The Sample Complexity of Meta Sparse Regression: Supplementary Materials

A COMPARISON ON RATES OF l

For (Jalali et al., 2010), we use $r := T, p := p, s := k, n := l$ (their notation := our notation) in the condition in Theorem 2. Therefore, the sample complexity in (Jalali et al., 2010) is $l \in O(\max(k\log(pT), kT(T+\log p))$. This is also supported by our Figure H.3.

For (Negahban and Wainwright, 2011), we use $r := T, p := p, |U| := |S| = k, n := l$ (their notation := our notation) in their equation (20) and (22) to obtain $l \in \Omega(k(T+\log(p)))$ and $l \in \Omega(T(T+\log(p)))$ for support recovery. Therefore, the sample complexity in (Negahban and Wainwright, 2011) is $l \in O(\max(k,T)(T+\log(p)))$.

For (Obozinski et al., 2011), we use $K := T, p := p, s := k, n := l$ (their notation := our notation). By their equation (22), the rate of $\psi(B^*)$ is between $O(k/T)$ (requiring a strong orthonormal assumption which we do not need) and $O(k)$. Here we use $\psi(B^*) \in O(k)$. From their equation (19) and (20), one needs $l \in \Omega(k\log(p-k))$ and $l \in \Omega(T\log(k))$. Therefore, the sample complexity in (Obozinski et al., 2011) is $l \in O(\max(k\log(p-k), T\log(k)))$. Note that the latter still grows with respect to T unless $T \in O(k\log(p-k)/\log(k))$. This is also supported by our Figure H.2.

B PROOF OF STEP 1 IN THE PRIMAL-DUAL WITNESS

We know that

$$[\nabla^2 \ell((\mathbf{w}_S, \mathbf{0}))]_{S,S} \succ 0 \Leftrightarrow \frac{1}{Tl}[\mathbf{X}_{[T]}^T \mathbf{X}_{[T]}]_{S,S} \succ 0. \tag{B.1}$$

We first show a useful theorem on bounding the difference between the sample covariance matrix and the population covariance matrix.

Lemma B.1 (Theorem 4.6.1 in Vershynin (2018))**.** *Let A be an $m \times n$ matrix whose rows $A_i \in \mathbb{R}^{1\times n}$ are independent, mean zero, sub-gaussian isotropic random vectors (i.e., $\mathbb{E}[A_i^T A_i] = I_n$). Then for any $t \geq 0$, we have*

$$\left\| \frac{1}{m} A^T A - I_n \right\| \leq K^2 \max(\delta, \delta^2) \quad \text{where } \delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right), K = \max_i \|A_i\|_{\psi_2}$$

with probability $1 - 2e^{-t^2}$.

To prove (B.1), we need to bound $\lambda_{\min}([\frac{1}{Tl}[\mathbf{X}_{[T]}^T \mathbf{X}_{[T]}]_{S,S})$ away from 0. We find independent isotropic random vectors Z_i such that for each row of $\mathbf{X}_{[T]}^T$ (denoted by X_i), $X_i = \Sigma_{S,S}^{1/2} Z_i$ (by the proof of Theorem 4.7.1 in Vershynin (2018), Z_i are also sub-Gaussian with $\|Z_i\|_{\psi_2} \leq K$ where K is a constant only depending on $\Sigma_{S,S}$.)

We let $m := Tl, n := k$, $A := [Z_1, Z_2, \cdots, Z_m]^T$. Then we use a similar technique as Lemma 4.1.5 in Vershynin (2018):

$$K^2 \max(\delta, \delta^2) \geq \left\| \frac{1}{m} A^T A - I_n \right\| \geq \left| \left\langle \left(\frac{1}{m} A^T A - I_n \right) x, x \right\rangle \right| = \left| \lambda_{min}\left(\frac{1}{m} A^T A \right) - 1 \right| \tag{B.2}$$

where we set $x = \arg\min_{a \in \mathbb{S}^{n-1}} a^T (A^T A) a$.

If $K \leq 1$, we let $t = \sqrt{m}/(6C)$, $m \geq 16nC^2$. If $K > 1$, we let $t = \sqrt{m}/(6K^2C)$, $m \geq 16nK^4C^2$. Under both cases, we have $K^2 \max(\delta, \delta^2) = K^2\delta < 1/2$. By (B.2), we further have $\lambda_{min}(\frac{1}{m} A^T A) > 1/2$ and also $\lambda_{min}(\frac{1}{Tl}[\mathbf{X}_{[T]}^T \mathbf{X}_{[T]}]_{S,S}) > \lambda_{min}(\Sigma_{S,S})/2 > 0$.

Therefore, if we have $Tl \in \Omega(k \log(p-k))$ tasks, (B.1) holds with probability greater than $1 - 2e^{-C'Tl}$, where C' is a constant.

C BOUND OF $\tilde{\mathbf{z}}_{S^c,1}$

Recall that

$$\tilde{\mathbf{z}}_{S^c,1} = \mathbf{X}_{[T],S^c}^T \left\{ \frac{1}{Tl} \mathbf{X}_{[T],S} (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S + \Pi_{\mathbf{X}_{[T],S}^{\perp}} \left(\frac{\epsilon_{[T]}}{\lambda Tl} \right) \right\}.$$

In order to bound $\|\tilde{\mathbf{z}}_{S^c,1}\|_\infty$, we consider each entry of the vector. That is, for $j \in S^c$, we need to bound

$$\tilde{\mathbf{z}}_{j,1} = \mathbf{X}_{[T],j}^T \left\{ \frac{1}{Tl} \mathbf{X}_{[T],S} (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S + \Pi_{\mathbf{X}_{[T],S}^{\perp}} \left(\frac{\epsilon_{[T]}}{\lambda Tl} \right) \right\}.$$

Since the entries in $\mathbf{X}_{[T],j} \in \mathbb{R}^n$ are independent and sub-Gaussian, and the rows in $\mathbf{X}_{[T],S}$ are also independent and sub-Gaussian, we define $E_{i,j}^T$ by decomposition:

$$\mathbf{X}_{[T],j} = \Sigma_{j,S} (\Sigma_{S,S})^{-1} \mathbf{X}_{[T],S} + E_{[T],j}^T$$

where $E_{i,j}^T$ is an independent sub-Gaussian random variable for all $i \in [T]$. We assume the variance proxy is σ_e^2 which is proportional to σ_x^2.

We can rewrite $\tilde{\mathbf{z}}_{j,1}$ based on $E_{[T],j}$:

$$\tilde{\mathbf{z}}_{j,1} = E_{[T],j} \underbrace{\left\{ \frac{1}{Tl} \mathbf{X}_{[T],S} (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S + \Pi_{\mathbf{X}_{[T],S}^{\perp}} \left(\frac{\epsilon_{[T]}}{\lambda Tl} \right) \right\}}_{A_j} + \underbrace{\Sigma_{j,S} (\Sigma_{S,S})^{-1} \tilde{\mathbf{z}}_S}_{B_j}.$$

By the mutual incoherence condition, we have $|B_j| = |\Sigma_{j,S}(\Sigma_{S,S})^{-1} \tilde{\mathbf{z}}_S| \le 1 - \gamma$. Therefore we only need to bound $\|A_j\|_2^2$ since the variance proxy for $E_{[T],j} A_j$ is $\|A_j\|_2^2 \sigma_e^2$ (and we can bound $E_{[T],j} A_j$ by the concentration inequality of sub-Gaussian random variables).

$$\|A_j\|_2^2 = A_j^T A_j = \frac{1}{Tl} \tilde{\mathbf{z}}_S^T (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S + \left\| \Pi_{\mathbf{X}_{[T],S}^{\perp}} \left(\frac{\epsilon_{[T]}}{\lambda Tl} \right) \right\|_2^2.$$

For the first part $\frac{1}{Tl} \tilde{\mathbf{z}}_S^T (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S$, by the techniques in appendix Section B, we have

$$\frac{1}{Tl} \tilde{\mathbf{z}}_S^T (\hat{\Sigma}_{S,S})^{-1} \tilde{\mathbf{z}}_S \le \frac{1}{Tl} \|\tilde{\mathbf{z}}_S\|_2^2 (\lambda_{\min}(\hat{\Sigma}_{S,S}))^{-1} \le \frac{1}{Tl} \frac{2k}{\lambda_{min}(\Sigma_{S,S})}$$

with probability $1 - 2e^{-C'Tl}$, where C' is a constant.

For the second part, we have

$$\left\| \Pi_{\mathbf{X}_{[T],S}^{\perp}} \left(\frac{\epsilon_{[T]}}{\lambda Tl} \right) \right\|_2^2 \le \frac{1}{\lambda^2 Tl} \frac{\|\epsilon_{[T]}\|_2^2}{Tl} = \frac{1}{\lambda^2 Tl} \frac{\sum_{i=1}^{Tl} \epsilon_i^2}{Tl} \le \frac{C_1 \sigma_\epsilon^2}{\lambda^2 Tl}.$$

The second inequality above is a direct result by the Orlicz norm of ϵ_i. By Lemma 5.7, we know that $\|\epsilon_i^2\|_{\psi_1} \le \sigma_\epsilon^2$. We let $Y_i = \frac{\epsilon_i^2}{\sigma_\epsilon^2} \ge 0$. Then we have

$$\mathbb{P}\left(\exp\left(\frac{\sum_{i=1}^{Tl} Y_i}{Tl} \right) \ge \exp(C_1) \right) \le \frac{\mathbb{E}[\exp(\sum_{i=1}^{Tl} Y_i])]}{\exp(C_1 Tl)} = \frac{\prod_{i=1}^{Tl} \mathrm{E}[\exp(|Y_i|)]}{\exp(C_1 Tl)} \le \frac{\sigma_\epsilon^{2Tl}}{\exp(C_1 Tl)} = e^{-C_2 Tl}.$$

where C_1, C_2 are constants. (The first inequality is by Markov's inequality, and the second inequality is by the definition of ψ_1-Orlicz norm.) Therefore we have that $\frac{\sum_{i=1}^{Tl} \epsilon_i^2}{Tl} \le C_1 \sigma_\epsilon^2$ holds with probability $1 - e^{-C_2 Tl}$.

Now we define $M(T,l,k) := \sigma_e^2\left(\frac{1}{Tl}\frac{2k}{\lambda_{min}(\Sigma_{S,S})} + \frac{C_1\sigma_\epsilon^2}{\lambda^2 Tl}\right)$ and the event $\mathcal{T}_j = \left\{E_{[T],j}A_j > \gamma/2\right\}$. We have

$$P\left(\bigcup_{j\in S^c}\mathcal{T}_j\right) \leq (p-k)\left(\exp\left(-\frac{\gamma^2}{8M(T,l,k)}\right) + 2e^{-C_3 Tl}\right).$$

Therefore, 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 that $\tilde{\mathbf{z}}_{S^c,1} < 1-\gamma/2$ holds with probability greater than $1-2e^{-C_4\log(p-k)}$.

D PROOF OF LEMMA 5.3

This is a generalization of Theorem 3.1.1 in Vershynin (2018).

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

Proof. By Lemma 5.8, we only need to show that there exists a constant c_9 such that $\|\|X\|_2\|_{\psi_2} \leq c_9\sigma_x\sqrt{d}$. From the definition of Orlicz norm, the latter is equivalent to

$$\mathrm{E}\left[\exp\left(\frac{\|X\|_2^2}{c_9^2\sigma_x^2 d}\right)\right] \leq 2.$$

We first show two useful lemmas:

Lemma D.2 (Maximal inequality. Lemma 2.2.2 in Van Der Vaart and Wellner (1996))**.** *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a convex, nondecreasing, nonzero function with $\psi(0) = 0$ and $\limsup_{x,y\to\infty}\psi(x)\psi(y)\psi(cxy) < \infty$ for some constant c. Then, for any random variables $X_1,\ldots,X_m$,*

$$\left\|\max_{1\leq i\leq m} X_i\right\|_\psi \leq K\psi^{-1}(m)\max_i\|X_i\|_\psi,$$

for a constant K depending only on ψ.

Lemma D.3 (Covering number. Lemma 5.7 and Example 5.8 in Wainwright (2019))**.** *A δ-cover of a set A with respect to a metric ρ is a set $\{\theta^1,\ldots,\theta^N\} \subseteq A$ such that for each $\theta \in A$, there exists some $i \in \{1,\ldots,N\}$ such that $\rho(\theta,\theta^i) \leq \delta$. The δ-covering number $N(\delta; A,\rho)$ is the cardinality of the smallest δ-cover. We let $B^d := \{x \in \mathbb{R}^d | \|x\|_2 \leq 1\}$. We have*

$$d\log(1/\delta) \leq \log N(\delta; B^d, \|\cdot\|_2) \leq d\log\left(1+\frac{2}{\delta}\right).$$

We let $N_{\frac{1}{2}}$ be the covering set that achieves the smallest $\frac{1}{2}$-covering number on set B^d. Therefore, for any $v \in B^d$, we can write $v = z + w$ where $z \in N_{\frac{1}{2}}$ and $\|w\| \leq \frac{1}{2}$ (i.e., $w \in \frac{1}{2}B^d$). Then we have

$$\max_{v\in B^d} v^T X \leq \max_{z\in N_{\frac{1}{2}}} z^T X + \max_{w\in\frac{1}{2}B^d} w^T X = \max_{z\in N_{\frac{1}{2}}} z^T X + \frac{1}{2}\max_{w\in B^d} w^T X.$$

Therefore, $\max_{v\in B^d} v^T X \leq 2\max_{z\in N_{\frac{1}{2}}} z^T X$.

We have

$$\mathbb{E}\left[\exp\left(\frac{\|X\|_2^2}{c_9^2\sigma_x^2 d}\right)\right] = \mathbb{E}\left[\exp\left(\frac{\max_{v\in B^d}(v^T X)^2}{c_9^2\sigma_x^2 d}\right)\right] \leq \mathbb{E}\left[\exp\left(\frac{\max_{z\in N_{\frac{1}{2}}}(z^T X)^2}{(c_9/2)^2\sigma_x^2 d}\right)\right].$$

From Lemma D.3, $|N_{\frac{1}{2}}| \le 5e^d$. We let $\psi(x) = \exp(x^2) - 1$ and $m = |N_{\frac{1}{2}}|$ in Lemma D.2. We have

$$\left\| \max_{1\le i\le |N_{\frac{1}{2}}|} z_i^T X \right\|_{\psi} \le K\psi^{-1}(|N_{\frac{1}{2}}|)\max_i \|z_i^T X\|_{\psi} \le K\sqrt{\log(5e^d+1)} \le K\sqrt{\log 6 + d}\;\sigma_x .$$

Since $d \ge 1$, we can find a constant c_9 such that

$$\left\| \max_{1\le i\le |N_{\frac{1}{2}}|} z_i^T X \right\|_{\psi} \le (c_9/2)\sigma_x\sqrt{d}.$$

Therefore, for this choice of c_9, we have $\|\|X\|_2\|_{\psi_2} \le c_9\sigma_x\sqrt{d}$.

□

E BOUND OF ESTIMATION ERROR IN THEOREM 4.1

The second part in Theorem 4.1 is about the estimation error. We write the estimation error in the following form:

$$\begin{aligned}\tilde{\mathbf{w}}_S - \mathbf{w}_S^* &= \hat{\Sigma}_{S,S}^{-1}\left(\frac{1}{Tl}\sum_{i=1}^{T}\mathbf{X}_{t_i,S}^T\epsilon_{t_i} - \lambda\tilde{\mathbf{z}}_S + \frac{1}{Tl}\sum_{i=1}^{T}\mathbf{X}_{t_i,S}^T\mathbf{X}_{t_i,S}\Delta_{t_i,S}^*\right)\\ &= \underbrace{\hat{\Sigma}_{S,S}^{-1}\frac{1}{Tl}\sum_{i=1}^{T}\mathbf{X}_{t_i,S}^T\epsilon_{t_i}}_{F_1} - \underbrace{\hat{\Sigma}_{S,S}^{-1}\lambda\tilde{\mathbf{z}}_S}_{F_2} + \underbrace{\hat{\Sigma}_{S,S}^{-1}\lambda\frac{1}{\lambda Tl}\sum_{i=1}^{T}\mathbf{X}_{t_i,S}^T\mathbf{X}_{t_i,S}\Delta_{t_i,S}^*}_{F_3}.\end{aligned}$$

In this section, we first show that in a general sub-Gaussian setting without the rotation invariance assumption **A6** on $\Delta_{t_i,S}^*$ and each row of $\mathbf{X}_{t_i,S}$, we have $\|\tilde{\mathbf{w}}_S - \mathbf{w}_S^*\|_\infty = O(\lambda\sqrt{k})$. Since the rate of λ needed for $S(\hat{\mathbf{w}}) \subseteq S$ could be as high as $\sqrt{k\log(p-k)/(Tl)}$, if we use $T \in \Omega(k\log(p-k)/l$, the estimation error bound $O(\lambda\sqrt{k}) = O(\sqrt{k})$ is not fully satisfactory. Therefore, we later show that with the rotation invariance assumption **A6**, we have a tighter bound $\|\tilde{\mathbf{w}}_S - \mathbf{w}_S^*\|_\infty = O(\lambda)$.

E.1 $O(\lambda\sqrt{k})$ bound without assumption A6

Since we know that $\|\tilde{\mathbf{z}}_S\|_\infty \le 1$, for bounding $\|F_2\|_\infty$, we need to bound $\|\hat{\Sigma}_{S,S}^{-1}\|_\infty$.

By the technique in appendix Section B, we know that

$$\|\hat{\Sigma}_{S,S}^{-1}\|_\infty \le \sqrt{k}\|\hat{\Sigma}_{S,S}^{-1}\|_2 \le \frac{2\sqrt{k}}{\lambda_{min}(\Sigma_{S,S})}$$

holds with probability greater than $1 - 2e^{-C'Tl}$, where C' is a constant.

Therefore,

$$\|F_2\|_\infty \le \frac{2\sqrt{k}\lambda}{\lambda_{min}(\Sigma_{S,S})}$$

holds with probability greater than $1 - 2e^{-C'Tl}$, where C' is a constant.

For $j \in S$, we have

$$\frac{1}{Tl}\sum_{i=1}^{T}\mathbf{X}_{t_i,j}^T\epsilon_{t_i} = \frac{1}{Tl}\sum_{i=1}^{T}\sum_{m=1}^{l}X_{t_i,j,m}\epsilon_{t_i,m}$$

which can be bounded by the concentration inequality of sub-exponential random variables. Here we let $\|X_{t_i,j,m}\epsilon_{t_i,m}\|_{\psi_1} = M$. By Lemma 5.7, we know that $M \in O(\sigma_x\sigma_\epsilon)$.

Now we use the Bernstein's inequality (Theorem 2.8.1 in Vershynin (2018)) to get

$$P\left(\left|\frac{1}{Tl}\sum_{i=1}^{T}\sum_{m=1}^{l} X_{t_i,j,m}\epsilon_{t_i,m}\right| \leq t\right) \geq 1 - 2\exp\left(C\min\left(\frac{t^2Tl}{M^2}, \frac{tTl}{M}\right)\right).$$

If we let $t = \lambda$, and $T \in \Omega\left(\frac{k\log(p-k)}{l}\right)$, then

$$\|F_1\|_\infty \leq \lambda\frac{2}{\lambda_{min}(\Sigma_{S,S})}$$

holds with probability greater than $1 - 2e^{-c_5 k\log(p-k)}$, where c_5 is a constant.

For F_3, we use the definition of ζ_S in Section 5.4 and we set $\gamma = 1$ in the ℓ_∞ bound of ζ_S. We know that

$$\|F_3\|_\infty \leq \frac{2\lambda\sqrt{k}}{\lambda_{min}(\Sigma_{S,S})}$$

holds with probability greater than $1 - c_6e^{-c_7\log(p-k)}$.

Therefore, we can bound the estimation error. That is, with probability greater than $1 - c_8e^{-c_9\log(p-k)}$, we have

$$\|\tilde{\mathbf{w}}_S - \mathbf{w}_S^*\|_\infty \leq \frac{6\lambda\sqrt{k}}{\lambda_{min}(\Sigma_{S,S})}.$$

E.2 $O(\lambda)$ bound with assumption A6

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. To obtain an $O(\lambda)$ bound, we need to assume that $\mathbf{X}_{t_i,S}$ and $\Delta^*_{t_i,S}$ are rotation invariant.

First, we can directly use the analysis on $\|F_1\|_\infty$ above since its bound is $O(\lambda)$ which is tight enough.

Second, we provide a lemma below similar to Lemma 5 in Wainwright (2009) and use it to tighten the bound on $\|F_2\|_\infty$. While Lemma 5 in Wainwright (2009) holds only for Gaussian variables, the lemma below holds for a more general case, i.e., rotation invariant sub-Gaussian variables.

Lemma E.1. *Consider a fixed nonzero vector $z \in \mathbb{R}^k$ and a random matrix $X \in \mathbb{R}^{n\times k}$ with i.i.d. rows X_i such that all X_i are rotation invariant, mean zero, sub-Gaussian isotropic random vectors with $\|X_i\|_{\psi_2} \leq K$. Under the scaling $n = \Omega(k\log(p-k))$, there are positive constants c_1 and c_2 such that for all $t > 0$,*

$$P\left(\left\|\left[\left(\frac{1}{n}X^TX\right)^{-1} - I_{k\times k}\right]z\right\|_\infty \geq c_1\|z\|_\infty\right) \leq 4e^{-c_2\min\{k,\log(p-k)\}}.$$

Proof. We begin by diagonalizing the random matrix: $(X^TX/n)^{-1} - I_{k\times k} = U^TDU$ where D is diagonal, and U is unitary. Since the distribution of X is rotation invariant, the matrices D and U are independent. Since $\|D\| = \|(X^TX/n)^{-1} - I_{k\times k}\|$, we use Lemma A.1. Thus, with probability $1 - 2e^{-k}$,

$$\|((X^TX/n) - I_{k\times k})s\|_2 \leq K^2C\sqrt{k/n}\|s\|_2,\ \forall s \in \mathbb{R}^k.$$

which is equivalent to

$$\|((X^TX/n)^{-1} - I_{k\times k})s'\|_2 \leq K^2C\sqrt{k/n}\|(X^TX/n)^{-1}s'\|_2,\ \forall s' \in \mathbb{R}^k$$

where we let $s' = (X^TX/n)s$.

We use the result in Appendix A and have $\|(X^TX/n)^{-1}s'\|_2 \leq 2\|s'\|_2$ with probability $1 - 2e^{-c_3k}$. Then we have

$$P\left(\left\|\left(\frac{1}{n}X^TX\right)^{-1} - I_{k\times k}\right\|_2 \geq c_4\sqrt{\frac{k}{n}}\right) \leq 4e^{-c_5k}.$$

We condition on the event $\|D\|_2 \leq c_4\sqrt{k/n}$ and follow similar arguments as in the proof of Lemma 5 in Wainwright (2009). □

To use this lemma, we need to also transform the $\tilde{\mathbf{z}}_S$ in F_2 into a fixed value sign($\mathbf{w}_S^*$). We define δ_S as

$$\delta_S := \hat{\Sigma}_{S,S}^{-1} \left(\frac{1}{Tl} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \epsilon_{t_i} - \lambda \text{sign}(\mathbf{w}_S^*) + \frac{1}{Tl} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \mathbf{X}_{t_i,S} \Delta_{t_i,S}^* \right).$$

Using the techniques in the proof of Lemma 3(b) in Wainwright (2009), we know that the sign consistency property $\tilde{\mathbf{z}}_S = \text{sign}(\mathbf{w}_S^*)$ is equivalent to $\text{sign}(\mathbf{w}_S^* + \delta_S) = \text{sign}(\mathbf{w}_S^*)$. Therefore we only need to bound the ℓ_∞ norm of δ_S. More specifically, for a fixed $\mathbf{w}_S^*$, we can choose our parameters to ensure $\text{sign}(\mathbf{w}_S^* + \delta_S) = \text{sign}(\mathbf{w}_S^*)$, then we have $\tilde{\mathbf{z}}_S = \text{sign}(\mathbf{w}_S^*)$ and $\tilde{\mathbf{w}}_S - \mathbf{w}_S^* = \delta_S$ being bounded by the same bound.

Now we redefine $F_2 = \hat{\Sigma}_{S,S}^{-1} \lambda \text{sign}(\mathbf{w}_S^*)$ by breaking it into two terms: $F_2 = \Sigma_{S,S}^{-1} \lambda \text{sign}(\mathbf{w}_S^*) + (\hat{\Sigma}_{S,S}^{-1} - \Sigma_{S,S}^{-1}) \lambda \text{sign}(\mathbf{w}_S^*)$ and bound their ℓ_∞ norm separately. We use the technique in Section V.B in Wainwright (2009) and have

$$P(\|F_2\|_\infty \geq c_6 \lambda \|\Sigma_{S,S}^{-1/2}\|_\infty^2) \leq 4e^{-c_2 \min\{k, \log(p-k)\}}.$$

Finally, we consider $F_3 = \hat{\Sigma}_{S,S}^{-1} \lambda \zeta_S$. From our Section 5.4, we have $\mathbb{P}(\|\zeta_S\|_\infty \geq 1) \leq \exp(-c_7 \log(p-k))$. We follow a similar procedure in Lemma C.1 and note that here $z = \zeta_S = \frac{1}{\lambda T l} \sum_{i=1}^{T} \mathbf{X}_{t_i,S}^T \mathbf{X}_{t_i,S} \Delta_{t_i,S}^*$ is not a constant vector. We show that with the rotation invariance condition on $\mathbf{X}_{t_i,S}$ and $\Delta_{t_i,S}^*$, we have that Uz is also rotation invariant where the unitary matrix U is defined by $(\mathbf{X}_S^T \mathbf{X}_S / n)^{-1} - I_{k \times k} = U^T D U$ with D being diagonal. This is because when we consider the distribution of a rotated Uz, i.e., $V(Uz)$ where V is another unitary matrix, $V(Uz)$ can be treated as $U(V'z)$ where V' is also a unitary matrix, then we can transform $\mathbf{X}_{t_i,S}$ and $\Delta_{t_i,S}^*$ as $V'^{-1/2} \mathbf{X}_{t_i,S}$ and $V'^{1/2} \Delta_{t_i,S}^*$ for all t_i to map $V(Uz)$ to $U'z'$ without changing its probability density function.

Since Uz is rotation invariant and $U = (u_1 \; u_2 \; \dots \; u_k)$ is unitary, we know that u_i is orthogonal to $g_i := \sum_{j \neq i} u_j z_j$ and u_i is uniformly distributed over a sphere of $k-1$ dimensions when conditioning on g_i. In the analysis of F_3, we further condition on two events: $A = \{\|D\|_2 \leq c_4 \sqrt{k/n}\}$ and $B = \{\|\zeta_S\|_\infty \leq 1\}$. Then we follow similar arguments as in the proof of Lemma 5 in Wainwright (2009) to claim that $F(u_i) := u_i^T D g_i$ is Lipschitz and use the concentration of measure for Lipschitz functions (Example 3.12 in Wainwright (2019)). Finally we have the same bound as in Lemma C.1 and

$$P(\|F_3\|_\infty \geq c_6 \lambda \|\Sigma_{S,S}^{-1/2}\|_\infty^2) \leq 4e^{-c_2 \min\{k, \log(p-k)\}} + e^{-c_7 \log(p-k)} + 4e^{-c_5 k} = c_8 e^{-c_9 \min\{k, \log(p-k)\}}.$$

Therefore, we can bound the estimation error by letting $c_3 := 2c_6 \|\Sigma_{S,S}^{-1/2}\|_\infty^2 + 2/\lambda_{min}(\Sigma_{S,S})$. Thus, with probability greater than $1 - c_8 e^{-c_9 \min\{k, \log(p-k)\}}$, we have

$$\|\tilde{\mathbf{w}}_S - \mathbf{w}_S^*\|_\infty \leq c_3 \lambda.$$

F PROOF OF THEOREM 4.4

We use the primal dual witness framework as in the proof of Theorem 4.1. Since for this novel $(T+1)$-th task, $\Delta_{t_i}^*, i = 1, 2, \cdots, T$ is not considered, the choice of l and λ can be more flexible. We set $l \in \Omega(k' \log(k - k'))$ and $\lambda \in \Omega(\sqrt{\log(k-k')/l})$.

For **step 1**, similar to the **step 1** in Theorem 4.1 (proved in the appendix Section B; here we replace Tl with l), with probability greater than $1 - 2e^{-C'l}$, we have

$$\frac{1}{l} [\mathbf{X}_{T+1}^T \mathbf{X}_{T+1}]_{S,S} \succ 0$$

For **step 5**, since $\Delta_{t_i}^*$ is no longer in $\tilde{\mathbf{z}}_{S^c}$, we have

$$\tilde{\mathbf{z}}_{S^c} = \mathbf{X}_{T+1,S^c}^T \left\{ \frac{1}{l} \mathbf{X}_{T+1,S} \left(\frac{1}{l} [\mathbf{X}_{T+1}^T \mathbf{X}_{T+1}]_{S,S} \right)^{-1} \tilde{\mathbf{z}}_S + \Pi_{\mathbf{X}_{T+1,S}^\perp} \left(\frac{\epsilon_{T+1}}{\lambda l} \right) \right\}$$

This is the same as the part $\tilde{\mathbf{z}}_{S^c,1}$ in the proof of **step 5** in Theorem 4.1. We can use the technique in appendix Section C to bound its ℓ_∞ norm. With probability greater than $1 - 2e^{-c_1 \log(k-k')}$, we have

$$\|\tilde{\mathbf{z}}_{S^c}\|_\infty \leq 1 - \gamma/2.$$

For the estimation error bound, we write it as below.

$$\begin{aligned}\hat{\mathbf{w}}_{T+1,S} - (\mathbf{w}^*_S + \Delta^*_{t_{T+1,S}}) &= \left(\frac{1}{l}[\mathbf{X}^T_{T+1}\mathbf{X}_{T+1}]_{S,S}\right)^{-1}\left(\frac{1}{l}\mathbf{X}^T_{T+1,S}\epsilon_{T+1} - \lambda\tilde{\mathbf{z}}_S\right) \\ &= \underbrace{\left(\frac{1}{l}[\mathbf{X}^T_{T+1}\mathbf{X}_{T+1}]_{S,S}\right)^{-1}\frac{1}{l}\mathbf{X}^T_{T+1,S}\epsilon_{T+1}}_{F_1} - \underbrace{\left(\frac{1}{l}[\mathbf{X}^T_{T+1}\mathbf{X}_{T+1}]_{S,S}\right)^{-1}\lambda\tilde{\mathbf{z}}_S}_{F_2}\end{aligned}$$

These two parts F_1, F_2 are similar to the F_1, F_2 in the appendix Section E, therefore we can use similar technique to bound its ℓ_∞ norm. We let $\Sigma_{S,S}$ be the population covariance matrix of task $T+1$. With probability greater than $1 - c_2 e^{-c_3 \log(k-k')}$, we have

$$\|\hat{\mathbf{w}}_{T+1} - (\mathbf{w}^* + \Delta^*_{t_{T+1}})\|_\infty \leq \frac{4\lambda\sqrt{k'}}{\lambda_{min}(\Sigma_{S,S})}.$$

To obtain an $O(\lambda)$ bound, we need to assume that all rows in $\mathbf{X}_{t_{T+1},S}$ are rotation invariant. The proof is the same as in Section E.2 (the only difference is that we do not have F_3). Thus, we replace $\tilde{z}_S$ with $\text{sign}(\mathbf{w}^* + \Delta^*_{t_{T+1}})$ in F_2, then we use Lemma C.1 to bound F_2.

$$P(\|F_2\|_\infty \geq c_4\lambda\|\Sigma^{-1/2}_{S,S}\|^2_\infty) \leq c_5 e^{-c_6 \min\{k', \log(k-k')\}}.$$

We let $c_7 := c_4\|\Sigma^{-1/2}_{S,S}\|^2_\infty + 2/\lambda_{min}(\Sigma_{S,S})$. With probability greater than $1 - c_8 e^{-c_9 \min\{k', \log(k-k')\}}$, we have

$$\|\hat{\mathbf{w}}_{T+1} - (\mathbf{w}^* + \Delta^*_{t_{T+1}})\|_\infty \leq c_7\lambda.$$

G PROOF OF THEOREM 4.5

We first introduce Fano's inequality (Fano, 1952; Yu, 1997) (the version below can also be found directly in Scarlett and Cevher (2019)).

Lemma G.1. *(Fano's inequality) With input dataset S, for any estimator $\hat{\theta}(S)$ with k possible outcomes, i.e., $\hat{\theta} \in \Theta, |\Theta| = k$, if S is generated from a model with true parameter θ^* chosen uniformly at random from the same k possible outcomes Θ, we have:*

$$\mathbb{P}[\hat{\theta}(S) \neq \theta^*] \geq 1 - \frac{\mathbb{I}(\theta^*, S) + \log 2}{\log k}.$$

Now we show that $\mathbb{I}(\theta^*, S) \leq Tl \cdot c_1 + l_{T+1} \cdot c_2$, where c_1, c_2 are constants, and θ^* represents the parameter $(\mathbf{w}^*, \Delta^*_{t_{T+1}})$ we want to recover. Here S is all the data in the $T+1$ tasks, $S_{[T]}$ is the data in the first T tasks, and S_i is the data of task t_i. The mutual information is bounded by the following steps.

$$\begin{aligned}\mathbb{I}(\theta^*, S) &= \frac{1}{k}\sum_{\theta^*\in\Theta}\int_S p_{S|\theta^*}(S)\log\frac{p_{S|\theta^*}(S)}{p_S(S)}dS = \frac{1}{k}\sum_{\theta^*\in\Theta}\int_S p_{S|\theta^*}(S)\log\frac{p_{S|\theta^*}(S)}{\frac{1}{k}\sum_{\theta'\in\Theta}p_{S|\theta'}(S)}dS \\ &\leq \frac{1}{k^2}\sum_{\theta^*\in\Theta}\sum_{\theta'\in\Theta}\int_S p_{S|\theta^*}(S)\log\frac{p_{S|\theta^*}(S)}{p_{S|\theta'}(S)}dS = \frac{1}{k^2}\sum_{\theta^*\in\Theta}\sum_{\theta'\in\Theta}\mathbb{KL}(P_{S|\theta^*}||P_{S|\theta'}).\end{aligned}$$

Given the common coefficient w^*, the data for each task is independent from each other. Therefore we have

$$\mathbb{KL}(P_{S|\theta^*}||P_{S|\theta'}) = \mathbb{KL}(P_{S_{[T]}|\theta^*}||P_{S_{[T]}|\theta'}) + \mathbb{KL}(P_{S_{T+1}|\theta^*}||P_{S_{T+1}|\theta'}). \quad \text{(G.3)}$$

First, we consider the first part in (G.3). We use S' to denote $S_{[T]}$. Let $P_{S'} = P_{S_{[T]}|\theta*}, P'_{S'} = P_{S_{[T]}|\theta'}$. Note that

$$\mathbb{KL}(P_{S'}||P'_{S'}) = \int_{S'} P_{S'}\log\frac{P_{S'}}{P'_{S'}}dS'.$$

Furthermore

$$\begin{aligned} P_{S'} &= \int_{\Delta^*_{t_1},\cdots,\Delta^*_{t_T}} P_{S'|\mathbf{w}^*,\Delta^*_{t_1},\cdots,\Delta^*_{t_T}} P_{\Delta^*_{t_1},\cdots,\Delta^*_{t_T}|\mathbf{w}^*} d\Delta^*_{t_1},\cdots, d\Delta^*_{t_T} \\ &= \int_{\Delta^*_{t_1}} P_{S_1|\mathbf{w}^*,\Delta^*_{t_1}} P_{\Delta^*_{t_1}|\mathbf{w}^*} d\Delta^*_{t_1} \cdots \int_{\Delta^*_{t_T}} P_{S_T|\mathbf{w}^*,\Delta^*_{t_T}} P_{\Delta^*_{t_T}|\mathbf{w}^*} d\Delta^*_{t_T}. \end{aligned}$$

This is because conditioning on $\mathbf{w}^*$,$\Delta^*_{t_1},\cdots,\Delta^*_{t_T}$ are independent, and conditioning on both of them, the data for each task is independent.

If we set $a_i = \int_{\Delta^*_{t_i}} P_{S_i|\mathbf{w}^*,\Delta^*_{t_i}} P_{\Delta^*_{t_i}|\mathbf{w}^*} d\Delta^*_{t_i}$, $a'_i = \int_{\Delta'_{t_i}} P_{S_i|\mathbf{w}',\Delta'_{t_i}} P_{\Delta'_{t_i}|\mathbf{w}'} d\Delta'_{t_i}$, we have

$$P_{S'} = a_1 a_2 \cdots a_T, \quad P'_{S'} = a'_1 a'_2 \cdots a'_T.$$

Therefore

$$\mathbb{KL}(P_{S'}||P'_{S'}) = \int_{S'} a_1 \cdots a_T \left(\log \frac{a_1}{a'_1} + \cdots + \log \frac{a_T}{a'_T} \right) dS'.$$

We know that a_i is a function of S_j only when $i = j$, and $\int_{S_j} a_j \; dS_j = 1$. Therefore, we have

$$\int_{S'} a_1 a_2 \cdots a_T \left(\log \frac{a_i}{a'_i} \right) dS' = \int_{S_i} a_i \log \frac{a_i}{a'_i} dS_i.$$

Therefore

$$\mathbb{KL}(P_{S'}||P'_{S'}) = \sum_{i=1}^{T} \int_{S_i} a_i \log \frac{a_i}{a'_i} dS_i = T \int_{S_i} a_i \log \frac{a_i}{a'_i} dS_i.$$

For any task t_i, conditioning on $(\mathbf{w}^*, \Delta^*_{t_i})$, we know that all samples in S_i are i.i.d. If we set $S_{i,j}$ to be the j-th sample in the task t_i, and $a_{i,j} = P_{S_{i,j}|\mathbf{w}^*,\Delta^*_{t_i}}$, we have

$$\int_{S_i} a_i \log \frac{a_i}{a'_i} dS_1 \leq l \int_{S_{i,1}} a_{i,1} \log \frac{a_{i,1}}{a'_{i,1}} dS_{i,1}.$$

Therefore,

$$\mathbb{KL}(P_{S'}||P'_{S'}) = Tl \int_{S_{i,1}} \int_{\Delta^*_{t_i}} P_{S_{i,1}|\mathbf{w}^*,\Delta^*_{t_i}} d\Delta^*_{t_i} \log \left(\frac{\int_{\Delta^*_{t_i}} P_{S_{i,1}|\mathbf{w}^*,\Delta^*_{t_i}} d\Delta^*_{t_i}}{\int_{\Delta'_{t_i}} P_{S_{i,1}|\mathbf{w}',\Delta'_{t_i}} d\Delta'_{t_i}} \right) dS_{i,1} = Tl \cdot c_1.$$

Then we consider the second part in (G.3). For the task t_{T+1}, conditioning on $(\mathbf{w}^*, \Delta^*_{t_{T+1}})$, since we know that all samples in S_{T+1} are i.i.d., we have

$$\mathbb{KL}(P_{S_{T+1}|\theta^*}||P_{S_{T+1}|\theta'}) = l_{T+1} \int_{S_{T+1,1}} P_{S_{T+1,1}|\mathbf{w}^*,\Delta^*_{t_{T+1}}} \frac{P_{S_{T+1,1}|\mathbf{w}^*,\Delta^*_{t_{T+1}}}}{P_{S_{T+1,1}|\mathbf{w}',\Delta'_{t_{T+1}}}} dS_{T+1,1} = l_{T+1} \cdot c_2.$$

Combining the results above, we have

$$\mathbb{I}(\theta^*, S) \leq Tl \cdot c_1 + l_{T+1} \cdot c_2.$$

Finally, from Fano's inequality, we know that

$$\mathbb{P}[\hat{\theta} \neq \theta^*] \geq 1 - \frac{\log 2 + Tl \cdot c_1 + l_{T+1} \cdot c_2}{\log |\Theta|}.$$

H ADDITIONAL EXPERIMENTS

In this section, we present simulations to show that Theorem 4.1 holds in the sense that for different choices of l and p, one only needs $T = c \cdot (k \log(p-k)/l)$ to recover the true common support S with high probability.

H.1 Simulations with fixed k

For all the experiments in this section, we let $k = |S| = 5$, and perform 100 repetitions for each setting. We compute the *empirical* probability of successful support recovery $P(\hat{S} = S)$ as the number of times we obtain exact support recovery among the 100 repetitions, divided by 100. We compute the standard deviation as $\sqrt{P(\hat{S} = S)(1 - P(\hat{S} = S))/100}$, that is, by using the formula of the standard deviation of the Binomial distribution. For the estimation error $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$, we calculate the mean and standard deviation by using the empirical results of the 100 repetitions.

H.1.1 Gaussian distribution setting

We first consider the setting of different sample size l. We choose $l \in \{3, 5, 7, 10\}$ and use $\lambda = \sqrt{k \log(p-k)/(5Tl)}$ for all the pairs of (T, l). We denote the set $\{1, 2, 3, \cdots, a\}$ by $[a]$. 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 $\mathbf{w}^*$. The results are shown in Figure H.1. The number of tasks T is rescaled to C defined by $\frac{Tl}{k \log(p-k)}$. For different choices of l, the curves overlap with each other perfectly (for both $P(\hat{S} = S)$ and $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$).

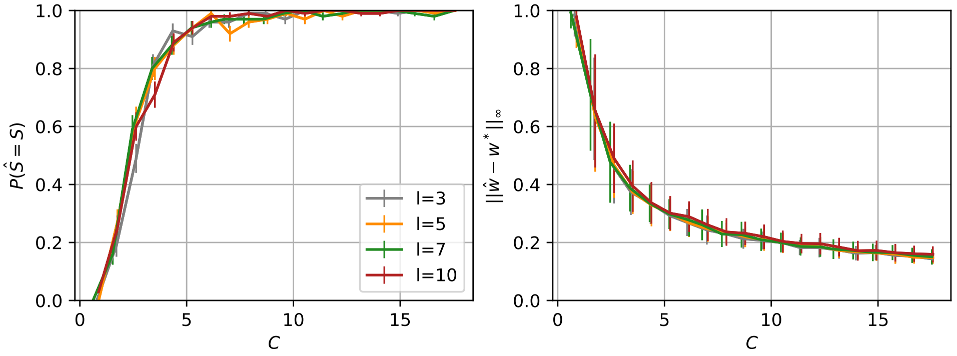

Figure H.1: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k \log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of sample size l. The x-axis is set by $C := \frac{Tl}{k \log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

Next we show that the problem described above cannot be solved by multi-task methods. We use two multi-task methods with regularization terms being $\ell_{1,2}$ Obozinski et al. (2011) and $\ell_1 + \ell_{1,\infty}$ Jalali et al. (2010) respectively. The results are shown in Figure H.2 and H.3, where we take $\hat{S} = \bigcup_{i=1}^T \hat{S}_i$. We show both $P(\hat{S} = S)$ and $P(\hat{S}_T = S_T)$ since the multi-task learning methods are not designed for recovering only the union of the supports of all tasks. As we claimed in Table 1, the multi-task methods require that l grows with T in order to retain the probability of support recovery. Therefore we see when l is fixed at $3, 5, 7, 10$, the probability of support recovery first increases then decreases to 0 as T increases. For the $\ell_{1,2}$ method of Obozinski et al. (2011), we use $\lambda_{1,2} = 30\sqrt{\log p/(Tl)}$ as the parameter for the $\ell_{1,2}$ norm; for the $\ell_1 + \ell_{1,\infty}$ method of Jalali et al. (2010), we use $\lambda_1 = 30\sqrt{\log p/(Tl)}$ as the parameter of the ℓ_1 norm and $\lambda_{1,\infty} = (1 + 1.5T)\lambda_1/2.5$ as the parameter of the $\ell_{1,\infty}$ norm. We also tried different choices of $\lambda_{1,2}, \lambda_1, \lambda_{1,\infty}$ and the trends of the results are similar.

The Figure 1 is from the results in Figure H.1, H.2, H.3.

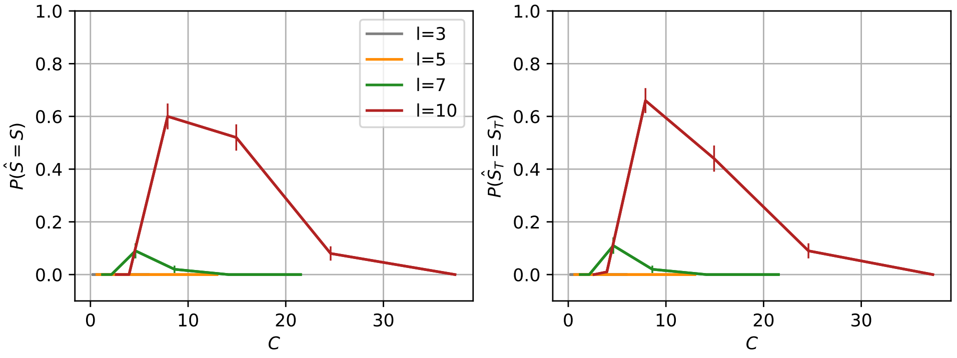

Figure H.2: Simulations with the multi-task method with $\ell_{1,2}$ regularization under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda_{1,2} = 30\sqrt{\log p/(Tl)}$. **Left:** Probability of exact support union recovery ($S = \hat{S} := \bigcup_{i=1}^{T} \hat{S}_i$) for different number of tasks under various settings of sample size l. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** Probability of exact support recovery of the last task ($\hat{S}_T = S_T$).

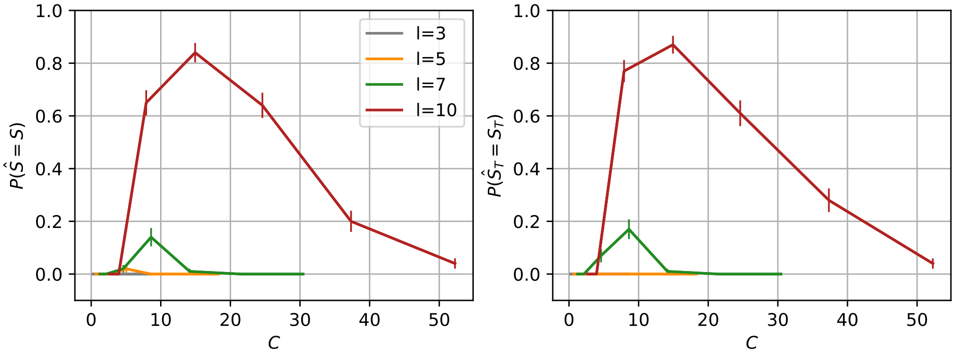

Figure H.3: Simulations with the multi-task method with $\ell_1 + \ell_{1,\infty}$ regularization under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda_1 = 30\sqrt{\log p/(Tl)}, \lambda_{1,\infty} = (1+1.5T)\lambda_1/2.5$. **Left:** Probability of exact support union recovery ($S = \hat{S} := \bigcup_{i=1}^{T} \hat{S}_i$) for different number of tasks under various settings of sample size l. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** Probability of exact support recovery of the last task ($\hat{S}_T = S_T$).

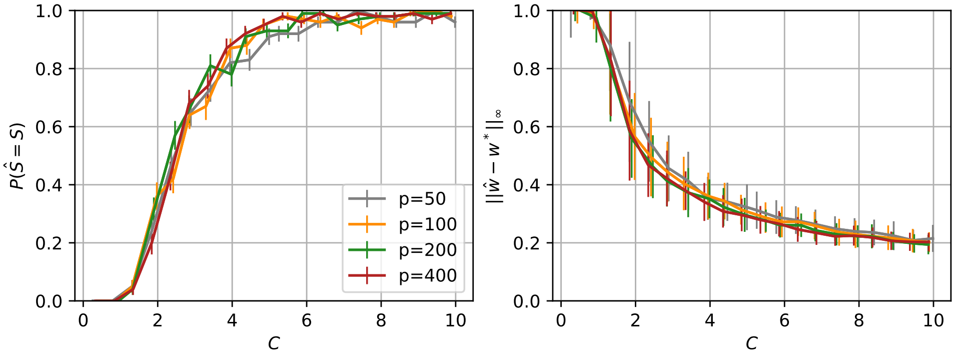

Figure H.4: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of number of parameters p. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

Then, for our method we consider the setting of different number of parameters p. We choose $p \in \{50, 100, 200, 400\}$ and use $\lambda = \sqrt{k \log(p-k)/(5Tl)}$ for all the pairs of (T, l). 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 $l = 5$, 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 $\mathbf{w}^*$. The results are shown in Figure H.4. The number of tasks T is rescaled to C defined by $\frac{Tl}{k \log(p-k)}$. For different choices of p, the curves overlap with each other perfectly (for both $P(\hat{S} = S)$ and $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$).

H.1.2 Uniform distribution setting

In this paper we only assume that the distributions are sub-Gaussian which includes the uniform distribution. Therefore in this section, we replace the Gaussian distribution setting in the appendix Section H.1.1 with a uniform distribution setting.

For all $i \in [T], j \in [l], m \in S$, we set $\epsilon_{t_i,j} \sim \text{Uniform}(-0.1\sqrt{3}, 0.1\sqrt{3})$, $\Delta^*_{t_i,m} \sim \text{Uniform}(-0.2\sqrt{3}, 0.2\sqrt{3})$, $X_{t_i,j,m} \sim \text{Uniform}(-\sqrt{3}, \sqrt{3})$, which are mutually independent. We consider the setting of different sample size l. We choose $l \in \{3, 5, 7, 10\}$ and use $\lambda = \sqrt{k \log(p-k)/(5Tl)}$ for all the pairs of (T, l). 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 $\mathbf{w}^*$. The results are shown in Figure H.5. The number of tasks T is rescaled to C defined by $\frac{Tl}{k \log(p-k)}$. For different choices of l, the curves overlap with each other perfectly (for both $P(\hat{S} = S)$ and $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$).

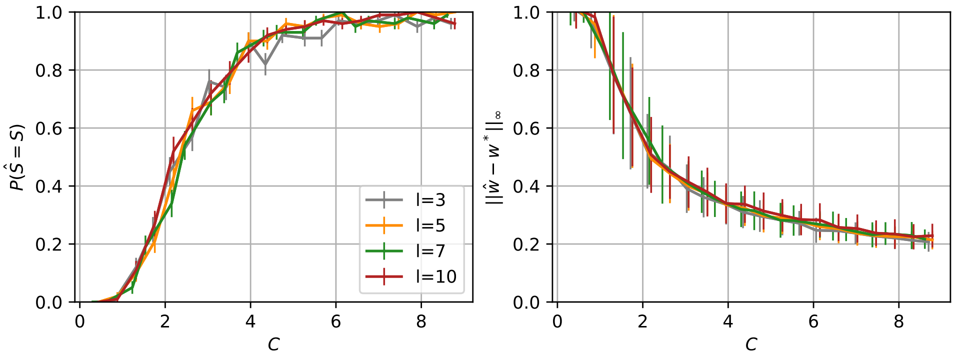

Figure H.5: Simulations with our meta sparse regression under uniform distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}\ \forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k \log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of l. The x-axis is set by $C := \frac{Tl}{k \log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

Then we consider the setting of different number of parameters p. We choose $p \in \{50, 100, 200, 400\}$ and use $\lambda = \sqrt{k \log(p-k)/(5Tl)}$ for all the pairs of (T, l). For all $i \in [T], j \in [l], m \in S$, we set $\epsilon_{t_i,j} \sim \text{Uniform}(-0.1\sqrt{3}, 0.1\sqrt{3})$, $\Delta^*_{t_i} \sim \text{Uniform}(-0.2\sqrt{3}, 0.2\sqrt{3})$, $X_{t_i,j,m} \sim \text{Uniform}(-\sqrt{3}, \sqrt{3})$, which are mutually independent. We set $l = 5$, and $\mathbf{w}^*$ having five entries equal to 1, and the rest of the entries being 0. The results are shown in Figure H.6. The number of tasks T is rescaled to C defined by $\frac{Tl}{k \log(p-k)}$. For different choices of p, the curves overlap with each other perfectly (for both $P(\hat{S} = S)$ and $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$).

H.1.3 Mixture of sub-Gaussian distribution setting

In Section 3.2, we state that we can consider the setting $S_i \subseteq S$ under the sub-Gaussian distribution assumption. Therefore in this section, we replace the Gaussian distribution setting of $\Delta^*_{t_i,m}$ in the appendix Section H.1.1 with a mixture of sub-Gaussian distribution setting. More specifically, we consider a mixture of a Dirac distribution and a Gaussian distribution.

For all $i \in [T], j \in [l], m \in S$, we set $\epsilon_{t_i,j} \sim N(\mu = 0, \sigma_\epsilon = 0.1)$, $X_{t_i,j,m} \sim N(\mu = 0, \sigma_x = 1)$, $\Delta^*_{t_i,m} \sim 0.5\ \delta_{-\mathbf{w}^*_m} + 0.5\ N(\mu = 0, \sigma_\Delta = 0.2)$, which are mutually independent. We consider the setting of different sample size l. We choose $l \in \{3, 5, 7, 10\}$ and use $\lambda = 4\sqrt{k \log(p-k)/(5Tl)}$ for all the pairs of (T, l). We set $p = 100$, and

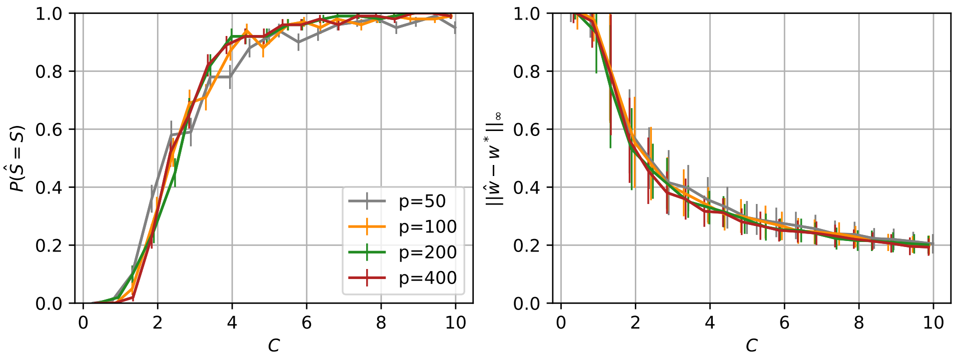

Figure H.6: Simulations with our meta sparse regression under uniform distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of number of parameters p. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

$\mathbf{w}^*$ having five entries equal to 2, and the rest of the entries being 0. The support of $\Delta^*_{t_i}$ is same as the support of $\mathbf{w}^*$ denoted by S while the support of $\mathbf{w}^* + \Delta^*_{t_i}$ could be a subset of S, i.e., $S_i \subseteq S$. More specifically, the distribution of $\Delta^*_{t_i,m}$ means that for the m-th parameter in the i-th task, i.e., $w_{i,m} := [\mathbf{w}^* + \Delta^*_{t_i}]_m, \forall i \in [T], m \in S$, there is a 50% probability that $w_{i,m} = 0$, and a 50% probability that $w_{i,m} \in N(2, 0.2)$.

The results are shown in Figure H.7. The number of tasks T is rescaled to C defined by $\frac{Tl}{k\log(p-k)}$. For different choices of l, the curves overlap with each other perfectly (for both $P(\hat{S} = S)$ and $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$).

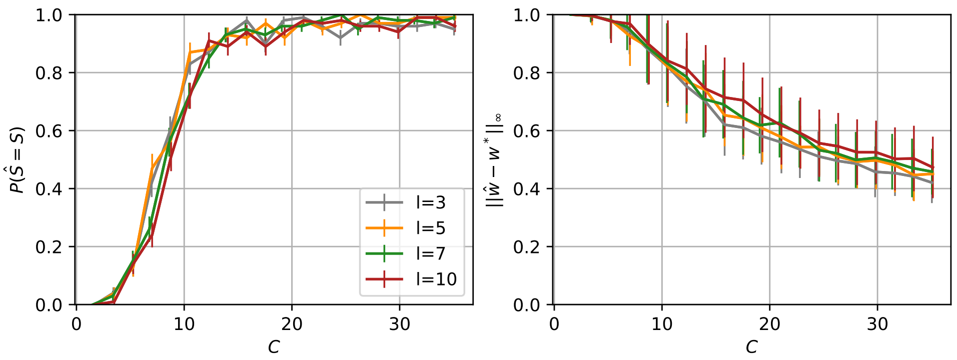

Figure H.7: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$ and a mixture of sub-Gaussian distributions of $\Delta^*_{t_i,m}$ $\forall i \in [T], m \in S$ such that the support of $\mathbf{w}^* + \Delta^*_{t_i}$ could be a subset of S, i.e., $S_i \subseteq S$. We use $\lambda = 4\sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of l. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

Then we consider the setting of different number of parameters p. We choose $p \in \{50, 100, 200, 400\}$ and use $\lambda = 4\sqrt{k\log(p-k)/(5Tl)}$ for all the pairs of (T, l). The distribution setting is same as in Figure H.7. We set $l = 5$, and $\mathbf{w}^* = (2, 2, 2, 2, 2, 0, 0, \cdots, 0)$. The results are shown in Figure H.8. The number of tasks T is rescaled to C defined by $\frac{Tl}{k\log(p-k)}$. For different choices of p, the curves overlap with each other perfectly (for both $P(\hat{S} = S)$ and $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_\infty$).

H.1.4 Gaussian distribution setting with entries in X being correlated

In this section, we consider three different correlation settings in X for our method:

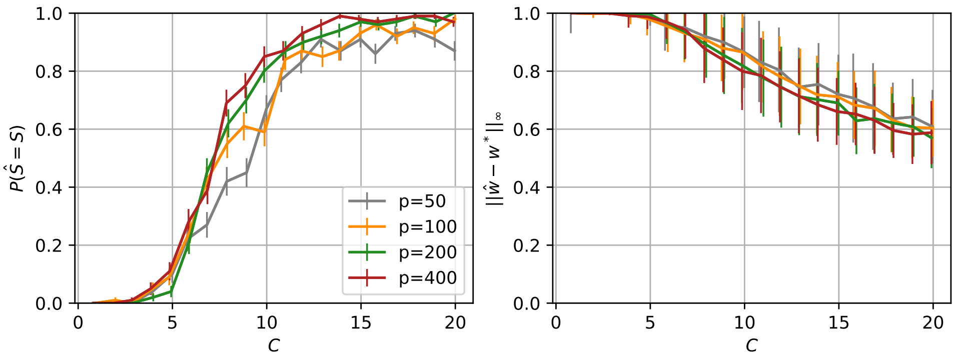

Figure H.8: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$ and a mixture of sub-Gaussian distributions of $\Delta^*_{t_i,m}$ $\forall i \in [T], m \in S$. We use $\lambda = 4\sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of number of parameters p. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

1. $X_{t_i,j}$ are i.i.d. from $N(0,\Sigma_x)$ where Σ_x is not a diagonal matrix;
2. $X_{t_i,j}$ are i.i.d. from $N(0,\Sigma_{x,t_i})$, and $\Sigma_{x,t_i} \sim F(\Sigma)$, i.e., for each task, the covariance matrix of X is different and sampled from a matrix distribution $F(\Sigma)$;
3. $X_{t_i,j}$ are i.i.d. from $N(\Delta^*_{t_i},\Sigma_{x,t_i})$, and $\Sigma_{x,t_i} \sim F_{\Delta^*_{t_i}}(\Sigma)$, i.e., for each task, the covariance matrix of X depends on the task specific coefficient $\Delta^*_{t_i}$.

First, we consider the setting of a nondiagonal Σ_x which leads to $\gamma < 1$ in the mutual incoherence condition, where $\gamma = 1 - |||\Sigma_{S^c,S}(\Sigma_{S,S})^{-1}|||_\infty$. For the simulations we present in the previous sections, the entries in $X_{t_i,j}$ are independent, therefore the covariance matrix of $X_{t_i,j}$ is diagonal and the corresponding $\gamma = 1$. Here we consider the case that the entries in $X_{t_i,j}$ are not independent. We choose $p = 100, l = 5$ and use $\lambda = \sqrt{k\log(p-k)/(5Tl)}$ for all the pairs of (T,l). We set $\mathbf{w}^*$ with 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 $\mathbf{w}^*$. For all $i \in [T], j \in [l], m \in S$, we set $\epsilon_{t_i,j} \sim N(\mu = 0, \sigma_\epsilon = 0.1)$, $\Delta^*_{t_i,m} \sim N(\mu = 0, \sigma_\Delta = 0.2)$, $X_{t_i,j} \sim N(\mu = 0, \Sigma = \Sigma_x)$, which are mutually independent. The covariance matrix $\Sigma_x = A^T A$ where A is a sum of a randomly generated orthonormal matrix U_0 and a matrix U_1 with each entry i.i.d. from Uniform$(-0.05, 0.05)$, i.e., $A = U_0 + U_1$. After we generate Σ_x, we calculate the corresponding γ. We generate 5 different Σ_x with 5 different γ. The results are shown in Figure H.9. The number of tasks T is rescaled to C defined by $\frac{Tl}{k\log(p-k)}$.

Then, we consider the setting of different Σ_x for each task, i.e., $X_{t_i,j} \sim N(0,\Sigma_{x,t_i}), \Sigma_{x,t_i} \sim F(\Sigma)$. We choose $p = 100, l = 5$ and use $\lambda = 2.5\sqrt{k\log(p-k)/(5Tl)}$ for all the pairs of (T,l). We set $\mathbf{w}^*$ with 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 $\mathbf{w}^*$. For all $i \in [T], j \in [l], m \in S$, we set $\epsilon_{t_i,j} \sim N(\mu = 0, \sigma_\epsilon = 0.1)$, $\Delta^*_{t_i,m} \sim N(\mu = 0, \sigma_\Delta = 0.2)$, $X_{t_i,j} \sim N(\mu = 0, \Sigma = \Sigma_{x,t_i})$, which are mutually independent. For each task, the covariance matrix $\Sigma_{x,t_i} = A^T_{t_i} A_{t_i}$ where A_{t_i} is a sum of a randomly generated orthonormal matrix U_{0,t_i} and a perturbation matrix U_{1,t_i} with each entry i.i.d. from Uniform$(-a,a)$, i.e., $A_{t_i} = U_{0,t_i} + U_{1,t_i}, [U_{1,t_i}]_{j,k} \sim$ Uniform$(-a,a)$. We choose the perturbation range a from $\{0.2, 0.1, 0.05, 0.01\}$. The results are shown in Figure H.10. The number of tasks T is rescaled to C defined by $\frac{Tl}{k\log(p-k)}$.

Finally, we consider the setting that for each task, the distribution of $X_{t_i,j}$ depends on the task specific coefficient $\Delta^*_{t_i}$. We choose $p = 100, l = 5$ and use $\lambda = 1.5\sqrt{k\log(p-k)/(5Tl)}$ for all the pairs of (T,l). We set $\mathbf{w}^*$ with 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 $\mathbf{w}^*$. For all $i \in [T], j \in [l], m \in S$, we set $\epsilon_{t_i,j} \sim N(\mu = 0, \sigma_\epsilon = 0.1)$, $\Delta^*_{t_i,m} \sim N(\mu = 0, \sigma_\Delta = 0.2)$, which are mutually independent. For each task, $X_{t_i,j} \sim N(\mu = \Delta^*_{t_i}, \Sigma = \Sigma_{x,t_i})$, and the covariance matrix $\Sigma_{x,t_i} = A^T_{t_i} A_{t_i}$ where A_{t_i} is a sum of a randomly generated orthonormal matrix U_{0,t_i} and a perturbation matrix $U_{1,t_i} = a\Delta^*_{t_i}(\Delta^*_{t_i})^T$, i.e., $A_{t_i} = U_{0,t_i} + U_{1,t_i}$. We choose the perturbation range a from $\{0.2, 0.1, 0.05, 0.01\}$. The results are shown in Figure H.11. The number of tasks T is rescaled to C defined by $\frac{Tl}{k\log(p-k)}$.

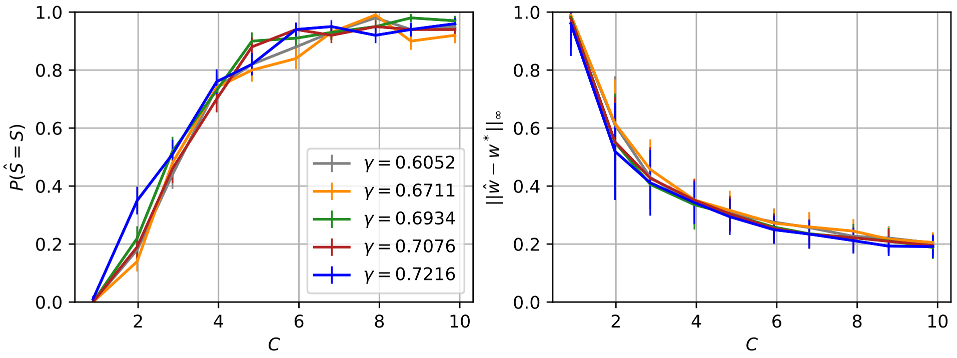

Figure H.9: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}$ and multivariate Gaussian distribution of $X_{t_i,j,m}$, $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k \log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of γ in the mutual incoherence condition, i.e., $\gamma = 1 - |||\Sigma_{S^c,S}(\Sigma_{S,S})^{-1}|||_\infty$. The x-axis is set by $C := \frac{Tl}{k \log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

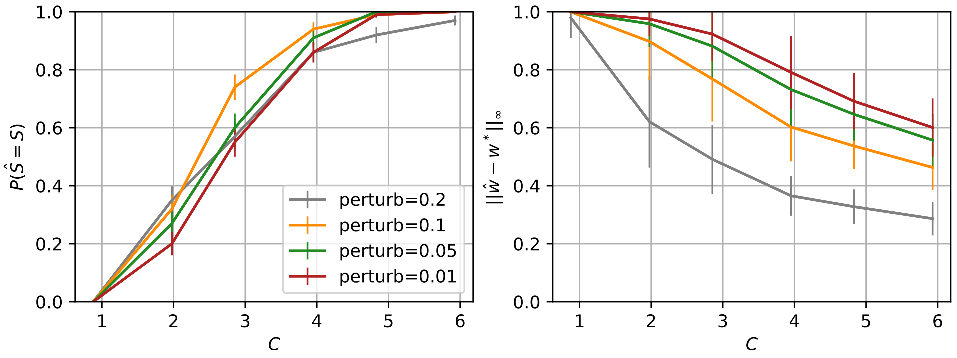

Figure H.10: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}$ and $X_{t_i,j,m} \sim N(0, \Sigma_{x,t_i})$, $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = 2.5\sqrt{\frac{k \log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of Σ_{x,t_i} where $\Sigma_{x,t_i} = A^T_{t_i} A_{t_i}$, $A_{t_i} = U_{0,t_i} + U_{1,t_i}$, U_{0,t_i} is randomly generated orthonormal matrix, U_{1,t_i} is perturbation matrix with each entry i.i.d. from Uniform$(-a, a)$, i.e., $[U_{1,t_i}]_{j,k} \sim$ Uniform$(-a, a)$. We choose the perturbation range a from $\{0.2, 0.1, 0.05, 0.01\}$. The x-axis is set by $C := \frac{Tl}{k \log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

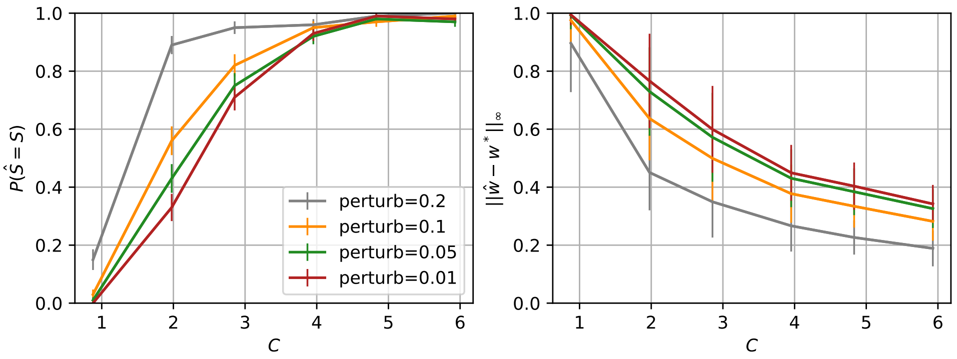

Figure H.11: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}$ and $X_{t_i,j,m} \sim N(\Delta^*_{t_i}, \Sigma_{x,t_i})$, $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = 1.5\sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of Σ_{x,t_i} where $\Sigma_{x,t_i} = A^T_{t_i}A_{t_i}, A = U_{0,t_i} + U_{1,t_i}$, U_{0,t_i} is randomly generated orthonormal matrix, $U_{1,t_i} = a\Delta^*_{t_i}(\Delta^*_{t_i})^T$ is perturbation matrix with a from $\{0.2, 0.1, 0.05, 0.01\}$. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

H.2 Simulations with changing k

Since our choice of λ is $O(\sqrt{k\log(p-k)/(Tl)})$ which has an extra $\sqrt{k}$ than the common choice of λ in LASSO, we perform experiments with changing k to support our theoretical result. We let $l = 5$ for the experiments in this section, and all the other settings are the same as in the previous section. More specifically, the results in Figure H.12, H.13, H.14, H.15 correspond to the results in Figure H.1, H.5, H.7, H.9 respectively. We can see that the curves overlap perfectly for different settings of k.

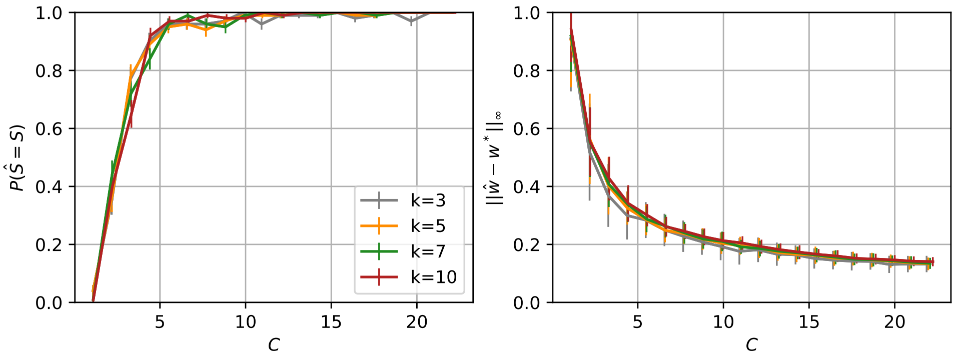

Figure H.12: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of k. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

H.3 Details about implementing CP-Regression

For CP-Regression, we choose its hyperparameters as $\mu = 100, \lambda_0 = 0, \lambda = 1$. More specifically, CP-Regression first uses ridge regression with λ_0 to fit each prior task to get models. Then it uses those models to predict the response of the data in the novel task. The predictions are scaled by μ and added to the covariate set. Therefore, if the original covariate in the novel task is $X_{t_{T+1}} \in \mathbb{R}^{l\times p}$, the new covariate set will be $X'_{t_{T+1}} \in \mathbb{R}^{l\times(p+T)}$ where T is the number of prior tasks. Finally, CP-Regression uses ridge regression with λ to fit the novel task with

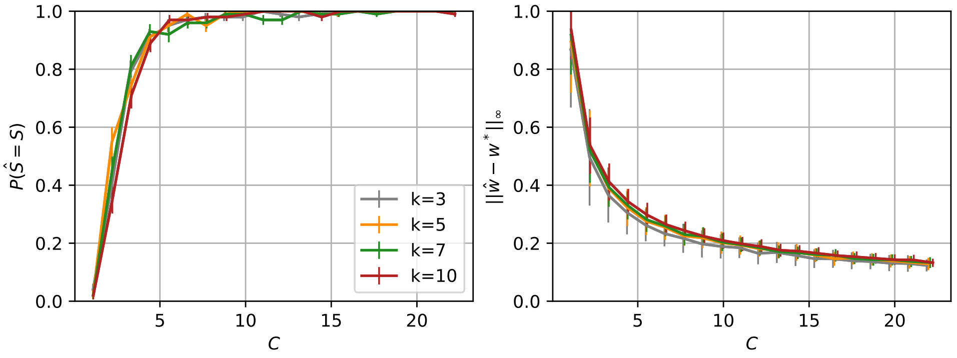

Figure H.13: Simulations with our meta sparse regression under uniform distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of k. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

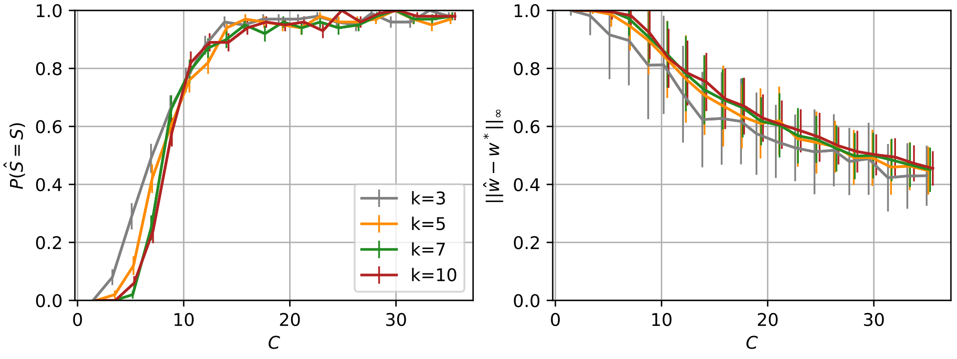

Figure H.14: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, X_{t_i,j,m}$ $\forall i \in [T], j \in [l], m \in S$ and a mixture of sub-Gaussian distributions of $\Delta^*_{t_i,m}$ $\forall i \in [T], m \in S$ such that the support of $\mathbf{w}^* + \Delta^*_{t_i}$ could be a subset of S, i.e., $S_i \subseteq S$. We use $\lambda = 4\sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of k. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

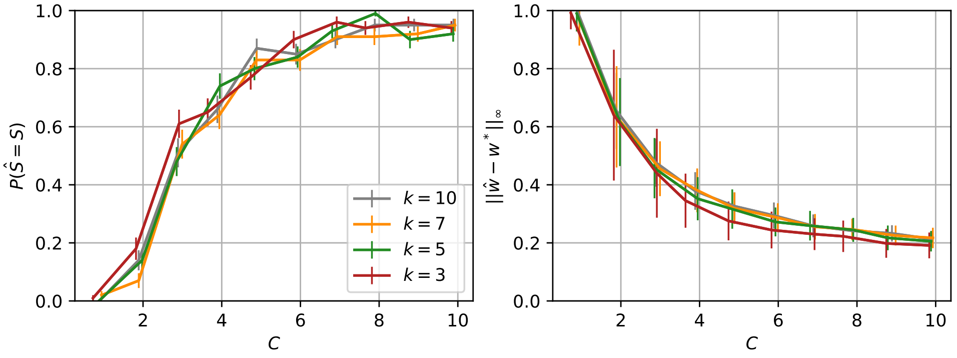

Figure H.15: Simulations with our meta sparse regression under Gaussian distributions of $\epsilon_{t_i,j}, \Delta^*_{t_i,m}$ and multivariate Gaussian distribution of $X_{t_i,j,m}$, $\forall i \in [T], j \in [l], m \in S$. We use $\lambda = \sqrt{\frac{k\log(p-k)}{5Tl}}$. **Left:** Probability of exact support recovery for different number of tasks under various settings of k. The x-axis is set by $C := \frac{Tl}{k\log(p-k)}$. **Right:** The corresponding estimation error of the common parameter $\mathbf{w}$ in ℓ_∞ norm.

$X'_{t_{T+1}}$. Maurer (2005) did not provide any strategy regarding the setting of the hyperparameters $(\lambda_0, \lambda, \mu)$ for CP-Regression, especially on how the hyperparameters depend on (T, l, p, k). Therefore, we select the best results for CP-Regression we have obtained.

I 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 contains 120 single cells for each time-point and was used in the experimental validation of Ollier and Viallon (2017). The original objective is to determine the associations among the transcription factors and how they vary over time. We formulate this as a meta-learning problem by setting the first 7 of the 8 time-points as the T tasks (for training) and the 8-th time-point as the novel task (for testing), i.e., $T = 7$. Similar to the analysis in Ollier and Viallon (2017), we pick one particular transcription factor, EGR2, as the response variable y, and the other 44 factors as the covariates in X, i.e., $p = 44$. The true value of the support size k is unknown. We choose $l \in \{5, 7, 10, 15\}$ to model this problem as few-shot learning.

We first randomly permute the 120 single cells (i.e., samples) while keeping their relative order in all of the 8 time points (i.e., tasks). Then we find a good choice of hyperparameters: λ in our method, $\lambda_{1,2}$ for the $\ell_{1,2}$ norm of the method in Obozinski et al. (2011); λ_1 and $\lambda_{1,\infty}$ for the ℓ_1 and $\ell_{1,\infty}$ norms, respectively of the method in Jalali et al. (2010). We use the tree-structured Parzen estimator approach (TPE) optimizing the criterion of expected improvement (EI) in the Python package `hyperopt` Bergstra et al. (2013). For CP-Regression, we choose its hyperparameters as $\mu = 100, \lambda_0 = 0, \lambda = 1$.

The search space is $[0, 100]$ for all these hyperparameters. For one choice of the hyperparameters, we choose l samples in each of the 7 tasks as training samples, and choose the rest $(120 - l)$ samples as validation samples. The TPE-EI algorithm evaluates 30 choices of hyperparameters to minimize the mean square error of the prediction on the validation samples.

After we determine the hyperparameters from all the three methods (ours, $\ell_{1,2}$, and $\ell_1 + \ell_{1,\infty}$), we choose l samples in each of the 7 tasks to train models by these methods to estimate S (for multi-task methods, $\hat{S} := \bigcup_{i=1}^{T} \hat{S}_i$.) The mean and standard deviation of the size of the estimated support are shown in the right panel of Figure 2.

When the estimated common supports are obtained, we can use LASSO constrained on the common support to solve for the new task, i.e., the 8-th time point. We determine the choice of hyperparameters using `hyperopt` in the same way shown above. Then we use LASSO with λ being set to those hyperparameters to estimate the support of the new task. Since the weight estimation of $(\mathbf{w}^* + \Delta^*_{t_{T+1}})$ by LASSO is not very accurate when the sample size l is small, we use linear regression to estimate $(\mathbf{w}^* + \Delta^*_{t_{T+1}})$ again with the support recovered by LASSO. The performance is measured by the mean square error (MSE) of prediction on the rest $(120 - l)$ samples. For one estimated common support, we take 6 random choices of the training l samples in the new task and calculate the mean of the the prediction error. The mean and standard deviation of MSE are shown in the left panel of Figure 2.

All the mean and standard deviation results (shown as error bars) in Figure 2 are obtained from 100 repetitions of the experiment setting above. From Figure 2 we can see that our method has lower MSE when l is small. Since T is not large and does not grow, the multi-task methods also perform well when l is large enough. We also show that the size of the estimated common support by our methods is not significantly larger than the ones by the other two multi-task methods, which suggests that our method produces a more accurate estimation of the common support set.

Bibliography

Bergstra, J., Yamins, D., and Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures.

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

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.

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.

Scarlett, J. and Cevher, V. (2019). An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*.

Van Der Vaart, A. and Wellner, J. (1996). Weak convergence and empirical processes: With applications to statistics springer series in statistics. *Springer*, 58:59.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Yu, B. (1997). *Assouad, Fano, and Le Cam*. Springer-Verlag.