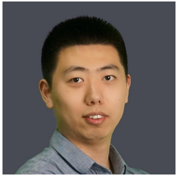
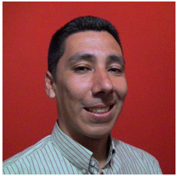
The Sample Complexity of Meta Sparse Regression


Zhanyu Wang
Purdue University


Jean Honorio
Purdue University

Background

- Few-shot learning relates to solving a task with only few training samples, e.g., training a multi-class classifier with only one image for each class in the training dataset.
- Meta-learning tackles this problem by gathering similar tasks instead of more samples from the same task.
- We propose one setting, meta sparse regression, and provide theoretical guarantee on few-shot learning under this setting using our proposed method. Our proof uses Primal-Dual Witness scheme[1].

[1] Martin J Wainwright. "Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso)". In: *IEEE transactions on information theory* 55.5 (2009), pp. 2183–2202.

Problem Setting

The dataset contains samples from multiple tasks, and is generated as follows:

$$y_{t_i,j} = X_{t_i,j}^T(\mathbf{w}^* + \Delta_{t_i}^*) + \epsilon_{t_i,j}, \quad i = 1, \cdots, T+1; j = 1, \cdots, l \tag{1}$$

where, t_i indicates the i-th task (solving t_{T+1} is our final goal), $\mathbf{w}^* \in \mathbb{R}^p$ is a constant across all tasks, and $\Delta_{t_i}^* \in \mathbb{R}^p$ is the individual parameter for each task.

Few-shot learning is the setting with small sample size l and large number of tasks T.

Our key assumptions: ($SG_p(\cdot)$ is a sub-Gaussian distribution of p-dimensional random vectors.)

1. $\Delta_{t_i}^* \sim SG_p(\sigma_\Delta^2)$. $\epsilon_{t_i,j} \sim SG_1(\sigma_\epsilon^2)$. $X_{t_i,j} \sim SG_p(\sigma_x^2)$. They are mutually independent and can come from different distributions for different tasks.
2. $S_i = Supp(\mathbf{w}^* + \Delta_{t_i}^*)$, and $S = Supp(\mathbf{w}^*)$. $S_i \subseteq S$, $|S| = k$.
3. The mixture distribution of covariates of all tasks satisfies the mutual incoherence condition, i.e., $|||\Sigma_{S^c,S}(\Sigma_{S,S})^{-1}|||_\infty \leq 1-\gamma, \gamma \in (0,1]$.
4. $\mathbf{X}_{t_i,S}$ and $\Delta_{t_i,S}^*$ are rotation invariant (only used for matching minimax optimal rates.)

Our Method

First, we determine the common support S over the prior tasks $\{t_i | i = 1, 2, \cdots, T\}$ by the support of $\hat{\mathbf{w}}$ formally introduced below, i.e., $\hat{S} = Supp(\hat{\mathbf{w}})$, where

$$\begin{aligned} \ell(\mathbf{w}) &= \frac{1}{2TI} \sum_{i=1}^{T} \sum_{j=1}^{I} \|y_{t_i,j} - X_{t_i,j}^T \mathbf{w}\|_2^2, \\ \hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \{\ell(\mathbf{w}) + \lambda \|\mathbf{w}\|_1\} \end{aligned} \tag{2}$$

Second, we use the support $\hat{S}$ as a constraint for recovering the parameters of the novel task t_{T+1}. That is

$$\begin{aligned} \ell_{T+1}(\mathbf{w}) &= \frac{1}{2I} \sum_{j=1}^{I} \|y_{t_{T+1},j} - X_{t_{T+1},j}^T \mathbf{w}\|_2^2, \\ \hat{\mathbf{w}}_{T+1} &= \arg\min_{\mathbf{w}, Supp(\mathbf{w}) \subseteq \hat{S}} \{\ell_{T+1}(\mathbf{w}) + \lambda_{T+1} \|\mathbf{w}\|_1\} \end{aligned} \tag{3}$$

Main results

Theorem (recovering the common support S)

Let $\hat{\mathbf{w}}$ be the solution of the optimization problem (2). Under assumptions **A1, A2, A3**, *if*

$$\lambda \in \Omega\left(\max\left(\sigma_\epsilon\sigma_x,\ \max(\sigma_x, \sigma_x^2)\sigma_\Delta\sqrt{k}\right)\sqrt{\frac{\log(p-k)}{Tl}}\right)$$

and $T \in \Omega\left(k\log(p-k)/l\right)$, with probability greater than $1 - c_1\exp(-c_2\log(p-k))$, we have that

1. *the support of $\hat{\mathbf{w}}$ is contained within S (i.e., $S(\hat{\mathbf{w}}) \subseteq S$);*
2. $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty \leq \begin{cases} c_3\sqrt{k}\lambda & \textit{without assumption } \mathbf{A4} \\ c_3\lambda & \textit{with assumption } \mathbf{A4} \end{cases}$

where c_1, c_2, c_3 are constants.

If $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty \in O(1)$, we have $S = S(\hat{\mathbf{w}})$ since $S \subseteq S(\hat{\mathbf{w}})$.

Main results

Theorem (the lower bound of sample complexity)

Let $\Theta := \{\theta = (\mathbf{w}, \Delta_{t_{T+1}}) | \mathbf{w} \in \{0,1\}^p,\ \|\mathbf{w}\|_0 = k,\ \Delta_{t_i} \in \{1,-1\}^p,\ Supp(\Delta_{t_i}) \subseteq Supp(\mathbf{w}),\ \|\mathbf{w} + \Delta_{t_i}\|_0 = k_i\}$*. Furthermore, assume that* $\theta^* = (\mathbf{w}^*, \Delta^*_{t_{T+1}})$ *is chosen uniformly at random from* Θ*. We have:*

$$\mathbb{P}[\hat{\theta} \neq \theta^*] \geq 1 - \frac{\log 2 + c_1'' \cdot Tl + c_2'' \cdot l_{T+1}}{\log |\Theta|}$$

where c_1'', c_2'' *are constants.*

Here $|\Theta| = \Omega\left(\binom{p}{k}\binom{k}{k_{T+1}}\right) = \Omega(p^k k^{k_{T+1}})$.

Therefore, if $T \in o(k \log p/l)$ and $l_{T+1} \in o(k_{T+1} \log k)$, then any algorithm will fail to recover the true parameter very likely.

Comparison on rates of sample size per task l

Table 1: Comparison among Our Method versus Different Multi-task Learning Methods.

Method	**Rate of l for support recovery**
(Ours) ℓ_1	$O(1)$ (only to recover the common support)
([2]) $\ell_1 + \ell_{1,\infty}$	$O(\max(k \log(pT), kT(T + \log p)))$
([3]) $\ell_{1,\infty}$	$O(\max(k, T)(T + \log p))$
([4]) $\ell_{1,2}$	$O(\max(k \log(p - k), T \log k))$

[2] Ali Jalali et al. "A dirty model for multi-task learning". In: *Advances in neural information processing systems*. 2010, pp. 964–972.

[3] Sahand N Negahban and Martin J Wainwright. "Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block ℓ_1/ℓ_∞-Regularization". In: *IEEE Transactions on Information Theory* 57.6 (2011), pp. 3841–3863.

[4] Guillaume Obozinski, Martin J Wainwright, Michael I Jordan, et al. "Support union recovery in high-dimensional multivariate regression". In: *The Annals of Statistics* 39.1 (2011), pp. 1–47.

Simulations

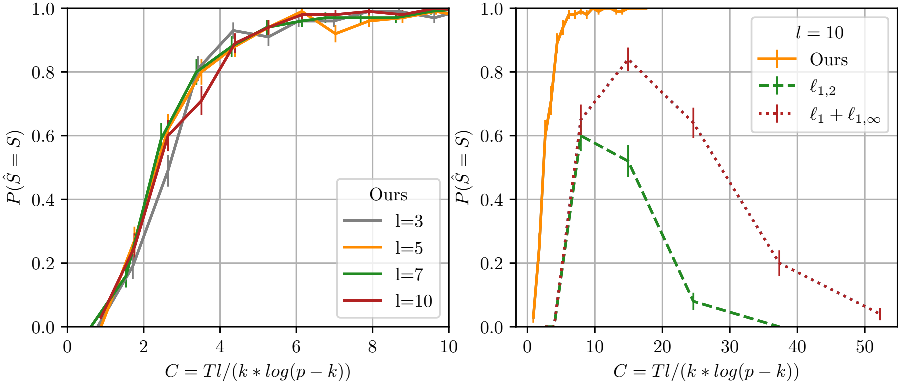


Figure 1: Simulations for Theorem 1 on the Probability of Exact Support Recovery with $\lambda = \sqrt{k \log(p-k)/(Tl)}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of l. We can see that $P(\hat{S} = S)$ depends on C but not on l, i.e., few-shot learning setting. **Right:** Our method outperforms multi-task methods especially when T is large ($\hat{S} := \bigcup_{i=1}^{T} \hat{S}_i$.)

Real-world experiments

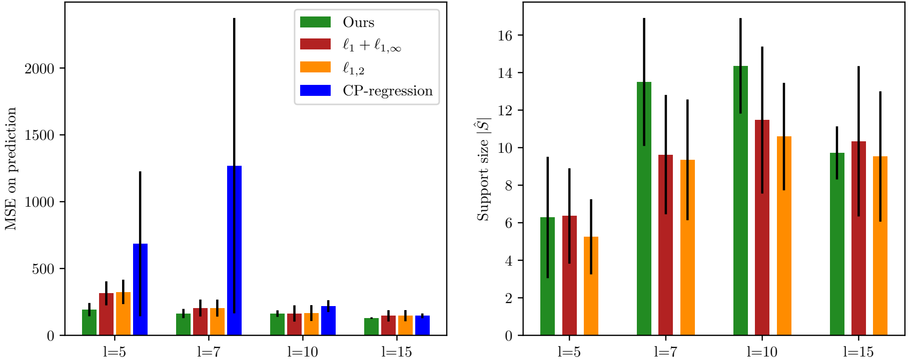


Figure 2: Results on the Single-Cell Gene Expression Dataset. **Left:** The mean square error (MSE) of prediction on the new task. **Right:** The size of the estimated common support $\hat{S}$. When l is small, our method has lower MSE and comparable $|\hat{S}|$ to others, which suggests that our $\hat{S}$ is more accurate.