

Preconditioned subgradient method for composite optimization: overparameterization and fast convergence

Mateo Díaz* Liwei Jiang[†] Abdel Ghani Labassi[‡]

Abstract

Composite optimization problems involve minimizing the composition of a smooth map with a convex function. Such objectives arise in numerous data science and signal processing applications, including phase retrieval, blind deconvolution, and collaborative filtering. The subgradient method achieves local linear convergence when the composite loss is well-conditioned. However, if the smooth map is, in a certain sense, ill-conditioned or overparameterized, the subgradient method exhibits much slower sublinear convergence even when the convex function is well-conditioned. To overcome this limitation, we introduce a Levenberg-Morrison-Marquardt subgradient method that converges linearly under mild regularity conditions at a rate determined solely by the convex function. Further, we demonstrate that these regularity conditions hold for several problems of practical interest, including square-variable formulations, matrix sensing, and tensor factorization. Numerical experiments illustrate the benefits of our method.

*Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA; <https://mateodd25.github.io>. MD was partially supported by NSF awards CCF 2442615 and DMS 2502377.

[†]Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN 47906, USA; <https://liweijiang97.github.io>.

[‡]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA; <https://aglabassi.github.io/website/>.

Contents

1	Introduction	3
2	Preliminaries	7
3	Algorithm and assumptions	8
3.1	Algorithmic description	8
3.2	Regularity of the parameterization	9
3.3	Regularity for nonsmooth outer functions	10
3.4	Regularity for smooth outer functions	11
4	General convergence guarantees	12
4.1	Guarantees for nonsmooth losses	13
4.2	Guarantees for smooth losses	16
4.3	Guarantees under local regularity	17
5	Consequences for statistical recovery problems	18
5.1	Squared-variable formulations	18
5.2	Matrix recovery problems	19
5.3	Tensor factorization	24
6	Numerical experiments	26
A	Missing proofs from Section 3	40
B	Missing proofs from Section 4	41
C	Missing proofs from Section 5	50
D	Auxiliary proofs and results	75
E	Computing the preconditioner	83

1 Introduction

The goal of *composite optimization problems* is to minimize

$$\min_{x \in \mathbf{R}^d} f(x) \quad \text{with} \quad f = h \circ F, \tag{1}$$

where $h: \mathbf{R}^m \rightarrow \mathbf{R}$ is a—possibly nonsmooth—convex function and $F: \mathbf{R}^d \rightarrow \mathbf{R}^m$ is a smooth mapping. Taking h or F as the identity map recovers convex and smooth optimization; thus, this formulation strictly extends both and amounts to a much richer class of nonsmooth nonconvex problems. Classical nonlinear least squares are a prominent example of this framework [9, 59, 83]. Recently, composite optimization has gained renewed interest due to its applications in data science, including phase retrieval, matrix completion, and tensor factorization [17, 38, 39, 97].

First-order methods, such as gradient descent, are the dominant algorithmic solution for large-scale composite problems. Under favorable growth conditions of the loss function, these methods converge linearly towards solutions provided good initialization [17, 23]. For instance, when the objective function f is β -smooth and is locally α -strongly convex, gradient descent converges linearly at a rate that depends on the condition number β/α . Inconveniently, this condition number might worsen drastically depending on the choice of the smooth map F . To illustrate this point, it is useful to think of the smooth map F as a *parameterization*: minimizing (1) is akin to solving a constrained problem

$$\min_{x \in \mathbf{R}^d} f(x) = \min_{z \in \text{Im } F} h(z), \tag{2}$$

where $\text{Im } F$ denotes the image of F . Intuitively, when $\text{Im } F$ is sufficiently “benign,” the intrinsic complexity of (2) should be dictated by the conditioning of h restricted to $\text{Im } F$ and not by the specific parameterization F .

Two factors concerning the parameterization F cause the conditioning of f to differ from that of h on $\text{Im } F$: (i) ill-conditionedness, or, even worse, (ii) an excess of parameters. For concreteness, consider a simple example, suppose we want to factorize a rank- r^* positive semidefinite (PSD) matrix M^* . In large-scale settings—where direct eigen- or singular-value decompositions are prohibitively costly—researchers turn to iterative schemes on low-rank parameterizations. The celebrated Burer-Monteiro approach [11, 12] parameterizes low-rank matrices via an explicit factorization, $F(U) = UU^\top$ with $U \in \mathbf{R}^{d \times r}$ and $r \geq r^*$, and aims to solve

$$\min_{U \in \mathbf{R}^{d \times r}} \frac{1}{2} \|UU^\top - M^*\|_F^2. \tag{3}$$

This can be seen as a nonconvex composite problem where $h(M) = \frac{1}{2} \|M - M^*\|_F^2$. A straightforward computation reveals that even though the condition number of the convex function h is one, the condition number of the composition $f = h \circ F$ near minimizers scales like $\sigma_1(M^*)/\sigma_r(M^*)$. This leads to two potential issues for the convergence of gradient descent. On the one hand, in the exactly parameterized regime, i.e., $r = r^*$, the condition number of f is proportional to the condition number $\kappa(M^*)$, which could lead to arbitrarily slow linear convergence depending on the matrix M^* .¹ On the other, in the overparameterized regime, i.e., $r > r^*$, $\sigma_r(M^*)$ is zero, leading to an infinite condition number, which in turn results in sublinear convergence [109]. Both of these situations happen in practice since it is common to encounter ill-conditioned matrices with unknown rank.

These issues go beyond this simple, smooth problem. Indeed, for nonsmooth functions that are sharp and Lipschitz, the Polyak subgradient method converges at a linear rate, yet the rate might drastically slow down depending on the parameterization [23, 26], and may even decay exponentially

¹The condition number of a matrix A is given by $\kappa(A) := \sigma_{\max}(A)/\sigma_{\min}(A)$ where $\sigma_{\min}(A)$ is the smallest nonzero singular value.

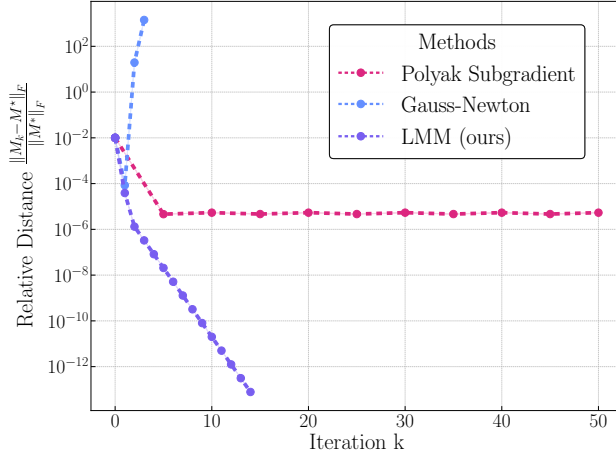


Figure 1: Relative distance to the solution against iteration count for Algorithm 4 applied to an overparameterized nonsmooth matrix factorization problem with $F(U) = UU^\top$, $h(M) = \|M - M^*\|_F$, $M^* \in \mathcal{S}_+^{50}$, and $U \in \mathbf{R}^{50 \times 3}$ with $\text{rank}(M^*) = 2 < r = 3$ and $\kappa(M^*) = 1$. All algorithms use the Polyak stepsize.

when overparameterization occurs; see Figure 1. Motivated by these drawbacks, numerous works have proposed ad-hoc preconditioned (sub)gradient methods that exhibit linear convergence at a rate independent of the parameterization F [26, 94, 104, 106]. Despite the breadth of this line of work, much of it focuses on concrete formulations, e.g., smooth low-rank matrix recovery, and the proposed methods do not systematically generalize to composite optimization problems, which motivates the main question of this work.

Is there a preconditioned subgradient method for general composite optimization problems that exhibits local linear convergence depending only on the convex outer function h ?

We answer this question in the affirmative under mild assumptions. Borrowing ideas that date back to the work of Levenberg [61], Morrison [78], and Marquardt [74] on nonlinear least squares, we propose a preconditioned subgradient method that updates

$$x_{k+1} \leftarrow x_k - \gamma_k (\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I)^{-1} \nabla F(x_k)^\top v_k, \quad (4)$$

with $v_k \in \partial h(F(x_k))$ where ∂h denotes the convex subdifferential of h . Let us comment on this algorithm and its underlying motivation. The method applies to both smooth and nonsmooth composite problems. The term $\nabla F(x_k)^\top v_k$ corresponds to a subgradient of f , thus $(\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I)^{-1}$ acts as a preconditioner. For structured problems, the cost of solving the linear system involved at each iteration is low. For instance, for low-rank matrix recovery problems, the cost is proportional to that of solving an $r \times r$ linear system. When the convex function h corresponds to the ℓ_2 norm squared, update (4) recovers the classical Levenberg-Morrison-Marquadt (LMM) method;² moreover, when $\lambda_k = 0$, it reduces to the Gauss-Newton method. Recently, Davis and Jiang [26] introduced a Gauss-Newton subgradient method (GNP) for general composite problems, which was the main inspiration for this work. Davis and Jiang showed that if ∇F has full rank near a minimizer, then GNP converges at a linear rate that only depends on the conditioning of h . However, for overparameterized problems, ∇F does not have full rank near minimizers, leading to

²LMM is often only attributed to Levenberg and Marquardt.

potentially ill-posed preconditioners. Indeed, even mild overparameterization in low-rank matrix factorization leads to the divergence of GNP; see Figure 1. To overcome this issue, we regularize the preconditioner, which improves numerical stability.

Main contributions. Let us summarize our three core contributions.

(Method) We propose a *Levenberg–Morrison–Marquardt* subgradient method (Algorithm 1) along with a concrete choice of stepsizes, γ_k , inspired by the Polyak stepsize [84], and damping coefficients, λ_k , that displays rapid local convergence universally across all combinations of smooth and nonsmooth, overparameterized and exactly parameterized settings. Since our parameter choice heavily relies on information about the function that may not be readily available to practitioners, we also present another parameter configuration based on geometrically decaying schedules [27, 45], which only requires rough bounds on the function parameters.

(General-purpose convergence guarantees) Under mild assumptions, we show that our parameter configurations guarantee linear convergence at a rate depending solely on the convex function h . Our results rely on nearly decoupled assumptions for h and the smooth map F , allowing one to combine these functions freely while still achieving rapid convergence. In particular, we require that h is in some sense well-conditioned on the image of F —quadratic growth with Lipschitz gradient in smooth settings, or sharp and Lipschitz in nonsmooth settings—while F must satisfy that its image and Jacobian are in a certain sense aligned near minimizers.

(Consequences for statistical recovery problems) To complement our convergence guarantees, we study their implications for various data science tasks. In particular, we show that the geometric assumptions required for our general-purpose convergence results hold for (i) nonnegative least squares formulations, (ii) (overparameterized) low-rank matrix recovery, and (iii) canonical polyadic (CP) tensor factorization problems. As a result, we establish the linear convergence of the Levenberg–Morrison–Marquardt subgradient method (4) for all these problems at a rate that only depends on the conditioning of the convex function h . This recovers existing convergence guarantees in certain settings and provides the first such results for others.

Outline of the paper. We conclude this section with related work. Section 2 sets out notation and necessary background. In Section 3, we formally introduce composite problems, our key assumptions, and the algorithm we propose. After that, Section 4 provides general-purpose convergence guarantees under suitable regularity conditions. In Section 5, we verify these conditions for several statistical recovery problems. Section 6 contains numerical experiments showcasing the benefits of our method. We defer long and technical proofs to the appendix.

1.1 Related work

Nonlinear least squares. Nonlinear least squares problems [9] form a widely studied instance of (1), where the outer function is the squared Euclidean norm. Although Newton’s method enjoys local quadratic convergence under mild regularity, forming or even applying the Hessian is often prohibitively expensive at scale. The Gauss-Newton algorithm is a computationally cheaper alternative that enjoys similar guarantees [81, 83] and has been widely used in the sciences and engineering [4, 19, 86]. Gauss–Newton can fail when the iteration’s linear system is singular, making the update ill-defined. To overcome this issue, Levenberg [61], Morrison [78], and Marquardt [74] independently introduced an additional damping term ensuring invertability. The LMM method

Algorithm	Low-rank matrix recovery					Converges to	Applicable beyond matrix recovery
	Overparam.	Symm.	Asymm.	Smooth	Nonsmooth		
ScaledGD [94]	✗	✓	✓	✓	✗	Solution	No*
ScaledGD(λ) [104]	✓	✓	✓	✓	✗	Neighborhood [‡]	No*
ScaledSM [95]	✗	✓	✓	✗	✓	Solution	No*
PreconditionedGD [106]	✓	✓	✗	✓	✗	Solution	No
Asymmetric PreconditionedGD [22]	✓	✗	✓	✓	✗	Solution	No
OPSA [44]	✓	✗	✓	✗	✓	Solution [‡]	No
APGD [69]	✓	✗	✓	✓	✗	Solution	No
Approximated GN [48]	✓	✓	✓	✓	✗	Solution	No
GNP [26]	✗	✓	✓	✓	✓	Solution	Yes
Algorithm 1 (ours)	✓	✓	✓	✓	✓	Solution	Yes

Table 1: Comparison of methods for low-rank matrix recovery. These are problems where $F(U) = UU^\top$ (symmetric) or $F(U, V) = UV^\top$ (asymmetric). A check mark ✓ indicates that the method exhibits local linear convergence (depending only on h) for that particular setting.

* The same authors modified ScaledGD to extend to tensor problems [97].

‡ Converges arbitrarily close to a solution, with the final distance controlled by a parameter.

‡ The method converges to the solution of a regularized problem, which might differ from the original one.

has been extensively analyzed for nonlinear least squares [40, 41, 47, 77] and is widely used in applications [3, 46, 85, 87].

Composite optimization. Splitting methods are a popular alternative for composite objectives. The term ‘composite optimization’ is often used to refer to the subclass of additive composite problems where the loss can be expressed as a sum of a convex and a smooth function. Classical schemes for this subclass include the forward-backward (proximal-gradient) splitting [18, 37, 68] and optimal accelerated versions [5, 55, 56]. For general composite objectives, the prox-linear method linearizes F and computes a proximal step of composition of the linearization with the convex outer function each iteration [16, 37, 38, 63, 81]. This scheme is closely related to classical trust region variants of Gauss-Newton [13, 14, 16, 42, 101]. Only recently has the local and global convergence of subgradient methods for composite problems been established [27, 28].

(Sub)gradient methods for matrix recovery. Low-rank matrix recovery via the factorization approach has been the subject of intensive study over the past decade [65, 73, 98, 105, 107]. For the exactly parameterized regime, it is known that the optimization landscape of smooth objectives is benign and randomly initialized gradient descent finds global minimizers [8, 20, 23, 43, 108]. Yet, all local convergence rates depend on the condition number of the ground truth [17, 71]. Recent work has focused on the rank-overparameterized setting, where subgradient methods still find global minimizers, yet they exhibit a sublinear local rate of convergence due to flattened local geometry caused by overparameterization [34, 109]. Several works have proposed strategies to accelerate the convergence of these methods based on small initialization with early stopping and alternating small and long steps [29, 35, 49, 50, 66, 72, 91, 92, 99, 103, 104]. Even though these methods achieve linear convergence, their rates still depend on the condition number of the solution matrix.

Overparameterized matrix recovery problems. The seminal works [94, 95] proposed preconditioned (sub)gradient methods to overcome the dependence on the conditioning of the ground truth matrix in the exactly parameterized setting. Soon after, [106] introduced a preconditioned gradient method that additionally handles overparameterization for smooth objectives with PSD matrices. Subsequently, many other works introduced methods based on preconditioning together with small

overparameterization [104], alternating minimization [69], and Gauss-Newton type methods [48, 58]. Closer to our work, [44] introduced the Overparameterized Preconditioned Subgradient Algorithm (OPSA), which achieves local convergence rates for nonsmooth, overparameterized, asymmetric objectives. However, OPSA converges to the solution to a regularized problem, which might differ from the ground truth. Beyond matrix problems, recent literature has also studied preconditioned methods for low Tucker-rank tensor recovery [70, 96, 97, 102]. Given that much of the existing literature focuses on matrix recovery problems, we include a comparison in Table 1.³

2 Preliminaries

Linear algebra. The symbol $[m]$ will be shorthand for the set $\{1, \dots, m\}$. Further, for a finite set S , the symbol $\#S$ denote its cardinality. We will use the symbol \mathbf{E} to denote a finite-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. The closed ball of radius $r > 0$ around $x \in \mathbf{E}$ will be denoted by $\mathbf{B}_r(x)$. For any point $x \in \mathbf{E}$ and a set $Q \subset \mathbf{E}$, the distance and the nearest-point projection (with respect to the Euclidean norm) are defined by

$$\text{dist}(x; Q) = \inf_{y \in Q} \|x - y\| \quad \text{and} \quad \text{proj}_Q(x) = \underset{y \in Q}{\text{argmin}} \|x - y\|,$$

respectively. Given a linear map between Euclidean spaces, $\mathcal{A}: \mathbf{E} \rightarrow \mathbf{Y}$, its adjoint map will be written as $\mathcal{A}^*: \mathbf{Y} \rightarrow \mathbf{E}$. We will use I_d for the d -dimensional identity matrix, while $\mathbf{0}_d$ denotes the d -dimensional origin. We use \mathbf{R} and \mathbf{R}_+ to denote the reals and the positive reals, respectively. We always endow the Euclidean space of vectors \mathbf{R}^d with the usual dot-product $\langle x, y \rangle = x^\top y$ and the induced ℓ_2 -norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. For any $x \in \mathbf{R}^d$, we use $\text{supp}(x)$ to denote the indices of nonzero entries of x . Given two vectors $x, y \in \mathbf{R}^d$, we let $x \odot y \in \mathbf{R}^d$ denote their Hadamard or component-wise product.

Similarly, we will equip the space of rectangular matrices $\mathbf{R}^{d_1 \times d_2}$ with the trace product $\langle X, Y \rangle = \text{tr}(X^\top Y)$ and the induced Frobenius norm $\|X\|_F = \sqrt{\text{tr}(X^\top X)}$. The operator norm of a matrix $X \in \mathbf{R}^{d_1 \times d_2}$ will be written as $\|X\|_{\text{op}}$. The symbol $\sigma(X)$ will denote the vector of singular values of a matrix X in nonincreasing order. We use $\sigma_{\max}(X)$ and $\sigma_{\min}(X)$ to denote the largest and smallest nonzero singular values. Similarly, for a given symmetric matrix The symbols \mathcal{S}^d and \mathcal{S}_+^d denote the sets of $d \times d$ symmetric and positive semidefinite, respectively. We use $O(d, r)$ to denote the set of matrices with orthogonal columns, i.e., $Q \in \mathbf{R}^{d \times r}$ such that $Q^\top Q = I$. Given a matrix $X \in \mathcal{S}^d$, the symbol $\lambda(X)$ denotes the vector of eigenvalues of X in nonincreasing order. Given two matrices A, B of potentially different sizes, we let $A \otimes_{\text{Kr}} B$ denote their Kronecker product. On the other hand, \otimes denotes the tensor product. Note that the inputs and outputs of Kronecker products are always matrices, while the inputs and outputs of the tensor product might be higher-order tensors. Given a matrix $X \in \mathbf{R}^{d \times r}$, we use X_i and X_j to denote its i th row and column, respectively, and X_S with $S \subseteq [r]$ to denote the submatrix of X with columns indexed by S .

Nonsmooth analysis. Nonsmooth functions are central to this work. Therefore, we will utilize a few basic constructions of generalized differentiation; we refer the interested reader to the monographs [10, 24, 76, 88]. Consider a function $f: \mathbf{E} \rightarrow \mathbf{R} \cup \{+\infty\}$ and a point x , with $f(x)$ finite. We use the convention $\frac{1}{0} = +\infty$. The *subdifferential* of f at x , denoted by $\partial f(x)$, is the set of all vectors $\xi \in \mathbf{E}$ satisfying

$$f(y) \geq f(x) + \langle \xi, y - x \rangle + o(\|y - x\|) \quad \text{as } y \rightarrow x, \quad (5)$$

³This table is an oversimplification for pedagogical purposes; each statement holds under potential additional assumptions of the respective paper.

Algorithm 1: Levenberg-Morrison-Marquardt Subgradient Method (LMM)**Input:** Initial $x_0 \in \mathbf{E}$, stepsizes $(\gamma_k)_{k \geq 0} \subset (0, \infty)$, and damping coefficients $(\lambda_k)_{k \geq 0} \subset (0, \infty)$ **Step** $k \geq 0$:Pick $v_k \in \partial h(F(x_k))$.Set $x_{k+1} \leftarrow x_k - \gamma_k \left(\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I \right)^{-1} \nabla F(x_k)^\top v_k$.

where $o(r)$ denotes any function satisfying $o(r)/r \rightarrow 0$ as $r \rightarrow 0$. Standard results show that for a convex function f the subdifferential $\partial f(x)$ reduces to the subdifferential in the sense of convex analysis, while for a differentiable function, it consists only of the gradient: $\partial f(x) = \{\nabla f(x)\}$. For any closed convex functions $h: \mathbf{Y} \rightarrow \mathbf{R}$ and C^1 -smooth map $F: \mathbf{E} \rightarrow \mathbf{Y}$, the chain rule holds [88, Theorem 10.6]:

$$\partial(h \circ F)(x) = \nabla F(x)^* \partial h(F(x)).$$

3 Algorithm and assumptions

In this section, we formally introduce the problem class we study, the different assumptions we make, and the Levenberg-Morrison-Marquardt subgradient method we use for the rest of the paper. As mentioned in the introduction, we consider

$$\min_{x \in \mathbf{E}} f(x) \quad \text{with} \quad f := h \circ F. \tag{6}$$

Here \mathbf{E} and \mathbf{Y} are finite-dimensional Euclidean spaces, $h: \mathbf{Y} \rightarrow \mathbf{R}$ is a convex function, and $F: \mathbf{E} \rightarrow \mathbf{Y}$ is a continuously differentiable map. Instantiations of this template include: low-rank symmetric matrix problems where $\mathbf{E} = \mathbf{R}^{d \times r}$, $\mathbf{Y} = \mathcal{S}_+^d$, and $F(U) = UU^\top$, and asymmetric matrix problems where $\mathbf{E} = \mathbf{R}^{d_0 \times r} \times \mathbf{R}^{d_2 \times r}$, $\mathbf{Y} = \mathbf{R}^{d_1 \times d_2}$, and $F(U, V) = UV^\top$. From now on, the symbol \mathcal{X}^* denotes the set of minimizers of (6).

We say that problem (6) is *overparameterized* if for some $x^* \in \mathcal{X}^*$, there is a sequence $(x_j)_j \subseteq \mathbf{E}$ converging to x^* such that the rank of $\nabla F(x_j)$ exceeds the rank of $\nabla F(x^*)$. Intuitively, this means that there are points arbitrarily close to a minimizer x^* for which a linear approximation of F requires more parameters than at x^* itself. This definition matches the natural notion of overparameterization for low-rank problems; we defer additional details to Section 5.

3.1 Algorithmic description

The Levenberg-Morrison-Marquardt subgradient method is summarized in Algorithm 1. For structured problems, the linear system at each iteration can often be solved efficiently. For instance, for low-rank matrix recovery, it reduces to solving a much smaller linear system; we defer the details to Appendix E. This method builds upon the GNP method introduced in [26], which sets $\lambda_k = 0$. The inspiration for GNP stems from a simple observation: when ∇F has constant rank near x^* —i.e., under exact parameterization—the image of F forms a manifold \mathcal{M} around $F(x^*)$. An elegant argument shows that in this regime, the mapped iterates of GNP $z_k = F(x_k)$ are akin to the iterates of a Riemannian subgradient method on \mathcal{M} with objective h [26]. Consequently, the iterates $F(x_k)$ are unaffected by the ill-conditioning of F . In contrast, overparameterization leads to problems where the image of F fails to form a manifold. In such cases, the Gauss-Newton preconditioner is

not even well-defined since $\nabla F(x_k)$ might not have constant rank. Our method bypasses this issue by adding a damping term to the preconditioner, thereby ensuring its invertibility and stability.

We propose two ways to set the hyperparameters γ_k and λ_k of Algorithm 1, which we dub “configurations.” The first configuration is based on the Polyak stepsize [84] and the damping parameter for the preconditioned gradient descent method in [106]. From now on, we will use h^\star and \mathcal{Z}^\star to denote the minimum value and set of minimizers of $\min_{z \in \text{Im } F} h(z)$. Further, we use Π^x as a shorthand for the projection matrix $\text{proj}_{\text{Im } \nabla F(x)}$.

Configuration 1 (Polyak). *Set the stepsize to $\gamma_k = \gamma \frac{h(z_k) - h^\star}{\|\Pi^{x_k} v_k\|^2}$ where $\gamma > 0$ is a tuning parameter. Additionally, set λ_k such that there exist constants $0 < C_{lb} \leq C_{ub}$ satisfying*

$$C_{lb} \text{dist}(z_k, \mathcal{Z}^\star) \leq \lambda_k \leq C_{ub} \text{dist}(z_k, \mathcal{Z}^\star).$$

This parameter choice relies on detailed information about the loss function and the distance to the solution set—quantities that may not be readily available in practice. To address this limitation, we introduce two alternative configurations tailored for nonsmooth and smooth problems, which do not depend as heavily on such information.

Configuration 2 (Nonsmooth). *Set $\gamma_k = \gamma q^k$ and $\lambda_k = \lambda q^k$, with $\gamma, \lambda > 0$ and $q \in (0, 1)$.*

Configuration 3 (Smooth). *Set $\gamma_k = \gamma$ and $\lambda_k = \lambda q^k$, with $\gamma, \lambda > 0$ and $q \in (0, 1)$.*

The primary difference between these two configurations is that the stepsize decreases geometrically for nonsmooth functions while it remains constant for smooth functions. These strategies are designed to emulate the behavior of the Polyak stepsize [27, 45]. In the following section, we demonstrate that these configurations achieve convergence rates comparable to those of the Polyak-based approach.

3.2 Regularity of the parameterization

Overparameterization and the addition of the damping term complicate the analysis and require more nuanced regularity conditions on the parameterization F . Conveniently, these conditions are independent of the conditioning assumptions for the outer function h , allowing us to pair any sufficiently regular parameterization F with any well-conditioned h , whether smooth or not. The remainder of this section introduces these assumptions and collects some algorithmic consequences. We begin with a standard assumption on F .

Assumption 1 (Smooth parameterization). *The map F is continuously differentiable, and there is a constant $L_{\nabla F} \geq 0$ such that*

$$\|\nabla F(x) - \nabla F(y)\|_{\text{op}} \leq L_{\nabla F} \|x - y\|, \quad \text{for all } x, y \in \mathbf{E}.$$

Henceforth, we use the following notation

$$P(x, \lambda) := \nabla F(x)(\nabla F(x)^\top \nabla F(x) + \lambda I)^{-1} \nabla F(x)^\top. \quad (7)$$

The matrix $P(x, \lambda)$ will play a crucial role in our analysis. It can be seen as a regularized projection matrix; indeed, when $\lambda = 0$ and $\nabla F(x)$ has full column rank, it reduces to the orthogonal projection onto the range of $\nabla F(x)$, which corresponds to Π^x . In what follows, we use the placeholders $z_k := F(x_k)$ and

$$P_k := P(x_k, \lambda_k).$$

The next result collects a few properties of P_k ; its proof is deferred to Appendix A.1. We introduce a bit of simplifying notation. Given any $x \in \mathbf{E}$, let U^x and σ^x denote respectively the matrix of left singular vectors and the vector of singular values of $\nabla F(x)$, moreover, we let $U_{1:j}^x$ be the matrix with the top j singular vectors of $\nabla F(x)$.

Lemma 3.1. *Let x_k and x_{k+1} be iterates from Algorithm 1. Let $z_k = F(x_k)$ and $z_{k+1} = F(x_{k+1})$. The following three hold true.*

1. (**Nonexpansiveness**) *The operator norm of both P_k and $I - P_k$ are less than or equal to one. Moreover, $\|P_k v\| \leq \|\Pi^{x_k} v\|$ for any $v \in \mathbf{Y}$.*
2. (**Restricted eigenvalues**) *The the eigenvalues of $I - P_k$ restricted to the span generated by the top j left singular vectors of $\nabla F(x_k)$ are bounded by $\frac{\lambda_k}{(\sigma_j^{x_k})^2 + \lambda_k}$, i.e.,*

$$\|(I - P_k)v\| \leq \frac{\lambda_k}{(\sigma_j^{x_k})^2 + \lambda_k} \|v\| \quad \text{for } v \in \text{span} \left(U_{1:j}^{x_k} \right).$$

3. (**Approximation error**) *If in addition ∇F is $L_{\nabla F}$ -Lipschitz when restricted onto the line segment connecting x_k and x_{k+1} , then $\|z_{k+1} - (z_k - \gamma_k P_k v_k)\| \leq \frac{L_{\nabla F} \gamma_k^2}{8\lambda_k} \|\Pi^{x_k} v_k\|^2$.*

In particular, the last item follows for any pair of consecutive iterates whenever Assumption 1 holds. Intuitively, this item states that an updated mapped iterate $z_{k+1} = F(x_{k+1})$ can be approximated by a linear step in \mathbf{Y} space. This approximation will play a critical role in our analysis.

The following two assumptions are crucial to obtain linear convergence for ill-conditioned and overparameterized problems, respectively. These assumptions are stated near a point z^* , which the reader can deem as a minimizer of h over the image of F . We let $\Pi_j^x = U_{1:j}^x (U_{1:j}^x)^\top$ be the projection onto the subspace spanned by the top j left singular vectors of $\nabla F(x)$.

Assumption 2 (Strong alignment). *For a given $z^* \in \text{Im } F$, there exist a function $\delta: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and a constant $s > 0$ such that for any $\rho > 0$ and $z = F(x) \in \mathbf{B}_{\delta(\rho)}(z^*)$ there is an index j for which*

$$\|(I - \Pi_j^x)(z - z^*)\| \leq \rho \|z - z^*\| \quad \text{and} \quad (\sigma_j^x)^2 \geq s.$$

Intuitively, this assumption amounts to the alignment between the image of F and a low-dimensional linear approximation around z^* . As alluded to earlier, in the exact parameterization regime when the rank of $\nabla F(\cdot)$ is locally constant, the image of F forms a manifold \mathcal{M} of dimension $j = \text{rank}(\nabla F(x^*))$. As such, Π_j^x corresponds to the projection onto the tangent of \mathcal{M} at $F(x)$. In turn, the error between centered manifold elements $z - z^*$ and the tangent increases at most quadratically in the norm of $z - z^*$, and one can easily derive that this assumption holds; see Lemma C.1 in the appendix.

However, for overparameterized problems, there is no manifold structure, and Assumption 2 cannot hold. To overcome this issue, we introduce a weaker condition, allowing the singular values of the linear approximation to decrease gracefully as we approach z^* .

Assumption 3 (Weak alignment). *For a given $z^* \in \text{Im } F$, there exist functions $\delta: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and $s: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ such that for any $\rho > 0$ and any $z = F(x) \in \mathbf{B}_{\delta(\rho)}(z^*)$ there is an index j for which*

$$\|(I - \Pi_j^x)(z - z^*)\| \leq \rho \|z - z^*\| \quad \text{and} \quad (\sigma_j^x)^2 \geq s(\rho) \|z - z^*\|.$$

In Section 5, we will see that this assumption holds for a variety of overparameterized problems. It is immediately clear that Assumption 2 implies Assumption 3. We emphasize that both assumptions on F are independent of the outer function h .

3.3 Regularity for nonsmooth outer functions

Next, we introduce the regularity conditions on the outer function h . Intuitively, regularity ensures that the function h is well-conditioned when restricted to the image of F . We present two different

notions of conditions depending on the smoothness of the problem. We start with conditions for nonsmooth losses. Recall that we use Π^x as a shorthand for the projection matrix $\text{proj}_{\text{Im } \nabla F(x)}$.

Assumption 4. *The function $h: \mathbf{Y} \rightarrow \mathbf{R}$ satisfies the following properties.*

1. (**Unique minimizer**) *The function h has a unique minimizer z^* over $\text{Im } F$.*
2. (**Convexity**) *The function h is convex.*
3. (**Restricted sharpness**) *The function h is μ -sharp on $\text{Im } F$. That is,*

$$h(z) - h^* \geq \mu \cdot \|z - z^*\| \quad \text{for all } z \in \text{Im } F,$$

where $h^* = h(z^*)$ is the minimum of $h|_{\text{Im } F}$.

4. (**Restricted Lipschitzness**) *There exists a constant $L \geq 0$ such that*

(a) *For any $x \in \text{dom } F$ and $v \in \partial h(F(x))$,*

$$\|\Pi^x v\| \leq L.$$

(b) *For any $x \in \mathbf{E}$, $z = F(x)$, $v \in \partial h(F(x))$, and $\lambda > 0$ we have*

$$|\langle v, (I - P(x, \lambda))(z - z^*) \rangle| \leq L \|(I - P(x, \lambda))(z - z^*)\|,$$

where the matrix $P(x, \lambda)$ is given by (7).

We point out that for our results, one can drop the uniqueness of the minimizer in Assumption 4, but we assume it for simplicity. Although the fourth condition might seem complicated, it is satisfied by any globally Lipschitz convex function h . Moreover, a simple argument shows that it ensures that $h(z) - h^* \leq 2L\|z - z^*\|$ for all $z \in \text{Im } F$. These regularity conditions are well-understood in the unconstrained setting where the parameterization is an identity, $F = I$. The seminal work [84] showed that the subgradient method coupled with the Polyak stepsize converges linearly to minimizers in this setting. The following is a direct consequence of Assumption 4.

Lemma 3.2 (Aiming towards solution). *Suppose that Assumption 4 holds. Then,*

$$\langle v, z - z^* \rangle \geq h(z) - h(z^*) \geq \mu \|z - z^*\| \quad \text{for all } z \in \text{Im } F \text{ and } v \in \partial h(z).$$

Thus, negative subgradients point towards the solution.

3.4 Regularity for smooth outer functions

Paralleling regularity for nonsmooth functions h , we now introduce analog regularity conditions for the smooth setting. Intuitively, they amount to quadratic lower and upper bounds.

Assumption 5. *The function $h: \mathbf{Y} \rightarrow \mathbf{R}$ satisfies the following properties.*

1. (**Unique minimizer**) *The function h has a unique minimizer z^* over $\text{Im } F$.*
2. (**Convexity**) *The function h is convex.*
3. (**Restricted quadratic growth**) *The function h exhibits α -quadratic growth on $\text{Im } F$; i.e.,*

$$h(z) - h^* \geq \frac{\alpha}{2} \|z - z^*\|^2 \quad \text{for all } z \in \text{Im } F,$$

where $h^* = h(z^*)$ is the minimum of $h|_{\text{Im } F}$.

4. (**Restricted Smoothness**) There exists a constant $\beta \geq 0$ such that for any $x \in \mathbb{E}$ and $z = F(x)$ the following hold true.

(a) We have

$$\|\Pi^x \nabla h(z)\| \leq \beta \|z - z^*\| \quad \text{and} \quad h(z) - h^* \geq \frac{1}{2\beta} \|\Pi^x \nabla h(z)\|^2.$$

(b) For any $\lambda > 0$

$$|\langle \nabla h(z), (I - P(x, \lambda))(z - z^*) \rangle| \leq \beta \|z - z^*\| \|(I - P(x, \lambda))(z - z^*)\|,$$

where $P(x, \lambda)$ is given in (7).

The first two conditions are exactly the same as in the nonsmooth setting. The third condition is satisfied by any globally α -strongly convex function, while the latter one holds for any globally β -smooth function [82]; e.g., both are trivially satisfied by $h(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. We collect an analog to Lemma 3.2; the proof follows from convexity and quadratic growth.

Lemma 3.3 (Aiming towards solution). *Suppose that Assumption 5 holds. Then,*

$$\langle \nabla h(z), z - z^* \rangle \geq h(z) - h(z^*) \geq \frac{\alpha}{2} \|z - z^*\|^2 \quad \text{for all } z \in \mathbf{E}.$$

We close this section with a bound on the progress made by the approximation from Lemma 3.1. We highlight that the next result applies regardless of the smoothness of h .

Lemma 3.4 (Linearization progress). *Let x_k be an iterate from Algorithm 1 with Configuration 1 (Polyak stepsizes) where we set the hyperparameter $\gamma \leq 1$. Suppose that h is convex and F is continuously differentiable. Letting $z_k = F(x_k)$, we have*

$$\|z_k - \gamma_k P_k v_k - z^*\|^2 \leq \|z_k - z^*\|^2 - \gamma \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} + 2\gamma_k \langle (I - P_k)v_k, z_k - z^* \rangle.$$

Proof. Expanding

$$\begin{aligned} \|z_k - \gamma_k P_k v_k - z^*\|^2 &= \|z_k - z^*\|^2 - 2\gamma_k \langle P_k v_k, z_k - z^* \rangle + \gamma_k^2 \|P_k v_k\|^2 \\ &= \|z_k - z^*\|^2 - 2\gamma_k \langle v_k, z_k - z^* \rangle + 2\gamma_k \langle (I - P_k)v_k, z_k - z^* \rangle + \gamma_k^2 \|P_k v_k\|^2 \\ &\leq \|z_k - z^*\|^2 - \gamma \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} + 2\gamma_k \langle (I - P_k)v_k, z_k - z^* \rangle, \end{aligned} \quad (8)$$

where the last inequality eliminates the term $\langle v_k, z_k - z^* \rangle$ via convexity of h , and upper bounds the last using Lemma 3.1, Configuration 1, and $\gamma \leq 1$. This completes the proof. \square

4 General convergence guarantees

In this section, we present our general-purpose guarantees for Algorithm 1 under the various parameter choices and smoothness assumptions introduced in Section 3. We also extend these guarantees to cases where the regularity assumptions for the parameterization F hold only locally, an extension that will be particularly useful for tensor problems in later sections. Since most proofs share the same structure, we provide only the simplest versions to illustrate the core ideas and defer technical details to Appendix B.

4.1 Guarantees for nonsmooth losses

We provide guarantees for both the exactly parameterized and overparameterized regimes. We start with the latter due to its novelty.

Theorem 4.1 (Convergence under weak alignment and nonsmoothness). *Suppose that Assumptions 1, 3 and 4 hold. Further assume that $z_0 = F(x_0)$ satisfies $\|z_0 - z^*\| \leq \delta \left(\frac{\mu}{8L}\right)$. The following two hold.*

1. (**Polyak stepsize**) *Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{1, \frac{C_{1b}}{L_{\nabla F}}\right\}$ and $C_{ub} \leq \frac{\mu}{8L} s \left(\frac{\mu}{8L}\right)$. Then, the iterates $z_k = F(x_k)$ satisfy*

$$\|z_k - z^*\| \leq \left(1 - \frac{\gamma\mu^2}{8L^2}\right)^{k/2} \|z_0 - z^*\| \quad \text{for all } k \geq 0.$$

2. (**Geometrically decaying stepsize**) *Suppose we ran Algorithm 1 initialized at x_0 using Configuration 2 with*

$$\lambda \leq \frac{Ms \left(\frac{\mu}{8L}\right) \mu}{128 L}, \quad \gamma \leq \frac{1}{L} \cdot \min \left\{ \frac{M\mu}{64L}, \sqrt{\frac{2\lambda M}{L_{\nabla F}}}, \frac{\lambda\mu}{2L_{\nabla F}L} \right\} \quad \text{and} \quad q \geq \max \left\{ 1 - \frac{\gamma\mu}{4M}, \frac{1}{\sqrt{2}} \right\}$$

where $M = \delta \left(\frac{\mu}{8L}\right)$. Then, the iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\| \leq Mq^k \quad \text{for all } k \geq 0.$$

Here, we only prove the statement concerning the Polyak stepsize and defer the proof for the geometrically decaying stepsize to Appendix B.1. The proofs of all our results follow the same template. Before proving Theorem 4.1, we introduce a proposition that provides the shared machinery underlying our argument for the Polyak stepsize.

Proposition 4.2 (One-step progress). *Suppose that x_k and x_{k+1} are iterates of Algorithm 1 with Configuration 1. Assume in addition that h is convex, F is continuously differentiable, and ∇F is $L_{\nabla F}$ -Lipschitz when restricted onto the line segment connecting x_k and x_{k+1} . Define $z_k = F(x_k)$ and $z_{k+1} = F(x_{k+1})$. If the stepsize hyperparameter satisfies $\gamma \leq \min \left\{1, \frac{C_{1b}}{L_{\nabla F}}\right\}$ and*

$$|\langle (I - P_k)v_k, z_k - z^* \rangle| \leq \frac{1}{4}(h(z_k) - h^*), \quad (9)$$

then, we have

$$\|z_{k+1} - z^*\|^2 \leq \|z_k - z^*\|^2 - \frac{\gamma (h(z_k) - h^*)^2}{8 \|\Pi^{x_k} v_k\|^2}.$$

Proof of Proposition 4.2. To derive this bound, we apply the triangle inequality with a one-step linear approximation

$$\|z_{k+1} - z^*\| \leq \underbrace{\|z_{k+1} - (z_k - \gamma_k P_k v_k)\|}_{T_1} + \underbrace{\|(z_k - \gamma_k P_k v_k) - z^*\|}_{T_2}. \quad (10)$$

We focus on bounding each of these terms separately. To bound T_1 we apply Lemma 3.1 and obtain

$$\begin{aligned} T_1 &\leq \frac{\gamma_k^2 L_{\nabla F}}{8\lambda_k} \|\Pi^{x_k} v_k\|^2 \\ &\leq \frac{\gamma_k^2 L_{\nabla F}}{8C_{1b} \|z_k - z^*\|} \|\Pi^{x_k} v_k\|^2. \end{aligned} \quad (11)$$

To bound T_2 , we compute

$$\begin{aligned}
T_2^2 &\leq \|z_k - z^*\|^2 - \gamma \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} + 2\gamma_k |\langle (I - P_k)v_k, z_k - z^* \rangle| \\
&\leq \|z_k - z^*\|^2 - \gamma \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} + \frac{\gamma_k}{2} (h(z_k) - h^*) \\
&= \|z_k - z^*\|^2 - \frac{\gamma}{2} \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2}, \tag{12}
\end{aligned}$$

where the first inequality follows from Lemma 3.4 and the second inequality follows from the bound (9). Next, we state a claim that we will use recurrently.

Claim 4.3. *If we have that $|\langle (I - P_k)v_k, z_k - z^* \rangle| \leq \frac{1}{4}(h(z_k) - h^*)$, then*

$$\frac{3}{4}(h(z_k) - h(z^*)) \leq \langle P_k v_k, z_k - z^* \rangle \leq \|\Pi^{x_k} v_k\| \|z_k - z^*\|.$$

Proof of the Claim 4.3. By the subgradient inequality

$$h(z_k) - h^* \leq \langle P_k v_k, z - z^* \rangle + \langle (I - P_k)v_k, z_k - z^* \rangle \leq \langle P_k v_k, z_k - z^* \rangle + \frac{1}{4}(h(z_k) - h^*),$$

rearranging the terms establishes the first inequality. The second inequality follows from Cauchy-Schwarz and Lemma 3.1. \square

In particular, we trivially obtain $T_2 \leq \|z_k - z^*\|$. Invoking (10) gives

$$\begin{aligned}
&\|z_{k+1} - z^*\|^2 \tag{13} \\
&\leq T_1^2 + T_2^2 + 2T_1 T_2 \\
&\leq \|z_k - z^*\|^2 - \frac{\gamma}{2} \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} + \frac{\gamma^4 L_{\nabla F}^2}{64C_{1b}^2 \|z_k - z^*\|^2} \frac{(h(z_k) - h^*)^4}{\|\Pi^{x_k} v_k\|^4} + \frac{\gamma^2 L_{\nabla F}}{4C_{1b}} \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} \\
&\leq \|z_k - z^*\|^2 - \frac{\gamma}{4} \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2} + \frac{\gamma^4 L_{\nabla F}^2}{64C_{1b}^2 \|z_k - z^*\|^2} \frac{(h(z_k) - h^*)^4}{\|\Pi^{x_k} v_k\|^4} \\
&\leq \|z_k - z^*\|^2 - \frac{\gamma}{8} \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k} v_k\|^2}, \tag{14}
\end{aligned}$$

where the second inequality uses (11), (12), and the definition of γ_k , while the third and the final inequalities follow from $\gamma \leq \min\left\{1, \frac{C_{1b}}{L_{\nabla F}}\right\}$ together with Claim 4.3. The proof of Proposition 4.2 is complete. \square

Armed with this proposition, we prove the main result of this section.

Proof of Theorem 4.1. By induction, it suffices to show that for z_k satisfying $\|z_k - z^*\| \leq \delta\left(\frac{\mu}{8L}\right)$,

$$\|z_{k+1} - z^*\|^2 \leq \left(1 - \frac{\gamma\mu^2}{8L^2}\right) \|z_k - z^*\|^2. \tag{15}$$

Let j be the index provided by Assumption 3 when applied to $\rho = \frac{\mu}{8L}$ and $z = z_k = F(x_k)$, i.e.,

$$\|(I - \Pi_j^{x_k})(z_k - z^*)\| \leq \frac{\mu}{8L} \|z_k - z^*\| \quad \text{and} \quad (\sigma_i^x)^2 \geq s \left(\frac{\mu}{8L}\right) \|z_k - z^*\|. \tag{16}$$

Invoking Item 4 of Assumption 4 and the triangle inequality, we derive

$$\begin{aligned} |\langle (I - P_k)v_k, z_k - z^* \rangle| &\leq L \|(I - P_k)(z_k - z^*)\| \\ &\leq L \left(\|(I - P_k)\Pi_j^{x_k}(z_k - z^*)\| + \|(I - P_k)(I - \Pi_j^{x_k})(z_k - z^*)\| \right). \end{aligned} \quad (17)$$

Lemma 3.1 ensures that the eigenvalues of $I - P_k$ restricted to the span generated by the top j left singular vectors of $\nabla F(x_k)$ are bounded by $\frac{\lambda_k}{(\sigma_j^x)^2 + \lambda_k}$. Using this fact in tandem with (16) gives

$$\begin{aligned} |\langle (I - P_k)v_k, z_k - z^* \rangle| &\leq L \left(\frac{\lambda_k}{(\sigma_j^x)^2 + \lambda_k} \|\Pi_j^{x_k}(z_k - z^*)\| + \frac{\mu}{8L} \|z_k - z^*\| \right) \\ &\leq L \left(\frac{C_{\text{ub}} \|z_k - z^*\|}{s \left(\frac{\mu}{8L}\right) \|z_k - z^*\|_2 + C_{\text{ub}} \|z_k - z^*\|} + \frac{\mu}{8L} \right) \|z_k - z^*\| \\ &= L \left(\frac{C_{\text{ub}}}{s \left(\frac{\mu}{8L}\right) + C_{\text{ub}}} + \frac{\mu}{8L} \right) \|z_k - z^*\| \\ &\leq \frac{\mu}{4} \|z_k - z^*\| \\ &\leq \frac{1}{4} (h(z_k) - h^*). \end{aligned} \quad (18)$$

The second and third inequalities use the fact that the function $b \mapsto \frac{b}{a+b}$ is strictly increasing on \mathbf{R}_+ for any given $a > 0$ together with the bounds $\lambda_k / \|z_k - z^*\|_2 \leq C_{\text{ub}} \leq s \left(\frac{\mu}{8L}\right) \frac{\mu}{8L}$ given by assumption. Invoking Proposition 4.2 and Assumption 4 gives

$$\|z_{k+1} - z^*\|^2 \leq \|z_k - z^*\|^2 - \frac{\gamma (h(z_k) - h^*)^2}{8 \|\Pi^{x_k} v_k\|^2} \leq \left(1 - \frac{\gamma \mu^2}{8L^2}\right) \|z_k - z^*\|^2,$$

completing the proof of Theorem 4.1. \square

Let us make a few remarks about the statement for the Polyak stepsize. While our conclusions also apply to geometrically decaying stepsizes, they are more transparent in the Polyak case. This result ensures that the iterates z_k will be within distance $\varepsilon > 0$ of the minimizer after $O\left(\frac{1}{\gamma} \frac{L^2}{\mu^2} \cdot \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations, which notably depends only on the conditioning of the outer function h . The parameter constraints enforce $\gamma \leq \frac{C_{\text{ub}}}{L \nabla F}$ and $C_{\text{ub}} \leq \frac{\mu}{L} s \left(\frac{\mu}{L}\right)$. In the context of low-rank matrix recovery problems, this simplifies to $\gamma \lesssim \frac{\mu^2}{L^2}$ and so the best rate one can get is $O\left(\frac{L^4}{\mu^4} \cdot \log(\varepsilon^{-1})\right)$.

The bound on C_{ub} is likely an artifact of our proof. Indeed, in our numerical experiments, we take γ and C_{ub} to be constants independent of μ/L and still observe linear convergence; see Section 6. Further, as we show in the next result, under strong alignment (Assumption 2)—which only holds for exactly parameterized problems—we can bypass the spurious bound on C_{ub} and prove a faster local rate.

Theorem 4.4 (Convergence under strong alignment and nonsmoothness). *Suppose that Assumptions 1, 2, and 4 hold. Further assume that $z_0 = F(x_0)$ satisfies $\|z_0 - z^*\| \leq \delta \left(\frac{\mu}{8L}\right)$. The following two hold.*

1. (**Polyak stepsize**) *Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min\left\{1, \frac{C_{\text{lb}}}{L \nabla F}\right\}$. Further, assume $\|z_0 - z^*\| \leq \frac{s\mu}{8C_{\text{ub}}L}$. Then, the iterates $z_k = F(x_k)$ satisfy*

$$\|z_k - z^*\|^2 \leq \left(1 - \frac{\gamma \mu^2}{8L^2}\right)^k \|z_0 - z^*\|^2 \quad \text{for all } k \geq 0.$$

2. (**Geometrically decaying stepsize**) Suppose we ran Algorithm 1 initialized at x_0 using Configuration 2 with

$$\lambda \leq \frac{s}{32} \frac{\mu}{L}, \quad \gamma \leq \frac{1}{L} \cdot \min \left\{ \frac{M\mu}{64L}, \sqrt{\frac{2\lambda M}{L_{\nabla F}}}, \frac{\lambda\mu}{2L_{\nabla F}L} \right\} \quad \text{and} \quad q \geq \max \left\{ 1 - \frac{\gamma\mu}{4M}, \frac{1}{\sqrt{2}} \right\},$$

where $M = \delta \left(\frac{\mu}{8L} \right)$. Then, the iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\| \leq Mq^k \quad \text{for all } k \geq 0.$$

We defer the proof of this result to Appendix B.2. There are two main differences between this result and Theorem 4.1 regarding the Polyak stepsize: (i) we replace Assumption 3 with Assumption 2 and (ii) we substitute the bound on C_{ub} with an additional constraint on the initial distance to optimum. By setting λ_k to ensure $\frac{C_{\text{lb}}}{L_{\nabla F}} = \Theta(1)$, we derive a local rate of $O\left(\frac{L^2}{\mu^2} \cdot \log(\epsilon^{-1})\right)$. In turn, this shows that a properly tuned Levenberg-Morrison-Marquardt method matches the rates of the Gauss-Newton method in the absence of overparameterization [26, Theorem 3.1].

4.2 Guarantees for smooth losses

Analogous arguments to the ones used for nonsmooth losses can be applied to derive linear convergence for composite losses where the outer function h is smooth and has quadratic growth. We state these guarantees here and defer the proof of the next result to Appendix B.3.

Theorem 4.5 (Convergence under weak alignment and smoothness). *Suppose that Assumptions 1, 3, and 5 hold. Further assume that $z_0 = F(x_0)$ satisfies $\|z_0 - z^*\| \leq \delta \left(\frac{\alpha}{16\beta} \right)$. The following two hold.*

1. (**Polyak stepsize**) Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{\text{lb}}}{L_{\nabla F}} \right\}$ and $C_{\text{ub}} \leq \frac{\alpha}{16\beta} s \left(\frac{\alpha}{16\beta} \right)$. Then, the iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\|_2^2 \leq \left(1 - \frac{\gamma\alpha}{32\beta} \right)^k \|z_0 - z^*\|_2^2 \quad \text{for all } k \geq 0.$$

2. (**Constant stepsize**) Suppose we ran Algorithm 1 initialized at x_0 using Configuration 3 with

$$\lambda \leq \frac{Ms \left(\frac{\alpha}{16\beta} \right)}{64} \frac{\alpha}{\beta}, \quad \gamma \leq \frac{1}{\beta} \cdot \min \left\{ \frac{1}{8}, \sqrt{\frac{32\lambda}{L_{\nabla F}M}}, \frac{\lambda}{2L_{\nabla F}M} \right\} \quad \text{and} \quad q \geq \max \left\{ \sqrt{1 - \frac{\gamma\alpha}{2}}, \frac{1}{\sqrt{2}} \right\},$$

where $M = \delta \left(\frac{\alpha}{16\beta} \right)$. Then, the iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\| \leq Mq^k \quad \text{for all } k \geq 0.$$

Again, the convergence rate depends solely on the conditioning of the outer function h . For matrix recovery problems, this results in a convergence rate of $O\left(\frac{\beta^3}{\alpha^3} \log(\epsilon^{-1})\right)$, which exhibits an undesirable cubic dependence on the condition number. As in the nonsmooth setting, we can further improve this convergence rate under conditions of strong alignment. The proof of the next result appears in Appendix B.4.

Theorem 4.6 (Convergence under strong alignment and smoothness). *Suppose that Assumptions 1, 2, and 5 hold. Further assume that $z_0 = F(x_0)$ satisfies $\|z_0 - z^*\| \leq \delta \left(\frac{\alpha}{16\beta} \right)$. The following two hold.*

1. (**Polyak stepsize**) Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{L_{\nabla F}} \right\}$. Additionally, suppose that $\|z_0 - z^*\|_2 \leq \frac{s\alpha}{16C_{ub}\beta}$. Then, the iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\|_2^2 \leq \left(1 - \frac{\gamma\alpha}{32\beta} \right)^k \|z_0 - z^*\|_2^2 \quad \text{for all } k \geq 0.$$

2. (**Constant stepsize**) Suppose we ran Algorithm 1 initialized at x_0 using Configuration 3 with

$$\lambda \leq \frac{s\alpha}{16\beta}, \quad \gamma \leq \frac{1}{\beta} \cdot \min \left\{ \frac{1}{8}, \sqrt{\frac{32\lambda}{L_{\nabla F}M}}, \frac{\lambda}{2L_{\nabla F}M} \right\} \quad \text{and} \quad q \geq \max \left\{ \sqrt{1 - \frac{\gamma\alpha}{2}}, \frac{1}{\sqrt{2}} \right\},$$

where $M = \delta \left(\frac{\alpha}{16\beta} \right)$. Then, the iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\| \leq Mq^k \quad \text{for all } k \geq 0.$$

4.3 Guarantees under local regularity

In some applications, such as tensor factorization, Assumptions 1 and 3 do not hold, i.e., there is no global Lipschitzness of ∇F or alignment. Instead, these two only hold locally. Notably, even under these weaker local conditions, all our previous rates still hold. In this section, we extend our guarantees to such a local regime. We start with local alternatives of Assumptions 1 and 3. Recall that $\mathcal{X}^* = \operatorname{argmin}_x h \circ F(x)$ is the set of minimizers.

Assumption 6 (Locally Lipschitz Jacobian). *The map F is continuously differentiable, and for a fixed $x^* \in \mathbf{R}^d$, there exists $\varepsilon_{\nabla F} > 0$ and $L_{\nabla F} \geq 0$ such that*

$$\|\nabla F(x) - \nabla F(y)\|_{\text{op}} \leq L_{\nabla F} \|x - y\|_2 \quad \text{for all } x, y \in B_{\varepsilon_{\nabla F}}(x^*).$$

Assumption 7 (Local weak alignment). *For a fixed $x^* \in \mathcal{X}^*$ and $z^* = F(x^*)$ there exist functions $\delta: \mathbf{R}_+ \rightarrow \mathbf{R}_+$, $s: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and a scalar $\varepsilon_{x^*} > 0$ such that for all $\rho > 0$ we have that if $x \in \mathbf{B}_{\varepsilon_{x^*}}(x^*)$, and $z = F(x) \in \mathbf{B}_{\delta(\rho)}(z^*)$ then there is an index j for which*

$$\|(I - \Pi_j^x)(z - z^*)\|_2 \leq \rho \|z - z^*\|_2 \quad \text{and} \quad (\sigma_j^x)^2 \geq s(\rho) \|z - z^*\|_2.$$

Next, we state a local guarantee under these local regularity assumptions. Although we only state it for weakly aligned and nonsmooth problems using the Polyak stepsize, there are similar guarantees for the other scenarios considered in this section. We defer those and the proof of the following result to Appendix B.5. The key idea to establish this result is to show that the iterates x_k stay in the region where the previous two assumptions hold, after which the argument follows precisely as it did for the global assumptions.

Theorem 4.7 (Convergence under local weak alignment and nonsmoothness). *Suppose Assumptions 4, 6 and 7 hold. Define $\tilde{q} := \sqrt{1 - \frac{\gamma\mu^2}{8L^2}}$, and let x_0 and $z_0 = F(x_0)$ be points satisfying*

$$\|x_0 - x^*\|_2 \leq \varepsilon/2 \quad \text{and} \quad \|z_0 - z^*\|_2 \leq \min \left\{ \delta \left(\frac{\mu}{8L} \right), \frac{(1 - \sqrt{\tilde{q}})^2 \varepsilon^2 C_{lb}}{2\gamma^2} \right\},$$

where $\varepsilon = \min \{ \varepsilon_{\nabla F}, \varepsilon_{x^*} \}$. Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{L_{\nabla F}} \right\}$ and $C_{ub} \leq \frac{\mu}{8L} s \left(\frac{\mu}{8L} \right)$. Then, the iterates x_k satisfy

$$\|x_k - x^*\|_2 < \varepsilon \quad \text{for all } k \geq 0,$$

and, moreover, the mapped iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\|^2 \leq \left(1 - \frac{\gamma\mu^2}{8L^2}\right)^k \|z_0 - z^*\|^2 \quad \text{for all } k \geq 0.$$

5 Consequences for statistical recovery problems

In this section, we instantiate the general convergence guarantees from Section 4 for concrete recovery problems in signal processing and data science. To this end, we show that our alignment assumptions hold for three families of parameterizations: squared-variable formulations, low-rank matrix factorizations, and CP tensor factorizations. Further, we establish that our restricted conditioning assumptions on the outer convex function are satisfied whenever well-established notions of strong identifiability, e.g., restricted isometry property, hold. Armed with these results, we derive local convergence rates for nonnegative least squares, robust matrix sensing, and tensor factorization under standard assumptions from the literature. All proofs are deferred to Appendix C.

5.1 Squared-variable formulations

Scientists dealing with unmixing problems often wish to minimize a convex function $h: \mathbf{R}^r \rightarrow \mathbf{R}$ over the positive orthant \mathbf{R}_+^r . A prominent example of this type of problem is nonnegative least squares [59]. These problems arise naturally across several domains, including acoustics, imaging, and genomics [64, 67, 93]. This type of problem can be reformulated as a composite optimization problem via the squared-variable map $c: x \mapsto x \odot x$ (where \odot denotes the component-wise product) [33, 62]. Although other algorithmic solutions might be preferable for this particular problem, e.g., the projected subgradient method, we cover this example as it provides a clear and simple illustration of our framework.

Regularity of the parameterization. Throughout we assume h has a unique minimizer z^* over \mathbf{R}_+^r and let $x^* \in \mathbf{R}^r$ be any vector such that $z^* = x^* \odot x^*$. It is immediate that $\nabla F(x) = 2 \text{diag}(x)$ and, the problem is ill-conditioned when $\max_{i \in [r]} |x_i^*| \gg \min_{i \in [r]} |x_i^*|$, and overparameterized whenever $r^* := \#\text{supp}(x^*) < r$. The next result establishes regularity for the squared-variable formulation with potential overparameterization; its proof appears in Appendix C.4.1.

Theorem 5.1 (Weak alignment of squared-variable map). *The map $F: \mathbf{R}^r \rightarrow \mathbf{R}^r$ given by $x \mapsto x \odot x$ satisfies Assumption 1 with $L_{\nabla F} = 2$ and Assumption 3 with*

$$s(\rho) = \frac{\rho}{\max\{\sqrt{r - r^*}, 1\}} \quad \text{and} \quad \delta(\rho) = \min \left\{ \min_{\substack{i, j \in [r] \\ z_i^* \neq z_j^*}} \frac{|z_i^* - z_j^*|}{2}, \min_{\substack{i \in [r] \\ z_i^* \neq 0}} \frac{z_i^*}{1 + s(\rho)}, \min_{\substack{i \in [r] \\ z_i^* \neq 0}} \frac{z_i^*}{2} \right\}$$

for any given $z^* \in \mathbf{R}_+^r$ with $r^* = \#\text{supp}(z^*)$.

Nonnegative least squares. We leverage the regularity of the squared-variable formulation to derive guarantees for nonnegative least squares [6, 30, 52, 79]. For a matrix $A \in \mathbf{R}^{m \times r}$ with $m \geq r$ and $b = Az^* \in \mathbf{R}^m$, define the smooth and nonsmooth formulations

$$\min_{x \in \mathbf{R}^r} \frac{1}{2} \|A(x \odot x) - b\|_2^2, \quad \text{and} \quad \min_{x \in \mathbf{R}^r} \|A(x \odot x) - b\|_2. \quad (19)$$

The following lemma is immediate, and so we omit its proof.

Lemma 5.2. *Suppose that A has full rank. Then, the function $z \mapsto \frac{1}{2}\|Az - b\|_2^2$ satisfies Assumption 5 with constants $\alpha = \sigma_{\min}(A)^2$ and $\beta = \sigma_{\max}(A)^2$. Similarly, the function $z \mapsto \|Az - b\|_2$ satisfies Assumption 4 with constants $\mu = \sigma_{\min}(A)$ and $L = \sigma_{\max}(A)$.*

Recall that $\kappa(A)$ denotes the condition number of A , i.e., $\kappa(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$. Equipped with these results, we are in good shape to derive a local convergence rate.

Corollary 5.3 (Smooth nonnegative least squares). *Suppose Algorithm 1 is applied to the first nonnegative least squares objective (19) (squared), initialized at some $x_0 \in \mathbf{R}^r$ using Configuration 1 with $\gamma \leq \min\left\{1, \frac{C_{lb}}{2}\right\}$, $C_{ub} \leq (2^8 \max\{\sqrt{r - r^*}, 1\})^{-1} \kappa^{-4}(A)$ and*

$$\|x_0 \odot x_0 - z^*\|_2 \leq \frac{1}{2} \min \left\{ \min_{\substack{i, j \in [r] \\ z_i^* \neq z_j^*}} \{z_i^* - z_j^*\}, \min_{i \in [r] | z_i^* \neq 0} z_i^* \right\}.$$

Then, the iterates x_k satisfy

$$\|x_k \odot x_k - z^*\|_2^2 \leq \left(1 - \frac{\gamma}{32} \kappa^{-2}(A)\right)^k \|x_0 \odot x_0 - z^*\|_2^2 \quad \text{for all } k \geq 0.$$

Corollary 5.4 (Nonsmooth nonnegative least squares). *Suppose Algorithm 1 is applied to the second nonnegative least squares objective in (19) (not squared), initialized at some $x_0 \in \mathbf{R}^r$ using Configuration 1 with $\gamma \leq \min\left\{1, \frac{C_{lb}}{2}\right\}$, $C_{ub} \leq \frac{1}{64 \max\{\sqrt{r - r^*}, 1\}} \kappa^{-2}(A)$ and*

$$\|x_0 \odot x_0 - z^*\|_2 \leq \frac{1}{2} \min \left\{ \min_{\substack{i, j \in [r] \\ z_i^* \neq z_j^*}} \{z_i^* - z_j^*\}, \min_{\substack{i \in [r] \\ z_i^* \neq 0}} z_i^* \right\}.$$

Then, the iterates x_k satisfy

$$\|x_k \odot x_k - z^*\|_2^2 \leq \left(1 - \frac{\gamma}{8} \kappa^{-2}(A)\right)^k \|x_0 \odot x_0 - z^*\|_2^2 \quad \text{for all } k \geq 0.$$

These corollaries follow directly from Theorem 4.1 and Theorem 4.5, respectively. While similar rates hold for geometrically decaying step sizes and better rates hold in the exactly parameterized case $r^* = r$, we omit them for brevity. The condition number of A appears in the convergence rate because we incorporated A into the definition of h ; this rate is standard for gradient descent applied to least squares. Interestingly, the convergence rate we derive for the nonsmooth formulation of the problem is faster than the convergence rate for its smooth counterpart. In particular, Corollary 5.4 only allows for $\gamma \asymp \kappa(A)^{-2}$ which translates to a rate of $O(\kappa(A)^4 \log(1/\varepsilon))$, while Corollary 5.3 imposes $\gamma \asymp \kappa(A)^{-4}$ translating to a rate of $O(\kappa(A)^6 \log(1/\varepsilon))$. We observe experimentally that our method is slightly faster when applied to the nonsmooth formulation; see Section 6.1.

5.2 Matrix recovery problems

Several modern data science tasks can be formulated as the problem of recovering a rank- r^* matrix Z^* from a small set of noisy measurements $b = \mathcal{A}(Z^*) + \varepsilon \in \mathbf{R}^m$ where \mathcal{A} is a known linear map and ε models noise. Applications arise in imaging, recommendation systems, control theory, and communications [17, 23, 25, 31, 100]. Remarkably, even though Z^* may be a large $d_1 \times d_2$ matrix, the number of measurements m required for recovery is often much lower, typically on the order of $O(r^*(d_1 + d_2))$. A popular approach to tackle this problem leverages low-rankness by solving one

of two formulations:

$$\min_{X \in \mathbf{R}^{d \times r}} \ell(\mathcal{A}(XX^\top) - b) \quad \text{or} \quad \min_{X \in \mathbf{R}^{d_1 \times r}, Y \in \mathbf{R}^{d_2 \times r}} \ell(\mathcal{A}(XY^\top) - b), \quad (20)$$

depending on whether the matrix is symmetric positive semidefinite $Z^* \in \mathcal{S}_+^d$ or asymmetric $Z^* \in \mathbf{R}^{d_1 \times d_2}$. Here r is an upper bound on the true rank r^* and $\ell(\cdot)$ is a measure of discrepancy. Common choices for $\ell(\cdot)$ include the ℓ_2 norm squared, which is effective against small unbiased noise [23], and the ℓ_1 norm, which is robust against gross outliers [17]. Iterative methods for these formulations are appealing since they do not need to project onto the set of low-rank matrices, which involves costly matrix factorizations that are prohibitively costly in large-scale settings.

In this section, we develop rates for Algorithm 1 applied to composite problems where the parameterization can be either $F_{\text{sym}}: \mathbf{R}^{d \times r} \rightarrow \mathcal{S}^d$ or $F_{\text{asym}}: \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r} \rightarrow \mathbf{R}^{d_1 \times d_2}$ given by

$$X \mapsto XX^\top \quad \text{and} \quad (X, Y) \mapsto XY^\top, \quad \text{respectively.} \quad (21)$$

We consider two concrete losses: $h(\cdot) = \frac{1}{2} \|\mathcal{A}(\cdot) - b\|_2^2$ and $h(\cdot) = \|\mathcal{A}(\cdot) - b\|_1$. In what follows, we develop theory for linear maps satisfying the standard restricted isometry property (RIP) or a modified version involving the ℓ_1 -norm—both of which hold for linear maps with appropriately normalized iid Gaussian entries. We leverage these results to derive guarantees for the ℓ_1 loss that hold even when gross outliers corrupt a constant fraction of the measurements.

5.2.1 Regularity of the parameterization

As a first step, we show that both parameterizations in (21) are smooth (Assumption 1) and establish weak alignment (Assumption 3) for the PSD factorization and its local analogue (Assumption 7) for the asymmetric factorization. The proofs of the following two results are rather technical and require carefully characterizing the spectrum of the Jacobians of these parametrizations; we defer these arguments to Appendices C.5.1 and C.5.2, respectively.

Theorem 5.5 (Weak alignment of PSD factorization). *The map $F_{\text{sym}}: \mathbf{R}^{d \times r} \rightarrow \mathcal{S}^d$ given by $X \mapsto XX^\top$ satisfies Assumption 1 with $L_{\nabla F} = 2$ and Assumption 3 with*

$$s(\rho) = \frac{4\rho}{\sqrt{2}(r - r^* + 1)}, \quad \text{and} \quad \delta(\rho) = \min \left\{ \frac{\rho}{\sqrt{2}}, \frac{1}{1 + s(\rho)}, \frac{1}{3} \right\} \lambda_{r^*}(Z^*).$$

for any $Z^* \in \mathcal{S}_+^d$ with $\text{rank}(Z^*) = r^*$.

Theorem 5.6 (Weak alignment of asymmetric factorization). *The map $F_{\text{asym}}: \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r} \rightarrow \mathbf{R}^{d_1 \times d_2}$ given by $(X, Y) \mapsto XY^\top$ satisfies Assumption 6 with $L_{\nabla F} = \sqrt{2}$ and $\varepsilon_{\nabla F} = \infty$, and Assumption 7 with*

$$\varepsilon_{x^*} = \frac{1}{16\sqrt{2}} \frac{\min \{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max \{\sigma_1(X^*), \sigma_1(Y^*)\}}, \quad s(\rho) = \frac{\rho}{10\sqrt{2}(r - r^* + 1)^2} \frac{\min \{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)},$$

$$\text{and} \quad \delta(\rho) = \min \left\{ \frac{\rho}{4}, \frac{1}{4s(\rho)} \right\} \min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\}$$

for any factorization $Z^* = X^*Y^{*\top}$ satisfying $\text{rank}(X^*) = \text{rank}(Y^*) = r^*$ and the right singular vectors of the two factors match $V^{X^*} = V^{Y^*}$.

These two regularity guarantees combined with the immediate fact that $h(Z) = \|Z - Z^*\|_F$ satisfies Assumption 4, can be used to derive fast convergence guarantees for Algorithm 1 applied to matrix factorization problems as the one we covered in the introduction; see Figure 1. We observe that for the asymmetric setting, the alignment is only local, and we require the right singular vectors of the two factors to be the same. Although this might sound restrictive, spectral initialization

procedures guarantee closedness to such balanced factors [23, 98]. We focus instead on more general matrix sensing problems where the input \mathcal{A} is not simply the identity.

5.2.2 Noiseless matrix sensing

In this section, we consider noiseless measurements $b = \mathcal{A}(Z^*)$ and the smooth objective $h(\cdot) = \frac{1}{2}\|\mathcal{A}(\cdot) - b\|_2^2$. Although we could instead work with the nonsquared loss, we restrict attention to the smooth objective since (i) the next section will explore an arguably more interesting nonsmooth loss, and (ii) most existing theory pertains to this setting. We will state definitions and some results only for the asymmetric case, since the extension to the positive semidefinite case follows immediately.

Our guarantees apply to maps satisfying the restricted isometry property (RIP)—a popular notion of strong identifiability that underpins most existing guarantees for linear inverse problems. A linear map $\mathcal{A}: \mathbf{R}^{d_1 \times d_2} \rightarrow \mathbf{R}^m$ satisfies RIP if there exists $\delta \in (0, 1)$ such that

$$(1 - \delta) \|Z\|_F^2 \leq \|\mathcal{A}(Z)\|_2^2 \leq (1 + \delta) \|Z\|_F^2 \quad (22)$$

for all matrices Z of rank at most r . In short, this property ensures that distances between low-rank matrices are approximately preserved after mapping by \mathcal{A} . While the identity map trivially satisfies this property, more interesting random maps with low-dimensional images also exhibit this behavior. We say that \mathcal{A} has i.i.d. entries if $\mathcal{A}(Z)_i = \langle A_i, Z \rangle$ where the entries of $A_i \in \mathbf{R}^{d_1 \times d_2}$ are drawn i.i.d. and the matrices A_i are independent of each other.

Lemma 5.7 (Theorem 2.3 in [15]). *Fix $r \leq \min(d_1, d_2)$ and $\delta \in (0, 1)$. Assume that \mathcal{A} has i.i.d. entries with distribution $N(0, 1/m)$. There exist universal constants $c_1, c_2, c_3 > 0$ such that if $m \geq c_1 r(d_1 + d_2)$, then \mathcal{A} satisfies (22) for all matrices Z of rank at most r with probability at least $1 - c_2 \exp(-c_3 m)$.*

In turn, RIP suffices for good conditioning. The proof of the next lemma is in Appendix C.5.3.

Lemma 5.8. *Suppose the map \mathcal{A} satisfies (22) for all matrices of rank at most $6r$, and that $b = \mathcal{A}(Z^*)$, with $Z^* \in \mathbf{R}^{d_1 \times d_2}$ a rank r matrix. Then, the function $h(\cdot) = \frac{1}{2}\|\mathcal{A}(\cdot) - b\|_2^2$ satisfies Assumption 5 with $\alpha = (1 - \delta)$ and $\beta = \frac{(1+\delta)^2}{(1-\delta)}$.*

Therefore, applying Theorem 5.5 (resp. Theorem 5.6) in tandem with the preceding lemma shows that the assumptions of the general convergence guarantee Theorem 4.5 (resp. Theorem B.13 in Appendix B.5) are satisfied in the symmetric (resp. asymmetric) case.⁴

Corollary 5.9 (Convergence for PSD matrix sensing). *Suppose that the measuring map $\mathcal{A}: \mathcal{S}^d \rightarrow \mathbf{R}^m$ satisfies (22) for all matrices Z of rank at most $6r$ and $b = \mathcal{A}(Z^*)$. Algorithm 1 is applied to the first objective in (20) with $\ell(z) = \|z\|_2^2$, initialized at X_0 using Configuration 1 with $\gamma \leq \min\{1, \frac{C_{ub}}{2}\}$, $C_{ub} \leq \frac{1}{64\sqrt{2(r-r^*+1)}} \frac{(1-\delta)^4}{(1+\delta)^4}$, and*

$$\|X_0 X_0^\top - Z^*\|_F \leq \frac{1}{16\sqrt{2}} \frac{(1-\delta)^2}{(1+\delta)^2} \lambda_{r^*}(Z^*).$$

Then, the iterates satisfy

$$\|X_k X_k^\top - Z^*\|_F^2 \leq \left(1 - \frac{\gamma (1-\delta)^2}{32(1+\delta)^2}\right)^k \|X_0 X_0^\top - Z^*\|_F^2 \quad \text{for all } k \geq 0.$$

⁴To derive the corollary in the asymmetric case we used that $1 - (1-x)^\alpha \geq \alpha x$ for all $x \in [0, 1]$ and $\alpha \in (0, 1)$

Corollary 5.10 (Convergence for asymmetric matrix sensing). *Suppose that the measuring map $\mathcal{A}: \mathbf{R}^{d_1} \times \mathbf{R}^{d_2} \rightarrow \mathbf{R}^m$ satisfies (22) for all matrices Z of rank at most $6r$ and $b = \mathcal{A}(Z^*)$. Let $X^*Y^{*\top} = Z^*$ be a factorization satisfying $\text{rank}(X^*) = \text{rank}(Y^*) = r^*$ and the right singular vectors of the two factors match $V^{X^*} = V^{Y^*}$. Assume Algorithm 1 is applied to the second objective in (20) with $\ell(z) = \|z\|_2^2$, initialized at (X_0, Y_0) using Configuration 1 with $\gamma \leq \min\{1, \frac{C_{ub}}{\sqrt{2}}\}$,*

$$C_{ub} \leq \frac{1}{2^{9.5}\sqrt{2}(r-r^*+1)^2} \frac{(1-\delta)^4 \min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{(1+\delta)^4 \sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)},$$

$$\|(X_0, Y_0) - (X^*, Y^*)\|_F \leq \frac{1}{32\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}}, \quad \text{and}$$

$$\|X_0 Y_0^\top - Z^*\|_F \leq \frac{1}{2^{23}} \frac{(1-\delta)^4 \min\{\sigma_{r^*}^4(X^*), \sigma_{r^*}^4(Y^*)\}}{(1+\delta)^4 \min\{\sigma_1^2(X^*), \sigma_1^2(Y^*)\}} C_{ub}.$$

Then, the iterates satisfy

$$\|X_k Y_k^\top - Z^*\|_F^2 \leq \left(1 - \frac{\gamma(1-\delta)^2}{32(1+\delta)^2}\right)^k \|X_0 Y_0^\top - Z^*\|_F^2 \quad \text{for all } k \geq 0.$$

The only dependency on the conditioning of Z^* appears in the size of the neighborhood where the algorithm exhibits linear convergence. The general guarantees under strong alignment can be used to derive faster rates in the exactly parameterized case; we omit such results for brevity.

5.2.3 Robust matrix sensing

In this section, we will study matrix sensing problems with gross outliers. That is, we consider corrupted measurements of the form

$$b = \begin{cases} \mathcal{A}(Z^*)_i & \text{if } i \in \mathcal{I}^c \\ \eta_i & \text{otherwise,} \end{cases} \quad (23)$$

where $\mathcal{I} \subseteq [m]$ is a subset of the entries and η_i is arbitrary. Inspired by [17], we consider (20) with $\ell(z) = \|z\|_1$. Before stating our results for this loss, we take a small detour to show that for nonsmooth matrix problems, the rather complicated Assumption 4 is implied by a more standard form of restricted conditioning. This matches the assumptions for ScaledGD [95]. We defer the proof of the next lemma to Appendix C.5.4.

Lemma 5.11. *Let $h: \mathbf{R}^{d_1 \times d_2} \rightarrow \mathbf{R}$ be a convex function and $Z^* \in \mathbf{R}^{d_1 \times d_2}$ satisfying the following two conditions.*

1. (**Restricted sharpness**) *For any $Z \in \mathbf{R}^{d_1 \times d_2}$ with $\text{rank } Z \leq r$ we have*

$$\mu \|Z - Z^*\|_F \leq |h(Z) - h(Z^*)|. \quad (24)$$

2. (**Restricted Lipschitzness**) *For any pair $Z, \tilde{Z} \in \mathbf{R}^{d_1 \times d_2}$ with $\text{rank}(Z - \tilde{Z}) \leq 2r$ we have*

$$|h(Z) - h(\tilde{Z})| \leq L \|Z - \tilde{Z}\|_F. \quad (25)$$

Then, h satisfies Assumption 4 with $F = F_{\text{asym}}$.

A completely analogous result holds for the symmetric parameterization; we omit it to avoid repetition. Next, we establish these notions of restricted Lipschitzness and sharpness. Just as in the noiseless case, we will enforce a restricted isometry property, but in this case, a mixed version with the ℓ_1 norm. In particular, we say that a linear map $\mathcal{A}: \mathbf{R}^{d_1 \times d_2} \rightarrow \mathbf{R}^m$ satisfies ℓ_1/ℓ_2 -RIP if

there exist constants $\omega_1, \omega_2 > 0$ such that

$$\omega_1 \|Z\|_F \leq \|\mathcal{A}(Z)\|_1 \leq \omega_2 \|Z\|_F \quad (26)$$

for all matrices Z of rank at most r . In turn, ℓ_1/ℓ_2 -RIP does not suffice to handle outliers. Instead, we require a slightly more restrictive condition. The map \mathcal{A} satisfies the \mathcal{I} -outlier bound if there exist a constant $\omega_0 > 0$ such that

$$\omega_0 \|Z\|_F \leq (\|\mathcal{A}_{\mathcal{I}^c}(Z)\|_1 - \|\mathcal{A}_{\mathcal{I}}(Z)\|_1) \quad (27)$$

for matrices Z of rank at most r , where $\mathcal{A}_{\mathcal{I}}(Z)$ and $\mathcal{A}_{\mathcal{I}^c}(Z)$ are the subvectors of $\mathcal{A}(Z)$ indexed by \mathcal{I} and \mathcal{I}^c . In turn, random Gaussian mappings also satisfy these properties.

Lemma 5.12 (Theorem 6.4 in [17]). *Fix $r \leq \min(d_1, d_2)$ and $\mathcal{I} \subseteq [m]$ with $\#\mathcal{I} < m/2$. Define $p_{\text{fail}} = \#\mathcal{I}/m$ and suppose that $\mathcal{A}: \mathbf{R}^{d_1 \times d_2} \rightarrow \mathbf{R}^m$ has i.i.d. Gaussian entries with distribution $N(0, 1/m^2)$. There exist universal constants $c_1, c_2, c_3 > 0$ such that if $m \geq \frac{c_1}{(1-2p_{\text{fail}})^2} \ln \left(c_2 + \frac{c_2}{(1-2p_{\text{fail}})^2} \right) r(d_1 + d_2 + 1)$, then \mathcal{A} satisfies (26) and (27) for matrices Z of rank at most r with probability at least $1 - 4 \exp(-c_3(1 - 2p_{\text{fail}})m)$.*

Several other random mappings satisfy (26) and (27), including those used for phase retrieval and blind deconvolution [17, Theorem 6.4]. The following lemma shows that whenever the measurement map satisfies RIP and the outlier bound, the loss function $h(\cdot) = \|\mathcal{A}(\cdot) - b\|_1$ satisfies the restricted Lipschitz continuity and sharpness. The proof of this lemma appeared in a slightly different form in [17]; we include it here for completeness.

Lemma 5.13. *Suppose that \mathcal{A} satisfies (26) and (27) for all matrices of rank at most $2r$ and that b is taken as in (23). Take the constants $\mu = \omega_0$ and $L = \omega_2$. Then, the function $h(\cdot) = \|\mathcal{A}(\cdot) - b\|_1$ satisfies (24) for all Z with $\text{rank} Z \leq r$ and (25) for all Z, \tilde{Z} with $\text{rank}(Z - \tilde{Z}) \leq 2r$.*

Proof. We start by establishing restricted sharpness. Label $\Delta = (\mathcal{A}(Z^*) - b)$, and let Z be an arbitrary matrix with rank at most r . Applying the reverse triangle inequality yields

$$\begin{aligned} |h(Z) - h(Z^*)| &= \left| \|\mathcal{A}(Z - Z^*) + \Delta\|_1 - \|\Delta\|_1 \right| \\ &= \left(\|\mathcal{A}_{\mathcal{I}^c}(Z - Z^*)\|_1 + \sum_{i \in \mathcal{I}} (|\mathcal{A}(Z - Z^*)|_i + |\Delta|_i| - |\Delta|_i|) \right) \\ &\geq (\|\mathcal{A}_{\mathcal{I}^c}(Z - Z^*)\|_1 - \|\mathcal{A}_{\mathcal{I}}(Z - Z^*)\|_1) \\ &\geq \omega_0 \|Z - Z^*\|_F, \end{aligned}$$

where the second inequality follows from (27).

Next, we demonstrate that the function h satisfies restricted Lipschitz continuity. Let Z and \tilde{Z} be two matrices such that $\text{rank}(Z - \tilde{Z}) \leq 2r$. Once more, the reverse triangle inequality yields

$$\begin{aligned} |h(Z) - h(\tilde{Z})| &= \left| \|\mathcal{A}(Z) - b\|_1 - \|\mathcal{A}(\tilde{Z}) - b\|_1 \right| \\ &\leq \|\mathcal{A}(Z - \tilde{Z})\|_1 \\ &\leq \omega_2 \|Z - \tilde{Z}\|_F, \end{aligned}$$

where the second inequality uses (26). This concludes the proof. \square

These results allow us to invoke Theorems 4.1 and 4.7 to derive the following two corollaries.

Corollary 5.14 (Convergence for robust PSD matrix sensing). *Suppose that the measurement map $\mathcal{A}: \mathcal{S}^d \rightarrow \mathbf{R}^m$ satisfies (26) and (27) for all matrices Z of rank at most $2r$, and that the vector*

$b \in \mathbf{R}^m$ is taken as in (23). Assume Algorithm 1 is applied to the first objective in (20) with $\ell(z) = \|z\|_1$, initialized at X_0 using Configuration 1 with $\gamma \leq \min\left\{1, \frac{C_{1b}}{2}\right\}$, $C_{ub} \leq \frac{1}{16\sqrt{2}(r-r^*+1)} \frac{\omega_0^2}{\omega_2^2}$, and

$$\|X_0 X_0^\top - Z^*\|_F \leq \frac{1}{8\sqrt{2}} \frac{\omega_0}{\omega_2} \lambda_{r^*}(Z^*).$$

Then, the iterates must satisfy

$$\|X_k X_k^\top - Z^*\|_F^2 \leq \left(1 - \frac{\gamma \omega_0^2}{8 \omega_2^2}\right)^k \|X_0 X_0^\top - Z^*\|_F^2 \quad \text{for all } k \geq 0.$$

Corollary 5.15 (Convergence for robust asymmetric matrix sensing). Suppose that $\mathcal{A}: \mathbf{R}^{d_1} \times \mathbf{R}^{d_2} \rightarrow \mathbf{R}^m$ satisfies (26) and (27) for all matrices Z of rank at most $2r$ and that the vector $b \in \mathbf{R}^m$ is taken as in (23). Let $X^* Y^{*\top} = Z^*$ be a factorization satisfying $\text{rank}(X^*) = \text{rank}(Y^*) = r^*$ and the right singular vectors of the two factors match $V^{X^*} = V^{Y^*}$. Assume Algorithm 1 is applied to the second objective in (20) with $\ell(z) = \|z\|_1$, initialized at (X_0, Y_0) using Configuration 1 with $\gamma \leq \min\left\{1, \frac{C_{1b}}{\sqrt{2}}\right\}$, $C_{ub} \leq \frac{1}{2^7 \cdot 5\sqrt{2}(r-r^*+1)^2} \frac{\omega_0^2}{\omega_2^2} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}$,

$$\|(X_0, Y_0) - (X^*, Y^*)\|_F \leq \frac{1}{32\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}}, \quad \text{and}$$

$$\|X_0 Y_0^\top - Z^*\|_F \leq \frac{1}{2^{19}} \frac{\omega_0^4}{\omega_2^4} \frac{\min\{\sigma_{r^*}^4(X^*), \sigma_{r^*}^4(Y^*)\}}{\min\{\sigma_1^2(X^*), \sigma_1^2(Y^*)\}} C_{1b}.$$

Then, the iterates satisfy

$$\|X_k Y_k^\top - Z^*\|_F^2 \leq \left(1 - \frac{\gamma \omega_0^2}{8 \omega_2^2}\right)^k \|X_0 Y_0^\top - Z^*\|_F^2 \quad \text{for all } k \geq 0.$$

5.3 Tensor factorization

Tensors are generalizations of matrices that store information in n modes as opposed to only two. They have numerous applications in recommender systems, biomedical imaging, quantum many-body simulations, and numerical linear algebra [2, 53, 54, 57, 75, 80, 90]. A major challenge in large-scale tensor analysis is the growth in storage and computation with increasing modes. To address this issue, practitioners typically employ low-rank tensor decompositions. Unlike the matrix SVD, tensor factorization admits no single canonical form; instead, a variety of models—such as canonical polyadic (CP), Tucker, and tensor train decompositions—are used, each with its properties and algorithmic trade-offs.

In this section, we focus on finding a CP tensor factorization. Although we will work only with third-order tensors, many results here likely extend to arbitrary tensors. We start by introducing some notation. Intuitively, a third-order tensor T can be viewed as a three-dimensional array of scalars. Given vectors $w \in \mathbf{R}^{d_1}$, $x \in \mathbf{R}^{d_2}$ and $y \in \mathbf{R}^{d_3}$ we use $w \otimes x \otimes y$ to denote a tensor with components given by $(w \otimes x \otimes y)_{ijk} = w_i x_j y_k$. A general tensor T has a CP decomposition of rank r if it can be written as $T = \sum_{i=1}^r w^{(i)} \otimes x^{(i)} \otimes y^{(i)}$; further the decomposition is symmetric if $w^{(i)} = x^{(i)} = y^{(i)}$ for all i . The *CP-rank* of T is the minimum r for which a CP decomposition exists; the *symmetric CP-rank* is defined analogously. We refer the interested reader to [53] for additional details. Our goal is, then, to factorize a three-dimensional tensor T^* with CP rank r^* . To do so, we

aim to fit the entries to one of the two explicit factorizations

$$\min_{X \in \mathbf{R}^{d \times r}} \left\| \sum_{j=1}^r X_j \otimes X_j \otimes X_j - T^* \right\|_F \quad \text{or} \quad \min_{\substack{W \in \mathbf{R}^{d_1 \times r}, X \in \mathbf{R}^{d_2 \times r}, \\ Y \in \mathbf{R}^{d_3 \times r}}} \left\| \sum_{j=1}^r W_j \otimes X_j \otimes Y_j - T^* \right\|_F, \quad (28)$$

depending on whether the tensor T^* is symmetric or not. Here, X_j denotes the j th column of X and the Frobenius norm is equal to the ℓ_2 norm of the vectorized tensor. These are instances of composite optimization with $h(T) = \|T - T^*\|_F$ and parameterizations

$$F_{\text{sym}}(X) := \sum_{j=1}^r X_j \otimes X_j \otimes X_j \quad \text{and} \quad F_{\text{asym}}(W, X, Y) := \sum_{j=1}^r W_j \otimes X_j \otimes Y_j. \quad (29)$$

Throughout, we assume the tensor of interest has CP-rank r^* .

Regularity of the parameterization. We show that the symmetric and asymmetric factorization maps satisfy local strong alignment (Assumption 8). Recall that to streamline the exposition, we present this assumption and its implications (Theorem B.12) only in Appendix B.5.2. We defer the proof of these theorems to Appendices C.6.1 and C.6.2, respectively.

Theorem 5.16 (Strong alignment of the symmetric CP map). *Let $X^* \in \mathbf{R}^{d \times r^*}$ be a full-rank matrix and set $T^* = F_{\text{sym}}(X^*)$. Then, the map F_{sym} with $r = r^*$ satisfies Assumption 6 at X^* with $\varepsilon_{\nabla F} = \|X^*\|_F$ and $L_{\nabla F} = 12 \|X^*\|_F$, and Assumption 8 at X^* with*

$$\varepsilon_{x^*} = \min \left\{ R, \frac{\sigma_{dr}(\nabla F_{\text{sym}}(X^*))}{24 \|X^*\|_F}, \|X^*\|_F \right\}, \quad \delta(\rho) = \frac{\rho}{C}, \quad \text{and} \quad s = \frac{1}{2} \sigma_{dr}(\nabla F_{\text{sym}}(X^*))$$

for some constants $R, C > 0$ that depend only on X^* .

Theorem 5.17 (Strong alignment of the asymmetric CP map). *Let $(W^*, X^*, Z^*) \in \mathbf{R}^{d_1 \times r^*} \times \mathbf{R}^{d_2 \times r^*} \times \mathbf{R}^{d_3 \times r^*}$ be full-rank matrices. Then, the map F_{asym} with $r = r^*$ satisfies Assumption 6 at (W^*, X^*, Y^*) with $\varepsilon_{\nabla F} = \|(W^*, X^*, Y^*)\|_F$ and $L_{\nabla F} = 4\sqrt{3} \|(W^*, X^*, Y^*)\|_F$, and Assumption 8 at (W^*, X^*, Y^*) with*

$$\varepsilon_{x^*} = \min \left\{ R, \frac{\sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Z^*))}{8\sqrt{3} \|(W^*, X^*, Y^*)\|_F}, \|(W^*, X^*, Y^*)\|_F \right\}, \quad \delta(\rho) = \frac{\rho}{C},$$

$$\text{and} \quad s = \frac{1}{2} \sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Y^*))$$

for some constants $R, C > 0$ that depend only on (W^*, X^*, Y^*) .

We observe that, unlike our results for matrix factorization, here, we only handle ill-conditioning and fail to capture the overparameterized settings. Nonetheless, our numerical experiments (Section 6) suggest that Algorithm 1 converges linearly even for overparameterized problems.

Convergence rates. The outer function for tensor factorization $h(T) = \|T - T^*\|_F$ is trivially well-conditioned, in particular, it satisfies Assumption 5 with $\mu = L = 1$. Thus, applying Theorem B.12 yields the following two corollaries.

Corollary 5.18 (Convergence rate for symmetric CP tensor factorization). *Let $T^* \in \mathbf{R}^d \otimes \mathbf{R}^d \otimes \mathbf{R}^d$ be a symmetric tensor with symmetric CP rank r^* and let $X^* \in \mathbf{R}^{d \times r^*}$ be such that $T^* = F_{\text{sym}}(X^*)$. Consider the first problem (28) with $r = r^*$ and suppose that we ran Algorithm 1*

initialized at X_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{12\|X^*\|_F} \right\}$ and

$$\|X_0 - X^*\|_F \leq \frac{1}{2} \min \left\{ R, \frac{\sigma_{dr}(\nabla F_{\text{sym}}(X^*))}{24\|X^*\|_F}, \|X^*\|_F \right\},$$

$$\|F_{\text{sym}}(X_0) - T^*\|_F \leq \min \left\{ \frac{1}{8C}, \frac{\sigma_{dr}(\nabla F_{\text{sym}}(X^*))}{16C_{ub}}, \frac{1}{2^{10}} C_{lb} \min \left\{ R, \frac{\sigma_{dr}(\nabla F_{\text{sym}}(X^*))}{24\|X^*\|_F}, \|X^*\|_F \right\}^2 \right\},$$

where $C, R > 0$ are constants depending only on X^* . Then, the iterates satisfy

$$\|F_{\text{sym}}(X_k) - T^*\|_F^2 \leq \left(1 - \frac{\gamma}{8} \right)^k \|F_{\text{sym}}(X_0) - T^*\|_F^2 \quad \text{for all } k \geq 0.$$

Corollary 5.19 (Convergence for asymmetric CP tensor factorization). Let $T^* \in \mathbf{R}^{d_1} \otimes \mathbf{R}^{d_2} \otimes \mathbf{R}^{d_3}$ be a tensor with CP rank r^* and let $(W^*, X^*, Y^*) \in \mathbf{R}^{d_1 \times r^*} \times \mathbf{R}^{d_2 \times r^*} \times \mathbf{R}^{d_3 \times r^*}$ be such that $T^* = F_{\text{asym}}(W^*, X^*, Y^*)$. Consider the second problem in (28) with $r = r^*$ and suppose that we ran Algorithm 1 initialized at W_0, X_0, Y_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{4\sqrt{3}\|(W^*, X^*, Y^*)\|_F} \right\}$,

$$\|(W_0, X_0, Y_0) - (W^*, X^*, Y^*)\|_F \leq \frac{1}{2} \min \left\{ R, \frac{\sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Y^*))}{8\sqrt{3}\|(W^*, X^*, Y^*)\|_F}, \|(W^*, X^*, Y^*)\|_F \right\},$$

and

$$\begin{aligned} & \|F_{\text{asym}}(W_0, X_0, Y_0) - T^*\|_F \\ & \leq \min \left\{ \frac{1}{8C}, \frac{\sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Y^*))}{16C_{ub}}, \right. \\ & \quad \left. \frac{1}{2^{10}} C_{lb} \min \left\{ R, \frac{\sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Y^*))}{8\sqrt{3}\|(W^*, X^*, Y^*)\|_F}, \|(W^*, X^*, Y^*)\|_F \right\}^2 \right\}, \end{aligned}$$

where $C, R > 0$ are constants depending only on (W^*, X^*, Y^*) . Then, the iterates satisfy

$$\|F_{\text{asym}}(W_k, X_k, Y_k) - T^*\|_F^2 \leq \left(1 - \frac{\gamma}{8} \right)^k \|F_{\text{asym}}(W_0, X_0, Y_0) - T^*\|_F^2 \quad \text{for all } k \geq 0.$$

Unlike our results for matrices, our tensor guarantees only handle exactly parameterized problems. Moreover, by invoking Theorem B.14, we can derive similar rates with worse constants for the smooth loss $h(T) = \frac{1}{2} \|T - T^*\|_2^2$.

6 Numerical experiments

In this section, we present numerical results that support our theoretical guarantees. Sections 6.1, 6.2, and 6.3 include experiments for nonnegative least squares, matrix sensing, and tensor factorization, respectively. The code for reproducing these experiments is available at

https://github.com/aglabassi/preconditioned_composite_opti.

Implementation details. We run all methods on a Google Colab compute unit with 12GB of system RAM, and a T4 GPU with 16GB of RAM (A100 for tensor experiments). We use Python 3.11.11 and Pytorch 2.5.1 paired with Cuda 12.4. Further, we use Pytorch's double-precision floating-point format. For almost all experiments, we solve the linear systems via the Conjugate Gradient method with a maximum of 100 iterations and a tolerance level of 10^{-25} .

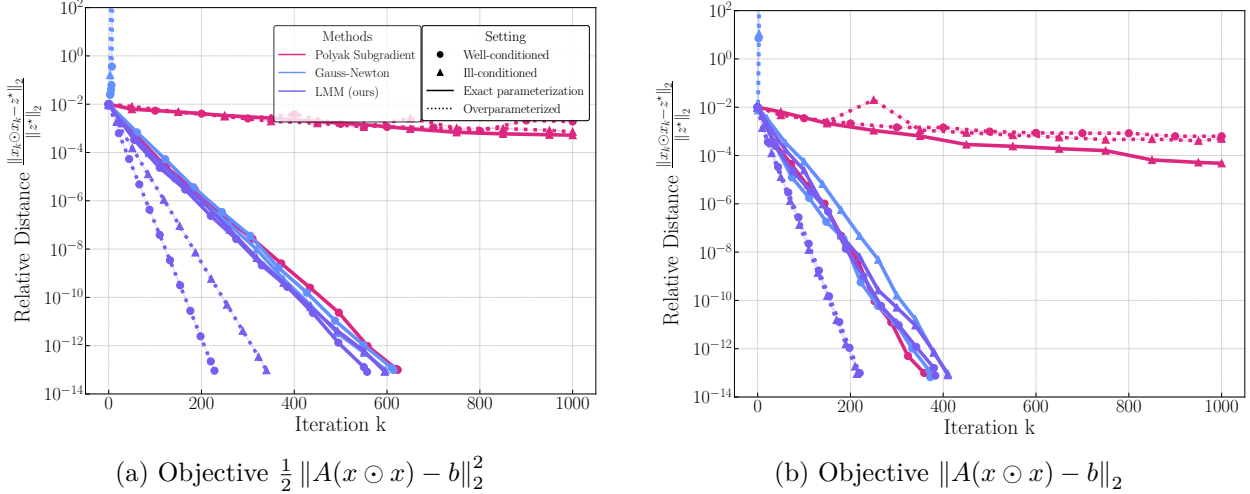


Figure 2: Relative distance against iteration count for nonnegative least squares losses (19).

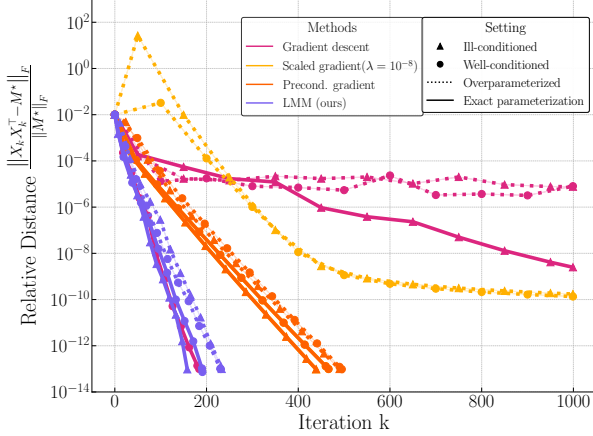
We use implicit evaluations of the matrix-vector products $\nabla F(x)^\top \nabla F(x)v$ using the derivations given in the Appendix E. The only expectation is nonnegative least squares, for which we use $(\nabla F(x)^\top \nabla F(x) + \lambda I)^{-1}v = v \odot \frac{1}{x \odot x + \lambda \mathbf{I}}$ directly.

6.1 Nonnegative least squares with smooth and nonsmooth losses

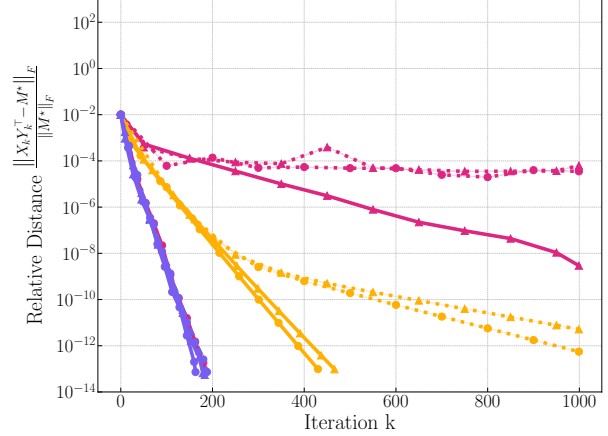
For our first experiment, we consider the two nonnegative least squares formulations(19) from Section 5.1. We generate the ground truth via $z^* = \left[1, \dots, \frac{1}{\tau}, \mathbf{0}_{r-r^*}\right]^\top \in \mathbb{R}^r$, where τ and $r - r^*$ respectively control the ill-conditionedness and overparameterization of the map $F(x) = x \odot x$. Ill-conditioning of the map $F(x) = x \odot x$ at z^* occurs when $\max_{i|z_i^* \neq 0} |z_i^*| \gg \min_{i|z_i^* \neq 0} |z_i^*|$, and overparameterization when $\dim(z^*) > \|z^*\|_0$. We vary $\tau \in \{1, 100\}$ and $r \in \{10, 100\}$, and take $r^* = 10$. We generate matrices $A \in \mathbb{R}^{m \times r}$ with $m = 2r$, and $\kappa(A) = 10$, and set $b = Az^*$. We initialize all methods at the same random x_0 satisfying $\|x_0 \odot x_0 - z^*\|_2 = 10^{-2} \|z^*\|_2$.

Baselines. We compare the performance of iterative methods applied to the smooth and nonsmooth formulations in (19). For both losses, we test Algorithm 1 against the standard subgradient method and the Gauss-Newton subgradient method from [26]. All methods use the Polyak-type stepsizes. For the subgradient method it is exactly the Polyak stepsize $(f(x_k) - \min f) / \|g_k\|^2$ with $g_k \in \partial f(x_k)$. For the other two methods, we use the stepsize from Configuration 1. For Algorithm 1 we use $\lambda_k = 10^{-2} \|Ax - b\|_2$ as an estimator of the quantity $\|x_k \odot x_k - z^*\|_2$, which emulates Configuration 1 without requiring access to z^* .

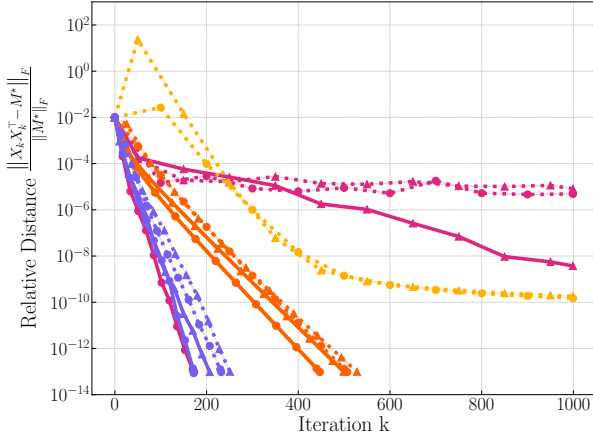
Discussion. Figure 2 displays the results. On the one hand, the Polyak subgradient method fails to converge linearly, whether there is ill-conditioning or overparameterization, and, further, Gauss-Newton diverges in the overparameterization case, which is expected as the precondition is ill-defined. On the other hand, Algorithm 1 is robust and converges linearly in all settings. Notably, the methods exhibit faster convergence when applied to the nonsmooth formulations, highlighting the benefit of using a nonsmooth loss for regression.



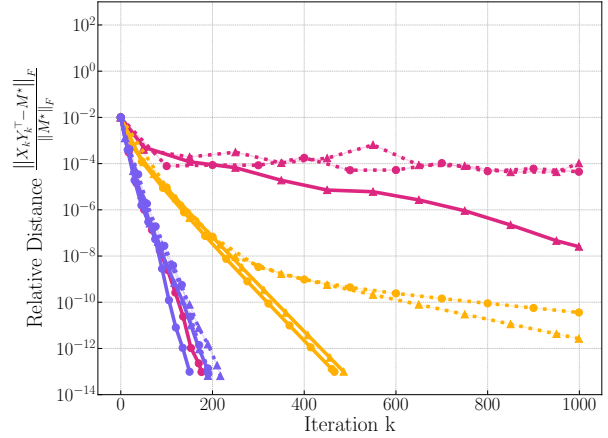
(a) Symmetric Matrix ($d = 100$)



(b) Asymmetric Matrix ($d = 100$)



(c) Symmetric Matrix ($d = 200$)



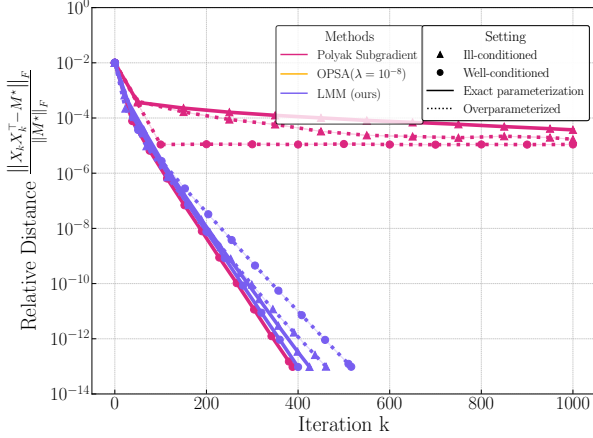
(d) Asymmetric Matrix ($d = 200$)

Figure 3: Smooth matrix sensing with the ℓ_2 -norm squared. We use $m = 4dr$ ($m = 2dr$ for symmetric), with $r^* = 2$, $r \in \{2, 5\}$.

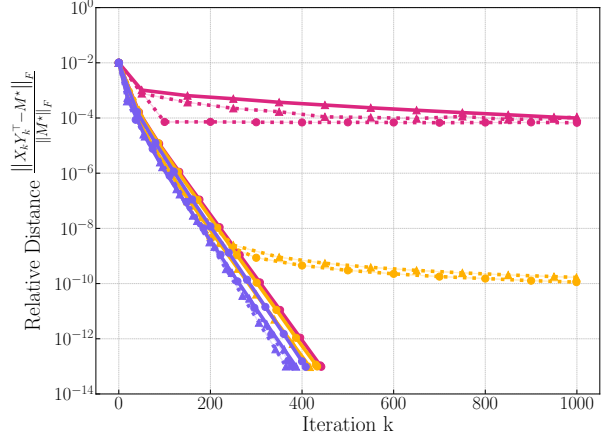
6.2 Matrix Sensing

For our second batch of experiments, we consider the matrix problems introduced in (20). We run three experiments to evaluate (i) convergence, (ii) hyperparameter sensitivity, and (iii) robustness to outliers. All three types of experiments use similar losses and parameter configurations, which we describe next.

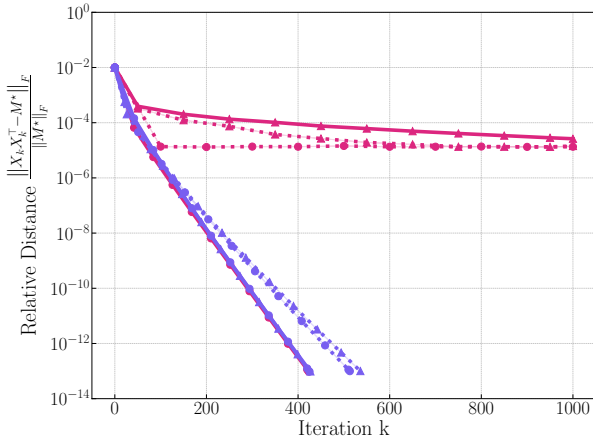
Setup. We solve matrix sensing using the squared ℓ_2 -loss $h(M) = \|\mathcal{A}(M) - b\|_2^2$ and the ℓ_1 -loss $h(M) = \|\mathcal{A}(M) - b\|_1$. We consider both PSD and general asymmetric ground truths: for PSD sensing we set $M^* = X^*X^{*\top}$ where $X^* = U D^{1/2}$ with $U \in \mathbb{R}^{d \times r^*}$ drawn at random satisfying $U^\top U = I$ and $D = \text{diag}(\xi_1, \dots, \xi_{r^*})$ with ξ_i linearly spaced in $[1/\tau, 1]$ for $\tau \in \{1, 100\}$; for asymmetric sensing we similarly draw $Y^* = V D^{1/2}$ and set $M^* = X^*(Y^*)^\top$ to ensure $\kappa(M^*) = \tau$. To test dimension-independent convergence we vary $d \in \{100, 200\}$, and to probe overparameterization we fix $r^* = 2$ while varying $r \in \{2, 5\}$. The map $\mathcal{A} : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^m$ has i.i.d. $N(0, \frac{1}{m})$ entries with $m = 2dr$ (or $m = 4dr$ for asymmetric), and all methods are initialized identically with relative error 10^{-2} .



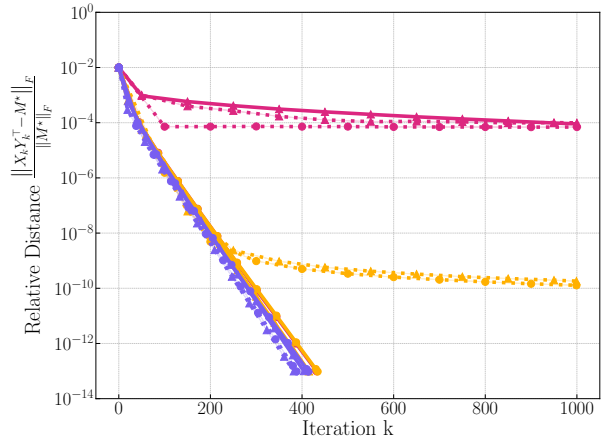
(a) Symmetric Matrix ($d = 100$)



(b) Asymmetric Matrix ($d = 100$)



(c) Symmetric Matrix ($d = 200$)



(d) Asymmetric Matrix ($d = 200$)

Figure 4: Matrix sensing with the ℓ_1 -norm. We use $m = 4dr$ ($m = 2dr$ for symmetric) with $r^* = 2$, $r \in \{2, 5\}$. OPSA [44] only applies to the asymmetric setting.

Baselines. For the smooth problems, we compare with gradient descent, **PrecGD** [106] (symmetric only), **ScaledGD**(λ) [104]. In the nonsmooth setting, we compare our method against the Polyak subgradient method, and **OPSA** [44] (asymmetric only). Unless otherwise stated, our method uses Polyak stepsizes (Configuration 1), where we set $\gamma = 1$. For constant-stepsize methods, **PrecGD** and **ScaledGD**(λ), we tune to select the largest parameter that leads to convergence, i.e., $\gamma_k = 1/2$. For **ScaledGD**(λ) we set $\lambda = 10^{-8}$. For **PrecGD** and Algorithm 1 we use a damping parameter of $\lambda_k = 2.5 \cdot 10^{-3} \sqrt{f(x_k)}$ in the smooth setting, or $\lambda_k = 10^{-5} \cdot f(x_k)$ in the nonsmooth one, as an estimator for the quantity $\|z_k - z^*\|_2$.

Experiment 1: convergence rates. We generate noiseless observations $b = \mathcal{A}(M^*)$ and solve the recovery problem using both the squared ℓ_2 -norm and the ℓ_1 -norm. Figures 3 and 4 report results for the smooth and nonsmooth formulations, respectively, benchmarked against established competitor methods. Algorithm 1 consistently matches or outperforms existing approaches across both problem classes. **ScaledGD** and **OPSA** employ a fixed damping parameter, which restricts their linear convergence to a neighborhood around the optimum; by contrast, **PrecGD** and our method

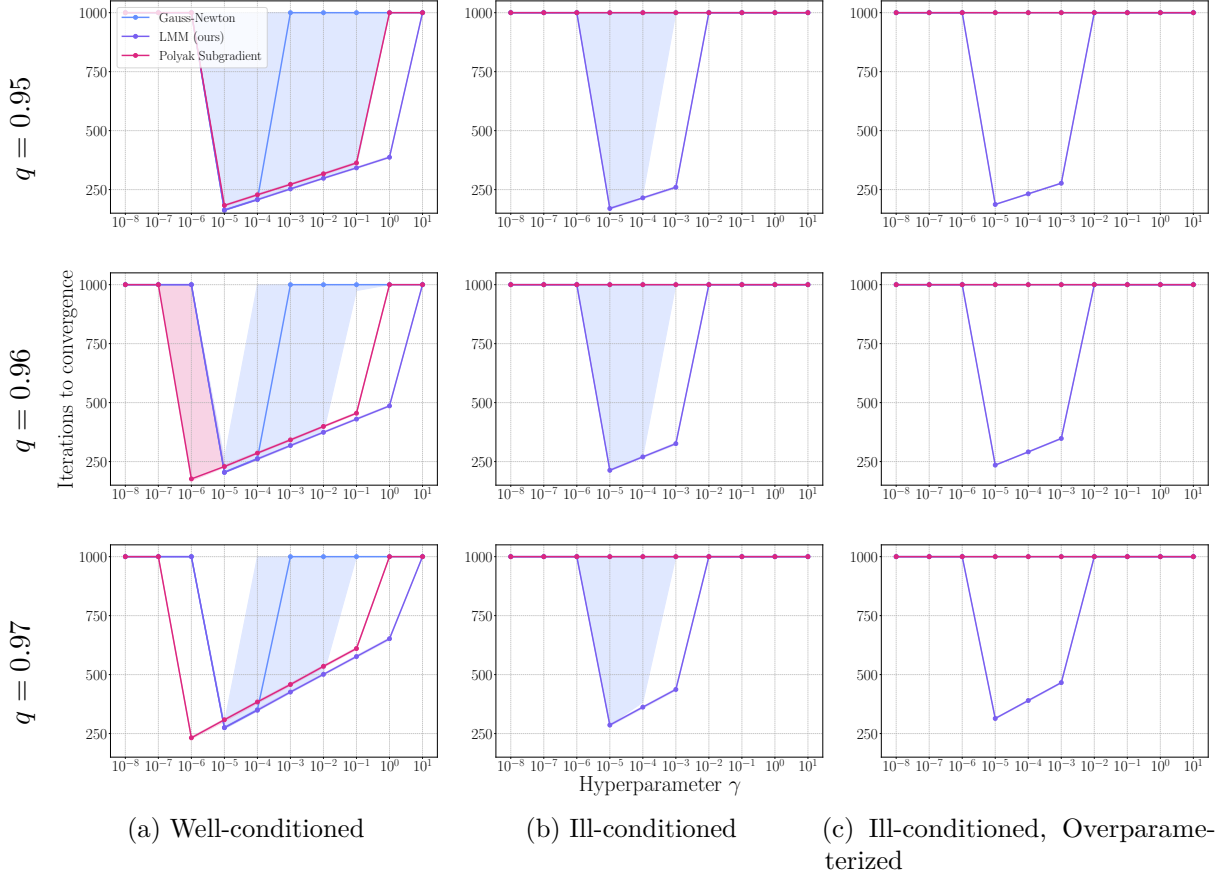


Figure 5: Median number of iterations to achieve convergence (100 draws) versus hyperparameter γ . We declare that a method converged when it reaches a relative error of 10^{-8} and cap the maximum number of iterations to 1000. The shaded area represents the 5th and 95th percentiles, respectively.

sustain linear convergence all the way to the exact solution. Furthermore, corroborate our theoretical finding that Algorithm 1 achieves a convergence rate independent of the problem dimension.

Experiment 2: hyperparameter sensitivity. In this experiment, we probe the robustness of Configuration 2 for the ℓ_1 norm loss. We set stepsizes to $\gamma_k = \gamma q^k$ and damping parameters to $\lambda_k = 10^{-5} q^k$ and vary $q \in \{0.95, 0.96, 0.97\}$ and $\gamma \in \{10^{-j} \mid j = 1, \dots, 8\}$. We cap the total number of iterations at 10^3 . Compared to the previous experiment, we take a smaller dimension $d = 30$ and consider more aggressive ill-conditioning by varying $\tau \in \{1, 10^4\}$. Figure 5 shows the median number of iterations needed to achieve a given relative error of 10^{-8} over 100 trials. This experiment suggests that Algorithm 1 converges efficiently across a broad spectrum of hyperparameter settings.

Experiment 3: robustness to gross outliers. For our last batch of experiments, we test the ability of our method to solve the PSD, ℓ_1 norm formulation, with different levels of gross outliers. We set the dimension to $d = 30$ and consider more aggressive ill-conditioning by varying $\tau \in \{1, 10^4\}$. We corrupt the vector b via (23) where the outliers are set to $\eta_i = \mathcal{A}(\overline{M})_i$ for some other random matrix $\overline{M} \in \mathcal{S}_+^d$. We vary the corruption level $p_{\text{fail}} = \#\mathcal{I}/m$ between 0 and $1/2$. Notice that when $p_{\text{fail}} > 1/2$, the solution switches to \overline{M} . We compare against the standard and Gauss-Newton subgradient methods [26]. The Polyak stepsize is not applicable because the true minimum value

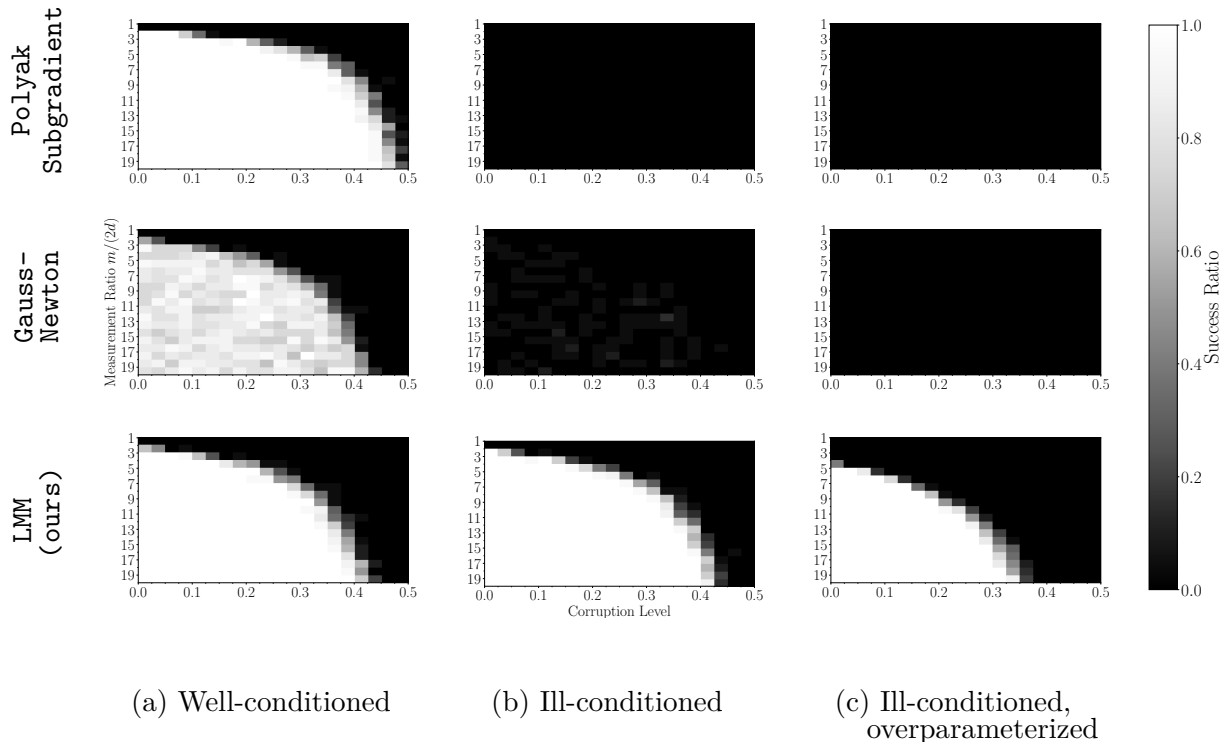


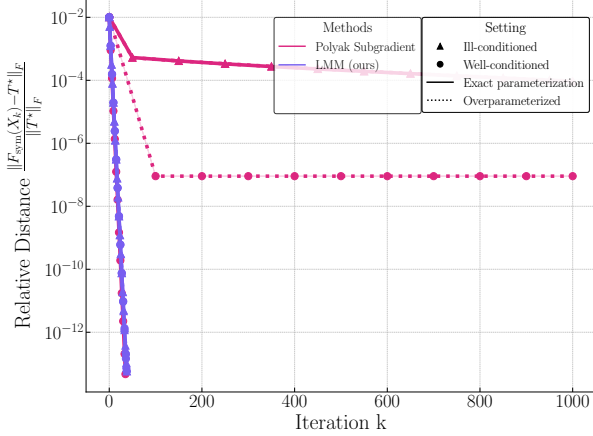
Figure 6: Matrix sensing transition plots of success rates (in %) over 20 trials for each (m, p_{fail}) . Success is declared when the relative error is below $\epsilon = 10^{-8}$ with an iteration budget of 500.

$\min f$ is unknown. Instead, we use Configuration 2 with $\lambda = 10^{-5}$, $\gamma = 10^{-4}$, and $q = 0.97$. Figure 6 displays the results with phase transition plots. For each pair (m, p_{fail}) , we run 20 problem instances and report the success ratio. A run is successful if, within 500 iterations, it achieves a relative error to fall below $\epsilon = 10^{-8}$. The Gauss-Newton preconditioned method exhibits unpredictable behavior when employing these geometrically decaying stepsizes; indeed, the guarantees in [26] do not cover this stepsize strategy. On the other hand, Algorithm 1 displays more stable performance, supporting our theory.

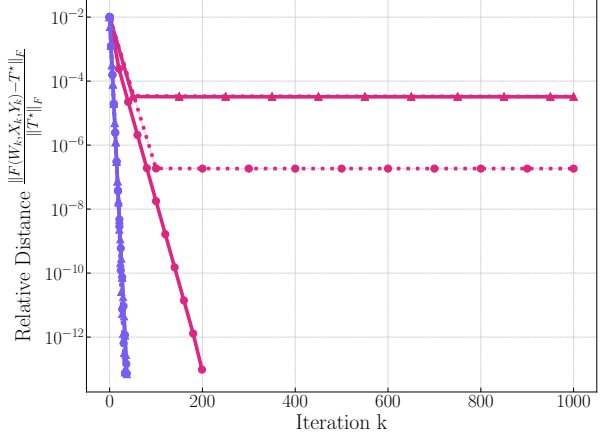
6.3 Tensor factorization and sensing

Finally, in our last batch of experiments, we evaluate Algorithm 1 on both tensor factorization (28) and robust tensor sensing. While our theoretical guarantees address only the factorization setting, the empirical results suggest that Algorithm 1 also works for tensor sensing. We leave the formal analysis of this case as an open question for future work.

Setup. For the factorization problem, we use the ℓ_2 -norm $h(T) = \|T - T^*\|_2$. For the sensing problem, we use the ℓ_1 -loss $h(T) = \|\mathcal{A}(T) - b\|_1$ with \mathcal{A} a linear measurement map and consider both symmetric and asymmetric CP-factorizations (29). We generate factor matrices $W^*, X^*, Y^* \in \mathbb{R}^{d \times r^*}$ by drawing $U \in \mathbb{R}^{d \times r^*}$ uniformly with $U^\top U = I$ and setting $X^* = U D^{1/3}$, where the diagonal matrix D has entries spanning $[1/\tau, 1]$; the ground-truth tensor is then $T^* = F_{\text{sym}}(X^*)$ or $T^* = F_{\text{asym}}(W^*, X^*, Y^*)$ depending on the experiment. All methods are initialized at random with relative error 10^{-2} . We set $d = 500$ for factorization and $d = 50$ for sensing, vary $\tau \in \{1, 100\}$ and $r \in \{2, 5\}$, and draw \mathcal{A} with i.i.d. $\mathcal{N}(0, 1/m)$ entries with $m = 5dr$ (or $30dr$ for asymmetric), taking

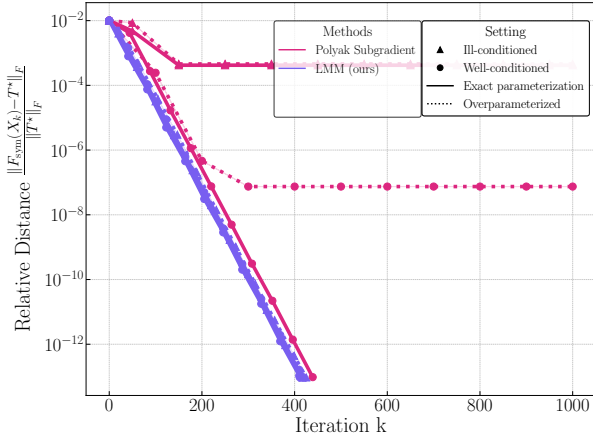


(a) Symmetric Tensor ($d = 500$)

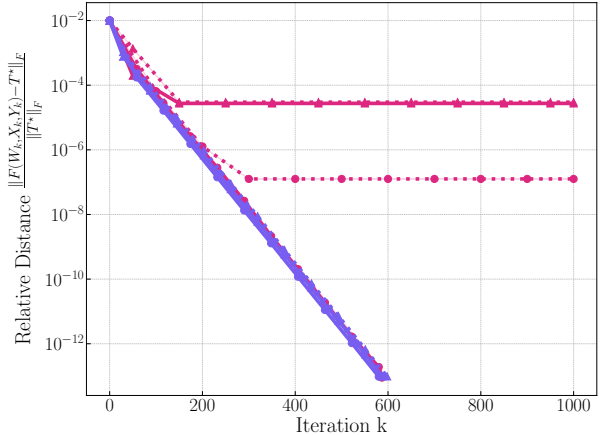


(b) Asymmetric Tensor ($d = 500$)

Figure 7: Tensor factorization with the ℓ_2 -norm. We use $r^* = 2$, $r \in \{2, 5\}$.



(a) Symmetric Tensor ($d = 50$)



(b) Asymmetric Tensor ($d = 50$)

Figure 8: Robust tensor sensing with the ℓ_1 -norm. We use $m = 5dr$ ($30dr$ for asymmetric) with $r^* = 2$, $r \in \{2, 5\}$ and 10% of gross outliers.

observations $b = \mathcal{A}(T^*)$.

Baselines. To our knowledge, no preconditioned first-order method offers convergence guarantees for CP tensor factorization.⁵ Thus, we only test against the subgradient method. For the factorization experiment, we use Configuration 1 with $\gamma = \frac{1}{2}$ and $\lambda_k = 10^{-3}f(x_k)$. For the robust sensing experiment, we use Configuration 2 with $\gamma = 10^{-3}$ (10^{-5} for asymmetric), $\lambda = 10^{-5}$ and $q = 0.94(0.96$ for asymmetric).

Discussion. Figure 7 shows the output for large tensor factorization using the ℓ_2 -norm. Algorithm 1 consistently displays fast convergence. Although we do not include a plot here, we observe that the convergence is much faster when using the unsquared ℓ_2 compared to its squared counterpart. This observation is consistent with the nonnegative least squares experiment. Figure 8 shows the

⁵A provably convergent version of `ScaledGD` exists for the Tucker asymmetric factorization [36].

convergence for tensor sensing using the ℓ_1 -norm with 10% gross outliers of the form $\eta_i = \mathcal{A}(\bar{T})_i$ for a spurious signal $\bar{T} \in \mathbf{R}^{d \times d \times d}$. Consistently, Algorithm 1 outperforms the subgradient method while remaining robust to ill-conditioning and overparameterization.

Acknowledgements

We thank Philippe Toint for pointing us to relevant literature, particularly for pointing us to the work of Morrison [78], whose contribution to the development of Algorithm 1 for nonlinear least-squares has too often gone unrecognized.

References

- [1] E. Acar, D. M. Dunlavy, and T. G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 25(2):67–86, 2011. doi: 10.1002/cem.1335.
- [2] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup. Scalable tensor factorizations for incomplete data. In *Proceedings of the 2011 SIAM International Conference on Data Mining (SDM)*, pages 701–712, Mesa, AZ, 2011.
- [3] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski. Bundle adjustment in the large. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, volume 6312 of *Lecture Notes in Computer Science*, pages 29–42. Springer, 2010.
- [4] I. K. Argyros and S. Hilout. On the gauss–newton method. *Journal of Applied Mathematics and Computing*, 35:537–550, 2011. doi: 10.1007/s12190-010-0377-8.
- [5] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.
- [6] M. H. V. Benthem and M. R. Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(10):441–450, 2004.
- [7] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.
- [8] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.
- [9] Å. Björck. *Numerical methods for least squares problems*. SIAM, 2024.
- [10] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3. Springer-Verlag, New York, 2000. ISBN 0-387-98940-4. Theory and examples.
- [11] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical programming*, 95(2):329–357, 2003.
- [12] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical programming*, 103(3):427–444, 2005.
- [13] J. V. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33:260–279, 1985.

- [14] J. V. Burke and M. C. Ferris. A gauss–newton method for convex composite optimization. *Mathematical Programming*, 71:179–194, 1995.
- [15] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.
- [16] C. Cartis, N. I. M. Gould, and P. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.
- [17] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.
- [18] G. H.-G. Chen and R. T. Rockafellar. Convergence rates in forward-backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.
- [19] P. Chen. Hessian matrix vs. gauss–newton hessian matrix. *SIAM Journal on Numerical Analysis*, 49(4):1417–1435, 2011. doi: 10.1137/100799988. URL <https://doi.org/10.1137/100799988>.
- [20] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.
- [21] Y. Chen, Y. Chi, J. Fan, and C. Ma. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021. ISSN 1935-8245. doi: 10.1561/22000000079. URL <http://dx.doi.org/10.1561/22000000079>.
- [22] C. Cheng and Z. Zhao. Accelerating gradient descent for over-parameterized asymmetric low-rank matrix sensing via preconditioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7705–7709. IEEE, 2024.
- [23] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- [24] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.
- [25] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.
- [26] D. Davis and T. Jiang. A linearly convergent gauss-newton subgradient method for ill-conditioned problems. *arXiv preprint arXiv:2212.13278*, 2022.
- [27] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179:962–982, 2018.
- [28] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.
- [29] D. Davis, D. Drusvyatskiy, and L. Jiang. Gradient descent with adaptive stepsize converges (nearly) linearly under fourth-order growth. *arXiv preprint arXiv:2409.19791*, 2024.

- [30] J. Diakonikolas, C. Li, S. Padmanabhan, and C. Song. A fast scale-invariant algorithm for non-negative least squares with non-negative data. In *NeurIPS*, 2022.
- [31] M. Díaz. The nonsmooth landscape of blind deconvolution. *NeurIPS Workshop: Optimization for Machine Learning*, 2019.
- [32] M. Díaz, A. J. Quiroz, and M. Velasco. Local angles and dimension estimation from data on manifolds. *Journal of Multivariate Analysis*, 173:229–247, 2019.
- [33] L. Ding and S. J. Wright. On squared-variable formulations. *arXiv preprint arXiv:2310.01784*, 2023. Available at <https://arxiv.org/abs/2310.01784>.
- [34] L. Ding, L. Jiang, Y. Chen, Q. Qu, and Z. Zhu. Rank overspecified robust matrix recovery: Subgradient method and exact recovery. *Advances in Neural Information Processing Systems*, 34:26767–26778, 2021.
- [35] L. Ding, Z. Qin, L. Jiang, J. Zhou, and Z. Zhu. A validation approach to over-parameterized matrix and image recovery. *arXiv preprint arXiv:2209.10675*, 2022.
- [36] H. Dong, T. Tong, C. Ma, and Y. Chi. Fast and provable tensor robust principal component analysis via scaled gradient descent. *arXiv preprint arXiv:2206.09109*, 2022.
- [37] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 2016. To appear; arXiv:1602.06661.
- [38] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.
- [39] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. Preprint, 2017.
- [40] J.-y. Fan and Y.-x. Yuan. On the quadratic convergence of the levenberg–marquardt method without nonsingularity assumption. *Computing*, 74(1):23–39, 2005.
- [41] A. Fischer, A. F. Izmailov, and M. V. Solodov. The levenberg–marquardt method: an overview of modern convergence theories and more. *Computational Optimization and Applications*, 89(1):33–67, 2024.
- [42] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. In *Nondifferential and variational techniques in optimization*, pages 67–76. Springer, 2009.
- [43] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.
- [44] P. Giampouras, H. Cai, and R. Vidal. Guarantees of a preconditioned subgradient algorithm for overparameterized asymmetric low-rank matrix recovery. *Preprint*, 2024.
- [45] J.-L. Goffin. On convergence rates of subgradient optimization methods. *Mathematical programming*, 13:329–347, 1977.
- [46] M. T. Hagan and M. B. Menhaj. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994.

- [47] A. F. Izmailov, M. V. Solodov, and E. I. Uskov. A globally convergent levenberg–marquardt method for equality-constrained optimization. *Computational Optimization and Applications*, 2019. doi: 10.1007/s10589-019-00123-0. Early access.
- [48] X. Jia, F. Fangchen, D. Meng, and D. Sun. Globally q-linear gauss-newton method for overparameterized non-convex matrix sensing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [49] L. Jiang, Y. Chen, and L. Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *SIAM Journal on Mathematics of Data Science*, 5(3):723–744, 2023.
- [50] J. Jin, Z. Li, K. Lyu, S. S. Du, and J. D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *International Conference on Machine Learning*, pages 15200–15238. PMLR, 2023.
- [51] R. G. Karim, D. Dulal, and C. Navasca. A modified levenberg-marquardt algorithm for tensor cp decomposition in image compression. In *2024 Data Compression Conference (DCC)*, page 563–563. IEEE, Mar. 2024. doi: 10.1109/dcc58796.2024.00080. URL <http://dx.doi.org/10.1109/DCC58796.2024.00080>.
- [52] D. Kim, S. Sra, and I. S. Dhillon. A non-monotonic method for large-scale non-negative least squares. *Optimization Methods and Software*, 28(5):1012–1039, 2013.
- [53] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. doi: 10.1137/07070111X.
- [54] K. K. Krishnan and K. P. Soman. Canonical polyadic decomposition of eeg image tensor for bci applications. In M. Tuba, S. Akashe, and A. Joshi, editors, *ICT Systems and Sustainability*, pages 819–826, Singapore, 2022. Springer Nature Singapore. ISBN 978-981-16-5987-4.
- [55] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.
- [56] G. Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159(1):201–235, 2016.
- [57] J. M. Landsberg. *Tensors: geometry and applications: geometry and applications*, volume 128. American Mathematical Soc., 2011.
- [58] E. V. Laufer and B. Nadler. Rgnmr: A gauss-newton method for robust matrix completion with theoretical guarantees. *arXiv preprint arXiv:2505.12919*, 2025.
- [59] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*, volume 15 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995. Revised reprint of the 1974 original.
- [60] J. M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, 2003. ISBN 0-387-95495-3.
- [61] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

- [62] E. Levin, J. Kileel, and N. Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *Mathematical Programming*, pages 1–49, 2024.
- [63] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, pages 1–46, 2015.
- [64] L. Li and T. P. Speed. Parametric deconvolution of positive spike trains. *The Annals of Statistics*, 28(5):1279–1301, 2000.
- [65] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal. Nonconvex robust low-rank matrix recovery. *SIAM Journal on Optimization*, 30(1):660–686, 2020.
- [66] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.
- [67] Y. Lin, D. D. Lee, and L. K. Saul. Nonnegative deconvolution for time of arrival estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii–377. IEEE, 2004.
- [68] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979. doi: 10.1137/0716071.
- [69] Z. Liu, Z. Han, Y. Tang, S. Tang, and Y. Wang. Efficient over-parameterized matrix sensing from noisy measurements via alternating preconditioned gradient descent. *arXiv preprint arXiv:2502.00463*, 2025.
- [70] Y. Luo and A. R. Zhang. Low-rank tensor estimation via riemannian gauss-newton: Statistical optimality and second-order convergence. *The Journal of Machine Learning Research*, 24(1): 18274–18321, 2023.
- [71] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 2019.
- [72] J. Ma and S. Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *Journal of Machine Learning Research*, 24(96):1–84, 2023.
- [73] Z. Ma, Y. Bi, J. Lavaei, and S. Sojoudi. Geometric analysis of noisy low-rank matrix recovery in the exact parametrized and the overparametrized regimes. *INFORMS Journal on Optimization*, 2023.
- [74] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.
- [75] M. Miranda. A canonical polyadic tensor basis for fast bayesian estimation of multi-subject brain activation patterns. *Frontiers in Neuroinformatics*, 18:1399391, Aug. 2024. doi: 10.3389/fninf.2024.1399391. Published Aug 12 2024.
- [76] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006. ISBN 3540254374.

- [77] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis: proceedings of the biennial Conference held at Dundee, June 28–July 1, 1977*, pages 105–116. Springer, 2006.
- [78] D. D. Morrison. Methods for nonlinear least squares problems and convergence proofs. In J. Lorell and F. Yagi, editors, *Proceedings of the Seminar on Tracking Programs and Orbit Determination*, pages 1–9, Pasadena, USA, 1960. Jet Propulsion Laboratory.
- [79] J. M. Myre, E. Frahm, D. J. Lilja, and M. O. Saar. Tnt-nn: A fast active set method for solving large non-negative least squares problems. In *Procedia Computer Science*, volume 108, pages 755–764, 2017.
- [80] M. Mørup, K. H. Madsen, and L. K. Hansen. Decomposing neuroimaging data sets with parallel factor analysis. *NeuroImage*, 33(3):1094–1109, 2006.
- [81] Y. Nesterov. Modified gauss–newton scheme with worst case guarantees for global performance. *Optimisation methods and software*, 22(3):469–483, 2007.
- [82] Y. Nesterov. *Lectures on convex optimization*. Springer, 2nd edition, 2018.
- [83] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.
- [84] B. T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.
- [85] R. Pourbagher, S. Y. Derakhshandeh, and M. E. H. Golshan. Application of high-order levenberg–marquardt method for solving the power flow problem in ill-conditioned systems. *IET Generation, Transmission & Distribution*, 10(12):3017–3022, 2016.
- [86] R. G. Pratt, C. Shin, and G. J. Hicks. Gauss–newton and full newton methods in frequency–space seismic waveform inversion. *Geophysical Journal International*, 133(2):341–362, 1998. doi: 10.1046/j.1365-246X.1998.00558.x.
- [87] J. Pujol. The solution of nonlinear inverse problems and the levenberg–marquardt method. *Geophysics*, 72(4):W1–W16, 2007.
- [88] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.
- [89] K. Schacke. On the kronecker product. *Master’s thesis, University of Waterloo*, 2004.
- [90] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis. Parallel factor analysis in sensor array processing. *IEEE Transactions on Signal Processing*, 48(8):2377–2388, 2000.
- [91] M. Soltanolkotabi, D. Stöger, and C. Xie. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. *IEEE Transactions on Information Theory*, 2025.
- [92] D. Stöger and M. Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

- [93] A. Szlam, Z. Guo, and S. Osher. A split bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing. In *2010 IEEE International Conference on Image Processing*, pages 1917–1920. IEEE, 2010.
- [94] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22:1–63, 2021.
- [95] T. Tong, C. Ma, and Y. Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. In *2021 IEEE Data Science and Learning Workshop (DSLW)*, pages 1–6, 2021. doi: 10.1109/DSLW51110.2021.9523407.
- [96] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned robust low-rank tensor regression. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9072–9076. IEEE, 2022.
- [97] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *Journal of Machine Learning Research*, 23(163):1–77, 2022.
- [98] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.
- [99] J. S. Wind. Asymmetric matrix sensing by gradient descent with small random initialization. *arXiv preprint arXiv:2309.01796*, 2023.
- [100] J. Wright and Y. Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.
- [101] S. J. Wright. Convergence of an inexact algorithm for composite nonsmooth optimization. *IMA Journal of Numerical Analysis*, 10(3):299–321, 1990.
- [102] T. Wu. Guaranteed nonconvex low-rank tensor estimation via scaled gradient descent. *arXiv preprint arXiv:2501.01696*, 2025.
- [103] N. Xiong, L. Ding, and S. S. Du. How over-parameterization slows down gradient descent in matrix sensing: The curses of symmetry and initialization. *arXiv preprint arXiv:2310.01769*, 2023.
- [104] X. Xu, Y. Shen, Y. Chi, and C. Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.
- [105] T. Ye and S. S. Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.
- [106] G. Zhang, S. Fattahi, and R. Y. Zhang. Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification. *Journal of Machine Learning Research*, 24:1–55, 2023.
- [107] H. Zhang, Y. Bi, and J. Lavaei. General low-rank matrix optimization: Geometric analysis and sharper bounds. *Advances in Neural Information Processing Systems*, 34:27369–27380, 2021.

- [108] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.
- [109] J. Zhuo, J. Kwon, N. Ho, and C. Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *Journal of Machine Learning Research*, 25(169):1–47, 2024.

A Missing proofs from Section 3

A.1 Proof of Lemma 3.1

First, we show the bound on the norm of P_k and $I - P_k$. Set $d = \dim(\mathbf{E})$, $m = \dim(\mathbf{Y})$, and $r = \text{rank}(\nabla F(x_k))$. Let $\nabla F(x_k) = U\Sigma V^\top$ be the economy SVD of matrix $\nabla F(x_k)$, where $U \in O(m, r)$, $V \in O(d, r)$, $\Sigma = \text{diag}(\sigma)$, and $\sigma = \sigma(\nabla F(x_k))$, for notational convenience we do not index these matrices with x_k . Then,

$$P_k = U\Sigma(\Sigma^\top \Sigma + \lambda_k I)^{-1} \Sigma^\top U^\top.$$

Let $w = \left(\frac{(\sigma_1^{x_k})^2}{(\sigma_1^{x_k})^2 + \lambda_k}, \dots, \frac{(\sigma_r^{x_k})^2}{(\sigma_r^{x_k})^2 + \lambda_k} \right)$. We know from linear algebra that $P_k = U \text{diag}(w) U^\top$. The eigenvalues of P_k are bounded by one since $\lambda_k > 0$, so $\|P_k\|_{\text{op}} \leq 1$. Similarly, let $v = \left(\frac{\lambda_k}{(\sigma_1^{x_k})^2 + \lambda_k}, \dots, \frac{\lambda_k}{(\sigma_r^{x_k})^2 + \lambda_k} \right)$ we can write

$$I - P_k = U \text{diag}(v) U^\top. \quad (30)$$

It's clear that $\|I - P_k\|_{\text{op}} \leq 1$. Moreover, for any $v \in Y$, we have

$$\|P_k v\| = \left\| U \text{diag}(w) U^\top v \right\| \leq \left\| U^\top v \right\| = \|\Pi^{x_k} v\|,$$

where the last equality follows since $\Pi^{x_k} = U U^\top$. Therefore, the first item holds.

Next, we note that the second item holds immediately from (30) and the monotonicity of singular values $\{\sigma_i^{x_k}\}_{i=1}^r$. Lastly, observe that $\gamma_k P_k v_k = \nabla F(x_k)(x_k - x_{k+1})$, thus

$$\begin{aligned} \|z_{k+1} - (z_k - \gamma_k P_k v_k)\| &= \|F(x_{k+1}) - F(x_k) - \nabla F(x_k)(x_{k+1} - x_k)\| \\ &\leq \frac{L_{\nabla F}}{2} \|x_{k+1} - x_k\|^2 \\ &= \frac{L_{\nabla F}}{2} \gamma_k^2 \left\| (\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I)^{-1} \nabla F(x_k)^\top v_k \right\|^2, \end{aligned} \quad (31)$$

where the inequality follows from Taylor's theorem. Just as before, we have that

$$(\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I)^{-1} \nabla F(x_k)^\top = V(\Sigma^\top \Sigma + \lambda_k I)^{-1} \Sigma^\top U^\top.$$

Once again, the nonzero singular values of this matrix correspond to $\frac{\sigma_i}{\sigma_i^2 + \lambda_k}$. By Young's inequality, we have $\sigma_i^2 + \lambda_k \geq 2\sigma_i \sqrt{\lambda_k}$, so $\frac{\sigma_i}{\sigma_i^2 + \lambda_k} \leq \frac{1}{2\sqrt{\lambda_k}}$, which implies

$$\left\| (\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I)^{-1} \nabla F(x_k)^\top v_k \right\| \leq \frac{1}{2\sqrt{\lambda_k}} \left\| U^\top v_k \right\| = \frac{1}{2\sqrt{\lambda_k}} \|\Pi^{x_k} v_k\|. \quad (32)$$

The conclusion follows directly from the estimates (31) and (32). This concludes the proof of Lemma 3.1. \square

B Missing proofs from Section 4

B.1 Proof of Theorem 4.1 (Geometric decaying stepsizes)

We start by establishing a couple of auxiliary lemmas.

Lemma B.1. *Suppose that Assumptions 1 and 4 hold. For any x_k, x_{k+1} generated by Algorithm 1, let $z_k = F(x_k)$ and $z_{k+1} = F(x_{k+1})$. Then we have*

$$\|z_{k+1} - (z_k - \gamma_k P_k v_k)\| \leq \frac{L_{\nabla F} L^2 \gamma_k^2}{8\lambda_k}.$$

Proof. A combination of Lemma 3.1 and Assumption 4 yields the desired bound. \square

Lemma B.2. *Suppose that Assumptions 1, 3, and 4 hold. For any z_k such that $\|z_k - z^*\| \leq \delta \left(\frac{\mu}{8L}\right)$, we have*

$$\begin{aligned} \|z_k - \gamma_k P_k v_k - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{3\mu\gamma_k}{2} \|z_k - z^*\| \\ &\quad + 2L\gamma_k \frac{\lambda_k}{s\left(\frac{\mu}{8L}\right) \|z_k - z^*\| + \lambda_k} \|z_k - z^*\| + \gamma_k^2 L^2. \end{aligned}$$

Proof. Note that

$$\begin{aligned} \|z_k - \gamma_k P_k v_k - z^*\|^2 &= \|z_k - z^*\|^2 - 2\gamma_k \langle v_k, z_k - z^* \rangle + 2\gamma_k \langle (I - P_k)v_k, z_k - z^* \rangle + \gamma_k^2 \|P_k v_k\|^2 \\ &\leq \|z_k - z^*\|^2 - 2\gamma_k \mu \|z_k - z^*\| + 2\gamma_k \langle v_k, (I - P_k)(z_k - z^*) \rangle + \gamma_k^2 L^2, \end{aligned} \quad (33)$$

where the inequality follows from Lemma 3.1, Lemma 3.2, and Item 4 of Assumption 4. On the other hand, by the same argument as in (17) and (18), we have

$$\|(I - P_k)(z_k - z^*)\| \leq \left(\frac{\lambda_k}{s\left(\frac{\mu}{8L}\right) \|z_k - z^*\| + \lambda_k} + \frac{\mu}{8L} \right) \|z_k - z^*\|.$$

By Item 4 of Assumption 4, we have

$$2\gamma_k \langle v_k, (I - P_k)(z_k - z^*) \rangle \leq 2L\gamma_k \left(\frac{\lambda_k}{s\left(\frac{\mu}{8L}\right) \|z_k - z^*\| + \lambda_k} + \frac{\mu}{8L} \right) \|z_k - z^*\|. \quad (34)$$

The desired inequality follows from a combination of (33) and (34). \square

We prove the theorem by induction. The conclusion holds for $k = 0$ by assumption. Next, suppose that the conclusion holds for some $k \geq 0$. We consider two cases:

Case 1. Suppose first that $\|z_k - z^*\| \leq \frac{M}{4} q^k$. We have

$$\begin{aligned} \|z_k - \gamma_k P_k v_k - z^*\|^2 &\leq \|z_k - z^*\|^2 + 2L\gamma_k \frac{\lambda_k}{s\left(\frac{\mu}{8L}\right) \|z_k - z^*\| + \lambda_k} \|z_k - z^*\| + \gamma_k^2 L^2 \\ &\leq \|z_k - z^*\|^2 + 2L\gamma_k \|z_k - z^*\| + \gamma_k^2 L^2 \\ &\leq \left(\frac{M^2}{16} + \frac{\gamma LM}{2} + \gamma^2 L^2 \right) q^{2k} \\ &\leq \frac{M^2}{4} q^{2k+2}, \end{aligned} \quad (35)$$

where the first inequality follows from Lemma B.2, the second inequality follows from the fact that

$$\frac{\lambda_k}{s\left(\frac{\mu}{8L}\right) \|z_k - z^*\| + \lambda_k} \leq 1,$$

the third inequality follows from the assumption that $\|z_k - z^*\| \leq \frac{M}{4}q^k$, and the last inequality follows from $q \geq \frac{1}{\sqrt{2}}$ and our assumption on $\gamma \leq M\mu/(64L^2) \leq M/(64L)$. As a result, $\|z_k - \gamma_k P_k v_k - z^*\| \leq \frac{M}{2}q^{k+1}$. Moreover, by Lemma B.1, the triangle inequality, and our assumption $\gamma^2 \leq 2\lambda M/(L_{\nabla F}L^2)$, we have

$$\begin{aligned} \|z_{k+1} - z^*\| &\leq \frac{M}{2}q^{k+1} + \frac{\gamma^2 L_{\nabla F} L^2}{8\lambda} q^k \\ &\leq Mq^{k+1}. \end{aligned}$$

Case 2. Now, suppose $\frac{M}{4}q^k \leq \|z_k - z^*\| \leq Mq^k$. We have

$$\begin{aligned} &\|z_k - \gamma_k P_k v_k - z^*\|^2 \\ &\leq \|z_k - z^*\|^2 - \frac{3\gamma\mu}{2}q^k \|z_k - z^*\| + 2\gamma L \frac{\lambda q^k}{s(\frac{\mu}{8L}) \|z_k - z^*\| + \lambda q^k} q^k \|z_k - z^*\| + \gamma^2 L^2 q^{2k} \\ &\leq \|z_k - z^*\|^2 - \frac{3\gamma\mu}{2M} \|z_k - z^*\|^2 + 2\gamma L \frac{4\lambda}{s(\frac{\mu}{8L}) M + 4\lambda} q^k \|z_k - z^*\| + \frac{16\gamma^2 L^2}{M^2} \|z_k - z^*\|^2 \\ &\leq \|z_k - z^*\|^2 - \frac{3\gamma\mu}{2M} \|z_k - z^*\|^2 + 2\gamma L \frac{16\lambda}{s(\frac{\mu}{8L}) M^2} \|z_k - z^*\|^2 + \frac{16\gamma^2 L^2}{M^2} \|z_k - z^*\|^2 \\ &\leq \left(1 - \frac{\gamma\mu}{M}\right) \|z_k - z^*\|^2, \end{aligned}$$

where the first inequality follows from Lemma B.2, the second and third inequalities follow from the assumed bound on $\|z_k - z^*\|$, and the last inequality follows from our assumption on λ and γ . Taking the square root of both sides, we have

$$\|z_k - \gamma_k P_k v_k - z^*\| \leq \sqrt{1 - \frac{\gamma\mu}{M}} \|z_k - z^*\| \leq \left(1 - \frac{\gamma\mu}{2M}\right) \|z_k - z^*\| \quad (36)$$

where the second inequality follows since $(1-x)^{1/2} \leq 1-x/2$ for all $x \leq 1$, which holds due to our constraints on γ . Then,

$$\begin{aligned} \|z_{k+1} - z^*\| &\leq \|z_k - \gamma_k P_k v_k - z^*\| + \|z_{k+1} - (z_k - \gamma_k P_k v_k)\| \\ &\leq \left(1 - \frac{\gamma\mu}{2M}\right) \|z_k - z^*\| + \frac{\gamma^2 L_{\nabla F} L^2}{8\lambda} q^k \\ &\leq \left(1 - \frac{\gamma\mu}{4M}\right) \|z_k - z^*\| \\ &\leq Mq^{k+1}, \end{aligned}$$

where the first inequality follows from Lemma B.1 and the triangle inequality, the second inequality follows from (36), Lemma B.1, and our assumption $\gamma \leq \lambda\mu/(2L_{\nabla F}L^2)$, and the last inequality follows from the inductive hypothesis and the fact that $1 - \gamma\mu/(4M) \leq q$.

The induction is complete, finishing the proof of Theorem 4.1. \square

B.2 Proof of Theorem 4.4

The proof of the following lemma is essentially the same as that of Lemma B.2. We omit the details.

Lemma B.3. *Suppose that Assumptions 1, 2, and 4 hold. For any z_k such that $\|z_k - z^*\| \leq \delta(\frac{\mu}{8L})$,*

we have

$$\|z_k - \gamma_k P_k v_k - z^*\|^2 \leq \|z_k - z^*\|^2 - \frac{3\mu\gamma_k}{2} \|z_k - z^*\| + 2L\gamma_k \frac{\lambda_k}{s + \lambda_k} \|z_k - z^*\| + \gamma_k^2 L^2.$$

Proof for Polyak stepsize. By induction, it suffices to prove the following claim:

Claim B.4. For any $z_k = F(x_k)$ with $\|z_k - z^*\| \leq \min\left\{\delta\left(\frac{\mu}{8L}\right), \frac{s\mu}{8C_{\text{ub}}L}\right\}$, we have

$$\|z_{k+1} - z^*\|^2 \leq \left(1 - \frac{\gamma\mu^2}{8L^2}\right) \|z_k - z^*\|^2.$$

To this end, we let j be the index provided by Assumption 2 when applied to $\rho = \frac{\mu}{8L}$ and $z = z_k = F(x_k)$, i.e.,

$$\|(I - \Pi_j^x)(z_k - z^*)\| \leq \frac{\mu}{8L} \|z_k - z^*\| \quad \text{and} \quad (\sigma_j^x)^2 \geq s.$$

Following the similar calculation as in (18), we have

$$\begin{aligned} |\langle (I - P_k)v_k, z_k - z^* \rangle| &\leq L \left(\|(I - P_k)\Pi_j^x(z_k - z^*)\| + \|(I - P_k)(I - \Pi_j^x)(z_k - z^*)\| \right) \\ &\leq L \left(\frac{\lambda_k}{(\sigma_j^x)^2 + \lambda_k} \|\Pi_j^x(z_k - z^*)\| + \frac{\mu}{8L} \|z_k - z^*\| \right) \\ &\leq L \left(\frac{C_{\text{ub}} \|z_k - z^*\|}{s + C_{\text{ub}} \|z_k - z^*\|} + \frac{\mu}{8L} \right) \|z_k - z^*\| \\ &\leq \frac{\mu}{4} \|z_k - z^*\| \\ &\leq \frac{h(z_k) - h^*}{4}, \end{aligned}$$

where the fourth inequality follows from the bound $\|z_k - z^*\| \leq \frac{s\mu}{8C_{\text{ub}}L}$. Invoking Proposition 4.2 and Assumption 4 gives

$$\begin{aligned} \|z_{k+1} - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{\gamma(h(z_k) - h^*)^2}{8 \|\Pi^{x_k} v_k\|^2} \\ &\leq \left(1 - \frac{\gamma\mu^2}{8L^2}\right) \|z_k - z^*\|^2, \end{aligned}$$

as desired.

Proof for geometrically decaying stepsize. We prove it by induction. First, the conclusion holds for $k = 0$. Now suppose that the conclusion holds for some $k \geq 0$. We consider two cases:

Case 1. Suppose $\|z_k - z^*\| \leq \frac{M}{4}q^k$. Using Lemma B.3 and the same argument in (35), we obtain $\|z_{k+1} - z^*\| \leq Mq^{k+1}$.

Case 2. Suppose $\frac{M}{4}q^k \leq \|z_k - z^*\| \leq Mq^k$. We have

$$\begin{aligned} \|z_k - \gamma_k P_k v_k - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{3\mu\gamma_k}{2} \|z_k - z^*\| + 2L\gamma_k \frac{\lambda_k}{s + \lambda_k} \|z_k - z^*\| + \gamma_k^2 L^2 \\ &\leq \|z_k - z^*\|^2 - \frac{3\gamma\mu}{2M} \|z_k - z^*\|^2 + \frac{8\gamma\lambda L}{sM} \|z_k - z^*\|^2 + \frac{16\gamma^2 L^2}{M^2} \|z_k - z^*\|^2 \\ &\leq \left(1 - \frac{\gamma\mu}{M}\right) \|z_k - z^*\|^2, \end{aligned}$$

where the first inequality follows from Lemma B.3, the second inequality follows from the lower bound on $\|z_k - z^*\|_2$, and the third inequality follows our bounds on λ and γ . The rest of the proof follows from the same argument as the proof of Theorem 4.1.

This completes the proof of Theorem 4.4. \square

B.3 Proof of Theorem 4.5

We start by stating a couple of auxiliary lemmas.

Lemma B.5. *Suppose Assumptions 1 and 5 hold, and let x_k and x_{k+1} be iterates generated by Algorithm 1 under Configuration 3. Define $z_k = F(x_k)$ and $z_{k+1} = F(x_{k+1})$. Then, we have*

$$\|z_{k+1} - (z_k - \gamma P_k \nabla h(z_k))\| \leq \frac{L_{\nabla F} \gamma^2}{8\lambda q^k} \|\Pi^{x_k} \nabla h(z_k)\|^2.$$

Proof. A combination of Lemma 3.1 and the choice of λ_k and γ_k in Configuration 3 yields the desired result. \square

Lemma B.6. *Suppose that Assumptions 1, 3, and 5 hold. Assume that we are under Configuration 3 and that $\gamma \leq \frac{1}{8\beta}$. For any z_k such that $\|z_k - z^*\| \leq \delta \left(\frac{\alpha}{16\beta} \right)$, we have*

$$\|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 \leq \|z_k - z^*\|^2 - \frac{7\gamma}{4} \langle \nabla h(z_k), z_k - z^* \rangle + 2\beta\gamma \frac{\lambda_k}{s \left(\frac{\alpha}{16\beta} \right) \|z_k - z^*\| + \lambda_k} \|z_k - z^*\|^2.$$

Proof. By expanding the square and adding and subtracting $2\gamma \langle \nabla h(z_k), z_k - z^* \rangle$ we get

$$\begin{aligned} & \|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 \\ & \leq \|z_k - z^*\|^2 - 2\gamma \langle \nabla h(z_k), z_k - z^* \rangle + 2\gamma \langle (I - P_k) \nabla h(z_k), z_k - z^* \rangle + \gamma^2 \|\Pi^{x_k} \nabla h(z_k)\|^2 \\ & \leq \|z_k - z^*\|^2 - \frac{7\gamma}{4} \langle \nabla h(z_k), z_k - z^* \rangle + 2\gamma \langle (I - P_k) \nabla h(z_k), z_k - z^* \rangle, \end{aligned}$$

where the first inequality follows from Lemma 3.1, and the second inequality follows from Lemma 3.2, Item 4 of Assumption 5, and $\gamma \leq \frac{1}{8\beta}$. We focus on bounding the inner product in the last term

$$\begin{aligned} |\langle (I - P_k) \nabla h(z_k), z_k - z^* \rangle| & \leq \beta \|z_k - z^*\| \|(I - P_k)(z_k - z^*)\| \\ & \leq \beta \left(\frac{\lambda_k}{s \left(\frac{\alpha}{16\beta} \right) \|z_k - z^*\| + \lambda_k} + \frac{\alpha}{16\beta} \right) \|z_k - z^*\|^2, \end{aligned}$$

where the first inequality follows from Item 4 of Assumption 5 and last inequality follows from the same calculation as (18) with v_k replaced by $\nabla h(z_k)$. This concludes the proof of the lemma. \square

Proof for Polyak stepsize. By induction, it suffices to prove the following claim.

Claim B.7. *For any $z_k = F(x_k)$ with $\|z_k - z^*\|_2 \leq \delta \left(\frac{\alpha}{16\beta} \right)$, we have*

$$\|z_{k+1} - z^*\|^2 \leq \left(1 - \frac{\gamma\alpha}{32\beta} \right) \|z_k - z^*\|^2.$$

Let j be the index provided by Assumption 3 when applied to $\rho = \frac{\alpha}{16\beta}$ and $z = z_k = F(x_k)$, i.e.,

$$\|(I - \Pi_j^{x_k})(z_k - z^*)\| \leq \frac{\alpha}{16\beta} \|z_k - z^*\| \quad \text{and} \quad (\sigma_i^x)^2 \geq s \left(\frac{\alpha}{16\beta} \right) \|z_k - z^*\|. \quad (37)$$

Thus, we have

$$\begin{aligned}
|\langle (I - P_k)\nabla h(z_k), z_k - z^* \rangle| &\leq \beta \left(\frac{C_{\text{ub}}}{s\left(\frac{\alpha}{16\beta}\right) + C_{\text{ub}}} + \frac{\alpha}{16\beta} \right) \|z_k - z^*\|^2 \\
&\leq \beta \frac{2\alpha}{16\beta} \|z_k - z^*\|^2 \\
&\leq \frac{1}{4} (h(z_k) - h^*),
\end{aligned} \tag{38}$$

where the first inequality follows from the same calculation as (18) and the second inequality is due to $C_{\text{ub}} \leq \frac{\alpha}{16\beta} s\left(\frac{\alpha}{16\beta}\right)$. Applying Proposition 4.2 and Assumption 5, we have

$$\begin{aligned}
\|z_{k+1} - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{\gamma}{8} \frac{(h(z_k) - h^*)^2}{\|\Pi^{x_k}\nabla h(z_k)\|^2} \\
&\leq \|z_k - z^*\|^2 - \frac{\gamma}{8} \frac{h(z_k) - h^*}{\|z_k - z^*\|^2} \frac{h(z_k) - h^*}{\|\Pi^{x_k}\nabla h(z_k)\|^2} \|z_k - z^*\|^2 \\
&\leq \left(1 - \frac{\gamma\alpha}{32\beta}\right) \|z_k - z^*\|^2,
\end{aligned}$$

concluding the proof for the Polyak stepsize.

Proof for geometrically decaying stepsize. We prove the rate by induction. Based on our assumption, the conclusion holds for $k = 0$. Now suppose that the conclusion holds for some $k \geq 0$. We consider two cases:

Case 1. $\|z_k - z^*\| \leq \frac{M}{4}q^k$. We have

$$\begin{aligned}
\|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 &\leq \|z_k - z^*\|^2 + 2\beta\gamma \frac{\lambda_k}{s\left(\frac{\alpha}{16\beta}\right) \|z_k - z^*\| + \lambda_k} \|z_k - z^*\|^2 \\
&\leq \|z_k - z^*\|^2 + 2\beta\gamma \|z_k - z^*\|^2 \\
&\leq \left(\frac{M^2}{16} + \frac{\beta\gamma M^2}{8}\right) q^{2k} \\
&\leq \frac{M^2}{4} q^{2k+2},
\end{aligned}$$

where the first inequality follows from Lemma B.6, the second inequality follows from

$$\frac{\lambda_k}{s\left(\frac{\alpha}{16\beta}\right) \|z_k - z^*\| + \lambda_k} \leq 1,$$

the third inequality follows from $\|z_k - z^*\| \leq \frac{M}{4}q^k$, and the last inequality follows from $q \geq \frac{1}{\sqrt{2}}$ and our assumption on γ . As a result, $\|z_k - \gamma P_k \nabla h(z_k) - z^*\| \leq \frac{M}{2}q^{k+1}$. Moreover, by Lemma B.5, Item 4 of Assumption 5, and our bound on γ , we derive

$$\begin{aligned}
\|z_{k+1} - z^*\| &\leq \|z_k - \gamma P_k \nabla h(z_k) - z^*\| + \frac{L_{\nabla F}\gamma^2}{8\lambda q^k} \|\Pi^{x_k}\nabla h(z_k)\|^2 \\
&\leq \frac{M}{2}q^{k+1} + \frac{\beta^2 L_{\nabla F}\gamma^2 M^2}{128\lambda} q^k \\
&\leq Mq^{k+1}.
\end{aligned}$$

Case 2. $\frac{M}{4}q^k \leq \|z_k - z^*\| \leq Mq^k$. We have

$$\begin{aligned} \|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{7\gamma}{4} \langle \nabla h(z_k), z_k - z^* \rangle + 2\beta\gamma \frac{\lambda q^k \|z_k - z^*\|^2}{s \left(\frac{\alpha}{16\beta}\right) \|z_k - z^*\| + \lambda q^k} \\ &\leq \|z_k - z^*\|^2 - \frac{7\gamma}{4} \langle \nabla h(z_k), z_k - z^* \rangle + 2\beta\gamma \frac{4\lambda}{s \left(\frac{\alpha}{16\beta}\right) M} \|z_k - z^*\|^2 \\ &\leq \|z_k - z^*\|^2 - \frac{3\gamma}{2} \langle \nabla h(z_k), z_k - z^* \rangle, \end{aligned}$$

where the first inequality follows from Lemma B.6, the second inequality follows from the assumed bound on $\|z_k - z^*\|$, and the last inequality follows from the bound on λ and Lemma 3.3. As a result of Lemma B.5 and triangle inequality, we have

$$\begin{aligned} \|z_{k+1} - z^*\|^2 &\leq \left(\|z_k - \gamma P_k \nabla h(z_k) - z^*\| + \frac{L_{\nabla F} \gamma^2}{8\lambda q^k} \|\Pi^{x_k} \nabla h(z_k)\|^2 \right)^2 \\ &\leq \left(\sqrt{\|z_k - z^*\|^2 - \frac{3\gamma}{2} \langle \nabla h(z_k), z_k - z^* \rangle} + \frac{L_{\nabla F} \gamma^2}{8\lambda q^k} \|\Pi^{x_k} \nabla h(z_k)\|^2 \right)^2 \\ &= \|z_k - z^*\|^2 - \frac{3\gamma}{2} \langle \nabla h(z_k), z_k - z^* \rangle + \frac{L_{\nabla F}^2 \gamma^4}{64\lambda^2 q^{2k}} \|\Pi^{x_k} \nabla h(z_k)\|^4 \\ &\quad + \frac{L_{\nabla F} \gamma^2}{4\lambda q^k} \|z_k - z^*\|_2 \|\Pi^{x_k} \nabla h(z_k)\|^2. \end{aligned}$$

Note that by Lemma 3.2 and Item 4 of Assumption 5, we have

$$\|\Pi^{x_k} \nabla h(z_k)\|^2 \leq 2\beta \langle \nabla h(z_k), z_k - z^* \rangle.$$

Combining with the upper bound on $\|z_k - z^*\|$ and our assumption on γ , we have

$$\begin{aligned} \|z_{k+1} - z^*\|^2 &\leq \|z_k - z^*\|^2 - \gamma \langle \nabla h(z_k), z_k - z^* \rangle \\ &\leq \left(1 - \frac{\gamma\alpha}{2}\right) \|z_k - z^*\|^2 \\ &\leq M^2 q^{2k+2}. \end{aligned}$$

The induction is complete, and so is the proof of Theorem 4.5. \square

B.4 Proof of Theorem 4.6

We start with an auxiliary result.

Lemma B.8. *Suppose that Assumptions 1, 2, and 5 hold. Suppose that we are under Configuration 3 and that $\gamma \leq \frac{1}{8\beta}$. For any z_k such that $\|z_k - z^*\| \leq \delta \left(\frac{\alpha}{16\beta}\right)$, we have*

$$\|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 \leq \|z_k - z^*\|^2 - \frac{7\gamma}{4} \langle \nabla h(z_k), z_k - z^* \rangle + 2\beta\gamma \frac{\lambda_k}{s + \lambda_k} \|z_k - z^*\|^2.$$

Proof. The proof follows the same argument as the proof of Lemma B.6. \square

Proof for Polyak stepsize. By induction, it suffices to prove the following claim.

Claim B.9. *For any $z_k = F(x_k)$ with $\|z_k - z^*\| \leq \left\{r \left(\frac{\alpha}{16\beta}\right), \frac{s\alpha}{16C_{ub}\beta}\right\}$, we have*

$$\|z_{k+1} - z^*\|^2 \leq \left(1 - \frac{\gamma\alpha}{32\beta}\right) \|z_k - z^*\|^2.$$

Let j be the index provided by Assumption 2 when applied to $\rho = \frac{\alpha}{16\beta}$ and $z = z_k = F(x_k)$, i.e.,

$$\left\| (I - \Pi_j^x)(z_k - z^*) \right\| \leq \frac{\alpha}{16\beta} \|z_k - z^*\| \quad \text{and} \quad (\sigma_j^x)^2 \geq s.$$

Following the similar calculation as (18) and (38), we have

$$\begin{aligned} |\langle (I - P_k)\nabla h(z_k), z_k - z^* \rangle| &\leq \beta \|z_k - z^*\| \left(\left\| (I - P_k)\Pi_j^x(z_k - z^*) \right\| + \left\| (I - P_k)(I - \Pi_j^x)(z_k - z^*) \right\| \right) \\ &\leq \beta \|z_k - z^*\| \left(\frac{\lambda_k}{(\sigma_j^x)^2 + \lambda_k} \left\| \Pi_j^x(z_k - z^*) \right\| + \frac{\alpha}{16\beta} \|z_k - z^*\| \right) \\ &\leq \beta \left(\frac{C_{\text{ub}} \|z_k - z^*\|}{s + C_{\text{ub}} \|z_k - z^*\|} + \frac{\alpha}{16\beta} \right) \|z_k - z^*\|^2 \\ &\leq \beta \left(\frac{1}{\frac{16\beta}{\alpha} + 1} + \frac{\alpha}{16\beta} \right) \|z_k - z^*\|^2 \\ &\leq \frac{\alpha}{8} \|z_k - z^*\|^2 \\ &\leq \frac{h(z_k) - h^*}{4}, \end{aligned}$$

where the fourth inequality follows from the bound $\|z_k - z^*\| \leq \frac{s\alpha}{16C_{\text{ub}}\beta}$ and for $\frac{1}{1+x^{-1}} + x \leq 2x$ for any $x \geq 0$. Applying Proposition 4.2 and Assumption 5, we get

$$\begin{aligned} \|z_{k+1} - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{\gamma}{8} \frac{(h(z_k) - h^*)^2}{\|(U^{x_k})^\top \nabla h(z_k)\|^2} \\ &\leq \left(1 - \frac{\gamma\alpha}{32\beta} \right) \|z_k - z^*\|^2, \end{aligned}$$

proving the result for the Polyak stepsize.

Proof for geometrically decaying stepsize. We prove the theorem by induction. Based on our assumption, the conclusion holds for $k = 0$. Now suppose that the conclusion holds for some $k \geq 0$. We consider two cases:

Case 1. $\|z_k - z^*\| \leq \frac{M}{4}q^k$. We have

$$\begin{aligned} \|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 &\leq \|z_k - z^*\|^2 + 2\beta\gamma \frac{\lambda_k}{s + \lambda_k} \|z_k - z^*\|^2 \\ &\leq \|z_k - z^*\|^2 + 2\beta\gamma \|z_k - z^*\|^2 \\ &\leq \left(\frac{M^2}{16} + \frac{\beta\gamma M^2}{8} \right) q^{2k} \\ &\leq \frac{M^2}{4} q^{2k+2}, \end{aligned}$$

where the first inequality follows from Lemma B.8, the second inequality follows from $\frac{\lambda_k}{s + \lambda_k} \leq 1$, the third inequality follows from the assumption that $\|z_k - z^*\| \leq \frac{M}{4}q^k$, and the last inequality follows from $q \geq \frac{1}{\sqrt{2}}$ and the bound on γ . As a result, $\|z_k - \gamma P_k \nabla h(z_k) - z^*\| \leq \frac{M}{2}q^{k+1}$. Moreover, by

Lemma B.5, Item 4 of Assumption 5, and our assumption on γ , we have

$$\begin{aligned} \|z_{k+1} - z^*\| &\leq \|z_k - \gamma P_k \nabla h(z_k) - z^*\| + \frac{L_{\nabla F} \gamma^2}{8\lambda q^k} \|\Pi^{x_k} \nabla h(z_k)\|^2 \\ &\leq \frac{M}{2} q^{k+1} + \frac{\beta^2 L_{\nabla F} \gamma^2 M^2}{128\lambda} q^k \\ &\leq M q^{k+1}. \end{aligned}$$

Case 2. $\frac{M}{4} q^k \leq \|z_k - z^*\|_2 \leq M q^k$. We have

$$\begin{aligned} \|z_k - \gamma P_k \nabla h(z_k) - z^*\|^2 &\leq \|z_k - z^*\|^2 - \frac{7\gamma}{4} \langle \nabla h(z_k), z_k - z^* \rangle + \frac{2\beta\gamma\lambda}{s} \|z_k - z^*\|^2 \\ &\leq \|z_k - z^*\|^2 - \frac{3\gamma}{2} \langle \nabla h(z_k), z_k - z^* \rangle, \end{aligned}$$

where the first inequality follows from Lemma B.8 and $q \leq 1$, and the last inequality follows from $\lambda \leq \frac{s\alpha}{16\beta}$ and Lemma 3.3. The rest of the proof follows the same as the proof of Theorem 4.5.

The induction is complete, and so is the proof of Theorem 4.6. \square

B.5 Additional results and proofs from Section 4.3

In this section, we prove Theorem 4.7 and state additional local guarantees we omitted in Section 4.3.

B.5.1 Proof of Theorem 4.7

The proofs of our local guarantees rely on the following two auxiliary results.

Lemma B.10. *Let x_k and x_{k+1} be iterates of Algorithm 1 under Configuration 1 and write $z_k = F(x_k)$ and $z_{k+1} = F(x_{k+1})$. Suppose that $|\langle (I - P_k)v_k, z_k - z^* \rangle| \leq \frac{1}{4}(h(z_k) - h^*)$ holds. Then, we have*

$$\|x_k - x_{k+1}\| \leq \frac{2\gamma \|z_k - z^*\|^{1/2}}{3 \sqrt{C_{1b}}}.$$

Proof. Recall from Claim 4.3 that

$$\frac{3}{4}(h(z_k) - h^*) \leq \|\Pi^{x_k} v_k\| \|z_k - z^*\|. \quad (39)$$

By the definition of Algorithm 1, we have

$$\begin{aligned} \|x_k - x_{k+1}\| &= \gamma_k \left\| (\nabla F(x_k)^\top \nabla F(x_k) + \lambda_k I)^{-1} \nabla F(x_k)^\top v_k \right\| \\ &\leq \frac{\gamma(h(z_k) - h^*)}{2\sqrt{\lambda_k} \|\Pi^{x_k} v_k\|} \\ &\leq \frac{2\gamma \|z_k - z^*\|^{1/2}}{3 \sqrt{C_{1b}}}, \end{aligned}$$

where the second line follows from (32) and the last line follows from (39) together with the lower bound on $\lambda_k \geq C_{1b} \|z_k - z^*\|$. \square

Proposition B.11. *Let $\{x_k\}_{k \geq 0} \subseteq \mathbf{E}$ and $\{z_k\}_{k \geq 0} \subseteq \mathbf{Y}$ be two sequences and let $x^* \in \mathbf{E}$ and $z^* \in \mathbf{Y}$ be two given points. Suppose that there exist constants $C > 0$, $\varepsilon > 0$, $r > 0$, and $q \in (0, 1)$ be constants such that the following two hold.*

1. For any $k \geq 0$, $\|x_k - x_{k+1}\| \leq C \|z_k - z^*\|^{1/2}$.

2. If $\|x_k - x^*\| \leq \varepsilon$, $\|x_{k+1} - x^*\| \leq \varepsilon$, and $\|z_k - z^*\| \leq r$, then $\|z_{k+1} - z^*\| \leq q\|z_k - z^*\|$.

Then, if $\|x_0 - x^*\|_2 \leq \frac{\varepsilon}{2}$ and $\|z_0 - z^*\|_2 \leq \min \left\{ \left(\frac{(1-q^{1/2})\varepsilon}{2C} \right)^2, r \right\}$ hold, we have

$$\|x_k - x^*\|_2 \leq \varepsilon \quad \text{and} \quad \|z_k - z^*\|_2 \leq \|z_0 - z^*\|_2 q^k, \quad \forall k \geq 0.$$

Proof. We apply induction to prove that for any $k \geq 0$,

$$\|x_k - x^*\|_2 \leq \frac{\varepsilon}{2} + \sum_{i=0}^{k-1} C \|z_0 - z^*\|_2^{1/2} q^{i/2} \leq \varepsilon, \quad \|z_k - z^*\|_2 \leq \|z_0 - z^*\|_2 q^k. \quad (40)$$

The bound (40) holds for $k = 0$ by our assumption. Suppose that (40) holds for k . Note that $\|z_0 - z^*\|_2 \leq r$, we have $\|z_k - z^*\| \leq r$. As a result of Item 1 and the induction hypothesis, we have

$$\begin{aligned} \|x_{k+1} - x^*\|_2 &\leq \|x_k - x^*\|_2 + \|x_k - x_{k+1}\|_2 \\ &\leq \frac{\varepsilon}{2} + \sum_{i=0}^{k-1} C \|z_0 - z^*\|_2^{1/2} q^{i/2} + C \|z_k - z^*\|_2^{1/2} \\ &\leq \frac{\varepsilon}{2} + \sum_{i=0}^k C \|z_0 - z^*\|_2^{1/2} q^{i/2} \\ &\leq \varepsilon. \end{aligned}$$

Additionally, by Item 2, we have

$$\|z_{k+1} - z^*\|_2 \leq q \|z_k - z^*\|_2 \leq \|z_0 - z^*\|_2 q^{k+1}.$$

The induction is complete, and the proof is finished. \square

Armed with these results, we can prove Theorem 4.7.

Proof of Theorem 4.7. Notice that the iterates satisfy the assumption in Lemma B.10 by the same argument we used in (18). With this, we will verify the two conditions required by Proposition B.11. First, notice that Lemma B.10 directly implies Item 1 of Proposition B.11, with $C = \frac{2\gamma}{3\sqrt{C_{1b}}}$. Next, we establish Item 2 using the same arguments from the proof of Theorem 4.1. Note that the derivation of (15) in that proof only relies on the Lipschitz continuity of ∇F along the line segment between x_k and x_{k+1} , rather than requiring the stronger global Lipschitz condition outlined in Assumption 1. Furthermore, due to Assumption 7, the inequalities (16) remain valid whenever $\|x_k - x^*\|_2 \leq \varepsilon_{x^*}$. Thus, if both points x_k and x_{k+1} are in the ball $B_{\varepsilon_{x^*}}(x^*)$, we have

$$\|z_{k+1} - z^*\|_2 \leq \left(1 - \frac{\gamma\mu^2}{8L^2} \right)^{1/2} \|z_k - z^*\|_2.$$

Consequently, Item 2 holds with parameters $r = \delta \left(\frac{\mu}{8L} \right)$ and $q = \left(1 - \frac{\gamma\mu^2}{8L^2} \right)^{1/2}$. Having established both conditions, the theorem follows directly from Proposition B.11. \square

B.5.2 Local convergence guarantees

Next, we present extensions to the other settings we considered.

Assumption 8 (Local strong alignment). *For fixed $x^* \in \mathcal{X}^*$ and $z^* = F(x^*)$ there exist functions $\delta: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and scalars $\varepsilon_{x^*}, s > 0$ such that for all $\rho > 0$, if $x \in \mathbf{B}_{\varepsilon_{x^*}}(x^*)$ and $z = F(x) \in \mathbf{B}_{\delta(\rho)}(z^*)$, then there is an index j for which*

$$\left\| (I - \Pi_j^x)(z - z^*) \right\|_2 \leq \rho \|z - z^*\|_2 \quad \text{and} \quad \left(\sigma_j^x \right)^2 \geq s.$$

We omit the proofs of the following three theorems since they follow an analogous argument to that in the proof of Theorem 4.7, with Claim 15 replaced by Claims B.4, B.7, and B.9, respectively.

Theorem B.12 (Convergence under local strong alignment and nonsmoothness). *Suppose Assumptions 4, 6 and 8 hold. Define $\tilde{q} := \sqrt{1 - \frac{\gamma\mu^2}{8L^2}}$, and let x_0 and $z_0 = F(x_0)$ be points satisfying*

$$\|x_0 - x^*\|_2 \leq \varepsilon/2 \quad \text{and} \quad \|z_0 - z^*\|_2 \leq \min \left\{ \delta \left(\frac{\mu}{8L} \right), \frac{s\mu}{8C_{ub}L}, \frac{(1 - \sqrt{\tilde{q}})^2 \varepsilon^2 C_{lb}}{2\gamma^2} \right\},$$

where $\varepsilon = \min \{\varepsilon_{\nabla F}, \varepsilon_{x^*}\}$. *Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{L_{\nabla F}} \right\}$. Then, the iterates x_k satisfy*

$$\|x_k - x^*\|_2 < \varepsilon \quad \text{for all } k \geq 0,$$

and, moreover, the mapped iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\|^2 \leq \left(1 - \frac{\gamma\mu^2}{8L^2} \right)^k \|z_0 - z^*\|^2 \quad \text{for all } k \geq 0.$$

Theorem B.13 (Convergence under local weak alignment and smoothness). *Suppose Assumptions 5, 6 and 7 hold. Define $\tilde{q} := \sqrt{1 - \frac{\gamma\alpha}{32\beta}}$ and let x_0 and $z_0 = F(x_0)$ be points satisfying*

$$\|x_0 - x^*\|_2 \leq \varepsilon/2 \quad \text{and} \quad \|z_0 - z^*\|_2 \leq \min \left\{ \delta \left(\frac{\alpha}{16\beta} \right), \frac{(1 - \sqrt{\tilde{q}})^2 \varepsilon^2 C_{lb}}{2\gamma^2} \right\},$$

where $\varepsilon = \min \{\varepsilon_{\nabla F}, \varepsilon_{x^*}\}$. *Suppose we ran Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{L_{\nabla F}} \right\}$ and $C_{ub} \leq \frac{\alpha}{16\beta} s \left(\frac{\alpha}{16\beta} \right)$. Then, the iterates x_k must satisfy*

$$\|x_k - x^*\|_2 < \varepsilon \quad \text{for all } k \geq 0$$

and, moreover, the mapped iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\|^2 \leq \left(1 - \frac{\gamma\alpha}{32\beta} \right)^k \|z_0 - z^*\|^2 \quad \text{for all } k \geq 0.$$

Theorem B.14 (Convergence under local strong alignment and smoothness). *Suppose Assumptions 5, 6 and 8 hold. Define $\tilde{q} := \left(1 - \frac{\gamma\alpha}{32\beta} \right)^{1/2}$ and let x_0 and $z_0 = F(x_0)$ be points satisfying*

$$\|x_0 - x^*\|_2 \leq \varepsilon/2 \quad \text{and} \quad \|z_0 - z^*\|_2 \leq \min \left\{ \delta \left(\frac{\alpha}{16\beta} \right), \frac{s\alpha}{16C_{ub}\beta}, \frac{(1 - \sqrt{\tilde{q}})^2 \varepsilon^2 C_{lb}}{2\gamma^2} \right\},$$

where $\varepsilon = \min \{\varepsilon_{\nabla F}, \varepsilon_{x^*}\}$. *If one runs Algorithm 1 initialized at x_0 using Configuration 1 with $\gamma \leq \min \left\{ 1, \frac{C_{lb}}{L_{\nabla F}} \right\}$, then, the iterates x_k must satisfy*

$$\|x_k - x^*\|_2 < \varepsilon \quad \text{for all } k \geq 0,$$

and, moreover, the mapped iterates $z_k = F(x_k)$ satisfy

$$\|z_k - z^*\|^2 \leq \left(1 - \frac{\gamma\alpha}{32\beta} \right)^k \|z_0 - z^*\|^2 \quad \text{for all } k \geq 0.$$

C Missing proofs from Section 5

In this section, we establish that weak and strong alignment hold for the parameterizations introduced in Section 5.

C.1 Sufficient conditions for alignment

Our proofs rely on establishing sufficient conditions for weak and strong alignment. In what follows, we present these conditions. We will use the symbols \bar{r}^* and \bar{r} instead of r^* and r to distinguish the rank of ∇F from the rank of its potential input, which will be particularly important when dealing with low-rank matrices.

C.2 Strong alignment

We start by showing that local strong alignment holds whenever the rank of the map F is constant near x^* , generalizing the assumptions for the Gauss-Newton subgradient method [26].

Lemma C.1 (Constant rank implies local strong alignment). *Let $x^* \in \mathbf{R}^d$ and $z^* = F(x^*)$. Assume that the map F satisfies Assumption 6 with $\varepsilon_{\nabla F} > 0$ and $L_{\nabla F} \geq 0$. Suppose there exists $\varepsilon > 0$ with*

$$\text{rank}(\nabla F(x^*)) = \text{rank}(\nabla F(x)) =: \bar{r}^* \quad \text{for all } x \in \mathbf{B}_\varepsilon(x^*).$$

Then, there exist positive constants R and C such that F satisfies Assumption 8 with

$$\delta(\rho) = \frac{\rho}{C}, \quad j = \bar{r}^*, \quad s = \frac{1}{2}\sigma_{\bar{r}^*}(\nabla F(x^*)) \quad \text{and} \quad \varepsilon_{x^*} = \min \left\{ R, \frac{\sigma_{\bar{r}^*}(\nabla F(x^*))}{2L_{\nabla F}}, \varepsilon_{\nabla F} \right\}.$$

Proof. We start with the first inequality in Assumption 8. By the Constant Rank Theorem [60, Theorem 4.12], there exists a constant $R' > 0$ such that the set $\mathcal{M} := F(\mathbf{B}_{R'}(x^*))$ is a C^1 -smooth manifold. It is well-known that near any point the distance between a manifold and its tangent grows quadratically [32, Lemma 3.2], that is, there are constants C and R'' such that for any $z := F(x) \in \mathbf{B}_{R''}(z^*)$ we have

$$\|(I - \text{proj}_{\mathcal{T}_{\mathcal{M}}(z)})(z^* - z)\| \leq C\|z^* - z\|^2, \quad (41)$$

where $\mathcal{T}_{\mathcal{M}}(z)$ is the tangent space of \mathcal{M} at z . Moreover, since $\mathcal{T}_{\mathcal{M}}(z) = \text{range}(\nabla F(x))$ [60, Chapter 5]. Therefore,

$$\left\| \left(I - \Pi_j^x \right) (z - z^*) \right\| \leq C \|z - z^*\|^2 \leq \rho \|z - z^*\|,$$

where the first inequality follows from $\Pi_j^x = \text{proj}_{\text{range}(\nabla F(x))} = \text{proj}_{\mathcal{T}_{\mathcal{M}}(z)}$, and the second inequality follows from $\|z - z^*\| \leq \delta(\rho) \leq \frac{\rho}{C}$.

To establish the lower bound on the singular value, we leverage Weyl's inequality. Since F satisfies Assumption 6 with $\varepsilon_{\nabla F}$ and $L_{\nabla F} \geq 0$, we have that $|\sigma_{\bar{r}^*}(\nabla F(x)) - \sigma_{\bar{r}^*}(\nabla F(x^*))| \leq L_{\nabla F} \|x - x^*\|$. Thus,

$$\sigma_{\bar{r}^*}(\nabla F(x)) \geq \sigma_{\bar{r}^*}(\nabla F(x^*)) - \varepsilon_{x^*} L_{\nabla F} \geq \frac{1}{2}\sigma_{\bar{r}^*}(\nabla F(x^*)).$$

Since F is continuous in $\mathbf{B}_{\varepsilon_{\nabla F}}(x^*)$ with a Lipschitz constant $\varepsilon_{\nabla F}L_{\nabla F} + \|\nabla F(x^*)\|_{op}$, we conclude upon taking $R = \min \left\{ R', \frac{R''}{\varepsilon_{\nabla F}L_{\nabla F} + \|\nabla F(x^*)\|_{op}} \right\}$. \square

C.3 Weak alignment

Recall that given a point x and a smooth map $F: \mathbf{E} \rightarrow \mathbf{Y}$, we let σ_j^x and Π_j^x denote the j th singular value of $\nabla F(x)$ and the projection onto the span of its top j left singular vectors. The following Proposition extends the proof idea in [106, Lemma 24].

Proposition C.2 (Sufficient condition for weak alignment). *Fix $z^* \in \text{Im } F$. Suppose there are functions $s: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and $\delta: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ satisfying that for any $\rho > 0$ and any $z = F(x)$ with*

$\|z - z^*\| \leq \delta(\rho)$, there exists an integer \bar{r}^* with $\bar{r}^* \leq \bar{r} := \text{rank}(\nabla F(x))$ such that the following statements hold.

1. Projected differences are bounded $\|(I - \Pi_{\bar{r}}^x)(z - z^*)\| \leq \rho \|z - z^*\|$.

2. For all $k \in \{\bar{r}^* + 1, \dots, \bar{r}\}$ we have

$$(\sigma_k^x)^2 \leq s(\rho) \|z - z^*\| \implies \|(I - \Pi_{k-1}^x)(z - z^*)\| \leq \rho \|z - z^*\|.$$

3. The \bar{r}^* -th singular value is lower bounded $(\sigma_{\bar{r}^*}^x)^2 \geq s(\rho) \|z - z^*\|$.

Then, the map F satisfies Assumption 3.

Proof. The result follows immediately from backward induction on k . \square

We also introduce a local version of this proposition.

Proposition C.3 (Sufficient condition for local weak alignment). *Let $x^* \in \mathbf{E}$ and $z^* \in \text{Im } F$ with $z^* = F(x^*)$. Suppose there is a scalar $\varepsilon_{x^*} > 0$, and functions $s: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ and $\delta: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ satisfying that for any $\rho > 0$ and any $z = F(x)$ with $\|z - z^*\| \leq \delta(\rho)$ and $\|x - x^*\| \leq \varepsilon_{x^*}$ there exists an integer \bar{r}^* with $\bar{r}^* \leq \bar{r} := \text{rank}(\nabla F(x))$ such that the following statements hold.*

1. Projected differences are bounded $\|(I - \Pi_{\bar{r}}^x)(z - z^*)\| \leq \rho \|z - z^*\|$.

2. For all $k \in \{\bar{r}^* + 1, \dots, \bar{r}\}$ we have

$$(\sigma_k^x)^2 \leq s(\rho) \|z - z^*\| \implies \|(I - \Pi_{k-1}^x)(z - z^*)\| \leq \rho \|z - z^*\|.$$

3. The \bar{r}^* -th singular value is lower bounded $(\sigma_{\bar{r}^*}^x)^2 \geq s(\rho) \|z - z^*\|$.

Then, the map F satisfies Assumption 7.

C.4 Proofs from Section 5.1

C.4.1 Proof of Theorem 5.1

Let us prove that the Hadamard map $x \mapsto x \odot x$ is smooth with parameter $L_{\nabla F} = 2$. We have

$$\|\nabla F(x) - \nabla F(y)\|_{\text{op}} = \|2 \text{diag}(x - y)\|_{\text{op}} \leq 2 \|x - y\|_2.$$

To prove that the map satisfies weak alignment Assumption 3, we leverage Proposition C.2. Thus, we will establish the three conditions stated in that proposition. To do so, we first introduce some auxiliary lemmas. These require a bit of extra notation. For a vector $x \in \mathbf{R}^r$, we let S_x be the set of permutations of the indexes $[r]$ that orders the entries of x in non ascending order, i.e., $\pi \in S_x$

$$x_{\pi(1)} \geq x_{\pi(2)} \geq \dots \geq x_{\pi(r)}.$$

Note that S is not a singleton whenever there are ties. We say that that two vectors $x, y \in \mathbf{R}^r$ are similarly ordered if $S_x \cap S_y \neq \emptyset$.

Lemma C.4. *Let $x^* \in \mathbf{R}^r$ be a fixed vector. Suppose $x \odot x \in \mathbf{B}_\varepsilon(x^* \odot x^*)$ with*

$$\varepsilon = \min_{i,j|x_i \neq x_j} \frac{|(x_i^*)^2 - (x_j^*)^2|}{2} \bigwedge_{i \in [r]|x_i^* \neq 0} \frac{(x_i^*)^2}{2}.$$

Then, $\#\text{supp}(x) \geq \#\text{supp}(x^)$ and, moreover, the component-wise squares $x \odot x$ and $x^* \odot x^*$ are similarly ordered. The result holds trivially when $\varepsilon = +\infty$, i.e., all components of x^* are equal.*

Proof. We defer the proof of $\#\text{supp}(x) \geq \#\text{supp}(x^*)$ to the end. Let us construct a permutation in the intersection of $S_{x^* \odot x^*}$ and $S_{x \odot x}$. We start with a base permutation $\pi \in S_{x^* \odot x^*}$, which we will modify inductively. Consider the partition B_1, \dots, B_ℓ of $[r]$ such that for any $i, j \in B_k$ we have $x_i^{*2} = x_j^{*2}$, and if $i \in B_n$ and $j \in B_m$ with $n < m$ then $x_i^{*2} > x_j^{*2}$. We claim that we also have $i \in B_n$ and $j \in B_m$ with $n < m$ then $x_i^2 > x_j^2$. To see this, take $\underline{i} = \text{argmin}_{i \in [B_n]} x_{\pi(i)}^2$ and $\bar{j} = \text{argmax}_{j \in [B_m]} x_{\pi(i)}^2$. Using the triangle inequality and the definition of the partition, we derive

$$x_{\pi(\underline{i})}^2 - x_{\pi(\bar{j})}^2 \geq x_{\pi(\underline{i})}^{*2} - x_{\pi(\bar{j})}^{*2} - 2\varepsilon > 0,$$

where the strict inequality follows since $x \odot x \in \mathbf{B}_\varepsilon(x^* \odot x^*)$; which proves our claim.

We construct π' from π as follows. Start with $\pi' = \pi$. We know that the indices in B_1 correspond to the top components in $x \odot x$ and $x^* \odot x^*$, so π sends them to the first $\#B_1$ components. We can modify π' to respect the ordering of the top $\#B_1$ entries of $x \odot x$. Since the new π' only differs from π in the B_1 block, it also belongs to $S_{x^* \odot x^*}$. We can apply the same update with B_2 , and so on until B_ℓ . After which, we have that $\pi' \in S_{x^* \odot x^*} \cap S_{x \odot x}$; proving that the squared vectors are similarly ordered.

Take $\pi \in S_{x^* \odot x^*} \cap S_{x \odot x}$ and recall that $r^* = \#\text{supp}(x^*)$. Thus, $(x_{\pi(r^*)}^*)^2 > 0$. By the definition of ε ,

$$x_{\pi(r^*)}^2 \geq \frac{(x_{\pi(r^*)}^*)^2}{2} > 0.$$

Therefore, $\#\text{supp}(x) \geq r^*$. This finishes the proof. \square

Recall that for $x \in \mathbf{R}^r$, we let Π_j^x and σ_j^x correspond to the projection onto the subspace generated by the top j singular vector of $\nabla F(x)$ and its j th top singular value, respectively.

Lemma C.5. *Let $x^* \in \mathbf{R}^r$ with r^* nonzero entries and $x \in \mathbf{R}^r$ such that $x \odot x \in B_\varepsilon(x^* \odot x^*)$ for*

$$\varepsilon = \min_{i,j|x_i \neq x_j} \frac{|(x_i^*)^2 - (x_j^*)^2|}{2} \bigwedge_{i \in [r]|x_i^* \neq 0} \frac{(x_i^*)^2}{2}.$$

Then, we have

$$\|(I - \Pi_k^x)(x \odot x - x^* \odot x^*)\|_2 \leq \sqrt{r-k} (\sigma_{k+1}^x)^2 \quad \text{for all } k \in \{r^*, \dots, r\}.$$

Proof. By Lemma C.4 there is a relabeling of the indexes $(j_i)_{i \in [r]}$ that simultaneously sorts the magnitude of the entries of x and x^* in nonascending order, i.e., $|x_{j_1}| \geq \dots \geq |x_{j_r}|$ and $|x_{j_1}^*| \geq \dots \geq |x_{j_r}^*|$. Since $\nabla F(x) = 2 \text{diag}(x)$, its SVD is given by

$$\begin{aligned} U &= (e_{j_1}, e_{j_2}, \dots, e_{j_r}), \\ \Sigma &= 2 \text{diag}(|x_{j_1}|, |x_{j_2}|, \dots, |x_{j_r}|), \\ V &= \text{diag}(\text{sign}(x_{j_1}), \text{sign}(x_{j_2}), \dots, \text{sign}(x_{j_r})), \end{aligned}$$

where e_k is the k -th standard vector basis. Hence, we have

$$[(I - \Pi_k^x)v]_i = \begin{cases} 0 & \text{if } i \in \{j_1, \dots, j_k\} \\ v_i & \text{otherwise,} \end{cases} \quad (42)$$

for an arbitrary vector $v \in \mathbf{R}^r$. Therefore, for any $k \in \{r^*, \dots, r\}$, we have

$$\begin{aligned} \|(I - \Pi_k^x)(x \odot x - x^* \odot x^*)\|_2^2 &= \|(I - \Pi_k^x)x \odot x - (I - \Pi_k^x)x^* \odot x^*\|_2^2 \\ &= \left\| (0, \dots, 0, x_{j_{k+1}}^2, \dots, x_{j_r}^2)^\top - (0, \dots, 0)^\top \right\|_2^2 \\ &= \left\| (x_{j_{k+1}}^2, \dots, x_{j_r}^2)^\top \right\|_2^2 \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=k+1}^r x_{j_i}^4 \\
&\leq (r-k)(\sigma_{k+1}^x)^4,
\end{aligned}$$

where the second equality follows from (42). This establishes the result. \square

We can now finish the proof of Theorem 5.1. Recall that $r^* = \#\text{supp}(x^*)$, $z = x \odot x$, and $z^* = x^* \odot x^*$. We show that $x \mapsto x \odot x$ satisfies the three conditions from Proposition C.2 and

$$s(\rho) = \frac{\rho}{\sqrt{r-r^*} \vee 1} \text{ and } \delta(\rho) = \left(\min_{i,j|x_i^* \neq x_j^*} \frac{|x_i^{*2} - x_j^{*2}|}{2} \right) \wedge \frac{\lambda_{r^*}(\text{diag}(z^*))}{1+s(\rho)} \wedge \frac{\lambda_{r^*}(\text{diag}(z^*))}{2}.$$

Suppose that $\|z - z^*\| \leq \delta(\rho)$.

1. By Lemma C.5, setting $k = r$, we have that $\|(I - \Pi_r^x)(z - z^*)\|_2^2 = 0 \leq \rho \|z - z^*\|_2$.
2. Take any $k \in \{r^* + 1, \dots, r\}$, by Lemma C.5, we have $\|(I - \Pi_{k-1}^x)(z - z^*)\|_2^2 \leq (r-k-1)\sigma_k^4$. Thus if $\sigma_k^2 \leq \frac{\rho}{\sqrt{r-r^*} \vee 1} \|z - z^*\|_2$, then,

$$\|(I - \Pi_{k-1}^x)(z - z^*)\|_2 \leq \rho \frac{\sqrt{r-k-1}}{\sqrt{r-r^*} \vee 1} \|z - z^*\|_2 \leq \rho \|z - z^*\|_2.$$

3. By Weyl's inequality, we have $\lambda_{r^*}(\text{diag}(z)) \geq \lambda_{r^*}(\text{diag}(z^*)) - \|z - z^*\|_2$. By the choice of the neighborhood $\delta(\rho)$, we have $\lambda_{r^*}(\text{diag}(z^*)) \geq (1 + s(\rho)) \|z - z^*\|_2$, we conclude that $\sigma_{r^*}^2 = \lambda_{r^*} \text{diag}(z) \geq s(\rho) \|z - z^*\|_2$.

Therefore, by Proposition C.2, Assumption 3 holds; completing the proof.

C.5 Proofs from Section 5.2

C.5.1 Proof of Theorem 5.5

We show that the symmetric Burer-Monteiro map $F_{\text{sym}}(X) = XX^\top$ is smooth with parameter $L_{\nabla F} = 2$. A straightforward computation reveals that the Jacobian of this map and its adjoint act on $Y \in \mathbf{R}^{d \times r}$ and $Z \in \mathcal{S}^d$ via

$$\nabla F_{\text{sym}}(X)[Y] = YX^\top + XY^\top \quad \text{and} \quad \nabla F_{\text{sym}}(X)^\top[Z] = (Z + Z^\top)X. \quad (43)$$

Therefore,

$$\begin{aligned}
\|\nabla F_{\text{sym}}(X) - \nabla F_{\text{sym}}(Y)\|_{\text{op}} &= \sup_{\|W\|_F=1} \|\nabla F_{\text{sym}}(X)[W] - \nabla F_{\text{sym}}(Y)[W]\|_F \\
&= \sup_{\|W\|_F=1} \left\| (WX^\top + XW^\top) - (WY^\top + YW^\top) \right\|_F \\
&\leq \sup_{\|W\|_F=1} 2 \left\| W(X - Y)^\top \right\|_F \\
&\leq 2 \|X - Y\|_F.
\end{aligned}$$

Thus, $L_{\nabla F} = 2$ as claimed.

Next, we prove weak alignment. Toward this goal, we state two auxiliary results. Consider any two $Z, Z^* \in \mathcal{S}_+^d$ with $r^* = \text{rank } Z^* \leq \text{rank } Z \leq r \leq d$ and let $X, X^* \in \mathbf{R}^{d \times r}$ be any matrices such that $Z = XX^\top$ and $Z^* = X^*(X^*)^\top$. We denote the SVD decompositions of X and $\nabla F_{\text{sym}}(X)$ as $U^X \Sigma^X (V^X)^\top$ and $U \Sigma V^\top$, respectively. We use U_i^X, U_i to denote the i -th column of U^X and U ,

respectively. With slight abuse of notation, we imagine completing the columns of U^X and U by choosing additional vectors in \mathbf{R}^d such that $\{U_i^X\}_{i=1}^d$ and $\{U_i\}_{i=1}^{\binom{d+1}{2}}$ forms an orthonormal basis of \mathbf{R}^d and $\mathbf{R}^{\binom{d+1}{2}}$ (we identify \mathcal{S}^d with $\mathbf{R}^{\binom{d+1}{2}}$), respectively. Further, we let Π_j^X be the orthogonal projection onto the span of the top j left singular vectors of $\nabla F_{\text{sym}}(X)$ and $\sigma_j = \Sigma_{jj}$ be its j -th singular value.

Proposition C.6. *Let $X^* \in \mathbf{R}^{d \times r}$. For any $\rho > 0$ and any $X \in \mathbf{R}^{d \times r}$ with*

$$\|XX^\top - X^*X^{*\top}\|_F \leq \min\left\{\frac{\rho}{\sqrt{2}}, \frac{1}{3}\right\} \sigma_{r^*}^2(X^*),$$

we have that

$$\|(I - \Pi_k^X)[XX^\top - X^*X^{*\top}]\|_F^2 \leq \frac{1}{16}(r - r^* + 1)\sigma_{k+1}^4 + \frac{\rho^2}{2}\|XX^\top - X^*X^{*\top}\|_F^2$$

for any $k \in \{\text{rank}(\nabla F_{\text{sym}}(X^)), \dots, \text{rank}(\nabla F_{\text{sym}}(X))\}$.*

Lemma C.7. *If $\|XX^\top - X^*X^{*\top}\|_{\text{op}} \leq \frac{1}{3}\sigma_{r^*}^2(X^*)$, then one has $\sigma_{\bar{r}^*}^2 \geq \sigma_{r^*}^2(X)$.*

We will prove these two results soon. Before delving into their proof, let us use these results to derive weak alignment. To this end, we show that the Burer-Monteiro map satisfies the three conditions from Proposition C.2 with

$$s(\rho) = \frac{4\rho}{\sqrt{2}(r - r^* + 1)} \text{ and } \delta(\rho) = \min\left\{\frac{\rho}{\sqrt{2}}, \frac{1}{1 + s(\rho)}, \frac{1}{3}\right\} \lambda_{r^*}(Z^*).$$

Define $\bar{r}^* = \text{rank}(\nabla F_{\text{sym}}(X^*))$ and take any $Z \in \text{Im } F_{\text{sym}}$ such that $\|Z - Z^*\|_F \leq \delta(\rho)$.

1. Using Proposition C.6 with $k = \text{rank } \nabla F_{\text{sym}}(X)$, we have

$$\|(I - \Pi_{\bar{r}}^X)[Z - Z^*]\|_F^2 \leq 0 + \frac{\rho^2}{2}\|Z - Z^*\|_F^2 \leq \rho^2\|Z - Z^*\|_F^2.$$

2. Let $k \in \{\bar{r}^* + 1, \dots, \text{rank } \nabla F_{\text{sym}}(X)\}$ and assume $\sigma_k^2 \leq s(\rho)\|Z - Z^*\|_F$. Again by Proposition C.6, we have

$$\begin{aligned} \|(I - \Pi_{k-1}^X)[Z - Z^*]\|_F^2 &\leq \frac{1}{16}(r - r^* + 1)\sigma_k^4 + \frac{\rho^2}{2}\|Z - Z^*\|_F^2 \\ &\leq \left(\frac{1}{16}(r - r^* + 1)s(\rho)^2 + \frac{\rho^2}{2}\right)\|Z - Z^*\|_F^2 \\ &= \rho^2\|Z - Z^*\|_F^2, \end{aligned}$$

where the last line follows from the definition of $s(\rho)$.

3. By Lemma C.7, we have that $\sigma_{\bar{r}^*}^2 \geq \sigma_{r^*}^2(X)$. By Weyl's inequality, we have $\sigma_{\bar{r}^*}^2(X) \geq \sigma_{\bar{r}^*}^2(X^*) - \|Z - Z^*\|_F$. By the choice of $\delta(\rho)$, we get $\sigma_{\bar{r}^*}^2(X^*) \geq (s(\rho) + 1 - 1)\|Z - Z^*\|_F$ and thus $\sigma_{\bar{r}^*}^2 \geq s(\rho)\|Z - Z^*\|_F$.

Then, invoking Proposition C.2 establishes Theorem 5.5. To complete the proof, we must still prove Proposition C.6 and Lemma C.7. The following are auxiliary results for such a purpose. Lemma C.7 follows directly from Lemma C.8 and Lemma C.11.

Lemma C.8 (Spectral characterization for Burer-Monteiro). *Let $F_{\text{sym}} : \mathbf{R}^{d \times r} \rightarrow \mathcal{S}^d$ be given by $F_{\text{sym}}(X) = XX^\top$. Then, the eigenpairs of $\nabla F_{\text{sym}}(X)\nabla F_{\text{sym}}(X)^\top$ are given by*

$$\left(2 \left(\sigma_i^2(X) + \sigma_j^2(X) \right), \frac{1}{c_{ij}} \left(U_i^X U_j^{X^\top} + U_j^X U_i^{X^\top} \right) \right)$$

for all $(i, j) \in [d] \times [d]$ with $i \leq j$. Here the normalizing constants are $c_{i,j} = 2$ if $i = j$ and $\sqrt{2}$ otherwise. Moreover, the eigenvectors form an orthonormal basis of \mathcal{S}^d .

Lemma C.8 likely already exists in the literature. We include a proof in Appendix D.1 for completeness. In turn, we need to understand how to conveniently index the eigenvalues and eigenvectors of $\nabla F_{\text{sym}}(X)^\top \nabla F_{\text{sym}}(X)$. The next few results develop such an indexing. A direct result is Lemma C.7.

Corollary C.9. *Let $\Delta = \{(i, j) \mid 1 \leq i \leq j \leq d\}$. Then, there exists a bijection $\tau: \Delta \rightarrow \binom{[d+1]}{2}$ such that for $(i, j) \in \Delta$, we have*

$$\begin{aligned} \lambda_{\tau(i,j)} \left(\nabla F_{\text{sym}}(X) \nabla F_{\text{sym}}(X)^\top \right) &= 2 \left(\sigma_i^2(X) + \sigma_j^2(X) \right), \quad \text{and} \\ U_{\tau(i,j)} &= \frac{1}{c_{ij}} U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X. \end{aligned} \quad (44)$$

Corollary C.10. *Let $X \in \mathbb{R}^{d \times r}$ of rank \tilde{r} . Then, the rank of $\nabla F_{\text{sym}}(X)$ is $d\tilde{r} - \binom{\tilde{r}}{2}$.*

These two are direct corollaries of Lemma C.8. In particular, $r^* = dr^* - \binom{r^*}{2}$ and the maximum rank of $\nabla F_{\text{sym}}(X)$ is $dr - \binom{r}{2}$. Consider the partition of Δ given by

$$\Delta_{r^*} = \{(i, j) \mid 1 \leq i \leq j \leq d \text{ and } i \leq r^*\} \quad \text{and} \quad \Delta_{r^*}^c = \Delta \setminus \Delta_{r^*}.$$

Lemma C.11. *If $\|XX^\top - X^*X^{*\top}\|_{\text{op}} \leq \frac{1}{3}\sigma_{r^*}^2(X^*)$, then there is a bijection $\tau: \Delta \rightarrow \binom{[d+1]}{2}$ satisfying (44) and*

$$\tau(i, j) < \tau(n, m) \quad \text{for all } (i, j) \in \Delta_{r^*} \text{ and } (n, m) \in \Delta_{r^*}^c. \quad (45)$$

Proof. Consider the bijection τ furnished by Corollary C.9. We claim that for any $i \in \{1, \dots, r^*\}$ and $n \in \{r^* + 1, \dots, d\}$,

$$\lambda_n(XX^\top) \leq \frac{\sigma_i^2(X)}{2}. \quad (46)$$

To show this inequality, we repeatedly apply Weyl's inequality

$$\begin{aligned} \lambda_n(XX^\top) &\leq \lambda_{r^*+1}(XX^\top) = \lambda_{r^*+1}(XX^\top) - \sigma_{r^*+1}^2(X^*) \\ &\leq \|XX^\top - X^*X^{*\top}\|_{\text{op}} \\ &\leq \frac{1}{2} \left(\sigma_{r^*}^2(X^*) - \|XX^\top - X^*X^{*\top}\|_{\text{op}} \right) \\ &\leq \frac{1}{2} \lambda_{r^*}(XX^\top) \leq \frac{1}{2} \sigma_i^2(X), \end{aligned}$$

where the second inequality follows since $\|XX^\top - X^*X^{*\top}\|_{\text{op}} \leq \frac{\sigma_{r^*}^2(X^*)}{3}$. Then, for any $(i, j) \in \Delta_{r^*}$ and $(n, m) \in \Delta_{r^*}^c$ we have

$$\begin{aligned} \lambda_{\tau(n,m)} \left(\nabla F_{\text{sym}}(X) \nabla F_{\text{sym}}(X)^\top \right) &= 2\lambda_n(XX^\top) + 2\lambda_m(XX^\top) \\ &\leq 2\sigma_i^2(X) \\ &\leq 2\sigma_i^2(X) + 2\sigma_j^2(X) = \lambda_{\tau(i,j)} \left(\nabla F_{\text{sym}}(X) \nabla F_{\text{sym}}(X)^\top \right), \end{aligned}$$

where the first inequality follows from (46). Therefore, by (44) we derive $\tau(i, j) \leq \tau(n, m)$; if this inequality does not hold strictly, we could modify τ to enforce strictness without breaking bijectivity. This establishes the result. \square

We are now ready to prove Proposition C.6.

Proof of Proposition C.6. Recall that U_i and σ_i denote the i th top left-singular vector and singular value of $\nabla F_{\text{sym}}(X)$, respectively. We use $X^* = U^{X^*} \Sigma^{X^*} (V^{X^*})^\top$ to denote the SVD of X^* , further we use $\sigma_i^X = \Sigma_{ii}^X$ and $\sigma_i^{X^*} = \Sigma_{ii}^{X^*}$. Let τ be the bijection provided by Lemma C.11 and expand

$$\begin{aligned} & (I - \Pi_k^X) [XX^\top - X^*X^{*\top}] \\ &= \left(\sum_{\ell=k+1}^{\binom{d+1}{2}} U_\ell U_\ell^\top \right) \text{vec} \left(XX^\top - X^*X^{*\top} \right) \\ &= \sum_{\substack{1 \leq i \leq j \leq d, \\ \tau(i,j) \geq k+1}} \frac{1}{c_{i,j}} \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right) \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right)^\top \text{vec} \left(XX^\top - X^*X^{*\top} \right), \end{aligned}$$

where the last equality follows from Corollary C.9, with $c_{i,j} = (2 + 2\mathbb{1}_{i=j})$. The next claim will help us understand this sum.

Claim C.12. *For any $1 \leq i \leq j \leq d$, we have*

$$\begin{aligned} & \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right) \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right)^\top \text{vec} \left(X^*X^{*\top} \right) \\ &= 2U_i^{X^\top} X^*X^{*\top} U_j^X \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right) \end{aligned}$$

and

$$\begin{aligned} & \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right) \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right)^\top \text{vec} \left(XX^\top \right) \\ &= \begin{cases} 4\sigma_i^2(X) \cdot U_i^X \otimes_{\text{Kr}} U_i^X & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

To prove this claim, for any $\tilde{X} \in \mathbf{R}^{d \times r}$, we apply properties of the Kronecker product, (67) and (68) to $\tilde{X}\tilde{X}^\top$ to derive

$$\begin{aligned} \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right)^\top \text{vec} \left(\tilde{X}\tilde{X}^\top \right) &= 2U_i^{X^\top} \tilde{X}\tilde{X}^\top U_j^X \\ &= 2 \sum_{\ell=1}^r \lambda_\ell(\tilde{X}\tilde{X}^\top) \langle U_\ell^{\tilde{X}}, U_i^X \rangle \langle U_\ell^{\tilde{X}}, U_j^X \rangle. \end{aligned}$$

The first and second equations imply the first and second statements, respectively.

Applying Claim C.12 and properties of the Kronecker product yields

$$\begin{aligned} (I - \Pi_k^X) [XX^\top - X^*X^{*\top}] &= 4 \underbrace{\sum_{\substack{1 \leq i \leq d, \\ \tau(i,i) \geq k+1}} \frac{1}{4} \left(U_i^X \otimes_{\text{Kr}} U_i^X \right) \sigma_i^2(X)}_{T_1:=} \\ &\quad - 2 \underbrace{\sum_{\substack{1 \leq i \leq j \leq d \\ \tau(i,j) \geq k+1}} \frac{1}{c_{i,j}} U_i^{X^\top} X^*X^{*\top} U_j^X \left(U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X \right)}_{T_2:=}. \end{aligned}$$

with $c_{i,j} = (2 + 2\mathbb{1}_{i=j})$. Taking the Frobenius norm and applying Young's inequality, the inequality yields

$$\left\| (I - \Pi_k^X) [XX^\top - X^*X^{*\top}] \right\|_F^2 \leq 2\|T_1\|_2^2 + 2\|T_2\|_2^2.$$

We focus on bounding each term.

Since $k \geq \text{rank } \nabla F_{\text{sym}}(X^*) = \#\Delta_{r^*}$, and τ satisfies (44) and (45) must have that

$$\tau(i, i) \geq k + 1 \quad \text{implies both} \quad i > r^* \quad \text{and} \quad \lambda_i(XX^\top) \leq \frac{1}{4}\sigma_{k+1}^2. \quad (47)$$

Equipped with these facts, we use the triangle inequality to obtain

$$\begin{aligned} \|T_1\|_2^2 &= \left\| \sum_{\substack{1 \leq i \leq d, \\ \tau(i, i) \geq k+1}} (U_i^X \otimes_{\text{Kr}} U_i^X) \sigma_i^2(X) \right\|_2^2 \\ &= \sum_{\substack{1 \leq i \leq d, \\ \tau(i, i) \geq k+1}} \lambda_i^2(XX^\top) \cdot \left\| (U_i^X \otimes_{\text{Kr}} U_i^X) \right\|_2^2 \\ &= \sum_{\substack{1 \leq i \leq d, \\ \tau(i, i) \geq k+1}} \frac{1}{16} \sigma_{k+1}^4 \\ &\leq \frac{1}{16} (r - r^* + 1) \sigma_{k+1}^4, \end{aligned}$$

where the second line follows from the orthonormality of $U_i^X \otimes_{\text{Kr}} U_i^X$, and the last two lines follows from (47).

Finally, we turn to the bounding T_2 . Thus, expanding T_2 , we get

$$\begin{aligned} \|T_2\|_2^2 &= \left\| 2 \sum_{\substack{1 \leq i \leq j \leq d \\ \tau(i, j) \geq k+1}} \frac{1}{(2 + 2\mathbb{1}_{i=j})} U_i^{X^\top} X^* X^{*\top} U_j^X (U_i^X \otimes_{\text{Kr}} U_j^X + U_j^X \otimes_{\text{Kr}} U_i^X) \right\|_2^2 \\ &\leq \left\| \sum_{\substack{1 \leq i \leq j \leq d \\ \tau(i, j) \geq k+1}} U_i^{X^\top} X^* X^{*\top} U_j^X (U_i^X \otimes_{\text{Kr}} U_j^X) \right\|_2^2 \\ &= \left\| \sum_{\substack{1 \leq i \leq j \leq d \\ \tau(i, j) \geq k+1}} U_i^X U_i^{X^\top} X^* X^{*\top} U_j^X U_j^{X^\top} \right\|_F^2, \end{aligned}$$

where the inequality follows from Cauchy-Schwarz inequality and $(2 + 2\mathbb{1}_{i=j}) \geq 2$. By the argument as (47), we have that

$$\tau(i, j) \geq k + 1 \quad \text{implies} \quad \min\{i, j\} > r^*.$$

Hence, we have

$$\begin{aligned}
\|T_2\|_2^2 &\leq \left\| \sum_{i=r^*+1}^d U_i^X U_i^{X^\top} X^* X^{*\top} \sum_{\substack{1 \leq j \leq d \\ \tau(i,j) \geq k+1}} U_j^X U_j^{X^\top} \right\|_F^2 \\
&\leq \left\| \left(\sum_{i=r^*+1}^d U_i^X U_i^{X^\top} \right) X^* X^{*\top} \left(\sum_{j=r^*+1}^d U_j^X U_j^{X^\top} \right) \right\|_F^2 \\
&\leq \frac{\rho^2}{2} \|X X^\top - X^* X^{*\top}\|_F^2,
\end{aligned}$$

where the second and third inequalities follow from Lemma D.2 and Lemma D.7, respectively. \square

This completes the proof of Theorem 5.5.

C.5.2 Proof of Theorem 5.6

We first show that the asymmetric matrix factorization map $F_{\text{asym}}(X, Y) = XY^\top$ satisfies Assumption 1 with parameter $L_{\nabla F} = \sqrt{2}$. A straight computation establishes that the Jacobian of this map and its adjoint acts on $(\tilde{X}, \tilde{Y}) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r}$ and $Z \in \mathbf{R}^{d_1 \times d_2}$ via

$$\nabla F_{\text{asym}}(X, Y)[(\tilde{X}, \tilde{Y})] = X\tilde{Y}^\top + \tilde{X}Y^\top \quad \text{and} \quad \nabla F_{\text{asym}}(X, Y)^\top[Z] = (ZY, Z^\top X). \quad (48)$$

Therefore,

$$\begin{aligned}
\left\| \nabla F_{\text{asym}}(X, Y) - \nabla F_{\text{asym}}(\tilde{X}, \tilde{Y}) \right\|_{\text{op}} &= \sup_{\|(A_1, A_2)\|_F=1} \left\| \left(\nabla F_{\text{asym}}(X, Y) - \nabla F_{\text{asym}}(\tilde{X}, \tilde{Y}) \right) [(A_1, A_2)] \right\|_F \\
&= \sup_{\|(A_1, A_2)\|_F=1} \left\| \left((X - \tilde{X})A_2^\top \right) + \left(A_1(Y - \tilde{Y})^\top \right) \right\|_F \\
&\leq \sup_{\|(A_1, A_2)\|_F=1} \left\| (X - \tilde{X})A_2^\top \right\|_F + \left\| A_1(Y - \tilde{Y}) \right\|_F \\
&\leq \|X - \tilde{X}\|_F + \|Y - \tilde{Y}\|_F \\
&\leq \sqrt{2} \left\| (X, Y) - (\tilde{X}, \tilde{Y}) \right\|_F,
\end{aligned}$$

where the last inequality comes from Young's inequality. Thus $L_{\nabla F} = \sqrt{2}$.

We turn to proving local weak alignment. Let us introduce some notation. Recall that we fixed a factorization $Z^* = X^*(Y^*)^\top$ with $\text{rank}(X^*) = \text{rank}(Y^*) = \text{rank}(Z^*) = r^*$. Consider any matrix Z with $r^* \leq \text{rank}(Z) \leq r$, and let $X \in \mathbf{R}^{d_1 \times r}$, $Y \in \mathbf{R}^{d_2 \times r}$ be any matrices such that $Z = XY^\top$. We denote the SVD decompositions of X, Y and $\nabla F_{\text{asym}}(X, Y)$ as $U^X \Sigma^X (V^X)^\top$, $U^Y \Sigma^Y (V^Y)^\top$, and $U \Sigma V^\top$, respectively. We use U_i^X, U_i^Y, U_i to denote the i -th column of U^X, U^Y , and U , respectively. With slight abuse of notation, we imagine completing the columns of U^X, U^Y , and U and adding additional vectors such that $\{U_i^X\}_{i=1}^{d_1}$, $\{U_i^Y\}_{i=1}^{d_2}$, and $\{U_i\}_{i=1}^{d_1 d_2}$ form an orthonormal basis of \mathbf{R}^{d_1} , \mathbf{R}^{d_2} , and $\mathbf{R}^{d_1 d_2}$, respectively. Further, we let $\Pi_j^{(X, Y)}$ be the orthogonal projection onto the span of the top j left singular vectors of $\nabla F_{\text{asym}}(X, Y)$ and σ_j be its j th singular value. Moreover, we denote by \bar{r} the rank of $\nabla F_{\text{asym}}(X, Y)$ and by \bar{r}^* the rank of $\nabla F_{\text{asym}}(X^*, Y^*)$. We state two key results that underline our arguments.

Proposition C.13. Let $(X^*, Y^*) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$ be a factorization of $Z^* = X^*(Y^*)^\top$ satisfying $\text{rank}(X^*) = \text{rank}(Y^*) = r^*$ and $V^{X^*} = V^{Y^*}$. Let $(X, Y) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$ be a pair of factors that satisfies $\|(X, Y) - (X^*, Y^*)\|_F \leq \frac{1}{16\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}}$. Then, for any $\rho > 0$, if

$$\|XY^\top - X^*Y^{*\top}\|_F \leq \rho \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{4},$$

we have that

$$\begin{aligned} & \left\| (I - \Pi_k^{(X, Y)}) [XY^\top - X^*Y^{*\top}] \right\|_F^2 \\ & \leq \left(5\sqrt{2} \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r - r^* + 1)^2 \right)^2 \sigma_{k+1}^4 + \frac{\rho^2}{2} \|XY^\top - X^*Y^{*\top}\|_F^2 \end{aligned}$$

for any $k \in \{\text{rank}(\nabla F_{\text{asym}}(X^*, Y^*)), \dots, \text{rank}(\nabla F_{\text{asym}}(X, Y))\}$.

Lemma C.14. Suppose $(X^*, Y^*) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$ satisfies $\text{rank}(X^*) = \text{rank}(Y^*) = r^*$. Further, let $(X, Y) \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$ be matrices satisfying $\|(X, Y) - (X^*, Y^*)\|_F \leq \frac{1}{16\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}}$. Then,

$$\sigma_{r^*}^2 \geq \min\{\sigma_{r^*}^2(X), \sigma_{r^*}^2(Y)\}.$$

We will soon prove these two results. Before that, let us use them to derive the local weak alignment property. To this end, we show that the asymmetric map satisfies the three conditions from Proposition C.3 with

$$\begin{aligned} \varepsilon_{x^*} &= \frac{1}{16\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}}, \\ s(\rho) &= \frac{\rho \min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{10\sqrt{2} (\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)) (r - r^* + 1)^2}, \quad \text{and} \\ \delta(\rho) &= \min\left\{\frac{\rho}{4}, \frac{1}{4s(\rho)}\right\} \min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}. \end{aligned}$$

Take any $XY^\top = Z \in \text{Im } F$ such that $\|Z - Z^*\|_F \leq \delta(\rho)$ and $\|(X, Y) - (X^*, Y^*)\|_F \leq \varepsilon_{x^*}$.

1. Applying Proposition C.13 with $\bar{r} = \text{rank } \nabla F(X, Y)$, we have

$$\left\| (I - \Pi_{\bar{r}}^{(X, Y)}) [Z - Z^*] \right\|_F^2 \leq \rho^2 \|Z - Z^*\|_F^2.$$

2. Let $k \in \{r^* + 1, \dots, \bar{r}\}$ and assume $\sigma_k^2 \leq s(\rho) \|Z - Z^*\|_F$. Again by Proposition C.13, we have

$$\begin{aligned} & \left\| (I - \Pi_{k-1}^{(X, Y)}) [Z - Z^*] \right\|_F^2 \\ & \leq \left(5\sqrt{2} \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r - r^* + 1)^2 \right)^2 \sigma_k^4 + \frac{\rho^2}{2} \|Z - Z^*\|_F^2 \\ & \leq \left(\left(5\sqrt{2} \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r - r^* + 1)^2 \right)^2 s(\rho)^2 + \frac{\rho^2}{2} \right) \|Z - Z^*\|_F^2 \\ & = \rho^2 \|Z - Z^*\|_F^2, \end{aligned}$$

where the last equality follows from the definition of $s(\rho)$.

3. Assume without loss of generality that $\min \{\sigma_{r^*}^2(X), \sigma_{r^*}^2(Y)\} = \sigma_{r^*}^2(X)$. We have:

$$\begin{aligned}\sigma_{\bar{r}^*}^2 &\geq \min \left\{ \sigma_{r^*}^2(X), \sigma_{r^*}^2(Y) \right\} \\ &\geq (\sigma_{r^*}(X^*) - \|X - X^*\|_F)^2 \\ &\geq \frac{1}{4} \min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\} \\ &\geq s(\rho) \left\| XY^\top - X^*Y^{*\top} \right\|_F,\end{aligned}$$

where the first line follows from Lemma C.14, the second line follows from Weyl's inequality and the choice of ε_{x^*} , and the last line follows from the choice of $\delta(\rho)$.

Then, the assumptions of Proposition C.3 hold, which establishes Theorem 5.6.

To complete the proof, we must still prove Proposition C.13 and Lemma C.14. The following are auxiliary results for such a purpose. Lemma C.14 follows directly from Lemma C.15 and Lemma C.19.

Lemma C.15 (Spectral Characterization). *The eigenpairs of $\nabla F(X, Y)\nabla F(X, Y)^\top$ are given by*

$$\sigma_i^2(X) + \sigma_j^2(Y) \text{ with eigenvector } U_j^Y U_i^{X^\top}$$

for all $i \in [d_1]$ and $j \in [d_2]$. Moreover, these eigenvectors form an orthonormal basis.

Lemma C.15 likely already exists in the literature. We include a proof in Appendix D.3.

Lemma C.16. *Suppose that $\|(X, Y) - (X^*, Y^*)\|_F \leq \frac{1}{16\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}}$. Then,*

$$\left\| XX^\top - X^*X^{*\top} \right\|_F + \left\| YY^\top - Y^*Y^{*\top} \right\|_F \leq \frac{1}{2\sqrt{2}} \min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\}.$$

We defer the proof of this lemma to Appendix D.2.

Corollary C.17. *There exists a bijection $\tau: [d_1] \times [d_2] \mapsto [d_1 d_2]$ such that*

$$\sigma_{\tau(i,j)}^2 = \sigma_i^2(X) + \sigma_j^2(Y) \text{ and } U_{\tau(i,j)} = U_j^Y U_i^{X^\top}. \quad (49)$$

Corollary C.18. *Let $X, Y \in \mathbb{R}^{d_1 \times r} \times \mathbb{R}^{d_2 \times r}$ be of the ranks r_1, r_2 . Then, the rank of $\nabla F(X, Y)$ is $d_1 r_2 + d_2 r_1 - r_1 r_2$.*

These two corollaries are direct consequences of Lemma C.15. In particular, $\bar{r}^* = (d_1 + d_2 - r^*)r^*$, and the maximum rank of $\nabla F(X, Y)$ is $(d_1 + d_2 - r)r$. Consider the partition of $[d_1] \times [d_2]$ given by

$$\Delta_{r^*} = \{(i, j) \mid i \leq r^* \text{ or } j \leq r^*\} \quad \text{and} \quad \Delta_{r^*}^c = \Delta \setminus \Delta_{r^*},$$

and observe that $\#\Delta_{r^*} = \text{rank}(\nabla F(X^*, Y^*))$. We derive a useful lemma for alignment.

Lemma C.19. *If $\left\| XX^\top - X^*X^{*\top} \right\|_{\text{op}} + \left\| YY^\top - Y^*Y^{*\top} \right\|_{\text{op}} \leq \frac{1}{2} \min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\}$, then there exists a bijection τ satisfying (49) such that*

$$\tau(i, j) < \tau(l, m) \quad \text{for all } (i, j) \in \Delta_{r^*} \text{ and } (l, m) \in \Delta_{r^*}^c.$$

Consequently, if $\tau(l, m) > \text{rank}(\nabla F(X^*, Y^*))$, then $l > r^*$ and $m > r^*$.

Proof of Lemma C.19. To establish that $\sigma_i^2(X) + \sigma_j^2(Y) \geq \sigma_l^2(X) + \sigma_m^2(Y)$ for any $(i, j) \in \Delta_{r^*}$ and $(l, m) \in \Delta_{r^*}^c$, it is sufficient to show

$$\sigma_{r^*+1}^2(X) + \sigma_{r^*+1}^2(Y) \leq \min \left\{ \sigma_{r^*}^2(X), \sigma_{r^*}^2(Y) \right\}. \quad (50)$$

The bound on $\|XX^\top - X^*X^{*\top}\|_{\text{op}} + \|YY^\top - Y^*Y^{*\top}\|_{\text{op}}$ implies that

$$\begin{aligned} & \|XX^\top - X^*X^{*\top}\|_{\text{op}} + \|YY^\top - Y^*Y^{*\top}\|_{\text{op}} \\ & \leq \min \left\{ \sigma_{r^*}^2(X^*) - \|XX^\top - X^*X^{*\top}\|_{\text{op}}, \sigma_{r^*}^2(Y^*) - \|YY^\top - Y^*Y^{*\top}\|_{\text{op}} \right\} \\ & \leq \min \left\{ \sigma_{r^*}^2(X), \sigma_{r^*}^2(Y) \right\}, \end{aligned} \quad (51)$$

where the second inequality follows from Weyl's inequality. To establish (50), we bound

$$\sigma_{r^*+1}^2(X) + \sigma_{r^*+1}^2(Y) \leq \|XX^\top - X^*X^{*\top}\|_{\text{op}} + \|YY^\top - Y^*Y^{*\top}\|_{\text{op}} \leq \min \left\{ \sigma_{r^*}^2(X), \sigma_{r^*}^2(Y) \right\},$$

where the first and second inequality follow from Weyl's and (51). This concludes the proof. \square

We are now ready to prove Proposition C.13.

Proof of Proposition C.13. We start by invoking the triangle inequality to decompose

$$\left\| \left(I - \Pi_k^{(X,Y)} \right) [XY^\top - X^*Y^{*\top}] \right\|_F \leq \underbrace{\left\| \left(I - \Pi_k^{(X,Y)} \right) [XY^\top] \right\|_F}_{T_1} + \underbrace{\left\| \left(I - \Pi_k^{(X,Y)} \right) [X^*Y^{*\top}] \right\|_F}_{T_2}.$$

We will provide upper bounds for both T_1 and T_2 . We begin with T_1 , Corollary C.17 yields

$$\begin{aligned} T_1^2 &= \left\| \sum_{i=k+1}^{d_1 d_2} U_i U_i^\top \text{vec} \left(XY^\top \right) \right\|_2^2 \\ &= \left\| \sum_{(i,j) | \tau(i,j) \geq k+1} \left(U_j^Y \otimes_{\text{Kr}} U_i^X \right) \left(U_j^Y \otimes_{\text{Kr}} U_i^X \right)^\top \text{vec} \left(XY^\top \right) \right\|_2^2 \\ &\stackrel{(i)}{=} \left\| \sum_{\substack{(i,j) \in [r] \times [r] \\ \tau(i,j) \geq k+1}} \sigma_i^X \sigma_j^Y \langle V_i^X, V_j^Y \rangle \left(U_j^Y \otimes_{\text{Kr}} U_i^X \right) \right\|_2^2 \\ &\stackrel{(ii)}{=} \sum_{\substack{(i,j) \in [r] \times [r] \\ \tau(i,j) \geq k+1}} \left\| \sigma_i^X \sigma_j^Y \langle V_i^X, V_j^Y \rangle \left(U_j^Y \otimes_{\text{Kr}} U_i^X \right) \right\|_2^2 \\ &\stackrel{(iii)}{\leq} \sum_{\substack{(i,j) \in [r] \times [r] \\ \tau(i,j) \geq k+1}} \frac{1}{4} \left(\sigma_i^2(X) + \sigma_j^2(Y) \right)^2 \\ &\stackrel{(iv)}{\leq} \frac{1}{4} \# \left\{ (i,j) \in [r]^2 \mid \text{rank} \nabla F(X^*, Y^*) + 1 \leq \tau(i,j) \leq \text{rank} \nabla F(X, Y) \right\} \sigma_{k+1}^4. \end{aligned}$$

Here, (i) follows since $XY^\top = \sum_{k=1}^r \sum_{\ell=1}^r \sigma_k^X \sigma_\ell^Y \langle V_k^X, V_\ell^Y \rangle U_k^X U_\ell^{Y^\top}$ and using (68) with (67) we derive

$$\left(U_j^Y \otimes_{\text{Kr}} U_i^X \right) \text{vec} \left(XY^\top \right) = \begin{cases} \sigma_i^X \sigma_j^Y \langle V_i^X, V_j^Y \rangle & \text{if } i \leq r \text{ and } j \leq r, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, (ii) follows from orthogonality and cancellation of cross-terms, (iii) from the Cauchy–Schwarz inequality and from Young's inequality, and (iv) follows from Lemma C.15.

Combining this bound with the fact that $\# \left\{ (i,j) \in [r]^2 \mid \text{rank} \nabla F(X^*, Y^*) + 1 \leq \tau(i,j) \leq \text{rank} \nabla F(X, Y) \right\} \leq$

$\text{rank } \nabla F(X, Y) \leq (r - r^* + 1)^2$ yields $T_1 \leq \frac{1}{2}(r - r^* + 1)\sigma_{k+1}^2$.

We continue by bounding the term T_2 . Let $(\underline{i}, \underline{j})$ be the pair such that $\tau(\underline{i}, \underline{j}) = k + 1$, then

$$\begin{aligned}
T_2 &= \left\| \sum_{(i,j) | \tau(i,j) \geq k+1} (U_j^Y \otimes_{\text{Kr}} U_i^X) (U_j^Y \otimes_{\text{Kr}} U_i^X)^\top \text{vec}(X^* Y^{*\top}) \right\|_2 \\
&\stackrel{(i)}{\leq} \left\| \sum_{i=\underline{i}}^{d_1} \sum_{j=\underline{j}}^{d_2} (U_j^Y \otimes_{\text{Kr}} U_i^X) (U_j^Y \otimes_{\text{Kr}} U_i^X)^\top \text{vec}(X^* Y^{*\top}) \right\|_2 \\
&\stackrel{(ii)}{=} \left\| \left(\sum_{i=\underline{i}}^{d_1} U_i^X U_i^{X^\top} \right) X^* Y^{*\top} \left(\sum_{j=\underline{j}}^{d_2} U_j^Y U_j^{Y^\top} \right) \right\|_F \\
&\stackrel{(iii)}{\leq} \frac{1}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \left\| \left(\sum_{i=1}^{\underline{i}-1} U_i^X U_i^{X^\top} \right) X Y^\top \left(\sum_{j=1}^{\underline{j}-1} U_j^Y U_j^{Y^\top} \right) - X^* Y^{*\top} \right\|_F^2 \\
&\stackrel{(iv)}{\leq} \frac{2 \left(\left\| \left(\sum_{i=1}^{\underline{i}-1} U_i^X U_i^{X^\top} \right) X Y^\top \left(\sum_{j=1}^{\underline{j}-1} U_j^Y U_j^{Y^\top} \right) - X Y^\top \right\|_F^2 + \left\| X Y^\top - X^* Y^{*\top} \right\|_F^2 \right)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \\
&\stackrel{(v)}{\leq} \underbrace{2 \left(\left\| \left(\sum_{i=1}^{\underline{i}-1} U_i^X U_i^{X^\top} \right) X Y^\top \left(\sum_{j=1}^{\underline{j}-1} U_j^Y U_j^{Y^\top} \right) - X Y^\top \right\|_F^2 \right)}_{T_3 :=} + \frac{\rho}{2} \left\| X Y^\top - X^* Y^{*\top} \right\|_F,
\end{aligned}$$

where (i) follows from the definition of $(\underline{i}, \underline{j})$ and from Lemma D.2, (ii) follows from the Kronecker product properties (67), (68), and (69), (iii) follows from Lemma D.7 together with Lemma C.16 and Lemma C.19, (iv) follows from adding and subtracting $X Y^\top$ in conjunction with Young's inequality, and (v) follows from the initial condition $\left\| X Y^\top - X^* Y^{*\top} \right\|_F \leq \rho \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{4}$.

Next, we provide a bound for the first term of the right-hand side of (vi). Let us denote $\mathcal{I}_{QQ} := \{(i, j) \in \mathbb{N}^2 : \underline{i} \leq i \leq r, \underline{j} \leq j \leq r\}$, $\mathcal{I}_{PQ} := \{(i, j) \in \mathbb{N}^2 : 1 \leq i < \underline{i}, \underline{j} \leq j \leq r\}$, $\mathcal{I}_{QP} := \{(i, j) \in \mathbb{N}^2 : \underline{i} \leq i \leq r, 1 \leq j < \underline{j}\}$, and $\mathcal{I} := \mathcal{I}_{QQ} \cup \mathcal{I}_{PQ} \cup \mathcal{I}_{QP}$. Then by orthonormality of the basis $\{U_j^Y \otimes U_i^X \mid (i, j) \in [d_1] \times [d_2]\}$ and since $\text{vec}(U_i^X U_j^{Y^\top}) = U_j^Y \otimes_{\text{Kr}} U_i^X$, we have that

$$T_3 = 2 \underbrace{\frac{\sum_{(i,j) \in \mathcal{I}_{QQ}} (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}}_{T_4 :=} + 2 \underbrace{\frac{\sum_{(i,j) \in \mathcal{I}_{PQ} \cup \mathcal{I}_{QP}} (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}}_{T_5 :=}.$$

We next bound T_4 and T_5 . For T_4 ,

$$\begin{aligned}
T_4 &\leq 2 \frac{\sigma_{r^*}(X) \sigma_{r^*}(Y) \sum_{i=\underline{i}}^r \sum_{j=\underline{j}}^r \sigma_i^X \sigma_j^Y}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \\
&\leq \frac{1}{2} \frac{\sigma_{r^*}^2(X) + \sigma_{r^*}^2(Y)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \sum_{i=\underline{i}}^r \sum_{j=\underline{j}}^r (\sigma_i^2(X) + \sigma_j^2(Y)) \\
&\leq 2 \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r - r^* + 1)^2 \sigma_{k+1}^2,
\end{aligned}$$

where the first inequality is due to Cauchy–Schwarz, the second is due to Young's inequality applied twice, and the third is due to Corollary C.17 and the initial conditions in conjunction with Weyl's

inequality. For the T_5 , we define $R := \operatorname{argmin}_{\tilde{R} \in O(r-r^*)} \left\| \left(V_{\{r^*+1\dots r\}}^Y - V_{\{r^*+1\dots r\}}^{Y^*} \tilde{R} \right) \right\|_F^2$. Next we only bound the terms in T_5 associated with the indices in \mathcal{I}_{PQ} , that is

$$\begin{aligned}
& \frac{\sum_{(i,j) \in \mathcal{I}_{PQ}} (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \stackrel{(i)}{=} 2 \frac{\sum_{i=1}^{r^*} \sum_{j=\underline{j}}^r (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2 + \sum_{i=r^*+1}^{i-1} \sum_{j=\underline{j}}^r (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \stackrel{(ii)}{\leq} 2 \frac{\sum_{i=1}^{r^*} \sum_{j=\underline{j}}^r (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2 + \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y) \sum_{i=r^*+1}^{i-1} \sum_{j=\underline{j}}^r 1}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \stackrel{(iii)}{\leq} 2 \frac{(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \left\| \left(V_{\{1\dots r^*\}}^X \right)^\top V_{\{r^*+1\dots r\}}^Y \right\|_F^2 + (r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \stackrel{(iv)}{\leq} 2 \frac{2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \left(\left\| \left(V_{\{1\dots r^*\}}^X \right)^\top V_{\{r^*+1\dots r\}}^{X^*} R \right\|_F^2 + \left\| \left(V_{\{1\dots r^*\}}^X \right)^\top \left(V_{\{r^*+1\dots r\}}^Y - V_{\{r^*+1\dots r\}}^{Y^*} R \right) \right\|_F^2 \right)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \quad + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \stackrel{(v)}{\leq} 2 \frac{2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \left\| \left(V_{\{1\dots r^*\}}^X \right)^\top V_{\{r^*+1\dots r\}}^{X^*} \right\|_F^2 + 2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \left\| V_{\{r^*+1\dots r\}}^Y - V_{\{r^*+1\dots r\}}^{Y^*} R \right\|_F^2}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \quad + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}},
\end{aligned}$$

where (i) follows from rearranging, (ii) follows from Weyl's inequality applied to $\sigma_{r^*+1}^X$ and from the Cauchy–Schwarz inequality, (iii) follows from Lemma C.19 and from adding nonnegative components to the Frobenius norm, (iv) follows from Young's inequality and from the assumption that $V^{X^*} = V^{Y^*}$. Finally, (v) follows from the Courant–Fisher theorem applied to orthogonal operators. We further bound,

$$\begin{aligned}
& 2 \frac{2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \left\| \left(V_{\{1\dots r^*\}}^X \right)^\top V_{\{r^*+1\dots r\}}^{X^*} \right\|_F^2 + 2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \left\| V_{\{r^*+1\dots r\}}^Y - V_{\{r^*+1\dots r\}}^{Y^*} R \right\|_F^2}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \quad + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \stackrel{(i)}{\leq} 2 \frac{2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \frac{4\|X-X^*\|_F^2}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} + 2(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2 \frac{8\|Y-Y^*\|_F^2}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}}}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \quad + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}} \\
& \leq \frac{32(\sigma_1^X)^2 (\sigma_{\underline{j}}^Y)^2}{\min \{ \sigma_{r^*}^4(X^*), \sigma_{r^*}^4(Y^*) \}} \| (X, Y) - (X^*, Y^*) \|_F^2 + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_{\underline{j}}^2(Y)}{\min \{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \}}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} \frac{1}{16} \frac{(\sigma_1^X)^2 (\sigma_j^Y)^2}{\max\{\sigma_1^2(X^*), \sigma_1^2(Y^*)\}} + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_j^2(Y)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \\
&\stackrel{(iii)}{\leq} \frac{\sigma_1^2(X^*) + \sigma_{r^*}^2(X^*)}{8 \max\{\sigma_1^2(X^*), \sigma_1^2(Y^*)\}} (\sigma_j^Y)^2 + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_j^2(Y)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \\
&\leq \frac{1}{4} (\sigma_j^Y)^2 + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_j^2(Y)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}. \\
&\leq \frac{1}{4} (\sigma_j^Y)^2 + 2 \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*) \sigma_j^2(Y)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \\
&\leq \frac{9}{4} \frac{(r-r^*+1)^2 \sigma_{r^*}^2(X^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \sigma_j^2(Y).
\end{aligned}$$

Here, (i) follows from a combination of results: first, we rewrite $\left\| \left(V_{\{1\dots r^*\}}^X \right)^\top V_{\{r^*+1\dots r\}}^{X^*} \right\|_F$ using Lemma D.3; second, we bound $\left\| V_{\{r^*+1\dots r\}}^Y - V_{\{r^*+1\dots r\}}^{Y^*} R \right\|_F$ using Lemma D.4; and third, we invoke Wedin's theorem [21, Theorem 2.9] on both of these terms, which is applicable since $\|(X, Y) - (X^*, Y^*)\|_F \leq \frac{1}{16\sqrt{2}} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\max\{\sigma_1(X^*), \sigma_1(Y^*)\}} \leq \frac{1}{4} \min\{\sigma_{r^*}(X^*), \sigma_{r^*}(Y^*)\}$. Finally, inequalities (ii) and (iii) are due to the bound on the condition $\|(X, Y) - (X^*, Y^*)\|_F$ together with Weyl's and Young's inequality. Using a similar argument, one can bound the rest of the terms in T_5 by

$$2 \frac{\sum_{(i,j) \in \mathcal{I}_{QP}} (\sigma_i^X)^2 (\sigma_j^Y)^2 \langle V_i^X, V_j^Y \rangle^2}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \leq \left(\frac{9}{4} \frac{(r-r^*+1)^2 \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \right) \sigma_i^2(X),$$

so that

$$\begin{aligned}
T_5 &\leq \left(\frac{9}{4} \frac{(r-r^*+1)^2 \max\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \right) (\sigma_i^2(X) + \sigma_j^2(Y)) \\
&= \left(\frac{9}{4} \frac{(r-r^*+1)^2 \max\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} \right) \sigma_{k+1}^2,
\end{aligned}$$

and thus, adding T_4 yields

$$T_4 + T_5 \leq \frac{17}{4} \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r-r^*+1)^2 \sigma_{k+1}^2.$$

To conclude, since $T_1 \leq \frac{1}{2} (r-r^*+1) \sigma_{k+1}^2 \leq \frac{1}{2} \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r-r^*+1)^2 \sigma_{k+1}^2$, we obtain

$$\begin{aligned}
&\| (I - \Pi_k^{(X,Y)}) [XY^\top - X^*Y^{*\top}] \|_F \\
&\leq \left(5 \frac{\sigma_{r^*}^2(X^*) + \sigma_{r^*}^2(Y^*)}{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}} (r-r^*+1)^2 \right) \sigma_{k+1}^2 + \frac{\rho}{2} \|XY^\top - X^*Y^{*\top}\|_F.
\end{aligned}$$

Taking the square of both sides and applying Young's inequality yields the desired result. \square

C.5.3 Proof of Lemma 5.8

The objective h is a composition of a linear map with a convex function, and so it's convex, proving Item 2 of Assumption 5. To establish Items 1 and 3, we establish quadratic growth with $z^* = M^*$. Notice that $\text{Im } F$ corresponds with the set of rank r matrices. Let an arbitrary $M \in \text{Im } F$. Applying

the reverse triangle inequality yields

$$\frac{1}{2} \|\mathcal{A}(M) - b\|_2^2 - \frac{1}{2} \|\mathcal{A}(M^*) - b\|_2^2 \geq \frac{1}{2} \|\mathcal{A}(M - M^*)\|_2^2 \geq \frac{1}{2} (1 - \delta) \|M - M^*\|_2^2, \quad (52)$$

where the second inequality follows from (22) since $M - M^*$ has rank at most $2r$.

We will use the following lemma for our proof of Item 4 in Assumption 5.

Lemma C.20 (Lemma 3.3 in [15] and Lemma 31 in [94]). *Assume that \mathcal{A} satisfies (22) for matrices of rank at most $2r$. Then one has*

$$\left\langle \mathcal{A}(M), \mathcal{A}(\tilde{M}) \right\rangle \leq \delta \|M\|_F \|\tilde{M}\|_F + \left| \langle M, \tilde{M} \rangle \right|,$$

for any matrices M and \tilde{M} of rank at most r .

We use Π as a shorthand for the projection $\Pi^{(X,Y)}$ onto the image of $\nabla F(X, Y)$. Recall from (48) that this image lies within the set of matrices of rank at most $2r$ and so

$$\text{Im } \Pi \subseteq \{Z \in \mathbf{R}^{d_1 \times d_2} \mid \text{rank}(Z) \leq 2r\}. \quad (53)$$

Therefore,

$$\begin{aligned} \|\Pi \nabla h(M)\|_2 &= \|\Pi [\mathcal{A}^* \mathcal{A}(M - M^*)]\|_F \\ &= \sup_{W \in \mathbf{R}^{d_1 \times d_2} \mid \|W\|_F=1} \langle \Pi (\mathcal{A}^* \mathcal{A}(M - M^*)), W \rangle \\ &= \sup_{W \in \mathbf{R}^{d_1 \times d_2} \mid \|W\|_F=1} \langle \mathcal{A}^* \mathcal{A}(M - M^*), \Pi[W] \rangle \\ &\leq \sup_{\substack{W \in \mathbf{R}^{d_1 \times d_2} \mid \|W\|_F=1 \\ \text{rank}(W) \leq 2r}} \langle \mathcal{A}^* \mathcal{A}(M - M^*), W \rangle \\ &= \sup_{\substack{W \in \mathbf{R}^{d_1 \times d_2} \mid \|W\|_F=1 \\ \text{rank}(W) \leq 2r}} \langle \mathcal{A}(M - M^*), \mathcal{A}(W) \rangle \\ &\leq \delta \|M - M^*\|_F + \sup_{W \in \mathbf{R}^{d_1 \times d_2} \mid \|W\|_F=1} \langle M - M^*, W \rangle \\ &= (1 + \delta) \|M - M^*\|_F, \end{aligned}$$

where the first inequality follows from (53), and the second inequality uses Lemma C.20. Using this result in tandem with (52) yields

$$\frac{(1 - \delta)}{2(1 + \delta)^2} \|\Pi \nabla h(M)\|_2^2 \leq h(M) - h(M^*),$$

thus, establishing part (a) of Item 4. For part (b), recall from the definition of $P((X, Y), \lambda)$, (7), that $\text{Im}(I - P((X, Y), \lambda)) = \text{Im } \nabla F(X, Y)$ and so $(I - P((X, Y), \lambda))[M] \in M + \text{Im } \nabla F(X, Y)$ and consequently

$$\text{rank}((I - P((X, Y), \lambda))[M]) \leq \text{rank } M + 2r. \quad (54)$$

Hence,

$$\begin{aligned} &\langle \mathcal{A}^* \mathcal{A}(M - M^*), (I - P((X, Y), \lambda))[M - M^*] \rangle \\ &= \langle \mathcal{A}(M - M^*), \mathcal{A}((I - P((X, Y), \lambda))[M - M^*]) \rangle \\ &\leq \delta \|M - M^*\|_F \|(I - P((X, Y), \lambda))[M - M^*]\|_F + |\langle M - M^*, (I - P((X, Y), \lambda))[M - M^*] \rangle| \\ &\leq (1 + \delta) \|M - M^*\|_F \|(I - P((X, Y), \lambda))(M - M^*)\|_F, \end{aligned}$$

where the first inequality follows from Lemma C.20, which applies due to (54), and the second inequality follows from Cauchy–Schwarz. The proof concludes by taking $\alpha = (1 - \delta)$ and $\beta =$

$$\max \left\{ \frac{(1+\delta)^2}{(1-\delta)}, (1+\delta) \right\} = \frac{(1+\delta)^2}{(1-\delta)}.$$

C.5.4 Proof of Lemma 5.11

Items 1, 2 and 3 hold automatically and so we focus on proving Item 4. Take $M \in \mathbf{R}^{d_1 \times d_2}$ and $V \in \partial f(M)$. A key ingredient to this proof is the fact that for any matrix W of rank at most $2r$ we have

$$\begin{aligned} \langle V, W \rangle &= \langle V, (W + M) - M \rangle \\ &\leq f(W + M) - f(M) \\ &\leq L \|W\|_F, \end{aligned} \tag{55}$$

where the first inequality holds since f is convex and $V \in \partial f(M)$, and the second holds since f satisfies restricted L Lipschitzness.

To establish Item 4a of the Assumption, we have

$$\begin{aligned} \left\| \Pi^{(X,Y)}[V] \right\|_F &\leq \sup_{\substack{W | \text{rank}(W) \leq 2r, \\ \|W\|_F = 1}} \langle W, V \rangle \\ &\leq \sup_{\|W\|_F = 1} L \|W\|_F = 1. \end{aligned}$$

Recall that $\Pi^{(X,Y)}$ denotes the projection onto $\text{range } \nabla F(X, Y)$, which is a subset of $\{W \mid W \in \mathbf{R}^{d_1 \times d_2} \text{ and } \text{rank}(W) \leq 2r\}$. Thus, establishing the first inequality, while the second inequality follows from (55).

To prove Item 4b, the matrix $W = (I - P((X, Y), \lambda))[M - M^*]$ has rank $4r$ due to (54). We can further decompose W into two rank- $2r$ matrices W_1 and W_2 such that $\langle W_1, W_2 \rangle = 0$. Therefore,

$$\begin{aligned} |\langle V, W \rangle|^2 &= |\langle V, W_1 \rangle + \langle V, W_2 \rangle|^2 \\ &= |\langle V, W_1 \rangle|^2 + 2 |\langle V, W_1 \rangle \langle V, W_2 \rangle| + |\langle V, W_2 \rangle|^2 \\ &\leq L^2 \|W_1\|_F^2 + 2 \|W_1\|_F \|W_2\|_F + L^2 \|W_2\|_F^2 \\ &= L^2 \left(\|W_1\|_F^2 + \|W_2\|_F^2 \right) \\ &= L^2 \|W\|_F^2, \end{aligned}$$

where the inequality follows from (55) and the last equality uses fact that $\|W\|_F^2 = \|W_1\|_F^2 + \|W_2\|_F^2$, since W_1 and W_2 are orthogonal. This concludes the proof.

C.6 Proofs from Section 5.3

We start with a few explicit definitions that will play a role in our arguments. In what follows, we use M_j and M_i to refer to the j -th row and i -th column, respectively.

Definition C.21 (Column-major vectorization of matrices and tensors). *Let $M \in \mathbf{R}^{d_1 \times d_2}$ and $T \in \mathbf{R}^{d_1 \times d_2 \times d_3}$. The vectors $\text{vec}(M) \in \mathbf{R}^{d_1 d_2}$ and $\text{vec}(T) \in \mathbf{R}^{d_1 d_2 d_3}$ are defined by*

$$\begin{aligned} \text{vec}(M)_{(i_2-1)d_1+i_1} &= M_{i_1, i_2} && \text{for } i_1 \in [d_1], i_2 \in [d_2], \\ \text{vec}(T)_{(i_3-1)d_1 d_2 + (i_2-1)d_1 + i_1} &= T_{i_1, i_2, i_3} && \text{for } i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3]. \end{aligned}$$

Definition C.22 (Matricization of tensors). *Let $T \in \mathbf{R}^{d_1 \times d_2 \times d_3}$. The mode-1 matricization $\mathcal{M}_1(T) \in \mathbf{R}^{d_1 \times (d_2 d_3)}$ is given by*

$$\mathcal{M}_1(T)_{i_1, (i_3-1)d_2+i_2} = T_{i_1, i_2, i_3} \quad \text{for } i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3].$$

Similarly, the mode-2 and mode-3 matricizations $\mathcal{M}_2(T) \in \mathbf{R}^{d_2 \times (d_1 d_3)}$ and $\mathcal{M}_3(T) \in \mathbf{R}^{d_3 \times (d_1 d_2)}$ are

$$\begin{aligned} \mathcal{M}_2(T)_{i_2, (i_3-1)d_1+i_1} &= T_{i_1, i_2, i_3} & \text{for } i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3], \\ \mathcal{M}_3(T)_{i_3, (i_2-1)d_1+i_1} &= T_{i_1, i_2, i_3} & \text{for } i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3]. \end{aligned}$$

Definition C.23 (Permutations of tensor matricization). Given $i \in \{2, 3\}$ and $T \in \mathbf{R}^{d_1 \times d_2 \times d_3}$, let $P_i \in \mathbf{R}^{d_1 d_2 d_3 \times d_1 d_2 d_3}$ be the permutation matrix such that

$$P_i \text{vec}(\mathcal{M}_i(T)) = \text{vec}(\mathcal{M}_1(T)) = \text{vec}(T).$$

Definition C.24 (Kronecker product for matrices and vectors). Given $M \in \mathbb{R}^{d_1 \times d_2}$ and $N \in \mathbb{R}^{d_3 \times d_4}$, their Kronecker product $M \otimes_{\text{Kr}} N \in \mathbb{R}^{d_1 d_3 \times d_2 d_4}$ is defined entrywise by

$$(M \otimes_{\text{Kr}} N)_{(i_1-1)d_3+i_3, (i_2-1)d_4+i_4} = M_{i_1, i_2} N_{i_3, i_4} \quad \text{for } i_1 \in [d_1], i_2 \in [d_2], i_3 \in [d_3], i_4 \in [d_4].$$

Moreover, for vectors $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$, one has

$$(u \otimes_{\text{Kr}} v)_{(i_1-1)d_2+i_2} = u_{i_1} v_{i_2} \quad \text{for } i_1 \in [d_1], i_2 \in [d_2].$$

Definition C.25. Given any two matrices $M \in \mathbf{R}^{d_1 \times r}$ and $N \in \mathbf{R}^{d_2 \times r}$, we define

$$\psi(M, N) = [M_1 \otimes_{\text{Kr}} N_1 \ \cdots \ M_r \otimes_{\text{Kr}} N_r] \in \mathbf{R}^{d_1 d_2 \times r} \quad \text{and} \quad \Psi(M, N) = \sum_{j=1}^r M_j M_j^\top \otimes_{\text{Kr}} N_j N_j^\top \in \mathbf{R}^{d_1 d_2 \times d_1 d_2},$$

where M_j and N_j denote the j th columns of M and N , respectively.

C.6.1 Proof of Theorem 5.16

We start with the actions of the Jacobian and its adjoint.

Lemma C.26. Let $X \in \mathbf{R}^{d \times r}$. Then, the action of $\nabla F_{\text{sym}}(X)$ on a direction $D \in \mathbf{R}^{d \times r}$ is given by

$$\begin{aligned} \nabla F_{\text{sym}}(X) \text{vec}(D) &= \sum_{\ell=1}^r (D_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} X_\ell + X_\ell \otimes_{\text{Kr}} D_\ell \otimes_{\text{Kr}} X_\ell + X_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} D_\ell) \\ &= (I + P_3 + P_2) \sum_{\ell=1}^r D_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} X_\ell \in \mathbf{R}^{d^3}. \end{aligned}$$

Moreover, the action of the adjoint $\nabla F_{\text{sym}}(X)^\top$ on a rank-1 tensor $a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c \in \mathbf{R}^{d^3}$ is given by

$$\nabla F_{\text{sym}}(X)^\top (a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c) = \text{vec} \left((a(b^\top \otimes_{\text{Kr}} c^\top) + b(a^\top \otimes_{\text{Kr}} c^\top) + c(a^\top \otimes_{\text{Kr}} b^\top)) \psi(X, X) \right) \in \mathbf{R}^{dr}.$$

The proof is deferred to Appendix D.4. Given this result, it is straightforward to establish Assumption 6. Consider an arbitrary pair $X, \tilde{X} \in \mathbf{R}^{d \times r}$ satisfying $\max \left\{ \|X - X^*\|_F, \|\tilde{X} - X^*\|_F \right\} \leq \|X^*\|_F$. Using the variational characterization of the operator norm, we have

$$\begin{aligned} \left\| \nabla F_{\text{sym}}(X) - \nabla F_{\text{sym}}(\tilde{X}) \right\|_{\text{op}} &= \sup_{A \in \mathbf{R}^{d \times r}, \|A\|_F=1} \left\| \left(\nabla F_{\text{sym}}(X) - \nabla F_{\text{sym}}(\tilde{X}) \right) \text{vec}(A) \right\|_2 \\ &\stackrel{(i)}{=} \sup_{\|A\|_F=1} \left\| (I + P_2 + P_3) \left(\sum_{\ell=1}^r A_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} X_\ell - A_\ell \otimes_{\text{Kr}} \tilde{X}_\ell \otimes_{\text{Kr}} \tilde{X}_\ell \right) \right\|_2 \\ &\stackrel{(ii)}{\leq} 3 \sup_{\|A\|_F=1} \left\| \sum_{\ell=1}^r A_\ell \otimes_{\text{Kr}} \left(X_\ell \otimes_{\text{Kr}} X_\ell - \tilde{X}_\ell \otimes_{\text{Kr}} \tilde{X}_\ell \right) \right\|_2 \\ &\stackrel{(iii)}{=} 3 \sup_{\|A\|_F=1} \left\| \left(\psi(X, X) - \psi(\tilde{X}, \tilde{X}) \right) A^\top \right\|_F \\ &\stackrel{(iv)}{\leq} 3 \left\| \psi(X, X) - \psi(\tilde{X}, \tilde{X}) \right\|_F, \end{aligned}$$

where (i) follows from the Lemma C.26, (ii) follows from the fact that permutation matrices have operator norm one, (iii) follows from $\sum_{\ell=1}^r A_\ell \otimes_{\text{Kr}} B_\ell = \text{vec}(BA^\top)$ for any A and B , and (iv) follows from the submultiplicativity of the Frobenius norm. Leveraging the fact that $\sum_{\ell=1}^r \|v_\ell\|_2^2 \leq (\sum_{\ell=1}^r \|v_\ell\|_2)^2$, we have

$$\begin{aligned}
\left\| \nabla F_{\text{sym}}(X) - \nabla F_{\text{sym}}(\tilde{X}) \right\|_{\text{op}} &\leq 3 \sum_{\ell=1}^r \left\| X_\ell \otimes_{\text{Kr}} X_\ell - \tilde{X}_\ell \otimes_{\text{Kr}} \tilde{X}_\ell \right\|_2 \\
&= 3 \sum_{\ell=1}^r \left\| X_\ell \otimes_{\text{Kr}} X_\ell - X_\ell \otimes_{\text{Kr}} \tilde{X}_\ell + X_\ell \otimes_{\text{Kr}} \tilde{X}_\ell - \tilde{X}_\ell \otimes_{\text{Kr}} \tilde{X}_\ell \right\|_2 \\
&\stackrel{(i)}{\leq} 3 \sum_{\ell=1}^r \left(\left\| X_\ell \otimes_{\text{Kr}} (X_\ell - \tilde{X}_\ell) \right\|_2 + \left\| (X_\ell - \tilde{X}_\ell) \otimes_{\text{Kr}} \tilde{X}_\ell \right\|_2 \right) \\
&\stackrel{(ii)}{\leq} 3 \left(\|X\|_F + \|\tilde{X}\|_F \right) \|X - \tilde{X}\|_F \\
&\stackrel{(iii)}{\leq} 12 \|X^*\|_F \|X - \tilde{X}\|_F,
\end{aligned}$$

where (i) follows from the triangle inequality and bilinearity of the Kronecker product, (ii) follows from the fact that $\|a \otimes b\|_2 = \|a\|_2 \|b\|_2$ for any a and b , together with the Cauchy-Schwarz inequality, and (iii) holds since by assumption $\max \left\{ \|X - X^*\|_F, \|\tilde{X} - X^*\|_F \right\} \leq \|X^*\|_F$, which implies $\max \left\{ \|X\|_F, \|\tilde{X}\|_F \right\} \leq 2 \|X^*\|_F$ by the reverse triangle inequality.

We now proceed to proving strong alignment for this map. Let $T^* \in \mathbf{R}^{d \times d \times d}$ be an arbitrary tensor and $X^* \in \mathbf{R}^{d \times r}$ be any full-rank matrix with $T^* = F_{\text{sym}}(X^*)$. To establish alignment, we use the following result.

Proposition C.27 (Constant rank of the symmetric CP map). *For any full rank matrix $X \in \mathbf{R}^{d \times r}$,*

$$\text{rank}(\nabla F_{\text{sym}}(X)) = dr.$$

Given this result, invoking Lemma C.1 gives us that the map F_{sym} satisfies Assumption 8 with $j = dr$ and with the quantities

$$\varepsilon_{x^*} = \min \left\{ R, \frac{\sigma_{dr}(\nabla F_{\text{sym}}(X^*))}{24 \|X^*\|_F}, \|X^*\|_F \right\}, \quad \delta(\rho) = \frac{\rho}{C}, \quad \text{and} \quad s = \frac{1}{2} \sigma_{dr}(\nabla F_{\text{sym}}(X^*))$$

for some positive constants R and C that depend only on X^* . This establishes the result, provided that we show Proposition C.27; we now proceed to prove it.

Proof of Proposition C.27. Let $X \in \mathbf{R}^{d \times r}$ be of rank r . The proof consists of two steps. Firstly, we will construct a set H of probing vectors, i.e., dr linearly independent vectors in \mathbf{R}^{d^3} . Secondly, we will prove that the probing set remains linearly independent after applying $\nabla F_{\text{sym}}(X)^\top$. This shows a lower bound $\text{rank}(\nabla F_{\text{sym}}(X)^\top) \geq dr$. Since $\nabla F_{\text{sym}}(X) \in \mathbf{R}^{d^3 \times dr}$, the rank of $\nabla F_{\text{sym}}(X)$ is at most dr , we must have $\text{rank}(\nabla F_{\text{sym}}(X)) = dr$.

We take the full SVD of X as $U^X \Sigma^X (V^X)^\top$, where $U^X \in \mathbf{R}^{d \times d}$, $\Sigma^X \in \mathbf{R}^{d \times r}$ and $(V^X)^\top \in \mathbf{R}^{r \times r}$. Consider the extended SVD

$$\tilde{\Sigma}^X := \text{diag}(\sigma_1^X, \dots, \sigma_r^X, 1, \dots, 1) \in \mathbf{R}^{d \times d}, \quad \text{and} \quad \tilde{V}^X := \begin{pmatrix} V^X & 0 \\ 0 & I \end{pmatrix} \in \mathbf{R}^{d \times d}.$$

Observe that with this notation we have $X_\ell = U^X \tilde{\Sigma}^X (\tilde{V}_\ell^X)^\top$ for all $\ell \in [r]$.

Probing set. First, since X is full rank, the vectors

$$T_k := U^X \left(\tilde{\Sigma}^X \right)^{-1} \left(\tilde{V}_{k:}^X \right)^\top \in \mathbf{R}^d \quad \text{are well-defined for all } k \in [d].$$

We construct $H = \{T_i \otimes_{\text{Kr}} T_j \otimes_{\text{Kr}} T_j \mid i \in [d], j \in [r]\}$. By (69), Kronecker products of invertible matrices are invertible. Then, the matrix $U^X \left(\tilde{\Sigma}^X \right)^{-1} \left(\tilde{V}^X \right)^\top \otimes_{\text{Kr}} U^X \left(\tilde{\Sigma}^X \right)^{-1} \left(\tilde{V}^X \right)^\top \otimes_{\text{Kr}} U^X \left(\tilde{\Sigma}^X \right)^{-1} \left(\tilde{V}^X \right)^\top \in \mathbf{R}^{d^3 \times d^3}$ is invertible. Since H is a subset of columns of this matrix, the vectors in H are linearly independent.

Rank lower bound. Let $i \in [d], j \in [r]$. By Lemma C.26, we have that

$$\begin{aligned} & \nabla F_{\text{sym}}(X)^\top (T_i \otimes_{\text{Kr}} T_j \otimes_{\text{Kr}} T_j) \\ &= \text{vec} \left(\left(T_i \left(T_j^\top \otimes_{\text{Kr}} T_j^\top \right) + T_j \left(T_i^\top \otimes_{\text{Kr}} T_j^\top \right) + T_j \left(T_i^\top \otimes_{\text{Kr}} T_j^\top \right) \right) \left[X_1 \otimes_{\text{Kr}} X_1 \quad \cdots \quad X_r \otimes_{\text{Kr}} X_r \right] \right) \\ &\stackrel{(i)}{=} \text{vec} \left(T_i \begin{bmatrix} \langle T_j, X_1 \rangle^2 \\ \vdots \\ \langle T_j, X_r \rangle^2 \end{bmatrix}^\top + 2T_j \begin{bmatrix} \langle T_i, X_1 \rangle \langle T_j, X_1 \rangle \\ \vdots \\ \langle T_i, X_r \rangle \langle T_j, X_r \rangle \end{bmatrix}^\top \right) \\ &\stackrel{(ii)}{=} \text{vec} \left(T_i \begin{bmatrix} \langle \tilde{V}_{j:}^X, \tilde{V}_{1:}^X \rangle^2 \\ \vdots \\ \langle \tilde{V}_{j:}^X, \tilde{V}_{r:}^X \rangle^2 \end{bmatrix}^\top + 2T_j \begin{bmatrix} \langle \tilde{V}_{i:}^X, \tilde{V}_{1:}^X \rangle \langle \tilde{V}_{j:}^X, \tilde{V}_{1:}^X \rangle \\ \vdots \\ \langle \tilde{V}_{i:}^X, \tilde{V}_{r:}^X \rangle \langle \tilde{V}_{j:}^X, \tilde{V}_{r:}^X \rangle \end{bmatrix}^\top \right) \\ &\stackrel{(iii)}{=} e_j \otimes_{\text{Kr}} T_i + 2\mathbb{1}_{i=j} (e_i \otimes_{\text{Kr}} T_j) \\ &= \begin{cases} 3e_i \otimes_{\text{Kr}} T_i & \text{if } i = j, \\ e_j \otimes_{\text{Kr}} T_i & \text{otherwise} \end{cases} \in \mathbf{R}^{dr}, \end{aligned}$$

where (i) follows from (69), (ii) follows from $\langle T_i, X_\ell \rangle = \tilde{V}_{i:}^X (\tilde{\Sigma}^X)^{-1} (U^X)^\top U^X (\tilde{\Sigma}^X) \left(\tilde{V}_{\ell:}^X \right)^\top = \langle \tilde{V}_{i:}^X, \tilde{V}_{\ell:}^X \rangle$, and (iii) follows from (70), here $\{e_j\}_{j \in [r]} \subseteq \mathbf{R}^r$ denotes the canonical basis for \mathbf{R}^r . The resulting dr vectors are scaled versions of different columns of the invertible matrix $I_r \otimes_{\text{Kr}} \left(U^X \left(\tilde{\Sigma}^X \right)^{-1} \left(\tilde{V}^X \right)^\top \right)$, thus they are linearly independent, which completes the proof of Proposition C.27. \square

This concludes the proof of Theorem 5.16.

C.6.2 Proof of Theorem 5.17

Let $W \in \mathbf{R}^{d_1 \times r}$, $X \in \mathbf{R}^{d_2 \times r}$ and $Y \in \mathbf{R}^{d_3 \times r}$ be arbitrary matrices. Denote the full SVD factorization of each one of these matrices as

$$W = U^W \Sigma^W \left(V^W \right)^\top, \quad X = U^X \Sigma^X \left(V^X \right)^\top \quad \text{and} \quad Y = U^Y \Sigma^Y \left(V^Y \right)^\top. \quad (58)$$

Moreover, have $\nabla F_{\text{asym}}(W, X, Y) = U \Sigma V^\top \in \mathbf{R}^{d_1 d_2 d_3 \times (d_1 + d_2 + d_3)r}$. We start with a basic result for the analytical expression of the Jacobian.

Lemma C.28 (Corollary 4.2 in [1], Lemma 1 in [51]). *The Jacobian of F_{asym} is given by*

$$\nabla F_{\text{asym}}(W, X, Y) = \begin{pmatrix} J^W & J^X & J^Y \end{pmatrix} \in \mathbf{R}^{d_1 d_2 d_3 \times (d_1 + d_2 + d_3)r}$$

where

$$\begin{aligned} J^W &= I_{d_1} \otimes_{\text{Kr}} \psi(X, Y) \in \mathbf{R}^{d_1 d_2 d_3 \times d_1 r}, \\ J^X &= P_2 (I_{d_2} \otimes_{\text{Kr}} \psi(W, Y)) \in \mathbf{R}^{d_1 d_2 d_3 \times d_2 r}, \text{ and} \\ J^Y &= P_3 (I_{d_3} \otimes_{\text{Kr}} \psi(W, X)) \in \mathbf{R}^{d_1 d_2 d_3 \times d_3 r}, \end{aligned}$$

with P_i and ψ introduced in Definitions C.23 and C.25, respectively.

Consider an arbitrary pair $(W, X, Y), (\tilde{W}, \tilde{X}, \tilde{Y}) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r} \times \mathbf{R}^{d_3 \times r}$ satisfying

$$\max \left\{ \|(W, X, Y) - (W^*, X^*, Y^*)\|_F, \|(\tilde{W}, \tilde{X}, \tilde{Y}) - (W^*, X^*, Y^*)\|_F \right\} \leq \|(W^*, X^*, Y^*)\|_F.$$

Using the variational characterization of the operator norm, we have

$$\begin{aligned} & \left\| \nabla F_{\text{asym}}(W, X, Y) - \nabla F_{\text{asym}}(\tilde{W}, \tilde{X}, \tilde{Y}) \right\|_{\text{op}} \\ &= \sup_{\|A\|_F=1} \left\| \left(\nabla F_{\text{asym}}(W, X, Y) - \nabla F_{\text{asym}}(\tilde{W}, \tilde{X}, \tilde{Y}) \right) \text{vec}(A) \right\|_2 \\ &= \sup_{\substack{(A_1, A_2, A_3) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r} \times \mathbf{R}^{d_3 \times r} \\ \|(A_1, A_2, A_3)\|_F=1}} \left\| \left(J^W - J^{\tilde{W}} \right) \text{vec}(A_1) + \left(J^X - J^{\tilde{X}} \right) \text{vec}(A_2) + \left(J^Y - J^{\tilde{Y}} \right) \text{vec}(A_3) \right\|_2 \\ &\stackrel{(i)}{\leq} \sup_{\substack{(A_1, A_2, A_3) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r} \times \mathbf{R}^{d_3 \times r} \\ \|(A_1, A_2, A_3)\|_F=1}} \left(\left\| \text{vec} \left(\left(\psi(X, Y) - \psi(\tilde{X}, \tilde{Y}) \right) A_1^\top \right) \right\|_2 \right. \\ &\quad \left. + \left\| P_2 \text{vec} \left(\left(\psi(W, Y) - \psi(\tilde{W}, \tilde{Y}) \right) A_2^\top \right) \right\|_2 \right. \\ &\quad \left. + \left\| P_3 \text{vec} \left(\left(\psi(W, X) - \psi(\tilde{W}, \tilde{X}) \right) A_3^\top \right) \right\|_2 \right), \end{aligned}$$

where (i) follows from applying Lemma C.28 in tandem with the triangle inequality and from (67). Leveraging the fact that the operator norm of a permutation matrix is at most one, we derive

$$\begin{aligned} & \left\| \nabla F_{\text{asym}}(W, X, Y) - \nabla F_{\text{asym}}(\tilde{W}, \tilde{X}, \tilde{Y}) \right\|_{\text{op}} \\ &\leq \sup_{\|(A_1, A_2, A_3)\|_F=1} \left(\left\| \left(\psi(X, Y) - \psi(\tilde{X}, \tilde{Y}) \right) A_1^\top \right\|_F \right. \\ &\quad \left. + \left\| \left(\psi(W, Y) - \psi(\tilde{W}, \tilde{Y}) \right) A_2^\top \right\|_F \right. \\ &\quad \left. + \left\| \left(\psi(W, X) - \psi(\tilde{W}, \tilde{X}) \right) A_3^\top \right\|_F \right) \\ &\leq \left\| \psi(X, Y) - \psi(\tilde{X}, \tilde{Y}) \right\|_F + \left\| \psi(W, Y) - \psi(\tilde{W}, \tilde{Y}) \right\|_F + \left\| \psi(W, X) - \psi(\tilde{W}, \tilde{X}) \right\|_F \\ &= \sum_{\ell=1}^r \left(\left\| \left(X_\ell - \tilde{X}_\ell \right) \otimes_{\text{Kr}} Y_\ell + \tilde{X}_\ell \otimes_{\text{Kr}} \left(Y_\ell - \tilde{Y}_\ell \right) \right\|_2 \right. \\ &\quad \left. + \left\| \left(W_\ell - \tilde{W}_\ell \right) \otimes_{\text{Kr}} Y_\ell + \tilde{W}_\ell \otimes_{\text{Kr}} \left(Y_\ell - \tilde{Y}_\ell \right) \right\|_2 \right. \\ &\quad \left. + \left\| \left(W_\ell - \tilde{W}_\ell \right) \otimes_{\text{Kr}} X_\ell + \tilde{W}_\ell \otimes_{\text{Kr}} \left(X_\ell - \tilde{X}_\ell \right) \right\|_2 \right) \\ &\stackrel{(i)}{\leq} \left(\|W\|_F + \|X\|_F + \|Y\|_F + \|\tilde{W}\|_F + \|\tilde{X}\|_F + \|\tilde{Y}\|_F \right) \left(\|W - \tilde{W}\|_F + \|X - \tilde{X}\|_F + \|Y - \tilde{Y}\|_F \right) \end{aligned}$$

$$\begin{aligned}
&\stackrel{(ii)}{\leq} \sqrt{3} \left(\|(W, X, Y)\|_F + \|(\tilde{W}, \tilde{X}, \tilde{Y})\|_F \right) \|(W, X, Y) - (\tilde{W}, \tilde{X}, \tilde{Y})\|_F \\
&\stackrel{(iii)}{\leq} 4\sqrt{3} \|(W^*, X^*, Y^*)\|_F \|(W, X, Y) - (\tilde{W}, \tilde{X}, \tilde{Y})\|_F,
\end{aligned}$$

where (i) follows from the fact that $\|v \otimes_{\mathbf{K}r} w\|_2 = \|v\|_2 \|w\|_2$ for any vectors v and w together with Cauchy-Schwarz and (ii) uses that $(|a| + |b| + |c|)^2 \leq 3(|a|^2 + |b|^2 + |c|^2)$ for $a, b, c \in \mathbf{R}$, and (iii) holds since by assumption $\max \left\{ \|(W, X, Y) - (W^*, X^*, Y^*)\|_F, \|(\tilde{W}, \tilde{X}, \tilde{Y}) - (W^*, X^*, Y^*)\|_F \right\} \leq \|(W^*, X^*, Y^*)\|_F$, which implies $\max \left\{ \|(W, X, Y)\|_F, \|(\tilde{W}, \tilde{X}, \tilde{Y})\|_F \right\} \leq 2 \|(W^*, X^*, Y^*)\|_F$ by the reverse triangle inequality.

We now proceed to prove Assumption 8 for this map. Let $T^* \in \mathbf{R}^{d_1 \times d_2 \times d_2}$ be an arbitrary tensor and let $W^* \in \mathbf{R}^{d_1 \times r}$, $X^* \in \mathbf{R}^{d_2 \times r}$ and $Y^* \in \mathbf{R}^{d_3 \times r}$ be any full-rank matrices such that $T^* = F_{\text{asym}}(W^*, X^*, Y^*)$.

Proposition C.29 (Constant Rank of Asymmetric Canonical Polyadic Map). *For any full rank matrices $W \in \mathbf{R}^{d_1 \times r}$, $X \in \mathbf{R}^{d_2 \times r}$ and $Y \in \mathbf{R}^{d_3 \times r}$, we have*

$$\text{rank}(\nabla F_{\text{asym}}(W, X, Y)) = (d_1 + d_2 + d_3 - 2)r.$$

Equipped with this constant rank result, we invoke Lemma C.1 to establish that the map F_{asym} satisfies Assumption 8 with the quantities

$$\begin{aligned}
\varepsilon_{x^*} &= \min \left\{ R, \frac{\sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Z^*))}{8\sqrt{3} \|(W^*, X^*, Y^*)\|_F}, \|(W^*, X^*, Y^*)\|_F \right\}, \quad \delta(\rho) = \frac{\rho}{C} \\
\text{and} \quad s &= \frac{1}{2} \sigma_{(d_1+d_2+d_3-2)r}(\nabla F_{\text{asym}}(W^*, X^*, Y^*))
\end{aligned}$$

for some positive constants R and C that depend only on the solution (W^*, X^*, Y^*) and with $j = \text{rank}(\nabla F_{\text{asym}}(W^*, X^*, Y^*)) = (d_1 + d_2 + d_3 - 2)r$. This establishes the result, if we prove Proposition C.29.

Proof of Proposition C.29. The proof structure involves three steps. For the first step, we will construct two sets H and H^c of vectors with cardinalities $(d_1 + d_3 + d_3 - 2)r$ and $d_1 d_2 d_3 - (d_1 + d_3 + d_3 - 2)r$, respectively, such that $H \cup H^c$ forms a basis for $\mathbf{R}^{d_1 d_2 d_3}$. For the second step, we will show that the set $H^c \subseteq \text{null}(\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top)$. This establishes an upper bound of $(d_1 + d_3 + d_3 - 2)r$ on $\text{rank}(\nabla F_{\text{asym}}(W, X, Y))$. For the final step, we will prove that the set H remains linearly independent after applying $\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top$. This establishes a lower bound of $(d_1 + d_3 + d_3 - 2)r$ on $\text{rank}(\nabla F_{\text{asym}}(W, X, Y))$, finishing the proof.

To start, we introduce some notation. Define the index sets

$$\begin{aligned}
I_0 &:= \{(l, l, l) \mid l \in [r]\}, \\
I_1 &:= \{(l, l, k) \mid l \in [r], k \in [d_3], l \neq k\}, \\
I_2 &:= \{(l, k, l) \mid l \in [r], k \in [d_2], l \neq k\}, \quad \text{and} \\
I_3 &:= \{(k, l, l) \mid l \in [r], k \in [d_1], l \neq k\}.
\end{aligned} \tag{59}$$

In what follows, we use M and N as placeholders for W, X , or Y . Further, d_M and d_N denote the number of rows of M and N , respectively. Define the sets

$$\begin{aligned}
\mathcal{T}_{\text{off}}^{M,N} &:= \{(i, j) \in [d_M] \times [d_N] \mid i > r \text{ or } j > r \text{ or } i \neq j\}, \quad \text{and} \\
\mathcal{T}_{\text{on}}^{M,N} &:= (\mathcal{T}_{\text{off}}^{M,N})^c = \{(i, j) \in [d_M] \times [d_N] \mid i = j \text{ and } i \leq r\}.
\end{aligned} \tag{60}$$

The sets in (59) form a partition of $I := \{(i, j, k) \mid (i, j) \in \mathcal{T}_{\text{on}}^{W,X} \text{ or } (j, k) \in \mathcal{T}_{\text{on}}^{X,Y} \text{ or } (i, k) \in \mathcal{T}_{\text{on}}^{W,Y}\}$.

Probing sets. For $M \in \{W, X, Y\}$, define

$$T_i^M = U^M (\tilde{\Sigma}^M)^{-1} (\tilde{V}_i^M)^\top, \quad (61)$$

where $\tilde{\Sigma}^M = \text{diag}(\sigma_1^M, \dots, \sigma_r^M, 1, \dots, 1) \in \mathbf{R}^{d_M \times d_M}$, and $\tilde{V}^M = \begin{pmatrix} V^M & 0 \\ 0 & I \end{pmatrix} \in \mathbf{R}^{d_M \times d_M}$, where U^M and $(V^M)^\top$ are the left and right eigenvectors of M , and σ_i 's are its singular values. We construct $H := \{T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \mid (i, j, k) \in I\}$ and $H^c := \{T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \mid (i, j, k) \in [d_1] \times [d_2] \times [d_3] \setminus I\}$. These sets are linearly independent since they correspond to column vectors of the matrix

$$U^W (\tilde{\Sigma}^W)^{-1} (\tilde{V}^W)^\top \otimes_{\text{Kr}} U^X (\tilde{\Sigma}^X)^{-1} (\tilde{V}^X)^\top \otimes_{\text{Kr}} U^Y (\tilde{\Sigma}^Y)^{-1} (\tilde{V}^Y)^\top \in \mathbf{R}^{d_1 d_2 d_3 \times d_1 d_2 d_3}$$

which again is invertible by (69). A relevant property about these vectors is

$$\langle M_i, T_j^M \rangle = \mathbb{1}_{i=j}. \quad (62)$$

Upper bound. We use the following lemma; whose proof is deferred to Appendix D.5.

Lemma C.30. *Let $(W, X, Y) \in \mathbf{R}^{d_1 \times r} \times \mathbf{R}^{d_2 \times r} \times \mathbf{R}^{d_3 \times r}$ be full-rank matrices. Then,*

$$\begin{aligned} & \nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \left(T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \right) \\ &= \mathbb{1}_{(i,j) \in \mathcal{T}_{\text{on}}^{W,X}} \left(W_i \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} T_k^Y \right) + \mathbb{1}_{(i,k) \in \mathcal{T}_{\text{on}}^{W,Y}} \left(W_i \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} Y_k \right) + \mathbb{1}_{(j,k) \in \mathcal{T}_{\text{on}}^{X,Y}} \left(T_i^W \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} Y_k \right). \end{aligned}$$

By Lemma C.30, it is easy to see that $\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \left(T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \right)$ is zero when $T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \in H^c$. Therefore,

$$\text{span}(H^c) \subseteq \text{null} \left(\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \right).$$

Since H^c is linearly independent, and since $\#H^c = d_1 d_2 d_3 - (d_1 + d_2 + d_3 - 2)r$, then $\dim(\text{span}(H^c)) = d_1 d_2 d_3 - (d_1 + d_2 + d_3 - 2)r$. Therefore,

$$d_1 d_2 d_3 - (d_1 + d_2 + d_3 - 2)r \leq \dim \left(\text{null} \left(\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \right) \right).$$

Then, the rank-nullity theorem yields

$$\text{rank}(\nabla F_{\text{asym}}(W, X, Y)) = \text{rank} \left(\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \right) \leq (d_1 + d_2 + d_3 - 2)r.$$

Lower bound. We will show that the elements in H remain linearly independent after applying $\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top$, which shows that $\text{rank} \left(\nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \right) \geq (d_1 + d_2 + d_3 - 2)r$. By Lemma C.30, we have that elements of such a set are given by

$$\begin{aligned} & \nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top \left(T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \right) \\ &= \begin{cases} W_i \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} T_k^Y + W_i \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} Y_k + T_i^W \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} Y_k, & \text{if } (i, j, k) \in I_0, \\ W_i \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} T_k^Y, & \text{if } (i, j, k) \in I_1, \\ W_i \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} Y_k, & \text{if } (i, j, k) \in I_2, \\ T_i^W \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} Y_k, & \text{if } (i, j, k) \in I_3, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Recall that $T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \in H$ if $(i, j, k) \in I = I_0 \cup I_1 \cup I_2 \cup I_3$. To show that these vectors are linearly independent, we will show that any linear combination of them that equates to zero has

to have zero coefficients. Thus, suppose that we have vectors of coefficients $\alpha, \beta, \gamma, \delta$ such that the following linear combination is equal to zero

$$\begin{aligned}
L^H &:= \sum_{l=1}^r \alpha_l \left(W_l \otimes_{\mathbf{K}_r} X_l \otimes_{\mathbf{K}_r} T_l^Y + W_l \otimes_{\mathbf{K}_r} T_l^X \otimes_{\mathbf{K}_r} Y_l + T_l^W \otimes_{\mathbf{K}_r} X_l \otimes_{\mathbf{K}_r} Y_l \right) \\
&\quad + \sum_{l \in [r], k \in [d_3] \setminus \{l\}} \beta_{l,k} \left(W_l \otimes_{\mathbf{K}_r} X_l \otimes_{\mathbf{K}_r} T_k^Y \right) \\
&\quad + \sum_{l \in [r], k \in [d_2] \setminus \{l\}} \gamma_{l,k} \left(W_l \otimes_{\mathbf{K}_r} T_k^X \otimes_{\mathbf{K}_r} Y_l \right) \\
&\quad + \sum_{l \in [r], k \in [d_1] \setminus \{l\}} \delta_{l,k} \left(T_k^W \otimes_{\mathbf{K}_r} X_l \otimes_{\mathbf{K}_r} Y_l \right) = 0.
\end{aligned} \tag{63}$$

For the rest of the proof, we focus on showing that the coefficients $\alpha_l, \beta_{l,k}, \gamma_{l,k}, \delta_{l,k} \in \mathbf{R}$ are all zero. To this end, we probe the equality above with several linear maps, which allows us to derive conclusions for specific coefficients. In particular, for fixed $i \in [r]$, $j \in [d_3] \setminus \{i\}$, we apply the linear transformation $I_{d_1} \otimes_{\mathbf{K}_r} T_i^X \otimes_{\mathbf{K}_r} T_j^Y$ on both sides of the equality, which yields

$$\begin{aligned}
0 &= \left(I_{d_1} \otimes_{\mathbf{K}_r} T_i^X \otimes_{\mathbf{K}_r} T_j^Y \right) L^H \\
&\stackrel{(i)}{=} \sum_{l=1}^r \alpha_l \left(\langle X_l, T_i^X \rangle \langle Y_l, T_j^Y \rangle + \langle X_l, T_i^X \rangle \langle T_l^Y, T_j^Y \rangle + \langle T_l^X, T_i^X \rangle \langle Y_l, T_j^Y \rangle \right) W_l \\
&\quad + \sum_{l \in [r], k \in [d_3] \setminus \{l\}} \beta_{l,k} \left(\langle X_l, T_i^X \rangle \langle T_k^Y, T_j^Y \rangle \right) W_l \\
&\quad + \sum_{l \in [r], k \in [d_2] \setminus \{l\}} \gamma_{l,k} \left(\langle T_k^X, T_i^X \rangle \langle Y_l, T_j^Y \rangle \right) W_l \\
&\quad + \sum_{l \in [r], k \in [d_1] \setminus \{l\}} \delta_{l,k} \left(\langle X_l, T_i^X \rangle \langle Y_l, T_j^Y \rangle \right) T_k^W \\
&\stackrel{(ii)}{=} \langle \alpha_i T_i^Y, T_j^Y \rangle W_i + \langle \alpha_j T_j^X, T_i^X \rangle W_j \mathbb{1}_{j \leq r} \\
&\quad + \left\langle \sum_{k \in [d_3] \setminus \{i\}} \beta_{i,k} T_k^Y, T_j^Y \right\rangle W_i + \left\langle \sum_{k \in [d_2] \setminus \{j\}} \gamma_{j,k} T_k^X, T_i^X \right\rangle W_j \mathbb{1}_{j \leq r},
\end{aligned} \tag{64}$$

where (i) follows from (69) and (ii) use the choice of indices together with (62). Since W has full column rank, the vectors $\{W_k \mid k \in [r]\}$ are linearly independent. Thus, (64) implies that the coefficient associated with the vector W_i is zero. Consequently, $\langle \alpha_i T_i^Y + \sum_{k \in [d_3] \setminus \{i\}} \beta_{i,k} T_k^Y, T_j^Y \rangle = 0$, which can be rewritten equivalently as

$$\langle T^Y \omega^{\alpha, \beta, i}, T_j^Y \rangle = 0 \quad \text{where} \quad \omega^{\alpha, \beta, i} := (\beta_{i,1}, \dots, \beta_{i,i-1}, \alpha_i, \beta_{i,i+1}, \dots, \beta_{i,d_3})^\top.$$

Since this holds for all $j \in [d_3] \setminus \{i\}$, then $T^Y \omega^{\alpha, \beta, i}$ is orthogonal to all columns of T^Y except the i -th column. The vector Y_i is also orthogonal to all these T_j^Y due to (62). Thus, $T^Y \omega^{\alpha, \beta, i}$ and Y_i are orthogonal to the same $d_3 - 1$ dimensional subspace, consequently colinear. Hence, for any $i \in [r]$, there exists $c^{\alpha, \beta, i} \in \mathbf{R} \setminus \{0\}$ such that $T^Y \omega^{\alpha, \beta, i} = c^{\alpha, \beta, i} Y_i$ and, consequently,

$$\omega^{\alpha, \beta, i} = c^{\alpha, \beta, i} \left(T^Y \right)^{-1} Y_i = c_i^{\alpha, \beta} \tilde{V}^Y \left(\tilde{\Sigma}^Y \right)^2 \left(\tilde{V}_{i:}^Y \right)^\top.$$

By an equivalent argument, for any $i \in [r]$, there exist $c^{\alpha, \gamma, i}, c^{\alpha, \delta, i} \in \mathbf{R}$ such that

$$\omega^{\alpha, \gamma, i} = c^{\alpha, \gamma, i} \tilde{V}^X \left(\tilde{\Sigma}^X \right)^2 \left(\tilde{V}_{i:}^X \right)^\top, \quad \text{and} \quad \omega^{\alpha, \delta, i} = c^{\alpha, \delta, i} \tilde{V}^W \left(\tilde{\Sigma}^W \right)^2 \left(\tilde{V}_{i:}^W \right)^\top. \tag{65}$$

If we prove these vectors are zero, then the original coefficients in (63) would also be zero. Equipped with this closed-form expression for the coefficients, we derive

$$\begin{aligned}
L^H &= \sum_{l=1}^r \left(W_l \otimes_{\text{Kr}} X_l \otimes_{\text{Kr}} \left(\alpha_l T_l^Y + \sum_{k \in [d_3] \setminus \{l\}} \beta_{l,k} T_k^Y \right) \right) \\
&\quad + \sum_{l=1}^r \left(W_l \otimes_{\text{Kr}} \left(\alpha_l T_l^X + \sum_{k \in [d_2] \setminus \{l\}} \gamma_{l,k} T_k^X \right) \otimes_{\text{Kr}} Y_l \right) \\
&\quad + \sum_{l=1}^r \left(\left(\alpha_l T_l^W + \sum_{k \in [d_1] \setminus \{l\}} \delta_{l,k} T_k^W \right) \otimes_{\text{Kr}} X_l \otimes_{\text{Kr}} Y_l \right) \\
&= \sum_{l=1}^r \left(W_l \otimes_{\text{Kr}} X_l \otimes_{\text{Kr}} T^Y \omega^{\alpha,\beta,l} \right) + \left(W_l \otimes_{\text{Kr}} T^X \omega^{\alpha,\gamma,l} \otimes_{\text{Kr}} Y_l \right) + \left(T^W \omega^{\alpha,\delta,l} \otimes_{\text{Kr}} X_l \otimes_{\text{Kr}} Y_l \right) \\
&= \sum_{l=1}^r \left(c^{\alpha,\beta,l} + c^{\alpha,\gamma,l} + c^{\alpha,\delta,l} \right) \left(W_l \otimes_{\text{Kr}} X_l \otimes_{\text{Kr}} Y_l \right), \tag{66}
\end{aligned}$$

where the first equality follows from distributing the sum over l and rearranging, the second equality rewrites $\alpha_l T_l^Y + \sum_{k \in [d_3] \setminus \{l\}} \beta_{l,k} T_k^Y$ as $T^Y \omega^{\alpha,\beta,l}$, and the third equality follows from the definition of $c^{\alpha,\beta,l}$, $c^{\alpha,\gamma,l}$, and $c^{\alpha,\delta,l}$ and the multilinearity of the Kronecker product.

Since the vectors $\{W_l \otimes_{\text{Kr}} X_l \otimes_{\text{Kr}} Y_l \mid l \in [r]\}$ are linearly independent, we conclude that $c^{\alpha,\beta,l} + c^{\alpha,\gamma,l} + c^{\alpha,\delta,l} = 0$ for all l . Let's show that each term is also zero. Using (65), we can extract the l th component from $\omega^{\alpha,\beta,l}$, $\omega^{\alpha,\gamma,l}$, and $\omega^{\alpha,\delta,l}$ via

$$\alpha_l = c^{\alpha,\beta,l} \tilde{V}_l^Y \left(\tilde{\Sigma}^Y \right)^2 \left(\tilde{V}_l^Y \right)^\top = c^{\alpha,\gamma,l} \tilde{V}_l^X \left(\tilde{\Sigma}^X \right)^2 \left(\tilde{V}_l^X \right)^\top = c^{\alpha,\delta,l} \tilde{V}_l^W \left(\tilde{\Sigma}^W \right)^2 \left(\tilde{V}_l^W \right)^\top.$$

Observe that all the quadratic forms in these equalities are strictly positive because $\tilde{\Sigma}^W$, $\tilde{\Sigma}^X$, and $\tilde{\Sigma}^Y$ are positive definite matrices. Thus $c^{\alpha,\beta,l}$, $c^{\alpha,\gamma,l}$ and $c^{\alpha,\delta,l}$ have the same sign, which in turn implies that $c^{\alpha,\beta,l} + c^{\alpha,\gamma,l} + c^{\alpha,\delta,l} = 0$ if, and only if, $c^{\alpha,\beta,l} = c^{\alpha,\gamma,l} = c^{\alpha,\delta,l} = 0$. Thus, by (65), that $\omega^{\alpha,\beta,l} = \omega^{\alpha,\gamma,l} = \omega^{\alpha,\delta,l} = 0$ and, consequently all the coefficients in (63) are also zero. Hence, the vectors $\{\nabla F_{\text{asym}}(W, X, Y)^\top \nabla F_{\text{asym}}(W, X, Y)v \mid v \in H\}$ are linearly independent; this finishes the proof of Proposition C.29. \square

This concludes the proof of Theorem 5.17.

D Auxiliary proofs and results

In this section, we summarize some auxiliary results that we use throughout the paper.

Lemma D.1 (Properties of the Kronecker Product). *[[89]] Let A, B, C, D be matrices (or vectors) of compatible dimensions. Then, the Kronecker product satisfies the following properties.*

1. *Multiplication from the right and left can be succinctly written as*

$$\text{vec}(ABC) = \left(C^\top \otimes_{\text{Kr}} A \right) \text{vec}(B). \tag{67}$$

2. *The transpose commutes with the Kronecker product*

$$(A \otimes_{\text{Kr}} B)^\top = A^\top \otimes_{\text{Kr}} B^\top. \tag{68}$$

3. *The matrix product commutes with the Kronecker product*

$$(AB) \otimes_{\text{Kr}} (CD) = (A \otimes_{\text{Kr}} C)(B \otimes_{\text{Kr}} D). \tag{69}$$

4. Trivially, if $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, the vectorized outer product is equal to the Kronecker product

$$y \otimes_{\text{Kr}} x = \text{vec} \left(xy^\top \right). \quad (70)$$

Lemma D.2. Let $U \in \mathbf{R}^{d \times d}$ be an orthogonal matrix. Then for any $I \subset J \subseteq [d]$ and vector $v \in \mathbf{R}^d$ we have

$$\left\| U_I U_I^\top v \right\|_2 \leq \left\| U_J U_J^\top v \right\|_2.$$

Proof. We have

$$\left\| U_J U_J^\top v \right\|_2^2 = \langle v, U_J U_J^\top v \rangle = \langle v, U_{J \setminus I} U_{J \setminus I}^\top v \rangle + \langle v, U_I U_I^\top v \rangle \geq \left\| U_I U_I^\top v \right\|_2^2,$$

where the first equality follows since orthogonal projections are symmetric and idempotent. \square

Lemma D.3 (Lemma 2.5 in [21]). Suppose $U = [U_0, U_1]$ and $V = [V_0, V_1]$ are square orthonormal matrices, where $U_0, V_0 \in \mathbb{R}^{d \times k}$. Let $\theta_1, \dots, \theta_k$ denote the principal angles between $\text{span}(U_0)$ and $\text{span}(V_0)$, and denote $\Theta(U_0, V_0) = \text{diag}(\theta_1, \dots, \theta_k)$. Then

$$\left\| U_0^\top V_1 \right\|_F = \left\| \sin \Theta(U_0, V_0) \right\|_F. \quad (71)$$

Lemma D.4 (Lemma 2.6 in [21]). Let $U, V \in \mathbf{R}^{d \times k}$ ($k \leq d$) have orthonormal columns. Write the principal angles between $\text{span}(U)$ and $\text{span}(V)$ as $\theta_1, \dots, \theta_k$ and denote $\Theta(U, V) = \text{diag}(\theta_1, \dots, \theta_k)$. Then,

$$\min_{Q \in O(k)} \|U - VQ\|_F \leq \sqrt{2} \|\sin \Theta(U, V)\|_F.$$

D.1 Proof of Lemma C.8

Eigenpairs. For any pair of indexes (i, j) with $i \leq j$, label $Z = U_i^X U_j^{X^\top} + U_j^X U_i^{X^\top}$. Then,

$$\begin{aligned} \nabla F_{\text{sym}}(X) \nabla F_{\text{sym}}(X)^\top [Z] &= 2 \nabla F_{\text{sym}}(X) [ZX] \\ &= 2ZXX^\top + 2XX^\top Z \\ &= 2 \left(Z \sum_{k=1}^n \sigma_k^2(X) U_k^X U_k^{X^\top} + \sum_{k=1}^n \sigma_k^2(X) U_k^X U_k^{X^\top} Z \right) \\ &= 2 \left(\sigma_i^2(X) \left(U_i^X U_j^{X^\top} + U_j^X U_i^{X^\top} \right) + \sigma_j^2(X) \left(U_i^X U_j^{X^\top} + U_j^X U_i^{X^\top} \right) \right) \\ &= 2 \left(\sigma_i^2(X) + \sigma_j^2(X) \right) Z, \end{aligned}$$

where the first two lines follow from (43) and the last two lines follow by definition of Z .

Orthonormal basis. Recall that the image of ∇F is \mathcal{S}^d , thus the number of eigenvectors above matches the number of dimensions of \mathcal{S}^d . It suffices to prove that they are orthogonal. Let $U'_{i,j}$ be a placeholder for

$$U'_{i,j} = \text{vec} \left(U_i^X U_j^{X^\top} + U_j^X U_i^{X^\top} \right).$$

Let $(i, j), (k, \ell) \in [d] \times [d]$ with $i \leq j$ and $k \leq \ell$. We shall show that $U'_{i,j}$ is nonzero, and if $(i, j) \neq (k, \ell)$, then $\langle U'_{i,j}, U'_{k,\ell} \rangle = 0$. Then,

$$\begin{aligned} \langle U'_{i,j}, U'_{k,\ell} \rangle &= \left\langle U_i^X U_j^{X^\top} + U_j^X U_i^{X^\top}, U_k^X U_\ell^{X^\top} + U_\ell^X U_k^{X^\top} \right\rangle \\ &= \text{trace} \left(U_i^X U_j^{X^\top} U_\ell^X U_k^{X^\top} \right) + \text{trace} \left(U_i^X U_j^{X^\top} U_k^X U_\ell^{X^\top} \right) \end{aligned}$$

$$\begin{aligned}
& + \text{trace} \left(U_j^X U_i^{X^\top} U_\ell^X U_k^{X^\top} \right) + \text{trace} \left(U_j^X U_i^{X^\top} U_k^X U_\ell^{X^\top} \right) \\
& = 2 \langle U_i^X, U_k^X \rangle \langle U_j^X, U_\ell^X \rangle + 2 \langle U_i^X, U_\ell^X \rangle \langle U_j^X, U_k^X \rangle.
\end{aligned}$$

First if $(i, j) = (k, \ell)$, then $\|U'_{i,j}\|_2^2 = \begin{cases} 2, & i \neq j \\ 4, & \text{otherwise} \end{cases} \neq 0$ and so $U'_{i,j}$ is nonzero. Second, if $(i, j) \neq (k, \ell)$, then either $i \neq k$ or $j \neq \ell$. Without loss of generality, assume that $i \neq k$; thus, the first term in the last expression is zero. Seeking contradiction, assume that the second term is nonzero. Since $\{U_i^X\}$ forms an orthonormal basis, we derive that $i = \ell \geq k = j$ and by assumption $i \leq j$, therefore, $i = j = k = \ell$, which is a contradiction. This completes the proof of the lemma.

D.2 Proof of Lemma C.16

Recall that we denote by r^* the ranks of X^* and Y^* . One has

$$\begin{aligned}
\|XX - X^*X^{*\top}\|_F & \leq \|(X - X^*)X^\top + X^*(X - X^*)^\top\|_F \\
& \stackrel{(i)}{\leq} \|(X - X^*)X^\top\|_F + \|X^*(X - X^*)^\top\|_F \\
& \stackrel{(ii)}{\leq} 2 \max\{\sigma_1(X^*), \sigma_1(X)\} \|X - X^*\|_F \\
& \stackrel{(iii)}{\leq} 2(\sigma_1(X^*) + \|X - X^*\|_F) \|X - X^*\|_F \\
& \stackrel{(iv)}{\leq} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{8\sqrt{2}} + \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{8\sqrt{2}\sigma_1(X^*)} \|X - X^*\|_F \\
& \stackrel{(v)}{\leq} \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{4\sqrt{2}},
\end{aligned}$$

where (i) follows from the triangle inequality, (ii) follows from the variational characterization of singular values, (iii) follows from Weyl's inequality, (iv) holds since $\|X - X^*\|_F \leq \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{16\sqrt{2}\sigma_1(X^*)}$ and (v) holds since $\frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{16\sqrt{2}\sigma_1(X^*)} \leq \sigma_1(X^*)$. A similar argument yields

$$\|YY - Y^*Y^{*\top}\|_F \leq \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{4\sqrt{2}}.$$

Adding both bounds, we conclude that $\|XX - X^*X^{*\top}\|_F + \|YY - Y^*Y^{*\top}\|_F \leq \frac{\min\{\sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*)\}}{2\sqrt{2}}$.

D.3 Proof of Lemma C.15

Eigenpairs. Let $i, j \in [d_1] \times [d_2]$ with corresponding eigenpairs $(\sigma_i^2(X), U_i^X)$ and $(\sigma_j^2(Y), U_j^Y)$ for respectively XX^\top and YY^\top . Let M be a placeholder for $U_i^X U_j^{Y^\top}$. We have

$$\begin{aligned}
\nabla F(X, Y) \nabla F(X, Y)^\top M & = MYY^\top + XX^\top M \\
& = \sigma_j^2(Y) \left(U_i^X U_j^{Y^\top} \right) + \sigma_i^2(X) \left(U_i^X U_j^{Y^\top} \right) \\
& = \left(\sigma_i^2(X) + \sigma_j^2(Y) \right) M,
\end{aligned}$$

where the first equality follows from (48) and the second equality is due to the orthonormality of the eigenvectors of both XX^\top and YY^\top .

Orthonormal basis. Let $(i, j), (k, \ell) \in [d_1] \times [d_2]$. We will show that if $(i, j) = (k, \ell)$, then $\langle U_i^X U_j^Y{}^\top, U_k^X U_\ell^Y{}^\top \rangle \neq 0$ and that if $(i, j) \neq (k, \ell)$, then $\langle U_i^X U_j^Y{}^\top, U_k^X U_\ell^Y{}^\top \rangle = 0$.

Case 1. $(i, j) = (k, \ell)$. Then $\langle U_i^X U_j^Y{}^\top, U_i^X U_j^Y{}^\top \rangle = \text{tr}(U_i^X U_j^Y{}^\top U_j^Y U_i^X{}^\top) = \langle U_i^X, U_i^X \rangle \langle U_j^Y, U_j^Y \rangle = 1 \neq 0$.

Case 2. $(i, j) \neq (k, \ell)$. Then $\langle U_i^X U_j^Y{}^\top, U_k^X U_\ell^Y{}^\top \rangle = \langle U_i^X, U_k^X \rangle \langle U_j^Y, U_\ell^Y \rangle = 0$.

D.4 Proof of Lemma C.26

Fix $X, D \in \mathbf{R}^{d \times r}$ arbitrary. For the action of the Jacobian, expanding the denominator in $\nabla F_{\text{sym}}(X) \text{vec}(D) = \lim_{t \downarrow 0} \frac{F_{\text{sym}}(X+tD) - F_{\text{sym}}(X)}{t}$ leads to

$$\nabla F_{\text{sym}}(X) \text{vec}(D) = \sum_{\ell=1}^r (D_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} X_\ell + X_\ell \otimes_{\text{Kr}} D_\ell \otimes_{\text{Kr}} X_\ell + X_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} D_\ell). \quad (72)$$

A straight computation shows that the permutation matrices P_2 and P_3 from Definition C.23 satisfy $P_2(a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} b) = b \otimes_{\text{Kr}} b \otimes_{\text{Kr}} a$ and $P_3(a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} b) = b \otimes_{\text{Kr}} a \otimes_{\text{Kr}} b$ for arbitrary vectors a, b , so that $\nabla F_{\text{sym}}(X) \text{vec}(D)$ can be written as $\sum_{\ell=1}^r (I + P_3 + P_2)(D_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} X_\ell)$.

For the action of the adjoint, recall that by definition, we have

$$\left\langle \nabla F_{\text{sym}}(X)^\top (a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c), \text{vec}(D) \right\rangle = \left\langle \nabla F_{\text{sym}}(X) \text{vec}(D), a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c \right\rangle.$$

Using (72), we have

$$\begin{aligned} & \left\langle \nabla F_{\text{sym}}(X) \text{vec}(D), a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c \right\rangle \\ &= \left\langle \sum_{\ell \in [r]} (X_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} D_\ell + X_\ell \otimes_{\text{Kr}} D_\ell \otimes_{\text{Kr}} X_\ell + D_\ell \otimes_{\text{Kr}} X_\ell \otimes_{\text{Kr}} X_\ell), a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c \right\rangle \\ &= \sum_{i,j,k \in [d]^3} \sum_{\ell \in [r]} (X_{i\ell} X_{j\ell} D_{k\ell} + X_{i\ell} D_{j\ell} X_{k\ell} + D_{i\ell} X_{j\ell} X_{k\ell}) a_i b_j c_k \\ &= \sum_{i,j,k \in [d]^3} \sum_{\ell \in [r]} (X_{i\ell} X_{j\ell} D_{k\ell} a_i b_j c_k + X_{i\ell} D_{j\ell} X_{k\ell} a_i b_j c_k + D_{i\ell} X_{j\ell} X_{k\ell} a_i b_j c_k). \end{aligned}$$

Next, we derive an expression for the first term of the above equation. Changing the order of summation, we have

$$\sum_{i,j,k \in [d]^3} \sum_{\ell \in [r]} X_{i\ell} X_{j\ell} D_{k\ell} a_i b_j c_k = \sum_{k \in [d], \ell \in [r]} D_{k\ell} \left(c_k \sum_{i,j \in [d]^2} X_{i\ell} X_{j\ell} a_i b_j \right).$$

The matrix M with components $M_{k\ell} = c_k \sum_{i,j \in [d]^2} X_{i\ell} X_{j\ell} a_i b_j$ can be compactly written as $M = c(a^\top \otimes_{\text{Kr}} b^\top)[X_1 \otimes_{\text{Kr}} X_1 \cdots X_r \otimes_{\text{Kr}} X_r]$. Therefore,

$$\sum_{i,j,k \in [d]^3} \sum_{\ell=1}^r X_{i\ell} X_{j\ell} D_{k\ell} a_i b_j c_k = \left\langle D, c(a^\top \otimes_{\text{Kr}} b^\top)[X_1 \otimes_{\text{Kr}} X_1 \cdots X_r \otimes_{\text{Kr}} X_r] \right\rangle.$$

Similarly, for the two remaining terms

$$\sum_{i,j,k \in [d]^3} \sum_{\ell=1}^r X_{i\ell} D_{j\ell} X_{k\ell} a_i b_j c_k = \left\langle D, b(a^\top \otimes_{\text{Kr}} c^\top)[X_1 \otimes_{\text{Kr}} X_1 \cdots X_r \otimes_{\text{Kr}} X_r] \right\rangle,$$

and

$$\sum_{i,j,k \in [d]^3} \sum_{\ell=1}^r D_{i\ell} X_{j\ell} X_{k\ell} a_i b_j c_k = \left\langle D, a(b^\top \otimes_{\text{Kr}} c^\top) [X_1 \otimes_{\text{Kr}} X_1 \cdots X_r \otimes_{\text{Kr}} X_r] \right\rangle.$$

By linearity of the inner product and matrix multiplication, we obtain

$$\left\langle \nabla F_{\text{sym}}(X)^\top (a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c), \text{vec}(D) \right\rangle = \left\langle \left(c(a^\top \otimes_{\text{Kr}} b^\top) + b(a^\top \otimes_{\text{Kr}} c^\top) + a(b^\top \otimes_{\text{Kr}} c^\top) \right) \psi(X, X), D \right\rangle.$$

Since D was arbitrary, one can show component-wise that $\nabla F_{\text{sym}}(X)^\top (a \otimes_{\text{Kr}} b \otimes_{\text{Kr}} c)$ matches $\left(c(a^\top \otimes_{\text{Kr}} b^\top) + b(a^\top \otimes_{\text{Kr}} c^\top) + a(b^\top \otimes_{\text{Kr}} c^\top) \right) \psi(X, X)$. This concludes the proof.

D.5 Proof of Lemma C.30

The proof relies heavily on the following two claims.

Claim D.5. *Let $u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w \in \mathbf{R}^{d_1 d_2 d_3}$ be arbitrary. Then,*

$$\begin{aligned} \nabla F_{\text{asym}}(W, X, Y) \nabla F_{\text{asym}}(W, X, Y)^\top (u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w) &= u \otimes_{\text{Kr}} [\Psi(X, Y) (v \otimes_{\text{Kr}} w)] \\ &\quad + P_2 (v \otimes_{\text{Kr}} [\Psi(W, Y) (u \otimes_{\text{Kr}} w)]) \\ &\quad + P_3 (w \otimes_{\text{Kr}} [\Psi(W, X) (u \otimes_{\text{Kr}} v)]), \end{aligned}$$

where P_i and Ψ are introduced in Definitions C.23 and C.25, respectively.

Proof of Claim D.5. By Lemma C.28, we have

$$\begin{aligned} \nabla F(W, X, Y) \nabla F(W, X, Y)^\top &= J^W (J^W)^\top + J^X (J^X)^\top + J^Y (J^Y)^\top \\ &= I_{d_1} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (X_l X_l^\top) \otimes_{\text{Kr}} (Y_l Y_l^\top) \right) \\ &\quad + P_2 \left(I_{d_2} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (W_l W_l^\top) \otimes_{\text{Kr}} (Y_l Y_l^\top) \right) \right) P_2^\top \\ &\quad + P_3 \left(I_{d_3} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (W_l W_l^\top) \otimes_{\text{Kr}} (X_l X_l^\top) \right) \right) P_3^\top, \end{aligned}$$

where the second equality we use the Kronecker property (68) to take the transpose and (69) to simplify the product. Given a vector $u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w$ we expand

$$\begin{aligned} &\nabla F(W, X, Y) \nabla F(W, X, Y)^\top (u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w) \\ &= \left(I_{d_1} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (X_l X_l^\top) \otimes_{\text{Kr}} (Y_l Y_l^\top) \right) \right) (u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w) \\ &\quad + P_2 \left(I_{d_2} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (W_l W_l^\top) \otimes_{\text{Kr}} (Y_l Y_l^\top) \right) \right) P_2^\top (u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w) \\ &\quad + P_3 \left(I_{d_3} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (W_l W_l^\top) \otimes_{\text{Kr}} (X_l X_l^\top) \right) \right) P_3^\top (u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w) \\ &\stackrel{(i)}{=} \left(I_{d_1} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (X_l X_l^\top) \otimes_{\text{Kr}} (Y_l Y_l^\top) \right) \right) (u \otimes_{\text{Kr}} v \otimes_{\text{Kr}} w) \\ &\quad + P_2 \left(I_{d_2} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (W_l W_l^\top) \otimes_{\text{Kr}} (Y_l Y_l^\top) \right) \right) (v \otimes_{\text{Kr}} u \otimes_{\text{Kr}} w) \end{aligned}$$

$$\begin{aligned}
& + P_3 \left(I_{d_3} \otimes_{\text{Kr}} \left(\sum_{l=1}^r (W_l W_l^\top) \otimes_{\text{Kr}} (X_l X_l^\top) \right) \right) (w \otimes_{\text{Kr}} u \otimes_{\text{Kr}} v) \\
\stackrel{(ii)}{=} & u \otimes_{\text{Kr}} \left(\left(\sum_{l=1}^r X_l X_l^\top \otimes_{\text{Kr}} Y_l Y_l^\top \right) (v \otimes_{\text{Kr}} w) \right) + P_2 \left[v \otimes_{\text{Kr}} \left(\left(\sum_{l=1}^r W_l W_l^\top \otimes_{\text{Kr}} Y_l Y_l^\top \right) (u \otimes_{\text{Kr}} w) \right) \right] \\
& + P_3 \left[w \otimes_{\text{Kr}} \left(\left(\sum_{l=1}^r W_l W_l^\top \otimes_{\text{Kr}} X_l X_l^\top \right) (u \otimes_{\text{Kr}} v) \right) \right] \\
= & u \otimes_{\text{Kr}} [\Psi(X, Y) (v \otimes_{\text{Kr}} w)] + P_2 (v \otimes_{\text{Kr}} [\Psi(W, Y) (u \otimes_{\text{Kr}} w)]) + P_3 (w \otimes_{\text{Kr}} [\Psi(W, X) (u \otimes_{\text{Kr}} v)]),
\end{aligned}$$

where (i) follows from the definition of P_i and (ii) follows from (69). This concludes the proof. \square

Claim D.6. *The following identity holds*

$$\Psi(M, N) \left(T_i^M \otimes_{\text{Kr}} T_j^N \right) = \mathbb{1}_{(i,j) \in \mathcal{T}_{\text{on}}^{M,N}} (M_i \otimes_{\text{Kr}} N_j).$$

Proof of Claim D.6. To establish the claim, we express the operator $\Psi(M, N)$ using the SVD of M and N . Since $M_l = \sum_{k=1}^r \sigma_k^M V_{kl}^M U_k^M$, then $\sum_{l=1}^r M_l M_l^\top = \sum_{l,k_1,k_2 \in [r]^3} \sigma_{k_1}^M \sigma_{k_2}^M V_{k_1 l}^M V_{k_2 l}^M U_{k_1}^M U_{k_2}^{M^\top}$. Moreover, by bilinearity of the Kronecker product, we have

$$M_l M_l^\top \otimes_{\text{Kr}} N_l N_l^\top = \sum_{k_1, k_2, \ell_1, \ell_2 \in [r]^4} \sigma_{k_1}^M \sigma_{k_2}^M \sigma_{\ell_1}^N \sigma_{\ell_2}^N V_{k_1 l}^M V_{k_2 l}^M V_{\ell_1 l}^N V_{\ell_2 l}^N \left(U_{k_1}^M U_{k_2}^{M^\top} \right) \otimes_{\text{Kr}} \left(U_{\ell_1}^N U_{\ell_2}^{N^\top} \right),$$

and so, taking a sum over ℓ yields

$$\Psi(M, N) = \sum_{l, k_1, k_2, \ell_1, \ell_2 \in [r]^5} \sigma_{k_1}^M \sigma_{k_2}^M \sigma_{\ell_1}^N \sigma_{\ell_2}^N V_{k_1 l}^M V_{k_2 l}^M V_{\ell_1 l}^N V_{\ell_2 l}^N \left(U_{k_1}^M U_{k_2}^{M^\top} \right) \otimes_{\text{Kr}} \left(U_{\ell_1}^N U_{\ell_2}^{N^\top} \right).$$

We are now ready to establish the action of this operator in a vector of the form $T_i^M \otimes_{\text{Kr}} T_j^N$. Recall the definition of $\mathcal{T}_{\text{on}}^{M,N}$ given in (60). First, assume that $\max\{i, j\} > r$, without loss of generality, suppose that $i > r$. Then, $T_i^M = U_i^M$, and $U_{k_2}^\top T_i^M = 0$ for all $k_2 \in [r]$ and, consequently, $\Phi(M, N)(T_i^M \otimes_{\text{Kr}} T_j^N) = 0$. Second, assuming both $i, j \in [r]$, we have that $T_i^M \otimes_{\text{Kr}} T_j^N = U^M (\Sigma^M)^{-1} (V_{:i}^M)^\top \otimes_{\text{Kr}} U^N (\Sigma^N)^{-1} (V_{:j}^N)^\top$. Then

$$\begin{aligned}
& \Psi(M, N) \left(T_i^M \otimes_{\text{Kr}} T_j^N \right) \\
= & \sum_{l, k_1, k_2, \ell_1, \ell_2 \in [r]^5} \sigma_{k_1}^M \sigma_{k_2}^M \sigma_{\ell_1}^N \sigma_{\ell_2}^N V_{l k_1}^M V_{l k_2}^M V_{l \ell_1}^N V_{l \ell_2}^N \left(U_{k_1}^M U_{k_2}^{M^\top} U^M (\Sigma^M)^{-1} (V_{:i}^M)^\top \right) \\
& \otimes_{\text{Kr}} \left(U_{\ell_1}^N U_{\ell_2}^{N^\top} U^N (\Sigma^N)^{-1} (V_{:j}^N)^\top \right) \quad (73)
\end{aligned}$$

$$\begin{aligned}
\stackrel{(i)}{=} & \sum_{l, k_1, k_2, \ell_1, \ell_2 \in [r]^5} \sigma_{k_1}^M \sigma_{k_2}^M \sigma_{\ell_1}^N \sigma_{\ell_2}^N V_{l k_1}^M V_{l k_2}^M V_{l \ell_1}^N V_{l \ell_2}^N V_{i k_2}^M (\sigma_{k_2}^M)^{-1} V_{j \ell_2}^N (\sigma_{\ell_2}^N)^{-1} \left(U_{k_1}^M \otimes_{\text{Kr}} U_{\ell_1}^N \right) \\
\stackrel{(ii)}{=} & \sum_{l, k_1, k_2, \ell_1, \ell_2 \in [r]^5} \sigma_{k_1}^M \sigma_{\ell_1}^N V_{l k_1}^M V_{l k_2}^M V_{l \ell_1}^N V_{l \ell_2}^N V_{i k_2}^M V_{j \ell_2}^N \left(U_{k_1}^M \otimes_{\text{Kr}} U_{\ell_1}^N \right) \\
\stackrel{(iii)}{=} & \sum_{l, k_1, \ell_1 \in [r]^3} \sigma_{k_1}^M \sigma_{\ell_1}^N V_{l k_1}^M V_{l \ell_1}^N \langle V_{:l}^M, V_{:i}^M \rangle \langle V_{:l}^N, V_{:j}^N \rangle \left(U_{k_1}^M \otimes_{\text{Kr}} U_{\ell_1}^N \right) \quad (74) \\
= & \mathbb{1}_{i=j} \sum_{k_1, \ell_1 \in [r]^2} \sigma_{k_1}^M \sigma_{\ell_1}^N V_{i k_1}^M V_{j \ell_1}^N \left(U_{k_1}^M \otimes_{\text{Kr}} U_{\ell_1}^N \right) \\
= & \mathbb{1}_{(i,j) \in \mathcal{T}_{\text{on}}^{M,N}} (M_i \otimes_{\text{Kr}} N_j),
\end{aligned}$$

where (i) follows since $U_{k_2}^{M\top} U^M (\Sigma^M)^{-1} (V_{:i}^M)^\top = (\sigma_{k_2}^M)^{-1} V_{ik_2}^M$ and $U_{\ell_2}^{M\top} U^N (\Sigma^N)^{-1} (V_{:i}^N)^\top = (\sigma_{\ell_2}^N)^{-1} V_{i\ell_2}^N$, (ii) follows from the fact that the columns of U^M and U^N form orthonormal bases, and (iii) follows from factorizing out the dot product. This finishes the proof of the claim. \square

We apply Claim D.5 with $u = T_i^W$, $v = T_j^X$, and $w = T_k^Y$, to derive

$$\begin{aligned} \nabla F_{\text{asym}}(W, X, Y) \nabla F(W, X, Y)^\top & \left(T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \right) = T_i^W \otimes \left[\Psi(X, Y) \left(T_j^X \otimes T_k^Y \right) \right] \\ & + P_2 \left[T_j^X \otimes \Psi(W, Y) \left(T_i^W \otimes T_k^Y \right) \right] \\ & + P_3 \left[T_k^Y \otimes \Psi(W, X) \left(T_i^W \otimes T_j^X \right) \right]. \end{aligned}$$

Using Lemma D.6 in tandem with Definition C.23 of P_2 and P_3 we obtain

$$\begin{aligned} \nabla F(W, X, Y) \nabla F(W, X, Y)^\top & \left(T_i^W \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} T_k^Y \right) = \mathbb{1}_{(i,j) \in \mathcal{T}_{\text{on}}^{W,X}} \left(W_i \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} T_k^Y \right) \\ & + \mathbb{1}_{(i,k) \in \mathcal{T}_{\text{on}}^{W,Y}} \left(W_i \otimes_{\text{Kr}} T_j^X \otimes_{\text{Kr}} Y_k \right) \\ & + \mathbb{1}_{(j,k) \in \mathcal{T}_{\text{on}}^{X,Y}} \left(T_i^W \otimes_{\text{Kr}} X_j \otimes_{\text{Kr}} Y_k \right). \end{aligned}$$

This completes the argument.

D.6 Alignment lemmas

Lemma D.7. *Let $X, X^* \in \mathbb{R}^{d_1 \times r}$ and $Y, Y^* \in \mathbb{R}^{d_2 \times r}$ with $\text{rank}(X^*) = \text{rank}(Y^*) = r^*$. Let $k_1 \in \{r^* \dots r\}$, $k_2 \in \{r^* \dots r\}$, and define $\mathcal{Q}_X := \sum_{i=k_1+1}^{d_1} U_i^X U_i^{X\top}$ and $\mathcal{Q}_Y := \sum_{i=k_2+1}^{d_2} U_i^Y U_i^{Y\top}$. Under the assumption that $V^{X^*} = V^{Y^*}$, if*

$$\left\| XX^\top - X^* X^{*\top} \right\|_F + \left\| YY^\top - Y^* Y^{*\top} \right\|_F \leq \frac{1}{2\sqrt{2}} \min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\},$$

one has

$$\left\| \mathcal{Q}_X X^* Y^{*\top} \mathcal{Q}_Y \right\|_F \leq \frac{1}{\min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\}} \left\| (I - \mathcal{Q}_X) X Y^\top (I - \mathcal{Q}_Y) - X^* Y^{*\top} \right\|_F^2.$$

Moreover, if $k_1 = k_2$ and $V^X = V^Y$, one has

$$\left\| \mathcal{Q}_X X^* Y^{*\top} \mathcal{Q}_Y \right\|_F \leq \frac{1}{\min \left\{ \sigma_{r^*}^2(X^*), \sigma_{r^*}^2(Y^*) \right\}} \left\| X Y^\top - X^* Y^{*\top} \right\|_F^2.$$

Proof. We will denote $E := (I - \mathcal{Q}_X) X Y^\top (I - \mathcal{Q}_Y) - X^* Y^{*\top}$. We can decompose the error E as

$$E = (I - \mathcal{Q}_X) E (I - \mathcal{Q}_Y) + \mathcal{Q}_X E (I - \mathcal{Q}_Y) + (I - \mathcal{Q}_X) E \mathcal{Q}_Y + \mathcal{Q}_X E \mathcal{Q}_Y.$$

Observe that all the terms in this sum are pairwise orthogonal in the Frobenius inner product. Therefore,

$$\begin{aligned} \|E\|_F^2 &= \|(I - \mathcal{Q}_X) E (I - \mathcal{Q}_Y)\|_F^2 + \|\mathcal{Q}_X E (I - \mathcal{Q}_Y)\|_F^2 + \|(I - \mathcal{Q}_X) E \mathcal{Q}_Y\|_F^2 + \|\mathcal{Q}_X E \mathcal{Q}_Y\|_F^2 \\ &\geq \|\mathcal{Q}_X E (I - \mathcal{Q}_Y)\|_F^2 + \|(I - \mathcal{Q}_X) E \mathcal{Q}_Y\|_F^2 \end{aligned}$$

$$\stackrel{(i)}{=} \left\| \mathcal{Q}_X X^* Y^{*\top} (I - \mathcal{Q}_Y) \right\|_F^2 + \left\| (I - \mathcal{Q}_X) X^* Y^{*\top} \mathcal{Q}_Y \right\|_F^2$$

$$\stackrel{(ii)}{=} \left\| \mathcal{Q}_X \left(X^* V^{X^*} \right)_{\{1, \dots, r^*\}} \left(\left(Y^* V^{Y^*} \right)_{\{1, \dots, r^*\}} \right)^\top (I - \mathcal{Q}_Y) \right\|_F^2$$

$$\begin{aligned}
& + \left\| (I - \mathcal{Q}_X) (X^* V^{X^*})_{\{1, \dots, r^*\}} \left((Y^* V^{Y^*})_{\{1, \dots, r^*\}} \right)^\top \mathcal{Q}_Y \right\|_F^2 \\
& \stackrel{(iii)}{\geq} \sigma_{r^*}^2 \left((I - \mathcal{Q}_Y) (Y^* V^{Y^*})_{\{1, \dots, r^*\}} \right) \left\| \mathcal{Q}_X (X^* V^{X^*})_{\{1, \dots, r^*\}} \right\|_F^2 \\
& \quad + \sigma_{r^*}^2 \left((I - \mathcal{Q}_X) (X^* V^{X^*})_{\{1, \dots, r^*\}} \right) \left\| \mathcal{Q}_Y (Y^* V^{Y^*})_{\{1, \dots, r^*\}} \right\|_F^2 \\
& \stackrel{(iv)}{\geq} \frac{1}{2} \min \left\{ \sigma_{r^*}^2 (X^*), \sigma_{r^*}^2 (Y^*) \right\} \left(\left\| \mathcal{Q}_X (X^* V^{X^*})_{\{1, \dots, r^*\}} \right\|_F^2 + \left\| \mathcal{Q}_Y (Y^* V^{Y^*})_{\{1, \dots, r^*\}} \right\|_F^2 \right) \\
& \stackrel{(v)}{\geq} \min \left\{ \sigma_{r^*}^2 (X^*), \sigma_{r^*}^2 (Y^*) \right\} \left\| \mathcal{Q}_X X^* Y^{*\top} \mathcal{Q}_Y \right\|_F,
\end{aligned}$$

where (i) follows from $\mathcal{Q}_X(I - \mathcal{Q}_X) = 0$ and $\mathcal{Q}_Y(I - \mathcal{Q}_Y) = 0$; (ii) follows from $V^{X^*} = V^{Y^*}$ and $r^* = r^*$; (iii) follows from the variational characterization of singular values; (iv) follows from Corollary D.9 and (v) follows from Young's inequality in conjunction with the Cauchy–Schwarz inequality. Moreover, if $V^X = V^Y$ and $k_1 = k_2$, then the same argument holds for $E := XY^\top - X^*Y^{*\top}$, since equality (i) holds. This concludes the proof. \square

Lemma D.8 (Lemma 33 in [106]). *Let $X^* \in \mathbf{R}^{d \times r^*}$ of rank r^* and let $X \in \mathbf{R}^{d \times r}$. Assume that $\|XX^\top - X^*X^{*\top}\|_F \leq \frac{1}{2\sqrt{2}}\sigma_{r^*}^2(X^*)$. Let $\mathcal{Q}_X = \sum_{i=k_1+1}^{d_1} U_i^X U_i^{X^\top}$ with $k_1 \geq r^*$. Then*

$$\lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* \right) \geq \lambda_1 \left(X^{*\top} \mathcal{Q}_X X^* \right).$$

Proof. We will use as placeholders $\alpha_1 := \lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* \right)$ and $\alpha_2 := \lambda_1 \left(X^{*\top} \mathcal{Q}_X X^* \right)$. Our argument follows by contradiction, we will prove that $\alpha_1 < \alpha_2$ implies $\frac{\|XX^\top - X^*X^{*\top}\|_F}{\sigma_{r^*}^2(X^*)} \geq \frac{1}{\sqrt{2}} > \frac{1}{2\sqrt{2}}$ which contradicts the hypothesis of the lemma. We bound

$$\left\| XX^\top - X^*X^{*\top} \right\|_F^2 \stackrel{(i)}{=} \left\| (I - \mathcal{Q}_X) XX^\top (I - \mathcal{Q}_X) - (I - \mathcal{Q}_X) X^* X^{*\top} (I - \mathcal{Q}_X) \right\|_F^2 \quad (75)$$

$$\begin{aligned}
& + 2 \left\| (I - \mathcal{Q}_X) X^* X^{*\top} \mathcal{Q}_X \right\|_F^2 \\
& \quad + \left\| \mathcal{Q}_X XX^\top \mathcal{Q}_X - \mathcal{Q}_X X^* X^{*\top} \mathcal{Q}_X \right\|_F^2 \\
& \stackrel{(ii)}{\geq} \left\| (I - \mathcal{Q}_X) XX^\top (I - \mathcal{Q}_X) - (I - \mathcal{Q}_X) X^* X^{*\top} (I - \mathcal{Q}_X) \right\|_F^2 \\
& \quad + \left\| \mathcal{Q}_X XX^\top \mathcal{Q}_X - \mathcal{Q}_X X^* X^{*\top} \mathcal{Q}_X \right\|_F^2 + 2\sigma_{r^*}^2 \left((I - \mathcal{Q}_X) X^* \right) \sigma_1^2 \left(\mathcal{Q}_X X^* \right)
\end{aligned} \quad (76)$$

$$\begin{aligned}
& \stackrel{(iii)}{=} \left\| (I - \mathcal{Q}_X) XX^\top (I - \mathcal{Q}_X) - (I - \mathcal{Q}_X) X^* X^{*\top} (I - \mathcal{Q}_X) \right\|_F^2 \\
& \quad + \left\| \mathcal{Q}_X XX^\top \mathcal{Q}_X - \mathcal{Q}_X X^* X^{*\top} \mathcal{Q}_X \right\|_F^2 + 2\alpha_1 \alpha_2,
\end{aligned} \quad (77)$$

where (i) follows by expanding the square and using orthogonality and (ii) follows from the variational characterization of singular values, and (iii) holds since $\sigma_k^2(PX^*) = \lambda_k \left(X^{*\top} P^\top P X^* \right) =$

$\lambda_k (X^{*\top} P X^*)$ for any $k \in \{1 \dots r^*\}$ and any orthogonal projection matrix $P \in \mathbf{R}^{d_1 \times d_1}$. We claim

$$\begin{aligned} & \|(I - \mathcal{Q}_X) X X^\top (I - \mathcal{Q}_X) - (I - \mathcal{Q}_X) X^* X^{*\top} (I - \mathcal{Q}_X)\|_F^2 + \|\mathcal{Q}_X X X^\top \mathcal{Q}_X - \mathcal{Q}_X X^* X^{*\top} \mathcal{Q}_X\|_F^2 \\ & \geq \min_{\beta_1, \beta_2 \in \mathbf{R}_+ | \beta_1 \geq \beta_2} \left\{ (\beta_1 - \alpha_1)^2 + (\beta_2 - \alpha_2)^2 \right\}, \end{aligned} \quad (78)$$

let us defer the proof of this inequality until after we establish the result. Given (78), if $\alpha_1 < \alpha_2$, then the optimal solution of the lower bound occurs at $\beta_1 = \beta_2 = \frac{\alpha_1 + \alpha_2}{2}$, so the minimum value becomes $\frac{1}{2}(\alpha_1 - \alpha_2)^2$. Substituting this into (75) gives

$$\|X X^\top - X^* X^{*\top}\|_F^2 \geq \frac{1}{2}(\alpha_1 - \alpha_2)^2 + 2\alpha_1\alpha_2 = \frac{1}{2}(\alpha_1 + \alpha_2)^2 \geq \frac{1}{2}\sigma_{r^*}^4(X^*),$$

where we used Weyl's inequality in the last step to bound

$$\alpha_1 + \alpha_2 \geq \lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* + X^{*\top} \mathcal{Q}_X X^* \right) = \lambda_{r^*} \left(X^{*\top} X^* \right) = \sigma_{r^*}^2(X^*).$$

Taking the square roots on both side implies $\|X X^\top - X^* X^{*\top}\|_F \geq \frac{1}{\sqrt{2}}\sigma_{r^*}^2(X^*)$, a contradicts the radius hypothesis. We turn to proving (78), we have that

$$\begin{aligned} & \|(I - \mathcal{Q}_X) X X^\top (I - \mathcal{Q}_X) - (I - \mathcal{Q}_X) X^* X^{*\top} (I - \mathcal{Q}_X)\|_F^2 + \|\mathcal{Q}_X X X^\top \mathcal{Q}_X - \mathcal{Q}_X X^* X^{*\top} \mathcal{Q}_X\|_F^2 \\ & \stackrel{(i)}{\geq} \min_{S_1 \geq 0, S_2 \geq 0 | \sigma_{r^*}(S_1) \geq \sigma_1(S_2)} \left\| S_1 - (I - \mathcal{Q}_X) X^* X^{*\top} (I - \mathcal{Q}_X) \right\|_F^2 + \left\| S_2 - \mathcal{Q}_X X^* X^{*\top} \mathcal{Q}_X \right\|_F^2 \\ & \stackrel{(ii)}{\geq} \min_{\beta_1, \beta_2 \in \mathbf{R}_+ | \beta_1 \geq \beta_2} (\beta_1 - \alpha_1)^2 + (\beta_2 - \alpha_2)^2, \end{aligned}$$

where (i) is due to $\sigma_{r^*} \left((I - \mathcal{Q}_X) X X^\top (I - \mathcal{Q}_X) \right) \geq \sigma_1 \left(\mathcal{Q}_X X X^\top \mathcal{Q}_X \right)$ and (ii) follows from the Hoffman-Wielandt Theorem [7, Problem III.6.15]. This concludes the proof. \square

Corollary D.9. *Let $X^* \in \mathbf{R}^{d \times r^*}$ of rank r^* . If $\|X X^\top - X^* X^{*\top}\|_F \leq \frac{1}{2\sqrt{2}}\sigma_{r^*}^2(X^*)$, we have that*

$$\lambda_{r^*}(X^{*\top} X^*) \leq 2\lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* \right).$$

Proof. One has

$$\begin{aligned} \lambda_{r^*} \left(X^{*\top} X^* \right) &= \lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* + X^{*\top} \mathcal{Q}_X^\top \mathcal{Q}_X X^* \right) \\ &\stackrel{(i)}{\leq} \lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* \right) + \lambda_1 \left(X^{*\top} \mathcal{Q}_X X^* \right) \\ &\stackrel{(ii)}{\leq} 2\lambda_{r^*} \left(X^{*\top} (I - \mathcal{Q}_X) X^* \right), \end{aligned}$$

where (i) follows from Weyl's inequality and (ii) follows from Lemma D.8. \square

E Computing the preconditioner

In this section, we elaborate on how to compute the preconditioners

$$\left(\nabla F(x)^\top \nabla F(x) + \lambda I \right)^{-1} g$$

given a fixed $g \in \mathbf{R}^d$. For this task, we use the conjugate gradients method (CG), which converges linearly at a rate that depends on the condition number of the matrix $P(x, \lambda) = \nabla F(x)^\top \nabla F(x) + \lambda I$. We have found empirically that executing around ten iterations of CG suffices to obtain fast

convergence of LMM. The main subroutine necessary for CG is the matrix-vector product $y \mapsto P(x, \lambda)y$; in what follows, we study the complexity of this subroutine in the examples we studied.

E.1 Square-variable map

For the component-wise square, we have that $m = d$, and that $\nabla F(x) = 2 \text{diag}(x)$. Thus, we have that the quantity $P(x, \lambda)y = 4x \odot x \odot y + \lambda y$ which can be computed with $\mathcal{O}(d)$ flops.

E.2 Burer-Monteiro factorization

For Burer-Monteiro map, given inputs $X \in \mathbf{R}^{d \times r}$ we have

$$P(X, \lambda)[\tilde{X}] = \text{vec} \left(\tilde{X}X^\top X + X\tilde{X}^\top X + \lambda\tilde{X} \right) \quad \text{for } \tilde{X} \in \mathbf{R}^{d \times r}.$$

The computation follows from (43). This action can be computed with $\mathcal{O}(dr^2)$ flops.

E.3 Asymmetric matrix factorization

For the asymmetric matrix factorization map, given inputs $X \in \mathbf{R}^{d_1 \times r}$ and $Y \in \mathbf{R}^{d_2 \times r}$ we have

$$P((X, Y), \lambda) \begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} = \begin{pmatrix} X\tilde{Y}^\top Y + \tilde{X}Y^\top Y \\ Y\tilde{X}^\top X + \tilde{Y}X^\top X \end{pmatrix} \quad \text{for } \tilde{X} \in \mathbf{R}^{d_1 \times r} \text{ and } \tilde{Y} \in \mathbf{R}^{d_2 \times r},$$

where the computation follows from (48). This action can be computed with $\mathcal{O}((d_1 + d_2)r^2)$ flops.

E.4 Symmetric CP factorization

For the symmetric canonical polyadic map, given an input $X \in \mathbf{R}^{d \times r}$ we have

$$P(X, \lambda)[\tilde{X}] = 3\tilde{X} \left(X^\top X \odot X^\top X \right) + 6X \left(\tilde{X}^\top X \odot X^\top X \right) + \lambda\tilde{X} \quad \text{for } \tilde{X} \in \mathbf{R}^{d \times r}.$$

This computation follows from Lemma C.26. As with the matrix case, this action can be computed with $\mathcal{O}(dr^2)$ flops.

E.5 CP factorization

For the canonical polyadic map, given inputs $W \in \mathbf{R}^{d_1 \times r}$, $X \in \mathbf{R}^{d_2 \times r}$, $Y \in \mathbf{R}^{d_3 \times r}$, we have

$$P((W, X, Y), \lambda) \begin{bmatrix} \tilde{W} \\ \tilde{X} \\ \tilde{Y} \end{bmatrix} = \begin{pmatrix} \tilde{W} \left(X^\top X \odot Y^\top Y \right) + W \left(\tilde{X}^\top X \odot Y^\top Y + X^\top X \odot \tilde{Y}^\top Y \right) + \lambda\tilde{W} \\ \tilde{X} \left(W^\top W \odot Y^\top Y \right) + X \left(\tilde{W}^\top W \odot Y^\top Y + W^\top W \odot \tilde{Y}^\top Y \right) + \lambda\tilde{X} \\ \tilde{Y} \left(W^\top W \odot X^\top X \right) + Y \left(\tilde{W}^\top W \odot X^\top X + W^\top W \odot \tilde{X}^\top X \right) + \lambda\tilde{Y} \end{pmatrix}$$

for $\tilde{W} \in \mathbf{R}^{d_1 \times r}$, $\tilde{X} \in \mathbf{R}^{d_2 \times r}$, $\tilde{Y} \in \mathbf{R}^{d_3 \times r}$. This computation follows from Lemma C.28. Once more, this action can be computed with $\mathcal{O}((d_1 + d_2 + d_3)r^2)$ flops.