The Sample Complexity of Meta Sparse Regression

Zhanyu Wang, Jean Honorio
wang4094@purdue.edu, jhonorio@purdue.edu
Purdue University, West Lafayette - IN, 47907.

Introduction

- Few-shot learning relates to solving a task with only few training samples. Often, there is no enough information in the data to solve the task by itself. Meta-learning tackles this problem by gathering similar tasks instead of more samples from the same task.
- There is a lack of theoretical understanding for how it is possible to learn and use the common knowledge among these similar tasks to reduce the complexity of learning a new task.
- **Contribution:** We propose one setting, meta sparse regression which contains T tasks, and provide a theoretical guarantee on few-shot learning under this setting. Let p be the dimension of the parameter vector for these tasks, k be the size of the support, and l be the sample size of each task. We propose an algorithm and show that $T \in O((k\log p)/l)$ tasks are sufficient in order to recover the common support of all tasks. We also prove that our rates are minimax optimal.

Assumptions

The dataset containing samples from multiple tasks is generated as follows:

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

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.

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

(A1) $\Delta_{t_i}^* \sim SG_p(\sigma_\Delta^2)$. $\epsilon_{t_i,j} \sim SG_1(\sigma_\epsilon^2)$. $X_{t_i,j} \sim SG_p(\sigma_x^2)$. They are mutually independent.

(A2) $S_i = Supp(\mathbf{w}^* + \Delta_{t_i}^*)$, and $S = Supp(\mathbf{w}^*)$. $S_i \subseteq S$, $|S| = k$.

(A3) The mixture distribution of covariates of all tasks has the second moment matrix Σ satisfying the mutual incoherence condition, i.e., $|||\Sigma_{S^c,S}(\Sigma_{S,S})^{-1}|||_\infty \le 1-\gamma, \gamma \in (0,1]$. Also, $|||\Sigma_{S,S}^{-1/2}|||_\infty^2 \le c_1$ and $\lambda_{\min}(\Sigma_{S,S}) \ge c_2$.

(A4) $\mathbf{X}_{t_i,S}$ and $\Delta_{t_i,S}^*$ are rotation invariant.

For the assumption **A1**, the distributions for different i, j can be different as long as they are all sub-Gaussian.

For the assumption **A2**, it is possible that $S_i \neq S$ as the sub-Gaussian distribution of $\Delta_{t_i}^*$ on the m-th entry can be a mixture of some sub-Gaussian distributions and a Dirac distribution $\delta_{-\mathbf{w}_m^*}$ that can cancel out the m-th entry in $\mathbf{w}^*$.

Assumption **A4** is only used for getting a tighter bound to match the minimax rate. Gaussian distribution naturally satisfies **A4**.

Our method and main results

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

$$\ell(\mathbf{w}) = \frac{1}{2Tl}\sum_{i=1}^{T}\sum_{j=1}^{l}\|y_{t_i,j} - X_{t_i,j}^T\mathbf{w}\|_2^2, \quad \hat{\mathbf{w}} = \arg\min_{\mathbf{w}}\{\ell(\mathbf{w}) + \lambda\|\mathbf{w}\|_1\} \quad (2)$$

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

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

Theorem 1 guarantees recovering the common support S.

Theorem 1 *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 $T \in \Omega(k\log(p-k)/l)$, 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. $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty \le \begin{cases} c_3\sqrt{k}\lambda & \text{without assumption } \mathbf{A4} \\ c_3\lambda & \text{with assumption } \mathbf{A4}\end{cases}$

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 $S \subseteq S(\hat{\mathbf{w}})$.

Theorem 2 guarantees recovering the novel task based on $\hat{S}$.

Theorem 2 *Let $\hat{\mathbf{w}}_{T+1}$ be the solution of the optimization problem (3). Under assumptions* ***A1, A2, A3***, *with the support $\hat{S}$ recovered from Theorem 1, if $k' := k_{T+1}$, $\mathbf{w}_{T+1}^* := \mathbf{w}^* + \Delta_{t_{T+1}}^*$, $\lambda' := \lambda_{T+1} \in \Theta\left(\sigma_\epsilon\sigma_x\sqrt{\log(k-k')/l}\right)$ and $l \in \Omega(k'\log(k-k'))$, with probability greater than $1 - c_1'\exp(-c_2'\log(k-k'))$, we have that*

1. the support of $\hat{\mathbf{w}}_{T+1}$ is contained within S_{T+1} (i.e., $S(\hat{\mathbf{w}}_{T+1}) \subseteq S_{T+1} \subseteq S$);

2. $\|\hat{\mathbf{w}}_{T+1} - \mathbf{w}_{T+1}^*\|_\infty \le \begin{cases} c_3'\sqrt{k'}\lambda' & \text{without } \mathbf{A4} \\ c_3'\lambda' & \text{with } \mathbf{A4}\end{cases}$

where c_1', c_2', c_3' are constants. If $\|\hat{\mathbf{w}}_{T+1} - \mathbf{w}_{T+1}^*\|_\infty \in O(1)$, we have $S_{T+1} = S(\hat{\mathbf{w}}_{T+1})$ since $S_{T+1} \subseteq S(\hat{\mathbf{w}}_{T+1})$.

Theorem 3 provides the lower bound of sample complexity for solving both the meta task and the novel task.

Theorem 3 *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:*

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

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.

Table 1: Comparison on Rates of l for Our Meta Sparse Regression Method versus Different Multi-task Learning Methods.

Model		**Rate of l for support recovery**
ℓ_1	Ours	$O(1)$ (only to recover the common support)
$\ell_1 + \ell_{1,\infty}$	Jalali et al. (2010)	$O(\max(k\log(pT), kT(T + \log p)))$
$\ell_{1,\infty}$	Negahban and Wainwright (2011)	$O(\max(k, T)(T + \log p))$
$\ell_{1,2}$	Obozinski et al. (2011)	$O(\max(k\log(p-k), T\log k))$

Simulations

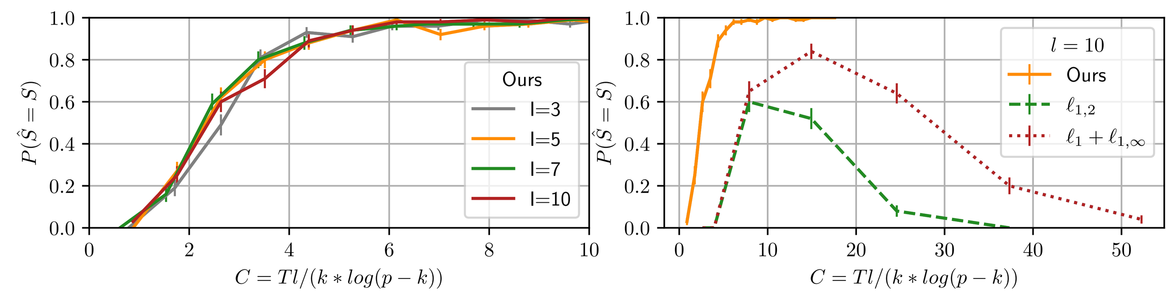

Figure 1: Simulations for Theorem 1 on the Probability of Exact Support Recovery with $\lambda = \sqrt{k\log(p-k)/(Tl)}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of l. We can see that $P(\hat{S} = S)$ depends on C but not on l, i.e., few-shot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

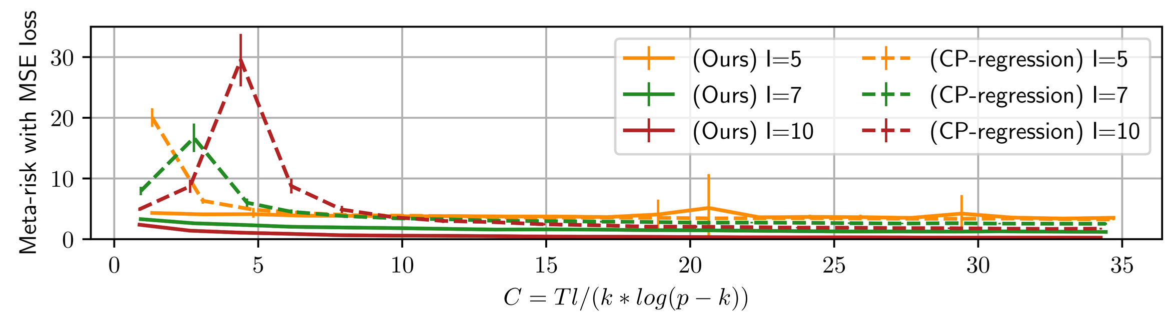

Figure 2: Comparison between our method and a meta learning method, CP-Regression (Maurer, 2005), under various settings of l. The y-axis is the expected MSE of prediction on the novel task. Our method is better.

Real-world experiments

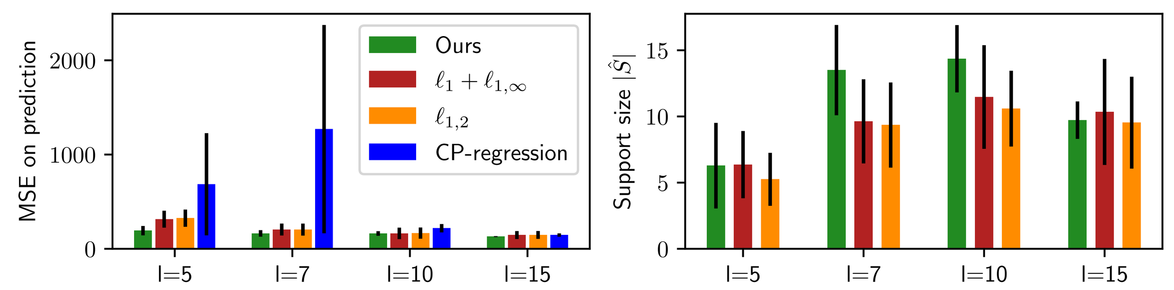

Figure 3: Results on the Single-Cell Gene Expression Dataset. **Left:** The mean square error (MSE) of prediction on the new task. **Right:** The size of the estimated common support $\hat{S}$. When l is small, our method has lower MSE and comparable $|\hat{S}|$ to others, which suggests that our $\hat{S}$ is more accurate.

References

Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in neural information processing systems*, pages 964–972, 2010.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6 (Jun):967–994, 2005.

Sahand N Negahban and Martin J Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Transactions on Information Theory*, 57 (6):3841–3863, 2011.

Guillaume Obozinski, Martin J Wainwright, Michael I Jordan, et al. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.