A common approach in multi-task learning is to encourage the tasks to share a low-dimensional representation by using trace norm regularization. In this paper, we extend this approach by allowing the tasks to partition into different groups, within which trace norm regularization is separately applied. We propose a smooth continuous bilevel optimization framework to simultaneously identify multi-task groups and learn a low-dimensional representation within each group.

2. Groupwise Trace Norm Regularization

**Goal:** Given a dataset \( \{y^{(i)}, X^{(i)}\}_{i=1}^{n} \) assumed organized into \( L \) groups of related tasks \( \{G_1, \ldots, G_L\} \), find a groupwise low-rank linear model \( \hat{W} \) such that

- (linear regression) for every \( t \in \{1, \ldots, T\} \), \( \hat{y}^{(i)} \approx X^{(i)} \hat{W} \)
- (low dimension) for every \( t \in \{1, \ldots, L\} \), the restriction \( \hat{W}_t \) is low rank

**Optimization problem:** Given \( \{y^{(i)}, X^{(i)}\}_{i=1}^{n} \) and a partition \( \mathcal{G} = \{G_1, \ldots, G_L\} \) of the \( T \) tasks in \( L \) groups, find\( \hat{W} \in \arg\min_{W} \sum_{t=1}^{T} \frac{1}{2} \|y^{(i)} - X^{(i)}W\|_{2,1}^2 \) + \frac{1}{2} \|W\|_{2,1} \text{ for some regularization parameter } \lambda > 0 \)

**Issues:** In general, \( \mathcal{G} \) might not be known a priori. Finding \( \mathcal{G} \) through an exhaustive search is a challenging combinatorial problem since there are \( (L^2/L)! \) possible partitions.

3. Proposed Setting

**Parameterization of the Groups:** Let \( \Theta = \{\theta_1, \ldots, \theta_L\} \subset [0,1]^T \) be the hyperparameter matrix encoding at most \( L \) groups, meaning that \( \theta_t = 1 \) if the \( t \)-th task belongs to \( G_t \) and 0 otherwise.

**Data:** Training sets \( \{y^{(i)}, X^{(i)}\}_{i=1}^{n} \) and validation sets \( \{y^{(i)}, X^{(i)}\}_{i=1}^{n} \) each one sample from a groupwise low-rank linear model \( Y = X\Theta + e \), where \( e \sim N(0, \sigma^2) \).

4. Exact Bilevel Problem

**Upper-level Problem:**

minimize \( \|y^{(i)} - X^{(i)}W\|_{2,1}^2 \) subject to \( W \in \mathbb{R}^{p \times q} \)

**Lower-level Problem:**

minimize \( \sum_{t=1}^{T} \frac{1}{2} \|y^{(i)} - X^{(i)}W_t\|_{2,1}^2 + \frac{1}{2} \|W_t\|_{2,1} \) subject to \( W_t \in \mathbb{R}^{p \times q} \), for all \( t \in \{1, \ldots, T\} \)

**Differences:**

- \( L \) is non-smooth (since \( \Theta \) is non-smooth)
- \( U \) is non-smooth (since \( W \) is non-smooth)
- \( \hat{W} = \arg\min_{W} \|y^{(i)} - X^{(i)}W\|_{2,1}^2 \) is not available in closed form

5. Approximate Bilevel Problem

**Upper-level Problem:**

minimize \( \|y^{(i)} - X^{(i)}W\|_{2,1}^2 \) subject to \( W \in \mathbb{R}^{p \times q} \)

**Lower-level Problem:**

minimize \( \sum_{t=1}^{T} \frac{1}{2} \|y^{(i)} - X^{(i)}W_t\|_{2,1}^2 + \frac{1}{2} \|W_t\|_{2,1} \) subject to \( W_t \in \mathbb{R}^{p \times q} \), for all \( t \in \{1, \ldots, T\} \)

6. Algorithmic Solution

We implement a forward-backward algorithm with Bregman distance [1, 2] for solving the dual problem.

**Mapping A:** smooth for specific Legendre function \( \Phi \)

\[
\Phi\left(\sum_{t=1}^{T} f_t(\theta_t)\right) = \min_{\theta} \sum_{t=1}^{T} \Phi\left(f_t(\theta_t)\right) + \lambda \sum_{t=1}^{T} \|\theta_t\|_{2,1}
\]

**Mapping B:** \( W^{(t)}(\theta_t) = \Phi\left(-\sum_{t=1}^{T} f_t(\theta_t)\right) \) smooth for \( W \)

**Uniform Convergence**

Theorem 2: For every \( \theta_{t,0} \) in \( \Theta \), \( \|W^{(t)}(\theta_{t,0}) - \hat{W}_t\|_{2,1}^2 \) converges to \( 0 \) as \( \lambda \to 0 \).

7. Choice of Legendre function

Separable Legendre function Since \( g \) is the sum of \( L \) trace norms, then \( g^* \) is separable and equal to the indicator function of \( B_1^1(\lambda) \), where \( B_1^1(\lambda) \) is the spectral ball of \( R_1^1 \) with radius \( \lambda \). Hence, we look at the function \( \Phi \) separable as well.

\[
\Phi(V) = \sum_{t=1}^{T} \Phi\left(f_t(\theta_t)\right) + \lambda \|\theta_t\|_{2,1}
\]

In order to find a smooth proximity function, we look at a Legendre function which does \( g^* = B_1^1(\lambda) \).

Legendre function acting on the singular values: Given the singular value decomposition \( Y_t = U_t \Sigma_t V_t^\top \) with \( \sigma_t = (\sigma_{t1}, \ldots, \sigma_{tL}) \) we define

\[
\phi(V_t) = \sum_{t=1}^{T} \phi(\sigma_t) \text{ and } \phi^\star_{\lambda}(\sigma_t) = \min_{\sigma_{t,0} \in \mathbb{R}^{+}} \|V_t - U_t \Sigma_t V_t^\top\|_{2,1}^2
\]

8. Synthetic Experiments

**Setting:** \( T = 30 \) tasks arranged in \( L = 3 \) groups of 10 tasks each. For each task, we have \( N = 10 \) noisy observations \( (y^{(i)} - 0,1) \) and \( F = 50 \) features.

9. Real Data Experiments

**Results:** We report the average over multiple splits and the standard deviation in parentheses.

Experimental results indicate the advantage of working with a variable number of groups over standard trace norm regularization (STL, STL2) and previous-sto of-the-art approaches.

Table: Results on benchmark data sets. We report the average over multiple splits and the standard deviation in parentheses.

References


