Robust Multi-task Regression with Grossly Corrupted Observations

Huan Xu
Department of Mechanical Engineering
National University of Singapore
mpexuh@nus.edu.sg

Chenlei Leng
Department of Statistics and Applied Probability
National University of Singapore
stalc@nus.edu.sg

Abstract

We consider the multiple-response regression problem, where the response is subject to *sparse gross errors*, in the high-dimensional setup. We propose a tractable regularized M-estimator that is robust to such error, where the sum of two individual regularization terms are used: the first one encourages row-sparse regression parameters, and the second one encourages a sparse error term. We obtain non-asymptotical estimation error bounds of the proposed method. To the best of our knowledge, this is the first analysis of the robust multi-task regression problem with gross errors.

1 Introduction

The past decade has witnessed a surge of research interest in analyzing high-dimensional data – data sets where the ambient dimension of the problem p is either close to or even substantially larger than the sample size n, due to a broad array of applications (e.g., Tibshirani, 1996; Candès et al., 2006; Candès & Tao, 2007; Donoho, 2006; Wainwright, 2009 and many others). Under such high-dimensional scaling, many standard statistical learning problems become ill-posed, and it is therefore vital to exploit any low-dimensional structure of the problems, such as sparsity (Candès et al., 2006; Tropp, 2006; Bickel et al., 2009), low-rank structure (Recht et al., 2010; Candès & Recht, 2009; Keshavan et al., 2010) or group sparsity (Yuan & Lin, 2006; Bach, 2008).

The focus of this paper is multi-task learning (Caruana, 1997; Argyriou et al., 2008), and specifically multiple-response regression. Here, we have $q > 1$ response variables to regress, and a common set of p covariates (i.e., features). Typically, the q tasks are believed to share certain structures, so that the performance of each learning task can be improved by exploiting this "intrinsic relatedness" while learning these tasks together. In particular, the setting we focus on is where the response variables have simultaneously sparse structure: each task involves a sparse set of relevant features, and there is a large overlap of these relevant features across the different tasks. This "simultaneous sparsity" model arises naturally in variety of applications including sparse signal recovery (Tropp et al., 2006), graphical models (Ravikumar et al., 2010) and learning with kernels (Bach, 2008). As standard, we represent the multiple regression parameters as a matrix Θ, where each column corresponds to a task, and each row to a feature. Because of the simultaneous sparse structure, one wants the matrix to be row-sparse (i.e., most rows are all-zero), which can be achieved using either the ℓ_1/ℓ_2 norm (Argyriou et al., 2008) or the ℓ_1/ℓ_∞ norm (Negahban & Wainwright, 2008) as a regularizer.

While powerful, this paradigm can be vulnerable to observation errors on the response variables. Particularly harmful are gross errors that may only affect a few observations, but in an otherwise uncontrolled and even adversarial manner. We address precisely this problem. Our approach is indeed intuitive: without gross errors, one would expect that $Y \approx X\Theta$ where Y is the responses and X the covariates. Hence, with gross errors the response should satisfy that $Y \approx X\Theta + G$, where the unknown matrix G is the gross error. Since the regression parameter Θ is row-sparse, and the gross error G is sparse, it is natural to consider the following formulation

$$\text{Minimize:}_{\Theta,G} \quad \|Y - X\Theta - G\|_F^2 + \lambda\|\Theta\|_{1,2} + \rho\|G\|_1,$$

where $\|\Theta\|_{1,2} = \sum_i \|\Theta_{i,\cdot}\|_2$ is the sum of the ℓ_2 norm of the rows, and $\|G\|_1 = \sum_{i,j} |G_{ij}|$ is the vector ℓ_1 norm of G. We show in this paper, through both theoretical analysis and numerical simulation, that this

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

intuitive formulation well addresses the multi-task regression with gross-error problem. To the best of our knowledge, this is the fist analysis of such problem.

Before concluding the introduction, we discuss some related work. A first thought to tackle the proposed problem would be to use regular $\ell_{1,2}$ based multi-task regression but with a robust ℓ_1 loss, which down-weights the effects of the outliers. While this method may provide an alternative way to estimate Θ, it does not attempt to estimate the outlier matrix G. Note that the latter is often of interest in application such as finance or computer vision. Using ℓ_1 norm to correct gross error is not a new idea. In Candès et al. (2005); Wright and Ma (2010), the authors considered the univariate response case (i.e., $q = 1$) with gross error. However, the formulation and the analysis is restricted to the noiseless case. Lee et al. (2011) studied adding case-specific parameters to the usual least-squares function in the univariate response case, which is indeed a special case of our formulation. They pointed out a connection of this formulation to the Huber's M-estimation. However, no theoretical results for their estimates were given. In a recent paper, Jalali et al. (2010) studied a related formulation in multi-task regression setup, where instead of having gross error in the response, they considered the case that the parameter matrix itself is subject to gross error. From a theoretic perspective, their analysis differed from ours as they focused on support recovery, which typically requires stronger assumptions.

Also relevant to this work is the recent study of decomposition of structured matrices from their summation. Chandrasekaran et al. (2011); Candès et al. (2011); Xu et al. (2010) studied the problem of decomposing a matrix into the sum of a low rank matrix and a sparse/column sparse matrix, in the noiseless case. The noisy case was investigated in Agarwal et al. (2011). While these works are close to ours in spirit, the regression setup investigated in this paper brings additional difficulties that need to be taken care of. Also, their models pose a different assumption that the coefficient matrix Θ is low-rank.

Notations: Most of our notations are standard. In addition, we frequently use subscripts to denote projecting a matrix to a subspace. For instance, if I is an index set of entries, then A_I stands for the matrix that set $A_{i,j}$ to zero for all $(i, j) \notin I$. The complementary of an index set I is denoted by $I^\perp$. We use $[a : b]$ to represent the set of all integers between a and b (inclusive).

2 Problem Formulation and the Main Results

We study the following multi-task regression problem with *gross errors*. We observe the response matrix $Y \in \mathbb{R}^{n\times q}$ and the covariate matrix $X \in \mathbb{R}^{n\times p}$ such that

$$Y = X\Theta^* + W + G^*.$$

Here, $\Theta^* \in \mathbb{R}^{p\times q}$ is an unknown linear relationship between the predictor and the response. Matrix $W \in \mathbb{R}^{n\times q}$ is the noise matrix, assumed to be "small"; and G^* is a matrix correspond to "gross" error. As discussed in the previous section, we assume that Θ^* is approximately *row sparse*, and G^* is entry-wise sparse. To estimate Θ^* and G^*, we propose the following M-estimator,

$$(\hat{\Theta}, \hat{G}) = \arg\min_{\Theta,G} \|Y - X\Theta - G\|_F^2 + \lambda\|\Theta\|_{1,2} + \rho\|G\|_1. \tag{1}$$

Here, $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_1$ is the entry-wise ℓ_1 norm, and $\|\cdot\|_{1,2}$ is the summation of ℓ_2 norm of *rows* of a matrix. Notice that such formulation is computational friendly, as it is a convex program involving only linear and quadratic functions.

Intuitively one would expect, due to the two regularization terms, that $\hat{\Theta}$ is row-sparse, and $\hat{G}$ is entry-wise sparse. Our hope is that this row sparse $\hat{\Theta}$ is close to Θ^*, and so is $\hat{G}$ to G^*. However, as in previous related work in matrix decomposition (Agarwal et al., 2011; Xu et al., 2010; Candès et al., 2011; Chandrasekaran et al., 2011), some additional assumptions are necessary. Indeed, suppose that $X\Theta^*$ itself is a sparse-matrix, and $G^* = -X\Theta^*$, then even in the noiseless case (i.e., $W = 0$, and consequently $Y = 0$), estimating Θ^* is impossible. To rule out such degenerate cases, we impose an "incoherence condition". We follow a similar line as Agarwal et al. (2011), and require that $\|X\Theta^*\|_\infty$ is not large: there exists $\tau \in \mathbb{R}$ such that

$$\|X\Theta^*\|_\infty \le \tau.$$

Consequently, we study in this paper a class of estimators as in Formulation (1), under the constraint that $\|X\Theta\|_\infty \le \tau$. Notice that the boundeness of infinity norm assumption is reasonable when the signal X^* is bounded, and can be weakened to the boundedness of the infinity norm with probability approaching one. In addition, we need the restricted eigenvalues of X to establish our results. Restricted maximum and minimum eigenvalue are defined as

$$\phi_{\min}(t) \triangleq \min_{\mathbf{z}\in\mathbb{R}^p\setminus\{0\}, \|\mathbf{z}\|_0\le t} \frac{\|X\mathbf{z}\|_2^2}{\|\mathbf{z}\|_2^2};$$

$$\phi_{\max}(t) \triangleq \max_{\mathbf{z}\in\mathbb{R}^p\setminus\{0\}, \|\mathbf{z}\|_0\le t} \frac{\|X\mathbf{z}\|_2^2}{\|\mathbf{z}\|_2^2}.$$

Main results

The main result of this paper is a non-asymptotic guarantee of the estimation error of the proposed method. That is, to bound

$$\hat{\Delta}^G \triangleq \hat{G} - G^*; \quad \text{and} \quad \hat{\Delta}^\Theta \triangleq \hat{\Theta} - \Theta^*.$$

Theorem 1. *Let A be a subset of row-index of Θ with $|A| = S$, and M be a subset of entry-index of G with $|M| = s$. Suppose $\lambda \geq 4\|X^\top W\|_{\infty,2}$ and $\rho \geq 4\|W\|_\infty + 8\tau$, then there exists a universal constant c such that for any $S' \in [1:p]$,*

$$\|\hat{\Delta}^G\|_F \leq 6\rho\sqrt{s} + \sqrt{6\rho\|G^*_{M^\perp}\|_1};$$

$$\|\hat{\Delta}^\Theta\|_F \leq \max\left\{ \frac{c[\lambda\sqrt{S} + \sqrt{\lambda\|\Theta^*_{A^\perp}\|_{1,2}} + \rho\sqrt{s} + \sqrt{\rho\|G^*_{M^\perp}\|_1}]}{\min(\kappa_1(S'),1)}, \right.$$
$$\frac{[6\rho\kappa_2(S')\sqrt{s}/\lambda - 3][6\rho\sqrt{s} + \sqrt{6\rho\|G^*_{M^\perp}\|_1}]}{\kappa_1(S')},$$
$$\left. \frac{16\kappa_2\left\{\|\Theta^*_{A^\perp}\|_{1,2} + \frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1\right\}}{\kappa_1(S')} \right\}.$$

Here, the coefficient $\kappa_1(S')$ and $\kappa_2(S')$ are defined as

$$\kappa_1(S') \triangleq \frac{\sqrt{\phi_{\min}(S+S')} - c_0\sqrt{\phi_{\max}(S')S/S'}}{1 + c_0\sqrt{S/S'}},$$
$$\kappa_2(S') \triangleq \frac{\sqrt{\phi_{\min}(S+S')} - c_0\sqrt{\phi_{\max}(S')S/S'}}{\sqrt{S'} + c_0\sqrt{S}} + \sqrt{\phi_{\max}(S')/S'}. \quad (2)$$

Note that A and M are arbitrary, and in particular need not coincide with the row-support of Θ^* and the support of G^*. Therefore, Theorem 1 also applies to the case where Θ^* is *approximately* row-sparse and G^* is *approximately* sparse. Similarly, we make no assumption on W either.

To illustrate Theorem 1, we provide results for the following specialized case:

Condition 1. The following holds: (1) Θ^* has S non-zero rows; (2) G^* has s non-zero elements; (3) W has i.i.d. $\mathcal{N}(0, \sigma^2/n)$ entries; (4) the ℓ_2 norm of each column of X is upper-bounded by 1.

Condition 2. There exists $S' \geq S$ such that $S'\phi_{\min}(S+S') \geq 16S\phi_{\max}(S')$.

Theorem 2. *Under Condition 1 and 2, let $\lambda_0 \triangleq \min\left(\frac{\sigma(\sqrt{q}+\sqrt{8\log pn})}{\sqrt{n}}, 7\sigma(1+\sqrt{q/n})\right)$. Set the parameters ρ and λ as $\rho = 16\sigma\sqrt{\frac{\log(nq)}{n}} + 8\tau$, and $\lambda = \max\left(4\lambda_0, \frac{\rho\sqrt{s}}{\sqrt{S}}\right)$. Then we have with probability $1 - 1/n^3$,*

$$\|\hat{\Delta}^G\|_F \leq c_1\sigma\sqrt{\frac{s\log(nq)}{n}} + c_1\tau,$$
$$\|\hat{\Delta}^\Theta\|_F \leq c_2\sigma\sqrt{\frac{s\log(nq)}{n}} + c_2\tau + c_2\sigma\sqrt{Sq/n} + c_2\sigma\sqrt{S}\min(\sqrt{\log(pn)/n}, 1).$$

Here, c_1, c_2 are upper-bounded by $c_0/\sqrt{\phi_{\min}(S+S')}$ for a universal constant c_0.

Note that Assumption 2 is standard in literature, and is satisfied by many design matrices, such as Gaussian random matrices. Notice that the rate of S, p and n are standard.

Example 1. Consider a random design case, where X_{ij} are IID following $\mathcal{N}(0, 1/n)$. If $S\log p \leq (1/1764)n$, then Assumption 2 holds for $S' = 48S$ with probability at least $1 - 1/p^2$. Furthermore, $\phi_{\min}(S+S') \geq 1/2$.

3 Simulation

We conduct a simple simulation study to compare the robust multiple-task regression with the usual multi-task regression that ignores gross errors. To this end, we generate X_{ij} from the normal distribution $\mathcal{N}(0, 1/n)$ and the entries of Θ are from iid $\mathcal{N}(0,1)$. The errors W_{ij} are generated from $\mathcal{N}(0, 1/n)$, here n is the number of samples, and we let $n = 50$. The gross errors are either uniform random numbers on $[0, 5]$ or fixed as 3. We either fix s or S. For the first case, we randomly select $s = 200$ entries of G^* and randomly select S rows of Θ^* with $S = 10,\ 15,\ 20,\ 25$ or 30 by setting the rest rows of Θ^* as zero. For the second case, we randomly select $s = 200,\ 400,\ 600,\ 800$ or 1000 entries in G^* as errors and choose $S = 20$ rows of Θ^*.

To choose λ and ρ, we generate randomly a new dataset (with same number of samples) $\tilde{Y} = \tilde{X}\Theta^* + \tilde{W}$ where $\tilde{X}$ and $\tilde{W}$ are generated in the same way as X and W respectively, that is devoted to tuning the parameters. We then choose these two tuning parameters that minimize the prediction error on $\tilde{Y}$ in terms of the squared Frobenius norm. This scheme is similar to cross validation with the extra benefit that unlike cross validation, the parameter is not tuned over the original data-set, hence avoids over-fitting the data.

The accuracy in terms of estimating Θ^* is measured by the ratio of the squared Frobenius norm of $\Theta^* - \hat{\Theta}$ for an estimate $\hat{\Theta}$. The robust multi-task regression is better if this ratio is smaller than one. From Figure 1, clearly, the proposed method outperforms the

standard multi-task regression by a large margin when either Θ is column-sparse or G is sparse, regardless of how the gross error is generated. This simulation result agrees with our theoretical analysis – that the proposed method is a promising approach to handle gross error in multi-task learning setup.

4 Conclusion

We have proposed a tractable approach for high-dimensional robust multiple-response regression. Our method exploits the joint sparsity of the features across the responses, and handles gross errors via regularization in a natural manner. Theoretical analysis confirms that this method possesses favorable properties even when the dimensionality is high.

A technical assumption we made is $\|X\Theta^*\|_\infty \leq \tau$, which can be restrictive. A future direction for research is to relax this assumption.

5 Proof of Theorem 1

This section is devoted to the proof of Theorem 1. Following a now-standard technique introduced in Negahban et al. (2009), it contains three parts: we first show that under appropriate selection of the regularization parameters, the estimator deviation satisfies "conic constraints". That is, the difference between the estimator and the ground truth will (approximately) belong to a cone. We then establish restricted eigenvalue condition and restricted strong convexity for all vectors satisfying such conic constraints. Based on these intermediate results, we establish Theorem 1.

5.1 Conic constraint

Note that we aim to estimate simultaneously G^* and Θ^*. Hence, we establish *two* conic constraints in this section. The first one bounds the direction of deviation of G, and the second one bounds the deviation of (Θ, G) jointly.

Lemma 1. *If $\rho \geq 4\|W\|_\infty + 8\tau$, then we have $\|\hat{\Delta}^G_{M^\perp}\|_1 \leq 3\|\hat{\Delta}^G_M\|_1 + 4\|G^*_{M^\perp}\|_1$.*

Proof. By optimality of $(\hat{\Theta}, \hat{G})$ we have

$$\begin{aligned}&\|Y - X\hat{\Theta} - \hat{G}\|_F^2 + \lambda\|\hat{\Theta}\|_{1,2} + \rho\|\hat{G}\|_1\\ \leq&\|Y - X\hat{\Theta} - G^*\|_F^2 + \lambda\|\hat{\Theta}\|_{1,2} + \rho\|G^*\|_1.\end{aligned}$$

Re-arranging the inequality, we have that

$$\begin{aligned}&\rho(\|\hat{G}\|_1 - \|G^*\|)\\ \leq&\langle \hat{G} - G^*,\, 2Y - 2X\hat{\Theta} - (\hat{G} + G^*)\rangle\\ =&2\langle \hat{G} - G^*,\, W\rangle + 2\langle \hat{G} - G^*,\, X(\Theta^* - \hat{\Theta})\rangle\\ &- \langle \hat{G} - G^*,\, \hat{G} - G^*\rangle \qquad (3)\\ \overset{(a)}{\leq}&2(\|W\|_\infty + 2\max_\Theta \|X\Theta\|_\infty)\|\hat{\Delta}^G\|_1\\ \leq&\frac{\rho}{2}\|\hat{\Delta}^G\|_1 \leq \frac{\rho}{2}(\|\hat{\Delta}^G_M\| + \|\hat{\Delta}^G_{M^\perp}\|),\end{aligned}$$

where (a) holds from the fact that $\|X\Theta\|_\infty \leq \tau$. Further notice that

$$\begin{aligned}&\|\hat{G}\|_1 - \|G^*\|_1\\ =&\|G^*_M + \hat{\Delta}^G_M\| + \|G^*_{M^\perp} + \hat{\Delta}^G_{M^\perp}\|_1 - \|G^*_M + G^*_{M^\perp}\|\\ \geq&\|G^*_M\|_1 - \|\hat{\Delta}^G_M\|_1 + \|\hat{\Delta}^G_{M^\perp}\|_1 - \|G^*_{M^\perp}\|_1\\ &- \|G^*_M\|_1 - \|G^*_{M^\perp}\|_1\\ =&\|\hat{\Delta}^G_{M^\perp}\|_1 - \|\hat{\Delta}^G_M\|_1 - 2\|G^*_{M^\perp}\|_1.\end{aligned} \qquad (4)$$

Combining Equation (3) and (4) yields $\|\hat{\Delta}^G_{M^\perp}\|_1 \leq 3\|\hat{\Delta}^G_M\|_1 + 4\|G^*_{M^\perp}\|_1$, as claimed. □

Denote $\Phi(\Theta, G) \triangleq \|\Theta\|_{1,2} + \frac{\rho}{\lambda}\|G\|_1$, which can be regarded as one joint regularization term of $(\hat{\Delta}^\Theta, \hat{\Delta}^G)$. We then establish a similar conic constraint.

Lemma 2. *The following holds*

$$\begin{aligned}\Phi(\Theta^*, G^*) - \Phi(\hat{\Theta}, \hat{G}) \leq \Phi(\hat{\Delta}^\Theta_A, \hat{\Delta}^G_M)\\ +\|\Theta^*_{A^\perp}\|_{1,2} + \frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1.\end{aligned} \qquad (5)$$

Furthermore, if $\lambda \geq 4\|X^\top W\|_{\infty,2}$ and $\rho \geq 4\|W\|_\infty + 8\tau$, then we have

$$\begin{aligned}\Phi(\hat{\Delta}^\Theta_{A^\perp}, \hat{\Delta}^G_{M^\perp}) \leq 3\Phi(\hat{\Delta}^\Theta_A, \hat{\Delta}^G_M)\\ +4\left\{\|\Theta^*_{A^\perp}\|_{1,2} + \frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1\right\}.\end{aligned} \qquad (6)$$

Proof. Equation (5) holds from the following algebra:

$$\begin{aligned}&\Phi(\Theta^*, G^*) - \Phi(\hat{\Theta}, \hat{G})\\ =&\Phi(\Theta^*_A, G^*_M) + \Phi(\Theta^*_{A^\perp}, G^*_{M^\perp}) - \Phi(\hat{\Theta}_A, \hat{G}_M)\\ &- \Phi(\hat{\Theta}_{A^\perp}, \hat{G}_{M^\perp})\\ \leq&\Phi(\Theta^*_A - \hat{\Theta}_A, G^*_M - \hat{G}_M) + \Phi(\Theta^*_{A^\perp}, G^*_{M^\perp})\\ =&\Phi(\hat{\Delta}^\Theta_A, \hat{\Delta}^G_M) + \|\Theta^*_{A^\perp}\|_{1,2} + \frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1.\end{aligned}$$

To show Equation (6), by optimality of $(\hat{\Theta}, \hat{G})$, we have

$$\begin{aligned}&\|Y - X\hat{\Theta} - \hat{G}\|_F^2 + \lambda\|\hat{\Theta}\|_{1,2} + \rho\|\hat{G}\|_1\\ \leq&\|Y - X\Theta^* - G^*\|_F^2 + \lambda\|\Theta^*\|_{1,2} + \rho\|G^*\|_1.\end{aligned}$$

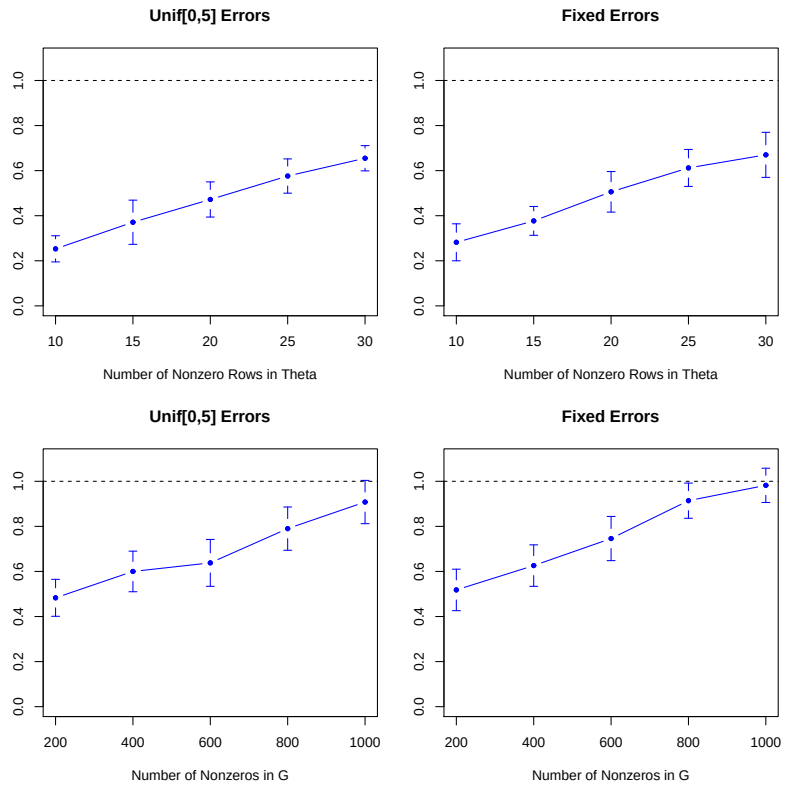

Figure 1: The ratio of error (measured by squared Frobenius norm) of the robust multi-task learning to that of the standard multi-task learning. The mean and the 95% confidence intervals are plotted.

Re-arranging the equation, we get

$$\begin{aligned}
&\lambda(\Phi(\hat{\Theta},\hat{G})-\Phi(\Theta^*,G^*))\\
\leq&\langle X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G},\, 2Y-X(\hat{\Theta}+\Theta^*)-(\hat{G}+G^*)\rangle\\
=&\langle X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G},\, 2W-[X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}]\rangle\\
\leq&\langle X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G},\, 2W\rangle\\
\leq&2\|X^\top W\|_{\infty,2}\|\hat{\Delta}^{\Theta}\|_{1,2}+2\|W\|_{\infty}\|\hat{\Delta}^{G}\|_1\\
\leq&\frac{\lambda}{2}\|\hat{\Delta}^{\Theta}\|_{1,2}+\frac{\rho}{2}\|\hat{\Delta}^{G}\|_1=\frac{\lambda}{2}\Phi(\hat{\Delta}^{\Theta},\hat{\Delta}^{G}).
\end{aligned}\tag{7}$$

On the other hand, notice the following holds

$$\begin{aligned}
&\Phi(\hat{\Theta},\hat{G})-\Phi(\Theta^*,G^*)\\
=&\Phi(\hat{\Theta}_A,\hat{G}_M)+\Phi(\hat{\Theta}_{A^\perp},\hat{G}_{M^\perp})-\Phi(\Theta^*_A,G^*_M)\\
&-\Phi(\Theta^*_{A^\perp},G^*_{M^\perp})\\
\geq&\Phi(\Theta^*_A,G^*_M)-\Phi(\hat{\Delta}^{\Theta}_A,\hat{\Delta}^{G}_M)+\Phi(\hat{\Delta}^{\Theta}_{A^\perp},\hat{\Delta}^{G}_{M^\perp})\\
&-\Phi(\Theta^*_{A^\perp},G^*_{M^\perp})-\Phi(\Theta^*_A,G^*_M)-\Phi(\Theta^*_{A^\perp},G^*_{M^\perp})\\
=&\Phi(\hat{\Delta}^{\Theta}_{A^\perp},\hat{\Delta}^{G}_{M^\perp})-\Phi(\hat{\Delta}^{\Theta}_A,\hat{\Delta}^{G}_M)-2\Phi(\Theta^*_{A^\perp},G^*_{M^\perp}).
\end{aligned}$$

Combining this with Equation (7), we get

$$\begin{aligned}
&\Phi(\hat{\Delta}^{\Theta}_{A^\perp},\hat{\Delta}^{G}_{M^\perp})-\Phi(\hat{\Delta}^{\Theta}_A,\hat{\Delta}^{G}_M)-2\Phi(\Theta^*_{A^\perp},G^*_{M^\perp})\\
\leq&\frac{1}{2}\Phi(\hat{\Delta}^{\Theta},\hat{\Delta}^{G})\leq\frac{1}{2}[\Phi(\hat{\Delta}^{\Theta}_A,\hat{\Delta}^{G}_M)+\Phi(\hat{\Delta}^{\Theta}_{A^\perp},\hat{\Delta}^{G}_{M^\perp})],
\end{aligned}$$

which by re-arranging implies Equation (6). □

5.2 Restricted Strong Convexity

We show in this section that when the conic constraints are satisfied, we can upper bound $\|\hat{\Delta}^{\Theta}\|_F$ by a function of $\|X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}\|_F$. Such kind of results are termed as *restricted strong convexity* in literature (Negahban et al., 2009). To establish this, we first show the following lemma that relates $\|\Theta\|_F$ to $\|X\Theta\|_F$ for any Θ that satisfies the conic constraints. This result extend the restricted eigenvalue condition in the single-task regression setup (Bickel et al., 2009) to the multi-task regression problem. The restricted strong convexity holds by specializing the restricted eigenvalue condition to $(\hat{\Delta}^{\Theta},\hat{\Delta}^{G})$.

Lemma 3. *Let $J_0\subset[1:p]$ with $|J_0|=S$ be a row-index set, and $\Theta\in\mathbb{R}^{p\times q}$ satisfies $\|\Theta_{J_0^\perp}\|_{1,2}\leq\alpha_0\|\Theta_{J_0}\|_{1,2}+\alpha_1$ for some α_0 and α_1. Then we have for any $S'\in[0:p]$,*

$$\begin{aligned}
\|X\Theta\|_F\geq&\frac{\sqrt{\phi_{\min}(S+S')}-\alpha_0\sqrt{\phi_{\max}(S')S/S'}}{1+\alpha_0\sqrt{S/S'}}\|\Theta\|_F\\
&-\left[\frac{\sqrt{\phi_{\min}(S+S')}-\alpha_0\sqrt{\phi_{\max}(S')S/S'}}{\sqrt{S'}+\alpha_0\sqrt{S}}\right.\\
&\left.+\sqrt{\phi_{\max}(S')/S'}\right]\alpha_1.
\end{aligned}\tag{8}$$

Proof. Fix $S' \in [1:p]$. Partition $J_0^\perp$ into K subsets, each of size S' and the last subset of size $\leq S'$, such that J_k is the set of indices corresponding to S' largest in ℓ_2 norm of rows of Θ outside $\bigcup_{j=0}^{k-1} J_j$. Let $J_{01} \triangleq J_0 \bigcup J_1$. We have

$$\begin{aligned}\|X\Theta\|_F &\geq \|X\Theta_{J_{01}}\|_F - \sum_{k=2}^{K} \|X\Theta_{J_k}\|_F \\ &\geq \|X\Theta_{J_{01}}\|_F - \sum_{k=2}^{K} \sqrt{\phi_{\max}(S')}\|\Theta_{J_k}\|_F.\end{aligned} \tag{9}$$

Here the last inequality holds from the following, where we denote the i-th column of Θ_{J_k} by $\Theta^i_{J_k}$

$$\begin{aligned}\|X\Theta_{J_k}\|_F &= \sqrt{\sum_{i=1}^{q} \|X\Theta^i_{J_k}\|_2^2} \\ &\leq \sqrt{\sum_{i=1}^{q} \phi_{\max}(S')\|\Theta^i_{J_k}\|_2^2} = \sqrt{\phi_{\max}(S')}\|\Theta_{J_k}\|_F.\end{aligned}$$

Now let $\overline{\theta}_{J_{k-1}}$ be the last row (i.e., the row with the smallest ℓ_2 norm) of $\Theta_{J_{k-1}}$, and we have

$$\begin{aligned}\|\Theta_{J_k}\|_F &\leq \sqrt{S'\|\overline{\theta}_{J_{k-1}}\|_2^2} \\ &= \sqrt{S'}\|\overline{\theta}_{J_{k-1}}\|_2 \leq \|\Theta_{J_{k-1}}\|_{1,2}/\sqrt{S'},\end{aligned}$$

which implies

$$\sum_{k=2}^{K} \|\Theta_{J_k}\|_F \leq \sum_{k=2}^{K} \|\Theta_{J_{k-1}}\|_{1,2}/\sqrt{S'} \leq \|\Theta_{J_0^\perp}\|_{1,2}/\sqrt{S'}. \tag{10}$$

Furthermore, we have the following

$$\begin{aligned}\|X\Theta_{J_{01}}\|_F &= \sqrt{\sum_{i=1}^{q} \|X\Theta^i_{J_01}\|_2^2} \\ &\geq \sqrt{\sum_{i=1}^{q} \phi_{\min}(S+S')\|\Theta^i_{J_01}\|_2^2} = \sqrt{\phi_{\min}(S+S')}\|\Theta_{J_{01}}\|_F.\end{aligned}$$

Substituting this and Equation (10) into Equation (9), we have

$$\begin{aligned}&\|X\Theta\|_F \\ \geq&\sqrt{\phi_{\min}(S+S')}\|\Theta_{J_{01}}\|_F - \sqrt{\phi_{\max}(S')}\|\Theta_{J_0^\perp}\|_{1,2}/\sqrt{S'} \\ \geq&\sqrt{\phi_{\min}(S+S')}\|\Theta_{J_{01}}\|_F \\ &- \sqrt{\phi_{\max}(S')/S'}(\alpha_0\|\Theta_{J_0}\|_{1,2} + \alpha_1) \\ \geq&\sqrt{\phi_{\min}(S+S')}\|\Theta_{J_{01}}\|_F \\ &- \sqrt{\phi_{\max}(S')/S'}(\sqrt{S}\alpha_0\|\Theta_{J_0}\|_F + \alpha_1) \\ \geq&\sqrt{\phi_{\min}(S+S')}\|\Theta_{J_{01}}\|_F \\ &- \sqrt{\phi_{\max}(S')/S'}(\sqrt{S}\alpha_0\|\Theta_{J_{01}}\|_F + \alpha_1) \\ =&\left[\sqrt{\phi_{\min}(S+S')} - \alpha_0\sqrt{\phi_{\max}(S')S/S'}\right]\|\Theta_{J_{01}}\|_F \\ &- \sqrt{\phi_{\max}(S')/S'}\alpha_1,\end{aligned} \tag{11}$$

where in the second inequality we used the assumption that Θ satisfies a conic constraint. Next notice that

$$\begin{aligned}\|\Theta_{J_{01}}\|_F \geq&\|\Theta\|_F - \sum_{k=2}^{K} \|\Theta_{J_k}\|_F \\ \geq&\|\Theta\|_F - \|\Theta_{J_0^c}\|_{1,2}/\sqrt{S'} \\ \geq&\|\Theta\|_F - (\alpha_0\|\Theta_{J_0}\|_{1,2} + \alpha_1)/\sqrt{S'} \\ \geq&\|\Theta\|_F - \alpha_0\sqrt{S/S'}\|\Theta_{J_0}\|_F - \alpha_1/\sqrt{S'} \\ \geq&\|\Theta\|_F - \alpha_0\sqrt{S/S'}\|\Theta_{J_{01}}\|_F - \alpha_1/\sqrt{S'},\end{aligned}$$

where in the second inequality we again use Equation (10). Re-arranging the terms we get,

$$\|\Theta_{J_{01}}\|_F \geq \frac{\|\Theta\|_F}{1+\alpha_0\sqrt{S/S'}} - \frac{\alpha_1}{\sqrt{S'}+\alpha_0\sqrt{S}}.$$

Substitute this into Equation (11), we conclude that Equation (8) holds, which proves the lemma. □

Recall definitions of $\kappa_1(\cdot)$ and $\kappa_2(\cdot)$. Thus, Equation (8) can be simplified as $\|X\Theta\|_F \geq \kappa_1(S')\|\Theta\|_F - \kappa_2(S')a_1$. The next lemma established an upper bound of $\|\hat{\Delta}^\Theta\|_F$.

Lemma 4 (Restricted Strong Convexity). *Fix S'. If $\Phi(\hat{\Delta}^\Theta_{A^\perp}, \hat{\Delta}^G_{M^\perp}) \leq c_0\Phi(\hat{\Delta}^\Theta_A, \hat{\Delta}^G_M) + c_1$, for some c_0, c_1 and*

$$\begin{aligned}&\|\hat{\Delta}^\Theta\|_F \geq \\ &\max\left\{\frac{2[c_0\rho\kappa_2(S')\sqrt{s}/\lambda - 1]}{\kappa_1(S')}\|\hat{\Delta}^G\|_F,\ \frac{4\kappa_2(S')c_1}{\kappa_1(S')}\right\},\end{aligned} \tag{12}$$

then we have

$$\|\hat{\Delta}^\Theta\|_F^2 \leq \frac{32}{\kappa_1(S')^2}\|X\hat{\Delta}^\Theta + \hat{\Delta}^G\|_F^2 + 2\|X\hat{\Delta}^\Theta\|_\infty\|\hat{\Delta}^G\|_1.$$

Proof. We first note the following,

$$\|X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}\|_F^2 \geq \|X\hat{\Delta}^{\Theta}\|_F^2+\|\hat{\Delta}^{G}\|_F^2+2\langle X\hat{\Delta}^{\Theta},\hat{\Delta}^{G}\rangle. \tag{13}$$

The assumption $\Phi(\hat{\Delta}^{\Theta}_{A^\perp},\hat{\Delta}^{G}_{M^\perp}) \leq c_0\Phi(\hat{\Delta}^{\Theta}_{A},\hat{\Delta}^{G}_{M})+c_1$ leads to

$$\|\hat{\Delta}^{\Theta}_{A^\perp}\|_{1,2} \leq c_0\|\hat{\Delta}^{\Theta}_{A}\|_{1,2}+\left\{\frac{\rho}{\lambda}\left[c_0\|\hat{\Delta}^{G}_{M}\|_1-\|\hat{\Delta}^{G}_{M^\perp}\|_1\right]+c_1\right\}$$

By Lemma 3, we thus have

$$\begin{aligned}&\|X\hat{\Delta}^{\Theta}\|_F \geq \kappa_1(S')\|\hat{\Delta}^{\Theta}\|_F\\ &\quad-\kappa_2(S')\left\{\frac{\rho}{\lambda}\left[c_0\|\hat{\Delta}^{G}_{M}\|_1-\|\hat{\Delta}^{G}_{M^\perp}\|_1\right]+c_1\right\}.\end{aligned}$$

Simplifying the right-hand-side using $c_0\|\hat{\Delta}^{G}_{M}\|_1-\|\hat{\Delta}^{G}_{M^\perp}\|_1 \leq c_0\sqrt{s}\|\hat{\Delta}^{G}\|_F$, we get

$$\begin{aligned}&\|X\hat{\Delta}^{\Theta}\|_F+\|\hat{\Delta}^{G}\|_F\\ \geq&\kappa_1(S')\|\hat{\Delta}^{\Theta}\|_F-(\frac{c_0\rho\kappa_2(S')\sqrt{s}}{\lambda}-1)\|\hat{\Delta}^{G}\|_F-\kappa_2(S')c_1\\ \geq&\frac{\kappa_1(S')}{4}\|\hat{\Delta}^{\Theta}\|_F,\end{aligned}$$

where the last inequality follows from Equation (12). This leads to

$$\begin{aligned}\|\hat{\Delta}^{\Theta}\|_F^2 &\leq \frac{16}{\kappa_1(S')^2}[\|X\hat{\Delta}^{\Theta}\|_F+\|\hat{\Delta}^{G}\|]^2\\ &\leq \frac{32}{\kappa_1(S')^2}\|X\hat{\Delta}^{\Theta}\|_F^2+\frac{32}{\kappa_1(S')^2}\|\hat{\Delta}^{G}\|_F^2.\end{aligned}$$

Substituting this into Equation (13) establishes the lemma. □

5.3 Proof of the Main Theorem

Proof of Theorem 1. We first bound $\|\hat{\Delta}^{G}\|_F$. From the optimality of $(\hat{\Theta},\hat{G})$, we have that

$$\begin{aligned}&\|Y-X\hat{\Theta}-\hat{G}\|_F^2+\lambda\|\hat{\Theta}\|_{1,2}+\rho\|\hat{G}\|_1\\ \leq&\|Y-X\hat{\Theta}-G^*\|_F^2+\lambda\|\hat{\Theta}\|_{1,2}+\rho\|G^*\|_1.\end{aligned}$$

Re-arranging the terms, we get

$$\begin{aligned}&\langle G^*-\hat{G},\,2Y-2X\hat{\Theta}-(\hat{G}+G^*)\rangle\\ &\quad+\rho(\|\hat{G}\|_1-\|G^*\|_1) \leq 0.\end{aligned}$$

Recall that $W=Y-X\Theta^*-G^*$, we have that

$$\begin{aligned}&2\langle-\hat{\Delta}^{G},W\rangle+2\langle-\hat{\Delta}^{G},-X\hat{\Delta}^{\Theta}\rangle\\ &\quad+\|\hat{\Delta}^{G}\|_F^2+\rho(\|\hat{G}\|_1-\|G^*\|_1) \leq 0,\end{aligned}$$

which implies

$$\begin{aligned}&\|\hat{\Delta}^{G}\|_F^2\\ \leq&2\|W\|_\infty\|\hat{\Delta}^{G}\|_1+2\|X\hat{\Delta}^{\Theta}\|_\infty\|\hat{\Delta}^{G}\|_1+\rho\|\hat{\Delta}^{G}\|_1\\ \leq&\frac{3}{2}\rho\|\hat{\Delta}^{G}\|_1\\ \leq&\frac{3}{2}\rho(4\|\hat{\Delta}^{G}_{M}\|_1+4\|G^*_{M^\perp}\|_1)\\ \leq&6\rho(\sqrt{s}\|\hat{\Delta}^{G}\|_F+\|G^*_{M^\perp}\|_1).\end{aligned}$$

Solving quadratic yields

$$\|\hat{\Delta}^{G}\|_F \leq 6\rho\sqrt{s}+\sqrt{6\rho\|G^*_{M^\perp}\|_1}. \tag{14}$$

Following similar steps, we bound the combined error in estimating $(X\Theta^*+G^*)$. By optimality of $(\hat{\Theta},\hat{G})$ we have

$$\begin{aligned}&\|Y-X\hat{\Theta}-\hat{G}\|_F^2+\lambda\|\hat{\Theta}\|_{1,2}+\rho\|\hat{G}\|_1\\ \leq&\|Y-X\Theta^*-G^*\|_F^2+\lambda\|\Theta^*\|_{1,2}+\rho\|G^*\|_1.\end{aligned}$$

Re-arranging leads to

$$\begin{aligned}&\|X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}\|_F^2\\ \leq&2\langle W,\,X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}\rangle+\lambda(\Phi(\Theta^*,G^*)-\Phi(\hat{\Theta},\hat{G}))\\ \leq&2\|X^\top W\|_{\infty,2}\|\hat{\Delta}^{\Theta}\|_{1,2}+2\|W\|_\infty\|\hat{\Delta}^{G}\|_1\\ &+\lambda(\Phi(\Theta^*,G^*)-\Phi(\hat{\Theta},\hat{G}))\\ \leq&\frac{\lambda}{2}\|\hat{\Delta}^{\Theta}\|_{1,2}+\frac{\rho}{2}\|\hat{\Delta}^{G}\|_1+\lambda(\Phi(\hat{\Delta}^{\Theta}_{A},\hat{\Delta}^{G}_{M})\\ &+\|\Theta^*_{A^\perp}\|_{1,2}+\frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1),\end{aligned}$$

where we used Equation (5) for the last inequality. Note that $\frac{\lambda}{2}\|\hat{\Delta}^{\Theta}\|_{1,2}+\frac{\rho}{2}\|\hat{\Delta}^{G}\|_1=\frac{\lambda}{2}\Phi(\hat{\Delta}^{\Theta},\hat{\Delta}^{G})$, hence by Equation (6) we have

$$\begin{aligned}&\|X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}\|_F^2\\ \leq&3\lambda(\Phi(\hat{\Delta}^{\Theta}_{A},\hat{\Delta}^{G}_{M})+\|\Theta^*_{A^\perp}\|_{1,2}+\frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1)\\ =&3\lambda\|\hat{\Delta}^{\Theta}_{A}\|_{1,2}+3\rho\|\hat{\Delta}^{G}_{M}\|_1+3\lambda\|\Theta^*_{A^\perp}\|_{1,2}\\ &+3\rho\|G^*_{M^\perp}\|_1\\ \leq&3\lambda\sqrt{S}\|\hat{\Delta}^{\Theta}\|_F+3\rho\sqrt{s}\|\hat{\Delta}^{G}\|_F+3\lambda\|\Theta^*_{A^\perp}\|_{1,2}\\ &+3\rho\|G^*_{M^\perp}\|_1.\end{aligned} \tag{15}$$

Now applying Lemma 4, we have that conditioned on

$$\begin{aligned}&\|\hat{\Delta}^{\Theta}\|_F \geq\\ &\max\left\{\frac{2[c_0\rho\kappa_2(S')\sqrt{s}/\lambda-1]}{\kappa_1(S')}\|\hat{\Delta}^{G}\|_F,\,\frac{4\kappa_2(S')c_1}{\kappa_1(S')}\right\},\end{aligned}$$

with $c_0=3$ and $c_1=4\left\{\|\Theta^*_{A^\perp}\|_{1,2}+\frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1\right\}$ (denote this event by $\mathcal{E}$), the following holds

$$\begin{aligned}&\|\hat{\Delta}^{\Theta}\|_F^2\\ \leq&\frac{32}{\kappa_1(S')^2}\|X\hat{\Delta}^{\Theta}+\hat{\Delta}^{G}\|_F^2+2\|X\hat{\Delta}^{\Theta}\|_\infty\|\hat{\Delta}^{G}\|_1\\ \leq&\frac{96}{\kappa_1(S')^2}\lambda\sqrt{S}\|\hat{\Delta}^{\Theta}\|_F+[\frac{96}{\kappa_1(S')^2}+1]\rho\sqrt{s}\|\hat{\Delta}^{G}\|_F\\ &+\frac{96}{\kappa_1(S')^2}\lambda\|\Theta^*_{A^\perp}\|_{1,2}+[\frac{96}{\kappa_1(S')^2}+1]\rho\|G^*_{M^\perp}\|_1,\end{aligned}$$

where in the last inequality we used Equation (15) and the fact that

$$\|X\hat{\Delta}^{\Theta}\|_\infty\|\hat{\Delta}^{G}\|_1 \leq \frac{\rho}{4}\|\hat{\Delta}^{G}\|_1 \leq \frac{\rho}{4}(4\sqrt{s}\|\hat{\Delta}^{G}\|_F+4\|G^*_{M^\perp}\|_1).$$

Using the bound in Equation (14), Equation (16) implies that for a universal c',

$$\|\hat{\Delta}^{\Theta}\|_F^2 \leq \frac{c'}{\min(\kappa_1(S')^2,1)} \times \left[\lambda\sqrt{S}\|\hat{\Delta}^{\Theta}\|_F + \lambda\|\Theta^*_{A^\perp}\|_{1,2} + \rho^2 s + \rho\|G^*_{M^\perp}\|_1\right],$$

which leads to under $\mathcal{E}$, for a universal constant c,

$$\|\hat{\Delta}^{\Theta}\|_F \leq \frac{c}{\min(\kappa_1(S'),1)} \times \left[\lambda\sqrt{S} + \sqrt{\lambda\|\Theta^*_{A^\perp}\|_{1,2}} + \rho\sqrt{s} + \sqrt{\rho\|G^*_{M^\perp}\|_1}\right].$$

Notice that $\mathcal{E}^c$ is the event that

$$\|\hat{\Delta}^{\Theta}\|_F \leq \max\left\{\frac{[6\rho\kappa_2(S')\sqrt{s}/\lambda - 3][6\rho\sqrt{s} + \sqrt{6\rho\|G^*_{M^\perp}\|_1}]}{\kappa_1(S')}, \frac{16\kappa_2(S')\left\{\|\Theta^*_{A^\perp}\|_{1,2} + \frac{\rho}{\lambda}\|G^*_{M^\perp}\|_1\right\}}{\kappa_1(S')}\right\}.$$

We thus establish the bound of $\|\hat{\Delta}^{\Theta}\|_F$ as claimed in the theorem. □

6 Proof of Theorem 2

In this section we prove Theorem 2, which involved two components: simplifying $\kappa_1(S')$ and $\kappa_2(S')$, and bounding ρ and λ. Due to space constraints, we defer the proofs of intermediate results to the supplementary material.

Lemma 5. *Under Condition 2, we have* $\kappa_1(S') \geq \frac{1}{16}\sqrt{\phi_{\min}(S+S')}$, *and* $\frac{\kappa_2(S')}{\kappa_1(S')} \leq \frac{5}{\sqrt{S}}$.

Lemma 6. *Under Condition 1, for any* $n, p \geq 2$, *we have with probability* $1 - 1/(2n^3)$

$$\|X^\top W\|_{\infty,2} \leq \min\left(\frac{\sigma(\sqrt{q} + \sqrt{8\log pn})}{\sqrt{n}}, 7\sigma(1+\sqrt{q/n})\right).$$

Proof of Theorem 2. We first show that ρ and λ, as set in the theorem, satisfies the condition of Theorem 1. Recall that for a standard Normal random variable $\bar{x} \sim \mathcal{N}(0,1)$, the following holds for all $x_0 > 0$

$$Pr(\bar{x} \geq x_0) \leq \frac{1}{x_0}\frac{1}{\sqrt{2\pi}}\exp(-x_0^2/2).$$

Thus, since $W_{ij} \sim \mathcal{N}(0, \sigma^2/n)$, we have

$$\begin{aligned} & Pr\big(\|W\|_\infty \geq 4\sigma\sqrt{\log(nq)/n}\big) \\ \leq\ & 2nqPr(\bar{x} \geq \sqrt{16\log(nq)}) \\ \leq\ & \frac{2nq}{\sqrt{32\pi\log(nq)}}\exp(-8\log(nq)) \leq 1/(2n^3). \end{aligned}$$

This, combined with Lemma 6, shows that with probability $1/n^3$ we have

$$\rho \geq 4\|W\|_\infty + 8\tau; \quad \lambda \geq 4\|X^\top W\|_{\infty,2}.$$

Apply Theorem 1 we have $\|\hat{\Delta}^G\|_F \leq 6\rho\sqrt{s} \leq c_1\sigma\sqrt{\frac{s\log(nq)}{n}} + c_1\alpha$, which establishes the first claim. We now turn to the second claim. Apply Theorem 1 we have

$$\|\hat{\Delta}^{\Theta}\|_F \leq \max\left(c[\lambda\sqrt{S} + \rho\sqrt{s}]/\min(\kappa_1,1),\ 36\rho^2 s\kappa_2/(\lambda\kappa_1)\right).$$

Since κ_1 is lower-bounded by $1/(16\sqrt{\phi_{\min}(S+S')})$, we have that

$$\begin{aligned} & c[\lambda\sqrt{S} + \rho\sqrt{s}]/\min(\kappa_1,1) \\ \leq & [c'\rho\sqrt{s} + c'\lambda\sqrt{S}]/\sqrt{\phi_{\min}(S+S')} \\ \leq & 2c'\rho\sqrt{s} + 4c'\lambda_0\sqrt{S}/\sqrt{\phi_{\min}(S+S')}, \end{aligned}$$

Here the last inequality is due to the value of λ. Recall $\kappa_2/\kappa_1 \leq 5/\sqrt{S}$ from Lemma 5, and hence

$$36\rho^2 s\kappa_2/(\lambda\kappa_1) \leq 180[\rho\sqrt{s}]\frac{\rho\sqrt{s}}{\lambda\sqrt{S}} \leq c''\rho\sqrt{s}.$$

Here the last inequality holds from definition of λ. Thus we conclude that for some universal constant c

$$\|\hat{\Delta}^{\Theta}\|_F \leq c[\rho\sqrt{s} + \lambda_0\sqrt{S}]/\sqrt{\phi_{\min}(S+S')}.$$

The Theorem follows by substituting ρ and λ_0 into the equation. □

Acknowledgements

The research of H. Xu is partially supported from the National University of Singapore under startup grant R-265-000-384-133. The research of C. Leng is partially supported from the Ministry of Education of Singapore, under Tier 1 grant R-155-000-107-112.

References

Agarwal, A., Negahban, S., & Wainwright, M. (2011). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. http://arxiv.org/abs/1102.4807.

Argyriou, A., Evgeniou, T., & Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, *73*, 243–272.

Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*, 1179–1225.

Bickel, P., Ritov, Y., & Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, *37*, 1705–1732.

Candès, E., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? To appear in *Journal of ACM*.

Candès, E., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*, 717–772.

Candès, E., Rudelson, M., Tao, T., & Vershynin, R. (2005). Error correction via linear programming. *FOCS* (pp. 295–308).

Candès, E. J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, *52*, 489–509.

Candès, E. J., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*, 2313–2351.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*.

Chandrasekaran, V., Sanghavi, S., Parrilo, P., & Willsky, A. (2011). Rank-sparsity incoherence for matrix decomposition. To appear in *SIAM Journal of Optimization* .

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*, 1289–1306.

Jalali, A., Ravikumar, P., Sanghavi, S., & Ruan, C. (2010). A dirty model for multi-task learning. *Advances in Neural Information Processing Systems*.

Keshavan, R. H., Montanari, A., & Oh, S. (2010). Matrix completion from noisy entries. *Journal of Machine Learning Research*, *11*, 2057–2078.

Lee, Y., MacEachern, S., & Jung, Y. (2011). Regularization of case-specific parameters for robustness and efficiency. Technical report, Ohio State University.

Negahban, S., Ravikumar, P., Wainwright, M., & Yu, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Advances in Neural Information Processing Systems*.

Negahban, S., & Wainwright, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. *Advances in Neural Information Processing Systems*.

Ravikumar, P., Wainwright, M., & Lafferty, J. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*.

Recht, B., Fazel, M., & Parrilo, P. (2010). Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. To appear in *SIAM Review*.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, *51*, 1030–1051.

Tropp, J. A., Gilbert, A. C., & Strauss, M. J. (2006). Algorithms for simultaneous sparse approximation. *Signal Processing*, *86*, 572–602.

Wainwright, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, *55*, 2183–2202.

Wright, J., & Ma, Y. (2010). Dense error correction via ℓ_1-minimization. *IEEE Transactions on Information Theory*, *56*, 3540–3560.

Xu, H., Caramanis, C., & Sanghavi, S. (2010). Robust PCA via outlier pursuit. *Advances in Neural Information Processing Systems*.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, *68*, 49–67.