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

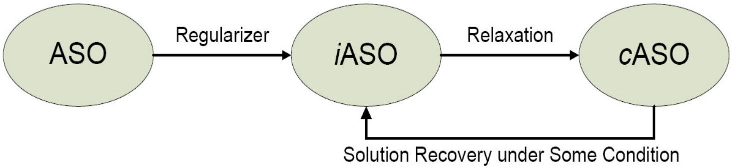
June 13, 2009

Our Motivation

- Multi-task learning aims to improve the generalization performance by learning from multiple related tasks.
    - Applied in the areas of machine learning, data mining, compute 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 (*i*ASO) using a novel regularizer.
- Convert *i*ASO into a (relaxed) convex formulation, which is not scalable to large data sets.
- Propose a convex alternating structure optimization (*c*ASO) algorithm to efficiently find the globally optimal solution for the convex relaxation.
- Present a theoretical condition under which *c*ASO finds a globally optimal solution to *i*ASO.



Problem Setting

- Given m supervised learning tasks, where the ℓ-th tasks is associated with training data

$$\{(x_1^\ell, y_1^\ell), \cdots, (x_{n_\ell}^\ell, y_{n_\ell}^\ell)\} \subset \mathbb{R}^d \times \{-1, 1\}, \ \ell \in \mathbb{N}_m,$$

  and a linear predictor denoted as

$$f_\ell(x) = u_\ell^\mathsf{T} x, \ \ u_\ell \in \mathbb{R}^d.$$

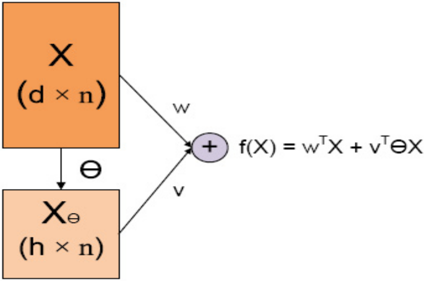
- Assume that the m learning tasks are related using some low dimensional feature space.

*i*ASO Formulation I

- We consider the linear predictor in the form of (Ando and Zhang, 2005)

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

  - Θ: the shared structure parameter
  - u_ℓ, w_ℓ, v_ℓ: the feature space weight vectors



*i*ASO Formulation II

- The proposed improved ASO formulation (*i*ASO) is given by:

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

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

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

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

- If $\alpha = 0$, *i*ASO reduces to m independent SVMs. If $\beta = 0$, *i*ASO reduces to the ASO formulation.

Equivalent Reformulation I

- The objective function in *i*ASO:

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

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

$$\sum_{\ell=1}^{m}\left(\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} L(u_\ell^{\mathsf{T}} x_i^\ell, y_i^\ell) + \alpha u_\ell^{\top}\left((1+\frac{\beta}{\alpha})I - \Theta^{\top}\Theta\right) u_\ell\right). \quad (4)$$

Equivalent Reformulation I

- **Reformulation 2:** Let $\eta = \beta/\alpha$. Following the equality

$$(1+\eta)I - \Theta^\mathsf{T}\Theta = \eta(1+\eta)\left(\eta I + \Theta^\mathsf{T}\Theta\right)^{-1}, \tag{5}$$

Eq. (4) can be further rewritten as

$$\sum_{\ell=1}^{m}\left(\frac{1}{n_\ell}\sum_{i=1}^{n_\ell} L(u_\ell^\mathsf{T} x_i^\ell, y_i^\ell) + \alpha\eta(1+\eta)u_\ell^\top \left(\eta I + \Theta^\mathsf{T}\Theta\right)^{-1} u_\ell\right). \tag{6}$$

Equivalent Reformulation III

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

$$(\mathbf{F_1}) \quad \min_{U,M} \quad \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^\mathsf{T} x_i^\ell, y_i^\ell) \right) + G_1(U, M)$$

$$\text{subject to} \quad M \in \left\{ M_e \mid M_e = \Theta^\mathsf{T}\Theta,\ \Theta\Theta^\mathsf{T} = I,\ \Theta \in \mathbb{R}^{h \times d} \right\},$$

where $G_1(U, M)$ is defined as

$$G_1(U, M) = \alpha\ \eta\ (1 + \eta)\ \mathrm{tr}\left(U^\mathsf{T} (\eta I + M)^{-1} U \right). \qquad (7)$$

Convex Relaxation I

- The convex hull of the set

$$\mathcal{M}_e = \left\{ M_e \,|\, M_e = \Theta^\mathsf{T}\Theta,\ \Theta\Theta^\mathsf{T} = I,\ \Theta \in \mathbb{R}^{h\times d} \right\} \tag{8}$$

can be precisely expressed as the convex set

$$\mathcal{M}_c = \left\{ M_c \,|\, \mathrm{tr}(M_c) = h,\ M_c \preceq I,\ M_c \in \mathbb{S}_+^d \right\}. \tag{9}$$

- 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 $\mathcal{M}_e$.

Convex Relaxation II

- We convert the non-convex problem F_1 into a convex formulation as

$$
\begin{aligned}
(\mathbf{F_2}) \quad & \min_{U,M} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^{\mathrm{T}} x_i^\ell, y_i^\ell) \right) + G_2(U, M) \\
\text{subject to} \quad & \mathrm{tr}(M) = h, \;\; M \preceq I, \;\; M \in \mathbb{S}_+^d, \qquad (10)
\end{aligned}
$$

where

$$
G_2(U, M) = \alpha \, \eta \, (1 + \eta) \, \mathrm{tr} \left(U^{\mathrm{T}} \left(\eta I + M \right)^{-1} U \right). \qquad (11)
$$

SDP Formulation

- Following the Schur complement Lemma, we rewrite F_2 as

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

$$\text{subject to} \quad \begin{pmatrix} \eta I + M & u_\ell \\ u_\ell^{\mathsf{T}} & t_\ell \end{pmatrix} \succeq 0, \ \forall \ell \in \mathbb{N}_m,$$

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

- If L is the hinge loss function, F_3 is a SDP, which is not scalable to high-dimensional data.

Convex ASO Algorithm I

- 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:

$$\min_{U} \quad \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(u_\ell^\mathrm{T} x_i^\ell, y_i^\ell) + \hat{g}(u_\ell) \right), \qquad (13)$$

where $\hat{g}(u_\ell)$ is given by

$$\hat{g}(u_\ell) = \alpha\, \eta\, (1+\eta)\, \mathrm{tr} \left(u_\ell^\mathrm{T} \left(\eta I + M\right)^{-1} u_\ell \right). \qquad (14)$$

- 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:

$$\begin{aligned} \min_{M} \quad & \operatorname{tr}\left(U^{\mathrm{T}} (\eta I + M)^{-1} U\right) \\ \text{subject to} \quad & \operatorname{tr}(M) = h, M \preceq I, M \in \mathbb{S}_{+}^{d}. \end{aligned} \tag{15}$$

- 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

$$M = P_1 \Gamma^* P_1^\top, \ \Gamma^* = \mathrm{diag}\left(\gamma_1^*, \cdots, \gamma_q^*\right). \tag{16}$$

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

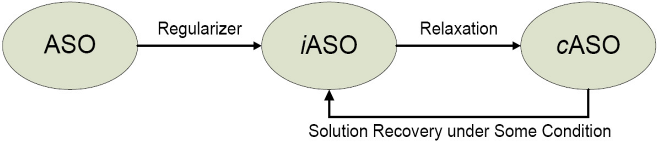
$$P_1 \in \mathbb{R}^{d \times d}, \Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_q) \in \mathbb{R}^{d \times m}, \mathrm{rank}(U) = q.$$

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

$$\begin{aligned} \min_{\{\gamma_i\}_{i=1}^q} \quad & \textstyle\sum_{i=1}^q \frac{\sigma_i^2}{\eta + \gamma_i} \\ \text{subject to} \quad & \textstyle\sum_{i=1}^q \gamma_i = h, \ 0 \le \gamma_i \le 1, \ \forall i \in \mathbb{N}_q. \end{aligned} \tag{17}$$

Computation of an Optimal Solution to *i*ASO

- We show under a theoretical condition a globally optimal solution to *i*ASO (F_1) can be obtained from *c*ASO (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 $\sigma_h / \sigma_{h+1} \geq 1 + 1/\eta$, the optimal solution to F_1 is given by (U^*, Θ^*), where Θ^* consist of the first h column of P_1.



Experimental Study I

Table: Performance comparison of competing algorithms.

Data Set (n, d, m)		Recreation (12797, 25095, 18)	Science (6345, 24002, 22)	Social (11914, 32492, 21)	Society (14507, 29189, 21)
Macro F1	SVM	43.01 ± 1.44	41.80 ± 1.45	35.87 ± 0.79	30.68 ± 0.94
	ASO	43.63 ± 1.29	39.26 ± 0.82	35.29 ± 0.67	29.42 ± 0.30
	*c*ASO	$\mathbf{47.12 \pm 0.73}$	$\mathbf{45.46 \pm 0.50}$	$\mathbf{39.30 \pm 1.28}$	$\mathbf{34.84 \pm 1.05}$
	*c*MTFL	46.13 ± 0.58	42.52 ± 0.59	38.94 ± 1.88	33.79 ± 1.43
Micro F1	SVM	49.15 ± 2.32	49.27 ± 4.64	63.05 ± 2.45	40.07 ± 3.42
	ASO	50.68 ± 0.18	49.05 ± 0.57	62.77 ± 3.59	46.13 ± 2.33
	*c*ASO	$\mathbf{53.34 \pm 0.90}$	$\mathbf{53.32 \pm 0.45}$	$\mathbf{66.04 \pm 0.62}$	$\mathbf{49.27 \pm 0.55}$
	*c*MTFL	52.52 ± 0.92	50.60 ± 0.76	65.60 ± 0.63	46.46 ± 0.87

Key Observation:

- *c*ASO outperforms or perform competitively with other competing algorithms.

Experimental Study II

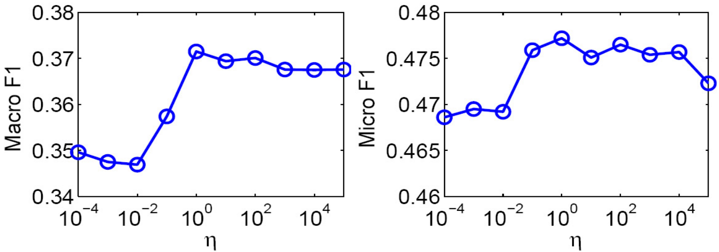


Figure 2. Sensitivity study of η on Micro/Micro F1.

Key Observation:

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

Experimental Study III

Table: Comparison of the optimal objective values of F_0 and F_2 with different choices of η.

η	1000	100	10	1	0.1	0.01	0.001
$1 + 1/\eta$	1.001	1.01	1.1	2	11	101	1001
σ_h/σ_{h+1}	1.23	1.25	1.34	1.75	3.07	13.79	89.49
OBJ_{F_0}	52.78	52.65	51.37	40.73	22.15	5.95	0.69
OBJ_{F_2}	52.78	52.65	51.37	40.71	20.73	4.11	0.41

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.