Supplementary Material: Minimum Volume Multi-Task Learning

Bo Liu* Ji Liu† Sridhar Mahadevan‡ Yong Ge§ Deguang Kong¶

First we define the operators necessary for the theoretical analysis.

1 Operators

Definition 1: (Operator P) Given a matrix $\Gamma \in \mathbb{R}^{m \times n}$ with the following representation

$$\Gamma = U \left[\begin{array}{cc} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{array} \right] V^T,$$

where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ are two orthonormal matrix, define the operator $(\mathrm{P}, \mathrm{P_c})$ pair as follows

$$\begin{aligned} \Gamma_1 &:= \mathrm{P}(\Gamma) = U \left[\begin{array}{cc} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & 0 \end{array} \right] V^T \\ \Gamma_2 &:= \mathrm{P_c}(\Gamma) = U \left[\begin{array}{cc} 0 & 0 \\ 0 & \Gamma_{22} \end{array} \right] V^T \end{aligned}$$

Lemma 1: Given an arbitrary matrix pair $(\Phi, \Gamma) \in \mathbb{R}^{m \times n}$ with $\mathrm{rank}(\Phi) = r$, and the SVD of Φ is

$$\Phi = U \left[\begin{array}{cc} \Sigma & 0 \\ 0 & 0 \end{array} \right] V^T,$$

where Σ is the diagonal matrix where the diagonal elements are the non-zero singular values of Φ. Let $\Gamma \in \mathbb{R}^{m \times n}$ represented as

$$\Gamma = U \left[\begin{array}{cc} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{array} \right] V^T,$$

where $\Gamma_{11} \in \mathbb{R}^{r \times r}$, $\Gamma_{12} \in \mathbb{R}^{(m-r) \times r}$, $\Gamma_{21} \in \mathbb{R}^{r \times (n-r)}$, $\Gamma_{22} \in \mathbb{R}^{(m-r) \times (n-r)}$, and for $\Gamma_1 := P(\Gamma), \Gamma_2 := P_c(\Gamma)$, the following hold for Γ_1, Γ_2 respectively:

1. For Γ_1, $\mathrm{rank}(\Gamma_1) \le 2r, \Phi\Gamma_2^T = 0, \Phi^T\Gamma_2 = 0$
2. For Γ_2, there is additive relation of the trace norm of (Φ, Γ_2) pair as $||\Phi + \Gamma_2||_* = ||\Phi||_* + ||\Gamma_2||_*$
3. for the matrix pair $(\Phi, \Gamma) \in \mathbb{R}^{m \times n}$, there is $||\Gamma||_* + ||\Phi||_* - ||\Phi + \Gamma||_* \le 2||\Gamma_1||_*$

Definition 2: (Operator Q) Given a matrix $\Gamma \in \mathbb{R}^{m \times n}$, $\mathrm{Q}(\Gamma)$ is defined as

$$\mathrm{Q}(\Gamma) = \Gamma(:, s_i), s_i = \{i | \forall j, \Gamma[j, i] \neq 0\}$$

namely, $\mathrm{Q}(\Gamma)$ is composed of the nonzero columns of Γ, and thus Q is used to extract the nonzero columns of a matrix. Q has the following property.

Lemma 2[1]: Given an arbitrary matrix pair $(\Psi, \Lambda) \in \mathbb{R}^{m \times n}$, there is

$$||\Psi||_{2,1} + ||\Lambda||_{2,1} - ||\Psi + \Lambda||_{2,1} \le ||\mathrm{Q}(\Psi)||_{2,1}$$

We then present Assumption 1, which is the foundation of later theoretical analysis. We will denote our solution W by the pair (U, V), where in Algorithm 1, $W = U = V$, and in Algorithm 2, $W = U + V$.

Assumption 1: For the solution (U, V) pair, and a constant pair p, q satisfying $p \le \min{(T, d)}, q \le T$, there exists a constant pair $(\kappa(p), \tau(q))$ such that

$$\begin{aligned} \kappa(p) &= \min_{U,V \in \mathcal{R}(p,q)} \frac{||L(U+V)||_F}{\sqrt{N}||\mathrm{P}(U)||_*} > 0 \\ \tau(q) &= \min_{U,V \in \mathcal{R}(p,q)} \frac{||L(U+V)||_F}{\sqrt{N}||\mathrm{Q}(V)||_{2,1}} > 0 \end{aligned} \tag{1.1}$$

where the restricted set $\mathcal{R}(p, q)$ is defined as

(1.2)
$$\mathcal{R}(p, q) = \{U, V | 0 < \mathrm{rank}(\mathrm{P}(U) \le p, 0 < |\mathrm{Q}(V)| \le q\}$$

2 Matrix Inversion

To cache the factorization for speeding up computation, we use the matrix inversion lemma stated as follows

(2.3)
$$(P + \rho A^T A)^{-1} = P^{-1} - \rho P^{-1} A^T (I + \rho A P^{-1} A^T)^{-1} A P^{-1}$$

In our computation, as in Algorithm 1, $\left(\rho_1 D^2 + \rho_3 \mathrm{I}\right)^{-1}$ and $\left(\frac{1}{Tn_i} {X_i}^T X_i + (\rho_2 + \rho_3) I\right)^{-1}$ can be computed likewise, and in Algorithm 2, $\left(\rho_1 D^2 + \rho_2 \mathrm{I}\right)^{-1}$ and $\left(\frac{1}{Tn_i} {X_i}^T X_i + \rho_2 I\right)^{-1}$ can be computed likewise.

*School of Computer Science, University of Massachusetts, boliu@cs.umass.edu

†Department of Computer Sciences, University of Rochester, jliu@cs.rochester.edu

‡School of Computer Science, University of Massachusetts, mahadeva@cs.umass.edu

§Department of Computer Science, University of North Carolina at Charlotte, yong.ge@uncc.edu

¶Samsung Research America, San Jose, CA, 95134, doogkong@gmail.com

3 Measurements

nMSE, aMSE, WMSE and WRSE are defined as follows,

$$\begin{aligned}
\text{WMSE} &= \frac{1}{T}\sum_{i=1}^{T}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{i,j}-x_{i,j}W_i)^2 \\
\text{WRSE} &= \frac{1}{N}\sum_{i=1}^{T}\left(n_i\sqrt{\sum_{j=1}^{n_i}(y_{i,j}-x_{i,j}W_i)^2}\right) \\
\text{nMSE} &= \frac{1}{T}\sum_{i=1}^{T}\frac{1}{\text{var}(y_i)n_i}\sum_{j=1}^{n_i}(y_{i,j}-x_{i,j}W_i)^2 \\
a\text{MSE} &= \frac{1}{T}\sum_{i=1}^{T}\frac{1}{||y_i||_2^2 n_i}\sum_{j=1}^{n_i}(y_{i,j}-x_{i,j}W_i)^2
\end{aligned}$$

(3.4)

where for (data, label) pair (X_i, Y_i) of the i-th task, $x_{i,j}$ is the j-th row of X_i, and $y_{i,j}$ is the j-th entry of Y_i.

References

[1] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *KDD*, pages 42–50, 2011.