and consider the optimization problem

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

Set $\nu := \sqrt{\frac{2\widehat{M}}{M}}$ and notice that the feasible region of (5.2) coincides with $\mathscr{S}_\nu$. 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 $m \geqslant c_3(d_1 + d_2)$ and $p_{\text{fail}} \leqslant c_0$, 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 (w_0, x_0) produce iterates that converge linearly to $\mathscr{S}_\nu^*$, that is

$$\frac{\text{dist}^2((w_k, x_k), \mathscr{S}_\nu^*)}{\|\bar{w}\bar{x}^\top\|_F} \leqslant c_4\left(1 - c_4\right)^k \qquad \forall k \geqslant 0.$$

---

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

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

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

Proof. We provide the proof under model **N1**. The proof under model N2 is completely analogous. Combining Propositions 5.2, 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 $m \geqslant c_3(d_1 + d_2)$ and $p_{\text{fail}} < c_0$, then for any $t \in [0, 1]$, with probability $1 - c_1 \exp\left(-c_2 mt\right)$, we have

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

and

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

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

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

Combining the estimates, we conclude

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

Thus to ensure the relative error assumption (4.6), it suffices to ensure the inequality

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

where c_5, c_6 are the constants from Corollary 4.2. Using the bound $\nu \leqslant \sqrt{3}$, 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 $\max\{d_1, d_2\} \leqslant d_1 + d_2 \leqslant \frac{m}{c_3}$ and applying Corollary 4.2 completes the proof. □

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 programming language `Julia` [8]. A reference implementation and code for the experiments is available in [21].

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

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

where $\operatorname{sign}(t)$ denotes the sign of t, with the convention $\operatorname{sign}(0) = 0$. 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 stepsize decay q; the experiment described in Section 6.1.2 demonstrates this sensitivity. Setting $\lambda = 1.0$ sufficed for all the experiments depicted hereafter.

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

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

Equivalently, one may rewrite this expression as a least absolute deviation objective:

$$\begin{aligned} f_{(w_k,x_k)}(w,x) &= \frac{1}{m}\sum_{i=1}^{m}\Bigg|\underbrace{\left(\,\langle x_k, r_i\rangle \ell_i^\top \,|\, \langle \ell_i, w_k\rangle r_i^\top\,\right)}_{A_i}\underbrace{\begin{pmatrix}w-w_k\\x-x_k\end{pmatrix}}_{z} - \underbrace{(y_i - \langle \ell_i, w_k\rangle\langle x_k, r_i\rangle)}_{\tilde{y}_i}\Bigg| \\ &= \frac{1}{m}\|Az-\tilde{y}\|_1 . \end{aligned}$$

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

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

Motivated by the work of [30] on robust phase retrieval, we solve this subproblem with the *graph splitting* variant of the alternating direction method of multipliers (ADMM), as described in [49]. This iterative method applies to problems of the form

$$\begin{aligned} &\min_{z\in\mathscr{X}}\ \frac{1}{m}\|t-\tilde{y}\|_1 + \frac{1}{2\alpha}\|z\|_2^2 \\ &\text{s.t.}\ \ t = Az. \end{aligned}$$

The ADMM method takes the form:

$$
\begin{aligned}
z' &\leftarrow \operatorname*{argmin}_{z\in\mathscr{S}_\nu}\left\{\frac{1}{2\alpha}\|z\|_2^2+\frac{\rho}{2}\|z-(z_k-\lambda_k)\|_2^2\right\}\\
t' &\leftarrow \operatorname*{argmin}_{t}\left\{\frac{1}{m}\|t-\tilde{y}\|_1+\frac{\rho}{2}\|t-(t_k-\nu_k)\|_2^2\right\}\\
\begin{pmatrix} z_+\\ t_+\end{pmatrix} &\leftarrow \begin{bmatrix} I_{d_1+d_2} & A^\top\\ A & -I_m\end{bmatrix}^{-1}\begin{bmatrix} I_{d_1+d_2} & A^\top\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{pmatrix} z'+\lambda\\ t'+\nu\end{pmatrix}\\
\lambda_+ &\leftarrow \lambda+(z'-z_+),\ \nu_+\leftarrow \nu+(t'-t_+),
\end{aligned}
$$

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 $\|(z_+,t_+)-(z,t)\|\leqslant\varepsilon_k\cdot\left(\sqrt{d_1+d_2}+\max\left\{\|z\|_2,\|t\|_2\right\}\right)$ and the dual residual satisfies $\|(\lambda_+,\nu_+)-(\lambda,\nu)\|\leqslant\varepsilon_k\cdot\left(\sqrt{d_1+d_2}+\max\left\{\|\lambda\|_2,\|\nu\|_2\right\}\right)$ with $\varepsilon_k=2^{-k}$.

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 Figs 2 and 3, we have depicted how the quantity

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

changes per iteration for the prox-linear and subgradient methods. We conducted tests in both the moderate corruption ($p_{\text{fail}}=0.25$) and high corruption ($p_{\text{fail}}=0.45$) regimes. For both methods, under moderate corruption ($p_{\text{fail}}=0.25$), we see that exact recovery is possible as long as $c\geqslant 5$. Likewise, even in high corruption regime ($p_{\text{fail}}=0.45$), exact recovery is still possible as long as $c\geqslant 8$. 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. Interestingly, Fig. 3 shows how the performance of the prox-linear method stagnates for the first few iterations before dropping at a quadratic rate. This might indicate that for these choices of c the initialization procedure outputs a point slightly outside of the

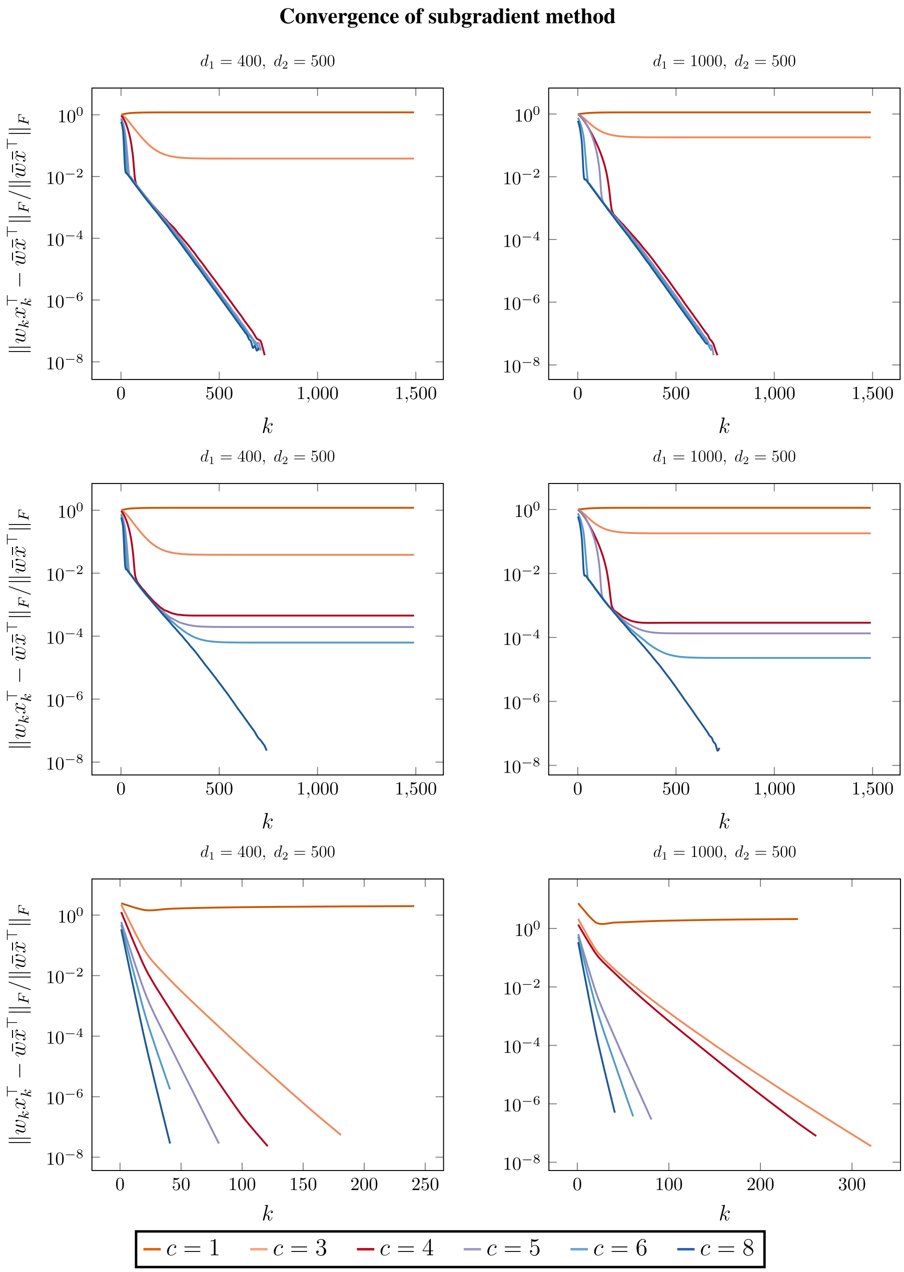

FIG. 2. Dimensions are $(d_1, d_2) = (400, 500)$ in the first column and $(d_1, d_2) = (1000, 500)$ in the second column. We plot the error $\|w_k x_k^\top - \bar{w}\bar{x}^\top\|_F / \|\bar{w}\bar{x}^\top\|_F$ vs. iteration count. Top row is using Algorithm 2 with $p_{\text{fail}} = 0.25$. Second row is using Algorithm 2 with $p_{\text{fail}} = 0.45$. Third row is using Algorithm 1 with $p_{\text{fail}} = 0$.

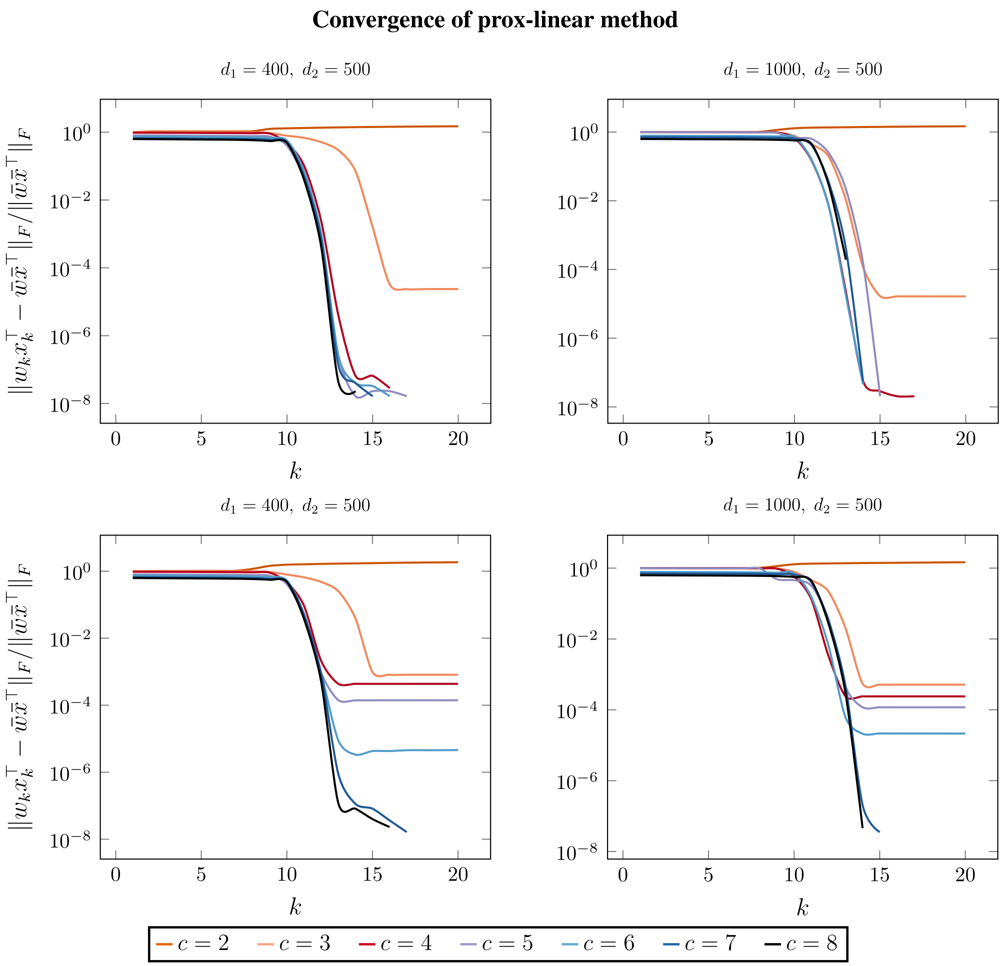

FIG. 3. Dimensions are $(d_1, d_2) = (400, 500)$ in the first column and $(d_1, d_2) = (1000, 500)$ in the second column. We plot the error $\|w_k x_k^\top - \bar{w}\bar{x}^\top\|_F / \|\bar{w}\bar{x}^\top\|_F$ vs. iteration count for an application of Algorithm 3 in the two settings: $p_{\text{fail}} = 0.25$ (top row) and $p_{\text{fail}} = 0.45$ (bottom row).

region of quadratic convergence. Another possibility is that the levels of accuracy set for solving the proximal subproblems, $\varepsilon_t := 2^{-t},\ t = 1, \ldots, T$, are not 'fine' enough for the first few iterations.

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 [49], 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 four 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

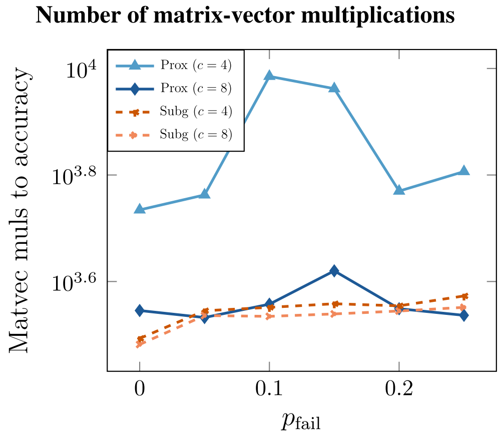

FIG. 4. Matrix-vector multiplications to reach rel. accuracy of 10^{-5}.

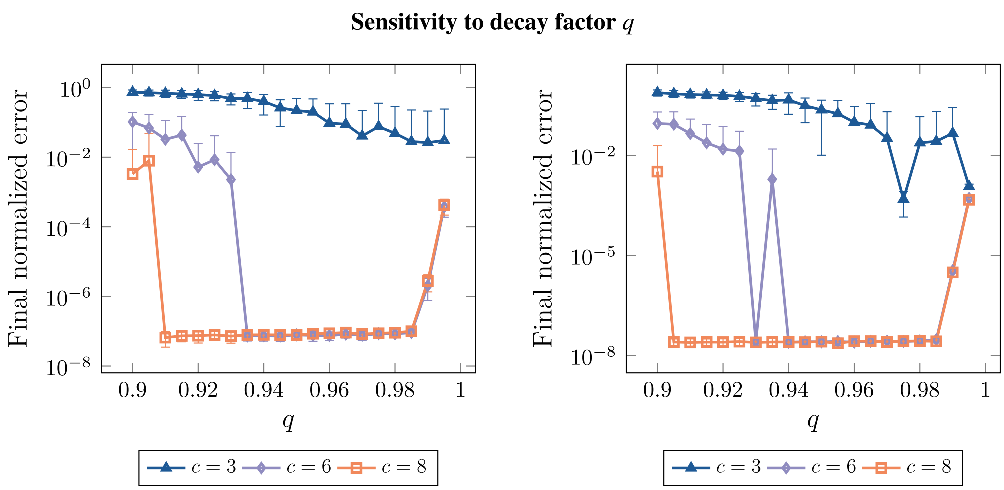

FIG. 5. Final normalized error $\|w_k x_k^\top - \bar{w}\bar{x}^\top\|_F / \|\bar{w}\bar{x}^\top\|_F$ for Algorithm 2 with different choices of q, in the settings $p_{\text{fail}} = 0$ (left) and $p_{\text{fail}} = 0.25$ (right).

we obtain $\frac{\|wx^\top - \bar{w}\bar{x}^\top\|_F}{\|\bar{w}\bar{x}^\top\|_F} \leqslant 10^{-5}$. 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 stepsize decay.* Due to the sensitivity of Algorithm 2 to the stepsize decay q, we experiment with different choices of q in order to find an empirical range of values that 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 over 50 test runs

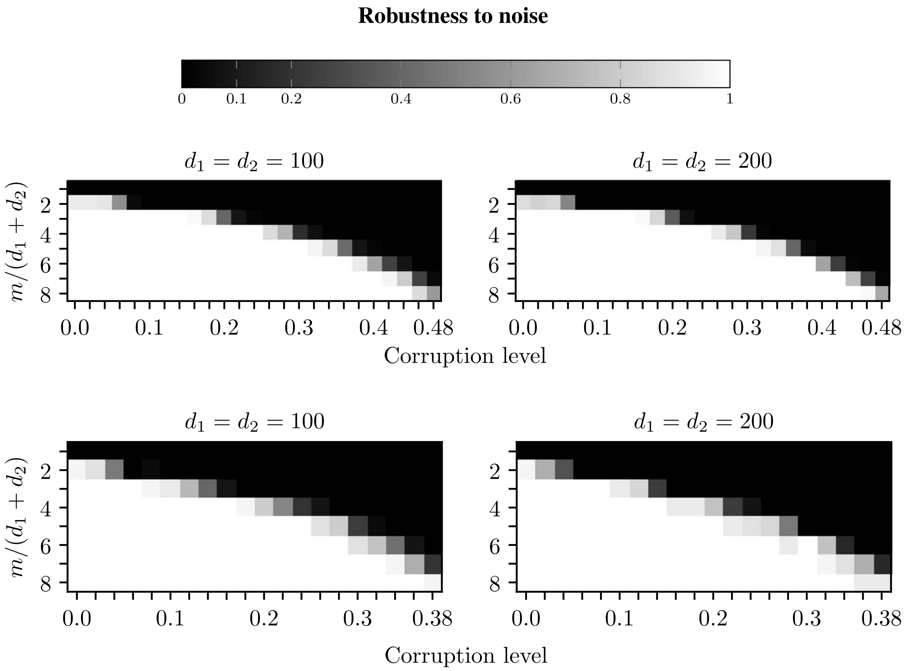

FIG. 6. Empirical recovery probabilities for matrix model **M** and noise models **N1** (**top**) and **N2** (**bottom**) across 100 independent runs using Algorithm 2. Lighter cells imply higher recovery probability.

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 [**5**,**30**], 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 the top row of Fig. 6.

The phase transition plots are similar for both dimensionality choices, revealing that in the moderate independent noise regime ($p_{\text{fail}} \leqslant 25\%$), setting $m \geqslant 4(d_1 + d_2)$ 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 $8 \cdot (d_1 + d_2)$.

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

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

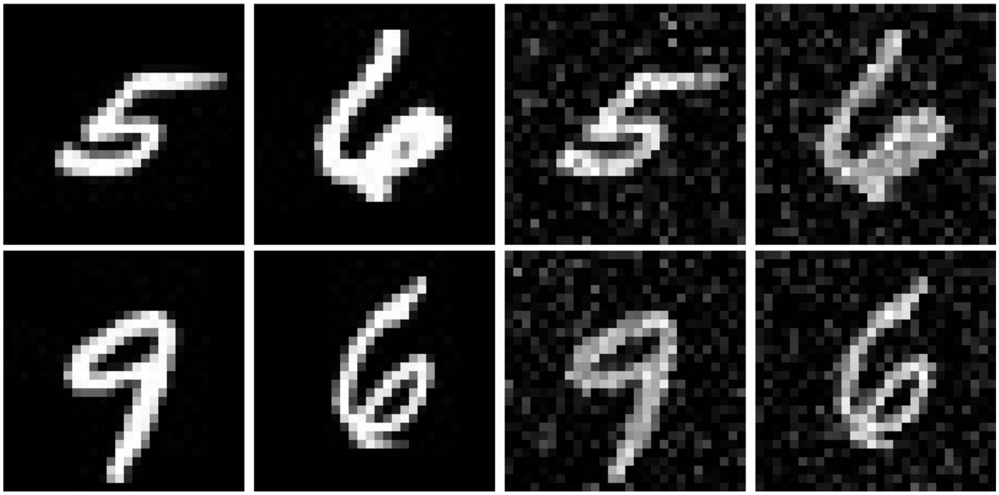

FIG. 7. Digits 5, 6 (top) and 9, 6 (bottom). Original images are shown on the left, estimates from spectral initialization on the right. Parameters: $p_{\text{fail}} = 0.45, m = 16 \cdot 784$.

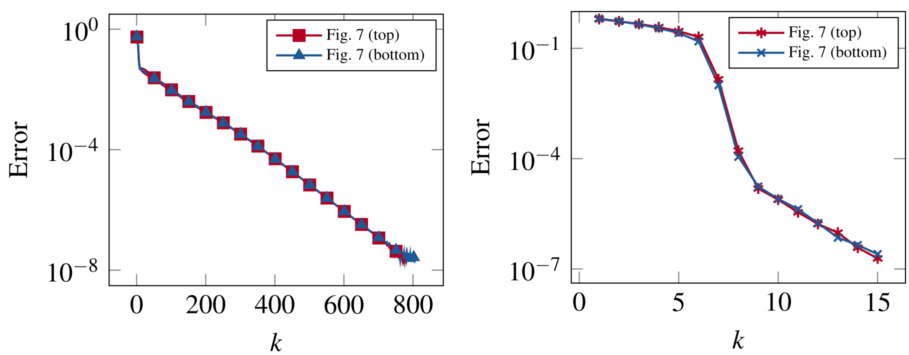

FIG. 8. Relative error vs. iteration count on `mnist` digits for subgradient method (left) and prox-linear method (right).

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 the bottom row of Fig. 6 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 data set [38]. 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. 7. Evidently, the initial estimates of the images are visually similar to the true digits, up to sign;

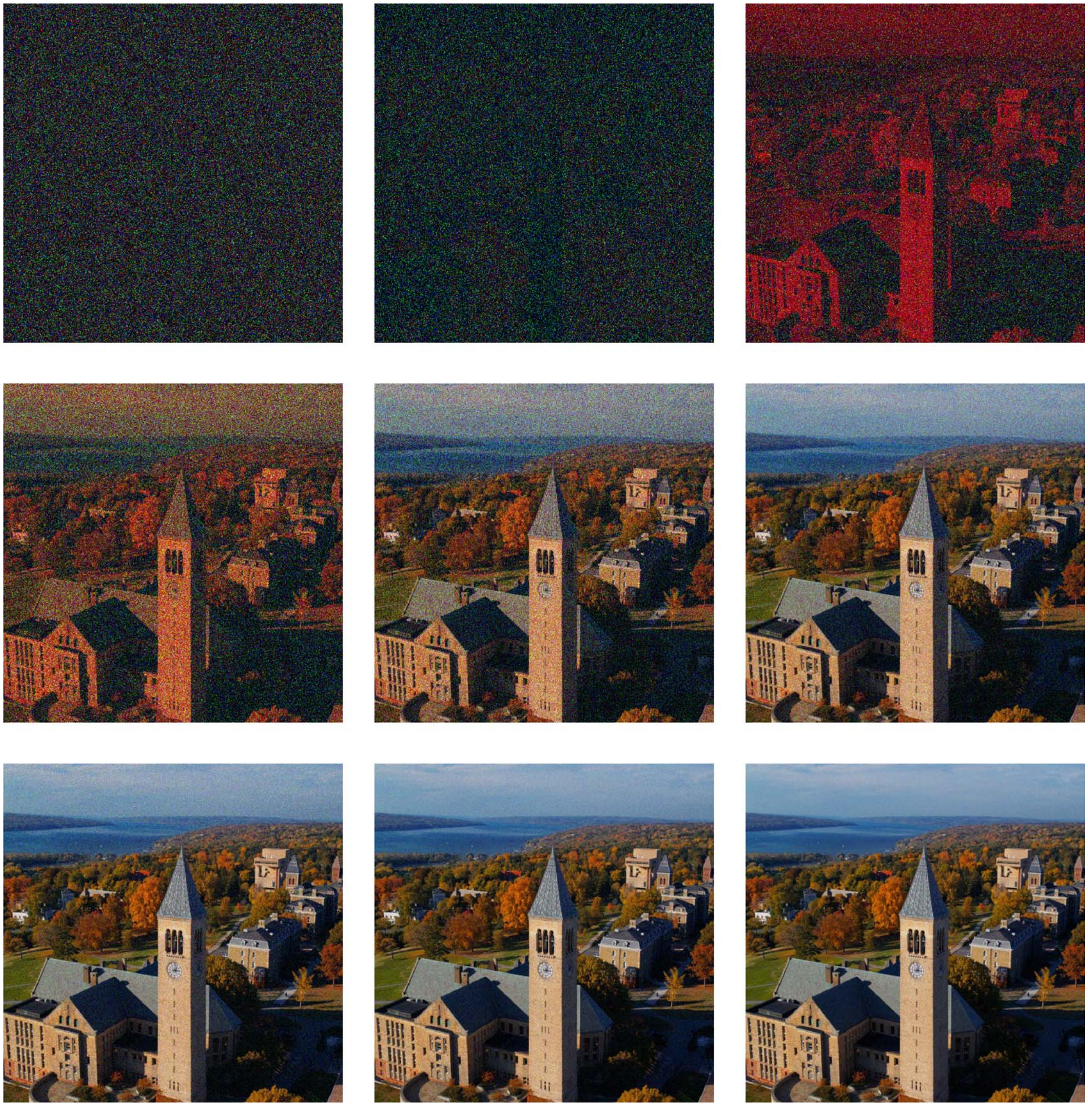

FIG. 9. Iterates w_{10i}, $i = 1, \ldots, 9$. $(m, k, d, n) = (2^{22}, 16, 2^{18}, 512)$.

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. 8.

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

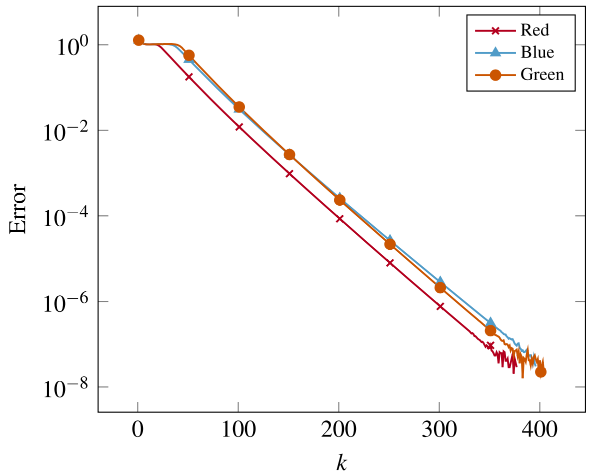

FIG. 10. Normalized error for different channels in image recovery.

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 $S_i = \operatorname{diag}(\xi_1, \ldots, \xi_d)$, where $\xi \sim_{\text{i.i.d}} \{-1, 1\}$, 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 element-wise 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 $p \mapsto L^\top p, \; q \mapsto R^\top q$, 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.

An example where we recover a pair of 512×512 color images using the Polyak subgradient method (Algorithm 1) is shown below; Fig. 9 shows the progression of the estimates w_k, up until the 90-th iteration, while Fig.10 depicts the normalized error at each iteration for the different channels of the images.

6.4 *Experiments on blind deconvolution*

In this section, we experiment on a realistic instance of the blind deconvolution problem, following [5,41]. Throughout, the measurement vectors ℓ_i and r_i are complex and the vectors ℓ_i are moreover deterministic. Note that this setting is outside the scope of our guarantees, which require all the vectors ℓ_i and r_i to be stochastic; nonetheless, we will see that the proposed methods work well even in this setting.

Recall that the complex vector space $\mathbb{C}^n$ is endowed with the Hermitian inner product $\langle x, y\rangle := x^H y = \sum_{i=1}^n \bar{x}_i y_i$, which satisfies $\langle x, y\rangle = \overline{\langle y, x\rangle}$. In the space of matrices $\mathbb{C}^{m\times n}$, the inner product is defined in an analogous fashion, with $\langle A, B\rangle := \mathrm{Tr}(A^H B)$, with A^H denoting the Hermitian transpose of A. Additionally, we write $\Re(z), \Im(z)$ for the real and imaginary parts of z, understood to hold elementwise if z is a vector.

In the blind deconvolution problem, we observe the circular convolution of two signals u and v, so that the measurements are

$$y_j = \sum_{i=1}^m u_i v_{(j-i+1) \bmod m}. \tag{6.2}$$

In (6.2), we assume that there is no observation noise for the sake of simplicity. A standard assumption is that u, v lie in known low-dimensional subspaces of $\mathbb{R}^m$ of dimensions d_1, d_2 respectively, so that

$$u = Bw_\sharp, \quad v = Cx_\sharp.$$

To recast this problem as a bilinear sensing problem, we may pass to the Fourier domain. Denote by F_m the $m \times m$ DFT matrix, with elements

$$(F_m)_{ij} := \exp\left(-\iota 2\pi \frac{(i-1)(j-1)}{m}\right),$$

where we set $\iota = \sqrt{-1}$, and also define

$$L = \overline{F_m B} \in \mathbb{C}^{m\times d_1}, \quad R = F_m C \in \mathbb{C}^{m\times d_2}.$$

Then following standard arguments (see e.g. [5]) the equivalent model to (6.2) in the Fourier domain is

$$\hat{y}_i = \langle \ell_i, w_\sharp\rangle \cdot \langle x_\sharp, r_i\rangle.$$

A common choice for B is the matrix $\begin{bmatrix} I_{d_1} \\ \mathbf{0} \end{bmatrix}$ [5,41,45], which leads to the partial DFT matrix $L \in \mathbb{R}^{m\times d_1}$ formed by taking the first d_1 columns of F_m and used in the experiments below. On the other hand, C is often assumed to have i.i.d. Gaussian entries, so that the entries of R are also i.i.d. and follow the complex Gaussian distribution. For simplicity, we relabel $\hat{y}$ to y in the sequel.

We therefore consider the nonsmooth formulation of the problem

$$\min_{\|w\|,\|x\|\leqslant \nu\sqrt{M}} \frac{1}{m} \sum_{i=1}^m \left|\langle \ell_i, w\rangle \langle x, r_i\rangle - y\right|, \tag{6.3}$$

where $|x|$ denotes the magnitude of the complex number x. With the help of Wirtinger calculus [35], we describe the extension of the subgradient method in the complex domain. The Wirtinger derivatives of a complex function $f(z)$ with $z = x + iy$, $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ are given by

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right)$$
$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right),$$

where $x \mapsto \bar{x}$ denotes complex conjugation. The chain rule of Wirtinger calculus, summarized below, is useful in formally defining a subgradient of the nonsmooth objective:

$$\frac{\partial (f \circ g)}{\partial z} = \left(\frac{\partial f}{\partial z} \circ g\right)\frac{\partial g}{\partial z} + \left(\frac{\partial f}{\partial \bar{z}} \circ g\right)\frac{\partial \bar{g}}{\partial z} \tag{6.4}$$

$$\frac{\partial (f \circ g)}{\partial \bar{z}} = \left(\frac{\partial f}{\partial z} \circ g\right)\frac{\partial g}{\partial \bar{z}} + \left(\frac{\partial f}{\partial \bar{z}} \circ g\right)\frac{\partial \bar{g}}{\partial \bar{z}}. \tag{6.5}$$

We now compute the Wirtinger derivative of the real-valued function

$$f_S(w, x) = \frac{1}{m}\sum_{i=1}^{m}\left|\langle \ell_i, w\rangle\langle x, r_i\rangle - y_i\right| = \frac{1}{m}\sum_{i=1}^{m}\left|\mathscr{A}(wx^H)_i - y_i\right|,$$

with $\mathscr{A}(X) = \left\{\ell_i^H X r_i\right\}_{i=1}^m$ the corresponding operator for the complex case. By the definition of the Wirtinger derivatives, it is easy to see that

$$\left.\frac{\partial\,|z|}{\partial \bar{z}}\right|_{z=z_k} = \begin{cases} 0, & z_k = \mathbf{0} + \mathbf{0}j \\ \frac{z_k}{2|z_k|}, & \text{otherwise} \end{cases}. \tag{6.6}$$

In this way, the linearization around z_k based on the Wirtinger gradient (see [35, pp. 20–21]) satisfies $|z| \geqslant \left|z_k\right| + 2\Re\left(\langle g(z_k), z - z_k\rangle\right),\ g(z) := \frac{\partial |z|}{\partial \bar{z}}$, after elementary calculations, much like its $\mathbb{R}^n$-counterpart.

With this in hand, the application of the chain rule from Equation (6.5) gives us that

$$\partial f_S(w, x) \ni \zeta_k := \begin{pmatrix} \frac{\partial f}{\partial \bar{w}} \\ \frac{\partial f}{\partial \bar{x}} \end{pmatrix} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\left|\langle \ell_i, w_k\rangle\langle x_k, r_i\rangle - y_i\right|}\begin{bmatrix} (\langle \ell_i, w\rangle\langle x_k, r_i\rangle - y_i)\langle r_i, x_k\rangle \ell_i \\ (\langle w_k, \ell_i\rangle\langle r_i, x_k\rangle - \bar{y}_i)\langle \ell_i, w_k\rangle r_i \end{bmatrix}.$$

In the above, we make the convention that when $z = \mathbf{0}$, we set $\frac{z}{|z|} = \mathbf{0}$, as in (6.6).

EXPERIMENTS. To avoid confusion due to the conjugate notation, in this section we will denote the ground truth signals by $w_\sharp$ and $x_\sharp$, respectively. We repeat the synthetic experiment under noise model **N1** for the case of complex measurements. Specifically, we form the left measurement matrix $L \in \mathbb{C}^{m\times d}$ by taking the first d columns of the (unnormalized) $m \times m$ DFT matrix, with $L^H L = mI_d$. For the right

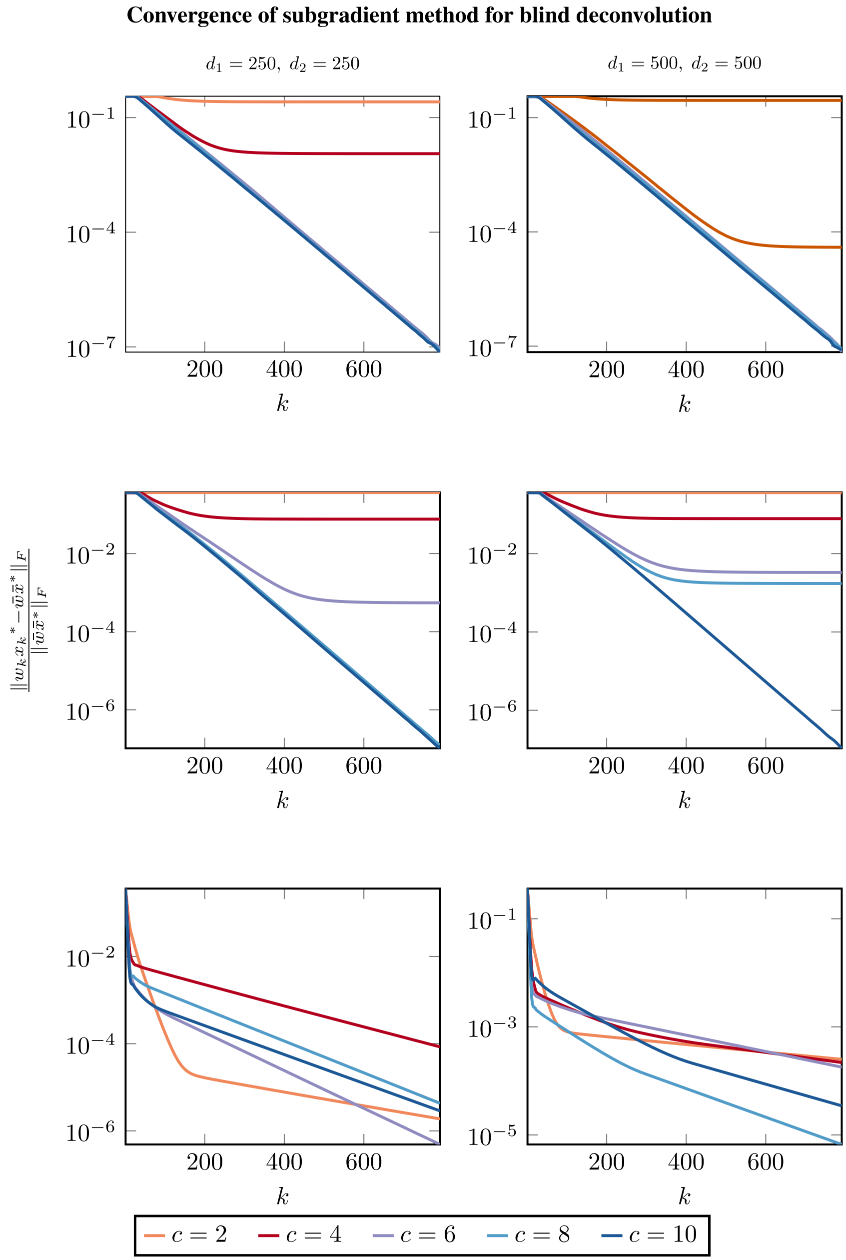

FIG. 11. Convergence plot for synthetic instances in the complex domain, using $m = c \cdot d$. Left: $d = 250$. Right: $d = 500$. Top 2 rows: Algorithm 2 with $p_{\text{fail}} \in \{0.25, 0.45\}$. Bottom row: Algorithm 1 with $p_{\text{fail}} = 0$.

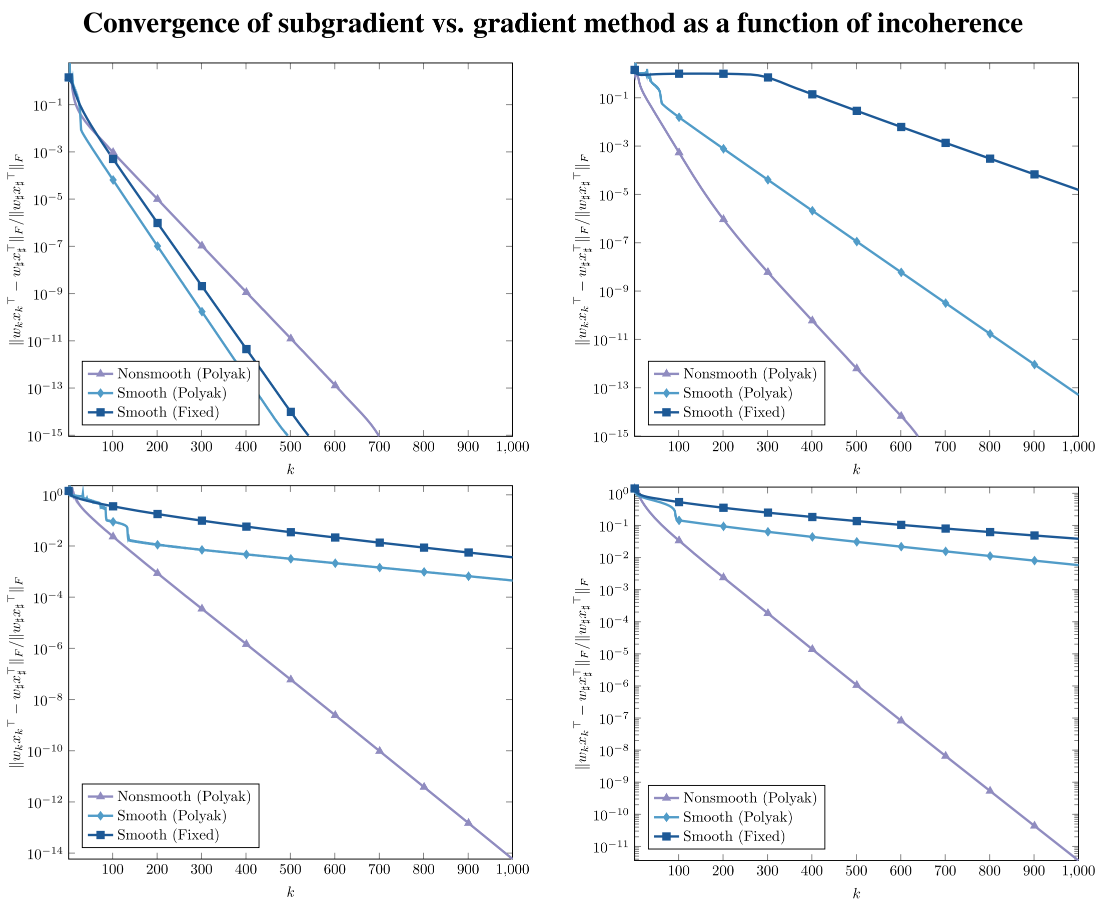

FIG. 12. Convergence behavior of Algorithm 1 for minimizing (1.3) vs. gradient descent for minimizing its smooth counterpart for $d = 100$ and incoherences $\mu_h^2 \in \{12, 23, 89, 100\}$ (clockwise, starting from top left).

measurement matrix $R \in \mathbb{C}^{m\times d}$, we set all entries equal to i.i.d. complex Gaussian random variables:

$$(R)_{i,k} = \sqrt{\frac{1}{2}}\left(X_{i,k} + jY_{i,k}\right), \quad X_{i,k}, Y_{i,k} \sim \mathsf{N}(0,1). \tag{6.7}$$

These are precisely the measurement matrices used in [**41**], the authors of which also provide a spectral initialization to find an ε-close initial estimate. However, this initialization requires $m \asymp d \log d$, so we opt for an artificial initialization as shown in (6.8), with $\delta := 0.25$:

$$w_0 := w_\sharp + \delta g_w, \quad x_0 := x_\sharp + \delta g_x, \quad g_x, g_w \sim \mathrm{Unif}(\mathbb{S}^{d-1}). \tag{6.8}$$

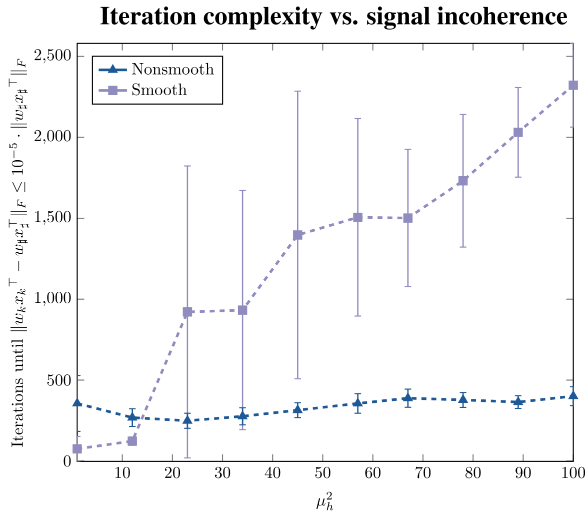

FIG. 13. Average number of iterations to reach normalized distance 10^{-5} for Algorithm 1 applied to (1.3) vs. gradient descent with Polyak stepsize applied to objective (6.10). Dashed lines are average over 25 independent realizations with error bars indicating one standard deviation.

We apply the subgradient methods from Algorithms 1 and 2, with the subgradient now calculated using Wirtinger calculus, as illustrated above. In Fig. 11, we generate synthetic instances with $\|w_\sharp\| = \|x_\sharp\| = 1$ and $p_{\text{fail}} \in \{0, 0.25, 0.45\}$ and evaluate the performance of our methods over a variety of measurement ratios $c := \frac{m}{d}$. We verify the linear rate of convergence of the projected subgradient method, as well as the effect of p_{fail} on the number of measurements required to converge to a minimizer. We observe that the partial DFT setting requires us to set m as big as $10 \cdot d$ for the highest corruption levels.

ROBUSTNESS TO SIGNAL INCOHERENCE. For completeness, we evaluate the sensitivity of the nonsmooth formulation (1.3) to the *incoherence* between $w_\sharp$ and the rows of L, given by

$$\mu_h^2 := \frac{\left\|Lw_\sharp\right\|_\infty^2}{\left\|w_\sharp\right\|_2^2}. \tag{6.9}$$

Intuitively, μ_h^2 captures the maximal correlation between rows of L and $\bar{w}$; in [**41**] the authors argue that signals with high μ_h^2 are the hardest to recover for smooth formulations. We generate noiseless instances where L is the partial $m \times d$ DFT matrix and R is a complex Gaussian matrix following (6.7), for a range of values of μ_h^2; for each such value, we set $x_\sharp \sim \text{Unif}(\mathbb{S}^{d-1})$ and $w_\sharp$ equal to a vector with μ_h^2 nonzero elements equal to 1 and all else equal to 0 (followed by normalization so that $w_\sharp \in \mathbb{S}^{d-1}$), which attains incoherence exactly μ_h^2 for this choice of L, following [**41**]. For simplicity, we set $w_0, x_0 \sim \text{Unif}(\mathbb{S}^{d-1})$, $d = 100$ and $m = 8 \cdot 2d$ and compare:

(i) the performance of Algorithm 1 applied to the objective (1.3),

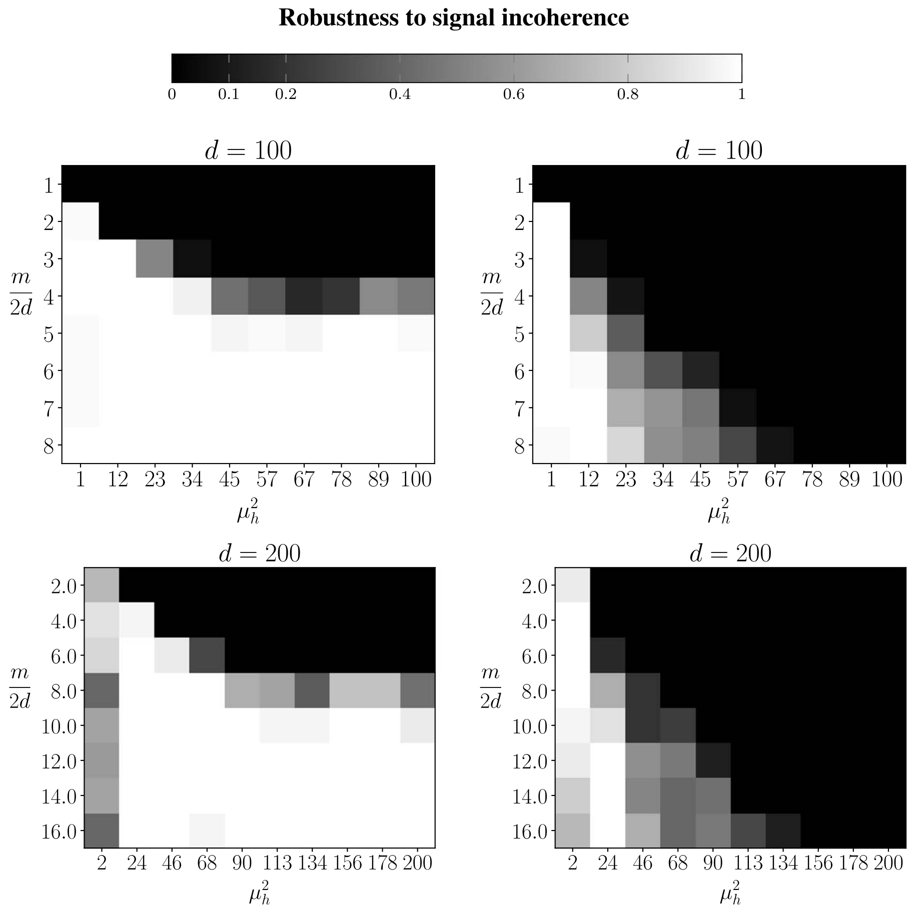

FIG. 14. Empirical recovery probabilities for various values of $\left(\frac{m}{2d}, \mu_h^2\right)$ over 50 independent trials. Lighter cells imply higher recovery probability. **Left**: Algorithm 1. **Right**: gradient descent with Polyak stepsize minimizing (6.10).

(ii) the performance of gradient descent (using the Wirtinger gradient) with Polyak stepsize on the smooth counterpart of (1.3), where we replace the ℓ_1-norm with the squared ℓ_2 loss as:

$$f_{\text{smooth}}(w, x) := \frac{1}{m} \sum_{i=1}^{m} \left| \langle \ell_i, w \rangle \langle r_i, x \rangle - y_i \right|^2, \tag{6.10}$$

(iii) the performance of gradient descent applied to (6.10) with a fixed stepsize η chosen among 2^{-i}, $i \in 1, \ldots, 15$ so that the final iterate distance is minimized.

Figure 12 illustrates that the nonsmooth objective is much more robust to variations on the incoherence of μ_h^2. As additional empirical evidence, Fig. 13 shows the average $\pm$ one standard deviation of the number of iterations required to reach normalized distance 10^{-5} for the two formulations, minimized using the Polyak stepsize. Perhaps surprisingly, the nonsmooth version remains practically constant over all choices of μ_h^2.

Finally, we generate a few transition plots for $d \in \{100, 200\}$ that illustrate the effects of incoherence on the nonsmooth and smooth flavors of the recovery objective. Following the setting of [41], we choose 10 equispaced values for $\mu_h^2 \in [1, d]$ and plot the empirical probability of recovery over 50 independent runs for various ratios $\frac{m}{2d}$. We consider the result of a run successful if it satisfies $\frac{\left\| w_t x_t^\top - w_\sharp x_\sharp^\top \right\|_F}{\left\| w_\sharp x_\sharp^\top \right\|_F} \leqslant 10^{-5}$ after at most 1000 iterations. Figure 14 shows that the nonsmooth objective is far more robust to signal incoherence, but it also reveals that it is not entirely unaffected by it; in particular, we can see that we need a higher threshold $\frac{m}{2d}$ to recover signals with higher incoherence after fixing the dimension d.

Data Availability Statement

No new data were generated or analyzed in support of this review.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions.

References

1. Aghasi, A., Ahmed, A. & Hand, P. (2017) BranchHull: convex bilinear inversion from the entrywise product of signals with known signs. arXiv preprint, arXiv:1702.04342.
2. Aghasi, A., Ahmed, A., Hand, P. & Joshi, B. (2018) A convex program for bilinear inversion of sparse vectors. arXiv preprint, arXiv:1809.08359.
3. Aghasi, A., Ahmed, A., Hand, P. & Joshi, B. (2019) Bilinear compressed sensing under known signs via convex programming. arXiv preprint, arXiv:1906.11636.
4. Ahmed, A., Aghasi, A. & Hand, P. (2018) Blind deconvolutional phase retrieval via convex programming. arXiv preprint, arXiv:1806.08091
5. Ahmed, A., Recht, B. & Romberg, J. (2014) Blind deconvolution using convex programming. *IEEE Trans. Inf. Theory*, **60**, 1711–1732.
6. Albano, P. & Cannarsa, P. (1999) Singularities of semiconcave functions in Banach spaces. *Stochastic Analysis, Control, Optimization and Applications. Systems Control Found. Appl*. Boston, MA: Birkhäuser Boston, pp. 171–190.
7. Bai, Y., Jiang, Q. & Sun, J. (2018) Subgradient descent learns orthogonal dictionaries. arXiv preprint, arXiv:1810.10702.
8. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. (2017) Julia: a fresh approach to numerical computing. *SIAM Rev.*, **59**, 65–98.
9. Borwein, J. & Zhu, Q. (2005) *Techniques of Variational Analysis*. New York: Springer.
10. Boucheron, S., Lugosi, G. & Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
11. Burke, J. & Ferris, M. (1995) A Gauss–Newton method for convex composite optimization. *Math. Program.*, **71**, 179–194.
12. Burke, J. V. (1985) Descent methods for composite nondifferentiable optimization problems. *Math. Program.*, **33**, 260–279.
13. Cai, T. T. & Zhang, A. (2015) ROP: matrix recovery via rank-one projections. *Ann. Stat.*, **43**, 102–138.
14. Candès, E., Li, X. & Soltanolkotabi, M. (2015) Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inf. Theory*, **61**, 1985–2007.
15. Candès, E. & Recht, B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**, 717–772.

16. Candès, E., Strohmer, T. & Voroninski, V. (2013) PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, **66**, 1241–1274.

17. Candes, E. J. & Plan, Y. (2011) Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, **57**, 2342–2359.

18. Chen, Y., Chi, Y. & Goldsmith, A. J. (2015) Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inf. Theory*, **61**, 4034–4059.

19. Chi, Y., Lu, Y. M. & Chen, Y. (2018) Nonconvex optimization meets low-rank matrix factorization: an overview. arXiv preprint, arXiv:1809.09573.

20. Choudhary, S. & Mitra, U. (2014) Sparse blind deconvolution: what cannot be done. *2014 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 3002–3006.

21. CorOpt (2019) Robust blind deconvolution: reference implementation. url https://github.com/COR-OPT/RobustBlindDeconv/releases/tag/imaiai.

22. Davenport, M. A. & Romberg, J. (2016) An overview of low-rank matrix recovery from incomplete observations. arXiv preprint, arXiv:1601.06422.

23. Davis, C. & Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, **7**, 1–46.

24. Davis, D., Drusvyatskiy, D., MacPhee, K. J. & Paquette, C. (2018) Subgradient methods for sharp weakly convex functions. arXiv preprint, arXiv:1803.02461.

25. Davis, D., Drusvyatskiy, D. & Paquette, C. (2017) The nonsmooth landscape of phase retrieval. arXiv preprint, arXiv:1711.03247.

26. Díaz, M. (2019) The nonsmooth landscape of blind deconvolution. arXiv preprint, arXiv:1911.08526.

27. Dontchev, A. & Rockafellar, R. (2009) Implicit functions and solution mappings. *Monographs in Mathematics*. Springer.

28. Drusvyatskiy, D. & Lewis, A. S. (2018) Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*

29. Drusvyatskiy, D. & Paquette, C. (2016) Efficiency of minimizing compositions of convex functions and smooth maps. arXiv preprint, arXiv:1605.00125

30. Duchi, J. & Ruan, F. (2017) Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. *Preprint arXiv*: 1705.02356.

31. Goffin, J. L. (1977) On convergence rates of subgradient optimization methods. *Math. Program.*, **13**, 329–347.

32. Huang, W. & Hand, P. (2018) Blind deconvolution by steepest descent algorithm on a quotient manifold. arXiv preprint, arXiv:1710.03309v2.

33. Ioffe, A. D. (2017) Variational Analysis of Regular Mappings: Theory and Applications. *Springer Monographs in Mathematics*. Cham: Springer.

34. Kech, M. & Krahmer, F. (2017) Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM J. Appl. Algebra Geom.*, **1**, 20–37.

35. Kreutz-Delgado, K. (2009) The complex gradient operator and the CR-calculus. arXiv preprint, arXiv:0906.4835.

36. Kuo, H.-W., Lau, Y., Zhang, Y. & Wright, J. (2019) Geometry and symmetry in short-and-sparse deconvolution. arXiv preprint, arXiv:1901.00256.

37. Lau, Y., Qu, Q., Kuo, H.-W., Zhou, P., Zhang, Y. & Wright, J. (2019) Short-and-sparse deconvolution–a geometric approach. arXiv preprint, arXiv:1908.10959.

38. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE*, **86**, 2278–2324.

39. Lewis, A. & Wright, S. (2015) A proximal method for composite minimization. *Math. Program.*, 1–46.

40. Lewis, A. S. & Pang, J.-S. (1998) Error bounds for convex inequality systems. *Generalized Convexity, Generalized Monotonicity: Recent Results (Luminy, 1996), vol. 27 of Nonconvex Optim. Appl*. Dordrecht: Kluwer Acad. Publ., pp. 75–110.

41. Li, X., Ling, S., Strohmer, T. & Wei, K. (2016) Rapid, robust, and reliable blind deconvolution via nonconvex optimization. arXiv preprint, arXiv:1606.04933.

42. Li, Y., Lee, K. & Bresler, Y. (2016) Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Trans. Inf. Theory*, **62**, 4266–4275.

43. Li, Y., Ma, C., Chen, Y. & Chi, Y. (2018) Nonconvex matrix factorization from rank-one measurements. arXiv preprint, arXiv:1802.06286.

44. Ling, S. & Strohmer, T. (2015) Self-calibration and biconvex compressive sensing. *Inverse Probl.*, **31**, 115002, 31.

45. Ma, C., Wang, K., Chi, Y. & Chen, Y. (2017) Implicit regularization in nonconvex statistical estimation: gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. arXiv preprint, arXiv:1711.10467.

46. Mordukhovich, B. (2006) Variational Analysis and Generalized Differentiation: Basic Theory. *I, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences],* vol. 330. Berlin: Springer.

47. Negahban, S. N., Ravikumar, P., Wainwright, M. J. & Yu, B. (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.*, **27**, 538–557.

48. Nurminskii, E. A. (1973) The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, **9**, 145–150.

49. Parikh, N. & Boyd, S. (2014) Block splitting for distributed optimization. *Math. Program. Comput.*, **6**, 77–102.

50. Penot, J.-P. (2013) *Calculus Without Derivatives*. Graduate Texts in Mathematics, vol. 266. New York: Springer.

51. Poliquin, R. & Rockafellar, R. (1996) Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, **348**, 1805–1838.

52. Recht, B., Fazel, M. & Parrilo, P. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**, 471–501.

53. Rockafellar, R. (1982) Favorable classes of Lipschitz-continuous functions in subgradient optimization. *Progress in nondifferentiable optimization, vol. 8 of IIASA Collaborative Proc. Ser. CP-82*. Laxenburg: Internat. Inst. Appl. Systems Anal., pp. 125–143.

54. Rockafellar, R. & Wets, R.-B. (1998) Variational Analysis. *Grundlehren der Mathematischen Wissenschaften*, vol. 317. Berlin: Springer.

55. Rolewicz, S. (1979) On paraconvex multifunctions. *Third Symposium on Operations Research (Univ. Mannheim, Mannheim, 1978), Section I, vol. 31 of Operations Res. Verfahren*. Königstein/Ts: Hain, pp. 539–546.

56. Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J. & Segev, M. (2015) Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Process. Mag.*, **32**, 87–109.

57. Sun, R. & Luo, Z.-Q. (2016) Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory*, **62**, 6535–6579.

58. Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M. & Recht, B. (2016) Low-rank solutions of linear matrix equations via procrustes flow. *Proceedings of the 33rd International Conference on International Conference on Machine Learning, vol. 48, ICML'16*, pp. 964–973. JMLR.org.

59. Vershynin, R. (2012) Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing*. Cambridge: Cambridge University Press, pp. 210–268.

60. Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

61. Wang, G., Giannakis, G. B. & Eldar, Y. C. (2018) Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Trans. Inf. Theory*, **64**, 773–794.

62. Zhang, Y., Lau, Y., Kuo, H.-w., Cheung, S., Pasupathy, A. & Wright, J. (2017) On the global geometry of sphere-constrained sparse blind deconvolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4894–4902.

A. Proofs of Section 4.1

A.1 *Proof of Proposition* 4.1

To see the first estimate, observe

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

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

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

The proof is complete.

A.2 *Proof of Theorem* 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 $\mathscr{S}_\nu^*$ may be written succinctly as

$$\operatorname{dist}((w,x), \mathscr{S}_\nu^*) = \sqrt{\inf_{(1/\nu) \leqslant |\alpha| \leqslant \nu} \left\{ \|w - \alpha\bar{w}\|_2^2 + \|x - (1/\alpha)\bar{x}\|_2^2 \right\}}.$$

Before we establish the general result, we first consider the simpler case, $d_1 = d_2 = 1$. The following bound holds:

$$|wx - 1| \geqslant \frac{1}{\sqrt{2}} \cdot \sqrt{\inf_{(1/\nu) \leqslant |\alpha| \leqslant \nu} \left\{ |w-\alpha|^2 + |x - (1/\alpha)|^2 \right\}},$$

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

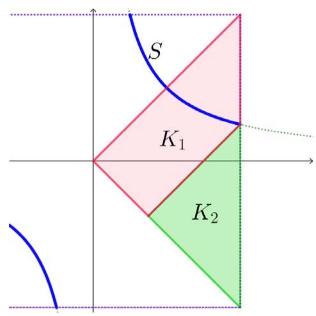

FIG. A15. The regions K_1, K_2 correspond to cases 1 and 2 of the proof of Claim A.1, respectively.

Proof of Claim. Consider a pair $(w,x) \in \mathbb{R}^2$ with $|w|, |x| \leqslant \nu$. It is easy to see that without loss of generality, we may assume $w \geqslant |x|$. We then separate the proof into two cases, which are graphically depicted in Fig. A.15.

CASE A:15 $w - x \leqslant \frac{\nu^2-1}{\nu}$. In this case, we will traverse from (w,x) to the $\mathscr{S}_\nu^*$ in the direction $(1,1)$. See Fig. A.15. 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 non-negative since $(w+x)^2 - (wx-1) = w^2 + x^2 + xw + 1 \geqslant 1$.

Set $\alpha = (w - t/\sqrt{2})$ and note the identity $1/\alpha = (x - t/\sqrt{2})$. Therefore,

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

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 $\alpha \geqslant 0, 1/\alpha \geqslant 0$, since

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

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

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

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

Case 2: $w - x \geqslant \frac{\nu^2-1}{\nu}$. 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 $|wx - 1| \geqslant \frac{t}{\sqrt{2}}$. Observe the following estimates:

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

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

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

To prove this bound, note that since $2b^2 \geqslant a^2 + b^2 = 1$, we have the $-\nu b - a/\nu \geqslant -\nu b \geqslant 1/\sqrt{2}$. 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 $p(0) \geqslant 1/\sqrt{2}$. Thus, it suffices to show that the solution s^* of the equation $p(s) = 1/\sqrt{2}$ satisfies $s^* \geqslant t$. To see this, we compute:

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

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

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

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

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 A.1, there exists an α satisfying $(1/\nu) \leqslant |\alpha| \leqslant \nu$, such that the following holds:

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

We now turn our attention to lower bounding the first term. Observe since $|w_1x_1 - \bar{w}_1\bar{x}_1| \leqslant \|wx^T - \bar{w}\bar{x}^T\|_F < 1/2$, we have

$$|w_1x_1| \geqslant |\bar{w}_1\bar{x}_1| - |w_1x_1 - \bar{w}_1\bar{x}_1| \geqslant (1/2)|\bar{w}_1\bar{x}_1| = 1/2.$$

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

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

Finally, we obtain the bound

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

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

B. RIP proofs

B.1 *Proof of Theorem* 4.6

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 $\mathscr{I} \subseteq \{1,\dots,m\}$ with $|\mathscr{I}| < m/2$. We begin with the following lemma.

Lemma B.1 (Pointwise concentration). The random variable $|\ell^\top X r|$ is sub-exponential with parameter $\sqrt{2}\eta^2$. Consequently, the estimate holds:

$$\mu_0 p_0 \leqslant \mathbb{E}|\ell^\top X r| \lesssim \eta^2. \tag{B.1}$$

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:

$$\frac{1}{m}\left|\|\mathscr{A}_{\mathscr{I}^c}(X)\|_1 - \|\mathscr{A}_{\mathscr{I}}(X)\|_1 - \mathbb{E}\left[\|\mathscr{A}_{\mathscr{I}^c}(X)\|_1 - \|\mathscr{A}_{\mathscr{I}}(X)\|_1\right]\right| \leqslant t. \tag{B.2}$$

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

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

which is the lower bound in (B.1). 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

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

where the second inequality follows since $\|\cdot\|_{\psi_1}$ is a norm and $\|XY\|_{\psi_1} \leqslant \|X\|_{\psi_2}\|Y\|_{\psi_2}$ [**60**, Lemma 2.7.7]. This bound has two consequences: first, $|\ell^\top Xr|$ is a sub-exponential random variable with parameter $\sqrt{2}\eta^2$ and second $\mathbb{E}|\ell^\top Xr| \leqslant \sqrt{2}\eta^2$, see [**60**, Exercise 2.7.2]. The first bound will be useful momentarily, while the second completes the proof of (B.1).

Next define the sub-exponential random variable

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

Standard results (e.g. [**60**, 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 D.4, to upper bound $\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^m Y_i\right| \geqslant t\right)$ completes the proof. □

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

$$\frac{1}{m}\left|\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}_{\mathcal{I}^c}(X)\|_1 - \|\mathcal{A}_{\mathcal{I}}(X)\|_1\right]\right| \leqslant t, \tag{B.3}$$

$$\frac{1}{m}\left|\|\mathcal{A}(X)\|_1 - \mathbb{E}\left[\|\mathcal{A}(X)\|_1\right]\right| \leqslant t. \tag{B.4}$$

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

$$\begin{aligned}\frac{1}{m}\|\mathcal{A}(X)\|_1 &\leqslant \frac{1}{m}\|\mathcal{A}(X_\star)\|_1 + \frac{1}{m}\|\mathcal{A}(X-X_\star)\|_1\\ &\leqslant \frac{1}{m}\mathbb{E}\|\mathcal{A}(X_\star)\|_1 + t + \frac{1}{m}\|\mathcal{A}(X-X_\star)\|_1 \qquad (B.5)\\ &\leqslant \frac{1}{m}\mathbb{E}\|\mathcal{A}(X)\|_1 + t + \frac{1}{m}\left(\mathbb{E}\|\mathcal{A}(X-X_\star)\|_1 + \|\mathcal{A}(X-X_\star)\|_1\right), \qquad (B.6)\end{aligned}$$

where (B.5) follows from (B.2) and (B.6) follows from the triangle inequality. To simplify the third term in (B.6), using singular value decompostion, 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

$$\frac{1}{m}\|\mathscr{A}(X - X_\star)\|_1 \leqslant \frac{1}{m}\|\mathscr{A}(X_1)\|_1 + \frac{1}{m}\|\mathscr{A}(X_2)\|_1$$
$$\leqslant c_2(\|X_1\|_F + \|X_2\|_F) \leqslant \sqrt{2}c_2\|X - X_\star\|_F \leqslant c_2\varepsilon, \tag{B.7}$$

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

$$\frac{1}{m}\|\mathscr{A}(X)\|_1 \leqslant \frac{1}{m}\mathbb{E}\|\mathscr{A}(X)\|_1 + t + 2c_2\varepsilon. \tag{B.8}$$

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

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

Assuming that $\varepsilon \leqslant 1/4$, we further deduce that

$$c_2 \leqslant \bar{\sigma} := \frac{2}{m}\sup_{X\in S_2}\mathbb{E}\|\mathscr{A}(X)\|_1 + 2t, \tag{B.9}$$

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

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

$$\frac{1}{m}\left\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X)\right\|_1 - \frac{1}{m}\left\|\mathscr{A}_{\hat{\mathscr{I}}}(X)\right\|_1$$
$$= \frac{1}{m}\left\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X_\star) + \mathscr{A}_{\hat{\mathscr{I}}^c}(X - X_\star)\right\|_1 - \frac{1}{m}\left\|\mathscr{A}_{\hat{\mathscr{I}}}(X_\star) + \mathscr{A}_{\hat{\mathscr{I}}}(X - X_\star)\right\|_1$$
$$\geqslant \frac{1}{m}\left\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X_\star)\right\|_1 - \frac{1}{m}\left\|\mathscr{A}_{\hat{\mathscr{I}}}(X_\star)\right\|_1 - \frac{1}{m}\left\|\mathscr{A}(X - X_\star)\right\|_1 \tag{B.10}$$
$$\geqslant \frac{1}{m}\mathbb{E}\left[\left\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X_\star)\right\|_1 - \left\|\mathscr{A}_{\hat{\mathscr{I}}}(X_\star)\right\|_1\right] - t - \frac{1}{m}\left\|\mathscr{A}(X - X_\star)\right\|_1 \tag{B.11}$$
$$\geqslant \frac{1}{m}\mathbb{E}\left[\left\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X)\right\|_1 - \left\|\mathscr{A}_{\hat{\mathscr{I}}}(X)\right\|_1\right] - t - \frac{1}{m}\left(\mathbb{E}\left\|\mathscr{A}(X - X_\star)\right\|_1 + \|\mathscr{A}(X - X_\star)\|_1\right) \tag{B.12}$$
$$\geqslant \frac{1}{m}\mathbb{E}\left[|\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X)\|_1 - \|\mathscr{A}_{\hat{\mathscr{I}}}(X)\|_1\right] - t - 2\bar{\sigma}\varepsilon, \tag{B.13}$$

where (B.10) uses the forward and reverse triangle inequalities, (B.11) follows from (B.3), the estimate (B.12) follows from the forward and reverse triangle inequalities and (B.13) follows from (B.7) and (B.9). Switching the roles of $\mathscr{I}$ and $\mathscr{I}^c$ in the above sequence of inequalities, and choosing $\varepsilon = t/4\bar{\sigma}$, we deduce

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

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

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

and therefore using (B.1), we conclude the RIP property

$$\mu_0 p_0 - \frac{3t}{2} \leqslant \frac{1}{m}\|\mathscr{A}(X)\|_1 \lesssim \eta^2 + \frac{3t}{2}, \qquad \forall X \in S_2. \tag{B.14}$$

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

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

Therefore, every $X \in S_2$ satisfies

$$\frac{1}{m}\left[\|\mathscr{A}_{\hat{\mathscr{I}}^c}(X)\|_1 - \|\mathscr{A}_{\hat{\mathscr{I}}}(X)\|_1\right] \geqslant \mu_0 p_0 \left(1 - \frac{2|\mathscr{I}|}{m}\right) - \frac{3t}{2}. \tag{B.15}$$

Setting $t = \frac{2}{3}\min\{\mu_0 p_0/2, \mu_0 p_0(1 - 2|\mathscr{I}|/m)/2\} = \frac{1}{3}\mu_0 p_0(1 - 2|\mathscr{I}|/m)$ in (B.14) and (B.15), we deduce the claimed estimates (4.4) and (4.5). Finally, let us estimate the probability of $\mathscr{E}$. Using Lemma B.1 and the union bound yields

$$\begin{aligned}
\mathbb{P}(\mathscr{E}^c) &\leqslant \sum_{X \in \mathscr{N}} \mathbb{P}\big\{(\text{B.3}) \text{ or } (\text{B.4}) \text{ fails at } X\big\} \\
&\leqslant 4|\mathscr{N}| \exp\left(-\frac{ct^2}{\eta^4} m\right) \\
&\leqslant 4\left(\frac{9}{\varepsilon}\right)^{2(d_1+d_2+1)} \exp\left(-\frac{ct^2}{\eta^4} m\right) \\
&= 4\exp\left(2(d_1+d_2+1)\ln(9/\varepsilon) - \frac{ct^2}{\eta^4} m\right),
\end{aligned}$$

where the second inequality follows from Lemma D.2 and c is a constant.

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

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

Hence as long as $m \geqslant \frac{18\eta^4 c_1 (d_1+d_2+1)\ln\left(c_2 + \frac{c_2}{1-2|\mathscr{I}|/m}\right)}{4c\mu_0^2 p_0^2 (1 - \frac{2|\mathscr{I}|}{m})^2}$, we can be sure $\mathbb{P}(\mathscr{E}^c) \leqslant 4\exp\left(-\frac{4c\mu_0^2 p_0^2(1-\frac{2|\mathscr{I}|}{m})^2}{18\eta^4} m\right)$. The result follows immediately. □

C. Initialization

C.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 [23] 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 C.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 $1 - c_1(\exp\left(-c_2 mt\right)$ 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 $\gamma_1 \geqslant c_3$ and $\gamma_2 \geqslant c_4$ and

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

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 $\gamma_1 \geqslant c_3$ and $\gamma_2 \geqslant c_4$ and

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

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:

$$L^{\text{init}} = \frac{1}{m}\Bigg(\underbrace{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}}_{m \cdot Y_0} + \underbrace{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} \left(P_{\bar{w}_\star} \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp + P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star}\right)}_{m \cdot Y_1} + \underbrace{\sum_{i \in \mathcal{I}_{\text{in}}^{\text{sel}}} P_{\bar{w}_\star}^\perp \ell_i \ell_i^\top P_{\bar{w}_\star}^\perp}_{m \cdot Y_2} + \underbrace{\sum_{i \in \mathcal{I}_{\text{out}}^{\text{sel}}} \ell_i \ell_i^\top}_{m \cdot Y_3}\Bigg). \tag{C1}$$

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

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

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

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

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

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

LEMMA C.2 For $t \geqslant 0$, we have

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

LEMMA C.3 There exist numerical constants $C, c > 0$ such that for any $t > 0$, we have

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

LEMMA C.4 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

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

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

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

The proof of the the theorem is complete. □

We now apply the Davis–Kahan $\sin\theta$ theorem [23] as stated in Lemma D.1. Throughout we assume that we are in the event described in Claim. C.1.

Proof of Proposition 5.1. We will use the notation from Theorem C.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 D.1 implies that

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

We will use these two inequalities to bound $\min_{s\in\{\pm1\}} \|\widehat{w}\widehat{x}^\top - s\bar{w}_\star\bar{x}_\star^\top\|_F$. To do so, we need to analyze $s_1 = \operatorname{argmin}_{s\in\{\pm1\}} \|\widehat{w} - s\bar{w}_\star\|$ and $s_2 = \operatorname{argmin}_{s\in\{\pm1\}} \|\widehat{x} - s\bar{x}_\star\|$. We split the argument into two cases.

Suppose first $s_1 = s_2$. Then

$$\begin{aligned}\|\widehat{w}\widehat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\|_F = \|\widehat{w}(\widehat{x} - s_2\bar{x}_\star)^\top - (\bar{w}_\star - s_1\widehat{w})\bar{x}_\star^\top\|_F &\leqslant \|\widehat{x} - s_2\bar{x}_\star\|_2 + \|\bar{w}_\star - s_1\widehat{w}\|_2 \\ &\leqslant \frac{2\sqrt{2}\max\{\|\Delta_1\|_{op}, \|\Delta_2\|_{op}\}}{\min\{\gamma_1,\gamma_2\}},\end{aligned}$$

as desired.

Suppose instead $s_1 = -s_2$. Then

$$\begin{aligned}\|\widehat{w}\widehat{x}^\top + \bar{w}_\star\bar{x}_\star^\top\|_F = \|\widehat{w}(\widehat{x} - s_2\bar{x}_\star)^\top + (\bar{w}_\star + s_2\widehat{w})\bar{x}_\star^\top\|_F &\leqslant \|\widehat{x} - s_2\bar{x}_\star\|_2 + \|\bar{w}_\star - s_1\widehat{w}\|_2 \\ &\leqslant \frac{2\sqrt{2}\max\{\|\Delta_1\|_{op}, \|\Delta_2\|_{op}\}}{\min\{\gamma_1,\gamma_2\}},\end{aligned}$$

as desired. Bounding $\max\{\|\Delta_1\|_{op}, \|\Delta_2\|_{op}\}$ using Theorem C.1 completes the proof. □

The next sections present the proof of Lemmas C.1–C.4. 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.

C.1.1 *Proof of Lemma C.1* 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 $\mathscr{I}_{\text{in}}^{\text{sel}}$.

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

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

Additionally, define the conditional expected value

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

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

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

Then we prove the following claim.

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

1. $\frac{|\mathscr{I}_{\text{in}}^{\text{sel}}|}{m} \geqslant \frac{1-2p_{\text{fail}}}{2}$.
2. $\mathbb{P}\left(\mathscr{I}_{\text{in}}^{Q} \supseteq \mathscr{I}_{\text{in}}^{\text{sel}}\right) \geqslant 1 - \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right)$.
3. $\mathbb{P}\left(|\mathscr{I}_{\text{in}}^{Q}| \geqslant \frac{6251m}{10000}\right) \leqslant \exp\left(-\frac{m}{2\cdot 10^8}\right)$.
4. $\mathbb{P}\left(\frac{1}{|\mathscr{I}_{\text{in}}^{Q}|}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \geqslant \omega_{\text{fail}} + t\right) \leqslant \exp\left(-c\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\frac{m(1-2p_{\text{fail}})}{2}\right) + \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right)$.

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

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

it follows that:

$$\begin{aligned} S - y_0 = \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}^{\text{sel}}}(1 - \ell_{i,1}^2) &\geqslant \frac{|\mathscr{I}_{\text{in}}^{\text{sel}}|}{m} - \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}}\ell_{i,1}^2 \geqslant \frac{1-2p_{\text{fail}}}{2} - \frac{|\mathscr{I}_{\text{in}}^{Q}|}{m\left|\mathscr{I}_{\text{in}}^{Q}\right|}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}}\ell_{i,1}^2 \\ &\geqslant \frac{1-2p_{\text{fail}}}{2} - \frac{631351}{1000000}\omega_{\text{fail}} \geqslant 0.04644344, \end{aligned}$$

where the first three inequalities follow by the definition of the event $\mathscr{E}$. The fourth inequality follows by the definition of $\mathscr{E}$ and Lemma D.5, which implies $\omega_{\text{fail}} \leqslant 0.56$ when $p_{\text{fail}} = 0.1$ and that the difference is minimized over $p_{\text{fail}} \in [0, 0.1]$ at the endpoint $p_{\text{fail}} = 0.1$. To get the claimed probabilities, we note that by Lemma D.5, we have $\omega_{\text{fail}} \geqslant 0.5$ 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

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

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

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

To that end, first note that

$$\begin{aligned}
\texttt{med}(\{|y_i|\}_i^m) &= \min\left\{ |y_j| : j \in [m], \sum_{i=1}^m \mathbf{1}\{|y_i| \leqslant |y_j|\} \geqslant \frac{m}{2} \right\} \\
&= \min\left\{ |y_j| : j \in [m], \sum_{i=1}^m \mathbf{1}\{|y_i| \leqslant |y_j|\} \geqslant \frac{|\mathscr{I}_{\text{in}}|}{2(1-p_{\text{fail}})} \right\} \\
&\leqslant \min\left\{ |y_j| : j \in \mathscr{I}_{\text{in}}, \sum_{i=1}^m \mathbf{1}\{|y_i| \leqslant |y_j|\} \geqslant \frac{|\mathscr{I}_{\text{in}}|}{2(1-p_{\text{fail}})} \right\} \\
&\leqslant \min\left\{ |y_j| : j \in \mathscr{I}_{\text{in}}, \sum_{i \in \mathscr{I}_{\text{in}}} \mathbf{1}\{|y_i| \leqslant |y_j|\} \geqslant \frac{|\mathscr{I}_{\text{in}}|}{2(1-p_{\text{fail}})} \right\} \\
&= \texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|\}_{i \in \mathscr{I}_{\text{in}}}\right),
\end{aligned}$$

where the first equality follows since $\frac{|\mathscr{I}_{\text{in}}|}{2(1-p_{\text{fail}})} = \frac{(1-p_{\text{fail}})m}{2(1-p_{\text{fail}})} = m/2$, 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

$$\begin{aligned}
\mathbb{P}\left(\texttt{med}(\{|y_i|\}_i^m) \geqslant Q_{\text{fail}} M\right) &\leqslant \mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|\}_{i \in \mathscr{I}_{\text{in}}}\right) \geqslant Q_{\text{fail}} M\right) \\
&= \mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|/M\}_{i \in \mathscr{I}_{\text{in}}}\right) \geqslant Q_{\text{fail}}\right),
\end{aligned}$$

and our remaining task is to bound this probability.

To bound this probability, we apply Lemma D.3 to the i.i.d. sample $\{|y_i|/M : i \in \mathscr{I}_{\text{in}}\}$, which is sampled from the distribution of $\mathscr{D}$ of $|ab|$ where $a, b \sim \mathsf{N}(0,1)$ and a, b are independent. Therefore, using the identities (for $i \in \mathscr{I}_{\text{in}}$)

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

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

$$\begin{aligned}\mathbb{P}\left(\texttt{quant}_{\frac{1}{2(1-p_{\text{fail}})}}\left(\{|y_i|/M\}_{i\in\mathscr{I}_{\text{in}}}\right) \geqslant Q_{\text{fail}}\right) &\leqslant \exp\left(\frac{m(q-p)^2}{2(q-p)/3+2q(1-q)}\right)\\ &= \exp\left(\frac{m(q-p)}{2/3+6q}\right)\\ &= \exp\left(-\frac{3(1-2p_{\text{fail}})m}{8(1-p_{\text{fail}})(2+18q)}\right)\\ &\leqslant \exp\left(-\frac{3(1-2p_{\text{fail}})}{160}m\right),\end{aligned}$$

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 $\mathscr{I}_{\text{in}}^{Q} \supseteq \mathscr{I}_{\text{in}}^{\text{sel}}$ with high probability, as desired.

Part 3. Since $\{|y_i|/M \colon i \in \mathscr{I}_{\text{in}}\}$ 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 $i \in \mathscr{I}_{\text{in}}$ that

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

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

$$\begin{aligned}\mathbb{P}\left(\frac{6251m}{10000} \leqslant |\mathscr{I}_{\text{in}}^{Q}|\right) = \mathbb{P}\left(\frac{m}{10000} \leqslant |\mathscr{I}_{\text{in}}^{Q}| - \frac{5m}{8}\right) &\leqslant \mathbb{P}\left(\frac{m}{10000} \leqslant |\mathscr{I}_{\text{in}}^{Q}| - \mathbb{E}|\mathscr{I}_{\text{in}}^{Q}|\right)\\ &\leqslant \exp\left(-\frac{(1/10000)^2 m}{2(1-p_{\text{fail}})}\right) \leqslant \exp\left(-\frac{m}{2\cdot 10^8}\right),\end{aligned}$$

as desired.

Part 4. First write

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{|\mathscr{I}_{\text{in}}^{Q}|}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \geqslant \omega_{\text{fail}} + t\right)\\ &= \mathbb{P}\left(\frac{1}{|\mathscr{I}_{\text{in}}^{Q}|}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \geqslant \omega_{\text{fail}} + t \text{ and } |\mathscr{I}_{\text{in}}^{Q}| \supseteq \mathscr{I}_{\text{in}}^{\text{sel}}\right) + \mathbb{P}\left(\frac{1}{|\mathscr{I}_{\text{in}}^{Q}|}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \geqslant \omega_{\text{fail}} + t \text{ and } \mathscr{I}_{\text{in}}^{Q} \not\supseteq \mathscr{I}_{\text{in}}^{\text{sel}}\right)\\ &\leqslant \mathbb{P}\left(\frac{1}{|\mathscr{I}_{\text{in}}^{Q}|}\sum_{i\in\mathscr{I}_{\text{in}}^{Q}} \ell_{i,1}^2 \geqslant \omega_{\text{fail}} + t \text{ and } |\mathscr{I}_{\text{in}}^{Q}| \geqslant \frac{m(1-2p_{\text{fail}})}{2}\right) + \exp\left(-\tfrac{3(1-2p_{\text{fail}})}{160}m\right),\end{aligned}$$

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

To that end, notice that

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

Observe that for any index $i\in\mathscr{I}_{\text{in}}$ and $t\geqslant 0$, we have $\mathbb{P}\left(\ell_{i,1}^{2}\geqslant t\mid i\in\mathscr{I}_{\text{in}}^{Q}\right)=\mathbb{P}\left(a^{2}\geqslant t\mid |ab|\leqslant Q_{\text{fail}}\right)$, where $a,b\sim\mathsf{N}(0,1)$ are independent. In addition, we have $q_{\text{fail}}=P(|ab|\leqslant Q_{\text{fail}})=\frac{5-2p_{\text{fail}}}{8(1-p_{\text{fail}})}\geqslant 5/8>1/2$, where we have used the fact that q_{fail} is an increasing function of p_{fail}. Therefore, applying Lemma D.4, we have the following bound:

$$\mathbb{P}\left(\ell_{i,1}^{2}\geqslant t\mid i\in\mathscr{I}_{\text{in}}^{Q}\right)\leqslant 2\exp(-t/2K_{1})\qquad\text{for all } t\geqslant 0 \text{ and } i\in\mathscr{I}_{\text{in}},$$

where K_1 is a numerical constant. In particular, by Theorem D.5 and the identity $\omega_{\text{fail}}=\mathbb{E}\left[a^{2}\geqslant t\mid |ab|\leqslant Q_{\text{fail}}\right]$, we have the following bound:

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

for numerical constants c and K, as desired. □

The proof is complete.

C.1.2 *Proof of Lemma C.2* Our goal is to bound the operator norm of the following matrix:

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

Simplifying, we find that

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

Evidently, $\|Y_{1}\|_{\text{op}}\leqslant\left\|\lambda_{2:d_{1}}\right\|_{2}$, 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.

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

$$F(a_{1},\dots,a_{m})=\left\|\frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}^{\text{sel}}}\ell_{i,1}a_{i}\right\|_{2}.$$

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

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

Moreover, the following bound holds:

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

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

$$\begin{aligned} |F(A) - F(B)| &\leqslant \frac{1}{m}\|(A - B)(\ell_{i,1}\mathbf{1}\{i \in \mathscr{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2 \\ &\leqslant \frac{1}{m}\|(A - B)\|_{\text{op}}\|(\ell_{i,1}\mathbf{1}\{i \in \mathscr{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2 \leqslant \widehat{\eta}\|A - B\|_F, \end{aligned}$$

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 D.6, imply that

$$\begin{aligned} &\mathbb{P}\Bigg(F(\ell_{1,2:d}, \dots, \ell_{m,2:d}) - \mathbb{E}\left[F(\ell_{1,2:d}, \dots, \ell_{m,2:d}) \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathscr{I}_{\text{in}}^{\text{sel}}\right] \\ &\geqslant t \;\Bigg|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathscr{I}_{\text{in}}^{\text{sel}}, \Bigg) \\ &\leqslant \exp\left(-\frac{mt^2}{8}\right). \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}\left[F(\ell_{1,2:d}, \dots, \ell_{m,2:d}) \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathscr{I}_{\text{in}}^{\text{sel}}\right] \\ &\leqslant \sqrt{\mathbb{E}\left[F(\ell_{1,2:d}, \dots, \ell_{m,2:d})^2 \;\middle|\; \hat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathscr{I}_{\text{in}}^{\text{sel}}\right]} \\ &= \sqrt{\frac{1}{m^2}\mathbb{E}\left[\sum_{i \in \mathscr{I}_{\text{in}}^{\text{sel}}} \ell_{i,1}^2(d_1 - 1) \;\middle|\; \hat{\eta} < \frac{2}{\sqrt{m}}\right]} \leqslant 2\sqrt{\frac{d_1 - 1}{m}}. \end{aligned}$$

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

$$E(a) = \frac{1}{m}\sqrt{\sum_{i \in \mathscr{I}_{\text{in}}^{\text{sel}}} a_i^2} = \frac{1}{m}\|(a_i\mathbf{1}\{i \in \mathscr{I}_{\text{in}}^{\text{sel}}\})_{i=1}^m\|_2$$

is m^{-1}-Lipschitz continuous. Moreover, we have that $\mathbb{E}\left[E(\ell_{1,i},\ldots,\ell_{1,d})\right] \leqslant \frac{1}{m}\mathbb{E}\left[\|(\ell_{1,i})_{i=1}^{m}\|_2\right] \leqslant m^{-1/2}$.

Therefore, by Gaussian concentration, we have

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

as desired. □

To complete the proof, observe that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\lambda_{2:d_1}\right\|_2 \geqslant 2\sqrt{\frac{d_1-1}{m}} + t\right) \\
&= \mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}^{\text{sel}}} \ell_{i,1}\ell_{i,2:d_1}\right\|_2 \geqslant 2\sqrt{\frac{d_1-1}{m}} + t\right) \\
&\leqslant \mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}^{\text{sel}}} \ell_{i,1}\ell_{i,2:d_1}\right\|_2 \geqslant 2\sqrt{\frac{d_1-1}{m}} + t \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}\right)\mathbb{P}\left(\widehat{\eta} < \frac{2}{\sqrt{m}}\right) + \mathbb{P}\left(\widehat{\eta} \geqslant \frac{2}{\sqrt{m}}\right) \\
&\leqslant \mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d}) \geqslant 2\sqrt{\frac{d_1-1}{m}} + t \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}\right) + \exp\left(-\frac{m}{2}\right),
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d}) \geqslant 2\sqrt{\frac{d_1-1}{m}} + t \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}\right) \\
&= \mathbb{E}_{\mathscr{I}_{\text{in}}^{\text{sel}},\{\ell_{i,1}\}_{i=1}^m}\left[\mathbb{P}\left(F(\ell_{1,2:d},\ldots,\ell_{m,2:d}) \geqslant 2\sqrt{\frac{d_1-1}{m}} + t \;\middle|\; \widehat{\eta} < \frac{2}{\sqrt{m}}, \{\ell_{1,i}\}_{i=1}^m, \mathscr{I}_{\text{in}}^{\text{sel}}\right)\right] \\
&\leqslant \exp\left(-\frac{mt^2}{8}\right),
\end{aligned}$$

which completes the proof.

C.1.3 *Proof of Lemma C.3* Observe the equality

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

Therefore, we seek to bound the following operator norm:

$$\left\|Y_2 - S_{\text{in}}\left(I_{d_1} - e_1e_1^\top\right)\right\|_{\text{op}} = \left\|\frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}^{\text{sel}}} (\ell_{i,2:d_1}\ell_{i,2:d_1}^\top - I_{d_1-1})\right\|_{\text{op}}.$$

Using the tower rule for expectations and appealing to Corollary D.1, we therefore deduce

$$\mathbb{P}\left(\left\|Y_2 - S_{\text{in}}\left(I_{d_1} - e_1e_1^\top\right)\right\|_{\text{op}} \geqslant C\sqrt{\frac{d_1}{m} + t}\right)$$

$$\leqslant \mathbb{E}_{\mathscr{I}_{\text{in}}^{\text{sel}}}\left[\mathbb{P}\left(\left\|Y_2 - S_{\text{in}}\left(I_{d_1} - e_1e_1^\top\right)\right\|_{\text{op}} \geqslant C\sqrt{\frac{d_1}{m} + t} \;\middle|\; \mathscr{I}_{\text{in}}^{\text{sel}} = \mathscr{I}\right)\right] \leqslant 2\exp(-cmt),$$

as desired.

C.1.4 *Proof of Lemma* C.4

NOISE MODEL N1. Under this noise model, we write

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

The proof follows by repeating the conditioning argument as in the proof of Lemma C.3.

NOISE MODEL N2. Observe that

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

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

Appealing to Corollary D.1, the result follows immediately.

C.2 *Proof of Proposition* 5.2

We will assume that $\|\hat{w}\hat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\|_F \leqslant \|\hat{w}\hat{x}^\top + \bar{w}_\star\bar{x}_\star^\top\|_F$. We will show that with high probability, $|\widehat{M} - M| \leqslant \delta M$, and moreover in this event if $\delta < 1$, we have $\widehat{M} > 0$. The other setting $\|\hat{w}\hat{x}^T - \bar{w}_\star\bar{x}_\star^\top\|_F \geqslant \|\hat{w}\hat{x}^T + \bar{w}_\star\bar{x}_\star^\top\|_F$ 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 $m \geqslant \frac{c_1(d_1+d_2+1)}{(1-\frac{2|\mathscr{I}|}{m})^2}\ln\left(c_2 + \frac{1}{1-2|\mathscr{I}|/m}\right)$, then with probability at least $1 - 4\exp\left(-c_3(1-\frac{2|\mathscr{I}|}{m})^2 m\right)$, we have

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

and

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

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

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

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

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

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

$$|\widehat{g}(0) - g(0)| \leqslant \frac{M}{m}\sum_{i=1}^{m}|\ell_i^\top \widehat{w}\widehat{x}^\top r_i - \ell_i^\top \bar{w}_\star \bar{x}_\star^\top r_i| \leqslant c_5 M \left\|\widehat{w}\widehat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F. \tag{C2}$$

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

$$\begin{aligned}
\hat{g}(a) &= \frac{1}{m}\sum_{i=1}^{m}\left|y_i - (1+a)M\ell_i^\top \widehat{w}\widehat{x}^\top r_i\right| \\
&= \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}}\left|\ell_i^\top \bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top \widehat{w}\widehat{x}^\top r_i\right| + \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{out}}}\left|y_i - (1+a)M\ell_i^\top \widehat{w}\widehat{x}^\top r_i\right| \\
&\geqslant \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{in}}}\left|\ell_i^\top \bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top \widehat{w}\widehat{x}^\top r_i\right| - \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{out}}}\left|\ell_i^\top \bar{w}\bar{x}^\top r_i - (1+a)M\ell_i^\top \widehat{w}\widehat{x}^\top r_i\right| \\
&\quad + \frac{1}{m}\sum_{i\in\mathscr{I}_{\text{out}}}\left|y_i - \ell_i^\top \bar{w}\bar{x}^\top r_i\right| \\
&\geqslant g(0) + \mu_3\|(1+a)M\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\|_F \\
&\geqslant \widehat{g}(0) + \mu_3\|(1+a)M\widehat{w}\widehat{x}^\top - \bar{w}\bar{x}^\top\|_F - c_5 M\left\|\widehat{w}\widehat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F \\
&\geqslant \widehat{g}(0) + \mu_3|a|M - \left(\mu_3 M + c_5 M\right)\left\|\widehat{w}\widehat{x}^\top - \bar{w}_\star\bar{x}_\star^\top\right\|_F,
\end{aligned}$$

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

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

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

D. Auxiliary lemmas

D.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 [23]). 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 D.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

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

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

Lemma D.2 (Lemma 3.1 in [17]). Let $S_r := \left\{X \in \mathbb{R}^{d_1\times d_2} \mid \text{rank}(X) \leqslant r, \|X\|_F = 1\right\}$. There exists an ε-net $\mathcal{N}$(with respect to $\|\cdot\|_F$) of S_r obeying

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

D.2 *Concentration inequalities*

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

Theorem D.1 (Hoeffding's inequality—Theorem 2.2.2 in [60]). Let $X_1, \ldots, X_N$ be independent symmetric Bernoulli random variables. Then for any $t \geqslant 0$, we have

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

Theorem D.2 (Bernstein's inequality—Theorem 2.8.4 in [60]). Let $X_1, \ldots, X_N$ be independent mean-zero random variables, such that for $|X_i| \leqslant K$ for all i. Then for any $t \geqslant 0$, we have

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

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

Theorem D.3 (Sub-gaussian concentration—Theorem 2.6.3 in [60]). 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 $t \geqslant 0$, we have

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

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

THEOREM D.4 (Sub-exponential concentration—Theorem 2.8.2 in [**60**]). 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 $t \geqslant 0$, we have that

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

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

THEOREM D.5 (Corollary 2.8.3 in [**60**]). Let $X_1, \ldots, X_m$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geqslant 0$, we have

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

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

THEOREM D.6 (Theorem 5.6 in [**10**]). 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_f-Lipschitz function. Then, for every $t \geqslant 0$, we have

$$\mathbb{P}\left(f(X) - \mathbb{E}f(X) \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2\mathsf{L}_f^2}\right).$$

The following concentration inequalities deal with quantiles of distributions:

LEMMA D.3 Let $X_1, \ldots, X_m$ be an i.i.d. sample with distribution $\mathscr{D}$, choose Q_q to be the q population quantile of the distribution $\mathscr{D}$, that is $q = \mathbb{P}\left(X_1 \leqslant Q_q\right)$, and let $p \in (0, 1)$ be any probability with $p < q$. Then,

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

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, $\mathtt{quant}_p(\{X_i\}_i^m) \geqslant Q_q$ if, and only if, $\frac{1}{m}\sum_{i=1}^m \mathbf{1}\{X_i \leqslant Q_q\} \leqslant p$. Notice that $\mathbf{1}\{X_i \leqslant Q_q\} \sim \mathrm{B}(q)$ are i.i.d. Bernoulli random variables and thus $\mathrm{Var}(\mathbf{1}\{X_i \leqslant Q_q\}) = q(1-q)$. Then, the result follows by applying Bernstein's inequality (Theorem D.4) to $\frac{1}{m}\sum \mathbf{1}\{X_i \leqslant Q_q\} - q$. □

LEMMA D.4 Let a, b be i.i.d. sub-Gaussian random variables. For any $Q > 0$ such that $q := \mathbb{P}(|ab| \leqslant Q) > 1/2$, consider the random variable c^2 defined as a^2 conditioned on the event $|ab| \leqslant Q$, namely for all t

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

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

$$\mathbb{P}\left(c^2 \geqslant t\right) \leqslant 2\exp(-t/2K),$$

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

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

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

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

With this, we can bound the probability

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

as claimed. □

The following theorem from [59] is especially useful in bounding the operator norm of random matrices:

THEOREM D.7 (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 $t \geqslant 0$, one has

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

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

Proof. The theorem is a direct corollary of [59, Theorem 5.39]. Specifically, the concavity of the square root gives us $\sqrt{a} + \sqrt{b} \leqslant \sqrt{2}\sqrt{a+b}$, implying that

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

Additionally, [59, Theorem 5.39] gives us that

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

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

Let us record the following elementary consequence.

COROLLARY D.1 Let $a_1, \ldots, a_m \in \mathbb{R}^d$ be independent, sub-Gaussian, isotropic random vectors in $\mathbb{R}^n$, and let $\mathscr{I} \subset \{1, \ldots, m\}$ be an arbitrary set. Then, for every $t \geqslant 0$, one has

$$\mathbb{P}\left(\left\|\frac{1}{m}\sum_{i\in\mathscr{I}}(a_ia_i^\top - I_d)\right\|_{\text{op}} \leqslant C\sqrt{\frac{d}{m} + t}\right) \geqslant 1 - 2\exp\left(-cmt\right),$$

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

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

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

Appealing to Theorem D.7, we therefore deduce for any $\gamma > 0$ the estimate

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

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

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

completes the proof. □

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 $\omega_{\mathrm{fail}} = \mathbb{E}[a^2 \mid |ab| \leqslant Q_{\mathrm{fail}}]$.

LEMMA D.5 The function $\omega : [0,1] \to \mathbb{R}_+$ given by

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

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

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

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

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

Let $0 \leqslant Q \leqslant Q'$ be arbitrary numbers, then

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

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

$$\mathbb{E}[a^2 \mid |ab| \leqslant Q] = \int_0^\infty \mathbb{P}(a^2 \geqslant t \mid |ab| \leqslant Q)\mathrm{d}t \leqslant \int_0^\infty \mathbb{P}(a^2 \geqslant t \mid |ab| \leqslant Q')\mathrm{d}t = \mathbb{E}[a^2 \mid |ab| \leqslant Q'],$$

where the inequality follows from the claim and the equalities follow from the identity $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geqslant t)\mathrm{d}t$ for non-negative random variables X. Hence, ω is a non-decreasing function.

The above implies that for any $p_{\mathrm{fail}} \in [0, 0.1]$, we have $\omega(0) \leqslant \omega_{\mathrm{fail}} \leqslant \omega(0.1)$. 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 $\mathbb{P}(|ab| \leqslant Q) \geqslant 0.6679 \geqslant 2/3 = q_{\mathrm{fail}}(0.1)$. Then computing numerically (with precision set to 32 digits), we obtain $\omega(0.1) \leqslant \mathbb{E}[a^2 \mid |ab| \leqslant Q] \leqslant 0.56$. Similarly, we find that if we set $Q = 0.5$ we get $\mathbb{P}(|ab| \leqslant Q) \leqslant 0.5903 \leqslant 5/8 = q_{\text{fail}}(0)$. Then evaluating we find $\omega(0) \geqslant \mathbb{E}[a^2 \mid |ab| \leqslant Q] \geqslant 0.5$.

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

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

is non-increasing. Our goal is to show that $h_t' \leqslant 0$. In order to prove this result, we proceed as follows. Define

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

and

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

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

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

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

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

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

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

Thus, we have

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

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

$$\left(\int_0^{\sqrt{t}}\int_0^{Q/x} \exp(-(x^2+y^2)/2)\,\mathrm{d}y\,\mathrm{d}x\right)$$
$$\geqslant \left(\int_0^{\sqrt{t}} \frac{\sqrt{t}}{x}\exp(-(Q^2/x^2 - Q^2/t)/2)\int_0^{Q/\sqrt{t}} \exp(-(x^2+y^2)/2)\mathrm{d}y\,\mathrm{d}x\right).$$

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

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

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

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

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

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

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

$$\frac{Q-x^2}{qx}\int_0^{Q/x}\exp(-y^2/2)\mathrm{d}y \geqslant \frac{Q-x^2}{Qx}\frac{Q}{x}\exp(-Q^2/2x^2) \geqslant -\exp(-Q^2/2x^2)$$

proving the claim. □

Thus, the proof is complete. □