The Sample Complexity of Meta Sparse Regression

Zhanyu Wang
wang4094@purdue.edu
Purdue University

Jean Honorio
jhonorio@purdue.edu
Purdue University

Abstract

This paper addresses the meta-learning problem in sparse linear regression with infinite tasks. We assume that the learner can access several similar tasks. The goal of the learner is to transfer knowledge from the prior tasks to a similar but novel task. For p parameters, size of the support set k, and l samples per task, we show that $T \in O((k \log p)/l)$ tasks are sufficient in order to recover the common support of all tasks. With the recovered support, we can greatly reduce the sample complexity for estimating the parameter of the novel task, i.e., $l \in O(1)$ with respect to T and p. We also prove that our rates are minimax optimal. A key difference between meta-learning and the classical multi-task learning, is that meta-learning focuses only on the recovery of the parameters of the novel task, while multi-task learning estimates the parameter of all tasks, which requires l to grow with T. Instead, our efficient meta-learning estimator allows for l to be constant with respect to T (i.e., few-shot learning).

1 INTRODUCTION

Current machine learning algorithms have shown great flexibility and representational power. On the downside, in order to obtain good generalization, a large amount of data is required for training. Unfortunately, in some scenarios, the cost of data collection is high. Thus, an inevitable question is how to train a model in the presence of few training samples. This is also called **Few-Shot Learning** (Wang et al., 2020). Indeed, there might not be much information about an underlying task when only few examples are available. A way to tackle this difficulty is **Meta-Learning** (Vanschoren, 2019): we gather many similar tasks instead of several examples in one task, and use the data from different tasks to train a model that can generalize well in the similar tasks. This hopefully also guarantees a good performance of the model for a novel task, even when only few examples are available for the new task. In this sense, the model can rapidly adapt to the novel task with prior knowledge extracted from other similar tasks.

Several meta-learning algorithms have been proposed for the particular model class of neural networks (Vinyals et al., 2016; Ravi and Larochelle, 2016; Finn et al., 2017; Snell et al., 2017). The aforementioned works are of experimental nature and unfortunately, there is a lack of theoretical understanding for the success of meta-learning given different tasks with only few samples for each task. For example, in few shot learning, the case of 5-way 1-shot classification requires the model to learn to classify images from 5 classes with only one example shown for each class. In this case, the model should be able to identify useful features (among a very large learned feature set) in the 5 examples instead of building the features from scratch.

There have been some efforts on building the theoretical foundation of meta-learning. Maurer (2005) gave a general method to prove generalization error bounds based on algorithmic stability. Finn et al. (2019) showed a regret bound of $O(\log T)$, and Fallah et al. (2020) showed a $O(1/\epsilon^2)$ convergence rate to an ϵ-first order stationary point. A natural question is how we can have a theoretical understanding of the meta-learning problem for any algorithm, i.e., the lower bound of the sample complexity of the problem. The upper and lower bounds of sample complexity is commonly analyzed in simple but well-defined statistical learning problems. Since we are learning a novel task with few samples, meta-learning falls in the same regime than sparse regression with large number of covariates p and a small sample size l, which is usually solved by ℓ_1 regularized (sparse) linear regression such as LASSO,

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

albeit for a single task. Even for a sample efficient method like LASSO, we still need the sample size l to be of order $\Omega(k \log p)$ to achieve correct support recovery, where k is the number of non-zero coefficients among the p coefficients. The $l \in \Theta(k \log p)$ rate has been proved to be optimal (Wainwright, 2009). If we consider meta-learning, we may be able to bring prior information from similar tasks to reduce the sample complexity of LASSO. In this respect, researchers have considered the multi-task problem, which assumes similarity among different tasks, e.g., tasks share a common support. Then, one learns for all tasks at once. While it seems that considering many similar tasks together can bring information to each single task, the noise or error is also introduced. In the results from previous papers, e.g., (Jalali et al., 2010; Obozinski et al., 2011; Negahban and Wainwright, 2011), in order to achieve good performance on all T tasks, one needs the number of samples l to scale with the number of tasks T. (See Table 1. Details can be found in appendix Section A.) More specifically, one requires $l \in \Omega(T)$ or $l \in \Omega(\log T)$ for each task, which is not useful in the regime where $l \in O(1)$ with respect to T. Results from other papers, e.g., (Lounici et al., 2009; Ollier and Viallon, 2017), only apply to deterministic (non-random) covariates.

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)	$O(\max(k \log(p-k), T \log k))$

Our contribution in this paper is as follows. First, we proposed a meta-sparse regression problem and a corresponding generative model that are amenable to solid statistical analysis and also capture the essence of meta-learning. Second, we prove the upper and lower bounds of the sample complexity of this problem, and show that they match in the sense that $T \in O((k \log p)/l)$ and $T \in \Omega((k \log p)/l)$. Here p is the number of coefficients in one task, k is the number of non-zero coefficients among the p coefficients, and l is the sample size of each task. In short, we assume that we have access to possibly an infinite number of tasks from a distribution of tasks, and for each task we only have limited number of samples. Our goal is to first recover the common support of all tasks and then use it for learning a novel task. The take-away message of our paper is that simply by merging all the data from different tasks and solving a ℓ_1 regularized (sparse) regression problem (LASSO), we can achieve the best sample complexity rate for identifying the common support and learning the novel task. The merge-and-solve method seems to be intuitive while its validity is not trivial. To the best of our knowledge, our results are the first to give upper and lower bounds of the sample complexity of meta-learning problems.

2 PRELIMINARY

For any set A, $|A|$ is the cardinality. We let $[p]$ be the set $\{1, 2, \cdots, p\}$. For any vector $X \in \mathbb{R}^p$ and set $S \subseteq [p]$, we let X_i be the ith entry of X, and let $X_S \in \mathbb{R}^{|S|}$ be a vector of the entries in X with indices in S. $Supp(X)$ is the set of indices of non-zero entries in X, i.e., $Supp(X) = \{i | X_i \neq 0\}$. For any matrix $X \in \mathbb{R}^{p \times q}$ and set $S \subseteq [p], S' \subseteq [q]$, we let $X_{i,j}$ be the entry at the ith row and the jth column, and let $X_{S,S'} \in \mathbb{R}^{|S| \times |S'|}$ be the submatrix of X with rows indexed by S and columns indexed by S'.

X is a sub-Gaussian random variable with variance proxy σ^2 if and only if $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\lambda^2 \sigma^2 / 2)$. We denote the latter by $X \in SG(\sigma^2)$. $X \in \mathbb{R}^p$ is a sub-Gaussian random vector with variance proxy σ^2 if and only if $v^T X \in SG(\sigma^2), \forall v \in \mathbb{S}^{p-1}$. We denote the latter by $X \in SG_p(\sigma^2)$. Note that for any $S \subseteq [p]$ and $|S| = k$, if $X \in SG_p(\sigma^2)$, then $X_S \in SG_k(\sigma^2)$.

Let $\psi_\alpha(x) := \exp(x^\alpha) - 1$. For any random variable X and $\alpha > 0$, the ψ_α-Orlicz norm of X is defined as

$$\|X\|_{\psi_\alpha} = \inf \{t > 0 : \mathbb{E}[\psi_\alpha(|X|/t)] \leq 1\}$$

We define $\inf \emptyset = \infty$. The above is a generalization of sub-Gaussianity since there is a constant c such that $\|X\|_{\psi_2} \leq c\sigma^2 \ \forall X \in SG(\sigma^2)$.

The notations $O(\cdot), o(\cdot), \Omega(\cdot), \Theta(\cdot)$ are defined as follows: $f(n) \in O(g(n))$ if there exist constants $M > 0$, $n_0 > 0$ such that $|f(n)| \leq Mg(n)$ for all $n \geq n_0$; $f(n) \in o(g(n))$ if for any $\epsilon > 0$, there exist a constant $n_0 > 0$ such that $|f(n)| \leq \epsilon g(n)$ for all $n \geq n_0$; $f(n) \in \Omega(g(n))$ if there exist constants $M > 0$, $n_0 > 0$ such that $f(n) \geq M|g(n)|$ for all $n \geq n_0$; $f(n) \in \Theta(g(n))$ if $f(n) \in O(g(n))$ and $f(n) \in \Omega(g(n))$.

3 METHOD

Here, we present the meta sparse regression problem as well as our ℓ_1 regularized regression method.

3.1 Problem setting

We consider the following meta sparse regression model. The dataset containing samples from multiple tasks $\{(X_{t_i,j}, y_{t_i,j}, t_i)|i = 1, 2, \cdots, T, T+1; j = 1, 2, \cdots, l\}$ is generated as follows:

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

where, t_i indicates the i-th task, $\mathbf{w}^* \in \mathbb{R}^p$ is a constant across all tasks, and $\Delta_{t_i}^* \in \mathbb{R}^p$ is the individual parameter for each task. Note that the tasks $\{t_i|i = 1, 2, \cdots, T\}$ are the related tasks we collect for helping solve the novel task t_{T+1}. Each task contains l training samples. The sample size of task t_{T+1} is denoted by l_{T+1}, which is equal to l in the setting above, but generally it could also be larger than l.

3.2 Assumptions

Our assumptions are as follows.

A1: $\Delta_{t_i}^*$ are mutually independent sub-Gaussian random vectors with mean 0 and variance proxy σ_Δ^2, i.e., $\Delta_{t_i}^* \in SG_p(\sigma_\Delta^2)$.

Note that we do not assume that the entries of $\Delta_{t_i}^*$ are mutually independent. Sub-Gaussianity is a very mild assumption, since the class of sub-Gaussian random variables includes for instance Gaussian random variables, any bounded random variable (e.g., Bernoulli, multinomial, uniform), any random variable with strictly log-concave density, and any finite mixture of sub-Gaussian variables.

We denote the support set of each task t_i as $S_i = Supp(\mathbf{w}^* + \Delta_{t_i}^*)$, and $S = Supp(\mathbf{w}^*)$.

A2: $S_i \subseteq S$ and $|S| = k \ll p$, $k \leq Tl$, $l \in O(k)$, $S_{T+1} \subseteq S$, $|S_{T+1}| = k_{T+1} \leq l$.

This is possible as the sub-Gaussian distribution of $\Delta_{t_i}^*$ on the m-th entry can be a mixture of some other sub-Gaussian distributions and a Dirac distribution $\delta_{-\mathbf{w}_m^*}$ that can cancel out the m-th entry in $\mathbf{w}^*$.

A3: $\epsilon_{t_i,j}$ are mutually independent and follow a sub-Gaussian distribution with mean 0 and variance proxy σ_ϵ^2, i.e., $\epsilon_{t_i,j} \in SG(\sigma_\epsilon^2)$. Sample covariates $X_{t_i,j} \in \mathbb{R}^p$ are mutually independent for any i, j. Each sample is a sub-Gaussian vector with variance proxy σ_x^2, i.e., $X_{t_i,j} \in SG_p(\sigma_x^2)$.

Note that the samples from different tasks can have different distributions.

A4: For every task t_i and any $q \in [p]$, the covariate $X := X_{t_i,j}$ has the second moment matrix Σ_{t_i} with $\|(\Sigma_{t_i})_{S,q}(\Sigma_{t_i})_{q,q}^{-1}\|_2 \in O(\sqrt{k} \cdot \max(1, \sigma_x))$, and the conditional random variable satisfies: $(X_S - (\Sigma_{t_i})_{S,q}(\Sigma_{t_i})_{q,q}^{-1}X_q)|X_q \in SG_k(\sigma_x^2)$.

A5: The mixture distribution of covariates of all tasks has the second moment matrix Σ satisfying the mutual incoherence condition, i.e., $|||\Sigma_{S^c,S}(\Sigma_{S,S})^{-1}|||_\infty \leq 1 - \gamma, \gamma \in (0, 1]$. In addition, there are constants c_1, c_2 such that $|||\Sigma_{S,S}^{-1/2}|||_\infty^2 \leq c_1$ and $\lambda_{\min}(\Sigma_{S,S}) \geq c_2$.

In this paper, we say that $X \in \mathbb{R}^k$ is rotation invariant if for any orthogonal matrix $Q \in \mathbb{R}^{k \times k}$, the distribution of X is the same as the distribution of QX. Note that the Gaussian distribution is rotation invariant while a general sub-Gaussian distribution is not.

A6: (This assumption is only used for getting an additional tighter bound.) $\mathbf{X}_{t_i,S}$ and $\Delta_{t_i,S}^*$ are rotation invariant.

Remark 3.1 (Difference between meta sparse regression and multitask learning)**.** Our setting and analysis focuses on the case that the sample size l of each task is fixed and small, and the number of tasks T goes to infinity, while the number of tasks in multitask learning is usually fixed, or grows with the sample size of each task. The mutual incoherence condition and $S_i \subseteq S = Supp(\mathbf{w}^*)$ are also common and mild assumptions in the multitask learning literature (Jalali et al., 2010; Negahban and Wainwright, 2011; Obozinski et al., 2011). Our problem focuses on recovering only S and S_{T+1} while multitask learning focuses on recovering S_i for all tasks which is much more difficult if the sample size of each task is fixed.

3.3 Our method

In meta sparse regression, our goal is to use the prior T tasks and their corresponding data to recover the common support of all tasks. We then estimate the parameters for the novel task. For the setting we explained above, this is equivalent to recover $(\mathbf{w}^*, \Delta_{t_{T+1}}^*)$.

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

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

Note that we have T tasks in total, and l samples for each task.

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

is

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

We point out that our method makes a proper application of ℓ_1 regularized (sparse) regression, and in that sense is somewhat intuitive. In what follows, we show that this method correctly recovers the common support and the parameter of the novel task. At the same time, our method is minimax optimal, i.e., it achieves the optimal sample complexity rate.

4 MAIN RESULTS

First, we state our result for the recovery of the common support among the prior T tasks.

Theorem 4.1. *Let $\hat{\mathbf{w}}$ be the solution of the optimization problem (2). Under assumptions **A1-A5**, 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 $T \in \Omega\left(\frac{k\log(p-k)}{l}\right)$, 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. $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty \le \begin{cases} c_3\sqrt{k}\lambda & \text{without assumption } \textbf{A6} \\ c_3\lambda & \text{with assumption } \textbf{A6} \end{cases}$

where c_1, c_2, c_3 are constants.

Remark 4.2 (Comparison to the result of traditional single-task LASSO)**.** The scale terms $k\log(p-k)$ in T and $\sqrt{\log(p-k)/Tl}$ in λ are typically encountered in the analysis of the single-task sparse regression or LASSO (Wainwright, 2009). The additional term $\max(\sigma_x, \sigma_x^2)\sigma_\Delta\sqrt{k}$ in λ is due to the difference in the coefficients among tasks. When we have larger k, the difference (noise) among the coefficients also becomes larger, therefore we need a larger λ to suppress the noise and extract the common coefficient (signal). In our technical analysis, the additional term comes from the concentration inequality of a random variable with finite $\psi_{\frac{2}{3}}$-Orlicz norm, which is the main novelty in our proof: bounding the product of three random variables.

Remark 4.3 (Exact support recovery $S(\hat{\mathbf{w}}) = S$)**.** We let $w_{\min}^* = \min_{i\in S}|\mathbf{w}_i^*|$ which represents the signal strength of the common parameter $\mathbf{w}^*$. The first result of Theorem 4.1 is $S(\hat{\mathbf{w}}) \subseteq S$. Now we show that when $w_{\min}^*$ is fixed, we can choose λ and T such that $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty < w_{\min}^*$, i.e., $S \subseteq S(\hat{\mathbf{w}})$. We choose $\lambda = c_4\sqrt{\frac{k\log(p-k)}{Tl}}, T = c_5\frac{k\log(p-k)}{l}$ where c_4, c_5 are constants only depending on $(\sigma_\epsilon, \sigma_x, \sigma_\Delta)$. We also let this choice satisfy the condition on λ and T in Theorem 4.1. Then we choose $T = \max\left(c_5, \frac{2c_3^2c_4^2}{(w_{\min}^*)^2}\right)\frac{k\log(p-k)}{l}$. Now we have $c_3\lambda < w_{\min}^*$. Therefore, under assumption **A6**, we can have the exact support recovery with $T \in O(k\log(p-k)/l)$.

Next, we state our result for the recovery of the parameters of the novel task. Note that in Theorem 4.4, the rate of l is not related to T or p, and if $k - k' \le$ e, we replace $\log(k-k')$ with 1. The proof can be found in appendix Section F.

Theorem 4.4. *Let $\hat{\mathbf{w}}_{T+1}$ be the solution of the optimization problem (3). Under assumptions **A1-A5**, with the support $\hat{S}$ recovered from Theorem 4.1, 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. $\|\hat{\mathbf{w}}_{T+1} - \mathbf{w}_{T+1}^*\|_\infty \le \begin{cases} c_3'\sqrt{k'}\lambda' & \text{without } \textbf{A6} \\ c_3'\lambda' & \text{with } \textbf{A6} \end{cases}$

where c_1', c_2', c_3' are constants.

The theorems above provide an upper bound of the sample complexity, which can be achieved by our method. The lower bound of the sample complexity is an information-theoretic result, and it relies on the construction of a restricted class of parameter vectors. We consider a special case of the setting we previously presented ($\Delta_{t_i}^*$ cancels out some non-zero entries in $\mathbf{w}^*$): all non-zero entries in $\mathbf{w}^*$ are 1, and the distribution of $(\Delta_{t_i}^*)_m$ is $\frac{1}{2}\delta_{-\mathbf{w}_m^*} + \frac{1}{2}\delta_{\mathbf{w}_m^*}$. Therefore all non-zero entries in $\mathbf{w}^* + \Delta_{t_i}^*$ must be 2. We use Θ to denote the set of all possible parameters $\theta^* = (\mathbf{w}^*, \Delta_{t_{T+1}}^*)$. Therefore the number of possible outcomes of the parameters $|\Theta| = \binom{p}{k}\binom{k}{k_{T+1}} \in O(p^k k^{k_{T+1}})$.

If the parameter θ^* is chosen uniformly at random from Θ, for any algorithm estimating this parameter by $\hat{\theta}$, the answer is wrong (i.e., $\hat{\theta} \ne \theta^*$) with probability greater than 1/2 if $(Tl + l_{T+1}) \in o(\log(|\Theta|))$. Here we use l_{T+1} to denote the sample size of task t_{T+1}. This fact is proved by the following theorem.

Theorem 4.5. *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.*

In the appendix Section G, we first prove that the mutual information $\mathbb{I}(\theta^*, S)$ between the true parameter θ^* and the data S is bounded by $c_1'' \cdot Tl + c_2'' \cdot l_{T+1}$. Then we use Fano's inequality (Fano, 1952) and the construction of a restricted class of parameter vectors to prove Theorem 4.5. The use of Fano's inequality and restricted ensembles is customary for information-theoretic lower bounds (Wang et al., 2010; Santhanam and Wainwright, 2012; Tandon et al., 2014).

Note that from Theorem 4.5, we know if $T \in o(\frac{k \log p}{l})$ and $l_{T+1} \in o(k_{T+1} \log k)$, then any algorithm will fail to recover the true parameter very likely. On the other hand, if we have $T \in \Omega(\frac{k \log p}{l})$ and $l_{T+1} \in \Omega(k_{T+1} \log k)$, by Theorem 4.1 and 4.4, we can recover the support of $\mathbf{w}^*$ and Δ^*_{T+1} (by $\mathbf{w}^*_{T+1} - \mathbf{w}^*$). Therefore we claim that our rates of sample complexity is minimax optimal.

5 SKETCH OF THE PROOF OF THEOREM 4.1

We use the primal-dual witness framework (Wainwright, 2009) to prove our results. First we construct the primal-dual candidate; then we show that the construction succeeds with high probability. Here we outline the steps in the proof. (See the supplementary materials for detailed proofs.)

We first introduce some useful notations:

$\mathbf{X}_{t_i} \in \mathbb{R}^{l \times p}$ is the matrix of collocated $X_{t_i,j}$ (covariates of all samples in the i-th task). Similarly, $\mathbf{y}_{t_i} \in \mathbb{R}^l$ and $\epsilon_{t_i} \in \mathbb{R}^l$. $\mathbf{X}_{[T]} \in \mathbb{R}^{Tl \times p}$ is the matrix of collocated $\mathbf{X}_{t_i}$ (covariates of all samples in all tasks). Similarly, $\epsilon_{[T]} \in \mathbb{R}^{Tl}$. $\mathbf{X}_{t_i,S} \in \mathbb{R}^{l \times k}$ is the sub-matrix of $\mathbf{X}_{t_i}$ containing only the rows corresponding to the support of $\mathbf{w}^*$, i.e., S with $|S| = k$. Similarly, $\mathbf{X}_{[T],S} \in \mathbb{R}^{Tl \times k}$, $\Delta^*_{t_i,S} \in \mathbb{R}^k$, and $\mathbf{w}_S \in \mathbb{R}^k$. $\mathbf{A}_{S,S} \in \mathbb{R}^{k \times k}$ is the sub-matrix of $\mathbf{A} \in \mathbb{R}^{p \times p}$ containing only the rows and columns corresponding to the support of $\mathbf{w}^*$.

5.1 Primal-dual witness

Step 1: Prove that the objective function has positive definite Hessian when restricted to the support, i.e., $\forall \mathbf{w}_{S^c} = \mathbf{0}, \forall \mathbf{w}_S \in \mathbb{R}^{|S|},\ [\nabla^2 \ell((\mathbf{w}_S, \mathbf{0}))]_{S,S} \succ 0$

Step 2: Set up a restricted problem:

$$\tilde{\mathbf{w}}_S = \underset{\mathbf{w}_S \in \mathbb{R}^{|S|}}{\arg\min}\, \ell((\mathbf{w}_S, \mathbf{0})) + \lambda \|\mathbf{w}_S\|_1 \tag{4}$$

Step 3: Choose the corresponding dual variable $\tilde{\mathbf{z}}_S$ to fulfill the complementary slackness condition:

$\forall i \in S,\ \tilde{\mathbf{z}}_i = sign(\tilde{\mathbf{w}}_i)$ if $\tilde{\mathbf{w}}_i \neq 0$, otherwise $\tilde{\mathbf{z}}_i \in [-1, +1]$

Step 4: Solve $\tilde{\mathbf{z}}_{S^c}$ to let $(\tilde{\mathbf{w}}, \tilde{\mathbf{z}})$ fulfill the stationarity condition:

$$[\nabla \ell((\tilde{\mathbf{w}}_S, \mathbf{0}))]_S + \lambda \tilde{\mathbf{z}}_S = 0 \tag{5}$$

$$[\nabla \ell((\tilde{\mathbf{w}}_S, \mathbf{0}))]_{S^c} + \lambda \tilde{\mathbf{z}}_{S^c} = 0 \tag{6}$$

Step 5: Verify that the strict dual feasibility condition is fulfilled for $\tilde{\mathbf{z}}_{S^c}$:

$$\|\tilde{\mathbf{z}}_{S^c}\|_\infty < 1$$

In order to prove support recovery, we only need to show that **step 1** and **step 5** hold. The proof of **step 1** being satisfied with high probability under the condition $T \in O(k/l)$ is in appendix Section B. Next we show that **step 5** also holds with high probability.

5.2 Strict dual feasibility condition

We first rewrite (5) as follows:

$$\frac{1}{Tl} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \mathbf{X}_{t_i,S} (\tilde{\mathbf{w}}_S - \mathbf{w}_S^*)$$
$$= -\lambda \tilde{\mathbf{z}}_S + \frac{1}{Tl} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \epsilon_{t_i} + \frac{1}{Tl} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \mathbf{X}_{t_i,S} \Delta^*_{t_i,S}$$

Then we solve for $(\tilde{\mathbf{w}}_S - \mathbf{w}_S^*)$. and plug it in (6). We have

$$\tilde{\mathbf{z}}_{S^c} = \underbrace{\mathbf{X}_{[T],S^c}^T \left\{ \frac{1}{Tl} \mathbf{X}_{[T],S} (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S + \Pi_{\mathbf{X}_{[T],S}^\perp} \left(\frac{\epsilon_{[T]}}{\lambda Tl} \right) \right\}}_{\tilde{\mathbf{z}}_{S^c,1}}$$
$$+ \underbrace{\frac{1}{\lambda Tl} \sum_{i=1}^{T} X_{t_i,S^c}^T X_{t_i,S} \Delta^*_{t_i,S}}_{\tilde{\mathbf{z}}_{S^c,2}}$$
$$- \underbrace{\frac{1}{\lambda (Tl)^2} \mathbf{X}_{[T],S^c}^T \mathbf{X}_{[T],S} (\hat{\Sigma}_{S,S})^{-1} \left(\sum_{i=1}^{T} X_{t_i,S}^T X_{t_i,S} \Delta^*_{t_i,S} \right)}_{\tilde{\mathbf{z}}_{S^c,3}}$$

where $\Pi_{\mathbf{X}_{[T],S}^\perp} := I_{n \times n} - \mathbf{X}_{[T],S} (\mathbf{X}_{[T],S}^T \mathbf{X}_{[T],S})^{-1} \mathbf{X}_{[T],S}^T$ is an orthogonal projection matrix, $\hat{\Sigma}_{S,S} = \frac{1}{Tl} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \mathbf{X}_{t_i,S}$ is the sample covariance matrix, and $\tilde{\mathbf{z}}_S$ is the dual variable chosen at step 3.

One can bound the ℓ_∞ norm of $\tilde{\mathbf{z}}_{S^c,1}$ by the techniques from Wainwright (2009): if $\lambda \in \Omega\left(\sigma_\epsilon \sigma_x \sqrt{\frac{\log(p-k)}{Tl}}\right)$ and $T \in \Omega\left(\frac{k\log(p-k)}{l}\right)$, we have

$$\mathbb{P}[\|\tilde{\mathbf{z}}_{S^c,1}\|_\infty \geq 1-\gamma/2] \leq 2\exp(-c_6 \log(p-k)).$$

where c_6 is a constant. The proof of this result is shown in the appendix Section C.

Note that the remaining two terms $\tilde{\mathbf{z}}_{S^c,2}, \tilde{\mathbf{z}}_{S^c,3}$ containing Δ_{t_i} are new to the meta-learning problem and need to be handled with novel proof techniques.

5.3 Bound on $\|\tilde{\mathbf{z}}_{S^c,2}\|_\infty$

We denote each of the T parts in the sum by $\tilde{\mathbf{z}}_{S^c,2,i}$ and each of the k entries in $\tilde{\mathbf{z}}_{S^c,2,i}$ by $\tilde{\mathbf{z}}_{q,2,i},\ q \in S^c$.

$$\|\tilde{\mathbf{z}}_{S^c,2}\|_\infty = \left\|\frac{1}{\lambda T l}\sum_{i=1}^T \tilde{\mathbf{z}}_{S^c,2,i}\right\|_\infty = \max_{q\in S^c}\left|\frac{1}{\lambda T l}\sum_{i=1}^T \tilde{\mathbf{z}}_{q,2,i}\right|,$$

$$\tilde{\mathbf{z}}_{q,2,i} = X_{t_i,q}^T X_{t_i,S}\Delta_{t_i,S}^*.$$

Here we let q be the index of the covariate, i.e., $X_{t_i,q} \in \mathbb{R}^l$. Since we know that $\Delta_{t_i,S}^*$ are mean 0 and independent of $X_{t_i,S}$ and $X_{t_i,q}$, we have $\mathbb{E}(\tilde{\mathbf{z}}_{q,2,i}) = 0$. In this section, we prove $\|\tilde{\mathbf{z}}_{q,2,i}\|_{\psi_{\frac{2}{3}}} \in O(\sigma_\Delta \max(\sigma_x^2, \sigma_x)\sqrt{kl})$ and use a concentration inequality to bound $\left|\frac{1}{\lambda T l}\sum_{i=1}^T \tilde{\mathbf{z}}_{q,2,i}\right|$.

For clarity of exposition, we hide t_i in the notations below since the analysis holds for all t_i.

Lemma 5.1. *For any $q \in S^c$, with our assumptions on random vectors $X_q \in \mathbb{R}^{l\times 1}, \Delta_S \in \mathbb{R}^{k\times 1}$ and random matrix $X_S \in \mathbb{R}^{l\times k}$, we have $\|X_q^T X_S \Delta_S^*\|_{\psi_{\frac{2}{3}}} \in O(\sigma_\Delta \max(\sigma_x^2, \sigma_x)\sqrt{kl})$.*

Proof. From the definition of Orlizc norm, we need to show that there exists $t = c_7\sigma_\Delta \max(\sigma_x^2, \sigma_x)\sqrt{kl} > 0$ such that

$$\forall i \in [T],\ \mathbb{E}\left[\exp\left(\left(\frac{X_q^T X_S \Delta_S^*}{t}\right)^{\frac{2}{3}}\right)\right] \leq 2$$

where c_7 is a constant. We further let $c_7 = c_{71}c_{72}c_{73}$ and $t_1 = c_{71}\sigma_\Delta$, $t_2 = c_{72}\max(1,\sigma_x)\sqrt{k}$, $t_3 = c_{73}\sigma_x\sqrt{l}$.

For any constant vector $a = (a_1, a_2, \cdots, a_l)$ with ℓ_2-norm being 1, i.e., $a \in \mathbb{S}^{l-1}$, we let $Y_S = \sum_{m=1}^l a_m (X_S)_m$ where $(X_S)_m \in \mathbb{R}^k$ is the mth row of X_S. From the Lemma 5.2 below, we know Y_S is also a sub-Gaussian vector with variance proxy σ_x^2.

Lemma 5.2 (Linear combination of independent sub–Gaussian vectors is a sub-Gaussian vector. Lemma 5.9 in (Vershynin, 2012))**.** *Let $\mathbf{a} = (a_1, a_2, \cdots, a_d) \in \mathbb{R}^d$. If $\{X_1, X_2, \cdots, X_d\}$ are independent random vectors and $X_i \in SG_p(\sigma_x^2)$, then $\sum_{i=1}^d a_i X_i \in SG_p(\|a\|_2 \cdot \sigma_x^2)$.*

Therefore, we define $Y_S^T := \frac{X_q^T}{\|X_q\|_2} X_S$ which is almost a sub-Gaussian vector when conditioning on X_q: By our assumption, the rows of $X_S - X_q(\Sigma_{S,q}(\Sigma_{q,q})^{-1})^T | X_q$ are sub-Gaussian vectors with variance proxy σ_x^2, and these l rows are also *mutually independent* since each of them is determined by only one of the l samples in task t_i. We let $Z_S = Y_S - \delta_q$, $\delta_q := (\Sigma_{S,q}(\Sigma_{q,q})^{-1})^T$, then $Z_S | X_q \in SG_k(\sigma_x^2)$ by the Lemma 5.2.

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\left(\frac{X_q^T X_S \Delta_S^*}{t}\right)^{\frac{2}{3}}\right)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\left(\frac{Y_S^T\Delta_S^* \cdot \|X_q\|_2}{t}\right)^{\frac{2}{3}}\right)\middle| X_q\right]\right] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}\left[\frac{2}{3}\mathbb{E}\left[\exp\left(\frac{Y_S^T\Delta_S^*}{t_1 t_2}\right)\middle| X_q\right]\right] + \frac{1}{3}\mathbb{E}\left[\exp\left(\frac{\|X_q\|_2^2}{t_3^2}\right)\right]
\end{aligned}$$

In (i) we use the definition of Y_S. In (ii) we use Young's inequality. Therefore we only need to show that $\|X_q\|_2$ has finite ψ_2-Orlicz norm, and $Y_S^T\Delta_S^* | X_q$ has finite ψ_1-Orlicz norm. We use the Lemma 5.3 below to choose a constant c_{73} such that $\|\|X_q\|_2\|_{\psi_2} \leq t_3$. The proof of this lemma is in the appendix Section D.

Lemma 5.3 (ℓ_2-norm of sub-Gaussian random vector is a sub-Gaussian random variable)**.** *Assume $X \in SG_d(\sigma_x^2)$, for any constant c_8, there exists a constant c_9 such that $\left\|\|X\|_2 + c_8\sigma_x\sqrt{d}\right\|_{\psi_2} \leq c_9\sigma_x\sqrt{d}$ and $\left\|\|X\|_2 + c_8\max(1,\sigma_x)\sqrt{d}\right\|_{\psi_2} \leq c_9\max(1,\sigma_x)\sqrt{d}$.*

Therefore, there exists c_{73} such that $\|\|X\|_2\|_{\psi_2} \leq c_{73}\sigma_x\sqrt{l}$. Now we consider $Y_S^T\Delta_S^* | X_q$.

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{Y_S^T\Delta_S^*}{t_1 t_2}\right)\middle| X_q\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{Y_S^T}{\|Y_S\|_2}\frac{\Delta_S^*}{t_1}\cdot\frac{\|Y_S\|_2}{t_2}\right)\middle| X_q\right]\right] \\
&\overset{\text{(iii)}}{\leq} \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\exp\left(\left(\frac{Y_S^T}{\|Y_S\|_2}\frac{\Delta_S^*}{t_1}\right)^2\right)\middle| Y_S\right]\middle| X_q\right]\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\|Y_S\|_2^2}{t_2^2}\right)\middle| X_q\right]\right]
\end{aligned}$$

In (iii) we use Young's inequality again. Since $\Delta_S^* \in SG_k(\sigma_\Delta^2)$ and Δ_S^* is *independent* of Y_S, we have $\left\|\frac{Y_S^T}{\|Y_S\|_2}\Delta_S^*\middle| Y_S\right\|_{\psi_2} \leq c\sigma_\Delta$ by the definition of sub-Gaussian vector. We can choose $c_{71} > c$, then the first term is then bounded by 1. For the second term,

by our assumption, there exists a constant c_{10} that $\|\delta_q\|_2 \leq c_{10}\max(1,\sigma_x)\sqrt{k}$. We use Lemma 5.3 again to choose a constant c_{72} such that $\|Y_S\|_{\psi_2} \leq t_2$. □

Remark 5.4 (Product of a random vector, a random matrix and another random vector)**.** In Lemma 5.1, we show that the product $X_q^T X_S \Delta_S^*$ has finite $\psi_{\frac{2}{3}}$-Orlicz norm with rate $O(\sqrt{kl})$. This is a stronger result than simply combining the Lemma 5.7 and 5.8 below since we used two levels of independence among (X_q, X_S, Δ_S^*).

Since we have $\|\tilde{\mathbf{z}}_{q,2,i}\|_{\psi_{\frac{2}{3}}} \leq c_7\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{kl}$, we can use the Lemma 5.5 below to bound $\left|\frac{1}{\lambda Tl}\sum_{i=1}^T \tilde{\mathbf{z}}_{q,2,i}\right|$.

Lemma 5.5 (Concentration inequality of random variables with finite ψ_α-Orlicz norm. Lemma 7 in (Hao et al., 2020))**.** *Suppose $0<\alpha<1$, $X_1, X_2, \cdots, X_n$ are independent random variables satisfying $\|X_i\|_{\psi_\alpha} \leq b$. Then there exists absolute constant $C(\alpha)$ only depending on α such that for any $\mathbf{a} = (a_1, a_2, \cdots, a_n) \in \mathbb{R}^n$ and $0<\delta<1/e^2$,*

$$\left|\sum_{i=1}^n a_i X_i - \mathbb{E}\left[\sum_{i=1}^n a_i X_i\right]\right| \leq C(\alpha)b\|\mathbf{a}\|_2(\log\delta^{-1})^{1/2} + C(\alpha)b\|\mathbf{a}\|_\infty(\log\delta^{-1})^{1/\alpha}$$

with probability at least $1-\delta$.

Here we let $\alpha = \frac{2}{3}$, $n = T$, $a_i = \frac{1}{\lambda Tl}$, $b = c_7\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{kl}$, then

$$\left|\frac{1}{\lambda Tl}\sum_{i=1}^T \tilde{\mathbf{z}}_{q,2,i}\right| \leq \frac{c_{11}b(-\log\delta)^{1/2}}{\lambda Tl}\left(\sqrt{T} - \log\delta\right)$$

with probability at least $1-\delta$ where c_{11} is a constant. We let $T \in \Omega\left(k\log(p-k)/l\right)$, $\log\delta^{-1} \in O\left(\log(p-k)\right)$, $\lambda \in \Omega\left(\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{\frac{k\log(p-k)}{Tl}}\right)$.

By the condition $l \in O(k)$, we have

$$\log\left(\mathbb{P}\left[\left|\frac{1}{\lambda Tl}\sum_{i=1}^T \tilde{\mathbf{z}}_{q,2,i}\right| \geq \frac{\gamma}{4}\right]\right) \in O(-\log(p-k)).$$

Therefore,

$$\mathbb{P}[\|\tilde{\mathbf{z}}_{S^c,2}\|_\infty \geq \gamma/4] \leq \exp(-c_{12}\log(p-k))$$

where c_{12} is a constant.

5.4 Bound on $\|\tilde{\mathbf{z}}_{S^c,3}\|_\infty$

We first transform $\tilde{\mathbf{z}}_{S^c,3}$ to the following form:

$$\begin{aligned}&\tilde{\mathbf{z}}_{S^c,3}\\ &= \frac{1}{\lambda(Tl)^2}\mathbf{X}_{[T],S^c}\mathbf{X}_{[T],S}(\hat{\Sigma}_{S,S})^{-1}\left(\sum_{i=1}^T X_{t_i,S}^T X_{t_i,S}\Delta_{t_i,S}^*\right)\\ &:= \frac{1}{Tl}\mathbf{X}_{[T],S^c}\mathbf{X}_{[T],S}(\hat{\Sigma}_{S,S})^{-1}\zeta_S\end{aligned}$$

where we define

$$\zeta_S := \frac{1}{\lambda Tl}\left(\sum_{i=1}^T X_{t_i,S}^T X_{t_i,S}\Delta_{t_i,S}^*\right).$$

In this section we use a similar technique for bounding $\tilde{\mathbf{z}}_{S^c,2}$ to bound $\|\zeta_S\|_\infty$. With the condition that $\|\zeta_S\|_\infty \leq \frac{\gamma}{2-\gamma}$, we can transform $\tilde{\mathbf{z}}_{S^c,3}$ into the first part in $\tilde{\mathbf{z}}_{S^c,1}$ by replacing ζ_S with $\tilde{\mathbf{z}}_S$, and use the same technique in appendix Section C to obtain the result: If $T \in \Omega\left(k\log(p-k)/l\right)$, we have

$$\mathbb{P}[\|\tilde{\mathbf{z}}_{S^c,3}\|_\infty \geq \gamma/2] \leq c_{13}\exp(-c_{14}\log(p-k))$$

where c_{13}, c_{14} are constants.

To obtain a bound of $\|\zeta_S\|_\infty$, we first need to bound $\|X_{t_i,q}^T X_{t_i,S}\Delta_{t_i,S}^*\|_{\psi_{\frac{2}{3}}}$ for $q \in S$. For clarity of exposition, we hide t_i in the notations below since the analysis holds for all t_i. We use $S^{\backslash q}$ to denote the set $S\backslash\{q\}$.

Lemma 5.6. *For any $q \in S$, with our assumptions on random vectors $X_q \in \mathbb{R}^{l\times 1}, \Delta_S \in \mathbb{R}^{k\times 1}$ and random matrix $X_S \in \mathbb{R}^{l\times k}$, we have $\|X_q^T X_S \Delta_S^*\|_{\psi_{\frac{2}{3}}} \in O(\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{kl})$.*

Proof. We first break $X_q^T X_S \Delta_S^*$ into two parts:

$$X_q^T X_S \Delta_S^* = \|X_q\|_2^2\Delta_q^* + X_q^T X_{S^{\backslash q}}\Delta_{S^{\backslash q}}^*.$$

The second part is similar to $X_{q'}^T X_S \Delta_S^*$ with $q' \in S^c$ since $q \notin S^{\backslash q}$, therefore we have $\|X_q^T X_{S^{\backslash q}}\Delta_{S^{\backslash q}}^*\|_{\psi_{\frac{2}{3}}} \in O(\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{(k-1)l})$ by the Lemma 5.1.

For the first part, we know $\|\|X_q\|_2\|_{\psi_2} \in O(\sigma_x\sqrt{l})$ and $\|\Delta_q^*\|_{\psi_2} \in O(\sigma_\Delta)$. Therefore we have $\|\|X_q\|_2^2\Delta_q^*\|_{\psi_{\frac{2}{3}}} \in O(\sigma_\Delta\sigma_x^2 l)$ by the Lemma 5.7 below.

Since $l \in O(k)$, by the Lemma 5.8 below, we have $\|X_q^T X_S\Delta_S^*\|_{\psi_{\frac{2}{3}}} \in O(\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{kl})$. □

Therefore, we can use the Lemma 5.5 again and let $T \in \Omega\left(k\log(p-k)/l\right)$, $\log\delta^{-1} \in O\left(\log(p-k)\right)$, $\lambda \in \Omega\left(\sigma_\Delta\max(\sigma_x^2,\sigma_x)\sqrt{\frac{k\log(p-k)}{Tl}}\right)$. Then we have

$$\mathbb{P}\left[\|\zeta_S\|_\infty \geq \frac{\gamma}{2-\gamma}\right] \leq \exp(-c_{15}\log(p-k))$$

where c_{15} is a constant.

Lemma 5.7 (Orlicz norm for Product of Random Variables; Lemma 8 in (Hao et al., 2020)). *Suppose $X_1, X_2, \cdots, X_m$ are m random variables (not necessarily independent) with ψ_α-Orlicz norm bounded by $\|X_j\|_{\psi_\alpha} \leq K_j$. Then the $\psi_{\alpha/m}$-Orlicz norm of $\prod_{j=1}^m X_j$ is bounded by*

$$\left\| \prod_{j=1}^m X_j \right\|_{\psi_{\frac{\alpha}{m}}} \leq \prod_{j=1}^m K_j.$$

Lemma 5.8 (Orlicz norm for Sum of Random Variables; Lemma A.3 in (Götze et al., 2019)). *For any $0 < \alpha < 1$ and any random variables X, Y, we have*

$$\|X + Y\|_{\psi_\alpha} \leq 2^{1/\alpha}(\|X\|_{\psi_\alpha} + \|Y\|_{\psi_\alpha}).$$

5.5 Bound on $\|\tilde{\mathbf{z}}_{S^c}\|_\infty$ and the estimation error $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$

Since we have bounded each part of $\tilde{\mathbf{z}}_{S^c}$, we have $\|\tilde{\mathbf{z}}_{S^c}\|_\infty < 1$ with high probability, therefore the first part of Theorem 4.1 about support recovery ($S(\hat{\mathbf{w}}) \subseteq S$) is proved through primal-dual witness by finishing **step 1** and **step 5**. The proof for the second part of Theorem 4.1 about the estimation error uses similar techniques. Details can be found in the appendix Section E.

6 EXPERIMENTS

In this section, we present simulations and a real-world experiment with a gene expression dataset to support Theorem 4.1 and show the advantage of our method.

6.1 Simulations

For all $i \in [T]$, $j \in [l]$, $m \in S$, we set $\epsilon_{t_i,j} \sim N(\mu = 0, \sigma_\epsilon = 0.1)$, $\Delta^*_{t_i,m} \sim N(\mu = 0, \sigma_\Delta = 0.2)$, $X_{t_i,j,m} \sim N(\mu = 0, \sigma_x = 1)$, which are mutually independent. We set $p = 100$ and $\mathbf{w}^*$ having five entries equal to 1, and the rest of the entries being 0. The support of $\Delta^*_{t_i}$ is same as the support of $\mathbf{w}^*$. We choose $l \in \{3, 5, 7, 10\}$ and use $\lambda = \sqrt{k \log(p-k)/(5Tl)}$ for all the pairs of (T, l). The results are shown in Figure 1. The number of tasks T is rescaled to $C := \frac{Tl}{k \log(p-k)}$. For different choices of l, the curves for $P(\hat{S} = S)$ overlap with each other perfectly. We compare our results to two multi-task methods (Obozinski et al., 2011; Jalali et al., 2010) (since they do not estimate S directly, we let $\hat{S} := \bigcup_{i=1}^T \hat{S}_i$.) Multi-task methods perform worse under larger T while our method performs better.

In the appendix Section H, we give the details of this simulation, and show more simulations with different settings (on changing p, changing k, random variables with Uniform distribution, random variables with mixture of sub-Gaussian distribution, and correlated Gaussian covariates in X) and more analyses (on $P(\hat{S} = S)$, $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$, and $P(\hat{S}_T = S_T)$ for multi-task methods). All the results support our theoretical sample complexity rate and estimation error bound $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty \in O(\lambda)$.

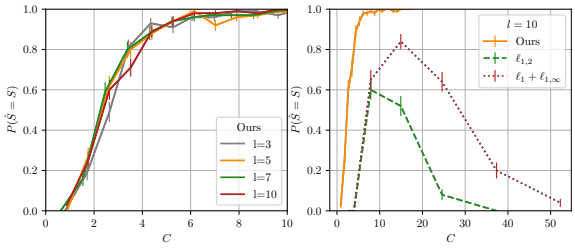

Figure 1: Simulations for Theorem 4.1 on the Probability of Exact Support Recovery with $\lambda = \sqrt{k \log(p-k)/(Tl)}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of l. The x-axis is set to $C := Tl/(k \log(p-k))$. **Right:** Our method outperforms multi-task methods ($\hat{S} := \bigcup_{i=1}^T \hat{S}_i$.)

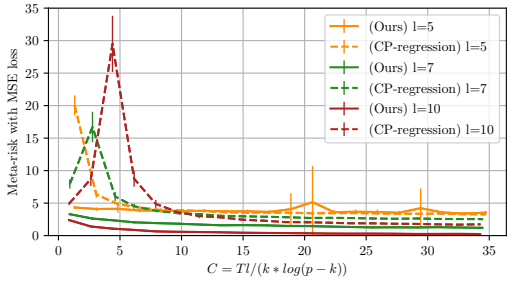

Figure 2: Comparison between our method and CP-Regression under various settings of l. The x-axis is set to $C := Tl/(k \log(p-k))$. The y-axis is the expected mean square error of prediction on the novel task.

We also compare our method with Chorus of Prototypes (CP)-Regression (Maurer, 2005). CP-Regression uses prior tasks to add new covariates to the novel task while our method uses S to remove inactive covariates. We measure the performance of the two methods by the expected mean square error of prediction on the novel task. The simulation results in Figure 2 show that our method performs better under $l = 7$ and $l = 10$. Details of this experiment can be found in the appendix Section H.

6.2 Real-world experiments

The single-cell gene expression dataset from (Kouno et al., 2013) contains expression levels of 45 transcription factors measured at 8 distinct time-points. This dataset was used in the experimental validation by Ollier and Viallon (2017). Similar to their analysis, we pick one transcription factor as the response variable y, and the other 44 factors as the covariates X, i.e., $p = 44$. Note that the true value of the support size k is unknown, and the distribution of the noise is also unknown (which may not be sub-Gaussian). We choose $l \in \{5, 7, 10, 15\}$ as the sample size of each task to model this problem as few-shot learning.

We compare our method to two multi-task methods (Obozinski et al., 2011; Jalali et al., 2010) and one meta-learning method, CP-Regression (Maurer, 2005). The results are shown in Figure 3. When l is small, our method has lower MSE and comparable $|\hat{S}|$ to others, which suggests that our $\hat{S}$ is more accurate. Details of this experiment can be found in the appendix Section I.

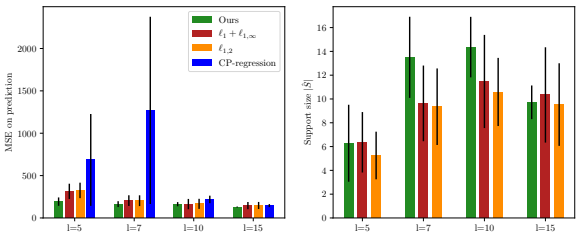

Figure 3: Results on the Single-Cell Gene Expression Dataset. **Left:** The mean square error (MSE) of prediction on the new task. **Right:** The size of the estimated common support $\hat{S}$. Note that CP-Regression does not estimate S.

7 DISCUSSIONS

Our problem setting and method are amenable to solid statistical analysis. By focusing on sparse regression, our analysis shows clearly the difference between meta-learning and multi-task learning. In meta-learning, we only need to recover $\mathbf{w}^*$ and $\Delta^*_{t_{T+1}}$, thus the number of samples needed for each task (including the novel task) is $l \in O((k \log p)/T + k_{T+1} \log k)$. When $T \to \infty$, meta-learning can recover $\mathbf{w}^*$ with high probability (shown in the left panel of Figure 1 where $C := \frac{Tl}{k \log(p-k)}$), therefore for the novel task, it only needs $l \in O(k_{T+1} \log k)$. For multi-task learning, one needs to recover $(\mathbf{w}^* + \Delta^*_{t_i})$ for all t_i, which requires the sample size at least $l \in O(k(T + \log p))$ (see Table 1.) When $T \to \infty$, the sample size of multi-task learning needed for support recovery goes to infinity which is supported by the right panel of Figure 1.

While meta sparse regression might apparently look similar to the classical sparse random effect model (Bondell et al., 2010), a key difference is that in the random effect model, the experimenter is interested on the distribution of the estimator $\mathbf{w}^*$ instead of support recovery. To the best of our knowledge, our results are the first to give upper and lower bounds of the sample complexity of meta-learning problems.

Although our paper shows that a proper application of ℓ_1 regularized (sparse) regression achieves the minimax optimal rate, it is still unclear whether there is a method that can improve the constants in our results. To have further theoretical understanding of meta-learning, one could consider other algorithms, such as nonparametric regression or neural networks. We believe that our results are a solid starting point for the sound statistical analysis of meta-learning.

Bibliography

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092.

Fano, R. (1952). Class notes for course 6.574: Transmission of information. *MIT*, 4:3.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930.

Götze, F., Sambale, H., and Sinulis, A. (2019). Concentration inequalities for polynomials in α-sub-exponential random variables. *arXiv preprint arXiv:1903.05964*.

Hao, B., Zhang, A., and Cheng, G. (2020). Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory*.

Jalali, A., Sanghavi, S., Ruan, C., and Ravikumar, P. K. (2010). A dirty model for multi-task learning. In *Advances in neural information processing systems*, pages 964–972.

Kouno, T., de Hoon, M., Mar, J. C., Tomaru, Y., Kawano, M., Carninci, P., Suzuki, H., Hayashizaki,

Y., and Shin, J. W. (2013). Temporal dynamics and transcriptional control using single-cell gene expression analysis. *Genome biology*, 14(10):R118.

Lounici, K., Pontil, M., Tsybakov, A. B., and Van De Geer, S. (2009). Taking advantage of sparsity in multi-task learning. In *the 22nd Annual Conference on Learning Theory*.

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

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

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

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1):83–96.

Ravi, S. and Larochelle, H. (2016). Optimization as a model for few-shot learning.

Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087.

Tandon, R., Shanmugam, K., Ravikumar, P. K., and Dimakis, A. G. (2014). On the information theoretic limits of learning ising models. In *Advances in Neural Information Processing Systems*, pages 2303–2311.

Vanschoren, J. (2019). Meta-learning. In *Automated Machine Learning*, pages 35–61. Springer, Cham.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3):1–34.