A Convex Formulation for Learning Shared Structures from Multiple Tasks

Jianhui Chen, Lei Tang, Jun Liu, and Jieping Ye

Department of Computer Science and Engineering
Center for Evolutionary Functional Genomics, The Biodesign Institute
Arizona State University

June 13, 2009
Multi-task learning aims to improve the generalization performance by learning from multiple related tasks.
- Applied in the areas of machine learning, data mining, computer vision, and bioinformatics.

Ando and Zhang (JMLR 05) propose the alternating structure optimization (ASO) algorithm to learn the predictive structure from multiple tasks.

The ASO formulation is non-convex and its algorithm is not guaranteed to find a global optimum.
Main Contribution

- Propose an improved ASO formulation (iASO) using a novel regularizer.
- Convert iASO into a (relaxed) convex formulation, which is not scalable to large data sets.
- Propose a convex alternating structure optimization (cASO) algorithm to efficiently find the globally optimal solution for the convex relaxation.
- Present a theoretical condition under which cASO finds a globally optimal solution to iASO.
Problem Setting

- Given $m$ supervised learning tasks, where the $\ell$-th task is associated with training data

$$\{(x^\ell_1, y^\ell_1), \cdots, (x^\ell_{n_\ell}, y^\ell_{n_\ell})\} \subset \mathbb{R}^d \times \{-1, 1\}, \ell \in \mathbb{N}_m,$$

and a linear predictor denoted as

$$f^\ell(x) = u^\ell^\top x, \; u^\ell \in \mathbb{R}^d.$$

- Assume that the $m$ learning tasks are related using some low dimensional feature space.
We consider the linear predictor in the form of (Ando and Zhang, 2005)

\[ f_\ell(x) = u_\ell^T x = w_\ell^T x + v_\ell^T \Theta x, \quad \Theta \Theta^T = I. \tag{1} \]

- \( \Theta \): the shared structure parameter
- \( u_\ell, w_\ell, v_\ell \): the feature space weight vectors
The proposed improved ASO formulation (iASO) is given by:

$$(F_0) \min_{\{u_\ell, v_\ell\}, \Theta} \sum_{\ell=1}^{m} \left( \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^T x_i^\ell, y_i^\ell) + g_\ell(u_\ell, v_\ell, \Theta) \right),$$

where $L$ is the loss function, and $g_\ell(u_\ell, v_\ell, \Theta)$ is defined as:

$$g_\ell(u_\ell, v_\ell, \Theta) = \alpha \| u_\ell - \Theta^T v_\ell \|^2 + \beta \| u_\ell \|^2. \tag{2}$$

- $\| u_\ell - \Theta^T v_\ell \|^2$: control the task relatedness
- $\| u_\ell \|^2$: control the complexity of the predictor functions

If $\alpha = 0$, iASO reduces to $m$ independent SVMs. If $\beta = 0$, iASO reduces to the ASO formulation.
The objective function in iASO:

$$\sum_{\ell=1}^{m} \left( \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^T x_i^\ell, y_i^\ell) + \alpha \| u_\ell - \Theta^T v_\ell \|_2^2 + \beta \| u_\ell \|_2^2 \right). \quad (3)$$

**Reformulation 1:** The optimal \( \{v_\ell\} \) to iASO is given by \( v_\ell = \Theta u_\ell \). By substitution, Eq. (3) can be written as

$$\sum_{\ell=1}^{m} \left( \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^T x_i^\ell, y_i^\ell) + \alpha u_\ell^T \left( (1 + \frac{\beta}{\alpha}) I - \Theta^T \Theta \right) u_\ell \right). \quad (4)$$
Reformulation 2: Let $\eta = \beta / \alpha$. Following the equality

$$(1 + \eta)I - \Theta^T\Theta = \eta(1 + \eta) \left( \eta I + \Theta^T\Theta \right)^{-1},$$

(5)

Eq. (4) can be further rewritten as

$$\sum_{\ell=1}^{m} \left( \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} L(u_{\ell}^T x_i^\ell, y_i^\ell) + \alpha \eta (1 + \eta) u_{\ell}^T (\eta I + \Theta^T\Theta)^{-1} u_{\ell} \right).$$

(6)
Equivalent Reformulation III

**Reformulation 3:** Let $U = [u_1, \cdots, u_m]$. By substituting the matrices product $\Theta^T \Theta$ using a matrix $M$, iASO can be reformulated as

$$(\mathbf{F}_1) \min_{U, M} \sum_{\ell=1}^m \left( \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_{\ell i}^T x_i, y_i^\ell) \right) + G_1(U, M)$$

subject to $M \in \{ M_e \mid M_e = \Theta^T \Theta, \Theta \Theta^T = I, \Theta \in \mathbb{R}^{h \times d} \}$, where $G_1(U, M)$ is defined as

$$G_1(U, M) = \alpha \eta (1 + \eta) \text{tr} \left( U^T (\eta I + M)^{-1} U \right). \quad (7)$$
The convex hull of the set

\[ \mathcal{M}_e = \left\{ M_e \mid M_e = \Theta^T \Theta, \, \Theta \Theta^T = I, \, \Theta \in \mathbb{R}^{h \times d} \right\} \]  

(8)

can be precisely expressed as the convex set

\[ \mathcal{M}_c = \left\{ M_c \mid \text{tr}(M_c) = h, \, M_c \preceq I, \, M_c \in \mathbb{S}^d_+ \right\}. \]  

(9)

Since \( \mathcal{M}_c \) consists of all convex combinations of the elements in \( \mathcal{M}_e \), \( \mathcal{M}_c \) is the smallest convex set that contains \( \mathcal{M}_e \).
We convert the non-convex problem $F_1$ into a convex formulation as

\[
(F_2) \quad \min_{U,M} \sum_{\ell=1}^{m} \left( \frac{1}{n_{\ell}} \sum_{i=1}^{n_{\ell}} L(u_{\ell}^T x_i^\ell, y_i^\ell) \right) + G_2(U, M)
\]

subject to \( \text{tr}(M) = h, \ M \preceq I, \ M \in \mathbb{S}_+^d, \) \hspace{1cm} (10)

where

\[
G_2(U, M) = \alpha \eta (1 + \eta) \text{tr} \left( U^T (\eta I + M)^{-1} U \right).
\] \hspace{1cm} (11)
Following the Schur complement Lemma, we rewrite $F_2$ as

$$
(F_3) \quad \min_{U,M,\{t_\ell\}} \sum_{\ell=1}^{m} \left( \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^T x_i, y_i^\ell) \right) + \alpha \eta (1 + \eta) \sum_{\ell=1}^{m} t_\ell
$$

subject to

$$
\left( \eta I + M \begin{pmatrix} u_\ell \\ u_\ell^T \end{pmatrix} \begin{pmatrix} u_\ell \\ t_\ell \end{pmatrix} \right) \succeq 0, \ \forall \ell \in \mathbb{N}_m,
$$

$$
\text{tr}(M) = h, \ M \preceq I, \ M \in \mathbb{S}_+^d. \quad (12)
$$

If $L$ is the hinge loss function, $F_3$ is a SDP, which is not scalable to high-dimensional data.
We propose a convex alternating structure optimization (cASO) algorithm to efficiently solve $F_2$, that is, recycling between the following two steps:

- Step 1: Given $M$, optimize $U$
- Step 2: Given $U$, optimize $M$

We can show that cASO finds the globally optimal solution to $F_2$ (Argyriou et al., 2007; Argyriou et al., 2008).
Convex ASO Algorithm II

- Given $M$, $U$ can optimized via the problem:

$$
\min_U \sum_{\ell=1}^{m} \left( \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(\mathbf{u}_\ell^T \mathbf{x}_i, y_i^\ell) + \hat{g}(\mathbf{u}_\ell) \right), \quad (13)
$$

where $\hat{g}(\mathbf{u}_\ell)$ is given by

$$
\hat{g}(\mathbf{u}_\ell) = \alpha \eta (1 + \eta) \text{tr} \left( \mathbf{u}_\ell^T (\eta \mathbf{I} + M)^{-1} \mathbf{u}_\ell \right). \quad (14)
$$

- If $L$ is the hinge loss, the problem in Eq. (13) decouples into $m$ independent quadratic programs (QP).
Convex ASO Algorithm III

Given $U$, $M$ can be optimized via the problem:

$$
\min_{M} \quad \text{tr} \left( U^T (\eta I + M)^{-1} U \right)
$$

subject to \( \text{tr}(M) = h, M \preceq I, M \in \mathbb{S}^d_+ \). \hspace{1cm} (15)

Although Eq. (15) can be recast into an SDP, we propose an efficient approach to find its optimal solution.
Efficient Computation of $M$

- Given $U \in \mathbb{R}^{d \times m}$, the optimal $M$ to Eq. (15) has an analytic form as

$$M = P_1 \Gamma^* P_1^T, \quad \Gamma^* = \text{diag} \left( \gamma_1^*, \cdots, \gamma_q^* \right). \quad (16)$$

- Step 1: Compute $P_1$ via the SVD of $U = P_1 \Sigma P_2^T$, where

$$P_1 \in \mathbb{R}^{d \times d}, \quad \Sigma = \text{diag}(\sigma_1, \cdots, \sigma_q) \in \mathbb{R}^{d \times m}, \quad \text{rank}(U) = q.$$

- Step 2: Compute $\{\gamma_i^*\}_{i=1}^q$ via solving:

$$\min_{\{\gamma_i\}_{i=1}^q} \sum_{i=1}^q \frac{\sigma_i^2}{\eta + \gamma_i}$$

subject to $\sum_{i=1}^q \gamma_i = h$, $0 \leq \gamma_i \leq 1$, $\forall i \in \mathbb{N}_q. \quad (17)$
We show under a theoretical condition a globally optimal solution to iASO ($F_1$) can be obtained from cASO ($F_2$).

Let $(U^*, M^*)$ be the optimal solution to $F_2$. Let $P_1 \in \mathbb{R}^{d \times d}$ and $\{\sigma_i\}_{i=1}^q$ be the left singular vectors and the non-zero singular values of $U^*$, respectively. If $\frac{\sigma_h}{\sigma_{h+1}} \geq 1 + 1/\eta$, the optimal solution to $F_1$ is given by $(U^*, \Theta^*)$, where $\Theta^*$ consist of the first $h$ column of $P_1$. 

![Diagram](attachment:image.png)
Table: Performance comparison of competing algorithms.

<table>
<thead>
<tr>
<th>Data Set (n, d, m)</th>
<th>Recreation (12797, 25095, 18)</th>
<th>Science (6345, 24002, 22)</th>
<th>Social (11914, 32492, 21)</th>
<th>Society (14507, 29189, 21)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Macro F1</td>
<td>SVM</td>
<td>43.01 ± 1.44</td>
<td>41.80 ± 1.45</td>
<td>35.87 ± 0.79</td>
</tr>
<tr>
<td></td>
<td>ASO</td>
<td>43.63 ± 1.29</td>
<td>39.26 ± 0.82</td>
<td>35.29 ± 0.67</td>
</tr>
<tr>
<td></td>
<td>cASO</td>
<td>47.12 ± 0.73</td>
<td>45.46 ± 0.50</td>
<td>39.30 ± 1.28</td>
</tr>
<tr>
<td></td>
<td>cMTFL</td>
<td>46.13 ± 0.58</td>
<td>42.52 ± 0.59</td>
<td>38.94 ± 1.88</td>
</tr>
<tr>
<td>Micro F1</td>
<td>SVM</td>
<td>49.15 ± 2.32</td>
<td>49.27 ± 4.64</td>
<td>63.05 ± 2.45</td>
</tr>
<tr>
<td></td>
<td>ASO</td>
<td>50.68 ± 0.18</td>
<td>49.05 ± 0.57</td>
<td>62.77 ± 3.59</td>
</tr>
<tr>
<td></td>
<td>cASO</td>
<td>53.34 ± 0.90</td>
<td>53.32 ± 0.45</td>
<td>66.04 ± 0.62</td>
</tr>
<tr>
<td></td>
<td>cMTFL</td>
<td>52.52 ± 0.92</td>
<td>50.60 ± 0.76</td>
<td>65.60 ± 0.63</td>
</tr>
</tbody>
</table>

Key Observation:
- cASO outperforms or perform competitively with other competing algorithms.
**Key Observation:**

- A small $\eta$ leads to lower F1, while $\eta \approx 1$ leads to the highest F1.
- cASO requires more computation time for convergence using a small $\eta$, while less computation time is required for a large $\eta$.

*Figure 2. Sensitivity study of $\eta$ on Micro/Macro F1.*
## Experimental Study III

**Table:** Comparison of the optimal objective values of $F_0$ and $F_2$ with different choices of $\eta$.

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>1000</th>
<th>100</th>
<th>10</th>
<th>1</th>
<th>0.1</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 + 1/\eta$</td>
<td>1.001</td>
<td>1.01</td>
<td>1.1</td>
<td>2</td>
<td>11</td>
<td>101</td>
<td>1001</td>
</tr>
<tr>
<td>$\sigma_h/\sigma_{h+1}$</td>
<td>1.23</td>
<td>1.25</td>
<td>1.34</td>
<td>1.75</td>
<td>3.07</td>
<td>13.79</td>
<td>89.49</td>
</tr>
<tr>
<td>OBJ$_{F_0}$</td>
<td>52.78</td>
<td>52.65</td>
<td>51.37</td>
<td>40.73</td>
<td>22.15</td>
<td>5.95</td>
<td>0.69</td>
</tr>
<tr>
<td>OBJ$_{F_2}$</td>
<td>52.78</td>
<td>52.65</td>
<td>51.37</td>
<td>40.71</td>
<td>20.73</td>
<td>4.11</td>
<td>0.41</td>
</tr>
</tbody>
</table>

**Key Observation:**

- We can observe that when $\eta \in \{1000, 100, 10\}$, the condition $\sigma_h/\sigma_{h+1} > 1 + 1/\eta$ is satisfied and hence $\text{OBJ}_{F_0} = \text{OBJ}_{F_2}$; otherwise, we observe $\text{OBJ}_{F_0} > \text{OBJ}_{F_2}$. 
Conclusion and Future Work

- Present $i$ASO for learning a shared feature representation from multiple related tasks.

- Convert $i$ASO into a relaxed convex formulation, and then develop the $c$ASO algorithm to compute its globally optimal solution efficiently.

- Present a theoretical condition, under which $c$ASO can find a globally optimal solution to $i$ASO.

- Plan to compare the $i$ASO formulation with the multi-task learning formulation using the trace-norm regularization.