Multi-Task Learning and Algorithmic Stability: Supplementary Material

The Proof of Theorem 1

In order to prove Theorem 1, we first introduce the Hoeffding's lemma.

Lemma 1 *Given a random variable X, $a \leq X \leq b$, and $\mathbb{E}[X] = 0$, then for any $s > 0$, we have*

$$\mathbb{E}\left[\exp\{sX\}\right] \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}.$$

Then we can prove Theorem 1.

Proof. We first define $X = \{X_1, ..., X_n\}$, $X_{i:j} = \{X_i, \ldots, X_j\}$, $Z_0 = \mathbb{E}[f(X)]$, $Z_i = \mathbb{E}[f(X)|X_1, \ldots, X_i]$, and $Z_n = f(X)$. Then let

$$\begin{aligned} U_k &= \sup_{\mathbf{u}} \{\mathbb{E}[f(X)|X_{1:k-q}, u_{1:q}] - \mathbb{E}[f(X)|X_{1:k-q}]\} \\ L_k &= \inf_{\mathbf{l}} \{\mathbb{E}[f(X)|X_{1:k-q}, l_{1:q}] - \mathbb{E}[f(X)|X_{1:k-q}]\}, \end{aligned}$$

where $\mathbf{u} = (u_1, \ldots, u_q)^T$, $\mathbf{l} = (l_1, \ldots, l_q)^T$, and k is required to be no smaller than q. Then we have

$$\begin{aligned} &U_k - L_k \\ &\leq \sup_{\mathbf{u},\mathbf{l}} \{\mathbb{E}[f(X)|X_{1:k-q}, u_{1:q}] - \mathbb{E}[f(X)|X_{1:k-q}, l_{1:q}]\} \\ &= \sup_{\mathbf{u},\mathbf{l}} \left\{ \int \left[f(X_{1:k-q}, u_{1:q}, y_{k+1:n}) - f(X_{1:k-q}, l_{1:q}, y_{k+1:n}) \right] \prod_{j=k+1}^{n} p(X_j = y_j) dy_{k+1:n} \right\} \\ &\leq \int \sup_{\mathbf{u},\mathbf{l}} \left[f(X_{1:k-q}, u_{1:q}, y_{k+1:n}) - f(X_{1:k-q}, l_{1:q}, y_{k+1:n}) \right] \prod_{j=k+1}^{n} p(X_j = y_j) dy_{k+1:n} \\ &\leq \int a \prod_{j=k+1}^{n} p(X_j = y_j) dy_{k+1:n} \\ &= a, \end{aligned}$$

where the first equality holds due to the independence among all variables, the second inequality holds due to the Jensen's inequality, and the last inequality holds due to the bounded differences condition. It is easy to show that $L_k \leq Z_k - Z_{k-q} \leq U_k$. Since

$$\begin{aligned} &\mathbb{E}[Z_k - Z_{k-q}|X_{1:k-q}] \\ &= \mathbb{E}_{X_{k-q+1:n}}[\mathbb{E}[f(X)|X_{1:k}] - \mathbb{E}[f(X)|X_{1:k-q}]] \\ &= \mathbb{E}[f(X)|X_{1:k-q}] - \mathbb{E}[f(X)|X_{1:k-q}] \\ &= 0, \end{aligned}$$

by applying Lemma 1, we can get

$$\mathbb{E}\left[\exp\{s(Z_k - Z_{k-q})\}|X_{1:k-q}\right] \leq \exp\left\{\frac{s^2 a^2}{8}\right\}. \tag{13}$$

By using Markov's inequality, we can bound $p(f(X) - \mathbb{E}[f(X)] \geq \varepsilon)$ as

$$
\begin{aligned}
&p(f(X) - \mathbb{E}[f(X)] \geq \varepsilon) \\
&\leq \exp\{-s\varepsilon\}\mathbb{E}[\exp\{s(f(X) - \mathbb{E}[f(X)])\}] \\
&= \exp\{-s\varepsilon\}\mathbb{E}[\exp\{s(Z_n - Z_0)\}] \\
&= \exp\{-s\varepsilon\}\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n/q}\left(Z_{iq} - Z_{(i-1)q}\right)\right\}\right] \\
&= \exp\{-s\varepsilon\}\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n/q-1}\left(Z_{iq} - Z_{(i-1)q}\right)\right\}\mathbb{E}\Big[\exp\{Z_n - Z_{n-q}\}|X_{1:n-q}\Big]\right] \\
&\leq \exp\{-s\varepsilon\}\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n/q-1}\left(Z_{iq} - Z_{(i-1)q}\right)\right\}\exp\left\{\frac{s^2a^2}{8}\right\}\right] \\
&\leq \dots \\
&\leq \exp\{-s\varepsilon\}\exp\left\{\frac{ns^2a^2}{8q}\right\} \\
&= \exp\left\{-s\varepsilon + \frac{na^2}{8q}s^2\right\},
\end{aligned}
$$

where the third equality follows due to one property of expectation that $\mathbb{E}[f(X)g(Y)] = \mathbb{E}_X[f(X)\mathbb{E}_Y[g(Y)|X]]$ and the second inequality holds due to the bound shown in Eq. (13). In order to obtain a tight bound, we minimize the last equation with respect to s and get a critical point at $s = \frac{4q\varepsilon}{na^2}$. Substituting, we can get

$$
\exp\left\{-s\varepsilon + \frac{na^2}{8q}s^2\right\} = \exp\left\{-\frac{2q\varepsilon^2}{na^2}\right\},
$$

which completes the proof. □

The Proof of Theorem 2

Proof. By removing one training data point for each task, we can bound the difference on the generalization error as

$$
\begin{aligned}
&|R(\mathcal{A}, \mathcal{S}) - R(\mathcal{A}, \mathcal{S}^{\setminus\mathcal{I}})| \\
&\leq \mathbb{E}_{\{\mathbf{z}_i\}}\left[\left|\sum_i \left(l(\mathcal{A}_S, \mathbf{z}_i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}}, \mathbf{z}_i)\right)\right|\right] \\
&\leq \tau,
\end{aligned}
$$

where the first inequality holds due to the Jensen's inequality and the second one follows because of the definition of the multi-task uniform stability. Then based on the above inequality, we have

$$
\begin{aligned}
&|R(\mathcal{A}, \mathcal{S}) - R(\mathcal{A}, \mathcal{S}^{\mathcal{I}})| \\
&\leq |R(\mathcal{A}, \mathcal{S}) - R(\mathcal{A}, \mathcal{S}^{\setminus\mathcal{I}})| + |R(\mathcal{A}, \mathcal{S}^{\mathcal{I}}) - R(\mathcal{A}, \mathcal{S}^{\setminus\mathcal{I}})| \\
&\leq 2\tau.
\end{aligned} \tag{14}
$$

Moreover, we have

$$
\begin{aligned}
&|R_{emp}(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S}^{\mathcal{I}})| \\
&\leq \frac{1}{n_0}\sum_{j\neq\mathcal{I}_0}|\sum_{i=1}^{m}\left(l(\mathcal{A}_S, \mathbf{z}_j^i) - l(\mathcal{A}_{S^{\mathcal{I}}}, \mathbf{z}_j^i)\right)| + \frac{1}{n_0}\sum_{i=1}^{m}|l(\mathcal{A}_S, \mathbf{z}_{\mathcal{I}_0}^i) - l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_{\mathcal{I}_0}^i)| \\
&\leq \frac{1}{n_0}\sum_{j\neq\mathcal{I}_0}|\sum_{i=1}^{m}\left(l(\mathcal{A}_S, \mathbf{z}_j^i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}}, \mathbf{z}_j^i)\right)| + \frac{1}{n_0}\sum_{j\neq\mathcal{I}_0}|\sum_{i=1}^{m}\left(l(\mathcal{A}_{S^{\mathcal{I}}}, \mathbf{z}_j^i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}}, \mathbf{z}_j^i)\right)| \\
&\quad + \frac{1}{n_0}\sum_{i=1}^{m}|l(\mathcal{A}_S, \mathbf{z}_{\mathcal{I}_0}^i) - l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_{\mathcal{I}_0}^i)| \\
&\leq 2\tau + \frac{mM}{n_0},
\end{aligned} \tag{15}
$$

where without loss of generality we assume different elements in $\mathcal{I}$ are identical (i.e., $\mathcal{I}_i = \mathcal{I}_0$ for $i = 1, \ldots, m$) due to the symmetry of the algorithm $\mathcal{A}$, making the derivation simple and clear. Combining the bounds in Eqs. (14) and (15), for any $\mathcal{I}$ we have

$$\begin{aligned}
&|(R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})) - (R(\mathcal{A}, \mathcal{S}^{\mathcal{I}}) - R_{emp}(\mathcal{A}, \mathcal{S}^{\mathcal{I}}))| \\
&\leq |R(\mathcal{A}, \mathcal{S}) - R(\mathcal{A}, \mathcal{S}^{\mathcal{I}})| + |R_{emp}(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S}^{\mathcal{I}})| \\
&\leq 4\tau + \frac{mM}{n_0}.
\end{aligned}$$

Thus the random variable $R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})$ satisfies the conditions of the generalized McDiarmid's inequality presented in Theorem 1 with $a = 4\tau + \frac{mM}{n_0}$ and hence we have

$$\begin{aligned}
p\Big(R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S}) - \mathbb{E}_S[R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})] \geq \varepsilon\Big) &\leq \exp\left\{-\frac{2\varepsilon^2}{n_0(4\tau + \frac{mM}{n_0})^2}\right\} \\
&= \exp\left\{-\frac{2n_0\varepsilon^2}{(4n_0\tau + mM)^2}\right\}
\end{aligned}$$

or equivalently

$$\begin{aligned}
&p\Big(R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S}) \leq \mathbb{E}_S[R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})] + \varepsilon\Big) \\
&\geq 1 - \exp\left\{-\frac{2n_0\varepsilon^2}{(4n_0\tau + mM)^2}\right\}. \qquad (16)
\end{aligned}$$

Then we need to bound $\mathbb{E}_S[R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})]$. Note that

$$\begin{aligned}
\mathbb{E}_S[R_{emp}(\mathcal{A}, \mathcal{S})] &= \sum_{i=1}^{m} \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbb{E}_S[l(\mathcal{A}_S, \mathbf{z}_j^i)] \\
&= \sum_{i=1}^{m} \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbb{E}_{S, \hat{\mathbf{z}}_j^i}[l(\mathcal{A}_S, \mathbf{z}_j^i)] \\
&= \sum_{i=1}^{m} \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbb{E}_{S, \hat{\mathbf{z}}_j^i}[l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_j^i)] \\
&= \sum_{i=1}^{m} \mathbb{E}_{S, \hat{\mathbf{z}}_i \sim \mathcal{D}_i}[l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_i)],
\end{aligned}$$

where the last second equality holds by renaming $\mathbf{z}_j^i$ as $\hat{\mathbf{z}}_j^i$ and the last one follows by renaming $\hat{\mathbf{z}}_j^i$ as $\hat{\mathbf{z}}_i$. Then we can bound $\mathbb{E}_S[R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})]$ as

$$\begin{aligned}
&\mathbb{E}_S[R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})] \\
&= \mathbb{E}_{S, \{\hat{\mathbf{z}}_i\}}\left[\sum_{i=1}^{m}\Big(l(\mathcal{A}_S, \hat{\mathbf{z}}_i) - l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_i)\Big)\right] \\
&\leq \mathbb{E}_{S, \{\hat{\mathbf{z}}_i\}}\left[\left|\sum_{i=1}^{m}\Big(l(\mathcal{A}_S, \hat{\mathbf{z}}_i) - l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_i)\Big)\right|\right] \\
&\leq \mathbb{E}_{S, \{\hat{\mathbf{z}}_i\}}\left[\left|\sum_{i=1}^{m}\Big(l(\mathcal{A}_S, \hat{\mathbf{z}}_i) - l(\mathcal{A}_{S^{\setminus \mathcal{I}}}, \hat{\mathbf{z}}_i)\Big)\right|\right] + \mathbb{E}_{S, \{\hat{\mathbf{z}}_i\}}\left[\left|\sum_{i=1}^{m}\Big(l(\mathcal{A}_{S^{\mathcal{I}}}, \hat{\mathbf{z}}_i) - l(\mathcal{A}_{S^{\setminus \mathcal{I}}}, \hat{\mathbf{z}}_i)\Big)\right|\right] \\
&\leq 2\tau,
\end{aligned}$$

where $\hat{\mathbf{z}}_i \sim \mathcal{D}_i$ for $i = 1, \ldots, m$ and the last inequality holds due to the definition of the multi-task uniform stability. Due to the above inequality, we have

$$\begin{aligned}
&p\Big(R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S}) \leq 2\tau + \varepsilon\Big) \\
&\geq p\Big(R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S}) \leq \mathbb{E}_S[R(\mathcal{A}, \mathcal{S}) - R_{emp}(\mathcal{A}, \mathcal{S})] + \varepsilon\Big) \\
&\geq 1 - \exp\left\{-\frac{2n_0\varepsilon^2}{(4n_0\tau + mM)^2}\right\},
\end{aligned}$$

where the second inequality holds due to Eq. (16). Then by setting $\exp\left\{-\frac{2n_0\varepsilon^2}{(4n_0\tau+mM)^2}\right\}$ to δ, we can reach the conclusion. □

The Proof of Eq. (6)

Proof. We denote an event $R_i^{ST} \leq R_{emp,i}^{ST} + 2\tau_i + (4n_0\tau_i + M)\sqrt{\frac{\ln(1/\rho)}{2n_0}}$ by $\mathcal{A}_i$. Then we have $p(\mathcal{A}_i) \geq 1-\rho$. Since $\mathcal{A}_i$'s are independent, we have $p(\bigcap_i \mathcal{A}_i) = \prod_{i=1}^m p(\mathcal{A}_i) = (1-\rho)^m$. Since it is well-known that $(1-\rho)^m \geq 1-m\rho$ and $(1-\rho)^m \approx 1-m\rho$ when ρ is very small, we have $p(\bigcap_i \mathcal{A}_i) \geq 1-m\rho$. Since $\bigcap_i \mathcal{A}_i$ implies $R^{ST} \leq R_{emp}^{ST} + 2\sum_{i=1}^m \tau_i + (4n_0\sum_{i=1}^m \tau_i + mM)\sqrt{\frac{\ln(1/\rho)}{2n_0}}$, then we have

$$p\left(R^{ST} \leq R_{emp}^{ST} + 2\sum_{i=1}^m \tau_i + (4n_0\sum_{i=1}^m \tau_i + mM)\sqrt{\frac{\ln(1/\rho)}{2n_0}}\right) \geq p(\bigcap_i \mathcal{A}_i) \geq 1-m\rho.$$

In order to compare with the multi-task bound in Eq. (3) which holds with probability at least $1-\delta$, we define $\delta = m\rho$ and then reach the conclusion. □

The Proof of Theorem 3

Proof. We first define the Bregman divergence associated to a function G of g to g' by

$$d_G(g, g') = G(g) - G(g') - \langle g - g', \bigtriangledown G(g')\rangle,$$

where $\bigtriangledown G(g')$ denotes the gradient of $G(\cdot)$. It is easy to show that $d_G(g, g')$ is nonnegative when $G(\cdot)$ is convex. We define

$$\begin{aligned}
R_r(\mathbf{W}) &= \sum_{i=1}^m \frac{1}{n_i}\sum_{j=1}^{n_i} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i) + \mathrm{tr}(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T)\\
R_r^{\setminus\mathcal{I}}(\mathbf{W}) &= \sum_{i=1}^m \frac{1}{n_i}\sum_{j\neq\mathcal{I}_i} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i) + \mathrm{tr}(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T)\\
R_{emp}(\mathbf{W}) &= \sum_{i=1}^m \frac{1}{n_i}\sum_{j=1}^{n_i} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i)\\
R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}) &= \sum_{i=1}^m \frac{1}{n_i}\sum_{j\neq\mathcal{I}_i} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i).
\end{aligned}$$

Let $\mathbf{W}$ and $\mathbf{W}^{\setminus\mathcal{I}}$ denote the minimizors of $R_r(\mathbf{W})$ and $R_r^{\setminus\mathcal{I}}(\mathbf{W})$ respectively. Then we have

$$\begin{aligned}
&d_{R_r}(\mathbf{W}^{\setminus\mathcal{I}}, \mathbf{W}) + d_{R_r^{\setminus\mathcal{I}}}(\mathbf{W}, \mathbf{W}^{\setminus\mathcal{I}})\\
=&R_r(\mathbf{W}^{\setminus\mathcal{I}}) - R_r(\mathbf{W}) + R_r^{\setminus\mathcal{I}}(\mathbf{W}) - R_r^{\setminus\mathcal{I}}(\mathbf{W}^{\setminus\mathcal{I}})\\
=&\sum_{i=1}^m \frac{1}{n_i}\left(c\left((\mathbf{w}_i^{\setminus\mathcal{I}})^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i\right) - c\left(\mathbf{w}_i^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i\right)\right),
\end{aligned}$$

where $\mathbf{w}_i^{\setminus\mathcal{I}}$ is the ith column of $\mathbf{W}^{\setminus\mathcal{I}}$ and the first equality holds since $\mathbf{W}$ and $\mathbf{W}^{\setminus\mathcal{I}}$ are the minimizors of $R_r(\mathbf{W})$ and $R_r^{\setminus\mathcal{I}}(\mathbf{W})$ respectively, implying that the gradients are zero. Due to the nonnegativity of the Bregman divergence, we have $d_{R_{emp}}(\mathbf{W}^{\setminus\mathcal{I}}, \mathbf{W}) + d_{R_{emp}^{\setminus\mathcal{I}}}(\mathbf{W}, \mathbf{W}^{\setminus\mathcal{I}}) \geq 0$. By denoting the regularizor (i.e. $\mathrm{tr}(\mathbf{W}\mathbf{\Omega}^{-1}\mathbf{W}^T)$) in problem (8) by a function $N(\mathbf{W})$ and using one fact that $d_{A+B}(g, g') = d_A(g, g') + d_B(g, g')$ for any functions A and B, we can get

$$d_N(\mathbf{W}^{\setminus\mathcal{I}}, \mathbf{W}) + d_N(\mathbf{W}, \mathbf{W}^{\setminus\mathcal{I}}) \leq \sum_{i=1}^m \frac{1}{n_i}\left(c\left((\mathbf{w}_i^{\setminus\mathcal{I}})^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i\right) - c\left(\mathbf{w}_i^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i\right)\right).$$

After some calculation, we can get $d_N(\mathbf{W}^{\setminus\mathcal{I}}, \mathbf{W}) = d_N(\mathbf{W}, \mathbf{W}^{\setminus\mathcal{I}}) = \mathrm{tr}\left((\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})\mathbf{\Omega}^{-1}(\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})^T\right)$. Then we can simplify the above inequality as

$$
\begin{aligned}
&2\mathrm{tr}\left((\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})\mathbf{\Omega}^{-1}(\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})^T\right) \\
&\leq \sum_{i=1}^{m} \frac{1}{n_i}\left(c\left((\mathbf{w}_i^{\setminus\mathcal{I}})^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i\right) - c\left(\mathbf{w}_i^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i\right)\right) \\
&\leq \sum_{i=1}^{m} \frac{\sigma}{n_i}|(\mathbf{w}_i^{\setminus\mathcal{I}})^T\mathbf{x}_{\mathcal{I}_i}^i - \mathbf{w}_i^T\mathbf{x}_{\mathcal{I}_i}^i| \text{ (σ-admissible loss function)} \\
&\leq \sum_{i=1}^{m} \frac{\sigma}{n_i}\|\mathbf{w}_i^{\setminus\mathcal{I}} - \mathbf{w}_i\|_2\|\mathbf{x}_{\mathcal{I}_i}^i\|_2 \text{ (Cauchy-Schwarz Inequality)} \\
&\leq \frac{\kappa\sigma}{\min_i n_i}\sum_{i=1}^{m}\|\mathbf{w}_i^{\setminus\mathcal{I}} - \mathbf{w}_i\|_2 \text{ (definition of κ)} \\
&\leq \frac{\kappa\sigma\sqrt{m}}{\min_i n_i}\|\mathbf{W}^{\setminus\mathcal{I}} - \mathbf{W}\|_F \text{ (Cauchy-Schwarz Inequality)}, \qquad (17)
\end{aligned}
$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm of a vector and $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Since $\lambda_1(\mathbf{\Omega})$ is the largest eigenvalue of $\mathbf{\Omega}$, we can get $\lambda_1(\mathbf{\Omega})\mathbf{I}_m \succeq \mathbf{\Omega}$ and $\mathbf{\Omega}^{-1} \succeq \frac{1}{\lambda_1(\mathbf{\Omega})}\mathbf{I}_m$ where $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is a positive semidefinite matrix and $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. Hence we have $\mathrm{tr}\left((\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})\mathbf{\Omega}^{-1}(\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})^T\right) \geq \frac{1}{\lambda_1(\mathbf{\Omega})}\mathrm{tr}\left((\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})(\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}})^T\right) = \frac{1}{\lambda_1(\mathbf{\Omega})}\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\|_F^2$. By plugging this inequality into the Eq. (17), we can get

$$
\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\|_F \leq \frac{\lambda_1(\mathbf{\Omega})\kappa\sigma\sqrt{m}}{2\min_i n_i}.
$$

Similar to the derivation in Eq. (17), we can easily get

$$
\left|\sum_{i=1}^{m}\left(l(\mathcal{A}_S, \mathbf{z}_i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}}, \mathbf{z}_i)\right)\right| \leq \kappa\sigma\sqrt{m}\|\mathbf{W}^{\setminus\mathcal{I}} - \mathbf{W}\|_F \leq \frac{\lambda_1(\mathbf{\Omega})\kappa^2\sigma^2 m}{2\min_i n_i},
$$

which holds for any $\mathbf{z}_i \sim \mathcal{D}_i$ $(i = 1, \ldots, m)$. By setting n_i to be n_0 for $i = 1, \ldots, m$, we can complete the proof. □

The Proof of Theorem 4

Proof. Since $\mathbf{W}^\star$ is the optimal solution of problem (8), we have

$$
\begin{aligned}
\mathrm{tr}(\mathbf{W}^\star\mathbf{\Omega}^{-1}(\mathbf{W}^\star)^T) \leq& \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c((\mathbf{w}_i^\star)^T\mathbf{x}_j^i, y_j^i) + \mathrm{tr}(\mathbf{W}^\star\mathbf{\Omega}^{-1}(\mathbf{W}^\star)^T) \\
\leq& \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c(0, y_j^i) \\
\leq& m\eta,
\end{aligned}
$$

where the last inequality holds due to the first property of the cost function $c(\cdot,\cdot)$. Since $\mathrm{tr}(\mathbf{W}^\star\mathbf{\Omega}^{-1}(\mathbf{W}^\star)^T) \leq \frac{1}{\lambda_1(\mathbf{\Omega})}\mathrm{tr}(\mathbf{W}^\star(\mathbf{W}^\star)^T) = \frac{1}{\lambda_1(\mathbf{\Omega})}\|\mathbf{W}^\star\|_F^2$, we can get $\|\mathbf{W}^\star\|_F \leq \sqrt{\lambda_1(\mathbf{\Omega})m\eta}$, which implies $\|\mathbf{w}_i^\star\|_2 \leq \sqrt{\lambda_1(\mathbf{\Omega})m\eta}$ for $i = 1, \ldots, m$. Then we can bound $c((\mathbf{w}_i^\star)^T\mathbf{x}, y)$ as

$$
\begin{aligned}
c((\mathbf{w}_i^\star)^T\mathbf{x}, y) &\leq |c((\mathbf{w}_i^\star)^T\mathbf{x}, y) - c(0, y)| + c(0, y) \\
&\leq \sigma|(\mathbf{w}_i^\star)^T\mathbf{x}| + \eta \\
&\leq \sigma\|\mathbf{w}_i^\star\|_2\|\mathbf{x}\|_2 + \eta \\
&\leq \sigma\kappa\sqrt{\lambda_1(\mathbf{\Omega})m\eta} + \eta,
\end{aligned}
$$

where the first inequality holds due to one fact that $|a+b| \leq |a|+|b|$ for any scalars a and b, and the second inequality holds due to the properties of the cost function, the third inequality holds due to the Cauchy-Schwarz inequality, and the last inequality holds due to the boundedness of $\|\mathbf{w}_i^\star\|_2$ and $\|\mathbf{x}\|_2$. Finally we reach the conclusion. □

The Proof of Theorem 5

For a general regularized multi-task algorithm defined in Eq. (7), we define

$$R_r(\mathbf{W}) = \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i) + g(\mathbf{W}) \tag{18}$$

$$R_r^{\setminus\mathcal{I}}(\mathbf{W}) = \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j\neq\mathcal{I}_i} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i) + g(\mathbf{W}) \tag{19}$$

$$R_{emp}(\mathbf{W}) = \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i)$$

$$R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}) = \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j\neq\mathcal{I}_i} c(\mathbf{w}_i^T\mathbf{x}_j^i, y_j^i).$$

Then we have the following result.

Lemma 2 *For any $t \in [0,1]$, we have*

$$g(\mathbf{W}_\star) - g(\mathbf{W}_\star + t\Delta\mathbf{W}_\star) + g(\mathbf{W}_\star^{\setminus\mathcal{I}}) - g(\mathbf{W}_\star^{\setminus\mathcal{I}} - t\Delta\mathbf{W}_\star) \leq \frac{t\sigma\kappa\sqrt{m}}{n_0}\left\|\mathbf{W}_\star - \mathbf{W}_\star^{\setminus\mathcal{I}}\right\|_F,$$

where $\mathbf{W}_\star$ is a minimizer of $R_r(\cdot)$, $\mathbf{W}_\star^{\setminus\mathcal{I}}$ is a minimizer of $R_r^{\setminus\mathcal{I}}(\cdot)$, and $\Delta\mathbf{W}_\star = \mathbf{W}_\star^{\setminus\mathcal{I}} - \mathbf{W}_\star$.

Proof. Since the cost function $c(\cdot,\cdot)$ is convex, $R_{emp}^{\setminus\mathcal{I}}(\cdot)$ is also convex, implying that

$$R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star + t\Delta\mathbf{W}_\star) - R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star) \leq t(R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}}) - R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star)),$$

which holds for any $t \in [0,1]$. Similarly we have

$$R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}} - t\Delta\mathbf{W}_\star) - R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}}) \leq t(R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star) - R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}})).$$

Summing the two preceding inequalities yields

$$R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star + t\Delta\mathbf{W}_\star) - R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star) + R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}} - t\Delta\mathbf{W}_\star) - R_{emp}^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}}) \leq 0. \tag{20}$$

According to the definitions of $\mathbf{W}_\star$ and $\mathbf{W}_\star^{\setminus\mathcal{I}}$, we have

$$\begin{aligned} &R_r(\mathbf{W}_\star) - R_r(\mathbf{W}_\star + t\Delta\mathbf{W}_\star) \leq 0 \\ &R_r^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}}) - R_r^{\setminus\mathcal{I}}(\mathbf{W}_\star^{\setminus\mathcal{I}} - t\Delta\mathbf{W}_\star) \leq 0. \end{aligned}$$

Summing the two preceding inequalities and Eq. (20), we can get

$$\begin{aligned} &g(\mathbf{W}_\star) - g(\mathbf{W}_\star + t\Delta\mathbf{W}_\star) + g(\mathbf{W}_\star^{\setminus\mathcal{I}}) - g(\mathbf{W}_\star^{\setminus\mathcal{I}} - t\Delta\mathbf{W}_\star) \\ &\leq \frac{1}{n_0}\sum_{i=1}^{m}(c((\tilde{\mathbf{w}}_i^\star)^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i) - c((\mathbf{w}_i^\star)^T\mathbf{x}_{\mathcal{I}_i}^i, y_{\mathcal{I}_i}^i)), \end{aligned}$$

where $\tilde{\mathbf{w}}_i^\star$ is the ith column of $\mathbf{W}_\star + t\Delta\mathbf{W}_\star$ and $\mathbf{w}_i^\star$ is the ith column of $\mathbf{W}_\star$. Since $c(\cdot,\cdot)$ is σ-admissible, we have

$$\begin{aligned} g(\mathbf{W}_\star) - g(\mathbf{W}_\star + t\Delta\mathbf{W}_\star) + g(\mathbf{W}_\star^{\setminus\mathcal{I}}) - g(\mathbf{W}_\star^{\setminus\mathcal{I}} - t\Delta\mathbf{W}_\star) &\leq \frac{t\sigma}{n_0}\sum_{i=1}^{m}\left|(\mathbf{w}_i^\star - \mathbf{w}_i^{\setminus\mathcal{I},\star})^T\mathbf{x}_{\mathcal{I}_i}^i\right| \\ &\leq \frac{t\sigma\kappa}{n_0}\sum_{i=1}^{m}\left\|\mathbf{w}_i^\star - \mathbf{w}_i^{\setminus\mathcal{I},\star}\right\|_2 \\ &\leq \frac{t\sigma\kappa\sqrt{m}}{n_0}\left\|\mathbf{W}_\star - \mathbf{W}_\star^{\setminus\mathcal{I}}\right\|_F, \end{aligned}$$

in which we reach the conclusion. □

Then based on Lemma 2, in the following we can prove Theorem 5.

Proof of Theorem 5. We define the regularization function $g(\mathbf{W})$ as $g(\mathbf{W}) = \alpha\|\mathbf{W}\|_* + \frac{\beta}{2}\|\mathbf{W}\|_F^2$. By denoting by $\mathbf{W}$ and $\mathbf{W}^{\setminus\mathcal{I}}$ the minimizers of $R_r(\cdot)$ and $R_r^{\setminus\mathcal{I}}(\cdot)$ respectively as defined in Eqs. (18) and (19), we have

$$g(\mathbf{W}) + g(\mathbf{W}^{\setminus\mathcal{I}}) - 2g\left(\frac{1}{2}(\mathbf{W} + \mathbf{W}^{\setminus\mathcal{I}})\right) \leq \frac{\sigma\kappa\sqrt{m}}{2n_0}\left\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\right\|_F,$$

which holds due to Lemma 2 by setting t to be $1/2$. Moreover, we have

$$\begin{aligned} & g(\mathbf{W}) + g(\mathbf{W}^{\setminus\mathcal{I}}) - 2g\left(\frac{1}{2}(\mathbf{W} + \mathbf{W}^{\setminus\mathcal{I}})\right) \\ = \;& \alpha\left(\|\mathbf{W}\|_* + \|\mathbf{W}^{\setminus\mathcal{I}}\|_* - \|\mathbf{W} + \mathbf{W}^{\setminus\mathcal{I}}\|_*\right) + \frac{\beta}{4}\left\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2 \\ \geq \;& \frac{\beta}{4}\left\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2, \end{aligned}$$

where the inequality holds due to the triangle inequality property of the trace norm. Combining the preceding two inequalities yields

$$\left\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\right\|_F \leq \frac{2\sigma\kappa\sqrt{m}}{\beta n_0}.$$

Similar to the derivation in Eq. (17), we can easily get

$$\left|\sum_{i=1}^{m}\left(l(\mathcal{A}_S, \mathbf{z}_i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}}, \mathbf{z}_i)\right)\right| \leq \kappa\sigma\sqrt{m}\|\mathbf{W}^{\setminus\mathcal{I}} - \mathbf{W}\|_F \leq \frac{2\kappa^2\sigma^2 m}{\beta n_0},$$

in which we reach the conclusion. □

The Proof of Theorem 6

Proof. Since $\mathbf{W}^\star$ is the optimal solution of problem (10), we have

$$\begin{aligned} \alpha\|\mathbf{W}^\star\|_* + \frac{\beta}{2}\|\mathbf{W}^\star\|_F^2 \leq & \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c((\mathbf{w}_i^\star)^T\mathbf{x}_j^i, y_j^i) + \alpha\|\mathbf{W}^\star\|_* + \frac{\beta}{2}\|\mathbf{W}^\star\|_F^2 \\ \leq & \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c(0, y_j^i) \\ \leq & m\eta, \end{aligned}$$

where the second inequality holds since $\mathbf{W}^\star$ is the minimizer of problem (10) and the last inequality holds due to the first property of the cost function $c(\cdot,\cdot)$. Since $\|\mathbf{W}^\star\|_* \geq \|\mathbf{W}^\star\|_F$, we can get $\alpha\|\mathbf{W}^\star\|_F + \frac{\beta}{2}\|\mathbf{W}^\star\|_F^2 \leq m\eta$, leading to $\|\mathbf{W}^\star\|_F \leq \frac{1}{\beta}(\sqrt{\alpha^2 + 2m\eta\beta} - \alpha)$ which implies that $\|\mathbf{w}_i^\star\|_2 \leq \frac{1}{\beta}(\sqrt{\alpha^2 + 2m\eta\beta} - \alpha)$ for $i = 1, \ldots, m$. Then we can bound $c((\mathbf{w}_i^\star)^T\mathbf{x}, y)$ as

$$\begin{aligned} c((\mathbf{w}_i^\star)^T\mathbf{x}, y) &\leq |c((\mathbf{w}_i^\star)^T\mathbf{x}, y) - c(0, y)| + c(0, y) \\ &\leq \sigma|(\mathbf{w}_i^\star)^T\mathbf{x}| + \eta \\ &\leq \sigma\|\mathbf{w}_i^\star\|_2\|\mathbf{x}\|_2 + \eta \\ &\leq \frac{\sigma\kappa}{\beta}(\sqrt{\alpha^2 + 2m\eta\beta} - \alpha) + \eta, \end{aligned}$$

where the first inequality holds due to one fact that $|a+b| \leq |a|+|b|$ for any scalars a and b, and the second inequality holds due to the properties of the cost function, the third inequality holds due to the Cauchy-Schwarz inequality, and the last inequality holds due to the boundedness of $\|\mathbf{w}_i^\star\|_2$ and $\|\mathbf{x}\|_2$. Finally we reach the conclusion. □

The Proof of Theorem 7

Proof. We define the regularization function $g(\mathbf{W})$ as $g(\mathbf{W}) = \frac{\alpha}{2}\|\mathbf{W}\|_*^2 + \frac{\beta}{2}\|\mathbf{W}\|_F^2$. By denoting by $\mathbf{W}$ and $\mathbf{W}^{\setminus\mathcal{I}}$ the minimizers of $R_r(\cdot)$ and $R_r^{\setminus\mathcal{I}}(\cdot)$ respectively as defined in Eqs. (18) and (19), we have

$$g(\mathbf{W}) + g(\mathbf{W}^{\setminus\mathcal{I}}) - 2g\left(\frac{1}{2}(\mathbf{W} + \mathbf{W}^{\setminus\mathcal{I}})\right) \leq \frac{\sigma\kappa\sqrt{m}}{2n_0}\left\|\mathbf{W} - \mathbf{W}^{\setminus\mathcal{I}}\right\|_F,$$

which holds due to Lemma 2 by setting t to be $1/2$. Moreover, we have

$$\begin{aligned}
& g(\mathbf{W}) + g(\mathbf{W}^{\setminus\mathcal{I}}) - 2g\left(\frac{1}{2}(\mathbf{W}+\mathbf{W}^{\setminus\mathcal{I}})\right) \\
= & \frac{\alpha}{2}\left(\|\mathbf{W}\|_*^2 + \|\mathbf{W}^{\setminus\mathcal{I}}\|_*^2 - \frac{1}{2}\|\mathbf{W}+\mathbf{W}^{\setminus\mathcal{I}}\|_*^2\right) + \frac{\beta}{4}\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2 \\
\geq & \frac{\alpha}{2}\left(\|\mathbf{W}\|_*^2 + \|\mathbf{W}^{\setminus\mathcal{I}}\|_*^2 - \frac{1}{2}(\|\mathbf{W}\|_* + \|\mathbf{W}^{\setminus\mathcal{I}}\|_*)^2\right) + \frac{\beta}{4}\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2 \\
= & \frac{\alpha}{4}\left(\|\mathbf{W}\|_*^2 + \|\mathbf{W}^{\setminus\mathcal{I}}\|_*^2 - 2\|\mathbf{W}\|_*\|\mathbf{W}^{\setminus\mathcal{I}}\|_*\right) + \frac{\beta}{4}\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2 \\
\geq & \frac{\alpha}{4}\left(\|\mathbf{W}\|_F^2 + \|\mathbf{W}^{\setminus\mathcal{I}}\|_F^2 - 2\|\mathbf{W}\|_*\|\mathbf{W}^{\setminus\mathcal{I}}\|_*\right) + \frac{\beta}{4}\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2 \\
\geq & \frac{\alpha}{4}\left(\|\mathbf{W}\|_F^2 + \|\mathbf{W}^{\setminus\mathcal{I}}\|_F^2 - 2\mathrm{tr}(\mathbf{W}^T\mathbf{W}^{\setminus\mathcal{I}})\right) + \frac{\beta}{4}\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2 \\
= & \frac{\alpha+\beta}{4}\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F^2
\end{aligned}$$

where the first inequality holds due to the triangle inequality property of the trace norm, the second inequality holds since $\|\mathbf{W}\|_* \geq \|\mathbf{W}\|_F$ for any matrix $\mathbf{W}$, and the last inequality holds due to the Von Neumanns trace inequality. Combining the preceding two inequalities yields

$$\left\|\mathbf{W}-\mathbf{W}^{\setminus\mathcal{I}}\right\|_F \leq \frac{2\sigma\kappa\sqrt{m}}{(\alpha+\beta)n_0}.$$

Similar to the derivation in Eq. (17), we can easily get

$$\left|\sum_{i=1}^{m}(l(\mathcal{A}_S,\mathbf{z}_i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}},\mathbf{z}_i))\right| \leq \kappa\sigma\sqrt{m}\|\mathbf{W}^{\setminus\mathcal{I}}-\mathbf{W}\|_F \leq \frac{2\kappa^2\sigma^2 m}{(\alpha+\beta)n_0},$$

in which we reach the conclusion. □

The Proof of Theorem 8

Proof. Since $\mathbf{W}^\star$ is the optimal solution of problem (11), we have

$$\begin{aligned}
\frac{\alpha+\beta}{2}\|\mathbf{W}^\star\|_F^2 \leq & \frac{\alpha}{2}\|\mathbf{W}^\star\|_*^2 + \frac{\beta}{2}\|\mathbf{W}^\star\|_F^2 \\
\leq & \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c((\mathbf{w}_i^\star)^T\mathbf{x}_j^i, y_j^i) + \frac{\alpha}{2}\|\mathbf{W}^\star\|_*^2 + \frac{\beta}{2}\|\mathbf{W}^\star\|_F^2 \\
\leq & \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c(0, y_j^i) \\
\leq & m\eta,
\end{aligned}$$

where the first inequality holds since $\|\mathbf{W}^\star\|_F \leq \|\mathbf{W}^\star\|_*$, the second inequality holds due to the nonnegative cost function, the third inequality holds since $\mathbf{W}^\star$ is the minimizer of problem (10), and the last inequality holds due to the first property of the cost function $c(\cdot,\cdot)$. Then we can get $\|\mathbf{W}^\star\|_F \leq \sqrt{\frac{2m\eta}{\alpha+\beta}}$, implying that $\|\mathbf{w}_i^\star\|_2 \leq \sqrt{\frac{2m\eta}{\alpha+\beta}}$ for $i = 1,\ldots,m$. Then we can bound $c((\mathbf{w}_i^\star)^T\mathbf{x}, y)$ as

$$\begin{aligned}
c((\mathbf{w}_i^\star)^T\mathbf{x}, y) &\leq |c((\mathbf{w}_i^\star)^T\mathbf{x}, y) - c(0,y)| + c(0,y) \\
&\leq \sigma|(\mathbf{w}_i^\star)^T\mathbf{x}| + \eta \\
&\leq \sigma\|\mathbf{w}_i^\star\|_2\|\mathbf{x}\|_2 + \eta \\
&\leq \sigma\kappa\sqrt{\frac{2m\eta}{\alpha+\beta}} + \eta,
\end{aligned}$$

where the first inequality holds due to one fact that $|a+b| \leq |a|+|b|$ for any scalars a and b, and the second inequality holds due to the properties of the cost function, the third inequality holds due to the Cauchy-Schwarz inequality, and the last inequality holds due to the boundedness of $\|\mathbf{w}_i^\star\|_2$ and $\|\mathbf{x}\|_2$. Finally we reach the conclusion. □

The Proof of Theorem 9

Proof. We define $\mathbf{\Theta}$ and the regularization function $g(\mathbf{\Theta})$ as $\mathbf{\Theta} = (\mathbf{L}, \mathbf{S})$ and $g(\mathbf{\Theta}) = \alpha\|\mathbf{L}\|_* + \beta\|\mathbf{S}\|_1 + \frac{\gamma}{2}\|\mathbf{L}+\mathbf{S}\|_F^2$. By denoting by $\mathbf{\Theta}$ and $\mathbf{\Theta}^{\setminus\mathcal{I}}$ the minimizers of $R_r(\cdot)$ and $R_r^{\setminus\mathcal{I}}(\cdot)$ respectively as defined in Eqs. (18) and (19), we have

$$g(\mathbf{\Theta}) + g(\mathbf{\Theta}^{\setminus\mathcal{I}}) - 2g\left(\frac{1}{2}(\mathbf{\Theta}+\mathbf{\Theta}^{\setminus\mathcal{I}})\right) \leq \frac{\sigma\kappa\sqrt{m}}{2n_0}\left\|\mathbf{\Theta}-\mathbf{\Theta}^{\setminus\mathcal{I}}\right\|_F,$$

which holds due to Lemma 2 by setting t to be $1/2$. Moreover, we have

$$\begin{aligned}
& g(\mathbf{\Theta}) + g(\mathbf{\Theta}^{\setminus\mathcal{I}}) - 2g\left(\frac{1}{2}(\mathbf{\Theta}+\mathbf{\Theta}^{\setminus\mathcal{I}})\right) \\
= \;& \alpha\left(\|\mathbf{L}\|_* + \|\mathbf{L}^{\setminus\mathcal{I}}\|_* - \|\mathbf{L}+\mathbf{L}^{\setminus\mathcal{I}}\|_*\right) + \beta\left(\|\mathbf{S}\|_1 + \|\mathbf{S}^{\setminus\mathcal{I}}\|_1 - \|\mathbf{S}+\mathbf{S}^{\setminus\mathcal{I}}\|_1\right) \\
& + \frac{\gamma}{4}\|\mathbf{\Theta}-\mathbf{\Theta}^{\setminus\mathcal{I}}\|_F^2 \\
\geq \;& \frac{\gamma}{4}\|\mathbf{\Theta}-\mathbf{\Theta}^{\setminus\mathcal{I}}\|_F^2,
\end{aligned}$$

where the inequality holds due to the triangle inequality property of the trace norm. Combining the preceding two inequalities yields

$$\left\|\mathbf{\Theta}-\mathbf{\Theta}^{\setminus\mathcal{I}}\right\|_F \leq \frac{2\sigma\kappa\sqrt{m}}{\gamma n_0}.$$

Similar to the derivation in Eq. (17), we can easily get

$$\begin{aligned}
& \left|\sum_{i=1}^{m}\left(l(\mathcal{A}_S, \mathbf{z}_i) - l(\mathcal{A}_{S^{\setminus\mathcal{I}}}, \mathbf{z}_i)\right)\right| \\
\leq \;& \kappa\sigma\sqrt{m}\|\mathbf{W}^{\setminus\mathcal{I}}-\mathbf{W}\|_F \\
= \;& \kappa\sigma\sqrt{m}\left\|\left(\mathbf{\Theta}-\mathbf{\Theta}^{\setminus\mathcal{I}}\right)\begin{pmatrix}\mathbf{I}_m \\ \mathbf{I}_m\end{pmatrix}\right\|_F \\
\leq \;& \kappa\sigma\sqrt{m}\|\mathbf{\Theta}-\mathbf{\Theta}^{\setminus\mathcal{I}}\|_F\left\|\begin{pmatrix}\mathbf{I}_m \\ \mathbf{I}_m\end{pmatrix}\right\|_F \\
\leq \;& \frac{2\sqrt{2}\kappa^2\sigma^2 m\sqrt{m}}{\gamma n_0},
\end{aligned}$$

where $\mathbf{W} = \mathbf{L}+\mathbf{S}$, $\mathbf{W}^{\setminus\mathcal{I}} = \mathbf{L}^{\setminus\mathcal{I}} + \mathbf{S}^{\setminus\mathcal{I}}$, and the second inequality holds due to the Cauchy-Schwarz inequality. Finally we reach the conclusion. □

The Proof of Theorem 10

Proof. We define $\mathbf{W}^\star$ as $\mathbf{W}^\star = \mathbf{L}^\star + \mathbf{S}^\star$. Since the pair $(\mathbf{L}^\star, \mathbf{S}^\star)$ is the optimal solution of problem (12), we have

$$\begin{aligned}
& \alpha\|\mathbf{L}^\star\|_* + \beta\|\mathbf{S}^\star\|_1 + \frac{\gamma}{2}\left(\|\mathbf{L}^\star\|_F^2 + \|\mathbf{S}^\star\|_F^2\right) \\
\leq & \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c((\mathbf{w}_i^\star)^T\mathbf{x}_j^i, y_j^i) + \alpha\|\mathbf{L}^\star\|_* + \beta\|\mathbf{S}^\star\|_1 + \frac{\gamma}{2}\left(\|\mathbf{L}^\star\|_F^2 + \|\mathbf{S}^\star\|_F^2\right) \\
\leq & \frac{1}{n_0}\sum_{i=1}^{m}\sum_{j=1}^{n_0} c(0, y_j^i) \\
\leq & m\eta,
\end{aligned}$$

where the first inequality holds due to the nonnegative cost function, the second inequality holds since $(\mathbf{L}^\star, \mathbf{S}^\star)$ is the minimizer of problem (10), and the last inequality holds due to the first property of the cost function $c(\cdot,\cdot)$. Moreover, since

$$\begin{aligned}
\alpha\|\mathbf{L}^\star\|_* + \beta\|\mathbf{S}^\star\|_1 \;&\geq\; \alpha\|\mathbf{L}^\star\|_F + \beta\|\mathbf{S}^\star\|_1 \\
&\geq\; \alpha\|\mathbf{L}^\star\|_F + \beta\|\mathbf{S}^\star\|_F \\
&\geq\; \theta(\|\mathbf{L}^\star\|_F + \|\mathbf{S}^\star\|_F) \\
&\geq\; \theta\|\mathbf{L}^\star + \mathbf{S}^\star\|_F \\
&=\; \theta\|\mathbf{W}^\star\|_F,
\end{aligned}$$

where the first inequality holds since $\|\mathbf{L}^\star\|_* \geq \|\mathbf{L}^\star\|_F$, the second inequality holds since $\|\mathbf{S}^\star\|_1 \geq \|\mathbf{S}^\star\|_F$, and the fourth inequality holds due to the triangle inequality of the matrix Frobenius norm, and

$$\|\mathbf{L}^\star\|_F^2 + \|\mathbf{S}^\star\|_F^2 \geq \frac{1}{2}\|\mathbf{L}^\star + \mathbf{S}^\star\|_F^2 = \frac{1}{2}\|\mathbf{W}^\star\|_F^2$$

then we can get $\frac{\gamma}{4}\|\mathbf{W}^\star\|_F^2 + \theta\|\mathbf{W}^\star\|_F - m\eta \leq 0$, implying that $\|\mathbf{W}^\star\|_F \leq \frac{2}{\gamma}(\sqrt{\theta^2 + m\eta\gamma} - \theta)$ and $\|\mathbf{w}_i^\star\|_2 \leq \frac{2}{\gamma}(\sqrt{\theta^2 + m\eta\gamma} - \theta)$ for $i = 1, \ldots, m$. Then we can bound $c((\mathbf{w}_i^\star)^T\mathbf{x}, y)$ as

$$\begin{aligned} c((\mathbf{w}_i^\star)^T\mathbf{x}, y) &\leq |c((\mathbf{w}_i^\star)^T\mathbf{x}, y) - c(0, y)| + c(0, y) \\ &\leq \sigma|(\mathbf{w}_i^\star)^T\mathbf{x}| + \eta \\ &\leq \sigma\|\mathbf{w}_i^\star\|_2\|\mathbf{x}\|_2 + \eta \\ &\leq \frac{2\sigma\kappa}{\gamma}(\sqrt{\theta^2 + m\eta\gamma} - \theta) + \eta, \end{aligned}$$

where the first inequality holds due to one fact that $|a+b| \leq |a|+|b|$ for any scalars a and b, and the second inequality holds due to the properties of the cost function, the third inequality holds due to the Cauchy-Schwarz inequality, and the last inequality holds due to the boundedness of $\|\mathbf{w}_i^\star\|_2$ and $\|\mathbf{x}\|_2$. Finally we reach the conclusion. □