THE UNIVERSITY OF CHICAGO

HIGH-DIMENSIONAL ESTIMATION AND OPTIMIZATION WITH MULTIPLE STRUCTURED SIGNALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
WOOSEOK HA

CHICAGO, ILLINOIS
AUGUST 2018

CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This thesis would not have been possible without the support and guidance from numerous people I've met in the past five years at the University of Chicago.

First and foremost, I am deeply grateful for my advisor, Rina Foygel Barber, in her support and encouragement. Her vast academic interests and technical knowledge inspired me to find and expand my research areas. Every discussion with her was incredibly insightful and pushed me to move on to the next step. Rina's generous encouragement and caring attitude throughout my graduate studies has been the driving force for me to successfully finish my thesis.

I would also like to express my gratitude to Emil Sidky, who has been advising me since I began the medical imaging projects. He helped me enjoy research outside of statistics and I benefited a lot from our discussions. Further I would like to thank my committee member John Lafferty for his help and support throughout the entire process. I appreciated his well-prepared courses, from which I was able to gain more insight on many areas of statistics and machine learning. Also, I wish to thank all of the faculty and staff in the Department of Statistics for their support and endeavors.

I am thankful to our HELIOS group members all of whom have actively discussed on a range of research topics every week. The comments and discussions from HELIOS were extremely constructive and insightful to develop my research ideas. I am thankful to my fellow graduate students for being great friends and sharing precious memories. It was a wonderful experience with them at the University of Chicago.

Finally, I would like to thank my parents, Nam Ho Ha and Doosoon Kwon, for their constant love, encouragement, and patience.

ABSTRACT

Statistical recovery in high-dimensional statistics and signal processing often requests a determination of multiple structured signals from massive data. And depending on its application, either one or both of the signals may be of primarily interest. While most classical statistical techniques focus on the recovery of a single signal with other parameters being pre-fixed, the recent advances in mathematical and computational tools have facilitated the development of estimating multiple structured signals simultaneously. This thesis details several such problems with different goals of signal recovery.

Chapter 2 describes a low-rank + sparse decomposition problem under data compression and we study rigorous statistical performance guarantee that is achievable using a joint convex optimization based estimator. It is well known that the convex relaxation of the structural constraints leads to a large bias on the strong signals, although it affords computationally tractable algorithms. Chapter 3 develops a new notion of local concavity coefficients to directly handle nonconvexity of structural constraints. Based upon these coefficients, Chapter 4 analyzes convergence of alternating minimization when nonconvex constraints are placed on each of the variables. The theory developed here is general enough to encompass a broad class of multiple structured statistical models such as low-rank + sparse decomposition, multitask regression, and Gaussian factor model under a single framework. Chapter 5 discusses a simultaneous framework for the calibration of an imaging system and image reconstruction in CT imaging. As a preliminary work, an efficient optimization-based approach is proposed for spectrum estimation.

CHAPTER 1

INTRODUCTION

High-dimensional data are routinely faced in modern statistics, in which knowledge of the underlying structure of the parameter space is typically leveraged to allow for consistent estimation of parameters. Complex data may exhibit multiple signals at the same time, each of which represents different structures or components of interest. In this thesis, we will cover three scenarios involving *multiple structured signals* in the data, with different goal and perspective on signal recovery for each of the scenarios.

Matrix decomposition is perhaps one of the most well known problems in high-dimensional statistics where multiple structured signals naturally arise. In the matrix decomposition problem, a data matrix is typically observed as a noisy realization of the sum of a low-rank matrix and a sparse matrix (or variants of sparsity) and the goal is to simultaneously recover both components. Problems of matrix decomposition are best motivated by the robust principal component analysis (RPCA) problem, which seeks to separate low-rank trends from sparse outliers within a data matrix [1], that is, to approximate a matrix D as the sum of a low-rank matrix L and a sparse matrix S. For example, in video surveillance, if we stack the video frames into a matrix D, the low-rank matrix L capture a background component, whereas the sparse matrix S capture the foreground objects. Other examples include face recognition [1], factor analysis [2], latent variable graphical model [3], among others.

In Chapter 2, we consider the robust principal component analysis problem under data compression, where the data Y is now approximately given by $(L+S)\cdot C$, that is, a low-rank + sparse data matrix that has been compressed to a size substantially smaller than the original dimension via multiplication with a compression matrix C. Typical applications include data compression for the purpose of preserving privacy or for reducing resources such as communication bandwidth and storage space, and recovering a motion component from a background component in dynamic MRI [4]. We propose a convex program for recovering the sparse component S along with the compressed low-rank component $L\cdot C$, and derive deterministic upper bounds on the error of this reconstruction that scales naturally with the compression dimension m and coincides with existing results for the uncompressed setting. We will also consider model errors introduced through additive noise or through missing data, under which non-asymptotic bounds on the reconstruction error are derived in terms of dimension, compression, signal complexity, and noise magnitudes.

Various estimators for statistical recovery are based on minimizing a loss function (for example, a negative log-likelihood) while constraining signals to the underlying structural constraints—a method often referred to as a constrained (or regularized) M-estimator in the high-dimensional statistics literature; here the loss function typically measures how well the model fits the data while the constraint encourages desirable structure. Accordingly, much of the literature in this area focuses on developing a computationally tractable algorithm for solving such optimization problems. Computational issue concerning optimization becomes even more crucial as larger data sets are collected in recent years. Although common constraints arising in high-dimensional statistics are naturally nonconvex, such as sparsity, and low rank, many estimators are based on convex relaxations of these structured constraints, such as the ℓ_1 norm (as a convex approximation to sparsity) or nuclear norm (as a convex approximation of low rank). Working with a convex penalty or convex constraint, as a proxy for the nonconvex structure of the variable of interest, allows for easier optimization from both a theoretical and a practical point of view.

Further, in settings where multiple structures may be present simultaneously in the data, we may need to optimize a function over several variables, which are each believed to exhibit some latent structure. For instance, we seek to optimize both a low-rank term and a sparse term in the compressed robust PCA setting in Chapter 2. Alternating minimization is a simple yet powerful algorithm for solving this type of problem, in which we iteratively minimize the loss as a function of one variable while the other variables are fixed. A large body of research has been devoted to understanding the method under classical settings with convex constraints [5, 6]. In contrast to optimization with a single variable which is now better understood in high-dimensional settings (e.g. [7]), however, alternating minimization still lacks the corresponding theoretical justifications, despite its superior empirical performance. For the special cases such as matrix factorization [8], multivariate regression [9], and phase retrieval [10], a fast convergence rate (i.e. linear rate) has been established.

In Chapters 3 and 4, we examine convergence of gradient descent, alternating minimization, and other optimization algorithms in high-dimensional settings when the constraints are *nonconvex*. With the presence of nonconvex constraints, we may face an fundamental challenges that are both theoretical and practical in nature. To handle nonconvexity arising from the constraints, in Chapter 3, we develop the notion of *local concavity coefficients* of the constraint set, a measure of the extent to which a constraint set violates convexity. The extent of violation can be seen as a measure of "concavity" for the set, and thereby, this concavity coefficient naturally extends the standard theory of projected gradient descent with convex constraints. Chapter 4 uses the concavity coefficients as a tool to develop conditions for alternating minimization that allows for fast convergence of the algorithm under nonconvex constraints. As a byproduct of our analysis, we also show the computational gain inherent in the alternating minimization method, compared to the non-alternating method.

Finally in Chapter 5, we turn to simultaneous spectrum estimation and image reconstruction in CT imaging. Computed tomography (CT) is an imaging technology using x-ray beams to create cross-sectional images based on the transmission measurements of the scanned objects from multiple view angles. Due to the polychromatic nature of the x-ray beams, the x-ray spectrum, which accounts for the energy spectrum of the x-ray radiation source and the detector response across different energy values (detector spectral response), is typically unknown and needs to be estimated when realizing CT imaging (also known as spectral calibration).

Reconstructing the x-ray spectrum from transmission measurements is a common strategy for spectrum calibration. In this approach, the problem can be concisely written as a linear inverse problem; solving such an inverse problem, however, poses challenges since the system matrix is highly ill-conditioned which effectively leads to high-dimensionality of the spectrum relative to transmission measurements.

To address this issue, in Chapter 5, we begin with designing a new regularization scheme for the task of spectral calibration and derive a constrained optimization problem for accurately recovering the x-ray spectrum from transmission measurements. We use the exponentiated-gradient (EG) algorithm [11] to solve the optimization problem, which is seen to be efficient. While our focus is mostly on spectrum estimation for given image values, the ultimate goal of this work will be to simultaneously estimating x-ray spectrum and unknown images. This simultaneous framework will allow for calibration of CT system and reduction of the image artifacts at the same time, potentially enhancing diagnostic accuracy in real applications. We combine our spectral calibration approach and previously developed MOCCA algorithm [12] for spectral CT image reconstruction, and employ alternating minimization to perform simultaneous estimation on small size simulated data. The result suggests promising research direction for further investigation.

1.1 Summary

A common theme underlying this thesis is to investigate statistical models with multiple structured signals arising in different problems. We take many perspectives on estimating the signals depending on the circumstances. In Chapter 2, we consider the robust PCA problem where a data matrix is compressed so that we have access only to the compressed data. In Chapters 3 and 4, our focus is on investigating the performance of alternating minimization when nonconvex constraints are placed on the variables. In Chapter 5, we work in the CT imaging where our goal is to eventually realize the simultaneous image reconstruction and spectrum estimation algorithm.

1.2 Notation

Throughout we will use the following notation. We write $[n] = \{1,\dots,n\}$ for any $n \geq 1$. We write $\|x\|_0$ or $\|X\|_0$ to denote the number of nonzero entries in a vector x or matrix X (note that this is not in fact a norm). X_{i*} and X_{*j} denote the ith row and jth column of a matrix X (always treated as column vectors) and X_{AB} denotes the submatrix of X indexed by $A \times B$. We We will use the matrix norms $\|X\|_{\mathsf{F}}$ (Frobenius norm), $\|X\|_1$ (elementwise ℓ_1 norm), $\|X\|_\infty$ (elementwise

ℓ_∞ norm), $\|X\|_{2,\infty}$ (largest row ℓ_2 norm), $\|X\|_{\text{sp}}$ (spectral norm, i.e. largest singular value), and $\|X\|_{\text{nuc}}$ (nuclear norm, also known as the trace norm, given by the sum of the singular values of X). For a function $\mathsf{f} : \mathbb{R}^d \mapsto \mathbb{R}$, we write $\nabla \mathsf{f}$ and $\nabla^2 \mathsf{f}$ to denote a gradient and a Hessian respectively. Similarly, for a function $\mathsf{f} : \mathbb{R}^{d\times m} \mapsto \mathbb{R}$, we write $\nabla \mathsf{f}$ and $\nabla^2 \mathsf{f}$ to denote a gradient and a Hessian with respect to a *vectorized* variable. For $\rho > 0$, we use $\mathbb{B}_2(x,\rho)$ to denote the ℓ_2-ball of radius ρ centered around x. For a set T, we use $\mathscr{P}_T(\cdot)$ to denote the Euclidean projection onto T.

CHAPTER 2

LOW RANK + SPARSE DECOMPOSITION WITH COMPRESSED DATA

Principal component analysis (PCA) is a tool for providing a low-rank approximation to a data matrix $D \in \mathbb{R}^{n\times d}$, with the aim of reducing dimension or capturing the main directions of variation in the data. More recently, there has been increased focus on more general forms of PCA, that is more robust to realistic flaws in the data such as heavy-tailed outliers. The robust PCA (RPCA) problem formulates a decomposition of the data,

$$D \approx L + S\,,$$

into a low-rank component L (capturing trends across the data matrix) and a sparse component S (capturing outlier measurements that may obscure the low-rank trends), which we seek to separate based only on observing the data matrix D [1, 13].

In this chapter,[1] we examine the possibility of demixing sparse and low rank structure, under the additional challenge of working with data that has been compressed,

$$Y = D \cdot C \approx (L+S) \cdot C \in \mathbb{R}^{n\times m},$$

where $L, S \in \mathbb{R}^{n\times d}$ comprise the (approximately) low-rank and (approximately) sparse components of the original data matrix D, while $C \in \mathbb{R}^{d\times m}$ is a random or fixed compression matrix. In the compressed robust PCA setting, we hope to learn about both the low-rank and sparse components. Unlike compressed sensing problems where sparse structure may be reconstructed perfectly with undersampling, here we face a different type of challenge: the sparse component S is potentially identifiable from the compressed component $S \cdot C$, using the tools of compressed sensing; however, the low-rank component L is *not* identifiable from its compression $L \cdot C$. Specifically, if we let $\mathscr{P}_C \in \mathbb{R}^{d\times d}$ be the projection operator onto the column span of C, then the two low-rank matrices

[1] The work presented in this chapter is published in Ha and Barber [14].

L and $L' = L \cdot \mathscr{P}_C$ cannot be distinguished after multiplication by C.

Therefore, our goal will be to recover both the sparse component S, and the *compressed* low-rank component $L \cdot C$. Note that recovering $L \cdot C$ is similar to the goal of recovering the column span of L, which may be a useful interpretation if we think of the columns of the data matrix D as data points lying in $\mathbb{R}^n$; the column span of L characterizes a low-rank subspace of $\mathbb{R}^n$ that captures the main trends in the data.

2.1 Problem formulation

Consider the data, which takes the form of a $n \times d$ matrix, that is well-approximated by a sum $L^\star + S^\star$, where $L^\star$ is low-rank and $S^\star$ is sparse. However, we can only access this data through a (noisy) compression: our observed data is the $n \times m$ matrix

$$Y = (L^\star + S^\star) \cdot C + Z, \tag{2.1}$$

where $C \in \mathbb{R}^{d \times m}$ is the compression matrix, and $Z \in \mathbb{R}^{n \times m}$ absorbs all sources of error and noise—we discuss specific models for Z later on.

While we can aim to recover the sparse component $S^\star$, as we mentioned earlier, there is no hope to recover the original low-rank component $L^\star$, since $L^\star$ is not identifiable in the compressed model. Therefore, we instead aim to recover the underlying *compressed* low-rank component $P^\star := L^\star \cdot C$ and the sparse component $S^\star$. Specifically, our data model is now expressed as

$$Y = P^\star + S^\star \cdot C + Z. \tag{2.2}$$

In the ordinary robust PCA setting, the task of separating the low-rank and sparse components has been known to be possible when the underlying low-rank component $L^\star$ satisfies certain conditions such as an incoherence condition as in [1] requiring certain bounds on the singular vectors, or a spikiness condition as in [2] which bounds the matrix entries themselves. In order to successfully decompose the low-rank and sparse component in the compressed data, we thus need the similar conditions to hold for the compressed low-rank component $P^\star$. As we will see, if $L^\star$ satisfies the spikiness condition, i.e. $\|L^\star\|_\infty \leq \alpha_0$, then the *compressed* low-rank component $P^\star$ satisfies the similar spikiness condition, i.e. a bound on $\|P^\star C^\top\|_\infty$. This motivates the possibility to recover both the low-rank and sparse components in the case of compressed data.

We define our estimators of the sparse component $S^\star$, and the low-rank product $P^\star$, as follows:

$$(\widehat{P}, \widehat{S}) = \mathop{\arg\min}_{(P,S): \|PC^\top\|_\infty \leq \alpha} \left\{ \frac{1}{2} \|Y - P - S \cdot C\|_{\mathsf{F}}^2 + \nu \|P\|_{\text{nuc}} + \lambda \|S\|_1 \right\}. \tag{2.3}$$

Note that we impose the spikiness condition $\|PC^\top\|_\infty \leq \alpha$ on P, in order to guarantee good performance for demixing two such superimposed components—later in section 2.2, we will see that the same condition holds for $P^\star$.

2.1.1 Related work

Existing methods to separate the sparse and low-rank components include convex [1, 13] and non-convex [15] methods, and can handle extensions or additional challenges such as missing data [1], column-sparse rather than elementwise-sparse structure [16], streaming data [17, 18], and different types of structures superimposed with a low-rank component [2].

Random projection methods have been shown to be highly useful for reducing dimensionality without much loss of accuracy for numerical tasks such as least squares regression [19] or low-rank matrix computations [20]. Here we use random projections to compress data while preserving the information about the underlying low-rank and sparse structure. Zhou and Tao [21] also applied random projection methods to the robust PCA problem, but their purpose is to accelerate the computational task of low-rank approximation, which is different from the aim of our work.

The most relevant work to ours is Mardani et al. [4] where they work directly with the compressed model (2.2) without assuming the underlying data model (2.1). They are working in the noiseless setting and prove exact recovery under restricted isometry condition on the compressed matrix. While our results in this chapter are stated in terms of the original data, the same results will hold without assuming the underlying model (2.1). In this regard, our work can be seen as an extension of Mardani et al. [4] into the noisy setting with relaxed conditions (i.e. restricted eigenvalue property). See the remark following Theorem 2.2.1 for a more detailed discussion of this distinction.

2.1.2 Motivating examples

Here we illustrate several applications that involve data model of the form (2.1) and (2.2), along with models for the compression matrix C.

Random compression In some settings, the original data naturally lies in $R^{n\times d}$, but is compressed by the user for some purpose. In general, we think of the compression dimension m as being significantly smaller than d, motivated by several considerations:

- Communication constraints: if the $n \times d$ data matrix consists of d-dimensional measurements taken at n remote sensors, compression would allow the sensors to transmit information of dimension $m \ll d$;
- Storage constraints: storing a matrix with nm many entries instead of nd many entries;
- Data privacy: if the data is represented as the $n \times d$ matrix, where n-dimensional features were collected from d individuals, we can preserve privacy by compressing the data by a random linear transformation and allow the access to database only through the compressed data. This privacy-preserving method has been called *matrix masking* in the privacy literature and studied by [22] in the context of high-dimensional linear regression.

In either case, we control the choice of the compression matrix C, and are free to use a simple random model. Here we consider two models:

$$\text{Gaussian model: the entries of } C \text{ are generated as } C_{ij} \overset{\text{iid}}{\sim} N(0,1/m). \tag{2.4}$$

$$\begin{gathered}\text{Orthogonal model: } C = \sqrt{d/m}\cdot U, \\ \text{where } U \in \mathbb{R}^{d\times m} \text{ is an orthonormal matrix chosen uniformly at random.}\end{gathered} \tag{2.5}$$

Note that in each case, $\mathbb{E}\left[CC^\top\right] = \mathbf{I}_d$.

Deterministic compression In other settings, compression matrix cannot be controlled by the user and is determined through observing the specific event or phenomenon. For instance, in a multitask learning, if the unknown regression matrix is approximately low-rank + sparse, the model can precisely be written in the form of (2.1) by taking the transpose: in this case, the compression matrix C is given by the transpose of the design matrix.

Another example is dynamic MRI, where measurements are acquired in a temporal series in order to resolve degradation of the quality of MRI due to respiratory motion [4]. Each image comprises of a background component and a motion component, and the motion component often admits a sparse representation under some dictionary D. If we stack the background component and the motion component of the dynamic MRI frames into matrices, which we denote by L and $D\cdot S$, the scanned temporal sequences of images in the frequency domain can be written as

$$Y \approx \Phi(L+DS),$$

where Φ is the partial FFT matrix consisting of a row subset of the full FFT matrix. Compare to the model (2.2), where we replace Y and S with $Y^\top$ and $S^\top$, and use the compression matrix $C = (\Phi D)^\top$. Then, if we set $P = (\Phi L)^\top$, the measurement model for dynamic MRI can be treated as a special case of our general model (2.2). For the purpose of dynamic MRI, since the motion component is a major concern, it suffices to recover S and subsequently $D\cdot S$, which coincides the aim of this study. More examples, such as traffic anomaly detection and face recognition, can be found in [4].

2.1.3 *Sources of errors and noise*

Next, we give several examples of models and interpretations for the error term Z in (2.1) and (2.2).

Random noise First, we may consider a model where the signal has an exact low-rank + sparse decomposition, with well-behaved additive noise added before and/or after the compression step:

$$Y = (L^\star + S^\star + Z_{\mathsf{pre}}) \cdot C + Z_{\mathsf{post}}, \tag{2.6}$$

where the entries of the pre- and post-compression noise, Z_{pre} and Z_{post}, are i.i.d. mean-zero subgaussian random variables. In this case, the noise term Z in (2.1) and (2.2) is given by $Z = Z_{\mathsf{pre}} \cdot C + Z_{\mathsf{post}}$.

Missing data Given an original data matrix $D = L^\star + S^\star$, we might have access only to a partial version of this matrix. We write D_Ω to denote the available data, where $\Omega \subset [n] \times [d]$ indexes the entries where data is available, and $(D_\Omega)_{ij} = D_{ij} \cdot \mathbb{1}_{ij \in \Omega}$. Then, a low-rank + sparse model for our compressed data is given by

$$Y = D_\Omega \cdot C = (L^\star + S^\star_\Omega) \cdot C + Z_{\mathsf{missing}} \cdot C,$$

where $Z_{\mathsf{missing}} = L^\star_\Omega - L^\star$. In some settings, we may first want to adjust D_Ω before compressing the data, for instance, by reweighting the observed entries in D_Ω to ensure a closer approximation to D. Denoting the reweighted matrix of partial observations by $\widetilde{D}_\Omega$, we have compressed data

$$Y = \widetilde{D}_\Omega \cdot C = (L^\star + \widetilde{S}^\star_\Omega) \cdot C + Z_{\mathsf{missing}} \cdot C, \tag{2.7}$$

with $Z_{\mathsf{missing}} = \widetilde{L}^\star_\Omega - L^\star$, and where $\widetilde{S}^\star_\Omega$ is the reweighted matrix of $S^\star_\Omega$. Then the error from the missing data can be absorbed into the Z term, i.e. $Z = Z_{\mathsf{missing}} \cdot C$.

2.1.4 Restricted eigenvalue condition

We state a version of the Restricted Eigenvalue property found in the compressed sensing and sparse regression literature [23], which plays a key role in our analysis:

Definition 2.1.1. For a matrix $X \in \mathbb{R}^{m\times d}$ and for $c_1, c_2 \geq 0$, X satisfies the restricted eigenvalue property with constants (c_1, c_2), denoted by $\mathsf{RE}_{m,d}(c_1, c_2)$, if

$$\|Xv\|_2 \geq c_1\|v\|_2 - c_2 \cdot \sqrt{\frac{\log(d)}{m}} \cdot \|v\|_1 \text{ for all } v \in \mathbb{R}^d. \tag{2.8}$$

2.2 Theoretical results

Now we develop theoretical error bounds for the compressed robust PCA problem under several of the scenarios described above.

2.2.1 Deterministic result

We first give a general deterministic result for the accuracy of the convex program (2.3).

Theorem 2.2.1. *Let $L^\star \in \mathbb{R}^{n\times d}$ be any matrix with $\operatorname{rank}(L^\star) \leq r$, and let $S^\star \in \mathbb{R}^{n\times d}$ be any matrix with at most s nonzero entries per row, that is, $\max_i \|S^\star_{i*}\|_0 \leq s$. Let $C \in \mathbb{R}^{d\times m}$ be any compression matrix and define the data Y and the error/noise term Z as in* (2.1)*. Let $P^\star = L^\star \cdot C$ as before. Suppose that $C^\top$ satisfies $\mathsf{RE}_{m,d}(c_1, c_2)$, where $c_0 := c_1 - c_2 \cdot \sqrt{16 s \log(d)/m} > 0$. If parameters (α, ν, λ) satisfy*

$$\alpha \geq \|L^\star C C^\top\|_\infty,\ \nu \geq 2\|Z\|_{\mathrm{sp}},\ \lambda \geq 2\|Z C^\top\|_\infty + 4\alpha, \tag{2.9}$$

then deterministically, the solution $(\widehat{P},\widehat{S})$ *to the convex program* (2.3) *satisfies*

$$\|\widehat{P}-P^\star\|_{\mathsf{F}}^2+c_0^2\|\widehat{S}-S^\star\|_{\mathsf{F}}^2 \le 18r\nu^2+9c_0^{-2}sn\lambda^2.$$

Remark on model assumptions It is worthwhile to mention that while Theorem 2.2.1 assumes the underlying model (2.1), analogous results can be obtained without assuming (2.1) but only with the model (2.2). In particular, if the spikiness condition holds for $P^\star$, namely$\|P^\star C^\top\|_\infty \le \alpha$, and $C^\top$ satisfies $\mathsf{RE}_{m,d}(c_1,c_2)$, under the same choice of parameters as in (2.9), the same error bounds holds, as long as $\mathrm{rank}(P^\star)\le r$ and $\max_i\|S^\star_{i*}\|_0\le s$.

2.2.2 Results for random compression with subgaussian noise

We specialize our main result to handle scenarios of pre- and post-compression noise, given in (2.6). We assume the compression matrix C is generated under either the Gaussian (2.4) or orthogonal (2.5) model, and the noise matrices $Z_{\mathrm{pre}},Z_{\mathrm{post}}$ are independent from each other and from C, with entries

$$(Z_{\mathrm{pre}})_{ij}\overset{\mathrm{iid}}{\sim}N(0,\sigma^2_{\mathrm{pre}}) \text{ and } (Z_{\mathrm{post}})_{ij}\overset{\mathrm{iid}}{\sim}N(0,\sigma^2_{\mathrm{post}}).$$

For this section, we assume $d\ge m$ without further comment (that is, the compression should reduce the dimension of the data). Let $\sigma^2_{\max}\ge\max\{\sigma^2_{\mathrm{pre}},\sigma^2_{\mathrm{post}}\}$. Specializing the result of Theorem 2.2.1 to this setting, we obtain the following probablistic guarantee:

Theorem 2.2.2. *Assume the model* (2.6)*. Suppose that* $\mathrm{rank}(L^\star)\le r$*,* $\max_i\|S^\star_{i*}\|_0\le s$*, and* $\|L^\star\|_\infty\le\alpha_0$*. Then there exist universal constants* $c,c',c''>0$ *such that if we define*

$$\alpha=5\alpha_0\sqrt{\frac{d\log(nd)}{m}},\ \nu=24\sigma_{\max}\sqrt{\frac{d(n+m)}{m}},\ \lambda=32\sigma_{\max}\sqrt{\frac{d\log(nd)}{m}}+4\alpha,$$

and if $m \geq c \cdot s \log(nd)$*, then the solution* $(\widehat{P}, \widehat{S})$ *to the convex program* (2.3) *satisfies*

$$\|\widehat{P} - P^\star\|_{\mathsf{F}}^2 + \|\widehat{S} - S^\star\|_{\mathsf{F}}^2 \leq c' \cdot \frac{d}{m}\left(\sigma_{\max}^2 \cdot r(n+m) + (\sigma_{\max}^2 + \alpha_0^2) \cdot sn \log(nd)\right) \tag{2.10}$$

with probability at least $1 - \frac{c''}{nd}$.

We remark that if the entries of Z_{pre} and Z_{post} are subgaussian rather than Gaussian, then the same result holds, except for a change in the constants appearing in the parameters (α, ν, λ). Theorem 2.2.2 shows the natural scaling: the first term $r(n+m)$ is the degree of freedom for *compressed* rank r matrix P whereas the term $sn\log(nd)$ is the signal complexity of sparse component S, which has sn many nonzero entries. The multiplicative factor $\frac{d}{m}\sigma_{\max}^2$ can be interpreted as the noise variance of the problem amplified by the compression.

2.2.3 Results for random compression with missing data

Next, we consider a missing data scenario where the original $n \times d$ matrix is only partially observed. We first specify a model for the missing data. For each $(i, j) \in [n] \times [d]$, let $\rho_{ij} \in [0, 1]$ be the probability that this entry is *observed*. Additionally, we assume that the sampling scheme is independent across all entries, and that the ρ_{ij}'s are known.[2]

Define reweighted versions of the partially observed data matrix and the low rank and sparse components:

$$(\widetilde{D}_\Omega)_{ij} = D_{ij}/\rho_{ij} \cdot \mathbb{1}_{ij \in \Omega} \quad \text{and} \quad (\widetilde{L}^\star_\Omega)_{ij} = L_{ij}/\rho_{ij} \cdot \mathbb{1}_{ij \in \Omega} \quad \text{and} \quad (\widetilde{S}^\star_\Omega)_{ij} = S_{ij}/\rho_{ij} \cdot \mathbb{1}_{ij \in \Omega},$$

and consider the model (2.7), where Y is approximated with a compression of $L^\star + \widetilde{S}^\star_\Omega$. The role of the reweighting step is to ensure that this noise term Z has mean zero. Note that, while the original sparse component $S^\star$, is not identifiable via the missing data model (since we have no

[2] In practice, the assumption that ρ_{ij}'s are known is not prohibitive. For example, we might model $\rho_{ij} = \alpha_i \beta_j$ (the row and column locations of the observed entries are chosen independently, e.g. see [24]), or a logistic model, $\log\left(\frac{\rho_{ij}}{1-\rho_{ij}}\right) = \alpha_i + \beta_j$. In either case, fitting a model using the observed set Ω is extremely accurate.

information to help us recover entries $S^\star_{ij}$ for $(i,j) \notin \Omega$), this new decomposition $L^\star + \widetilde{S}^\star_\Omega$ now has a sparse component that *is* identifiable, since by definition, $\widetilde{S}^\star_\Omega$ preserves the sparsity of $S^\star$ but has no nonzero entries in unobserved locations, that is, $(\widetilde{S}^\star_\Omega)_{ij} = 0$ whenever $(i,j) \notin \Omega$.

With this model in place, we obtain the following probabilistic guarantee for this setting, which is another specialized version of Theorem 2.2.1.

Theorem 2.2.3. *Assume the model* (2.7). *Suppose that* $\mathrm{rank}(L^\star) \leq r$, $\max_i \|S^\star_{i*}\|_0 \leq s$, *and* $\|L^\star\|_\infty \leq \alpha_0$. *If the sampling scheme satisfies* $\rho_{ij} \geq \rho_{\mathsf{min}}$ *for all* $(i,j) \in [n] \times [d]$ *for some positive constant* $\rho_{\mathsf{min}} > 0$, *then there exist universal constants* $c, c', c'' > 0$ *such that if we define*

$$\alpha = 5\alpha_0\sqrt{\frac{d\log(nd)}{m}},\ \nu = 10\rho_{\mathsf{min}}^{-1}\alpha_0\sqrt{\frac{d(n+m)\log(nd)}{m}},\ \lambda = 12\rho_{\mathsf{min}}^{-1}\alpha_0\sqrt{\frac{d\log^2(nd)}{m}} + 4\alpha,$$

and if $m \geq c \cdot s\log(nd)$, *then the solution* $(\widehat{P}, \widehat{S})$ *to the convex program* (2.3) *satisfies*

$$\|\widehat{P} - P^\star\|_{\mathsf{F}}^2 + \|\widehat{S} - \widetilde{S}^\star_\Omega\|_{\mathsf{F}}^2 \leq c' \cdot \frac{d}{m} \cdot \rho_{\mathsf{min}}^{-2}\alpha_0^2 \left(r(n+m)\log(nd) + sn\log^2(nd)\right)$$

with probability at least $1 - \frac{c''}{nd}$.

2.3 Empirical results

Now we use simulated data to study the behavior of the convex program (2.3) for different compression dimensions, signal complexities and missing levels. We generate the compression matrix C under the orthogonal model (2.5). We solve the convex program (2.3) via alternating minimization over L and S, selecting the regularization parameters ν and λ that minimizes the squared Frobenius error. For simplicity, in all experiments, we select $\alpha = \infty$, which is easier for optimization and generally results in a solution that still has low spikiness (that is, the solution is the same as if we had imposed a bound with finite α). All results are averaged over 5 trials.

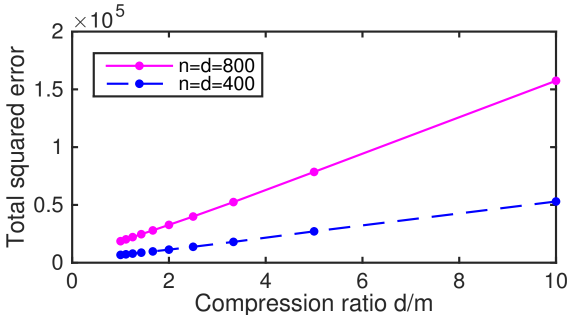

Figure 2.1: The total squared error, calculated as in Theorem 2.2.2, is plotted against the compression ratio d/m.

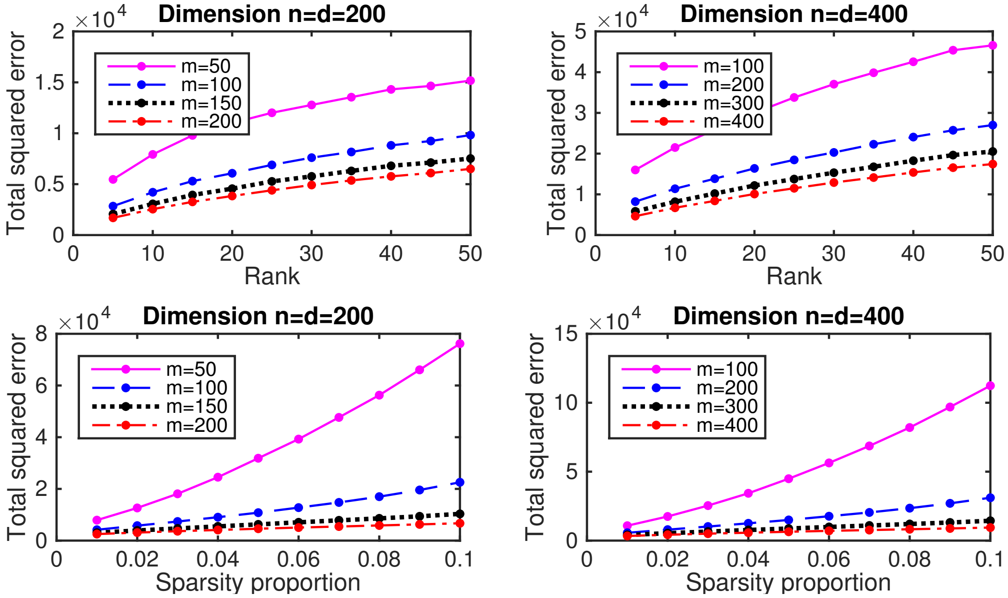

Figure 2.2: The total squared error, calculated as in Theorem 2.2.2, is plotted against the rank r or sparsity proportion s/d.

2.3.1 *Compression ratio.*

First we examine the role of the compression dimension m. We fix the matrix dimension $n = d \in \{400, 800\}$. The low-rank component is given by $L^\star = \sqrt{r} \cdot UV^\top$, where U and V are $n \times r$ and $d \times r$ matrices with i.i.d. $N(0,1)$ entries, for rank $r = 10$. The sparse component $S^\star$ has 1% of its entries generated as $5 \cdot N(0,1)$, that is, $s = 0.01d$. The data is $D = L^\star + S^\star + Z$, where $Z_{ij} \overset{\text{iid}}{\sim} N(0, 0.25)$. Figure 2.1 shows the squared Frobenius error $\|\widehat{P} - P^\star\|_{\mathsf{F}}^2 + \|\widehat{S} - S^\star\|_{\mathsf{F}}^2$ plotted against the compression ratio d/m. We see error scaling linearly with the compression ratio, which supports our theoretical results.

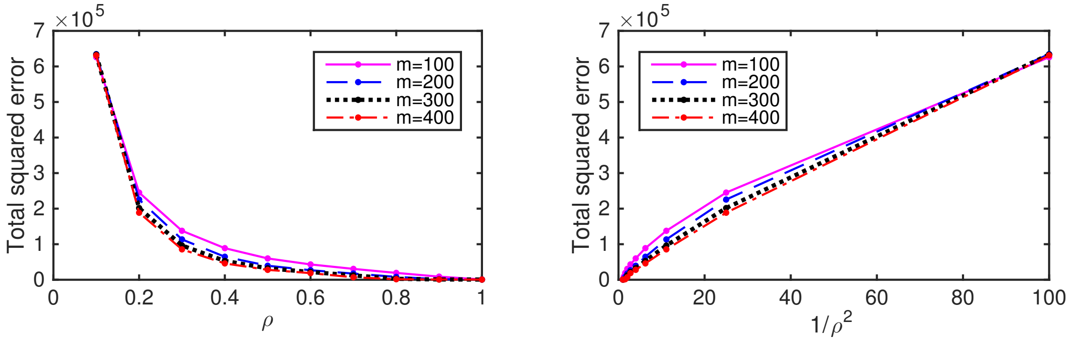

Figure 2.3: The total squared error, calculated as in Theorem 2.2.3, is plotted against ρ (proportion of observed data) or against $1/\rho^2$, for various values of m, based on one trial.

2.3.2 *Rank and sparsity.*

Next we study the role of rank and sparsity, for a matrix of size $n = d = 200$ or $n = d = 400$. We generate the data D as before, but we either vary the rank $r \in \{5, 10, \ldots, 50\}$, or we vary the sparsity s with $s/d \in \{0.01, 0.02, \ldots, 0.1\}$. Figure 2.2 shows the squared Frobenius error plotted against either the varying rank or the varying sparsity. We repeat this experiment for several different compression dimensions m. We see a little deviation from linear scaling for the smallest m, which can be due to the fact that our theorems give upper bounds rather than tight matching upper and lower bounds (or perhaps the smallest value of m does not satisfy the condition stated in the theorems). However, for all but the smallest m, we see error scaling nearly linearly with rank or with sparsity, which is consistent with our theory.

2.3.3 *Missing data.*

Finally, we perform experiments under the existence of missing entries in the data matrix $D = L^\star + S^\star$. We fix dimensions $n = d = 400$ and generate $L^\star$ and $S^\star$ as before, with $r = 10$ and $s = 0.01d$, but do not add noise. To introduce the missing entries in the data, we use a uniform sampling scheme, where each entry of D is observed with probability ρ, with $\rho \in \{0.1, 0.2, \ldots, 1\}$. Figure 2.3 shows the squared Frobenius error $\|\widehat{P} - P^\star\|_{\mathsf{F}}^2 + \|\widehat{S} - \widetilde{S}^\star_\Omega\|_{\mathsf{F}}^2$ (see Theorem 2.2.3 for details) across a range of probabilities ρ. We see that the squared error scales approximately linearly with $1/\rho^2$, as predicted by our theory.

2.4 Proofs

2.4.1 Background

First we introduce a few definitions using the decomposability of the ℓ_1 norm and the nuclear norm. Let $\Omega \subset [n] \times [d]$ be the support of the true sparse component $S^\star$, and let $\Omega_i \subset [d]$ be the i-th row of Ω, i.e. $\Omega_i = \{j : S^\star_{ij} \neq 0\}$. Let T be the tangent space to the nuclear norm at $P^\star$, which is given by [1]

$$T = \{AV^\top + UB^\top : \text{ any matrices } A \in \mathbb{R}^{n\times r}, B \in \mathbb{R}^{m\times r}\},$$

where $P^\star = U\Sigma V^\top$ is a singular value decomposition of $P^\star$ with $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$. It is known [23] that, for any $S \in \mathbb{R}^{n\times d}$, for each row $i \in [n]$,

$$\|S^\star_{i*}\|_1 - \|S_{i*}\|_1 \leq \|(S-S^\star)_{i\Omega_i}\|_1 - \|(S-S^\star)_{i\Omega_i^c}\|_1, \tag{2.11}$$

which trivially yields

$$\|S^\star\|_1 - \|S\|_1 \leq \|\mathscr{P}_\Omega(S-S^\star)\|_1 - \|\mathscr{P}_\Omega^\perp(S-S^\star)\|_1, \tag{2.12}$$

where $\mathscr{P}_\Omega()$ and $\mathscr{P}_\Omega^\perp()$ denote projection onto the subspace of matrices supported on Ω, and onto the orthogonal subspace. Furthermore for any $P \in \mathbb{R}^{n\times m}$,

$$\|P^\star\|_{\text{nuc}} - \|P\|_{\text{nuc}} \leq \|\mathscr{P}_T(P-P^\star)\|_{\text{nuc}} - \|\mathscr{P}_T^\perp(P-P^\star)\|_{\text{nuc}}, \tag{2.13}$$

where $\mathscr{P}_T()$ and $\mathscr{P}_T^\perp()$ denote projection onto the subspace $T \subset \mathbb{R}^{n\times m}$, and onto its orthogonal complement $T^\perp$. Throughout, we will use the facts that $\|M\|_{\text{nuc}} \leq \|\mathscr{P}_T(M)\|_{\text{nuc}} + \|\mathscr{P}_T^\perp(M)\|_{\text{nuc}}$ and similarly $\|M\|_1 \leq \|\mathscr{P}_\Omega(M)\|_1 + \|\mathscr{P}_\Omega^\perp(M)\|_1$ without comment.

2.4.2 Proofs of Theorems

Proof of Theorem 2.2.1. By optimality,

$$\frac{1}{2}\|Y-\widehat{P}-\widehat{S}C\|_{\mathsf{F}}^2+\nu\|\widehat{P}\|_{\mathrm{nuc}}+\lambda\|\widehat{S}\|_1 \le \frac{1}{2}\|Y-P^\star-S^\star G\|_{\mathsf{F}}^2+\nu\|P^\star\|_{\mathrm{nuc}}+\lambda\|S^\star\|_1\ . \tag{2.14}$$

Define errors $\Delta^P=\widehat{P}-P^\star$ and $\Delta^S=\widehat{S}-S^\star$. Using our model (2.2) for Y, and applying (2.12) and (2.13), we rearrange terms to obtain

$$\begin{aligned}
\frac{1}{2}\|\Delta^P+\Delta^S C\|_{\mathsf{F}}^2 &\le \langle Z,\Delta^P+\Delta^S C\rangle+\nu\left(\|\mathscr{P}_T(\Delta^P)\|_{\mathrm{nuc}}-\|\mathscr{P}_T^\perp(\Delta^P)\|_{\mathrm{nuc}}\right)\\
&\qquad\qquad\qquad\qquad\qquad +\lambda\left(\|\mathscr{P}_\Omega(\Delta^S)\|_1-\|\mathscr{P}_\Omega^\perp(\Delta^S)\|_1\right)\\
&\le \|Z\|_{\mathrm{sp}}\cdot\|\Delta^P\|_{\mathrm{nuc}}+\|ZC^\top\|_\infty\cdot\|\Delta^S\|_1+\nu\left(\|\mathscr{P}_T(\Delta^P)\|_{\mathrm{nuc}}-\|\mathscr{P}_T^\perp(\Delta^P)\|_{\mathrm{nuc}}\right)\\
&\qquad\qquad\qquad\qquad\qquad +\lambda\left(\|\mathscr{P}_\Omega(\Delta^S)\|_1-\|\mathscr{P}_\Omega^\perp(\Delta^S)\|_1\right)\\
&\le \|\mathscr{P}_T(\Delta^P)\|_{\mathrm{nuc}}(\nu+\|Z\|_{\mathrm{sp}})-\|\mathscr{P}_T^\perp(\Delta^P)\|_{\mathrm{nuc}}(\nu-\|Z\|_{\mathrm{sp}})\\
&\qquad +\|\mathscr{P}_\Omega(\Delta^S)\|_1(\lambda+\|ZC^\top\|_\infty)-\|\mathscr{P}_\Omega^\perp(\Delta^S)\|_1(\lambda-\|ZC^\top\|_\infty)\ .
\end{aligned}$$

Now we consider the left-hand side. We have

$$\begin{aligned}
\frac{1}{2}\|\Delta^P+\Delta^S C\|_{\mathsf{F}}^2 &= \frac{1}{2}\|\Delta^P\|_{\mathsf{F}}^2+\frac{1}{2}\|\Delta^S C\|_{\mathsf{F}}^2+\langle\Delta^P,\Delta^S C\rangle\\
&\ge \frac{1}{2}\|\Delta^P\|_{\mathsf{F}}^2+\frac{1}{2}\|\Delta^S C\|_{\mathsf{F}}^2-\|\Delta^P C^\top\|_\infty\cdot\|\Delta^S\|_1\\
&\ge \frac{1}{2}\|\Delta^P\|_{\mathsf{F}}^2+\frac{1}{2}\|\Delta^S C\|_{\mathsf{F}}^2-2\alpha\|\Delta^S\|_1\ ,
\end{aligned}$$

where the last step uses $\|\Delta^P C^\top\|_\infty\le\|\widehat{P}C^\top\|_\infty+\|P^\star C^\top\|_\infty\le 2\alpha$ by the assumption $\|P^\star C^\top\|_\infty\le\alpha$ (2.9) and the constraint $\|\widehat{P}C^\top\|_\infty\le\alpha$ in the optimization problem (2.3). Including this into the work above, then,

$$\frac{1}{2}\|\Delta^P\|_{\mathsf{F}}^2+\frac{1}{2}\|\Delta^S C\|_{\mathsf{F}}^2 \leq \|\mathscr{P}_T(\Delta^P)\|_{\mathrm{nuc}}(\nu+\|Z\|_{\mathrm{sp}})-\|\mathscr{P}_T^{\perp}(\Delta^P)\|_{\mathrm{nuc}}(\nu-\|Z\|_{\mathrm{sp}})$$
$$+\|\mathscr{P}_\Omega(\Delta^S)\|_1(\lambda+\|ZC^\top\|_\infty+2\alpha)-\|\mathscr{P}_\Omega^{\perp}(\Delta^S)\|_1(\lambda-\|ZC^\top\|_\infty-2\alpha)$$
$$\leq \nu(1.5\|\mathscr{P}_T(\Delta^P)\|_{\mathrm{nuc}}-0.5\|\mathscr{P}_T^{\perp}(\Delta^P)\|_{\mathrm{nuc}})+\lambda(1.5\|\mathscr{P}_\Omega(\Delta^S)\|_1-0.5\|\mathscr{P}_\Omega^{\perp}(\Delta^S)\|_1)\,, \quad (2.15)$$

where the last step uses the assumptions (2.9) on the parameters (α,ν,λ).

Next, we need to use the restricted strong convexity assumption on C. First, we consider the rows of $\widehat{S}$ individually. Fixing $\widehat{P}$, we note that the optimization problem (2.3) separates over the rows of $\widehat{S}$: ignoring the term $\nu\|\widehat{P}\|_{\mathrm{nuc}}$ which is constant with respect to S, we have

$$\frac{1}{2}\|Y-\widehat{P}-\widehat{S}C\|_{\mathsf{F}}^2+\lambda\|\widehat{S}\|_1=\sum_i\left(\frac{1}{2}\|Y_{i*}-\widehat{P}_{i*}-C^\top\widehat{S}_{i*}\|_2^2+\lambda\|\widehat{S}_{i*}\|_1\right)\,.$$

Therefore, $\widehat{S}_{i*}$ is the minimizer of the term in parentheses, for each i, and in particular we have

$$\frac{1}{2}\|Y_{i*}-\widehat{P}_{i*}-C^\top\widehat{S}_{i*}\|_2^2+\lambda\|\widehat{S}_{i*}\|_1 \leq \frac{1}{2}\|Y_{i*}-\widehat{P}_{i*}-C^\top S^\star_{i*}\|_2^2+\lambda\|S^\star_{i*}\|_1\,.$$

Rearranging terms and applying (2.11), we get

$$\frac{1}{2}\|C^\top(\widehat{S}_{i*}-S^\star_{i*})\|_2^2 \leq \langle Y_{i*}-\widehat{P}_{i*}-C^\top S^\star_{i*},\Delta^S_{i*}\rangle+\lambda\left(\|\Delta^S_{i\Omega_i}\|_1-\|\Delta^S_{i\Omega_i^c}\|_1\right)$$
$$\leq \|C(Y_{i*}-\widehat{P}_{i*}-C^\top S^\star_{i*})\|_\infty\cdot\|\Delta^S_{i*}\|_1+\lambda\left(\|\Delta^S_{i\Omega_i}\|_1-\|\Delta^S_{i\Omega_i^c}\|_1\right)\,.$$

We also have

$$\|C(Y_{i*}-\widehat{P}_{i*}-C^\top S^\star_{i*})\|_\infty=\|C(Z_{i*}-(\widehat{P}-P^\star)_{i*})\|_\infty \leq \|(Z-(\widehat{P}-P^\star))C^\top\|_\infty$$
$$\leq \|ZC^\top\|_\infty+\|\widehat{P}C^\top\|_\infty+\|P^\star C^\top\|_\infty \leq \|ZC^\top\|_\infty+2\alpha \leq \lambda/2\,,$$

by the assumption (2.9) on λ. Combining this with the above, we then have

$$\frac{1}{2}\|C^\top \Delta^S_{i*}\|_2^2 \le \lambda\left(1.5\|\Delta^S_{i\Omega_i}\|_1 - 0.5\|\Delta^S_{i\Omega_i^c}\|_1\right) ,$$

and since the left-hand side is nonnegative, we therefore have

$$\|\Delta^S_{i\Omega_i^c}\|_1 \le 3\|\Delta^S_{i\Omega_i}\|_1 ,$$

that is, for *every* row of the sparse matrix, a substantial portion of the ℓ_1 norm of the error is located on the correct support. Therefore,

$$\|\Delta^S_{i*}\|_1 = \|\Delta^S_{i\Omega_i^c}\|_1 + \|\Delta^S_{i\Omega_i}\|_1 \le 4\|\Delta^S_{i\Omega_i}\|_1 \le 4\sqrt{s}\|\Delta^S_{i\Omega_i}\|_2 \le 4\sqrt{s}\|\Delta^S_{i*}\|_2 ,$$

where the next-to-last inequality holds because $|\Omega_i| \le s$ by assumption on the sparsity of the row $S^\star_{i*}$. Next, by assumption of the theorem, $C^\top$ satisfies $\mathsf{RE}_{m,d}(c_1,c_2)$. We then have

$$\|C^\top \Delta^S_{i*}\|_2 \ge c_1\|\Delta^S_{i*}\|_2 - c_2\cdot\sqrt{\frac{\log(d)}{m}}\|\Delta^S_{i*}\|_1 \ge \left(c_1 - c_2\cdot 4\sqrt{s}\cdot\sqrt{\frac{\log(d)}{m}}\right)\|\Delta^S_{i*}\|_2 = c_0\|\Delta^S_{i*}\|_2 ,$$

where the last step uses the definition of c_0 in the theorem. (Recall that $c_0 > 0$ by assumption.) Summing over the rows, we then have

$$\|\Delta^S C\|_{\mathsf{F}}^2 = \sum_i \|C^\top \Delta^S_{i*}\|_2^2 \ge \sum_i c_0^2\|\Delta^S_{i*}\|_2^2 = c_0^2\|\Delta^S\|_{\mathsf{F}}^2 . \tag{2.16}$$

Now we return to (2.15) and plug in our result in (2.16), to obtain

$$\begin{aligned}&\frac{1}{2}\|\Delta^P\|_{\mathsf{F}}^2 + \frac{c_0^2}{2}\|\Delta^S\|_{\mathsf{F}}^2\\ &\qquad\le \nu(1.5\|\mathscr{P}_T(\Delta^P)\|_{\mathrm{nuc}} - 0.5\|\mathscr{P}_T^\perp(\Delta^P)\|_{\mathrm{nuc}}) + \lambda(1.5\|\mathscr{P}_\Omega(\Delta^S)\|_1 - 0.5\|\mathscr{P}_\Omega^\perp(\Delta^S)\|_1) .\end{aligned}$$

Removing negative terms from the right-hand side and multiplying by 2,

$$\|\Delta^P\|_{\mathsf{F}}^2 + c_0^2\|\Delta^S\|_{\mathsf{F}}^2 \leq 3\nu\|\mathscr{P}_T(\Delta^P)\|_{\text{nuc}} + 3\lambda\|\mathscr{P}_\Omega(\Delta^S)\|_1 \; .$$

Since $\text{rank}(\mathscr{P}_T(\Delta^P)) \leq 2r$ by definition of T, and similarly since $\|\mathscr{P}_\Omega(\Delta^S)\|_0 \leq sn$ by definition of Ω, we have

$$\begin{aligned}\|\Delta^P\|_{\mathsf{F}}^2 + c_0^2\|\Delta^S\|_{\mathsf{F}}^2 &\leq 3\nu\|\mathscr{P}_T(\Delta^P)\|_{\mathsf{F}} \cdot \sqrt{2r} + 3\lambda\|\mathscr{P}_\Omega(\Delta^S)\|_{\mathsf{F}} \cdot \sqrt{sn} \\ &\leq 3\nu\|\Delta^P\|_{\mathsf{F}} \cdot \sqrt{2r} + 3\lambda\|\Delta^S\|_{\mathsf{F}} \cdot \sqrt{sn} \\ &\leq \sqrt{\|\Delta^P\|_{\mathsf{F}}^2 + c_0^2\|\Delta^S\|_{\mathsf{F}}^2} \cdot \sqrt{18r\nu^2 + 9c_0^{-2}sn\lambda^2}\;,\end{aligned}$$

where the last step uses the Cauchy-Schwarz inequality. In particular, this implies that

$$\|\Delta^P\|_{\mathsf{F}}^2 + c_0^2\|\Delta^S\|_{\mathsf{F}}^2 \leq 18r\nu^2 + 9c_0^{-2}sn\lambda^2\;,$$

which proves the desired result.

□

Proof of Theorem 2.2.2. This result is a straightforward application of Theorem 2.2.1. It will be sufficient to check that, with the stated probability, the following statements all hold:

$$C^\top \text{ satisfies } \mathsf{RE}_{m,d}(c_1, c_2), \text{ with } c_0 := c_1 - c_2\sqrt{\frac{16s\log(d)}{m}} > 0, \tag{2.17}$$

and

$$\alpha \geq \|L^\star CC^\top\|_\infty,\; \nu \geq 2\|Z\|_{\text{sp}},\; \lambda \geq 2\|ZC^\top\|_\infty + 4\alpha. \tag{2.18}$$

To prove that (2.17) holds, the following lemma is sufficient (along with the assumption $m \geq c \cdot s\log(nd)$):

Lemma 2.4.1. *Under either the Gaussian model* (2.4) *or the orthogonal model* (2.5) *for the com-*

pression matrix C, for any $\delta > 0$, $C^\top$ satisfies $\mathsf{RE}_{m,d}(c_1, c_2)$ for constants

$$c_1 = \frac{1}{4(2+\sqrt{2})} \quad \textit{and} \quad c_2 = \frac{9}{2+\sqrt{2}}$$

with probability at least $1 - c'e^{-cm}$, where $c, c' > 0$ are universal constants.

To prove (2.18), we consider the first inequality by treating $L^\star$ as fixed and analysing the random model for C:

Lemma 2.4.2. *Under either the Gaussian model* (2.4) *or the orthogonal model* (2.5) *for C, for any fixed matrix $L^\star \in \mathbb{R}^{n\times d}$ and fixed $\delta > 0$, if $m \geq 16\log(nd)$, then*

$$\mathbb{P}\left\{ \|L^\star CC^\top\|_\infty > \|L^\star\|_\infty \cdot \left(1 + \sqrt{\frac{16d\log(nd)}{m}}\right)\right\} \leq \frac{4}{nd} \, .$$

For the second and third inequalities in (2.18), we first have the following bound on C:

Lemma 2.4.3. *Under either the Gaussian* (2.4) *or orthogonal* (2.5) *model for C, with probability at least $1 - 2de^{-m/8}$,*

$$\|C\|_{\mathrm{sp}} \leq \sqrt{12d/m} \textit{ and } \|C\|_{2,\infty} \leq 2 \, .$$

Next, we consider C as fixed and analyse the random model for the noise terms Z_{pre} and Z_{post} (we can treat C as fixed since the noise is generated independently from C). Fixing C, the rows of $Z = Z_{\mathrm{pre}}C + Z_{\mathrm{post}}$ are i.i.d. draws from the distribution $N(0, \sigma_{\mathrm{pre}}^2 C^\top C + \sigma_{\mathrm{post}}^2 \mathbf{I}_m)$. Then, writing $\Sigma = \sigma_{\mathrm{pre}}^2 C^\top C + \sigma_{\mathrm{post}}^2 \mathbf{I}_m$, we have

$$\|Z\|_{\mathrm{sp}} \leq \|Z\cdot\Sigma^{-1/2}\|_{\mathrm{sp}} \cdot \sqrt{\|\Sigma\|_{\mathrm{sp}}} \leq 3\sqrt{n+m}\cdot\sqrt{\|\Sigma\|_{\mathrm{sp}}} \, ,$$

with probability at least $1 - e^{-m}$, where the last step uses the fact that $Z \cdot \Sigma^{-1/2}$ is a $n \times m$ matrix with i.i.d. standard normal entries, and applies [25, Theorem II.13]. Furthermore,

$$\|\Sigma\|_{\mathrm{sp}} \leq \sigma_{\mathrm{pre}}^2 \|C\|_{\mathrm{sp}}^2 + \sigma_{\mathrm{post}}^2 \leq \sigma_{\max}^2 \cdot (12d/m + 1) \leq \left(4\sigma_{\max}\sqrt{d/m}\right)^2 \, ,$$

where the last step follows from Lemma 2.4.3. Combining these steps,

$$\|Z\|_{\mathrm{sp}} \leq 12\sigma_{\max}\sqrt{n+m}\cdot\sqrt{d/m}\,.$$

Next, we need to bound $\|ZC^\top\|_\infty$. Note that the entries are distributed as

$$(ZC^\top)_{ij} \sim N(0, \sigma_{\mathrm{pre}}^2(CC^\top CC^\top)_{jj} + \sigma_{\mathrm{post}}^2(CC^\top)_{jj})\,,$$

and this variance term is bounded as

$$\begin{aligned}\sigma_{\mathrm{pre}}^2(CC^\top CC^\top)_{jj} + \sigma_{\mathrm{post}}^2(CC^\top)_{jj} &= \sigma_{\mathrm{pre}}^2\mathbf{e}_j^\top CC^\top CC^\top\mathbf{e}_j + \sigma_{\mathrm{post}}^2\mathbf{e}_j^\top CC^\top\mathbf{e}_j\\ &\leq \sigma_{\max}^2\left(\|C\|_{\mathrm{sp}}^2+1\right)\|C^\top\mathbf{e}_j\|_2^2 \leq \sigma_{\max}^2\cdot(12d/m+1)\cdot 2^2 \leq \left(8\sigma_{\max}\sqrt{d/m}\right)^2\,,\end{aligned}$$

where the last step follows from Lemma 2.4.3. Therefore, using standard tail bounds on the normal distribution, with probability at least $1-\frac{2}{nd}$,

$$\|ZC^\top\|_\infty = \max_{ij}\left|(ZC^\top)_{ij}\right| \leq 8\sigma_{\max}\sqrt{d/m}\cdot 2\sqrt{\log(nd)}\,.$$

□

Proof of Theorem 2.2.3. This result is another immediate consequence of Theorem 2.2.1, with $\widetilde{S}^\star_\Omega$ in place of $S^\star$ (note that $\max_i\|(\widetilde{S}^\star_\Omega)_{i*}\|_0 \leq \max_i\|S^\star_{i*}\|_0 \leq s$ by assumption) . Since the restricted eigenvalue property and the condition $\alpha \geq \|L^\star CC^\top\|_\infty$ follow from Lemma 2.4.1 and Lemma 2.4.2 respectively, it is sufficient to check that, with the stated probability, the following statements both hold:

$$\nu \geq 2\|Z\|_{\mathrm{sp}},\ \lambda \geq 2\|ZC^\top\|_\infty + 4\alpha, \tag{2.19}$$

where $Z = (\widetilde{L}^\star_\Omega - L^\star)\cdot C$ as defined before. Let $B_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathsf{Bernoulli}(\rho_{ij})$ be an indicator variable for

$(i,j) \in \Omega$, that is, for whether we observe entry (i,j). Then we can write $\widetilde{L}^{\star}_{\Omega}$ as

$$(\widetilde{L}^{\star}_{\Omega})_{ij} = \frac{B_{ij}}{\rho_{ij}} \cdot L^{\star}_{ij}$$

for each $(i,j) \in [n] \times [d]$, and so Z can be written as

$$Z = \sum_{ij} \left(\frac{B_{ij}}{\rho_{ij}} - 1 \right) \cdot E_{ij} \tag{2.20}$$

where $E_{ij} = L^{\star}_{ij} \cdot \mathbf{e}_i C^{\top}_{j*} \in \mathbb{R}^{n \times m}$, and where $\mathbf{e}_i \in \mathbb{R}^n$ is the i-th standard basis vector and $C_{j*} \in \mathbb{R}^m$ is j-th row of the compression matrix C. To prove the first inequality in (2.19), we consider C as fixed and analyse the random model for B_{ij}'s. We first have the following bound on the sum of random scalars times fixed matrices:

Lemma 2.4.4 (Adapted from [26, Theorem 4.1.1]). *Let $A_1, \ldots, A_L \in \mathbb{R}^{d_1 \times d_2}$ be fixed matrices, and let $B_1, \ldots, B_L$ be independent mean-zero random variables, such that for each $\ell = 1, \ldots, L$, B_ℓ is σ^2-subgaussian, that is,*

$$\mathbb{E}\left[e^{tB_\ell} \right] \le e^{\sigma^2 t^2 / 2} \textit{ for all } t \in \mathbb{R} .$$

Then

$$\mathbb{P}\left\{ \left\| \sum_{\ell=1}^{L} B_\ell A_\ell \right\|_{\mathrm{sp}} \ge t \right\} \le (d_1 + d_2) \exp\left\{ - \frac{t^2}{2\sigma^2 \max\left\{ \|\sum_{\ell=1}^{L} A_\ell A_\ell^{\top}\|_{\mathrm{sp}}, \|\sum_{\ell=1}^{L} A_\ell^{\top} A_\ell\|_{\mathrm{sp}} \right\}} \right\} .$$

To apply Lemma 2.4.4 to the error term expression Z in (2.20), we first show that the random scalar, defined by

$$\widetilde{B}_{ij} = \frac{B_{ij}}{\rho_{ij}} - 1 ,$$

is σ^2-subgaussian with $\sigma^2 = 2\mu^2$ for all $(i,j) \in [n] \times [d]$. To see this, first note that $\mathbb{E}\left[\widetilde{B}_{ij} \right] = 0$

and $|\widetilde{B}_{ij}|$ is bounded by μ for all $(i,j) \in [n] \times [d]$. If $|t| \geq (2\mu)^{-1}$, then

$$\mathbb{E}\left[e^{t\widetilde{B}_{ij}}\right] \leq \mathbb{E}\left[e^{(2\mu^2t^2+\widetilde{B}_{ij}^2/2\mu^2)/2}\right] = e^{\mu^2t^2}\mathbb{E}\left[e^{\widetilde{B}_{ij}^2/4\mu^2}\right] \leq e^{\mu^2t^2}e^{1/4} \leq e^{2\mu^2t^2}$$

where the last inequality holds due to $|t| \geq (2\mu)^{-1}$. If $|t| \leq (2\mu)^{-1}$, we have $|t\widetilde{B}_{ij}| \leq 1/2$, and so

$$\mathbb{E}\left[e^{t\widetilde{B}_{ij}}\right] \leq 1 + t\mathbb{E}\left[\widetilde{B}_{ij}\right] + t^2\mathbb{E}\left[\widetilde{B}_{ij}^2\right] = 1 + t^2\mathbb{E}\left[\widetilde{B}_{ij}^2\right] \leq e^{t^2\mathbb{E}\left[\widetilde{B}_{ij}^2\right]} \leq e^{\mu^2t^2}$$

where the first inequality follows from the fact that $e^x \leq 1 + x + x^2$ for $|x| \leq 1/2$. Therefore, we apply Lemma 2.4.4 to the error term expression (2.20) so that, with probability at least $1 - \frac{1}{nd}$ (with respect to the randomness of the B_{ij}'s),

$$\|Z\|_{\mathrm{sp}} \leq \sqrt{4\mu^2 \max\left\{\|\sum_{ij} E_{ij}E_{ij}^\top\|_{\mathrm{sp}}, \|\sum_{ij} E_{ij}^\top E_{ij}\|_{\mathrm{sp}}\right\} \log\left(nd \cdot (n+m)\right)}\ .$$

Next, we derive the probabilistic bound on $\max\left\{\|\sum_{ij} E_{ij}E_{ij}^\top\|_{\mathrm{sp}}, \|\sum_{ij} E_{ij}^\top E_{ij}\|_{\mathrm{sp}}\right\}$. We first state the following bound on C:

Lemma 2.4.5. *Under either the Gaussian* (2.4) *or orthogonal* (2.5) *model for C, with probability at least* $1 - 2e^{-m}$*,*

$$\|C\|_{\mathrm{sp}} \leq \sqrt{12d/m} \text{ and } \|C\|_{\mathsf{F}} \leq \sqrt{3d}\ .$$

Direct calculation shows that

$$\|\sum_{ij} E_{ij}E_{ij}^\top\|_{\mathrm{sp}} = \max_i \left(\sum_{j=1}^d \|C_{j*}\|_2^2 {L_{ij}^\star}^2\right) \leq \alpha_0^2 \cdot \|C\|_{\mathsf{F}}^2$$

and

$$\|\sum_{ij} E_{ij}^\top E_{ij}\|_{\mathrm{sp}} = \|\sum_{ij} {L_{ij}^\star}^2 C_{j*}C_{j*}^\top\|_{\mathrm{sp}} \leq \alpha_0^2 \cdot n\|C\|_{\mathrm{sp}}^2.$$

Then, applying Lemma 2.4.5, with probability at least $1-2e^{-m}$,

$$\max\left\{\|\sum_{ij} E_{ij}E_{ij}^\top\|_{\mathrm{sp}}, \|\sum_{ij} E_{ij}^\top E_{ij}\|_{\mathrm{sp}}\right\} \leq \alpha_0^2 \max\left\{\|C\|_{\mathsf{F}}^2, n\|C\|_{\mathrm{sp}}^2\right\} \leq \alpha_0^2 \cdot 12\frac{d(n+m)}{m}\ .$$

In total, we have with probability at least $1-\frac{2}{nd}$,

$$\|Z\|_{\mathrm{sp}} \leq \mu\alpha_0\sqrt{48\frac{d(n+m)}{m}\log\left(nd(n+m)\right)}\ .$$

Since $m \leq d$, we can write $\log(nd(n+m)) \leq \log(nd(n+d)) \leq \max\{\log(2n^2d), \log(2nd^2)\} \leq 2\log(nd)$, where we assume $n, d \geq 2$ to avoid triviality. So,

$$\|Z\|_{\mathrm{sp}} \leq 10\mu\alpha_0\sqrt{\frac{d(n+m)}{m}\log(nd)}\ . \tag{2.21}$$

Next, we need to bound on $\|ZC^\top\|_\infty$. Note that

$$\|ZC^\top\|_\infty = \|(\widetilde{L}_\Omega^\star - L^\star)CC^\top\|_\infty \leq \|\widetilde{L}_\Omega^\star - L^\star\|_\infty + \|(\widetilde{L}_\Omega^\star - L^\star)(CC^\top - \mathbf{I}_d)\|_\infty\ .$$

By our assumptions, we can immediately bound $\|\widetilde{L}_\Omega^\star - L^\star\|_\infty \leq \mu\alpha_0$. Next consider the term $\|(\widetilde{L}_\Omega^\star - L^\star)(CC^\top - \mathbf{I}_d)\|_\infty$. We first consider C as fixed and analyse the random model for B_{ij}'s. The (i,ℓ)-th entry of $(\widetilde{L}_\Omega^\star - L^\star)(CC^\top - \mathbf{I}_d)$ can be written as

$$\left[(\widetilde{L}_\Omega^\star - L^\star)(CC^\top - \mathbf{I}_d)\right]_{i\ell} = \sum_j \widetilde{B}_{ij}\cdot L_{ij}^\star(CC^\top - \mathbf{I}_d)_{j\ell}\ ,$$

which is mean zero random scalar and bounded above by $\mu\alpha_0\|CC^\top - \mathbf{I}_d\|_\infty$. Therefore, applying Hoeffding's Lemma and union bound, with probability at least $1-\frac{1}{nd}$ (with respect to the randomness of the B_{ij}'s),

$$\|(\widetilde{L}_\Omega^\star - L^\star)(CC^\top - \mathbf{I}_d)\|_\infty \leq \sqrt{2d(\alpha_0+\alpha_1)^2\mu^2\|CC^\top - \mathbf{I}_d\|_\infty^2\log\left(2n^2d^2\right)}\ . \tag{2.22}$$

For the bound on $\|CC^\top - \mathbf{I}_d\|_\infty$, we have the following result:

Lemma 2.4.6. *Under either the Gaussian* (2.4) *or orthogonal* (2.5) *model for C, with probability at least* $1-\frac{4}{nd}$*,*

$$\|CC^\top - \mathbf{I}_d\|_\infty \le \sqrt{\frac{24\log(nd)}{m}} .$$

Combining (2.22) with Lemma 2.4.6, we have with probability at least $1-\frac{5}{nd}$,

$$\|ZC^\top\|_\infty \le 7\mu\alpha_0\sqrt{\frac{d\log(nd)\log(2n^2d^2)}{m}} \le 12\mu\alpha_0\sqrt{\frac{d\log^2(nd)}{m}} .$$

□

2.4.3 Concentration lemma

We first state a concentration result under the Gaussian model (2.4) or the orthogonal model (2.5):

Lemma 2.4.7. *Under either the Gaussian model* (2.4) *or the orthogonal model* (2.5)*, for any fixed vector* $w \in \mathbb{R}^d$ *and any* $\varepsilon > 0$*,*

$$\mathbb{P}\left\{\frac{\|C^\top w\|_2^2}{\|w\|_2^2} - 1 > \varepsilon\right\} \le \exp\left\{-\frac{m}{8}\cdot\min\{\varepsilon,\varepsilon^2\}\right\}, \mathbb{P}\left\{\frac{\|C^\top w\|_2^2}{\|w\|_2^2} - 1 < -\varepsilon\right\} \le \exp\left\{-\frac{m}{4}\varepsilon^2\right\} . \tag{2.23}$$

Proof. Under the Gaussian model,

$$m\cdot\frac{\|C^\top w\|_2^2}{\|w\|_2^2} \sim \chi_m^2$$

and therefore, by the χ^2 tail bounds of [27, Lemma 1], for any $t > 0$,

$$\mathbb{P}\left\{m\cdot\frac{\|C^\top w\|_2^2}{\|w\|_2^2} > m + 2\sqrt{mt} + 2t\right\} \le e^{-t} \text{ and } \mathbb{P}\left\{m\cdot\frac{\|C^\top w\|_2^2}{\|w\|_2^2} < m - 2\sqrt{mt}\right\} \le e^{-t} .$$

Setting $t = \frac{m}{8}\cdot\min\{\varepsilon,\varepsilon^2\}$, we obtain the desired result (2.23). Next, turning to the orthogonal model, we have $G = \sqrt{\frac{d}{m}}\cdot U$ where $U \in \mathbb{R}^{d\times m}$ is an orthonormal matrix chosen uniformly at random. Let $v \in \mathbb{R}^d$ be a random unit vector. Then $\|U^\top w\|_2^2$ is equal in distribution to $v_1^2 + \cdots + v_m^2$. In this setting, [28, Lemma 2.4] states that, for any $0 < \beta_0 < 1$,

$$\mathbb{P}\left\{v_1^2 + \cdots + v_m^2 < \beta_0 \frac{m}{d}\right\} \leq \exp\left\{\frac{m}{2}(1 - \beta_0 + \log(\beta_0))\right\}$$

and for any $\beta_1 > 1$,

$$\mathbb{P}\left\{v_1^2 + \cdots + v_m^2 > \beta_1 \frac{m}{d}\right\} \leq \exp\left\{\frac{m}{2}(1 - \beta_1 + \log(\beta_1))\right\} .$$

Next, set $\beta_1 = 1 + \varepsilon$. Then, since for all $x > 0$ we have $\log(1+x) \leq x - \frac{\min\{x, x^2\}}{4}$, then

$$1 - \beta_1 + \log(\beta_1) \leq 1 - (1+\varepsilon) + \varepsilon - \frac{\min\{\varepsilon, \varepsilon^2\}}{4} = -\frac{\min\{\varepsilon, \varepsilon^2\}}{4} .$$

Therefore,

$$\mathbb{P}\left\{\frac{\|C^\top w\|_2^2}{\|w\|_2^2} > 1 + \varepsilon\right\} \leq \exp\left\{-\frac{m}{8} \cdot \min\{\varepsilon, \varepsilon^2\}\right\} .$$

Next we want to bound the probability of the event $\frac{\|C^\top w\|_2^2}{\|w\|_2^2} < 1 - \varepsilon$. If $\varepsilon \geq 1$ then trivially this cannot occur. If instead $\varepsilon < 1$, then we set $\beta_0 = 1 - \varepsilon$. Since $\log(1-x) \leq -x - \frac{x^2}{2}$ for all $0 < x < 1$, we have

$$1 - \beta_0 + \log(\beta_0) = 1 - (1-\varepsilon) - \varepsilon - \frac{\varepsilon^2}{2} = -\frac{\varepsilon^2}{2} ,$$

and so

$$\mathbb{P}\left\{\frac{\|C^\top w\|_2^2}{\|w\|_2^2} < 1 - \varepsilon\right\} \leq \exp\left\{-\frac{m}{4} \cdot \varepsilon^2\right\} .$$

This is sufficient to prove the desired bound. □

2.4.4 Proofs of supporting lemmas

Proof of Lemma 2.4.2. Set $\varepsilon = \sqrt{\frac{16\log(nd)}{m}}$ and note that $\varepsilon \leq 1$ by assumption. For each $i \in [n]$, define the unit vector $v_i = \frac{L^\star_{i*}}{\|L^\star_{i*}\|_2}$ (treated as a column vector). Now fix any $i \in [n]$ and any $j \in [d]$.

Then

$$\left(L^\star CC^\top\right)_{ij} = \|L^\star_{i*}\|_2 \cdot v_i^\top CC^\top \mathbf{e}_j = \|L^\star_{i*}\|_2 \cdot \frac{1}{4}\left(\|C^\top(v_i+\mathbf{e}_j)\|_2^2 - \|C^\top(v_i-\mathbf{e}_j)\|_2^2\right) \ .$$

By Lemma 2.4.7, with probability at least $1-4e^{-m\varepsilon^2/8}$,

$$\left|\frac{\|C^\top(v_i+\mathbf{e}_j)\|_2^2}{\|v_i+\mathbf{e}_j\|_2^2} - 1\right| \le \varepsilon \quad \text{and} \quad \left|\frac{\|C^\top(v_i-\mathbf{e}_j)\|_2^2}{\|v_i-\mathbf{e}_j\|_2^2} - 1\right| \le \varepsilon \ .$$

If these bounds hold, then

$$\begin{aligned}
\left(L^\star CC^\top\right)_{ij} &= \|L^\star_{i*}\|_2 \cdot \frac{1}{4}\left(\|C^\top(v_i+\mathbf{e}_j)\|_2^2 - \|C^\top(v_i-\mathbf{e}_j)\|_2^2\right) \\
&\le \|L^\star_{i*}\|_2 \cdot \frac{1}{4}\left((1+\varepsilon)\cdot\|v_i+\mathbf{e}_j\|_2^2 - (1-\varepsilon)\cdot\|v_i-\mathbf{e}_j\|_2^2\right) \\
&= \|L^\star_{i*}\|_2 \cdot \frac{1}{4}\left(\left(\|v_i+\mathbf{e}_j\|_2^2 - \|v_i-\mathbf{e}_j\|_2^2\right) + \varepsilon\left(\|v_i+\mathbf{e}_j\|_2^2 + \|v_i-\mathbf{e}_j\|_2^2\right)\right) \\
&= \|L^\star_{i*}\|_2 \cdot \frac{1}{4}\left(4\langle v_i,\mathbf{e}_j\rangle + \varepsilon\left(2\|v_i\|_2^2 + 2\|\mathbf{e}_j\|_2^2\right)\right) \\
&= \|L^\star_{i*}\|_2 \cdot \left(\langle v_i,\mathbf{e}_j\rangle + \varepsilon\right) \quad \text{since } \|v_i\|_2 = \|\mathbf{e}_j\|_2 = 1 \\
&= \langle L^\star_{i*},\mathbf{e}_j\rangle + \varepsilon\|L^\star_{i*}\|_2 \quad \text{by definition of } v_i \\
&= L^\star_{ij} + \varepsilon\|L^\star_{i*}\|_2 \\
&\le \|L^\star\|_\infty\left(1+\varepsilon\sqrt{d}\right) \ .
\end{aligned}$$

Using the same arguments, the same bound holds for $-(L^\star CC^\top)_{ij}$, and therefore,

$$\left|(L^\star CC^\top)_{ij}\right| \le \|L^\star\|_\infty\left(1+\varepsilon\sqrt{d}\right) \ .$$

Applying the union bound over each $i \in [n]$ and each $j \in [d]$, we see that

$$\|L^\star CC^\top\|_\infty \le \|L^\star\|_\infty(1+\varepsilon\sqrt{d})$$

with probability at least

$$1 - nd \cdot 4e^{-m\varepsilon^2/8} = 1 - 4nd \exp\left\{ \frac{m}{8} \left(\sqrt{\frac{16\log(nd)}{m}} \right)^2 \right\} = 1 - \frac{4}{nd} \, .$$

□

Proof of Lemma 2.4.3. First we treat $\|C\|_{\mathrm{sp}}$. Under the orthogonal model, $\|C\|_{\mathrm{sp}} \le \sqrt{d/m}$ trivially, while under the Gaussian model for C (2.4), $\|C\|_{\mathrm{sp}} \le \sqrt{d/m}(2+\sqrt{2})$ with probability at least $1-e^{-m}$ by again applying [25, Theorem II.13]. Next consider $\|C\|_{2,\infty} = \max_{i=1,\dots,d} \|C^\top \mathbf{e}_i\|_2$. For each i, by Lemma 2.4.7,

$$\mathbb{P}\left\{ \|C^\top \mathbf{e}_j\|_2 > 2 \right\} \le e^{-m/8} \, .$$

Therefore,

$$\mathbb{P}\left\{ \|C\|_{2,\infty} > 2 \right\} \le d \cdot e^{-m/8} \, .$$

□

Proof of Lemma 2.4.4. [26, Theorem 4.1.1] proves this exact statement for the special case that either $B_\ell \overset{\text{iid}}{\sim} N(0,1)$ (Gaussian variables) or $B_\ell \overset{\text{iid}}{\sim} \{\pm 1\}$ (Rademacher variables). To see why the statement holds in this more general case, we observe that for Corollary 4.2 in Tropp, the distribution of the B_ℓ's is used only once: to prove the bound

$$\mathbb{E}\left[e^{tB_\ell A} \right] \preceq e^{t^2 A^2/2}$$

for each ℓ and for any fixed Hermitian matrix A. For the general case, take a fixed Hermitian matrix A, with $A = Q\Lambda Q^\top$ its eigendecomposition. We have

$$\begin{aligned}
\mathbb{E}\left[e^{tB_\ell A}\right] &= \mathbb{E}\left[e^{Q\cdot(tB_\ell\Lambda)\cdot Q^\top}\right] \\
&= Q\cdot \mathrm{diag}\{\mathbb{E}\left[e^{tB_\ell\lambda_i}\right]\}\cdot Q^\top \\
&\preceq Q\cdot \mathrm{diag}\{e^{\sigma^2t^2\lambda_i^2/2}\}\cdot Q^\top \\
&= e^{Q\cdot(\sigma^2t^2\Lambda^2/2)\cdot Q^\top} \\
&= e^{\sigma^2t^2A^2/2}\ .
\end{aligned}$$

Therefore, this is sufficient to see that Corollary 4.2 of Tropp holds in this case also. □

Proof of Lemma 2.4.5. The result for $\|C\|_{\text{sp}}$ follows from Lemma 2.4.3. Next consider $\|C\|_{\mathsf{F}}^2$. Under the orthogonal model, $\|C\|_{\mathsf{F}}^2 = \mathsf{tr}(C^\top C) = d$ holds. Under the Gaussian model for C, we note that $\|C\|_{\mathsf{F}}^2 \sim \chi^2_{md}/m$. By the χ^2 tail bounds of [27, Lemma 1], we have

$$\mathbb{P}\left\{\|C\|_{\mathsf{F}}^2 \geq d + 2\sqrt{d} + 2\right\} \leq e^{-m}\ .$$

Since $3d \geq d + 2\sqrt{d} + 2$ for $d \geq 1$, with probability at least $1 - e^{-m}$, we have $\|C\|_{\mathsf{F}}^2 \leq 3d$. □

Proof of Lemma 2.4.6. This result is the consequence of Lemma 2.4.7 and union bound. Set $\varepsilon = \sqrt{\frac{24\log(nd)}{m}}$ and $\varepsilon \leq 1$. By Lemma 2.4.7, with probability at least $1 - 2e^{-m\varepsilon^2/8}$, for $i \neq j$,

$$\begin{aligned}
(CC^\top - \mathbf{I}_d)_{ij} = \mathbf{e}_i^\top(CC^\top - \mathbf{I}_d)\mathbf{e}_j &= \frac{1}{4}(\|C^\top(\mathbf{e}_i + \mathbf{e}_j)\|_2^2 - \|C^\top(\mathbf{e}_i - \mathbf{e}_j)\|_2^2) \\
&\leq \frac{1}{4}((1+\varepsilon)\|\mathbf{e}_i + \mathbf{e}_j\|_2^2 - (1-\varepsilon)\|\mathbf{e}_i - \mathbf{e}_j\|_2^2) \leq \varepsilon\ .
\end{aligned}$$

The same bound holds for $-(CC^\top - \mathbf{I}_d)_{ij}$ if we use the same arguments, and so with probability at least $1 - 4e^{-m\varepsilon^2/8}$,

$$|(CC^\top - \mathbf{I}_d)_{ij}| \leq \varepsilon\ .$$

For $i = j$, applying Lemma 2.4.7 again, with probability at least $1 - 2e^{-m\varepsilon^2/8}$,

$$|(CC^\top - \mathbf{I}_d)_{ij}| = |\mathbf{e}_i^\top (CC^\top - \mathbf{I}_d)\mathbf{e}_j| = |\|C^\top \mathbf{e}_j\|_2^2 - 1| \le \varepsilon \ .$$

Applying the union bound over each $(i, j) \in [d] \times [d]$, we have that

$$\|CC^\top - \mathbf{I}_d\|_\infty \le \sqrt{\frac{24\log(nd)}{m}}$$

with probability at least $1 - 4d^2 e^{-m\varepsilon^2/8} \ge 1 - \frac{4}{nd}$. □

Proof of Lemma 2.4.1 (restricted strong convexity). First, for the Gaussian model (2.4), by [29, Theorem 1], for universal constants $c, c' > 0$,

$$\mathbb{P}\left\{ \|C^\top x\|_2 \ge \frac{1}{4}\|x\|_2 - 9\sqrt{\frac{\log(d)}{m}}\|x\|_1 \text{ for all } x \in \mathbb{R} \right\} \ge 1 - c'e^{-cm} \ .$$

Next, we turn to the orthogonal model (2.5). Let $H \in \mathbb{R}^{d\times m}$ be a matrix with $H_{ij} \overset{\text{iid}}{\sim} N(0, 1/m)$, let $H = UDV^\top$ be its singular value decomposition, and without loss of generality take $C = \sqrt{\frac{d}{m}} \cdot U$ (since H is rotation invariant and so U is uniformly distributed over the space of uniform matrices, this satisfies the orthogonal model (2.5)). Then for any $x \in \mathbb{R}^d$,

$$\|H^\top x\|_2^2 = \|VDU^\top x\|_2^2 \le \|VD\|_{\text{sp}}^2 \|U^\top x\|_2^2 = \|H\|_{\text{sp}}^2 \cdot \frac{m}{d} \cdot \|C^\top x\|_2^2 \ .$$

By the work above for the Gaussian model, with probability at least $1 - c'e^{-cm}$,

$$\|H^\top x\|_2 \ge \frac{1}{4}\|x\|_2 - 9\sqrt{\frac{\log(d)}{m}}\|x\|_1 \text{ for all } x \in \mathbb{R}^d \ ,$$

and by [25, Theorem II.13], with probability at least $1 - e^{-m}$,

$$\|H\|_{\text{sp}} \le \sqrt{\frac{d}{m}} + 1 + \sqrt{\frac{2m}{d}} \le \sqrt{\frac{d}{m}}\left(2 + \sqrt{2}\right) \ .$$

Combining all these bounds, with probability at least $1 - c'e^{-cm} - e^{-m} \geq 1 - (c'+1)e^{-\min\{c,1\}\cdot m}$, for all $x \in \mathbb{R}^d$,

$$\|C^\top x\|_2 \geq \frac{1}{4(2+\sqrt{2})}\|x\|_2 - \frac{9}{2+\sqrt{2}}\sqrt{\frac{\log(d)}{m}}\|x\|_1 .$$

Clearly, this statement holds also for the Gaussian model as well (since this is a strictly weaker result than the one stated above.)

□

CHAPTER 3

CONCAVITY COEFFICIENTS FOR A NONCONVEX SPACE

A convex relaxation is a common technique in many applications with high-dimensional data, and enables consistent estimation of signals from fewer measurements than the ambient dimension. Examples include the ℓ_1 norm and nuclear norm as a proxy to sparsity and low rank, which we have seen in the compressed RPCA setting of Chapter 2. While estimation via convex relaxations often enjoys near-optimal sample complexity and global convergence guarantee [30], there also exists a well-known tradeoff between shrinkage and bias in the accuracy in which convex relaxations lead to increased bias of the large signals. Nonconvex optimization can avoid such loss of accuracy, but now we are faced with the possibility of becoming trapped in a local minimum or failing to converge.

As a first step to study nonconvex optimization algorithm over multiple signals, this chapter[1] explores local geometric properties of a nonconvex constraint set $\mathscr{C}$ and develops *local concavity coefficients*, characterizing the extent to which $\mathscr{C}$ is nonconvex relative to each of its points. These coefficients, a generalization of the notion of *prox-regular sets* in the analysis literature, bound the set's violations of four different characterizations of convexity—e.g. convex combinations of points must lie in the set, and the first-order optimality conditions for minimization over the set—with respect to a structured norm, such as the ℓ_1 norm for sparse problems, chosen to capture the natural structure of the problem. As we will see later on, these multiple notions of nonconvexity are in fact exactly equivalent. The local concavity coefficient allow us to characterize the geometric properties of the constraint set $\mathscr{C}$ that are favorable for analyzing the convergence of projected gradient descent.

[1] The work presented in this chapter is published in Barber and Ha [31].

3.1 Global concavity coefficients

Consider the constraint set $\mathscr{C} \subset \mathbb{R}^d$, and we quantify the concavity of $\mathscr{C}$ by describing the extent to which the constraint set $\mathscr{C}$ deviates from convexity. Concretely, we consider four properties that would hold if $\mathscr{C}$ were convex, and define the (global) concavity coefficient of $\mathscr{C}$, denoted $\gamma = \gamma(\mathscr{C})$, to characterize the extent to which these properties are violated. Since we are interested in developing flexible tools for high-dimensional optimization problems, several different norms will appear in the definitions of the concavity coefficients:

- The Euclidean ℓ_2 norm, $\|\cdot\|_2$. Projections to $\mathscr{C}$ will always be taken with respect to the ℓ_2 norm. If our variable is a matrix $X \in \mathbb{R}^{n\times m}$, the Euclidean ℓ_2 norm is known as the Frobenius norm, $\|X\|_{\mathsf{F}} = \sqrt{\sum_{ij} X_{ij}^2}$.

- A "structured" norm $\|\cdot\|$, which can be chosen to be any norm on $\mathbb{R}^d$. In some cases it may be the ℓ_2 norm, but often it will be a different norm reflecting natural structure in the problem. For instance, for a low-rank estimation problem, if $\mathscr{C}$ is a set of rank-constrained matrices then we will work with the nuclear norm, $\|\cdot\| = \|\cdot\|_{\mathsf{nuc}}$. For sparse signals, we will instead use the ℓ_1 norm, $\|\cdot\| = \|\cdot\|_1$.

- A norm $\|\cdot\|^*$, which is the dual norm to the structured norm $\|\cdot\|$. For low-rank matrix problems, if we work with the nuclear norm, $\|\cdot\| = \|\cdot\|_{\mathsf{nuc}}$, then the dual norm is given by the spectral norm, $\|\cdot\|^* = \|\cdot\|_{\mathsf{sp}}$. For sparse problems, if $\|\cdot\| = \|\cdot\|_1$ then its dual is given by the ℓ_∞ norm, $\|\cdot\|^* = \|\cdot\|_\infty$.

When we take projections to the constraint set $\mathscr{C}$, if the minimizer $\mathscr{P}_{\mathscr{C}}(z) \in \arg\min_{x\in\mathscr{C}} \|x - z\|_2$ is non-unique, then we write $\mathscr{P}_{\mathscr{C}}(z)$ to denote any point chosen from this set. We will assume without comment that $\mathscr{C}$ is closed and nonempty so that the set $\arg\min_{x\in\mathscr{C}} \|x-z\|_2$ is nonempty for any z. In the following, we present several definitions of the concavity coefficient of $\mathscr{C}$.

Curvature First, we define γ as a bound on the extent to which a convex combination of two elements of $\mathscr{C}$ may lie outside of $\mathscr{C}$: for $x, y \in \mathscr{C}$,

$$\limsup_{t \searrow 0} \frac{\min_{z \in \mathscr{C}} \|z - ((1-t)x + ty)\|}{t} \leq \gamma \|x - y\|_2^2. \tag{3.1}$$

Approximate contraction Second, we define γ via a condition requiring that the projection operator $\mathscr{P}_{\mathscr{C}}$ is approximately contractive in a neighborhood of the set $\mathscr{C}$, that is, $\|\mathscr{P}_{\mathscr{C}}(z) - \mathscr{P}_{\mathscr{C}}(w)\|_2$ is not much larger than $\|z - w\|_2$: for $x, y \in \mathscr{C}$,

For any $z, w \in \mathbb{R}^d$ with $\mathscr{P}_{\mathscr{C}}(z) = x$ and $\mathscr{P}_{\mathscr{C}}(w) = y$,

$$\big(1 - \gamma \|z - x\|^* - \gamma \|w - y\|^*\big) \cdot \|x - y\|_2 \leq \|z - w\|_2. \tag{3.2}$$

For convenience in our theoretical analysis we will also consider a weaker "one-sided" version of this property, where one of the two points is assumed to already lie in $\mathscr{C}$: for $x, y \in \mathscr{C}$,

$$\text{For any } z \in \mathbb{R}^d \text{ with } \mathscr{P}_{\mathscr{C}}(z) = x, \quad (1 - \gamma \|z - x\|^*) \cdot \|x - y\|_2 \leq \|z - y\|_2. \tag{3.3}$$

First-order optimality For our third characterization of the concavity coefficient, we consider the standard first-order optimality conditions for minimization over a convex set, and measure the extent to which they are violated when optimizing over $\mathscr{C}$:[2] for $x, y \in \mathscr{C}$,

For any differentiable $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$ such that x is a local minimizer of f over $\mathscr{C}$,

$$\langle y - x, \nabla \mathsf{f}(x) \rangle \geq -\gamma \|\nabla \mathsf{f}(x)\|^* \|y - x\|_2^2. \tag{3.4}$$

[2] A more general form of this condition, with f Lipschitz but not necessarily differentiable, appears in (3.32) (see Section 3.5.3 for further details).

Inner products Fourth, we introduce an inner product condition, requiring that projection to the constraint set $\mathscr{C}$ behaves similarly to a convex projection: for $x,y \in \mathscr{C}$,

$$\text{For any } z \in \mathbb{R}^d \text{ with } \mathscr{P}_{\mathscr{C}}(z) = x, \quad \langle y-x, z-x \rangle \leq \gamma \|z-x\|^* \|y-x\|_2^2. \tag{3.5}$$

We emphasize the distinction between the structured norm $\|\cdot\|$ (and its dual norm $\|\cdot\|^*$) and the ℓ_2 norm $\|\cdot\|_2$ in the definitions of the concavity coefficients. We will see later that, by choosing $\|\cdot\|$ to reflect the structure in the signal (rather than working only with the ℓ_2 norm), we are able to obtain a more favorable scaling in our concavity coefficients, and hence to prove meaningful convergence results in high-dimensional settings. On the other hand, regardless of our choice of $\|\cdot\|$, note that the ℓ_2 norm also appears in the definition of the concavity coefficients, as is natural when working with inner products (recall the projection operator $\mathscr{P}_{\mathscr{C}}(\cdot)$ is defined with respect to the ℓ_2 norm).

Now we show that the above conditions are in fact exactly equivalent:

Theorem 3.1.1. *The properties* (3.1)*,* (3.2)*,* (3.3)*,* (3.4)*, and* (3.5) *are equivalent; that is, for a fixed choice $\gamma \in [0,\infty]$, they either all hold for every $x,y \in \mathscr{C}$, or all fail to hold for some $x,y \in \mathscr{C}$.*

Formally, we will define $\gamma(\mathscr{C})$ to be the smallest value such that the above properties hold:

$$\gamma(\mathscr{C}) := \min\left\{\gamma \in [0,\infty] : \text{Properties (3.1), (3.2), (3.3), (3.4), (3.5) hold for all } x,y \in \mathscr{C}\right\}.$$

This global coefficient $\gamma(\mathscr{C})$ is often of limited use in practical settings, since many sets are well-behaved locally but not globally. For instance, the set $\mathscr{C} = \{X \in \mathbb{R}^{n\times m} : \text{rank}(X) \leq r\}$ has $\gamma(\mathscr{C}) = \infty$, but exhibits smooth curvature as long as we stay away from rank-degenerate matrices (that is, matrices with $\text{rank}(X) < r$). This motivates us to expand to a local version of the same concavity bounds.

3.2 Local concavity coefficients

Next we consider the *local concavity coefficients* $\gamma_x(\mathscr{C})$, measuring the concavity in a set $\mathscr{C}$ relative to a specific point x in the set. First we define a set of "degenerate points",

$$\mathscr{C}_{\mathsf{dgn}} = \{x \in \mathscr{C} : \ \mathscr{P}_{\mathscr{C}} \text{ is not continuous over any neighborhood of } x\},$$

and then let

$$\gamma_x(\mathscr{C}) = \begin{cases} \infty, & x \in \mathscr{C}_{\mathsf{dgn}}, \\ \min\{\gamma \in [0,\infty] : \text{Property (*) holds for this point } x \text{ and any } y \in \mathscr{C}\}, & x \not\in \mathscr{C}_{\mathsf{dgn}}, \end{cases} \tag{3.6}$$

where the property (*) may refer to any of the four definitions of the concavity coefficients,[3] namely (3.1), (3.3), (3.4), or (3.5). We will see shortly why it is necessary to make an exception for the degenerate points $x \in \mathscr{C}_{\mathsf{dgn}}$ in the definition of these coefficients.

We show that the equivalence between the four properties (3.1), (3.3), (3.4), and (3.5) in terms of the global concavity coefficient $\gamma(\mathscr{C})$, holds also for the local coefficients:

Theorem 3.2.1. *For all $x \in \mathscr{C}$, the definition* (3.6) *of $\gamma_x(\mathscr{C})$ is equivalent for all four choices of the property* (*)*, namely the conditions* (3.1)*,* (3.3)*,* (3.4)*, or* (3.5).

To develop an intuition for the global and local concavity coefficients, we give a simple example in $\mathbb{R}^2$ (relative to the ℓ_2 norm, i.e. $\|\cdot\| = \|\cdot\|^* = \|\cdot\|_2$), displayed in Figure 3.1. Define $\mathscr{C} = \{x \in \mathbb{R}^2 : x_1 \leq 0 \text{ or } x_2 \leq 0\}$. Due to the degenerate point $x = (0,0)$, we can see that $\gamma(\mathscr{C}) = \infty$ in this

[3] In this definition, we only consider the "one-sided" formulation (3.3) of the contraction property, since the two-sided formulation (3.2) would involve the local concavity coefficient at both x and y due to symmetry—we will see in Lemma 3.2.4 below that a version of the two-sided contraction property still holds using local coefficients.

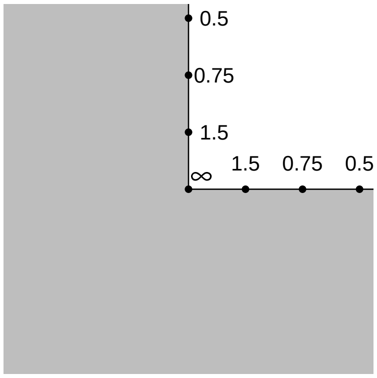

Figure 3.1: A simple example of the local concavity coefficients on $\mathscr{C} = \{x \in \mathbb{R}^2 : x_1 \leq 0 \text{ or } x_2 \leq 0\}$. The gray shaded area represents $\mathscr{C}$ while the numbers give the local concavity coefficients at each marked point.

case. The local concavity coefficients are given by

$$\begin{cases} \gamma_x(\mathscr{C}) = \infty, & \text{if } x = (0,0), \\ \gamma_x(\mathscr{C}) = \frac{1}{2t}, & \text{if } x = (t,0) \text{ or } (0,t) \text{ for } t > 0, \\ \gamma_x(\mathscr{C}) = 0, & \text{if } x_1 < 0 \text{ or } x_2 < 0. \end{cases}$$

Note that at the degenerate point $x = (0,0)$, $\mathscr{C}$ actually contains all convex combinations of this point x with any $y \in \mathscr{C}$, and so the curvature condition (3.1) is satisfied with $\gamma = 0$. However, $x \in \mathscr{C}_{\mathsf{dgn}}$, so we nonetheless set $\gamma_x(\mathscr{C}) = \infty$.

Practical high-dimensional examples, such as a rank constraint, will be discussed in depth in Section 3.3. For example we will see that, for the rank-constrained set $\mathscr{C} = \{X \in \mathbb{R}^{n \times m} : \text{rank}(X) \leq r\}$, the local concavity coefficients satisfy $\gamma_X(\mathscr{C}) = \frac{1}{2\sigma_r(X)}$ relative to the nuclear norm.

In general, a rough intuition for the local coefficients is that:

- If x lies in the interior of $\mathscr{C}$, or if $\mathscr{C}$ is convex, then $\gamma_x(\mathscr{C}) = 0$;
- If x lies on the boundary of $\mathscr{C}$, which is a nonconvex set with a smooth boundary, then we will typically see a finite but nonzero $\gamma_x(\mathscr{C})$;
- $\gamma_x(\mathscr{C}) = \infty$ can indicate a nonconvex cusp or other degeneracy at the point x.

3.2.1 Properties

We examine some properties of the local coefficients $\gamma_x(\mathscr{C})$ that will be useful for gaining intuition for these coefficients.

First, the global and local coefficients are related in the natural way:

Lemma 3.2.1. *For any $\mathscr{C}$, $\gamma(\mathscr{C}) = \sup_{x\in\mathscr{C}} \gamma_x(\mathscr{C})$.*

Next, observe that $x \mapsto \gamma_x(\mathscr{C})$ is not continuous in general (in particular, since $\gamma_x(\mathscr{C}) = 0$ in the interior of $\mathscr{C}$ but is often positive on the boundary). However, this map does satisfy upper semi-continuity:

Lemma 3.2.2. *The function $x \mapsto \gamma_x(\mathscr{C})$ is upper semi-continuous over $x \in \mathscr{C}$.*

Furthermore, setting $\gamma_x(\mathscr{C}) = \infty$ at the degenerate points $x \in \mathscr{C}_{\mathsf{dgn}}$ is natural in the following sense: the resulting map $x \mapsto \gamma_x(\mathscr{C})$ is the minimal upper semi-continuous map such that the relevant local concavity properties are satisfied. We formalize this with the following lemma:

Lemma 3.2.3. *For any $u \in \mathscr{C}_{\mathsf{dgn}}$, for any of the four conditions,* (3.1)*,* (3.3)*,* (3.4)*, or* (3.5)*, this property does not hold in any neighborhood of u for any finite γ. That is, for any $r > 0$,*

$$\min\Big\{\gamma \geq 0 : \textit{ Property (*) holds for all } x \in \mathscr{C} \cap \mathbb{B}_2(u,r) \textit{ and for all } y \in \mathscr{C}\Big\} = \infty,$$

where (*) may refer to any of the four equivalent properties, i.e. (3.1), (3.3), (3.4), and (3.5). (Here $\mathbb{B}_2(u,r)$ is the ball of radius r around the point u, with respect to the ℓ_2 norm.)

Finally, the next result shows that two-sided contraction property (3.2) holds using local coefficients, meaning that all five definitions of concavity coefficients are equivalent:

Lemma 3.2.4. *For any $z, w \in \mathbb{R}^d$,*

$$\left(1 - \gamma_{\mathscr{P}_{\mathscr{C}}(z)}(\mathscr{C})\|z - \mathscr{P}_{\mathscr{C}}(z)\|^* - \gamma_{\mathscr{P}_{\mathscr{C}}(w)}(\mathscr{C})\|w - \mathscr{P}_{\mathscr{C}}(w)\|^*\right) \cdot \|\mathscr{P}_{\mathscr{C}}(z) - \mathscr{P}_{\mathscr{C}}(w)\|_2 \leq \|z - w\|_2$$

In particular, for any fixed $c \in (0,1)$, Lemma 3.2.4 proves that

$$\mathscr{P}_{\mathscr{C}} \text{ is } c\text{-Lipschitz over the set } \Big\{ z \in \mathbb{R}^d : 2\gamma_{\mathscr{P}_{\mathscr{C}}(z)}(\mathscr{C}) \|z - \mathscr{P}_{\mathscr{C}}(z)\|^* \leq 1 - c \Big\}, \tag{3.7}$$

where the Lipschitz constant is defined with respect to the ℓ_2 norm. This provides a sort of converse to our definition of the degenerate points, where we set $\gamma_x(\mathscr{C}) = \infty$ for all $x \in \mathscr{C}_{\mathsf{dgn}}$, i.e. all points x where $\mathscr{P}_{\mathscr{C}}$ is *not* continuous in any neighborhood of x.

3.2.2 Connection to prox-regular sets

The notion of prox-regular sets and sets of positive reach arises in the literature on nonsmooth analysis in Hilbert spaces, for instance see [32] for a comprehensive overview of the key results in this area.

A prox-regular set is a set $\mathscr{C} \subset \mathbb{R}^d$ that satisfies

$$\langle y - x, z - x \rangle \leq \frac{1}{2\rho} \|z - x\|_2 \|y - x\|_2^2, \tag{3.8}$$

for all $x, y \in \mathscr{C}$ and all $z \in \mathbb{R}^d$ with $\mathscr{P}_{\mathscr{C}}(z) = x$, for some constant $\rho > 0$. To capture the local variations in concavity over the set $\mathscr{C}$, $\mathscr{C}$ is prox-regular with respect to a continuous function $\rho : \mathscr{C} \to (0, \infty]$ if

$$\langle y - x, z - x \rangle \leq \frac{1}{2\rho(x)} \|z - x\|_2 \|y - x\|_2^2 \tag{3.9}$$

for all $x, y \in \mathscr{C}$ and all $z \in \mathbb{R}^d$ with $\mathscr{P}_{\mathscr{C}}(z) = x$ (see e.g. [32, Theorem 3b]). Prox-regularity was first formulated via the notion of “positive reach” [33]: the parameter ρ appearing in (3.8) is the largest radius such that the projection operator $\mathscr{P}_{\mathscr{C}}$ is unique for all points z within distance ρ of the set $\mathscr{C}$; in the local version (3.9), the radius is allowed to vary locally as a function of $x \in \mathscr{C}$.

These definitions (3.8) and (3.9) exactly coincide with our inner product condition (3.5), in the special case that $\|\cdot\|$ is the ℓ_2 norm, by taking $\gamma = \frac{1}{2\rho}$ or, for the local coefficients, $\gamma = \frac{1}{2\rho(x)}$. The distinctions between our definitions and results on local concavity coefficients, and the literature on prox-regularity, center on two key differences: the role of continuity, and the flexibility of the structured norm $\|\cdot\|$ (rather than the ℓ_2 norm).

First, for prox-regular sets, the "reach" function $x \mapsto \rho(x) \in (0, \infty]$ is assumed to be continuous [32, Definition 1]. Equivalently, we could take a continuous function $x \mapsto \gamma_x = \frac{1}{2\rho(x)} \in [0, \infty)$ to agree with the notation of our local concavity coefficients. However, we do not enforce continuity of the map $x \mapsto \gamma_x$ in our definitions, and instead define $\gamma_x(\mathscr{C})$ as the smallest value such that the conditions are satisfied. This leads to substantial challenges in proving the equivalence of the various conditions; in Lemma 3.2.2 we prove that the map is naturally upper semi-continuous, which allows us to show the desired equivalences. In addition, if we do require a continuity assumption on the function $x \mapsto \gamma_x$, then we would be forced to have $\gamma_x > 0$ for some points $x \in \mathsf{Int}(\mathscr{C})$ since we must have $\gamma_x > 0$ for at least some of the points x on the boundary of $\mathscr{C}$. This means that γ_x would not give a precise quantification of the concavity in its interior $\mathsf{Int}(\mathscr{C})$.

Next, prox-regularity is defined with respect to the ℓ_2 norm, whereas we define local concavity coefficients with respect to a general structured norm $\|\cdot\|$, such as the ℓ_1 norm in a sparse signal estimation setting. While the equivalence of all norms on $\mathbb{R}^d$ means that if $\gamma(\mathscr{C})$ is finite when defined with respect to the ℓ_2 norm (i.e. $\mathscr{C}$ is prox-regular), then it is finite with respect to any other norm, the distinction is that in optimization problems arising in high-dimensional settings (for instance, high-dimensional regression in statistics), structured norms such as the ℓ_1 norm (for problems involving sparse signals) or the nuclear norm (for low-rank signals) allow for statistical and computational analyses that would not be possible with the ℓ_2 norm. In particular, we will see later on that convergence for the minimization problem $\min_{x \in \mathscr{C}} \mathsf{g}(x)$ will depend on bounding $\|\nabla \mathsf{g}(x)\|^*$. If $\|\cdot\|$ is the ℓ_1 norm, for instance, then $\|\nabla \mathsf{g}(x)\|^* = \|\nabla \mathsf{g}(x)\|_\infty$ will in general be much smaller than $\|\nabla \mathsf{g}(x)\|_2$. For instance, in a statistical problem, if $\nabla \mathsf{g}(x)$ consists of Gaussian or subgaussian noise at the true parameter vector x, then $\|\nabla \mathsf{g}(x)\|_\infty \sim \sqrt{\log(d)}$ while $\|\nabla \mathsf{g}(x)\|_2 \sim \sqrt{d}$. Therefore, being able to bound the concavity of $\mathscr{C}$ with respect to the ℓ_1 norm rather than the ℓ_2 norm is crucial for analyzing convergence in a high-dimensional setting.

3.3 Examples

In this section we consider a range of nonconvex constraints arising naturally in high-dimensional statistics, and show that these sets come equipped with well-behaved local concavity coefficients.

3.3.1 Low rank

Estimating a matrix with low rank structure arises in a variety of problems in high-dimensional statistics and machine learning. A partial list includes PCA (principal component analysis), factor models, matrix completion, and reduced rank regression.

Here we will study the set of rank-constrained matrices

$$\mathscr{C} = \{X \in \mathbb{R}^{n\times m} : \operatorname{rank}(X) \leq r\}$$

to determine how our general framework of local concavity applies to this specific low rank setting. To avoid triviality, we assume $r < \min\{n,m\}$. Writing $\sigma_1(X) \geq \sigma_2(X) \geq \dots$ to denote the sorted singular values of any matrix X, we compute the curvature condition of $\mathscr{C}$:

Lemma 3.3.1. *Let $\mathscr{C} = \{X \in \mathbb{R}^{n\times m} : \operatorname{rank}(X) \leq r\}$. Then $\mathscr{C}$ has local concavity coefficients given by $\gamma_X(\mathscr{C}) = \frac{1}{2\sigma_r(X)}$ for all $X \in \mathscr{C}$, with respect to norms $\|\cdot\| = \|\cdot\|_{\text{nuc}}$ and $\|\cdot\|^* = \|\cdot\|_{\text{sp}}$.*

3.3.2 Sparsity

In many applications in high-dimensional statistics, the signal of interest is believed to be sparse or approximately sparse. Using an ℓ_1 penalty or constraint serves as a convex relaxation to the sparsity constraint, i.e. the Lasso method [34], in the case of a linear regression problem. However, the ℓ_1 norm penalty also leads to undesirable shrinkage bias on the large coefficients of x, e.g. [35]. The shrinkage problem can be alleviated by turning to nonconvex regularization functions, including the SCAD penalty [36], the MCP penalty [37], and the adaptive Lasso / reweighted ℓ_1 method [38] (which is related to a nonconvex "log-ℓ_1" penalty).

Loh and Wainwright [39] considers a class of nonconvex sparse regularizers, which takes the form

$$\mathrm{Pen}(x) = \sum_i \mathsf{p}(|x_i|) \text{ where } \begin{cases} \mathsf{p}(0) = 0 \text{ and } \mathsf{p} \text{ is nondecreasing,} \\ t \mapsto \mathsf{p}(t)/t \text{ is nonincreasing (i.e. } \mathsf{p} \text{ is concave),} \\ t \mapsto \mathsf{p}(t) + \frac{\mu}{2}t^2 \text{ is convex,} \\ \mathsf{p} \text{ is differentiable on } t > 0, \text{ with } \lim_{t \searrow 0} \mathsf{p}'(t) = 1. \end{cases} \tag{3.10}$$

Essentially, this means that $\mathrm{Pen}(x)$ behaves like a nonconvex version of the ℓ_1 norm, shrinking small coefficients to zero but avoiding heavy shrinkage on large coefficients; the SCAD, MCP, and log-ℓ_1 penalties are all examples.

Now consider the sparsity-inducing constraint set $\mathscr{C} = \{x : \mathrm{Pen}(x) \le c\}$, then the following result calculates the local concavity coefficients for $\mathscr{C}$.

Lemma 3.3.2. *Suppose that* $\mathrm{Pen}(x) = \sum_i \mathsf{p}(|x_i|)$ *where* p *satisfies conditions* (3.10)*. Then*

$$\begin{cases} \gamma_x(\mathscr{C}) \le \frac{\mu/2}{\mathsf{p}'(x_{\min})}, & \textit{if } \mathrm{Pen}(x) = c, \\ \gamma_x(\mathscr{C}) = 0, & \textit{if } \mathrm{Pen}(x) < c, \end{cases}$$

with respect to the norm $\|\cdot\| = \|\cdot\|_1$ *and its dual* $\|\cdot\|^* = \|\cdot\|_\infty$*, where for any* $x \in \mathbb{R}^d \backslash \{0\}$ *we define* $x_{\min}$ *to be the magnitude of its smallest nonzero entry.*

3.3.3 Spheres, orthogonal groups, and orthonormal matrices

We next consider a constraint set given by $\mathscr{C} = \{X \in \mathbb{R}^{n\times r} : X^\top X = \mathbf{I}_r\}$, the space of all orthonormal $n \times r$ matrices. This constraint set arises in PCA type problems where we would like to find the basis vectors that span the best rank-r subspace of a data set.

Lemma 3.3.3. *Let* $\mathscr{C} = \{X \in \mathbb{R}^{n\times r} : X^\top X = \mathbf{I}_r\}$*, the space of orthogonal $n \times r$ matrices. Then* $\mathscr{C}$ *has local concavity coefficients* $\gamma_X(\mathscr{C}) = \frac{1}{2}$ *with respect to* $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$ *and dual norm* $\|\cdot\|^* =$

$\|\cdot\|_{\mathrm{sp}}$.

Observe that the sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ is a special case, obtained when $r = 1$.

Next, in many problems we may aim to find a rank-r subspace that is optimal in some regard, but the exact choice of basis for this subspace does not matter; that is, an orthonormal basis $X \in \mathbb{R}^{n\times r}$ is identifiable only up to a rotation of its columns. In this case, we can instead choose to work with rank-r projection matrices:

Lemma 3.3.4. *Let $\mathscr{C} = \{X \in \mathbb{R}^{n\times n} : \mathrm{rank}(X) = r, X \succeq 0, X^2 = X\}$, the space of rank-r projection matrices. Then $\mathscr{C}$ has local concavity coefficients $\gamma_X(\mathscr{C}) \leq 2$ with respect to $\|\cdot\| = \|\cdot\|_{\mathrm{nuc}}$ and $\|\cdot\|^* = \|\cdot\|_{\mathrm{sp}}$.*

A special case is the setting $r = n$, when $\mathscr{C}$ is the orthogonal group, in which case Lemma 3.3.3 proves the concavity coefficient is equal to 1/2

3.4 Convergence of projected gradient descent

In this section, we briefly explore how the concavity coefficients allow us to extend the standard analysis of projected gradient descent to incorporate the nonconvexity of the constraint set.

Consider an optimization problem constrained to a nonconvex set, $\min\{\mathrm{g}(x) : x \in \mathscr{C}\}$, where $\mathrm{g} : \mathbb{R}^d \to \mathbb{R}$ is a differentiable function. After choosing some initial point $x_0 \in \mathscr{C}$, for each $t \geq 0$ we define

$$\begin{cases} x'_{t+1} = x_t - \eta \nabla \mathrm{g}(x_t), \\ x_{t+1} = \mathscr{P}_{\mathscr{C}}(x'_{t+1}), \end{cases} \tag{3.11}$$

where if $\mathscr{P}_{\mathscr{C}}(x'_{t+1})$ is not unique then any closest point may be chosen.

3.4.1 Assumptions

Let $\widehat{x}$ be the target of our optimization procedure, $\widehat{x} \in \arg\min_{x \in \mathscr{C}} \mathrm{g}(x)$. We assume that g satisfies restricted strong convexity (RSC) and restricted smoothness (RSM) conditions over $x, y \in \mathscr{C}$,

$$\mathrm{g}(y) \geq \mathrm{g}(x) + \langle y - x, \nabla \mathrm{g}(x) \rangle + \frac{\alpha}{2} \|x - y\|_2^2 - \frac{\alpha}{2} \varepsilon_{\text{stat}}^2, \tag{3.12}$$

and

$$\mathrm{g}(y) \leq \mathrm{g}(x) + \langle y - x, \nabla \mathrm{g}(x) \rangle + \frac{\beta}{2} \|x - y\|_2^2 + \frac{\alpha}{2} \varepsilon_{\text{stat}}^2. \tag{3.13}$$

Without loss of generality we can take $\alpha \leq \beta$. The term $\varepsilon_{\text{stat}}$, often referred to as the "statistical error" in the high-dimensional statistics literature [7, 39], gives some "slack" in our assumption on g, and is intended to capture some vanishingly small error level. See Section 4.2 below for a more detailed discussion of statistical error in the context of high-dimensional optimization.

Next, we assume a *norm compatibility condition*,

$$\|z - \mathscr{P}_{\mathscr{C}}(z)\|^* \leq \phi \min_{x \in \mathscr{C}} \|z - x\|^* \text{ for all } z \in \mathbb{R}^d, \tag{3.14}$$

for some constant $\phi \geq 1$. The norm compatibility condition is trivially true with $\phi = 1$ if $\|\cdot\|$ is the ℓ_2 norm, since $\mathscr{P}_{\mathscr{C}}$ is a projection with respect to the ℓ_2 norm. In many natural settings it holds even for other norms, often with $\phi = 1$.

Finally, we assume a gradient condition that reveals the connection between the curvature of the nonconvex set $\mathscr{C}$ and the target function g: we require that

$$2\phi \cdot \max_{x, x' \in \mathscr{C} \cap \mathbb{B}_2(\widehat{x}, \rho)} \gamma_x(\mathscr{C}) \|\nabla \mathrm{g}(x')\|^* \leq (1 - c_0) \cdot \alpha. \tag{3.15}$$

(Since $x \mapsto \gamma_x(\mathscr{C})$ is upper semi-continuous, if g is continuously differentiable, then we can find some radius $\rho > 0$ and some constant $c_0 > 0$ satisfying this condition, as long as $2\phi \gamma_{\widehat{x}}(\mathscr{C}) \|\nabla \mathrm{g}(\widehat{x})\|^* < \alpha$.) Our projected gradient descent algorithm will then succeed if initialized within this radius

ρ from the target point $\widehat{x}$, with an appropriate step size. For a detailed discussion of the necessity of this type of initialization condition, we refer the reader to [31].

3.4.2 Fast convergence guarantee

We now state our convergence result of projected gradient descent. The inner product condition (3.5) as well as the initialization condition (3.15) ensure fast convergence to $\widehat{x}$ as long as initialized at some $x_0 \in \mathscr{C}$ sufficiently close to $\widehat{x}$.

Theorem 3.4.1. *Let $\mathscr{C} \subset \mathbb{R}^d$ be a constraint set and let* g *be a differentiable function, with minimizer $\widehat{x} \in \arg\min_{x\in\mathscr{C}} \mathrm{g}(x)$. Suppose $\mathscr{C}$ satisfies the norm compatibility condition* (3.14) *with parameter ϕ, and* g *satisfies restricted strong convexity* (3.12) *and restricted smoothness* (3.13) *with parameters $\alpha, \beta, \varepsilon_{\text{stat}}$ for all $x, y \in \mathscr{C}$, and the initialization condition* (3.15) *for some $c_0 > 0$. If the initial point $x_0 \in \mathscr{C}$ and the error level $\varepsilon_{\text{stat}}$ satisfy $\|x_0 - \widehat{x}\|_2^2 < \rho^2$ and $\varepsilon_{\text{stat}}^2 < \frac{c_0\rho^2}{1.5}$, then for each step $t \geq 0$ of the projected gradient descent algorithm* (3.11) *with step size $\eta = 1/\beta$,*

$$\|x_t - \widehat{x}\|_2^2 \leq \left(1 - c_0 \cdot \frac{2\alpha}{\alpha+\beta}\right)^t \|x_0 - \widehat{x}\|_2^2 + \frac{1.5\varepsilon_{\text{stat}}^2}{c_0}.$$

In other words, the iterates x_t converge linearly to the minimizer $\widehat{x}$, up to precision level $\varepsilon_{\text{stat}}$. To compare this result to the convex setting, if $\mathscr{C}$ is a convex set and g is α-strongly convex and β-smooth, then we can set $c_0 = 1$ and $\varepsilon_{\text{stat}} = 0$. Our result then yields matching known rates for the convex setting (see e.g. [40, Theorem 3.10]).

Proof of Theorem 3.4.1. For $t = 0$, the statement holds trivially. To prove that the bound holds for subsequent steps, we will proceed by induction. Choose any $\rho_0 \in (0, \rho)$ such that

$$\rho_0 \geq \max\left\{ \|x_0 - \widehat{x}\|_2, \sqrt{\frac{1.5\varepsilon_{\text{stat}}^2}{c_0}} \right\},$$

where this maximum is $< \rho$ by assumption of the theorem. We will prove that

$$\begin{cases} \|x_{t+1} - \widehat{x}\|_2^2 \le \left(1 - \frac{2c_0\alpha}{\alpha+\beta}\right) \|x_t - \widehat{x}\|_2^2 + \frac{3\alpha}{\alpha+\beta}\varepsilon_{\text{stat}}^2, \\ \|x_{t+1} - \widehat{x}\|_2 \le \rho_0, \end{cases} \tag{3.16}$$

for all $t \ge 0$. Assuming that this holds, then applying the first bound of (3.16) iteratively, we will then have

$$\|x_t - \widehat{x}\|_2^2 \le \left(1 - \frac{2c_0\alpha}{\alpha+\beta}\right)^t \|x_0 - \widehat{x}\|_2^2 + \frac{1.5}{c_0}\varepsilon_{\text{stat}}^2,$$

which proves the theorem.

Now we turn to proving (3.16), assuming that it holds at the previous time step. First, we have

$$\|x'_{t+1} - x_{t+1}\|^* = \|x'_{t+1} - \mathscr{P}_{\mathscr{C}}(x'_{t+1})\|^* \le \phi\|x'_{t+1} - x_t\|^* = \phi\|-\eta\nabla \mathrm{g}(x_t)\|^* \le \frac{\eta\alpha(1-c_0)}{2\gamma_{x_{t+1}}(\mathscr{C})}, \tag{3.17}$$

where first inequality uses the norm compatibility condition (3.14) while the second uses the initialization condition (3.15), since $\|x_t - \widehat{x}\|_2 \le \rho$.

Next, the inner product condition yields

$$\langle \widehat{x} - x_{t+1}, x'_{t+1} - x_{t+1}\rangle \ \le \ \gamma_{x_{t+1}}(\mathscr{C})\|x'_{t+1} - x_{t+1}\|^*\|x_{t+1} - \widehat{x}\|_2^2 \ = \ \frac{\eta\alpha(1-c_0)}{2}\|x_{t+1} - \widehat{x}\|_2^2, \tag{3.18}$$

where the last step applies (3.17).

We will now apply the first-order optimality conditions (3.4) at the point $x = \widehat{x}$. We have

$$
\begin{aligned}
&\mathrm{g}(x_{t+1}) - \mathrm{g}(\widehat{x}) \\
&\geq \langle x_{t+1} - \widehat{x}, \nabla \mathrm{g}(\widehat{x}) \rangle + \frac{\alpha}{2}\|x_{t+1} - \widehat{x}\|_2^2 - \frac{\alpha\varepsilon_{\text{stat}}^2}{2} \text{ by restricted strong convexity (3.12)} \\
&\geq -\gamma_{x_{t+1}}(\mathscr{C})\|\nabla \mathrm{g}(\widehat{x})\|^* \|x_{t+1} - \widehat{x}\|_2^2 + \frac{\alpha}{2}\|x_{t+1} - \widehat{x}\|_2^2 - \frac{\alpha\varepsilon_{\text{stat}}^2}{2} \text{ by first-order optimality} \\
&\geq -\frac{\alpha(1-c_0)}{2}\|x_{t+1} - \widehat{x}\|_2^2 + \frac{\alpha}{2}\|x_{t+1} - \widehat{x}\|_2^2 - \frac{\alpha\varepsilon_{\text{stat}}^2}{2} \\
&= \frac{c_0\alpha}{2}\|x_{t+1}(s) - \widehat{x}\|_2^2 - \frac{\alpha\varepsilon_{\text{stat}}^2}{2}, \qquad (3.19)
\end{aligned}
$$

where the next-to-last step applies the initialization condition (3.15) (plus the fact that $\phi \geq 1$) to bound $\|\nabla \mathrm{g}(\widehat{x})\|^*$. On the other hand, we have

$$
\begin{aligned}
&\mathrm{g}(x_{t+1}) - \mathrm{g}(\widehat{x}) = \mathrm{g}(x_{t+1}) - \mathrm{g}(x_t) + \mathrm{g}(x_t) - \mathrm{g}(\widehat{x}) \\
&\leq \langle x_{t+1} - x_t, \nabla \mathrm{g}(x_t) \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2 + \frac{\alpha\varepsilon_{\text{stat}}^2}{2} + \langle x_t - \widehat{x}, \nabla \mathrm{g}(x_t) \rangle - \frac{\alpha}{2}\|x_t - \widehat{x}\|_2^2 + \frac{\alpha\varepsilon_{\text{stat}}^2}{2} \\
&= \langle x_{t+1} - \widehat{x}, \nabla \mathrm{g}(x_t) \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|_2^2 - \frac{\alpha}{2}\|x_t - \widehat{x}\|_2^2 + \alpha\varepsilon_{\text{stat}}^2, \qquad (3.20)
\end{aligned}
$$

where the inequality applies restricted strong convexity (3.12) and restricted smoothness (3.13). To bound the remaining inner product term, we have

$$
\begin{aligned}
\langle x_{t+1} - \widehat{x}, \nabla \mathrm{g}(x_t) \rangle &= \frac{1}{\eta}\langle x_{t+1} - \widehat{x}, x_t - x'_{t+1} \rangle = \frac{1}{\eta}\langle x_{t+1} - \widehat{x}, x_t - x_{t+1} \rangle + \frac{1}{\eta}\langle x_{t+1} - \widehat{x}, x_{t+1} - x'_{t+1} \rangle \\
&\leq \frac{1}{\eta}\langle x_{t+1} - \widehat{x}, x_t - x_{t+1} \rangle + \frac{\alpha(1-c_0)}{2}\|\widehat{x} - x_{t+1}(s)\|_2^2, \quad (3.21)
\end{aligned}
$$

where the last step applies (3.18). For the first term on the right-hand side, we can trivially check that

$$
\frac{1}{\eta}\langle x_{t+1} - \widehat{x}, x_t - x_{t+1} \rangle = \frac{1}{2\eta}\|x_t - \widehat{x}\|_2^2 - \frac{1}{2\eta}\|x_{t+1} - \widehat{x}\|_2^2 - \frac{1}{2\eta}\|x_{t+1} - x_t\|_2^2. \qquad (3.22)
$$

Combining steps (3.19), (3.20), (3.21), and (3.22), then, since $\frac{1}{2\eta} = \frac{\beta}{2}$,

$$\frac{c_0\alpha}{2}\|x_{t+1}-\widehat{x}\|_2^2 \leq \frac{1}{2\eta}\|x_t-\widehat{x}\|_2^2 - \frac{1}{2\eta}\|x_{t+1}-\widehat{x}\|_2^2 + \frac{\alpha(1-c_0)}{2}\|\widehat{x}-x_{t+1}(s)\|_2^2 \\ - \frac{\alpha}{2}\|x_t-\widehat{x}\|_2^2 + 1.5\alpha\varepsilon_{\mathrm{stat}}^2.$$

Rearranging terms we obtain

$$\|x_{t+1}-\widehat{x}\|_2^2 \leq \left(1-\frac{2c_0\alpha}{\alpha+\beta}\right)\|x_t-\widehat{x}\|_2^2 + \frac{3\alpha}{\alpha+\beta}\varepsilon_{\mathrm{stat}}^2. \tag{3.23}$$

In particular, since $\|x_t-\widehat{x}\|_2 \leq \rho_0$ and $\varepsilon_{\mathrm{stat}}^2 \leq \frac{c_0\rho_0^2}{1.5}$ by assumption, this proves that

$$\|x_{t+1}-\widehat{x}\|_2 \leq \rho_0. \tag{3.24}$$

This proves that the inductive step (3.16) holds for x_{t+1}, as desired, which completes the proof of Theorem 3.4.1.

□

3.5 Proofs of local concavity coefficient results

In this section we prove the equivalence of the multiple notions of the (local or global) concavity of the constraint set $\mathscr{C}$, given in Theorems 3.1.1 and 3.2.1, as well as some properties of these coefficients (Lemmas 3.2.1, 3.2.2, 3.2.3, and 3.2.4).

Notation Before the equivalence is established, we begin by introducing notation for the local concavity coefficients defined using each of these four properties: for all $x \in \mathscr{C}$, define

$$\gamma_x^{\mathrm{curv}}(\mathscr{C}) = \min\{\gamma \in [0,\infty] : \text{The curvature condition (3.1) holds for this point } x \text{ and any } y \in \mathscr{C}\},$$

$$\gamma_x^{\mathrm{contr}}(\mathscr{C}) = \min\{\gamma \in [0,\infty] : \text{The contraction condition (3.3) holds for this point } x \text{ and any } y \in \mathscr{C}\},$$

$$\gamma_x^{\mathrm{FO}}(\mathscr{C}) = \min\{\gamma \in [0,\infty] : \text{The first-order condition (3.4) holds for this point } x \text{ and any } y \in \mathscr{C}\},$$

$$\gamma_x^{\mathrm{IP}}(\mathscr{C}) = \min\{\gamma \in [0,\infty] : \text{The inner product condition (3.5) holds for this point } x \text{ and any } y \in \mathscr{C}\}.$$

We emphasize that here we are *not* explicitly setting these coefficients to equal ∞ at degenerate points $x \in \mathscr{C}_{\mathsf{dgn}}$—they may take finite values (we will need this distinction for some technical parts of our proofs later on). We will prove that these four definitions are all equal for all $x \notin \mathscr{C}_{\mathsf{dgn}}$, which is sufficient for the equivalence result Theorem 3.2.1 since the local concavity coefficients are set to ∞ at degenerate points.

Next, we define a constant $B_{\mathrm{norm}} > 0$ such that

$$\text{For all } z \in \mathbb{R}^d, \begin{cases} B_{\mathrm{norm}}^{-1}\|z\|_2 \leq \|z\| \leq B_{\mathrm{norm}}\|z\|_2, \\ B_{\mathrm{norm}}^{-1}\|z\|_2 \leq \|z\|^* \leq B_{\mathrm{norm}}\|z\|_2. \end{cases} \tag{3.25}$$

By equivalence of norms on $\mathbb{R}^d$, B_{norm} will be always finite.

We state a well-known fact about projections, which we will use throughout our proofs:

$$\text{For any } z \in \mathbb{R}^d \text{ and } x \in \mathscr{C} \text{ with } \mathscr{P}_{\mathscr{C}}(z) = x, \text{ for any } t \in [0,1],\ \mathscr{P}_{\mathscr{C}}((1-t)x + tz) = x. \tag{3.26}$$

3.5.1 Proof outline of Theorem 3.2.1

To prove the equivalence of the four definitions of the local coefficients in (3.6), we first need to show that these coefficients are upper semi-continuous, as claimed in Lemma 3.2.2. Since we do not yet know that the four definitions are equivalent, we first show that the map $x \mapsto \gamma_x^{\mathrm{IP}}$ is upper semi-continuous over $x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$ (see Lemma 3.5.1 below).

Using upper semi-continuity of $\gamma_x^{\mathrm{IP}}(\mathscr{C})$ over $x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$, we next show that

$$\gamma_x^{\mathrm{curv}}(\mathscr{C}) = \gamma_x^{\mathrm{contr}}(\mathscr{C}) = \gamma_x^{\mathrm{IP}}(\mathscr{C}) = \gamma_x^{\mathrm{FO}}(\mathscr{C}) \text{ for all } x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}. \tag{3.27}$$

In particular, we prove the equivalence between the inner products property (3.5) and other three conditions, namely the curvature condition (3.1), the (one-sided) contraction property (3.3), and the first-order condition (3.4). Recall that if $x \in \mathscr{C}_{\mathsf{dgn}}$, then $\gamma_x(\mathscr{C}) = \infty$ under all four definitions. Combining with (3.27), we conclude the equivalence results, as claimed in Theorem 3.2.1.

In fact, we will also show that a weaker statement holds for all $x \in \mathscr{C}$ (i.e. without excluding degenerate points), namely

$$\gamma_x^{\mathrm{IP}}(\mathscr{C}) \leq \min\{\gamma_x^{\mathrm{curv}}(\mathscr{C}), \gamma_x^{\mathrm{contr}}(\mathscr{C}), \gamma_x^{\mathrm{FO}}(\mathscr{C})\} \text{ for all } x \in \mathscr{C}. \tag{3.28}$$

This additional bound will be useful later in our characterization of the degenerate points, when we prove Lemma 3.2.3.

3.5.2 Upper semi-continuity

First we prove upper semi-continuity of the map $x \mapsto \gamma_x^{\mathrm{IP}}(\mathscr{C})$ for $x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$:

Lemma 3.5.1. *The map $x \mapsto \gamma_x^{\mathrm{IP}}(\mathscr{C})$ is upper semi-continuous over $x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$.*

Once Theorem 3.2.1 is proved, then Lemma 3.5.1 becomes equivalent to the original lemma, Lemma 3.2.2, since $\gamma_x(\mathscr{C}) = \infty$ by definition on the subset $\mathscr{C}_{\mathsf{dgn}} \subset \mathscr{C}$, which is a closed subset by definition, while Lemma 3.5.1 proves that $x \mapsto \gamma_x(\mathscr{C})$ is upper semi-continuous over the open subset $\mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}} \subset \mathscr{C}$.

Proof of Lemma 3.5.1. Take any sequence $x_n \to x$, with $x, x_1, x_2, \cdots \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$. We want to prove that

$$\gamma := \limsup_{n \to \infty} \gamma_{x_n}^{\mathrm{IP}}(\mathscr{C}) \leq \gamma_x^{\mathrm{IP}}(\mathscr{C}). \tag{3.29}$$

Since $x \notin \mathscr{C}_{\mathsf{dgn}}$ by assumption, we know that $\mathscr{P}_{\mathscr{C}}$ is continuous in some neighborhood of x. Let $r > 0$ be some radius so that $\mathscr{P}_{\mathscr{C}}$ is continuous on $\mathbb{B}_*(x, r)$, where $\mathbb{B}_*(x, r)$ is the ball of radius r around the point x in the dual norm $\|\cdot\|^*$. Assume also that $\gamma > 0$, otherwise again the claim is trivial.

Taking a subsequence of the points $x_1, x_2, \ldots$ if necessary, we can assume without loss of generality that

$$\gamma^{\mathrm{IP}}_{x_n}(\mathscr{C}) \to \gamma.$$

Fix any $\varepsilon > 0$ such that $\varepsilon < \gamma$. For each n, by definition of the local concavity coefficient $\gamma^{\mathrm{IP}}_{x_n}(\mathscr{C})$, there must exist some $y_n \in \mathscr{C}$ and some $z'_n \in \mathbb{R}^d$ with $\mathscr{P}_{\mathscr{C}}(z'_n) = x_n$, such that

$$\langle y_n - x_n, z'_n - x_n \rangle > \left(\gamma^{\mathrm{IP}}_{x_n}(\mathscr{C}) - \varepsilon \right) \|z'_n - x_n\|^* \|y_n - x_n\|_2^2. \tag{3.30}$$

Define

$$z_n = \begin{cases} z'_n, & \text{if } \|z'_n - x_n\|^* \le r/2, \\ x_n + (z'_n - x_n) \cdot \frac{r/2}{\|z'_n - x_n\|^*}, & \text{if } \|z'_n - x_n\|^* > r/2, \end{cases}$$

so that $\|z_n - x_n\|^* \le r/2$. By (3.26), $\mathscr{P}_{\mathscr{C}}(z_n) = x_n$. Furthermore, rescaling both sides of the inequality (3.30),

$$\langle y_n - x_n, z_n - x_n \rangle > \left(\gamma^{\mathrm{IP}}_{x_n}(\mathscr{C}) - \varepsilon \right) \|z_n - x_n\|^* \|y_n - x_n\|_2^2. \tag{3.31}$$

Since the left-hand side is bounded by $\|y_n - x_n\| \|z_n - x_n\|^*$, we see that

$$\|y_n - x_n\|_2^2 < \frac{\|y_n - x_n\|}{\gamma^{\mathrm{IP}}_{x_n}(\mathscr{C}) - \varepsilon} \le \frac{\|y_n - x_n\|}{(\gamma - \varepsilon)/2}$$

for all n sufficiently large so that $\gamma^{\mathrm{IP}}_{x_n}(\mathscr{C}) > \gamma - \frac{\gamma - \varepsilon}{2}$. Therefore, since $\|y_n - x_n\| \le B_{\mathrm{norm}} \|y_n - x_n\|_2$ for some finite B_{norm}, then for all large n, y_n lies in some ball of finite radius around x. The same is true for z_n since $\|z_n - x_n\|^* \le r/2$ by construction. Thus we can find a convergent subsequence, that is, $n_1, n_2, \ldots$ such that

$$\begin{cases} y_{n_i} \to y \text{ for some point } y, \\ z_{n_i} \to z \text{ for some point } z. \end{cases}$$

Since $\mathscr{C}$ is closed, we must have $y \in \mathscr{C}$. And, since $x_{n_i} \to x$, for sufficiently large i we have $\|x_{n_i} - x\|^* \leq r/2$, so that $z_{n_i} \in \mathbb{B}_*(x,r)$. Since $\mathscr{P}_{\mathscr{C}}$ is continuous on the ball $\mathbb{B}_*(x,r)$, then, $\mathscr{P}_{\mathscr{C}}(z_{n_i}) = x_{n_i} \to x$ implies that we must have $\mathscr{P}_{\mathscr{C}}(z) = x$. And,

$$\begin{aligned} \langle y - x, z - x\rangle = \lim_{i\to\infty} \langle y_{n_i} - x_{n_i}, z_{n_i} - x_{n_i}\rangle \geq \lim_{i\to\infty} \left(\gamma^{\mathrm{IP}}_{x_{n_i}}(\mathscr{C}) - \varepsilon\right) \|z_{n_i} - x_{n_i}\|^* \|y_{n_i} - x_{n_i}\|_2^2 \\ = (\gamma - \varepsilon)\|z - x\|^* \|y - x\|_2^2, \end{aligned}$$

where the inequality applies (3.31) for each n_i. Therefore, $\gamma^{\mathrm{IP}}_x(\mathscr{C}) \geq \gamma - \varepsilon$. Since $\varepsilon > 0$ was chosen to be arbitrarily small, this proves that $\gamma^{\mathrm{IP}}_x(\mathscr{C}) \geq \gamma$, as desired. □

3.5.3 Equivalence for local concavity

Now we prove the equivalence results Theorem 3.2.1.

Inner products ⇒ First-order optimality. Fix any $u \in \mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$. Let $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$ be differentiable, and suppose that u is a local minimizer of f over $\mathscr{C}$. By [41, Theorem 6.12], this implies that $-\nabla\mathsf{f}(u) \in N_{\mathscr{C}}(u)$, where $N_{\mathscr{C}}(u)$ is the normal cone to $\mathscr{C}$ at the point u (see [41, Definition 6.3]). By [32, (12)], we know that the normal cone can be obtained by a limit of proximal normal cones,

$$N_{\mathscr{C}}(u) = \lim \sup_{x\in\mathscr{C}, x\to u} \underbrace{\left\{w \in \mathbb{R}^d : \mathscr{P}_{\mathscr{C}}(x + \varepsilon \cdot w) = x \text{ for some } \varepsilon > 0\right\}}_{\text{Proximal normal cone to } \mathscr{C} \text{ at } x}.$$

Therefore, we can find some sequences $u_1, u_2, \cdots \in \mathscr{C}$, $w_1, w_2, \cdots \in \mathbb{R}^d$, and $\varepsilon_1, \varepsilon_2, \cdots > 0$, such that $\mathscr{P}_{\mathscr{C}}(u_n + \varepsilon_n \cdot w_n) = u_n$ for all $n \geq 1$, with $u_n \to u$ and $w_n \to -\nabla\mathsf{f}(u)$.

Now fix any $y \in \mathscr{C}$. By the inner product condition (3.5), for each $n \geq 1$,

$$\langle y - u_n, w_n \rangle = \langle y - u_n, (u_n + w_n) - u_n \rangle \leq \gamma^{\mathrm{IP}}_{u_n}(\mathscr{C}) \|w_n\|^* \|y - u_n\|_2^2.$$

Taking limits on both sides, since $u_n \to u$ and $w_n \to -\nabla \mathsf{f}(u)$,

$$\langle y - u, -\nabla \mathsf{f}(u) \rangle \leq \left(\limsup_{t \to \infty} \gamma^{\mathrm{IP}}_{u_n}(\mathscr{C}) \right) \cdot \|\nabla \mathsf{f}(u)\|^* \|y - u\|_2^2.$$

Finally, recall that Lemma 3.5.1 proves that $x \mapsto \gamma^{\mathrm{IP}}_x(\mathscr{C})$ is upper semi-continuous over $x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$, and $\mathscr{C}_{\mathsf{dgn}} \subset \mathscr{C}$ is a closed subset. Since $u \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$, we therefore have $u_n \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$ for all sufficiently large t, and therefore $\limsup_{t \to \infty} \gamma^{\mathrm{IP}}_{u_n}(\mathscr{C}) \leq \gamma^{\mathrm{IP}}_u(\mathscr{C})$. This proves that $\gamma^{\mathrm{FO}}_u(\mathscr{C}) \leq \gamma^{\mathrm{IP}}_u(\mathscr{C})$, as desired.

In fact, we can formulate a more general version of the first-order optimality condition:

For any Lipschitz continuous $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$ such that x is a local minimizer of f over $\mathscr{C}$,

$$\langle y - x, v \rangle \geq -\gamma \|v\|^* \|y - x\|_2^2 \text{ for some } v \in \partial \mathsf{f}(x), \quad (3.32)$$

where $\partial \mathsf{f}(x)$ is the subdifferential to f at x (see [41, Definition 8.3]). To see why (3.32) holds, [41, Theorem 8.15] guarantees that, since f is Lipschitz and x is a local minimizer of f over the closed set $\mathscr{C}$, then we must have $-v \in N_{\mathscr{C}}(x)$ for some subgradient $v \in \partial \mathsf{f}(x)$.[4] The remainder of the proof is identical to the differentiable case treated above, with v in place of $\nabla \mathsf{f}(x)$; this proves that, for any $x \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$ and any $y \in \mathscr{C}$, the stronger first-order optimality condition (3.32) holds with $\gamma = \gamma^{\mathrm{IP}}_x(\mathscr{C})$. □

First-order optimality ⇒ Inner products. This direction of the equivalence is immediate: setting $\mathsf{f}(w) = \frac{1}{2}\|w - z\|_2^2$, we can easily see that $\gamma^{\mathrm{IP}}_x(\mathscr{C}) \leq \gamma^{\mathrm{FO}}_x(\mathscr{C})$ for all $x \in \mathscr{C}$, while previously we

[4] More precisely, [41, Theorem 8.15] assumes only that f is proper and lower semi-continuous, but additionally requires the condition that $\partial^\infty \mathsf{f}(x) \cap \big(-N_{\mathscr{C}}(x)\big) = \{0\}$ (see [41, Chapter 8] for definitions). Since the horizon subdifferential $\partial^\infty \mathsf{f}(x)$ contains only the zero vector for any Lipschitz function f, this condition must be satisfied once we assume that f is Lipschitz.

showed that the reverse inequality holds over $x \in \mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$. Therefore, $\gamma_x^{\mathrm{IP}}(\mathscr{C}) = \gamma_x^{\mathrm{FO}}(\mathscr{C})$ for $x \in \mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$. □

Curvature ⇒ Inner products. Fix any $x, y \in \mathscr{C}$ and any $z \in \mathbb{R}^d$ with $\mathscr{P}_{\mathscr{C}}(z) = x$. For all $t \in (0,1)$, let $x_t = (1-t)x + ty$, and choose

$$\tilde{x}_t \in \operatorname*{arg\,min}_{x \in \mathscr{C}} \|x - x_t\| \text{ such that } \limsup_{t \searrow 0} \frac{\|\widetilde{x}_t - x_t\|}{t} \le \gamma_x^{\mathrm{curv}}(\mathscr{C})\|x-y\|_2^2,$$

as in the definition of $\gamma_x^{\mathrm{curv}}(\mathscr{C})$. Fix any $\varepsilon > 0$. Then for some $t_0 > 0$, for all $t < t_0$,

$$\frac{\|\widetilde{x}_t - x_t\|}{t} \le \gamma_x^{\mathrm{curv}}(\mathscr{C})\|x-y\|_2^2 + \varepsilon.$$

Since $x = \mathscr{P}_{\mathscr{C}}(z)$, this means that for all $t \in (0,1)$,

$$\|z-x\|_2^2 \le \|z - \widetilde{x}_t\|_2^2 = \|z - x_t\|_2^2 + \|\widetilde{x}_t - x_t\|_2^2 + 2\langle z - x_t, x_t - \widetilde{x}_t\rangle.$$

We can also calculate

$$\|z - x_t\|_2^2 = \|z - (1-t)x - ty\|_2^2 = \|z-x\|_2^2 - 2t\langle y-x, z-x\rangle + t^2\|x-y\|_2^2.$$

We rearrange terms to obtain

$$\langle y-x, z-x\rangle \le \frac{1}{2t}\left(\|\widetilde{x}_t - x_t\|_2^2 + 2\langle z - x_t, x_t - \widetilde{x}_t\rangle + t^2\|x-y\|_2^2\right).$$

Recalling that $\|\cdot\|_2 \le B_{\mathrm{norm}}\|\cdot\|$ for some finite constant B_{norm} by (3.25), we then have

$$\begin{aligned}
\langle y-x, z-x\rangle &\le \frac{1}{2t}\left((B_{\mathrm{norm}})^2\|\widetilde{x}_t - x_t\|^2 + 2\|z-x_t\|^*\|\widetilde{x}_t - \widetilde{x}\| + t^2\|x-y\|_2^2\right) \\
&\le \frac{1}{2t}\left((B_{\mathrm{norm}})^2\left((\gamma_x^{\mathrm{curv}}(\mathscr{C})\|x-y\|_2^2 + \varepsilon)\cdot t\right)^2 + 2\|z-x_t\|^*(\gamma_x^{\mathrm{curv}}(\mathscr{C})\|x-y\|_2^2 + \varepsilon)\cdot t + t^2\|x-y\|_2^2\right) \\
&= \|z-x_t\|^*(\gamma_x^{\mathrm{curv}}(\mathscr{C})\|x-y\|_2^2 + \varepsilon) + \frac{t}{2}\left((B_{\mathrm{norm}})^2\left((\gamma_x^{\mathrm{curv}}(\mathscr{C})\|x-y\|_2^2 + \varepsilon)\right)^2 + \|x-y\|_2^2\right).
\end{aligned}$$

Taking a limit as t approaches zero,

$$\langle y-x, z-x\rangle \leq (\gamma_x^{\text{curv}}(\mathscr{C})\|x-y\|_2^2+\varepsilon)\cdot\|z-x\|^*.$$

Since $\varepsilon>0$ was chosen to be arbitrarily small, therefore, $\gamma_x^{\text{IP}}(\mathscr{C})\leq\gamma_x^{\text{curv}}(\mathscr{C})$, for any $x\in\mathscr{C}$. $\square$

Inner products $\Rightarrow$ Curvature. To prove the curvature condition, we will actually need to use the stronger form (3.32) of the first-order optimality condition—as proved previously, this condition holds with $\gamma=\gamma_x^{\text{IP}}(\mathscr{C})$ for all $x\in\mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$.

Fix any $u\in\mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$ and $y\in\mathscr{C}$. Let $u_t=(1-t)\cdot u+t\cdot y$, and define $\mathsf{f}(x)=\|x-u_t\|$. Note that f is a Lipschitz function. Since $\mathscr{C}$ is closed, and f is continuous and nonnegative, it must attain a minimum over $\mathscr{C}$, $x_t\in\arg\min_{x\in\mathscr{C}}\mathsf{f}(x)$. Since $\mathscr{C}_{\mathsf{dgn}}$ is a closed subset of $\mathscr{C}$, this means that $x_t\in\mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$ for any sufficiently small $t>0$, since

$$\|x_t-u\|\leq\|x_t-u_t\|+\|u_t-u\|\leq 2\|u-u_t\|=2t\|u-y\|$$

(where the second inequality uses the definition of x_t), and so $x_t\rightarrow u$.

Next, consider the subdifferential $\partial\mathsf{f}(x_t)$. It is well known that this subdifferential is not empty, and any element $v\in\partial\mathsf{f}(x_t)$ must satisfy $\|v\|^*\leq 1$ and $\langle v, x_t-u_t\rangle=\|x_t-u_t\|$. Now, applying the stronger form of the first-order optimality condition given in (3.32), we have

$$\langle v, y-x_t\rangle\geq-\gamma_{x_t}^{\text{IP}}(\mathscr{C})\|v\|^*\|y-x_t\|_2^2=-\gamma_{x_t}^{\text{IP}}(\mathscr{C})\|y-x_t\|_2^2$$

and similarly, replacing $y\in\mathscr{C}$ with $u\in\mathscr{C}$,

$$\langle v, u-x_t\rangle\geq-\gamma_{x_t}^{\text{IP}}(\mathscr{C})\|u-x_t\|_2^2.$$

Taking the appropriate linear combination of these two inequalities,

$$\langle v, x_t - u_t\rangle \leq \gamma^{\mathrm{IP}}_{x_t}(\mathscr{C})\left((1-t)\|u-x_t\|_2^2 + t\|y-x_t\|_2^2\right) = \gamma^{\mathrm{IP}}_{x_t}(\mathscr{C})\left(t(1-t)\|u-y\|_2^2 + \|u_t - x_t\|_2^2\right),$$

where the last step simply uses the definition $u_t = (1-t)u + ty$ and rearranges terms. Finally, $\|u_t - x_t\|_2 \leq B_{\mathrm{norm}}\|u_t - x_t\| \leq B_{\mathrm{norm}}\|u - u_t\| = tB_{\mathrm{norm}}\|u-y\|$, by definition of u_t and x_t, so combining everything we can write

$$\min_{x\in\mathscr{C}}\|x-u_t\| = \|x_t - u_t\| = \langle v, x_t - u_t\rangle \leq \gamma^{\mathrm{IP}}_{x_t}(\mathscr{C})\left(t(1-t)\|u-y\|_2^2 + t^2 B_{\mathrm{norm}}^2\|u-y\|^2\right).$$

Dividing by t and taking a limit,

$$\lim_{t\searrow 0}\frac{\min_{x\in\mathscr{C}}\|x-u_t\|}{t} \leq \left(\limsup_{t\searrow 0}\gamma^{\mathrm{IP}}_{x_t}(\mathscr{C})\right)\cdot\|u-y\|_2^2.$$

Finally, recall that $x \mapsto \gamma^{\mathrm{IP}}_x(\mathscr{C})$ is upper semi-continuous by Lemma 3.5.1, and $x_t \to u$ as proved above. We thus have $\limsup_{t\searrow 0}\gamma^{\mathrm{IP}}_{x_t}(\mathscr{C}) \leq \gamma^{\mathrm{IP}}_u(\mathscr{C})$. This proves that $\gamma^{\mathrm{curv}}_u(\mathscr{C}) \leq \gamma^{\mathrm{IP}}_u(\mathscr{C})$, for any $u \in \mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$.

Combining with our previous steps, we now have

$$\gamma^{\mathrm{FO}}_x(\mathscr{C}) = \gamma^{\mathrm{IP}}_x(\mathscr{C}) = \gamma^{\mathrm{curv}}_x(\mathscr{C})$$

for all $x \in \mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$, while for $x \in \mathscr{C}$ we have the weaker statement $\gamma^{\mathrm{IP}}_x(\mathscr{C}) \leq \min\{\gamma^{\mathrm{curv}}_x(\mathscr{C}), \gamma^{\mathrm{FO}}_x(\mathscr{C})\}$.

□

Approximate contraction ⇔ Inner products. This proof, for the case of a general norm $\|\cdot\|$, proceeds identically as the proof for the case where $\|\cdot\| = \|\cdot\|_2$ (presented e.g. in [32, Theorem 3(b,d)]). For completeness, we reproduce the argument here.

First, we show that $\gamma^{\mathrm{IP}}_x(\mathscr{C}) \leq \gamma^{\mathrm{contr}}_x(\mathscr{C})$. Fix any $x \in \mathscr{C}$, and any $z \in \mathbb{R}^d$ with $x = \mathscr{P}_{\mathscr{C}}(z)$. Define $z_t = t\cdot z + (1-t)\cdot x$ for $t \in [0,1]$. By (3.26), $x = \mathscr{P}_{\mathscr{C}}(z_t)$ for all $t \in [0,1]$.

Then for any $y \in \mathscr{C}$, since $\|z_t - x\|^* = t\|z-x\|^*$,

$$\|y-x\|_2 \left(1 - \gamma_x^{\text{contr}}(\mathscr{C}) \cdot t\|z-x\|^*\right) \leq \|y - z_t\|_2$$

by the approximate contraction property (3.3). For sufficiently small t, the left-hand side is nonnegative (except for the trivial case $\gamma_x^{\text{contr}}(\mathscr{C}) = \infty$, in which case there is nothing to prove). Squaring both sides and rearranging some terms,

$$\|y-x\|_2^2 \leq \|y-z_t\|_2^2 + (2\gamma_x^{\text{contr}}(\mathscr{C}) \cdot t\|z-x\|^* - (\gamma_x^{\text{contr}}(\mathscr{C}) \cdot t\|z-x\|^*)^2)\|y-x\|_2^2.$$

And,

$$\|y-z_t\|_2^2 = \|y-x\|_2^2 + \|x-z_t\|_2^2 + 2\langle y-x, x-z_t\rangle$$

so rearranging terms again,

$$2\langle y-x, z_t - x\rangle \leq \|x-z_t\|_2^2 + (2\gamma_x^{\text{contr}}(\mathscr{C}) \cdot t\|z-x\|^* - (\gamma_x^{\text{contr}}(\mathscr{C}) \cdot t\|z-x\|^*)^2)\|y-x\|_2^2.$$

Plugging in the definition of z_t,

$$2t\langle y-x, z-x\rangle \leq t^2\|x-z\|_2^2 + (2\gamma_x^{\text{contr}}(\mathscr{C}) \cdot t\|z-x\|^* - \gamma_x^{\text{contr}}(\mathscr{C})^2 \cdot t^2(\|z-x\|^*)^2)\|y-x\|_2^2.$$

Dividing by $2t$, then taking the limit as $t \searrow 0$,

$$\langle y-x, z-x\rangle \leq \gamma_x^{\text{contr}}(\mathscr{C})\|z-x\|^*\|y-x\|_2^2.$$

Therefore, for any $x \in \mathscr{C}$, $\gamma_x^{\text{IP}}(\mathscr{C}) \leq \gamma_x^{\text{contr}}(\mathscr{C})$.

Now we prove the reverse inequality, i.e. $\gamma_x^{\text{contr}}(\mathscr{C}) \leq \gamma_x^{\text{IP}}(\mathscr{C})$. Fix any $x, y \in \mathscr{C}$ and any $z \in \mathbb{R}^d$ with $x = \mathscr{P}_{\mathscr{C}}(z)$. Then

$$\|y-x\|_2^2 + \langle y-x, z-y\rangle = \langle y-x, z-x\rangle \leq \gamma_x^{\text{IP}}(\mathscr{C})\|z-x\|^*\|y-x\|_2^2.$$

Simplifying,

$$\left(1-\gamma_x^{\mathrm{IP}}(\mathscr{C})\|z-x\|^*\right)\|y-x\|_2^2 \le -\langle y-x, z-y\rangle \le \|y-x\|_2\|z-y\|_2,$$

and so

$$\left(1-\gamma_x^{\mathrm{IP}}(\mathscr{C})\|z-x\|^*\right)\|y-x\|_2 \le \|z-y\|_2.$$

Therefore, for any $x \in \mathscr{C}$ $\gamma_x^{\mathrm{contr}}(\mathscr{C}) \le \gamma_x^{\mathrm{IP}}(\mathscr{C})$.

Combining everything, we have now proved

$$\gamma_x^{\mathrm{contr}}(\mathscr{C}) = \gamma_x^{\mathrm{IP}}(\mathscr{C}) = \gamma_x^{\mathrm{FO}}(\mathscr{C}) = \gamma_x^{\mathrm{curv}}(\mathscr{C})$$

for all $x \in \mathscr{C}\backslash\mathscr{C}_{\mathsf{dgn}}$, in addition to the weaker bound (3.28) for all $x \in \mathscr{C}$, as desired. This completes the proof of Theorem 3.2.1.

□

3.5.4 Characterization of degenerate points

Proof of Lemma 3.2.3. Next we prove that the degenerate points $u \in \mathscr{C}_{\mathsf{dgn}}$ are precisely those points where any of the four local concavity conditions would fail to hold, in any neighborhood of u and for any finite γ. First, the characterization of prox-regularity given in [42, Proposition 1.2, Theorem 1.3(i)] proves that, if the projection operator $\mathscr{P}_{\mathscr{C}}$ is *not* continuous in a neighborhood of $u \in \mathscr{C}$, then there are no constants $\varepsilon > 0$ and $\gamma < \infty$ such that the inner product condition (3.5) holds for all $x \in \mathscr{C} \cap \mathbb{B}_2(u,\varepsilon)$. Therefore, for any $r > 0$, $\sup_{x\in\mathscr{C}\cap\mathbb{B}_2(u,r)} \gamma_x^{\mathrm{IP}}(\mathscr{C}) = \infty$.

Finally, in proving Theorem 3.2.1, we proved (3.28), i.e.

$$\gamma_x^{\mathrm{IP}}(\mathscr{C}) \le \min\{\gamma_x^{\mathrm{curv}}(\mathscr{C}), \gamma_x^{\mathrm{contr}}(\mathscr{C}), \gamma_x^{\mathrm{FO}}(\mathscr{C})\}$$

for all $x \in \mathscr{C}$. This implies that,

$$\lim_{r \to 0} \left\{ \sup_{x \in \mathscr{C} \cap \mathbb{B}_2(u,r)} \gamma_x^{(*)}(\mathscr{C}) \right\} = \infty,$$

where (*) denotes any of the four properties, i.e. $\gamma_x^{\text{curv}}(\mathscr{C})$ for the curvature condition (3.1), $\gamma_x^{\text{contr}}(\mathscr{C})$ for the contraction property (3.3), $\gamma_x^{\text{IP}}(\mathscr{C})$ for the inner product condition (3.5), or $\gamma_x^{\text{FO}}(\mathscr{C})$ for the first-order optimality condition (3.4). This proves the lemma. □

3.5.5 Two-sided contraction property

Proof of Lemma 3.2.4. This proof, for the case of a general norm $\|\cdot\|$, proceeds identically as the proof for the case where $\|\cdot\| = \|\cdot\|_2$ (presented e.g. in [32, Theorem 3(b,d)]). For completeness, we reproduce the argument here.

Take any $x, y \in \mathscr{C}$ and any $z, w \in \mathbb{R}^d$ with $\mathscr{P}_{\mathscr{C}}(z) = x$ and $\mathscr{P}_{\mathscr{C}}(w) = x$. By definition of the local concavity coefficients, applying the inner product bound (3.5) we have

$$\langle y - x, z - x \rangle \leq \gamma_x(\mathscr{C}) \|z - x\|^* \|x - y\|_2^2.$$

Applying the same property with the roles of the variables reversed,

$$\langle x - y, w - y \rangle \leq \gamma_y(\mathscr{C}) \|w - y\|^* \|x - y\|_2^2.$$

Adding these two inequalities together,

$$\langle y - x, z - x - w + y \rangle \leq \gamma_x(\mathscr{C}) \|z - x\|^* \|x - y\|_2^2 + \gamma_y(\mathscr{C}) \|w - y\|^* \|x - y\|_2^2.$$

Rearranging terms and simplifying,

$$\left(1 - \gamma_x(\mathscr{C}) \|z - x\|^* - \gamma_y(\mathscr{C}) \|w - y\|^*\right) \|x - y\|_2^2 \leq \langle y - x, w - z \rangle.$$

Since the right-hand side is bounded by $\|x-y\|_2\|z-w\|_2$ by the Cauchy–Schwarz inequality, this proves the lemma. □

3.5.6 Equivalence for global concavity and local vs global coefficients

Proofs of Theorem 3.1.1 and Lemma 3.2.1. We prove Theorem 3.1.1, which states that the five definitions for the global concavity coefficient $\gamma(\mathscr{C})$ are equivalent, alongside Lemma 3.2.1, which states that $\gamma(\mathscr{C}) = \sup_{x\in\mathscr{C}} \gamma_x(\mathscr{C})$.

First, suppose that $\mathscr{C}$ contains one or more degenerate points, i.e. $\mathscr{C}_{\mathsf{dgn}} \neq \varnothing$, in which case $\sup_{x\in\mathscr{C}} \gamma_x(\mathscr{C}) = \infty$. By definition of $\mathscr{C}_{\mathsf{dgn}}$, the projection operator $\mathscr{P}_{\mathscr{C}}$ is not continuous on any neighborhood of $\mathscr{C}$. [42, Theorem 4.1] prove that this implies $\mathscr{C}$ is not prox-regular, and so $\gamma(\mathscr{C}) = \infty$ as discussed in Section 3.2.2.

Next, suppose that $\mathscr{C}$ contains no degenerate points. Let $\gamma^* = \sup_{x\in\mathscr{C}} \gamma_x(\mathscr{C})$. Then clearly, by definition of the local coefficients $\gamma_x(\mathscr{C})$,

$$\gamma^* = \min\{\gamma \in [0,\infty] : \text{ Property (*) holds for all } x,y \in \mathscr{C}\}$$

where (*) may refer to any of the four equivalent properties, namely the curvature condition (3.1), the (one-sided) contraction property (3.3), the inner product condition (3.5), and the first-order condition (3.4). Next, let

$$\gamma^\sharp = \min\{\gamma \in [0,\infty] : \text{ The two-sided contraction property (3.2) holds for all } x,y \in \mathscr{C}\}.$$

Clearly, the two-sided contraction property (3.2) is stronger than its one-sided version (3.3), and so $\gamma^* \leq \gamma^\sharp$. However, Lemma 3.2.4 shows that they are in fact equal, proving that

$$\big(1-\gamma_x(\mathscr{C})\|z-x\|^* - \gamma_y(\mathscr{C})\|w-y\|^*\big)\cdot\|x-y\|_2 \leq \|z-w\|_2$$

for all $z,w \in \mathbb{R}^d$ with $x = \mathscr{P}_{\mathscr{C}}(z)$, $y = \mathscr{P}_{\mathscr{C}}(w)$. Since $\gamma_x(\mathscr{C}), \gamma_y(\mathscr{C}) \leq \gamma^*$ for all $x,y \in \mathscr{C}$, this implies that

$$\left(1-\gamma^*\|z-x\|^*-\gamma^*\|w-y\|^*\right)\cdot\|x-y\|_2 \le \|z-w\|_2,$$

that is, (3.2) holds for all $x,y\in\mathscr{C}$ with constant $\gamma=\gamma^*$. So, we have $\gamma^\sharp\le\gamma^*$. Therefore, the five conditions defining $\gamma(\mathscr{C})$ are equivalent, and $\gamma(\mathscr{C})=\gamma^\sharp=\gamma^*=\sup_{x\in\mathscr{C}}\gamma_x(\mathscr{C})$, proving Theorem 3.1.1 and Lemma 3.2.1.

□

3.6 Proofs for examples

In this section we prove results calculating the local concavity coefficients $\gamma_x(\mathscr{C})$ for the constraint sets considered in Section 3.3.

Tangent space For any rank-r matrix X, let T_X be the tangent space of low-rank matrices at X, given by[5]

$$T_X=\left\{UA^\top+BV^\top\ :\ A\in\mathbb{R}^{m\times r}, B\in\mathbb{R}^{n\times r}\text{ are any matrices}\right\}, \tag{3.33}$$

where $U\in\mathbb{R}^{n\times r}$, $V\in\mathbb{R}^{m\times r}$ are orthonormal bases for the column and row span of X. This tangent space has frequently been studied in the context of nuclear norm minimization, see for instance [43]. This definition has also appeared in Chapter 2 when we analyze error bound of the estimator for compressed RPCA.

3.6.1 Low rank constraints

Recalling the subspace T_X defined in (3.33) for any rank-r matrix X, we begin with an auxiliary lemma:

[5] For $X\in\mathscr{C}$ which is of rank strictly lower than r, we can define T_X by taking $U\in\mathbb{R}^{n\times r}$, $V\in\mathbb{R}^{m\times r}$ to be any orthonormal matrices which contain the column and row span of X; this choice is not unique, but formally we assume that we have fixed some choice of space T_X for each $X\in\mathscr{C}$.

Lemma 3.6.1. *Let* $X, Y \in \mathbb{R}^{n\times m}$ *satisfy* $\mathrm{rank}(X), \mathrm{rank}(Y) \leq r$. *Then*

$$\|\mathscr{P}_{T_X}^{\perp}(Y)\|_{\mathrm{nuc}} \leq \frac{1}{2\sigma_r(X)}\|X-Y\|_{\mathsf{F}}^2.$$

Proof of Lemma 3.6.1. Assume $\sigma_r(X) > 0$ (otherwise the statement is trivial). For any matrix $M \in (T_X)^{\perp}$ with $\|M\|_{\mathrm{sp}} \leq 1$, define a function

$$\mathsf{f}_M(Z) = \frac{1}{2\sigma_r(X)}\|Z-X\|_{\mathsf{F}}^2 - \langle Z, M\rangle$$

over matrices $Z \in \mathbb{R}^{n\times m}$. We can rewrite this as

$$\mathsf{f}_M(Z) = \frac{1}{2\sigma_r(X)}\|Z-(X+\sigma_r(X)M)\|_{\mathsf{F}}^2 + \langle X, M\rangle - \frac{\sigma_r(X)}{2}\|M\|_{\mathsf{F}}^2.$$

Now, we minimize $\mathsf{f}_M(Z)$ over a rank constraint:

$$\mathop{\arg\min}_{\mathrm{rank}(Z)\leq r} \mathsf{f}_M(Z) = \mathop{\arg\min}_{Z}\{\|Z-(X+\sigma_r(X)M)\|_{\mathsf{F}}^2 : \mathrm{rank}(Z) \leq r\} = \mathscr{P}_{\mathscr{C}}(X+\sigma_r(X)M).$$

Since $\sigma_1(X), \ldots, \sigma_r(X) \geq \sigma_r(X)$ while $\|\sigma_r(X)M\|_{\mathrm{sp}} \leq \sigma_r(X)$, and $M \in (T_X)^{\perp}$, we see that

$$X = \mathscr{P}_{\mathscr{C}}(X+\sigma_r(X)M).$$

(It may be the case that X and $\sigma_r(X)M$ both have one or more singular values exactly equal to $\sigma_r(X)$, in which case the projection is not unique, but X is always one of the values of the projection.) So, $Z = X$ minimizes $\mathsf{f}_M(Z)$ over rank-r matrices, and therefore, for any Z with $\mathrm{rank}(Z) \leq r$,

$$\mathsf{f}_M(Z) \geq \mathsf{f}_M(X) = \frac{1}{2\sigma_r(X)}\|X-X\|_{\mathsf{F}}^2 + \langle X, M\rangle = 0,$$

since $\langle X, M\rangle = 0$ due to $M \in (T_X)^\perp$. Therefore, in particular, $\mathsf{f}_M(Y) \geq 0$ which implies that

$$\langle Y, M\rangle \leq \frac{1}{2\sigma_r(X)}\|Y - X\|_\mathsf{F}^2.$$

Since this is true for all $M \in (T_X)^\perp$ with $\|M\|_\mathrm{sp} \leq 1$, we have proved that

$$\|\mathscr{P}_{T_X}^\perp(Y)\|_\mathrm{nuc} = \max_{M\in(T_X)^\perp, \|M\|_\mathrm{sp}=1} \langle Y, M\rangle \leq \frac{1}{2\sigma_r(X)}\|Y - X\|_\mathsf{F}^2,$$

as desired. □

Proof of Lemma 3.3.1. First, let $\mathscr{P}_\mathscr{C}(Z)$ be any closest rank-r matrix to Z (not necessarily unique), and let $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$ be orthonormal bases for the column span and row span of $\mathscr{P}_\mathscr{C}(Z)$ (that is, if $\mathscr{P}_\mathscr{C}(Z)$ is unique then the columns of U and V are the top r left and right singular vectors of Z). Regardless of uniqueness we will have $Z - \mathscr{P}_\mathscr{C}(Z)$ orthogonal to U on the left and to V on the right, i.e. we can write

$$Z - \mathscr{P}_\mathscr{C}(Z) = (\mathbf{I} - UU^\top)\cdot(Z - \mathscr{P}_\mathscr{C}(Z))\cdot(\mathbf{I} - VV^\top).$$

We then have

$$\begin{aligned}\langle Y - \mathscr{P}_\mathscr{C}(Z), Z - \mathscr{P}_\mathscr{C}(Z)\rangle &= \langle Y - \mathscr{P}_\mathscr{C}(Z), (\mathbf{I} - UU^\top)\cdot(Z - \mathscr{P}_\mathscr{C}(Z))\cdot(\mathbf{I} - VV^\top)\rangle\\ &= \langle(\mathbf{I} - UU^\top)\cdot(Y - \mathscr{P}_\mathscr{C}(Z))\cdot(\mathbf{I} - VV^\top), Z - \mathscr{P}_\mathscr{C}(Z)\rangle\\ &\leq \|(\mathbf{I} - UU^\top)\cdot(Y - \mathscr{P}_\mathscr{C}(Z))\cdot(\mathbf{I} - VV^\top)\|_\mathrm{nuc}\cdot\|Z - \mathscr{P}_\mathscr{C}(Z)\|_\mathrm{sp}\\ &\leq \|(\mathbf{I} - UU^\top)\cdot Y\cdot(\mathbf{I} - VV^\top)\|_\mathrm{nuc}\cdot\|Z - \mathscr{P}_\mathscr{C}(Z)\|_\mathrm{sp},\end{aligned}$$

where the last step holds since $\mathscr{P}_\mathscr{C}(Z)$ is spanned by U on the left and V on the right. Applying Lemma 3.6.1 with $X = \mathscr{P}_\mathscr{C}(Z)$, which trivially has U, V as its left and right singular vectors, we obtain

$$\|(\mathbf{I} - UU^\top)\cdot Y\cdot(\mathbf{I} - VV^\top)\|_\mathrm{nuc} \leq \frac{1}{2\sigma_r(\mathscr{P}_\mathscr{C}(Z))}\|Y - \mathscr{P}_\mathscr{C}(Z)\|_\mathsf{F}^2.$$

Therefore,

$$\langle Y - \mathscr{P}_{\mathscr{C}}(Z), Z - \mathscr{P}_{\mathscr{C}}(Z)\rangle \leq \frac{1}{2\sigma_r(\mathscr{P}_{\mathscr{C}}(Z))}\|Z - \mathscr{P}_{\mathscr{C}}(Z)\|_{\mathsf{sp}}\|Y - \mathscr{P}_{\mathscr{C}}(Z)\|_{\mathsf{F}}^2.$$

This proves that $\gamma_X(\mathscr{C}) \leq \frac{1}{2\sigma_r(X)}$ for all $x \in \mathscr{C}$ by the inner product condition (3.5).

To prove equality, take any $X \in \mathscr{C} \backslash \mathscr{C}_{\mathsf{dgn}}$ (that is, we assume that $\operatorname{rank}(X) = r$), and let $X = \sigma_1 u_1 v_1^\top + \cdots + \sigma_r u_r v_r^\top$ be a singular value decomposition with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Let $u' \in \mathbb{R}^n, v' \in \mathbb{R}^m$ be unit vectors orthogonal to the left and right singular vectors of X, respectively. Define

$$Y = \sigma_1 u_1 v_1^\top + \cdots + \sigma_{r-1} u_{r-1} v_{r-1}^\top + \sigma_r u' v'^\top$$

and

$$Z = X + cu'v'^\top,$$

for some fixed $c \in (0, \sigma_r)$. Then $\mathscr{P}_{\mathscr{C}}(Z) = X$, and we have

$$\langle Y - X, Z - X\rangle = \langle \sigma_r u' v'^\top - \sigma_r u_r v_r^\top, cu'v'^\top\rangle = c\sigma_r$$

while

$$\|Z - X\|_{\mathsf{sp}}\|Y - X\|_{\mathsf{F}}^2 = \|cu'v'^\top\|_{\mathsf{sp}}\|\sigma_r u' v'^\top - \sigma_r u_r v_r^\top\|_{\mathsf{F}}^2 = 2c\sigma_r^2,$$

therefore by the inner product condition (3.5), we must have $\gamma_X(\mathscr{C}) \geq \frac{1}{2\sigma_r(X)}$. Combining with our previous steps, we now have $\gamma_X(\mathscr{C}) = \frac{1}{2\sigma_r(X)}$ for all $x \in \mathscr{C}$, proving Lemma 3.3.1.

□

3.6.2 Sparsity

Proof of Lemma 3.3.2. We check the local concavity coefficients. Fix any $x \in \mathscr{C}$. As before, if x is in the interior (i.e. $\operatorname{Pen}(x) < c$) then $\gamma_x(\mathscr{C}) = 0$, so we turn to the case that $\operatorname{Pen}(x) = c$, and in particular, $x \neq 0$. Without loss of generality, assume that $x_1 > 0$ and that x_1 is the smallest nonzero coordinate of x (and then $x_{\min} = x_1$). Choose any $y \in \mathscr{C}$ and $t \in [0,1]$. Let

$$x_t = (1-t)x + ty \text{ and } z_t = x_t - s_t \mathbf{e}_1$$

where $\mathbf{e}_1 = (1,0,\ldots,0)$ and

$$s_t = t \cdot \frac{\mu/2}{\mathsf{p}'((x_t)_1)} \cdot \|x-y\|_2^2.$$

Since $\lim_{t \searrow} x_t = x$, and p is continuously differentiable (since it is both concave and differentiable on the positive real line), we have

$$\lim_{t \searrow 0} \frac{s_t}{t} = \frac{\mu/2}{\mathsf{p}'(x_1)} \cdot \|x-y\|_2^2.$$

In particular, this implies that, for sufficiently small t, we have $(x_t)_1 > 0$ and $(z_t)_1 > 0$.

We claim that $\operatorname{Pen}(z_t) \leq c$, in which case

$$\lim_{t \searrow 0} \frac{\min_{x' \in \mathscr{C}} \|x_t - x'\|_1}{t} \leq \lim_{t \searrow 0} \frac{\|x_t - z_t\|_1}{t} = \lim_{t \searrow 0} \frac{s_t}{t} = \frac{\mu/2}{\mathsf{p}'(x_{\min})} \cdot \|x-y\|_2^2,$$

which proves the lemma.

It now remains to check that $\operatorname{Pen}(z_t) \leq c$. We have, for coordinate $i = 1$,

$$\mathsf{p}(|z_t|_i) = \mathsf{p}((x_t)_1 - s_t) \leq \mathsf{p}((x_t)_1) - s_t \mathsf{p}'((x_t)_1),$$

since $0 < (x_t)_1 - s_t < (x_t)_1$ and ρ is concave over $\mathbb{R}_+$. And, for every coordinate i,

$$\begin{aligned}
\mathsf{p}(|(x_t)_i|) &= \mathsf{p}\left(|(1-t)x_i + ty_i|\right) \\
&\leq \mathsf{p}\left((1-t)|x_i| + t|y_i|\right) \quad \text{since } \rho \text{ is nondecreasing} \\
&\leq (1-t)\mathsf{p}(|x_i|) + t\mathsf{p}(|y_i|) + \frac{\mu}{2}t(1-t)(|x_i| - |y_i|)^2 \quad \text{since } t \mapsto \mathsf{p}(t) + \mu t^2/2 \text{ is convex} \\
&\leq (1-t)\mathsf{p}(|x_i|) + t\mathsf{p}(|y_i|) + \frac{\mu}{2}t(1-t)(x_i - y_i)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Pen}(z_t) = \sum_i \mathsf{p}(|z_t|_i) &\le \left(\sum_i (1-t)\mathsf{p}(|x_i|) + t\mathsf{p}(|y_i|) + \frac{\mu}{2}t(1-t)(x_i-y_i)^2\right) - s_t\mathsf{p}'((x_t)_1)\\ &\le (1-t)\mathrm{Pen}(x) + t\mathrm{Pen}(y) + \frac{\mu}{2}\|x-y\|_2^2 - s_t\mathsf{p}'((x_t)_1) \le c + \frac{\mu}{2}\|x-y\|_2^2 - s_t\mathsf{p}'((x_t)_1) = c,\end{aligned}$$

where the last step holds by definition of s_t. □

3.6.3 Other examples

Proof of Lemma 3.3.3. Let $X, Y \in \mathscr{C}$. For a fixed $t \in (0,1)$, let $(1-t)X + tY = ADB^\top$ be a singular value decomposition. Since $AB^\top \in \mathbb{R}^{n\times r}$ is an orthonormal matrix, we then have

$$\begin{aligned}\min_{Z\in\mathscr{C}} \|Z - ((1-t)X + tY)\|_{\mathrm{nuc}} \le \|AB^\top - ((1-t)X+tY)\|_{\mathrm{nuc}} &= \|AB^\top - ADB^\top\|_{\mathrm{nuc}}\\ &= \sum_{i=1}^r (1-D_{ii}).\end{aligned}$$

Furthermore,

$$\begin{aligned}\|D\|_{\mathsf{F}}^2 &= \|(1-t)X+tY\|_{\mathsf{F}}^2\\ &= (1-t)^2\|X\|_{\mathsf{F}}^2 + t^2\|Y\|_{\mathsf{F}}^2 + 2t(1-t)\langle X,Y\rangle\\ &= (1-t)^2\|X\|_{\mathsf{F}}^2 + t^2\|Y\|_{\mathsf{F}}^2 + t(1-t)\left(\|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2 - \|X-Y\|_{\mathsf{F}}^2\right)\\ &= (1-t)^2 r + t^2 r + t(1-t)\left(r + r - \|X-Y\|_{\mathsf{F}}^2\right)\\ &= r - t(1-t)\|X-Y\|_{\mathsf{F}}^2.\end{aligned}$$

A trivial calculation shows that $1-D_{ii} = \frac{1-D_{ii}^2}{2} + \frac{(1-D_{ii})^2}{2}$, so we have

$$\begin{aligned}\min_{Z\in\mathscr{C}} \|Z-((1-t)X+tY)\|_{\text{nuc}} &\leq \sum_{i=1}^{r}(1-D_{ii}) = \sum_{i=1}^{r} \frac{1-D_{ii}^2}{2} + \frac{(1-D_{ii})^2}{2} \\ &= \frac{r-\|D\|_{\mathsf{F}}^2}{2} + \sum_{i=1}^{r} \frac{(1-D_{ii})^2}{2} = \frac{1}{2}t(1-t)\|X-Y\|_{\mathsf{F}}^2 + \sum_{i=1}^{r} \frac{(1-D_{ii})^2}{2}.\end{aligned}$$

Furthermore, we can show that the last term is $o(t)$, as follows. For any unit vector $u \in \mathbb{R}^r$,

$$\|((1-t)X+tY)u\|_2 \geq (1-t)\|Xu\|_2 - t\|Yu\|_2 \geq 1-2t$$

since X,Y are both orthonormal. Therefore $(1-t)X+tY$ has all its singular values $\geq 1-2t$, that is, $D_{ii} \geq 1-2t$ for all i. And trivially $\|((1-t)X+tY)u\|_2 \leq 1$ so $D_{ii} \leq 1$. Then $\sum_{i=1}^r (1-D_{ii})^2 \leq \sum_{i=1}^r (2t)^2 = 4t^2 r$, so we have

$$\min_{Z\in\mathscr{C}} \|Z-((1-t)X+tY)\|_{\text{nuc}} \leq \frac{1}{2}t(1-t)\|X-Y\|_{\mathsf{F}}^2 + 2t^2 r.$$

Dividing by t and taking a limit,

$$\lim_{t\searrow 0} \frac{\min_{Z\in\mathscr{C}} \|Z-((1-t)X+tY)\|_{\text{nuc}}}{t} \leq \frac{1}{2}\|X-Y\|_{\mathsf{F}}^2.$$

Comparing to the curvature condition (3.1) we see that $\gamma_X(\mathscr{C}) \leq \frac{1}{2}$, as desired.

Next, to obtain equality, take any $X \in \mathscr{C}$. Fix any $c \in (0,1)$. Let $Y = -X \in \mathscr{C}$ and $Z = cX \in \mathbb{R}^{n\times r}$. Clearly, $\mathscr{P}_{\mathscr{C}}(Z) = X$. By the contraction property (3.3), we must have

$$(1-\gamma_X(\mathscr{C})\|Z-X\|^*)\|Y-X\|_{\mathsf{F}} \leq \|Y-Z\|_{\mathsf{F}}.$$

Plugging in our choices for Y and Z, we obtain

$$(1-\gamma_X(\mathscr{C})\cdot(1-c))\cdot 2\sqrt{r} \leq (1+c)\sqrt{r},$$

and so $\gamma_X(\mathscr{C}) \geq \frac{1}{2}$.

$\square$

Proof of Lemma 3.3.4. For $X, Y \in \mathscr{C}$, write $X = UU^\top$, and $Y = VV^\top$ for some orthonormal matrices $U, V \in \mathbb{R}^{n \times r}$. For $t \in (0,1)$, let $U_t = (1-t)U + tV$, and let $U_t = ADB^\top$ be a singular value decomposition. Then $AB^\top$ is the projection of U_t onto the set of orthonormal $n \times r$ matrices. Since $A \in \mathbb{R}^{n \times r}$ is orthonormal, we have $AA^\top \in \mathscr{C}$, and so

$$\begin{aligned}\min_{Z \in \mathscr{C}} \|Z - ((1-t)X + tY)\|_{\text{nuc}} &\leq \|AA^\top - ((1-t)X + tY)\|_{\text{nuc}} \\ &\leq \underbrace{\|AA^\top - U_t U_t^\top\|_{\text{nuc}}}_{\text{(Term 1)}} + \underbrace{\|U_t U_t^\top - ((1-t)X + tY)\|_{\text{nuc}}}_{\text{(Term 2)}}.\end{aligned}$$

For (Term 1),

$$\begin{aligned}\|AA^\top - U_t U_t^\top\|_{\text{nuc}} &= \|AA^\top - ADB^\top \cdot BDA^\top\|_{\text{nuc}} \\ &= \|A(\mathbf{I}_r - D^2)A^\top\|_{\text{nuc}} \\ &= r - \|D\|_{\mathsf{F}}^2 = r - \|U_t\|_{\mathsf{F}}^2 \\ &= r - \|(1-t)U + tV\|_{\mathsf{F}}^2 \\ &= r - (1-t)^2\|U\|_{\mathsf{F}}^2 - t^2\|V\|_{\mathsf{F}}^2 - 2t(1-t)\langle U, V\rangle \\ &= r - (1-t)^2\|U\|_{\mathsf{F}}^2 - t^2\|V\|_{\mathsf{F}}^2 - t(1-t)\left(\|U\|_{\mathsf{F}}^2 + \|V\|_{\mathsf{F}}^2 - \|U - V\|_{\mathsf{F}}^2\right) \\ &= r - (1-t)^2 r - t^2 r - t(1-t)\left(2r - \|U - V\|_{\mathsf{F}}^2\right) \\ &= t(1-t)\|U - V\|_{\mathsf{F}}^2.\end{aligned}$$

For (Term 2),

$$\begin{aligned}
&\|U_tU_t^\top - ((1-t)X + tY)\|_{\text{nuc}} \\
&= \|((1-t)U + tV)((1-t)U + tV)^\top - (1-t)UU^\top - tVV^\top\|_{\text{nuc}} \\
&= \|{-t}(1-t)UU^\top - t(1-t)VV^\top + t(1-t)UV^\top + t(1-t)VU^\top\|_{\text{nuc}} \\
&= \|{-t}(1-t)(U-V)(U-V)^\top\|_{\text{nuc}} \\
&= t(1-t)\|U-V\|_{\mathsf{F}}^2.
\end{aligned}$$

Combining the two, then,

$$\min_{Z\in\mathscr{C}} \|Z - ((1-t)X + tY)\|_{\text{nuc}} \le 2t(1-t)\|U-V\|_{\mathsf{F}}^2.$$

Next, note that the choice of U and V is not unique. Fixing any factorizations $X = UU^\top$ and $Y = VV^\top$, let $U^\top V = ADB^\top$ be a singular value decomposition, and let $\widetilde{V} = VBA^\top$. Then $Y = \widetilde{V}\widetilde{V}^\top$, and following the same steps as above we can calculate

$$\min_{Z\in\mathscr{C}} \|Z - ((1-t)X + tY)\|_{\text{nuc}} \le 2t(1-t)\|U-\widetilde{V}\|_{\mathsf{F}}^2.$$

Furthermore,

$$\begin{aligned}
\|U-\widetilde{V}\|_{\mathsf{F}}^2 = \|U\|_{\mathsf{F}}^2 + \|\widetilde{V}\|_{\mathsf{F}}^2 - 2\,\text{trace}(U^\top\widetilde{V}) = 2r - 2\,\text{trace}(U^\top VBA^\top)& \\
= 2r - 2\,\text{trace}(ADB^\top BA^\top) = 2r - 2\,\text{trace}(D).&
\end{aligned}$$

And,

$$\begin{aligned}
\|X-Y\|_{\mathsf{F}}^2 = \|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2 - 2\,\text{trace}(XY) = 2r - 2\,\text{trace}(UU^\top\widetilde{V}\widetilde{V}^\top)& \\
= 2r - 2\|U^\top\widetilde{V}\|_{\mathsf{F}}^2 = 2r - 2\|D\|_{\mathsf{F}}^2 \ge 2r - 2\,\text{trace}(D),&
\end{aligned}$$

since $\|D\|_\mathsf{F}^2 = \sum_i (D_{ii})^2 \leq \sum_i D_{ii}$, as $0 \leq D_{ii} \leq 1$ for all i since U, V are both orthonormal matrices. Therefore, this proves that $\|U - \widetilde{V}\|_\mathsf{F}^2 \leq \|X - Y\|_\mathsf{F}^2$, and so

$$\min_{Z \in \mathscr{C}} \|Z - ((1-t)X + tY)\|_{\mathrm{nuc}} \leq 2t(1-t)\|X - Y\|_\mathsf{F}^2.$$

Based on the curvature condition characterization (3.1) of the local concavity coefficients, we have therefore computed $\gamma_X(\mathscr{C}) \leq 2$, as desired. □

CHAPTER 4

ALTERNATING MINIMIZATION AND INEXACT ALTERNATING MINIMIZATION OVER NONCONVEX CONSTRAINTS

Consider the problem of optimizing a function over two variables, one or both of which is constrained to lie in some constraint set:

$$(\widehat{x}, \widehat{y}) = \arg\min\{\mathscr{L}(x,y) : x \in \mathscr{X}, y \in \mathscr{Y}\}, \tag{4.1}$$

where $\mathscr{X} \subset \mathbb{R}^{d_x}$ and $\mathscr{Y} \subset \mathbb{R}^{d_y}$ reflect our beliefs or desired properties for the x and y variables, while $\mathscr{L}$ is the target function to minimize (for example, a negative log-likelihood, in which case we are searching for the constrained maximum likelihood estimator). This type of problem arises naturally in multiple structured statistical models, including multitask regression and robust principle component analysis, and we have already seen one realization of this setting in compressed RPCA (recall we were working with a regularized form in (2.3), but converting to the equivalent constrained form is straightforward).

Due to its simplicity and effectiveness, alternating minimization has long been a popular optimization method to solve the problem (4.1). Based on the tool of concavity coefficients developed in Chapter 3, we assume that the constraint sets $\mathscr{X}$ and $\mathscr{Y}$ may potentially be *nonconvex*. In particular, this chapter[1] studies the convergence behavior of the (inexact) alternating minimization method over (possibly) nonconvex sets, where we iterate the steps

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

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,

[1] The work presented in this chapter is published in Ha and Barber [44].

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.

In this chapter, we explore theoretical properties of (exact or inexact) alternating minimization as well as its convergence behavior in numerical simulation. We also examine a range of specific examples with rank-constrained variables, including multitask regression, robust principal component analysis, and factor models.

4.1 Handling nonconvex constraints in alternating minimization

Before proceeding, we explain how concavity coefficients can be useful for establishing convergence of the alternating minimization method. One main challenge of working over nonconvex regions is that, since $\mathscr{X}$ and $\mathscr{Y}$ are potentially nonconvex sets, the standard first-order optimality conditions under convex setting do not apply. Specifically, fixing any $y \in \mathscr{Y}$ and defining

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

the nonconvexity of $\mathscr{X}$ means that we cannot assume that $\langle x - x_y, \nabla_x \mathscr{L}(x_y, y)\rangle \geq 0$ for all other $x \in \mathscr{X}$ (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.

Recall our definition of local concavity coefficients, $\gamma_x(\mathscr{X})$ for $x \in \mathscr{X}$, given in (3.6). The equivalent results, Theorem 3.2.1, tells that the concavity coefficient equivalently characterizes the extent to which the usual first-order optimality conditions are violated when minimizing over the set $\mathscr{X}$. Now fix some structured 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 $\|\cdot\|_x, \|\cdot\|_y \geq \|\cdot\|_2$, 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 $\gamma_x(\mathscr{X})$ and $\gamma_y(\mathscr{Y})$ be defined with respect to these potentially different norms. The first-order condition (3.4) 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 $\min\{\mathscr{L}(x,y) : x \in \mathscr{X}\}$ (i.e. the x update step of alternating minimization), then for all $x \in \mathscr{X}$,

$$\langle x - x_y, \nabla_x \mathscr{L}(x_y, y)\rangle \geq -\gamma_{x_y}(\mathscr{X}) \|\nabla_x \mathscr{L}(x_y, y)\|_x^* \|x - x_y\|_2^2, \tag{4.2}$$

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

4.2 Problem formulation

Given the loss function $\mathscr{L}(x,y)$ which is differentiable, we consider an optimization problem given in (4.1). Formally we require the target of our optimization problem $(\widehat{x},\widehat{y})$ only to be a *local* minimizer of $\mathscr{L}(x,y)$—this is because $\mathscr{L}(x,y)$ may potentially be highly nonconvex or degenerate in regions (x,y) far from the origin, and we may even have $\lim \mathscr{L}(x,y) = -\infty$ 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 $\mathscr{X}_0 = \mathscr{X} \cap \mathbb{B}_2(x_0, \rho_x)$ and $\mathscr{Y}_0 = \mathscr{Y} \cap \mathbb{B}_2(y_0, \rho_y)$, where (x_0, y_0) is our initialization point. These neighborhoods of the original constraint sets $\mathscr{X}$ and $\mathscr{Y}$ 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 $\mathscr{L}(x,y)$ is well-behaved over this small region $\mathscr{X}_0 \times \mathscr{Y}_0$.

We then define

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

and run the alternating minimization algorithm locally by iterating the steps

$$\begin{cases} x_t = \arg\min_{x \in \mathscr{X}_0} \mathscr{L}(x, y_{t-1}), \\ y_t = \arg\min_{y \in \mathscr{Y}_0} \mathscr{L}(x_t, y). \end{cases} \tag{4.3}$$

For our intuition, we should interpret these radius constraints, i.e. working in $\mathscr{X}_0$ and $\mathscr{Y}_0$ rather than in $\mathscr{X}$ and $\mathscr{Y}$, 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.

Recall the *global concavity coefficient* $\gamma(\mathscr{X}_0) = \sup_{x \in \mathscr{X}_0} \gamma_x(\mathscr{X}_0)$ and $\gamma(\mathscr{Y}_0) = \sup_{y \in \mathscr{Y}_0} \gamma_y(\mathscr{Y}_0)$ given by Lemma 3.2.1. The following lemma proves that, if the radii ρ_x, ρ_y are chosen to be small, the curvature conditions of $\mathscr{X}$ and $\mathscr{Y}$ are inherited by $\mathscr{X}_0$ and $\mathscr{Y}_0$:

Lemma 4.2.1. *If $\rho_x < \frac{1}{2\max_{x \in \mathscr{X}_0} \gamma_x(\mathscr{X})}$, then $\gamma_x(\mathscr{X}_0) \leq \gamma_x(\mathscr{X})$ for all $x \in \mathscr{X}_0$, and in particular,*

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

The analogous statement holds for y.

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 $x_y = \arg\min\{\mathscr{L}(x,y) : x \in \mathscr{X}_0\}$. Then the first-order condition (3.4) shows the following bound (which we can compare to (4.2)),

$$\langle x' - x_y, \nabla_x \mathscr{L}(x_y, y)\rangle \geq -\gamma(\mathscr{X}_0) \|\nabla_x \mathscr{L}(x_y, y)\|_x^* \|x' - x_y\|_2^2 \text{ for all } x' \in \mathscr{X}_0, \tag{4.4}$$

while Lemma 4.2.1 proves a useful bound for $\gamma(\mathscr{X}_0)$ as long as ρ_x is sufficiently small (and similarly for y).

4.2.1 Related work

Alternating minimization is a classical topic in the optimization literature, dating back to early work (e.g. [45]), and a large body of research has been devoted to understanding the method under various settings (e.g. [46, 6]). 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.

Luo and Tseng [6] prove linear convergence for alternating minimization under convex constraints, assuming that the loss function $\mathscr{L}$ is β-smooth and α-strongly convex in each variable (and, in the case of more than two variables, for the analogous coordinate descent algorithm).

In some settings, the loss function $\mathscr{L}$ may be more well-behaved with respect to one of the variables than the other, in terms of its smoothness and convexity properties. Beck [5] studies alternating minimization for a convex loss $\mathscr{L}(x,y)$ under convex constraints on x and on y, proving that the gap in the loss function values, i.e. the difference $\mathscr{L}(x_t,y_t) - \mathscr{L}(\widehat{x},\widehat{y})$, decays according to the rate $\mathscr{O}\left(\frac{\min\{\beta_x,\beta_y\}}{t}\right)$, where β_x and β_y represent the smoothness parameters of the loss $\mathscr{L}$ 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 [5]'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.

4.2.2 Assumptions

Next we formally establish our assumptions on the loss function $\mathscr{L}(x,y)$ 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:

$$D^2(x;\widehat{x}) = \langle x-\widehat{x}, \nabla_x\mathscr{L}(x,\widehat{y}) - \nabla_x\mathscr{L}(\widehat{x},\widehat{y})\rangle, \text{ and} \tag{4.5}$$

$$D^2(y;\widehat{y}) = \langle y-\widehat{y}, \nabla_y\mathscr{L}(\widehat{x},y) - \nabla_y\mathscr{L}(\widehat{x},\widehat{y})\rangle. \tag{4.6}$$

This divergence has been used also in [39] to prove statistical errors of any local minimum in the sparse regression setting. Note that, if $\mathscr{L}$ 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

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

to accommodate the case where the divergences may be negative.

Throughout we will write $\varepsilon_x, \varepsilon_y \geq 0$ 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 $\sim \varepsilon_x$ 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 $\varepsilon_x, \varepsilon_y$ 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 $\mathscr{L}(x,y)$. Our optimization method works locally in neighborhoods of the initialization point (x_0, y_0), and consequently it is sufficient for us to require the assumptions on $\mathscr{L}(x,y)$ to hold only locally in the regions $\mathscr{X}_0$ and $\mathscr{Y}_0$.

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

$$\langle \begin{pmatrix} x - \widehat{x} \\ y - \widehat{y} \end{pmatrix}, \nabla \mathscr{L}(x,y) - \nabla \mathscr{L}(\widehat{x}, \widehat{y}) \rangle \geq \alpha_x \|x - \widehat{x}\|_2^2 + \alpha_y \|y - \widehat{y}\|_2^2 - \alpha_x \varepsilon_x^2 - \alpha_y \varepsilon_y^2. \tag{4.7}$$

Note that we require joint RSC to hold only at the target $(\widehat{x}, \widehat{y})$. In other regions of $\mathscr{X} \times \mathscr{Y}$, 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 4.2.2. (Restricted smoothness (RSM).) *For all $x \in \mathscr{X}_0$ and all $y \in \mathscr{Y}_0$,*

$$D^2(x; \widehat{x}) \leq \beta_x \|x - \widehat{x}\|_2^2 + \alpha_x \varepsilon_x^2 \quad \text{and} \quad D^2(y; \widehat{y}) \leq \beta_y \|y - \widehat{y}\|_2^2 + \alpha_y \varepsilon_y^2. \tag{4.8}$$

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

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

Assumption 4.2.3. (Cross-product bound.) *For all $x \in \mathscr{X}_0$ and all $y \in \mathscr{Y}_0$,*

$$\begin{aligned}\left|\langle x-\widehat{x}, \nabla_x \mathscr{L}(x,y) - \nabla_x \mathscr{L}(x,\widehat{y})\rangle - \langle y-\widehat{y}, \nabla_y \mathscr{L}(x,y) - \nabla_y \mathscr{L}(\widehat{x},y)\rangle\right| \\ \leq \frac{1}{2}\mu_x \|x-\widehat{x}\|_2^2 + \frac{1}{2}\mu_y \|y-\widehat{y}\|_2^2 + \alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2,\end{aligned}$$

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

To understand this assumption, suppose that $\mathscr{L}$ is twice differentiable. In this case, applying Taylor's theorem to rewrite the above expression in terms of $\nabla^2 \mathscr{L}$, we find that Assumption 4.2.3 holds with

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

where the norm $\|\cdot\|_{\mathrm{sp}}$ is the matrix operator norm (the largest singular value). Since $\mathscr{X}_0$ and $\mathscr{Y}_0$ are bounded via the radii ρ_x, ρ_y, then, this condition is satisfied whenever $\nabla^2_{xy}\mathscr{L}$ is Lipschitz. As a special case, if $\mathscr{L}(x,y)$ is quadratic, then we can trivially take $\mu_x = \mu_y = 0$ since $\nabla^2_{xy}\mathscr{L}$ 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 4.2.4. (Initialization condition.)

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

and

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

Recall that Lemma 4.2.1 provides easy bounds on $\gamma(\mathscr{X}_0)$ and $\gamma(\mathscr{Y}_0)$, as long as the radii ρ_x, ρ_y are chosen to be sufficiently small; furthermore, if $\mathscr{X}$ is convex, then $\gamma(\mathscr{X}_0) = 0$ and so the first bound holds trivially, and similarly for the second bound if $\mathscr{Y}$ is convex. In the nonconvex setting where $\gamma(\mathscr{X}_0)$ and/or $\gamma(\mathscr{Y}_0)$ are nonzero, see [31] for a discussion of the necessity of this type of initialization condition for the related problem of gradient descent in a single variable (see also (3.15) in Section 3.4.1). Roughly speaking, the condition requires that algorithm must be initialized within some neighborhood of the global minimizer—sufficiently close so that, locally, the (restricted) convexity of the loss function $\mathscr{L}(x,y)$ is sufficient to outweigh nonconvexity in the constraints.

4.3 Convergence guarantee

4.3.1 Exact alternating minimization

Now we prove convergence by working with the first-order divergence defined in (4.5), (4.6) above. According to Assumption 4.2.1, the divergence will be always nonnegative in the regions $\mathscr{X}_0$ and $\mathscr{Y}_0$, up to the statistical error.

Theorem 4.3.1. *Suppose that Assumptions 4.2.1, 4.2.2, 4.2.3, and 4.2.4 hold. Then the iterations of the alternating minimization algorithm* (4.3) *satisfy the recursive bounds*

$$D(x_t;\widehat{x}) \leq \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot D(y_{t-1};\widehat{y}) + \sqrt{3(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)}, \text{ and} \tag{4.9}$$

$$D(y_t;\widehat{y}) \leq \sqrt{1-\frac{\alpha_x}{2\beta_x}}\cdot D(x_t;\widehat{x}) + \sqrt{3(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)}, \tag{4.10}$$

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

$$\|(x_t,y_t)-(\widehat{x},\widehat{y})\|_2 \leq \left(\sqrt{1-\frac{\alpha_x}{2\beta_x}}\cdot\sqrt{1-\frac{\alpha_y}{2\beta_y}}\right)^t \cdot \frac{\sqrt{6\beta_y}\rho_y}{\sqrt{\min\{\alpha_x,\alpha_y\}}} + C\cdot\max\{\varepsilon_x,\varepsilon_y\} \tag{4.11}$$

for all $t \geq 1$, where

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

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 (4.11), 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.

4.3.2 Dependence on condition number

Examining the bound (4.11) 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 $C \cdot \max\{\varepsilon_x, \varepsilon_y\}$). 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,

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

and the joint condition number

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

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

In (4.11), 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 $\min\{\kappa_x(\mathscr{L}), \kappa_y(\mathscr{L})\}$ 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 $\kappa(\mathscr{L}) \geq \max\{\kappa_x(\mathscr{L}), \kappa_y(\mathscr{L})\}$. For example, if $\mathscr{L}$ is α-strongly convex and β-smooth in the joint variable (x,y), standard results (see e.g. [40]) 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, $\kappa(\mathscr{L}) \geq \max\{\kappa_x(\mathscr{L}), \kappa_y(\mathscr{L})\}$.

Therefore, in settings where one of the two—$\kappa_x(\mathscr{L})$ or $\kappa_y(\mathscr{L})$—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 $\min\{\kappa_x(\mathscr{L}), \kappa_y(\mathscr{L})\}$. (We mention that [5] finds 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. $\min\{\beta_x, \beta_y\}$.)

4.3.3 Inexact alternating minimization

In some settings, it may be impractical to solve the alternating minimization steps exactly, i.e. when $\mathscr{L}(x,y)$ 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 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 $t \geq 1$,

$$\begin{cases} x_t^{\text{exact}} = \arg\min_{x \in \mathscr{X}_0} \mathscr{L}(x, y_{t-1}), \quad x_t \in \mathscr{X}_0 \cap \mathbb{B}_2(x_t^{\text{exact}}, \delta_t^x), \\ y_t^{\text{exact}} = \arg\min_{y \in \mathscr{Y}_0} \mathscr{L}(x_t, y), \quad y_t \in \mathscr{X}_0 \cap \mathbb{B}_2(y_t^{\text{exact}}, \delta_t^y). \end{cases} \tag{4.12}$$

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 (4.12), we require an additional assumption:

Assumption 4.3.1. (Relaxed triangle inequality.) *For all $x, x' \in \mathscr{X}_0$,*

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

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

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

It can be shown that a stronger form of the restricted smoothness condition (Assumption 4.2.2) 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 4.3.1, along with the relaxed triangle inequality (Assumption 4.3.1).

Theorem 4.3.2. *Suppose that Assumptions 4.2.1, 4.2.2, 4.2.3, 4.2.4, and 4.3.1 hold. Then, the steps of the inexact alternating minimization algorithm satisfy*

$$D(x_t;\widehat{x}) \leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1};\widehat{y}) + \sqrt{\beta_x}\delta_t^x + \sqrt{8(\alpha_x\varepsilon_x^2 + \alpha_y\varepsilon_y^2)} \quad \text{and} \tag{4.13}$$

$$D(y_t;\widehat{y}) \leq \sqrt{1 - \frac{\alpha_x}{2\beta_x}} \cdot D(x_t;\widehat{x}) + \sqrt{\beta_y}\delta_t^y + \sqrt{8(\alpha_x\varepsilon_x^2 + \alpha_y\varepsilon_y^2)}, \tag{4.14}$$

for all $t \geq 1$.

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 $\varepsilon_x, \varepsilon_y$). 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

$$\delta_t^x \leq c_x\|x_{t-1} - x_t^{\text{exact}}\|_2 + C_x\varepsilon_x, \quad \delta_t^y \leq c_y\|y_{t-1} - y_t^{\text{exact}}\|_2 + C_y\varepsilon_y, \tag{4.15}$$

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 projected gradient descent.

Lemma 4.3.1. *Suppose that, for all $t \geq 1$, the slack terms δ_t^x, δ_t^y satisfy* (4.15). *Then, under the assumptions of Theorem 4.3.2, if*

$$r := \left(\sqrt{1 - \frac{\alpha_x}{2\beta_x}} + 3c_y\sqrt{\frac{\beta_y}{\alpha_y}}\right) \cdot \left(\sqrt{1 - \frac{\alpha_y}{2\beta_y}} + 3c_x\sqrt{\frac{\beta_x}{\alpha_x}}\right) < 1, \tag{4.16}$$

then the iterations of the inexact alternating minimization algorithm (4.12) *satisfy*

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

for all $t \geq 1$, where

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

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.

4.3.4 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 (4.15). Here we give informal statment for this claim.

In Theorem 3.4.1 of Chapter 3, it is shown that under the framework of local concavity measure, projected gradient descent converges rapidly to the optimal as long as it is initialized near the target. Translating this result to the alternating minimization setting, we can bound the tolerance parameters δ_t^x, δ_t^y appearing in the inexact alternating minimization algorithm (4.12), when the inexact steps are computed via gradient descent. Here we state the statements informally to avoid overly complicated assumptions and constants:

Lemma 4.3.2 (Adapted from [Theorem 3.4.1])**.** *Under appropriate assumptions on the loss $\mathscr{L}$, initialization, and step size, the output of m_x many gradient descent steps on the x variable satisfies*

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

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

Examining the conditions (4.15) and (4.16) on the allowed size of the slack terms δ_t^x and δ_t^y, we can see that taking $c_x = \mathscr{O}\left(\left(\alpha_x/\beta_x\right)^{1.5}\right)$ and $c_y = \mathscr{O}\left(\left(\alpha_y/\beta_y\right)^{1.5}\right)$ is sufficient to ensure that the condition (4.16) will hold. This yields the following corollary:

Corollary 4.3.1. *Under the assumptions of Lemmas 4.3.1 and 4.3.2, for some radius* $\text{Rad} < 1$,

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

for all $t \geq 1$ as long as either the x update and the y update are exact, or are approximated via $m_x = \mathscr{O}\left(\frac{\beta_x}{\alpha_x} \log\left(\frac{\beta_x}{\alpha_x}\right)\right)$ and $m_y = \mathscr{O}\left(\frac{\beta_y}{\alpha_y} \log\left(\frac{\beta_y}{\alpha_y}\right)\right)$ many steps of gradient descent.

4.4 Examples

In this section, we highlight applications of our general theory to two classes of low-rank estimation problems, namely matrix decomposition and multitask regression.

4.4.1 *Matrix decomposition*

We revisit the low rank + sparse decomposition studied in Chapter 2; here we change our notation slightly and assume $X^\star$ is a low-rank matrix and $Y^\star$ is a sparse matrix. We further assume that the underlying matrices $X^\star$ and $Y^\star$ belong to the following constraint sets:[2]

$$\mathscr{X} = \left\{X \in \mathbb{S}_+^{d\times d} : \operatorname{rank}(X) \le r,\ \|X\|_\infty \le \frac{\alpha_{\mathsf{sp}}}{d}\right\},\ \mathscr{Y} = \left\{Y \in \mathbb{S}^{d\times d} : \|Y\|_1 \le \|Y^\star\|_1\right\}, \quad (4.17)$$

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 $\mathscr{X}$ makes sure that the low-rank update by the algorithm is at most α_{sp}-spiky at every iteration. The following lemma computes the upper bound on the concavity coefficient $\gamma_X(\mathscr{X})$:

Lemma 4.4.1. *For the constraint set* $\mathscr{X} = \{X \in \mathbb{S}_+^{d\times d} : \operatorname{rank}(X) \le r, \|X\|_\infty \le \frac{\alpha_{\mathsf{sp}}}{d}\}$*, we have* $\gamma_X(\mathscr{X}) \le \frac{5}{4\sigma_r(X)}$ *with respect to the nuclear norm* $\|\cdot\|_X = \|\cdot\|_{\mathrm{nuc}}$.

While we prove in Lemma 3.3.1 that the set of rank-constrained matrices without spikiness constraint has the local concavity coefficient $\gamma_X(\mathscr{X}) = \frac{1}{2\sigma_r(X)}$, the lemma above shows that the coefficient for $\mathscr{X}$ can be upper bounded with a larger constant factor.

Now we aim to recover the underlying matrices $(X^\star, Y^\star)$ by solving the constrained optimization problem

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

via (inexact) alternating minimization.

[2] For our analysis, we assume that $\|Y^\star\|_1$ is known *exactly*; this is a common assumption for the constrained problem, e.g. see [47]. 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.

Robust principal component analysis (RPCA) We study the robust PCA problem as formulated in [2], where the data matrix $Z \in \mathbb{S}^{n\times n}$ is generated from the model

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

where $\mathscr{A} : \mathbb{S}^{d\times d} \to \mathbb{S}^{n\times n}$ 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. We consider the least squares based estimator,

$$(\widehat{X}, \widehat{Y}) = \arg\min\left\{\frac{1}{2}\|Z - \mathscr{A}(X+Y)\|_{\mathsf{F}}^2 : (X,Y) \in \mathscr{X} \times \mathscr{Y}\right\}. \tag{4.18}$$

We follow the notion of RSC as introduced in [2, Definition 2] and require the *restricted eigenvalue* condition on $\mathscr{A}$:

Assumption 4.4.1. (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 $\operatorname{rank}(\Delta_X) \leq 2r$,*

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

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

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

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 $\|\mathscr{A}^*\mathscr{A}(\Delta_X)\|_\infty \approx \sqrt{\frac{d^2\log d}{n^2}}\|\Delta_X\|_\infty$ —for instance, an i.i.d. Gaussian ensemble will satisfy this property with high probability.

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

Lemma 4.4.2. *Suppose that the sample size is large enough to satisfy*

$$\frac{32\tau \cdot sd\log d}{n^2} \le \alpha_A. \tag{4.19}$$

Then, under the previously stated conditions, if $\|\mathscr{A}^*(W)\|_{\mathsf{sp}} \le c_1 \cdot \alpha_A \sigma_r(X^\star)$*, the steps* $(X_t, Y_t)_{t=1}^\infty$ *produced by the alternating minimization algorithm* (4.3) *satisfy*

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

Here, $\{c_i > 0, i = 1, 2\}$ *are universal constants, and* $C > 0$ *is defined in Theorem 4.3.1.*

We remark that 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. With extra effort, we can also prove that each step of the alternating minimization update can be replaced by several steps of the gradient descent updates, which we do not pursue here.

Gaussian factor model We next consider a Gaussian factor model, where our data consists of observations

$$z_i = U^\star w_i + \varepsilon_i,$$

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 $w_i \overset{\text{iid}}{\sim} \mathscr{N}(0, \mathbf{I}_r)$ and the independent noise $\varepsilon_i \overset{\text{iid}}{\sim} \mathscr{N}(0, Y^\star)$.

We further assume that the covariance structure of the noise, $Y^\star$, is sparse. We can calculate $\mathrm{Cov}(z_i) = U^\star U^{\star\top} + Y^\star$, 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

$$(\widehat{X}, \widehat{Y}) = \arg\min \left\{ \langle S_n, (X+Y)^{-1} \rangle - \log\det(X+Y)^{-1} : (X,Y) \in \mathscr{X} \times \mathscr{Y} \right\}, \tag{4.20}$$

for $S_n = \frac{1}{n} \sum_{i=1}^n z_i z_i^\top$, the sample covariance matrix of z_i's, where $\mathscr{X}$ and $\mathscr{Y}$ are defined as in (4.17).

Zwiernik et al. [48] studies the related loss function $\mathscr{L}(\Sigma) = \mathscr{L}(X+Y)$ in the context of a linear Gaussian covariance model. They prove that this loss is in fact convex in the region[3] $\{\Sigma \in \mathbb{R}^{d\times d} : 0 \prec \Sigma \prec 2S_n\}$ 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, showing that the loss (4.20) satisfies all the assumptions of Theorem 4.3.1 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 $\mathscr{X}_0 = \mathscr{X} \cap \mathbb{B}_2(X_0, \rho_X), \mathscr{Y}_0 = \mathscr{Y} \cap \mathbb{B}_2(Y_0, \rho_Y)$, 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 $\mathscr{X}_0 \times \mathscr{Y}_0$ contain the pair of true matrices $(X^\star, Y^\star)$ and the global optimal $(\widehat{X}, \widehat{Y})$, i.e. $(X^\star, Y^\star), (\widehat{X}, \widehat{Y}) \in \mathscr{X}_0 \times \mathscr{Y}_0$. With this setup, we now establish the following probabilistic guarantee:

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

Lemma 4.4.3. *Suppose that*

$$\sqrt{\frac{d}{n}} \leq c_1 \cdot \min\{\sigma_r(X^\star)\lambda_{\min}^{-1}(\Sigma^\star)\kappa^{-4}(\Sigma^\star), \kappa^{-1}(\Sigma^\star)\}. \tag{4.21}$$

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* (4.3) *satisfy*

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

Here, $\{c_i > 0, i = 1, \ldots, 4\}$ *are universal constants, and* $C > 0$ *is defined in Theorem 4.3.1.*

The discussion following Lemma 4.4.2 is also valid in this setting—in particular, the error due to the nonidentifiability of the model now appears as the term $\alpha_{\mathsf{sp}}^2 \frac{s}{d}$.

4.4.2 Multitask regression

Next we consider the multitask regression model where the response takes multiple output values. Suppose we are given m different tasks, namely each response is of the form $z_i \in \mathbb{R}^m$. Denoting the feature vector by $\phi_i \in \mathbb{R}^d$, the response is generated through the linear model

$$z_i = X^\star \phi_i + \varepsilon_i, \tag{4.22}$$

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 $\varepsilon_i \in \mathbb{R}^m$ is the measurement error from a centered multivariate normal distribution, with an unknown covariance matrix $\mathrm{Cov}(\varepsilon_i) = \Theta^{\star -1}$.

In the reduced regression setting [49], the true matrix $X^\star$ is assumed to be low rank. We then optimize the constrained negative log-likelihood function,

$$(\widehat{X},\widehat{\Theta}) = \arg\min\left\{-\log\det(\Theta)+\frac{1}{n}\sum_{i=1}^{n}(z_i - X\phi_i)^\top\Theta(z_i - X\phi_i) : X\in\mathscr{X},\Theta\succeq 0\right\}, \qquad (4.23)$$

where $\mathscr{X} = \{X \in \mathbb{R}^{m\times d} : \mathrm{rank}(X) \leq \mathrm{rank}(X^\star) = r\}$ represents the rank constraint on the coefficients X.[4] While this problem is generally nonconvex in (X,Θ), in addition to the nonconvex constraint $X \in \mathscr{X}$, we show that this problem satisfies all the assumptions that we require for our results.

For the purpose of our analysis, we consider a random design model, i.e. the feature vectors are sampled from a Gaussian distribution $\phi_i \overset{\text{iid}}{\sim} \mathscr{N}(0,\Sigma_\phi)$. Let (X_0,Θ_0) be the initialization point, and denote the local neighborhoods of the constraint sets around (X_0,Θ_0) by $\mathscr{X}_0 = \mathscr{X}\cap\mathbb{B}_2(X_0,\rho_X)$ and $\mathscr{Q}_0 = \mathbb{S}_+^{m\times m}\cap\mathbb{B}_2(\Theta_0,\rho_\Theta)$. We choose the radii ρ_X,ρ_Θ to satisfy $\rho_X \leq c_0\cdot\sigma_r(X^\star)\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)$ and $\rho_\Theta \leq c_0\cdot\lambda_{\min}(\Theta^\star)\kappa^{-1}(\Sigma_\phi)$ for some $c_0 > 0$, 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. $(X^\star,\Theta^\star),(\widehat{X},\widehat{\Theta})\in\mathscr{X}_0\times\mathscr{Q}_0$.

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

Lemma 4.4.4. *Suppose that*

$$\sqrt{\frac{1}{\lambda_{\min}(\Theta^\star)\lambda_{\max}(\Sigma_\phi)}}\sqrt{\frac{m+d}{n}} \leq c_1\cdot\sigma_r(X^\star)\kappa^{-1}(\Theta^\star)\kappa^{-1}(\Sigma_\phi). \qquad (4.24)$$

Then, under the previously stated conditions, with probability at least $1-c_2\exp(-c_3(m+d))$*, the*

[4] 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.

steps $(X_t, \Theta_t)_{t=1}^{\infty}$ *produced by the alternating minimization algorithm* (4.3) *satisfy*

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

for all $t \geq 1$*, 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, $\{c_i > 0, i = 1,\ldots,6\}$ *are universal constants, and* $C > 0$ *is defined in Theorem 4.3.1.*

We remark that the result in Lemma 4.4.4 assumes updating Θ first. While Lemma 4.4.4 is stated in terms of the *exact* alternating minimization, by working with Lemmas 4.3.1 and 4.3.2, 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.) We also refer the reader to [9] for similar results under the context of the pooled model.

4.5 Empirical results

We perform a numerical experiment on the multitask regression problem (Section 4.4.2) to examine the empirical performance of the alternating algorithm, as compared to performing gradient descent when treating (x,y) as a single variable. 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 $\varepsilon_i \overset{\text{iid}}{\sim}$

$\mathcal{N}(0,\Theta^{\star-1})$, 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 $\Theta^{\star,-1}_{ij} = \sigma^2 \cdot \rho^{|i-j|}$, where ρ is a local correlation parameter that we vary, while $\sigma^2 = \frac{\mathrm{Mean}(\|X^\star\phi_i\|^2_{\mathsf{F}}/m)}{3}$ is chosen to obtain a moderately difficult signal-to-noise ratio. The responses, z_i, are then drawn according to the model (4.22).

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 (4.23) with respect to the variables X and Θ, i.e. $\kappa_X(\mathcal{L})$ versus $\kappa_\Theta(\mathcal{L})$. As discussed in Section 4.3.2, 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 (4.23) 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:

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

  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

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

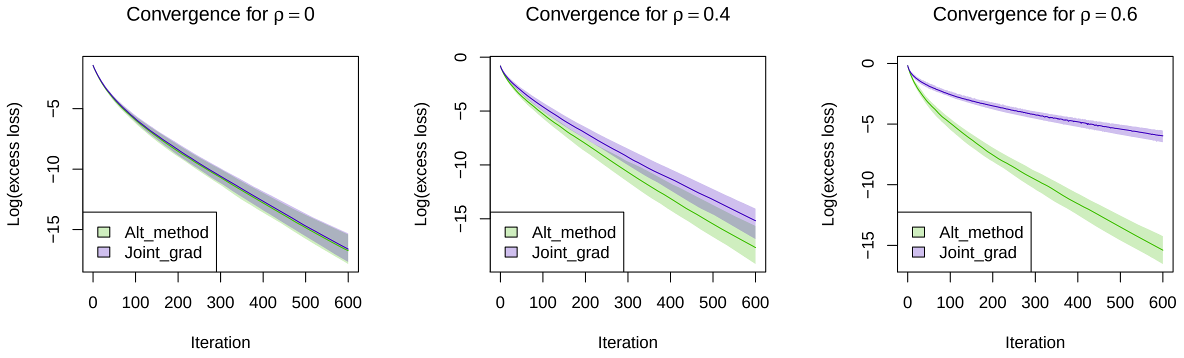

Figure 4.1: Comparison of the alternating algorithm and the joint gradient descent method applied to the simulated multitask regression problem. Results are shown for three settings of the noise correlation parameter ρ, across iterations $t = 0, 1, \ldots, 600$. In each plot, the solid line indicates the median loss over 100 trials, and the light band shows the interquartile range.

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, $\min_{t=1,\ldots,T} \mathscr{L}(X_t, \Theta_t)$, for the first $T = 1200$ iterations.

Figure 4.1 shows the excess loss at each iteration (on a log scale), where the excess loss is given by

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

where $\mathscr{L}_{\min}$ 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,[5] 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

[5] 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.

scales with the better of the two condition numbers, i.e. $\min\{\kappa_X(\mathscr{L}), \kappa_\Theta(\mathscr{L})\}$, while the joint gradient descent method is known to scale with the maximum. Since $\kappa_X(\mathscr{L}) \sim \kappa(\Sigma_\phi)\kappa(\Theta^\star)$ while $\kappa_\Theta(\mathscr{L}) \sim \kappa^2(\Theta^\star)$, 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.

4.6 Proofs of Theorems

In this section, we prove our main result on linear convergence for the exact alternating minimization algorithm, Theorem 4.3.1, and for the inexact algorithm, Theorem 4.3.2.

Proof of Theorem 4.3.1. First we prove the bound on the x update step, given in (4.9), for iteration number t. By definition of x_t, we can apply the first-order optimality condition (4.2), with $\mathscr{X}_0$ in place of $\mathscr{X}$, to obtain

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

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

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

Adding these two inequalities together, applying the initialization condition (Assumption 4.2.4) and rearranging terms several times, we have

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

where the last step holds by applying joint restricted strong convexity (Assumption 4.2.1) to the first term, and the cross-product condition (Assumption 4.2.3) to the expression in square brackets. (Note that these assumptions can be applied since we have $x_t \in \mathscr{X}_0$ and $y_{t-1} \in \mathscr{Y}_0$). Combining terms and simplifying, multiplying by 2, and using the assumption that $\mu_x \geq 0$ while $\mu_y \leq \alpha_y$, we obtain

$$0 \geq D^2(x_t;\widehat{x}) - D^2(y_{t-1};\widehat{y}) + \frac{\alpha_y}{2}\|y_{t-1}-\widehat{y}\|_2^2 - 2\alpha_x\varepsilon_x^2 - 2\alpha_y\varepsilon_y^2. \tag{4.25}$$

Now, by restricted smoothness (Assumption 4.2.2) and using the assumption $\alpha_y \leq \beta_y$,

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

Returning to (4.25) and rearranging terms,

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

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

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

thus proving that the bound (4.9) 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 (4.10) 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 (4.11) showing linear convergence in the ℓ_2 norm; details are given in Section 4.7.1. □

Proof of Theorem 4.3.2. This proof is a straightforward combination of the relaxed triangle inequality (Assumption 4.3.1) with Theorem 4.3.1, the contraction result for the exact alternating minimization algorithm. First, since x_t^{exact} exactly solves the alternating minimization step, i.e. $\arg\min_{x\in\mathscr{X}_0}\mathscr{L}(x,y_{t-1})$, Theorem 4.3.1 proves that

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

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 4.3.1),

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

This proves the bound (4.13). The bound (4.14) on $D(y_t;\widehat{y})$ is proved analogously.

□

4.7 Proofs of lemmas

4.7.1 Details on ℓ_2 convergence bound (4.11)

Here we give details for the ℓ_2 convergence bound (4.11) in Theorem 4.3.1. 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 (4.9) and (4.10) can be rewritten as

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

and

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

Then, applying these bounds recursively, we have

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

and

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

Let $\alpha_{\min} = \min\{\alpha_x, \alpha_y\}, \alpha_{\max} = \max\{\alpha_x, \alpha_y\}$. By joint restricted strong convexity (4.7),

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

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

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

and by restricted smoothness (4.8),

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

Combining everything, and simplifying, we obtain the overall convergence guarantee (4.11).

Proof of Lemma 4.3.1. 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 (4.11) for the exact algorithm, given in Section 4.7.1, 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 $t \geq 1$,

$$\begin{cases} D(x_t; \widehat{x}) \leq r_x \cdot (r_x r_y)^{t-1} \cdot D_0 + \frac{1+r_x}{1-r_x r_y} \cdot C' \max\{\varepsilon_x, \varepsilon_y\}, \\ D(y_t; \widehat{y}) \leq (r_x r_y)^t \cdot D_0 + \frac{1+r_y}{1-r_x r_y} \cdot C' \max\{\varepsilon_x, \varepsilon_y\}, \end{cases} \tag{4.26}$$

where

$$C' = 4\left(1 + c_x \sqrt{\frac{\beta_x}{\alpha_x}} + c_y \sqrt{\frac{\beta_y}{\alpha_y}}\right) \sqrt{\alpha_x + \alpha_y} + C_x \sqrt{\beta_x} + C_y \sqrt{\beta_y}. \tag{4.27}$$

First we prove the bounds (4.26) at time $t = 1$. For the x bound,

$$\begin{aligned}
D(x_1;\widehat{x}) &\le \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot D(y_0;\widehat{y}) + \sqrt{\beta_x}\delta_t^x + \sqrt{8(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)} \text{ by Theorem 4.3.2}\\
&\le \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot D(y_0;\widehat{y}) + \sqrt{\beta_x}\left(c_x\|x_0 - x_1^{\text{exact}}\|_2 + C_x\varepsilon_x\right) + \sqrt{8(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)} \text{ by (4.15)}\\
&\le \sqrt{1-\frac{\alpha_y}{2\beta_y}}\cdot (\sqrt{\beta_y}\rho_y + \sqrt{\alpha_y}\varepsilon_y) + \sqrt{\beta_x}(c_x\cdot\rho_x + C_x\varepsilon_x) + \sqrt{8(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)},
\end{aligned}$$

where the last step holds since $x_1^{\text{exact}} \in \mathscr{X}_0 \subset \mathbb{B}_2(x_0,\rho_x)$, and $D(y_0;\widehat{y})$ can be bounded by restricted smoothness (Assumption 4.2.2). Simplifying,

$$D(x_1;\widehat{x}) \le r_x D_0 + C' \max\{\varepsilon_x,\varepsilon_y\},$$

which proves the bound (4.26) on $D(x_1;\widehat{x})$ at time $t = 1$. Similarly, for the y bound,

$$\begin{aligned}
D(y_1;\widehat{y}) &\le \sqrt{1-\frac{\alpha_x}{2\beta_x}}\cdot D(x_1;\widehat{x}) + \sqrt{\beta_y}\delta_t^y + \sqrt{8(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)} \text{ by Theorem 4.3.2}\\
&\le \sqrt{1-\frac{\alpha_x}{2\beta_x}}\cdot D(x_1;\widehat{x}) + \sqrt{\beta_y}\left(c_y\|y_0 - y_1^{\text{exact}}\|_2 + C_y\varepsilon_y\right) + \sqrt{8(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)} \text{ by (4.15)}\\
&\le \sqrt{1-\frac{\alpha_x}{2\beta_x}}\cdot \left(r_x D_0 + C'\max\{\varepsilon_x,\varepsilon_y\}\right) + \sqrt{\beta_y}\left(c_y\rho_y + C_y\varepsilon_y\right) + \sqrt{8(\alpha_x\varepsilon_x^2+\alpha_y\varepsilon_y^2)}\\
&\le r_x r_y D_0 + (1+r_y)\cdot C'\max\{\varepsilon_x,\varepsilon_y\},
\end{aligned}$$

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

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

$$\begin{aligned}
&\|x_{t-1} - x_t^{\text{exact}}\|_2 \leq \|x_{t-1} - \widehat{x}\|_2 + \|x_t^{\text{exact}} - \widehat{x}\|_2 \\
&\leq \frac{1}{\sqrt{\alpha_x}} \left(D(x_{t-1}; \widehat{x}) + D(x_t^{\text{exact}}; \widehat{x}) \right) + \frac{2\sqrt{\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2}}{\sqrt{\alpha_x}} \text{ by joint restricted strong convexity (4.7)} \\
&\leq \frac{1}{\sqrt{\alpha_x}} \left(D(x_{t-1}; \widehat{x}) + D(y_{t-1}; \widehat{y}) + \sqrt{3(\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2)} \right) + \frac{2\sqrt{\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2}}{\sqrt{\alpha_x}} \text{ by Theorem 4.3.1} \\
&\leq \frac{1}{\sqrt{\alpha_x}} \left(D(x_{t-1}; \widehat{x}) + D(y_{t-1}; \widehat{y}) + 4\sqrt{\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2} \right).
\end{aligned}$$

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

$$\begin{aligned}
&D(x_t; \widehat{x}) \leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{\beta_x} \delta_t^x + \sqrt{8(\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2)} \text{ by Theorem 4.3.2} \\
&\leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{\beta_x} \left(c_x \|x_{t-1} - x_t^{\text{exact}}\|_2 + C_x \varepsilon_x \right) + \sqrt{8(\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2)} \text{ by (4.15)} \\
&\leq \sqrt{1 - \frac{\alpha_y}{2\beta_y}} \cdot D(y_{t-1}; \widehat{y}) + \sqrt{\beta_x} \left(c_x \left[\frac{1}{\sqrt{\alpha_x}} \left(D(x_{t-1}; \widehat{x}) + D(y_{t-1}; \widehat{y}) + 4\sqrt{\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2} \right) \right] + C_x \varepsilon_x \right) \\
&\qquad\qquad + \sqrt{8(\alpha_x \varepsilon_x^2 + \alpha_y \varepsilon_y^2)} \\
&\leq \left(\sqrt{1 - \frac{\alpha_y}{2\beta_y}} + c_x \sqrt{\frac{\beta_x}{\alpha_x}} \right) D(y_{t-1}; \widehat{y}) + c_x \sqrt{\frac{\beta_x}{\alpha_x}} D(x_{t-1}; \widehat{x}) + C' \max\{\varepsilon_x, \varepsilon_y\},
\end{aligned}$$

where C' is defined as in (4.27) above. Assuming by induction that the bounds (4.26) hold with $t-1$ in place of t, we obtain

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

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

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

Plugging in the definition of r_x, then,

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

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 (4.26), as desired. The bound on $D(y_t;\widehat{y})$ is proved similarly.

Finally, by joint restricted strong convexity (4.7), we know that

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

Combining this with the bounds (4.26) proves the result.

□

Proof of Lemma 4.2.1. Take any $x,x' \in \mathscr{X}_0 \subset \mathscr{X}$ with $x \neq x'$ and take $t \in [0,1]$. By the curvature condition (3.1) on the larger set $\mathscr{X}$, we can find a family of points $\widetilde{x}_t \in \mathscr{X}$, indexed by $t \in [0,1]$, and some sequence $\delta_t \to 0$, where

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

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

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

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

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

Combining everything,

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

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

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

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

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

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

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

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

$$\gamma_x(\mathscr{X}_0) \leq \gamma_x(\mathscr{X}).$$

□

4.8 Proofs for examples

Throughout the section, if f is a function 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 $\mathrm{vec}\,(A) \in \mathbb{R}^{mn}$.

Proof of Lemma 4.4.1. We first reparametrize the variable $X \in \mathscr{X}$ by $X = \mathsf{g}(U) = UU^\top$ with the corresponding convex set

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

where U_{i*} represents ith row of U. Note that under such reparametrization, we trivially have $\mathscr{X} = \mathsf{g}(\mathscr{U})$. Now take $X,X' \in \mathscr{X}$ 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, for some $s,s' \in [0,1]$, and $U_* = (1-s)U + sU_t$,

$U_\# = (1-s)U' + sU_t$, we have

$$\begin{aligned}
X_t - \mathrm{g}(U_t) &= (1-t)\mathrm{g}(U) + t\mathrm{g}(U') - \mathrm{g}(U_t) \\
&= (1-t)(\mathrm{g}(U) - \mathrm{g}(U_t)) + t(\mathrm{g}(U') - \mathrm{g}(U_t)) \\
&= (1-t)\left[\nabla \mathrm{g}(U_t)(U-U_t) + \frac{1}{2}\nabla^2 \mathrm{g}(U_*)(U-U_t, U-U_t)\right] \\
&\qquad\qquad + t\left[\nabla \mathrm{g}(U_t)(U'-U_t) + \frac{1}{2}\nabla^2 \mathrm{g}(U_\#)(U'-U_t, U'-U_t)\right] \\
&= (1-t)\left[t\nabla \mathrm{g}(U_t)(U-U') + \frac{t^2}{2}\nabla^2 \mathrm{g}(U_*)(U-U', U-U')\right] \\
&\qquad\qquad + t\left[(1-t)\nabla \mathrm{g}(U_t)(U'-U) + \frac{(1-t)^2}{2}\nabla^2 \mathrm{g}(U_\#)(U'-U, U'-U)\right] \\
&= \frac{t^2(1-t)}{2}\nabla^2 \mathrm{g}(U_*)(U-U', U-U') + \frac{t(1-t)^2}{2}\nabla^2 \mathrm{g}(U_\#)(U'-U, U'-U). \qquad (4.28)
\end{aligned}$$

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

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

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 (4.28),

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

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

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

where the last inequality follows from [50, Lemma 5.4]. This completes the proof of the lemma.

□

Proof of Lemma 4.4.2. Recalling the constrained least squares problem (4.18) for the robust PCA problem, we verify that under the conditions of Lemma 4.4.2, the loss function satisfies the assumptions of Theorem 4.3.1, i.e. Assumptions 4.2.1, 4.2.2, 4.2.3, and 4.2.4, with parameters specified below. Before proceeding, we observe that for all $Y \in \mathscr{Y}_0$,

$$\|Y - \widehat{Y}\|_1 \leq \|Y - Y^\star\|_1 + \|\widehat{Y} - Y^\star\|_1 \leq 2\sqrt{sd}\|Y - \widehat{Y}\|_\mathsf{F} + 4\sqrt{sd}\|\widehat{Y} - Y^\star\|_\mathsf{F}, \tag{4.29}$$

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

Note that the cross-product condition (Assumption 4.2.3) trivially holds with $\mu_x = \mu_y = 0$, since the Hessian $\nabla^2_{XY}\mathscr{L}(X,Y)$ 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 $X \in \mathscr{X}_0$ and $Y \in \mathscr{Y}_0$. Invoking the restricted eigenvalue property (Assumption 4.4.1), we have

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

Applying the inequality (4.29) and the spikiness constraint, and using the fact that $\frac{32\tau sd\log d}{n^2} \leq$

α_A, the last term can be bounded by

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

where the second step uses the identity $ab\le\frac{ca^2}{2}+\frac{b^2}{2c}$ for any $c>0$. Combining the pieces, then

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

Restricted smoothness A similar calculation shows that the marginal restricted smoothness condition holds, that is, by the restricted eigenvalue property (Assumption 4.4.1), we have

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

and similarly,

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

Initialization condition Since $\mathscr{Y}$ is convex, the initialization condition is trivial for the set $\mathscr{Y}_0$. For $\mathscr{X}$, we first bound $\|\nabla_X\mathscr{L}(X,Y)\|_{\mathsf{sp}}$ for all $X\in\mathscr{X}_0$ and $Y\in\mathscr{Y}_0$. Given the observational model

$Z = \mathscr{A}(X^\star + Y^\star) + W$, we have the decomposition of $\|\nabla_X \mathscr{L}(X,Y)\|_{\mathrm{sp}}$ into the sums

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

We can find some X' with $\mathrm{rank}(X') = 1$ and $\|X'\|_{\mathsf{F}} \leq 1$ so that

$$\|\mathscr{A}^*\mathscr{A}(X - X^\star)\|_{\mathrm{sp}} = \langle \mathscr{A}(X'), \mathscr{A}(X - X^\star) \rangle.$$

By the restricted eigenvalue condition (Assumption 4.4.1), this proves

$$\|\mathscr{A}(X')\|_{\mathsf{F}}^2 \leq \beta_A \;\text{ and }\; \|\mathscr{A}(X - X^\star)\|_{\mathsf{F}}^2 \leq \beta_A \|X - X^\star\|_{\mathsf{F}}^2.$$

So, we have

$$\text{(Term 1)} = \|\mathscr{A}(X')\|_{\mathsf{F}} \|\mathscr{A}(X - X^\star)\|_{\mathsf{F}} \leq \beta_A \|X - X^\star\|_{\mathsf{F}} \leq 2\beta_A \rho_X,$$

where the last inequality follows from $X, X^\star \in \mathbb{B}_2(X_0, \rho_X)$. Again, by Assumption 4.4.1, we can bound $\|\mathscr{A}(Y - Y^\star)\|_{\mathsf{F}}^2 \leq \beta_A \|Y - Y^\star\|_{\mathsf{F}}^2 + \frac{4\tau s d \log d}{n^2} \|Y - Y^\star\|_{\mathsf{F}}^2 \leq \frac{9\beta_A}{8} \|Y - Y^\star\|_{\mathsf{F}}^2$, and so for some X'' with $\mathrm{rank}(X'') = 1$ and $\|X''\|_{\mathsf{F}} \leq 1$,

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

Putting these bounds together, we have $\|\nabla_X \mathscr{L}(X,Y)\|_{\mathrm{sp}} \leq 3\beta_A(\rho_X + \rho_Y) + \|\mathscr{A}^*(W)\|_{\mathrm{sp}}$. Now, by Lemma 4.4.1, we know $\gamma_X(\mathscr{X}) \leq \frac{5}{4\sigma_r(X)}$, and so

$$\max_{X \in \mathscr{X}_0} \gamma_X(\mathscr{X}) \leq \frac{5}{4\sigma_r - 8\rho_X} \leq \frac{5}{2\sigma_r}, \tag{4.30}$$

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 $\rho_X, \rho_Y \leq c_0 \cdot \sigma_r \kappa^{-1}(\mathscr{A})$ for some sufficiently small $c_0 > 0$, this implies that the conditions of Lemma 4.2.1 hold, i.e. $\rho_X < \frac{1}{2\max_{X\in\mathscr{X}_0}\gamma_X(\mathscr{X})}$, and in particular, we have $\gamma(\mathscr{X}_0) \le \frac{5}{2\sigma_r}$. Now combining all the pieces, we have

$$\begin{aligned}2\gamma(\mathscr{X}_0)\cdot\left[\|\nabla_X\mathscr{L}(\widehat{X},\widehat{Y})\|_{\mathrm{sp}}+\max_{Y\in\mathscr{Y}_0}\|\nabla_Y\mathscr{L}(X_Y,Y)\|_{\mathrm{sp}}\right] &\le 4\gamma(\mathscr{X}_0)\cdot\max_{X\in\mathscr{X}_0,Y\in\mathscr{Y}_0}\|\nabla_X\mathscr{L}(X,Y)\|_{\mathrm{sp}}\\ &\le \frac{10}{\sigma_r}\cdot\left(3\beta_A\rho_X+3\beta_A\rho_Y+\|\mathscr{A}^*(W)\|_{\mathrm{sp}}\right)\le\alpha_A,\end{aligned}$$

where we use $\|\mathscr{A}^*(W)\|_{\mathrm{sp}} \le \sigma_r\cdot\frac{\alpha_A}{30}$ in the last step. This establishes the initialization condition.

Now by specializing to the robust PCA problem (4.18), Theorem 4.3.1 immediately yields the result of Lemma 4.4.2, as desired. □

Proof of Lemma 4.4.3. Next we turn to prove our claims for the Gaussian factor model, as presented in (4.20). First we calculate the gradient and Hessian of $\mathscr{L}(X,Y)$: for all $\Delta_X,\Delta_Y\in\mathbb{R}^{d\times d}$,

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

and

$$\begin{aligned}\begin{pmatrix}\Delta_X\\ \Delta_Y\end{pmatrix}^\top\nabla^2\mathscr{L}(X,Y)\begin{pmatrix}\Delta_X\\ \Delta_Y\end{pmatrix}&=\mathrm{vec}\left(\Delta_X\right)^\top\mathscr{H}(X,Y)\mathrm{vec}\left(\Delta_X\right)+\mathrm{vec}\left(\Delta_Y\right)^\top\mathscr{H}(X,Y)\mathrm{vec}\left(\Delta_Y\right)\\ &\quad+2\mathrm{vec}\left(\Delta_X\right)^\top\mathscr{H}(X,Y)\mathrm{vec}\left(\Delta_Y\right),\end{aligned}$$

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

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

Throughout this proof, we use the following concentration inequality: 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

$$\|S_n - \Sigma^\star\|_{\text{sp}} \le \|\Sigma^\star\|_{\text{sp}} \|\Sigma^{\star -1/2} S_n \Sigma^{\star -1/2} - \mathbf{I}_d\|_{\text{sp}} \le 3\lambda_{\max}(\Sigma^\star)\sqrt{\frac{d}{n}}, \tag{4.31}$$

where the second step holds by a concentration bound on the extreme singular values of a standard Gaussian ensemble [25].

We calculate a few inequalities to use later. Recall

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

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 $X^\star \in \mathscr{X}_0, Y^\star \in \mathscr{Y}_0$, then we have

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

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

$$\frac{3}{4}\lambda_{\min}(\Sigma^\star) \le \lambda_{\min}(X + Y) \le \lambda_{\max}(X + Y) \le \frac{5}{4}\lambda_{\max}(\Sigma^\star). \tag{4.32}$$

Applying Weyl's inequality again and using the inequality (4.31), we also have

$$\frac{1}{2}\lambda_{\min}(\Sigma^\star) \le \lambda_{\min}(2S_n - X - Y) \le \lambda_{\max}(2S_n - X - Y) \le \frac{3}{2}\lambda_{\max}(\Sigma^\star), \tag{4.33}$$

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

$$\frac{32}{125}\frac{\kappa^{-1}(\Sigma^\star)}{\lambda_{\max}^2(\Sigma^\star)} \le \lambda_{\min}(\mathscr{H}(X, Y)) \le \lambda_{\max}(\mathscr{H}(X, Y)) \le \frac{32}{9}\frac{\kappa(\Sigma^\star)}{\lambda_{\min}^2(\Sigma^\star)}. \tag{4.34}$$

Finally, due to the spikiness constraint and the ℓ_1 norm inequality (4.29), we have the following finite bound on the inner product between the low-rank and sparse components: for all $X \in \mathscr{X}_0$, $Y \in \mathscr{Y}_0$,

$$\langle X - \widehat{X}, Y - \widehat{Y} \rangle \leq \|X - \widehat{X}\|_\infty \|Y - \widehat{Y}\|_1 \leq 4\alpha_{\mathsf{sp}} \sqrt{\frac{s}{d}} \|\Delta_Y\|_{\mathsf{F}} + 8\alpha_{\mathsf{sp}} \sqrt{\frac{s}{d}} \|\widehat{Y} - Y^\star\|_{\mathsf{F}}. \tag{4.35}$$

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

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

$$\begin{aligned} &\begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix}^\top \nabla^2 \mathscr{L}(X(t), Y(t)) \begin{pmatrix} X - \widehat{X} \\ Y - \widehat{Y} \end{pmatrix} \\ &\qquad = \left(\operatorname{vec}\left(X - \widehat{X}\right) + \operatorname{vec}\left(Y - \widehat{Y}\right) \right)^\top \mathscr{H}(X(t), Y(t)) \left(\operatorname{vec}\left(X - \widehat{X}\right) + \operatorname{vec}\left(Y - \widehat{Y}\right) \right), \end{aligned}$$

where $X(t) = (1-t)X + t\widehat{X}$ and $Y(t) = (1-t)Y + t\widehat{Y}$ for some $t \in [0,1]$. Using (4.34), and applying the inequality (4.35) and $ab \leq \frac{ca^2}{2} + \frac{b^2}{2c}$, we can lower bound as

$$\begin{aligned} &\frac{32}{125} \frac{\kappa^{-1}(\Sigma^\star)}{\lambda_{\max}^2(\Sigma^\star)} \left\| \operatorname{vec}\left(X - \widehat{X}\right) + \operatorname{vec}\left(Y - \widehat{Y}\right) \right\|_2^2 \\ &\qquad\qquad \geq \frac{32}{125} \frac{\kappa^{-1}(\Sigma^\star)}{\lambda_{\max}^2(\Sigma^\star)} \left[\|X - \widehat{X}\|_{\mathsf{F}}^2 + \frac{1}{2} \|Y - \widehat{Y}\|_{\mathsf{F}}^2 - 16 \|\widehat{Y} - Y^\star\|_{\mathsf{F}}^2 - 36\alpha_{\mathsf{sp}}^2 \frac{s}{d} \right]. \end{aligned}$$

Restricted smoothness By Taylor's theorem and using the inequality (4.34), we have

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

and similarly with the roles of X and Y reversed.

Cross-product bound As discussed in Section 4.2.2 following the Assumption 4.2.3, in order to establish the cross-product condition, it suffices to bound

$$\|\nabla^2_{XY}\mathscr{L}(X,Y(t)) - \nabla^2_{XY}\mathscr{L}(X(t'),Y)\|_{\mathrm{sp}} \tag{4.36}$$

for all $X \in \mathscr{X}_0$, $Y \in \mathscr{Y}_0$, where $X(t') = (1-t')X + t'\widehat{X}$ and $Y(t) = (1-t)Y + t\widehat{Y}$. Also $\nabla^2_{XY}\mathscr{L}(X,Y)$ is symmetric in X and Y, so we will only bound $\|\nabla^2_{XY}\mathscr{L}(X,Y(t)) - \nabla^2_{XY}\mathscr{L}(X,Y)\|_{\mathrm{sp}}$; by the triangle inequality, similar bound would hold for the other term. We can also see that the operator norms of $\mathscr{H}_1$ and $\mathscr{H}_2$ are same, and so

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

Therefore we only work with the term $\|\mathscr{H}_1(X,Y(t)) - \mathscr{H}_1(X,Y)\|_{\mathrm{sp}}$. Let

$$\Delta\mathscr{H}_1 = (X+Y(t))^{-1}(2S_n - (X+Y(t)))(X+Y(t))^{-1} - (X+Y)^{-1}(2S_n - (X+Y))(X+Y)^{-1}.$$

Then simple algebra yields

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

$\Delta\mathscr{H}_1$ is further decomposed as

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

Meanwhile, by the inequalities (4.32) and (4.33), we have

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

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

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

This implies

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

and hence that

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

Returning to the above equation, this implies

$$\|\nabla_{XY}^2\mathscr{L}(X,Y(t)) - \nabla_{XY}^2\mathscr{L}(X,Y)\|_{\mathrm{sp}} \leq \frac{448}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda_{\min}^4(\Sigma^\star)}\|Y-\widehat{Y}\|_{\mathrm{sp}},$$

and in particular, we have

$$\|\nabla_{XY}^2\mathscr{L}(X,Y(t)) - \nabla_{XY}^2\mathscr{L}(X(t'),Y)\|_{\mathrm{sp}} \leq \frac{448}{27}\frac{\lambda_{\max}(\Sigma^\star)}{\lambda_{\min}^4(\Sigma^\star)}\left(\|X-\widehat{X}\|_{\mathrm{sp}} + \|Y-\widehat{Y}\|_{\mathrm{sp}}\right).$$

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

Initialization condition To prove the initialization condition, it suffices to bound the quantity $4\gamma(\mathscr{X}_0) \cdot \max_{X\in\mathscr{X}_0, Y\in\mathscr{Y}_0} \|\nabla_X \mathscr{L}(X,Y)\|_{\mathrm{sp}}$. Note that for any $X \in \mathscr{X}_0, Y \in \mathscr{Y}_0$,

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

where the first step uses the inequality (4.32), and the second step uses the inequality $\|X+Y-\Sigma^\star\|_{\mathrm{sp}} \leq 2(\rho_X + \rho_Y)$ and the concentration bound (4.31) as well as our assumptions on the radii and the sample size (4.21) (by choosing $c_0, c_1 > 0$ sufficiently small). Moreover, by the same reasoning to the equation (4.30), we also have $\gamma(\mathscr{X}_0) \leq \frac{5}{2\sigma_r}$. Therefore,

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

completing the proof of Lemma 4.4.3.

□

Proof of Lemma 4.4.4. Recall the loss function, given by the negative log-likelihood function

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

We calculate the gradient

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

and the Hessian operators

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

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 $\Theta \in \mathscr{Q}_0$, its minimum and maximum eigenvalues are bounded by

$$\frac{\lambda_{\min}(\Theta^\star)}{2} \leq \lambda_{\min}(\Theta) \leq \lambda_{\max}(\Theta) \leq \frac{3\lambda_{\max}(\Theta^\star)}{2}, \tag{4.37}$$

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

We will use the following two concentration results: first, following [51, Lemma 2], with probability at least $1 - 4\exp(-n/2)$, we have the bound of the form:

$$\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i\phi_i^\top\right) \geq \frac{\lambda_{\min}(\Sigma_\phi)}{9} \text{ and } \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\phi_i\phi_i^\top\right) \leq 9\lambda_{\max}(\Sigma_\phi). \tag{4.38}$$

Next, letting $\widetilde{\varepsilon}_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \mathbf{I}_m)$, it has been shown in [51, Lemma 3] that for some $c, c' > 0$, with probability at least $1 - c\exp(-c'(m+d))$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i\phi_i^\top\right\|_{\mathrm{sp}} \leq 5\sqrt{\lambda_{\max}(\Sigma_\phi)}\sqrt{\frac{m+d}{n}}. \tag{4.39}$$

Now, we turn to verifying Lemma 4.4.4:

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

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

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 $z_i = X^\star \phi_i + \varepsilon$, we have the following decomposition:

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

For (Term 1), we lower bound as

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

where the second step uses the inequalities (4.37) and (4.38). For (Term 2), we further decompose 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

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

where the inequality uses the bound on the radius ρ_Θ (by choosing $c_0 \le \frac{1}{36\cdot 54}$) and (4.38). Meanwhile, we can bound the second part of (Term 2) as

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

Combining the two yields

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

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

$$\begin{aligned}
\text{(Term 3)} \le 2\|X-\widehat{X}\|_{\mathrm{nuc}}\left\|\frac{1}{n}\sum_{i=1}^{n}(\Theta-\widehat{\Theta})\varepsilon_i\phi_i^\top\right\|_{\mathrm{sp}} &\le \frac{\rho_\Theta}{\sqrt{\lambda_{\min}(\Theta^\star)}}\cdot 2\sqrt{2r}\|X-\widehat{X}\|_{\mathsf{F}}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\varepsilon}_i\phi_i^\top\right\|_{\mathrm{sp}} \\
&\le \frac{\rho_\Theta\sqrt{\lambda_{\max}(\Sigma_\phi)}}{\sqrt{\lambda_{\min}(\Theta^\star)}}\cdot 10\sqrt{2r}\|X-\widehat{X}\|_{\mathsf{F}}\sqrt{\frac{m+d}{n}},
\end{aligned}$$

where the second step follows since $X-\widehat{X}$ is of rank $2r$ and $\varepsilon_i = (\Theta^\star)^{-1/2}\cdot\widetilde{\varepsilon}_i$ for $\widetilde{\varepsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\mathbf{I}_m)$, and the next step uses the concentration bound (4.39). Using the identity $ab \le \frac{ca^2}{2}+\frac{b^2}{2c}$ and the bound on ρ_Θ, then

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

Lastly, by (4.37), the minimum eigenvalue of $\Theta(t)^{-1}$ is lower bounded by $\frac{2}{3\lambda_{\max}(\Theta^\star)}$, so it follows that

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

Combining all the bounds together, we have

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

Restricted smoothness For the X variable, we apply the inequalities (4.37) and (4.38) to obtain

$$\begin{aligned}\langle X - \widehat{X}, \nabla_X \mathscr{L}(X, \widehat{\Theta}) - \nabla_X \mathscr{L}(\widehat{X}, \widehat{\Theta})\rangle &= \frac{2}{n}\sum_{i=1}^n \phi_i^\top (X - \widehat{X})^\top \widehat{\Theta}(X - \widehat{X})\phi_i \\ &\leq 27\lambda_{\max}(\Theta^\star)\lambda_{\max}(\Sigma_\phi)\|X - \widehat{X}\|_{\mathsf{F}}^2.\end{aligned}$$

For the Θ variable, by Taylor's theorem combined with the bound (4.37), for some $t \in [0,1]$,

$$\begin{aligned}\langle \Theta - \widehat{\Theta}, \nabla_\Theta \mathscr{L}(\widehat{X}, \Theta) - \nabla_\Theta \mathscr{L}(\widehat{X}, \widehat{\Theta})\rangle &= \text{vec}\left(\Theta - \widehat{\Theta}\right)^\top \nabla_{\Theta\Theta}^2 \mathscr{L}(\widehat{X}, (1-t)\Theta + t\widehat{\Theta})\text{vec}\left(\Theta - \widehat{\Theta}\right) \\ &\leq \frac{4}{\lambda_{\min}^2(\Theta^\star)}\|\Theta - \widehat{\Theta}\|_{\mathsf{F}}^2.\end{aligned}$$

Cross-product bound Take $X \in \mathscr{X}_0$, $\Theta \in \mathscr{Q}_0$. Then, by Taylor's theorem, for some $t, t' \in [0,1]$,

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

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 We already know that $\gamma_X(\mathscr{X}) = \frac{1}{2\sigma_r(X)}$ (see Lemma 3.3.1), so

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

where the first inequality holds due to Weyl's inequality, and the next inequality holds since $\rho_X \leq \frac{1}{4}\sigma_r$. This also shows that the conditions of Lemma 4.2.1 is satisfied, so we have $\gamma(\mathscr{X}_0) \leq \frac{1}{\sigma_r}$.

Next, we bound the gradient term $\|\nabla_X \mathscr{L}(X,\Theta)\|_{\text{sp}}$. Given the observational model $z_i = X^\star\phi_i + \varepsilon_i$, we can decompose the gradient as

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

Using the inequalities (4.37) and (4.38), 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

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

where the steps use the inequalities (4.37) and (4.39). Combining the two and using the bound on ρ_X and the assumption (4.24), for sufficiently small $c_0, c_1 > 0$, we have

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

and therefore

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

completing the proof of Lemma 4.4.4.

□

CHAPTER 5

SPECTRAL CALIBRATION AND IMAGE RECONSTRUCTION IN CT IMAGING

X-ray computed tomography (CT) is a medical imaging modality that allows to image the internal structure of human body in a non-invasive way. Generally speaking, x-ray tube in CT scanner generates a high flux of x-rays which traverse an object from multiple directions, and detectors observe the incoming x-rays as the outcome of interaction between the x-rays and the object. The forward model describes the interaction between the x-rays and the object, in which it is typically assumed that the x-rays travel through the object along straight lines (i.e. ignoring x-ray scatter) and the intensity of the x-rays are attenuated, while traveling, with a rate of decay depending on the properties of the materials. Based on these measurements of the object, reconstruction algorithms aim to recover the structure of the objects that interacted with the x-rays.

In addition to acquiring the x-ray transmission measurements, reconstruction algorithms typically need knowledge of the x-ray source spectrum and the detector response as the x-ray beams generated for medical CT is polychromatic in nature. While we may assume that the x-ray source spectrum can be modeled to a certain degree, the detector spectral response is often unknown due to many non-ideal physical effects of the detector. For instance, photon-counting detectors can discriminate incident x-ray photons based on their energies, allowing for the CT data acquisition in each energy window, and thus are useful for material decompositions to more than two basis functions; however, they also exhibit undesirable technical issues such as pulse pile-up and charge sharing [52], potentially resulting in serious artifacts in the reconstructed images—here we refer to the artifacts as the discrepancy between the reconstructed values in the images and the true image values of the object. Therefore, in reconstructing CT images, it is crucial to accurately calibrate the spectral response of the detectors for further reduction of image artifacts.

In this chapter, we present a new x-ray spectrum reconstruction method based on transmission measurements of a calibration phantom—a material with known thicknesses and compositions.[1] Our aim is to formulate spectrum estimation as an optimization problem, for which an efficient first-order iterative algorithm is employed to solve the resulting optimization problem rapidly. The proposed method is capable of incorporating prior information about the physical shape of an x-ray spectrum, which enables accurate and realistic estimation of x-ray spectrum by including the characteristic lines of the target spectrum in the final estimation.

Although the method studied here can be used for any spectrum estimation task, the focus here is on photon-counting CT. Interest in estimating the x-ray spectrum of a CT system is recently growing due to the development of spectral CT with a photon-counting detector [52, 54, 55]. For spectral CT, the effective spectrum estimate, which includes the source spectrum and detector response, is needed for material decomposition into basis material sinograms [54] and for direct inversion into basis material images [55, 12]. This suggests that our optimization-based approach can be useful when the spectral calibration of an imaging system is combined with other optimization-based algorithms for spectral CT image reconstruction. As a preliminary study, here we perform alternating minimization based algorithm on a two-material phantom derived from the FORBILD head phantom[2] to demonstrate the utility of the method on the task of simultaneous spectrum estimation and spectral CT image reconstruction.

5.1 Background on spectrum estimation

This section presents the discretized forward model that relates the expected photon counts to the x-ray spectral distribution, and provides a brief overview of expectation-maximization (EM) and other related methods for the spectrum estimation problem.

[1] The work presented in this chapter is published in [53].

[2] http://www.imp.uni-erlangen.de/phantoms/

5.1.1 Transmission measurement model

We assume the standard transmission measurement model for an x-ray imaging system—writing $\hat{c}_\ell$ to denote the number of transmitted photon counts along ray ℓ which encodes different source positions, then the forward data model after discretization is expressed as

$$\hat{c}_\ell = N_\ell \sum_i s_i \exp\left\{ -\sum_m x_{m\ell}\mu_{mi} \right\}, \tag{5.1}$$

where N_ℓ is the expected number of photon counts detected along ray ℓ in the absence of an object, s_i is the normalized distribution of x-ray photons at frequency i in the absence of an object (i.e. $\sum_i s_i$ is equal to 1), μ_{mi} is the linear x-ray attenuation coefficient for material m at frequency i, and $x_{m\ell}$ is the total amount of material m lying along ray ℓ. The model given in (5.1) is idealized and neglects numerous physical factors such as x-ray scatter.

The x-ray spectrum, given by s_i across frequencies indexed by i, comprises the energy spectrum of the x-ray source and the spectral response of the detector. In the task of spectrum estimation, transmisison measurements are acquired through known dimensions of known materials, so the attenuation functions $\{\mu_{mi}\}$ and the path lengths $\{x_{m\ell}\}$ are known and the only uknowns in (5.1) are the x-ray energy spectrum. The approach for reconstructing th spectrum studied in this section inverts the forward model (5.1) to estimate the x-ray spectrum, $\{s_i\}$, from noisy transmission measurements, $\{c_\ell\}$.

The difficulty inherenet in inverting (5.1), however, is twofold. First, the system matrix describing the attenuation of x-ray photons is highly low-rank, leading to the ill-conditioned linear system of the spectrum estimation problem. In particular, some form of regularization is necessary for reliable estimation of the x-ray spectrum. Second, the physical nature of an x-ray spectrum involves multiple structures in its shape, namely, the low-frequency component arising from bremsstrahlung radiation, which covers the entire range of the energy bins, and the high-frequency component arising from characteristic radiation, which produces sharp peaks at certain energy locations—for instance, see Figure 5.1 in Section 5.1.3 for a typical x-ray spectra. The challenge is to recover both structures simultaneously, so that the estimated spectrum accurately represents the spectral response of the x-ray imaging system.

Here we exploit prior knowledge of the x-ray spectrum to design a suitable regularizer when we formulate an optimization problem; in this way, we can allows for recovering both structures and at the same time overcome the ill-conditioning of the problem.

5.1.2 Related work

For an energy resolved CT system, the x-ray spectrum represents the product of the polychromatic source spectrum and the detector spectral response. Due to its importance in x-ray imaging, a number of methods have been proposed for obtaining a stable and accurate x-ray spectrum.

Using a physical model with few parameters effectively reduces the degrees of freedom of the problem, and allows for stable estimation of the spectrum by expressing a low-dimensional representation of the x-ray spectrum [56, 57, 58]. The parameters are fitted with least squares or other data discrepancy objectives. Meanwhile, [59] investigates an iterative perturbation method that minimizes differences between measured and calculated transmission curves using low-Z attenuators.

Various forms of regularization have been also employed to avoid the ill-conditioning of the problem and ensure stable spectrum estimation. For instance, [60] performs a minimization of the sum of a χ^2 objective term and a nonlinear regularization term to stabilize the final solution. [61] uses the expectation-maximization (EM) method to iteratively solve the ill-conditioned linear system and truncates the iteration of the algorithm at some finite iteration. Here early stopping serves as a sort of regularization as it prevents overfitting of the model. Singular value decomposition (SVD) is a more direct approach that attempts to directly invert the linear system to estimate each bin contents of spectrum [62, 63]. The SVD method often involves truncating smaller singular values and singular vectors of the system matrix, also known as truncated singular value decomposition (TSVD), since these components make almost no contribution to the measured data and are susceptible to the noise [64]. The obtained spectrum from TSVD is sufficiently accurate to model the measured transmission curve, but it has the drawback that positivity of the spectrum is not guaranteed and the solution exhibits negative values in some energy positions. Recently, an extension of the TSVD method, called prior truncated singular value decomposition (PTSVD), has been proposed in [65] to further incorporate prior information about the statistical nature of the transmission data and about high-frequency spectral components such as characteristic peaks. In particular, by exploiting basis vectors for the null space of the system matrix, the authors reconstruct an x-ray spectrum that accurately reproduces the physical shape of the ground truth spectrum.

5.1.3 EM method

For the purpose of comparison, here we describe the expectation-maximization (EM) approach which has frequently been applied to the problem of x-ray spectrum estimation (e.g. [61, 66]).

Broadly speaking, EM is a general framework for solving a maximum likelihood estimation problem when the obtained data is incomplete. In the setting of spectrum determination, the incompleteness of the data arises from the fact that the detected photon count along ray ℓ is observed through a sum of transmitted photon counts across frequencies i, namely, $c_\ell \approx \sum_i X_{\ell i} s_i$ for the system matrix $\{X_{\ell i}\}$. Under the Poisson noise, EM then finds the maximum likelihood estimate

$$\widehat{s} = \arg\min_s \left[\sum_\ell \left\{ \sum_i X_{\ell i} s_i - c_\ell \log \left(\sum_i X_{\ell i} s_i \right) \right\} \right],$$

by applying the iterations

$$s_i^{(t+1)} = \frac{s_i^{(t)}}{\sum_\ell X_{\ell i}} \sum_\ell \frac{X_{\ell i} c_\ell}{\sum_{i'} X_{\ell i'} s_{i'}^{(t)}} \text{ for all } i. \tag{5.2}$$

Here the update equation is derived by minimizing the EM objective function $Q(s; s^{(t)})$, given by

$$Q(s; s^{(t)}) = \sum_{\ell i} \left\{ X_{\ell i} s_i - c_\ell \frac{X_{\ell i} s_i^{(t)}}{\sum_{i'} X_{\ell i'} s_{i'}^{(t)}} \log(X_{\ell i} s_i) \right\}. \tag{5.3}$$

Note that the multiplicative form of the update equation (5.2) automatically guarantees non-negativity of the solution as long as the initial value is chosen to be non-negative. We also note that the EM iterations does not satisfy the normalization assumption, namely $\sum_i s_i^{(t)} = 1$ is not necessarily true.

For an underdetermined and noisy linear system, the maximum likelihood solution is known to overfit to the data and yield undesirable structure of the x-ray spectrum. Hence, it is desirable to minimize the data discrepancy function (the transmission Poisson likelihood function in this case) but possibly subject to constraints and/or regularization on the spectrum. Meanwhile, it is known that the iterations (5.2) are guaranteed to converge to the solution $\{\widehat{s_i}\}$ for any initial point. Therefore, if run to convergence, the EM iterations should reach the solution $\{\widehat{s_i}\}$ and thus fail to deliver accurate estimation of the x-ray spectrum. In particular, if we try to incorporate prior information about the x-ray spectrum into the initialization of EM, we would expect it to still end up at the same solution, $\{\widehat{s_i}\}$. To avoid this issue, early stopping of EM is often employed (e.g. [61]) to regularize the algorithm path and avoid overfitting to the data. Note that, due to the global convergence property of EM, the idea of incorporating prior information via initialization makes sense only in the context of early stopping.

In Figure 5.1, the spectral curves estimated by EM are shown for different numbers of iterations. As seen in the figure, by stopping after 500 iterations, the EM method recovers the ground truth spectrum remarkably well and both spectra are indistinguishable in the plot; however, for low iteration number (e.g. 10 iterations), the resulting spectrum is still biased towards the initial value, and for high iteration number (e.g. $50,000$ iterations), the EM method appears to overfit to the transmission data and therefore cannot generalize to transmission measurements beyond the given data set. While determining a good iteration number is crucial to implement the EM method, the authors in [61] further demonstrates the robustness of the EM method, i.e. the estimated x-ray spectrum is not strongly sensitive to the choice of number of iterations.

While EM enjoys many empirical advantages for spectrum reconstruction, our motivation to derive a spectrum reconstruction method from an optimization framework is to enhance interpretability and flexibility of the reconstruction procedure; for EM, it is not clear what kind of

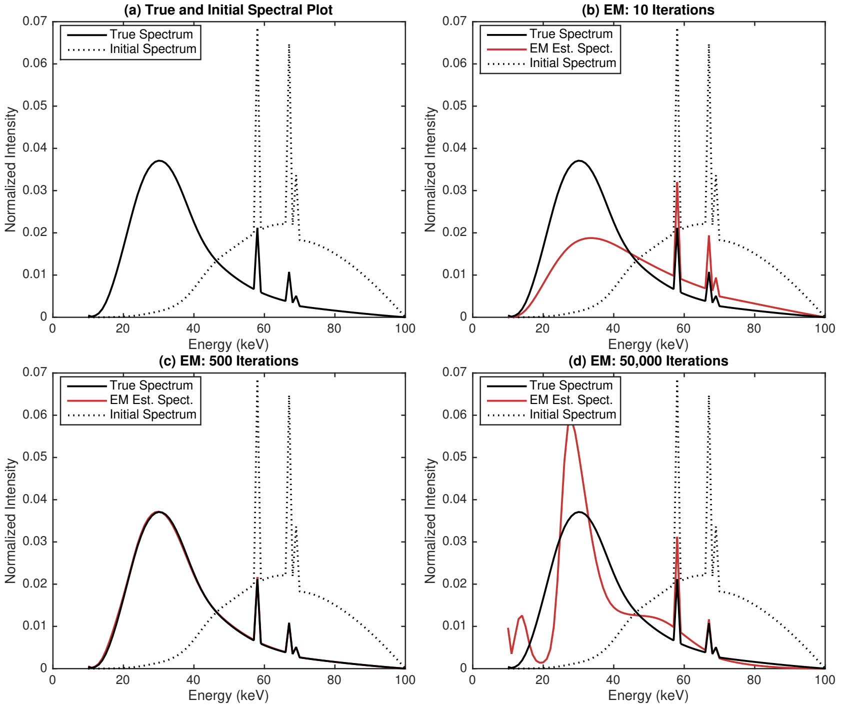

Figure 5.1: Spectrum estimation from simulated transmission measurements by use of EM. A detailed description of the simulation setting is given in the simulation study (Section 5.3.1). Panel (a) shows the ground truth x-ray spectrum (black solid line) and the initial x-ray spectrum (black dotted line). The remaining three panels show the estimated spectra by EM for different number of iterations.

regularization early stopping is performing for the algorithm, and other desirable constraints on the spectrum, such as a normalization constraint, cannot be easily incorporated. On the other hand, our framework is capable of including multiple constraints on the spectrum, and moreover, we do not require any form of early stopping but rather fully minimize target optimization problem for accurate reconstruction of x-ray spectrum. Our approach can also allow us to build towards simultaneous spectrum estimation and basis material maps reconstruction in spectral CT.

5.2 Spectrum estimation via a KL-divergence constraint

Now we turn to the development of our method to estimate the x-ray spectrum from transmission measurements through an optimization problem.

We assume that an initial spectrum, namely a prior estimate of the x-ray spectrum, is available such that the initial value exactly captures the characteristic peaks of the target spectrum (without such information, we cannot hope to recover details of the spectral curves such as the characteristic peaks.) Denoting the initial value by $\{s_i^{\text{ini}}\}$, we measure the distance of the x-ray spectrum $\{s_i\}$ to the initial value via Kullbeck-Leibler (KL) divergence, i.e. $\mathrm{d}_{\mathrm{KL}}(s\,;s^{\text{ini}})$, where for positive vectors $x \geq 0, y > 0$, the KL divergence is defined by

$$\mathrm{d}_{\mathrm{KL}}(x\,;y) = \sum_i \{x_i \log(x_i/y_i) + y_i - x_i\}\,. \tag{5.4}$$

(Note that the definition (5.4) reduces to the usual definition of KL-divergence over probability vectors, when the vectors x, y satisfy $\sum_i x_i = \sum_i y_i = 1$.) The KL-divergence is convex in (x,y) and satisfies $\mathrm{d}_{\mathrm{KL}}(x\,;y) \geq 0$ for $x \geq 0, y > 0$, and $\mathrm{d}_{\mathrm{KL}}(x\,;y) = 0$ if and only if $x = y$. In order to stabilize inversion of the data model, a KL-divergence constraint, i.e. a bound on $\mathrm{d}_{\mathrm{KL}}(s\,;s^{\text{ini}})$, is placed on the estimated x-ray spectrum $\{s_i\}$ to control the deviation from the initial value.

Specifically, the x-ray spectrum is reconstructed through the following constrained minimization problem:

$$\begin{aligned} \underset{s}{\text{minimize}} \quad & \mathrm{d}_{\mathrm{KL}}(c\,;Xs) \\ \text{subject to} \quad & \mathrm{d}_{\mathrm{KL}}(s\,;s^{\text{ini}}) \leq c, \\ & \sum_i s_i = 1, s_i \geq 0 \text{ for all } i, \end{aligned} \tag{5.5}$$

for a constraint parameter $c \geq 0$, where the KL-divergence is employed for both the data discrepancy function and the constraint function. Note that the data discrepancy function here, namely, KL-divergence between measured data and calculated photon counts, is equivalent to the transmission Poisson likelihood (TPL) function up to constant terms [67] hence the solution of the problem (5.5) is equivalent to a constrained maximum likelihood estimate of the counts data under a Poisson noise assumption. The TPL function can be useful even when the measured counts data is inconsistent with the Poisson assumption, since it assigns more weight to higher count measurements [12]. The constraint $\sum_i s_i = 1$ ensures normalization of the resulting solution, which endows physical meaning to the reconstructed x-ray spectrum.

Although the description of the data model (5.1) is idealized, the proposed optimization-based approach is flexible and can include other physical effects, such as x-ray scatter as well as other non-ideal detector effects, in the estimation process by adding constraints or modifying the objective function. The use of KL-divergence as a constraint function can be valid for any given optimization formulations. In terms of computation, the problem (5.5) is a convex program, so any convex solver can be applied to solve the problem efficiently. For instance, we have implemented the method using the "cvx" package in Matlab with solver MOSEK which solves the problem (5.5) in less than a second. Alternatively, we can apply the exponentiated-gradient (EG) algorithm, which is a simple first-order algorithm that iteratively performs a descent step followed by projection onto the feasible region of x-ray spectra. See Section 5.2.2 for a detailed discussion of the EG algorithm and convergence guarantees for obtaining optimal solution of the problem (5.5).

Care must be taken in specifying the initial value $\{s^{\text{ini}}\}$ as it has a great influence on the final estimation of the x-ray spectrum. If the employed initial value reflects realistic structure of a spectral curve, the resulting solution can provide accurate estimation of the target spectrum and therefore accurately reproduce transmission measurements. The robustness of the method with respect to the initial spectrum is also investigated in the simulation study (see Section 5.3.1).

5.2.1 Connection to maximum entropy method

The proposed method based on KL-divergence is closely related to the well known *principle of maximum entropy* in the existing literature. This principle state that, of all possible solutions that are consistent with the data, we choose the one with the largest entropy $-\sum_i s_i \log s_i$, or with the least divergence (or relative entropy) $\sum_i s_i \log(s_i/s_i^{\text{ini}})$ if the prior information $\{s_i^{\text{ini}}\}$ is known. The maximum entropy principle has been widely studied in the following decades, with applications to a broad range of problems including image reconstruction from incomplete and noisy data [68]. We refer the reader to [69] for justification of the principle.

In the context of spectrum estimation, applying the maximum entropy principle with prior information $\{s_i^{\text{ini}}\}$ leads to the following constrained optimization problem:

$$\begin{aligned} \underset{s}{\text{minimize}} \quad & \mathrm{d}_{\mathrm{KL}}(s\,;s^{\text{ini}}) \\ \text{subject to} \quad & \mathrm{d}_{\mathrm{KL}}(c\,;Xs) \leq C, \\ & \sum_i s_i = 1, s_i \geq 0 \text{ for all } i, \end{aligned} \tag{5.6}$$

where we again employ the TPL discrepancy function as a measure of the fit to the data, and $C > 0$ is a parameter that limits the amount of this discrepancy.

Now, since the problem is convex in the variable $\{s_i\}$, we can find a one-to-one correspondence between the parameters c in (5.5) and C in (5.6) such that the solutions from both optimization problems exactly match; this, in turn, implies that the problem (5.5) is equivalent to the problem (5.6), and particularly shows the equivalence between the proposed approach and the maximum entropy principle. This provides a justification of the use of KL-divergence as a constraint function for spectrum estimation. On the other hand, note that the convexity of the TPL discrepancy function is essential here. While the KL-divergence constraint can be applied to the data models including other physical factors, the resulting data discrepancy function can generally be nonconvex in which case the equivalence property is no longer guaranteed to hold. Even in such case, however, we believe that a similar kind of interpretation can be useful in gaining insight into the constrained approach with KL-divergence.

5.2.2 Exponentiated-gradient algorithm

While the problem (5.5) can generally be solved by any convex solver, in some applications, it is useful to have an iterative algorithm that solves the problem more explicitly. In this work, we solve this optimization problem using the exponentiated-gradient (EG) algorithm [11], that is designed to solve general convex objectives over the simplex $\{s : \sum_i s_i = 1, s_i \geq 0 \text{ for all } i\}$. Exponentiated-gradient algorithm can also be viewed as a special case of mirror descent with the mirror map given as the negative entropy function [40].

First, we write the constrained problem (5.5) in the equivalent Lagrangian form:

$$\begin{aligned} \underset{s}{\text{minimize}} \quad & \mathrm{d_{KL}}(c\,;Xs) + \lambda \cdot \mathrm{d_{KL}}(s\,;s^{\text{ini}}) \\ \text{subject to} \quad & \sum_i s_i = 1, s_i \geq 0 \text{ for all } i, \end{aligned} \tag{5.7}$$

where λ is a regularization parameter that controls the amount of regularizing effect, and the constraints represent the feasible region of x-ray spectra. Again, there is a one-to-one correspondence between c in (5.5) and λ in (5.7), due to the convexity of the problem.

The EG algorithm applied to the above problem yields the following iterations: initialize $s^{(0)} = s^{\text{ini}}$, fix the step size $\eta > 0$, then for steps $t = 0, 1, 2, \ldots$,

$$\begin{cases} \text{Set } g^{(t)} = \nabla_s \mathrm{d_{KL}}(c\,;Xs^{(t)}) + \lambda \nabla_s \mathrm{d_{KL}}(s^{(t)}\,;s^{\text{ini}}); \\ \text{Set } s_i^{(t+1)} \leftarrow s_i^{(t)} \exp\left(-\eta \cdot g_i^{(t)}\right) \text{ for all } i; \\ \text{Set } s^{(t+1)} \leftarrow s^{(t+1)} / \sum_i s_i^{(t+1)}. \end{cases} \tag{5.8}$$

Now examining the steps given in (5.8), we see that the update equation of $s_i^{(t+1)}$ is multiplicative as analogous to the EM iterations (5.2). Particularly, this guarantees automatic inclusion of non-negativity constraints in the estimated spectrum, as long as the initial spectrum is non-negative. On the other hand, a distinct feature of the EG algorithm is that at every iteration the normalization constraint is enforced by the projection step $s^{(t+1)} / \sum_i s_i^{(t+1)}$ (more precisely, the projection is performed with respect to the KL-divergence), whereas the EM method can give no such guarantees on the final solution. The projection step can be optional, and is not needed if the normalization constraint is not included in (5.7). To compare the EG and EM algorithms, while the EM algorithm seeks to minimize (5.3) at each iteration to reach the maximum likelihood solution (if EM is run to convergence), EG instead seeks to take each step that monotonically decreases (5.3) with additional KL-divergence regularization term. Both algorithms will produce a sequence of estimates that will decrease the (regularized) data discrepancy at each iteration.

5.2.3 Convergence test

The convergence of the EG algorithm has been well established in the literature—for instance, Bubeck [40] shows that the objective gap between the point at iteration t and the optimal solution decays with the rate $\mathscr{O}\left(\frac{1}{t}\right)$. The convergence test of the EG algorithm has appeared in [70], which we provide here for the sake of completeness. The Lagrangian function associated with the problem (5.7) is given by

$$\mathscr{L}(s,\nu,\gamma) = \mathrm{d_{KL}}(c;Xs) + \lambda \cdot \mathrm{d_{KL}}(s;s^{\mathrm{ini}}) - \sum_i \nu_i s_i + \gamma \cdot (\sum_i s_i - 1).$$

By the KKT condition, the optimal solution satisfies the following conditions:

(a) $\sum_i s_i = 1$ and $s_i \geq 0 \; \forall i$.

(b) $\nu_i \geq 0 \; \forall i$.

(c) $\nu_i s_i = 0 \; \forall i$.

(d) $\nabla_s \mathscr{L}(s,\nu,\gamma) = 0$.

Set $\gamma = -\min(\nabla_s \mathrm{d_{KL}}(c;Xs) + \lambda \nabla_s \mathrm{d_{KL}}(s;s^{\mathrm{ini}}))$ and $\nu = \nabla_s \mathrm{d_{KL}}(c;Xs) + \lambda \nabla_s \mathrm{d_{KL}}(s;s^{\mathrm{ini}}) + \gamma \cdot \mathbf{1}$, where min is taken componentwise. Then it can be checked that the conditions (b),(d) are satisfied. Also the condition (a) is trivial since the optimal solution is always feasible from the update equation (5.8). It remains to check the complementary slackness condition (c). By the conditions (a),(b), we know that $\nu_i \cdot s_i$ is non-negative, so the condition (c) is implied if $\sum_i \nu_i s_i = 0$. Therefore, we can test convergence of the algorithm by checking $\sum_i \nu_i s_i < \varepsilon$ for a predefined threshold $\varepsilon > 0$.

5.3 Numerical analysis

5.3.1 Simulation study

Now we perform a numerical experiment on the simulated transmission measurements to examine the empirical performance of the proposed method, as well as compare to the EM method.

A step wedge phantom is modeled and simulated, consisting of Aluminum and Polymethyl Methacrylate (PMMA). The thicknesses of Aluminum and PMMA are each selected in the range of $\{0, 0.635, 1.270, 1.905, 2.540\}$ and $\{0, 2.540, 5.080, 7.620, 10.160\}$ respectively, giving a total of 25 combinations across the step wedge. The linear attenuation coefficients are obtained using the NIST table [71]. Three kinds of polychromatic spectra, sampled at 1 keV intervals between 10 keV and 100 keV, are employed to either generate transmission measurements, or to serve as an initial value for the effective spectrum estimation; those spectra are determined from the experimental data described in Section 5.3.2, and represent a typical spectral response of the photon-counting CT system for energy windows with thresholds at 25 keV, 40 keV, and 60 keV. Using the experimentally determined spectra allows us to model the rational shape of the x-ray spectrum.

Given the true spectrum, the expected total transmitted photon counts $\{\hat{c}_\ell\}$ are computed according to the data model (5.1) with expected incident photon counts $N_\ell = 10^5$ for each ray ℓ. The noisy measurements $\{c_\ell\}$ are then generated with an independent Poisson model from which the true x-ray spectrum is reconstructed. Additionally, we generate another set of noisy transmission measurements through 20 different thicknesses of water which are varied from 0 to 20 centimeters at equal intervals, and where the NIST values are used to obtain the energy dependent attenuation coefficients. These measurements are not included in the reconstruction of the x-ray spectrum, but will serve as a "validation" set to assess the reproducibility of the spectrum estimation methods.

The x-ray spectrum is reconstructed by solving the optimization problem (5.7) with an implementation of the EG algorithm, as described in Section 5.2.2. Recall that λ is the user-defined parameter to control the trade-off between the data fidelity of the model and the regularization on the KL-divergence of the solution. We vary λ over $\lambda \in \{20, 30, \ldots, 1000\}$, and select the value

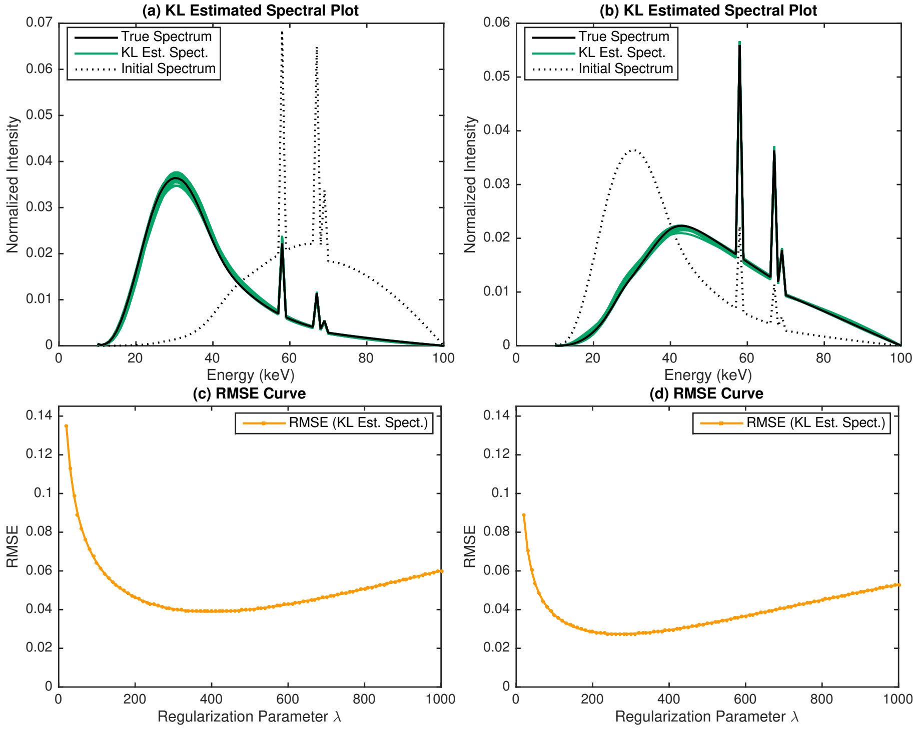

Figure 5.2: Spectrum estimation from simulated transmission measurements by use of KL. Different types of true and initial x-ray spectrum are employed shown in black solid line and black dotted lines respectively. In each setting, spectrum reconstruction is performed for 20 independent sets of transmission measurements. Panels (a),(b): Spectral curves for 20 different trials. The band formed by the curves shows variation between the reconstructed x-ray spectra. Panels (c),(d): The RMSE curves computed by (5.9) for different regularization parameters. Each point represents an average over 20 trials.

that minimizes the root mean square error (RMSE)

$$\text{RMSE}(\lambda) = \sqrt{\frac{\sum_i (s_i(\lambda) - s_i^{\text{true}})^2}{\sum_i (s_i^{\text{true}})^2}}, \tag{5.9}$$

where $\{s_i(\lambda)\}$ is the estimated spectrum given this choice of λ, and $\{s_i^{\text{true}}\}$ is the true spectrum. The spectrum achieving the minimum RMSE will be close to the true spectrum in shape, and thus can reliably reproduce transmission curves for any configurations of materials. For step size, we fix $\eta = 1.3 \cdot 10^{-5}$ throughout the simulation. We run the EG algorithm (5.8) until convergence, where we check the convergence of the algorithm as given in Section 5.2.3. For the present work, we set the threshold $\varepsilon = 10^{-8}$.

Figure 5.2(a,b) show the spectral curves reconstructed from transmission measurements by employing the ground truth and initial spectrum shown in the figures, respectively. For each given ground truth and initial spectrum, we simulate 20 independent sets of transmission measurements and obtain the best spectrum solutions by running the EG algorithm. Hence, each plot of Figure 5.2(a,b) shows reconstructed x-ray spectrum for 20 different sets of measurements. Due to the noise, there exist some variation between the spectral curves. As seen in the figures, however, the spectra generated by the method are concentrated near the respective true spectra and furthermore every single spectrum resembles the shape of its target with a high precision. More importantly, the results further show the robustness to the shapes of the chosen ground truth and initial spectra; the method continues to perform well, as long as the initial spectrum shares the same locations of the characteristic peaks as the true spectrum even though the relative intensities can be substantially different. This property can be particularly favorable for spectral calibration of a photon-counting detector, since spectral information from one energy window can be useful for estimating the spectral response of other windows.

The lower row of Figure 5.2 displays the RMSE plots, averaged over 20 trials, with respect to regularization parameter λ. For the two plots, the method yields larger error at first, but drops rapidly thereafter and achieves a minimum at λ in the range of 200–400. The error remains relatively lower in a broad range of λ's around the minimum, which illustrates that the method is numerically stable relative to the choice of λ. At larger values of λ, bias is induced in the solution and the error from the true spectrum begins to grow again. In comparison to the other case, the RMSE curve is placed higher in Figure 5.2(c), which results from the fact that the employed initial spectrum is farther from the truth than the other case.

Figure 5.3(a) and (b) show comparison of the spectra fitted by the KL-divergence based method and the EM method from simulated transmission measurements. For EM, the number of iterations is varied from 10 to 10^4 and the optimal number is chosen based on the RMSE rule described in (5.9). While it is seen that EM tends to estimate the true spectrum more faithfully (the averaged RMSE values by the best case KL and EM solutions are 0.0350 and 0.0184 respectively), the

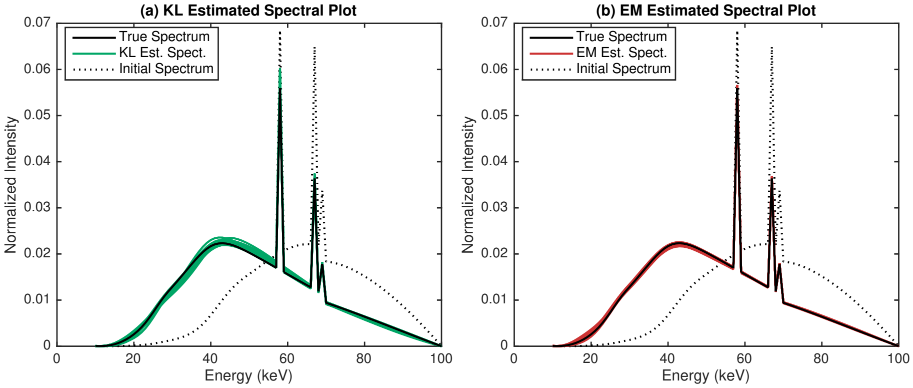

Figure 5.3: Comparison of spectrum estimation from simulated transmission measurements by use of KL and EM. Results for spectral curves, fitted by KL and EM respectively, are displayed for 20 different trials.

spectrum representations by both methods generally exhibit comparable performance in recovering physical shape of the true spectrum. Moreover, the utility of the KL-divergence approach lies in the mathematical formulation of spectrum estimation as an optimization problem.

Next, we evaluate the prediction of the transmission curves using the spectrum estimates based on water transmission measurements at 20 thicknesses. We use the ℓ_2-distance for log counts

$$\sum_{\ell'} (\log(c_{\ell'}) - \log(\hat{c}_{\ell'}))^2 \tag{5.10}$$

to measure the prediction performance. Figure 5.4(a) shows the prediction error of the KL-divergence approach plotted against the varying regularization parameter, as well as the prediction by the best case EM solution (which, recall, minimizes the RMSE criterion in (5.9)) and the true spectrum for reference (note that even the true spectrum cannot perfectly reproduce the transmission data due to the noise). For small values of λ, the KL-divergence approach performs nearly as well as the best case EM solution and slightly less than the true spectrum, demonstrating its capability to represent the measurement process; for higher values of λ, however, the performance rapidly degrades which results from the underfitting of the model. Figure 5.4(b) displays the actual transmission curves predicted by both methods, as well as the simulated water transmission data and the transmission curve predicted by the initial spectrum. Without loss of generality, here

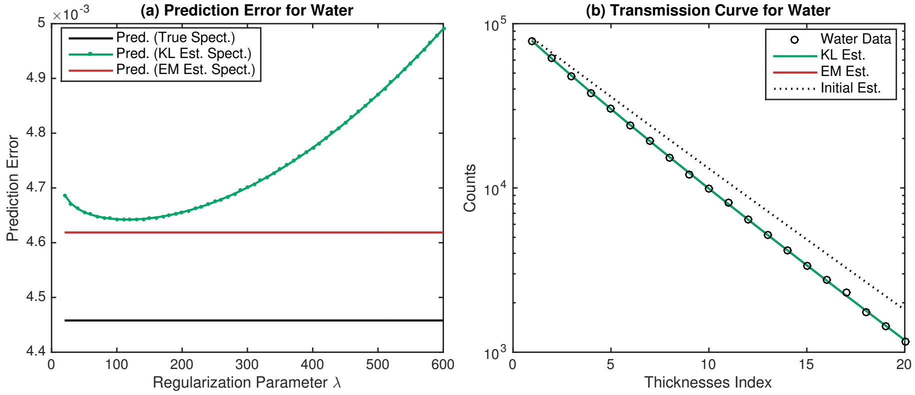

Figure 5.4: Panel (a): Prediction error in the transmission curves derived from the x-ray spectra using KL and EM. For the EM error, the best case solution is used to produce the transmission curve, which is irrelevant with respect to the regularization parameters. Each point represents an average over 20 trials. Panel (b): Plot of the predicted transmission curves for the reference material water. The x-axis indicates the thicknesses index ℓ' for water, and the y-axis is plotted on a logarithmic scale. The results for KL and EM are nearly identical and cannot be distinguished in the plot.

we only give a representative result from different trials. Again it is clearly seen that both predicted transmission curves are accurate enough to predict the water transmission data and show the significant improvement over the transmission curve predicted by the initial spectrum.

5.3.2 *Experimental study*

The proposed KL-divergence approach is evaluated on the experimental data which is performed on a bench-top x-ray system consisting of a microfocus x-ray tube and a photon-counting Cadmium-Zinc-Telluride (CZT) detector comprised of 128 detector pixels, of which 96 are usable. A step wedge phantom made of Aluminum and PMMA, shown in Figure 5.5(a), are measured at the

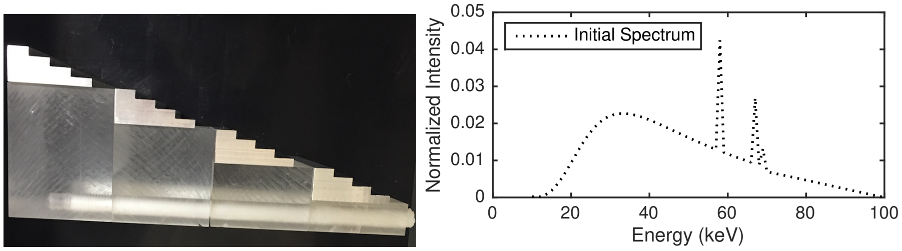

Figure 5.5: Left: (a) Step wedge phantom used for spectral calibration in x-ray imaging. Right: (b) The initial x-ray spectrum.

same dimensions with the simulated step wedge as described in Section 5.3.1. We refer the reader to Schmidt et al. [55] for more details of the experimental setup.

The initial spectrum is generated with the SPEC78 software from the IPEM78 report [72], which contains the expected spectrum exiting the tube for a 100 kV beam with 1-mm of aluminum filtration. Based on the measurement sets, reconstruction is performed with the KL-divergence regularized problem (5.7) to estimate effective spectral response of the photon-counting detectors for each energy window and detector pixel. Determining a good regularization parameter is critical in obtaining an accurate x-ray spectrum. The RMSE rule (5.9) cannot be applied here, since the true spectrum is unknown in the experimental setting. A validation method is another attractive option to choose a good value of λ, for which we randomly partition the transmission measurements into the training data and test data and select λ that best predicts the test data using the x-ray spectrum reconstructed from the training data. While the validation method is observed to perform well in the simulation setting, we find that when applied to the experimental setting, the estimated spectra tend to highly ovefit the experimental data and show unphysical fluctuations in the resulting curves. This is attributed to the systematic dependencies present in the measured photon counts, which can arise from various non-ideal physical effects of photon-counting detectors that have not been included in the data model (5.1).

For the current experiment, we instead rely on *ad hoc* procedure for selecting the optimal value of λ. The selection rule is based on the observation that the bremsstrahlung spectrum typically reveals unimodal structure in the corresponding energy region. The initial spectrum, shown in Figure 5.5(b), exhibits characteristic peaks at $58, 67, 69$ keV, but in other regions, the curve is smooth and nearly unimodal—it has a local minimum at s_{11}^{ini} (not visible in the figure), and a local maximum at s_{33}^{ini}. We expect to see this type of simple structure in the true spectrum as well. We therefore choose regularization parameters λ that yield the spectrum whose bremsstrahlung part reflects the same unimodal structure as the initial spectrum. More specifically, consider the spectrum $\{s_i\}$ constrained to the bremsstrahlung part of the frequency curve, by removing the characteristic

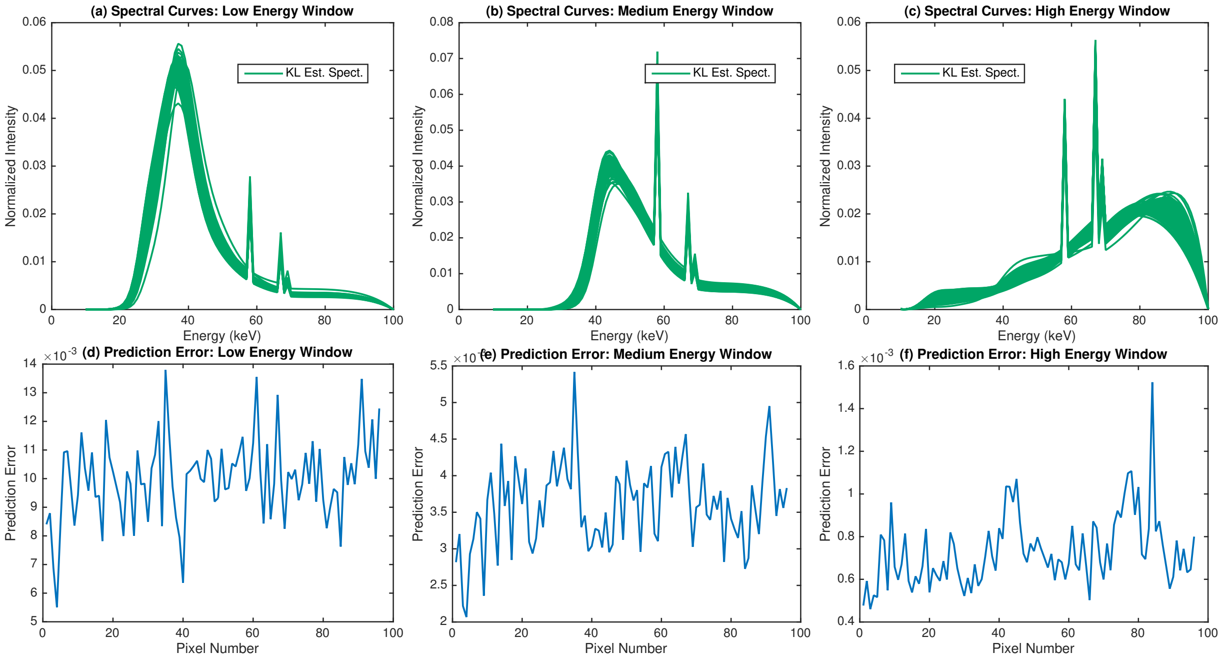

Figure 5.6: Spectrum estimation from the measured transmission data by use of KL. Each column represents the results for different spectral windows. Panels (a),(b),(c): Spectral curves for each energy window. The plots show the solution curves for 96 different detector pixels. Panels (d),(e),(f): Prediction error in the transmission curves derived from the x-ray spectra shown above across detector pixels.

peaks at $58, 67, 69$ keV:

$$s_{\text{brem}} = (s_{10}, \ldots, s_{57}, s_{59}, \ldots, s_{66}, s_{68}, s_{70}, \ldots, s_{100}),$$

i.e. the energy spectrum of photons is decomposed into s_{brem} and $s_{\text{char}} = (s_{58}, s_{67}, s_{69})$. We choose the regularization parameter by taking the smallest value of λ such that the estimated spectrum, $s(\lambda)$, exhibits at most one local minimum and one local maximum, when the characteristic peaks are removed—that is, at most one local minimum and one local maximum in the vector $(s(\lambda))_{\text{brem}}$, the bremsstrahlung part of the estimated spectrum. We expect that values of λ which are too small, leading to insufficient regularization, would yield an estimated spectrum $s(\lambda)$ that overfits to the data, which would typically exhibit many local minima and maxima; therefore, our procedure ensures that we choose a value of λ that is not too small, to avoid overfitting.

Results for experimental data are shown in Figure 5.6. Each panel in the upper row shows the reconstructed x-ray spectra for three different energy windows, as well as the initial spectrum depicted as the dotted line. Within each panel, the curves are obtained by running the EG algorithm from 96 different detector pixels, where step size is set to $\eta = 1.3 \cdot 10^{-5}$. While there is substantial variation in the reconstructed x-ray spectra across the detector pixels, the selection method based on the unimodality consisitently yields spectra that resemble realistic shapes of the bremsstrahlung and characteristic lines. Compared to the results for high energy window, the spectra estimated for low and medium energy windows appear to follow the realistic shape more faithfully. The spectral curves displayed in high energy window seem to be less stable and exhibit more fluctuations in the bremsstrahlung region. Similar results are also observed by comparing the prediction errors for different energy windows shown in the lower row of Figure 5.6, where it is suspected that the method tends to overfit to the data for high energy window in comparison to the other windows. Of course we can increase the penalization parameter λ to avoid this problem of overfitting, but the resulting spectra will now be strongly biased towards the initial spectrum. In principle, the problem of calibrating spectral response for high energy window is more difficult than the other cases, because the consecutive photons with low energies can be wrongly counted as the single photon with high energy, leading to a degradation of the spectral measurements in the high energy window.

To improve the stability of the estimated spectra for high energy window, we implement a simple variant of the proposed method that imposes KL-divergence regularization on the spectrum with different weights on each component of the spectral density s_i. In particular, we solve the following regularized optimization problem:

$$\begin{aligned} \underset{s}{\text{minimize}} \quad & \mathrm{d}_{\mathrm{KL}}(c; Xs) + \lambda \cdot \mathrm{d}^{\mathrm{w}}_{\mathrm{KL}}(s; s^{\mathrm{ini}}) \\ \text{subject to} \quad & \sum_i s_i = 1, s_i \geq 0 \text{ for all } i, \end{aligned} \tag{5.11}$$

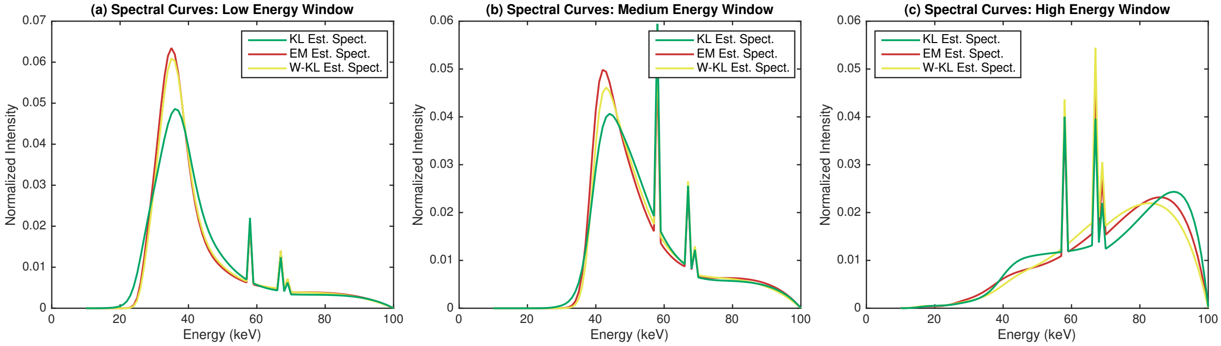

Figure 5.7: Comparison of spectrum estimation from the measured transmission data by use of KL, EM, and weighted KL. Results are shown for one particular detector pixel (pixel number = 34). Each panel shows the reconstructed spectra for different spectral windows, along with the initial spectrum.

where $\mathrm{d}_{\mathrm{KL}}^{\mathrm{w}}(x;y)$ is a weighted KL-divergence given by

$$\mathrm{d}_{\mathrm{KL}}^{\mathrm{w}}(x;y) = \sum_i w_i \left\{ x_i \log\left(x_i/y_i\right) + y_i - x_i \right\}, \tag{5.12}$$

for a weight vector $\{w_i\}$. In the current setting, each column of the system matrix $\{X_{\ell i}\}$ that contributes to the measured photon counts has different scalings, so we choose to use the weights such that $w_i \propto \sum_\ell X_{\ell i}$ for each i. This helps to treat different spectral densities s_i on a more equal basis. The EG algorithm can also be applied to solve the problem (5.11).

Figure 5.7 shows spectral curves reconstructed by the three methods, the original KL-divergence based method, its weighted version given in (5.11), and the EM method, from the measured counts data for different spectral windows. Here we fix the detector pixel (pixel number = 34) such that the spectrum returned by the KL-divergence approach exhibits some fluctuation in the high energy window. We can see that employing the weighted KL-divergence removes such unphysical shape in the resulting curve and makes the spectrum more smooth in the bremsstrahlung energy region. Moreover, it is interesting to see that in all energy windows, the weighted KL-divergence and the EM method yield x-ray spectra that are close in shape, but have some deviations from the x-ray spectra generated by the KL-divergence based method. We observe this phenomenon not only for the measured data at this particular detector pixel, but across all detector pixels. This is in sharp

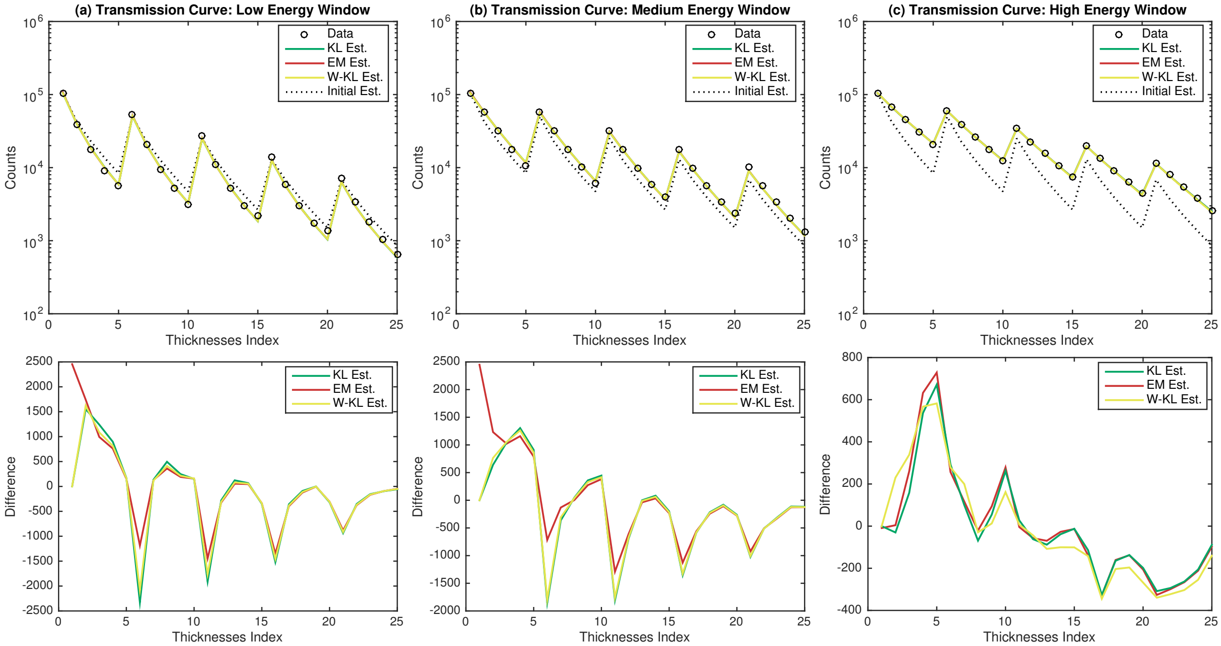

Figure 5.8: Comparison of spectrum estimation from the measured transmission data by use of KL, EM, and weighted KL. Results are shown for one particular detector pixel (pixel number = 34). The x-axis indicates the thicknesses index ℓ for the step wedge. The upper row of each panel shows prediction error in the transmission curves derived from the x-ray spectra shown in Figure 5.7. The lower row of each panel shows the residuals between the measured transmission curve and the predicted transmission curves.

contrast to the results shown in simulation study where both the KL-divergence approach and the EM method yield x-ray spectra that closely resemble the ground truth. Under the presence of inconsistency between the data model (5.1) and the physical transmission model, the KL-divergence based method can perform quite differently in comparison to EM and the weighted KL-divergence approach.

In Figure 5.8, the prediction performance is evaluated using the fitted x-ray spectra shown in Figure 5.7, where the error is computed according to the squared log count distance (5.10). We can see that all three methods significantly improve the prediction of the transmission curves compared to the initial spectrum. The residuals between the measured and predicted transmission curves are shown in the lower row of Figure 5.8. While the residuals generally behave similarly between the three methods, in the case of low and medium energy windows, the EM method generates larger residual errors for small thicknesses indexes; this is attributed to the fact that spectrum normalization constraint is not imposed in the EM solutions, which leads to errors in the transmission curves when there is no object in the scan system (thickness index 1 in the figure corresponds to the absence of an object). For high energy window, the three methods appear to perform similarly in terms of predicting the transmission data, though the curves between the KL-divergence approach and EM are more similar than the weighted KL-divergence. It is also worth noting that the plots displayed in Figure 5.8 clearly show visible trends in the residual errors, indicating the presence of systematic errors due to the unmodeled physics in the measurement process. In particular, this suggests the need for employing more realistic modeling of physical factors in order to account for the limitation of the method and enable more accurate and effective spectrum estimation.

5.4 Alternating minimization based framework for simultaneous spectrum estimation and image reconstruction

A recent development in detector technology has enabled the use of energy resolved photon-counting detectors. Photon-counting detectors acquire spectral information for the scanned object by separating incoming photons into pre-defined energy windows based on their energies. Using this energy information in CT imaging, also called spectral CT imaging, can mitigate beam-hardening artifacts and allows to estimate more than two basis material maps from the measured counts data.

In this section, we explore the use of KL-divergence approach for spectrum estimation to allow for auto-calibration of the spectral response of the imaging system in the spectral CT image reconstruction. Specifically, we incorporate unknown spectral components in the spectral CT data model and formulate simultaneous image reconstruction and estimation of these spectral components into the framework of alternating minimization. A simulation study is carried out to show how the algorithm can be implemented on spectral CT data.

5.4.1 Spectral CT data model

The spectral CT data model employed for the measured photon counts is given by

$$\hat{c}_{w\ell} = N_{w\ell} \sum_i s_{w\ell i} \exp\left\{ -\sum_{km} x_{\ell k} \mu_{mi} f_{km} \right\}, \tag{5.13}$$

where all the terms here are defined analogously as in (5.1), except that the density map f_{km} of material m in pixel k is now unknown and the measurements are acquired for each energy window w and each ray ℓ (here ℓ encodes different source and detector positions); $x_{\ell k}$ is the total length of the intersection between ray ℓ and pixel k which can be calculated from the scanning configuration of the CT scanner. For simplicity, we assume throughout this section that the spectral density $\{s_{w\ell i}\}$ is independent of ray ℓ, i.e. for each energy window w, the detector spectral response is equal across the detector pixels encoded by ℓ. Hereafter, we write $\{s_{wi}\}$ to denote the x-ray spectrum for energy window w.

In the idealized setting where the spectral components $\{s_{wi}\}$ are known, the only unknowns in the model (5.13) are the pixelized material maps $\{f_{km}\}$, and reconstruction algorithm determines these unknowns from noisy measured data $\{c_{w\ell}\}$. In particular, the one step reconstruction approach directly estimates the basis material maps from photon counts data by inverting the model (5.13),

$$\widehat{f} = \arg\min_{f \geq 0} D_{\mathrm{TPL}}(c, \hat{c}(f)) \ \text{ subject to } \ \|f_m\|_{\mathrm{TV}} \leq \gamma_m, \tag{5.14}$$

where the total variation norm $\|\cdot\|_{\mathrm{TV}}$ reflects the fact that the images exhibit locally constant or nearly-constant regions across pixels indexed by k. The transmission Poisson likelihood function $D_{\mathrm{TPL}}(\cdot,\cdot)$ is again employed for data discrepancy function. The optimization problem (5.14) is highly nonconvex, due to the nonlinear dependence of x-ray attenuations in the transmission model, and thus standard convex optimization techniques do not apply. To resolve this issue, Barber et al. [12] develops the mirrored convex/concave (MOCCA) algorithm, a nonconvex generalization of the Chambolle-Pock (CP) primal-dual algorithm [73], which can handle the optimization problem of the form (5.14). The details of the MOCCA algorithm are further explained in the next section.

In practice, the spectral response of the detectors are affected by various physical processes involved in a photon-counting detector and thus the spectral components $\{s_{wi}\}$ are never known exactly. Accurately calibrating the spectral response is an important task in the spectral CT image reconstruction as mis-calibrated detector elements can lead to strong ring artifacts in the reconstructed image. To account for such nonideal detector effects, we propose to calibrate the spectral response while performing image reconstruction, to adjust for the incorrectly estimated spectra in the spectral CT data model. Specifically, the optimization problem employed for simultaneous recovery of the material maps $\{f_{km}\}$ and the x-ray spectrum $\{s_{wi}\}$ is

$$(\widehat{f},\widehat{s}) = \underset{f\geq 0, s\in\mathscr{S}}{\arg\min}\ D_{\text{TPL}}(c,\hat{c}(f,s)) \ \text{ subject to } \ \|f_m\|_{\text{TV}} \leq \gamma_m,\ \text{d}_{\text{KL}}(s_w;s_w^{\text{ini}}) \leq c_w, \tag{5.15}$$

where $\mathscr{S}$ indicates the feasible set of the x-ray spectrum for each energy window w, namely $\mathscr{S} = \{s : \sum_i s_{wi} = 1, s_{wi} \geq 0\}$, and $c_w > 0$ is the KL-divergence constraint parameter. This problem is a natural extension of the one step inversion approach in (5.14) by jointly estimating the unknown spectral response with the basis material maps. The KL-divergence constraint is placed on the spectrum variables in order to capture important features of x-ray spectra and prevent overfitting to the measured counts. The initial estimates $\{s_{wi}^{\text{ini}}\}$ can be obtained from previous direct measurement of the spectrum, or from preliminary calibration step using transmission measurements before conducting image reconstruction.

The optimization problem (5.15) fits within the framework of alternating minimization since the TPL function can easily be minimized with respect to each variable while fixing the other variable. In particular, we can alternate between updating the material maps $\{f_{km}\}$ and the spectrum variables $\{s_{wi}\}$ at every iteration as:

$$\begin{cases} \text{For a fixed } \{s_{wi}\}\text{, take one step of the MOCCA algorithm to update } \{f_{km}\}; \\ \text{For a fixed } \{f_{km}\}\text{, take one step of the EG algorithm to update } \{s_{wi}\}. \end{cases}$$

In other words, given the current estimated spectra, the material maps are iteratively refined to find better images, while the update steps on the spectra components always reduces the data discrepancy function as the minimization problem of (5.15) is convex with respect to $\{s_{wi}\}$.

5.4.2 *An algorithm for simultaneous spectral calibration image reconstruction*

We now derive each step of the alternating minimization algorithm for simultaneous spectrum estimation and image reconstruction implemented in this chapter.

First, we fix $\{s_{wi}^{(n)}\}$. The problem (5.15) then reduces to the exact problem studied in [12] with known x-ray spectra. Specifically, denoting $L(f) = D_{\text{TPL}}(c, \hat{c}(f, s^{(n)}))$, it can be shown that $L(f)$ can be approximated by a quadratic convex function $Q(f; f_0) = F(z; z_0)$ at $Kf_0 = z_0$,

$$L(f) \approx \frac{1}{2} f^\top K^\top DKf - f^\top K^\top (b + EKf_0) := Q(f; f_0),$$

where the matrices K, D, E and the vector b depend on c and/or f (explicit formulas can be found in [12]). Denoting $\widetilde{F}(z) = \sum_m \mathbf{1}_{\|z_m\|_1 \leq \gamma_m}$ by indicator functions for the sparsity constraint of the gradient images, the MOCCA algorithm invokes primal-dual algorithm on the problem with the convex approximation $F(z; z_0)$ at the *mirrored* expansion point $z_0^{(n+1)} = \nabla_y F^*(y^{(n)}, z_0^{(n)})$

$$\min_{f \geq 0} \max_{y, \widetilde{y}} \langle Kf, y \rangle + \langle \nabla_{TV} f, \widetilde{y} \rangle - F^*(y; z_0^{(n+1)}) - \widetilde{F}^*(\widetilde{y}),$$

where $\nabla_{TV} f$ denotes the gradient operator applied to each of the material maps, and $F^*(y)$ and $\widetilde{F}^*(\widetilde{y})$ represent convex conjugates of $F(z)$ and $\widetilde{F}(z)$, respectively. Therefore, at iteration $(n+1)$, the iterations steps for MOCCA are given by

$$\begin{aligned}
y^{(n+1)} &= \operatorname*{arg\,min}_{y} \, \langle K\bar{f}^{(n)}, y\rangle - F^*(y; z_0^{(n+1)}) + \frac{1}{2}\|y - y^{(n)}\|^2_{\Sigma^{-1}}, \\
\widetilde{y}^{(n+1)} &= \operatorname*{arg\,min}_{y'} \, \langle \nabla_{TV}\bar{f}^{(n)}, \widetilde{y}\rangle - \widetilde{F}^*(\widetilde{y}) + \frac{1}{2}\|\widetilde{y} - \widetilde{y}^{(n)}\|^2_{\widetilde{\Sigma}^{-1}}, \\
f^{(n+1)} &= \operatorname*{arg\,min}_{f \geq 0} \, \langle Kf, y^{(n+1)}\rangle + \langle \nabla_{TV} f, \widetilde{y}^{(n+1)}\rangle + \mathrm{d}_{\mathrm{KL}}^{T^{-1}}(f\,; f^{(n)}), \\
\bar{f}^{(n+1)} &= 2f^{(n+1)} - f^{(n)},
\end{aligned}$$

where for a diagonal matrix W, we define $\mathrm{d}_{\mathrm{KL}}^{W}(f\,; f') = \mathrm{d}_{\mathrm{KL}}^{\mathrm{diag}(W)}(f\,; f')$ (see (5.12) for the definition of the weighted KL-divergence). Here in the update step of $f^{(n+1)}$, the (weighted) KL-divergence is used as the proximity term rather than the (weighted) ℓ_2-distance, which has the benefit of automatically including positiveness in the material maps without explicitly enforcing the positive constraints.

Next we fix $\{f_{km}^{(n+1)}\}$. We directly solve the constrained problem (5.15) with respect to the spectrum variables $\{s_{wi}\}$ without relying on the equivalent regularized form as in (5.7), i.e.

$$\text{Minimize } \{L(s) : s_{wi} \geq 0, \sum_i s_{wi} = 1\} \text{ subject to } \mathrm{d}_{\mathrm{KL}}(s_w\,; s_w^{\mathrm{ini}}) \leq c_w,$$

where $L(s) = D_{\mathrm{TPL}}(c, \hat{c}(f^{(n+1)}, s))$ denotes the TPL function at $f = f^{(n+1)}$. We use an iterative algorithm based on the alternating direction method of multipliers (ADMM) [74], whose preconditioned form is known to be closely related to the primal-dual algorithm [73]. Specifically, we employ a slight variant form of ADMM, called Bregman ADMM [75], which is known to perform well over the simplex $\mathscr{S} = \{s : \sum_i s_{wi} = 1, s_{wi} \geq 0\}$. The Bregman ADMM yields the following iterations when applied to the above minimization problem:

$$
\begin{aligned}
s^{(n+1)} &= \operatorname*{arg\,min}_{s\in\mathscr{S}} \; L(s) + \langle u^{(n)}, s\rangle + \rho \cdot \mathrm{d}_{\mathrm{KL}}(s; z^{(n)}) \\
&\approx \operatorname*{arg\,min}_{s\in\mathscr{S}} \; \langle \nabla L(s^{(n)}) + u^{(n)}, s\rangle + \rho \cdot \mathrm{d}_{\mathrm{KL}}(s; z^{(n)}) + \beta \cdot \mathrm{d}_{\mathrm{KL}}(s; s^{(n)}), \\
z^{(n+1)} &= \operatorname*{arg\,min}_{z} \; -\langle u^{(n)}, z\rangle + \rho \cdot \mathrm{d}_{\mathrm{KL}}(z; s^{(n+1)}) + \sum_{w} \mathbf{1}_{\mathrm{d}_{\mathrm{KL}}(z_w; s_w^{\mathrm{ini}}) \le c_w}, \\
u^{(n+1)} &= u^{(n)} + \rho(s^{(n+1)} - z^{(n+1)}).
\end{aligned}
$$

Note that in the step for $s^{(n+1)}$, we approximately solve the sub-problem using the Taylor expansion at the current points and adding the proximity terms. The $z^{(n+1)}$ step can be updated separately for each energy window w.

5.4.3 Simulation study

Here we implement the algorithm derived in Section 5.4.2 on simulated transmission measurements to investigate the potential of calibrating the detector spectral response during image reconstruction.

A pixelized two-material phantom from the FORBILD head phantom, shown in Figure 5.9, is simulated based on the spectral CT data model (5.13). Each of the true material maps, bone and brain maps, consists of 64×64 pixels and the linear attenuation coefficients for the corresponding materials are obtained from the NIST attenuation functions [71]. The number of detector bins is 64. Further details on the simulation setup, including the scanning configuration, can be found in [12].

The photon-counting detectors are simulated with two energy windows in the ranges of [20-70] keV and [70-120] keV and each energy window exhibits different spectral response as shown in Figure 5.10. We assume that the spectral response varies only with energy windows and are otherwise same across the detector pixels. For the spectral CT data, the number of expected total counts are set to 10^6 for each ray ℓ and 64 views are acquired for each detector pixel, giving a total

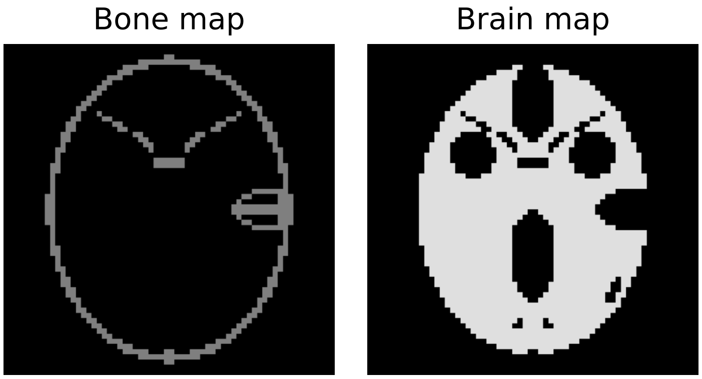

Figure 5.9: True bone and brain images shown in the gray scale window $[0.9, 1.1]$.

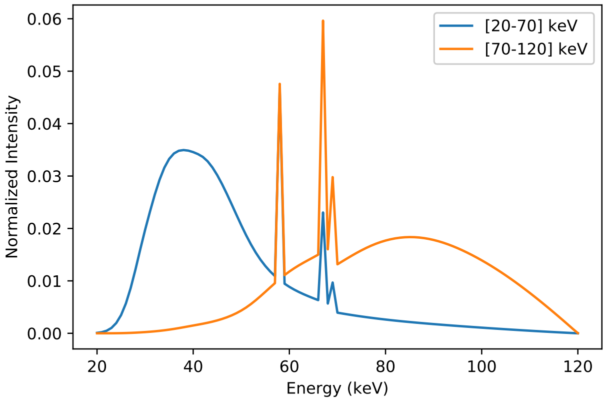

Figure 5.10: Spectral curves used for CT simulations for energy windows of [20-70] keV and [70-120] keV.

of 64×64 measurements for each energy window. The Poisson noise is added in the simulated counts data.

The image reconstruction is performed with bone and brain as the basis material maps. The simultaneous estimation of the detector response and image values depends on the choice of constraint parameters γ_m and c_w in (5.15). In this study, we fix the constraint parameters as the known actual values as both the true material maps and x-ray spectra are available in the simulation study. Further, the algorithm described in Section 5.4.2 needs to specify the tuning parameters, namely the diagonal matrices $\Sigma, \widetilde{\Sigma}, T \succ 0$ for the image reconstruction step and the step size parameters $\rho, \beta > 0$ for the calibration step. We follow the same strategy of Barber et al. [12] for choosing the diagonal preconditioners $\Sigma, \widetilde{\Sigma}, T \succ 0$, while we fix $\rho = 5 \cdot 10^8$ and $\beta = 6 \cdot 10^8$ throughout the simulation.

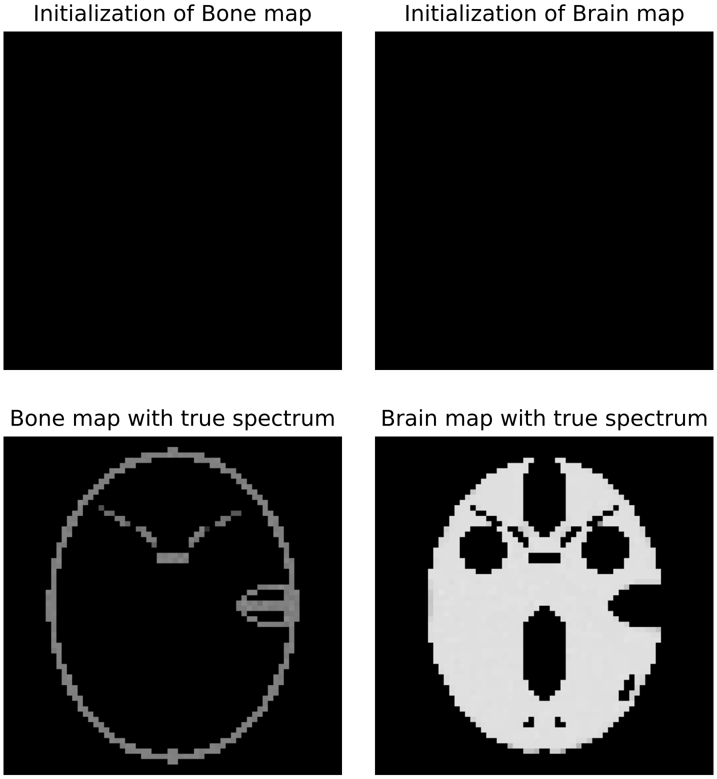
Initialization of Bone map
Initialization of Brain map
Bone map with true spectrum
Brain map with true spectrum

Figure 5.11: The upper row of each panel shows the initialized maps that are fed into the MOCCA algorithm. The lower row of each panel shows reconstructed material maps from noisy simulated measurements. The x-ray spectra that generate the transmission measurements are assumed to be known.

We first show the results for the ideal setting where we assume knowledge of the true x-ray spectra that generate the transmission measurements, as shown in Figure 5.10. In this setting, the MOCCA algorithm with TV constraints has been demonstrated to be effective even for undersampling. We apply the MOCCA algorithm to solve the problem (5.14) while initializing the material maps as depicted in the upper row of Figure 5.11. The reconstructed material maps from the two energy window CT data are presented in the lower row of Figure 5.11. As seen in the figure, the MOCCA algorithm accurately recovers the underlying structure of the true phantom material maps. Figure 5.12 indicates that the algorithm has reached nearly convergence to the solution after few thousands iterations.

Next we perform image reconstruction without knowledge of the exact distribution of spectral

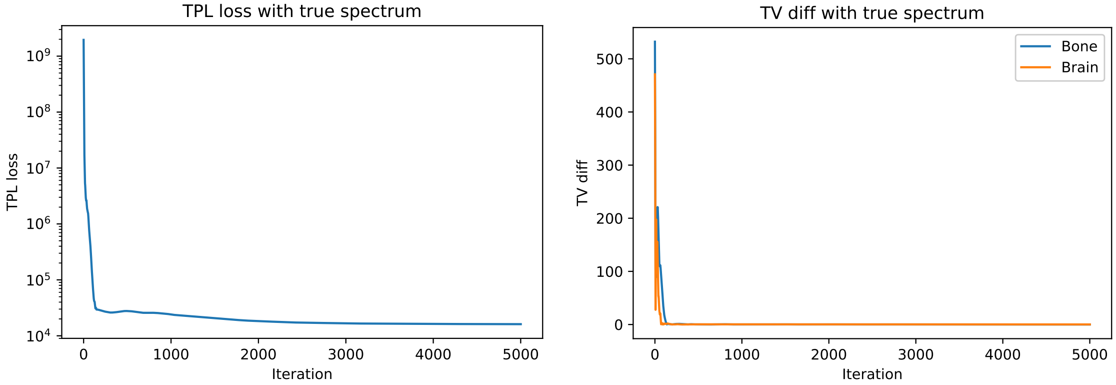

Figure 5.12: Plots of the transmission Poisson likelihood and difference between the TV values of the estimated maps and the true material maps across iterations $n = 0, 1, \ldots, 5000$.

components, but prior information is available to capture important features of the x-ray spectra. In particular, for each energy window, the initial x-ray spectrum is obtained by perturbing the corresponding true spectrum, which are then used in the reconstruction formula (5.15) to simultaneously determine the x-ray spectrum and the image values. For comparison, the MOCCA algorithm without estimating the spectrum variables is also performed, for which the initial estimates are treated as the truth throughout the iterations.

Figure 5.13 presents x-ray spectra estimated from the alternating minimization algorithm using the initial estimates displayed in the figures. Multiple results are shown varying the initial estimates employed for spectrum estimation. The reconstructed bone maps and brain maps that are estimated simultaneously with the spectra are also shown in Figure 5.14 and Figure 5.15. As clearly seen in the figures, the algorithm accurately recovers realistic estimate of the spectra and at the same time reduce artifacts in the reconstructed images compared to the results without spectrum calibration. Moreover, better reconstruction results are obtained when the initial estimate is more precise. This is confirmed by visually comparing the recontruction results in the figures, and also by examining the convergence behavior of the TPL function values shown in Figure 5.16, in which the algorithm enters the near convergence region more rapidly if the employed initial estimate is closer to the true spectrum. The qualitatively same behavior is also observed in the TV plot of Figure 5.17.

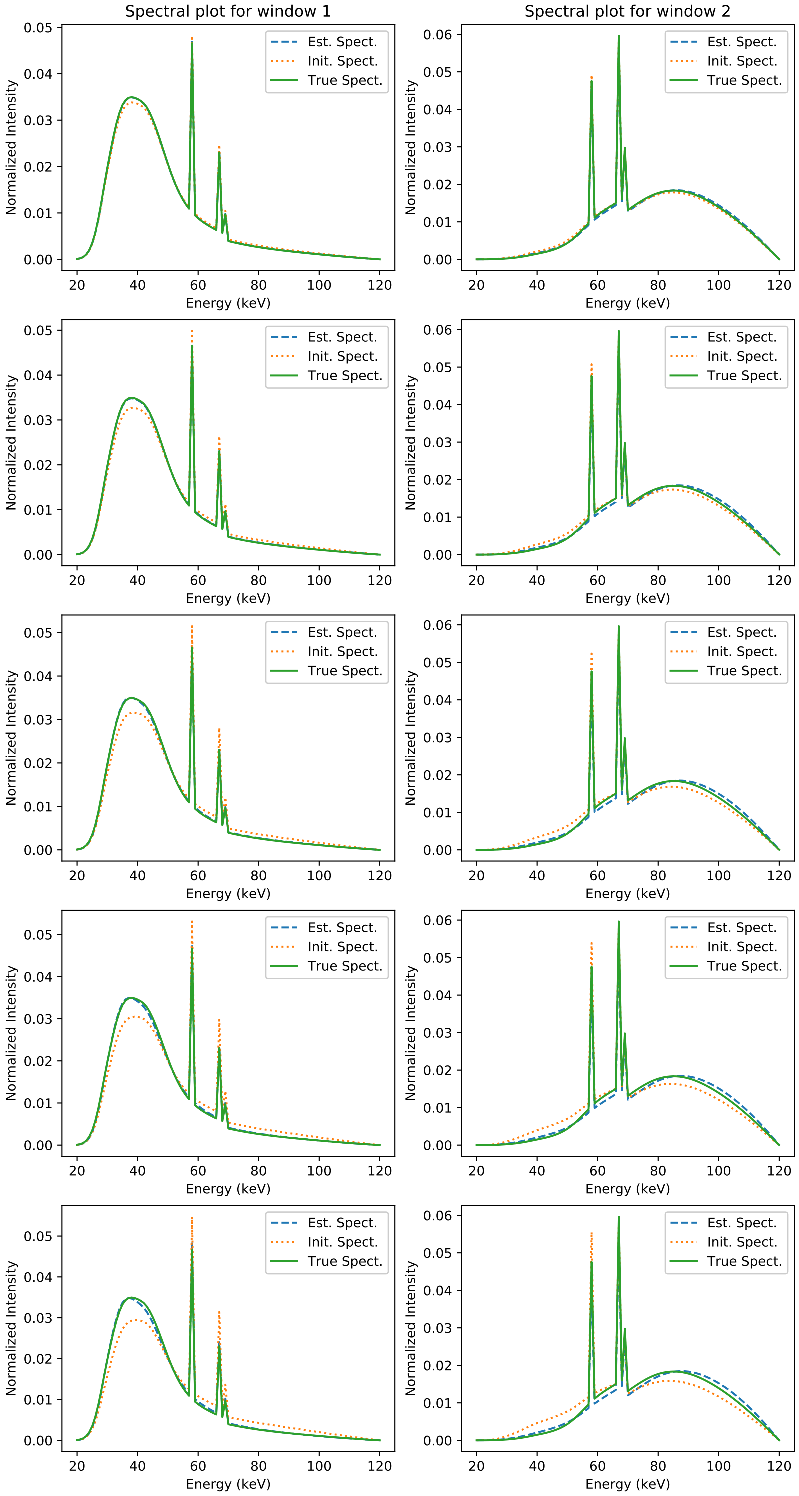

Figure 5.13: Spectrum estimation from the measured transmission data using alternating minimization for simultaneous spectrum estimation and image reconstruction. The true spectrum (green solid line) and the initial spectrum (yellow dotted line) are also displayed in the figures. Each column represents each of the energy window sensitivity of the detector for varying initial estimates.

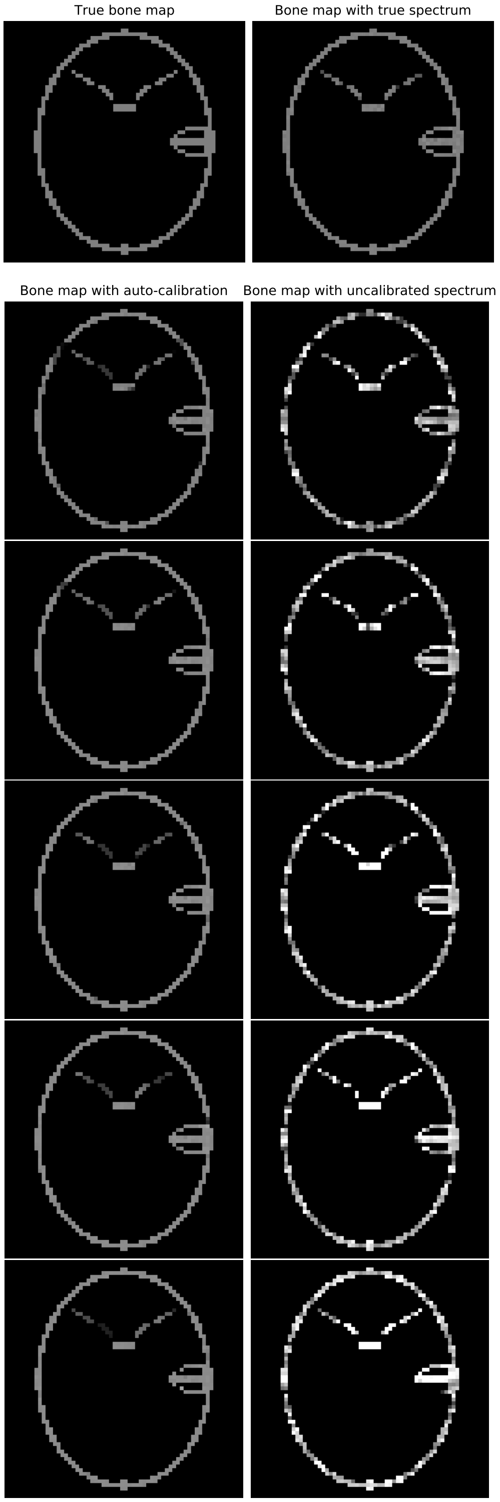
True bone map
Bone map with true spectrum
Bone map with auto-calibration
Bone map with uncalibrated spectrum

Figure 5.14: Reconstructed bone maps with and without spectral calibration. Each row corresponds to the different initial estimates of the spectrum as shown in Figure 5.13. For comparison, the true image and the reconstructed map with true spectrum are also displayed in the top panels.

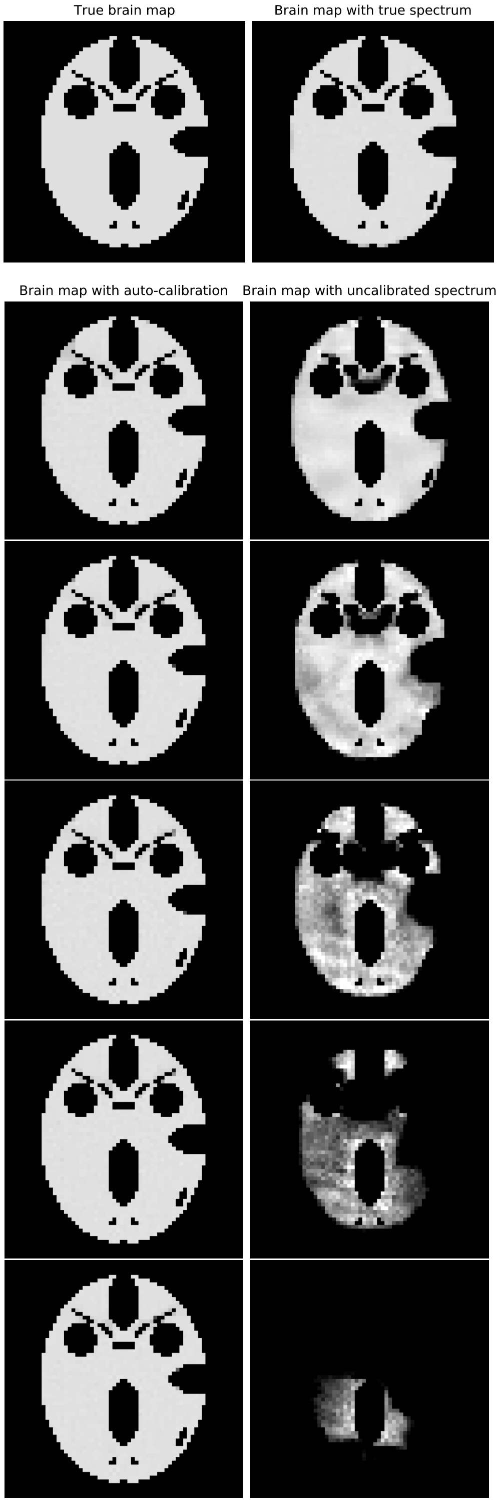
True brain map
Brain map with true spectrum
Brain map with auto-calibration
Brain map with uncalibrated spectrum

Figure 5.15: Reconstructed brain maps with and without spectral calibration. Each row corresponds to the different initial estimates of the spectrum as shown in Figure 5.13. For comparison, the true image and the reconstructed map with true spectrum are also displayed in the top panels.

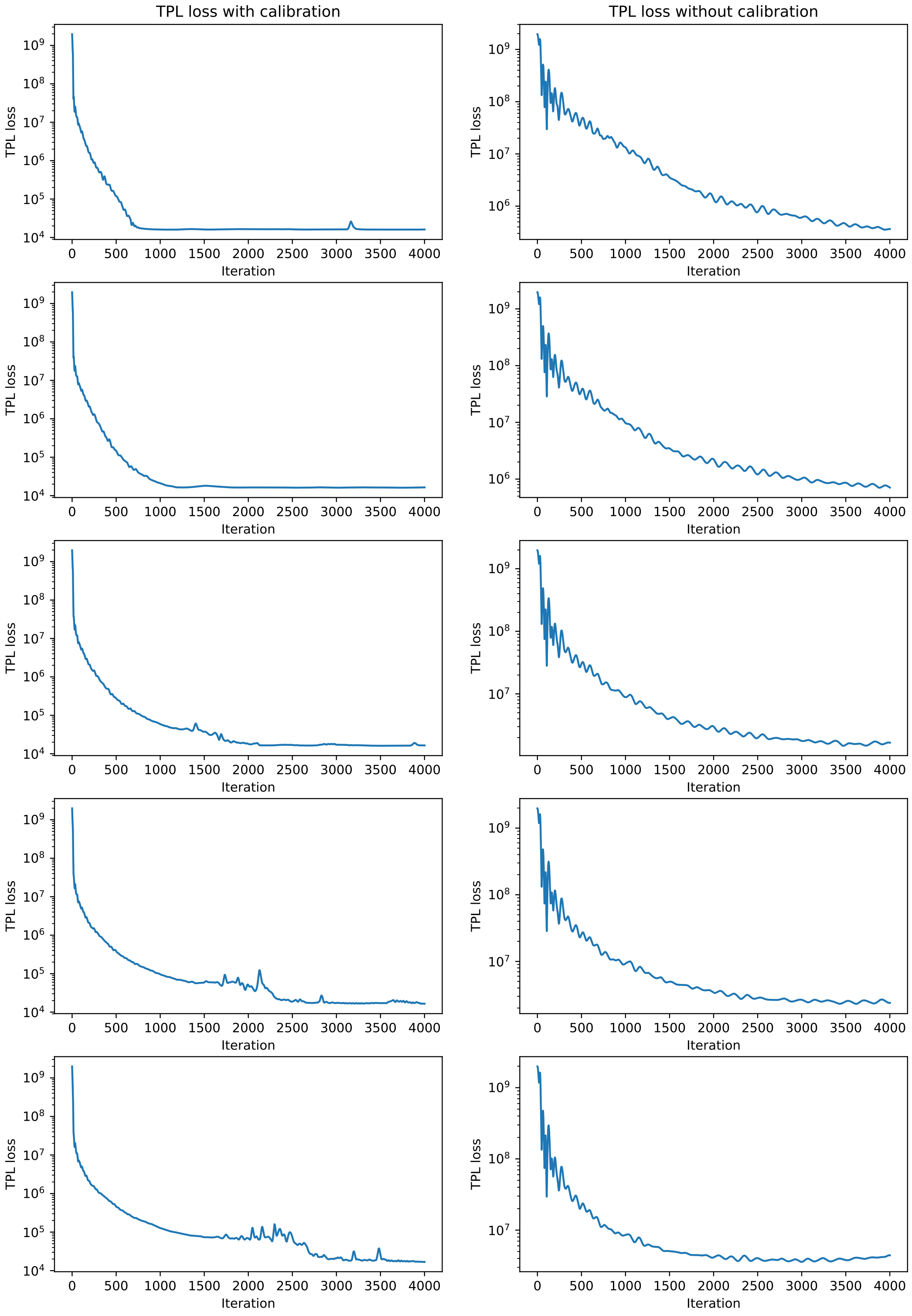

Figure 5.16: Plots of the transmission Poisson likelihood against iteration number for image reconstruction with and without spectrum calibration. Each row corresponds to the different initial estimates of the spectrum as shown in Figure 5.13.

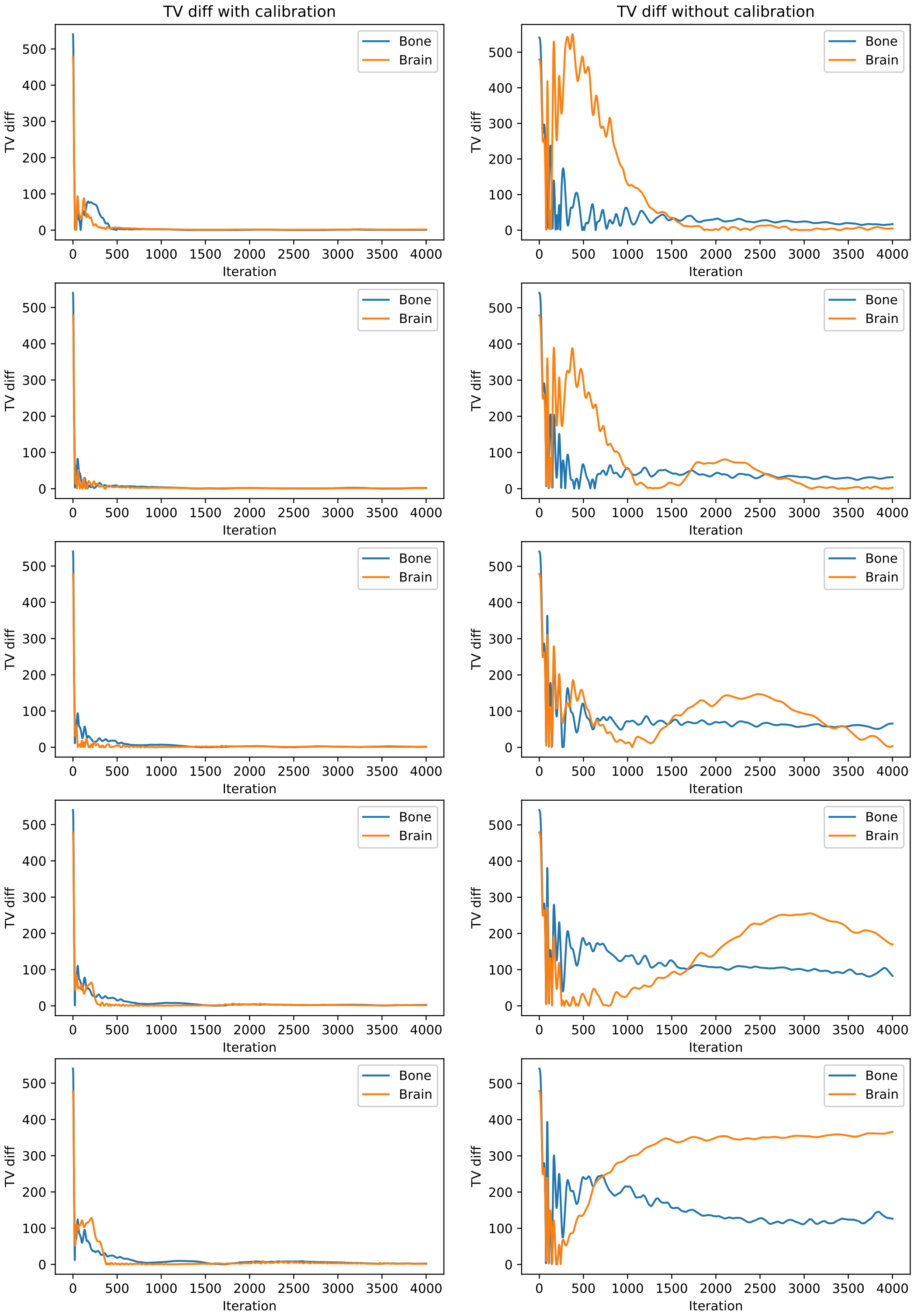

Figure 5.17: Plots of the difference between the TV values of the estimated maps and the true material maps against iteration number for image reconstruction with and without spectrum calibration. Each row corresponds to the different initial estimates of the spectrum as shown in Figure 5.13.

5.5 Discussion and conclusions

In this chapter, we have developed a constrained optimization problem for reconstructing x-ray spectrum from transmission measurements through known thicknesses of known materials. The proposed method places a KL-divergence constraint on the spectrum variable which improves numerical stability of the inversion process and allows to incorporate prior knowledge on the spectrum. The formulated optimization problem is a convex program over the simplex, which we propose to solve based on the exponentiated-gradient algorithm. Both numerical simulations and experimental results show that the method can yield realistic x-ray spectra that can accurately reproduce the spectral response of the CT system.

Here we emphasize two benefits of our approach relative to other methods. First, the proposed approach using a KL-divergence constraint provides the benefit of interpreting spectrum determination from transmission measurements in relation to the maximum entropy principle. Second, our formulation is a general optimization framework for spectrum estimation that can support different data discrepancy functions and incorporate other desirable constraints on the x-ray spectrum. More importantly, the flexibility of the method allows to easily incorporate the calibration procedure in the framework of simultaneous spectral calibration and spectral CT image reconstruction. In Section 5.4.3, we investigated the possibility of combining image reconstruction with the KL-divergence approach in the simple alternating minimization based framework which is seen to reduce artifacts in the reconstructed images. We hope that the work here may inspire future work to build a general framework for auto-calibration of the spectral response of the imaging system during the spectral CT image reconstruction.

BIBLIOGRAPHY

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, art. no. 11, 2011.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," *The Annals of Statistics*, pp. 1171–1197, 2012.

[3] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 1610–1613.

[4] M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Transactions on Information Theory*, vol. 59, pp. 5186–5205, 2013.

[5] A. Beck, "On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes," *SIAM Journal on Optimization*, vol. 25, pp. 185–209, 2015.

[6] Z.-Q. Luo and P. Tseng, "Error bounds and convergence analysis of feasible descent methods: a general approach," *Annals of Operations Research*, vol. 46, pp. 157–178, 1993.

[7] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Advances in Neural Information Processing Systems*, 2010, pp. 37–45.

[8] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 665–674.

[9] P. Jain and A. Tewari, "Alternating minimization for regression problems with vector-valued outputs," in *Advances in Neural Information Processing Systems*, 2015, pp. 1126–1134.

[10] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[11] J. Kivinen and M. K. Warmuth, “Exponentiated gradient versus gradient descent for linear predictors,” *Inf. Comput.*, vol. 132, pp. 1–63, 1997.

[12] R. F. Barber, E. Y. Sidky, T. G. Schmidt, and X. Pan, “An algorithm for constrained one-step inversion of spectral CT data,” *Phys. Med. Biol.*, vol. 61, pp. 3784–3818, 2016.

[13] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, “Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization,” in *Advances in Neural Information Processing Systems*, 2009, pp. 2080–2088.

[14] W. Ha and R. F. Barber, “Robust pca with compressed data,” in *Advances in Neural Information Processing Systems*, 2015, pp. 1936–1944.

[15] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, “Non-convex robust PCA,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1107–1115.

[16] H. Xu, C. Caramanis, and S. Sanghavi, “Robust PCA via outlier pursuit,” in *Advances in Neural Information Processing Systems*, 2010, pp. 2496–2504.

[17] J. He, L. Balzano, and J. Lui, “Online robust subspace tracking from partial information,” *arXiv preprint arXiv:1109.3827*, 2011.

[18] J. He, L. Balzano, and A. Szlam, “Incremental gradient on the grassmannian for online foreground and background separation in subsampled video,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1568–1575.

[19] O. Maillard and R. Munos, “Compressed least-squares regression,” in *Advances in Neural Information Processing Systems*, 2009, pp. 1213–1221.

[20] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, pp. 217–288, 2011.

[21] T. Zhou and D. Tao, "Godec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 33–40.

[22] S. Zhou, J. Lafferty, and L. Wasserman, "Compressed and privacy-sensitive sparse regression," *IEEE Transactions on Information Theory*, vol. 55, pp. 846–866, 2009.

[23] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of lasso and dantzig selector," *The Annals of Statistics*, pp. 1705–1732, 2009.

[24] R. Foygel, O. Shamir, N. Srebro, and R. R. Salakhutdinov, "Learning with the weighted trace-norm under arbitrary sampling distributions," in *Advances in Neural Information Processing Systems*, 2011, pp. 2133–2141.

[25] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and banach spaces," *Handbook of the geometry of Banach spaces*, vol. 1, pp. 317–366, 2001.

[26] J. A. Tropp *et al.*, "An introduction to matrix concentration inequalities," *Foundations and Trends® in Machine Learning*, vol. 8, pp. 1–230, 2015.

[27] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[28] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[29] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *The Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[30] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Transactions on Information Theory*, vol. 56, pp. 2053–2080, 2010.

[31] R. F. Barber and W. Ha, “Gradient descent with nonconvex constraints: local concavity determines convergence,” *arXiv preprint arXiv:1703.07755*, 2017.

[32] G. Colombo and L. Thibault, “Prox-regular sets and applications,” *Handbook of Nonconvex Analysis*, pp. 978–1, 2010.

[33] H. Federer, “Curvature measures,” *Transactions of the American Mathematical Society*, vol. 93, pp. 418–491, 1959.

[34] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[35] W. Su, M. Bogdan, E. Candes *et al.*, “False discoveries occur early on the lasso path,” *The Annals of Statistics*, vol. 45, pp. 2133–2150, 2017.

[36] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[37] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, vol. 38, pp. 894–942, 2010.

[38] E. J. Candès, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted ℓ_1 minimization,” *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008.

[39] P.-L. Loh and M. J. Wainwright, “Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima,” in *Advances in Neural Information Processing Systems*, 2013, pp. 476–484.

[40] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Foundations and Trends® in Machine Learning*, vol. 8, pp. 231–357, 2015.

[41] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis.* Springer Science & Business Media, 2009, vol. 317.

[42] R. Poliquin, R. Rockafellar, and L. Thibault, “Local differentiability of distance functions,” *Transactions of the American Mathematical Society*, vol. 352, pp. 5231–5249, 2000.

[43] E. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Communications of the ACM*, vol. 55, pp. 111–119, 2012.

[44] W. Ha and R. F. Barber, “Alternating minimization and alternating descent over nonconvex sets,” *arXiv preprint arXiv:1709.04451*, 2017.

[45] J. M. Ortega and W. C. Rheinboldt, “Iterative solution of nonlinear equations in several variables,” 1970.

[46] A. Auslender, “Optimisation: méthodes numériques,” 1976.

[47] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, “Living on the edge: Phase transitions in convex programs with random data,” *Information and Inference: A Journal of the IMA*, vol. 3, pp. 224–294, 2014.

[48] P. Zwiernik, C. Uhler, and D. Richards, “Maximum likelihood estimation for linear gaussian covariance models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

[49] A. J. Izenman, “Reduced-rank regression for the multivariate linear model,” *Journal of multivariate analysis*, vol. 5, pp. 248–264, 1975.

[50] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, “Low-rank solutions of linear matrix equations via procrustes flow,” *arXiv preprint arXiv:1507.03566*, 2015.

[51] S. Negahban and M. J. Wainwright, “Estimation of (near) low-rank matrices with noise and high-dimensional scaling,” *The Annals of Statistics*, pp. 1069–1097, 2011.

[52] K. Taguchi and J. S. Iwanczyk, “Vision 20/20: Single photon counting x-ray detectors in medical imaging,” *Med. Phys.*, vol. 40, art. no. 100901, 2013.

[53] W. Ha, E. Y. Sidky, R. F. Barber, T. G. Schmidt, and X. Pan, “Estimating the spectrum in computed tomography via kullback-leibler divergence constrained optimization,” *arXiv preprint arXiv:1805.00162*, 2018.

[54] J. P. Schlomka, E. Roessl, R. Dorscheid, S. Dill, G. Martens, T. Istel, C. Bäumer, C. Herrmann, R. Steadman, G. Zeitler, A. Livne, and R. Proksa, “Experimental feasibility of multi-energy photon-counting K-edge imaging in pre-clinical computed tomography,” *Phys. Med. Biol.*, vol. 53, pp. 4031–4047, 2008.

[55] T. Schmidt, R. Barber, and E. Sidky, “A spectral CT method to directly estimate basis material maps from experimental photon-counting data,” *IEEE Trans. Med. Imaging*, vol. 36, pp. 1808–1819, 2017.

[56] L. Silberstein, “Determination of the spectral composition of X-ray radiation from filtration data,” *JOSA*, vol. 22, pp. 265–280, 1932.

[57] B. Perkhounkov, J. Stec, E. Y. Sidky, and X. Pan, “X-ray spectrum estimation from transmission measurements by an exponential of a polynomial model,” in *SPIE Med. Imaging*, vol. 9783, 2016, art. no. 97834W.

[58] W. Zhao, L. Xing, Q. Zhang, Q. Xie, and T. Niu, “Segmentation-free x-ray energy spectrum estimation for computed tomography using dual-energy material decomposition,” *J. Med. Imaging*, vol. 4, art. no. 023506, 2017.

[59] R. G. Waggener, M. M. Blough, J. A. Terry, D. Chen, N. E. Lee, S. Zhang, and W. D. McDavid, “X-ray spectra estimation using attenuation measurements from 25 kVp to 18 MV,” *Med. Phys.*, vol. 26, pp. 1269–1278, 1999.

[60] C. Ruth and P. M. Joseph, "Estimation of a photon energy spectrum for a computed tomography scanner," *Med. Phys.*, vol. 24, pp. 695–702, 1997.

[61] E. Y. Sidky, L. Yu, X. Pan, Y. Zou, and M. Vannier, "A robust method of x-ray source spectrum estimation from transmission measurements: Demonstrated on computer simulated, scatter-free transmission data," *J. Appl. Phys.*, vol. 97, art. no. 124701, 2005.

[62] P. Francois, A. Catala, and C. Scouarnec, "Simulation of x-ray spectral reconstruction from transmission data by direct resolution of the numeric system AF= T," *Med. Phys.*, vol. 20, pp. 1695–1703, 1993.

[63] M. Stampanoni, M. Fix, P. Francois, and P. Rüegsegger, "Computer algebra for x-ray spectral reconstruction between 6 and 25 MV," *Med. Phys.*, vol. 28, pp. 325–327, 2001.

[64] B. Armbruster, R. J. Hamilton, and A. K. Kuehl, "Spectrum reconstruction from dose measurements as a linear inverse problem," *Phys. Med. Biol.*, vol. 49, pp. 5087–5099, 2004.

[65] C. Leinweber, J. Maier, and M. Kachelrieß, "X-ray spectrum estimation for accurate attenuation simulation," *Med. Phys.*, vol. 44, pp. 6183–6194, 2017.

[66] X. Duan, J. Wang, L. Yu, S. Leng, and C. H. McCollough, "CT scanner x-ray spectrum estimation from transmission measurements," *Med. Phys.*, vol. 38, pp. 993–997, 2011.

[67] H. H. Barrett and K. J. Myers, *Foundations of image science*. John Wiley & Sons, 2013.

[68] S. F. Gull and G. J. Daniell, "Image reconstruction from incomplete and noisy data," *Nature*, vol. 272, pp. 686–690, 1978.

[69] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. 26, pp. 26–37, 1980.

[70] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *ICML*, 2013, pp. 280–288.

[71] M. J. Berger, J. H. Hubbell, S. Seltzer, J. Chang, J. Coursey, and R. Sukumar, "XCOM: Photon cross sections database," *NIST Standard Reference Database*, vol. 8, pp. 3587–3597, 1998.

[72] K. Cranley, B. Gilmore, G. Fogarty, and L. Desponds, "IPEM report 78: Catalogue of diagnostic x-ray spectra and other data," Inst. Phys. Eng. Med., York, U.K., Tech. Rep., 1997.

[73] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of mathematical imaging and vision*, vol. 40, pp. 120–145, 2011.

[74] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, pp. 1–122, 2011.

[75] H. Wang and A. Banerjee, "Bregman alternating direction method of multipliers," in *Advances in Neural Information Processing Systems*, 2014, pp. 2816–2824.