École Polytechnique Fédérale de Lausanne

Robust Multi-Task Learning

Master's Semester Project

Author:
Luca Rossi

Professors:
Nicolas Flammarion
Nicolas Boumal
Supervisor:
Etienne Boursier

January 6, 2023

Abstract

Multi-Task Learning (MTL) is a Machine Learning paradigm aiming to leverage useful information contained in multiple related tasks to help improve the generalization performance of all the tasks.
In the past few years, Multi-Task Learning has gained popularity due to the possibility to train models, despite very few samples per task. However, there is still a lacking of theoretical understanding of this paradigm. The main hypothesis is that, when a single model is trained jointly on different tasks, a shared representation of the tasks is learned. This representation will allow us to perform better (than training a different model for each task) when the data is scarce, as well as generalize better to unseen tasks. In this work, we investigate this hypothesis in the convenient setting of multi-task linear regression, where the parameters of multiple linear regression models share a common linear, low-dimensional representation except for a few components. We propose to retrieve the parameters solving an optimization problem with low-rank and sparse penalization. We exhibit, under the restricted strong convexity assumption, bounds on the estimation errors. Finally, we show how to solve the problem efficiently and we confirm empirically our results on synthetic datasets.

Contents

1. Introduction

Humans can learn different tasks simultaneously and can apply the knowledge of a task to help the learning process of other related tasks. For instance, learning to play a second instrument is easier once we know how to play a first one. Inspired by such human ability, *Multi-Task Learning* is a learning paradigm that aims to learn multiple related tasks jointly, with the hope that the knowledge of a task can be transferred to others. Multi-Task learning usually outperforms learning each task separately when the number of training samples per task is scarce, and it has been used in a wide variety of applications including natural language processing [Ando and Zhang (2005)], computer vision [Donahue et al. (2013)], and image segmentation [Moeskops et al. (2016)].
Multi-task learning is strongly linked with *Transfer Learning* [Zhuang et al. (2019)]. In Transfer learning, weights from a model trained on one task are used as weight initialization for similar tasks.
Another closely related paradigm named *Meta-Learning* has also been introduced. Its aim is to use the information from some training tasks to perform well on new, unseen tasks. Several algorithms such as MAML [Finn et al. (2017)] and Reptile [Nichol et al. (2018)], have shown incredible empirical success. Therefore, researchers have wondered if the effectiveness of these methods was due to *rapid learning*: the ability to learn faster new representations or *feature reuse*: the efficacy of the representation learned with the training tasks. [Raghu et al. (2019)] found that feature reuse is the dominant factor. However, we have very few algorithms that can learn *provably* from different tasks when we have a small number of samples per task, even in simple settings. Most of them, focus on simple multi-task linear regression and assume that the parameters to learn, share a low-dimensional linear representation [Boursier et al. (2022)], [Tripuraneni et al. (2020)]. However, there is a whole line of work in which the authors assume that the parameter matrix can be decomposed as the sum of low-rank and sparse matrices [Chen et al. (2012)],[Chen et al. (2011)]. This allows us to model the tasks' parameters as lying almost on a low-dimensional subspace a part from a few components that are allowed to vary. This is in line with the idea of solving a robust version of the low-dimensional shared representation problem. Our work aims to show theoretical guarantees for such a simple decomposable model, always in a multi-task linear regression setting. Borrowing the theory from [Negahban et al. (2012)] and [Agarwal et al. (2012)], we will show bounds in norm between our estimates and the true parameters and we will confirm our results empirically on synthetic datasets.

1.1 Notations

For the reader's convenience, we summarize here some of the notation that we will use throughout the report. For a matrix $A \in \mathbb{R}^{d\times T}$ we will denote by $||A||_F := \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{T} A_{ij}^2}$

the Frobenius norm of the matrix. Given $A, B \in \mathbb{R}^{d\times T}$, we denote by $\langle A, B\rangle \coloneqq \sum_{i=1}^{d}\sum_{j=1}^{T} A_{ij}B_{ij}$ the Frobenius inner product between two matrices. We denote the singular values of A by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_h(A)$ with $h = \min\{d, T\}$. We denote by $||A||_* \coloneqq \sum_{i=1}^{h}\sigma_i(A)$ the nuclear norm of A, by $||A||_1 \coloneqq \sum_{i=1}^{d}\sum_{j=1}^{T}|A_{ij}|$ the 1-norm of a matrix and by $||A||_{\text{op}} \coloneqq \sup_{v\in\mathbb{R}^T: v\neq 0}\frac{||Av||_2}{||v||_2}$ the operator norm of A. The $L_{p,q}$ norm of A as $||A||_{p,q} = \left(\sum_{j=1}^{T}\left(\sum_{i=1}^{d}|A_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}$. Moreover, we will indicate by A_i the i-th column of A and by a_i the i-th row of A. If we suppose that A has rank r, the SVD of A is such that $A = \tilde{U}\Sigma\tilde{V}^T$, we can define the Burer-Monteiro factorization of A as $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{T\times r}$ such that $A = UV^T$. The factorization is not unique. However, if we define Σ_{tr} the matrix in $\mathbb{R}^{r\times r}$ corresponding to the non zero diagonal entries of Σ (submatrix of non-zero singular values), $\tilde{U}_{\text{tr}}$, the matrix of the corresponding left singular vectors. We could select U as $\tilde{U}_{\text{tr}}$, similarly $V^T = \Sigma_{\text{tr}}\tilde{V}_{\text{tr}}^T$, with $\tilde{V}_{\text{tr}}$ is the matrix of non-zero right singular vectors. Finally, we will often speak about estimation errors. If we estimate A^* by $\hat{A}$, a bound on the estimation error is an upper-bound on $||\hat{A} - A^*||_F^2$, when two matrices have to be estimated it will be an upper-bound on $||\hat{A} - A^*||_F^2 + ||\hat{B} - B^*||_F^2$.

1.2 Model

In the following, we consider a multi-task linear regression setting. Let T be the number of tasks and $(X^{(t)}, y^{(t)})$ be the feature matrices and labels for the t-th task, clearly $t \in [T] = \{1, 2, \ldots, T\}$. For simplicity, we assume that each task has the same number of samples n and the dimension space is d. Therefore each matrix $X^{(t)}$ is in $\mathbb{R}^{n\times d}$ and each $y^{(t)}$ in $\mathbb{R}^n$. In simple linear regression, $T = 1$ and we normally assume that each observation y_i is generated by

$$y_i = x_i^T\beta^* + \epsilon_i \quad \text{such that} \quad \epsilon_i \sim \mathcal{N}(0, \nu^2) \quad \text{and} \quad \beta^* \in \mathbb{R}^d$$

The goal is to learn the parameter β^* which we assume to have generated the labels y. Notice that the same problem in vector form can be written as

$$y = X\beta + \epsilon \quad \text{such that} \quad \epsilon \sim \mathcal{N}(0, \nu^2 I)$$

This model has been widely studied in the statistics literature.
Defining $\hat{\beta}^{\text{ls}} \in \arg\min \beta \in \mathbb{R}^d ||y - X\beta||_2^2$, the usual least squares estimator, it is easy to show that if X is full column rank, $\hat{\beta}^{\text{ls}} = (X^TX)^{-1}X^Ty$. Additionally, It can be shown that if $d < n$ and $B = \frac{X^TX}{n}$ has rank r, then we have that with probability at least $1 - \delta$

$$||\beta^* - \hat{\beta}^{\text{ls}}||_2^2 \leq \frac{\nu^2}{\lambda_{\min}(B)}\frac{r + \log(\frac{1}{\delta})}{n} \tag{1.1}$$

This bound shows that we have generally a good recovery on the true β^* when the number of samples n is bigger than d. In case the opposite is true: $d > n$ (high dimensional case), we need to make structural assumptions on the true β^* to come to a similar conclusion (e.g. sparsity). We defer to [Rigollet and Hutter (2017)] for a complete discussion on the bound of the estimation error in those settings. In the case of multi-task linear regression, we assume the same for each task:

$$y^{(t)} = X^{(t)}\beta^{*(t)} + \epsilon^{(t)} \quad \text{such that} \quad \epsilon^{(t)} \sim \mathcal{N}(0, \nu^2 I) \quad \forall t \in [T] \tag{1.2}$$

our goal is to estimate the parameters $\beta^{*(t)}$. In multi-task linear regression, the difficulty is two-fold, on the one hand, we are trying to estimate T vectors of parameters instead

of a single one, on the other hand, we should assume that $d > n$ since this is where the multi-task paradigm is more useful. Indeed, in the case $d < n$, we could estimate each parameter $\beta^{*(t)}$ with the ordinary least squares estimator on task t-th and we could have good parameter recovery as shown by equation (1.1).
First of all, notice that in our setting (multi-task linear regression), we can express equation (1.2) in a more compact form as follows:

$$Y = \mathcal{X}(B^*) + W \quad \text{such that} \quad Y, W \in \mathbb{R}^{n \times T}, \; B^* \in \mathbb{R}^{d \times T}$$

Where W is a matrix in which each entry follows an independent (from the other) normal distribution with mean zero and variance ν^2, Y is a matrix whose t-th column is $y^{(t)}$, similarly B^* is a matrix whose t-th column is $\beta^{*(t)}$. Finally $\mathcal{X} : \mathbb{R}^{d \times T} \to \mathbb{R}^{n \times T}$ is a linear operator such that

$$\mathcal{X}(A) = \begin{bmatrix} | & | & | & | \\ X^{(1)}A_1 & X^{(2)}A_2 & \dots & X^{(T)}A_T \\ | & | & | & | \end{bmatrix}$$

where A_i denotes the i-th column of the matrix $A \in \mathbb{R}^{d \times T}$. It is therefore natural to try to estimate B as the solution to the following optimization problem.

$$\hat{B} = \underset{B \in \mathbb{R}^{d \times T}}{\arg\min} \sum_{i \in [n]} \sum_{t \in [T]} \left(y_i^t - x_i^{tT} \beta^{*(t)} \right)^2 = \underset{B \in \mathbb{R}^{d \times T}}{\arg\min} ||Y - \mathcal{X}(B)||_F^2$$

where x_i^t is the i-th row of the feature matrix $X^{(t)}$ and we assumed to use the squared error loss.
As in the case of a single task, we need to estimate B^* but it is convenient to make some structural assumptions on the latter. A natural assumption is that the matrix B^* lies on a low dimensional subspace i.e. it is low rank. Indeed, imagine the extreme case in which all the tasks are the same, in this case, $\beta^{*(t_1)} = \beta^{*(t_2)}$ for each couple of tasks (t_1, t_2), and the rank of the matrix B^* would be one. Relaxing this case, we can imagine some tasks to be related to each other and therefore the rank of the matrix B^* to be small. Geometrically, this is equivalent to assuming that the parameters lie on a low-dimensional hyper-plane spanned by the basis of the columns of B^*, in formulae $B_i^* = U^* v_i^*$ with $U \in \mathbb{R}^{d \times r}$ (r is the dimension of the subspace). This model has already been studied in the literature and bounds on the estimation errors have been proved among others by [Tripuraneni et al. (2020)], [Boursier et al. (2022)], [Rohde and Tsybakov (2011)]. However, we are interested in a similar but slightly different scenario. We imagine that B^* is the sum of two matrices: $B^* = \Theta^* + \Gamma^*$ with Θ^* a low-rank matrix and Γ^* an entry-wise or column-wise sparse one. The structure allows us to solve a more robust version of the original problem. In the entry-wise sparse version, two regression parameters $\beta^{*(t_1)}$ and $\beta^{*(t_2)}$ are not necessarily on a smaller linear subspace but are allowed to lie almost entirely on the subspace a part from a few components, which can be seen as components that are automatically fine-tuned by Γ^*. In the column-wise sparse setting, most of the parameters lie on the low-dimensional subspace but we still allow them to be outside this subspace thanks to the additional matrix Γ^*. Indeed, it is reasonable to think that there might be some outlier tasks that are not related to the previous. In this case, the expressive power of the model comes at hand and the corresponding columns of Γ^* will not be zero. These structures have been proven successful also on real datasets as shown by [Chen et al. (2012)] and [Adeli (2017)]. Therefore, with the same notations, our assumption is that

$$\begin{aligned} &Y = \mathcal{X}(\Theta^* + \Gamma^*) + W \\ &\text{s.t.} \quad \text{rank}(\Theta^*) \leq r < T \quad \text{and} \quad \text{Sp}(\Gamma^*) \leq s \end{aligned} \tag{1.3}$$

And $\text{Sp}(\Gamma^*)$ denotes the number of non-zero entries, in the case of entry-wise sparsity, or columns, in the case of column sparsity of the matrix Γ^*.

1.3 Related work

Multi-task linear regression is one of the easiest, theoretically tractable problems in multi-task learning. The section aims to provide an overview of the related literature with the hope of further motive our work.
One of the first papers which proposed a low-rank structure is the one of [Ando and Zhang (2005)], who supposed that the model's parameters of different tasks share in part a low-rank subspace. In particular $\beta^{*(t)} = c^{(t)} + U^T v^{(t)}$ with $U \in \mathbb{R}^{r\times d}$. A more similar (to our setting) model has been proposed by [Pong et al. (2010)], who proposed B^* as a low-rank matrix and propose to enforce the low-rank structure using the nuclear norm penalization. In the simpler setting in which B^* was low-rank only, a first theoretical study has been carried out by [Rohde and Tsybakov (2011)], under the assumption of restricted isometry property. They again propose to solve the optimization problem with an additional low-rank regularizer. However, the assumption is not totally suitable for Multi-Task learning since it assumes the number of sample n to be bigger than the dimension d, we will discuss more it in the section (2.3). A different approach has been followed by [Tripuraneni et al. (2020)] who made use of the Burer-Monteiro factorization, rewriting the low-rank matrix B as UV^T with $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{T\times r}$. Therefore, they recovered some estimators of U and V as the solution to the optimization problem

$$\min_{U\in\mathbb{R}^{d\times r}, V\in\mathbb{R}^{T\times r}} \sum_{i\in[n]}\sum_{t\in[T]} \left(y_i^t - x_i^{tT}(U(V^T)_t)\right)^2 = ||Y - \mathcal{X}(UV^T)||_F^2 \tag{1.4}$$

As we can see, the low-rank matrix is factorized as $B = UV^T$, with the rank of B determined (equal to) by the dimension r appearing in the matrices U and V. Since the problem 1.4 is not jointly convex in (U, V), we could only aspire to find a local minimum of the problem. However, the authors showed that each local minimum $\hat{B}$ has good guarantees bounding the errors in Frobenius norm between the estimate $\hat{B}$ and the true B^* when the number of samples per task is $\Omega(r^4 \log(T))$. Always [Tripuraneni et al. (2020)], propose also an algorithm that they call the method of moments which has a slightly looser bound but requires a lower number of samples per task ($\Omega(r\log(r))$). Finally, [Boursier et al. (2022)], after proving that a restricted strong convexity assumption hold (refer to (2.8)), could show similar bounds compared to [Tripuraneni et al. (2020)] even when the number of samples per task is $\Omega(1)$ when solving:

$$\hat{B} \in \underset{B\in\mathbb{R}^{d\times T}}{\arg\min} ||Y - \mathcal{X}(B)||_F^2 + \lambda||B||_* \tag{1.5}$$

Notice that the problem 1.5 is quite natural since it is the convex relaxation to the optimization problem

$$\hat{B} \in \underset{B\in\mathbb{R}^{d\times T}}{\arg\min} ||Y - \mathcal{X}(B)||_F^2 + \lambda \underbrace{||\Lambda(B)||_0}_{\text{rank}(B)}$$

If we indicate by $\Lambda(B)$ the vector of singular values of B. Exactly as the 1-norm convexifies the 0-norm, the nuclear norm convexifies the zero norm of the vector of singular values of B. Therefore the matrix $\hat{B}$ is forced to be low-rank when λ is large.
However, there is a whole line of work in which the parameter matrix B^* is assumed to be the sum of two or more matrices. [Chen et al. (2012)], assume $B^* = \Theta^* + \Gamma^*$ with Θ^*

low rank and Γ^* entrywise sparse. In a subsequent work [Chen et al. (2011)] assume Γ^* to be column-wise sparse instead of simply sparse. This acts as a way to regularize the low-rank only problem when there are tasks whose parameters don't lie on the low-dimensional subspace. In this case, the corresponding column of Γ^* will not be zero and this structure will still recover the correct parameter. This also acts as a way to identify automatically the outlier tasks: those which don't have the corresponding column of Γ^* equal to zero (since the parameters don't lie on the shared subspace). Note that more complex structures have been proposed [Jalali et al. (2010)], [Jacob et al. (2008)]. However, despite a proven empirical success, there was a general lack of theoretical analysis for the case $B^* = \Theta^* + \Gamma^*$ with Θ^* low-rank and Γ^* sparse. [Agarwal et al. (2012)] show bounds on the estimation error when the feature's matrices $X^{(t)}$ are common to each task: $X^{(t_1)} = X^{(t_2)}$ for all $t_1, t_2 \in [T]$, assuming them to be invertible for the case in which Γ^* is entrywise sparse. More recently, [Pal et al. (2022)] approached the problem using again the Burer-Monteiro factorization and estimated U, V, and B as some local minimum of the following problem, which was solved by alternating minimization.

$$\min_{U \in \mathbb{R}^{d \times r}, V \in \mathbb{R}^{T \times r}, B \in \mathbb{R}^{d \times T}} \sum_{i \in [n]} \sum_{t \in [T]} \left(y_i^t - x_i^{tT} (U V_t + B_t) \right)^2$$
$$\text{s.t. } U^T U = I, \ ||B_i||_0 \leq k \ \ \forall i \in [T]$$

Again they bound the estimation error for each local minimum of the problem. However, their algorithm has only local convergence and many hyperparameters to finetune, making it impractical to use. In the following section, we will show how we can derive, under loose assumptions, bounds on the estimation error borrowing the theory from [Agarwal et al. (2012)].

1.4 Contribution

Our contribution is mainly to further adapt the theory from [Agarwal et al. (2012)] to the multi-task setting. We adapt the corollaries of the main theorem 2.5 to the case in which we have different design matrices for different tasks (which was not the case for [Agarwal et al. (2012)] and very often the case in Multi-Task Learning). Moreover, we provide bounds of estimation errors for the case in which Γ^* is column-wise sparse in corollary 2.7.2. We show how to reduce the cost of solving the optimization problem by adapting and implementing the Frank-Wolfe-Thresholding Algorithm 3. Finally, we complement our work with an empirical analysis of our work on synthetic datasets

2. Bound on the estimation error

2.1 Definition of the problem

As we discussed in the previous section, we are interested in solving the following minimization problem:

$$\min_{\Theta\in\mathbb{R}^{d\times T}:\text{rank}(\Theta)\leq r,\ \Gamma\in\mathbb{R}^{d\times T}:\text{Sp}(\Gamma)\leq s} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2$$

Since the problem is constrained and non-convex, we could turn it into a convex, unconstrained problem by penalizing the objective when Θ is not low-rank and Γ is not sparse. A natural choice could be to use the nuclear norm to enforce the low-rank structure (since it convexifies the zero norm of the singular values) and another regularizer $\mathcal{R}$ (which we could think of as the L_1-norm or the $L_{2,1}$ norm) to enforce the sparse structure. Therefore, the following problem has been proposed as a convex relaxation of the previous one.

$$\min_{\Theta\in\mathbb{R}^{d\times T},\ \Gamma\in\mathbb{R}^{d\times T}} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2+\lambda_d||\Theta||_*+\mu_d\mathcal{R}(\Gamma) \tag{2.1}$$

The optimization problem 2.1 has mainly two advantages:

- It is convex, therefore it is easier to optimize, see section 3 for a complete discussion.
- We don't need to know exactly the rank r and the sparsity of Γ: the two depends now only on the hyperparameters λ_d and μ_d.

However, it turns out that the model 1.3 is unidentifiable if we don't add a further constraint. Indeed, for the sake of example, consider a matrix B^* which is both low-rank and sparse for instance a matrix B^* with a single entry different from zero. In this case, we cannot recover Θ^* and Γ^* univocally. Therefore [Agarwal et al. (2012)] introduce a further constraint to the optimization problem. For the given regularizer $\mathcal{R}$ we define $\kappa_d(\mathcal{R})=\sup_{V\neq 0}\frac{||V||_F}{\mathcal{R}(V)}$ measuring how the Frobenius norm of a matrix varies with respect to the regularizer. Moreover, the associated dual norm of the regularizer is defined as $\mathcal{R}^*(U)\coloneqq\sup_{\mathcal{R}(V)\leq 1}\langle V,U\rangle$. We will define $\varphi_{\mathcal{R}}(\Theta)\coloneqq\kappa_d(\mathcal{R}^*)\mathcal{R}^*(\Theta)$. We will be interested in the family of estimators that have $\varphi_{\mathcal{R}}(\Theta)\leq\alpha$, therefore the problem becomes

$$\min_{\Theta\in\mathbb{R}^{d\times T},\ \Gamma\in\mathbb{R}^{d\times T}} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2+\lambda_d||\Theta||_*+\mu_d\mathcal{R}(\Gamma)\ \text{ s.t. }\ \varphi_{\mathcal{R}}(\Theta)\leq\alpha \tag{2.2}$$

Even if the additional constraint does not seem very intuitive at first sight, in the proofs it will allow us to bound the radius of non-identifiability. To get some familiarity with this condition, we will consider two examples of regularizers that are common in multi-task learning.

Proposition 2.1. *Let $V \in \mathbb{R}^{d\times T}$ if $\mathcal{R}(V) = ||V||_1$, we have that $\kappa_d(\mathcal{R}^*) = \sqrt{dT}$ and $\mathcal{R}^*(V) = ||V||_\infty$.*

Proof. First of all, notice that each matrix $V \in \mathbb{R}^{d\times T}$ can be transformed into a vector in $\mathbb{R}^{dT}$ by the vectorization operator. Rigorously,

$$\text{Vec}(V) = [V_{1,1}, \dots, V_{d,1}, V_{1,2}, \dots, V_{d,2}, \dots, V_{1,T}, \dots, V_{d,T}]^T$$

Let $v = \text{Vec}(V), u = \text{Vec}(U)$, then $\mathcal{R}^*(V) \coloneqq \sup_{\mathcal{R}(U)\leq 1}\langle V, U\rangle = \sup_{||u||_1\leq 1} v^T u$. Notice that

$$\sup_{||u||_1\leq 1} v^T u = \sup_{||u||_1\leq 1} \sum_{i=1}^{dT} |v_i u_i| = \sup_{||u||_1\leq 1} \sum_{i=1}^{dT} |v_i||u_i| \leq \sup_{||u||_1\leq 1} ||v||_\infty \sum_{i=1}^{dT} |u_i| = ||v||_\infty$$

Therefore, we have proved that $\mathcal{R}^*(V) \leq ||v||_\infty$. We will now show that the bound is actually attained. Let i the coordinate such that $||v||_\infty = |v_i|$, let $u = \text{sign}(v_i)e_i$, we have that $v^T u = |v_i| = ||v||_\infty$. Therefore the bound is attained and so $\mathcal{R}^*(V) = ||v||_\infty = ||V||_\infty$. Now, $\kappa_d(\mathcal{R}^*) = \sup_{V\neq 0} \frac{||V||_F}{||V||_\infty} = \sup_{v\neq 0} \frac{||v||_2}{||v||_\infty} = \sqrt{dT}$. Where the last inequality follows from the equivalence between the two and the infinity norm of a vector. □

Proposition 2.2. *Let $V \in \mathbb{R}^{d\times T}$ if $\mathcal{R}(V) = ||V||_{2,1}$, we have that $\kappa_d(\mathcal{R}^*) = \sqrt{T}$ and $\mathcal{R}^*(V) = ||V||_{2,\infty}$.*

Proof. With the same notations as the previous exercise, $\mathcal{R}^*(V) \coloneqq \sup_{||U||_{2,1}\leq 1}\langle V, U\rangle$. Notice that

$$\sup_{||U||_{2,1}\leq 1} \langle V, U\rangle = \sup_{U:\sum_{k=1}^T ||U_k||_2\leq 1} \langle U, V\rangle \leq \sup_{U:\sum_{k=1}^T ||U_k||_2\leq 1} \sum_{k=1}^T \langle U_k, V_k\rangle \leq$$

$$\leq \sup_{U:\sum_{k=1}^T ||U_k||_2\leq 1} \sum_{k=1}^T ||U_k||_2||V_k||_2 \leq ||V||_{2,\infty}$$

Therefore, we have proved that $\mathcal{R}^*(V) \leq ||V||_{2,\infty}$. We will now show that the bound is actually attained. Let i the coordinate such that $||V_i||_2$ is maximized, let

$$U = \begin{bmatrix} | & | & | & | & | & | \\ 0 & 0 & \dots & \frac{V_i}{||V_i||} & 0 & \dots \\ | & | & | & | & | & | \end{bmatrix}$$

We have that $||U||_{2,1} = 1$ and $\langle V, U\rangle = ||V||_{2,\infty}$, therefore $\mathcal{R}^*(V) = ||V||_{2,\infty}$. Finally, $\kappa_d(\mathcal{R}^*) = \sup_{V\neq 0} \frac{||V||_F}{||V||_{2,\infty}} = \sqrt{T}$ since the expression is maximised if $||V_k||_2 = ||V_i||_2$ for all $i \neq j$. □

2.2 Decomposable regularizers

The notion of decomposability of the regularizers has been introduced by [Negahban et al. (2012)] and it is defined in terms of a pair of subspaces. For the sake of simplicity, we will consider a special case of decomposability which will be sufficient for our work.

Definition 2.1. *Given a subspace $\mathbb{M} \subseteq \mathbb{R}^{d\times T}$ and its orthogonal complement $\mathbb{M}^\perp$, a regularizer $\mathcal{R}$ based on the norm is decomposable with respect to the pair $(\mathbb{M}, \mathbb{M}^\perp)$ if*

$$\mathcal{R}(U + V) = \mathcal{R}(U) + \mathcal{R}(V) \quad \forall \quad U \in \mathbb{M},\ V \in \mathbb{M}^\perp$$

Notice that it is always true that $\mathcal{R}(U+V) \leq \mathcal{R}(U)+\mathcal{R}(V)$ by the triangle inequality of the norm. However, equality does not necessarily hold. $\mathbb{M}$ represents the model subspace, i.e. we will choose $\mathbb{M}$ such that Γ^* lies in $\mathbb{M}$. Therefore, the equality in the definition shows that we penalize deviation from $\mathbb{M}$ as much as possible.

Proposition 2.3. *The regularizer $\mathcal{R}(.) = ||.||_1$ is decomposable.*

Proof. Assuming the regularizer is such that $\mathcal{R}(\Gamma) = ||\Gamma||_1$ for $\Gamma \in \mathbb{R}^{d\times T}$. Let S be an arbitrary subspace of the indices: $S \subseteq \{1,\dots,d\} \times \{1,\dots,T\}$, representing the support of the true matrix Γ^*. We define,

$$\mathbb{M}(S) \coloneqq \{\Gamma \in \mathbb{R}^{d\times T} : \Gamma_{ij} = 0 \ \ \forall\ (i,j) \ \notin S\}, \qquad \mathbb{M}^{\perp}(S) = (\mathbb{M}(S))^{\perp}$$

It is verified that for $A \in \mathbb{M}(S),\ B \in \mathbb{M}^{\perp}(S)$, we have $||A+B||_1 = ||A||_1 + ||B||_1$. □

Proposition 2.4. *The regularizer $\mathcal{R}(.) = ||.||_{2,1}$ is decomposable.*

Proof. Assuming the regularizer is such that $\mathcal{R}(\Gamma) = ||\Gamma||_{2,1}$ for $\Gamma \in \mathbb{R}^{d\times T}$. Let S be an arbitrary subspace of the indices of the columns: $S \subseteq \{1,\dots,T\}$, representing the column support of the true matrix Γ^*. We define,

$$\mathbb{M}(S) \coloneqq \{\Gamma \in \mathbb{R}^{d\times T} : \Gamma_{i} = 0 \ \ \forall\ i \ \notin S\}, \qquad \mathbb{M}^{\perp}(S) = (\mathbb{M}(S))^{\perp}$$

It is verified that for $A \in \mathbb{M}(S),\ B \in \mathbb{M}^{\perp}(S)$, we have $||A+B||_{2,1} = ||A||_{2,1} + ||B||_{2,1}$. Indeed,

$$||A+B||_{2,1} = \sum_{k=1}^{T} ||A_k + B_k||_2 = \sum_{k=1}^{T} ||A_k||_2 + \sum_{k=1}^{T} ||B_k||_2 = ||A||_{2,1} + ||B||_{2,1}$$

Where the second to last equality holds because $A \in \mathbb{M}(S)$ and $B \in \mathbb{M}^{\perp}(S)$ □

Notice that many other norm-based regularizers are decomposable, other examples are the nuclear norm (for a slighlty more general definition of decomposable regularizers) and the other group-based norms. For any decomposable regularizer and subspace $\mathbb{M}$ the compatibility constant is defined as

$$\Psi(\mathbb{M},\mathcal{R}) \coloneqq \sup_{U\in\mathbb{M}, U\neq 0} \frac{\mathcal{R}(U)}{||U||_F}$$

The constant measures the compatibility of the regularizer over $\mathbb{M}$ and the Frobenius norm. For example, for $\mathcal{R}$ being the $||.||_1$ and $\mathbb{M}$ being defined as in proposition 2.3, it is not difficult to show that $\Psi(\mathbb{M},\mathcal{R}) = \sqrt{s}$ with s the number of non-zero entries of the true sparse matrix Γ^* (cardinality of the set S in the proof of proposition 2.3). The same holds for $||.||_{2,1}$, with s being the number of non-zero columns of Γ^* (cardinality of the set S in the proof of proposition 2.4). We will refer to s as the sparsity of the matrix Γ^*.

2.3 Restricted Strong Convexity

Given a loss function $\mathcal{L}$, a strong convexity assumption consists of establishing a quadratic lower bound on the error in the first-order Taylor approximation. In formulae, the strong convexity assumption asks:

$$\mathcal{L}(\Omega+\Delta) - \mathcal{L}(\Omega) - \langle \nabla\mathcal{L}(\Omega), \Delta\rangle \geq \frac{\gamma}{2}||\Delta||^2 \ \ \forall\ \Delta,\Omega, \quad \gamma > 0 \tag{2.3}$$

First of all, notice that in our case the assumption is equivalent to

$$\frac{1}{2}||\mathcal{X}(\Delta)||_F^2 \geq \frac{\gamma}{2}||\Delta||_F^2 \quad \forall\ \Delta \tag{2.4}$$

Indeed, simple algebra shows that

$$\begin{aligned}
&\mathcal{L}(\Omega+\Delta) - \mathcal{L}(\Omega) - \langle \nabla\mathcal{L}(\Omega), \Delta\rangle = \\
&= \frac{1}{2}||Y - \mathcal{X}(\Omega+\Delta)||_F^2 - \frac{1}{2}||Y-\mathcal{X}(\Omega)||_F^2 + \langle \mathcal{X}^*\big(Y-\mathcal{X}(\Omega)\big), \Delta\rangle = \\
&\overset{(1)}{=} \frac{1}{2}||Y - \mathcal{X}(\Omega) - \mathcal{X}(\Delta)||_F^2 - \frac{1}{2}||Y-\mathcal{X}(\Omega)||_F^2 + \langle \mathcal{X}^*\big(Y-\mathcal{X}(\Omega)\big), \Delta\rangle = \\
&\overset{(2)}{=} \frac{1}{2}||\mathcal{X}(\Delta)||_F^2 - \langle Y-\mathcal{X}(\Omega), \mathcal{X}(\Delta)\rangle + \langle Y-\mathcal{X}(\Omega), \mathcal{X}(\Delta)\rangle = \frac{1}{2}||\mathcal{X}(\Delta)||_F^2
\end{aligned}$$

Where in (1) we have used the linearity of the operator $\mathcal{X}$ and in (2) we have used the definition of the adjoint operator $\mathcal{X}^*$. Therefore the left-hand side of equation (2.4) is exactly the first-order Taylor approximation of the loss function.
This condition is strong and is equivalent, in the case of $\mathcal{L} \in C^2$, to imposing a positive lower bound on the curvature of the loss function ($\nabla^2\mathcal{L}(\Omega) \succeq \mu$ for all Ω). This assumption is very useful when we want to have a bound on Δ if we have an upper bound on the first-order Taylor approximation of $\mathcal{L}$.
The strong convexity assumption is directly linked to the restricted isometry condition used by [Rohde and Tsybakov (2011)]. In particular, she assumed (for a specific class of matrices to which Δ belongs) that

$$(1-\delta_r)||\Delta||_F \leq \eta||\mathcal{X}(\Delta)||_F \leq (1+\delta_r)||\Delta||_F \tag{2.5}$$

As we can see, (restricted) isometry condition (2.5) is stronger than strong convexity (2.4). In fact, in isometry condition not only a lower bound on the first order Taylor approximation of $\mathcal{L}$ is requested, but also an upper bound on the latter. However, what we aim to consider is the high dimensional case: $d > n$ (small number of samples). In this case, the restricted isometry condition would imply the injectiveness of $\mathcal{X}$. But by the rank-nullity theorem applied to $\mathcal{X} : \mathbb{R}^{d\times T} \to \mathbb{R}^{n\times T}$, we have that dT =dim(Ker) + dim(Im) and therefore dim(Ker)$\geq (d-m)T > 0$, showing that the operator is not injective and so the condition cannot be applied to the high dimensional case.
Therefore restricted isometry condition is not a reasonable assumption in our case. Similarly, always in the case $d > n$, strong convexity assumption is unreasonable too because the Hessian matrix $\nabla^2\mathcal{L}(\Omega)$ is often singular. As a concrete example, consider our setting, in the specific case in which $\mathcal{X} = X \in \mathbb{R}^{n\times d}$ since the loss is $\mathcal{L}(\Theta,\Gamma) = \frac{1}{2}||Y - X(\Theta+\Gamma)||_F^2$. The Hessian matrix is

$$\nabla^2_{(\Theta,\Gamma)}\left(\frac{1}{2}||Y-X(\Theta+\Gamma)||_F^2\right) = \left[\begin{array}{c|c} X^TX & X^TX \\ \hline X^TX & X^TX \end{array}\right]$$

Which is positive semidefinite but not positive definite. In fact, notice that

$$(x_1,x_2)^T \left[\begin{array}{c|c} X^TX & X^TX \\ \hline X^TX & X^TX \end{array}\right] (x_1,x_2) = (x_1+x_2)^T X^TX(x_1+x_2) \geq 0$$

Now $X^TX \in \mathbb{R}^{d\times d}$ is positive semidefinite but not positive definite. Indeed X^TX has rank at most $n < d$ (because $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$), therefore X^TX is simply positive semidefinite and the same holds for $\nabla^2\mathcal{L}(\Theta,\Gamma)$. As the final consequence, $\mathcal{L}$ cannot be strongly convex.

Despite in the high dimensional case, the Hessian matrix is flat along some directions, we can still hope to have positive curvature along some directions of interest for example where $(\hat{\Theta}, \hat{\Gamma}) - (\Theta^*, \Gamma^*)$ lies. This is basically what restricted strong convexity is assuming. Formally, we can define a weighted combination of the two regularizers by

$$\mathcal{Q}(\Theta, \Gamma) \coloneqq ||\Theta||_N + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma) \tag{2.6}$$

And, we can define the associated norm to this weighted combination by

$$\Phi(\Delta) \coloneqq \inf_{\Theta+\Gamma=\Delta} \mathcal{Q}(\Theta + \Gamma)$$

After having defined $\hat{\Delta} = \hat{\Delta}^\Theta + \hat{\Delta}^\Gamma$ with $\hat{\Delta}^\Theta = \hat{\Theta} - \Theta^*$ and $\hat{\Delta}^\Gamma = \hat{\Gamma} - \Gamma^*$, we will now define their decomposition $\hat{\Delta}^\Theta = \hat{\Delta}^\Theta_A + \hat{\Delta}^\Theta_B$ and $\hat{\Delta}^\Gamma = \hat{\Delta}^\Gamma_{\mathbb{M}} + \hat{\Delta}^\Gamma_{M^\perp}$.

- Decomposition $\hat{\Delta}^\Theta$: If $\Theta^* = U\Sigma^* V^T$, is the SVD decomposition of Θ^*, we let

$$\left[\begin{array}{c|c} \hat{B}_{11} & \hat{B}_{12} \\ \hline \hat{B}_{21} & \hat{B}_{22} \end{array}\right] = U^T \hat{\Delta}^\Theta V$$

  with $\hat{B}_{11} \in \mathbb{R}^{r\times r}, \hat{B}_{12} \in \mathbb{R}^{r\times(T-r)}, \hat{B}_{21} \in \mathbb{R}^{(T-r)\times r}$ and $\hat{B}_{22} \in \mathbb{R}^{(T-r)\times(T-r)}$. We define now

$$\hat{\Delta}^\Theta_A \coloneqq U \left[\begin{array}{c|c} \hat{B_{11}} & \hat{B_{12}} \\ \hline \hat{B_{21}} & 0 \end{array}\right] V^T \quad \text{and} \quad \hat{\Delta}^\Theta_B \coloneqq U \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \hat{B_{22}} \end{array}\right] V^T$$

  This is a valid decomposition with properties given by the first point of lemma 2.6.

- Decomposition of $\hat{\Delta}^\Gamma$: $\hat{\Delta}^\Gamma_{\mathbb{M}}$ is the projection of $\hat{\Delta}^\Gamma$ onto the subspace $\mathbb{M}$, similarly $\hat{\Delta}^\Gamma_{\mathbb{M}^\perp}$ is the projection of $\hat{\Delta}^\Gamma$ onto the subspace $\mathbb{M}^\perp$.

We define a cone $\mathcal{C}$ as

$$\mathcal{C} = \left\{ (\hat{\Delta}^\Theta, \hat{\Delta}^\Gamma) \mid \mathcal{Q}(\hat{\Delta}^\Theta_B, \hat{\Delta}^\Gamma_{\mathbb{M}}) \leq 3\mathcal{Q}(\hat{\Delta}^\Theta_A, \hat{\Delta}^\Gamma_{\mathbb{M}^\perp}) + 4\left[\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\right] \right\} \tag{2.7}$$

Finally, the restricted strong convexity assumption is the following.

Definition 2.2. *The squared error loss with linear operator $\mathcal{X} : \mathbb{R}^{d\times T} \to \mathbb{R}^{d\times n}$ satisfies restricted strong convexity with parameters $(\gamma > 0, \tau_n \geq 0)$ over the cone $\mathcal{C}$ if*

$$\frac{1}{2}||\mathcal{X}(\Delta)||_F^2 \geq \frac{\gamma}{2}||\Delta||_F^2 - \tau_n\Phi(\Delta) \quad \textit{for all} \quad \Delta \in \mathcal{C} \tag{2.8}$$

The left-hand side is the error in the first order Taylor approximation of the loss, the right-hand side consists of two terms: $\frac{\gamma}{2}||\Delta||_F^2$ is the one that appeared in the strong convexity assumption too, while $-\tau_n\Phi(\Delta)$ is called tolerance term and weakens the assumption.

The restricted strong convexity assumption will be crucial in our analysis. Even if we will not prove explicitly that this condition holds in our setting, [Boursier et al. (2022)] proved that a form of restricted strong convexity holds for the case of low-rank only. However, an extension of (2.8) to the case of low-rank plus sparse was considered out of the scope of the work. The definition given here for the restricted strong convexity assumption is quite different from the one of [Agarwal et al. (2012)]. Indeed they asked the condition to hold for each $\Delta \in \mathbb{R}^{d\times T}$ while we ask it to hold for each Δ in a specific cone $\mathcal{C}$. Their condition is of course stronger than ours but as we will see in the proof of lemma 2.7, we don't need this condition to hold everywhere but it is enough if it holds for $(\hat{\Delta}^\Theta, \hat{\Delta}^\Gamma)$ in the cone $\mathcal{C}$.

2.4 Main result for general regularizers

We state and prove here the main result of [Agarwal et al. (2012)], and we will later specialize it in the case of multi-task learning for the specific case in which the regularizer is the L_1 or the $L_{2,1}$ norm. The theorem will give upper bounds on the error, measured using the squared Frobenius norm, between the estimates $\hat{\Theta}$, $\hat{\Gamma}$ and the true Θ^*, Γ^*. In particular, the theorem will upper bound

$$e^2(\hat{\Theta}, \hat{\Gamma}) = ||\hat{\Theta} - \Theta^*||_F^2 + ||\hat{\Gamma} - \Gamma^*||_F^2$$

As a function of three terms:

- $\mathcal{K}_{\Theta^*} = \frac{\lambda_d^2}{\gamma^2}\left(r + \frac{\gamma}{\lambda_d}\sum_{j=r+1}^{h}\sigma_j(\Theta^*)\right)$
- $\mathcal{K}_{\Gamma^*} = \frac{\mu_d^2}{\gamma^2}\left(\Psi^2(\mathbb{M}, \mathcal{R}) + \frac{\gamma}{\mu_d}\mathcal{R}(\Pi_{\mathbb{M}^\perp}(\Gamma^*))\right)$
- $\mathcal{K}_{\tau_n} = \frac{\tau_n}{\gamma}\left(\sum_{j=r+1}^{h}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Pi_{\mathbb{M}^\perp}(\Gamma^*))\right)^2$

Where we denoted by $\Pi_V(A)$ (sometimes to abbreviate also A_V) the projection of the matrix A onto the subspace V and by $h = \min\{d, T\}$. As we will discuss shortly, the first term corresponds to the error associated with the low-rank matrix Θ^*, the second with the sparse matrix Γ^* and the third term with the non-zero tolerance τ_n (in the restricted strong convexity (2.8)).

Theorem 2.5. *Suppose that the operator $\mathcal{X}$ satisfies the restricted strong convexity condition (2.8) with curvature $\gamma > 0$ and tolerance τ_n and there exist integers $r = 1, 2, \ldots, \min\{d, T\}$ for which*

$$128\tau_n r < \frac{\gamma}{4} \quad \textit{and} \quad 64\tau_n\left(\Psi(\mathbb{M}, \mathcal{R})\frac{\mu_d}{\lambda_d}\right)^2 < \frac{\gamma}{4}$$

Then, if $\hat{\Theta}$ and $\hat{\Gamma}$ are the solution to the convex problem 2.2 with regularization parameters

$$\lambda_d \geq 4||\mathcal{X}^*(W)||_{op} \quad \textit{and} \quad \mu_d \geq 4\mathcal{R}^*(\mathcal{X}^*(W)) + \frac{4\gamma\alpha}{\kappa_d}$$

There are universal (not depending on the size of the problem) constants c_1, c_2, c_3 such that for any matrix pair (Θ^*, Γ^*) satisfying $\varphi_{\mathcal{R}}(\Theta^*) \leq \alpha$ and any decomposable regularizer $\mathcal{R}$ with respect to the subspace pair $(\mathbb{M}, \mathbb{M}^\perp)$, any optimal solution $(\hat{\Theta}, \hat{\Gamma})$ satisfies

$$e^2(\hat{\Theta}, \hat{\Gamma}) \leq c_1\mathcal{K}_{\Theta^*} + c_2\mathcal{K}_{\Gamma^*} + c_3\mathcal{K}_{\tau_n}$$

Remark. Each term corresponds to a different error, the first one is the error corresponding to the estimation of Θ^*, the second is the one of the estimation of Γ^*, and the third one is associated to the non-zero tolerance τ_n (in the restricted strong convexity (2.8)). However, for each of the first two terms, we can have a further interpretation. Consider for example $\mathcal{K}^*_\Theta$, the first part: (term $\lambda_d^2 r$) is the estimation error while the second (term $\lambda_d\sum_{j=r+1}^{h}\sigma_j(\Theta^*)$) is the approximation error due to the fact that we are representing Θ^* by a matrix of rank r.
Notice moreover, that in the case $\tau_n = 0$ or the case in which Θ^* is exactly low rank with rank r and Γ^* lies on $\mathbb{M}$, the term $\mathcal{K}_{\tau_n}$ vanishes.

Proof. The proof of the theorem is based on two lemmas:

Lemma 2.6. *For any $r = 1, 2, \ldots, \min\{d, T\}$ there is a decomposition $\hat{\Delta}^\Theta = \hat{\Delta}^\Theta_A + \hat{\Delta}^\Theta_B$ such that:*

1. *The decomposition satisfies*

$$rank(\hat{\Delta}^\Theta_A) \leq 2r \quad and \quad (\hat{\Delta}^\Theta_A)^T \hat{\Delta}^\Theta_B = (\hat{\Delta}^\Theta_B)^T \hat{\Delta}^\Theta_A = 0 \tag{2.9}$$

2.

$$\mathcal{Q}(\Theta^*, \Gamma^*) - \mathcal{Q}(\Theta^* + \hat{\Delta}^\Theta, \Gamma^* + \hat{\Delta}^\Gamma) \leq \mathcal{Q}(\hat{\Delta}^\Theta_A, \hat{\Delta}^\Gamma_{\mathbb{M}}) - \mathcal{Q}(\hat{\Delta}^\Theta_B, \hat{\Delta}^\Gamma_{\mathbb{M}^\perp}) + 2\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{2\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp}) \tag{2.10}$$

3. *under the assumption on the regularization parameters of the main theorem 2.5, $\hat{\Delta}^\Theta$ and $\hat{\Delta}^\Gamma$ satisfy*

$$\mathcal{Q}(\hat{\Delta}^\Theta_B, \hat{\Delta}^\Gamma_{\mathbb{M}}) \leq 3\mathcal{Q}(\hat{\Delta}^\Theta_A, \hat{\Delta}^\Gamma_{\mathbb{M}^\perp}) + 4\left\{\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\right\} \tag{2.11}$$

Proof. See the appendix A.1 for the proof of this result. □

Lemma 2.7. *Under the conditions of Theorem 2.5, we have the following bound*

$$\frac{1}{2}||\mathcal{X}(\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma)||_F^2 \geq \frac{\gamma}{4}(||\hat{\Delta}^\Theta||_F^2 + ||\hat{\Delta}^\Gamma||_F^2) - \frac{\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^\Theta, \hat{\Delta}^\Gamma) \tag{2.12}$$

$$- 32\tau_n\left\{\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\right\}^2 \tag{2.13}$$

Proof. See the appendix A.1 for the proof of this result. □

We can now prove the main theorem using the two lemmas. By the optimality of $(\hat{\Theta}, \hat{\Gamma})$ and feasibility of (Θ^*, Γ^*) in the convex problem 2.2

$$\frac{1}{2}||Y - \mathcal{X}(\hat{\Theta} + \hat{\Gamma})||_F^2 + \lambda_d||\hat{\Theta}||_* + \mu_d\mathcal{R}(\hat{\Gamma}) \leq \frac{1}{2}||Y - \mathcal{X}(\Theta^* + \Gamma^*)||_F^2 + \lambda_d||\Theta^*||_* + \mu_d\mathcal{R}(\Gamma^*)$$

Now, using the fact that $Y = \mathcal{X}(\Theta^* + \Gamma^*) + W$ and naming $\hat{\Delta}^\Theta := \hat{\Theta} - \Theta^*$ and $\hat{\Delta}^\Gamma := \hat{\Gamma} - \Gamma^*$, we obtain

$$\frac{1}{2}||\mathcal{X}(\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma)||_F^2 \leq \langle \hat{\Delta}^\Theta + \hat{\Delta}^\Gamma, \mathcal{X}^*(W)\rangle + \lambda_d\mathcal{Q}(\Theta^*, \Gamma^*) - \lambda_d\mathcal{Q}(\Theta^* + \hat{\Delta}^\Theta, \Gamma^* + \hat{\Delta}^\Gamma),$$

Where $\mathcal{Q}$ is the weighted norm defined in 2.6. Now, we use 2.10 to obtain

$$\frac{1}{2}||\mathcal{X}(\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma)||_F^2 \leq \langle \hat{\Delta}^\Theta + \hat{\Delta}^\Gamma, \mathcal{X}^*(W)\rangle + \lambda_d\mathcal{Q}(\hat{\Delta}^\Theta_A, \hat{\Delta}^\Gamma_{\mathbb{M}}) - \lambda_d\mathcal{Q}(\hat{\Delta}^\Theta_B, \hat{\Delta}^\Gamma_{\mathbb{M}^\perp}) + 2\lambda_d\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + 2\mu_d\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})$$

Now, applying Holder's inequality to $\langle \hat{\Delta}^\Theta + \hat{\Delta}^\Gamma, \mathcal{X}^*(W)\rangle$, we obtain

$$\begin{aligned}\langle \hat{\Delta}^{\Theta} + \hat{\Delta}^{\Gamma}, \mathcal{X}^*(W)\rangle &= \langle \hat{\Delta}^{\Theta}_A + \hat{\Delta}^{\Theta}_B + \hat{\Delta}^{\Gamma}_{\mathbb{M}} + \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp}, \mathcal{X}^*(W)\rangle \\ &= \langle \hat{\Delta}^{\Theta}_A + \hat{\Delta}^{\Theta}_B, \mathcal{X}^*(W)\rangle + \langle \hat{\Delta}^{\Gamma}_{\mathbb{M}} + \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp}, \mathcal{X}^*(W)\rangle \\ &\leq ||\hat{\Delta}^{\Theta}_A + \hat{\Delta}^{\Theta}_B||_* ||\mathcal{X}^*(W)||_{\text{op}} + \mathcal{R}(\hat{\Delta}^{\Gamma}_{\mathbb{M}} + \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp})\mathcal{R}^*(\mathcal{X}^*(W)) \\ &\leq (||\hat{\Delta}^{\Theta}_A||_* + ||\hat{\Delta}^{\Theta}_B||_*)||\mathcal{X}^*(W)||_{\text{op}} + (\mathcal{R}(\hat{\Delta}^{\Gamma}_{\mathbb{M}}) + \mathcal{R}(\hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp}))\mathcal{R}^*(\mathcal{X}^*(W))\end{aligned}$$

Therefore, substituting in the previous inequality, we get the following bound

$$\begin{aligned}\frac{1}{2}||\mathcal{X}(\hat{\Delta}^{\Theta} + \hat{\Delta}^{\Gamma})||_F^2 \leq &(||\hat{\Delta}^{\Theta}_A||_* + ||\hat{\Delta}^{\Theta}_B||_*)||\mathcal{X}^*(W)||_{\text{op}} + (\mathcal{R}(\hat{\Delta}^{\Gamma}_{\mathbb{M}}) + \mathcal{R}(\hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp}))\mathcal{R}^*(\mathcal{X}^*(W)) \\ &+ \lambda_d \mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) - \lambda_d \mathcal{Q}(\hat{\Delta}^{\Theta}_B, \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp}) + 2\lambda_d \sum_{j=r+1}^{h} \sigma_j(\Theta^*) + 2\mu_d \mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\end{aligned}$$

Using now the inequalities on λ_d and μ_d: $\lambda_d \geq 4||\mathcal{X}^*(W)||_{\text{op}}$ and $\mu_d \geq 4\mathcal{R}^*(\mathcal{X}^*(W))$, we can simplify the previous equation in

$$\frac{1}{2}||\mathcal{X}(\hat{\Delta}^{\Theta} + \hat{\Delta}^{\Gamma})||_F^2 \leq \frac{3\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + 2\lambda_d \sum_{j=r+1}^{h} \sigma_j(\Theta^*) + 2\mu_d \mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})$$

Now, we lower bound the right-hand side using 2.12, and we obtain:

$$\begin{aligned}\frac{\gamma}{4}\big(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma})||_F^2\big) \leq \frac{3\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + \frac{\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) + \\ + 32\tau_n\{\sum_{j=r+1}^{d} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2 + 2\lambda_d \sum_{j=r+1}^{h} \sigma_j(\Theta^*) + 2\mu_d \mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\end{aligned}$$

Now, by triangle inequality $\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) \leq \mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + \mathcal{Q}(\hat{\Delta}^{\Theta}_B, \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp})$, moreover, combining it with lemma 2.11, we obtain

$$\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) \leq 4\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + 4\{\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}$$

Substituting in the previous equation, we get

$$\begin{aligned}\frac{\gamma}{4}\big(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma})||_F^2\big) \leq \frac{3\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + 32\tau_n\{\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2 \\ + 4\lambda_d \sum_{j=r+1}^{h} \sigma_j(\Theta^*) + 4\mu_d \mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\end{aligned} \tag{2.14}$$

Now we use one side of the equivalence between the nuclear and the Frobenius norm i.e. $||M||_*^2 \leq \text{rank}(M)||M||_F^2$, moreover, by definition of $\Psi(\mathbb{M})$, $\mathcal{R}(\hat{\Delta}^{\Gamma}_{\mathbb{M}}) \leq \Psi(\mathbb{M})||\hat{\Delta}^{\Gamma}_{\mathbb{M}}||_F^2$. Finally, by lemma 2.9 $\hat{\Delta}^{\Theta}_A$ has rank at most $2r$. Therefore combining,

$$\begin{aligned}\lambda_d \mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) &\leq \sqrt{2r}\lambda_d ||\hat{\Delta}^{\Theta}_A||_F + \Psi(\mathbb{M})\mu_d ||\hat{\Delta}^{\Gamma}_{\mathbb{M}}||_F \\ &\leq \sqrt{2r}\lambda_d ||\hat{\Delta}^{\Theta}||_F + \Psi(\mathbb{M})\mu_d ||\hat{\Delta}^{\Gamma}||_F\end{aligned}$$

Substituting that in the inequality 2.14 we get

$$||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma}||_F^2 \leq \frac{16}{\gamma}\{\sqrt{2r}\lambda_d||\hat{\Delta}^{\Theta}||_F + \Psi(\mathbb{M})\mu_d||\hat{\Delta}^{\Gamma}||_F\} + \frac{128}{\gamma}\tau_n\{\sum_{j=r+1}^{h}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2$$
$$+ \frac{16}{\gamma}\{\lambda_d\sum_{j=r+1}^{h}\sigma_j(\Theta^*) + \mu_d\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}$$

Now naming $x = ||\hat{\Delta}^{\Theta}||_F$, $y = ||\hat{\Delta}^{\Gamma}||_F$, rearranging we get

$$(x - \frac{8}{\gamma}\sqrt{2r}\lambda_d)^2 + (y - \frac{8}{\gamma}\Psi(\mathbb{M})\mu_d)^2 \leq \frac{128}{\gamma}\tau_n\{\sum_{j=r+1}^{h}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2$$
$$+ \frac{16}{\gamma}\{\lambda_d\sum_{j=r+1}^{h}\sigma_j(\Theta^*) + \mu_d\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}$$
$$+ \frac{128r}{\gamma^2}\lambda_d^2 + \frac{64}{\gamma^2}\mu_d^2\Psi(\mathbb{M})^2$$

Now clearly for $(x, y) = (0, 0)$ the bound is satisfied, this means that $(0, 0)$ is inside the circle centered at $(\frac{8}{\gamma}\sqrt{2r}\lambda_d, \frac{8}{\gamma}\Psi(\mathbb{M})\mu_d)$. Therefore, we can upper bound $x^2 + y^2$ by two times the radius of the circle centered at $(\frac{8}{\gamma}\sqrt{2r}\lambda_d, \frac{8}{\gamma}\Psi(\mathbb{M})\mu_d)$. Doing that, and renaming conveniently c_1, c_2, c_3, the result of the theorem follows. □

The result proved is general and applies not only to Multi-Task regression but also to a wide variety of problems such as factor analysis with sparse noise, robust PCA, and robust covariance estimator.

2.5 Application to Multi-Task Learning

In this section, we will specialize the theorem in the case in which the regularizer $\mathcal{R}$ is the L_1 or the $L_{2,1}$ norm. As we showed previously, the regularizer is decomposable with respect to careful choices of subspaces $\mathbb{M}$ and $\mathbb{M}^\perp$. The first result will be specific to the L_1 norm and it has already been proved in a similar fashion by [Agarwal et al. (2012)] (despite we relax some of the assumptions). However, the second is specific to the $L_{2,1}$ norm and, to our knowledge, is the first result providing a bound of estimation for this structure.

Corollary 2.7.1 (L1 penalization)**.** *Suppose that the matrix Θ^* has rank at most r and satisfies $||\Theta^*||_\infty \leq E := \frac{\alpha}{\sqrt{dT}}$, the matrix Γ^* has at most s non-zero entries. Assuming restricted strong convexity with positive γ and tolerance τ_n holds, and*

$$128\tau_n r < \frac{\gamma}{4} \quad \textit{and} \quad 64\tau_n\left(\sqrt{s}\,\frac{\mu_d}{\lambda_d}\right)^2 < \frac{\gamma}{4}$$

If the entries of W are i.i.d $\mathcal{N}(0, \nu^2)$, each column of each $\tilde{X}^{(t)} := \frac{X^{(t)}}{\sqrt{n}}$ is bounded in norm by k_{max}, let $\sigma_{max} = \max_{t\in[T]}\sigma_{max}(\tilde{X}^{(t)})$ and we solve the convex problem 2.2 with regularization parameters

$$\lambda_d = 16\nu\sigma_{max}\sqrt{n}(\sqrt{d} + \sqrt{T}) \quad \textit{and} \quad \mu_d = 16\nu k_{max}\sqrt{n\log\left(dT\right)} + 4\gamma\sqrt{n}\underbrace{\frac{\alpha}{\sqrt{dT}}}_{E}$$

Then we have that with probability higher than $1 - 4\max\{e^{-7\log(dT)}, e^{-c(d+T)}\}$ *that any optimal solution* $(\hat{\Theta}, \hat{\Gamma})$ *satisfies*

$$e^2(\hat{\Theta}, \hat{\Gamma}) \leq c_1 \frac{\sigma_{max}^2 \nu^2}{\gamma^2}\left(\frac{r(d+T)}{n}\right) + c_2\left(\frac{\nu^2 k_{max}^2}{\gamma^2}\frac{s\log(dT)}{n} + \underbrace{\frac{\alpha^2}{dT}}_{E^2} s\right) \tag{2.15}$$

Where c_1, c_2 *are universal positive constants.*

Proof. In order to prove the result, we will verify that the hypothesis of the main theorem 2.5 holds and the result will follow as a corollary of the latter. In order to do that, we will need to verify that $\lambda_d \geq 4||\mathcal{X}^*(W)||_{\text{op}}$ and $\mu_d \geq 4||\mathcal{X}^*(W)||_\infty + \frac{4\gamma\alpha}{\sqrt{dT}}$ with probability higher than $1-4\max\{e^{-7\log(dT)}, e^{-c(d+T)}\}$. First of all, recall that the operator $\mathcal{X} : \mathbb{R}^{d\times T} \to \mathbb{R}^{n\times T}$ is such that

$$\mathcal{X}(A) = \begin{bmatrix} | & | & | & | \\ X^{(1)}A_1 & X^{(2)}A_2 & \dots & X^{(T)}A_T \\ | & | & | & | \end{bmatrix}$$

Therefore, knowing that the adjoint operator, denoted by $\mathcal{X}^* : \mathbb{R}^{n\times T} \to \mathbb{R}^{d\times T}$ is such that $\langle \mathcal{X}(A), B\rangle = \langle A, \mathcal{X}^*(B)\rangle$, then we can compute it explicitely by its definition.

$$\begin{aligned}
\langle \mathcal{X}(A), B\rangle &= \langle \begin{pmatrix} X^{(1)}A_1 & X^{(2)}A_2 & \dots & X^{(T)}A_T \end{pmatrix}, B\rangle = \\
&= \langle X^{(1)}A_1, B_1\rangle + \langle X^{(2)}A_2, B_2\rangle + \dots \langle X^{(T)}A_T, B_T\rangle = \\
&= \langle A_1, {X^{(1)}}^T B_1\rangle + \langle A_2, {X^{(2)}}^T B_2\rangle + \dots \langle A_T, {X^{(T)}}^T B_T\rangle = \\
&= \langle A, \mathcal{X}^*(B)\rangle
\end{aligned}$$

Therefore,

$$\mathcal{X}^*(B) = \begin{bmatrix} | & | & | & | \\ {X^{(1)}}^T B_1 & {X^{(2)}}^T B_2 & \dots & {X^{(T)}}^T B_T \\ | & | & | & | \end{bmatrix}$$

is the adjoint operator. We will denote by $\tilde{\mathcal{X}}$, $\tilde{W}$, $\tilde{Y}$ the operator and the matrices such that $\tilde{\mathcal{X}}(A) = \frac{1}{\sqrt{n}}\mathcal{X}(A)$, $\tilde{W} = \frac{1}{\sqrt{n}}W$ and $\tilde{Y} = \frac{1}{\sqrt{n}}Y$, let $\tilde{\lambda_d} = \frac{\lambda_d}{n}$ and $\tilde{\mu_d} = \frac{\mu_d}{n}$.
We will verify that $\tilde{\lambda_d} \geq 4||\tilde{\mathcal{X}}^*(\tilde{W})||_{\text{op}}$ with high probability. Now, recall that $(\mathcal{X}^*(W))_j = Z_j \sim \mathcal{N}(0, \nu^2 {X^{(j)}}^T X^{(j)})$, so we have that $(\tilde{\mathcal{X}}^*(\tilde{W}))_j = Z_j \sim \mathcal{N}(0, \frac{\nu^2}{n}{\tilde{X}^{(j)}}^T \tilde{X}^{(j)})$, moreover $||{\tilde{X}^{(j)}}^T \tilde{X}^{(j)}||_{\text{op}} = \sigma^2_{\max}(\tilde{X}^{(j)})$, therefore, by results on the singular values of Gaussian random matrices,

$$\mathbb{P}\left[||\tilde{\mathcal{X}}(\tilde{W})||_{\text{op}} \geq \frac{4\nu\sigma_{\max}\left(\sqrt{d}+\sqrt{T}\right)}{\sqrt{n}}\right] \leq 2e^{-c(d+T)}$$

Therefore, setting $\tilde{\lambda_d} = \frac{16\nu\sigma_{\max}\left(\sqrt{d}+\sqrt{T}\right)}{\sqrt{n}}$ ensure that the hypothesis of the theorem is satisfied.
Now, since we supposed that the norm of each column of each feature matrix $X^{(j)}$ is bounded by $k_{\max}\sqrt{n}$, it follows that $(\tilde{\mathcal{X}}^*(\tilde{W}))_{ij} = \sum_k \tilde{X}_{ki}\tilde{W}_{kj}$ is a normal random variable with mean zero and variance bounded by $\frac{\nu^2 k^2_{\max}}{n}$. Now we need to bound the probability

that $\mathbb{P}\left[||\tilde{\mathcal{X}}^*(\tilde{W})||_\infty > \frac{1}{4}\tilde{\mu_d} - \frac{\gamma\alpha}{k_d}\right]$. Recall that $k_d = \sup_{V\neq 0}\frac{||V||_F}{||V||_1} = \sqrt{dT}$, therefore, choosing $\tilde{\mu_d} = 16\frac{\nu k_{\max}}{\sqrt{n}}\sqrt{\log\left(dT\right)} + \frac{4\alpha\gamma}{\sqrt{dT}}$, we simply need to bound

$$\mathbb{P}\left[||\tilde{\mathcal{X}}^*(\tilde{W})||_\infty > 4\frac{\nu k_{\max}}{\sqrt{n}}\sqrt{\log\left(dT\right)}\right]$$

Now by union bound,

$$\mathbb{P}\left[||\tilde{\mathcal{X}}^*(\tilde{W})||_\infty > \frac{4\nu k_{\max}}{\sqrt{n}}\sqrt{\log\left(dT\right)}\right] \leq dT\ \mathbb{P}\left[|(\sqrt{n}\tilde{\mathcal{X}}^*(\tilde{W}))_{i,j}| > 4\nu k_{\max}\sqrt{\log\left(dT\right)}\right]$$

Using the standard concentration inequality for subgaussian random variables: $\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$, we can conclude that

$$\mathbb{P}\left[|(\sqrt{n}\tilde{\mathcal{X}}^*(\tilde{W}))_{i,j}| > 4\nu k_{\max}\sqrt{\log\left(dT\right)}\right] \leq 2e^{-8\log(dT)}$$

All together,

$$\mathbb{P}\left[||\tilde{\mathcal{X}}^*(\tilde{W})||_\infty > \frac{1}{4}\tilde{\mu_d} - \frac{\gamma\alpha}{k_d}\right] \leq 2e^{\log(dT)-8\log(dT)} \leq 2e^{-7\log(dT)}.$$

By union bound,

$$\mathbb{P}\left[||\tilde{\mathcal{X}}^*(\tilde{W})||_{\text{op}} > \frac{1}{4}\tilde{\lambda_d} \text{ or } ||\tilde{\mathcal{X}}^*(\tilde{W})||_\infty > \frac{1}{4}\tilde{\mu_d} - \frac{\gamma\alpha}{\sqrt{dT}}\right] \leq 4\max\{e^{-7\log(dT)}, e^{-c(d+T)}\}$$

Therefore, the hypotheses of the theorem are verified with high probability, and by direct application, we conclude that, letting $\hat{\Theta}$ and $\hat{\Gamma}$ the solution to the optimization problem

$$\arg\min_{\Theta,\Gamma}\frac{1}{2}||\tilde{Y} - \tilde{X}(\Theta+\Gamma)||_F^2 + \tilde{\lambda_d}||\Theta||_* + \tilde{\mu_d}||\Gamma||_1 \quad \text{s.t.} \quad \varphi_{||.||_1}(\Theta) \leq \alpha \tag{2.16}$$

We have that with probability at least $1 - 4\max\{e^{-7\log(dT)}, e^{-(\sqrt{d}+\sqrt{T})^2}\}$,

$$e^2(\hat{\Theta},\hat{\Gamma}) \leq \frac{c_1\sigma_{\max}^2\nu^2}{\gamma^2}\left(\frac{r(d+T)}{n}\right) + c_2\left(\frac{\nu^2k_{\max}^2}{\gamma^2}\frac{s\log(dT)}{n} + \frac{\alpha^2 s}{dT}\right)$$

We conclude by noticing that 2.16 is equivalent to the original optimization problem 2.1 when $\lambda_d = n\tilde{\lambda_d}$ and $\mu_d = n\tilde{\mu_d}$. □

Corollary 2.7.2 (L2,1 penalization)**.** *Suppose that the matrix Θ^* has rank at most r and satisfies $||\Theta^*||_\infty \leq E := \frac{\alpha}{\sqrt{T}}$, the matrix Γ^* has at most s non-zero columns. Assuming restricted strong convexity with positive γ and tolerance τ_n holds and*

$$128\tau_n r < \frac{\gamma}{4} \quad \textit{and} \quad 64\tau_n\left(\sqrt{s}\,\frac{\mu_d}{\lambda_d}\right)^2 < \frac{\gamma}{4}$$

If the entries of W are i.i.d $\mathcal{N}(0,\nu^2)$, each column of each $\tilde{X}^{(t)} := \frac{X^{(t)}}{\sqrt{n}}$ is bounded in norm by k_{max}, let $\sigma_{max} = \max_{j\in[T]}\{\sigma_{max}(\tilde{X}^{(t)})\}$ and we solve the convex problem 2.2 with regularization parameters

$$\lambda_d = 16\nu\sigma_{max}\sqrt{n}(\sqrt{d}+\sqrt{T}) \quad \textit{and} \quad \mu_d = 16\nu k_{max}\sqrt{n\log\left(dT\right)} + 4\gamma\sqrt{n}\underbrace{\frac{\alpha}{\sqrt{T}}}_{E}$$

Then we have that with probability higher than $2\max\{e^{-3\log(T)}, e^{-c(d+T)}\}$, *any optimal solution* $(\hat{\Theta}, \hat{\Gamma})$ *satisfies*

$$e^2(\hat{\Theta}, \hat{\Gamma}) \leq \frac{c_1 \sigma_{max}^2 \nu^2}{\gamma^2}\left(\frac{r(d+T)}{n}\right) + c_2\left(\frac{\nu^2\sigma_{max}^2}{\gamma^2}\frac{s(d+\log(T))}{n} + s\underbrace{\frac{\alpha^2}{T}}_{E^2}\right)$$

Where c_1, c_2 *are universal positive constants.*

Proof. Using the results of the previous corollary and defining again $\tilde{\mu_d} = \frac{\mu_d}{n}$ and $\tilde{\lambda_d} = \frac{\lambda_d}{n}$, we already have that setting $\tilde{\lambda_d} = \frac{16\nu\sigma_{\max}\left(\sqrt{d}+\sqrt{T}\right)}{\sqrt{n}}$ ensures

$$\mathbb{P}\left[||\tilde{\mathcal{X}}(\tilde{W})||_{\text{op}} \geq \frac{1}{4}\tilde{\lambda_d}\right] \leq 2e^{-c(d+T)} \quad \text{for some positive c}$$

Now, as shown previously, if $\mathcal{R} = ||.||_{2,1}$, we have $\mathcal{R}^* = ||.||_{2,\infty}$, therefore we are interested in proving that $\mathbb{P}\left[||\tilde{\mathcal{X}}(\tilde{W})||_{2,\infty} > \frac{1}{4}\tilde{\mu_d} + \frac{\gamma\alpha}{\sqrt{T}}\right]$ is sufficiently small. Now notice that $\sqrt{n}\left(\tilde{\mathcal{X}}(\tilde{W})\right)_j \sim \mathcal{N}(0, \nu^2 \tilde{X}^{(j)^T}\tilde{X}^{(j)})$. This implies that the random vector $\sqrt{n}\left(\tilde{\mathcal{X}}(\tilde{W})\right)_j$ is subgaussian with parameter $\nu^2||\tilde{X}^{(j)^T}\tilde{X}^{(j)}||_{\text{op}}$, therefore using the results in [Rinaldo (2019)], and letting $\sigma_{\max}^{(j)} := \sigma_{\max}(\tilde{X}^{(j)})$ we have that for each $t > 0$:

$$\mathbb{P}[||\sqrt{n}\left(\tilde{\mathcal{X}}(\tilde{W})\right)_j||_2 \geq 4\sigma_{\max}^{(j)}\nu\sqrt{d} + t\Big] \leq e^{-\frac{t^2}{4(\sigma_{\max}^{(j)})^2\nu^2}}$$

As a consequence, after a union bound and denoting by $\sigma_{\max} = \max_j\{\sigma_{\max}(\tilde{X}^{(j)})\}$ we have,

$$\mathbb{P}[||\sqrt{n}\left(\tilde{\mathcal{X}}(\tilde{W})\right)||_{2,\infty} \geq 4\sigma_{\max}\nu\sqrt{d} + t\Big] \leq e^{-\frac{t^2}{4(\sigma_{\max})^2\nu^2} + \log(T)}$$

Choosing now $t = 4\nu\sigma_{\max}\sqrt{\log(T)}$, we have

$$\mathbb{P}[||\sqrt{n}\left(\tilde{\mathcal{X}}(\tilde{W})\right)||_{2,\infty} \geq 4\sigma_{\max}\nu\sqrt{d} + 4\nu\sigma_{\max}\sqrt{\log(T)}\Big] \leq e^{-3\log(T)}$$

Therefore we can choose $\frac{1}{4}\tilde{\mu_d} - \frac{\gamma\alpha}{\sqrt{T}} = 4\frac{\sigma_{\max}\nu\sqrt{d}}{\sqrt{n}} + 4\frac{\nu\sigma_{\max}\sqrt{\log(T)}}{\sqrt{n}}$ so that

$$\mathbb{P}\left[4||\tilde{\mathcal{X}}(\tilde{W})||_{2,\infty} > \tilde{\mu_d} - \frac{4\gamma\alpha}{\sqrt{T}}\right] \leq e^{-3\log(T)}$$

Again by union bound,

$$\mathbb{P}\left[||\tilde{\mathcal{X}}^*(\tilde{W})||_{\text{op}} > \frac{1}{4}\tilde{\lambda_d} \;\text{ or }\; ||\tilde{\mathcal{X}}^*(\tilde{W})||_{2,\infty} > \frac{1}{4}\tilde{\mu_d} - \frac{\gamma\alpha}{\sqrt{T}}\right] \leq 2\max\{e^{-3\log(T)}, e^{-c(d+T)}\}$$

Therefore, the hypotheses of the theorem 2.5 are satisfied with high probability, applying it we obtain

$$e^2(\hat{\Theta}, \hat{\Gamma}) \leq \frac{c_1\sigma_{\max}^2\nu^2}{\gamma^2}\left(\frac{r(d+T)}{n}\right) + c_2\left(\frac{\nu^2\sigma_{\max}^2}{\gamma^2}\frac{s(d+\log(T))}{n} + s\frac{\alpha^2}{T}\right)$$

The proof ends by noticing that a new optimization problem

$$\arg\min_{\Theta,\Gamma}\frac{1}{2}||\tilde{Y} - \tilde{X}(\Theta+\Gamma)||_F^2 + \tilde{\lambda_d}||\Theta||_* + \tilde{\mu_d}||\Gamma||_{2,1} \quad \text{s.t.} \;\; \varphi_{||.||_{2,1}}(\Theta) \leq \alpha$$

is equivalent to the original optimization problem 2.1 when $\lambda_d = n\tilde{\lambda_d}$ and $\mu_d = n\tilde{\mu_d}$. □

Remark. The results obtained are in line with the best known bounds on the estimation error for the low-rank part only and the sparse part only. In fact, as can be see in table 1 of [Boursier et al. (2022)], the best results known for the estimation error of Θ^* is, omitting logarithmic factors, of the form

$$||\hat{\Theta} - \Theta^*||_F \leq c\nu \left(\frac{r(d+T)}{n} \right)$$

For some universal constant c. This is exactly what we could get if $s = 0$ and $\frac{\sigma_{\max}}{\gamma}$ was a constant in our corollaries. Notice that [Boursier et al. (2022)] proved restricted strong convexity in the low-rank case with γ being a constant (not depending on the parameters). Therefore, hoping that this still holds in our setting, we only need $\sigma_{\max}$ to be a constant to get the same bound on the estimation error in both corollaries when $s = 0$ (Γ^* is the null matrix).
On the other hand if $r = 0$ i.e. Θ^* is the null matrix, again as long as $\frac{\sigma_{\max}}{\gamma}$ is a constant, we recover for both corollaries the same estimation errors (by constant factors) of [Wainwright (2019)] chapters 9.5 and 9.6 and these estimation errors are later proved to be optimal in chapter 15. Therefore, the estimation errors proved combine the best known bounds for the low-rank and sparse recovery only.

3. Optimization and experiments

3.1 Properties of the problem

From a practical perspective, it is important to learn how to solve efficiently the optimization problem 2.2. We recall it here that the problem is

$$\min_{\Theta \in \mathbb{R}^{d\times T},\ \Gamma \in \mathbb{R}^{d\times T}} \frac{1}{2}||Y - \mathcal{X}(\Theta + \Gamma)||_F^2 + \lambda_d||\Theta||_* + \mu_d \mathcal{R}(\Gamma) \quad \text{s.t.} \quad \varphi_{\mathcal{R}}(\Theta) \leq \alpha$$

For simplicity, in this section, we will focus on the scenario in which $\mathcal{R}(\Gamma) = ||\Gamma||_1$ but the optimization procedures that we show can be extended to other regularizers for instance $||.||_{2,1}$ or even $||.||_{1,2}$. As a consequence, the problem becomes

$$\min_{\Theta \in \mathbb{R}^{d\times T},\ \Gamma \in \mathbb{R}^{d\times T}} \frac{1}{2}||Y - \mathcal{X}(\Theta + \Gamma)||_F^2 + \lambda_d||\Theta||_* + \mu_d||\Gamma||_1 \quad \text{s.t.} \quad ||\Theta||_\infty \leq \frac{\alpha}{\sqrt{dT}}$$

First of all, we have the following proposition.

Proposition 3.1. *The problem is convex.*

Proof. Recall that a constrained minimization problem is convex when both the objective and the set constraint are convex.

- The term $\frac{1}{2}||Y - \mathcal{X}(\Theta+\Gamma)||_F^2$ is convex indeed, it is a composition of a convex function ($f : \mathbb{R}^{d\times T} \to \mathbb{R}$ with $f = \frac{1}{2}||.||_F^2$) with a linear function (recall that the operator is linear).
- $\lambda_d||\Theta||_* + \mu_d||\Gamma||_1$ is convex because it is a linear combination with positive coefficients of convex functions (norms are convex).
- The constraint set is also convex since the L_∞ ball with each radius r is convex.

Therefore, the problem is convex. □

Notice now that the constraint $||\Theta||_\infty \leq \frac{\alpha}{\sqrt{dT}}$ is always automatically satisfied for a big enough α coefficient. Moreover, it was introduced in the analysis mainly to restrict the class of matrices and to provide estimation bounds in this class. However, if we solve the convex problem without the constraint, the constraint will be automatically satisfied for some choice of α (for both our estimate and the true Θ^*) and the theorem will still apply for this choice of α, even if we will not control the magnitude of the latter. Therefore, we

can assume without loss of generality that the problem we are interested in solving is the following

$$\min_{\Theta\in\mathbb{R}^{d\times T},\ \Gamma\in\mathbb{R}^{d\times T}} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2+\lambda_d||\Theta||_*+\mu_d||\Gamma||_1 \tag{3.1}$$

Now, the problem is still convex since it is the unconstrained version of a convex problem. Moreover, we notice that the objective is not necessarily strongly convex. Indeed, if for instance $\mathcal{X}=X$ (i.e. all design matrices are equal), then

$$\nabla^2_{(\Theta,\Gamma)}\left(\frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2\right)=\left[\begin{array}{c|c} X^TX & X^TX \\ \hline X^TX & X^TX \end{array}\right]$$

Which is positive semidefinite but not necessarily positive definite. Indeed,

$$(x_1,x_2)^T\left[\begin{array}{c|c} X^TX & X^TX \\ \hline X^TX & X^TX \end{array}\right](x_1,x_2)=(x_1+x_2)^TX^TX(x_1+x_2)\geq 0$$

Because X^TX is positive semidefinite. Moreover, the nuclear norm and the 1-norm are not strictly convex and, therefore, not even strongly convex. It follows that the objective is not strongly convex. Moreover, the function to minimize is not differentiable but it is the sum of a differentiable part ($\frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2$) and a non-differentiable part ($\lambda_d||\Theta||_*+\mu_d||\Gamma||_1$). Finally, we notice that $\frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2$ is a smooth function (in the optimization sense not in the analysis sense, see [Bernd Gartner (2020)] for a proper definition). The proof is straightforward

Proposition 3.2. *The function f defined as $f(\Theta,\Gamma)=\frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2$ is smooth.*

Proof. Because the function f is convex and differentiable and the domain is the whole space $\mathbb{R}^{d\times T}\times\mathbb{R}^{d\times T}$, smoothness is equivalent to having Lipschitz gradient, see lemma 2 in [Bernd Gartner (2020)]. In our case

$$\begin{aligned}
||\nabla f(\Theta,\Gamma)-\nabla f(\Theta',\Gamma')||_F &= ||\left[\begin{array}{c} \mathcal{X}^*\mathcal{X}(\Theta+\Gamma-\Theta'-\Gamma') \\ \hline \mathcal{X}^*\mathcal{X}(\Theta+\Gamma-\Theta'-\Gamma') \end{array}\right]||_F \\
&= \sqrt{2}||\mathcal{X}^*\mathcal{X}(\Theta+\Gamma-\Theta'-\Gamma')||_F \\
&\leq \sqrt{2}||\mathcal{X}^*\mathcal{X}||_{\text{op}}||\Theta+\Gamma-\Theta'-\Gamma'||_F^2 \\
&\leq \sqrt{2}||\mathcal{X}^*\mathcal{X}||_{\text{op}}\big(||\Theta-\Theta'||_F^2+||\Gamma-\Gamma'||_F^2\big)
\end{aligned}$$

As a consequence, the gradient is Lipschitz, and f is smooth. □

3.2 Optimization

We could think about using subgradient descent to optimize 3.1 but in this case, we will not have any theoretical guarantees since the objective is not strongly convex nor Lipschitz [Bernd Gartner (2020)].
On the other hand, we could turn the optimization problem into a theoretically equivalent one:

$$\min_{\Theta\in\mathbb{R}^{d\times T},\ \Gamma\in\mathbb{R}^{d\times T}} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2 \quad \text{s.t.} \quad ||\Gamma||_1\leq r_1, \quad ||\Theta||_*\leq r_2$$

And optimize it with projected gradient descent, since now the objective is convex and smooth. Despite we could have theoretical convergence in $O(\frac{1}{\epsilon})$ [Bernd Gartner (2020)],

it is not practical to choose the hyperparameters r_1 and r_2 and, it would not be easy to relate them to μ_d and λ_d and verify empirically the results proved in the corollary 2.7.1. However, the problem has a nice structure, indeed, the objective is the sum of two functions: $f(\Theta,\Gamma) = \frac{1}{2}||Y - \mathcal{X}(\Theta+\Gamma)||_F^2$ and $g(\Theta,\Gamma) = \lambda_d||\Theta||_* + \mu_d||\Gamma||_1$. Both are convex functions but f is smooth and differentiable. Therefore, it is a natural choice to apply the Proximal Gradient algorithm since convergence in $O(\frac{1}{\epsilon})$ is assured. The idea behind this algorithm is simple, if $h : \mathbb{R}^d \to \mathbb{R}$ is $h = g + f$, the iterates of the algorithm are defined as

$$\begin{aligned} x_{t+1} &= \arg\min_{y\in\mathbb{R}^d} f(x_t) + \nabla f(x_t)^T(y - x_t) + \frac{1}{2\gamma}||y - x_t||^2 + g(y) \\ &= \arg\min_{y\in\mathbb{R}^d} \frac{1}{2\gamma}||y - (x_t - \gamma\nabla f(x_t))||^2 + g(y) \end{aligned}$$

The last formulation makes clear the goal of the update: staying close to the classical gradient update, as well as minimizing g. We will now show how to compute the iterates in our problem. Following the scheme proposed and calling $x_t = (\Theta_t, \Gamma_t)$ the iterate at iteration t, the update is

$$\begin{aligned} (\Theta_{t+1},\Gamma_{t+1}) = \arg\min_{(\Theta,\Gamma)\in\mathbb{R}^{d\times T}\times\mathbb{R}^{d\times T}} &\frac{1}{2}||Y - \mathcal{X}(\Theta_t + \Gamma_t)||_F^2 - \langle Y - \mathcal{X}(\Theta_t + \Gamma_t), \mathcal{X}(\Theta - \Theta_t)\rangle \\ &- \langle Y - \mathcal{X}(\Theta_t + \Gamma_t), \mathcal{X}(\Gamma - \Gamma_t)\rangle + \frac{1}{2\gamma}||\Theta - \Theta_t||_F^2 \\ &+ \frac{1}{2\gamma}||\Gamma - \Gamma_t||_F^2 + \lambda_d||\Theta||_* + \mu_d||\Gamma||_1 \end{aligned}$$

As we can see, the problem can be nicely decoupled into two different optimization problems:

$$\begin{aligned} \Theta_{t+1} &= \arg\min_{\Theta\in\mathbb{R}^{d\times T}} -\langle Y - \mathcal{X}(\Theta_t + \Gamma_t), \mathcal{X}(\Theta - \Theta_t)\rangle + \frac{1}{2\gamma}||\Theta - \Theta_t||_F^2 + \lambda_d||\Theta||_* \\ \Gamma_{t+1} &= \arg\min_{\Gamma\in\mathbb{R}^{d\times T}} -\langle Y - \mathcal{X}(\Theta_t + \Gamma_t), \mathcal{X}(\Gamma - \Gamma_t)\rangle + \frac{1}{2\gamma}||\Gamma - \Gamma_t||_F^2 + \mu_d||\Gamma||_1 \end{aligned}$$

Notice that the two problems can be reformulated (after some algebra) equivalently as

$$\Theta_{t+1} = \arg\min_{\Theta\in\mathbb{R}^{d\times T}} \frac{1}{2\gamma}||\Theta - \{\Theta_t + \gamma\mathcal{X}^*(Y - \mathcal{X}(\Theta_t + \Gamma_t))\}||_F^2 + \lambda_d||\Theta||_* \tag{3.2}$$

$$\Gamma_{t+1} = \arg\min_{\Gamma\in\mathbb{R}^{d\times T}} \frac{1}{2\gamma}||\Gamma - \{\Gamma_t + \gamma\mathcal{X}^*(Y - \mathcal{X}(\Theta_t + \Gamma_t))\}||_F^2 + \mu_d||\Gamma||_1 \tag{3.3}$$

The two problems have already been studied [Chambolle et al. (1998)], [Cai et al. (2008)] and have a closed-form solution. In particular, it is not difficult to verify that

$$\begin{aligned} &\Gamma_{t+1} = S_{\gamma\mu_d}(\Gamma_t + \gamma\mathcal{X}^*(Y - \mathcal{X}(\Theta_t + \Gamma_t))) \\ &\Theta_{t+1} = US_{\gamma\lambda_d}(\Sigma)V^T \quad \text{where} \quad \Theta_t + \gamma\mathcal{X}^*(Y - \mathcal{X}(\Theta_t + \Gamma_t)) = U\Sigma V^T \end{aligned}$$

And where S_λ is the thresholding operator and is defined as $S_\lambda : \mathbb{R}^{d\times T} \to \mathbb{R}^{d\times T}$ such that

$$(S_\lambda(A))_{ij} = \begin{cases} A_{ij} - \lambda & \text{if } A_{ij} > \lambda \\ 0 & \text{if } |A_{ij}| \le \lambda \\ A_{ij} + \lambda & \text{if } A_{ij} < -\lambda \end{cases}$$

Algorithm 1: Proximal Gradient Descent

Data: Initial guesses (Θ_0, Γ_0), step-size γ, regularization parameters (λ_d, μ_d), max-iters n

$t = 0$
while $t \leq n$ **do**
  $H_t = -\gamma \mathcal{X}^*(Y - \mathcal{X}(\Theta_t + \Gamma_t))$
  $K_\Gamma = \Gamma_t - \gamma H_t$
  $K_\Theta = \Theta_t - \gamma H_t$
  $\Gamma_{t+1} = S_{\gamma\mu_d}(K_\Gamma)$
  $U, \Sigma, V^T = \text{SVD}(K_\Theta)$
  $\Theta_{t+1} = U S_{\gamma\lambda_d}(\Sigma) V^T$
  $t \leftarrow t + 1$
end
return *The matrices* Θ_n, Γ_n

Therefore, the algorithm using proximal gradient descent can be summarized in Algorithm 1. The algorithm has theoretical guarantees of convergence in $O(\frac{1}{\epsilon})$ iterations but the cost per iteration is significant when the dimensions n, d and T are huge. In particular, the cost per iteration is dominated by a full SVD on K_Θ, which has cost $O(\min(dT^2, d^2T))$ and by the gradient computation which costs $O(dTn)$. Since we suppose $n \leq d$ in our setting, then the cost per iteration is $O(\min(dT^2, d^2T))$. As this cost can get significant in our case, the algorithm could be slow in practice. Notice that the Proximal Gradient Algorithm can be accelerated. In this case, we could converge in $\frac{1}{\sqrt{\epsilon}}$ iterations [Beck and Teboulle (2009)], each of cost $O(\min(dT^2, d^2T))$. We have thought about different ways to reduce this cost per iteration. One possibility would be to approximate the Singular Value Thresholding operator applied to the matrix K_Θ [Oh et al. (2018)] by approximating K_Θ with a smaller matrix of rank s, in this way, the total cost per iteration would reduce to $O(dTs)$. However, K_Θ is not necessarily low-rank and therefore the speed-up could be minimal. Moreover, even if K_Θ were low-rank, the solution would still be approximate and therefore the results of our corollaries could not be verified. Another approach would be to cast the original proximal subproblem 3.2 into a bilinear factorization form [Liu and Yan (2012)], however in this case we don't have guarantees to reach a global minimizer too and, using the factorization to solve the subproblem is inconvenient: we would have used it directly on the original problem 3.1. Therefore, we decided to adapt the work of [Mu et al. (2016)] which leverages the use of the Frank-Wolfe algorithm to our setting. We defer to the appendix A.2 for a complete discussion of the algorithm and its analysis of convergence. We could reach convergence in $O(\frac{1}{\epsilon})$ iterations, with a cost per iteration of $O(dTn)$ which is significantly lower than the cost of a full SVD if d and T are significantly larger than the number of samples per task. This improvement in the cost per iteration has been proved experimentally and the time per iteration is indeed lower when using Frank-Wolfe for large values of d and T compared to n.

3.3 Experiments

To complete our analysis, we verify the correctness of our work on synthetic datasets. The setting is similar to the one of [Boursier et al. (2022)] but differs slightly because of the additional sparse structure (Γ^*). The codes are available at https://github.com/lucarossi9/Robust-Multi-Task-Learning.git. We generate the ground-truth matrices Θ^* and Γ^* in the following

way:

- For the low-rank matrix Θ^*, we first sample a matrix $C^* \in \mathbb{R}^{d\times d}$ from i.i.d. $\sim \mathcal{N}(0,1)$. Then, if the rank of Θ^* is r, we select the top r left singular vectors to form $B^* \in \mathbb{R}^{d\times r}$. We sample each entry of $\alpha^* \in \mathbb{R}^{r\times T}$ always from i.i.d. $\sim \mathcal{N}(0,1)$. Finally, we set $\Theta^* = B^*\alpha^*$.

- For Γ^*, first we sample each entry of A^* from i.i.d. $\sim \mathcal{N}(0,1)$. Then, if the sparsity of the matrix is set to s (the matrix has s non-zero entries), we generate a random mask of s entries and apply it to the matrix A^* to obtain Γ^*.

- We generate the matrix W so that each entry is drawn from i.i.d. $\sim \mathcal{N}(0,\nu^2)$, with $\nu = 0.1$.

- We generate the design matrices $X^{(t)} \in \mathbb{R}^{n\times T}$ always such that each entry is sampled from $\sim \mathcal{N}(0,1)$. And each matrix is sampled independently from the others.

- The matrix Y is computed as $Y = \mathcal{X}(\Theta^* + \Gamma^*) + W$.

We set $d = 50$, $n = 25$ and we will vary the number of tasks T. The rank of Θ^* is set to 5 and the sparsity of Γ^* is set to 8% of the entries in figures 3.1 and 3.2. We will vary the rank and the sparsity in figure 3.3 while keeping $d = 50$, $n = 25$, $T = 200$. We set μ_d and λ_d such that

$$\lambda_d = \propto \left(16\nu\sigma_{\max}\sqrt{n}(\sqrt{d}+\sqrt{T})\right) \quad \text{and} \quad \mu_d = \propto \left(16\nu k_{\max}\sqrt{n\log\left(dT\right)} + \frac{4\alpha\gamma\sqrt{n}}{\sqrt{dT}}\right)$$

as we proved in corollary 2.7.1. In the following experiment section, we used the Proximal Gradient Descent method to recover the estimates since we did not observe a significant time improvement when using the Frank-Wolfe-Thresholding algorithm 3 when d and n take the above-mentioned values. Notice, however, that each method is equivalent since we always reach (if we train for a sufficiently large number of iterations) a global minimum, for which we have the theoretical guarantees of corollary 2.7.1. Letting $(\hat{\Theta}, \hat{\Gamma})$ be the estimates and (Θ^*, Γ^*) be the true matrices, we will define

$$\text{normalized ratio} = \frac{||\hat{\Theta} - \Theta^*||_F^2 + ||\hat{\Gamma} - \Gamma^*||_F^2}{||\Theta^*||_F^2 + ||\Gamma^*||_F^2}$$

This quantity captures the relative error of the estimates normalized with respect to the true parameters. Notice that the zero matrices attain a normalized ratio of 1. Therefore, it is desirable to have a normalized ratio well below 1 to conclude that the estimates are reliable. We show here how the normalized ratio varies with the number of tasks for our low-rank + entrywise-sparse model.

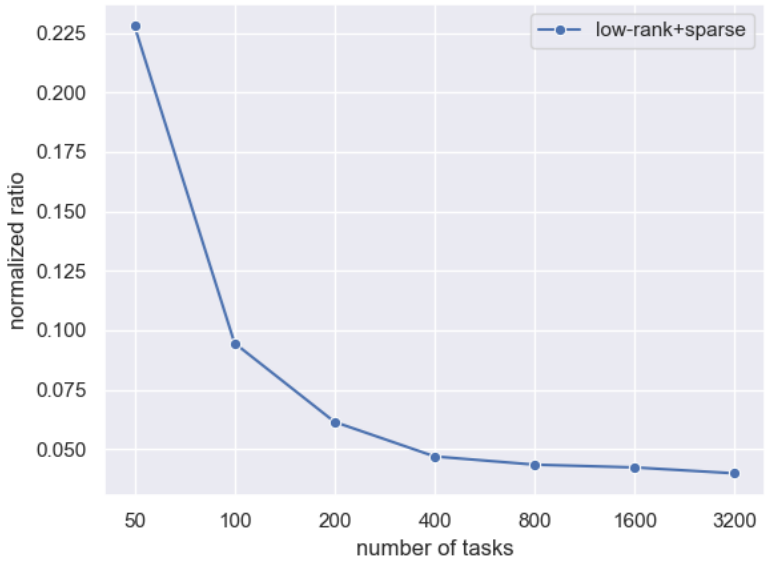

Figure 3.1: Normalized Ratio low-rank+sparse

When the number of tasks gets significant, we have a ratio very close to zero. We wanted to compare our normalized ratio with the one of [Pal et al. (2022)]. However, they could not release their code, and, despite we are trying to implement it ourselves, we could not tune properly the hyperparameters and therefore the comparison wouldn't have been fair. However, we can still highlight here the main differences, our algorithm assures global convergence while their algorithm can only reach a local minimum. Nevertheless, their algorithm can be directly extended to non-linear models (e.g. Neural Networks) while ours is still constrained to linear models. Finally, our algorithm (as described here) is mostly for Multi-task learning since we are interested in recovering the true parameters Θ^* and Γ^*. On the other hand, in their case it is more towards Meta-Learning, since they aim to make good predictions in new unseen tasks. Regarding the time complexity their algorithm has a complexity per iteration of $O((dr)^3 + (nT)(dr)^2)$ (or $O(nTdr)$ if a closed form solution for U is replaced by a gradient step) while ours has a complexity per iteration of $O(\min\{dT^2, d^2T\})$ (or $O(dTn)$ if FWT algorithm 3 is used).
In addition, we define similarly to [Boursier et al. (2022)]

$$\text{Normalized Frobenius Distance} = \frac{||\hat{\Theta} + \hat{\Gamma} - \Theta^* - \Gamma^*||_F}{\sqrt{T}}$$

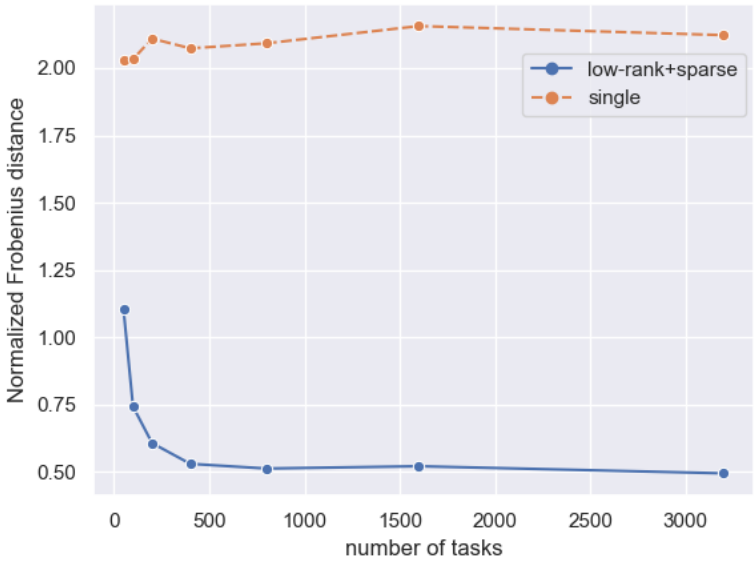

Figure 3.2: Normalized Frobenius distance

In Figure 3.2, we compare the normalized Frobenius distance for our "low-rank+sparse" model to a model ("single") that fits a different least squares estimator for each task. As we can see, our model gets better and better when T is increased while, of course, this does not happen for the "single" model.

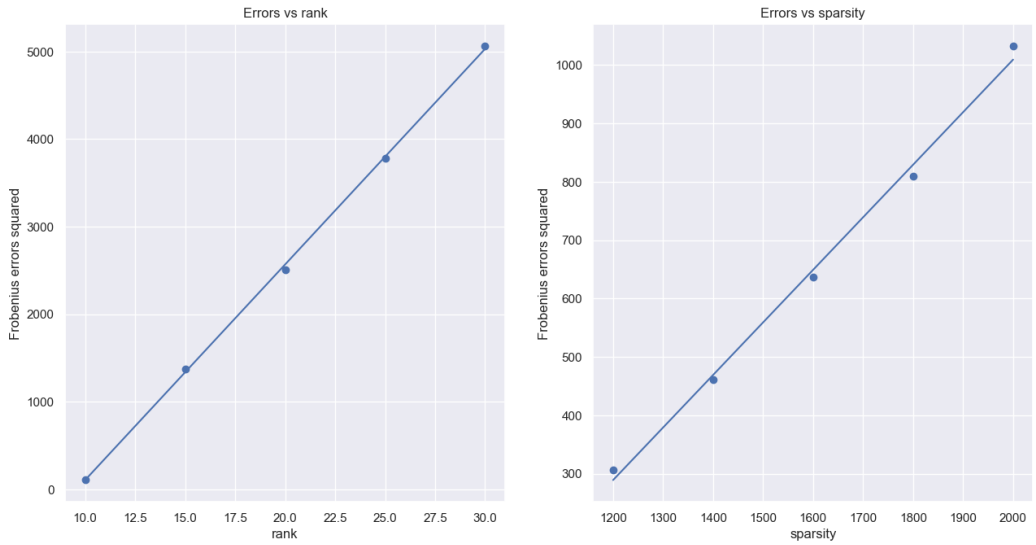

Figure 3.3: parameters dependence

Figure 3.3 shows how the experimental results match the theoretical bounds of equation (2.15) since a linear dependence of the Frobenius errors squared in both r and s is observed for our model.

4. Discussion

In this work, we studied the robust version of the already well-known multi-task linear regression model assuming the parameters to lie on a lower-dimensional subspace. We showed how to reformulate the problem in a convex manner, with low-rank and sparse regularizers. We provided bounds on the estimation error if the sparse regularizer is the 1-norm or the 2-1-norm. We showed how to solve the optimization problem with proximal gradient descent and we adapted the Frank-Wolfe Thresholding method to it for the L-1 norm. Finally, we tested our model on synthetic datasets on which the latter outperforms a model which learns each task independently when the number of tasks grows. Finally, we proved empirically some of the dependencies of the errors bound previously defined.
Our work heavily relies on the restricted strong convexity assumption at least on a cone $\mathcal{C}$ where the errors on the parameters lie. It could be interesting to prove formally that this assumption holds in our setting. Moreover, we hoped to compare our method with the one of [Pal et al. (2022)] but the code was not available yet. It would be interesting to study which algorithm performs better for different values of d, T, n. Finally, future works could also require to analyze which of the two penalizations (L-1 or L-21) would perform better on real datasets and to extend our work to the Meta-Learning setting.

A. Appendix

The following material is complementary to the main work and aims to give further details to the reader. It includes proofs of the lemma used to prove the Theorem 2.5 and a complete discussion on the Frank-Wolfe Algorithm to optimize problem 3.1.

A.1 Proofs of the lemmas

We include here the proof of the two lemmas used when proving the main theorem. We start by stating and proving the first lemma.

Lemma A.1. *For any $r = 1, 2, \dots, \min\{d, T\}$ there is a decomposition $\hat{\Delta}^\Theta = \hat{\Delta}^\Theta_A + \hat{\Delta}^\Theta_B$ such that:*

1. *The decomposition satisfies*

$$rank(\hat{\Delta}^\Theta_A) \leq 2r \quad and \quad (\hat{\Delta}^\Theta_A)^T \hat{\Delta}^\Theta_B = (\hat{\Delta}^\Theta_B)^T \hat{\Delta}^\Theta_A = 0 \tag{A.1}$$

2.

$$\begin{aligned} \mathcal{Q}(\Theta^*, \Gamma^*) - \mathcal{Q}(\Theta^* + \hat{\Delta}^\Theta, \Gamma^* + \hat{\Delta}^\Gamma) \leq \mathcal{Q}(\hat{\Delta}^\Theta_A, \hat{\Delta}^\Gamma_{\mathbb{M}}) - \mathcal{Q}(\hat{\Delta}^\Theta_B, \hat{\Delta}^\Gamma_{\mathbb{M}^\perp}) \\ + 2 \sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{2\mu_d}{\lambda_d} \mathcal{R}(\Gamma^*_{\mathbb{M}^\perp}) \end{aligned} \tag{A.2}$$

3. *under the assumption on the regularization parameters of the main theorem 2.5, $\hat{\Delta}^\Theta$ and $\hat{\Delta}^\Gamma$ satisfy*

$$\mathcal{Q}(\hat{\Delta}^\Theta_B, \hat{\Delta}^\Gamma_{\mathbb{M}}) \leq 3\mathcal{Q}(\hat{\Delta}^\Theta_A, \hat{\Delta}^\Gamma_{\mathbb{M}^\perp}) + 4\{ \sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d} \mathcal{R}(\Gamma^*_{\mathbb{M}^\perp}) \} \tag{A.3}$$

Proof. 2.9 was established in lemma 3.4 of [Recht et al. (2010)].
Let $\Theta^* = U \left[\begin{array}{c|c} \Sigma & 0 \\ \hline 0 & 0 \end{array}\right] V^T$ be the SVD decomposition of Θ^*, let

$$\hat{B} = U^T(\hat{\Delta}^\Theta)V = \left[\begin{array}{c|c} \hat{B_{11}} & \hat{B_{12}} \\ \hline \hat{B_{21}} & \hat{B_{22}} \end{array}\right]$$

with $\hat{B}_{11} \in \mathbb{R}^{r\times r}, \hat{B}_{12} \in \mathbb{R}^{r\times(T-r)}, \hat{B}_{21} \in \mathbb{R}^{(T-r)\times r}$ and $\hat{B}_{22} \in \mathbb{R}^{(T-r)\times(T-r)}$. Define now

$$\hat{\Delta}^\Theta_A = U \left[\begin{array}{c|c} \hat{B_{11}} & \hat{B_{12}} \\ \hline \hat{B_{21}} & 0 \end{array}\right] V^T \quad \text{and} \quad \hat{\Delta}^\Theta_B = U \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \hat{B_{22}} \end{array}\right] V^T$$

The decomposition satisfies the lemma. We will now prove 2.10.

$$
\begin{aligned}
&\mathcal{Q}(\Theta^*, \Gamma^*) - \mathcal{Q}(\Theta^* + \hat{\Delta}^\Theta, \Gamma^* + \hat{\Delta}^\Gamma) \\
&= \mathcal{Q}(\Theta_A^* + \Theta_B^*, \Gamma_{\mathbb{M}}^* + \Gamma_{\mathbb{M}^\perp}^*) - \mathcal{Q}(\Theta_A^* + \hat{\Delta}_A^\Theta + \Theta_B^* + \hat{\Delta}_B^\Theta, \Gamma_{\mathbb{M}}^* + \Gamma_{\mathbb{M}^\perp}^* + \hat{\Delta}_{\mathbb{M}}^\Gamma + \hat{\Delta}_{\mathbb{M}^\perp}^\Gamma) \\
&\overset{(1)}{=} \mathcal{Q}(\Theta_A^*, \Gamma_{\mathbb{M}}^*) + \mathcal{Q}(\Theta_B^*, \Gamma_{\mathbb{M}^\perp}^*) - \mathcal{Q}(\Theta_A^* + \hat{\Delta}_A^\Theta, \Gamma_{\mathbb{M}}^* + \hat{\Delta}_{\mathbb{M}}^\Gamma) - \mathcal{Q}(\Theta_B^* + \hat{\Delta}_B^\Theta, \Gamma_{\mathbb{M}^\perp}^* + \hat{\Delta}_{\mathbb{M}^\perp}^\Gamma) \\
&\overset{(2)}{\leq} \mathcal{Q}(\hat{\Delta}_A^\Theta, \hat{\Delta}_{\mathbb{M}}^\Gamma) + \mathcal{Q}(\Theta_B^*, \Gamma_{\mathbb{M}^\perp}^*) - \mathcal{Q}(\Theta_B^* + \hat{\Delta}_B^\Theta, \Gamma_{\mathbb{M}^\perp}^* + \hat{\Delta}_{\mathbb{M}^\perp}^\Gamma) \\
&\overset{(3)}{\leq} \mathcal{Q}(\hat{\Delta}_A^\Theta, \hat{\Delta}_{\mathbb{M}}^\Gamma) + 2\mathcal{Q}(\Theta_B^*, \Gamma_{\mathbb{M}^\perp}^*) - \mathcal{Q}(\hat{\Delta}_B^\Theta, \hat{\Delta}_{\mathbb{M}^\perp}^\Gamma) \\
&= \mathcal{Q}(\hat{\Delta}_A^\Theta, \hat{\Delta}_{\mathbb{M}}^\Gamma) - \mathcal{Q}(\hat{\Delta}_B^\Theta, \hat{\Delta}_{\mathbb{M}^\perp}^\Gamma) + 2\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{2\mu_d}{\lambda_d}\mathcal{R}(\Gamma_{\mathbb{M}^\perp}^*)
\end{aligned}
$$

Where in (1) we used the decomposability of the two $||.||_*$ and $\mathcal{R}(.)$, in (2) we used the triangle inequality applied to $\mathcal{Q}(\Theta_A^*, \Gamma_{\mathbb{M}}^*)$ and in (3) we used the inverse triangle inequality applied to $\mathcal{Q}(\Theta_B^* + \hat{\Delta}_B^\Theta, \Gamma_{\mathbb{M}^\perp}^* + \hat{\Delta}_{\mathbb{M}^\perp}^\Gamma)$.
Finally, we prove 2.11. This part is a generalization of lemma 1 in [Negahban et al. (2012)]. This lemma applies to optimization problems of the form

$$\min_{\theta \in \mathbb{R}^p} \{\mathcal{L}(\theta) + \gamma_n r(\theta)\}$$

and their lemma requires the regularization parameter γ_n to be chosen bigger or equal to $2r^*(\nabla\mathcal{L}(\theta^*))$. In our case, we can apply the lemma with

$$\mathcal{L}(\Theta, \Gamma) = \frac{1}{2}||Y - \mathcal{X}(\Theta + \Gamma)||_F^2 \quad r(\theta) = \mathcal{Q}(\Theta, \Gamma) = ||\Theta||_* + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma)$$

and with the regularization parameter $\lambda_d = \gamma_n$. Notice that $\mathcal{Q}(\Theta, \Gamma)$ is still decomposable because both the nuclear norm and $\mathcal{R}$ are decomposable. Therefore, to prove the lemma it is sufficient to show that $\lambda_d \geq 2\mathcal{Q}^*(\nabla\mathcal{L}(\Theta, \Gamma))$. It is easy to prove that

$$\mathcal{Q}^*(A, B) = ||A||_{\text{op}} + \frac{\lambda_d}{\mu_d}\mathcal{R}^*(B) \quad \text{and} \quad \nabla\mathcal{L}(\Theta^*, \Gamma^*) = -\mathcal{X}^*[W, W]^T$$

Therefore, we just need to prove that

$$\lambda_d \geq 2||\mathcal{X}^*(W)||_{\text{op}} + \frac{2\lambda_d}{\mu_d}\mathcal{R}^*(\mathcal{X}^*(W)) \tag{A.4}$$

By the choice of the regularization parameters $\frac{2\lambda_d}{\mu_d}\mathcal{R}^*(\mathcal{X}^*(W)) \leq \frac{\lambda_d}{2}$ and $\lambda_d \geq 4||\mathcal{X}^*(W)||_{\text{op}}$ which together show that A.4 is satisfied. □

We will now prove the second lemma

Lemma A.2. *Under the conditions of Theorem 2.5, we have the following bound*

$$\frac{1}{2}||\mathcal{X}(\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma)||_F^2 \geq \frac{\gamma}{4}(||\hat{\Delta}^\Theta||_F^2 + ||\hat{\Delta}^\Gamma||_F^2) - \frac{\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^\Theta, \hat{\Delta}^\Gamma) - 32\tau_n\{\sum_{j=r+1}^{h} \sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma_{\mathbb{M}^\perp}^*)\}^2 \tag{A.5}$$

Proof. By the RSC condition (2.8), we have

$$\frac{1}{2}||\mathcal{X}(\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma)||_F^2 - \frac{\gamma}{2}||\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma||_F^2 \geq -\tau_n\Phi^2(\hat{\Delta}^\Theta + \hat{\Delta}^\Gamma) \geq -\tau_n\mathcal{Q}^2(\hat{\Delta}^\Theta, \hat{\Delta}^\Gamma) \tag{A.6}$$

By definition of RSC and by definition of Φ. Now, we know that

$$\frac{\gamma}{2}\left(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma}||_F^2\right) - \frac{\gamma}{2}||\hat{\Delta}^{\Theta} + \hat{\Delta}^{\Gamma}||_F^2 = -\gamma\langle\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}\rangle$$

Moreover, we can use Holder to bound

$$\gamma|\langle\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}\rangle| \leq \gamma\mathcal{R}^*(\hat{\Delta}^{\Theta})\mathcal{R}(\hat{\Delta}^{\Gamma})$$

Furthermore,

$$\gamma\mathcal{R}^*(\hat{\Delta}^{\Theta}) \leq \gamma(\mathcal{R}^*(\hat{\Theta}) + \mathcal{R}^*(\Theta^*)) \leq \frac{2\alpha\gamma}{\kappa_d} \leq \frac{\mu_d}{2}$$

Where the first inequality follows from triangle inequality, the second for the feasibility of both $\hat{\Theta}$ and Θ^*, and the third by the choice of μ_d. Combining we get

$$\begin{aligned}
\frac{\gamma}{2}||\hat{\Delta}^{\Theta} + \hat{\Delta}^{\Gamma}||_F^2 &\geq \frac{\gamma}{2}(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma}||_F^2) - \frac{\mu_d}{2}\mathcal{R}(\hat{\Delta}^{\Gamma}) \\
&\geq \frac{\gamma}{2}(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma}||_F^2) - \frac{\mu_d}{2}\mathcal{R}(\hat{\Delta}^{\Gamma}) - \frac{\lambda_d}{2}||\hat{\Delta}^{\Theta}||_* \\
&\geq \frac{\gamma}{2}(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma}||_F^2) - \frac{\lambda_d}{2}\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) \qquad \text{(A.7)}
\end{aligned}$$

Where the last inequality follows by the definition of $\mathcal{Q}$. Now, we upper bound the term $\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma})$ using the triangle inequality:

$$\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) \leq \mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + \mathcal{Q}(\hat{\Delta}^{\Theta}_B, \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp})$$

Now, using 2.11, we get

$$\mathcal{Q}(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) \leq 4\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}^\perp}) + 4\{\sum_{j=r+1}^{h}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}$$

Keeping in mind that the rank of $\hat{\Delta}^{\Theta}_A$ is at most $2r$,

$$\begin{aligned}
\lambda_d\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) = \lambda_d||\hat{\Delta}^{\Theta}_A||_* + \mu_d||\hat{\Delta}^{\Gamma}_{\mathbb{M}}||_F &\leq \sqrt{2r}\lambda_d||\hat{\Delta}^{\Theta}_A||_F + \Psi(\mathbb{M})\mu_d||\hat{\Delta}^{\Gamma}_{\mathbb{M}}||_F \\
&\leq \sqrt{2r}\lambda_d||\hat{\Delta}^{\Theta}||_F + \Psi(\mathbb{M})\mu_d||\hat{\Delta}^{\Gamma}||_F
\end{aligned}$$

Now, putting it all together

$$\begin{aligned}
\tau_n\mathcal{Q}^2(\hat{\Delta}^{\Theta}, \hat{\Delta}^{\Gamma}) &\leq \tau_n\left(4\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}}) + 4\{\sum_{j=r+1}^{d}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}\right)^2 \\
&\leq 2\tau_n\left(16\mathcal{Q}(\hat{\Delta}^{\Theta}_A, \hat{\Delta}^{\Gamma}_{\mathbb{M}})^2 + 16\{\sum_{j=r+1}^{d}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2\right) \\
&\leq 2\tau_n\left(16(\sqrt{2r}||\hat{\Delta}^{\Theta}||_F + \Psi(\mathbb{M})\frac{\mu_d}{\lambda_d}||\hat{\Delta}^{\Gamma}||_F)^2 + 16\{\sum_{j=r+1}^{d}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2\right) \\
&\leq 128\tau_n r||\hat{\Delta}^{\Theta}||_F^2 + 64\Psi(\mathbb{M})^2\frac{\mu_d}{\lambda_d}||\hat{\Delta}^{\Gamma}||_F^2 + 32\tau_n\{\sum_{j=r+1}^{d}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2 \\
&\overset{(1)}{\leq} \frac{\gamma}{4}\left(||\hat{\Delta}^{\Theta}||_F^2 + ||\hat{\Delta}^{\Gamma}||_F^2\right) + 32\tau_n\{\sum_{j=r+1}^{d}\sigma_j(\Theta^*) + \frac{\mu_d}{\lambda_d}\mathcal{R}(\Gamma^*_{\mathbb{M}^\perp})\}^2 \qquad \text{(A.8)}
\end{aligned}$$

Where we used repeatedly that $(a+b)^2 \leq 2(a^2+b^2)$ and in (1) we used the hypothesis on τ_n of theorem 2.5. Finally, inserting A.7 and A.8 in A.6 yields the claim. □

A.2 Frank-Wolfe for robust MTL

The goal of the section is to show how we can leverage the use of the Frank-Wolfe algorithm to solve the following optimization problem more efficiently

$$\min_{\Theta\in\mathbb{R}^{d\times T},\ \Gamma\in\mathbb{R}^{d\times T}} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2+\lambda_d||\Theta||_*+\mu_d||\Gamma||_1$$

The idea is to reformulate the problem equivalently as

$$\min_{\Theta\in\mathbb{R}^{d\times T},\ \Gamma\in\mathbb{R}^{d\times T},t_l>0,t_s>0} g(\Theta,\Gamma,t_l,t_s) \coloneqq \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2+\lambda_d t_l+\mu_d t_s \tag{A.9}$$
$$\text{such that } \ ||\Theta||_*\leq t_l,\ ||\Gamma||_1\leq t_s$$

Calling g the objective function, it is easy to show that the objective function has Lipschitz gradient. In fact, we have already proved that $\frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2$ had Lipschitz gradient as a function of (Θ,Γ), therefore it still has Lipschitz gradient as a function of (Θ,Γ,t_l,t_s). Now, noticing that the objective is simply the sum of the latter and a linear function we have that g is gradient Lipschitz. However, the Frank-Wolfe method still cannot be applied to the reformulation above, since the region of feasibility is unbounded. Nevertheless, we can make the region bounded with a simple observation. Notice that $(\Theta,\Gamma,t_l,t_s)=(0,0,0,0)$ is a feasible solution to the reformulation A.9. Plugging them in, the value of the objective function of the problem, in this case, is $\frac{1}{2}||Y||_F^2$ and, therefore, this tells us that $t_l\leq\frac{1}{2\lambda_d}||Y||_F^2$ and $t_s\leq\frac{1}{2\mu_d}||Y||_F^2$. As a consequence, calling $U_L=\frac{1}{2\lambda_d}||Y||_F^2$ and $U_S=\frac{1}{2\mu_d}||Y||_F^2$, we can reformulate the problem as

$$\min_{\Theta,\Gamma,t_l,t_s} \frac{1}{2}||Y-\mathcal{X}(\Theta+\Gamma)||_F^2+\lambda_d t_l+\mu_d t_s \tag{A.10}$$
$$\text{s.t. } \ ||\Theta||_*\leq t_l\leq U_L,\ ||\Gamma||_1\leq t_s\leq U_S$$

The following proposition shows that the feasible set has a bounded diameter.

Proposition A.3. *The feasible set* $\mathcal{D}=\{(\Theta,\Gamma,t_l,t_s):||\Theta^*||_*\leq t_l\leq U_L,||\Gamma||_1\leq t_s\leq U_s\}$ *has diameter bounded by* $\sqrt{5}\sqrt{U_L^2+U_S^2}$.

Proof. We refer to lemma 4.2 of [Mu et al. (2016)] for a proof of this fact, the proof still holds in our setting since the feasible region is the same. □

It is straightforward to prove that the feasible region is also convex and compact. As a consequence, we could apply the Frank-Wolfe Algorithm directly to A.10 to obtain convergence in $O(\frac{1}{\epsilon})$ iterations. We will now analyze how the sequence of iterates is created and why the cost per iteration is smaller compared to Proximal-methods. If we let the sequence of iterates be $x^k=(\Theta^k,\Gamma^k,t_l^k,t_s^k)$ and the solution to the k-th call of the Linear Minimization Oracle $v^k=(V_\Theta^k,V_\Gamma^k,V_l^k,V_s^k)$, naming $H_k=-\mathcal{X}^*(Y-\mathcal{X}(\Theta^k+\Gamma^k)$ we have that

$$v^k\in\arg\min_{v\in\mathcal{D}}\langle H_k,V_\Theta+V_\Gamma\rangle+\lambda_d V_l+\mu_d V_s$$

And again the problem can be decoupled into two subproblems.

$$(V_\Theta,V_l)\in\arg\min_{||V_\Theta||_*\leq V_l\leq U_L} g_L(V_\Theta,V_l)\coloneqq\langle H_k,V_\Theta\rangle+\lambda_d V_l \tag{A.11}$$
$$(V_\Gamma,V_s)\in\arg\min_{||V_\Gamma||_1\leq V_s\leq U_S} g_S(V_\Gamma,V_s)\coloneqq\langle H_k,V_\Gamma\rangle+\mu_d V_s \tag{A.12}$$

We consider first the problem A.11, since g_L is a homogeneous function, i.e. $g_L(\alpha V_\Theta, \alpha V_l) = \alpha g_L(V_\Theta, V_l)$, its optimal value is $g_L(V_\Theta^k, V_l^k) = V_l^k \hat{g_L}(D_l^k)$ with

$$D_l^k \in \underset{||D_l||_* \leq 1}{\arg\min}\, \hat{g_L}(D_l) \coloneqq \langle H_k, D_l \rangle + \lambda_d \tag{A.13}$$

Therefore, when $\hat{g}_L(D_l^k) > 0$, $V_l^k = 0$, if $\hat{g}_L(D_l^k) < 0$, $V_l^k = U_L$. As a consequence, the solution to A.11 can be summarized as

$$(V_\Theta^k, V_l^k) \in \begin{cases} (0,0) & \text{if } \hat{g}_L(D_l^k) > 0 \\ \text{conv}\{(0,0), U_L(D_l^k, 1)\} & \text{if } \hat{g}_L(D_l^k) = 0 \\ U_L(D_l^k, 1) & \text{if } \hat{g}_L(D_l^k) < 0 \end{cases}$$

Where Conv stands for convex combination. Similarly, the solution for A.12 is the following

$$(V_\Gamma^k, V_s^k) \in \begin{cases} (0,0) & \text{if } \hat{g}_S(D_s^k) > 0 \\ \text{conv}\{(0,0), U_S(D_s^k, 1)\} & \text{if } \hat{g}_S(D_s^k) = 0 \\ U_S(D_s^k, 1) & \text{if } \hat{g}_S(D_s^k) < 0 \end{cases}$$

Where

$$D_s^k \in \underset{||D_s||_1 \leq 1}{\arg\min}\, \hat{g_S}(D_s) \coloneqq \langle H_k, D_s \rangle + \mu_d \tag{A.14}$$

Notice now that, $D_l^k = -uv^T$ where u and v are the left and right singular vectors corresponding to the largest singular value of H_k. Therefore, contrarily to the full-SVD requested by the proximal method, we just need to compute the top-1 largest singular vectors which have a smaller cost of $O(dT)$ and can be done efficiently by the power method. Similarly, $D_s^k = -\text{sng}(H_{k_{i^*j^*}}) e_{i^*} e_{j^*}^T$ with $(i^*, j^*) \in \text{argmax}_{(i,j)} |H_{k_{ij}}|$ which again can be computed in $O(dT)$. The cost per iteration is therefore dominated by the computation of H_k which cost $O(dTn)$. Therefore, running the Frank-Wolfe algorithm 2 yields convergence in $O(\frac{1}{\epsilon})$ iterations, with a cost per iteration of $O(dTn)$. We formalize the rate of convergence in this proposition.

Proposition A.4. *Let $x^* = (\Theta^*, \Gamma^*, t_l^*, t_s^*)$ be an optimal solution to the problem A.9, let $x^k = (\Theta^k, \Gamma^k, t_l^k, t_s^k)$ be a sequence produced by Algorithm 2, then*

$$g(x^k) - g(x^*) \leq \frac{20(U_S^2 + U_L^2)}{k+2}$$

Proof. It follows from the convergence analysis of the Frank-Wolfe Algorithm. □

Despite the algorithm being theoretically attractive for our case, it is slow in practice. This is mainly due to two reasons:

1. The update of the sparse matrix Γ differs from the previous one by at most a single entry.

2. The rate of convergence depends on U_L and U_S which may be huge in practice.

To solve these issues [Mu et al. (2016)] propose a different algorithm combining the Frank-Wolfe Algorithm with the proximal method to overcome the problem of the single entry update and decreasing U_L and U_S at each iteration, hoping for a better convergence rate. They call their algorithm Frank-Wolfe-Thresholding (FWT). FWT is shown in Algorithm 3. The main differences compared to the Frank-Wolfe algorithm 2 are the following:

Algorithm 2: Frank-Wolfe

Data: $\Theta^0 = 0$, $\Gamma^0 = 0$, $t_l = t_s = 0$, regularization parameters (λ_d, μ_d), max-iters n

$t = 0$
$U_L = g(\Theta^0, \Gamma^0, t_l^0, t_s^0)/\lambda_d$
$U_S = g(\Theta^0, \Gamma^0, t_l^0, t_s^0)/\mu_d$
while $k \leq n$ **do**
  $H_k = -\mathcal{X}^*(Y - \mathcal{X}(\Theta^k + \Gamma^k))$
  Compute $D_l^k \in \arg\min_{||D_l||_* \leq 1} \hat{g_L}(D_l) \coloneqq \langle H_k, D_l \rangle + \lambda_d$
  Compute $D_s^k \in \arg\min_{||D_s||_1 \leq 1} \hat{g_S}(D_s) \coloneqq \langle H_k, D_s \rangle + \mu_d$
  if $\lambda_d \geq -\langle H_k, D_l^k \rangle$ **then**
    $V_\Theta^k = 0, V_t^k = 0;$
  else
    $V_\Theta^k = U_L D_l^k, V_l^k = U_L$;
  end
  if $\mu_d \geq -\langle H_k, D_s^k \rangle$ **then**
    $V_\Gamma^k = 0, V_s^k = 0;$
  else
    $V_\Gamma^k = U_S D_s^k, V_s^k = U_S$;
  end
  $\gamma = \frac{2}{k+2}$
  $\Theta^{k+1} = (1-\gamma)\Theta^k + \gamma V_\Theta^k$
  $t_l^{k+1} = (1-\gamma)t_l^k + \gamma V_l^k$
  $\Gamma^{k+1} = (1-\gamma)\Gamma^k + \gamma V_\Gamma^k$
  $t_s^{k+1} = (1-\gamma)t_s^k + \gamma V_s^k$
  $k = k + 1$
end
return *The matrices* Θ^n, Γ^n

- Instead of using the fixed step size $\gamma = \frac{2}{k+2}$ used in the Frank-Wolfe, we select it with a line search. It guarantees that the new iterates $x^{k+\frac{1}{2}}$ are such that $g(x^{k+\frac{1}{2}}) \leq g(x^k)$. This fact will be crucial in the convergence analysis.
- We perform an additional gradient proximal step on Γ, after the usual step of the Frank-Wolfe algorithm, to avoid updating a single entry at the time.
- We update U_L^k and U_S^k, since g is decreasing with the iterations, also U_L^k and U_S^k are decreasing with the iterations, allowing faster convergence.

We can still formally prove convergence in $O(\frac{1}{\epsilon})$ iterations.

Proposition A.5. *Let r_L^* and r_S^* be the smallest radii such that*

$$\left\{(\Theta, \Gamma) \mid \frac{1}{2}||Y - \mathcal{X}(\Theta + \Gamma)||_F^2 + \lambda_d||\Theta||_* + \mu_d||\Gamma||_1 \leq \frac{1}{2}||Y||_F^2\right\} \subseteq \bar{B}(r_L^*) \times \bar{B}(r_S^*)$$

where $\bar{B}(r)$ is the closed ball of Frobenius-norm or radius r. Then for a sequence $\{\Theta^k, \Gamma^k, t_l^k, t_s^k\}$ generated by Algorithm 3, we have

$$g(\Theta^k, \Gamma^k, t_l^k, t_s^k) - g(\Theta^*, \Gamma^*, t_l^*, t_s^*) \leq \frac{\min\{4(t_l^* + r_l^*)^2 + 4(t_s^* + r_s^*)^2, 16(U_L^0)^2 + 16(U_S^0)^2\}}{k+2}$$

Proof. Refer to theorem 4.5 of [Mu et al. (2016)]. The proof still holds in our case as long as the stepsize γ is chosen as $1/L$ with L the Lipschitz constant of $\frac{1}{2}||Y - \mathcal{X}(\Theta + \Gamma)||_F^2$. □

Algorithm 3: Frank-Wolfe-Thresholding

Data: $\Theta_0 = 0$, $\Gamma_0 = 0$, $t_l = t_s = 0$, regularization parameters (λ_d, μ_d), max-iters n, learning rate γ

$t = 0$

$U_L^0 = g(\Theta^0, \Gamma^0, t_l^0, t_s^0)/\lambda_d$

$U_S^0 = g(\Theta^0, \Gamma^0, t_l^0, t_s^0)/\mu_d$

while $k \leq n$ **do**

  Same as lines 5-17 of Algorithm 2

  $(\Theta^{k+\frac{1}{2}}, \Gamma^{k+\frac{1}{2}}, t_l^{k+\frac{1}{2}}, t_s^{k+\frac{1}{2}})$ computed as

$$\arg\min_{\Theta,\Gamma,t_l,t_s} \frac{1}{2}||Y - \mathcal{X}(\Theta + \Gamma)||_F^2 + \lambda_d t_l + \mu_d t_s$$
$$\text{s.t.} \begin{pmatrix}\Theta \\ t_l\end{pmatrix} \in \text{conv}\left\{\begin{pmatrix}\Theta^k \\ t_l^k\end{pmatrix}, \begin{pmatrix}V_\Theta^k \\ V_l^k\end{pmatrix}\right\}, \begin{pmatrix}\Gamma \\ t_s\end{pmatrix} \in \text{conv}\left\{\begin{pmatrix}\Gamma^k \\ t_s^k\end{pmatrix}, \begin{pmatrix}V_\Gamma^k \\ V_s^k\end{pmatrix}\right\}$$

  $H_k = -\mathcal{X}^*(Y - \mathcal{X}(\Theta^{k+\frac{1}{2}} + \Gamma^{k+\frac{1}{2}}))$

  $\Gamma^{k+1} = S_{\gamma\lambda_d}(\Gamma^{k+\frac{1}{2}} - \gamma H_k)$

  $\Theta^{k+1} = \Theta^{k+\frac{1}{2}}$

  $t_s^{k+1} = ||\Gamma^{k+1}||_1$

  $t_l^{k+1} = t_l^{k+\frac{1}{2}}$

  $U_L^{k+1} = g(\Theta^{k+1}, \Gamma^{k+1}, t_l^{k+1}, t_s^{k+1})/\lambda_d$

  $U_S^{k+1} = g(\Theta^{k+1}, \Gamma^{k+1}, t_l^{k+1}, t_s^{k+1})/\mu_d$

  $k = k + 1$

end

return *The matrices* Θ^n, Γ^n

In this case, the cost per iteration is still $O(dTn)$ and is still dominated by the computation of H_k, we still need $O(\frac{1}{\epsilon})$ iterations to reach an objective error of ϵ but the algorithm is faster in practice.

Bibliography

Li Adeli, Meng. Joint sparse and low-rank regularized multitask multi-linear regression for prediction of infant brain development with incomplete data. *Med Image Comput Comput Assist Interv*, 2017. doi: 10.1007/978-3-319-66182-7_5.

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2), apr 2012. doi: 10.1214/12-aos1000. URL https://doi.org/10.1214%2F12-aos1000.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, 2005.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2:183–202, 2009.

Martin Jaggi Bernd Gartner. Lectures in optimization for machine learning, 2020. https://github.com/epfml/OptML_course/blob/master/lecture_notes/lecture-notes.pdf.

Etienne Boursier, Mikhail Konobeev, and Nicolas Flammarion. Trace norm regularization for multi-task learning with scarce data, 2022. URL https://arxiv.org/abs/2202.06742.

Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen. A singular value thresholding algorithm for matrix completion, 2008. URL https://arxiv.org/abs/0810.3286.

A. Chambolle, R.A. De Vore, Nam-Yong Lee, and B.J. Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Transactions on Image Processing*, 7(3):319–335, 1998. doi: 10.1109/83.661182.

Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. KDD '11, page 42–50, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450308137. doi: 10.1145/2020408.2020423. URL https://doi.org/10.1145/2020408.2020423.

Jianhui Chen, Ji Liu, and Jieping Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM Trans. Knowl. Discov. Data*, 5(4), feb 2012. ISSN 1556-4681. doi: 10.1145/2086737.2086742. URL https://doi.org/10.1145/2086737.2086742.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition, 2013. URL https://arxiv.org/abs/1310.1531.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. URL https://arxiv.org/abs/1703.03400.

Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: A convex formulation, 2008. URL https://arxiv.org/abs/0809.2085.

Ali Jalali, Pradeep Ravikumar, Sujay Sanghavi, and Chao Ruan. A dirty model for multi-task learning. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 964–972, Red Hook, NY, USA, 2010. Curran Associates Inc.

Guangcan Liu and Shuicheng Yan. Active Subspace: Toward Scalable Low-Rank Learning. *Neural Computation*, 24(12):3371–3394, 12 2012. ISSN 0899-7667. doi: 10.1162/NECO_a_00369. URL https://doi.org/10.1162/NECO_a_00369.

Pim Moeskops, Jelmer M. Wolterink, Bas H. M. van der Velden, Kenneth G. A. Gilhuijs, Tim Leiner, Max A. Viergever, and Ivana Iš gum. Deep learning for multi-task medical image segmentation in multiple modalities. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 478–486. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46723-8_55. URL https://doi.org/10.1007%2F978-3-319-46723-8_55.

Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank–wolfe meets proximal methods. *SIAM Journal on Scientific Computing*, 38(5):A3291–A3317, jan 2016. doi: 10.1137/15m101628x. URL https://doi.org/10.1137%2F15m101628x.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4), nov 2012. doi: 10.1214/12-sts400. URL https://doi.org/10.1214%2F12-sts400.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms, 2018. URL https://arxiv.org/abs/1803.02999.

Tae-Hyun Oh, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Fast randomized singular value thresholding for low-rank optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):376–391, 2018. doi: 10.1109/TPAMI.2017.2677440.

Soumyabrata Pal, Prateek Varshney, Prateek Jain, Abhradeep Guha Thakurta, Gagan Madan, Gaurav Aggarwal, Pradeep Shenoy, and Gaurav Srivastava. Private and efficient meta-learning with low rank and sparse decomposition, 2022. URL https://arxiv.org/abs/2210.03505.

Ting Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM Journal on Optimization*, 20: 3465–3489, 01 2010. doi: 10.1137/090763184.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml, 2019. URL https://arxiv.org/abs/1909.09157.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, jan 2010. doi: 10.1137/070697835. URL https://doi.org/10.1137%2F070697835.

Philippe Rigollet and Jan-Christian Hutter. Lecture notes on high dimensional statistics, 2017. https://math.mit.edu/~rigollet/PDFs/RigNotes17.pdf.

Alessandro Rinaldo. Lectures in advanced statistical theory, 2019. https://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures/Feb21_Shenghao.pdf.

Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2), apr 2011. doi: 10.1214/10-aos860. URL https://doi.org/10.1214%2F10-aos860.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations, 2020. URL https://arxiv.org/abs/2002.11684.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2019. URL https://arxiv.org/abs/1911.02685.