Alternating minimization and alternating descent over nonconvex sets

Wooseok Ha and Rina Foygel Barber

11.01.17

Abstract

We analyze the performance of alternating minimization for loss functions optimized over two variables, where each variable may be restricted to lie in some potentially nonconvex constraint set. This type of setting arises naturally in high-dimensional statistics and signal processing, where the variables often reflect different structures or components within the signals being considered. Our analysis depends strongly on the notion of local concavity coefficients, which have been recently proposed in Barber and Ha [2017] to measure and quantify the concavity of a general nonconvex set. Our results further reveal important distinctions between alternating and non-alternating methods. Since computing the alternating minimization steps may not be tractable for some problems, we also consider an inexact version of the algorithm and provide a set of sufficient conditions to ensure fast convergence of the inexact algorithms. We demonstrate our framework on several examples, including low rank + sparse decomposition and multitask regression, and provide numerical experiments to validate our theoretical results.

1 Introduction

Many methods in modern statistics use structured constraints to improve signal recovery in a high-dimensional setting. Common constraints in the high-dimensional statistics literature include sparsity, requiring that a signal consists of mostly zero values; variants on sparsity, such as total variation sparsity, requiring that the first-order differences of a signal are locally constant; and low rank, where the signal is a matrix expressed as the sum of only a few linear factors. In many settings, multiple structures may be present simultaneously in the data, in which case we may need to optimize a function over several variables, which are each believed to exhibit some latent structure—for instance, a low-rank term and a sparse term.

Much of the literature in this area focuses on convex relaxations of these structured constraints, such as the ℓ_1 norm (as a convex approximation to sparsity). Working with a convex penalty or convex constraint, as a proxy for the nonconvex structure of the variable(s) of interest, allows for easier optimization from both a theoretical and a practical point of view. In recent years, however, attention has turned to nonconvex optimization problems, aiming to avoid the loss of accuracy that is often the cost of taking a convex relaxation.

In this work, we consider the problem of optimizing over two variables, one or both of which is constrained to lie in some potentially nonconvex set:

$$(\widehat{x}, \widehat{y}) = \arg\min\{\mathcal{L}(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\},$$

where $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ reflect our beliefs or desired properties for the x and y variables, while $\mathcal{L}$ is the target function to minimize (for example, a negative log-likelihood, in which case we are searching for the constrained maximum likelihood estimator).

Our aim is to study the convergence behavior of the (inexact) *alternating minimization* method for this problem, where we iterate the steps

$$\begin{cases} \text{Fix } y, \text{ and choose } x \in \mathcal{X} \text{ to (approximately) minimize the function } x \mapsto \mathcal{L}(x, y); \\ \text{Fix } x, \text{ and choose } y \in \mathcal{Y} \text{ to (approximately) minimize the function } y \mapsto \mathcal{L}(x, y). \end{cases}$$

This type of method can be practical in scenarios where the loss function is relatively simple to minimize when viewed as a function of either x or y only—for instance, in multitask regression, where x represents the coefficients and y represents the covariance structure. In other settings, even the marginal minimization steps are expensive to calculate, but we can instead consider approximating each one with other iterative procedures, such as gradient descent.

Our main results derive conditions under which (exact or inexact) alternating minimization converges linearly, under certain assumptions:

- Loss function: The assumptions on the loss function $\mathcal{L}(x, y)$ are familiar in the high-dimensional statistics literature, namely, restricted strong convexity (RSC) and restricted smoothness (RSM) assumptions, which essentially require that $\mathcal{L}(x, y)$ behaves like a smooth and strongly convex function when restricted to the constrained domains $\mathcal{X}$ and $\mathcal{Y}$.
- Nonconvex constraints: For the constraint sets $\mathcal{X}$ and $\mathcal{Y}$, we work in the framework established in Barber and Ha [2017], requiring bounded *local concavity coefficients* for each set (see Section 2 for further details). This geometric condition is a natural relaxation of convexity and is satisfied by many commonly used nonconvex constraints, such as a low-rank constraint.
- Initialization: In order to ensure convergence to an optimal point, we need to assume that the algorithm is initialized within some neighborhood of the global minimizer—sufficiently close so that, locally, the (restricted) convexity of the loss function $\mathcal{L}(x, y)$ is sufficient to outweigh nonconvexity in the constraints. Details are given in our results below.

To demonstrate the utility of our results, we also consider a range of specific examples with rank-constrained variables, including multitask regression, robust principal component analysis, and factor models.

1.1 Related work

Alternating minimization is a classical topic in the optimization literature, and a large body of research has been devoted to understanding the method under various settings. On the other hand, nonconvex constraints have recently received a lot of attention from community, including many results treating sparsity-constrained or rank-constrained problems specifically. Our work combines both settings, and thus is naturally related to many existing works. Here, we summarize some of the key recent results, and describe how they relate to our contributions; for brevity, we only focus on the papers most relevant to our work.

Nonconvex constraints on a single variable The past few years have witnessed extensive results on the optimization problem over a nonconvex set. For instance, Jain et al. [2014] consider the problem of minimizing a loss function over a sparsity constraint or a rank constraint, and show that the iterative hard thresholding method can achieve global convergence to a target point, as long as the sparsity or rank of the target is far smaller than the given threshold of the constraint. For a rank constraint specifically, Grussler et al. [2016] study the duality-gap for a certain class of problems and present some situations under which there exists no duality-gap; therefore, solutions to the original problem and its convex relaxation (largest convex minorizer) must coincide under these situations. Turning to a more general setting, Oymak et al. [2015] study projected gradient descent scheme for least squares objective when constraining to a nonconvex regularizer set, establishing linear convergence of the algorithm from any initialization point. In particular, by introducing a descent cone at the target point, the authors characterize the convergence rate in terms of the singular values of the Hessian matrix restricted to this cone. The work of Barber and Ha [2017] takes a different approach and instead develops a way of measuring local concavity of the constraint set at any given point. This measure of local concavity is then used to analyze the local convergence of projected gradient descent.

In the setting of two variables, x and y, as considered in this paper, Barber and Ha [2017]'s approach can be applied by performing gradient descent on the joint variable (x, y), constraining to the space $\mathcal{X} \times \mathcal{Y}$—that is, any problem with multiple variables can of course be reformulated as a single variable problem. In this work, however, we find that separating the variables and alternating their updates can provide substantial benefits, both theoretically and empirically. We discuss these issues in detail in Section 3.

Alternating methods for convex constraints Due to its simplicity and effectiveness, alternating minimization has long been a popular optimization method, dating back to early work in the optimization literature (e.g. Ortega and Rheinboldt [1970]), and has been widely studied under various assumptions (e.g. Auslender [1976], Luo and Tseng [1993]). For instance, assuming that the loss function $\mathcal{L}$ is β-smooth and α-strongly convex in each variable, Luo and Tseng [1993] prove linear convergence for alternating minimization under convex constraints (and, in the case of more than two variables, for the analogous coordinate descent algorithm).

In some settings, the loss function $\mathcal{L}$ may be more well-behaved with respect to one of the variables than the other, in terms of its smoothness and convexity properties. Beck [2015] studies alternating minimization for a convex loss $\mathcal{L}(x, y)$ under convex constraints on x and on y, proving that the gap in the loss function values, i.e. the difference $\mathcal{L}(x_t, y_t) - \mathcal{L}(\widehat{x}, \widehat{y})$, decays according to the rate $\mathcal{O}\left(\frac{\min\{\beta_x, \beta_y\}}{t}\right)$, where β_x and β_y represent the smoothness parameters of the loss $\mathcal{L}$ with respect to the variables x and y respectively. Interestingly, this rate is controlled by the better of the two smoothness parameters—that is, the algorithm will converge rapidly as long as *at least one* of the two smoothness parameters is bounded.

Our main results demonstrate an analogous phenomenon under an additional (restricted) strong convexity assumption—in this setting, we find a *linear* convergence rate, with the convergence radius determined by $\min\left\{\frac{\beta_x}{\alpha_x}, \frac{\beta_y}{\alpha_y}\right\}$, where β_x, β_y are smoothness parameters as before, while α_x, α_y are the (restricted) strong convexity parameters with respect to x and y, respectively. That is, the linear convergence rate depends on the better of the two condition numbers, while in Beck [2015]'s result, without strong convexity, the sublinear convergence rate depends on the better of the two smoothness parameters. Thus, while a main focus of our work is to establish convergence results in a nonconvex setting, even in the convex setting our results reveal the interesting role of the two relative condition numbers (i.e. for the x and the y variables) in determining the overall convergence rate.

2 Optimization over nonconvex constraints

In this section, we briefly review the notion of local concavity coefficients introduced in Barber and Ha [2017], measuring concavity of the set at any given point.

2.1 Local concavity coefficients

One main challenge of working over nonconvex regions is that, since $\mathcal{X}$ and $\mathcal{Y}$ are potentially nonconvex sets, the standard first-order optimality conditions under convex setting do not apply. Specifically, fixing any $y \in \mathcal{Y}$ and defining

$$x_y = \arg\min\{\mathcal{L}(x, y) : x \in \mathcal{X}\},$$

the nonconvexity of $\mathcal{X}$ means that we cannot assume that $\langle x - x_y, \nabla_x \mathcal{L}(x_y, y)\rangle \geq 0$ for all other $x \in \mathcal{X}$ (and same when we reverse the roles of x and y). This makes the analysis of optimization problem with nonconvex constraints difficult, since the first-order optimality condition is crucial for understanding convergence behavior.

In order to overcome this obstacle, Barber and Ha [2017] recently proposed the notion of *local concavity coefficients* for any nonconvex set, related to the notion of prox-regular sets in the analysis literature. These concavity coefficients measure the extent to which the set deviates from convexity using four different properties of a convex set, and prove that these multiple definitions are all equivalent. Here we consider the *curvature condition*, which can be seen as a natural relaxation of the geometric characterization of a convex set:

Definition 1. (Curvature condition.) Let $\mathcal{Z} \subset \mathbb{R}^d$ be a closed subset, containing a point $z \in \mathcal{Z}$. We say that $\mathcal{Z}$ satisfies the *curvature condition* with respect to a norm $\|\cdot\|$ at the point z with parameter γ_z if, for all $z' \in \mathcal{Z}$,

$$\limsup_{t \to 0} \frac{\min_{w \in \mathcal{Z}} \|w - ((1-t)z + tz')\|}{t} \leq \gamma_z \|z - z'\|_2^2.$$

The norm $\|\cdot\|$ used in the definition of the curvature condition is not necessarily the ℓ_2 norm, and may be chosen to suit the problem at hand—it is intended to reflect the natural structure arising in the setting of the problem. For example, in high-dimensional setting, we may instead choose a more structured norm such as the ℓ_1 norm ($\|\cdot\| = \|\cdot\|_1$) for a

sparsity-inducing constraint or the nuclear norm (the sum of the singular values of a matrix) for a low-rank constraint ($\|\cdot\| = \|\cdot\|_{\text{nuc}}$).

Based on Definition 1, we next define the local concavity coefficient $\gamma_z(\mathcal{Z})$ for any point $z \in \mathcal{Z}$. We write $\mathcal{P}_{\mathcal{Z}}$ to denote (possibly non-unique) projection to $\mathcal{Z}$ with respect to the ℓ_2 norm.

Definition 2. (Concavity coefficients.) For a closed subset $\mathcal{Z} \subset \mathbb{R}^d$, let $\mathcal{D} \subset \mathcal{Z}$ be a set of *degenerate points*,

$$\mathcal{D} = \{z \in \mathcal{Z} : \mathcal{P}_{\mathcal{Z}} \text{ is not continuous in any neighborhood of } z\}.$$

The *local concavity coefficient* $\gamma_z(\mathcal{Z})$ at $z \in \mathcal{Z}$, with respect to a norm $\|\cdot\|$ and its dual norm $\|\cdot\|^*$, is then given by

$$\gamma_z(\mathcal{Z}) = \begin{cases} \infty, & z \in \mathcal{D}, \\ \min\{\gamma_z \in [0,\infty] : \text{Condition (1) holds at } z \in \mathcal{Z} \text{ with } \gamma_z\}, & z \notin \mathcal{D}. \end{cases}$$

The *global concavity coefficient* is given by

$$\gamma(\mathcal{Z}) = \sup_{z \in \mathcal{Z}} \gamma_z(\mathcal{Z}).$$

Note that any convex set $\mathcal{Z}$ trivially satisfies $\gamma_z(\mathcal{Z}) = 0$ for all $z \in \mathcal{Z}$. Moreover, Barber and Ha [2017] showed that these coefficients are easy to compute or bound for many nonconvex sets that are commonly used in high-dimensional statistical models—for instance, for the rank-constrained set $\mathcal{Z} = \{Z \in \mathbb{R}^{d_1 \times d_2} : \text{rank}(Z) \le r\}$, the coefficients are given by $\gamma_z(\mathcal{Z}) = \frac{1}{2\sigma_r(Z)}$, where $\sigma_r(Z)$ is the rth singular value of the matrix Z. Intuitively, the curvature condition ensures that while $\mathcal{Z}$ may have nonconvex boundaries in general, this nonconvexity must be fairly "smooth" wherever the coefficient $\gamma_z(\mathcal{Z})$ is small. For more details on local concavity coefficient and its application to nonconvex optimization, see Barber and Ha [2017]. The connection to the notion of prox-regular sets from the nonsmooth analysis literature is also discussed in Barber and Ha [2017, Section 2.3].

One useful property of the local concavity coefficients $\gamma_z(\mathcal{Z})$ is that they equivalently characterize the extent to which the usual first-order optimality conditions are violated when minimizing over the set $\mathcal{Z}$:

Lemma 1 (Barber and Ha [2017, Theorem 2])**. (First-order optimality.)** *Let $\gamma_z(\mathcal{Z})$ be a local concavity coefficient, as in Definition 2, with respect to a norm $\|\cdot\|$ and its dual $\|\cdot\|^*$. Then, for any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ such that z is a local minimizer of f over $\mathcal{Z}$,*

$$\langle z' - z, \nabla f(z)\rangle \ge -\gamma_z(\mathcal{Z})\|\nabla f(z)\|^* \|z' - z\|_2^2 \text{ for all } z' \in \mathcal{Z}.$$

Comparing to a convex setting, where first-order optimality properties ensure that $\langle z' - z, \nabla f(z)\rangle \ge 0$ whenever z is a local minimizer of f over $\mathcal{Z}$, we see that a small coefficient $\gamma_z(\mathcal{Z})$ ensures that $\mathcal{Z}$ behaves "almost" like a convex set in this regard.

Returning to the alternating minimization setting, we first fix norms $\|\cdot\|_x$ and $\|\cdot\|_y$ for the x and y variables—for instance, for a low-rank + sparse problem, we might choose $\|\cdot\|_x$ and $\|\cdot\|_y$ to be the nuclear norm and the ℓ_1 norm, respectively. To simplify our exposition, we will assume that our structured norms $\|\cdot\|_x, \|\cdot\|_y$ are scaled to satisfy $\|\cdot\|_x, \|\cdot\|_y \ge \|\cdot\|_2$, which is the case for many of the structured norms that arise in various applications (such as the ℓ_1 norm and nuclear norm).

Let the local concavity coefficients $\gamma_x(\mathcal{X})$ and $\gamma_y(\mathcal{Y})$ be defined with respect to these potentially different norms. Lemma 1 allows us to obtain approximate first-order optimality conditions for the steps of the alternating minimization algorithm—for instance, letting x_y be a local minimum of the problem $\min\{\mathcal{L}(x,y) : x \in \mathcal{X}\}$ (i.e. the x update step of alternating minimization), then for all $x \in \mathcal{X}$,

$$\langle x - x_y, \nabla_x \mathcal{L}(x_y, y)\rangle \ge -\gamma_{x_y}(\mathcal{X})\|\nabla_x \mathcal{L}(x_y, y)\|_x^* \|x - x_y\|_2^2, \tag{1}$$

and similarly for y. These bounds provide a critical ingredient for our convergence analysis.

3 Convergence analysis of alternating minimization

We now turn to our convergence result on the alternating minimization method. Given the loss function $\mathcal{L}(x,y)$ which is differentiable, we consider an optimization problem,

$$\text{Minimize } \mathcal{L}(x,y) \text{ over } x \in \mathcal{X}, y \in \mathcal{Y},$$

where the sets $\mathcal{X} \subset \mathbb{R}^{d_x}$ and $\mathcal{Y} \subset \mathbb{R}^{d_y}$ represent the structural constraints on the variables x and y respectively.

Let $(\widehat{x}, \widehat{y})$ be the target of our optimization problem, which formally we require only to be a *local* minimizer of $\mathcal{L}(x, y)$—this is because $\mathcal{L}(x, y)$ may potentially be highly nonconvex or degenerate in regions (x, y) far from the origin, and we may even have $\lim \mathcal{L}(x, y) = -\infty$ as (x, y) tends to infinity in some direction. If this is the case, then the steps of the alternating minimization algorithm could potentially diverge, and it may instead be necessary to choose our update steps locally.

To formalize this, define new constraint sets $\mathcal{X}_0 = \mathcal{X} \cap \mathbb{B}_2(x_0, \rho_x)$ and $\mathcal{Y}_0 = \mathcal{Y} \cap \mathbb{B}_2(y_0, \rho_y)$, where (x_0, y_0) is our initialization point. These neighborhoods of the original constraint sets $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be sufficiently large so as to contain the target point $(\widehat{x}, \widehat{y})$ (in other words, our initialization point (x_0, y_0) was chosen to be close to the target $(\widehat{x}, \widehat{y})$), but sufficiently small so that the loss function $\mathcal{L}(x, y)$ is well-behaved over this small region $\mathcal{X}_0 \times \mathcal{Y}_0$.

We then define

$$(\widehat{x}, \widehat{y}) = \arg\min \{\mathcal{L}(x, y) : x \in \mathcal{X}_0, y \in \mathcal{Y}_0\},$$

and run the alternating minimization algorithm locally by iterating the steps

$$\begin{cases} x_t = \arg\min_{x \in \mathcal{X}_0} \mathcal{L}(x, y_{t-1}), \\ y_t = \arg\min_{y \in \mathcal{Y}_0} \mathcal{L}(x_t, y). \end{cases} \tag{2}$$

For our intuition, we should interpret these radius constraints, i.e. working in $\mathcal{X}_0$ and $\mathcal{Y}_0$ rather than in $\mathcal{X}$ and $\mathcal{Y}$, as a technicality for the theory, which we do not need to actually implement in practice. In particular, for many settings, the alternating minimization steps are implemented with some kind of local search procedure, such as gradient descent in the x or the y variable, which will move towards a nearby local minimizer without enforcing a radius constraint. In other settings, even the *global* minimizer for the x or the y variable (while the other variable is fixed), stays within a small neighborhood, without enforcing a radius constraint. In other words, the radius constraint will generally not be active, and thus we can often ignore it in our implementation of the algorithm. However, for the theoretical results obtained here, we require it in order to be able to handle a broader range of problems.

The following lemma proves that, if the radii ρ_x, ρ_y are chosen to be small, the curvature conditions (Definition 1) of $\mathcal{X}$ and $\mathcal{Y}$ are inherited by $\mathcal{X}_0$ and $\mathcal{Y}_0$:

Lemma 2. *If $\rho_x < \frac{1}{2 \max_{x \in \mathcal{X}_0} \gamma_x(\mathcal{X})}$, then $\gamma_x(\mathcal{X}_0) \leq \gamma_x(\mathcal{X})$ for all $x \in \mathcal{X}_0$, and in particular,*

$$\gamma(\mathcal{X}_0) \leq \max_{x \in \mathcal{X}_0} \gamma_x(\mathcal{X}).$$

The analogous statement holds for y.

The proof of this result is given in Appendix B.3.

To see how this result will play a role in the convergence analysis for alternating minimization, consider a single update step for the x variable. Let $x_y = \arg\min\{\mathcal{L}(x, y) : x \in \mathcal{X}_0\}$. Then Lemma 1 proves the following bound (which we can compare to (1)),

$$\langle x' - x_y, \nabla_x \mathcal{L}(x_y, y) \rangle \geq -\gamma(\mathcal{X}_0) \|\nabla_x \mathcal{L}(x_y, y)\|_x^* \|x' - x_y\|_2^2 \text{ for all } x' \in \mathcal{X}_0, \tag{3}$$

while Lemma 2 proves a useful bound for $\gamma(\mathcal{X}_0)$ as long as ρ_x is sufficiently small (and similarly for y).

3.1 Assumptions

Next we formally establish our assumptions on the loss function $\mathcal{L}(x, y)$ as well as initialization condition.

Loss function We first define some notation. Our convergence results will be derived in terms of the *first-order divergence*, a measure of distance to the optimal points $\widehat{x}$ and $\widehat{y}$ that is defined relative to the loss function:

$$D^2(x; \widehat{x}) = \langle x - \widehat{x}, \nabla_x \mathcal{L}(x, \widehat{y}) - \nabla_x \mathcal{L}(\widehat{x}, \widehat{y}) \rangle, \text{ and} \tag{4}$$

$$D^2(y; \widehat{y}) = \langle y - \widehat{y}, \nabla_y \mathcal{L}(\widehat{x}, y) - \nabla_y \mathcal{L}(\widehat{x}, \widehat{y}) \rangle. \tag{5}$$

This divergence has been used also in Loh and Wainwright [2013] to prove statistical errors of any local minimum in the sparse regression setting. Note that, if $\mathcal{L}$ is nonconvex, then potentially $D^2(x;\widehat{x})$ or $D^2(y;\widehat{y})$ may be negative. Abusing notation, we define the square root of the divergence as

$$D(x;\widehat{x}) = \sqrt{\max\{0, D^2(x;\widehat{x})\}} \quad \text{and} \quad D(y;\widehat{y}) = \sqrt{\max\{0, D^2(y;\widehat{y})\}},$$

to accommodate the case where the divergences may be negative.

Throughout we will write $\epsilon_x, \epsilon_y \geq 0$ to indicate vanishing error terms that allow a small amount of slack in the convexity and smoothness conditions. In the high-dimensional statistics literature, these terms often represent the "statistical error"—meaning, if the global minimizer $\widehat{x}$ approximates some "true" parameter $x^\star$ only up to an error level of ϵ_x, then as soon as our iterative algorithm reaches a solution x_t within distance $\sim \epsilon_x$ of $\widehat{x}$, we are already optimal (up to a constant) in terms of estimating the underlying parameters $x^\star$. While our work in this paper is not based in a concrete statistical model, we will still refer to ϵ_x, ϵ_y as the statistical error terms, as this is often the case for many of the applications of our result.

We now state our assumptions on the loss $\mathcal{L}(x, y)$. As mentioned earlier, our optimization method works locally in neighborhoods of the initialization point (x_0, y_0). Consequently, it is sufficient for us to require the assumptions on $\mathcal{L}(x, y)$ to hold only locally in the regions $\mathcal{X}_0$ and $\mathcal{Y}_0$.

First, since (x, y) are being optimized jointly, we need to ensure that these two variables are identifiable, and require a joint restricted strong convexity (RSC) condition at the target point $(\widehat{x}, \widehat{y})$:

Assumption 1. (Joint restricted strong convexity (RSC).) *For all $x \in \mathcal{X}_0$ and all $y \in \mathcal{Y}_0$,*

$$\langle \begin{pmatrix} x - \widehat{x} \\ y - \widehat{y} \end{pmatrix}, \nabla\mathcal{L}(x, y) - \nabla\mathcal{L}(\widehat{x}, \widehat{y})\rangle \geq \alpha_x\|x - \widehat{x}\|_2^2 + \alpha_y\|y - \widehat{y}\|_2^2 - \alpha_x\epsilon_x^2 - \alpha_y\epsilon_y^2. \tag{6}$$

Note that we require joint RSC to hold only at the target $(\widehat{x}, \widehat{y})$. In other regions of $\mathcal{X} \times \mathcal{Y}$, we may not have joint convexity if the variables x and y are not identifiable from each other in general (for instance, this arises in low-rank + sparse decomposition problems).

Next, we assume that, marginally in x and in y, the loss function satisfies the restricted smoothness (RSM) property near the optimal point $(\widehat{x}, \widehat{y})$:

Assumption 2. (Restricted smoothness (RSM).) *For all $x \in \mathcal{X}_0$ and all $y \in \mathcal{Y}_0$,*

$$D^2(x;\widehat{x}) \leq \beta_x\|x - \widehat{x}\|_2^2 + \alpha_x\epsilon_x^2 \quad \text{and} \quad D^2(y;\widehat{y}) \leq \beta_y\|y - \widehat{y}\|_2^2 + \alpha_y\epsilon_y^2. \tag{7}$$

Comparing to the restricted strong convexity assumption, we see that we need to choose constants $\alpha_x \leq \beta_x$ and $\alpha_y \leq \beta_y$.

We also mention that we allow one of the smoothness parameters to be much larger relative to the other, i.e. $\beta_x \gg \beta_y$ or vice versa, unlike methods that work jointly in the combined (x, y) variable (e.g. gradient descent on this single combined variable), whose performance is closely tied to the smoothness of the total problem, $\max\{\beta_x, \beta_y\}$. We will discuss this distinction in more detail in Section 3.2.1.

Finally, we require a "cross-product" condition (explained below):

Assumption 3. (Cross-product bound.) *For all $x \in \mathcal{X}_0$ and all $y \in \mathcal{Y}_0$,*

$$|\langle x - \widehat{x}, \nabla_x\mathcal{L}(x, y) - \nabla_x\mathcal{L}(x, \widehat{y})\rangle - \langle y - \widehat{y}, \nabla_y\mathcal{L}(x, y) - \nabla_y\mathcal{L}(\widehat{x}, y)\rangle| \\ \leq \frac{1}{2}\mu_x\|x - \widehat{x}\|_2^2 + \frac{1}{2}\mu_y\|y - \widehat{y}\|_2^2 + \alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2,$$

where $0 \leq \mu_x \leq \alpha_x$ and $0 \leq \mu_y \leq \alpha_y$.

To understand this assumption, suppose that $\mathcal{L}$ is twice differentiable. In this case, applying Taylor's theorem to rewrite the above expression in terms of $\nabla^2\mathcal{L}$, we find that Assumption 3 holds with

$$\mu_x = \mu_y = \sup_{\substack{x\in\mathcal{X}_0; y\in\mathcal{Y}_0 \\ t,t'\in[0,1]}} 2\|\nabla^2_{xy}\mathcal{L}(x, ty + (1-t)\widehat{y}) - \nabla^2_{xy}\mathcal{L}(t'x + (1-t')\widehat{x}, y)\|_{\text{op}},$$

where the norm $\|\cdot\|_{\text{op}}$ is the matrix operator norm (the largest singular value). Since $\mathcal{X}_0$ and $\mathcal{Y}_0$ are bounded via the radii ρ_x, ρ_y, then, this condition is satisfied whenever ∇^2_{xy} is Lipschitz. As a special case, if $\mathcal{L}(x, y)$ is quadratic, then we can trivially take $\mu_x = \mu_y = 0$ since ∇^2_{xy} is constant.

Initialization As our theoretical results mainly concern the local behavior of the alternating minimization method, the initialization scheme is crucial to ensure the success of the procedure. Our results require the following initialization condition:

Assumption 4. (Initialization condition.)

$$2\gamma(\mathcal{X}_0) \cdot \left(\|\nabla_x \mathcal{L}(\widehat{x}, \widehat{y})\|_x^* + \max_{y \in \mathcal{Y}_0} \|\nabla_x \mathcal{L}(x_y, y)\|_x^* \right) \leq \alpha_x - \mu_x,$$

and

$$2\gamma(\mathcal{Y}_0) \cdot \left(\|\nabla_y \mathcal{L}(\widehat{x}, \widehat{y})\|_y^* + \max_{x \in \mathcal{X}_0} \|\nabla_y \mathcal{L}(y, y_x)\|_y^* \right) \leq \alpha_y - \mu_y.$$

Recall that Lemma 2 provides easy bounds on $\gamma(\mathcal{X}_0)$ and $\gamma(\mathcal{Y}_0)$, as long as the radii ρ_x, ρ_y are chosen to be sufficiently small; furthermore, if $\mathcal{X}$ is convex, then $\gamma(\mathcal{X}_0) = 0$ and so the first bound holds trivially, and similarly for the second bound if $\mathcal{Y}$ is convex. In the nonconvex setting where $\gamma(\mathcal{X}_0)$ and/or $\gamma(\mathcal{Y}_0)$ are nonzero, see Barber and Ha [2017] for a discussion of the necessity of this type of initialization condition for the related problem of gradient descent in a single variable; we believe that this type of condition is necessary for alternating minimization as well, in the absence of additional assumptions.

3.2 Convergence guarantee

Now we show that by alternating optimization over x and over y, we obtain fast convergence to the target $(\widehat{x}, \widehat{y})$ up to the level of a small statistical error term. We prove convergence by working with the first-order divergence defined in (4), (5) above.

While the divergence may take negative values in general, according to Assumption 1, it will be always nonnegative in the regions $\mathcal{X}_0$ and $\mathcal{Y}_0$, up to the statistical error. The following result then provides guarantee on convergence of alternating minimization (2) as measured in the square-root divergences $D(x; \widehat{x})$ and $D(y; \widehat{y})$:

Theorem 1. *Suppose that Assumptions 1, 2, 3, and 4 hold. Then the iterations of the alternating minimization algorithm* (2) *satisfy the recursive bounds*

$$D(x_t; \widehat{x}) \leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{3(\alpha_x \epsilon_x^2 + \alpha_y \epsilon_y^2)}, \text{ and} \tag{8}$$

$$D(y_t; \widehat{y}) \leq \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot D(x_t; \widehat{x}) + \sqrt{3(\alpha_x \epsilon_x^2 + \alpha_y \epsilon_y^2)}, \tag{9}$$

for all $t \geq 1$. *In particular, this implies a linear rate of convergence:*

$$\|(x_t, y_t) - (\widehat{x}, \widehat{y})\|_2 \leq \left(\sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \right)^t \cdot \frac{\sqrt{6\beta_y}\rho_y}{\sqrt{\min\{\alpha_x, \alpha_y\}}} + C \cdot \max\{\epsilon_x, \epsilon_y\} \tag{10}$$

for all $t \geq 1$, *where*

$$C = \frac{18}{1 - \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot \sqrt{1 - \frac{\alpha_y}{2\beta_y}}} \cdot \sqrt{\frac{\max\{\alpha_x, \alpha_y\}}{\min\{\alpha_x, \alpha_y\}}}.$$

This theorem is proved in Section 7.

Before proceeding, we remark that the order of the updates—that is, after initializing at time $t = 0$ with points x_0, y_0, at time $t = 1$ we then update first x and then y—is arbitrary. In particular, the term $\sqrt{\beta_y}\rho_y$ appearing in the numerator of (10), can of course be replaced instead by $\sqrt{\beta_x}\rho_x$ if we switch the order of the updates. This suggests that it may be best to first update the variable with *poorer* smoothness parameter—that is, if the y variable is more well-conditioned, at our first step we should fix y and update x.

3.2.1 Dependence on condition number

Examining the bound (10) for the convergence rate in the ℓ_2 norm, we see that the convergence rate is dominated by the radius

$$\sqrt{1-\frac{\alpha_x}{2\beta_x}}\cdot\sqrt{1-\frac{\alpha_y}{2\beta_y}}$$

(here we ignore the negligible statistical error term $C\cdot\max\{\epsilon_x,\epsilon_y\}$). We now discuss the implications of this result, in terms of its dependence on the convexity and smoothness parameters, α_x,α_y and β_x,β_y. To help us discuss the conditioning of this problem, we define the two marginal condition numbers of the loss function with respect to the x and the y variables,

$$\kappa_x(\mathcal{L})=\frac{\beta_x}{\alpha_x} \text{ and } \kappa_y(\mathcal{L})=\frac{\beta_y}{\alpha_y},$$

and the joint condition number

$$\kappa(\mathcal{L})=\frac{\max\{\beta_x,\beta_y\}}{\min\{\alpha_x,\alpha_y\}}\geq\max\{\kappa_x(\mathcal{L}),\kappa_y(\mathcal{L})\},$$

which, up to constant factors, gives the condition number of the loss function $\mathcal{L}$ as a function of the joint variable (x,y).

In (10), we see that our convergence radius is strictly smaller than 1, as long as *either* of the two marginal condition numbers is bounded from above, that is, if $\min\{\kappa_x(\mathcal{L}),\kappa_y(\mathcal{L})\}$ is bounded. On the other hand, if we consider optimization algorithms that work with the combined joint variable (x,y), the performance of such algorithms typically relies heavily on the joint condition number $\kappa(\mathcal{L})\geq\max\{\kappa_x(\mathcal{L}),\kappa_y(\mathcal{L})\}$. For example, if $\mathcal{L}$ is α-strongly convex and β-smooth in the joint variable (x,y), standard results (see e.g. Bubeck [2015]) prove that gradient descent in (x,y) yields

$$\|(x_t,y_t)-(\widehat{x},\widehat{y})\|_2\leq\left(\sqrt{1-\alpha/\beta}\right)^t\|(x_0,y_0)-(\widehat{x},\widehat{y})\|_2.$$

Comparing to our notation, it can be shown that $\alpha\leq\min\{\alpha_x,\alpha_y\}$ and $\beta\geq\max\{\beta_x,\beta_y\}$, and so the radius of covergence for (joint) gradient descent is controlled by the joint condition number, $\kappa(\mathcal{L})\geq\max\{\kappa_x(\mathcal{L}),\kappa_y(\mathcal{L})\}$.

Therefore, in settings where one of the two—$\kappa_x(\mathcal{L})$ or $\kappa_y(\mathcal{L})$—is much larger than the other, we may expect that (joint) gradient descent, or other non-alternating algorithms, might perform poorly, while alternating minimization will continue to perform well, since its linear convergence rate depends only on the best of the two condition numbers, i.e. on $\min\{\kappa_x(\mathcal{L}),\kappa_y(\mathcal{L})\}$. (As discussed earlier in Section 1.1, Beck [2015] find an analogous result without strong convexity assumptions, demonstrating that the sublinear rate of convergence for alternating minimization method is driven by minimum of the two smoothness parameters, i.e. $\min\{\beta_x,\beta_y\}$.)

We will explore this phenomenon empirically when we present numerical experiments with simulated data (see Section 6 below).

4 Inexact alternating minimization

In some settings, it may be impractical to solve the alternating minimization steps exactly, i.e. when $\mathcal{L}(x,y)$ is difficult to minimize even as a function of only x or only y. In these cases, we may want to solve each step of the alternating minimization algorithm inexactly. We first state a general result for this inexact setting, then discuss the specific strategy of taking the approximate steps via alternating gradient descent.

To study the convergence behavior of alternating minimization where the steps are computed only approximately, we formulate an inexact algorithm where, at each step, we choose x_t and y_t to be within some tolerance parameters δ_t^x,δ_t^y of the exact alternating minimization steps at that time: for all $t\geq1$,

$$\begin{cases}x_t^{\text{exact}}=\arg\min_{x\in\mathcal{X}_0}\mathcal{L}(x,y_{t-1}), & x_t\in\mathcal{X}_0\cap\mathbb{B}_2(x_t^{\text{exact}},\delta_t^x),\\ y_t^{\text{exact}}=\arg\min_{y\in\mathcal{Y}_0}\mathcal{L}(x_t,y), & y_t\in\mathcal{X}_0\cap\mathbb{B}_2(y_t^{\text{exact}},\delta_t^y).\end{cases}\tag{11}$$

Here x_t and y_t can be chosen arbitrarily (or even adversarially) as long as they are within the required distance of the true solutions x_t^{exact} and y_t^{exact}.

In order to establish the convergence of the inexact alternating minimization algorithm (11), we require an additional assumption:

Assumption 5. (Relaxed triangle inequality.) *For all $x, x' \in \mathcal{X}_0$,*

$$D(x;\widehat{x}) \leq D(x';\widehat{x}) + \sqrt{\beta_x}\|x - x'\|_2 + \sqrt{\alpha_x}\epsilon_x,$$

and for all $y, y' \in \mathcal{Y}_0$,

$$D(y;\widehat{y}) \leq D(y';\widehat{y}) + \sqrt{\beta_y}\|y - y'\|_2 + \sqrt{\alpha_y}\epsilon_y,$$

It can be shown that a stronger form of the restricted smoothness condition (Assumption 2) implies this type of relaxed triangle inequality, but for simplicity we state it as an assumption.

The following theorem states that the inexact alternating minimization inherits fast convergence of the alternating minimization steps to the target $(\widehat{x}, \widehat{y})$, under the same assumptions as the original result Theorem 1, along with the relaxed triangle inequality (Assumption 5).

Theorem 2. *Suppose that Assumptions 1, 2, 3, 4, and 5 hold. Then, the steps of the inexact alternating minimization algorithm satisfy*

$$D(x_t;\widehat{x}) \leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1};\widehat{y}) + \sqrt{\beta_x}\delta_t^x + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)} \quad \text{and} \tag{12}$$

$$D(y_t;\widehat{y}) \leq \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot D(x_t;\widehat{x}) + \sqrt{\beta_y}\delta_t^y + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)}, \tag{13}$$

for all $t \geq 1$.

This theorem is proved in Section 7.

Of course, in order for this result to be meaningful, the slack terms δ_t^x, δ_t^y need to be sufficiently small, so that the errors $D(x_t;\widehat{x})$ and $D(y_t;\widehat{y})$ are able to converge to zero (or, at least, to the level of the statistical error terms ϵ_x, ϵ_y). As a special case, consider the setting where the slack terms δ_t^x, δ_t^y decrease as the solution converges, via the rule

$$\delta_t^x \leq c_x\|x_{t-1} - x_t^{\text{exact}}\|_2 + C_x\epsilon_x, \quad \delta_t^y \leq c_y\|y_{t-1} - y_t^{\text{exact}}\|_2 + C_y\epsilon_y, \tag{14}$$

for some sufficiently small $c_x, c_y \geq 0$ and for some $C_x, C_y < \infty$.

In fact, we will see momentarily that this recursive bound arises naturally when the approximate iterative solutions x_t and y_t are obtained via alternating gradient descent. First, however, we prove that the recursive rule (14) is sufficient to ensure linear convergence as long as the constants c_x, c_y are sufficiently small.

Lemma 3. *Suppose that, for all $t \geq 1$, the slack terms δ_t^x, δ_t^y satisfy* (14)*. Then, under the assumptions of Theorem 2, if*

$$r := \left(\sqrt{1 - \frac{\alpha_x}{2\beta_x}} + 3c_y\sqrt{\frac{\beta_y}{\alpha_y}}\right) \cdot \left(\sqrt{1 - \frac{\alpha_y}{2\beta_y}} + 3c_x\sqrt{\frac{\beta_x}{\alpha_x}}\right) < 1, \tag{15}$$

then the iterations of the inexact alternating minimization algorithm (11) *satisfy*

$$\|(x_t, y_t) - (\widehat{x}, \widehat{y})\|_2 \leq r^t \cdot \frac{\sqrt{6(\alpha_x\rho_x^2 + \beta_y\rho_y^2)}}{\sqrt{\min\{\alpha_x, \alpha_y\}}} + C \cdot \max\{\epsilon_x, \epsilon_y\}$$

for all $t \geq 1$, where

$$C = \frac{39}{1 - r} \cdot \sqrt{\frac{\alpha_x + \alpha_y + C_x^2\beta_x + C_y^2\beta_y}{\min\{\alpha_x, \alpha_y\}}}.$$

The proof of this lemma is given in Appendix B.2.

We should interpret this lemma as covering two distinct scenarios:

- First, if the loss is well-conditioned in both the x and the y variables—that is, both $\frac{\beta_x}{\alpha_x}$ and $\frac{\beta_y}{\alpha_y}$ are bounded—then we can afford inexact update steps for both variables, allowing c_x, c_y to both be small positive constants while still obtaining linear convergence.

- Alternately, if the loss is well-conditioned in one variable only—without loss of generality, if $\frac{\beta_x}{\alpha_x}$ is large (or even $\beta_x = \infty$) while $\frac{\beta_y}{\alpha_y}$ is bounded—then we can allow the y variable update to be performed inexactly, while the x variable should be updated with the exact alternating minimization step (that is, $c_x = C_x = 0$, i.e. $\delta_x^t = 0$ at each update iteration t). In this case, we can still obtain a linear convergence rate.

We will see examples of this second setting in the simulations.

4.1 Alternating gradient descent

As mentioned above, if the alternating minimization update for x is approximated via gradient descent in x (and same for y), then we may expect the errors in each step to scale linearly as in (14). We now give details for this claim, relying on earlier work Barber and Ha [2017], which we summarize here.

To run the alternating descent algorithm, we first initialize at $x_0 \in \mathcal{X}$, $y_0 \in \mathcal{Y}$, then, at each iteration $t = 1, 2, \ldots,$

1. Perform m_x many gradient descent steps for x:

$$\begin{cases} \text{Set } x_{t;0} = x_{t-1}; \\ \text{For } m = 1, \ldots, m_x, \text{ set } x_{t;m} = \mathcal{P}_{\mathcal{X}_0}\left(x_{t;m-1} - \eta_x \nabla_x \mathcal{L}(x_{t;m-1}, y_{t-1})\right); \\ \text{Set } x_t = x_{t;m_x}. \end{cases} \tag{16}$$

2. Perform m_y many gradient descent steps for y:

$$\begin{cases} \text{Set } y_{t;0} = y_{t-1}; \\ \text{For } m = 1, \ldots, m_y, \text{ set } y_{t;m} = \mathcal{P}_{\mathcal{Y}_0}\left(y_{t;m-1} - \eta_y \nabla_y \mathcal{L}(x_t, y_{t;m-1})\right); \\ \text{Set } y_t = y_{t;m_y}. \end{cases} \tag{17}$$

Here $\mathcal{P}_{\mathcal{X}_0}$ and $\mathcal{P}_{\mathcal{Y}_0}$ denote projection, with respect to the ℓ_2 norm, to the sets $\mathcal{X}_0$ and $\mathcal{Y}_0$. (Of course, this projection may not be unique in the presence of nonconvexity—if this is the case, we can take any one of the closest points.)

Often, the number of steps in each "inner loop", namely m_x and m_y, can be taken to be a small constant (or even 1) to obtain good empirical performance. In our theoretical analysis, under restricted strong convexity and restricted smoothness, constant values for m_x and m_y suffice to guarantee convergence.

Convergence results Gradient descent (and its variants) now serves as a popular tool in large-scale optimization problem, due to its scalablity to the high-dimensional setting. For convex constraints, the convergence behavior of the gradient descent has been well established (e.g. Nesterov [2013], Nesterov et al. [2007], Bubeck [2015]) and also generalized to the high-dimensional setting (e.g. Agarwal et al. [2010]). In contrast to the convex setting, gradient descent under nonconvex constraints is more challenging to analyze theoretically, although it often performs well empirically.

In our own recent work, Barber and Ha [2017], we have shown that, under the framework of local concavity measure (Definition 2), gradient descent converges rapidly to the optimal as long as it is initialized near the target. Here we follow the same setup, but slightly modify the assumptions to better match the setting of the alternating minimization problem. We assume that:

- The "marginal" loss functions $x \mapsto \mathcal{L}(x, y)$ and $y \mapsto \mathcal{L}(x, y)$ satisfy restricted strong convexity (RSC) uniformly over $\mathcal{X}_0$ and over $\mathcal{Y}_0$, respectively. In other words, for all $x, x' \in \mathcal{X}_0$ and all $y \in \mathcal{Y}_0$,

$$\mathcal{L}(x, y) - \mathcal{L}(x', y) - \langle \nabla_x \mathcal{L}(x', y), x - x' \rangle \geq \frac{\alpha_x}{2} \|x - x'\|_2^2 - \frac{\alpha_x}{2} \cdot \epsilon_x^2, \tag{18}$$

  and analogously with the roles of x and y reversed.

- The "marginal" functions $x \mapsto \mathcal{L}(x, y)$ and $y \mapsto \mathcal{L}(x, y)$ satisfy restricted smoothness (RSM) in $\mathcal{X}_0$ and $\mathcal{Y}_0$, that is,
$$\mathcal{L}(x, y) - \mathcal{L}(x', y) - \langle \nabla_x \mathcal{L}(x', y), x - x' \rangle \leq \frac{\beta_x}{2} \|x - x'\|_2^2 + \frac{\alpha_x}{2} \cdot \epsilon_x^2, \tag{19}$$
and analogously with the roles of x and y reversed.

- There exist parameters $\phi_x, \phi_y \geq 1$ such that for any $z \in \mathbb{R}^{d_x}$,
$$\|z - \mathcal{P}_{\mathcal{X}_0}(z)\|_x^* \leq \phi_x \min_{x' \in \mathcal{X}_0} \|z - x'\|_x^*, \tag{20}$$
and analogously with the roles of x and y reversed (note that this condition holds trivially with $\phi_x = 1$ if $\|\cdot\|_x = \|\cdot\|_2$, but $\phi_x = 1$ often suffices even for other norms).

- For all $y \in \mathcal{Y}_0$,
$$2\phi_x \cdot \max_{x, x' \in \mathcal{X}_0} \gamma_x(\mathcal{X}) \|\nabla_x \mathcal{L}(x', y)\|_x^* \leq (1 - a_x) \cdot \alpha_x \tag{21}$$
for some constant $a_x > 0$, and analogously with the roles of x and y reversed with some constant $a_y > 0$.

We remark that the assumptions (18) and (19) are made in terms of objective function, as opposed to the gradient forms in the corresponding conditions from earlier, given in (6) for RSC and (7) for RSM. However, in general, the two forms of these conditions are roughly interchangeable with each other. For detailed discussion on the rest of the assumptions, see Barber and Ha [2017] and references therein.

With these assumptions in place, we can bound the tolerance parameters δ_t^x, δ_t^y appearing in the inexact alternating minimization algorithm (11), when the inexact steps are computed via gradient descent.

Lemma 4 (Barber and Ha [2017, Theorem 3])**.** *Suppose that conditions* (18)*,* (19)*,* (20) *and* (21) *hold. Then the output of m_x many gradient descent steps on the x variable, given in* (16)*, satisfies*

$$\|x_t - x_t^{\text{exact}}\|_2^2 \leq \left(1 - a_x \frac{2\alpha_x}{\alpha_x + \beta_x}\right)^{m_x} \cdot \|x_{t-1} - x_t^{\text{exact}}\|_2^2 + \frac{1.5}{a_x} \cdot \epsilon_x^2.$$

The analogous statement holds with the roles of x and y reversed.

Examining the conditions (14) and (15) on the allowed size of the slack terms δ_t^x and δ_t^y, we can see that taking $c_x = \mathcal{O}\left((\alpha_x/\beta_x)^{1.5}\right)$ and $c_y = \mathcal{O}\left((\alpha_y/\beta_y)^{1.5}\right)$ is sufficient to ensure that the condition (15) will hold. This yields the following corollary, which we state informally to avoid overly complicated constants:

Corollary 1. *Under the assumptions of Lemmas 3 and 4, for some radius* Rad < 1*,*

$$\|(x_t, y_t) - (\widehat{x}, \widehat{y})\|_2 \leq \mathcal{O}\left(\text{Rad}^t \cdot \max\{\rho_x, \rho_y\} + \max\{\epsilon_x, \epsilon_y\}\right)$$

for all $t \geq 1$ as long as

$$\begin{cases} \text{Either the } x \text{ update is exact, or is approximated via } m_x = \mathcal{O}\left(\frac{\beta_x}{\alpha_x} \log\left(\frac{\beta_x}{\alpha_x}\right)\right) \text{ many steps of gradient descent in } x; \\ \text{Either the } y \text{ update is exact, or is approximated via } m_y = \mathcal{O}\left(\frac{\beta_y}{\alpha_y} \log\left(\frac{\beta_y}{\alpha_y}\right)\right) \text{ many steps of gradient descent in } y. \end{cases}$$

5 Examples

In this section, we highlight applications of our general theory to two classes of low-rank estimation problems: matrix decomposition and multitask regression. In each setting, we present a target optimization problem and verify the assumptions of our main results, under the suitable choices of the radii ρ_x, ρ_y. All results in this section are proved in Appendix C.

5.1 Matrix decomposition

In robust principal component analysis (RPCA) and in Gaussian factor models, it is common to assume that the matrix of interest is formed by a sum of low-rank and sparse components. Given the data generated through a matrix $X^\star + Y^\star$, where $X^\star$ is low-rank and $Y^\star$ is sparse, the task of matrix decomposition is to recover both the low-rank and sparse components simultaneously.

While this problem is generally known to be ill-posed, certain incoherence conditions on the low-rank component have been shown to guarantee exact or approximate recovery of both low-rank and sparse components (e.g. Candès et al. [2011], Chandrasekaran et al. [2011]). Following Negahban and Wainwright [2012], we rely on the condition called *spikiness condition*, which restricts the class of low-rank matrices by imposing the constraint $\|X^\star\|_\infty \leq \frac{\alpha_{\mathsf{sp}}}{d}$.

Suppose that the underlying low-rank matrix $X^\star$ is positive semidefinite with rank r and the sparse matrix $Y^\star$ is symmetric with at most s nonzero entries per each row. This gives rise to the following constraint sets:[1]

$$\mathcal{X} = \left\{ X \in \mathbb{S}_+^{d\times d} : \operatorname{rank}(X) \leq r,\ \|X\|_\infty \leq \frac{\alpha_{\mathsf{sp}}}{d} \right\},\ \mathcal{Y} = \left\{ Y \in \mathbb{S}^{d\times d} : \|Y\|_1 \leq \|Y^\star\|_1 \right\}, \tag{22}$$

where $\mathbb{S}$ and $\mathbb{S}_+$ denote the sets of symmetric or positive semidefinite $d \times d$ matrices, respectively. The ℓ_∞-ball constraint in the set $\mathcal{X}$ makes sure that the low-rank update by the algorithm is at most α_{sp}-spiky at every iteration. To compute the local concavity coefficient of the set $\mathcal{X}$, the following lemma provides the upper bound on the concavity coefficient $\gamma_X(\mathcal{X})$:

Lemma 5. *For the constraint set $\mathcal{X} = \{X \in \mathbb{S}_+^{d\times d} : \operatorname{rank}(X) \leq r, \|X\|_\infty \leq \frac{\alpha_{\mathsf{sp}}}{d}\}$, we have $\gamma_X(\mathcal{X}) \leq \frac{5}{4\sigma_r(X)}$ with respect to the nuclear norm $\|\cdot\|_X = \|\cdot\|_{\text{nuc}}$.*

Note that while the set of rank-constrained matrices without spikiness constraint has the local concavity coefficient $\gamma_X(\mathcal{X}) = \frac{1}{2\sigma_r(X)}$ (see Barber and Ha [2017, Lemma 7]), the lemma above shows that the coefficient for $\mathcal{X}$ can be upper bounded with a larger constant factor.

For a given loss function $\mathcal{L}(X, Y)$, we can then recover the underlying matrices $(X^\star, Y^\star)$ by solving the constrained optimization problem

$$(\widehat{X}, \widehat{Y}) = \arg\min\{\mathcal{L}(X, Y) : (X, Y) \in \mathcal{X} \times \mathcal{Y}\}$$

via alternating minimization or alternating gradient descent. Two specific instances of the loss function $\mathcal{L}(X, Y)$ arise from robust PCA and from the Gaussian factor model.

Robust principal component analysis (RPCA) We study the robust PCA problem as formulated in Agarwal et al. [2012], where the data matrix $Z \in \mathbb{S}^{n\times n}$ is generated from the model

$$Z = \mathcal{A}(X^\star + Y^\star) + W,$$

where $\mathcal{A} : \mathbb{S}^{d\times d} \to \mathbb{S}^{n\times n}$ is a linear operator mapping matrices from $\mathbb{S}^{d\times d}$ to $\mathbb{S}^{n\times n}$, and $W \in \mathbb{S}^{n\times n}$ represents a symmetric noise matrix.

Our estimators are defined based on the least squares loss,

$$(\widehat{X}, \widehat{Y}) = \arg\min \left\{ \frac{1}{2}\|Z - \mathcal{A}(X + Y)\|_{\text{F}}^2 : (X, Y) \in \mathcal{X} \times \mathcal{Y} \right\}. \tag{23}$$

In determining the properties of the loss function, it is crucial that the operator $\mathcal{A}$ satisfies certain conditions. Following the notion of RSC as introduced in Agarwal et al. [2012, Definition 2], we require $\mathcal{A}$ to satisfy the following property:

Assumption 6. (Restricted Eigenvalue.) *There exist constants α_A, β_A and $\tau \geq 0$ such that for all $\Delta_X, \Delta_Y \in \mathbb{R}^{d\times d}$ with $\operatorname{rank}(\Delta_X) \leq 2r$,*

$$\alpha_A \left(\|\Delta_X\|_{\text{F}}^2 + \|\Delta_Y\|_{\text{F}}^2 \right) - \tau_{n,d} \leq \|\mathcal{A}(\Delta_X + \Delta_Y)\|_{\text{F}}^2 \leq \beta_A \left(\|\Delta_X\|_{\text{F}}^2 + \|\Delta_Y\|_{\text{F}}^2 \right) + \tau_{n,d},$$

where $\tau_{n,d}$ is given by

$$\tau_{n,d} = \tau \cdot \left(\frac{\log d}{n^2} \|\Delta_Y\|_1^2 + \sqrt{\frac{d^2 \log d}{n^2}} \|\Delta_X\|_\infty \|\Delta_Y\|_1 \right).$$

[1] For our analysis, we assume that $\|Y^\star\|_1$ is known *exactly*; this is a common assumption for the constrained problem, e.g. see Amelunxen et al. [2014]. On the other hand, $\|X^\star\|_\infty$ needs only to be bounded by some known value α_{sp}/d; we do not need to know it exactly.

Marginally in X and in Y, by taking $\Delta_Y = 0$ or $\Delta_X = 0$, we see that this assumption simply reduces to the well-known restricted eigenvalue property. The expression $\sqrt{\frac{d^2 \log d}{n^2}}\|\Delta_X\|_\infty \|\Delta_Y\|_1$ reflects the restriction on the degree of interaction between Δ_X and Δ_Y, which would hold if $\|\mathcal{A}^*\mathcal{A}(\Delta_X)\|_\infty \approx \sqrt{\frac{d^2 \log d}{n^2}}\|\Delta_X\|_\infty$ —for instance, an i.i.d. Gaussian ensemble will satisfy this property with high probability. We make this more general assumption on the operator $\mathcal{A}$ to give a deterministic result on the least squares loss $\mathcal{L}(X, Y)$.

Now let the radii ρ_X, ρ_Y satisfy $\rho_X, \rho_Y \leq c_0 \cdot \sigma_r(X^\star)\kappa^{-1}(\mathcal{A})$ for some $c_0 > 0$, where $\sigma_r(X^\star)$ is the smallest singular value of $X^\star$, and where $\kappa(\mathcal{A}) = \frac{\beta_A}{\alpha_A}$, which we can think of as a restricted condition number of the linear operator $\mathcal{A}$ (i.e. characterizing the action of $\mathcal{A}$ restricted to low-rank and sparse matrices). Given the initialization point (X_0, Y_0), denote the corresponding neighborhoods by $\mathcal{X}_0 = \mathcal{X} \cap \mathbb{B}_2(X_0, \rho_X), \mathcal{Y}_0 = \mathcal{Y} \cap \mathbb{B}_2(Y_0, \rho_Y)$, and further assume that both the underlying matrices $(X^\star, Y^\star)$ and the global optimal $(\widehat{X}, \widehat{Y})$ belong to these local neighborhoods $\mathcal{X}_0 \times \mathcal{Y}_0$. With this setup, we then have the following guarantee:

Lemma 6. *Suppose that the sample size is large enough to satisfy*

$$\frac{32\tau \cdot sd \log d}{n^2} \leq \alpha_A. \tag{24}$$

Then, under the previously stated conditions, if $\|\mathcal{A}^*(W)\|_{\text{op}} \leq c_1 \cdot \alpha_A \sigma_r(X^\star)$*, the steps* $(X_t, Y_t)_{t=1}^\infty$ *produced by the alternating minimization algorithm* (2) *satisfy*

$$\|(X_t, Y_t) - (\widehat{X}, \widehat{Y})\|_{\text{F}} \leq \overbrace{\left(1 - \frac{\kappa^{-1}(\mathcal{A})}{3}\right)^t}^{\text{linear convergence}} \cdot c_0 \sigma_r(X^\star)\sqrt{\kappa^{-1}(\mathcal{A})} + \underbrace{c_2 \cdot C\left(\|\widehat{Y} - Y^\star\|_{\text{F}}^2 + \frac{\alpha_{\mathsf{sp}}^2}{\alpha_A^2}\frac{sd \log d}{n^2}\right)}_{\text{statistical error term}}.$$

Here, $\{c_i > 0, i = 1, 2\}$ *are universal constants, and* $C > 0$ *is defined in Theorem 1.*

The proof of this lemma appears in Appendix C.2.

The result given in the lemma is the bound obtained by updating the Y variable first instead of the X variable. The statistical error involves the term $\alpha_{\mathsf{sp}}^2 \frac{sd \log d}{n^2}$, which appears as a consequence of the nonidentifiaibility of the model—see Agarwal et al. [2012] for a detailed discussion on the nonidentifiability of the matrix decomposition problem. With more effort, we can also prove the assumptions of Lemmas 3 and 4 for the least square loss, implying that each step of the alternating minimization update can be replaced by several steps of the gradient descent updates.

Several works also study alternating minimization and its variants such as alternating gradient descent employed on the robust PCA problem—for instance, see Chen and Wainwright [2015], Yi et al. [2016], Gu et al. [2016]. In contrast to our approach, these methods are all based on the factorized approach, working on a highly nonconvex loss function arising from the factorization of the low-rank matrix $X = UU^\top$ (in the case of positive semidefinite matrix), which is identifiable only up to rotations. In fact, our framework can cover this factorization scenario as well, where the restricted strong convexity (18) & restricted smoothness (19) of the original loss $\mathcal{L}(X, Y)$ can be shown to be inherited by the reparametrized loss $\mathcal{L}(UU^\top, Y)$, up to rotations, as long as the initialization condition (21) holds in the neighborhood $\mathcal{X}_0$. Although this approach is not the focus of our current paper, we refer the reader to Chen and Wainwright [2015], Tu et al. [2015] and Zheng and Lafferty [2015] for the related topic; see also Gu et al. [2016] and the references therein.

Gaussian factor model We next consider a Gaussian factor model, where our data consists of observations $z_i = U^\star w_i + \epsilon_i$, for $i = 1, \ldots, n$. Here $U^\star \in \mathbb{R}^{d \times r}$ represents the latent structure present in the data, while the other terms in the model are the random factors $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_r)$ and the independent noise $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, Y^\star)$. We further assume that the covariance structure of the noise, $Y^\star$, is sparse. We can calculate $\text{Cov}(z_i) = U^\star U^{\star\top} + Y^\star$, a low-rank + sparse decomposition, and can then estimate the unknown components $X^\star = U^\star U^{\star\top}$ and $Y^\star$ by solving the constrained optimization problem

$$(\widehat{X}, \widehat{Y}) = \arg\min\left\{\langle S_n, (X + Y)^{-1}\rangle - \log\det(X + Y)^{-1} : (X, Y) \in \mathcal{X} \times \mathcal{Y}\right\},$$

for $S_n = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top$, the sample covariance matrix of z_i's, where $\mathcal{X}$ and $\mathcal{Y}$ are defined as in (22). Full details for results of the Gaussian factor model appear in Appendix A.

5.2 Multitask regression

Next we consider a class of regression problems where the response contains more than a single observation, that is, it takes multiple output values. Each component of the output corresponds to a specific task.

Suppose we are given m different tasks, where for each observation the response is of the form $z_i \in \mathbb{R}^m$. In the multitask regression model, the m tasks are assumed to share the same feature vector, which we denote by $\phi_i \in \mathbb{R}^d$, and the response is generated through the linear model

$$z_i = X^\star \phi_i + \epsilon_i, \tag{25}$$

where $X^\star \in \mathbb{R}^{m\times d}$ is an unknown matrix whose rows correspond to the underlying coefficient vectors for each task, and $\epsilon_i \in \mathbb{R}^m$ is the measurement error from a centered multivariate normal distribution, with an unknown covariance matrix $\mathrm{Cov}(\epsilon_i) = \Theta^{\star -1}$. Given such model, we would like to recover the unknown matrices $X^\star$ and $\Theta^\star$ from the data $(z_i, \phi_i)_{i=1}^n$. In the high-dimensional setting, it is common to estimate the matrix $X^\star$ under a low-rank constraint; this method is also referred to as "reduced rank regression" in the literature, e.g. Izenman [1975].

We would then like to optimize the constrained negative log-likelihood function,

$$(\widehat{X}, \widehat{\Theta}) = \arg\min \left\{ -\log\det(\Theta) + \frac{1}{n}\sum_{i=1}^n (z_i - X\phi_i)^\top \Theta (z_i - X\phi_i) : X \in \mathcal{X}, \Theta \succeq 0 \right\}, \tag{26}$$

where $\mathcal{X} = \{X \in \mathbb{R}^{m\times d} : \mathrm{rank}(X) \le \mathrm{rank}(X^\star) = r\}$ represents the rank constraint on the coefficients X.[2] A challenge here is that this problem is nonconvex in (X, Θ), in addition to the nonconvex constraint $X \in \mathcal{X}$, so it cannot be easily solved by standard convex optimization methods. Nevertheless, we will show that this problem satisfies all the assumptions that we require for our results, which then allows to apply the (inexact) alternating minimization algorithm to solve the problem efficiently.

For the purpose of our analysis, we consider a random design model, i.e. the feature vectors are sampled from a Gaussian distribution $\phi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_\phi)$. Let (X_0, Θ_0) be the initialization point, and denote the local neighborhoods of the constraint sets around (X_0, Θ_0) by $\mathcal{X}_0 = \mathcal{X} \cap \mathbb{B}_2(X_0, \rho_X)$ and $\mathcal{Q}_0 = \mathbb{S}_+^{m\times m} \cap \mathbb{B}_2(\Theta_0, \rho_\Theta)$. We choose the radii ρ_X, ρ_Θ to satisfy $\rho_X \le c_0 \cdot \sigma_r(X^\star)\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)$ and $\rho_\Theta \le c_0 \cdot \lambda_{\min}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)$ for some $c_0 > 0$, where $\sigma_r(X^\star)$ is the smallest singular value of $X^\star$, $\lambda_{\min}(\Theta^\star)$ is the smallest eigenvalue of $\Theta^\star$, and where $\kappa(\Theta^\star), \kappa(\Sigma_\phi)$ are the condition numbers of $\Theta^\star$ and Σ_ϕ, respectively. Assume also that the initialization point (X_0, Θ_0) lies within these radii ρ_X, ρ_Θ to the unknown matrices $(X^\star, \Theta^\star)$ and the global optimal $(\widehat{X}, \widehat{\Theta})$, i.e. $(X^\star, \Theta^\star), (\widehat{X}, \widehat{\Theta}) \in \mathcal{X}_0 \times \mathcal{Q}_0$.

With these definitions in place, we have the following probabilistic guarantee:

Lemma 7. *Suppose that*

$$\sqrt{\frac{1}{\lambda_{\min}(\Theta^\star)\lambda_{\max}(\Sigma_\phi)}}\sqrt{\frac{m+d}{n}} \le c_1 \cdot \sigma_r(X^\star)\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi). \tag{27}$$

Then, under the previously stated conditions, with probability at least $1 - c_2\exp(-c_3(m+d))$*, the steps* $(X_t, \Theta_t)_{t=1}^\infty$ *produced by the alternating minimization algorithm* (2) *satisfy*

$$\|(X_t, \Theta_t) - (\widehat{X}, \widehat{\Theta})\|_{\mathrm{F}} \le \overbrace{\left(1 - c_4(\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi) + \kappa^{-2}(\Theta^\star))\right)^t}^{\text{linear convergence}} \cdot (\text{Const}) + \underbrace{c_5 \cdot C\left(\|\widehat{X} - X^\star\|_{\mathrm{F}}^2 + \frac{r(m+d)}{n}\frac{1}{\lambda_{\min}(\Theta^\star)\lambda_{\max}(\Sigma_\phi)}\right)}_{\text{statistical error term}}$$

for all $t \ge 1$*, where* (Const) *is given by*

$$(\text{Const}) = c_6\sigma_r(X^\star)\sqrt{\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)} \cdot \min\left\{1, \lambda_{\min}^3(\Theta^\star)\kappa^2(\Theta^\star)\lambda_{\min}(\Sigma_\phi)\right\}.$$

Here, $\{c_i > 0, i = 1, \dots, 6\}$ *are universal constants, and* $C > 0$ *is defined in Theorem 1.*

[2] It is also possible to consider structural constraints on Θ or $\Sigma = \Theta^{-1}$ such as "sparsity" or "low-rank + diagonal" structure. For simplicity, we don't pursue this direction further, but our framework can be also applied to this general setting.

While Lemma 7 is stated in terms of the *exact* alternating minimization, by working with the inexact versions Lemmas 3 and 4, we can also obtain a linear rate of convergence for the alternating method when the minimization step for X is approximated by successive iterates of gradient descent. (The alternating minimization update for Θ has a closed form solution, $\Theta = \left(\frac{1}{n}\sum_{i=1}^n (z_i - X\phi_i)(z_i - X\phi_i)^\top\right)^{-1}$, i.e. the inverse of the sample covariance matrix.) The result in Lemma 7 assumes updating Θ first, but the analogous result is also available if the algorithm begins with X updates instead.

As a final remark, an alternative approach to estimate $X^\star$ is by treating each task as a separate regression problem and simply performing least square procedure. However, it is generally known that the solution resulting from (26) is significantly better than the least squares estimators, because the estimator defined in (26) exploits the correlation across tasks, whereas the least squares estimators ignore such correlation. We refer the reader to Jain and Tewari [2015] for similar results under the context of the pooled model.

6 Empirical results

We perform a numerical experiment on the multitask regression problem (Section 5.2) to examine the empirical performance of the alternating algorithm, as compared to performing gradient descent when treating (x, y) as a single variable.[3] Fix the number of tasks $m = 20$, the dimension of features $d = 50$, and set the low-rank component $X^\star = U^\star V^{\star\top}$ for rank $r = 3$, where $U^\star \in \mathbb{R}^{20\times 3}$ and $V^\star \in \mathbb{R}^{50\times 3}$ are orthonormal matrices drawn uniformly at random. The features ϕ_i are drawn i.i.d. from the Gaussian distribution $\phi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_\phi)$, and the noise terms ϵ_i are generated as $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \Theta^{\star -1})$, where Σ_ϕ and $\Theta^{\star -1}$ are both defined to have a tapered covariance structure: we set $\Sigma_{\phi,ij} = 0.3^{|i-j|}$ and $\Theta^\star_{ij}{}^{-1} = \sigma^2 \cdot \rho^{|i-j|}$, where ρ is a local correlation parameter that we vary, while $\sigma^2 = \frac{\text{Mean}(\|X^\star\phi_i\|_{\text{F}}^2/m)}{3}$ is chosen to obtain a moderately difficult signal-to-noise ratio. The responses, z_i, are then drawn according to the model (25).

The parameter ρ controls the strength of the correlation of the noise (i.e. correlation among entries of ϵ_i, for a single observation i, across the $m = 20$ tasks). By varying ρ, we can vary the relative condition numbers of the loss function $\mathcal{L}(X, \Theta)$ given in (26) with respect to the variables X and Θ, i.e. $\kappa_X(\mathcal{L})$ versus $\kappa_\Theta(\mathcal{L})$. As discussed in Section 3.2.1, convergence rates for alternating minimization type methods are expected to scale with the *minimum* of these two condition numbers, while non-alternating methods (i.e. gradient descent in the joint variable (X, Θ)) will scale with the *maximum* of the two.

Given the data $(\phi_i, z_i)_{i=1}^n$ with sample size $n = 200$, we solve the constrained minimization problem (26) based on two iterative methods:

- The alternating method, which alternates between updating X and Θ at every iteration. For the X update, fixing Θ we approximately minimize $\mathcal{L}(X, \Theta)$ by taking one gradient descent step, while for the Θ update, fixing X we minimize $\mathcal{L}(X, \Theta)$ exactly:

$$\begin{cases} X_t = \mathcal{P}_{\{\text{rank}(X)\le r\}}\left(X_{t-1} + \eta_X \cdot 2\Theta_{t-1}\left(\frac{1}{n}\sum_{i=1}^n (z_i - X_{t-1}\phi_i)\phi_i^\top\right)\right), \\ \Theta_t = \arg\min_{\Theta\succeq 0} \mathcal{L}(X_t, \Theta) = \left(\frac{1}{n}\sum_{i=1}^n (z_i - X_t\phi_i)(z_i - X_t\phi_i)^\top\right)^{-1}, \end{cases}$$

  with step size $\eta_X = 0.001$.

- The joint gradient method, where we take gradient descent steps in the joint variable (X, Θ). The update step is given by

$$\begin{cases} X_t = \mathcal{P}_{\{\text{rank}(X)\le r\}}\left(X_{t-1} + \eta_X \cdot 2\Theta_{t-1}\left(\frac{1}{n}\sum_{i=1}^n (z_i - X_{t-1}\phi_i)\phi_i^\top\right)\right), \\ \Theta_t = \mathcal{P}_{\{\Theta\succeq 0\}}\left(\Theta_{t-1} + \eta_\Theta \cdot \left(\Theta_{t-1}^{-1} - \frac{1}{n}\sum_{i=1}^n (z_i - X_{t-1}\phi_i)(z_i - X_{t-1}\phi_i)^\top\right)\right), \end{cases}$$

  where we allow different step sizes on the two variables X and Θ. We set $\eta_X = 0.001$ as for the alternating method, and select $\eta_\Theta \in \{\eta_1, \ldots, \eta_{30}\}$, where $\eta_1, \ldots, \eta_{30}$ is a geometric sequence from $\eta_1 = 5$ to $\eta_{30} = 400$. For each trial, we then retain only the step size η_Θ that yields the lowest loss over any iteration, $\min_{t=1,\ldots,T} \mathcal{L}(X_t, \Theta_t)$, for the first $T = 1200$ iterations.

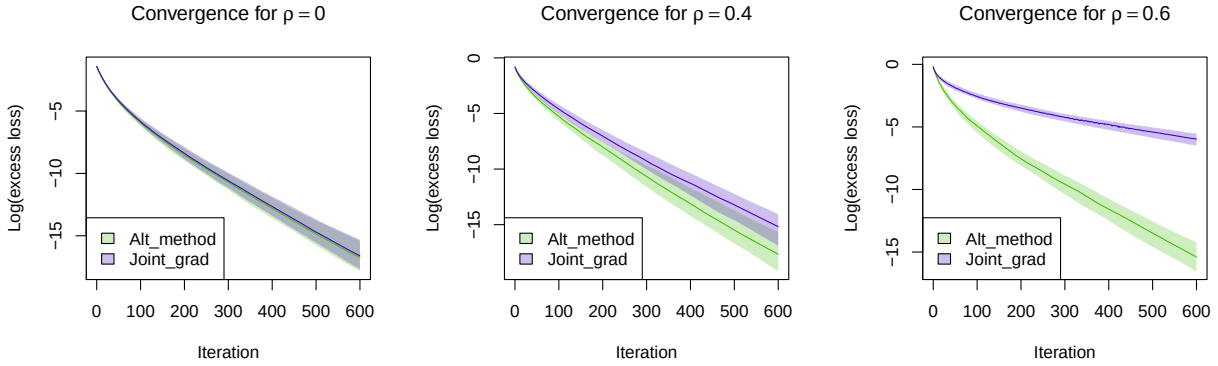

Figure 1: Comparison of the alternating algorithm and the joint gradient descent method applied to the simulated multitask regression problem. Results are shown for three settings of the noise correlation parameter ρ, across iterations $t = 0, 1, \ldots, 600$. In each plot, the solid line indicates the median loss over 100 trials, and the light band shows the interquartile range.

Figure 1 shows the excess loss at each iteration (on a log scale), where the excess loss is given by

$$\mathcal{L}(X_t, \Theta_t) - \mathcal{L}_{\min},$$

where $\mathcal{L}_{\min}$ is the minimum loss achieved by either method over $T = 1200$ iterations (calculated for each individual trial and each choice of ρ). As clearly seen in the figure, for both methods, the errors scale linearly with the iteration number. Furthermore, comparing the two methods, we see that they perform nearly identically when there is no correlation in the noise, i.e. $\rho = 0$; for low correlation, $\rho = 0.4$, the alternating method is moderately faster,[4] and for high correlation, $\rho = 0.6$, the alternating method still shows rapid linear convergence while joint gradient descent does not appear to converge well. This is consistent with our theoretical results, since the alternating method scales with the better of the two condition numbers, i.e. $\min\{\kappa_X(\mathcal{L}), \kappa_\Theta(\mathcal{L})\}$, while the joint gradient descent method is known to scale with the maximum. Since $\kappa_X(\mathcal{L}) \sim \kappa(\Sigma_\phi)\kappa(\Theta^\star)$ while $\kappa_\Theta(\mathcal{L}) \sim \kappa^2(\Theta^\star)$, and $\kappa(\Sigma_\phi)$ is constant with respect to ρ while $\kappa(\Theta)$ increases as the noise correlation parameter ρ increases, we see that the *minimum* condition number is less affected by increasing ρ, than the *maximum* condition number.

7 Proofs of theorems

In this section, we prove our main result on linear convergence for the exact alternating minimization algorithm, Theorem 1, and for the inexact algorithm, Theorem 2. All other results are proved in the Appendix.

7.1 Proof of Theorem 1

First we prove the bound on the x update step, given in (8), for iteration number t. By definition of x_t, we can apply the first-order optimality condition (1), with $\mathcal{X}_0$ in place of $\mathcal{X}$, to obtain

$$\langle \widehat{x} - x_t, \nabla_x \mathcal{L}(x_t, y_{t-1}) \rangle \geq -\gamma(\mathcal{X}_0) \|\nabla_x \mathcal{L}(x_t, y_{t-1})\|_x^* \|x_t - \widehat{x}\|_2^2.$$

Meanwhile, since $\widehat{x}$ is the minimizer of the problem $\min\{\mathcal{L}(x, \widehat{y}) : x \in \mathcal{X}_0\}$, we also have

$$\langle x_t - \widehat{x}, \nabla_x \mathcal{L}(\widehat{x}, \widehat{y}) \rangle \geq -\gamma(\mathcal{X}_0) \|\nabla_x \mathcal{L}(\widehat{x}, \widehat{y})\|_x^* \|x_t - \widehat{x}\|_2^2.$$

[3] Code available at `http://www.stat.uchicago.edu/~rina/code/altmin_simulation.R`.

[4] Note that the shaded bands in the plots are *not* standard error bars, but rather interquartile range over 100 trials, so the difference between the two lines is indeed significant.

Adding these two inequalities together, applying the initialization condition (Assumption 4) and rearranging terms several times, we have

$$
\begin{aligned}
&\frac{\alpha_x - \mu_x}{2}\|x_t - \widehat{x}\|_2^2 \\
&\geq \langle x_t - \widehat{x}, \nabla_x \mathcal{L}(x_t, y_{t-1}) - \nabla_x \mathcal{L}(\widehat{x}, \widehat{y})\rangle \\
&= \frac{1}{2}\left\langle \begin{pmatrix} x_t - \widehat{x} \\ y_{t-1} - \widehat{y} \end{pmatrix}, \nabla \mathcal{L}(x_t, y_{t-1}) - \nabla \mathcal{L}(\widehat{x}, \widehat{y}) \right\rangle \\
&\qquad + \frac{1}{2}\langle x_t - \widehat{x}, \nabla_x \mathcal{L}(x_t, y_{t-1}) - \nabla_x \mathcal{L}(\widehat{x}, \widehat{y})\rangle - \frac{1}{2}\langle y_{t-1} - \widehat{y}, \nabla_y \mathcal{L}(x_t, y_{t-1}) - \nabla_y \mathcal{L}(\widehat{x}, \widehat{y})\rangle \\
&= \frac{1}{2}\left\langle \begin{pmatrix} x_t - \widehat{x} \\ y_{t-1} - \widehat{y} \end{pmatrix}, \nabla \mathcal{L}(x_t, y_{t-1}) - \nabla \mathcal{L}(\widehat{x}, \widehat{y}) \right\rangle \\
&\qquad + \frac{1}{2}\langle x_t - \widehat{x}, \nabla_x \mathcal{L}(x_t, \widehat{y}) - \nabla_x \mathcal{L}(\widehat{x}, \widehat{y})\rangle - \frac{1}{2}\langle y_{t-1} - \widehat{y}, \nabla_y \mathcal{L}(\widehat{x}, y_{t-1}) - \nabla_y \mathcal{L}(\widehat{x}, \widehat{y})\rangle \\
&\qquad\quad + \frac{1}{2}\Big[\langle x_t - \widehat{x}, \nabla_x \mathcal{L}(x_t, y_{t-1}) - \nabla_x \mathcal{L}(x_t, \widehat{y})\rangle - \langle y_{t-1} - \widehat{y}, \nabla_y \mathcal{L}(x_t, y_{t-1}) - \nabla_y \mathcal{L}(\widehat{x}, y_{t-1})\rangle\Big] \\
&\geq \frac{\alpha_x}{2}\|x_t - \widehat{x}\|_2^2 + \frac{\alpha_y}{2}\|y_{t-1} - \widehat{y}\|_2^2 - \frac{\alpha_x \epsilon_x^2 + \alpha_y \epsilon_y^2}{2} \\
&\qquad + \frac{1}{2}\langle x_t - \widehat{x}, \nabla_x \mathcal{L}(x_t, \widehat{y}) - \nabla_x \mathcal{L}(\widehat{x}, \widehat{y})\rangle - \frac{1}{2}\langle y_{t-1} - \widehat{y}, \nabla_y \mathcal{L}(\widehat{x}, y_{t-1}) - \nabla_y \mathcal{L}(\widehat{x}, \widehat{y})\rangle \\
&\qquad\quad - \frac{1}{2}\left(\frac{1}{2}\mu_x\|x_t - \widehat{x}\|_2^2 + \frac{1}{2}\mu_y\|y_{t-1} - \widehat{y}\|_2^2 + \alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2\right),
\end{aligned}
$$

where the last step holds by applying joint restricted strong convexity (Assumption 1) to the first term, and the cross-product condition (Assumption 3) to the expression in square brackets. (Note that these assumptions can be applied since we have $x_t \in \mathcal{X}_0$ and $y_{t-1} \in \mathcal{Y}_0$). Combining terms and simplifying, multiplying by 2, and using the assumption that $\mu_x \geq 0$ while $\mu_y \leq \alpha_y$, we obtain

$$
0 \geq D^2(x_t; \widehat{x}) - D^2(y_{t-1}; \widehat{y}) + \frac{\alpha_y}{2}\|y_{t-1} - \widehat{y}\|_2^2 - 2\alpha_x\epsilon_x^2 - 2\alpha_y\epsilon_y^2. \tag{28}
$$

Now, by restricted smoothness (Assumption 2) and using the assumption $\alpha_y \leq \beta_y$,

$$
\frac{\alpha_y}{2}\|y_{t-1} - \widehat{y}\|_2^2 \geq \frac{\alpha_y}{2\beta_y} D^2(y_{t-1}; \widehat{y}) - \frac{\alpha_y}{2}\epsilon_y^2.
$$

Returning to (28) and rearranging terms,

$$
D^2(x_t; \widehat{x}) \leq \left(1 - \frac{\alpha_y}{2\beta_y}\right) D^2(y_{t-1}; \widehat{y}) + 3\left(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2\right).
$$

If $D^2(x_t; \widehat{x}) \geq 0$, then by taking a square root on both sides, we obtain

$$
D(x_t; \widehat{x}) \leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{3(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)},
$$

thus proving that the bound (8) holds at time t, while if $D^2(x_t; \widehat{x}) \leq 0$, then $D(x_t; \widehat{x}) = 0$ and so the bound holds trivially. The proof that the analogous bound (9) for the y update step, proceeds similarly.

By applying these bounds recursively, along with the restricted strong convexity and restricted smoothness conditions, we can obtain the result (10) showing linear convergence in the ℓ_2 norm; details are given in Appendix B.1.

7.2 Proof of Theorem 2

This proof is a straightforward combination of the relaxed triangle inequality (Assumption 5) with Theorem 1, the contraction result for the exact alternating minimization algorithm. First, since x_t^{exact} exactly solves the alternating

minimization step, i.e. $\arg\min_{x\in\mathcal{X}_0} \mathcal{L}(x, y_{t-1})$, Theorem 1 proves that

$$D(x_t^{\text{exact}}; \widehat{x}) \leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} D(y_{t-1}; \widehat{y}) + \sqrt{3(\alpha_x \epsilon_x^2 + \alpha_y \epsilon_y^2)}.$$

Next, we use this to bound $D(x_t; \widehat{x})$, using only the assumption that x_t is chosen to be within radius δ_t^x of x_t^{exact}. By the relaxed triangle inequality (Assumption 5),

$$\begin{aligned} D(x_t; \widehat{x}) &\leq D(x_t^{\text{exact}}; \widehat{x}) + \sqrt{\beta_x}\|x_t - x_t^{\text{exact}}\|_2 + \sqrt{\alpha_x}\epsilon_x \\ &\leq \left(\sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{3(\alpha_x \epsilon_x^2 + \alpha_y \epsilon_y^2)}\right) + \sqrt{\beta_x}\delta_t^x + \sqrt{\alpha_x}\epsilon_x \\ &\leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{\beta_x}\delta_t^x + \sqrt{8(\alpha_x \epsilon_x^2 + \alpha_y \epsilon_y^2)}. \end{aligned}$$

This proves the bound (12). The bound (13) on $D(y_t; \widehat{y})$ is proved analogously.

8 Discussion

In this paper, we present a general convergence result for the alternating minimization method and its inexact variants in the presence of nonconvex constraints. We have shown that under standard assumptions on the loss function $\mathcal{L}(x, y)$, these methods can offer linear convergence, as long as certain initialization condition is satisfied. A major tool allowing to handle the nonconvex constraints is the local concavity coefficient, which enables us to bound the concavity arising from the nonconvexity of the sets. The important implication of our result is the computational gain inherent in the alternating methods that relies on the variable with better conditioned, in contrast to the methods that work on the combined variable. This phenomenon has been further demonstrated through numerical experiments.

There are many open questions related to our problem. First, the initialization condition presented in this paper requires the careful choice of initialization point to begin with, but in some applications, a valid initialization procedure may not be available. Characterizing the convergence behavior of alternating minimization without strong initialization assumptions is therefore of practical importance for real applications. Furthermore, as seen in the low-rank factorization approach, it would be interesting to generalize our current result to more general setup, including loss function that may only satisfy convexity up to identifiability issues (such as matrix factorization methods), which would extend our understanding of alternating minimization and descent algorithms to a broader range of problems.

Acknowledgements

R.F.B. was supported by an Alfred P. Sloan Fellowship and by NSF award DMS-1654076.

References

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, pages 1171–1197, 2012.

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

Alfred Auslender. Optimisation: méthodes numériques. 1976.

Rina Foygel Barber and Wooseok Ha. Gradient descent with nonconvex constraints: local concavity determines convergence. *arXiv preprint arXiv:1703.07755*, 2017.

Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM Journal on Optimization*, 25(1):185–209, 2015.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

Christian Grussler, Anders Rantzer, and Pontus Giselsson. Low-rank optimization with convex constraints. *arXiv preprint arXiv:1606.01793*, 2016.

Quanquan Gu, Zhaoran Wang Wang, and Han Liu. Low-rank and sparse structure pursuit via alternating minimization. In *Artificial Intelligence and Statistics*, pages 600–609, 2016.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5 (2):248–264, 1975.

Prateek Jain and Ambuj Tewari. Alternating minimization for regression problems with vector-valued outputs. In *Advances in Neural Information Processing Systems*, pages 1126–1134, 2015.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Yurii Nesterov et al. Gradient methods for minimizing composite objective function, 2007.

James M Ortega and Werner C Rheinboldt. Iterative solution of nonlinear equations in several variables. 1970.

Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Xinyang Yi, Dohyung Park, Yudong Chen, and Constantine Caramanis. Fast algorithms for robust pca via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

Piotr Zwiernik, Caroline Uhler, and Donald Richards. Maximum likelihood estimation for linear gaussian covariance models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

A Gaussian factor model

In this section we provide details for our results on the Gaussian factor model described in Section 5.1. Consider random samples drawn from a non-degenerate multivariate normal distribution, i.e.

$$z_1, \dots, z_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma^\star),$$

where $\text{Cov}(z_i) = \Sigma^\star \succ 0$. In the factor model, we assume an underlying loading matrix $U^\star \in \mathbb{R}^{d\times r}$ representing the latent low-dimensional structure present in the data. Rewriting the model in the generative form,

$$z_i = U^\star w_i + \epsilon_i \text{ for } i = 1, \dots, n,$$

where $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, Y^\star)$ denote the random factors and the observational noise, respectively, and are independent from each other.

In the simplest setting, the covariance matrix of the noise, $Y^\star$, is assume to be proportional to identity matrix in which case the dependence of the random vector z_i is determined entirely by the latent structure $U^\star$. Here we allow more flexibility on the covariance structure and assume that the covariance matrix $Y^\star$ is sparse. Under this assumption, we can compute $\Sigma^\star = U^\star U^{\star\top} + Y^\star$, exhibiting a low-rank + sparse decomposition of the covariance of the data.

We estimate the unknown components $X^\star = U^\star U^{\star\top}$ and $Y^\star$ by minimizing the constrained negative log-likelihood function

$$(\widehat{X}, \widehat{Y}) = \arg\min\{\mathcal{L}(X,Y) : (X,Y) \in \mathcal{X}\times\mathcal{Y}\} \text{ where } \mathcal{L}(X,Y) = \langle S_n, (X+Y)^{-1}\rangle - \log\det(X+Y)^{-1}, \quad (29)$$

for $S_n = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top$, the sample covariance matrix of z_i's. Note that the loss function (29) is highly nonconvex due to the presence of the matrix inverse.

Zwiernik et al. [2016] study the related loss function $\mathcal{L}(\Sigma) = \mathcal{L}(X+Y)$ in the context of a linear Gaussian covariance model. The authors prove that this loss is in fact convex in the region[5] $\{\Sigma \in \mathbb{R}^{d\times d} : 0 \prec \Sigma \prec 2S_n\}$ and furthermore this region contains both the true covariance matrix $\Sigma^\star$ and the maximum likelihood estimator $\widehat{\Sigma}$ with high probability, as long as the sample size is large enough, $n \gtrsim d$. In this regard, our setting can be seen as imposing different structure on the covariance matrix.

In the lemma to follow, we verify analogous results as in Zwiernik et al. [2016], showing that the loss (29) satisfies all the assumptions of Theorem 1 in the local region, ensuring fast convergence of the alternating minimization algorithm. Suppose that the algorithm is initialized at the point (X_0, Y_0) with the corresponding neighborhoods $\mathcal{X}_0 = \mathcal{X} \cap \mathbb{B}_2(X_0, \rho_X), \mathcal{Y}_0 = \mathcal{Y} \cap \mathbb{B}_2(Y_0, \rho_Y)$, where for some $c_0 > 0$ the radii are defined to satisfy

$$\rho_X, \rho_Y \le c_0 \cdot \min\{\sigma_r(X^\star)\kappa^{-3}(\Sigma^\star), \lambda_{\min}(\Sigma^\star)\kappa^{-4}(\Sigma^\star)\},$$

where $\sigma_r(X^\star)$ is the smallest singular value of $X^\star$, $\lambda_{\min}(\Sigma^\star)$ (and $\lambda_{\max}(\Sigma^\star)$ resp.) is the minimum (and maximum resp.) eigenvalue of $\Sigma^\star$, and where $\kappa(\Sigma^\star) = \lambda_{\max}(\Sigma^\star)/\lambda_{\min}(\Sigma^\star)$ is the condition number of $\Sigma^\star$. Assume also that these neighborhoods $\mathcal{X}_0 \times \mathcal{Y}_0$ contain the pair of true matrices $(X^\star, Y^\star)$ and the global optimal $(\widehat{X}, \widehat{Y})$, i.e. $(X^\star, Y^\star), (\widehat{X}, \widehat{Y}) \in \mathcal{X}_0 \times \mathcal{Y}_0$. With this setup, we now establish the following probabilistic guarantee:

Lemma 8. *Suppose that*

$$\sqrt{\frac{d}{n}} \le c_1 \cdot \min\{\sigma_r(X^\star)\lambda_{\min}^{-1}(\Sigma^\star)\kappa^{-4}(\Sigma^\star), \kappa^{-1}(\Sigma^\star)\}. \quad (30)$$

[5]Specifically they show that the Hessian matrix of the loss function $\mathcal{L}(\Sigma)$ is positive semidefinite in the region $\{\Sigma \in \mathbb{R}^{d\times d} : 0 \prec \Sigma \prec 2S_n\}$.

Then, under the previously stated conditions, with probability at least $1-2e^{-d}$*, the steps* $(X_t, Y_t)_{t=1}^{\infty}$ *produced by the alternating minimization* (2) *satisfy*

$$\|(X_t,Y_t)-(\widehat{X},\widehat{Y})\|_{\mathrm{F}} \leq \overbrace{\left(1-c_2\kappa^{-4}(\Sigma^\star)\right)^t}^{\text{linear convergence}}\cdot c_3\min\{\sigma_r(X^\star)\kappa^{-1}(\Sigma^\star),\lambda_{\min}(\Sigma^\star)\kappa^{-2}(\Sigma^\star)\} + \underbrace{c_4\cdot C\left(\|\widehat{Y}-Y^\star\|_{\mathrm{F}}^2+\alpha_{\mathsf{sp}}^2\frac{s}{d}\right)}_{\text{statistical error term}}.$$

Here, $\{c_i>0, i=1,\dots,4\}$ *are universal constants, and* $C>0$ *is defined in Theorem 1.*

This lemma is proved in Appendix C.4.

The discussion following Lemma 6 is also valid in this setting—in particular, the error due to the nonidentifiability of the model now appears as the term $\alpha_{\mathsf{sp}}^2\frac{s}{d}$.

B Additional proofs

B.1 Proof of ℓ_2 convergence bound (10)

Here we give details for the ℓ_2 convergence bound (10) in Theorem 1. We first write $r_x=\sqrt{1-\frac{\alpha_y}{2\beta_y}}$ and $r_y=\sqrt{1-\frac{\alpha_x}{2\beta_x}}$ for simplicity; then (8) and (9) can be rewritten as

$$D(x_t;\widehat{x})\leq r_x D(y_{t-1};\widehat{y})+\sqrt{3(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)},$$

and

$$D(y_t;\widehat{y})\leq r_y D(x_t;\widehat{x})+\sqrt{3(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)}.$$

Then, applying these bounds recursively, we have

$$D(x_t;\widehat{x})\leq r_x(r_xr_y)^{t-1}D(y_0;\widehat{y})+\sqrt{3(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)}\cdot\frac{1+r_x}{1-r_xr_y},$$

and

$$D(y_t;\widehat{y})\leq (r_xr_y)^{t}D(y_0;\widehat{y})+\sqrt{3(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)}\cdot\frac{1+r_y}{1-r_xr_y}.$$

Let $\alpha_{\min}=\min\{\alpha_x,\alpha_y\}, \alpha_{\max}=\max\{\alpha_x,\alpha_y\}$. By joint restricted strong convexity (6),

$$\begin{aligned}\|(x_t,y_t)-(\widehat{x},\widehat{y})\|_2 &= \sqrt{\|x_t-\widehat{x}\|_2^2+\|y_t-\widehat{y}\|_2^2}\\ &\leq \sqrt{\frac{(D^2(x_t;\widehat{x})+D^2(y_t;\widehat{y}))}{\alpha_{\min}}+\frac{2\alpha_{\max}(\epsilon_x^2+\epsilon_y^2)}{\alpha_{\min}}}\leq \frac{D(x_t;\widehat{x})+D(y_t;\widehat{y})}{\sqrt{\alpha_{\min}}}+\frac{\sqrt{2\alpha_{\max}(\epsilon_x^2+\epsilon_y^2)}}{\sqrt{\alpha_{\min}}}\\ &\leq (r_xr_y)^t\cdot D(y_0;\widehat{y})\cdot\frac{(1+r_y^{-1})}{\sqrt{\alpha_{\min}}}+\frac{\sqrt{3(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)}\cdot\left(\frac{2+2r_x}{1-r_xr_y}\right)}{\sqrt{\alpha_{\min}}}+\frac{\sqrt{2\alpha_{\min}(\epsilon_x^2+\epsilon_y^2)}}{\sqrt{\alpha_{\min}}}\end{aligned}$$

and, by definition of r_x and r_y and the fact that $r_xr_y\leq 1$, we see that $r_x, r_y\in[1/\sqrt{2},\sqrt{2}]$. Simplifying,

$$\|(x_t,y_t)-(\widehat{x},\widehat{y})\|_2\leq (r_xr_y)^t\cdot D(y_0;\widehat{y})\cdot\frac{\sqrt{6}}{\sqrt{\alpha_{\min}}}+\frac{\sqrt{3(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)}\cdot\frac{2+2\sqrt{2}}{1-r_xr_y}}{\sqrt{\alpha_{\min}}}+\frac{\sqrt{2\alpha_{\max}(\epsilon_x^2+\epsilon_y^2)}}{\sqrt{\alpha_{\min}}},$$

and by restricted smoothness (7),

$$D^2(y_0;\widehat{y})\leq\beta_y\|y_0-\widehat{y}\|_2^2+\alpha_y\epsilon_y^2.$$

Combining everything, and simplifying, we obtain the overall convergence guarantee (10).

B.2 Proof of Lemma 3

For convenience define

$$r_x = \sqrt{1 - \frac{\alpha_y}{2\beta_y}} + (1+\sqrt{2}) \cdot c_x \sqrt{\frac{\beta_x}{\alpha_x}} \text{ and } r_y = \sqrt{1 - \frac{\alpha_x}{2\beta_x}} + (1+\sqrt{2}) \cdot c_y \sqrt{\frac{\beta_y}{\alpha_y}}.$$

(Comparing to the proof of the ℓ_2 convergence bound (10) for the exact algorithm, given in Appendix B.1, we see that these definitions coincide with the previous ones in the special case that $c_x = c_y = 0$, i.e. when our updates are exact.) Define also $D_0 = \sqrt{\alpha_x}\rho_x + \sqrt{\beta_y}\rho_y$.

We will first show, by induction, that for each $t \geq 1$,

$$\begin{cases} D(x_t; \widehat{x}) \leq r_x \cdot (r_x r_y)^{t-1} \cdot D_0 + \frac{1+r_x}{1-r_x r_y} \cdot C' \max\{\epsilon_x, \epsilon_y\}, \\ D(y_t; \widehat{y}) \leq (r_x r_y)^t \cdot D_0 + \frac{1+r_y}{1-r_x r_y} \cdot C' \max\{\epsilon_x, \epsilon_y\}, \end{cases} \tag{31}$$

where

$$C' = 4\left(1 + c_x\sqrt{\frac{\beta_x}{\alpha_x}} + c_y\sqrt{\frac{\beta_y}{\alpha_y}}\right)\sqrt{\alpha_x + \alpha_y} + C_x\sqrt{\beta_x} + C_y\sqrt{\beta_y}. \tag{32}$$

First we prove the bounds (31) at time $t = 1$. For the x bound,

$$\begin{aligned} D(x_1; \widehat{x}) &\leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_0; \widehat{y}) + \sqrt{\beta_x}\delta_t^x + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)} \text{ by Theorem 2} \\ &\leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_0; \widehat{y}) + \sqrt{\beta_x}\left(c_x\|x_0 - x_1^{\text{exact}}\|_2 + C_x\epsilon_x\right) + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)} \text{ by (14)} \\ &\leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot (\sqrt{\beta_y}\rho_y + \sqrt{\alpha_y}\epsilon_y) + \sqrt{\beta_x}\left(c_x \cdot \rho_x + C_x\epsilon_x\right) + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)}, \end{aligned}$$

where the last step holds since $x_1^{\text{exact}} \in \mathcal{X}_0 \subset \mathbb{B}_2(x_0, \rho_x)$, and $D(y_0; \widehat{y})$ can be bounded by restricted smoothness (Assumption 2). Simplifying,

$$D(x_1; \widehat{x}) \leq r_x D_0 + C' \max\{\epsilon_x, \epsilon_y\},$$

which proves the bound (31) on $D(x_1; \widehat{x})$ at time $t = 1$. Similarly, for the y bound,

$$\begin{aligned} D(y_1; \widehat{y}) &\leq \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot D(x_1; \widehat{x}) + \sqrt{\beta_y}\delta_t^y + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)} \text{ by Theorem 2} \\ &\leq \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot D(x_1; \widehat{x}) + \sqrt{\beta_y}\left(c_y\|y_0 - y_1^{\text{exact}}\|_2 + C_y\epsilon_y\right) + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)} \text{ by (14)} \\ &\leq \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot (r_x D_0 + C' \max\{\epsilon_x, \epsilon_y\}) + \sqrt{\beta_y}\left(c_y\rho_y + C_y\epsilon_y\right) + \sqrt{8(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)} \\ &\leq r_x r_y D_0 + (1 + r_y) \cdot C' \max\{\epsilon_x, \epsilon_y\}, \end{aligned}$$

where for the last step we use the fact that $r_y \geq \sqrt{1 - \frac{\alpha_x}{2\beta_x}}$ by definition.

Next, take any $t \geq 2$. For the x bound, we first calculate

$$\begin{aligned} \|x_{t-1} - x_t^{\text{exact}}\|_2 &\leq \|x_{t-1} - \widehat{x}\|_2 + \|x_t^{\text{exact}} - \widehat{x}\|_2 \\ &\leq \frac{1}{\sqrt{\alpha_x}}\left(D(x_{t-1}; \widehat{x}) + D(x_t^{\text{exact}}; \widehat{x})\right) + \frac{2\sqrt{\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2}}{\sqrt{\alpha_x}} \text{ by joint restricted strong convexity (6)} \\ &\leq \frac{1}{\sqrt{\alpha_x}}\left(D(x_{t-1}; \widehat{x}) + D(y_{t-1}; \widehat{y}) + \sqrt{3(\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2)}\right) + \frac{2\sqrt{\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2}}{\sqrt{\alpha_x}} \text{ by Theorem 1} \\ &\leq \frac{1}{\sqrt{\alpha_x}}\left(D(x_{t-1}; \widehat{x}) + D(y_{t-1}; \widehat{y}) + 4\sqrt{\alpha_x\epsilon_x^2 + \alpha_y\epsilon_y^2}\right). \end{aligned}$$

We now bound $D(x_t; \widehat{x})$:

$$\begin{aligned}
D(x_t;\widehat{x}) &\le \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot D(y_{t-1};\widehat{y}) + \sqrt{\beta_x}\delta_t^x + \sqrt{8(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)} \text{ by Theorem 2} \\
&\le \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot D(y_{t-1};\widehat{y}) + \sqrt{\beta_x}\left(c_x\|x_{t-1}-x_t^{\text{exact}}\|_2 + C_x\epsilon_x\right) + \sqrt{8(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)} \text{ by (14)} \\
&\le \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot D(y_{t-1};\widehat{y}) + \sqrt{\beta_x}\left(c_x\left[\frac{1}{\sqrt{\alpha_x}}\left(D(x_{t-1};\widehat{x})+D(y_{t-1};\widehat{y})+4\sqrt{\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2}\right)\right] + C_x\epsilon_x\right) \\
&\qquad + \sqrt{8(\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2)} \\
&\le \left(\sqrt{1-\frac{\alpha_y}{2\beta_y}} + c_x\sqrt{\frac{\beta_x}{\alpha_x}}\right)D(y_{t-1};\widehat{y}) + c_x\sqrt{\frac{\beta_x}{\alpha_x}}D(x_{t-1};\widehat{x}) + C'\max\{\epsilon_x,\epsilon_y\},
\end{aligned}$$

where C' is defined as in (32) above. Assuming by induction that the bounds (31) hold with $t-1$ in place of t, we obtain

$$\begin{aligned}
D(x_t;\widehat{x}) \le \left(\sqrt{1-\frac{\alpha_y}{2\beta_y}} + c_x\sqrt{\frac{\beta_x}{\alpha_x}}\right)\cdot\left((r_xr_y)^{t-1}D_0 + \frac{1+r_y}{1-r_xr_y}\cdot C'\max\{\epsilon_x,\epsilon_y\}\right) \\
+ c_x\sqrt{\frac{\beta_x}{\alpha_x}}\left(r_x(r_xr_y)^{t-2}D_0 + \frac{1+r_x}{1-r_xr_y}\cdot C'\max\{\epsilon_x,\epsilon_y\}\right) + C'\max\{\epsilon_x,\epsilon_y\}.
\end{aligned}$$

Since $r_y \ge 1/\sqrt{2}$ we can rewrite this as

$$\begin{aligned}
D(x_t;\widehat{x}) \le \left(\sqrt{1-\frac{\alpha_y}{2\beta_y}} + c_x\sqrt{\frac{\beta_x}{\alpha_x}}\right)\cdot\left((r_xr_y)^{t-1}D_0 + \frac{1+r_y}{1-r_xr_y}\cdot C'\max\{\epsilon_x,\epsilon_y\}\right) \\
+ c_x\sqrt{\frac{\beta_x}{\alpha_x}}\left(\sqrt{2}(r_xr_y)^{t-1}D_0 + \frac{1+r_x}{1-r_xr_y}\cdot C'\max\{\epsilon_x,\epsilon_y\}\right) + C'\max\{\epsilon_x,\epsilon_y\}.
\end{aligned}$$

Plugging in the definition of r_x, then,

$$D(x_t;\widehat{x}) \le r_x\cdot(r_xr_y)^{t-1}\cdot D_0 + \left[c_x\sqrt{\frac{\beta_x}{\alpha_x}}\cdot\frac{1+r_x}{1-r_xr_y} + \left(\sqrt{1-\frac{\alpha_y}{2\beta_y}} + c_x\sqrt{\frac{\beta_x}{\alpha_x}}\right)\cdot\frac{1+r_y}{1-r_xr_y} + 1\right]\cdot C'\max\{\epsilon_x,\epsilon_y\}.$$

Plugging in the definition of r_x, and the assumption that $r_xr_y < 1$, we see that the term in square brackets is bounded by $\frac{1+r_x}{1-r_xr_y}$, which proves the desired bound on $D(x_t;\widehat{x})$ as in (31), as desired. The bound on $D(y_t;\widehat{y})$ is proved similarly.

Finally, by joint restricted strong convexity (6), we know that

$$\|x_t-\widehat{x}\|_2 \le \frac{D(x_t;\widehat{x})}{\sqrt{\alpha_x}} + \frac{\sqrt{\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2}}{\sqrt{\alpha_x}} \text{ and } \|y_t-\widehat{y}\|_2 \le \frac{D(y_t;\widehat{y})}{\sqrt{\alpha_y}} + \frac{\sqrt{\alpha_x\epsilon_x^2+\alpha_y\epsilon_y^2}}{\sqrt{\alpha_y}}.$$

Combining this with the bounds (31) proves the result.

B.3 Proof of Lemma 2

Take any $x, x' \in \mathcal{X}_0 \subset \mathcal{X}$ and take $t \in [0,1]$. By the curvature condition (Definition 1) on the larger set $\mathcal{X}$, we can find a family of points $\widetilde{x}_t \in \mathcal{X}$, indexed by $t \in [0,1]$, such that $\delta_t \to 0$, where

$$\|\big((1-t)x+tx'\big)-\widetilde{x}_t\|_x \le t\cdot\Big[\gamma_x(\mathcal{X})\cdot\|x-x'\|_2^2+\delta_t\Big].$$

Next, we show that $\widetilde{x}_t \in \mathcal{X}_0$ for sufficiently small $t > 0$. Recall that $\mathcal{X}_0 = \mathcal{X} \cap \mathbb{B}_2(x_0, \rho_x)$, and therefore we only need to check that $\|\widetilde{x}_t - x_0\|_2 \le \rho_x$. Since $\|\cdot\|_2 \le \|\cdot\|_x$ by assumption, we have

$$\begin{aligned}\|\widetilde{x}_t - x_0\|_2 &\le \|\widetilde{x}_t - \big((1-t)x + tx'\big)\|_2 + \|\big((1-t)x + tx'\big) - x_0\|_2 \\ &\le \|\widetilde{x}_t - \big((1-t)x + tx'\big)\|_x + \|\big((1-t)x + tx'\big) - x_0\|_2 \\ &\le t \cdot \Big[\gamma_x(\mathcal{X}) \cdot \|x - x'\|_2^2 + \delta_t\Big] + \|\big((1-t)x + tx'\big) - x_0\|_2.\end{aligned}$$

Next, a simple calculation shows that

$$\begin{aligned}\|\big((1-t)x + tx'\big) - x_0\|_2 &= \|(1-t)\cdot(x - x_0) + t\cdot(x' - x_0)\|_2 \\ &= \sqrt{(1-t)\|x - x_0\|_2^2 + t\cdot\|x' - x_0\|_2^2 - t(1-t)\|x - x'\|_2^2},\end{aligned}$$

and since $x, x' \in \mathcal{X}_0 \subset \mathbb{B}_2(x_0, \rho_x)$, we obtain

$$\|\big((1-t)x + tx'\big) - x_0\|_2 \le \sqrt{\rho_x^2 - t(1-t)\|x - x'\|_2^2} \le \rho_x - \frac{t(1-t)\|x - x'\|_2^2}{2\rho_x}.$$

Combining everything,

$$\|\widetilde{x}_t - x_0\|_2 \le \rho_x - t\|x - x'\|_2^2 \cdot \left[\frac{1}{2\rho_x} - \gamma_x(\mathcal{X}) - \frac{t}{2\rho_x} - \frac{\delta_t}{\|x - x'\|_2^2}\right].$$

Since $\gamma_x(\mathcal{X}) < \frac{1}{2\rho_x}$ by assumption, and $\delta_t \to 0$, we can find some $t_0 > 0$ such that, for all $t \in [0, t_0]$,

$$\frac{t}{2\rho_x} + \frac{\delta_t}{\|x - x'\|_2^2} \le \frac{1}{2\rho_x} - \gamma_x(\mathcal{X}).$$

Therefore, $\widetilde{x}_t \in \mathcal{X}_0$ for all $t \in [0, t_0]$, and so

$$\frac{\min_{x'' \in \mathcal{X}_0}\|\big((1-t)x + tx'\big) - x''\|_x}{t} \le \frac{\|\big((1-t)x + tx'\big) - \widetilde{x}_t\|_x}{t} \le \gamma_x(\mathcal{X}) \cdot \|x - x'\|_2^2 + \delta_t$$

for all $t \in [0, t_0]$. This proves that

$$\lim_{t \to 0} \frac{\min_{x'' \in \mathcal{X}_0}\|\big((1-t)x + tx'\big) - x''\|}{t} \le \gamma_x(\mathcal{X}) \cdot \|x - x'\|_2^2.$$

Since $x, x' \in \mathcal{X}_0$ were chosen arbitrarily, then, we have shown that

$$\gamma_x(\mathcal{X}_0) \le \gamma_x(\mathcal{X}).$$

C Proofs for examples (Section 5)

In this section, we provide the proofs of the lemmas displayed in Section 5. Throughout the section, given any function $\mathsf{f}(A)$ over a matrix variable $A \in \mathbb{R}^{m \times n}$, we write $\nabla^2_{AA}\mathsf{f}(A) \in \mathbb{R}^{mn \times mn}$ to refer to the second derivative of $\mathsf{f}(A)$ with respect to the *vectorized* variable $\mathrm{vec}\,(A) \in \mathbb{R}^{mn}$.

C.1 Proof of Lemma 5

We first reparametrize the variable $X \in \mathcal{X}$ by $X = \mathsf{g}(U) = UU^\top$ with the corresponding convex set

$$\mathcal{U} = \left\{U \in \mathbb{R}^{d \times r} : \max_{i=1,\dots,d} \|U_{i*}\|_2 \le \sqrt{\frac{\alpha_{\mathsf{sp}}}{d}}\right\},$$

where U_{i*} represents ith row of U. Note that under such reparametrization, we trivially have $\mathcal{X} = \mathrm{g}(\mathcal{U})$. Now take $X, X' \in \mathcal{X}$ with $X = UU^\top, X' = U'U'^\top$. For $t > 0$, let $X_t = (1-t)X + tX'$ and $U_t = (1-t)U + tU'$. Then, by Taylor's theorem,

$$\begin{aligned}
&X_t - \mathrm{g}(U_t) = (1-t)\mathrm{g}(U) + t\mathrm{g}(U') - \mathrm{g}(U_t) \\
&= (1-t)(\mathrm{g}(U) - \mathrm{g}(U_t)) + t(\mathrm{g}(U') - \mathrm{g}(U_t)) \\
&= (1-t)\left[\nabla \mathrm{g}(U_t)(U - U_t) + \frac{1}{2}\nabla^2 \mathrm{g}(U_*)(U - U_t, U - U_t)\right] + t\left[\nabla \mathrm{g}(U_t)(U' - U_t) + \frac{1}{2}\nabla^2 \mathrm{g}(U_\#)(U' - U_t, U' - U_t)\right] \\
&= (1-t)\left[t\nabla \mathrm{g}(U_t)(U - U') + \frac{t^2}{2}\nabla^2 \mathrm{g}(U_*)(U - U', U - U')\right] \\
&\qquad + t\left[(1-t)\nabla \mathrm{g}(U_t)(U' - U) + \frac{(1-t)^2}{2}\nabla^2 \mathrm{g}(U_\#)(U' - U, U' - U)\right]. \qquad (33)
\end{aligned}$$

Meanwhile, some calculation yields that for $i, j = 1, \ldots, d$,

$$\nabla^2 \mathrm{g}_{ij}(U) = (e_i e_j^\top \otimes \mathbf{I}_r + e_j e_i^\top \otimes \mathbf{I}_r) \in \mathbb{R}^{dr \times dr},$$

where $e_i \in \mathbb{R}^d$ denotes the ith standard basis vector. Hence, we have

$$\nabla^2 \mathrm{g}(U_*)(U - U', U - U') = \nabla^2 \mathrm{g}(U_\#)(U - U', U - U') = 2(U - U')(U - U')^\top.$$

Combining with (33),

$$\min_{X'' \in \mathcal{X}} \|X'' - X_t\|_{\text{nuc}} \le \|\mathrm{g}(U_t) - X_t\|_{\text{nuc}} = t(1-t)\|(U - U')(U - U')^\top\|_{\text{nuc}} = t(1-t)\|U - U'\|_{\text{F}}^2,$$

so dividing out by t and taking $t \to 0$,

$$\limsup_{t \to 0} \frac{\min_{X'' \in \mathcal{X}} \|X'' - X_t\|_x}{t} \le \|U - U'\|_{\text{F}}^2 \le \frac{5}{4\sigma_r(X)}\|X - X'\|_{\text{F}}^2,$$

where the last inequality follows from Tu et al. [2015, Lemma 5.4]. This completes the proof of the lemma.

C.2 Proof of Lemma 6

Recalling the constrained least squares problem (23) for the robust PCA problem, we verify that under the conditions of Lemma 6, the loss function satisfies the assumptions of Theorem 1, i.e. Assumptions 1, 2, 3, and 4, with parameters specified below. Before diving into the proof, observe that for all $Y \in \mathcal{Y}_0$, we have

$$\|Y - \widehat{Y}\|_1 \le \|Y - Y^\star\|_1 + \|\widehat{Y} - Y^\star\|_1 \le 2\sqrt{sd}\|Y - \widehat{Y}\|_{\text{F}} + 4\sqrt{sd}\|\widehat{Y} - Y^\star\|_{\text{F}}, \qquad (34)$$

where the last step holds thanks to the triangle inequality and the fact that $Y^\star$ is sd-sparse by our assumption.

We now establish the joint restricted strong convexity, restricted smoothness, and initialization conditions, for the least squares loss $\mathcal{L}(X, Y) = \frac{1}{2}\|Z - \mathcal{A}(X + Y)\|_{\text{F}}^2$. Note that the cross-product condition (Assumption 3) trivially holds with $\mu_x = \mu_y = 0$, since the Hessian $\nabla^2_{XY}\mathcal{L}(X, Y)$ is constant over all (X, Y). We also use the shorthand $\sigma_r = \sigma_r(X^\star)$ to denote the smallest singular value of $X^\star$.

Joint restricted strong convexity Let $X \in \mathcal{X}_0$ and $Y \in \mathcal{Y}_0$. Invoking the restricted eigenvalue property (Assumption 6),

$$\begin{aligned}
\left\langle \begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix}, \nabla\mathcal{L}(X, Y) - \nabla\mathcal{L}(\widehat{X}, \widehat{Y}) \right\rangle &= \|\mathcal{A}(X - \widehat{X} + Y - \widehat{Y})\|_{\text{F}}^2 \\
&\ge \alpha_A(\|X - \widehat{X}\|_{\text{F}}^2 + \|Y - \widehat{Y}\|_{\text{F}}^2) - \tau\left(\frac{\log d}{n^2}\|Y - \widehat{Y}\|_1^2 + \sqrt{\frac{d^2 \log d}{n^2}}\|X - \widehat{X}\|_\infty \|Y - \widehat{Y}\|_1\right).
\end{aligned}$$

Applying the inequality (34) and the spikiness constraint, the terms in the parenthesis of the above equation can be bounded by

$$\begin{aligned}
&\frac{\log d}{n^2}\|Y-\widehat{Y}\|_1^2+\sqrt{\frac{d^2\log d}{n^2}}\|X-\widehat{X}\|_\infty\|Y-\widehat{Y}\|_1 \\
&\qquad\le \frac{4sd\log d}{n^2}\left(\|Y-\widehat{Y}\|_{\mathrm{F}}+2\|\widehat{Y}-Y^\star\|_{\mathrm{F}}\right)^2+4\alpha_{\mathsf{sp}}\sqrt{\frac{sd\log d}{n^2}}\left(\|Y-\widehat{Y}\|_{\mathrm{F}}+2\|\widehat{Y}-Y^\star\|_{\mathrm{F}}\right) \\
&\le\left(\frac{8sd\log d}{n^2}\|Y-\widehat{Y}\|_{\mathrm{F}}^2+\frac{32sd\log d}{n^2}\|\widehat{Y}-Y^\star\|_{\mathrm{F}}^2\right)+\left(\frac{\alpha_A}{4}\|Y-\widehat{Y}\|_{\mathrm{F}}^2+\alpha_A\|\widehat{Y}-Y^\star\|_{\mathrm{F}}^2+\frac{32\alpha_{\mathsf{sp}}^2}{\alpha_A}\frac{sd\log d}{n^2}\right),
\end{aligned}$$

where the second step uses the identity $ab\le\frac{ca^2}{2}+\frac{b^2}{2c}$ for any $c>0$. Combining the pieces and using the fact that $\frac{32\tau sd\log d}{n^2}\le\alpha_A$, then

$$\left\langle\begin{pmatrix}X-\widehat{X}\\ Y-\widehat{Y}\end{pmatrix},\nabla\mathcal{L}(X,Y)-\nabla\mathcal{L}(\widehat{X},\widehat{Y})\right\rangle\ge\alpha_A\|X-\widehat{X}\|_{\mathrm{F}}^2+\frac{\alpha_A}{2}\left(\|Y-\widehat{Y}\|_{\mathrm{F}}^2-4\|\widehat{Y}-Y^\star\|_{\mathrm{F}}^2-\frac{64\alpha_{\mathsf{sp}}^2}{\alpha_A^2}\frac{sd\log d}{n^2}\right).$$

Restricted smoothness A similar calculation shows that the marginal restricted smoothness condition holds, that is, by the restricted eigenvalue property (Assumption 6),

$$\langle X-\widehat{X},\nabla_X\mathcal{L}(X,\widehat{Y})-\nabla_X\mathcal{L}(\widehat{X},\widehat{Y})\rangle=\|\mathcal{A}(X-\widehat{X})\|_{\mathrm{F}}^2\le\beta_A\|X-\widehat{X}\|_{\mathrm{F}}^2,$$

and furthermore

$$\begin{aligned}
\langle Y-\widehat{Y},\nabla_Y\mathcal{L}(\widehat{X},Y)-\nabla_Y\mathcal{L}(\widehat{X},\widehat{Y})\rangle&=\|\mathcal{A}(Y-\widehat{Y})\|_{\mathrm{F}}^2 \\
&\le\frac{3\beta_A}{2}\|Y-\widehat{Y}\|_{\mathrm{F}}^2+\frac{\alpha_A}{2}\left(4\|\widehat{Y}-Y^\star\|_{\mathrm{F}}^2+\frac{64\alpha_{\mathsf{sp}}^2}{\alpha_A^2}\frac{sd\log d}{n^2}\right),
\end{aligned}$$

proving the RSM condition for the loss $\mathcal{L}(X,Y)$.

Initialization condition Since $\mathcal{Y}$ is convex, the initialization condition is trivial for the set $\mathcal{Y}_0$. For $\mathcal{X}$, we first bound $\|\nabla_X\mathcal{L}(X,Y)\|_{\mathrm{op}}$ for all $X\in\mathcal{X}_0$ and $Y\in\mathcal{Y}_0$. Given the observational model $Z=\mathcal{A}(X^\star+Y^\star)+W$, we have the decomposition of $\|\nabla_X\mathcal{L}(X,Y)\|_{\mathrm{op}}$ into the sums

$$\|\nabla_X\mathcal{L}(X,Y)\|_{\mathrm{op}}\le\underbrace{\|\mathcal{A}^*\mathcal{A}(X-X^\star)\|_{\mathrm{op}}}_{\text{(Term 1)}}+\underbrace{\|\mathcal{A}^*\mathcal{A}(Y-Y^\star)\|_{\mathrm{op}}}_{\text{(Term 2)}}+\|\mathcal{A}^*(W)\|_{\mathrm{op}}.$$

Note that $\|\mathcal{A}^*\mathcal{A}(X-X^\star)\|_{\mathrm{op}}=\langle\mathcal{A}(X'),\mathcal{A}(X-X^\star)\rangle$ for some X' with $\mathrm{rank}(X')=1$ and $\|X'\|_{\mathrm{F}}\le1$. By the restricted eigenvalue condition (Assumption 6), then, $\|\mathcal{A}(X')\|_{\mathrm{F}}^2\le\beta_A$, and moreover, $\|\mathcal{A}(X-X^\star)\|_{\mathrm{F}}^2\le\beta_A\|X-X^\star\|_{\mathrm{F}}^2$. So, we have

$$\text{(Term 1)}=\|\mathcal{A}(X')\|_{\mathrm{F}}\|\mathcal{A}(X-X^\star)\|_{\mathrm{F}}\le\beta_A\|X-X^\star\|_{\mathrm{F}}\le2\beta_A\rho_X,$$

where the last inequality uses $X,X^\star\in\mathbb{B}_2(X_0,\rho_X)$. Again, by Assumption 6, we can bound $\|\mathcal{A}(Y-Y^\star)\|_{\mathrm{F}}^2\le\beta_A\|Y-Y^\star\|_{\mathrm{F}}^2+\frac{4\tau sd\log d}{n^2}\|Y-Y^\star\|_{\mathrm{F}}^2\le\frac{9\beta_A}{8}\|Y-Y^\star\|_{\mathrm{F}}^2$, and so for some X'' with $\mathrm{rank}(X'')=1$ and $\|X''\|_{\mathrm{F}}\le1$,

$$\text{(Term 2)}=\langle\mathcal{A}(X''),\mathcal{A}(Y-Y^\star)\rangle\le\frac{3\sqrt{2}}{4}\beta_A\|Y-Y^\star\|_{\mathrm{F}}\le3\beta_A\rho_Y.$$

Putting these bounds together, we have $\|\nabla_X\mathcal{L}(X,Y)\|_{\mathrm{op}}\le3\beta_A(\rho_X+\rho_Y)+\|\mathcal{A}^*(W)\|_{\mathrm{op}}$. Now, by Lemma 5, we know $\gamma_X(\mathcal{X})\le\frac{5}{4\sigma_r(X)}$, and so

$$\max_{X\in\mathcal{X}_0}\gamma_X(\mathcal{X})\le\frac{5}{4\sigma_r-8\rho_X}\le\frac{5}{2\sigma_r},\tag{35}$$

where the first inequality holds due to Weyl's inequality, while the second inequality follows from $\rho_X \leq \frac{1}{4}\sigma_r$. Recalling $\rho_X, \rho_Y \leq c_0 \cdot \sigma_r \kappa^{-1}(\mathcal{A})$ for some sufficiently small $c_0 > 0$, this implies that the conditions of Lemma 2 hold, i.e. $\rho_X < \frac{1}{2\max_{X\in\mathcal{X}_0}\gamma_X(\mathcal{X})}$, and in particular, we have $\gamma(\mathcal{X}_0) \leq \frac{5}{2\sigma_r}$. Now combining all the pieces, then,

$$2\gamma(\mathcal{X}_0)\cdot\left(\|\nabla_X\mathcal{L}(\widehat{X},\widehat{Y})\|_{\text{op}} + \max_{Y\in\mathcal{Y}_0}\|\nabla_Y\mathcal{L}(X_Y,Y)\|_{\text{op}}\right) \leq 4\gamma(\mathcal{X}_0)\cdot \max_{X\in\mathcal{X}_0,Y\in\mathcal{Y}_0}\|\nabla_X\mathcal{L}(X,Y)\|_{\text{op}}$$
$$\leq \frac{10}{\sigma_r}\cdot(3\beta_A\rho_X + 3\beta_A\rho_Y + \|\mathcal{A}^*(W)\|_{\text{op}}) \leq \alpha_A,$$

where we use $\|\mathcal{A}^*(W)\|_{\text{op}} \leq \sigma_r \cdot \frac{\alpha_A}{30}$ in the last step. This establishes the initialization condition.

Now by specializing to the robust PCA problem (23), Theorem 1 immediately yields the result of Lemma 6. This completes the proof of Lemma 6.

C.3 Proof of Lemma 7

Recall the loss function is given by the negative log-likelihood function with respect to (X, Θ), i.e.

$$\mathcal{L}(X,\Theta) = -\log\det(\Theta) + \frac{1}{n}\sum_{i=1}^{n}(z_i - X\phi_i)^\top\Theta(z_i - X\phi_i).$$

Simple algebra then shows that the gradients are given by

$$\nabla_X\mathcal{L}(X,\Theta) = \frac{2}{n}\sum_{i=1}^{n}\Theta(X\phi_i - z_i)\phi_i^\top \text{ and } \nabla_\Theta\mathcal{L}(X,\Theta) = -\Theta^{-1} + \frac{1}{n}\sum_{i=1}^{n}(z_i - X\phi_i)(z_i - X\phi_i)^\top,$$

and the Hessian operators take the following form:

$$\langle\Delta_X, \nabla^2_{XX}\mathcal{L}(X,\Theta)\Delta_X\rangle = \frac{2}{n}\sum_{i=1}^{n}\phi_i^\top\Delta_X^\top\Theta\Delta_X\phi_i,$$
$$\langle\Delta_X, \nabla^2_{X\Theta}\mathcal{L}(X,\Theta)\Delta_\Theta\rangle = \frac{2}{n}\sum_{i=1}^{n}\phi_i^\top\Delta_X^\top\Delta_\Theta(X\phi_i - z_i),$$
$$\langle\Delta_\Theta, \nabla^2_{\Theta\Theta}\mathcal{L}(X,\Theta)\Delta_\Theta\rangle = \text{vec}\,(\Delta_\Theta)^\top\left(\Theta^{-1}\otimes\Theta^{-1}\right)\text{vec}\,(\Delta_\Theta).$$

Throughout we use the shorthand notation $\sigma_r = \sigma_r(X^\star)$. Recall that the radii are chosen to satisfy $\rho_X \leq c_0 \cdot \sigma_r\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)$ and $\rho_\Theta \leq c_0\cdot\lambda_{\min}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)$ for some small $c_0 > 0$. Then, according to Weyl's inequality, for any $\Theta \in \mathcal{Q}_0$, its minimum and maximum eigenvalues are bounded by

$$\frac{\lambda_{\min}(\Theta^\star)}{2} \leq \lambda_{\min}(\Theta) \leq \lambda_{\max}(\Theta) \leq \frac{3\lambda_{\max}(\Theta^\star)}{2}, \tag{36}$$

since we have that $\|\Theta - \Theta^\star\|_{\text{F}} \leq 2\rho_\Theta$ and $\rho_\Theta \leq \frac{\lambda_{\min}(\Theta^\star)}{4}$.

We will use the following two concentration results: first, following Negahban and Wainwright [2011, Lemma 2], with probability at least $1 - 4\exp(-n/2)$, we have the bound of the form:

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i\phi_i^\top\right) \geq \frac{\lambda_{\min}(\Sigma_\phi)}{9} \text{ and } \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i\phi_i^\top\right) \leq 9\lambda_{\max}(\Sigma_\phi). \tag{37}$$

Next, letting $\widetilde{\epsilon}_i \overset{\text{iid}}{\sim} \text{N}(0, \mathbf{I}_m)$, it has been shown in Negahban and Wainwright [2011, Lemma 3] that for some $c, c' > 0$, with probability at least $1 - c\exp(-c'(m+d))$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\epsilon}_i\phi_i^\top\right\|_{\text{op}} \leq 5\sqrt{\lambda_{\max}(\Sigma_\phi)}\sqrt{\frac{m+d}{n}}. \tag{38}$$

Now, we turn to verifying Lemma 7:

Joint restricted strong convexity Take $X \in \mathcal{X}_0$ and $\Theta \in \mathcal{Q}_0$. By Taylor's theorem, we have

$$\left\langle \begin{pmatrix} X - \widehat{X} \\ \Theta - \widehat{\Theta} \end{pmatrix}, \nabla \mathcal{L}(X, \Theta) - \nabla \mathcal{L}(\widehat{X}, \widehat{\Theta}) \right\rangle = \begin{pmatrix} X - \widehat{X} \\ \Theta - \widehat{\Theta} \end{pmatrix}^\top \nabla^2 \mathcal{L}(X(t), \Theta(t)) \begin{pmatrix} X - \widehat{X} \\ \Theta - \widehat{\Theta} \end{pmatrix},$$

where $X(t) = (1-t)X + t\widehat{X}$ and $\Theta(t) = (1-t)\Theta + t\widehat{\Theta}$ for some $t \in (0,1)$. Using the expression for the Hessian operator and substituting our observational model $z_i = X^\star \phi_i + \epsilon$, we have the following decomposition:

$$\begin{pmatrix} X - \widehat{X} \\ \Theta - \widehat{\Theta} \end{pmatrix}^\top \nabla^2 \mathcal{L}(X(t), \Theta(t)) \begin{pmatrix} X - \widehat{X} \\ \Theta - \widehat{\Theta} \end{pmatrix} = \underbrace{\frac{2}{n} \sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top \Theta(t) (X - \widehat{X}) \phi_i}_{\text{(Term 1)}}$$
$$+ \underbrace{\frac{2}{n} \sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top (\Theta - \widehat{\Theta}) (X(t) - X^\star) \phi_i}_{\text{(Term 2)}} - \underbrace{\frac{2}{n} \sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top (\Theta - \widehat{\Theta}) \cdot \epsilon_i}_{\text{(Term 3)}}$$
$$+ \underbrace{\operatorname{vec}\left(\Theta - \widehat{\Theta}\right)^\top \left(\Theta(t)^{-1} \otimes \Theta(t)^{-1}\right) \operatorname{vec}\left(\Theta - \widehat{\Theta}\right)}_{\text{(Term 4)}}.$$

For (Term 1), we can lower bound it as

$$\text{(Term 1)} \geq 2\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n \phi_i \phi_i^\top\right) \cdot \lambda_{\min}(\Theta(t)) \|X - \widehat{X}\|_{\mathrm{F}}^2 \geq \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{9} \|X - \widehat{X}\|_{\mathrm{F}}^2,$$

where the second step uses the inequalities (36) and (37). For (Term 2), we further decompose it into the sum

$$\text{(Term 2)} = (1-t) \cdot \frac{2}{n} \sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top (\Theta - \widehat{\Theta}) (X - \widehat{X}) \phi_i + \frac{2}{n} \sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top (\Theta - \widehat{\Theta}) (\widehat{X} - X^\star) \phi_i,$$

then the first term is bounded by

$$4\rho_\Theta \cdot \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n \phi_i \phi_i^\top\right) \|X - \widehat{X}\|_{\mathrm{F}}^2 \leq \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{54} \|X - \widehat{X}\|_{\mathrm{F}}^2,$$

where the inequality uses the bound on the radius ρ_Θ (by choosing $c_0 \leq \frac{1}{36 \cdot 54}$) and the inequality (37). Meanwhile, we can bound the second part of (Term 2) as

$$\frac{2}{n} \sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top (\Theta - \widehat{\Theta}) (\widehat{X} - X^\star) \phi_i \leq \frac{4\rho_\Theta}{n} \sum_{i=1}^n \|(X - \widehat{X})\phi_i\|_2 \|(\widehat{X} - X^\star)\phi_i\|_2$$
$$\leq \frac{2\rho_\Theta}{n} \sum_{i=1}^n \|(X - \widehat{X})\phi_i\|_2^2 + \frac{2\rho_\Theta}{n} \sum_{i=1}^n \|(\widehat{X} - X^\star)\phi_i\|_2^2 \leq \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{108} \|X - \widehat{X}\|_{\mathrm{F}}^2 + \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{108} \|\widehat{X} - X^\star\|_{\mathrm{F}}^2.$$

Combining the two then yields

$$\text{(Term 2)} \leq \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{36} \|X - \widehat{X}\|_{\mathrm{F}}^2 + \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{108} \|\widehat{X} - X^\star\|_{\mathrm{F}}^2.$$

Next, using the inequality $\langle a, b \rangle \leq \|a\|_{\mathrm{nuc}} \|b\|_{\mathrm{op}}$, we find that

$$\text{(Term 3)} \leq 2\|X - \widehat{X}\|_{\mathrm{nuc}} \left\| \frac{1}{n} \sum_{i=1}^n (\Theta - \widehat{\Theta}) \epsilon_i \phi_i^\top \right\|_{\mathrm{op}} \leq \frac{\rho_\Theta}{\sqrt{\lambda_{\min}(\Theta^\star)}} \cdot 2\sqrt{2r} \|X - \widehat{X}\|_{\mathrm{F}} \left\| \frac{1}{n} \sum_{i=1}^n \widetilde{\epsilon}_i \phi_i^\top \right\|_{\mathrm{op}} \leq$$
$$\frac{\rho_\Theta \sqrt{\lambda_{\max}(\Sigma_\phi)}}{\sqrt{\lambda_{\min}(\Theta^\star)}} \cdot 10\sqrt{2r} \|X - \widehat{X}\|_{\mathrm{F}} \sqrt{\frac{m+d}{n}},$$

where the second step follows since $X - \widehat{X}$ is of rank $2r$ and $\epsilon_i = (\Theta^\star)^{-1/2} \cdot \widetilde{\epsilon}_i$ for $\widetilde{\epsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_m)$, and the next step uses the concentration bound (38). Using the identity $ab \le \frac{ca^2}{2} + \frac{b^2}{2c}$ and the bound on ρ_Θ, then

$$\text{(Term 3)} \le \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{36}\|X - \widehat{X}\|_{\mathrm{F}}^2 + \frac{25}{13122}\frac{r(m+d)}{n}\frac{\lambda_{\min}(\Sigma_\phi)}{\lambda_{\max}(\Sigma_\phi)}.$$

Lastly, by (36), the minimum eigenvalue of $\Theta(t)^{-1}$ is lower bounded by $\frac{2}{3\lambda_{\max}(\Theta^\star)}$, so it follows that

$$\text{(Term 4)} \ge \frac{4}{9\lambda_{\max}^2(\Theta^\star)}\|\Theta - \widehat{\Theta}\|_{\mathrm{F}}^2.$$

Putting all the bounds together, we have

$$\begin{aligned}\langle \begin{pmatrix} X - \widehat{X} \\ \Theta - \widehat{\Theta} \end{pmatrix}, \nabla\mathcal{L}(X,\Theta) - \nabla\mathcal{L}(\widehat{X},\widehat{\Theta})\rangle \ge \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{18}\Big(\|X - \widehat{X}\|_{\mathrm{F}}^2 - \frac{1}{6}\|\widehat{X} - X^\star\|_{\mathrm{F}}^2 \\ - \frac{25}{729}\frac{r(m+d)}{n}\frac{1}{\lambda_{\min}(\Theta^\star)\lambda_{\max}(\Sigma_\phi)}\Big) + \frac{4}{9\lambda_{\max}^2(\Theta^\star)}\|\Theta - \widehat{\Theta}\|_{\mathrm{F}}^2.\end{aligned}$$

Restricted smoothness To establish the restricted smoothness for the X variable, we apply the inequalities (36) and (37) to obtain

$$\langle X - \widehat{X}, \nabla_X\mathcal{L}(X,\widehat{\Theta}) - \nabla_X\mathcal{L}(\widehat{X},\widehat{\Theta})\rangle = \frac{2}{n}\sum_{i=1}^n \phi_i^\top(X - \widehat{X})^\top\widehat{\Theta}(X - \widehat{X})\phi_i \le 27\lambda_{\max}(\Theta^\star)\lambda_{\max}(\Sigma_\phi)\|X - \widehat{X}\|_{\mathrm{F}}^2.$$

Meanwhile, by Taylor's theorem combined with the bound (36), for some $t \in [0,1]$,

$$\begin{aligned}\langle \Theta - \widehat{\Theta}, \nabla_\Theta\mathcal{L}(\widehat{X},\Theta) - \nabla_\Theta\mathcal{L}(\widehat{X},\widehat{\Theta})\rangle = \mathrm{vec}\left(\Theta - \widehat{\Theta}\right)^\top \nabla^2_{\Theta\Theta}\mathcal{L}(\widehat{X},(1-t)\Theta + t\widehat{\Theta})\mathrm{vec}\left(\Theta - \widehat{\Theta}\right) \\ \le \frac{4}{\lambda_{\min}^2(\Theta^\star)}\|\Theta - \widehat{\Theta}\|_{\mathrm{F}}^2,\end{aligned}$$

proving the desired results.

Cross-product bound Take $X \in \mathcal{X}_0$ and $\Theta \in \mathcal{Q}_0$. Then, by Taylor's theorem, it is sufficient to bound the difference of cross Hessian terms, i.e. for some $t, t' \in [0,1]$,

$$\begin{aligned}&|\langle X - \widehat{X}, \nabla_X\mathcal{L}(X,\Theta) - \nabla_X\mathcal{L}(X,\widehat{\Theta})\rangle - \langle \Theta - \widehat{\Theta}, \nabla_\Theta\mathcal{L}(X,\Theta) - \nabla_\Theta\mathcal{L}(\widehat{X},\Theta)\rangle| \\ &\quad\le \mathrm{vec}\left(X - \widehat{X}\right)^\top\left(\nabla^2_{X\Theta}\mathcal{L}(X, t\Theta + (1-t)\widehat{\Theta}) - \nabla^2_{X\Theta}\mathcal{L}(t'X + (1-t')\widehat{X},\Theta)\right)\mathrm{vec}\left(\Theta - \widehat{\Theta}\right). \\ &\quad= \frac{2(1-t')}{n}\sum_{i=1}^n \phi_i^\top(X - \widehat{X})^\top(\Theta - \widehat{\Theta})(X - \widehat{X})\phi_i \le \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{54}\|X - \widehat{X}\|_{\mathrm{F}}^2.\end{aligned}$$

This proves the cross-product condition with $\mu_X = \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{27}$ and $\mu_\Theta = 0$.

Initialization condition It has been proved in Barber and Ha [2017] that $\gamma_X(\mathcal{X}) = \frac{1}{2\sigma_r(X)}$, so we have

$$\max_{X \in \mathcal{X}_0}\gamma_X(\mathcal{X}) \le \frac{1}{2\sigma_r - 4\rho_X} \le \frac{1}{\sigma_r},$$

where the first inequality holds due to Weyl's inequality, and the next inequality holds since $\rho_X \le \frac{1}{4}\sigma_r$. Moreover, this shows that the conditions of Lemma 2 is satisfied, and so we also have $\gamma(\mathcal{X}_0) \le \frac{1}{\sigma_r}$.

Next, we bound the gradient term $\|\nabla_X \mathcal{L}(X,\Theta)\|_{\text{op}}$. Given the observational model $z_i = X^\star \phi_i + \epsilon_i$, we can decompose the gradient into the sum

$$\|\nabla_X \mathcal{L}(X,\Theta)\|_{\text{op}} \leq \left\| \frac{2}{n}\sum_{i=1}^n \Theta(X - X^\star)\phi_i\phi_i^\top \right\|_{\text{op}} + \left\| \frac{2}{n}\sum_{i=1}^n \Theta \cdot \epsilon_i \phi_i^\top \right\|_{\text{op}}.$$

Using the inequalities (36) and (37), the first term is upper bounded by $54\rho_X \cdot \lambda_{\max}(\Theta^\star) \cdot \lambda_{\max}(\Sigma_\phi)$, whereas we can bound the second term as

$$\left\| \frac{2}{n}\sum_{i=1}^n \Theta \cdot \epsilon_i \phi_i^\top \right\|_{\text{op}} \leq \frac{3\lambda_{\max}(\Theta^\star)}{\sqrt{\lambda_{\min}(\Theta^\star)}} \cdot \left\| \frac{1}{n}\sum_{i=1}^n \widetilde{\epsilon}_i \phi_i^\top \right\|_{\text{op}} \leq \frac{15\lambda_{\max}(\Theta^\star)\sqrt{\lambda_{\max}(\Sigma_\phi)}}{\sqrt{\lambda_{\min}(\Theta^\star)}}\sqrt{\frac{m+d}{n}},$$

where the steps use the inequalities (36) and (38). Combining the two and using the bound on ρ_X and the assumption (27), for sufficiently small $c_0, c_1 > 0$, we have

$$\max_{X\in\mathcal{X}_0,\Theta\in\mathcal{Q}_0} \|\nabla_X \mathcal{L}(X,\Theta)\|_{\text{op}} \leq \sigma_r \cdot \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{216},$$

and therefore

$$4\gamma(\mathcal{X}_0) \cdot \max_{X\in\mathcal{X}_0,\Theta\in\mathcal{Q}_0} \|\nabla_X \mathcal{L}(X,\Theta)\|_{\text{op}} \leq \frac{\lambda_{\min}(\Theta^\star)\lambda_{\min}(\Sigma_\phi)}{54} = \alpha_X - \mu_X,$$

which completes the proof of Lemma 7.

C.4 Proof of Lemma 8

Next we turn to prove our claims for the Gaussian factor model, as presented in (29). First, with some algebra, we have the following expression for the gradient and Hessian of $\mathcal{L}(X,Y)$: for all $\Delta_X, \Delta_Y \in \mathbb{R}^{d\times d}$,

$$\left\langle \begin{pmatrix} \Delta_X \\ \Delta_Y \end{pmatrix}, \nabla\mathcal{L}(X,Y) \right\rangle = \text{tr}((\Delta_X+\Delta_Y)^\top (X+Y)^{-1}(X+Y-S_n)(X+Y)^{-1}),$$

and

$$\begin{pmatrix} \Delta_X \\ \Delta_Y \end{pmatrix}^\top \nabla^2\mathcal{L}(X,Y) \begin{pmatrix} \Delta_X \\ \Delta_Y \end{pmatrix} = \text{vec}\,(\Delta_X)^\top \mathcal{H}(X,Y)\text{vec}\,(\Delta_X) + \text{vec}\,(\Delta_Y)^\top \mathcal{H}(X,Y)\text{vec}\,(\Delta_Y) + 2\text{vec}\,(\Delta_X)^\top \mathcal{H}(X,Y)\text{vec}\,(\Delta_Y),$$

where $\mathcal{H}(X,Y)$ is a d^2-by-d^2 matrix, given by

$$\mathcal{H}(X,Y) = \underbrace{\frac{1}{2}(X+Y)^{-1}(2S_n-(X+Y))(X+Y)^{-1}\otimes(X+Y)^{-1}}_{\mathcal{H}_1(X,Y)} + \underbrace{\frac{1}{2}(X+Y)^{-1}\otimes(X+Y)^{-1}(2S_n-(X+Y))(X+Y)^{-1}}_{\mathcal{H}_2(X,Y)}.$$

In the proof, the following concentration inequality will be used: since $z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\Sigma^\star)$ and S_n is a sample covariance matrix formed by $\{z_i\}_{i=1}^n$, with probability at least $1-2e^{-d}$, we have

$$\|S_n - \Sigma^\star\|_{\text{op}} \leq \|\Sigma^\star\|_{\text{op}} \|\Sigma^{\star -1/2} S_n \Sigma^{\star -1/2} - \mathbf{I}_d\|_{\text{op}} \leq 3\lambda_{\max}(\Sigma^\star)\sqrt{\frac{d}{n}}, \tag{39}$$

where the second step holds by a concentration bound on the extreme singular values of a standard Gaussian ensemble [Davidson and Szarek, 2001].

We calculate a few inequalities to use later. Recall $\rho_X, \rho_Y \leq c_0 \cdot \min\{\sigma_r(X^\star)\kappa^{-3}(\Sigma^\star), \lambda_{\min}(\Sigma^\star)\kappa^{-4}(\Sigma^\star)\}$ for a sufficiently small $c_0 > 0$. For $X \in \mathbb{B}_2(X_0, \rho_X)$ and $Y \in \mathbb{B}_2(Y_0, \rho_Y)$, assuming $X^\star \in \mathcal{X}_0, Y^\star \in \mathcal{Y}_0$, then

$$\|X + Y - \Sigma^\star\|_{\mathsf{op}} \leq \|X + Y - X_0 - Y_0\|_{\mathsf{op}} + \|X_0 + Y_0 - \Sigma^\star\|_{\mathsf{op}} \leq 2\rho_X + 2\rho_Y \leq \frac{\lambda_{\min}(\Sigma^\star)}{4},$$

where the last inequality follows from $\rho_X, \rho_Y \leq \frac{\lambda_{\min}(\Sigma^\star)}{16}$. Applying Weyl's inequality, this yields

$$\frac{3}{4}\lambda_{\min}(\Sigma^\star) \leq \lambda_{\min}(X + Y) \leq \lambda_{\max}(X + Y) \leq \frac{5}{4}\lambda_{\max}(\Sigma^\star). \tag{40}$$

Applying Weyl's inequality again and using the inequality (39), we also have

$$\frac{1}{2}\lambda_{\min}(\Sigma^\star) \leq \lambda_{\min}(2S_n - X - Y) \leq \lambda_{\max}(2S_n - X - Y) \leq \frac{3}{2}\lambda_{\max}(\Sigma^\star), \tag{41}$$

where we use the assumption $\sqrt{\frac{d}{n}} \leq \frac{\kappa^{-1}(\Sigma^\star)}{24}$. In particular, putting these bounds together and using standard properties of the Kronecker product, we further have

$$\frac{32}{125}\frac{\kappa^{-1}(\Sigma^\star)}{\lambda^2_{\max}(\Sigma^\star)} \leq \lambda_{\min}(\mathcal{H}(X, Y)) \leq \lambda_{\max}(\mathcal{H}(X, Y)) \leq \frac{32}{9}\frac{\kappa(\Sigma^\star)}{\lambda^2_{\min}(\Sigma^\star)}. \tag{42}$$

Finally, due to the spikiness constraint and the ℓ_1 norm inequality (34), we have the following finite bound on the inner product between the low-rank and sparse components: for all $X \in \mathcal{X}_0$ and all $Y \in \mathcal{Y}_0$,

$$\langle X - \widehat{X}, Y - \widehat{Y}\rangle \leq \|X - \widehat{X}\|_\infty \|Y - \widehat{Y}\|_1 \leq 4\alpha_{\mathsf{sp}}\sqrt{\frac{s}{d}}\|\Delta_Y\|_{\mathrm{F}} + 8\alpha_{\mathsf{sp}}\sqrt{\frac{s}{d}}\|\widehat{Y} - Y^\star\|_{\mathrm{F}}. \tag{43}$$

We are now prepared to prove the desired properties for the loss function of the factor model, $\mathcal{L}(X, Y) = \langle S_n, (X + Y)^{-1}\rangle - \log\det(X + Y)^{-1}$. Throughout the proof, we use the shorthand notation $\sigma_r = \sigma_r(X^\star)$.

Joint restricted strong convexity Take $X \in \mathcal{X}_0$ and $Y \in \mathcal{Y}_0$. By Taylor's theorem, it is sufficient to lower bound the term

$$\begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix}^\top \nabla^2\mathcal{L}(X(t), Y(t)) \begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix}$$
$$= \left(\mathrm{vec}\left(X - \widehat{X}\right) + \mathrm{vec}\left(Y - \widehat{Y}\right)\right)^\top \mathcal{H}(X(t), Y(t)) \left(\mathrm{vec}\left(X - \widehat{X}\right) + \mathrm{vec}\left(Y - \widehat{Y}\right)\right),$$

where $X(t) = (1 - t)X + t\widehat{X}$ and $Y(t) = (1 - t)Y + t\widehat{Y}$ for some $t \in [0, 1]$. By (42), this term is lower bounded by

$$\frac{32}{125}\frac{\kappa^{-1}(\Sigma^\star)}{\lambda^2_{\max}(\Sigma^\star)}\left\|\mathrm{vec}\left(X - \widehat{X}\right) + \mathrm{vec}\left(Y - \widehat{Y}\right)\right\|_2^2,$$

and furthermore, applying the inequality (43) and $ab \leq \frac{ca^2}{2} + \frac{b^2}{2c}$, we have

$$\begin{aligned}\left\|\mathrm{vec}\left(X - \widehat{X}\right) + \mathrm{vec}\left(Y - \widehat{Y}\right)\right\|_2^2 &\geq \|X - \widehat{X}\|_{\mathrm{F}}^2 + \|Y - \widehat{Y}\|_{\mathrm{F}}^2 - 8\alpha_{\mathsf{sp}}\sqrt{\frac{s}{d}}\|Y - \widehat{Y}\|_{\mathrm{F}} - 16\alpha_{\mathsf{sp}}\sqrt{\frac{s}{d}}\|\widehat{Y} - Y^\star\|_{\mathrm{F}} \\ &\geq \|X - \widehat{X}\|_{\mathrm{F}}^2 + \frac{1}{2}\|Y - \widehat{Y}\|_{\mathrm{F}}^2 - 16\|\widehat{Y} - Y^\star\|_{\mathrm{F}}^2 - 36\alpha_{\mathsf{sp}}^2\frac{s}{d}.\end{aligned}$$

Substituting this to the work above then yields

$$\begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix}^\top \nabla^2\mathcal{L}(X(t), Y(t)) \begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix}$$
$$\geq \frac{32}{125}\frac{\kappa^{-1}(\Sigma^\star)}{\lambda^2_{\max}(\Sigma^\star)}\left(\|X - \widehat{X}\|_{\mathrm{F}}^2 + \frac{1}{2}\|Y - \widehat{Y}\|_{\mathrm{F}}^2 - 16\|\widehat{Y} - Y^\star\|_{\mathrm{F}}^2 - 36\alpha_{\mathsf{sp}}^2\frac{s}{d}\right).$$

Restricted smoothness Analogous to the proof of joint RSC, it is straightforward to establish the restricted smoothness condition by using Taylor's theorem and the inequality (42); we have

$$\langle X - \widehat{X}, \nabla_X \mathcal{L}(X, \widehat{Y}) - \nabla_X \mathcal{L}(\widehat{X}, \widehat{Y})\rangle \leq \frac{32}{9} \frac{\kappa(\Sigma^\star)}{\lambda_{\min}^2(\Sigma^\star)} \|X - \widehat{X}\|_{\mathrm{F}}^2,$$

and analogously with the roles of X and Y reversed.

Cross-product bound As discussed in Section 3.1 following the Assumption 3, in order to establish the cross-product condition, it is sufficient to bound

$$\|\nabla_{XY}^2 \mathcal{L}(X, Y(t)) - \nabla_{XY}^2 \mathcal{L}(X(t'), Y)\|_{\mathrm{op}} \tag{44}$$

for all $X \in \mathcal{X}_0$ and $Y \in \mathcal{Y}_0$, where $X(t') = (1-t')X + t'\widehat{X}$ and $Y(t) = (1-t)Y + t\widehat{Y}$. Furthermore, since $\nabla_{XY}^2 \mathcal{L}(X, Y)$ is symmetric in X and Y, we only focus on bounding the term $\|\nabla_{XY}^2 \mathcal{L}(X, Y(t)) - \nabla_{XY}^2 \mathcal{L}(X, Y)\|_{\mathrm{op}}$; by the triangle inequality, this is sufficient to give a bound on the term (44). We also find that the operator norms of $\mathcal{H}_1$ and $\mathcal{H}_2$ are equivalent by the property of Kronecker product, and therefore,

$$\begin{aligned}\|\nabla_{XY}^2 \mathcal{L}(X, Y(t)) - \nabla_{XY}^2 \mathcal{L}(X, Y)\|_{\mathrm{op}} &= \|\mathcal{H}(X, Y(t)) - \mathcal{H}(X, Y)\|_{\mathrm{op}} \\ &\leq \|\mathcal{H}_1(X, Y(t)) - \mathcal{H}_1(X, Y)\|_{\mathrm{op}} + \|\mathcal{H}_2(X, Y(t)) - \mathcal{H}_2(X, Y)\|_{\mathrm{op}} \leq 2\|\mathcal{H}_1(X, Y(t)) - \mathcal{H}_1(X, Y)\|_{\mathrm{op}}.\end{aligned}$$

Now we work on bounding $\|\mathcal{H}_1(X, Y(t)) - \mathcal{H}_1(X, Y)\|_{\mathrm{op}}$. First, letting $\Delta\mathcal{H}_1 = (X + Y(t))^{-1}(2S_n - (X + Y(t)))(X + Y(t))^{-1} - (X + Y)^{-1}(2S_n - (X + Y))(X + Y)^{-1}$, then simple algebra yields

$$\begin{aligned}\mathcal{H}_1(X, Y(t)) - \mathcal{H}_1(X, Y) = \frac{1}{2}\Delta\mathcal{H}_1 \otimes (X + Y(t))^{-1} \\ + \frac{1}{2}(X + Y)^{-1}(2S_n - (X + Y))(X + Y)^{-1} \otimes \left((X + Y(t))^{-1} - (X + Y)^{-1}\right).\end{aligned}$$

$\Delta\mathcal{H}_1$ is further decomposed into the sum

$$\begin{aligned}\Delta\mathcal{H}_1 = &\left((X + Y(t))^{-1} - (X + Y)^{-1}\right)(2S_n - (X + Y(t)))(X + Y(t))^{-1} \\ &+ (X + Y)^{-1}(Y - Y(t))(X + Y(t))^{-1} + (X + Y)^{-1}(2S_n - (X + Y(t)))\left((X + Y(t))^{-1} - (X + Y)^{-1}\right).\end{aligned}$$

Meanwhile, by the inequalities (40) and (41), we have

$$\|(X + Y(t))^{-1}\|_{\mathrm{op}}, \|(X + Y)^{-1}\|_{\mathrm{op}} \leq \frac{4}{3\lambda_{\min}(\Sigma^\star)},\ \|2S_n - (X + Y(t))\|_{\mathrm{op}}, \|2S_n - (X + Y)\|_{\mathrm{op}} \leq \frac{3\lambda_{\max}(\Sigma^\star)}{2},$$

and so from the identity $(X + Y(t))^{-1} - (X + Y)^{-1} = (X + Y(t))^{-1}(Y - Y(t))(X + Y)^{-1}$, we have

$$\|(X + Y(t))^{-1} - (X + Y)^{-1}\|_{\mathrm{op}} \leq \frac{16}{9\lambda_{\min}^2(\Sigma^\star)} t\|Y - \widehat{Y}\|_{\mathrm{op}}.$$

This implies

$$\|\Delta\mathcal{H}_1\|_{\mathrm{op}} \leq \frac{64}{9}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda_{\min}^3(\Sigma^\star)}\|Y - \widehat{Y}\|_{\mathrm{op}} + \frac{16}{9}\frac{1}{\lambda_{\min}^2(\Sigma^\star)}\|Y - \widehat{Y}\|_{\mathrm{op}},$$

and hence that

$$\begin{aligned}\|\mathcal{H}_1(X, Y(t)) - \mathcal{H}_1(X, Y)\|_{\mathrm{op}} &\leq \frac{1}{2}\|\Delta\mathcal{H}_1\|_{\mathrm{op}}\|(X + Y(t))^{-1}\|_{\mathrm{op}} \\ &+ \frac{1}{2}\|(X + Y)^{-1}(2S_n - (X + Y))(X + Y)^{-1}\|_{\mathrm{op}}\|(X + Y(t))^{-1} - (X + Y)^{-1}\|_{\mathrm{op}} \\ &\leq \frac{192}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda_{\min}^4(\Sigma^\star)}\|Y - \widehat{Y}\|_{\mathrm{op}} + \frac{32}{27}\frac{1}{\lambda_{\min}^3(\Sigma^\star)}\|Y - \widehat{Y}\|_{\mathrm{op}} \leq \frac{224}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda_{\min}^4(\Sigma^\star)}\|Y - \widehat{Y}\|_{\mathrm{op}}.\end{aligned}$$

Returning to the cross product condition, this implies

$$\|\nabla^2_{XY}\mathcal{L}(X,Y(t)) - \nabla^2_{XY}\mathcal{L}(X,Y)\|_{\text{op}} \leq \frac{448}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda^4_{\min}(\Sigma^\star)}\|Y-\widehat{Y}\|_{\text{op}},$$

and therefore, by applying the symmetry argument from above, we have

$$\|\nabla^2_{XY}\mathcal{L}(X,Y(t)) - \nabla^2_{XY}\mathcal{L}(X(t'),Y)\|_{\text{op}} \leq \frac{448}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda^4_{\min}(\Sigma^\star)}\left(\|X-\widehat{X}\|_{\text{op}} + \|Y-\widehat{Y}\|_{\text{op}}\right).$$

Summarizing so far, we have shown that $\mu_X = \mu_Y = \frac{896}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda^4_{\min}(\Sigma^\star)}(\rho_X+\rho_Y)$. By choosing c_0 sufficiently small, this gives the claim $\mu_X = \mu_Y \leq \frac{16}{125}\frac{\lambda_{\min}(\Sigma^\star)}{\lambda^3_{\max}(\Sigma^\star)}$ as desired.

Initialization condition To prove the initialization condition, it is sufficient for us to bound the quantity $4\gamma(\mathcal{X}_0)\cdot\max_{X\in\mathcal{X}_0,Y\in\mathcal{Y}_0}\|\nabla_X\mathcal{L}(X,Y)\|_{\text{op}}$. Note that for any $X\in\mathcal{X}_0, Y\in\mathcal{Y}_0$,

$$\begin{aligned}\|\nabla_X\mathcal{L}(X,Y)\|_{\text{op}} &= \|(X+Y)^{-1}(X+Y-S_n)(X+Y)^{-1}\|_{\text{op}} \\ &\leq \frac{16}{9\lambda^2_{\min}(\Sigma^\star)}\cdot(\|X+Y-\Sigma^\star\|_{\text{op}} + \|S_n-\Sigma^\star\|_{\text{op}}) \leq \frac{8\sigma_r}{625}\frac{\lambda_{\min}(\Sigma^\star)}{\lambda^3_{\max}(\Sigma^\star)},\end{aligned}$$

where the first step uses the inequality (40), and the second step uses the inequality $\|X+Y-\Sigma^\star\|_{\text{op}} \leq 2(\rho_X+\rho_Y)$ and the concentration bound (39) as well as our assumptions on the radii and the sample size (30) (by choosing $c_0, c_1 > 0$ sufficiently small). Moreover, by the same reasoning to the equation (35), we also have $\gamma(\mathcal{X}_0) \leq \frac{5}{2\sigma_r}$. Therefore,

$$4\gamma(\mathcal{X}_0)\cdot\max_{X\in\mathcal{X}_0,Y\in\mathcal{Y}_0}\|\nabla_X\mathcal{L}(X,Y)\|_{\text{op}} \leq \frac{10}{\sigma_r}\cdot\frac{8\sigma_r}{625}\frac{\lambda_{\min}(\Sigma^\star)}{\lambda^3_{\max}(\Sigma^\star)} = \frac{16}{125}\frac{\lambda_{\min}(\Sigma^\star)}{\lambda^3_{\max}(\Sigma^\star)} = \alpha_X - \mu_X,$$

completing the proof of Lemma 8.