8 Appendix

A Proof of Theorem 1

We first introduce the following lemma.

Lemma 3. *When the rows of $X_1,\ldots,X_t$ are independent subgaussian random vectors, with mean zero, covariance $\Sigma_1,\ldots,\Sigma_t$, respectively. Let*

$$C_M = \max_{t\in[m]}\max_{j\in[p]}\left(M_t^T\left(\frac{X_t^TX_t}{n}\right)M_t\right)_{jj}.$$

Then with probability at least $1-2mp\exp(-cn)-2mp^{-2}$ for some constant c, we have

$$C_M \le 2\max_{t\in[m]}\max_{j\in[p]}(\Sigma_t^{-1})_{jj}.$$

Proof. As shown in Theorem 2.4 of [25], Σ_t^{-1} will be a feasible solution for the problem of estimating M_t. Since we're minimizing $(M_t^T\hat{\Sigma}_tM_t)_{jj}$, we must have

$$\max_{j\in[p]}(M_t^T\hat{\Sigma}_tM_t)_{jj} \le \max_{j\in[p]}(\Sigma_t^{-1}\hat{\Sigma}_t\Sigma_t^{-1})_{jj}.$$

Based on the concentration results of sub-exponential random variable [26], also Lemma 3.3 of [17], we know with probability at least $1-2p\exp(-cn)$ for some constant c, we have

$$\max_{j\in[p]}(\Sigma_t^{-1}\hat{\Sigma}_t\Sigma_t^{-1})_{jj} \le 2\max_{j\in[p]}(\Sigma_t^{-1})_{jj}.$$

Take an union bound over $t\in[m]$, we obtain with probability at least $1-2mp\exp(-cn)$,

$$C_M \le \max_{t\in[m]}\max_{j\in[p]}(M_t^T\hat{\Sigma}_tM_t)_{jj} \le \max_{t\in[m]}\max_{j\in[p]}(\Sigma_t^{-1}\hat{\Sigma}_t\Sigma_t^{-1})_{jj} \le 2\max_{t\in[m]}\max_{j\in[p]}(\Sigma_t^{-1})_{jj}.$$

□

Now we are ready to prove Theorem 1, recall the model assumption

$$y_t = X_t\beta_t^* + \varepsilon_t, \qquad t=1,\ldots,m, \tag{17}$$

and the debiased estimation

$$\hat{\beta}_t^u = \hat{\beta}_t + n^{-1}M_tX_t^T(y_t - X_t\hat{\beta}_t), \tag{18}$$

we have

$$\begin{aligned}\hat{\beta}_t^u =& \hat{\beta}_t + \frac{1}{n}M_tX_t^T(X_t\beta_t^* - X_t\hat{\beta}_t) + \frac{1}{n}M_tX_t^T\varepsilon_t\\ =& \beta_t^* + (M_t\hat{\Sigma}_t - I)(\beta_t^* - \hat{\beta}_t) + \frac{1}{n}M_tX_t^T\varepsilon_t.\end{aligned}$$

For the term $(M_t\hat{\Sigma}_t - I)(\beta_t^* - \hat{\beta}_t)$, define

$$C_\mu = 10e\sigma_X^4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}},$$

we have the following bound

$$\begin{aligned}\|(M_t\hat{\Sigma}_t - I)(\beta_t^* - \hat{\beta}_t)\|_\infty \le& \max_j\|\hat{\Sigma}_tm_{tj} - e_j\|_\infty\|\beta_t^* - \hat{\beta}_t\|_1\\ \le_P& C_\mu\sqrt{\frac{\log p}{n}}\cdot\frac{16A}{\kappa}\sigma|S|\sqrt{\frac{\log p}{n}}\\ =& \frac{16AC_\mu\sigma|S|\log p}{\kappa n}.\end{aligned} \tag{19}$$

Noticed that

$$n^{-1}M_t X_t^T \varepsilon_t \sim \mathcal{N}\left(0, \frac{\sigma^2 M_t \hat{\Sigma}_t {M_t}^T}{n}\right).$$

Our next step uses a result on the concentration of χ^2 random variables. For any coordinate j, we have

$$\sum_{i=1}^{m} \left(n^{-1} e_j^T M_t X^T \varepsilon_t\right)^2 \leq \frac{C_M^2 \sigma^2}{n} \sum_{i=1}^{m} \xi_i^2,$$

where $(\xi_i)_{i\in[m]}$ are standard normal random variables. Using Lemma 9 with a weight vector

$$v = \left(\frac{C_M^2\sigma^2}{n}, \frac{C_M^2\sigma^2}{n}, \ldots, \frac{C_M^2\sigma^2}{n}\right)$$

and choosing $t = \sqrt{m} + \frac{\log p}{\sqrt{m}}$, we have

$$P\left\{\frac{\left(\frac{C_M^2\sigma^2}{n}\right)\sum_{i=1}^m \xi_i^2}{\sqrt{2m}\left(\frac{C_M^2\sigma^2}{n}\right)} - \sqrt{\frac{m}{2}} > \sqrt{m} + \frac{\log p}{\sqrt{m}}\right\} \leq 2\exp\left(-\frac{\left(\sqrt{m}+\frac{\log p}{\sqrt{m}}\right)^2}{2+2\sqrt{2}(1+\frac{\log p}{m})}\right).$$

A union bound over all $j \in [p]$ gives us that with probability at least $1 - p^{-1}$

$$\sum_{i\in[m]} \left(n^{-1} e_j^T M_t X^T \varepsilon_t\right)^2 \leq 3m\left(\frac{C_M^2\sigma^2}{n}\right) + \sqrt{2}\log p\left(\frac{C_M^2\sigma^2}{n}\right), \qquad \forall j \in [p]. \tag{20}$$

Combining (19) and (20), we get the following estimation error bound:

$$\begin{aligned}
\|\hat{B}_j - B_j\|_2 &= \sqrt{\sum_{i\in[m]} \left([M_t\hat{\Sigma}_t - I)(\beta_t^* - \hat{\beta}_t)]_j + \left[n^{-1}M_t X_t^T \varepsilon_t\right]_j\right)^2} \\
&\leq \sqrt{\sum_{i\in[m]} 2\left([M_t\hat{\Sigma}_t - I)(\beta_t^* - \hat{\beta}_t)]_j^2 + \left[n^{-1}M_t X_t^T \varepsilon_t\right]_j^2\right)} \\
&\leq \sqrt{\sum_{i\in[m]} \left(\frac{512A^2C_\mu^2\sigma^2|S|^2(\log p)^2}{\kappa^2 n^2}\right) + 6m\left(\frac{C_M^2\sigma^2}{n}\right) + 2\sqrt{2}\log p\left(\frac{C_M^2\sigma^2}{n}\right)} \\
&= \frac{\sigma}{\sqrt{n}}\sqrt{\frac{512A^2C_\mu^2 m|S|^2(\log p)^2}{\kappa^2 n} + 6C_M^2 m + 2\sqrt{2}C_M^2\log p} \\
&\leq \frac{91C_\mu\sigma|S|\sqrt{m}\log p}{\kappa n} + 3C_M\sigma\sqrt{\frac{m+\log p}{n}},
\end{aligned} \tag{21}$$

where the first inequality uses the fact $(a+b)^2 \leq 2a^2+2b^2$, and the second inequality uses (19) and (20)), the last inequality uses the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. For every variable $j \notin S$, we have

$$\|\hat{B}_j\|_2 \leq \frac{91C_\mu\sigma|S|\sqrt{m}\log p}{\kappa n} + 3C_M\sigma\sqrt{\frac{m+\log p}{n}}.$$

plug in $\kappa \geq \frac{1}{2}\lambda_{\min}$, $C_\mu = 10e\sigma_X^4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$, $C_M \leq 2K$, we obtain

$$\|\hat{B}_j\|_2 \leq \frac{1820e\sigma_X^4\lambda_{\max}^{1/2}\sigma|S|\sqrt{m}\log p}{\lambda_{\min}^{3/2}n} + 6K\sigma\sqrt{\frac{m+\log p}{n}}.$$

From (21) and the choice of Λ^*, we see that all variables not in S will be excluded from $\hat{S}$ as well. For every variable $j \in S$, we have

$$\|\hat{B}_j\|_2 \geq \|B_j\|_2 - \|\tilde{B}_j - B_j\|_2 \geq 2\Lambda^* - \Lambda^* = \Lambda^*.$$

Therefore, all variables in S will correctly stay in $\hat{S}$ after the group hard thresholding.

B Proof of Corollary 2

From Theorem 2 we have that $\hat{S}(\Lambda^*) \subseteq S$ and

$$\|\tilde{B}_j - B_j\|_2 \leq \frac{1820e\sigma_X^4\lambda_{\max}^{1/2}\sigma|S|\sqrt{m}\log p}{\lambda_{\min}^{3/2}n} + 6K\sigma\sqrt{\frac{m+\log p}{n}}, \tag{22}$$

with high probability. Summing over $j \in S$, we obtain the ℓ_1/ℓ_2 estimation error bound. For the prediction risk bound, we have

$$\begin{aligned}\frac{1}{nm}\sum_{t=1}^{m}\|X_t(\tilde{\beta}_t - \beta_t^*)\|_2^2 \leq & \frac{\lambda_{\max}}{m}\sum_{i=1}^{m}\|\tilde{\beta}_t - \beta_t^*\|_2^2 \\ = & \frac{\lambda_{\max}}{m}\sum_{j=1}^{p}\|\tilde{B}_j - B_j\|_2^2.\end{aligned}$$

Using (22) and the fact that $\tilde{B} - B$ is row-wise $|S|$-sparse, we obtain the prediction risk bound.

C Fixed design analysis

In this section, we present our theoretical results for the DSML procedure for fixed design, we will state the results without proof since the process is essentially the same as the case for random design. The results and comparisons are summarized in Table 3 and 4. We start by describing assumptions that we make on the model in (1). We assume the following condition on the design matrices $\{X_t\}_{t=1}^m$.

Approach	Communication	Assumptions	Min signal strength	Strength type
Lasso	0	Mutual Incoherence Sparse Eigenvalue	$\sqrt{\frac{\log p}{n}}$	Element-wise
Group lasso	$O(np)$	Mutual Incoherence Sparse Eigenvalue	$\sqrt{\frac{1}{n}\left(1+\frac{\log p}{m}\right)}$	Row-wise
DSML	$O(p)$	Generalized Coherence Restricted Eigenvalue	$\sqrt{\frac{1}{n}\left(1+\frac{\log p}{m}\right)} + \frac{\lvert S\rvert\log p}{n}$	Row-wise

Table 3: Conditions on the design matrix, and corresponding lower bound on coefficients required to ensure support recovery with p variables, m tasks, n samples per task and a true support of size $|S|$.

Approach	Assumptions	ℓ_1/ℓ_2 estimation error	Prediction error
Lasso	Restricted Eigenvalue	$\sqrt{\frac{\lvert S\rvert^2\log p}{n}}$	$\frac{\lvert S\rvert\log p}{n}$
Group lasso	Restricted Eigenvalue	$\frac{\lvert S\rvert}{\sqrt{n}}\sqrt{1+\frac{\log p}{m}}$	$\frac{\lvert S\rvert}{n}\left(1+\frac{\log p}{m}\right)$
DSML	Generalized Coherence Restricted Eigenvalue	$\frac{\lvert S\rvert}{\sqrt{n}}\sqrt{1+\frac{\log p}{m}} + \frac{\lvert S\rvert^2\log p}{n}$	$\frac{\lvert S\rvert}{n}\left(1+\frac{\log p}{m}\right) + \frac{\lvert S\rvert^3(\log p)^2}{n^2}$

Table 4: Conditions on the design matrix, and comparison of parameter estimation errors and prediction errors. The DSML guarantees improve over Lasso and have the same leading term as the Group lasso as long as $m < n/(|S|^2\log p)$.

A1 (Restricted Eigenvalues): Let $\mathcal{C}(s,L) = \{\Delta \in \mathbb{R}^p \mid \|\Delta_{U^c}\|_1 \leq L\|\Delta_U\|_1, U \subseteq [p], |U| \leq s\}$. There exists a constant $\kappa > 0$ such that

$$\min_{\Delta \in \mathcal{C}(|S|,L)} \Delta^T\hat{\Sigma}_t\Delta \geq \kappa\|\Delta_U\|_2^2, \qquad t = 1,\ldots,m.$$

There exists a constant $\phi_{\max} < \infty$ such that

$$\max_{\Delta \in \mathbb{R}^p} \Delta_S^T\hat{\Sigma}_t\Delta_S < \phi_{\max}\|\Delta_S\|_2^2.$$

The above assumption is commonly assumed in the literature in order to establish consistent estimation in high-dimensions [36]. **A1** imposes restrictions directly on the sample covariances $\hat{\Sigma}_t$, however, it is well known that the assumption will

hold with high-probability when rows of X_t are i.i.d. sub-gaussian or sub-exponential random vectors with population covariance satisfying **A1** [37, 38].

We will also need the following notion of coherence of the design matrices.

Definition 4 (Generalized Coherence)**.** *For matrices $X \in \mathbb{R}^{n\times p}$ and $M = (m_1,\ldots,m_p) \in \mathbb{R}^{p\times p}$, let*

$$\mu(X,M) = \max_{j\in[p]} \|\Sigma m_j - e_j\|_\infty$$

be the generalized coherence parameter between X and M, where $\Sigma = n^{-1}X^TX$. Furthermore, let $\mu^* = \min_{t\in[m]} \min_{M\in\mathbb{R}^{p\times p}} \mu(X_t,M)$ be the minimum generalized coherence.

This assumption is more relaxed than the mutual coherence parameter [39]. As shown in Theorem 2.4 of [25], $\mu(X_t,\Sigma^{-1}) \le 2\sqrt{\log(p)/n}$ with high-probability when the rows of X_t are i.i.d. sub-gaussian vectors with covariance matrix Σ.

The following theorem is our main result, which is proved in appendix.

Theorem 5. *Assume that* **A1** *holds and that the generalized coherence condition satisfies $\mu^* \le C_\mu\sqrt{\frac{\log p}{n}}$ for some constant C_μ. Suppose λ in (2) was chosen as $\lambda_t = A\sigma\sqrt{\frac{\log p}{n}}$ with constant $A > \sqrt{2}$. Furthermore, suppose that the multi-task coefficients in (1) satisfy the following bound on the signal strength*

$$\min_{j\in S}\sqrt{\sum_{t\in[m]}(\beta^*_{tj})^2} \ge \frac{2\sigma}{\sqrt{n}}\sqrt{\frac{512A^2C_\mu^2m|S|^2(\log p)^2}{\kappa^2 n} + 6C_M^2m + 2\sqrt{2}C_M^2\log p} := 2\Lambda^*, \tag{23}$$

where C_M is a constant that only depends on $\{M_t\}_{t=1}^m$. Then the support estimated by the master node satisfies $\hat{S}(\Lambda^*) = S$ with probability at least $1 - mp^{1-A^2/2} - p^{-1}$.

Based on Theorem 1, we have the following corollary that characterizes estimation error and prediction risk of DSML, with the proof given in the appendix.

Corollary 6. *Suppose the conditions of Theorem 5 hold. With probability at least $1 - mp^{1-A^2/2} - p^{-1}$, we have*

$$\sum_{j=1}^p \|\tilde{B}_j - B_j\|_2 \le \frac{|S|\sigma}{\sqrt{n}}\sqrt{\frac{512A^2C_\mu^2|S|^2(\log p)^2}{\kappa^2 n} + 6C_M^2m + 2\sqrt{2}C_M^2\log p}$$

and

$$\frac{1}{nm}\sum_{t=1}^m \|X_t(\tilde{\beta}_t - \beta_t^*)\|_2^2 \le \frac{\phi_{\max}|S|\sigma^2}{n}\left(\frac{512A^2C_\mu^2m|S|^2(\log p)^2}{\kappa^2 n} + 6C_M^2 + \frac{2\sqrt{2}C_M^2\log p}{m}\right).$$

D Collection of known results

For completeness, we first give the definition of subgaussian norm, details could be found at [26].

Definition 7 (Subgaussian norm)**.** *The subgaussian norm $\|X\|_{\psi_2}$ of a subgaussian p-dimensional random vector X, is defined as*

$$\|X\|_{\psi_2} = \sup_{x\in\mathbb{S}^{p-1}}\sup_{q>1} q^{-1/2}(\mathbb{E}|\langle X,x\rangle|^q)^{1/q},$$

where $\mathbb{S}^{p-1}$ is the p-dimensional unit sphere.

We then define the restricted set $\mathcal{C}(|S|,3)$ as

$$\mathcal{C}(|S|,3) = \{\Delta\in\mathbb{R}^p \,|\, \|\Delta_{U^c}\|_1 \le 3\|\Delta_U\|_1, U\subset[p], |U|\le|S|\}.$$

The following proposition is a simple extension of Theorem 6.2 in [36].

Proposition 8. *Let*

$$\lambda_t = A\sigma\sqrt{\frac{\log p}{n}}$$

with some constant $A > 2\sqrt{2}$ be the regularization parameter in lasso. With probability at least $1 - mp^{1-A^2/8}$,

$$\|\hat{\beta}_t - \beta_t^*\|_1 \leq \frac{16A}{\kappa'}\sigma|S|\sqrt{\frac{\log p}{n}},$$

where κ is the minimum restricted eigenvalue of design matrix $X_1, \ldots, X_m$:

$$\kappa = \min_{t\in[m]} \min_{\Delta\in C(|S|,3)} \frac{\Delta^T \left(\frac{X_t^T X_t}{n}\right)\Delta}{\|\Delta_S\|_2^2}.$$

Proof. Using Theorem 6.2 in [36] and take an union bound over $1, \ldots, m$ we obtain the result. □

Lemma 9 (Equation (27) in [40]; Lemma B.1 in [8])**.** *Let $\xi_1, \xi_2, \ldots \xi_m$ be i.i.d. standard normal random variables, let $v = (v_1, \ldots, v_m) \neq 0$, $\eta_v = \frac{1}{\sqrt{2}\|v\|_2}\sum_{i=1}^m (\xi_i^2 - 1)v_i$ and $m(v) = \frac{\|v\|_\infty}{\|v\|_2}$. We have, for all $t > 0$, that*

$$P(|\eta_v| > t) \leq 2\exp\left(-\frac{t^2}{2 + 2\sqrt{2}tm(v)}\right).$$

The next lemma relies on the generalized coherence parameter:

Definition 10 (Generalized Coherence)**.** *For matrices $X \in \mathbb{R}^{n\times p}$ and $M = (m_1, \ldots, m_p) \in \mathbb{R}^{p\times p}$, let*

$$\mu(X, M) = \max_{j\in[p]} \|\Sigma m_j - e_j\|_\infty$$

be the generalized coherence parameter between X and M, where $\Sigma = n^{-1}X^T X$. Furthermore, let $\mu^* = \min_{t\in[m]} \min_{M\in\mathbb{R}^{p\times p}} \mu(X_t, M)$ be the minimum generalized coherence.

Lemma 11 (Theorem 2.4 in [25])**.** *When X_t are drawn from subgaussian random vectors with covariance matrix Σ_t, and $X_t\Sigma_t^{-1/2}$ has bounded subgaussian norm $\|X_t\Sigma_t^{-1/2}\|_{\psi_2} \leq \sigma_X$. When $n \geq 24\log p$, then with probability at least $1 - 2p^{-2}$, we have*

$$\mu(X_t, \Sigma_t^{-1}) < 10e\sigma_X^4\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sqrt{\frac{\log p}{n}}.$$

For subgaussian design, we also have the following restricted eigenvalue condition [38, 17].

Lemma 12. *When X_t are drawn from subguassian random vectors with covariance matrix Σ_t, and bounded subgaussian norm σ_X. When $n \geq 4000s'\sigma_X \log\left(\frac{60\sqrt{2}ep}{s'}\right)$ where $s' = \left(1 + 30000\frac{\lambda_{\max}}{\lambda_{\min}}\right)|S|$, and $p > s'$, then with probability at least $1 - 2\exp^{(-n/4000C_\kappa^4)}$, for any vector $\Delta \in C(|S|, 3)$ where we have*

$$\Delta^T\left(\frac{X_t^T X_t}{n}\right)\Delta \geq \frac{1}{2}\lambda_{\min}\|\Delta_S\|_2^2.$$