Low-rank Interaction with Sparse Additive Effects Model for Large Data Frames

Geneviève Robin, Hoi-To Wai, Julie Josse, Olga Klopp, Éric Moulines

Low-rank Interaction with Sparse Additive Effects Model for Large Data Frames

Geneviève Robin
Centre de Mathématiques Appliquées
École Polytechnique, XPOP, INRIA
91120 Palaiseau, France
genevieve.robin@polytechnique.edu

Hoi-To Wai
Department of SE&EM
The Chinese University of Hong Kong
Shatin, Hong Kong
htwai@se.cuhk.edu.hk

Julie Josse
Centre de Mathématiques Appliquées
École Polytechnique, XPOP, INRIA
91120 Palaiseau, France
julie.josse@polytechnique.edu

Olga Klopp
ESSEC Business School
CREST, ENSAE
95021 Cergy, France
klopp@essec.edu

Éric Moulines
Centre de Mathématiques Appliquées
École Polytechnique, XPOP, INRIA
91120 Palaiseau, France
eric.moulines@polytechnique.edu

Abstract

Many applications of machine learning involve the analysis of large data frames – matrices collecting heterogeneous measurements (binary, numerical, counts, etc.) across samples – with missing values. Low-rank models, as studied by Udell et al. [30], are popular in this framework for tasks such as visualization, clustering and missing value imputation. Yet, available methods with statistical guarantees and efficient optimization do not allow explicit modeling of main additive effects such as row and column, or covariate effects. In this paper, we introduce a *low-rank interaction and sparse additive effects* (LORIS) model which combines matrix regression on a dictionary and low-rank design, to estimate main effects and interactions simultaneously. We provide statistical guarantees in the form of upper bounds on the estimation error of both components. Then, we introduce a *mixed coordinate gradient descent* (MCGD) method which provably converges sub-linearly to an optimal solution and is computationally efficient for large scale data sets. We show on simulated and survey data that the method has a clear advantage over current practices, which consist in dealing separately with additive effects in a preprocessing step.

1 Introduction

Recently, a lot of effort has been devoted towards the efficient analysis of large data frames, a term coined by Udell et al. [30]. A data frame is a large table of heterogeneous data (binary, numerical, counts) with missing entries, where each row represents an example and each column a feature. In order to analyze them, a powerful technique is to use *low-rank models* that embed rows and columns

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

of data frames into low-dimensional spaces [18, 28, 30], enabling effective data analytics such as clustering, visualization and missing value imputation; see also [22] and the references therein.

Characterizing additive effects of side information – such as covariates, row or column effects – *simultaneously* with low rank interactions is an important extension to plain low-rank models. For example, in data frames obtained from recommender systems, user information and item characteristics are known to influence the ratings in addition to interactions between users and items [9]. These modifications to the low rank model have been advocated in the statistics literature, but they have been implemented only for small data frames [1].

In the large-scale low-rank matrix estimation literature, available methods either do not take additive effects into account [8, 24, 30, 26, 10], or only handle the numerical data [15, 14]. As a common heuristics for preprocessing, prior work such as [24, 30] remove the row and column means and apply some normalization of the row and column variance. We show in numerical experiments this apparently benign operation is not appropriate for large and heterogenous data frames, and can cause severe impairments in the analysis.

The present work investigates a generalization of previous contributions in the analysis of data frames. Our contributions can be summarized as follows.

Contributions We present a new framework that is *statistically* and *computationally* efficient for analyzing large and incomplete heterogeneous data frames.

- We describe in Section 2 the *low-rank interaction with sparse additive effects* (LORIS) model, which combines matrix regression on a dictionary with low rank approximation. We propose a convex doubly penalized quasi-maximum likelihood approach, where the rank constraint is relaxed with a nuclear norm penalty, to estimate the regression coefficients and the low rank component simultaneously. We establish non-asymptotic upper bounds on the estimation errors.
- We propose in Section 3 a Mixed Coordinate Gradient Descent (MCGD) method to solve efficiently the LORIS estimation problem. It uses a mixed update strategy including a proximal update for the sparse component and a conditional gradient (CG) for the low-rank component. We show that the MCGD method converges to an ϵ-optimal solution in $\mathcal{O}(1/\epsilon)$ iterations. We also outline an extension to efficient distributed implementation.
- We demonstrate in Section 4 the efficacy of our method both in terms of estimation and imputation quality on simulated and survey data examples.

Related work Our statistical model and analysis are related to prior work on *low-rank plus sparse matrix decomposition* [32, 5, 6, 16, 21]; these papers provide statistical results for a particular case where the loss function is quadratic and the sparse component is entry-wise sparse. In comparison, the originality of the present work is two-fold. First, the sparsity pattern of the main effects is not restricted to entry-wise sparsity. Second, the data fitting term is not quadratic, but a heterogeneous exponential family quasi log-likelihood. This new framework enables us to tackle many more data sets combining heterogeneous data, main effects and interactions.

For the algorithmic development, our proposed method is related to the prior work such as [25, 29, 7, 14, 33, 17, 27, 11, 23, 4, 12]. These are based on various first-order optimization methods and shall be reviewed in detail in Section 3. Among others, the MCGD method is mostly related to the recent FW-T method by Mu et al. [27] that uses a mixed update rule to tackle a similar estimation problem. There are two differences: first, FW-T is focused on a quadratic loss which is a special case of the statistical estimation problem that we analyze; second, the per-iteration complexity of MCGD is lower as the update rules are simpler. Despite the simplifications, using a new proof technique, we prove that the convergence rate of MCGD is strictly faster than FW-T.

Notations: For any $m \in \mathbb{N}$, $[m] := \{1, ..., m\}$. The operator $\mathcal{P}_\Omega(\cdot) : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ is the projection operator on the set of entries in $\Omega \subset [n] \times [p]$, and $(\cdot)_+ : \mathbb{R} \to \mathbb{R}_+$ is the projection operator on the non-negative orthant $(x)_+ := \max\{0, x\}$. For matrices, we denote by $\|\cdot\|_F$ the Frobenius norm, $\|\cdot\|_\star$ the nuclear norm, $\|\cdot\|$ the operator norm, and $\|\cdot\|_\infty$ the entry-wise infinity norm. For vectors, we denote by $\|\cdot\|_1$ is the ℓ_1-norm, $\|\cdot\|_2$ the Euclidean norm, $\|\cdot\|_\infty$ the infinity norm, and $\|\cdot\|_0$ the number of non zero coefficients. The binary operator $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle$ denotes the Frobenius

inner product. A function $f : \mathbb{R}^q \to \mathbb{R}$ is said to be σ-smooth if f is continuously differentiable and $\|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq \sigma \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^q$.

2 Problem Formulation

Heterogenous Data Model Let (Y, X) be a probability space equipped with a σ-finite measure μ. The canonical exponential family distribution $\{\mathrm{Exp}_{h,g}(m), m \in \mathsf{X}\}$ with base measure $h : \mathsf{Y} \to \mathbb{R}^+$, link function $g : \mathsf{X} \to \mathbb{R}$, and scalar parameter, $m \in \mathsf{X}$, has a density given by

$$f_m(y) = h(y) \exp\left(ym - g(m)\right) . \tag{1}$$

The exponential family is a flexible framework to model different types of data. For example, $(\mathsf{Y} = \mathbb{R}, g(m) = m^2\sigma^2/2, h(y) = (2\pi\sigma^2)^{-1/2} \exp(-y^2/2\sigma^2))$ yields a Gaussian distribution with mean m and variance σ^2 for numerical data; $(\mathsf{Y} = \{0, 1\}, g(m) = \log(1 + \exp(m)), h(y) = 1)$ yields a Bernoulli distribution with success probability $1/(1 + \exp(-m))$ for binary data; $(\mathsf{Y} = \mathbb{N}, g(m) = \exp(am), h(y) = 1/y!)$ where $a \in \mathbb{R}$ yields a Poisson distribution with intensity $\exp(am)$ for count data. In these cases, the parameter space is $\mathsf{X} = \mathbb{R}$.

Let $\{(\mathsf{Y}_j, g_j, h_j),\ j \in [p]\}$ be a collection of observation spaces, base and link functions corresponding to the column types of a data frame $\mathbf{Y} = [\mathbf{Y}_{ij}]_{(i,j)\in[n]\times[p]} \in \mathsf{Y}_1^n \times \ldots \times \mathsf{Y}_p^n$. For each $i \in [n]$ and $j \in [p]$, we denote by $\mathbf{M}^0_{ij}$ the target parameter minimizing the Kullback-Leibler divergence between the distribution of $\mathbf{Y}_{ij}$ and the exponential family Exp_{h_j,g_j}, $j \in [p]$, given by

$$\mathbf{M}^0_{ij} = \arg\max_m \ \mathbb{E}_{\mathbf{Y}_{ij}}[\log(h_j(\mathbf{Y}_{ij})) + \mathbf{Y}_{ij}m - g_j(m)] . \tag{2}$$

We propose the following model to estimate $\mathbf{M}^0 = [\mathbf{M}^0_{ij}]_{(i,j)\in[n]\times[p]}$ in the presence of additive effects and interactions.

LOw-rank Interaction with Sparse additive effects (LORIS) model For every entry $\mathbf{Y}_{ij}$, assume a vector of covariates $\mathbf{x}_{ij} \in \mathbb{R}^q$ is also available, e.g., user information and item characteristics. Denote $\mathbf{x}_{ij}(k)$, $k \in [q]$ the k-th component of $\mathbf{x}_{ij}$ and define the matrix $\mathbf{X}(k) = [\mathbf{x}_{ij}(k)]_{(i,j)\in[n]\times[p]}$. We introduce the following decomposition of the parameter matrix $\mathbf{M}^0$:

$$\mathbf{M}^0 = \sum_{k=1}^{q} \boldsymbol{\alpha}^0_k \mathbf{X}(k) + \boldsymbol{\Theta}^0. \tag{3}$$

We call (3) the LORIS model, where $\boldsymbol{\alpha} \in \mathbb{R}^q$ is a sparse vector with unknown support modeling additive effects and $\boldsymbol{\Theta}^0 \in \mathbb{R}^{n\times p}$ a low-rank matrix modeling the interactions.
In fact, LORIS is a generalization of *robust* matrix completion [5], where the parameter matrix can be decomposed as the sum of two matrices, one is low-rank and the other has some complementary low-dimensional structure such as entry-wise or column-wise sparsity. Statistical recoverability results in robust matrix estimation under a noiseless setting can be found in [32, 5, 6, 16]; the additive noise setting can be found in a recent work [21].

Estimation Problem Denote $\Omega = \{(i, j) \in [n] \times [p] : \mathbf{Y}_{ij} \text{ is observed}\}$ as the observation set. For $\mathbf{M} \in \mathbb{R}^{n\times p}$, $\mathcal{L}(\mathbf{M})$ is the negative log-likelihood of the observed data $(\mathbf{Y}, \Omega)$ parameterized by $\mathbf{M}$. Up to an additive constant,

$$\mathcal{L}(\mathbf{M}) = \sum_{(i,j)\in\Omega} \left\{-\mathbf{Y}_{ij}\mathbf{M}_{ij} + g_j(\mathbf{M}_{ij})\right\} . \tag{4}$$

For $a > 0$, we consider the following estimation problem:

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}) \in \underset{\substack{\|\boldsymbol{\alpha}\|_\infty \leq a \\ \|\boldsymbol{\Theta}\|_\infty \leq a}}{\operatorname{argmin}} \ \mathcal{L}\left(\sum_{k=1}^{q} \boldsymbol{\alpha}_k \mathbf{X}(k) + \boldsymbol{\Theta}\right) + \lambda_S \|\boldsymbol{\alpha}\|_1 + \lambda_L \|\boldsymbol{\Theta}\|_\star . \tag{5}$$

We denote by $\hat{\mathbf{M}} = \sum_{k=1}^{q} \hat{\boldsymbol{\alpha}}_k \mathbf{X}(k) + \hat{\boldsymbol{\Theta}}$ the estimated parameter matrix. The ℓ_1 and nuclear norm penalties are convex relaxations of the sparsity and low-rank constraints, and the regularization parameters λ_S and λ_L serve as trade-offs between fitting the data and enforcing sparsity of $\boldsymbol{\alpha}$ and controlling the "effective rank" of $\boldsymbol{\Theta}$.

Statistical Guarantees Here we establish convergence rates for the joint estimation of $\boldsymbol{\alpha}^0$ and $\boldsymbol{\Theta}^0$; the proofs can be found in the supplementary material. Consider the following assumptions.

H1 $\left\|\boldsymbol{\Theta}^0\right\|_\infty \leq a$, $\left\|\boldsymbol{\alpha}^0\right\|_\infty \leq a$ *and for all* $k \in [q]$ *such that* $\boldsymbol{\alpha}_k^0 \neq 0$, $\langle \boldsymbol{\Theta}^0, \mathbf{X}(k) \rangle = 0$.

In particular, H1 guarantees the uniqueness of the decomposition in the LORIS model (3).

H2 *For* $\nu > 0$, *all* $k \in [q]$ *and* $(i,j) \in [n] \times [p]$, $\mathbf{X}(k)_{ij} \in [-1,1]$. *Furthermore for all* $(i,j) \in [n] \times [p]$, $\sum_{k=1}^q |\mathbf{X}(k)_{ij}| \leq \nu$.

In particular, H2 guarantees that for all $(\boldsymbol{\Theta}, \boldsymbol{\alpha})$ satisfying H1, the matrix $\mathbf{M} = \sum_{k=1}^q \boldsymbol{\alpha}_k \mathbf{X}(k) + \boldsymbol{\Theta}$ satisfies $\|\mathbf{M}\|_\infty \leq (1+\nu)a$. Let $\mathbf{G}$ be the $q \times q$ Gram matrix of the dictionary $(\mathbf{X}(1), \ldots, \mathbf{X}(q))$ defined by $\mathbf{G} = [\langle \mathbf{X}(k), \mathbf{X}(l) \rangle]_{(k,l) \in [q] \times [q]}$.

H3 *For* $\kappa > 0$ *and all* $\boldsymbol{\alpha} \in \mathbb{R}^q$, $\boldsymbol{\alpha}^\top \mathbf{G} \boldsymbol{\alpha} \geq \kappa^2 \|\boldsymbol{\alpha}\|_2^2$.

Note we do not consider the case where the Gram matrix is singular, e.g., $q > np$. For $0 < \sigma_- \leq \sigma_+ < +\infty$ and $0 < \gamma < \infty$ consider the following assumption on the link functions g_j:

H4 *The functions* g_j *are twice differentiable, and for all* $x \in [-(1+\nu)a - \gamma, (1+\nu)a + \gamma]$,

$$\sigma_-^2 \leq g_j''(x) \leq \sigma_+^2, \; j \in [p].$$

H4 implies the data fitting term $\mathcal{L}(\mathbf{M})$ is smooth and satisfies a restricted strong convexity property.

H5 *For all* $(i,j) \in [n] \times [p]$, Y_{ij} *is a sub-exponential random variable with scale and variance parameters* $1/\gamma$ *and* σ_+^2.

If the random variables Y_{ij} are actually distributed according to an exponential family distribution of the form (1), then H4 implies H5.

H6 *For* $(i,j) \in [n] \times [p]$, *the events* $\omega_{ij} = \{(i,j) \in \Omega\}$ *are independent with occurrence probability* π_{ij}. *Furthermore, there exists* $0 < \pi \leq 1$ *such that for all* $(i,j) \in [n] \times [p]$, $\pi_{ij} \geq \pi$.

H6 implies a data missing-at-random scenario where $\mathbf{Y}_{ij}$ is observed with probability at least π.

Theorem 1 *Assume H1-6. Set*

$$\lambda_L = 2C\sigma_+ \sqrt{\pi \max(n,p) \log(n+p)}, \text{ and } \lambda_S = 24 \max_k \|\mathbf{X}(k)\|_1 \log(n+p)/\gamma, \tag{6}$$

where C *is a positive constant. Assume that* $\max(n,p) \geq 4\sigma_+^2/\gamma^6 \log^2(\sqrt{\min(n,p)/(\pi\gamma\sigma_-)}) + 2\exp(\sigma_+^2/\gamma^2 + 2\sigma_+^2\gamma a)$. *Then, with probability at least* $1 - 9(n+p)^{-1}$,

$$\begin{aligned} \left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\right\|_2^2 &\leq C_1 \frac{s \max_k \|\mathbf{X}(k)\|_1 \log(n+p)}{\kappa^2 \pi} + \mathsf{D}_{\boldsymbol{\alpha}}, \\ \left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^0\right\|_F^2 &\leq C_2 \left(\frac{r \max(n,p)}{\pi} + \frac{s \max_k \|\mathbf{X}(k)\|_1}{\pi} \right) \log(n+p) + \mathsf{D}_{\boldsymbol{\Theta}}. \end{aligned} \tag{7}$$

In (23), $s := \|\boldsymbol{\alpha}^0\|_0$, $r := \operatorname{rank}(\boldsymbol{\Theta}^0)$. C_1 *and* C_2 *are positive constants and* $\mathsf{D}_{\boldsymbol{\alpha}}$ *and* $\mathsf{D}_{\boldsymbol{\Theta}}$ *are residuals of lower order whose exact values are given in Appendix A.*

The proof can be found in Appendix A. In Theorem 1, the rate obtained for $\boldsymbol{\alpha}^0$ is the same as the bound obtained in [21] in the special case of robust matrix completion. Examples satisfying $\max_k \|\mathbf{X}(k)\|_1 / \kappa^2 = \mathcal{O}(1)$ include the case where the elements of the dictionary are matrices are all zeros except a row or a column of one, (to model row and column effects) and the number of rows n and columns p are of the same order; or when the covariates $\mathbf{x}_{ij}$ are categorical and the categories are balanced, *i.e.,* the number of samples per category is of the same order.

The rate obtained for $\boldsymbol{\Theta}^0$ is the sum of the standard low-rank matrix completion rate of order $r \max(n,p)/\pi$, e.g., [19], and of a term which boils down to sparse vector estimation rate as long as $\max_k \|\mathbf{X}(k)\|_1 = \mathcal{O}(1)$. Again, the latter can be satisfied by the special case of robust matrix completion, for which our rates match the results of [21].

3 A Mixed Coordinate Gradient Descent Method for LORIS

This section introduces a mixed coordinate gradient descent (MCGD) method to solve the LORIS estimation problem (5). We assume that a is sufficiently large such that the constraints $\|\boldsymbol{\alpha}\|_\infty \leq a, \|\boldsymbol{\Theta}\|_\infty \leq a$ are always inactive. To simplify notation, we denote the log-likelihood function as $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta}) := \mathcal{L}\left(\sum_{k=1}^q \boldsymbol{\alpha}_k \mathbf{X}(k) + \boldsymbol{\Theta}\right)$. We assume

H7 *(a) $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is $\sigma_{\boldsymbol{\Theta}}$-smooth* w.r.t. Θ_{ij} *for $(i,j) \in \Omega$ and (b) $\sigma_{\boldsymbol{\alpha}}$-smooth* w.r.t. $\boldsymbol{\alpha}$*; (c) the gradient $\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is $\hat{\sigma}_{\boldsymbol{\Theta}}$-Lipschitz* w.r.t. $\boldsymbol{\Theta}$*. Moreover, the gradient $\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is bounded as long as $\boldsymbol{\alpha}, \boldsymbol{\Theta}$ are bounded.*

The above is implied by H4 for bounded $(\boldsymbol{\alpha}, \boldsymbol{\Theta})$. We consider the augmented objective function:

$$F(\boldsymbol{\alpha}, \boldsymbol{\Theta}, R) := \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta}) + \lambda_S \|\boldsymbol{\alpha}\|_1 + \lambda_L R \,. \tag{8}$$

For some $R_{\mathsf{UB}} \geq 0$, if an optimal solution $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}})$ to (5) satisfies $\|\hat{\boldsymbol{\Theta}}\|_\star \leq R_{\mathsf{UB}}$, then any optimal solution to the following problem

$$\mathsf{P}(R_{\mathsf{UB}}): \quad \min_{\boldsymbol{\alpha}\in\mathbb{R}^q, \boldsymbol{\Theta}\in\mathbb{R}^{n\times p}, R\in\mathbb{R}_+} F(\boldsymbol{\alpha}, \boldsymbol{\Theta}, R) \;\text{ s.t. }\; R_{\mathsf{UB}} \geq R \geq \|\boldsymbol{\Theta}\|_\star \,, \tag{9}$$

will also be optimal to (5). For example, $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R})$ with $\hat{R} = \|\hat{\boldsymbol{\Theta}}\|_\star$ is an optimal solution to (9). We have defined the problem as $\mathsf{P}(R_{\mathsf{UB}})$ to emphasize its dependence on the upper bound R_{UB}. Later we shall describe a simple strategy to estimate R_{UB}. We fix the set $\Xi \subseteq [n] \times [p]$ where $\Omega \subseteq \Xi$ is the target coordinate set for the low rank matrix $\hat{\boldsymbol{\Theta}}$ that we are interested in.

Proposed Method A natural way to exploit structure in $\mathsf{P}(R_{\mathsf{UB}})$ is to apply coordinate gradient descent to update $\boldsymbol{\alpha}$ and $(\boldsymbol{\Theta}, R)$ separately. While the trace-norm constraint on $(\boldsymbol{\Theta}, R)$ can be handled by the conditional gradient (CG) method [17], the ℓ_1 norm penalization on $\boldsymbol{\alpha}$ is more efficiently tackled by the proximal gradient method in practice. In addition, we tighten the upper bound R_{UB} on-the-fly as the algorithm proceeds. The MCGD method goes as follows. At the tth iteration, we are given the previous iterate $(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$ and the upper bound $R_{\mathsf{UB}}^{(t)}$ is computed. The first block $\boldsymbol{\alpha}$ is updated with a proximal gradient step:

$$\begin{aligned} \boldsymbol{\alpha}^{(t)} &= \mathrm{prox}_{\gamma\lambda_S\|\cdot\|_1}\left(\boldsymbol{\alpha}^{(t-1)} - \gamma\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)})\right) \\ &= \mathsf{T}_{\gamma\lambda_S}\left(\boldsymbol{\alpha}^{(t-1)} - \gamma\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha}^{(t-1)}, \boldsymbol{\Theta}^{(t-1)})\right) . \end{aligned} \tag{10}$$

In (10), $\nabla_{\boldsymbol{\alpha}}\mathcal{L}(\cdot)$ is the gradient of the log-likelihood function taken *w.r.t.* $\boldsymbol{\alpha}$, $\gamma > 0$ is a pre-defined step size parameter and $\mathsf{T}_\lambda(\boldsymbol{x}) := \mathrm{sign}(\boldsymbol{x}) \odot (\boldsymbol{x} - \lambda\mathbf{1})_+$ is the component-wise soft thresholding operator. Alternatively, we can exactly solve the problem

$$\boldsymbol{\alpha}^{(t)} \in \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^q} \; F(\boldsymbol{\alpha}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)}) \,, \tag{11}$$

for which closed-form solution can be obtained in certain special cases (see below).

The second block $(\boldsymbol{\Theta}, R)$ is updated with a CG step

$$(\boldsymbol{\Theta}^{(t)}, R^{(t)}) = (\boldsymbol{\Theta}^{(t-1)}, R^{(t-1)}) + \beta_t(\hat{\boldsymbol{\Theta}}^{(t)} - \boldsymbol{\Theta}^{(t-1)}, \hat{R}^{(t)} - R^{(t-1)}) \,, \tag{12}$$

where $\beta_t \in [0,1]$ is a step size to be defined later. $(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{R}^{(t)})$ is a direction evaluated as

$$(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{R}^{(t)}) \in \arg\min_{\boldsymbol{Z}, R} \; \langle \boldsymbol{Z}, \nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})\rangle + \lambda_1 R \;\text{ s.t. }\; \|\boldsymbol{Z}\|_\star \leq R \leq R_{\mathsf{UB}}^{(t)} \,, \tag{13}$$

and $\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\cdot)$ is the gradient of $\mathcal{L}(\cdot)$ taken *w.r.t.* $\boldsymbol{\Theta}$. If $(\boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$ is feasible to $\mathsf{P}(R_{\mathsf{UB}}^{(t)})$, then $(\boldsymbol{\Theta}^{(t)}, R^{(t)})$ must also be feasible to $\mathsf{P}(R_{\mathsf{UB}}^{(t)})$. Furthermore, if we let $\boldsymbol{u}_1, \boldsymbol{v}_1$ be the top left and right singular vectors of the gradient matrix $\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})$ and $\sigma_1(\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)}))$ be the top singular value, then $(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{R}^{(t)})$ admits a simple closed form solution:

$$(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{R}^{(t)}) = \begin{cases} (\mathbf{0}, 0), & \text{if } \lambda_L \geq \sigma_1(\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})) \,, \\ (-R_{\mathsf{UB}}^{(t)}\boldsymbol{u}_1\boldsymbol{v}_1^\top, R_{\mathsf{UB}}^{(t)}), & \text{if } \lambda_L < \sigma_1(\nabla_{\boldsymbol{\Theta}}\mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})) \,. \end{cases} \tag{14}$$

Lastly, the step size β_t is determined by:

$$\beta_t = \min\left\{1, \frac{\langle \boldsymbol{\Theta}^{(t-1)} - \hat{\boldsymbol{\Theta}}^{(t)}, \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})\rangle + \lambda_L (R^{(t-1)} - \hat{R}^{(t)})}{\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^{(t)} - \boldsymbol{\Theta}^{(t-1)})\|_{\mathrm{F}}^2}\right\}. \tag{15}$$

The step size strategy ensures decrease in the objective value between successive iterations. This is essential for establishing convergence of the proposed method [cf. Theorem 2]. We remark that the arithmetics in the MCGD method are not affected when we restrict the update of $\boldsymbol{\Theta}^{(t)}$ in (12) to the entries in Ξ only. This is due to $\mathcal{L}(\boldsymbol{X}) = \mathcal{L}(\mathcal{P}_\Omega(\boldsymbol{X}))$ and the CG update direction (13) only involves the gradient of $\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})$ *w.r.t.* entries of $\boldsymbol{\Theta}$ in Ω, where $\Omega \subseteq \Xi$.

Computing the Upper Bound $R_{\mathsf{UB}}^{(t)}$ We describe a strategy for computing a valid upper bound $R_{\mathsf{UB}}^{(t)}$ for $\hat{R}$ and $\|\hat{\boldsymbol{\Theta}}\|_\star$ during the updates in the MCGD method. Let us assume that:

H8 *For all $\boldsymbol{\Theta}$ and $\boldsymbol{\alpha}$, we have $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta}) \geq 0$.*

The above can be enforced as the log-likelihood function is lower bounded [cf. H4]. From (5) and using the above assumption, it is obvious that

$$F_0(\mathbf{0}, \mathbf{0}) = \mathcal{L}(\mathbf{0}, \mathbf{0}) \geq \mathcal{L}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}) + \lambda_S \|\hat{\boldsymbol{\alpha}}\|_1 + \lambda_L \|\hat{\boldsymbol{\Theta}}\|_\star \geq \lambda_L \|\hat{\boldsymbol{\Theta}}\|_\star, \tag{16}$$

and thus $R_{\mathsf{UB}}^0 := \lambda_L^{-1} \mathcal{L}(\mathbf{0} + f_{\boldsymbol{U}}(\mathbf{0}))$ is a valid upper bound to $\|\hat{\boldsymbol{\Theta}}\|_\star$; furthermore it can be tightened as we progress in the MCGD method. In particular, observe that $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R})$ with $\hat{R} = \|\hat{\boldsymbol{\Theta}}\|_\star$ is an optimal solution to $\mathsf{P}(R_{\mathsf{UB}}^0)$, we have

$$F(\boldsymbol{\alpha}, \boldsymbol{\Theta}, R) \geq F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R}) = \mathcal{L}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}) + \lambda_S \|\hat{\boldsymbol{\alpha}}\|_1 + \lambda_L \hat{R} \geq \lambda_L \hat{R}. \tag{17}$$

In other words, for all feasible $(\boldsymbol{\alpha}, \boldsymbol{\Theta}, R)$ to $\mathsf{P}(R_{\mathsf{UB}}^0)$, $\lambda_L^{-1} F(\boldsymbol{\alpha}, \boldsymbol{\Theta}, R)$ is an upper bound to $\hat{R}$ and $\|\hat{\boldsymbol{\Theta}}\|_\star$. The above motivates us to select $R_{\mathsf{UB}}^{(t)} := \lambda_L^{-1} F(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$ at iteration t, where we observe that $R_{\mathsf{UB}}^{(t)} \geq R^{(t-1)}$. That is, $(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$ is feasible to both $\mathsf{P}(R_{\mathsf{UB}}^{(t)})$ and $\mathsf{P}(R_{\mathsf{UB}}^{(t-1)})$. Lastly, we summarize the MCGD method in Algorithm 1.

Computation Complexity Consider the MCGD method in Algorithm 1. Observe that line 3 requires computing the gradient *w.r.t.* $\boldsymbol{\alpha}$ which involves $|\Omega|q$ Floating Points Operations (FLOPS) and the soft thresholding operator involves $\mathcal{O}(q)$ FLOPS. As the log-likelihood function $\mathcal{L}(\cdot)$ is evaluated element-wisely on $\boldsymbol{\Theta}$, evaluating the objective value and the derivative *w.r.t.* $\boldsymbol{\Theta}$ requires $\mathcal{O}(|\Omega|)$ FLOPS. As such, line 4 can be evaluated in $\mathcal{O}(|\Omega|)$ FLOPS and line 5 requires $\mathcal{O}(|\Omega| \max\{n, p\} \log(1/\delta))$ FLOPS where the additional complexity is due to the top SVD computation and δ is a preset accuracy level of SVD computation. Lastly, line 6 requires $\mathcal{O}(|\Xi|)$ FLOPS since we only need to update the entries of $\boldsymbol{\Theta}$ in Ξ [cf. see the remark after (15)]. The overall per-iteration complexity is $\mathcal{O}(|\Xi| + |\Omega|(\max\{n, p\} \log(1/\delta) + q))$.

Algorithm 1 MCGD Method for (9).

1: **Initialize:** — $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\alpha}^{(0)}, R^{(0)}$. E.g., $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\alpha}^{(0)}, R^{(0)} = (\mathbf{0}, \mathbf{0}, 0)$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: *// Update for $\boldsymbol{\alpha}$ //*
Compute the proximal update using (10) [or exact update via (11)] to obtain $\boldsymbol{\alpha}^{(t)}$.
4: *// Update for $(\boldsymbol{\Theta}, R)$ //*
Compute the upper bound as $R_{\mathsf{UB}}^{(t)} := \lambda_L^{-1} F(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$.
5: Compute the update direction, $(\hat{\boldsymbol{\Theta}}^{(t)}, \hat{R}^{(t)})$, using Eq. (14).
6: Compute the CG update using (12), where the step size β_t is set as Eq. (15).
7: **end for**
8: **Return:** $\boldsymbol{\Theta}^{(T)}, \boldsymbol{\alpha}^{(T)}, R^{(T)}$.

From the above, the per-iteration computation complexity of the MCGD method scales linearly with the problem dimension $\max\{n, p\}$ and $|\Omega|$. This is comparable to [27, 11], where the former focuses only on the least square loss case. The following theorem, whose proof can be found in Appendix C, shows that the MCGD method converges at a sublinear rate.

Theorem 2 *Assume H7 and H8. Define the quantity*

$$C(t) := \max\left\{\frac{24(Q^{(t)})^2}{\gamma}, \frac{24\hat{\sigma}_{\boldsymbol{\Theta}}^2 (Q^{(t)})^2}{\sigma_{\boldsymbol{\Theta}}} + \max\{6R_{\mathsf{UB}}^{(t)}(\lambda_L + M^{(t)}), 24\sigma_{\boldsymbol{\Theta}}(R_{\mathsf{UB}}^{(t)})^2\}\right\}, \tag{18}$$

where $Q^{(t)} := \lambda_S^{-1} F(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t)}, R^{(t)})$, $M^{(t)} := \|\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)})\|_2$ *and* $R_{\mathsf{UB}}^{(t)} := \lambda_L^{-1} F(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t-1)}, R^{(t-1)})$. *If we choose the step sizes as* $\gamma \leq 1/\sigma_{\boldsymbol{\alpha}}$ *and* β_t *as in* (15)*, then (i) the above quantity is upper bounded as* $C(t) \leq \overline{C}$ *for all* $t \geq 1$*, where*

$$\overline{C} := \max\Big\{ \frac{24(Q^{(0)})^2}{\gamma}, \frac{24\hat{\sigma}_{\boldsymbol{\Theta}}^2 (Q^{(0)})^2}{\sigma_{\boldsymbol{\Theta}}} + \max\{6R_{\mathsf{UB}}^{(0)}(\lambda_L + \bar{M}), 24\sigma_{\boldsymbol{\Theta}}(R_{\mathsf{UB}}^{(0)})^2\} \Big\}, \tag{19}$$

such that $\bar{M}$ *is an upper bound to* $M^{(t)}$*, and (ii) the MCGD method converges to an* ϵ*-optimal solution to* (5) *in* T *iterations,* i.e., $F_0(\boldsymbol{\alpha}^{(T)}, \boldsymbol{\Theta}^{(T)}) - F_0(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}) \leq \epsilon$*, where*

$$T \geq \overline{C}(T)\Big(\frac{1}{\epsilon} - \frac{1}{F_0(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\Theta}^{(0)}) - F_0(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}})}\Big)_+ \quad \textit{with} \quad \overline{C}(T) := \Big(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{C(t)}\Big)^{-1}. \tag{20}$$

In particular, as $\overline{C}(T) \leq \overline{C}$*, at most* $\overline{C}(\epsilon^{-1} - (F_0(\boldsymbol{\alpha}^{(0)}, \boldsymbol{\Theta}^{(0)}) - F_0(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}))^{-1})_+$ *iterations are required for the MCGD method to reach an* ϵ*-optimal solution to* (5)*.*

Detailed Comparison to Prior Algorithms Previous contributions have focused on the special case of (5) where $q = np$, the dictionary $(\mathbf{X}(1), \ldots, \mathbf{X}(q))$ is the canonical basis of $\mathbb{R}^{n \times p}$, and the link functions are quadratic. In this particular case, (5) becomes the estimation problem solved in sparse plus low-rank matrix decomposition. Popular examples are the alternating direction method of multiplier [25, 29] or the projected gradient method on a reformulated problem [7]. These methods either require computing a complete SVD or knowing the optimal rank number of $\boldsymbol{\Theta}$ a priori. When $n, p \gg 1$, it is computationally prohibitive to evaluate the complete SVD since each iteration would require $\mathcal{O}(\max\{n^2 p, p^2 n\})$ FLOPS. Other related work rely on factorizing the low-rank component, yielding nonconvex problems [14]; see also [33] and references therein.

Similar to the development of MCGD, a natural alternative is to apply algorithms based on the CG (a.k.a. Frank-Wolfe) method [17], whose iterations only require the computation of a top SVD. The present work is closely related to the efforts in [27, 11] which focused on the quadratic setting. Mu et al. [27] combines the CG method with proximal update as a two-steps procedure; Garber et al. [11] combines a CD method with CG updates on both the sparse and low-rank components. The work in [11] is also related to [23, 4] which combine CD with CG updates for solving constrained problems, instead of penalized problems like (5). Sublinear convergence rates are proven for the above methods. Finally, Fithian and Mazumder [10] also suggested to apply CD on (5), yet the convergence properties were not discussed.

In fact, when the MCGD's result is specialized to the same setting as [27], our worst-case bound on iteration number computed with $\overline{C}$ match the bound in [27]. As shown in the supplementary material, we have $C(t) \to C^\star$, where $C^\star$ depends on the optimal objective value of (9) and is smaller than $\overline{C}$. Since the quantity $\overline{C}(T)$ in (20) is an average of $\{C(t)\}_{t=1}^T$, this implies that the MCGD method requires less number of iterations for convergence than that is required by [27]. Such reduction is possible due to the on-the-fly update for $R_{\mathsf{UB}}^{(t)}$. Moreover, our analysis in Theorem 2 holds when the MCGD method is implemented with a few practical modifications.

Exact Partial Minimization for $\boldsymbol{\alpha}$ Consider the special case of (5) where the link functions are either quadratic or exponential and the dictionary matrices satisfy:

$$\operatorname{supp}(\boldsymbol{X}(k)) \cap \operatorname{supp}(\boldsymbol{X}(k')) = \emptyset, \ k \neq k' \ \text{ and } \ [\boldsymbol{X}(k)]_{i,j} = c_k, \ \forall\, (i,j) \in \operatorname{supp}(\boldsymbol{X}(k)). \tag{21}$$

In this case, the partial minimization (11) can be decoupled into q scalar optimizations involving one coordinate of $\boldsymbol{\alpha}$, which can be solved in closed form. Note that this modification to the MCGD method is supported by Theorem 2 and the sublinear convergence rate holds. On the contrary, closed form update of $\boldsymbol{\alpha}$ is not supported by prior works such as [27, 11, 23, 4].

Distributed MCGD Optimization Consider the case where the observed data entries are stored across K workers, each of them communicating with a central server. It is natural to distribute the MCGD optimization over these workers to offload computation burden, or for privacy protection. Formally, we divide Ω into K disjoint partitions such that $\Omega = \Omega_1 \cup \cdots \cup \Omega_K$ and worker k holds Ω_k. In this way, $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta}) = \sum_{k=1}^K \mathcal{L}_k(\boldsymbol{\alpha}, \boldsymbol{\Theta})$, where $\mathcal{L}_k(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is defined by replacing the summation over Ω with Ω_k in (4). Clearly, when $\boldsymbol{\alpha}$ and $\mathcal{P}_{\Omega_k}(\boldsymbol{\Theta})$ are given to the kth worker, the worker will be

able to evaluate the *local* loss function and its gradient.
As shown in Appendix D, the MCGD method can be easily extended to utilize distributed computation. The proximal update in line 3 is replaced by the following procedure. First, the *local* gradients computed by the workers are aggregated, then the soft thresholding operation is performed at the central server. Meanwhile, as the CG update in line 5 essentially requires computing the top singular vectors of the gradient matrix $\nabla_{\boldsymbol{\Theta}} L(\boldsymbol{\alpha}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \nabla_{\boldsymbol{\Theta}} \mathcal{L}_k(\boldsymbol{\alpha}, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}))$, the latter can be implemented through a distributed version of the power method exploiting the decomposable structure of the gradient, such as described in [34]. It only requires $\mathcal{O}(\log(1/\delta))$ power iterations to compute a top SVD solution of accuracy δ. Thus, for a sufficiently small $\delta > 0$, the overall per-iteration complexity of the distributed method at the tth iteration is reduced to $\mathcal{O}(|\Xi| + \max\{n, p\}\log(1/\delta))$ at the central server, and $\mathcal{O}(|\Omega_k|(\max\{n, p\}\log(1/\delta) + q))$ at the kth worker.

4 Numerical Experiments

Experimental Setup We first generate the target parameter $\mathbf{M}^0$ according to the LORIS model in (3). For the sparse additive effects component, we consider $q = pn/5$ where we set $(\mathbf{X}(k))_{ij} = 1$ if $j(n-1) + i \in \{5(k-1)+1, ..., 5k\}$. This models a categorical variable containing $n/5$ categories. Furthermore, the target sparse component $\boldsymbol{\alpha}^0$ has a sparsity level of 10%. For the low-rank component, the target parameter $\boldsymbol{\Theta}^0$ is generated as a rank-4 matrix formed by the outer product of random orthogonal vectors. Notice that due to the structure of sparse additive effects, the surveyed prior methods [25, 14, 7] cannot be applied directly.

Gaussian Design To compare our framework to a reasonable benchmark, we focus on a homogenous setting with numerical data modeled with the quadratic link function $g(m) = m^2$. We set the regularization parameters λ_S and λ_L to the theoretical values given in Theorem 1. We compare our result with a common two-step procedure where the components $\boldsymbol{\alpha}_{kj}$ are first estimated in a preprocessing step as the means of the variables taken by group; then $\boldsymbol{\Theta}$ is estimated using the softImpute method proposed in [15]. The regularization parameter for [15] is set to the same value λ_L. We compare the results in terms of estimation error and computing time in Table 1, after letting the two methods converge to the same precision of 10^{-5}. We observe the two methods perform equally well in terms of estimating $\boldsymbol{\Theta}$. LORIS yields constant estimation errors of $\boldsymbol{\alpha}^0$ as the dimension increases and the support of $\boldsymbol{\alpha}^0$ is kept constant, contrary to the two-step procedure for which the estimation error of $\boldsymbol{\alpha}^0$ increases with the dimension. As expected, the two-step method is faster for small data sets, whereas for large data sizes LORIS is superior in computational time. The above results are consistent with our theoretical findings.

problem size ($n \times p$)	**time** (secs)		$\left\|\boldsymbol{\Theta}^0 - \hat{\boldsymbol{\Theta}}\right\|_F^2$		$\left\|\boldsymbol{\alpha}^0 - \hat{\boldsymbol{\alpha}}\right\|_2^2$	
	LORIS	two-step	LORIS	two-step	LORIS	two-step
150×30	0.17	0.02	52	52	1.8	3.0
$1,500 \times 300$	13.8	10.7	175.5	234	0.95	17.1
$15,000 \times 300$	130.2	136.6	675	720	0.95	16.2
$15,000 \times 3,000$	348	528	2.7×10^3	2.6×10^3	2.34	180

Table 1: Comparison of proposed method with a two-step method in terms of computation time and estimation error for increasing dimensions (averaged over 10 experiments).

Survey data To test the efficacy of our framework with heterogeneous data, we examine a survey conducted by the French National Institute of Statistics (Insee: `http://www.insee.fr/`) concerning the hobbies of French people. The data set contains $n = 8,403$ individuals and $p = 19$ binary and quantitative variables, indicating whether or not the person has been involved in different activities (reading, fishing, etc.), the number of hours spent watching TV and the overall number of hobbies of the individuals. Individuals are grouped by age category ($15-25$, $25-35$, etc.): this categorical variable is used as a predictor of the survey responses in the subsequent experiment. We introduce 30% of missing values in the data set, and compare the imputation error of LORIS with a mixed data model (using a quadratic loss for numeric columns, a logistic loss for binary columns and a Poisson loss for counts) and LORIS with a Gaussian data model, with the imputation error of softImpute. The

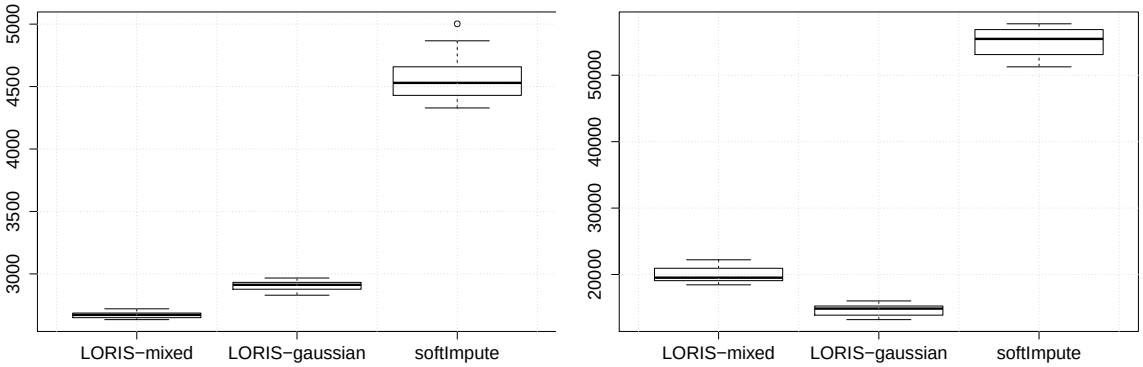

Figure 2: Imputation error of LORIS with mixed data model and Gaussian data model, and softImpute (10 replications) for categorical variables (left) and quantitative variables (right).

results are given in Figure 2 across 10 replications of the experiment, and show that, for this example, both LORIS models improve on the baseline softImpute by a factor 2. We also observe that modeling explicitly the binary variables leads to better imputation.

Finally, we apply LORIS with a mixed data model to the original data set. A subset of the resulting $\boldsymbol{\alpha}$ vector is given in Table 2. There is a coefficient in $\boldsymbol{\alpha}_{kj}$ for every age category k and every variable j. The coefficients in Table 2 indicate that young individuals engage in activities such as music and sport more than older people, and the opposite trend for collecting, knitting and fishing. Some coefficients are set to zero, indicating the absence of effect of the age category on the variable. We also observe that younger people engage overall in more activities than older people.

Age category	Music	Sport	Collecting	Mechanic	Knitting	Fishing	Nb activities
25-35	2.2	0.4	-2.1	0	-1.7	-1.9	10.0
35-45	2.0	0.3	-2.7	0	-2.3	-2.3	13.0
45-55	1.1	-0.8	-2.1	0	-2.7	-2.7	13.8
55-65	0	-2.2	-1.9	0	-1.0	-1.6	8.8
65-75	0	-2.1	-1.4	-1.1	-0.7	-1.3	5.5
75-85	-0.1	-0.9	-0.6	-0.5	-0.1	-0.6	2.2

Table 2: Estimated age category effects ($\boldsymbol{\alpha}$).

Conclusion In this paper, we proposed a new framework for handling large data frames with heterogeneous data and missing values which incorporates additive effects. It consists of a doubly penalized quasi-maximum likelihood estimator and a new optimization algorithm to implement the estimator. We examined both the statistical and computational efficiency of the framework and derived worst case bounds of its performance. Future work includes the incorporation of qualitative features with more than two categories and of missing values in the dictionary matrices.

5 Acknowledgement

The authors would like to thank for the useful comments from three anonymous reviewers. HTW's work was supported by the grant NSF CCF-BSF 1714672.

References

[1] A. Agresti. *Categorical Data Analysis, 3rd Edition*. Wiley, 2013.

[2] J.-P. Aubin and I. Ekeland. *Applied nonlinear analysis*. Pure and applied mathematics. John Wiley, New-York, 1984. A Wiley-Interscience publication.

[3] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

[4] A. Beck, E. Pauwels, and S. Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM Journal on Optimization*, 25(4):2024–2049, 2015.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL http://doi.acm.org/10.1145/1970392.1970395.

[6] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. doi: 10.1137/090761793. URL https://doi.org/10.1137/090761793.

[7] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *CoRR*, abs/1509.03025, 2015.

[8] J. de Leeuw. Principal component analysis of binary data by iterated singular value decomposition. *Comput. Stat. Data Anal.*, 50(1):21–39, Jan. 2006. ISSN 0167-9473. doi: 10.1016/j.csda.2004.07.010. URL http://dx.doi.org/10.1016/j.csda.2004.07.010.

[9] A. Feuerverger, Y. He, and S. Khatri. Statistical significance of the netflix challenge. *Statist. Sci.*, 27(2): 202–231, 05 2012. doi: 10.1214/11-STS368. URL http://dx.doi.org/10.1214/11-STS368.

[10] W. Fithian and R. Mazumder. Flexible Low-Rank Statistical Modeling with Missing Data and Side Information. *Statistical Science*, 33(2):238–260, 2018.

[11] D. Garber, S. Sabach, and A. Kaplan. Fast generalized conditional gradient method with applications to matrix recovery problems. *arXiv preprint arXiv:1802.05581*, 2018.

[12] G. Gidel, F. Pedregosa, and S. Lacoste-Julien. Frank-wolfe splitting via augmented lagrangian method. In *OPTML 2017: 10th NIPS Workshop on Optimization for Machine Learning (NIPS 2017)*, page 21, 2017. URL http://opt-ml.org/papers/OPT2017_paper_21.pdf.

[13] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[14] Q. Gu, Z. W. Wang, and H. Liu. Low-rank and sparse structure pursuit via alternating minimization. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 600–609, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/gu16.html.

[15] T. Hastie, R. Mazumder, J. Lee, and R. Zadeh. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *The Journal of Machine Learning Research*, 16:3367–3402, jan 2015.

[16] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *EEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

[17] M. Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[18] H. A. L. Kiers. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56(2):197–212, Jun 1991. ISSN 1860-0980. doi: 10.1007/BF02294458. URL https://doi.org/10.1007/BF02294458.

[19] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[20] O. Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics* , 9(2):2348–2369, 2015. URL https://hal.archives-ouvertes.fr/hal-01111757.

[21] O. Klopp, K. Lounici, and A. B. Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1):523–564, Oct 2017. doi: 10.1007/s00440-016-0736-y. URL https://doi.org/10.1007/s00440-016-0736-y.

[22] N. K. Kumar and J. Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017. doi: 10.1080/03081087.2016.1267104. URL https://doi.org/10.1080/03081087.2016.1267104.

[23] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate frank-wolfe optimization for structural svms. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages I–53. JMLR. org, 2013.

[24] A. J. Landgraf and Y. Lee. Generalized principal component analysis: Projection of saturated model parameters. Technical report, The Ohio State University, Department of Statistics, 06 2015.

[25] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *Advances in neural information processing systems*, pages 612–620, 2011.

[26] L. T. Liu, E. Dobriban, and A. Singer. e pca: High dimensional exponential family pca. *Annals of Applied Statistics, to appear*, 2018.

[27] C. Mu, Y. Zhang, J. Wright, and D. Goldfarb. Scalable robust matrix recovery: Frank–wolfe meets proximal methods. *SIAM Journal on Scientific Computing*, 38(5):A3291–A3317, 2016. doi: 10.1137/15M101628X. URL https://doi.org/10.1137/15M101628X.

[28] J. Pagès. *Multiple factor analysis by example using R*. Chapman and Hall/CRC, 2014.

[29] M. Tao and X. Yuan. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.

[30] M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1), 2016. doi: 10.1561/2200000055. URL http://dx.doi.org/10.1561/2200000055.

[31] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

[32] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'10, pages 2496–2504, USA, 2010. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2997046.2997174.

[33] X. Zhang, L. Wang, and Q. Gu. A unified framework for nonconvex low-rank plus sparse matrix recovery. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1097–1107, 2018. URL http://proceedings.mlr.press/v84/zhang18c.html.

[34] W. Zheng, A. Bellet, and P. Gallinari. A distributed frank-wolfe framework for learning low-rank matrices with the trace norm. *arXiv preprint arXiv:1712.07495*, 2017.

A Statistical guarantees

A.1 Main result

We recall the convergence rates for the Frobenius norm of the errors $\Delta\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^0$ and $\Delta\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0$ given in Section 2. Define $d_{\mathbf{X}} = \max_k \|\mathbf{X}(k)\|_1$ and the following quantities:

$$\mathsf{D}_{\boldsymbol{\alpha}} = \frac{\|\boldsymbol{\alpha}^0\|_1}{\pi}\frac{\log(n+p)}{\sigma_-^2\gamma} + \left(\frac{a}{\pi}\right)^2 \log(n+p),$$

$$\mathsf{D}_{\boldsymbol{\Theta}} = \mathsf{D}_{\boldsymbol{\alpha}} + d_{\mathbf{X}}\|\boldsymbol{\alpha}^0\|_1 \left\{ \frac{12\pi\sqrt{\log(n+p)}}{\gamma(1+\nu)a\sigma_+\sqrt{\beta}} + \frac{1}{\pi\sigma_-^2}\left(\frac{\log(n+p)}{\gamma}\right) + 1 \right\}.$$

We assume that $M = (n \vee p)$ is large enough, that is

$$M \geq \left\{ \frac{4\sigma_+^2}{\gamma^6}\log^2\left(\frac{\sqrt{n\wedge p}}{p\gamma\sigma_-}\right) \vee 2\exp\left(\sigma_+^2/\gamma^2 \vee \sigma_+^2\gamma(1+\nu a)\right) \right\}.$$

Theorem 3 *Assume H1-6. Set*

$$\lambda_L = 2C\sigma_+\sqrt{\pi\max(n,p)\log(n+p)}, \text{ and } \lambda_S \geq 24\max_k \|\mathbf{X}(k)\|_1 \log(n+p)/\gamma, \tag{22}$$

where C is a positive constant. Assume that $\max(n,p) \geq 4\sigma_+^2/\gamma^6 \log^2(\sqrt{\min(n,p)/(\pi\gamma\sigma_-)}) + 2\exp(\sigma_+^2/\gamma^2 + 2\sigma_+^2\gamma a)$. *Then, with probability at least* $1-9(n+p)^{-1}$,

$$\begin{aligned} \left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0\right\|_2^2 &\leq C_1 \frac{s d_{\mathbf{X}} \log(n+p)}{\kappa^2 \pi} + \mathsf{D}_{\boldsymbol{\alpha}}, \\ \left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^0\right\|_F^2 &\leq C_2 \left(\frac{r\max(n,p)}{\pi} + \frac{s d_{\mathbf{X}}}{\pi}\right) \log(n+p) + \mathsf{D}_{\boldsymbol{\Theta}}. \end{aligned} \tag{23}$$

In (23), $s := \|\boldsymbol{\alpha}^0\|_0$, $r := \operatorname{rank}(\boldsymbol{\Theta}^0)$. *$C_1$ and C_2 are positive constants and $\mathsf{D}_{\boldsymbol{\alpha}}$ and $\mathsf{D}_{\boldsymbol{\Theta}}$ are residuals of lower order whose exact values are given in Appendix A.*

Denoting by $\lesssim$ the inequality up to constant and logarithmic factors, the order of magnitude of the bounds are therefore:

$$\begin{aligned} \|\Delta\boldsymbol{\alpha}\|_2^2 &\lesssim \frac{s d_{\mathbf{X}}}{p\kappa^2}, \\ \|\Delta\boldsymbol{\Theta}\|_F^2 &\lesssim \frac{r\beta}{p^2} + \frac{s d_{\mathbf{X}}}{p}, \end{aligned}$$

where $s = \left\|\boldsymbol{\alpha}^0\right\|_0$ and $r = \operatorname{rank}(\boldsymbol{\Theta}^0)$. In the case of almost uniform sampling, *i.e.* $c_1\pi \leq \pi_{ij} \leq c_2\pi$ for all $(i,j) \in [n] \times [p]$ and two positive constants c_1 and c_2, we obtain that $\beta \leq c_2(n \vee p)\pi$, which yields the following simplified bound:

$$\|\Delta\boldsymbol{\Theta}\|_F^2 \lesssim \frac{rM}{\pi} + \frac{s d_{\mathbf{X}}}{\pi}. \tag{24}$$

The rate given in (24) is the sum of the usual low-rank convergence rate rM/p and, when $d_{\mathbf{X}}$ is a constant, of the usual sparse vector convergence rate.

A.2 Sketch of the proof

Let $\{\epsilon_{ij}\}$ be an i.i.d. Rademacher sequence independent of Y and Ω. We define

$$\Sigma_R = \sum_{i=1}^{n}\sum_{j=1}^{p} \omega_{ij}\epsilon_{ij}E_{ij}.$$

In Theorem 4 we give a general result under some assumptions on the regularization parameters λ_L and λ_S, which depend on the random matrices $\nabla\mathcal{L}(\mathbf{M}^0)$ and Σ_R. Then, Lemma 4 and 5 allow us to compute values of λ_L and λ_S that satisfy the assumptions of Theorem 4 with high probability. Finally we combining these results yield Theorem 3. Define

$$\Psi_{\boldsymbol{\alpha}} = \frac{\left\|\boldsymbol{\alpha}^0\right\|_1}{\pi}\left\{\frac{\lambda_S}{\sigma_-^2} + a^2 d_{\mathbf{X}}\mathbb{E}\|\Sigma_R\|_\infty\right\} + \left(\frac{a}{\pi}\right)^2 \log(n+p), \tag{25}$$

$$\Psi_{\boldsymbol{\Theta}} = \frac{r}{\pi^2}\mathbb{E}\|\Sigma_R\|^2 + \frac{\|\boldsymbol{\alpha}\|_1}{\pi}\left\{\frac{\lambda_S}{(1+\nu)a\lambda_L} + d_{\mathbf{X}}\mathbb{E}\|\Sigma_R\|_\infty\right\} + \Psi_{\boldsymbol{\alpha}}. \tag{26}$$

Theorem 4 *Let*

$$\lambda_L \geq 2\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|, \quad \lambda_S \geq 2d_{\mathbf{X}}\left(\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2(1+\nu)a\right),$$

*and assumptions **H** 2-6 hold. Then, with probability at least* $1-8(n+p)^{-1}$

$$\begin{aligned} &(i) \qquad \|\Delta\boldsymbol{\alpha}\|_2^2 \leq \frac{C}{\kappa^2}\Psi_{\boldsymbol{\alpha}}, \textit{ and} \\ &(ii) \qquad \|\Delta\boldsymbol{\Theta}\|_F^2 \leq C\left\{\frac{r\lambda_L^2}{\pi^2\sigma_-^4} + (1+\nu)a\Psi_{\boldsymbol{\Theta}}\right\}. \end{aligned} \tag{27}$$

Denote $\Delta\mathbf{M} = \hat{\mathbf{M}} - \mathbf{M}^0$. We first derive an upper bound on the Frobenius error restricted to the observed entries $\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2$. Then we show some restricted strong convexity property, meaning that $\mathbb{E}\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2$ is upper bounded by $\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2$ up to a residual term defined later.

Upper bound on $\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2$. By definition of $\hat{\mathbf{\Theta}}$ and $\hat{\boldsymbol{\alpha}}$:

$$\mathcal{L}(\hat{\mathbf{M}}) - \mathcal{L}(\mathbf{M}^0) \leq \lambda_L \left(\left\|\mathbf{\Theta}^0\right\|_\star - \left\|\hat{\mathbf{\Theta}}\right\|_\star \right) + \lambda_S \left(\left\|\boldsymbol{\alpha}^0\right\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1 \right).$$

Recall that, for $\boldsymbol{\alpha} \in \mathbb{R}^q$, we use the notation $f_U(\boldsymbol{\alpha}) = \sum_{k=1}^q \boldsymbol{\alpha}_k \mathbf{X}(k)$. Adding $\langle \nabla\mathcal{L}(\mathbf{M}^0), \Delta\mathbf{M}\rangle$ on both sides of the last inequality, we get

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{M}}) - \mathcal{L}(\mathbf{M}^0) + \langle \nabla\mathcal{L}(\mathbf{M}^0), \Delta\mathbf{M}\rangle \leq \lambda_L \left(\left\|\mathbf{\Theta}^0\right\|_\star - \left\|\hat{\mathbf{\Theta}}\right\|_\star \right) - \langle \nabla\mathcal{L}(\mathbf{M}^0), \Delta\mathbf{\Theta}\rangle \\ + \lambda_S \left(\left\|\boldsymbol{\alpha}^0\right\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1 \right) - \langle \nabla\mathcal{L}(\mathbf{M}^0), f_U(\Delta\boldsymbol{\alpha})\rangle.\end{aligned}$$

The strong convexity of the link functions g_j, $j \in [p]$, allows us to lower bound the left hand side term and obtain

$$\begin{aligned}\frac{\sigma_-^2}{2} \|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2 \leq \lambda_L \left(\left\|\mathbf{\Theta}^0\right\|_\star - \left\|\hat{\mathbf{\Theta}}\right\|_\star \right) - \langle \nabla\mathcal{L}(\mathbf{M}^0), \Delta\mathbf{\Theta}\rangle \\ + \lambda_S \left(\left\|\boldsymbol{\alpha}^0\right\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1 \right) - \langle \nabla\mathcal{L}(\mathbf{M}^0), f_U(\Delta\boldsymbol{\alpha})\rangle.\end{aligned}$$

We now upper bound the right hand side using the following three agruments: the duality of the norms $\|\cdot\|_\star$ and $\|\cdot\|$ on the one hand and of the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on the other hand, the triangular inequality and the following assumptions:

$$\lambda_L \geq 2 \left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|, \quad \lambda_S \geq 2 \left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty d_{\mathbf{X}}.$$

We obtain

$$\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2 \leq \frac{3\lambda_L}{\sigma_-^2} \sqrt{2\,\mathrm{rank}(\mathbf{M}^0)}\, \|\Delta\mathbf{\Theta}\|_F + \frac{3\lambda_S}{\sigma_-^2} \left\|\boldsymbol{\alpha}^0\right\|_1. \tag{28}$$

Restricted strong convexity We now show that when the errors $\Delta\mathbf{\Theta}$ and $\Delta\boldsymbol{\alpha}$ belong to a subspace $\mathcal{C}$ and for a residual D - both defined later on - the following holds with high probability:

$$\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2 \geq \mathbb{E}\,\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2 - \mathsf{D}. \tag{29}$$

We start by defining the set $\mathcal{C}$ and prove that it contains the errors $\Delta\mathbf{\Theta}$ and $\Delta\boldsymbol{\alpha}$ with high probability (Lemma 1-2); then we show that restricted strong convexity holds on this subspace (Lemma 3).

For non-negative constants d_1, d_Π, $\rho < m$ and ε that will be specified later on, define the two following sets:

$$\mathcal{A}(d_1, d_\Pi) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^q : \ \|\boldsymbol{\alpha}\|_1 \leq d_1, \ \|\mathcal{P}_\Omega(f_U(\boldsymbol{\alpha}))\|_F^2 \leq d_\Pi \right\}. \tag{30}$$

The constants d_1 and d_Π define the constraints on the ℓ_1 norm of $\boldsymbol{\alpha}$ and weighted Frobenius norm of $f_U(\boldsymbol{\alpha})$.

$$\begin{aligned}\mathcal{L}(\rho, \varepsilon) \quad = \Bigg\{ \mathbf{\Theta} \in \mathbb{R}^{n\times p}, \boldsymbol{\alpha} \in \mathbb{R}^q : \|\mathcal{P}_\Omega(\mathbf{\Theta} + f_U(\boldsymbol{\alpha}))\|_F^2 \geq \frac{72\log(n+p)}{\pi\log(6/5)}, \\ \|\mathbf{\Theta} + f_U(\boldsymbol{\alpha})\|_\infty \leq 1, \|\mathbf{\Theta}\|_\star \leq \sqrt{\rho}\,\|\mathbf{\Theta}\|_F + \varepsilon \Bigg\}\end{aligned} \tag{31}$$

Condition $\|\mathbf{\Theta}\|_\star \leq \sqrt{\rho}\,\|\mathbf{\Theta}\|_F + \varepsilon$ is a relaxed form of the condition $\|\mathbf{\Theta}\|_\star \leq \sqrt{\rho}\,\|\mathbf{\Theta}\|_F$ satisfied for matrices of rank ρ. Finally, we define the constrained set of interest:

$$\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) = \mathcal{L}(\rho, \varepsilon) \cap \left\{ \mathbb{R}^{n\times p} \times \mathcal{A}(d_1, d_\Pi) \right\}.$$

Let

$$\begin{aligned} d_1 \quad &= 4\,\|\boldsymbol{\alpha}\|_1, \\ d_\Pi \quad &= \frac{3\lambda_S}{\sigma_-^2} \left\|\boldsymbol{\alpha}^0\right\|_1 + 64a^2 d_{\mathbf{X}} \mathbb{E}\|\Sigma_R\|_\infty \|\boldsymbol{\alpha}\|_1 + 3072a^2\pi^{-1} + \frac{72a^2\log(n+p)}{\log(6/5)}. \end{aligned}$$

The following Lemma, proved in Appendix B.1 states that with high probability, $\Delta\boldsymbol{\alpha} \in \mathcal{A}(d_1, d_\Pi)$.

Lemma 1 *Let* $\lambda_S \geq 2d_{\mathbf{X}} \left(\left\| \nabla \mathcal{L}(\mathbf{M}^0) \right\|_\infty + 2\sigma_+^2(1+d_{\mathbf{X}})a \right)$ *and assume* ***H*** *2-6 hold. Then, with probability at least* $1-8(n+p)^{-1}$,

$$\Delta\boldsymbol{\alpha} \in \mathcal{A}(d_1, d_\Pi);$$

Lemma 1 (proved in Appendix B.2) implies (i) of Theorem 4. Thus, we only need to prove (ii).

Lemma 2 *Let*

$$\lambda_L \geq 2 \left\| \nabla \mathcal{L}(\mathbf{M}^0) \right\|, \quad \lambda_S \geq 2d_{\mathbf{X}} \left(\left\| \nabla \mathcal{L}(\mathbf{M}^0) \right\|_\infty + 2\sigma_+^2(1+d_{\mathbf{X}})a \right),$$

and assumption ***H*** *4 hold. Then, for* $\rho = 32r$ *and* $\varepsilon = 3\lambda_S/\lambda_L \left\| \boldsymbol{\alpha}^0 \right\|_1$,

$$\|\Delta\boldsymbol{\Theta}\|_\star \leq \sqrt{\rho} \|\Delta\boldsymbol{\Theta}\|_F + \varepsilon.$$

A proof of Lemma 2 can be found in Appendix B.2. As a consequence, under the conditions on the regularization parameters λ_L and λ_S given in Lemma 2 and whenever

$$\mathbb{E} \|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 \geq \frac{72\log(n+p)}{\pi\log(6/5)},$$

the error terms $(\Delta\boldsymbol{\Theta}, \Delta\boldsymbol{\alpha})$ belong to the constrained set $\mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$ with high probability. We therefore consider the two possible cases: $\mathbb{E} \|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 < \frac{72\log(n+p)}{\pi\log(6/5)}$ and $\mathbb{E} \|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 \geq \frac{72\log(n+p)}{\pi\log(6/5)}$.

Case 1: Suppose $\mathbb{E} \|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 < \frac{72\log(n+p)}{\pi\log(6/5)}$. Then, Lemma 1 combined with the fact that $\|\mathbf{M}\|_F^2 \leq \pi^{-1} \|\mathcal{P}_\Omega(\mathbf{M})\|_F^2$ for all $\mathbf{M}$, and the identity $(a+b)^2 \geq a^2/4 - 4b^2$ ensures that

$$\|\Delta\boldsymbol{\Theta}\|_F^2 \leq 4 \|\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha})\|_F^2 + 16 \|f_U(\Delta\boldsymbol{\alpha})\|_F^2,$$

therefore

$$\|\Delta\boldsymbol{\Theta}\|_F^2 \leq \frac{288a^2\log(n+p)}{\log(6/5)} + 16\Phi_{\boldsymbol{\alpha}},$$

which implies (ii) of Theorem 4.

Case 2: Suppose $\mathbb{E} \|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 \geq \frac{72\log(n+p)}{\pi\log(6/5)}$. Then, Lemma 1 and 2 yield that with probability at least $1-8(n+p)^{-1}$,

$$\left(\frac{\Delta\boldsymbol{\Theta}}{2(1+\nu)a}, \frac{\Delta\boldsymbol{\alpha}}{2(1+\nu)a} \right) \in \mathcal{C}(d_1', d_\Pi', \rho', \varepsilon'), \text{ with}$$

$$\begin{aligned} d_1' &= \frac{d_1}{2(1+\nu)a}, & d_\Pi' &= \frac{d_\Pi}{4(1+\nu)^2a^2}, \\ \rho' &= \rho, & \varepsilon' &= \frac{\varepsilon}{2(1+\nu)a}, \end{aligned}$$

where d_1, d_Π, ρ and ε are defined in Lemma 1 and 2. We use the following result, proved in Appendix B.3. Define the set $\tilde{\mathcal{A}}(d_1)$ as follows:

$$\tilde{\mathcal{A}}(d_1) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^q : \quad \|\boldsymbol{\alpha}\|_\infty \leq 1; \quad \|\boldsymbol{\alpha}\|_1 \leq d_1; \quad \|\mathcal{P}_\Omega(f_U\boldsymbol{\alpha})\|_F^2 \geq \frac{18\log(n+p)}{\pi\log(6/5)} \right\}.$$

Let d_1, d_Π, ρ and ε be positive constants, and

$$\begin{aligned} \mathsf{D}_{\boldsymbol{\alpha}} &= 8\nu d_1 d_{\mathbf{X}} \mathbb{E}\|\Sigma_R\|_\infty + 768\pi^{-1}, \\ \mathsf{D}_X &= \frac{112\rho}{\pi} \mathbb{E}\|\Sigma_R\|^2 + 8\nu\varepsilon\mathbb{E}\|\Sigma_R\| + 8\nu d_1 d_{\mathbf{X}} \mathbb{E}\|\Sigma_R\|_\infty + d_\Pi + 768\pi^{-1}. \end{aligned} \tag{32}$$

Lemma 3 *Assume* ***H*** *6. Then, the following properties hold:*

(i) For any $\boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1)$, with probability at least $1 - 8(n+p)^{-1}$,

$$\|\mathcal{P}_\Omega(f_U(\boldsymbol{\alpha}))\|_F^2 \geq \frac{1}{2}\mathbb{E}\,\|\mathcal{P}_\Omega(f_U(\boldsymbol{\alpha}))\|_F^2 - \mathsf{D}_{\boldsymbol{\alpha}}.$$

(ii) For any pair $(\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon)$, with probability at least $1 - 8(n+p)^{-1}$

$$\|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 \geq \frac{1}{2}\mathbb{E}\,\|\mathcal{P}_\Omega(\Delta\boldsymbol{\Theta} + f_U(\Delta\boldsymbol{\alpha}))\|_F^2 - \mathsf{D}_X. \tag{33}$$

Lemma 3 is proved in Appendix B.3. We apply Lemma 3 (ii) to $\left(\frac{\Delta\boldsymbol{\Theta}}{2(1+\nu)a}, \frac{\Delta\boldsymbol{\alpha}}{2(1+\nu)a}\right)$ which implies that with probability at least $1 - 8(n+p)^{-1}$, $\mathbb{E}\mathbb{E}\,\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2 \leq 2\,\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2 + 2(1+\nu)a\Psi_{\boldsymbol{\Theta}}$. Combined with (28) and $\|\Delta\mathbf{M}\|_F^2 \leq \pi^{-1}\mathbb{E}\,\|\mathcal{P}_\Omega(\Delta\mathbf{M})\|_F^2$, it implies that

$$\|\Delta\mathbf{M}\|_F^2 \leq \frac{6\sqrt{2r}\lambda_L}{p\sigma_-^2}\,\|\Delta\boldsymbol{\Theta}\|_F + \frac{6\lambda_S}{\pi\sigma_-^2}\,\left\|\boldsymbol{\alpha}^0\right\|_1 + 2(1+\nu)a\Psi_{\boldsymbol{\Theta}}.$$

Now using $\|\Delta\mathbf{M}\|_F^2 \geq \frac{\|\Delta\boldsymbol{\Theta}\|_F^2}{2} - \|f_U(\Delta\boldsymbol{\alpha})\|_F^2$ and $\frac{6\sqrt{2r}\lambda_L}{\pi\sigma_-^2}\,\|\Delta\boldsymbol{\Theta}\|_F \leq \frac{\|\Delta\boldsymbol{\Theta}\|_F^2}{4} + \frac{288r\lambda_L^2}{p^2\sigma_-^4}$, we obtain

$$\|\Delta\boldsymbol{\Theta}\|_F^2 \leq \frac{1152r\lambda_L^2}{p^2\sigma_-^4} + \frac{24\lambda_S\left\|\boldsymbol{\alpha}^0\right\|_1}{\pi\sigma_-^2} + 2(1+\nu)a\Psi_{\boldsymbol{\Theta}} + 4\Psi_{\boldsymbol{\alpha}},$$

which gives the result of Theorem 4 (ii).

We now give deterministic upper bounds on $\mathbb{E}\|\Sigma_R\|$ and $\mathbb{E}\|\Sigma_R\|_\infty$, and probabilistic upper bounds on $\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|$ and $\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty$. We will use them to select values of λ_L and λ_S which satisfy the assumptions of Theorem 4 and compute the corresponding upper bounds.

Lemma 4 *[21, Lemma 10] Let assumption **H** 6 hold. Then, there exists an absolute constant $C^\star$ such that the two following inequalities hold*

$$\mathbb{E}\|\Sigma_R\|_\infty \leq 1, \text{ and}$$
$$\mathbb{E}\|\Sigma_R\| \leq C^\star\left\{\sqrt{\beta} + \sqrt{\log(\min(n,p))}\right\}.$$

Lemma 5 *[21, Lemma 10] Let assumptions **H** 1-6 hold. Then, there exists an absolute constant $c^\star$ such that the following two inequalities hold with probability at least $1 - (n+p)^{-1}$.*

$$\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty \leq 6\max\left\{\sigma_+\sqrt{\log(n+p)}, \frac{\log(n+p)}{\gamma}\right\}, \tag{34}$$

$$\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\| \leq c^\star\max\left\{\sigma_+\sqrt{\beta\log(n+p)}, \frac{\log(n+p)}{\gamma}\log\left(\frac{1}{\sigma_-}\sqrt{\frac{np}{\beta}}\right)\right\}. \tag{35}$$

From Theorem 4, Lemma 4 and 5 combined with a union bound argument, we deduce result given in Section 2.

B Technical results

B.1 Proof of Lemma 1

We start by proving $\|\Delta\boldsymbol{\alpha}\|_1 \leq 4\left\|\boldsymbol{\alpha}^0\right\|_1$. By the optimality conditions over a convex set [2, Chapter 4, Section 2, Proposition 4], there exist two subgradients $\hat{f}_{\boldsymbol{\Theta}}$ in the subdifferential of $\|\cdot\|_\star$ taken at $\hat{\boldsymbol{\Theta}}$ and $\hat{f}_{\boldsymbol{\alpha}}$ in the subdifferential of $\|\cdot\|_1$ taken at $\hat{\boldsymbol{\alpha}}$, such that for all feasible pairs $(\boldsymbol{\Theta}, \boldsymbol{\alpha})$ we have

$$\langle\nabla\mathcal{L}(\hat{\mathbf{M}}), \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}} + \sum_{k=1}^{q}(\boldsymbol{\alpha}_k - \hat{\boldsymbol{\alpha}}_k)\mathbf{X}(k)\rangle + \lambda_L\langle\hat{f}_{\boldsymbol{\Theta}}, \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}\rangle + \lambda_S\langle\hat{f}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\rangle \geq 0. \tag{36}$$

Applying inequality (36) to the pair $(\hat{\mathbf{\Theta}}, \boldsymbol{\alpha}^0)$ we obtain

$$\langle \nabla \mathcal{L}(\hat{\mathbf{M}}), \sum_{k=1}^{q} \Delta\boldsymbol{\alpha}_k \mathbf{X}(k) \rangle + \lambda_S \langle \hat{f}_{\boldsymbol{\alpha}}, \Delta\boldsymbol{\alpha} \rangle \geq 0.$$

Denote $\tilde{\mathbf{M}} = \hat{\mathbf{\Theta}} + \sum_{k=1}^{q} \boldsymbol{\alpha}_k^0 \mathbf{X}(k)$. The last inequality is equivalent to

$$\underbrace{\langle \nabla \mathcal{L}(\mathbf{M}^0), f_U(\Delta\boldsymbol{\alpha}) \rangle}_{\mathsf{B}_1} + \underbrace{\langle \nabla \mathcal{L}(\tilde{\mathbf{M}}) - \nabla \mathcal{L}(\mathbf{M}^0), f_U(\Delta\boldsymbol{\alpha}) \rangle}_{\mathsf{B}_2} + \underbrace{\langle \nabla \mathcal{L}(\hat{\mathbf{M}}) - \nabla \mathcal{L}(\tilde{\mathbf{M}}), f_U(\Delta\boldsymbol{\alpha}) \rangle}_{\mathsf{B}_3} + \lambda_S \langle \hat{f}_{\boldsymbol{\alpha}}, \Delta\boldsymbol{\alpha} \rangle \geq 0.$$

We now derive upper bounds on the three terms B_1, B_2 and B_3 separately. Recall that we denote $d_{\mathbf{X}} = \max_k \|\mathbf{X}(k)\|_1$ and bound B_1 as follows:

$$\mathsf{B}_1 \leq \|\Delta\boldsymbol{\alpha}\|_1 \left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty d_{\mathbf{X}}. \tag{37}$$

Similarly, the duality between $\|\cdot\|_\infty$ and $\|\cdot\|_1$ gives

$$\mathsf{B}_2 \leq \|\Delta\boldsymbol{\alpha}\|_1 \left\|\nabla \mathcal{L}(\tilde{\mathbf{M}}) - \nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty d_{\mathbf{X}}.$$

Moreover, $\nabla \mathcal{L}(\tilde{\mathbf{M}}) - \nabla \mathcal{L}(\mathbf{M}^0)$ is a matrix with entries $g_j'(\tilde{\mathbf{M}}_{ij}) - g_j'(\mathbf{M}^0_{ij})$, therefore assumption **H** 4 ensures

$$\left\|\nabla \mathcal{L}(\tilde{\mathbf{M}}) - \nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty \leq 2\sigma_+^2(1+\nu)a,$$

and finally we obtain

$$\mathsf{B}_2 \leq \|\Delta\boldsymbol{\alpha}\|_1 2\sigma_+^2(1+\nu)a d_{\mathbf{X}}. \tag{38}$$

We finally bound B_3 as follows. We have that

$$\mathsf{B}_3 = \sum_{i=1}^{n}\sum_{j=1}^{p} \omega_{ij} \left(g_j'(\hat{\mathbf{M}}_{ij}) - g_j'(\tilde{\mathbf{M}}_{ij})\right)\left(\tilde{\mathbf{M}}_{ij} - \hat{\mathbf{M}}_{ij}\right).$$

Now, for all $j \in [p]$, g_j' is increasing therefore

$$\left(g_j'(\hat{\mathbf{M}}_{ij}) - g_j'(\tilde{\mathbf{M}}_{ij})\right)\left(\tilde{\mathbf{M}}_{ij} - \hat{\mathbf{M}}_{ij}\right) \leq 0,$$

which implies $\mathsf{B}_3 \leq 0$. Combined with (37) and (38) this yields

$$\lambda_S \langle \hat{f}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \rangle \leq \|\Delta\boldsymbol{\alpha}\|_1 d_{\mathbf{X}} \left(\left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2(1+\nu)a\right).$$

Besides, the convexity of $\|\cdot\|_1$ gives $\langle \hat{f}_{\boldsymbol{\alpha}}, \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \rangle \geq \|\hat{\boldsymbol{\alpha}}\|_1 - \|\boldsymbol{\alpha}\|_1$, therefore

$$\left\{\lambda_S - d_{\mathbf{X}} \left(\left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2(1+\nu)a\right)\right\} \|\hat{\boldsymbol{\alpha}}\|_1 \leq \left\{\lambda_S + d_{\mathbf{X}} \left(\left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2(1+\nu)a\right)\right\} \|\boldsymbol{\alpha}\|_1,$$

and the condition $\lambda_S \geq 2\left\{d_{\mathbf{X}} \left(\left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2(1+\nu)a\right)\right\}$ gives $\|\hat{\boldsymbol{\alpha}}\|_1 \leq 3\|\boldsymbol{\alpha}\|_1$ and finally

$$\|\Delta\boldsymbol{\alpha}\|_1 \leq 4\|\boldsymbol{\alpha}\|_1. \tag{39}$$

We consider the two following cases.

Case 1: $\mathbb{E}\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\|_F^2 < \frac{72a^2 \log(n+p)}{\pi \log(6/5)}$. Then the result holds trivially.

Case 2: $\mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 \geq \frac{72a^2\log(n+p)}{\pi\log(6/5)}$. For $d_1 > 0$ recall the definition of the set

$$\tilde{\mathcal{A}}(d_1) = \left\{\boldsymbol{\alpha}\in\mathbb{R}^q : \quad \|\boldsymbol{\alpha}\|_\infty \leq 1; \quad \|\boldsymbol{\alpha}\|_1 \leq d_1; \quad \mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 \geq \frac{18\log(n+p)}{\pi\log(6/5)}\right\}.$$

Inequality (39) and $\|\Delta\boldsymbol{\alpha}\|_\infty \leq 2a$ imply that

$$\frac{\Delta\boldsymbol{\alpha}}{2a} \in \tilde{\mathcal{A}}\left(\frac{2\|\boldsymbol{\alpha}\|_1}{a}\right).$$

Therefore we can apply Lemma 3(i) and obtain that with probability at least $1-8(n+p)^{-1}$,

$$\mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|^2 \leq 2\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 + 64\nu a\|\boldsymbol{\alpha}\|_1 d_{\mathbf{X}}\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{40}$$

We now must upper bound the quantity $\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\|_F^2$. Recall that $\tilde{\mathbf{M}} = \sum_{k=1}^q \boldsymbol{\alpha}_k\mathbf{X}(k) + \hat{\mathbf{M}}$. By definition,

$$\mathcal{L}(\hat{\mathbf{X}}) + \lambda_L\left\|\hat{\boldsymbol{\Theta}}\right\|_\star + \lambda_S\|\hat{\boldsymbol{\alpha}}\|_1 \leq \mathcal{L}(\tilde{\mathbf{M}}) + \lambda_L\left\|\hat{\boldsymbol{\Theta}}\right\|_\star + \lambda_S\|\boldsymbol{\alpha}\|_1,$$

i.e.

$$\mathcal{L}(\hat{\mathbf{M}}) - \mathcal{L}(\tilde{\mathbf{M}}) \leq \lambda_S\left(\|\boldsymbol{\alpha}\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1\right).$$

Substracting $\langle\nabla\mathcal{L}(\tilde{\mathbf{M}}), \hat{\mathbf{M}}-\tilde{\mathbf{M}}\rangle$ on both sides and by strong convexity of $\mathcal{L}$ we obtain

$$\begin{aligned}\frac{\sigma_-^2}{2}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|^2 &\leq \lambda_S\left(\|\boldsymbol{\alpha}\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1\right) + \langle\nabla\mathcal{L}(\tilde{\mathbf{M}}), \mathsf{f}_U(\Delta\boldsymbol{\alpha})\rangle \\ &\leq \lambda_S\left(\|\boldsymbol{\alpha}\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1\right) + \underbrace{\left|\langle\nabla\mathcal{L}(\mathbf{M}^0), \mathsf{f}_U(\Delta\boldsymbol{\alpha})\rangle\right|}_{\mathsf{C}_1} \\ &\quad + \underbrace{\left|\langle\nabla\mathcal{L}(\mathbf{M}^0) - \nabla\mathcal{L}(\tilde{\mathbf{M}}), \mathsf{f}_U(\Delta\boldsymbol{\alpha})\rangle\right|}_{\mathsf{C}_2}.\end{aligned} \tag{41}$$

The duality of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ yields $\mathsf{C}_1 \leq \left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty d_{\mathbf{X}}\|\Delta\boldsymbol{\alpha}\|_1$, and

$$\mathsf{C}_2 \leq \left\|\nabla\mathcal{L}(\mathbf{M}^0) - \nabla\mathcal{L}(\tilde{\mathbf{M}})\right\|_\infty d_{\mathbf{X}}\|\Delta\boldsymbol{\alpha}\|_1.$$

Furthermore,

$$\left\|\nabla\mathcal{L}(\mathbf{M}^0) - \nabla\mathcal{L}(\tilde{\mathbf{M}})\right\|_\infty \leq 2\sigma_+^2 a,$$

since for all $(i,j)\in[n]\times[p]$ $|\tilde{\mathbf{M}}_{ij} - \mathbf{M}^0_{ij}| \leq 2a$ and $g_j''(\tilde{\mathbf{M}}_{ij}) \leq \sigma_+^2$. The last three inequalities plugged in (41) give

$$\frac{\sigma_-^2}{2}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 \leq \lambda_S\left(\|\boldsymbol{\alpha}\|_1 - \|\hat{\boldsymbol{\alpha}}\|_1\right) + d_{\mathbf{X}}\|\Delta\boldsymbol{\alpha}\|_1\left\{\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2 a\right\}.$$

The triangular inequality gives

$$\begin{aligned}\frac{\sigma_-^2}{2}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 &\leq \left\{d_{\mathbf{X}}\left(\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2 a\right) + \lambda_S\right\}\|\boldsymbol{\alpha}\|_1 \\ &\quad + \left\{d_{\mathbf{X}}\left(\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2 a\right) - \lambda_S\right\}\|\hat{\boldsymbol{\alpha}}\|_1.\end{aligned}$$

Then, the assumption $\lambda_S \geq 2d_{\mathbf{X}}\left(\left\|\nabla\mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2(1+\nu)a\right)$ gives

$$\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 \leq \frac{3\lambda_S}{\sigma_-^2}\|\boldsymbol{\alpha}\|_1.$$

Plugged into (40), this last inequality implies that with probability at least $1-8(n+p)^{-1}$

$$\mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\Delta\boldsymbol{\alpha}))\right\|_F^2 \leq \frac{3\lambda_S}{\sigma_-^2}\|\boldsymbol{\alpha}\|_1 + 64\nu a\|\boldsymbol{\alpha}\|_1 d_{\mathbf{X}}\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + 3072a^2p^{-1}. \tag{42}$$

Combining (39) and (42) gives the result.

B.2 Proof of Lemma 2

Using (36) for $L = \mathbf{\Theta}^0$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}$ we obtain

$$\langle \nabla \mathcal{L}(\hat{\mathbf{M}}), \Delta\mathbf{\Theta} + \sum_{k=1}^{q} (\Delta\boldsymbol{\alpha}_k)\mathbf{X}(k) \rangle + \lambda_L \langle \hat{f}_L, \Delta\mathbf{\Theta} \rangle + \lambda_S \langle \hat{f}_{\boldsymbol{\alpha}}, \Delta\boldsymbol{\alpha} \rangle \geq 0.$$

Then, the convexity of $\|\cdot\|_*$ and $\|\cdot\|_1$ imply that

$$\begin{aligned} \left\|\mathbf{\Theta}^0\right\|_\star &\geq \left\|\hat{\mathbf{\Theta}}\right\|_\star + \langle \partial \left\|\hat{\mathbf{\Theta}}\right\|_\star, \Delta\mathbf{\Theta} \rangle, \\ \left\|\boldsymbol{\alpha}\right\|_1 &\geq \left\|\hat{\boldsymbol{\alpha}}\right\|_\star + \langle \partial \left\|\hat{\boldsymbol{\alpha}}\right\|_1, \Delta\boldsymbol{\alpha} \rangle. \end{aligned}$$

The last three inequalities yield

$$\begin{aligned} \lambda_L \left(\left\|\hat{\mathbf{\Theta}}\right\|_\star - \left\|\mathbf{\Theta}^0\right\|_\star \right) + \lambda_S \left(\left\|\hat{\boldsymbol{\alpha}}\right\|_1 - \left\|\boldsymbol{\alpha}\right\|_1 \right) &\leq \langle \nabla \mathcal{L}(\hat{\mathbf{M}}), \Delta\mathbf{\Theta} \rangle \\ &\quad + \langle \nabla \mathcal{L}(\hat{\mathbf{M}}), \sum_{k=1}^{q} (\Delta\boldsymbol{\alpha}_k)\mathbf{X}(k) \rangle \\ &\leq \left\|\nabla \mathcal{L}(\hat{\mathbf{M}})\right\| \left\|\Delta\mathbf{\Theta}\right\|_\star + d_{\mathbf{X}} \left\|\nabla \mathcal{L}(\hat{\mathbf{M}})\right\|_\infty \left\|\Delta\boldsymbol{\alpha}\right\|_1. \end{aligned}$$

Using the conditions

$$\lambda_L \geq 2 \left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|, \quad \lambda_S \geq 2 d_{\mathbf{X}} \left\{ \left\|\nabla \mathcal{L}(\mathbf{M}^0)\right\|_\infty + 2\sigma_+^2 (1+\nu) a \right\},$$

we get

$$\begin{aligned} \lambda_L \left(\left\|P_{\mathbf{\Theta}^0}^{\perp}(\Delta\mathbf{\Theta})\right\|_\star - \left\|P_{\mathbf{\Theta}^0}(\Delta\mathbf{\Theta})\right\|_\star \right) + \lambda_S \left(\left\|\hat{\boldsymbol{\alpha}}\right\|_1 - \left\|\boldsymbol{\alpha}\right\|_1 \right) \leq \\ \frac{\lambda_L}{2} \left(\left\|P_{\mathbf{\Theta}^0}^{\perp}(\Delta\mathbf{\Theta})\right\|_\star + \left\|P_{\mathbf{\Theta}^0}(\Delta\mathbf{\Theta})\right\|_\star \right) + \frac{\lambda_S}{2} \left\|\Delta\boldsymbol{\alpha}\right\|_1, \end{aligned}$$

which implies

$$\left\|P_{\mathbf{\Theta}^0}^{\perp}(\Delta\mathbf{\Theta})\right\|_\star \leq 3 \left\|P_{\mathbf{\Theta}^0}(\Delta\mathbf{\Theta})\right\|_\star + 3\lambda_S/\lambda_L \left\|\boldsymbol{\alpha}\right\|_1.$$

Now, using

$$\left\|\Delta\mathbf{\Theta}\right\|_\star \leq \left\|P_{\mathbf{\Theta}^0}^{\perp}(\Delta\mathbf{\Theta})\right\|_\star + \left\|P_{\mathbf{\Theta}^0}(\Delta\mathbf{\Theta})\right\|_\star, \quad \left\|P_{\mathbf{\Theta}^0}(\Delta\mathbf{\Theta})\right\|_F \leq \left\|\Delta\mathbf{\Theta}\right\|_F$$

and $\operatorname{rank}(P_{\mathbf{\Theta}^0}(\Delta\mathbf{\Theta})) \leq 2r$, we get

$$\left\|\Delta\mathbf{\Theta}\right\|_\star \leq \sqrt{32r} \left\|\Delta\mathbf{\Theta}\right\|_F + 3\lambda_S/\lambda_L \left\|\boldsymbol{\alpha}\right\|_1.$$

This completes the proof of Lemma 2.

B.3 Proof of Lemma 3

Proof of (i): Recall

$$\mathsf{D}_{\boldsymbol{\alpha}} = 8\nu d_1 d_{\mathbf{X}} \mathbb{E}\left[\left\|\Sigma_R\right\|_\infty\right] + 768 p^{-1}$$

and

$$\tilde{\mathcal{A}}(d_1) = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^q : \quad \left\|\boldsymbol{\alpha}\right\|_\infty \leq 1; \quad \left\|\boldsymbol{\alpha}\right\|_1 \leq d_1; \quad \mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\right\|_F^2 \geq \frac{18\log(n+p)}{\pi \log(6/5)} \right\}.$$

We will show that the probability of the following event is small:

$$\mathcal{B} = \left\{ \exists \boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1) \text{ such that } \left| \left\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\right\|_F^2 - \mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\right\|_F^2 \right| > \frac{1}{2}\mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\right\|_F^2 + \mathsf{D}_{\boldsymbol{\alpha}} \right\}.$$

Indeed, $\mathcal{B}$ contains the complement of the event we are interested in. We use a peeling argument to upper bound the probability of event $\mathcal{B}$. Let $\nu = \frac{18\log(n+p)}{\pi\log(6/5)}$ and $\eta = 6/5$. For $l \in \mathbb{N}$ set

$$\mathcal{S}_l = \left\{ \boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1) : \quad \eta^{l-1}\nu \leq \mathbb{E}\left\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\right\|_F^2 \leq \eta^l \nu \right\}.$$

Under the event $\mathcal{B}$, there exists $l \geq 1$ and $\boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1) \cap S_l$ such that

$$\begin{aligned} \left| \|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right| &> \frac{1}{2}\mathbb{E}\,\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 + \mathsf{D}_{\boldsymbol{\alpha}} \\ &> \frac{1}{2}\eta^{l-1}\nu + \mathsf{D}_{\boldsymbol{\alpha}} \\ &= \frac{5}{12}\eta^l\nu + \mathsf{D}_{\boldsymbol{\alpha}}. \end{aligned} \tag{43}$$

For $T > \nu$, consider the set of vectors

$$\tilde{\mathcal{A}}(d_1, T) = \left\{ \boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1) : \mathbb{E}\,\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \leq T \right\}$$

and the event

$$\mathcal{B}_l = \left\{ \exists \boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1, \eta^l\nu) : \left| \|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right| > \frac{5}{12}\eta^l\nu + \mathsf{D}_{\boldsymbol{\alpha}} \right\}.$$

If $\mathcal{B}$ holds, then (43) implies that $\mathcal{B}_l$ holds for some $l \leq 1$. Therefore ,$\mathcal{B} \subset \cup_{l=1}^{+\infty}\mathcal{B}_l$, and it is enough to estimate the probability of the events $\mathcal{B}_l$ and then apply the union bound. Such an estimation is given in the following Lemma, adapted from Lemma 10 in [20].

Lemma 6 *Define* $Z_T = \sup_{\boldsymbol{\alpha} \in \tilde{\mathcal{A}}(d_1,T)} \left| \|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right|$. *Then,*

$$\mathbb{P}\left(Z_T \geq \mathsf{D}_{\boldsymbol{\alpha}} + \frac{5}{12}T \right) \leq 4\mathrm{e}^{-\pi T/18}.$$

Lemma 6 gives that $\mathbb{P}(\mathcal{B}_l) \leq 4\exp(-\pi\eta^l\nu/18)$. Applying the union bound we obtain

$$\begin{aligned} \mathbb{P}(\mathcal{B}) &\leq \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{B}_l) \\ &\leq 4\sum_{l=1}^{\infty} \exp(-\pi\eta^l\nu/18) \\ &\leq 4\sum_{l=1}^{\infty} \exp(-\pi\log(\eta)l\nu/18), \end{aligned}$$

where we used $e^x \geq x$. Finally, for $\nu = \frac{18\log(n+p)}{\pi\log(6/5)}$ we obtain

$$\mathbb{P}(\mathcal{B}) \leq \frac{4\exp(-\pi\nu\log(\eta)/18)}{1-\exp(-\pi\nu\log(\eta)/18)} \leq \frac{4\exp(-\log(n+p))}{1-\exp(-\log(n+p))} \leq \frac{8}{n+p},$$

since $d - 1 \geq (n+p)/2$, which concludes the proof of (i).

Proof of (ii): The proof is very similar to that of (i); we recycle some of the notations for simplicity. Recall

$$\mathsf{D}_X = \frac{112\rho}{\pi}\mathbb{E}\left[\|\Sigma_R\|\right]^2 + 8\nu\varepsilon\mathbb{E}\left[\|\Sigma_R\|\right] + 8\nu d_1 d_{\mathbf{X}}\mathbb{E}\left[\|\Sigma_R\|_\infty\right] + d_\Pi + 768p^{-1}.$$

Let

$$\begin{aligned} \mathcal{B} = \Big\{ &\exists(\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon); \\ &\left| \|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right| > \frac{1}{2}\mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 + \mathsf{D}_X \Big\}, \end{aligned}$$

$\nu = \frac{72\log(n+p)}{\pi\log(6/5)}$, $\eta = \frac{6}{5}$ and for $l \in \mathbb{N}$

$$\mathcal{S}_l = \left\{ (\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) : \quad \eta^{l-1}\nu \leq \mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \leq \eta^l\nu \right\}.$$

As before, if $\mathcal{B}$ holds, then there exist $l \geq 2$ and $(\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) \cap S_l$ such that

$$\left| \|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right| \quad > \frac{5}{12}\eta^l \nu + \mathsf{D}_X. \tag{44}$$

For $T > \nu$, consider the set $\tilde{\mathcal{C}}(T) = \left\{ (\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \mathcal{C}(d_1, d_\Pi, \rho, \varepsilon) : \mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \leq T \right\}$, and the event

$$\mathcal{B}_l = \left\{ \exists (\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \tilde{\mathcal{C}}(\eta^l \nu) : \quad \left| \|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right| > \frac{5}{12}\eta^l \nu + \mathsf{D}_X \right\}.$$

Then, (44) implies that $\mathcal{B}_l$ holds and $\mathcal{B} \subset \cup_{l=1}^{+\infty} \mathcal{B}_l$. Thus, we estimate in Lemma 7 the probability of the events $\mathcal{B}_l$, and then apply the union bound.

Lemma 7 *Let* $W_T = \sup_{(\boldsymbol{\Theta}, \boldsymbol{\alpha}) \in \tilde{\mathcal{C}}(T)} \left| \|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 - \mathbb{E}\,\|\mathcal{P}_\Omega(\boldsymbol{\Theta} + \mathsf{f}_U(\boldsymbol{\alpha}))\|_F^2 \right|$.

$$\mathbb{P}\left(W_T \geq \mathsf{D}_X + \frac{5}{12} T \right) \leq 4\mathrm{e}^{-\pi T/72}.$$

Lemma 7 gives that $\mathbb{P}(\mathcal{B}_l) \leq 4\exp(-\pi\eta^l\nu/72)$. Applying the union bound we obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{B}) \quad &\leq \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{B}_l) \\
&\leq 4 \sum_{l=1}^{\infty} \exp(-\pi\eta^l\nu/72) \\
&\leq 4 \sum_{l=1}^{\infty} \exp(-\pi \log(\eta) l\nu/72),
\end{aligned}$$

where we used $\mathrm{e}^x \geq x$. Finally, for $\nu = \frac{72\log(n+p)}{\pi \log(6/5)}$ we obtain

$$\mathbb{P}(\mathcal{B}) \leq \frac{4\exp(-\pi\nu\log(\eta)/72)}{1 - \exp(-\pi\nu\log(\eta)/72)} \leq \frac{4\exp(-\log(n+p))}{1 - \exp(-\log(n+p))} \leq 8(n+p)^{-1},$$

since $n + p - 1 \geq (n+p)/2$, which concludes the proof of (ii).

C Proof of Theorem 2

To prove the theorem, we first lower bound on the progress made by the algorithm at the two blocks between the iterations. With a slight abuse of notations, in the following we shall denote the iterates without the bracket in the superscripts, e.g., we denote $\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t)}, R^{(t)}$ by $\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t$, respectively, to simplify our discussions.

For the first block on $\boldsymbol{\alpha}$, in Section C.1 we show that

$$F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - \frac{\gamma}{2} \frac{\left(g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}; Q^{t-1})\right)^2}{(2Q^{t-1})^2}\,, \tag{45}$$

where $Q^{t-1} := \lambda_S^{-1} F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1})$ as defined in the main paper and

$$g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}; Q^{t-1}) := \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \boldsymbol{\alpha}^{t-1} - \hat{\boldsymbol{\alpha}}^{t-1} \rangle + \lambda_S(\|\boldsymbol{\alpha}^{t-1}\|_1 - \|\hat{\boldsymbol{\alpha}}^{t-1}\|_1)\,, \tag{46}$$

such that

$$\hat{\boldsymbol{\alpha}}^{t-1} := \arg\min_{\boldsymbol{\alpha}} \left(\langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \boldsymbol{\alpha} \rangle + \lambda_S \|\boldsymbol{\alpha}\|_1 \right) \text{ s.t. } \|\boldsymbol{\alpha}\|_1 \leq Q^{t-1}\,. \tag{47}$$

For the second block on $(\boldsymbol{\Theta}, R)$, Section C.2 shows that

$$F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) \leq F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - \frac{(g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t))^2}{\max\{2R_{\mathsf{UB}}^t(\lambda_L + M^t), 8\sigma_{\boldsymbol{\Theta}}(R_{\mathsf{UB}}^t)^2\}}\,, \tag{48}$$

where $M^t := \|\nabla_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})\|_2$ and we recall that $R^t_{\mathsf{UB}} := \lambda_L^{-1} F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1})$ and we have defined

$$g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R^t_{\mathsf{UB}}) := \langle \boldsymbol{\Theta}^{t-1} - \hat{\boldsymbol{\Theta}}^t, \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \rangle + \lambda_L (R^{t-1} - \hat{R}^t) \; . \tag{49}$$

Moreover, Section C.2 shows that

$$F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) - F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq -\frac{\sigma_{\boldsymbol{\Theta}}}{2} \|\mathcal{P}_\Omega(\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1})\|_F^2 \; . \tag{50}$$

Statement (i). The above results show that the objective values for the iterates produced by the MCGD method are non-increasing, *i.e.,*

$$F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) \leq F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \tag{51}$$

Now, consider the time varying part in the quantity $C(t)$ [cf. (18)] — Q^t, R^t_{UB}, M^t. The first two quantities are defined from the objective values and are thus bounded by $\lambda_S^{-1} F(\boldsymbol{\alpha}^0, \boldsymbol{\Theta}^0, R^0)$, $\lambda_L^{-1} F(\boldsymbol{\alpha}^0, \boldsymbol{\Theta}^0, R^0)$, respectively. Moreover, from the monotonicity of $F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1})$, we have $\lambda_L \|\boldsymbol{\Theta}^{t-1}\|_\star + \lambda_S \|\boldsymbol{\alpha}^t\|_1 \leq F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq F(\boldsymbol{\alpha}^0, \boldsymbol{\Theta}^0, R^0)$ for all $t \geq 1$. As the gradient $\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is bounded whenever $\boldsymbol{\alpha}, \boldsymbol{\Theta}$ are bounded, we conclude that M^t is bounded, e.g., $M^t \leq \bar{M}$. Finally, this shows for all $t \geq 1$ that

$$C(t) \leq \overline{C} := \max\left\{ \frac{24(Q^0)^2}{\gamma}, \frac{24\hat{\sigma}^2_{\boldsymbol{\Theta}}(Q^0)^2}{\sigma_{\boldsymbol{\Theta}}} + \max\{6R^0_{\mathsf{UB}}(\lambda_L + \bar{M}), 24\sigma_{\boldsymbol{\Theta}}(R^0_{\mathsf{UB}})^2\} \right\} . \tag{52}$$

Statement (ii). To characterize the convergence rate of the MCGD method, let us consider the Lyapunov function, $g^t(Q^t, R^t_{\mathsf{UB}})$, defined as:

$$g^t(Q^t, R^t_{\mathsf{UB}}) := g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}; Q^t) + g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R^t_{\mathsf{UB}}) \; . \tag{53}$$

Note that as the loss function $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ is convex and $\|\hat{\boldsymbol{\alpha}}\|_1 \leq Q^t$, $\|\hat{\boldsymbol{\Theta}}\|_\star \leq R^t_{\mathsf{UB}}$, it is possible to lower bound $g^t(Q^t, R^t_{\mathsf{UB}})$ by:

$$g^t(Q^t, R^t_{\mathsf{UB}}) \geq F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R}) \; . \tag{54}$$

Furthermore, we can obtain an upper bound to $g^t(Q^t, R^t_{\mathsf{UB}})$ in terms of the objective values:

$$\begin{aligned}
&g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}; Q^t) = \max_{\|\boldsymbol{\alpha}\| \leq Q^t} \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}), \boldsymbol{\alpha}^t - \boldsymbol{\alpha} \rangle + \lambda_S(\|\boldsymbol{\alpha}^t\|_1 - \|\boldsymbol{\alpha}\|_1) \\
&= \max_{\|\boldsymbol{\alpha}\|_1 \leq Q^t} \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t), \boldsymbol{\alpha}^t - \boldsymbol{\alpha} \rangle + \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) - \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t), \boldsymbol{\alpha}^t - \boldsymbol{\alpha} \rangle \\
&\qquad + \lambda_S\big(\|\boldsymbol{\alpha}^t\|_1 - \|\boldsymbol{\alpha}\|_1\big) \\
&\leq \max_{\|\boldsymbol{\alpha}\|_1 \leq Q^t} \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t), \boldsymbol{\alpha}^t - \boldsymbol{\alpha} \rangle + \lambda_S\big(\|\boldsymbol{\alpha}^t\|_1 - \|\boldsymbol{\alpha}\|_1\big) \\
&\qquad + \|\nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) - \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t)\|_2 \|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}\|_2 \\
&\leq g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t; Q^t) + 2\hat{\sigma}_{\boldsymbol{\Theta}} Q^t \|\mathcal{P}_\Omega(\boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t)\|_F \; .
\end{aligned} \tag{55}$$

Consequently, we have

$$\begin{aligned}
&\big(g^t(Q^t, R^t_{\mathsf{UB}})\big)^2 \\
&\leq 3\big((g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R^t_{\mathsf{UB}}))^2 + (g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t; Q^t))^2 + 4\hat{\sigma}^2_{\boldsymbol{\Theta}}(Q^t)^2 \|\mathcal{P}_\Omega(\boldsymbol{\Theta}^{t-1} - \boldsymbol{\Theta}^t)\|_F^2\big) \\
&\leq 3\big(C_1^t\big(F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) - F(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\Theta}^t, R^t)\big) + C_2^t\big(F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t)\big)\big)
\end{aligned}$$

where

$$C_1^t := \frac{8(Q^t)^2}{\gamma}, \;\; C_2^t = \frac{8\hat{\sigma}^2_{\boldsymbol{\Theta}}(Q^t)^2}{\sigma_{\boldsymbol{\Theta}}} + \max\{2R^t_{\mathsf{UB}}(\lambda_L + M), 8\sigma_{\boldsymbol{\Theta}}(R^t_{\mathsf{UB}})^2\} \tag{56}$$

Observe that $C(t)$ is defined by $C(t) = 3\max\{C_1^t, C_2^t\}$ as the upper bound of the above constants, we get

$$\big(g^t(Q^t, R^t_{\mathsf{UB}})\big)^2 \leq C(t)\big(F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\Theta}^t, R^t)\big) \; . \tag{57}$$

Using the shorthand notation $\Delta^t := F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R})$ and notice that $\big(g^t(Q^t, R^t_{\mathsf{UB}})\big)^2 \geq (\Delta^t)^2$, we arrive at the following inequality:

$$\Delta^{t+1} \leq \Delta^t - \frac{1}{C(t)}(\Delta^t)^2 \tag{58}$$

Applying Lemma 8 in Section C.3, we can show that

$$\Delta^{t+1} \leq \frac{1}{(\Delta^1)^{-1} + \sum_{i=1}^t \frac{1}{C(i)}}\,, \tag{59}$$

Note that $\Delta^1 \leq \tilde{\Delta}^0 := F(\boldsymbol{\alpha}^0, \boldsymbol{\Theta}^0, R^0) - F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R})$, we have

$$\Delta^{t+1} \leq \frac{1}{(\tilde{\Delta}^0)^{-1} + \sum_{i=1}^t \frac{1}{C(i)}} \leq \frac{1}{(\tilde{\Delta}^0)^{-1} + t\overline{C}(t)},\ \forall\, t \geq 0\,. \tag{60}$$

The proof is concluded by the straightforward inequality $F_0(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\Theta}^{t+1}) - F_0(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}) \leq F(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\Theta}^{t+1}, R^{t+1}) - F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R}) \leq \Delta^{t+1}$.

Comment on $\lim_{t\to\infty} C(t)$. Since both $F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t)$ and $F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1})$ converge to $F^\star := F(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}, \hat{R})$, *i.e.,* the optimal objective value. It is clear that $Q^t \to \hat{Q} := \lambda_S^{-1} F^\star$ and $R^t_{\mathsf{UB}} \to \hat{R}_{\mathsf{UB}} := \lambda_L^{-1} F^\star$ as well. Furthermore, by continuity of the gradient, we have $M^t \to \|\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}})\|_2$. This shows that the limit $C^\star = \lim_{t\to\infty} C(t)$ exists.

To obtain a computable bound for $C^\star$, note that $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}})$ is also an optimal solution to (5) and the optimality condition shows that

$$\mathbf{0} \in \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}}) + \lambda_L \partial \|\hat{\boldsymbol{\Theta}}\|_\star \tag{61}$$

By [31, P. 41], we know that $\partial \|\hat{\boldsymbol{\Theta}}\|_\star = \{\boldsymbol{U}_1 \boldsymbol{V}_1^\top + \boldsymbol{W} \ :\ \|\boldsymbol{W}\|_2 \leq 1,\ \boldsymbol{U}_1^\top \boldsymbol{W} = \mathbf{0},\ \boldsymbol{W}\boldsymbol{V}_1 = \mathbf{0}\}$ such that $\boldsymbol{U}_1 \in \mathbb{R}^{m_1 \times r}, \boldsymbol{V}_1 \in \mathbb{R}^{m_2 \times r}$ are the left/right singular vectors of $\hat{\boldsymbol{\Theta}}$ corresponding the $r := \mathrm{rank}(\hat{\boldsymbol{\Theta}})$ non-zero singular values of $\hat{\boldsymbol{\Theta}}$. Importantly, this implies that $\|\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\Theta}})\|_2 \leq 2\lambda_L$ and

$$C^\star \leq \overline{C}^\star := \max\left\{\frac{24(\hat{Q})^2}{\gamma}, \frac{24\hat{\sigma}^2_{\boldsymbol{\Theta}}(\hat{Q})^2}{\sigma_{\boldsymbol{\Theta}}} + \max\{18\hat{R}_{\mathsf{UB}}\lambda_L, 24\sigma_{\boldsymbol{\Theta}}(\hat{R}_{\mathsf{UB}})^2\}\right\}. \tag{62}$$

C.1 Proof of Eq. (45)

Suppose $\boldsymbol{\alpha}^t$ is obtained by the proximal update in (10), we observe that

$$\begin{aligned} F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq\ & F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) + \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t-1}\rangle \\ & + \frac{\sigma_{\boldsymbol{\alpha}}}{2}\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t-1}\|_2^2 + \lambda_S\big(\|\boldsymbol{\alpha}^t\|_1 - \|\boldsymbol{\alpha}^{t-1}\|_1\big)\,. \end{aligned} \tag{63}$$

On the other hand, when $\boldsymbol{\alpha}^t$ is obtained by the exact minimization in (11), denoted by $\boldsymbol{\alpha}^t_{\mathsf{exact}}$ to avoid confusion, we have $F(\boldsymbol{\alpha}^t_{\mathsf{exact}}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1})$ since the latter is an exact minimizer. Thus, $F(\boldsymbol{\alpha}^t_{\mathsf{exact}}, \boldsymbol{\Theta}^{t-1}, R^{t-1})$ is upper bounded by the right hand side in the above inequality.

Using the property of the proximal operator, it can be shown that

$$\boldsymbol{\alpha}^t \in \arg\min_{\boldsymbol{\alpha}} \Big(\langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}\rangle + \frac{1}{2\gamma}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}\|_2^2 + \lambda_S(\|\boldsymbol{\alpha}\|_1 - \|\boldsymbol{\alpha}^{t-1}\|_1)\Big) \tag{64}$$

Due to our choice of step size, we have $\sigma_{\boldsymbol{\alpha}} \leq 1/\gamma$. Combining this with the above inequality implies that

$$\begin{aligned} F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \leq\ & F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) + \langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}\rangle \\ & + \frac{1}{2\gamma}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^{t-1}\|_2^2 + \lambda_S\big(\|\boldsymbol{\alpha}\|_1 - \|\boldsymbol{\alpha}^{t-1}\|_1\big),\ \forall\, \boldsymbol{\alpha} \in \mathbb{R}^K\,. \end{aligned} \tag{65}$$

Furthermore, for all $b \in \mathbb{R}$ it holds that

$$\begin{aligned} F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) &\leq F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) + b\langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \hat{\boldsymbol{\alpha}}^{t-1} - \boldsymbol{\alpha}^{t-1} \rangle \\ &\quad + \frac{b^2}{2\gamma} \|\hat{\boldsymbol{\alpha}}^{t-1} - \boldsymbol{\alpha}^{t-1}\|_2^2 + \lambda_S \big(\|b\hat{\boldsymbol{\alpha}}^{t-1} + (1-b)\boldsymbol{\alpha}^{t-1}\|_1 - \|\boldsymbol{\alpha}^{t-1}\|_1 \big) \\ &\leq F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) + b\langle \nabla_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}), \hat{\boldsymbol{\alpha}}^{t-1} - \boldsymbol{\alpha}^{t-1} \rangle \\ &\quad + \frac{b^2}{2\gamma} \|\hat{\boldsymbol{\alpha}}^{t-1} - \boldsymbol{\alpha}^{t-1}\|_2^2 + b\lambda_S \big(\|\hat{\boldsymbol{\alpha}}^{t-1}\|_1 - \|\boldsymbol{\alpha}^{t-1}\|_1 \big) \,, \end{aligned} \tag{66}$$

where we have limited our search space from $\boldsymbol{\alpha} \in \mathbb{R}^K$ to $\boldsymbol{\alpha} = b\hat{\boldsymbol{\alpha}}^t + (1-b)\boldsymbol{\alpha}^{t-1}$ for $b \in \mathbb{R}$. Minimizing the right hand side of the above with respect to b yields

$$\begin{aligned} &F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \\ &\leq -\frac{\gamma}{2} \frac{(g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}; Q^{t-1}))^2}{\|\hat{\boldsymbol{\alpha}}^{t-1} - \boldsymbol{\alpha}^{t-1}\|_2^2} \leq -\frac{\gamma}{2} \frac{(g_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}; Q^{t-1}))^2}{(2Q^{t-1})^2} \,, \end{aligned} \tag{67}$$

where we have used $\|\hat{\boldsymbol{\alpha}}^{t-1} - \boldsymbol{\alpha}^{t-1}\|_2^2 \leq (2Q^{t-1})^2$ in the last inequality.

C.2 Proof of Eq. (48) and (50)

Let us observe that

$$\begin{aligned} F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) &= F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - \beta_t g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) \\ &\quad + \frac{\beta_t^2}{2} \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1}) \\ \hat{R}^t - R^{t-1} \end{pmatrix}^\top \nabla^2_{\boldsymbol{\Theta},R}(\boldsymbol{\xi}) \begin{pmatrix} \mathrm{vec}(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1}) \\ \hat{R}^t - R^{t-1} \end{pmatrix} \,, \end{aligned} \tag{68}$$

where $\boldsymbol{\xi}$ is any point that lies on the line $[(\mathrm{vec}(\boldsymbol{\Theta}^{t-1}); R^{t-1}), (\mathrm{vec}(\boldsymbol{\Theta}^t); R^t)]$. From the property of F, we observe that

$$\nabla^2_{\boldsymbol{\Theta},R}(\boldsymbol{\xi}) \preceq \begin{pmatrix} \sigma_{\boldsymbol{\Theta}} \mathrm{Diag}(\mathcal{P}_\Omega(\mathbf{J})) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \,, \tag{69}$$

where $\mathbf{J}$ is the $m_1 \times m_2$ all-ones matrix. The above implies that

$$\begin{aligned} F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) &\leq F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - \beta_t g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) \\ &\quad + \frac{\beta_t^2 \sigma_{\boldsymbol{\Theta}}}{2} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2 \,. \end{aligned} \tag{70}$$

Recall that $\beta_t = \min\{1, g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t)/(\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2)\}$. If $g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) \geq \sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2$, then we choose $\beta_t = 1$ and observe:

$$\begin{aligned} &F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) - F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \\ &\leq -\frac{1}{2} g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) = -\frac{1}{2} \frac{(g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t))^2}{g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t)} \\ &\leq -\frac{1}{2} \frac{(g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t))^2}{R_{\mathsf{UB}}^t(\lambda_L + 2M^t)} \,, \end{aligned} \tag{71}$$

where we have used the upper bound to $g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t)$ as follows:

$$\begin{aligned} g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) &\leq \lambda_L R_{\mathsf{UB}}^t + \langle \boldsymbol{\Theta}^{t-1} - \hat{\boldsymbol{\Theta}}^t, \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \rangle \\ &\leq R_{\mathsf{UB}}^t \big(\lambda_L + 2M^t \big) \,, \end{aligned} \tag{72}$$

with $M^t := \|\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})\|_2$ being the spectral norm of the gradient.

Otherwise, we choose $\beta_t = g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t)/(\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2)$ and observe:

$$\begin{aligned} &F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) - F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) \\ &\leq -\frac{1}{2} \frac{(g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t))^2}{\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2} \leq -\frac{1}{2} \frac{(g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t))^2}{\sigma_{\boldsymbol{\Theta}} (2R_{\mathsf{UB}}^t)^2} \,, \end{aligned} \tag{73}$$

where we have used $\|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2 \leq \|\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1}\|_F^2 \leq \|\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1}\|_\star^2 \leq (2R_{\mathsf{UB}}^t)^2$.

To prove (50), we observe that

$$\|\mathcal{P}_\Omega(\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1})\|_F^2 = \beta_t^2 \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2 \,. \tag{74}$$

If $\beta_t = 1$, then we have $g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) \geq \sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2$ and therefore we can upper bound $\|\mathcal{P}_\Omega(\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1})\|_F^2$ by:

$$\frac{1}{\sigma_{\boldsymbol{\Theta}}} g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t) \leq \frac{2}{\sigma_{\boldsymbol{\Theta}}} \Big(F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) \Big) \tag{75}$$

where the last inequality follows from (72). Otherwise, we choose $\beta_t = g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t)/\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2$ and therefore,

$$\begin{aligned} \|\mathcal{P}_\Omega(\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^{t-1})\|_F^2 &= \frac{1}{\sigma_{\boldsymbol{\Theta}}} \frac{(g_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R_{\mathsf{UB}}^t))^2}{\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_\Omega(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2} \\ &\leq \frac{2}{\sigma_{\boldsymbol{\Theta}}} \Big(F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}) - F(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t) \Big) \,, \end{aligned} \tag{76}$$

where the last inequality follows from (73).

C.3 Additional Lemma

The following lemma is modified from [3, Lemma 3.5].

Lemma 8 *Let $\{A_k\}_{k\geq 1}$ be a non-negative sequence satisfying:*

$$A_{k+1} \leq A_k - \gamma_k A_k^2, \; k \geq 1 \,, \tag{77}$$

where γ_k is some positive number for all $k \geq 1$. Then,

$$A_{k+1} \leq \frac{1}{\frac{1}{A_1} + \sum_{i=1}^k \gamma_i}, \; k \geq 1 \,. \tag{78}$$

Proof: Consider the following chain of inequality:

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \gamma_k \frac{A_k}{A_{k+1}} \geq \gamma_k \,, \tag{79}$$

where the last inequality is due to the fact that $A_{k+1} \leq A_k$. Consequently, we have

$$\frac{1}{A_{k+1}} - \frac{1}{A_1} = \sum_{i=1}^k \Big(\frac{1}{A_{i+1}} - \frac{1}{A_i} \Big) \geq \sum_{i=1}^k \gamma_i \,. \tag{80}$$

Reshuffling terms shows the desired result in (78). **Q.E.D.**

D Distributed MCGD Optimization

Similar to the previous section, in the following we shall denote the iterates without the bracket in the superscripts, e.g., we denote $\boldsymbol{\alpha}^{(t)}, \boldsymbol{\Theta}^{(t)}, R^{(t)}$ by $\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^t, R^t$, respectively, to simplify our discussions.

Let us describe a distributed version of the MCGD method under a *master-slave* architecture setting where there exists K workers and each of them is connected to a central server. Our goal is to offload the computation required by MCGD method to the workers, while protecting the privacy sensitive data owned by the workers. To describe our setting, the set of observed data $\mathbf{Y}_{ij}$, $(i,j) \in \Omega$ are stored in K different workers, where the kth worker holds $\mathbf{Y}_{ij}$ with $(i,j) \in \Omega_k \subset \Omega$. Particularly, we have $\Omega = \Omega_1 \cup \cdots \cup \Omega_K$ with $\Omega_k \cap \Omega_{k'} = \emptyset$ for all $k \neq k'$. In this way, we can write

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta}) = \sum_{(i,j)\in\Omega} \{-\mathbf{Y}_{ij}\mathbf{M}_{ij} + g_j(\mathbf{M}_{ij})\} = \sum_{k=1}^K \underbrace{\sum_{(i,j)\in\Omega_k} \{-\mathbf{Y}_{ij}\mathbf{M}_{ij} + g_j(\mathbf{M}_{ij})\}}_{:=\mathcal{L}_k(\boldsymbol{\alpha}, \boldsymbol{\Theta})} \tag{81}$$

such that the log-likelihood function can be decomposed as $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\Theta}) := \sum_{k=1}^{K} \mathcal{L}_k(\boldsymbol{\alpha}, \boldsymbol{\Theta})$. Moreover, notice that $\mathcal{L}_k(\boldsymbol{\alpha}, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta})) = \mathcal{L}_k(\boldsymbol{\alpha}, \boldsymbol{\Theta})$ since the kth local function is evaluated only on the entries in Ω_k. For simplicity, we assume that computation can be done synchronously among the workers.

We can implement the MCGD method in a distributed setting as follows. We focus on the tth iteration where $\boldsymbol{\alpha}^{t-1}, \boldsymbol{\Theta}^{t-1}, R^{t-1}$ have been previously computed and worker k now holds $\boldsymbol{\alpha}^{t-1}, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1}), R^{t-1}$.

Firstly, the proximal update step of line 3 is replaced by a natural distributed implementation where the workers compute and transmit the local gradients of the log-likelihood function, $\nabla_{\boldsymbol{\alpha}} \mathcal{L}_k(\boldsymbol{\alpha}^{t-1}, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1}))$, to the master node; the master node can then *aggregate* the received local gradients to form the update in (10), yielding $\boldsymbol{\alpha}^t$ which is then transmitted back to the workers.

Secondly, the CG update of line 5 requires the top SVD of $\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})$ whose complexity is $\mathcal{O}(|\Omega| \max\{n, p\} \log(1/\delta))$ using a centralized implementation, where $\delta > 0$ is the desired accuracy of SVD. In a distributed setting, we can replace the step by a *distributed power method* for offloading the complexity. Importantly, we observe that the top singular vectors of $\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})$ can be *approximated* by the following power method recursions:

Algorithm 2 Distributed Power Method for MCGD.

1: **Initialize:** initialization — $\boldsymbol{u}(0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \in \mathbb{R}^n$, and the parameter $P \in \mathbb{Z}$.
2: **for** $p = 1, 2, \ldots, P$ **do**
3: The central server sends the vector $\boldsymbol{u}(p-1)$ to workers.
4: For all k, worker k computes the vector:

$$\boldsymbol{v}_k(p) = \nabla_{\boldsymbol{\Theta}} \mathcal{L}_k(\boldsymbol{\alpha}^t, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1})) \boldsymbol{u}(p-1) \tag{82}$$

and transmit it to the central server.
5: The central server forms the next iterate by $\boldsymbol{v}(p) = \sum_{k=1}^{K} \boldsymbol{v}_k(p)$ and sends the vector $\boldsymbol{v}(p)$ to workers.
6: For all k, worker k computes the vector:

$$\boldsymbol{u}_k(p) = \nabla_{\boldsymbol{\Theta}} \mathcal{L}_k(\boldsymbol{\alpha}^t, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1}))^\top \boldsymbol{v}(p) \tag{83}$$

and transmit it to the central server.
7: The central server forms the next iterate by $\boldsymbol{u}(p) = \sum_{k=1}^{K} \boldsymbol{u}_k(p)$.
8: **end for**
9: At the central server, compute the top left and right singular vector as $\boldsymbol{u}_{(1)}^t = \boldsymbol{u}(P)/\|\boldsymbol{u}(P)\|$ and $\boldsymbol{v}_{(1)}^t = \boldsymbol{v}(P)/\|\boldsymbol{v}(P)\|$.
10: **Return:** the top singular vectors $\boldsymbol{u}_{(1)}^t, \boldsymbol{v}_{(1)}^t$.

Line 4 and 5 in the above pseudo code implement the following power iterations:

$$\boldsymbol{v}(p) = \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \boldsymbol{u}(p-1) = \sum_{k=1}^{K} \nabla_{\boldsymbol{\Theta}} \mathcal{L}_k(\boldsymbol{\alpha}^t, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1})) \boldsymbol{u}(p-1) \tag{84}$$

$$\boldsymbol{u}(p) = \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})^\top \boldsymbol{v}(p) = \sum_{k=1}^{K} \nabla_{\boldsymbol{\Theta}} \mathcal{L}_k(\boldsymbol{\alpha}^t, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1}))^\top \boldsymbol{v}(p) \ , \tag{85}$$

where we have exploited the decomposable structure of the log-likelihood function in the distributed setting. Upon computing $\boldsymbol{u}_{(1)}^t, \boldsymbol{v}_{(1)}^t$, we can estimate the top singular value by $(\boldsymbol{v}_{(1)}^t)^\top \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \, \boldsymbol{u}_{(1)}^t$ which can also be computed distributively using similar scheme as in the above. Consequently, the update direction $(\hat{\boldsymbol{\Theta}}^t, \hat{R}^t)$ can be computed at the central server using

$$(\hat{\boldsymbol{\Theta}}^t, \hat{R}^t) = \begin{cases} (\boldsymbol{0}, 0), & \text{if } \lambda_L \geq (\boldsymbol{v}_{(1)}^t)^\top \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \, \boldsymbol{u}_{(1)}^t \ , \\ (-R_{\mathsf{UB}}^t \boldsymbol{u}_{(1)}^t (\boldsymbol{v}_{(1)}^t)^\top, R_{\mathsf{UB}}^t), & \text{if } \lambda_L < (\boldsymbol{v}_{(1)}^t)^\top \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \, \boldsymbol{u}_{(1)}^t \ . \end{cases} \tag{86}$$

Lastly, to compute the step size β_t required in line 6, an efficient way is to observe the following decomposition of the inner product:

$$\langle \boldsymbol{\Theta}^{t-1} - \hat{\boldsymbol{\Theta}}^t, \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \rangle = \sum_{k=1}^{K} \langle \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1} - \hat{\boldsymbol{\Theta}}^t), \nabla_{\boldsymbol{\Theta}} \mathcal{L}_k(\boldsymbol{\alpha}^t, \mathcal{P}_{\Omega_k}(\boldsymbol{\Theta}^{t-1})) \rangle \ . \tag{87}$$

This implies that the inner product on the left hand side can be computed by aggregating the K terms on the right hand side, where each of the K terms can be computed at the kth worker once $\mathcal{P}_{\Omega_k}(\hat{\boldsymbol{\Theta}}^t)$ is available. As such, the central server also sends $\mathcal{P}_{\Omega_k}(\hat{\boldsymbol{\Theta}}^t)$ to the workers after (86). Consequently, the step size is given by:

$$\beta_t = \min \left\{ 1, \frac{(\widehat{g}_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R^t_{\mathsf{UB}}))_+}{\sigma_{\boldsymbol{\Theta}} \|\mathcal{P}_{\Omega}(\hat{\boldsymbol{\Theta}}^t - \boldsymbol{\Theta}^{t-1})\|_F^2} \right\} , \tag{88}$$

where

$$\widehat{g}_{\boldsymbol{\Theta}}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}, R^{t-1}; R^t_{\mathsf{UB}}) := \langle \boldsymbol{\Theta}^{t-1} - \hat{\boldsymbol{\Theta}}^t, \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \rangle + \lambda_L \big(R^{t-1} - \hat{R}^t \big) \ . \tag{89}$$

Note that unlike the function $g_{\boldsymbol{\Theta}}(\cdot)$ defined in (49), the function $\widehat{g}_{\boldsymbol{\Theta}}(\cdot)$ can be negative since the matrix $\hat{\boldsymbol{\Theta}}^t$ herein is computed from an inexact pair of top singular vectors.

Several remarks are in order. Throughout the optimization, the central server is unaware of the local gradient matrix *w.r.t.* $\boldsymbol{\Theta}$, instead only its corresponding matrix-vector products are transmitted from the workers to the server. In this way, the privacy-sensitive data from the workers will not be revealed to the server.

For any $\delta > 0$, it is well known that in high probability (with respect to the random initialization), the power method in Algorithm 2 converges [13] to an δ-accurate top SVD solution in $P = \mathcal{O}(\log(1/\delta))$ steps[1]. Therefore, for the distributed MCGD method, the overall complexity required per iteration is $\mathcal{O}(|\Xi| + \max\{n, p\} \log(1/\delta))$ at the central server, and it is $\mathcal{O}(|\Omega_k| \max\{n, p\} \log(1/\delta))$ for the kth worker. The overall complexity is lower than a centralized implementation especially when $|\Omega_k| \ll |\Omega|$, e.g., when the number of workers increases.

[1]For example, an δ-accurate top SVD solution satisfies

$$\left| (\boldsymbol{u}^t_{(1)})^\top \nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1}) \boldsymbol{v}^t_{(1)} - \sigma_1(\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})) \right| \leq \delta \ . \tag{90}$$

In the complexity measure, we have hidden the dependency on the spectral gap $\Delta := \sigma_2(\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})) / \sigma_1(\nabla_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\alpha}^t, \boldsymbol{\Theta}^{t-1})) \leq 1$ in the big-O notation.