Electronic Journal of Statistics
Vol. 0 (0000) 1–0
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Supplement to "Double Fused Lasso Regularized Regression with Both Matrix and Vector Valued Predictors"

Mei Li* **and Lingchen Kong***

Department of Applied Mathematics, Beijing Jiaotong University
e-mail: 18118016@bjtu.edu.cn; lchkong@bjtu.edu.cn

Zhihua Su‡

Department of Statistics, University of Florida
e-mail: zhihuasu@ufl.edu

Appendix A: Appendix section

A.1. Moreau-envelope function and proximal mapping

We present Moreau envelope function and proximal mapping. In particular, we list the explicit form of proximal mapping for specific functions, such as the indictor function, L_1-norm regularization function, fused Lasso regularization function, nuclear norm regularization function and matrix indictor function. Let $p:\mathbb{R}^n\to(-\infty,+\infty]$ be a closed proper convex function such that for a given $\nu>0$, the Moreau envelope function $\psi_p(\cdot)$ of p [11] is defined by

$$\psi_{p/\nu}(x)=\min_{z\in\mathbb{R}^n}\left\{p(z)+\frac{\nu}{2}\|z-x\|^2\right\},\ \ \forall\, x\in\mathbb{R}^n, \tag{1}$$

and the corresponding solution is called as the proximal mapping:

$$Prox_{p/\nu}(x)=\arg\min_{z\in\mathbb{R}^n}\left\{p(z)+\frac{\nu}{2}\|z-x\|^2\right\},\ \ \forall\, x\in\mathbb{R}^n.$$

Let $p:\mathbb{R}^n\to(-\infty,+\infty]$ be a closed proper convex function, then the Fenchel conjugate function of p is defined as $p^*(x):=\sup_{x'\in\mathbb{R}^n}\{\langle x,x'\rangle-p(x')\},\forall x\in\mathbb{R}^n$.

Proposition 1. *[12] Let $p:\mathbb{R}^n\to(-\infty,+\infty]$ be a closed proper convex function, and $p^*(x)$ be its Fenchel conjugate function. Then for any $t>0$,*

$$Prox_{tp}(x)+tProx_{p^*/t}(x/t)=x,\ \ \forall x\in\mathbb{R}^n. \tag{2}$$

The equality (2) is often referred to as the Moreau identity.

*Some comment
†First supporter of the project
‡Second supporter of the project

Now we discuss the proximal mapping of problem (1) with $\nu = 1$. We can obtain the explicit form of proximal mapping for some special functions. For example, if $p(z) = \delta(z;\Omega)$, where Ω is a nonempty closed convex set. The proximal mapping of the indicator function δ_Ω is the projection operator on the set Ω:

$$Prox_{\delta_\Omega}(x) = \arg\min_{z\in\mathbb{R}^n}\left\{\delta(z;\Omega) + \frac{\nu}{2}\|z-x\|^2\right\} = \arg\min_{z\in\Omega}\left\{\|z-x\|^2\right\} = \Pi(x;\Omega).$$

If $\Omega = B_{\|\cdot\|_\infty}(0;r)$, the proximal mapping of $\delta(x;\Omega)$ is

$$Prox_{\delta_\Omega}(x) = \Pi(x;\Omega) = x - sign(x)\cdot\max\{|x|-r,0\}. \tag{3}$$

If $p(z) = \lambda\|z\|_1$, the proximal mapping of p is $Prox_p(x) = shrink(x,\lambda) := sign(x)\cdot\max\{|x|-\lambda,0\}$, which is called as the soft-thresholding operator in [7].

If $p(z) = \lambda_1\|z\|_1 + \lambda_2\|Az\|_1$, where $\lambda_1,\lambda_2 \geq 0$ are given parameters and $A \in \mathbb{R}^{(p-1)\times p}$,

$$A = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}, \tag{4}$$

then the proximal mapping of p is

$$Prox_p(x) = \arg\min_{z\in\mathbb{R}^n}\left\{\lambda_1\|z\|_1 + \lambda_2\|Az\|_1 + \frac{1}{2}\|z-x\|^2\right\},\ \ \forall\, x\in\mathbb{R}^n. \tag{5}$$

If $\lambda_1 = 0$ in (5), we denote the proximal mapping of $p(z) = \lambda_2\|Az\|_1$ by $z_{\lambda_2}(x)$, and $z_{\lambda_2}(x)$ is

$$z_{\lambda_2}(x) = \arg\min_{z\in\mathbb{R}^n}\left\{\lambda_2\|Az\|_1 + \frac{1}{2}\|z-x\|^2\right\},\ \ \forall x\in\mathbb{R}^n.$$

Proposition 2. *[7] Let $p(z) = \lambda_1\|z\|_1 + \lambda_2\|Az\|_1$, $A \in \mathbb{R}^{(p-1)\times p}$ has the form as in (4), then we have*

$$Prox_p(x) = Prox_{\lambda_1\|\cdot\|_1}(z_{\lambda_2}(x)) = sign(z_{\lambda_2}(x))\cdot max\{|z_{\lambda_2}(x)|-\lambda_1,0\},\ \ \forall\, x\in\mathbb{R}^n. \tag{6}$$

Now we present the proximal mapping for the matrix-form function. Let $p(M) = \|M\|_*$, then

$$Prox_p(D) = \arg\min_{M\in\mathbb{R}^{m\times q}}\left\{\|M\|_* + \frac{\nu}{2}\|M-D\|_F^2\right\},\ \ \forall\, D\in\mathbb{R}^{m\times q},$$

and it has a closed-form solution, which is given by

$$Prox_p(D) = U_D Diag(\hat{\varsigma})V_D^{\mathrm{T}},\ \ \hat{\varsigma} = shrink(\varsigma,1/\nu) = sign(\varsigma)\cdot\max\{|\varsigma|-1/\nu,0\},$$

where U_D, V_D, Σ_D are from the singular value decomposition of D, i.e., $D = U_D\Sigma_D V_D^{\mathrm{T}}$, and ς is a vector that contains the diagonal element of Σ_D. The proof can be found in [2].

Let $\nu > 0$, and $p(M) = \delta(M;\Omega^*)$, then the proximal mapping of p is

$$Prox_p(D) = \underset{M\in\mathbb{R}^{m\times q}}{\arg\min}\left\{\delta(M;\Omega^*) + \frac{\nu}{2}\|M-D\|_F^2\right\},\ \ \forall\, D\in\mathbb{R}^{m\times q}$$

with $\Omega^* = B_{\|\cdot\|_2(\mathbf{0};\lambda)}$ and it also has a closed-form solution, which is

$$Prox_p(D) = U_D Diag(\hat{\varsigma})V_D^{\mathrm{T}},\ \ \hat{\varsigma} = \Pi(\varsigma; B_{\|\cdot\|_\infty(0;\lambda)}). \tag{7}$$

In special cases, the proximal mapping is a projection and plays an important role in solving the problem. Based on it, we derive an efficient sGS-ADMM algorithm to solve DFMR and DFMLR.

A.2. An introduction to sGS-ADMM algorithm

Now we give a brief introduction on sGS-ADMM algorithm for a general convex composite programming model as discussed in [3]. Let m and n be two nonnegative integers, $\mathscr{X}, \mathscr{Y}_i, 1\le i\le m$ and $\mathscr{Z}_j, 1\le j\le n$, be finite dimensional Euclidean spaces. Define $\mathscr{Y} := \mathscr{Y}_1\times\cdots\times\mathscr{Y}_m$ and $\mathscr{Z} := \mathscr{Z}_1\times\cdots\times\mathscr{Z}_n$. Consider the following general convex composite programming model:

$$\min_{y\in\mathscr{Y},z\in\mathscr{Z}}\{p_1(y_1) + f(y_1,\cdots,y_m) + q_1(z_1) + g(z_1,\cdots,z_n)\mid \mathscr{A}^*y + \mathscr{B}^*z = c\}, \tag{8}$$

where $p_1 : \mathscr{Y}_1\to(-\infty,+\infty]$ and $q_1 : \mathscr{Z}_1\to(-\infty,+\infty]$ are two closed proper convex functions, $f : \mathscr{Y}\to(-\infty,+\infty)$ and $g : \mathscr{Z}\to(-\infty,+\infty)$ are continuously differentiable convex functions whose gradients are Lipschitz continuous. The linear mappings $\mathscr{A} : \mathscr{Y}\to\mathscr{X}$ and $\mathscr{B} : \mathscr{Z}\to\mathscr{X}$ are defined such that their adjoints are given by $\mathscr{A}^*y = \sum_{i=1}^{m}\mathscr{A}_i^*y_i$ for $y = (y_1,\cdots,y_m)\in\mathscr{Y}$ and $\mathscr{B}^*z = \sum_{j=1}^{n}\mathscr{B}_j^*z_j$ for $z = (z_1,\cdots,z_n)\in\mathscr{Z}$, where $\mathscr{A}_i^* : \mathscr{Y}_i\to\mathscr{X}$ and $\mathscr{B}_j^* : \mathscr{Z}_j\to\mathscr{X}$ are the adjoints of the linear mappings $\mathscr{A}_i : \mathscr{X}\to\mathscr{Y}_i$ and $\mathscr{B}_j : \mathscr{X}\to\mathscr{Z}_j$, respectively.

The augmented Lagrangian function of problem (8) is defined as follows.

$$\mathscr{L}_\sigma(y,z;x) := p_1(y_1) + f(y) + q_1(z_1) + g(z) + \langle x, \mathscr{A}^*y + \mathscr{B}^*z - c\rangle + \frac{\sigma}{2}\|\mathscr{A}^*y + \mathscr{B}^*z - c\|^2,$$

where $\sigma > 0$ is a penalty parameter, and x is the Lagrange multiplier. With initial point $(y^0,z^0,x^0)\in dom p_1\times dom q_1\times\mathscr{X}$, where $dom p_1$ and $dom q_1$ denote the domain of p_1 and q_1, the iterative scheme of sGS-ADMM algorithm for (8) in the $(k+1)$th $(k = 0,1,2,\ldots)$ iteration is

$$\left\{\begin{array}{rcl}
\tilde{y}_i^{k+1} &=& \arg\min\{\mathscr{L}_\sigma(y_{\le i-1}^k, y_i, \tilde{y}_{\ge i+1}^{k+1}, z^k; x^k)\}, i = m,\cdots,2,\\
y_i^{k+1} &=& \arg\min\{\mathscr{L}_\sigma(y_{\le i-1}^{k+1}, y_i, \tilde{y}_{\ge i+1}^{k+1}, z^k; x^k)\}, i = 1,\cdots,m,\\
\tilde{z}_j^{k+1} &=& \arg\min\{\mathscr{L}_\sigma(y^{k+1}, z_{\le j-1}^k, z_j, \tilde{z}_{\ge j+1}^{k+1}; x^k)\}, j = n,\cdots,2,\\
z_j^{k+1} &=& \arg\min\{\mathscr{L}_\sigma(y^{k+1}, z_{\le j-1}^{k+1}, z_j, \tilde{z}_{\ge j+1}^{k+1}; x^k)\}, j = 1,\cdots,n,\\
x^{k+1} &=& x^k - \tau\sigma(\mathscr{A}^*y^{k+1} + \mathscr{B}^*z^{k+1} - c),
\end{array}\right.$$

where $y_{\le i-1} := (y_1,\ldots,y_{i-1})$, $\tilde{y}_{\ge i+1} := (\tilde{y}_{i+1},\ldots,\tilde{y}_m)$, $z_{\le j-1} := (z_1,\ldots,z_{j-1})$, $\tilde{z}_{\ge j+1} := (\tilde{z}_{j+1},\ldots,\tilde{z}_n)$ and τ is the step length. Now, we present the convergence theorem of sGS-ADMM algorithm.

Theorem 3. *Suppose that the solution set $\bar{W}$ to the KKT system of problem (8) is nonempty and the sequence $\{y^k, z^k, x^k\}$ is generated by the sGS-ADMM in the kth iteration. Let Σ_f, Σ_g, S and T be self-adjoint positive semidefinite linear operators such that $\Sigma_f + S + \sigma\mathscr{A}\mathscr{A}^* \succ 0$, $\Sigma_g + T + \sigma\mathscr{B}\mathscr{B}^* \succ 0$. Then the sequence $\{y^k, z^k, x^k\}$ converges to a point in $\bar{W}$.*

A.3. Iterative scheme sGS-ADMM algorithm for DFMR

The each subproblems in Table 1 have closed-form solutions, which are obtained from the derivative of the augmented Lagrangian function or the properties of proximal mapping. Now let us look at the resulting subproblems in Table 1 one by one. The $u-$subproblem can be written as

$$u^{k+1} = \arg\min_u \left\{ \frac{1}{2}\|u\|_2^2 - u^{\mathrm{T}}y - \langle x_1^k, \mathbb{X}^{\mathrm{T}}u\rangle - \langle x_2^k, \mathbb{Z}^{\mathrm{T}}u\rangle + \frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}}u - w^k\|^2 + \frac{\sigma}{2}\|\mathbb{X}^{\mathrm{T}}u + C^{\mathrm{T}}v^k - vec(D^k)\|^2 \right\}.$$

It is a quadratic form of u and has a unique closed-form solution

$$u^{k+1} = (I + \sigma\mathbb{X}\mathbb{X}^{\mathrm{T}} + \sigma\mathbb{Z}\mathbb{Z}^{\mathrm{T}})^{-1}(y + \mathbb{X}x_1^k + \mathbb{Z}x_2^k - \sigma(\mathbb{X}C^{\mathrm{T}}v^k - \mathbb{X}vec(D^k) - \mathbb{Z}w^k)).$$

Similarly, the unique closed-form solution of the $v-$subproblem is

$$v^{k+1} = \frac{1}{\sigma}(I + CC^{\mathrm{T}})^{-1}(Cx_1^k + x_3^k - \sigma(C\mathbb{X}^{\mathrm{T}}u^{k+1} - Cvec(D^k) + t^k)).$$

The $D-$subproblem can be written as

$$\begin{aligned} D^{k+1} &= \arg\min_D \left\{ \delta(D; B_{\|\cdot\|_2}(\mathbf{0};\lambda_1)) + \frac{\sigma}{2}\|\mathbb{X}^{\mathrm{T}}u^{k+1} + C^{\mathrm{T}}v^{k+1} - vec(D) - \frac{x_1^k}{\sigma}\|_2^2 \right\} \\ &= \arg\min_D \left\{ \delta(D; B_{\|\cdot\|_2}(\mathbf{0};\lambda_1)) + \frac{\sigma}{2}\|D - E\|_F^2 \right\}, \end{aligned}$$

where $vec(E) = \mathbb{X}^{\mathrm{T}}u^{k+1} + C^{\mathrm{T}}v^{k+1} - x_1^k/\sigma$. By (7) of the Supplementary material, the closed-form solution is $D^{k+1} = UDiag(e^*)V^{\mathrm{T}}$, where U, V, e satisfy the singular value decomposition of E, i.e., $E = U\Sigma V^{\mathrm{T}}$. The vector e contains the diagonal element of Σ, and $e^* = \Pi(e; B_{\|\cdot\|_\infty(0;\lambda_1)})$.

The $t-$subproblem can be written as

$$\begin{aligned} t^{k+1} &= \arg\min_t \left\{ \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) - \langle x_3^k, v^{k+1} + t\rangle + \frac{\sigma}{2}\|v^{k+1} + t\|_2^2 \right\} \\ &= \arg\min_t \left\{ \delta(t; B_{\|\cdot\|_\infty(0;\lambda_2)}) + \frac{\sigma}{2}\|v^{k+1} + t - x_3^k/\sigma\|_2^2 \right\}. \end{aligned}$$

By (3), the closed-form solution can be obtained from the soft-thresholding operator

$$\begin{aligned} t^{k+1} &= \Pi(x_3^k/\sigma - v^{k+1}; B_{\|\cdot\|_\infty(0;\lambda_2)}) \\ &= (x_3^k/\sigma - v^{k+1}) - sign(x_3^k/\sigma - v^{k+1}) \cdot \max\left\{ |v^{k+1} - x_3^k/\sigma| - \lambda_2, 0 \right\}. \end{aligned}$$

The $w-$subproblem can be written as

$$\begin{aligned} w^{k+1} &= \arg\min_{w}\left\{P^*(w)-\langle x_2^k,\mathbb{Z}^{\mathrm{T}}u^{k+1}-w\rangle+\frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}}u^{k+1}-w\|_2^2\right\} \\ &= \arg\min_{w}\left\{P^*(w)+\frac{\sigma}{2}\|\mathbb{Z}^{\mathrm{T}}u^{k+1}-w-x_2^k/\sigma\|_2^2\right\}. \end{aligned}$$

Applying the Moreau identity (2), we have

$$w^{k+1}=Prox_{P^*/\sigma}(\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k/\sigma)=(\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k/\sigma)-1/\sigma Prox_{\sigma P}(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k).$$

The closed-form solution for $Prox_{\sigma P}(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k)$ is given by

$$Prox_{\sigma P}(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k)=sign(x_{\sigma\lambda_4}(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k))\cdot\max\left\{|x_{\sigma\lambda_4}(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k)|-\sigma\lambda_3,0\right\},$$

where $x_{\sigma\lambda_4}(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k)=\arg\min_{x}\left\{\sigma\lambda_4\|Ax\|_1+\frac{1}{2}\|x-(\sigma\mathbb{Z}^{\mathrm{T}}u^{k+1}-x_2^k)\|_2^2\right\}$.

Finally, the stopping criterion eta for DFMR estimator is derived from the KKT condition.

$$\eta_P=\max\left\{\frac{\|u^{k+1}-y-\mathbb{X}x_1^{k+1}-\mathbb{Z}x_2^{k+1}\|}{1+\|u^{k+1}\|+\|x_2^{k+1}\|},\frac{\|Cx_1^{k+1}+x_3^{k+1}\|}{1+\|x_1^{k+1}\|+\|x_3^{k+1}\|}\right\},$$

$$\eta_D=\max\left\{\frac{\|v^{k+1}+t^{k+1}\|}{1+\|v^{k+1}\|+\|t^{k+1}\|},\frac{\|\mathbb{Z}^{\mathrm{T}}u^{k+1}-w^{k+1}\|}{1+\|w^{k+1}\|},\frac{\|\mathbb{X}^{\mathrm{T}}u^{k+1}+C^{\mathrm{T}}v^{k+1}-vec(D^{k+1})\|}{1+\|vec(D^{k+1})\|}\right\},$$

$$eta=\max\{\eta_P,\eta_D\}<tol.$$

The maximum number of iterations k is set to be 50000.

A.4. Iterative scheme of sGS-ADMM algorithm for DFMLR

The iterative scheme of sGS-ADMM algorithm for solving (11) is summarized in Table 1.

TABLE 1
Iterative scheme of sGS-ADMM algorithm for solving (11)

Algorithm 2:

Input: X, Z, y and tolerance level tol. Choose $\lambda_1 > 0, \lambda_2 > 0, \lambda_3 > 0, \lambda_4 > 0$ and $\sigma > 0$.

Let $\tau \in (0, (1+\sqrt{5})/2)$ be the step-length. Set the initial point $(u^0, v^0, \alpha^0, x^0)$.

For $k = 0, 1, \cdots$, perform the following steps:

Step 1a. (Backward GS sweep) Compute $u^{k+\frac{1}{2}}$ and $v^{k+\frac{1}{2}}$,

$$u^{k+\frac{1}{2}} = \arg\min_u \mathscr{L}_\sigma(u, v^k, \alpha^k; x^k),$$
$$v^{k+\frac{1}{2}} = \arg\min_v \mathscr{L}_\sigma(u^{k+\frac{1}{2}}, v, \alpha^k; x^k).$$

Step 1b. (Forward GS sweep) Compute u^{k+1}, v^{k+1} and α^{k+1},

$$\alpha^{k+1} = \arg\min_\alpha \mathscr{L}_\sigma(u^{k+\frac{1}{2}}, v^{k+\frac{1}{2}}, \alpha; x^k),$$
$$v^{k+1} = \arg\min_v \mathscr{L}_\sigma(u^{k+\frac{1}{2}}, v, \alpha^{k+1}; x^k),$$
$$u^{k+1} = \arg\min_u \mathscr{L}_\sigma(u, v^{k+1}, \alpha^{k+1}; x^k).$$

Step 2. Update Lagrange multipliers $x_1^{k+1}, x_2^{k+1}, x_3^{k+1}$ and x_4^{k+1},

$$x_1^{k+1} = x_1^k - \tau\sigma(\mathbb{X}^{\mathrm{T}}u^{k+1} + C^{\mathrm{T}}v^{k+1} - vec(D^{k+1})),$$
$$x_2^{k+1} = x_2^k - \tau\sigma(\mathbb{Z}^{\mathrm{T}}u^{k+1} - w^{k+1}),$$
$$x_3^{k+1} = x_3^k - \tau\sigma(u^{k+1} + s^{k+1} - y),$$
$$x_4^{k+1} = x_4^k - \tau\sigma(v^{k+1} + t^{k+1}).$$

If $eta < tol$ **stop**

The D, w, t subproblems have the same solutions as in the DFMR. Now we give the closed-form solutions for $u-$subproblem and $v-$ subproblem. The solution of u-subproblem is

$$u^{k+1} = \frac{1}{\sigma}(I + \mathbb{X}\mathbb{X}^{\mathrm{T}} + \mathbb{Z}\mathbb{Z}^{\mathrm{T}})^{-1}\left(\mathbb{X}x_1^k + \mathbb{Z}x_2^k + x_3^k - \sigma\big(\mathbb{X}C^{\mathrm{T}}v^k - \mathbb{X}vec(D^k) - \mathbb{Z}w^k\big) + s^k - y\right).$$

Similar to the u-subproblem, $v-$subproblem has a unique closed-form solution

$$v^{k+1} = \frac{1}{\sigma}(I + CC^{\mathrm{T}})^{-1}(Cx_1^k + x_4^k - \sigma(C\mathbb{X}^{\mathrm{T}}u^{k+1} - Cvec(D^k) + t^k)).$$

The stopping criterion eta for DFMLR estimator is also derived from the KKT condition.

$$\eta_P = \max\left\{\frac{\|\mathbb{X}x_1^{k+1} + \mathbb{Z}x_2^{k+1} + x_3^{k+1}\|}{1 + \|x_1^{k+1}\| + \|x_2^{k+1}\| + \|x_3^{k+1}\|}, \frac{\|Cx_1^{k+1} + x_4^{k+1}\|}{1 + \|x_1^{k+1}\| + \|x_4^{k+1}\|}\right\},$$
$$\eta_D = \max\left\{\frac{\|v^{k+1} + t^{k+1}\|}{1 + \|v^{k+1}\| + \|t^{k+1}\|}, \frac{\|\mathbb{Z}^{\mathrm{T}}u^{k+1} - w^{k+1}\|}{1 + \|w^{k+1}\|}, \frac{\|\mathbb{X}^{\mathrm{T}}u^{k+1} + C^{\mathrm{T}}v^{k+1} - vec(D^{k+1})\|}{1 + \|vec(D^{k+1})\|}, \right.$$
$$\left.\frac{\|u^{k+1} + s^{k+1} - y\|}{1 + \|u^{k+1}\| + \|s^{k+1}\|}\right\},$$
$$eta = \max\{\eta_P, \eta_D\} < tol.$$

The maximum number of iterations k was set to be 50000.

A.5. Corollaries for Theorem 4 and Theorem 7

Corollaries 4 and 5 give risk bounds of the estimators from two special cases of the DFMR problem (3): the matrix-type fused Lasso and the fused Lasso.

Corollary 4. *Suppose that the data* $(y,\mathbb{X})$ *satisfies* $y=\mathbb{X}vec(B^*)+\varepsilon$. *Define* $W^*\in\partial\|B^*\|_*$ *and*

$$\hat{B}(\lambda_1,\lambda_2)=\arg\min_B\left\{\frac{1}{2}\|y-\mathbb{X}vec(B)\|_2^2+\lambda_1\|B\|_*+\lambda_2\|Cvec(B)\|_1\right\}.$$

Assume that $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ *is non-singular, and* $0<\underline{\sigma}\le\lambda_{\min}(\frac{1}{n}\mathbb{X}^{\mathrm{T}}\mathbb{X})\le\lambda_{\max}(\frac{1}{n}\mathbb{X}^{\mathrm{T}}\mathbb{X})\le\overline{\sigma}$. *Then*

$$E(\|\hat{B}(\lambda_1,\lambda_2)-B^*\|_F^2)\le 16\frac{\lambda_2^2(m-1)q\|C\|_F^2+mq\overline{\sigma}n\sigma^2+\lambda_1^2\|W^*\|_F^2}{(\underline{\sigma}n)^2}.$$

Corollary 5. *Suppose that the data* $(y,\mathbb{Z})$ *satisfies* $y=\mathbb{Z}\gamma^*+\varepsilon$. *Define*

$$\hat{\gamma}(\lambda_3,\lambda_4)=\arg\min_\gamma\left\{\frac{1}{2}\|y-\mathbb{Z}\gamma\|_2^2+\lambda_3\|\gamma\|_1+\lambda_4\|A\gamma\|_1\right\}.$$

Assume that $\mathbb{Z}^{\mathrm{T}}\mathbb{Z}$ *is non-singular, and* $0<\underline{\sigma}\le\lambda_{\min}(\frac{1}{n}\mathbb{Z}^{\mathrm{T}}\mathbb{Z})\le\lambda_{\max}(\frac{1}{n}\mathbb{Z}^{\mathrm{T}}\mathbb{Z})\le\overline{\sigma}$. *Then*

$$E(\|\hat{\gamma}(\lambda_3,\lambda_4)-\gamma^*\|_2^2)\le 16\frac{\lambda_4^2(p-1)\|A\|_F^2+p\overline{\sigma}n\sigma^2+\lambda_3^2}{(\underline{\sigma}n)^2}.$$

Corollaries 6 and 7 provide risk bounds for two special cases of DFMLR, i.e., matrix-type fused Lasso regularized logistic regression and fused Lasso regularized logistic regression. Under the matrix-type fused Lasso regularized logistic regression, the optimal solution is

$$\hat{B}(\lambda_1,\lambda_2)=\arg\min_B\left\{\sum_{i=1}^n\log(1+e^{\langle X_i,B\rangle})-y_i\langle X_i,B\rangle+\lambda_1\|B\|_*+\lambda_2\|Cvec(B)\|_1\right\}.$$

Corollary 6. *Assume that* $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ *is non-singular, and* $0<\underline{\sigma}\le\lambda_{\min}(\frac{1}{n}\mathbb{X}^{\mathrm{T}}\mathbb{X})\le\lambda_{\max}(\frac{1}{n}\mathbb{X}^{\mathrm{T}}\mathbb{X})\le\overline{\sigma}$. *Suppose that* $\nabla^2R(\hat{B})\succeq L\lambda_{\min}(\mathbb{X}^{\mathrm{T}}\mathbb{X})I\succ 0$, *where* $R(B)=\sum_{i=1}^n\log(1+e^{\langle X_i,B\rangle})-y_i\langle X_i,B\rangle$. *Then*

$$E(\|\hat{B}(\lambda_1,\lambda_2)-B^*\|_F^2)\le 4\frac{\lambda_2^2(m-1)q\|C\|_F^2+mq\overline{\sigma}n+\lambda_1^2\|W^*\|_F^2}{(\underline{\sigma}nL)^2}.$$

Under the fused Lasso regularized logistic regression, the optimal solution is obtained by

$$\hat{\gamma}(\lambda_3,\lambda_4)=\arg\min_\gamma\left\{\sum_{i=1}^n\log(1+e^{\langle z_i,\gamma\rangle})-y_i\langle z_i,\gamma\rangle+\lambda_3\|\gamma\|_1+\lambda_4\|A\gamma\|_1\right\}.$$

Corollary 7. *Assume that* $\mathbb{Z}^{\mathrm{T}}\mathbb{Z}$ *is non-singular, and* $0<\underline{\sigma}\leq\lambda_{\min}(\frac{1}{n}\mathbb{Z}^{\mathrm{T}}\mathbb{Z})\leq\lambda_{\max}(\frac{1}{n}\mathbb{Z}^{\mathrm{T}}\mathbb{Z})\leq\overline{\sigma}$. *Suppose that* $\nabla^2 R(\hat{\gamma})\succeq L\lambda_{\min}(\mathbb{Z}^{\mathrm{T}}\mathbb{Z})I\succ 0$, *where* $R(\gamma)=\sum_{i=1}^{n}\log(1+e^{\langle z_i,\gamma\rangle})-y_i\langle z_i,\gamma\rangle$. *Then*

$$E(\|\hat{\gamma}(\lambda_3,\lambda_4)-\gamma^*\|_2^2)\leq 4\frac{\lambda_4^2(p-1)\|A\|_F^2+p\overline{\sigma}n+\lambda_3^2}{(\underline{\sigma}nL)^2}.$$

A.6. Proofs

Proof of Theorem 1. The global convergence of sGS-ADMM algorithm is established by [3]. For the problem (8) in our paper is a convex optimization problem which has better structures where $I+\sigma\mathbb{X}\mathbb{X}^{\mathrm{T}}+\sigma\mathbb{Z}\mathbb{Z}^{\mathrm{T}}$ and $I+CC^{\mathrm{T}}$ are positive definite matrices. The global convergence of the sGS-ADMM algorithm for solving problem (8) is easily satisfied. □

In order to prove the Theorem 2, we established two lemmas which give the upper bound of the KKT system of the iterative point and investigate the distance between iterative points and the optimal solution, respectively. We give the definition of metric subregularity from [4]. Let $\mathscr{F}:\mathscr{X}\to\mathscr{Y}$ be a multi-valued mapping. Denote its inverse by $\mathscr{F}^{-1}$. Define the graph of multi-valued functions $\mathscr{F}$ as follows

$$graph\mathscr{F}:=\{(x,y)\in\mathscr{X}\times\mathscr{Y}|y\in\mathscr{F}(x)\}.$$

Definition 8. *A multi-valued mapping* $\mathscr{F}:\mathscr{X}\to\mathscr{Y}$ *is said to be metric subregular at* $\tilde{x}\in\mathscr{X}$ *for* $\tilde{y}\in\mathscr{Y}$ *with modulus* $\kappa>0$ *if* $(\bar{x},\bar{y})\in gragh\mathscr{F}$ *and there exist neighborhood* $\mathscr{U}$ *of* $\bar{x}$ *and* $\mathscr{V}$ *of* $\bar{y}$ *such that*

$$dist(x,\mathscr{F}^{-1}(\bar{y}))\leq\kappa dist(\bar{y},\mathscr{F}(x)\cap\mathscr{V}),\forall x\in\mathscr{U}.$$

Let $\Theta:=\mathbb{R}^n\times\mathbb{R}^{(m-1)q}\times\mathscr{Y}\times\mathscr{X}$, and define the KKT mapping $R:\Theta\to\Theta$ as

$$R(\theta):=\begin{pmatrix} u-y-\mathbb{X}x_1-\mathbb{Z}x_2\\ -Cx_1-x_3\\ w-Prox_{P^*}(w-x_2)\\ D-Prox_{\delta_{B_{\|\cdot\|_2\leq\lambda_1}}}(D-\Xi)\\ t-Prox_{\delta_{B_{\|\cdot\|_\infty\leq\lambda_2}}}(t+x_3)\\ \mathbb{X}^{\mathrm{T}}u+C^{\mathrm{T}}v-vec(D)\\ \mathbb{Z}^{\mathrm{T}}u-w\\ v+t\end{pmatrix},\quad\forall\theta\in\Theta,$$

where $vec(\Xi)=x_1$. We know that $R(\theta)=0\Leftrightarrow\theta\in\bar{\Omega}$. According to Moreau identity (2) $Prox_{\delta_{B_{\|\cdot\|_2\leq\lambda_1}}}(D-\Xi)+Prox_{\lambda_1\|\cdot\|_*}(D-\Xi)=D-\Xi$, we have $D-Prox_{\delta_{B_{\|\cdot\|_2\leq\lambda_1}}}(D-\Xi)=$

$\Xi + Prox_{\lambda_1\|\cdot\|_*}(D-\Xi)$. So we obtain that

$$R(\theta) := \begin{pmatrix} u-y-\mathbb{X}x_1-\mathbb{Z}x_2 \\ -Cx_1-x_3 \\ w-Prox_{P^*}(w-x_2) \\ D-Prox_{\delta_{B_{\|\cdot\|_2\le\lambda_1}}}(D-\Xi) \\ t-Prox_{\delta_{B_{\|\cdot\|_\infty\le\lambda_2}}}(t+x_3) \\ \mathbb{X}^{\mathrm{T}}u+C^{\mathrm{T}}v-vec(D) \\ \mathbb{Z}^{\mathrm{T}}u-w \\ v+t \end{pmatrix} = \begin{pmatrix} u-y-\mathbb{X}x_1-\mathbb{Z}x_2 \\ -Cx_1-x_3 \\ w-Prox_{P^*}(w-x_2) \\ \Xi+Prox_{\lambda_1\|\cdot\|_*}(D-\Xi) \\ t-Prox_{\delta_{B_{\|\cdot\|_\infty\le\lambda_2}}}(t+x_3) \\ \mathbb{X}^{\mathrm{T}}u+C^{\mathrm{T}}v-vec(D) \\ \mathbb{Z}^{\mathrm{T}}u-w \\ v+t \end{pmatrix}.$$

Define $\kappa_1 := (12\sigma+4\tau)\lambda_{\max}(\mathbb{X}^{\mathrm{T}}\mathbb{X})$, $\kappa_2 := (8\sigma+4\tau)\lambda_{\max}(\mathbb{Z}^{\mathrm{T}}\mathbb{Z})$, $\kappa_3 := (12\sigma+4\tau)\lambda_{\max}(C^{\mathrm{T}}C)$, and $\kappa_4 := \max\{\kappa_1+\kappa_3+2\sigma+2\tau+1/\sigma, \kappa_2+2\sigma+2\tau+1/\sigma, 14\sigma+5\tau+1/\sigma\}$. Let $\mathscr{H}_0$ be the block-diagonal linear operator defined by $\mathscr{H}_0 := \kappa_4 Diag(0, CC^{\mathrm{T}}, \Sigma_I, (\tau^2\sigma)^{-1}I_x)$. We provide the following lemma which is useful in proving Theorem 2.

Lemma 1. *Let $\{\theta^k := (u^k, v^k, \alpha^k, x^k)\}$ is generated by the sGS-ADMM. Then for any $k \ge 1$,*

$$\|\theta^{k+1}-\theta^k\|^2_{\mathscr{H}_0} \ge \|R(\theta^{k+1})\|^2. \tag{9}$$

Proof The optimal condition for every subproblem in Table 1 is

$$\begin{cases} 0 = u^{k+1}-y-\mathbb{X}x_1^k-\mathbb{Z}x_2^k+\sigma\mathbb{X}(\mathbb{X}^{\mathrm{T}}u^{k+1}+C^{\mathrm{T}}v^k-vec(D^k))+\sigma\mathbb{Z}(\mathbb{Z}^{\mathrm{T}}u^{k+1}-w^k), \\ 0 = -Cx_1^k-x_3^k+\sigma C(\mathbb{X}u^{k+1}+C^{\mathrm{T}}v^{k+1}-vec(D^k))+\sigma(v^{k+1}+t^k), \\ 0 \in \partial P^*(w^{k+1})+x_2^k-\sigma(\mathbb{Z}^{\mathrm{T}}u^{k+1}-w^{k+1}), \\ 0 \in \partial\delta_{B_{\|\cdot\|_2\le\lambda_1}}(D^{k+1})+\Xi^k-\sigma\Lambda^{k+1}, \\ 0 \in \partial\delta_{B_{\|\cdot\|_\infty\le\lambda_2}}(t^{k+1})-x_3^k+\sigma(v^{k+1}+t^{k+1}), \end{cases}$$

where $vec(\Lambda^{k+1}) = \mathbb{X}u^{k+1}+C^{\mathrm{T}}v^{k+1}-vec(D^{k+1})$. We obtain from the definition of $R(\cdot)$ that

$$\begin{aligned} \|R(\theta^{k+1})\|^2 \quad \le \quad & \|u^{k+1}-y-\mathbb{X}x_1^{k+1}-\mathbb{Z}x_2^{k+1}\|^2+\|Cx_1^{k+1}+x_3^{k+1}\|^2+\|\sigma(\mathbb{Z}^{\mathrm{T}}u^{k+1}-w^{k+1})-x_2^k+x_2^{k+1}\|^2 \\ & +\|\sigma\Lambda^{k+1}-\Xi^k+\Xi^{k+1}\|_F^2+\|\sigma(v^{k+1}+t^{k+1})-x_3^k+x_3^{k+1}\|^2+\|v^{k+1}+t^{k+1}\|^2 \\ & +\|\mathbb{X}^{\mathrm{T}}u^{k+1}+C^{\mathrm{T}}v^{k+1}-vec(D^{k+1})\|^2+\|\mathbb{Z}^{\mathrm{T}}u^{k+1}-w^{k+1}\|^2. \end{aligned}$$

According to schemes of Lagrange multipliers and the definition of Λ^{k+1}, we have

$$\begin{aligned}
\|R(\theta^{k+1})\|^2 \leq\ & 12\sigma^2\lambda_{max}(\mathbb{X}^{\mathbb{T}}\mathbb{X})\|v^k-v^{k+1}\|^2_{CC^{\mathrm{T}}}+8\sigma^2\lambda_{max}(\mathbb{Z}^{\mathbb{T}}\mathbb{Z})\|w^{k+1}-w^k\|^2\\
&+(12\sigma^2\lambda_{max}(\mathbb{X}^{\mathbb{T}}\mathbb{X})+12\sigma^2\lambda_{max}(C^TC))\|vec(D^{k+1})-vec(D^k)\|^2\\
&+12\sigma^2\|t^k-t^{k+1}\|^2+((12\sigma^2+4\tau\sigma)\lambda_{max}(\mathbb{X}^{\mathbb{T}}\mathbb{X})+(12\sigma^2+3\tau\sigma)\\
&\lambda_{max}(C^TC)+2\sigma^2+2\tau\sigma+1)\|(\tau\sigma)^{-1}(x_1^k-x_1^{k+1})\|^2\\
&+((8\sigma^2+4\tau\sigma)\lambda_{max}(\mathbb{Z}^{\mathbb{T}}\mathbb{Z})+2\sigma^2+2\tau\sigma+1)\|(\tau\sigma)^{-1}(x_2^k-x_2^{k+1})\|^2\\
&+(14\sigma^2+5\tau\sigma+1)\|(\tau\sigma)^{-1}(x_3^k-x_3^{k+1})\|^2\\
\leq\ & \kappa_1\|v^k-v^{k+1}\|^2_{CC^{\mathrm{T}}}+\kappa_2\|w^{k+1}-w^k\|^2+(\kappa_1+\kappa_3)\|vec(D^{k+1})-vec(D^k)\|^2\\
&+12\sigma^2\|t^k-t^{k+1}\|^2+(\kappa_1+\kappa_3+2\sigma^2+2\tau\sigma+1)\|(\tau\sigma)^{-1}(x_1^k-x_1^{k+1})\|^2\\
&+(\kappa_2+2\sigma^2+2\tau\sigma+1)\|(\tau\sigma)^{-1}(x_2^k-x_2^{k+1})\|^2\\
&+(14\sigma^2+5\tau\sigma+1)\|(\tau\sigma)^{-1}(x_3^k-x_3^{k+1})\|^2.
\end{aligned}$$

Thus we can immediately imply (9). □

Define $t_\tau := \frac{1}{2}(1-\tau+\min\{\tau,\tau^{-1}\})$ and the self-adjoint linear operator $\mathscr{H} := Diag(\frac{1}{2}I, 2t_\tau\tau\sigma(I+CC^{\mathrm{T}}), 2t_\tau\tau\sigma\Sigma_I+\frac{1}{2}\Sigma_\alpha, t_\tau(\tau^2\sigma)^{-1}I_x)+\frac{1}{4}t_\tau\sigma\Phi\Phi^*$. We easily know that $1/4 \leq s_\tau \leq 5/4$ and $0 \leq t_\tau \leq 1/2$. We next give inequality which plays a key role in the Q-linear rate of convergence for the sGS-ADMM algorithm of DFMR estimator.

Lemma 2. *Let $\{\theta^k := (u^k, v^k, \alpha^k, x^k)\}$ be an infinite sequence generated by the sGS-ADMM. Then for any $\bar{\theta} = (\bar{u}, \bar{v}, \bar{\alpha}, \bar{x}) \in \bar{\Omega}$ and any $k \geq 1$,*

$$\|\theta^{k+1}-\bar{\theta}\|^2_{\mathscr{M}} \leq \|\theta^k-\bar{\theta}\|^2_{\mathscr{M}} - \|\theta^{k+1}-\theta^k\|^2_{\mathscr{H}}. \tag{10}$$

Consequently, we have

$$dist^2_{\mathscr{M}}(\theta^{k+1}, \bar{\Omega}) \leq dist^2_{\mathscr{M}}(\theta^k, \bar{\Omega}) - \|\theta^{k+1}-\theta^k\|^2_{\mathscr{H}}. \tag{11}$$

Proof For any θ and θ', we define the function $J(\theta,\theta') := (\tau\sigma)^{-1}\|x-x'\|^2+\sigma\|v-v'\|^2_{I+CC^{\mathrm{T}}}+\sigma\|\Sigma_I(\alpha-\alpha')\|^2$. Let $\bar{\theta} = (\bar{u}, \bar{v}, \bar{\alpha}, \bar{x}) \in \bar{\Omega}$. Following Appendix B of [6], for any $k \geq 1$ the inequality holds that

$$\begin{aligned}
&J(\theta^{k+1}, \bar{\theta}) + (1-\min\{\tau,\tau^{-1}\})\sigma\|\Phi^*(u^{k+1}, v^{k+1}, \alpha^{k+1}, 0)\|^2\\
&-[J(\theta^k, \bar{\theta}) + (1-\min\{\tau,\tau^{-1}\})\sigma\|\Phi^*(u^k, v^k, \alpha^k, 0)\|^2]\\
\leq\ & -\tau(1-\tau+\min\{\tau,\tau^{-1}\})\sigma(\|v^{k+1}-v^k\|^2_{I+CC^{\mathrm{T}}}+\|\Sigma_I(\alpha^{k+1}-\alpha^k)\|^2)\\
&-2\|u^{k+1}-\bar{u}\|^2-2\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}-(1-\tau+\min\{\tau,\tau^{-1}\})\\
&\sigma\|\Phi^*(u^{k+1}, v^{k+1}, \alpha^{k+1}, 0)\|^2,
\end{aligned} \tag{12}$$

by reorganizing the terms in (12), we obtain

$$
\begin{aligned}
&(\tau\sigma)^{-1}\|x^{k+1}-\bar{x}\|^2+\sigma\|v^{k+1}-\bar{v}\|^2_{I+CC^{\mathrm{T}}}+\sigma\|\Sigma_I(\alpha^{k+1}-\bar{\alpha})\|^2\\
&+s_\tau\sigma\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2+\|u^{k+1}-\bar{u}\|^2+\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}\\
\le\ &(\tau\sigma)^{-1}\|x^{k}-\bar{x}\|^2+\sigma\|v^{k}-\bar{v}\|^2_{I+CC^{\mathrm{T}}}+\sigma\|\Sigma_I(\alpha^{k}-\bar{\alpha})\|^2\\
&+s_\tau\sigma\|\Phi^*(u^{k},v^{k},\alpha^{k},0)\|^2+\|u^{k}-\bar{u}\|^2+\|\alpha^{k}-\bar{\alpha}\|^2_{\Sigma_\alpha}\\
&-\{2t_\tau\tau\sigma[\|v^{k+1}-v^k\|^2_{I+CC^{\mathrm{T}}}+\|\Sigma_I(\alpha^{k+1}-\alpha^k)\|^2]+\|u^{k+1}-\bar{u}\|^2\\
&+\|u^k-\bar{u}\|^2+\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}+\|\alpha^k-\bar{\alpha}\|^2_{\Sigma_\alpha}+t_\tau\sigma\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2\\
&+\frac{1}{2}t_\tau\sigma[\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2+\|\Phi^*(u^k,v^k,\alpha^k,0)\|^2]\}.
\end{aligned}
$$

Using equalities

$$
\begin{aligned}
\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0) &= (\tau\sigma)^{-1}(x^k-x^{k+1}),\\
\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2 &= \|\theta^{k+1}-\bar{\theta}\|^2_{\Phi\Phi^*}
\end{aligned}
$$

and inequalities

$$
\begin{aligned}
\|u^{k+1}-\bar{u}\|^2+\|u^k-\bar{u}\|^2 &\ge \frac{1}{2}\|u^{k+1}-u^k\|^2,\\
\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}+\|\alpha^k-\bar{\alpha}\|^2_{\Sigma_\alpha} &\ge \frac{1}{2}\|\alpha^{k+1}-\alpha^k\|^2_{\Sigma_\alpha},\\
\|\Phi^*(u^{k+1},v^{k+1},\alpha^{k+1},0)\|^2+\|\Phi^*(u^k,v^k,\alpha^k,0)\|^2 &\ge \frac{1}{2}\|\theta^{k+1}-\theta^k\|^2_{\Phi\Phi^*},
\end{aligned}
$$

we obtain

$$
\begin{aligned}
&(\tau\sigma)^{-1}\|x^{k+1}-\bar{x}\|^2+\sigma\|v^{k+1}-\bar{v}\|^2_{I+CC^{\mathrm{T}}}+\sigma\|\alpha^{k+1}-\bar{\alpha}\|^2\\
&+s_\tau\sigma\|\theta^{k+1}-\bar{\theta}\|^2_{\Phi\Phi^*}+\|u^{k+1}-\bar{u}\|^2+\|\alpha^{k+1}-\bar{\alpha}\|^2_{\Sigma_\alpha}\\
\le\ &(\tau\sigma)^{-1}\|x^{k}-\bar{x}\|^2+\sigma\|v^{k}-\bar{v}\|^2_{I+CC^{\mathrm{T}}}+\sigma\|\alpha^{k}-\bar{\alpha}\|^2+s_\tau\sigma\|\theta^k-\bar{\theta}\|^2_{\Phi\Phi^*}\\
&+\|u^k-\bar{u}\|^2+\|\alpha^k-\bar{\alpha}\|^2_{\Sigma_\alpha}-\{2t_\tau\tau\sigma[\|v^{k+1}-v^k\|^2_{I+CC^{\mathrm{T}}}\\
&+\|\alpha^{k+1}-\alpha^k\|^2]+\frac{1}{2}\|u^{k+1}-u^k\|^2+\frac{1}{2}\|\alpha^{k+1}-\alpha^k\|^2_{\Sigma_\alpha}\\
&+t_\tau(\tau^2\sigma)^{-1}\|x^k-x^{k+1}\|^2+\frac{1}{4}t_\tau\sigma\|\theta^{k+1}-\theta^k\|^2_{\Phi\Phi^*}\}.
\end{aligned}
$$

It shows that (10) holds. Note that $\bar{\Omega}$ is a nonempty closed convex set and (10) holds for any $\bar{\theta}\in\bar{\Omega}$, we immediately get (11). □

Based on Lemma 1 and Lemma 2, we give a specific proof for the Q-linear rate of convergence of the sGS-ADMM algorithm.

Proof of Theorem 2. We know that L_1 norm and fused Lasso regularization are polyhedral convex functions [10, 13], their Fenchel conjugate functions are also polyhedral convex functions. According to [8], $Prox_{P^*}(\cdot)$ and $Prox_{\delta_{B_{\|\cdot\|_\infty\le\lambda_2}}}(\cdot)$ are piecewise polyhedral. The multi-valued mapping $\partial\|\cdot\|_*:\mathbb{R}^{m\times q}\to\mathbb{R}^{m\times q}$ is metrically subregular at the KKT point for origin [15]. So multi-valued mapping $R(\theta)$ is metrically subregular at the KKT point for origin. There exist positive constants $\hat{\eta}>0$

and $\hat{\delta} > 0$ such that $dist(\theta^k, \bar{\Omega}) \leq \hat{\eta}\|R(\theta^k)\|, \forall \theta \in \{\theta : \|\theta - \bar{\theta}\| \leq \hat{\delta}\}$. According to Lemma 1, it holds that $\|R(\theta^{k+1})\|^2 \leq \|\theta^{k+1} - \theta^k\|^2_{\mathscr{H}_0}$. Thus, we get that for all $k \geq 1$, $dist^2(\theta^{k+1}, \bar{\Omega}) \leq \hat{\eta}^2\|R(\theta^{k+1})\|^2 \leq \hat{\eta}^2\|\theta^{k+1} - \theta^k\|^2_{\mathscr{H}_0}$. We have for all $k \geq 1$, $\|\theta^{k+1} - \theta^k\|^2_{\mathscr{H}} \geq 0$ and

$$\begin{aligned} \|\theta^{k+1} - \theta^k\|^2_{\mathscr{H}} &\geq \min\{2\tau, 1\} t_\tau \kappa_4^{-1} \|\theta^{k+1} - \theta^k\|^2_{\mathscr{H}_0} \\ &\geq \min\{2\tau, 1\} t_\tau \kappa_4^{-1} \hat{\eta}^{-2} dist^2(\theta^{k+1}, \bar{\Omega}) \\ &\geq \kappa dist^2_{\mathscr{M}}(\theta^{k+1}, \bar{\Omega}), \end{aligned} \tag{13}$$

where $\kappa = \min\{2\tau, 1\} t_\tau \kappa_4^{-1} \hat{\eta}^{-2} > 0$. According to (11) and (13) we have $dist^2_{\mathscr{M}}(\theta^{k+1}, \bar{\Omega}) - dist^2_{\mathscr{M}}(\theta^k, \bar{\Omega}) \leq -\|\theta^{k+1} - \theta^k\|^2_{\mathscr{H}} \leq -\kappa dist^2_{\mathscr{M}}(\theta^{k+1}, \bar{\Omega})$, it holds that $(1+\kappa) dist^2_{\mathscr{M}}(\theta^{k+1}, \bar{\Omega}) \leq dist^2_{\mathscr{M}}(\theta^k, \bar{\Omega})$. Denote $\mu = (1+\kappa)^{-1} < 1$, The proof of Theorem 2 has been completed. □

The proof of Theorem 3 is similar to Theorem 2, we will omit it.

Proof of Theorem 4. For convenience, we denote

$$\bar{\lambda} := (\lambda_2, \lambda_4), \tilde{\lambda} := (\lambda_1, \lambda_3), \bar{C} := (C, A), \bar{X} := \mathscr{G} = (\mathbb{X}, \mathbb{Z}),$$

$$\beta := \begin{pmatrix} B \\ \gamma \end{pmatrix}, \beta^* := \begin{pmatrix} B^* \\ \gamma^* \end{pmatrix}, vec(\beta) := \begin{pmatrix} vec(B) \\ \gamma \end{pmatrix},$$

and define

$$\hat{\beta}(0, \tilde{\lambda}) = \begin{pmatrix} \hat{B}(0, \tilde{\lambda}) \\ \hat{\gamma}(0, \tilde{\lambda}) \end{pmatrix} = \arg\min_{B, \gamma} \left\{ \frac{1}{2}\|y - \mathbb{X}vec(B) - \mathbb{Z}\gamma\|_2^2 + \lambda_1\|B\|_* + \lambda_3\|\gamma\|_1 \right\}.$$

By the definition $\hat{\beta}(0, \tilde{\lambda})$ and $\hat{\beta}(\bar{\lambda}, \tilde{\lambda})$, we know

$$\begin{aligned} &\frac{1}{2}\|y - \mathbb{X}vec(\hat{B}(0, \tilde{\lambda})) - \mathbb{Z}\hat{\gamma}(0, \tilde{\lambda})\|_2^2 + \lambda_1\|\hat{B}(0, \tilde{\lambda})\|_* + \lambda_3\|\hat{\gamma}(0, \tilde{\lambda})\|_1 \\ \leq\ &\frac{1}{2}\|y - \mathbb{X}vec(\hat{B}(\bar{\lambda}, \tilde{\lambda})) - \mathbb{Z}\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_2^2 + \lambda_1\|\hat{B}(\bar{\lambda}, \tilde{\lambda})\|_* + \lambda_3\|\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_1 \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{2}\|y - \mathbb{X}vec(\hat{B}(\bar{\lambda}, \tilde{\lambda})) - \mathbb{Z}\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_2^2 + \lambda_1\|\hat{B}(\bar{\lambda}, \tilde{\lambda})\|_* + \lambda_2\|Cvec(\hat{B}(\bar{\lambda}, \tilde{\lambda}))\|_1 \\ &+ \lambda_3\|\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_1 + \lambda_4\|A\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_1 \\ \leq\ &\frac{1}{2}\|y - \mathbb{X}vec(\hat{B}(0, \tilde{\lambda})) - \mathbb{Z}\hat{\gamma}(0, \tilde{\lambda})\|_2^2 + \lambda_1\|\hat{B}(0, \tilde{\lambda})\|_* + \lambda_2\|Cvec(\hat{B}(0, \tilde{\lambda}))\|_1 \\ &+ \lambda_3\|\hat{\gamma}(0, \tilde{\lambda})\|_1 + \lambda_4\|A\hat{\gamma}(0, \tilde{\lambda})\|_1. \end{aligned}$$

According to two inequalities, we have

$$\begin{aligned} &\lambda_2\|Cvec(\hat{B}(0, \tilde{\lambda}))\|_1 - \lambda_2\|Cvec(\hat{B}(\bar{\lambda}, \tilde{\lambda}))\|_1 + \lambda_4\|A\hat{\gamma}(0, \tilde{\lambda})\|_1 - \lambda_4\|A\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_1 \\ \geq\ &\frac{1}{2}\|y - \mathbb{X}vec(\hat{B}(\bar{\lambda}, \tilde{\lambda})) - \mathbb{Z}\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_2^2 + \lambda_1\|\hat{B}(\bar{\lambda}, \tilde{\lambda})\|_* + \lambda_3\|\hat{\gamma}(\bar{\lambda}, \tilde{\lambda})\|_1 \\ &- \frac{1}{2}\|y - \mathbb{X}vec(\hat{B}(0, \tilde{\lambda})) - \mathbb{Z}\hat{\gamma}(0, \tilde{\lambda})\|_2^2 - \lambda_1\|\hat{B}(0, \tilde{\lambda})\|_* - \lambda_3\|\hat{\gamma}(0, \tilde{\lambda})\|_1. \end{aligned}$$

Denote $P(\beta)=\lambda_1\|B\|_*+\lambda_3\|\gamma\|_1$, $R(\beta)=\frac{1}{2}\|y-\bar{X}vec(\beta)\|_2^2=\frac{1}{2}\|y-\mathbb{X}vec(B)-\mathbb{Z}\gamma\|_2^2$. For the optimization problem $\min_{\beta}\{R(\beta)+P(\beta)\}$, the loss function $R(\beta)$ is differentiable. Following the proof of Lemma 3.1 in [5], we can imply that for all β the formula $\langle-\nabla R(\hat{\beta}(0,\tilde{\lambda})),\beta-\hat{\beta}(0,\tilde{\lambda})\rangle\leq P(\beta)-P(\hat{\beta}(0,\tilde{\lambda}))$ holds. We conclude that

$$\begin{aligned}&\frac{1}{2}\|y-\mathbb{X}vec(\hat{B}(\bar{\lambda},\tilde{\lambda}))-\mathbb{Z}\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_2^2+\lambda_1\|\hat{B}(\bar{\lambda},\tilde{\lambda})\|_*+\lambda_3\|\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1\\&-\frac{1}{2}\|y-\mathbb{X}vec(\hat{B}(0,\tilde{\lambda}))-\mathbb{Z}\hat{\gamma}(0,\tilde{\lambda})\|_2^2-\lambda_1\|\hat{B}(0,\tilde{\lambda})\|_*-\lambda_3\|\hat{\gamma}(0,\tilde{\lambda})\|_1\\&\geq(vec(\hat{\beta}(0,\tilde{\lambda}))-vec(\hat{\beta}(\bar{\lambda},\tilde{\lambda})))^{\mathrm{T}}(\frac{1}{2}\bar{X}^{\mathrm{T}}\bar{X})(vec(\hat{\beta}(0,\tilde{\lambda}))-vec(\hat{\beta}(\bar{\lambda},\tilde{\lambda}))).\end{aligned}\tag{14}$$

On the one hand, denote $\|\beta\|_{F2}^2:=\|B\|_F^2+\|\gamma\|_2^2$. Then it holds that

$$\begin{aligned}&\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))\|_1-\lambda_2\|Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1+\lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})\|_1-\lambda_4\|A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1\\&\leq\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))-Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1+\lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})-A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1\\&\leq\lambda_2\|C\|_F\sqrt{(m-1)q}\|vec(\hat{B}(0,\tilde{\lambda})-vec(\hat{B}(\bar{\lambda},\tilde{\lambda})\|_2+\lambda_4\|A\|_F\sqrt{p-1}\|\hat{\gamma}(0,\tilde{\lambda})-\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_2\\&\leq\max\{\lambda_2,\lambda_4\}\sqrt{2}\sqrt{(m-1)q+p-1}\|\bar{C}\|_F\|\hat{\beta}(0,\tilde{\lambda})-\hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}.\end{aligned}$$

By (14), we obtain that

$$\begin{aligned}&\frac{1}{2}(\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X}))\|\hat{\beta}(0,\tilde{\lambda})-\hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}^2\\&\leq(vec(\hat{\beta}(0,\tilde{\lambda}))-vec(\hat{\beta}(\bar{\lambda},\tilde{\lambda})))^{\mathrm{T}}(\frac{1}{2}\bar{X}^{\mathrm{T}}\bar{X})(vec(\hat{\beta}(0,\tilde{\lambda}))vec(\hat{\beta}(\bar{\lambda},\tilde{\lambda})))\\&\leq\max\{\lambda_2,\lambda_4\}\sqrt{(m-1)q+p-1}\|\bar{C}\|_F\|\hat{\beta}(0,\tilde{\lambda})-\hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}.\end{aligned}$$

Therefore

$$\|\hat{\beta}(0,\tilde{\lambda})-\hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}\leq\frac{2\sqrt{2}\sqrt{(m-1)q+p-1}\hat{\lambda}\|\bar{C}\|_F}{\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X})}.\tag{15}$$

On the other hand, denote $\tilde{\lambda}\|\beta\|_{*1}:=\lambda_1\|B\|_*+\lambda_3\|\gamma\|_1$ and

$$L(\beta)=\frac{1}{2}\sum_{i=1}^{n}(y_i-\langle B,X_i\rangle-\langle\gamma,z_i\rangle)^2+\lambda_1\|B\|_*+\lambda_3\|\gamma\|_1=\frac{1}{2}\sum_{i=1}^{n}\left(y_i-(X_i^{\mathrm{T}},z_i^{\mathrm{T}})\beta\right)^2+\tilde{\lambda}\|\beta\|_{*1}.$$

The $L(\beta)$ is a convex function. It holds that

$$L(\hat{\beta}(0,\tilde{\lambda}))-L(\beta^*)\geq\langle D^*,\hat{\beta}(0,\tilde{\lambda})-\beta^*\rangle,\tag{16}$$

where

$$D^*\in\partial L(\beta^*)=-\sum_{i=1}^{n}\begin{pmatrix}X_i\\z_i\end{pmatrix}\left(y_i-(X_i^{\mathrm{T}},z_i^{\mathrm{T}})\beta^*\right)+\tilde{\lambda}\partial\|\beta^*\|_{*1}.$$

Combining (14) and (16), we have

$$\begin{aligned}&(vec(\beta^*)-vec(\hat{\beta}(0,\tilde{\lambda})))^{\mathrm{T}}(\frac{1}{2}\bar{X}^{\mathrm{T}}\bar{X})(vec(\beta^*)-vec(\hat{\beta}(0,\tilde{\lambda})))\\&\leq L(\beta^*)-L(\hat{\beta}(0,\tilde{\lambda}))\leq\langle D^*,\beta^*-\hat{\beta}(0,\tilde{\lambda})\rangle.\end{aligned}$$

For $D\in\partial\|\beta^*\|_{*1}$, let

$$D^*=-\sum_{i=1}^{n}\begin{pmatrix}X_i\\z_i\end{pmatrix}\left(y_i-(X_i^{\mathrm{T}},z_i^{\mathrm{T}})\beta^*\right)+\tilde{\lambda}D,$$

where

$$D=\left\{\begin{pmatrix}W^*\\\xi^*\end{pmatrix}\mid W^*\in\partial\|B^*\|_*,\xi^*\in\partial\|\gamma^*\|_1\right\}.$$

It holds that

$$\begin{aligned}&\frac{1}{2}\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X})\|vec(\beta^*)-vec(\hat{\beta}(0,\tilde{\lambda}))\|_2^2\\&\leq\|(vec(\beta^*)-vec(\hat{\beta}(0,\tilde{\lambda})))^{\mathrm{T}}(\frac{1}{2}\bar{X}^{\mathrm{T}}\bar{X})(vec(\beta^*)-vec(\hat{\beta}(0,\tilde{\lambda})))\|_2\\&\leq\|D^*\|_{F2}\|vec(\beta^*)-vec(\hat{\beta}(0,\tilde{\lambda}))\|_2,\end{aligned}$$

so we have

$$\begin{aligned}&E(\|\beta^*-\hat{\beta}(0,\tilde{\lambda})\|_{F2}^2)\\&\leq 4(\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}(2E(\|\bar{X}^{\mathrm{T}}(y-\bar{X}\beta^*)\|_2^2)+2\lambda^2\|D\|_{F2}^2)\\&=8(\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}(E\|\bar{X}^{\mathrm{T}}\varepsilon\|_2^2+\lambda_1^2\|W^*\|_F^2+\lambda_3^2\|\xi^*\|_2^2)\\&=8(\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}(E(\varepsilon^{\mathrm{T}}\bar{X}\bar{X}^{\mathrm{T}}\varepsilon)+\lambda_1^2\|W^*\|_F^2+\lambda_3^2\|\xi^*\|_2^2)\\&\leq 8(\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}((mq+p)\lambda_{max}(\bar{X}^{\mathrm{T}}\bar{X})\sigma^2+\lambda_1^2\|W^*\|_F^2+\lambda_3^2\|\xi^*\|_2^2)\\&\leq 8(\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}((mq+p)\lambda_{max}(\bar{X}^{\mathrm{T}}\bar{X})\sigma^2+\lambda_1^2\|W^*\|_F^2+\lambda_3^2).\end{aligned}\tag{17}$$

Combing (15) and (17), we conclude that

$$\begin{aligned}&E(\|\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\beta^*\|_{F2}^2)\\&\leq 2E(\|\beta^*-\hat{\beta}(0,\tilde{\lambda})\|_{F2}^2)+2E(\|\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda})\|_{F2}^2)\\&\leq\frac{16(mq+p)\lambda_{max}(\bar{X}^{\mathrm{T}}\bar{X})\sigma^2+\lambda_1^2\|W^*\|_F^2+\lambda_3^2}{\lambda_{min}^2(\bar{X}^{\mathrm{T}}\bar{X})}+\frac{16((m-1)q+p-1)\hat{\lambda}^2\|\bar{C}\|_F^2}{\lambda_{min}^2(\bar{X}^{\mathrm{T}}\bar{X})},\end{aligned}$$

therefore we have

$$\begin{aligned}&E(\|\hat{B}(\bar{\lambda},\tilde{\lambda})-B^*\|_F^2)+E(\|\hat{\gamma}(\bar{\lambda},\tilde{\lambda})-\gamma^*\|_2^2)\\&\leq 16\frac{\hat{\lambda}^2((m-1)q+p-1)\|\bar{C}\|_F^2+(mq+p)\overline{\sigma}n\sigma^2+\lambda_1^2\|W^*\|_F^2+\lambda_3^2}{(\underline{\sigma}n)^2}.\end{aligned}$$

□

In order to prove Theorem 5, we give the following Lemmas.

Lemma 3. *There is a strictly convex function G with* $G(u) = \frac{u^2}{2mq}$ *such that for all* B, B' *we have*

$$R(B) - R(B') \geq \langle \nabla R(B'), B - B' \rangle + G(\|B - B'\|_F). \tag{18}$$

Proof of Lemma 3.

$$R(B) = ER_n(B) = \frac{1}{2n}\sum_{i=1}^{n} E((y_i - \langle X_i, B\rangle)^2) = \frac{1}{2n}\sum_{i=1}^{n} E_{X_i}[E((y_i - \langle X_i, B\rangle)^2)|X_i].$$

Suppose that X_i has its only 1 at entry (k, j), F is the distribution function of noise. Then $\langle X_i, B\rangle = B_{kj}$. Define

$$\begin{aligned}
r(x,B): &= E((y_i - \langle X_i, B\rangle)^2 | X_i = x) = E((y_i - B_{kj})^2), \\
\nabla_{B_{kj}} r(x,B) &= -2E(y_i - B_{kj}) = -2\int_{-\infty}^{+\infty} (\bar{y} - B_{kj}) dF(\bar{y}) \\
&= -2\int_{-\infty}^{+\infty} \bar{y} dF(\bar{y}) + 2\int_{-\infty}^{+\infty} B_{kj} dF(\bar{y}) \\
&= -2\int_{-\infty}^{+\infty} \bar{y} dF(\bar{y}) + 2B_{kj}, \\
\nabla^2_{B_{kj}} r(x,B) &= 2.
\end{aligned}$$

The Taylor expansion around B' is given by

$$r(x,B) = r(x,B') + \langle \nabla r(x,B'), B_{kj} - B'_{kj} \rangle + \frac{\nabla^2 r(x,\tilde{B})}{2}(B_{kj} - B'_{kj})^2,$$

where $\tilde{B}$ is an intermediate point. We can see that inequality (18) holds with $G(u) = \frac{u^2}{2mq}$. □

Lemma 4. *For all* $B \in R^{m\times q}$, *we have*

$$\langle -\nabla R_n(\hat{B}), B - \hat{B} \rangle \leq \lambda_1 \|B\|_* + \lambda_2 \|Cvec(B)\|_1 - (\lambda_1 \|\hat{B}\|_* + \lambda_2 \|Cvec(\hat{B})\|_1).$$

Proof of Lemma 4. Define for $0 < t < 1$ $B_t = (1-t)\hat{B} + tB$. Since $\hat{B}$ is the minimizer of the objective function, i.e., $R_n(\hat{B}) + \lambda_1\|\hat{B}\|_* + \lambda_2\|Cvec(\hat{B})\|_1 \leq R_n(B) + \lambda_1\|B\|_* + \lambda_2\|Cvec(B)\|_1$. By the convexity of $\lambda_1\|B\|_* + \lambda_2\|Cvec(B)\|_1$ we have

$$\begin{aligned}
&R_n(\hat{B}) + \lambda_1\|\hat{B}\|_* + \lambda_2\|Cvec(\hat{B})\|_1 \leq R_n(B_t) + \lambda_1\|B_t\|_* + \lambda_2\|Cvec(B_t)\|_1 \\
&\leq R_n(B_t) + t(\lambda_1\|B\|_* + \lambda_2\|Cvec(B)\|_1) + (1-t)(\lambda_1\|\hat{B}\|_* + \lambda_2\|Cvec(\hat{B})\|_1).
\end{aligned}$$

We conclude that $\frac{R_n(\hat{B}) - R_n(B_t)}{t} \leq \lambda_1\|B\|_* + \lambda_2\|Cvec(B)\|_1 - (\lambda_1\|\hat{B}\|_* + \lambda_2\|Cvec(\hat{B})\|_1)$. Letting $t \to 0$ the proof is completed. □

Proof of Theorem 5. The first order Taylor expansion of R at $\hat{B}$ is given by

$$R(B) = R(\hat{B}) + \langle \nabla R(\hat{B}), B - \hat{B} \rangle + Rem(\hat{B}, B).$$

Case 1
If

$$\langle \nabla R(\hat{B}), B-\hat{B}\rangle \geq \delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1-2\lambda_1\|B^{-}\|_{*} \\ -\lambda_{*}-\lambda_2\|Cvec(B)\|_1,$$

then we find that

$$\begin{aligned} R(B)-R(\hat{B}) &\geq \langle \nabla R(\hat{B}), B-\hat{B}\rangle + G(\|B-\hat{B}\|_F) \\ &\geq \delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1 \\ &-2\lambda_1\|B^{-}\|_{*}-\lambda_{*}-\lambda_2\|Cvec(B)\|_1+G(\|B-\hat{B}\|_F). \end{aligned}$$

Because of $0 \leq G(\|B-B'\|_F)$, we imply that

$$\begin{aligned} &\delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1+R(\hat{B})-R(B) \\ &\leq 2\lambda_1\|B^{-}\|_{*}+\lambda_{*}+\lambda_2\|Cvec(B)\|_1. \end{aligned}$$

Case 2
If

$$\langle \nabla R(\hat{B}), B-\hat{B}\rangle \leq \delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1-2\lambda_1\|B^{-}\|_{*} \\ -\lambda_{*}-\lambda_2\|Cvec(B)\|_1.$$

By Lemma 4, for all $B \in R^{m\times q}$, we have

$$\langle -\nabla R_n(\hat{B}), B-\hat{B}\rangle \leq \lambda_1\|B\|_{*}+\lambda_2\|Cvec(B)\|_1-(\lambda_1\|\hat{B}\|_{*}+\lambda_2\|Cvec(\hat{B})\|_1),$$

which implies that

$$0 \leq \langle \nabla R_n(\hat{B}), B-\hat{B}\rangle + \lambda_1\|B\|_{*}+\lambda_2\|Cvec(B)\|_1-(\lambda_1\|\hat{B}\|_{*}+\lambda_2\|Cvec(\hat{B})\|_1).$$

Hence,

$$\begin{aligned} &\langle \nabla R_n(\hat{B})-\nabla R(\hat{B}), B-\hat{B}\rangle + \lambda_1\|B\|_{*}+\lambda_2\|Cvec(B)\|_1-(\lambda_1\|\hat{B}\|_{*}+\lambda_2\|Cvec(\hat{B})\|_1) \\ &+\delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1 \geq 2\lambda_1\|B^{-}\|_{*}+\lambda_{*}+\lambda_2\|Cvec(B)\|_1. \end{aligned}$$

According to the condition in Theorem 5 and $\|B\|_{*}-\|B'\|_{*} \leq \Omega^{+}(B'-B)-\Omega^{-}(B'-B)+2\|B^{-}\|_{*}$, we have

$$\begin{aligned} & \langle -\nabla R(\hat{B}), B-\hat{B}\rangle + \delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B) \\ \leq\ & \langle \nabla R_n(\hat{B})-\nabla R(\hat{B}), B-\hat{B}\rangle + \lambda_1\|B\|_{*}+\lambda_2\|Cvec(B)\|_1-(\lambda_1\|\hat{B}\|_{*}+\lambda_2\|Cvec(\hat{B})\|_1) \\ & +\delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B) \\ \leq\ & \lambda_{\varepsilon}\underline{\Omega}(\hat{B}-B)+\lambda_{*}+\lambda_1\|B\|_{*}+\lambda_2\|Cvec(B)\|_1-(\lambda_1\|\hat{B}\|_{*}+\lambda_2\|Cvec(\hat{B})\|_1) \\ & +\delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B) \\ \leq\ & (\lambda_{\varepsilon}+\delta\underline{\lambda})\Omega^{+}(\hat{B}-B)+(\lambda_{\varepsilon}+\delta\underline{\lambda})\Omega^{-}(\hat{B}-B)+\lambda_{*}+\lambda_1\Omega^{+}(\hat{B}-B)-\lambda_1\Omega^{-}(\hat{B}-B) \\ & +2\lambda_1\|B^{-}\|_{*}+\lambda_2\|Cvec(B)\|_1-\lambda_2\|Cvec(\hat{B})\|_1 \\ \leq\ & \bar{\lambda}\Omega^{+}(\hat{B}-B)-(1-\delta)\underline{\lambda}\Omega^{-}(\hat{B}-B)+\lambda_{*}+2\lambda_1\|B^{-}\|_{*}+\lambda_2\|Cvec(B)\|_1-\lambda_2\|Cvec(\hat{B})\|_1. \end{aligned}$$

Therefore, $\bar{\lambda}\Omega^+(\hat{B}-B)-(1-\delta)\underline{\lambda}\Omega^-(\hat{B}-B)\geq 0$, $\Omega^-(\hat{B}-B)\leq \frac{\bar{\lambda}}{(1-\delta)\underline{\lambda}}\Omega^+(\hat{B}-B)$. Let H be the convex conjugate of G. Then we have the convex conjugate inequality

$$\Omega^+(\hat{B}-B)\leq \|\hat{B}-B\|_F 3\sqrt{s}\leq H(3\sqrt{s})+G(\|\hat{B}-B\|_F),$$

which implies that

$$\begin{aligned}
& \langle -\nabla R(\hat{B}), B-\hat{B}\rangle + \delta\underline{\lambda}\Omega^+(\hat{B}-B)+\delta\underline{\lambda}\Omega^-(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1 \\
= \;& R(\hat{B})-R(B)+Rem(\hat{B},B)+\delta\underline{\lambda}\Omega^+(\hat{B}-B)+\delta\underline{\lambda}\Omega^-(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1 \\
\leq \;& H(\bar{\lambda}3\sqrt{s})+G(\|\hat{B}-B\|_F)+\lambda_*+2\lambda_1\|B^-\|_*+\lambda_2\|Cvec(B)\|_1 \\
\leq \;& H(\bar{\lambda}3\sqrt{s})+Rem(\hat{B},B)+\lambda_*+2\lambda_1\|B^-\|_*+\lambda_2\|Cvec(B)\|_1.
\end{aligned}$$

So

$$\begin{aligned}
& R(\hat{B})-R(B)+\delta\underline{\lambda}\Omega^+(\hat{B}-B)+\delta\underline{\lambda}\Omega^-(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1 \\
& \leq H(\bar{\lambda}3\sqrt{s})+\lambda_*+2\lambda_1\|B^-\|_*+\lambda_2\|Cvec(B)\|_1.
\end{aligned}$$

□

The proof relies on the property of nuclear norm, Concentration [1], Symmetrization [14] and Contraction Theorems [9].

Proof of Theorem 6. Define for all $M>0$

$$Z_M := \sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \nabla R_n(B')-\nabla R(B'), B-B'\rangle|.$$

$$\rho(B)=(y_i-\langle X_i,B\rangle)^2, \bar{\rho}(B)=y_i-\langle X_i,B\rangle, R_n(B)=\frac{1}{2n}\sum_{i=1}^n \rho(B),$$

$$\nabla R_n(B)=\frac{1}{n}\sum_{i=1}^n (y_i-\langle X_i,B\rangle)(-X_i)=\frac{1}{n}\sum_{i=1}^n \bar{\rho}(B)(-X_i).$$

We have

$$\begin{aligned}
EZ_M: \;=\; & E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \nabla R_n(B')-\nabla R(B'), B-B'\rangle| \\
=\; & E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \frac{1}{n}\sum_{i=1}^n \bar{\rho}(B')(-X_i)-E(\frac{1}{n}\sum_{i=1}^n \bar{\rho}(B')(-X_i)), B-B'\rangle| \\
\leq\; & 2E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i\bar{\rho}(B')(-X_i), B-B'\rangle| \\
\leq\; & 4EL\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle -\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i, B-B'\rangle| \\
\leq\; & 4L\underline{\Omega}(B-B')E\underline{\Omega}_*(\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i) \\
\leq\; & 4MLE\underline{\Omega}_*(\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i).
\end{aligned}$$

The first inequality follows from Symmetrization of Expectations in [14], the second inequality form Contraction Theorem in [9], and the third inequality follows form the dual norm inequality. $\bar{\rho}(B)$ is Lipschitz continuous with Lipschitz constant L. The $\tilde{\varepsilon}_i$ is i.i.d. Randemacher random variables independent of X_i. The 2-Orlicz norm of a Rademacher random variable is equal to $\|\tilde{\varepsilon}\|_{\psi_2} = \sqrt{\frac{1}{log2}}$. $E(\tilde{\varepsilon}_i X_i) = E\tilde{\varepsilon}_i EX_i = 0, E(\tilde{\varepsilon}_i X_i \tilde{\varepsilon}_i X_i^T) = E\tilde{\varepsilon}_i^2 EX_i X_i^T = EX_i X_i^T, E(\tilde{\varepsilon}_i X_i^T \tilde{\varepsilon}_i X_i) = E\tilde{\varepsilon}_i^2 EX_i^T X_i = EX_i^T X_i$.

Assume that $\lambda_{max}(X_i)$ and $\lambda_{max}(EX_i X_i^T), \lambda_{max}(EX_i^T X_i)$ is bounded. Denote $\lambda_{max}(X_i) \leq b_1$, $\max\left\{\lambda_{max}(EX_i X_i^T), \lambda_{max}(EX_i^T X_i)\right\} := b_2$. $\lambda_{max}(\tilde{\varepsilon}_i X_i) = |\tilde{\varepsilon}_i|\lambda_{max}(X_i) \leq b_1|\tilde{\varepsilon}_i|$. $\|\tilde{\varepsilon}_i X_i\|_{\psi_2} \leq b_1^2\|\tilde{\varepsilon}\|_{\psi_2} = b_1^2\sqrt{\frac{1}{log2}}$.

$$S^2 = \max\left\{\frac{\lambda_{max}(\sum_{i=1}^n E\tilde{\varepsilon}_i^2 X_i X_i^T)}{n}, \frac{\lambda_{max}(\sum_{i=1}^n E\tilde{\varepsilon}_i^2 X_i^T X_i)}{n}\right\} = b_2.$$

According to [5, Theorem 2.1], for a constant C and all $t > 0$ we have

$$P\left(\frac{\lambda_{max}(\sum_{i=1}^n \tilde{\varepsilon}_i X_i)}{n} \geq CS\sqrt{\frac{t+log(m+q)}{n}} + Clog^{\frac{1}{2}}(\frac{K}{S})\frac{t+log(m+q)}{n}\right) \leq e^{-t},$$

where $S = \sqrt{b_2}, K = b_1^2\sqrt{\frac{1}{log2}}$. Then

$$E\left(\frac{\lambda_{max}(\sum_{i=1}^n \tilde{\varepsilon}_i X_i)}{n}\right) \leq C\left(\sqrt{b_2}\sqrt{\frac{log(m+q)}{n}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)})\frac{log(m+q)}{n}\right),$$

so

$$E\underline{\Omega}_*(\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i) \leq C\left(\sqrt{b_2}\sqrt{\frac{log(m+q)}{n}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)})\frac{log(m+q)}{n}\right).$$

Define $f(X_i^T B') = \langle \nabla\rho(B'), B - B'\rangle = \langle (-X_i)(y_i - \langle X_i, B'\rangle), B - B'\rangle = \langle (-X_i)\bar{\rho}(B'), B - B'\rangle$, we know $Ef(B') = 0$.

$$\begin{aligned}
& \sup_{B':\underline{\Omega}(B'-B)\leq M} Var(f(X_1^T B')) \\
= & \sup_{B':\underline{\Omega}(B'-B)\leq M} Var\left(\sum_{l=1}^m\sum_{j=1}^q X_{1_{lj}}(B_{lj} - B'_{lj})\bar{\rho}(B')\right) \\
\leq & \sup_{B':\underline{\Omega}(B'-B)\leq M} \bar{L}^2 E\left[\sum_{l=1}^m\sum_{j=1}^q (B_{lj} - B'_{lj})^2\right] \\
\leq & \frac{\bar{L}^2 M^2}{mq},
\end{aligned}$$

where $\|X_i\bar{\rho}(B')\| \leq \bar{L}$. Therefore, assume that $\|B - B'\|_\infty \leq 2\vartheta$, $\|f(X_i^T B')\|_\infty = \|\langle (-X_i)(y_i - \langle X_i, B'\rangle), B - B'\rangle\|_\infty \leq 2\vartheta\bar{L}$. According to Bousequet's Concentration Theorem in [1] we obtain that for all $t > 0$

$$P\left(Z_M \geq 8MLC\left(\sqrt{b_2}\sqrt{\frac{log(m+q)}{n}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)})\frac{log(m+q)}{n}\right) + \frac{\bar{L}M}{\sqrt{mq}}\sqrt{\frac{2t}{n}} + \frac{8t\vartheta\bar{L}}{3n}\right) \leq e^{-t}.$$

Replacing t by $mlog(m+q)$ we obtain

$$P\left(Z_M \geq 8MLC\left(\sqrt{b_2}\sqrt{\frac{log(m+q)}{n}}+(\sqrt{2logb_1}+\sqrt{log(1+b_2)})\frac{log(m+q)}{n}\right)\right.$$
$$\left.+\frac{\bar{L}M}{\sqrt{mq}}\sqrt{\frac{2mlog(m+q)}{n}}+\frac{8mlog(m+q)\vartheta\bar{L}}{3n}\right)\leq e^{-mlog(m+q)}.$$

There exist constant $C_0=8LC+\sqrt{2}\bar{L}, C_1=8\vartheta\bar{L}/3$ and

$$\lambda_\varepsilon = C_0\left(\sqrt{b_2}\sqrt{\frac{log(m+q)}{n}}+\sqrt{\frac{log(m+q)}{nq}}+(\sqrt{2logb_1}+\sqrt{log(1+b_2)})\frac{log(m+q)}{n}\right)$$
$$\lambda_* = \frac{C_1 mlog(m+q)}{n}.$$

We can imply that $P(Z_M \geq M\lambda_\varepsilon+\lambda_*)\leq e^{-mlog(m+q)}$. According to Theorem 5 and [5, Lemma B.5], we have probability at least $1-(j_0+2)e^{-mlog(m+q)}$ such that

$$R(\hat{B})-R(B)+\delta\underline{\lambda}\Omega^+(\hat{B}-B)+\delta\underline{\lambda}\Omega^-(\hat{B}-B)+\lambda_2\|Cvec(\hat{B})\|_1$$
$$\leq H(\bar{\lambda}3\sqrt{s})+\lambda_*+2\lambda_1\|B^-\|_*+\lambda_2\|Cvec(B)\|_1.$$

Assume that $q=o(\frac{n}{log(m+q)})$, we can obtain that $\lambda_\varepsilon \asymp \sqrt{\frac{log(m+q)}{nq}}$. We further assume that $\lambda_2 \asymp \sqrt{\frac{log(m+q)}{nq}}$, then we have

$$R(\hat{B})-R(B)\leq R(B)-R(B)+\mathbb{O}_P\left(\frac{mslog(m+q)}{n}+\sqrt{\frac{log(m+q)}{nq}}(\|B^-\|_*+\|Cvec(B)\|_1)\right). \quad (19)$$

Let $B=B^*$ in (19), the convergence rate for the optimal solution $\hat{B}$ is given by

$$\|\hat{B}-B^*\|_F^2=\mathbb{O}_P\left(\frac{2m^2qslog(m+q)}{n}\right).$$

□

Proof of Theorem 7. Define

$$\hat{\beta}(0,\tilde{\lambda}) = \begin{pmatrix}\hat{B}(0,\tilde{\lambda})\\ \hat{\gamma}(0,\tilde{\lambda})\end{pmatrix}$$
$$= \underset{B,\gamma}{\arg\min}\left\{\sum_{i=1}^n \log(1+e^{\langle\mathbb{X}_i,B\rangle+\langle\mathbb{Z}_i,\gamma\rangle})-y_i(\langle\mathbb{X}_i,B\rangle+\langle\mathbb{Z}_i,\gamma\rangle)+\lambda_1\|B\|_*+\lambda_3\|\gamma\|_1\right\}.$$

By the definition $\hat{\beta}(0,\tilde{\lambda})$ and $\hat{\beta}(\bar{\lambda},\tilde{\lambda})$, we know

$$\sum_{i=1}^n \log(1+e^{\langle\mathbb{X}_i,\hat{B}(0,\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(0,\tilde{\lambda})\rangle})-y_i(\langle\mathbb{X}_i,\hat{B}(0,\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(0,\tilde{\lambda})\rangle)+\lambda_1\|\hat{B}(0,\tilde{\lambda})\|_*+\lambda_3\|\hat{\gamma}(0,\tilde{\lambda})\|_1$$
$$\leq \sum_{i=1}^n \log(1+e^{\langle\mathbb{X}_i,\hat{B}(\bar{\lambda},\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\rangle})-y_i(\langle\mathbb{X}_i,\hat{B}(\bar{\lambda},\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\rangle)+\lambda_1\|\hat{B}(\bar{\lambda},\tilde{\lambda})\|_*+\lambda_3\|\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1$$

and

$$\begin{aligned}
&\sum_{i=1}^{n}\log(1+e^{\langle\mathbb{X}_i,\hat{B}(\bar{\lambda},\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\rangle})-y_i(\langle\mathbb{X}_i,\hat{B}(\bar{\lambda},\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\rangle)+\lambda_1\|\hat{B}(\bar{\lambda},\tilde{\lambda})\|_*\\
&+\lambda_2\|Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1+\lambda_3\|\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1+\lambda_4\|A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1\\
\le&\sum_{i=1}^{n}\log(1+e^{\langle\mathbb{X}_i,\hat{B}(0,\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(0,\tilde{\lambda})\rangle})-y_i(\langle\mathbb{X}_i,\hat{B}(0,\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(0,\tilde{\lambda})\rangle)+\lambda_1\|\hat{B}(0,\tilde{\lambda})\|_*\\
&+\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))\|_1+\lambda_3\|\hat{\gamma}(0,\tilde{\lambda})\|_1+\lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})\|_1.
\end{aligned}$$

According to two inequalities, we have

$$\begin{aligned}
&\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))\|_1-\lambda_2\|Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1+\lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})\|_1-\lambda_4\|A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1\\
\ge&\sum_{i=1}^{n}\log(1+e^{\langle\mathbb{X}_i,\hat{B}(\bar{\lambda},\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\rangle})-y_i(\langle\mathbb{X}_i,\hat{B}(\bar{\lambda},\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\rangle)\\
&+\lambda_1\|\hat{B}(\bar{\lambda},\tilde{\lambda})\|_*+\lambda_3\|\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1-\sum_{i=1}^{n}\log(1+e^{\langle\mathbb{X}_i,\hat{B}(0,\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(0,\tilde{\lambda})\rangle})\\
&-y_i(\langle\mathbb{X}_i,\hat{B}(0,\tilde{\lambda})\rangle+\langle\mathbb{Z}_i,\hat{\gamma}(0,\tilde{\lambda})\rangle)-\lambda_1\|\hat{B}(0,\tilde{\lambda})\|_*-\lambda_3\|\hat{\gamma}(0,\tilde{\lambda})\|_1. \qquad (20)
\end{aligned}$$

Denote

$$P(\beta)=\lambda_1\|B\|_*+\lambda_3\|\gamma\|_1,\ R(\beta)=\sum_{i=1}^{n}\log(1+e^{\langle\mathbb{X}_i,B\rangle+\langle\mathbb{Z}_i,\gamma\rangle})-y_i(\langle\mathbb{X}_i,B\rangle+\langle\mathbb{Z}_i,\gamma\rangle).$$

For the optimization problem $\min_{\beta}\{R(\beta)+P(\beta)\}$, the loss function $R(\beta)$ is differentiable. Following the proof of Lemma 3.1 in [5], we imply that for all β $\langle-\nabla R(\hat{\beta}(0,\tilde{\lambda})),\beta-\hat{\beta}(0,\tilde{\lambda})\rangle\le P(\beta)-P(\hat{\beta}(0,\tilde{\lambda}))$, so we have

$$\begin{aligned}
&R(\hat{\beta}(\bar{\lambda},\tilde{\lambda}))+P(\hat{\beta}(\bar{\lambda},\tilde{\lambda}))-[R(\hat{\beta}(0,\tilde{\lambda}))+P(\hat{\beta}(0,\tilde{\lambda}))] \qquad (21)\\
=&R(\hat{\beta}(\bar{\lambda},\tilde{\lambda}))-R(\hat{\beta}(0,\tilde{\lambda}))+P(\hat{\beta}(\bar{\lambda},\tilde{\lambda}))-R(\hat{\beta}(0,\tilde{\lambda}))\\
\ge&R(\hat{\beta}(\bar{\lambda},\tilde{\lambda}))-R(\hat{\beta}(0,\tilde{\lambda}))-\langle\nabla R(\hat{\beta}(0,\tilde{\lambda})),\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda})\rangle\\
\ge&\langle\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda}),\frac{1}{2}\nabla^2R(\hat{\beta}(0,\tilde{\lambda}))(\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda}))\rangle.
\end{aligned}$$

According to (20) and (21), we have

$$\begin{aligned}
&\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))\|_1-\lambda_2\|Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1+\lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})\|_1-\lambda_4\|A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1\\
\ge&\langle\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda}),\frac{1}{2}\nabla^2R(\hat{\beta}(0,\tilde{\lambda}))(\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda}))\rangle.
\end{aligned}$$

We know

$$\nabla R(\beta)=\sum_{i=1}^{n}\bar{X}_i^T(\frac{e^{\langle\bar{X}_i,vec(\beta)\rangle}}{1+e^{\langle\bar{X}_i,vec(\beta)\rangle}}-y_i),\ \nabla^2R(\beta)=\sum_{i=1}^{n}\bar{X}_i^T\bar{X}_i\frac{e^{\langle\bar{X}_i,vec(\beta)\rangle}}{(1+e^{\langle\bar{X}_i,vec(\beta)\rangle})^2}.$$

Assume the second-order derivative of $R(\hat{\beta}(0,\tilde{\lambda}))$ has a low bound $L\lambda_{\min}(\bar{X}^T\bar{X})I$, i.e.,

$$\nabla^2 R(\hat{\beta}(0,\tilde{\lambda})) = \bar{X}_i^T\bar{X}_i \frac{e^{\langle \bar{X}_i, vec(\hat{\beta}(0,\tilde{\lambda}))\rangle}}{(1+e^{\langle \bar{X}_i, vec(\hat{\beta}(0,\tilde{\lambda}))\rangle})^2} \succeq L\sum_{i=1}^{n}\bar{X}_i^T\bar{X}_i = L\bar{X}^T\bar{X} \succeq L\lambda_{\min}(\bar{X}^T\bar{X})I.$$

The following formula holds, i.e.,

$$\begin{aligned}&\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))\|_1 - \lambda_2\|Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1 + \lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})\|_1 - \lambda_4\|A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1 \\ &\geq L\lambda_{\min}(\bar{X}^T\bar{X})\|\hat{\beta}(\bar{\lambda},\tilde{\lambda}) - \hat{\beta}(0,\tilde{\lambda})\|_{F_2}^2.\end{aligned} \tag{22}$$

On the one hand, denote $\|\beta\|_{F2}^2 := \|B\|_F^2 + \|\gamma\|_2^2$. It holds that

$$\begin{aligned}&\lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda}))\|_1 - \lambda_2\|Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1 + \lambda_4\|A\hat{\gamma}(0,\tilde{\lambda})\|_1 - \lambda_4\|A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1 \\ &\leq \lambda_2\|Cvec(\hat{B}(0,\tilde{\lambda})) - Cvec(\hat{B}(\bar{\lambda},\tilde{\lambda}))\|_1 + \lambda_4\|A\hat{\gamma}(0,\tilde{\lambda}) - A\hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_1 \\ &\leq \lambda_2\|C\|_F\sqrt{(m-1)q}\|vec(\hat{B}(0,\tilde{\lambda}) - vec(\hat{B}(\bar{\lambda},\tilde{\lambda})\|_2 + \lambda_4\|A\|_F\sqrt{p-1}\|\hat{\gamma}(0,\tilde{\lambda}) - \hat{\gamma}(\bar{\lambda},\tilde{\lambda})\|_2 \\ &\leq \max\{\lambda_2,\lambda_4\}\sqrt{2}\sqrt{(m-1)q+p-1}\|\bar{C}\|_F\|\hat{\beta}(0,\tilde{\lambda}) - \hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}.\end{aligned} \tag{23}$$

By (22) and (23) we have

$$L\lambda_{\min}(\bar{X}^{\mathrm{T}}\bar{X})\|\hat{\beta}(0,\tilde{\lambda}) - \hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}^2 \leq \max\{\lambda_2,\lambda_4\}\sqrt{(m-1)q+p-1}\|\bar{C}\|_F\|\hat{\beta}(0,\tilde{\lambda}) - \hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2}.$$

We can obtain that

$$\|\hat{\beta}(0,\tilde{\lambda}) - \hat{\beta}(\bar{\lambda},\tilde{\lambda})\|_{F2} \leq \frac{\sqrt{2}\sqrt{(m-1)q+p-1}\hat{\lambda}\|\bar{C}\|_F}{L\lambda_{\min}(\bar{X}^{\mathrm{T}}\bar{X})}. \tag{24}$$

On the other hand, denote $\tilde{\lambda}\|\beta\|_{*1} := \lambda_1\|B\|_* + \lambda_3\|\gamma\|_1$ and $L(\beta) = R(\beta) + P(\beta) = R(\beta) + \tilde{\lambda}\|\beta\|_{*1}$. The $L(\beta)$ is a convex function. It holds that $L(\hat{\beta}(0,\tilde{\lambda})) - L(\beta^*) \geq \langle D^*, \hat{\beta}(0,\tilde{\lambda}) - \beta^*\rangle$, where $D^* \in \partial L(\beta^*) = \nabla R(\beta^*) + \tilde{\lambda}\partial\|\beta^*\|_{*1}$. We have

$$L\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X})\|\hat{\beta}(0,\tilde{\lambda}) - \hat{\beta^*}\|_{F2}^2 \leq L(\beta^*) - L(\hat{\beta}(0,\tilde{\lambda})) \leq \langle D^*, \beta^* - \hat{\beta}(0,\tilde{\lambda})\rangle.$$

For $D \in \partial\|\beta^*\|_{*1}$, let $D^* = \nabla R(\beta^*) + \tilde{\lambda}D$, where $D = \left\{ \begin{pmatrix} W^* \\ \xi^* \end{pmatrix} \mid W^* \in \partial\|B^*\|_*, \xi^* \in \partial\|\gamma^*\|_1 \right\}$. It holds that

$$L\lambda_{min}(\bar{X}^{\mathrm{T}}\bar{X})\|vec(\beta^*) - vec(\hat{\beta}(0,\tilde{\lambda}))\|_2^2 \leq \|D^*\|_{F2}\|vec(\beta^*) - vec(\hat{\beta}(0,\tilde{\lambda}))\|_2.$$

Hence, we obtain that

$$\begin{aligned}&E(\|\beta^* - \hat{\beta}(0,\tilde{\lambda})\|_{F2}^2) \\ &\leq (L\lambda_{\min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}(2E(\|\nabla R(\beta^*)\|_2^2) + 2\lambda^2\|D\|_{F2}^2) \\ &\leq 2(L\lambda_{\min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}((mq+p)\lambda_{\max}(\bar{X}^{\mathrm{T}}\bar{X}) + \lambda_1^2\|W^*\|_F^2 + \lambda_3^2\|\xi^*\|_2^2) \\ &\leq 2(L\lambda_{\min}(\bar{X}^{\mathrm{T}}\bar{X}))^{-2}((mq+p)\lambda_{\max}(\bar{X}^{\mathrm{T}}\bar{X}) + \lambda_1^2\|W^*\|_F^2 + \lambda_3^2).\end{aligned} \tag{25}$$

Combing (24) and (25), we conclude that

$$\begin{aligned}
&E(\|\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\beta^*\|_{F2}^2)\\
&\leq 2E(\|\beta^*-\hat{\beta}(0,\tilde{\lambda})\|_{F2}^2)+2E(\|\hat{\beta}(\bar{\lambda},\tilde{\lambda})-\hat{\beta}(0,\tilde{\lambda})\|_{F2}^2)\\
&\leq \frac{4(mq+p)\lambda_{max}(\bar{X}^{\mathrm{T}}\bar{X})\sigma^2+\lambda_1^2\|W^*\|_F^2+\lambda_3^2}{L^2\lambda_{\min}^2(\bar{X}^{\mathrm{T}}\bar{X})}+\frac{4((m-1)q+p-1)\hat{\lambda}^2\|\bar{C}\|_F^2}{L^2\lambda_{\min}^2(\bar{X}^{\mathrm{T}}\bar{X})},
\end{aligned}$$

therefore we have

$$\begin{aligned}
&E(\|\hat{B}(\bar{\lambda},\tilde{\lambda})-B^*\|_F^2)+E(\|\hat{\gamma}(\bar{\lambda},\tilde{\lambda})-\gamma^*\|_2^2)\\
&\leq 4\frac{\hat{\lambda}^2((m-1)q+p-1)\|\bar{C}\|_F^2+(mq+p)\overline{\sigma}n+\lambda_1^2\|W^*\|_F^2+\lambda_3^2}{L^2(\underline{\sigma}n)^2}.
\end{aligned}$$

□

Lemma 5. *There is a strictly convex function G with $G(u)=\frac{Lu^2}{2mq}$ such that for all B,B' we have*

$$R(B)-R(B')\geq\langle\nabla R(B'),B-B'\rangle+G(\|B-B'\|_F). \tag{26}$$

Proof of Lemma 5.

$$\begin{aligned}
R(B)=ER_n(B)&=\frac{1}{n}\sum_{i=1}^{n}E(\log(1+e^{\langle X_i,B\rangle})-y_i\langle X_i,B\rangle)\\
&=\frac{1}{n}\sum_{i=1}^{n}E_{X_i}[E(\log(1+e^{\langle X_i,B\rangle})-y_i\langle X_i,B\rangle)|X_i].
\end{aligned}$$

Suppose that X_i has its only 1 at entry (k,j), F is the distribution function of noise. Then $\langle X_i,B\rangle=B_{kj}$. Define

$$\begin{aligned}
r(x,B): &= E(\log(1+e^{\langle X_i,B\rangle})-y_i\langle X_i,B\rangle|X_i=x)=E(\log(1+e^{B_{kj}})-y_iB_{kj}),\\
\nabla_{B_{kj}}r(x,B) &= E(\frac{e^{B_{kj}}}{1+e^{B_{kj}}}-y_i)=\int_{-\infty}^{+\infty}(\frac{e^{B_{kj}}}{1+e^{B_{kj}}}-\bar{y})dF(\bar{y})\\
&= -\int_{-\infty}^{+\infty}\bar{y}dF(\bar{y})+\int_{-\infty}^{+\infty}\frac{e^{B_{kj}}}{1+e^{B_{kj}}}dF(\bar{y})\\
&= -\int_{-\infty}^{+\infty}\bar{y}dF(\bar{y})+\frac{e^{B_{kj}}}{1+e^{B_{kj}}},\\
\nabla_{B_{kj}}^2r(x,B) &= \frac{e^{B_{kj}}}{(1+e^{B_{kj}})^2}.
\end{aligned}$$

The Taylor expansion around B' is given by

$$r(x,B)=r(x,B')+\langle\nabla r_{B_{kj}}(x,B'),B_{kj}-B'_{kj}\rangle+\frac{\nabla^2r_{B_{kj}}(x,\tilde{B})}{2}(B_{kj}-B'_{kj})^2,$$

where $\tilde{B}$ is an intermediate point. Suppose that the second-order derivative $\nabla^2 R(B) \succeq LI \succ 0$. We can see that inequality (26) holds with $G(u) = \frac{Lu^2}{2mq}$. □

According to the proof of Theorem 5, we can derive the following formula holds

$$\begin{aligned} & R(\hat{B}) - R(B) + \delta\underline{\lambda}\Omega^+(\hat{B}-B) + \delta\underline{\lambda}\Omega^-(\hat{B}-B) + \lambda_2\|Cvec(\hat{B})\|_1 \\ \leq & \frac{9mqs\bar{\lambda}^2}{2L} + \lambda_* + 2\lambda_1\|B^-\|_* + \lambda_2\|Cvec(B)\|_1. \end{aligned}$$

□

Proof of Theorem 8. Define for all $M > 0$

$$Z_M := \sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \nabla R_n(B') - \nabla R(B'), B - B' \rangle|.$$

$$\rho(B) = \log(1+e^{\langle X_i,B\rangle}) - y_i\langle X_i, B\rangle, \quad \bar{\rho}(B) = X_i^T(\frac{1}{1+e^{-\langle X_i,B\rangle}} - y_i),$$
$$R_n(B) = \frac{1}{n}\sum_{i=1}^n \rho(B), \quad \nabla R_n(B) = \frac{1}{n}\sum_{i=1}^n X_i^T(\frac{1}{1+e^{-\langle X_i,B\rangle}} - y_i) = \frac{1}{n}\sum_{i=1}^n \bar{\rho}(B).$$

We have

$$\begin{aligned} EZ_M: &= E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \nabla R_n(B') - \nabla R(B'), B - B' \rangle| \\ &= E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \frac{1}{n}\sum_{i=1}^n \bar{\rho}(B') - E(\frac{1}{n}\sum_{i=1}^n \bar{\rho}(B')), B - B' \rangle| \\ &\leq 2E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i\bar{\rho}(B'), B - B' \rangle| \\ &= 2E\sup_{B':\underline{\Omega}(B'-B)\leq M} |\langle \frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i^T(\frac{1}{1+e^{-\langle X_i,B'\rangle}} - y_i), B - B' \rangle| \\ &\leq 2\underline{\Omega}(B-B')E\underline{\Omega}_*(\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i) \\ &\leq 4ME\underline{\Omega}_*(\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i). \end{aligned}$$

The first inequality follows from Symmetrization of Expectations in [14], the second inequality follows the dual norm inequality and $\frac{1}{1+e^{-\langle X_i,B'\rangle}} - y_i < 1$. According to the proof of Theorem 6, we have

$$E\underline{\Omega}_*(\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i X_i) \leq C\left(\sqrt{b_2}\sqrt{\frac{log(m+q)}{n}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)})\frac{log(m+q)}{n}\right).$$

Define $f(X_i^T B') = \langle \nabla \rho(B'), B - B' \rangle = \langle \bar{\rho}(B'), B - B' \rangle$. We know $Ef(B') = 0$.

$$\begin{aligned} & \sup_{B': \underline{\Omega}(B'-B) \le M} Var(f(X_1^T B')) \\ = & \sup_{B': \underline{\Omega}(B'-B) \le M} Var\left(\sum_{l=1}^{m} \sum_{j=1}^{q} X_{1_{lj}} (B_{lj} - B'_{lj}) \bar{\rho}(B') \right) \\ \le & \sup_{B': \underline{\Omega}(B'-B) \le M} \bar{L}^2 E \left[\sum_{l=1}^{m} \sum_{j=1}^{q} (B_{lj} - B'_{lj})^2 \right] \\ \le & \frac{\bar{L}^2 M^2}{mq}, \end{aligned}$$

where $\|X_i\| \le \bar{L}$. Therefore, assume that $\|B - B'\|_\infty \le 2\vartheta$, $\|f(X_i^T B')\|_\infty \le 2\vartheta \bar{L}$. According to Bousequet's Concentration Theorem in [1] we obtain that for all $t > 0$

$$P\left(Z_M \ge 8M\bar{L}C \left(\sqrt{b_2} \sqrt{\frac{log(m+q)}{n}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)}) \frac{log(m+q)}{n} \right) + \frac{\bar{L}M}{\sqrt{mq}} \sqrt{\frac{2t}{n}} + \frac{8t\vartheta\bar{L}}{3n} \right) \le e^{-t}.$$

Replacing t by $mlog(m+q)$ we obtain

$$\begin{aligned} P\Bigg(Z_M \ge 8M\bar{L}C & \left(\sqrt{b_2} \sqrt{\frac{log(m+q)}{n}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)}) \frac{log(m+q)}{n} \right) \\ & + \frac{\bar{L}M}{\sqrt{mq}} \sqrt{\frac{2mlog(m+q)}{n}} + \frac{8mlog(m+q)\vartheta\bar{L}}{3n} \Bigg) \le e^{-mlog(m+q)}. \end{aligned}$$

There exist constant $C_0 = 8\bar{L}C + \sqrt{2}\bar{L}, C_1 = 8\vartheta\bar{L}/3$ and

$$\begin{aligned} \lambda_\varepsilon &= C_0 \left(\sqrt{b_2} \sqrt{\frac{log(m+q)}{n}} + \sqrt{\frac{log(m+q)}{nq}} + (\sqrt{2logb_1} + \sqrt{log(1+b_2)}) \frac{log(m+q)}{n} \right) \\ \lambda_* &= \frac{C_1 mlog(m+q)}{n} \end{aligned}$$

such that $P(Z_M \ge M\lambda_\varepsilon + \lambda_*) \le e^{-mlog(m+q)}$. According to Theorem 5 and [5, Lemma B.5], then we have probability at least $1 - (j_0+2)e^{-mlog(m+q)}$ such that

$$\begin{aligned} & R(\hat{B}) - R(B) + \delta \underline{\lambda} \Omega^+(\hat{B} - B) + \delta \underline{\lambda} \Omega^-(\hat{B} - B) + \lambda_2 \|Cvec(\hat{B})\|_1 \\ & \le H(\bar{\lambda} 3\sqrt{s}) + \lambda_* + 2\lambda_1 \|B^-\|_* + \lambda_2 \|Cvec(B)\|_1. \end{aligned}$$

Assume that $q = o(\frac{n}{log(m+q)})$, then $\lambda_\varepsilon \asymp \sqrt{\frac{log(m+q)}{nq}}$. We further assume that $\lambda_2 \asymp \sqrt{\frac{log(m+q)}{nq}}$, then we obtain

$$R(\hat{B}) - R(B) \le R(B) - R(B) + \mathbb{O}_P \left(\frac{mslog(m+q)}{n} + \sqrt{\frac{log(m+q)}{nq}} (\|B^-\|_* + \|Cvec(B)\|_1) \right). \quad (27)$$

Let $B = B^*$ in (27), the convergence rate for the optimal solution $\hat{B}$ is given by

$$\|\hat{B} - B^*\|_F^2 = \mathbb{O}_P\left(\frac{2m^2 qslog(m+q)}{n}\right).$$

□

References

[1] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[2] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[3] Liang Chen, Defeng Sun, and Kim-Chuan Toh. An efficient inexact symmetric gauss–seidel based majorized admm for high-dimensional convex composite conic programming. *Mathematical Programming*, 161(1-2):237–270, 2017.

[4] Asen L Dontchev and R Tyrrell Rockafellar. Implicit functions and solution mappings. *Springer Monographs in Mathematics. Springer*, 208, 2009.

[5] Andreas Elsener, Sara van de Geer, et al. Robust low-rank matrix estimation. *The Annals of Statistics*, 46(6B):3481–3509, 2018.

[6] Maryam Fazel, Ting Kei Pong, Defeng Sun, and Paul Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

[7] Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007.

[8] Deren Han, Defeng Sun, and Liwei Zhang. Linear rate convergence of the alternating direction method of multipliers for convex composite programming. *Mathematics of Operations Research*, 43(2):622–637, 2017.

[9] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[10] Xudong Li, Defeng Sun, and Kim-Chuan Toh. On efficiently solving the subproblems of a level-set method for fused lasso problems. *SIAM Journal on Optimization*, 28(2):1842–1866, 2018.

[11] Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. 1962.

[12] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[13] R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1970.

[14] Aad W Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

[15] Zirui Zhou and Anthony Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.