Supplements to "Targeting Underrepresented Populations in Precision Medicine: A Federated Transfer Learning Approach"

This supplementary file contains alternative algorithms (Sections A and B), conditions, proofs to theorems, corollaries and additional lemmas (Sections C and D), additional simulation results (Section E) and data processing details (Section F).

A Federated learning algorithm to obtain the target-only estimator

In this section, we present the federated algorithm we use to obtain population-specific estimator.

Algorithm A.1: Federated algorithm for population-specific learning

```
Input  : Data from the $k$-th population$\{\boldsymbol{X}^{(m,k)}, \boldsymbol{y}^{(m,k)}\}_{m=1}^{M}$.
Initial value $\hat{\boldsymbol{w}}_0^{(k)}$. Note that if $k = 0$, $\hat{\boldsymbol{w}}_0^{(k)} = \hat{\boldsymbol{\beta}}_0$.
Output: $\hat{\boldsymbol{w}}_T^{(k)}$
for $t = 1, \dots, T$ do
    Threshold $\check{\boldsymbol{w}}_{t-1}^{(k)} = \mathcal{H}_{\sqrt{N^{(k)}}}(\hat{\boldsymbol{w}}_{t-1}^{(k)})$.
    for $m = 1, \dots, M$ do
        Transmit $\nabla L^{(m,k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})$ and $\nabla^2 L^{(m,k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})$ to the leading site.
    end
    Compute the combined first-order information $\nabla L^{(0)}(\check{\boldsymbol{\beta}}_{t-1})$, $\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})$
     according to (2.4).

    $$\hat{\boldsymbol{w}}_t = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \left\{ \widehat{R}^{(k)}(\boldsymbol{b}; \check{\boldsymbol{\beta}}_{t-1}) + \lambda_\beta \|\boldsymbol{b}\|_1 \right\}.$$
end
```

The algorithm is used to obtain the *target-only*, *source-only* estimators. To obtain the *combined* estimator which is fitted using all data, we also apply this algorithm treating data from the source and target population indistinctly.

B Leveraging local Hessian under design homogeneity

When the distribution of $\boldsymbol{x}$ in the k-th population is the same across sites, we introduce a modified version of Algorithm 1 in the main paper which only requires each participating site sharing only the gradients. This method generalizes the surrogate likelihood approach proposed by Wang et al. (2017); Jordan et al. (2018) to the transfer learning framework and it enjoys communication efficiency. The idea of this algorithm is to use the local data to approximate the Hessian matrices across multiple sites. We require that the leading site (the m^*-th site) has data from all the $(K+1)$ populations. We will use the empirical Hessian matrix obtained at the leading site to the approximate of the global Hessian in each population. For $k = 0, \ldots, K$, denote

$$R^{(local,k)}(\boldsymbol{b};\mathring{\boldsymbol{b}}) = \frac{1}{2}(\boldsymbol{b}-\mathring{\boldsymbol{b}})^\intercal \widehat{\boldsymbol{H}}^{(m^*,k)}(\mathring{\boldsymbol{b}})(\boldsymbol{b}-\mathring{\boldsymbol{b}}) + \langle \boldsymbol{b}-\mathring{\boldsymbol{b}}, \nabla L^{(k)}(\mathring{\boldsymbol{b}})\rangle, \text{ where}$$

where

$$\widehat{\boldsymbol{H}}^{(m^*,k)}(\mathring{\boldsymbol{b}}) = \frac{1}{n^{(m^*,k)}}\nabla^2 L^{(m^*,k)}(\mathring{\boldsymbol{b}})$$

is the empirical Hessian for the k-th population at $\boldsymbol{b}'$ based on the samples in the leading site.

Algorithm B.1: Federated transfer learning leveraging local Hessian

```
Input  : Target population$\{\boldsymbol{X}^{(m,0)}, \boldsymbol{y}^{(m,0)}\}_{m=1}^M$ and source populations
         $\{\{\boldsymbol{X}^{(m,k)}, \boldsymbol{y}^{(m,k)}\}_{m=1}^M\}_{k=1}^K$.
Initial values $\hat{\boldsymbol{\beta}}_0$, $\{\hat{\boldsymbol{w}}_0^{(k)}\}_{k=1}^K$.
Output: $\hat{\boldsymbol{\beta}}_T$
for $t = 1, \ldots, T$ do
    Threshold $\check{\boldsymbol{w}}_{t-1}^{(k)} = \mathcal{H}_{c_n}(\hat{\boldsymbol{w}}_{t-1}^{(k)})$ and $\check{\boldsymbol{\beta}}_{t-1} = \mathcal{H}_{c_n}(\hat{\boldsymbol{\beta}}_{t-1})$.
    for $m = 1, \ldots, M$ do
        Transmit $\nabla L^{(m,0)}(\check{\boldsymbol{\beta}}_{t-1})$ and $\{\nabla L^{(m,k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})\}_{k=1}^K$ to the leading site.
    end
    Compute the combined first-order information $\nabla L^{(0)}(\check{\boldsymbol{\beta}}_{t-1})$, $\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})$
      according to (2.4).
    In (2.5), (2.6), and (2.7) of Algorithm 1, we replace $\widehat{R}^{(k)}(\boldsymbol{b};\boldsymbol{b}')$ with
      $\widehat{R}^{(local,k)}(\boldsymbol{b};\boldsymbol{b}')$ and replace $\lambda^{(k)}, \lambda_\delta, \lambda_\beta$ with $\lambda_t^{(k)}, \lambda_{\delta,t}^{(k)}, \lambda_{\beta,t}$, respectively.
end
```

Without sharing the Hessian matrices, Algorithm B.1 largely reduces the communication cost. However, one limitation is that it requires the distribution $f(\boldsymbol{x}_i)$ in the k-th population are homogeneous across sites for any fixed k. Second, its reliable performance requires the existence of a single site that has relatively large samples from all $K+1$ populations. Otherwise, the local Hessian approximation can be inaccurate and lead to large estimation errors. In practice, however, such a desirable local site may not always exist. We provide a theoretical comparison in Section 3 showing that larger T might be needed in Algorithm B.1 to achieve the same estimation accuracy compared to 1.

B.1 Convergence rate of Algorithm B.1

In this section, we provide theoretical guarantees for Algorithm B.1, which leverages local Hessian and only transmits first-order information across sites. As we discussed before, it relies on the homogeneity assumption on the distribution of $\boldsymbol{x}^{(m,k)}$ for $m=1,\ldots,M$ at each given k. In the next theorem, we analyze the error contraction behavior of Algorithm B.1.

Theorem B.1 (Error contraction of Algorithm B.1)**.** *Assume Conditions 1, 2, Condition D.1 and true parameters are in* $\Theta(s,h)$*. Assume that* $h\leq s$, $\min_{1\leq k\leq K} n^{(m^*,k)}\geq n^{(m^*,0)}$, $\max_{0\leq k\leq K} s^2\log p/n^{(m^*,k)}=o(1)$*. Suppose that event* E_0' *in (D.1) holds and tuning parameters satisfy (D.2). Then with probability at least* $1-\exp(-c_1\log p)$*, it holds that*

$$\|\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}\|_2^2\lesssim\frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+\big(\max_{1\leq k\leq K}s(\lambda_0^{(k)})^2+\|\hat{\boldsymbol{\beta}}_0^{(0)}-\boldsymbol{\beta}\|_2^2\big)(\frac{s^2\log p}{n^{(m^*,0)}})^T.$$

Theorem B.1 provides the error contraction analysis of Algorithm B.1. The event E_0' in (D.1) assumes the consistency of initial estimators and specifies the tuning parameters. In fact, the tuning parameters of Algorithm B.1 depend on the convergence rates of initial estimators and hence depend on the unknown s and h. In the single-task first-order method with ℓ_1-regularization (Section 3.2 in Jordan et al. (2018)), the tuning parameters also depend on unknown parameters. In practice, specifying these tuning parameters can be challenging and the practical performance can be less accurate without proper tuning.

In the following two corollaries, we provide convergence rate analysis of Algorithm B.1 under two initializations proposed in Section 2.3.

Corollary B.1 (Convergence rate of Algorithm B.1 with single-site initialization)**.** *Assume Conditions 1, 2, and Condition D.1. Assume that* $h\leq s$, $\min_{1\leq k\leq K} n^{(m^*,k)}\geq n^{(m^*,0)}$, $\max_{0\leq k\leq K} s^2\log p/n^{(m^*,k)}=o(1)$*. Suppose that tuning parameters satisfy (D.2). Then with probability at least* $1-\exp(-c_1\log p)$*, it holds that for any finite* $T\geq 1$*,*

$$\|\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}\|_2^2\lesssim\frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+\left(\frac{s\log p}{\min_{1\leq k\leq K} n^{(m^*,k)}}+\frac{h\log p}{n^{(m^*,0)}}\right)(\frac{s^2\log p}{n^{(m^*,0)}})^T.$$

For $\hat{\boldsymbol{\beta}}_T$ obtained from Algorithm B.1, we see that it requires $O(\ln N/\ln n^{(m^*,0)})$ iterations to achieve the minimax optimal rate. We now compare the theoretical performance of Algorithm 1 and Algorithm B.1 with single-site initialization. In comparison to the upper bound derived in Corollary 1, we see that the convergence rate of Algorithm 1 is always no worse than the rate of Algorithm B.1 for any given T. Hence, to reach comparable performance, the local Hessian algorithm requires more iterations and hence more rounds of communication. This implies that transmitting Hessian matrices not only allows heterogeneous covariates but can accelerate the convergence of federated estimators.

Corollary B.2 (Convergence rate of Algorithm B.1 with multi-site initialization)**.** *Assume Conditions 1, 2, and Condition D.1. Assume that* $h \leq s$, $\min_{1\leq k\leq K} n^{(m^*,k)} \geq n^{(m^*,0)}$, $\max_{0\leq k\leq K} s^2 \log p/n^{(m^*,k)} = o(1)$. *Suppose that tuning parameters satisfy (D.2). Then with probability at least* $1-\exp(-c_1 \log p)$, *it holds that*

$$\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}\|_2^2 \lesssim \frac{s\log p}{N} + \frac{h\log p}{N^{(0)}} + \frac{s\log p}{\min_{0\leq k\leq K} n^{(I_k,k)}}(\frac{s^2\log p}{n^{(m^*,0)}})^T.$$

In Corollary B.2, we provide the convergence rate of Algorithm B.1 with multi-site initialization. In comparison to Corollary 2, the local Hessian algorithm has slower convergence rate at any given T. It requires $O(\ln N/\ln n^{(m^*,0)})$ iterations to achieve the minimax optimal rate.

C Proofs of theorems and lemmas in Section 3

We then prove the theorems and lemmas in the main paper.

Notations. Let $\boldsymbol{w}^{(0)} = \boldsymbol{\beta}$ and $\hat{\boldsymbol{w}}_t^{(0)} = \hat{\boldsymbol{\beta}}_t$. Let $\hat{\boldsymbol{u}}_t^{(k)} = \hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{w}^{(k)}$, $\hat{\boldsymbol{v}}_t^{(k)} = \hat{\boldsymbol{\delta}}_t^{(k)} - \boldsymbol{\delta}^{(k)}$. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$, define

$$\Delta^{(m,k)}(\boldsymbol{a},\boldsymbol{b}) = L^{(m,k)}(\boldsymbol{a}) - L^{(m,k)}(\boldsymbol{b}) - \langle \boldsymbol{a}-\boldsymbol{b}, \nabla L^{(m,k)}(\boldsymbol{b})\rangle.$$

We say that the restricted strong convexity (RSC) holds for $\Delta^{(k)}$ at $\boldsymbol{b}$ if

$$\Delta^{(k)}(\boldsymbol{a},\boldsymbol{b}) \geq n_k c_1 \|\boldsymbol{a}-\boldsymbol{b}\|_2^2 - \log p\|\boldsymbol{a}-\boldsymbol{b}\|_1^2. \tag{C.1}$$

For simplicity, let $\widehat{\boldsymbol{H}}_t^{(k)} = \widehat{\boldsymbol{H}}^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})$ and $\widehat{\boldsymbol{H}}_t^{(0)} = \widehat{\boldsymbol{H}}^{(0)}(\check{\boldsymbol{\beta}}_{t-1})$.

C.1 Proof of Lemma 1

Proof of Lemma 1. Step (i). Under the conditions of Lemma 1, it is easy to show that

$$\|\hat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}\|_2^2 \leq \frac{(s+h)\log p}{N^{(k)}},\ \|\hat{\boldsymbol{w}}^{(k)} - \boldsymbol{w}\|_1 \leq (s+h)\sqrt{\frac{\log p}{N^{(k)}}},\ k = 1,\dots,K \tag{C.2}$$

with probability at least $1-\exp(-c_1\log p)$.

Step (ii). Next, we show that for $k=1,\ldots,K$,

$$\|\check{\boldsymbol{\delta}}^{(k)}-\boldsymbol{\delta}^{(k)}\|_2^2 \lesssim \frac{h\log p}{N^{(0)}}. \tag{C.3}$$

The oracle inequality for $\hat{\boldsymbol{\delta}}^{(k)}$ is

$$\begin{aligned}
&\frac{1}{N^{(0)}}L^{(0)}(\hat{\boldsymbol{w}}^{(k)}+\hat{\boldsymbol{\delta}}^{(k)})-\frac{1}{N^{(0)}}L^{(0)}(\hat{\boldsymbol{w}}^{(k)}+\boldsymbol{\delta}^{(k)})-\frac{1}{N^{(0)}}|\langle\hat{\boldsymbol{v}}^{(k)},\nabla L^{(0)}(\boldsymbol{\beta})\rangle| \\
&\le \frac{1}{N^{(0)}}|\langle\hat{\boldsymbol{v}}^{(k)},\nabla L^{(0)}(\boldsymbol{\beta})\rangle|+\lambda_\delta\|\boldsymbol{\delta}^{(k)}\|_1-\lambda_\delta\|\hat{\boldsymbol{\delta}}^{(k)}\|_1.
\end{aligned} \tag{C.4}$$

Notice that

$$\text{LHS of (C.4)}=\frac{1}{N^{(0)}}\Delta^{(0)}(\boldsymbol{w}^{(k)}+\hat{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta})+Q(\hat{\boldsymbol{w}}^{(k)})-Q(\boldsymbol{w}^{(k)}),$$

where

$$Q(\boldsymbol{b})=\frac{1}{N^{(0)}}L^{(0)}(\boldsymbol{b}+\hat{\boldsymbol{\delta}}^{(k)})-\frac{1}{N^{(0)}}L^{(0)}(\boldsymbol{b}+\boldsymbol{\delta}^{(k)}).$$

By Taylor's expansion and the boundedness of $\ddot{\psi}(\cdot)$, for some constant $\rho_1\in[0,1]$,

$$\begin{aligned}
&|Q(\hat{\boldsymbol{w}}^{(k)})-Q(\boldsymbol{w}^{(k)})|=|\langle\hat{\boldsymbol{u}}^{(k)},\nabla Q(\boldsymbol{w}^{(k)}+\rho_1\hat{\boldsymbol{u}}^{(k)})\rangle| \\
&\quad=|\langle\hat{\boldsymbol{u}}^{(k)},\frac{1}{N^{(0)}}\nabla L^{(0)}(\boldsymbol{w}^{(k)}+\rho_1\hat{\boldsymbol{u}}^{(k)}+\hat{\boldsymbol{\delta}}^{(k)})-\frac{1}{N^{(0)}}\nabla L^{(0)}(\boldsymbol{w}^{(k)}+\rho_1\hat{\boldsymbol{u}}^{(k)}+\boldsymbol{\delta}^{(k)})\rangle| \\
&\quad\le\frac{C}{N^{(0)}}\sum_{i\in\mathcal{N}^{(0)}}|\boldsymbol{x}_i^\intercal\hat{\boldsymbol{u}}^{(k)}||\boldsymbol{x}_i^\intercal\hat{\boldsymbol{v}}^{(k)}|\le C\|\hat{\boldsymbol{v}}^{(k)}\|_1\|\boldsymbol{X}^{(0)}\|_{\infty,\infty}\frac{1}{N^{(0)}}\sum_{i\in\mathcal{N}^{(0)}}|\boldsymbol{x}_i^\intercal\hat{\boldsymbol{u}}^{(k)}|.
\end{aligned}$$

As $\hat{\boldsymbol{u}}^{(k)}$ is independent of $\boldsymbol{X}^{(0)}$, conditioning on $\hat{\boldsymbol{u}}^{(k)}$, $\boldsymbol{x}_i^\intercal\hat{\boldsymbol{u}}^{(k)}$ are independent sub-Gaussian with sub-Gaussian norm no larger than $\|\hat{\boldsymbol{u}}^{(k)}\|_2$. Hence, with probability at least $1-\exp(-c_1N^{(0)})$,

$$|Q(\hat{\boldsymbol{w}}^{(k)})-Q(\boldsymbol{w}^{(k)})|\le C\|\hat{\boldsymbol{v}}^{(k)}\|_1\sqrt{\frac{\|\hat{\boldsymbol{u}}^{(k)}\|_2^2}{N^{(0)}}}=o(1)\|\hat{\boldsymbol{v}}^{(k)}\|_1\lambda_\delta.$$

given that $\max_{1\le k\le K}\|\hat{\boldsymbol{u}}^{(k)}\|_2=O(1)$.

We arrive at the following oracle inequality

$$\frac{1}{N^{(0)}}\Delta^{(0)}(\boldsymbol{w}^{(k)}+\hat{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta})\le\frac{\lambda_\delta}{2}\|\hat{\boldsymbol{v}}^{(k)}\|_1+\lambda_\delta\|\boldsymbol{\delta}^{(k)}\|_1-\lambda_\delta\|\hat{\boldsymbol{\delta}}^{(k)}\|_1.$$

Standard analysis gives that with probability at least $1-\exp(-c_1\log p)$,

$$\|\hat{\boldsymbol{\delta}}^{(k)}-\boldsymbol{\delta}^{(k)}\|_2^2\lesssim\frac{h\log p}{N^{(0)}}.$$

By Lemma 17 in Yuan et al. (2018) and the condition $h \lesssim \sqrt{N^{(0)}/\log p}$, the proof for (C.3) is complete.

Step (iii). To ease our notation, define $\check{\boldsymbol{\delta}}^{(0)} = \boldsymbol{\delta}^{(0)} = 0$. Finally, the oracle inequality for $\hat{\boldsymbol{\beta}}$ is

$$\frac{1}{N}\sum_{k=0}^{K}\Delta^{(k)}(\hat{\boldsymbol{\beta}}+\check{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta}+\check{\boldsymbol{\delta}}^{(k)}) \quad \leq \frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)}+\check{\boldsymbol{v}}^{(k)})\rangle| + \lambda_\beta\|\boldsymbol{\beta}\|_1 - \lambda_\beta\|\hat{\boldsymbol{\beta}}\|_1. \tag{C.5}$$

Under the RSC of $\Delta^{(k)}(\hat{\boldsymbol{\beta}}+\check{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta}+\check{\boldsymbol{\delta}}^{(k)})$, we have

$$\text{LHS of (C.5)} \geq c_1\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 - c_2\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1^2\frac{K\log p}{N}.$$

For RHS of (C.5),

$$\begin{aligned}
\frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)}+\check{\boldsymbol{v}}^{(k)})\rangle| &\leq \frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)})\rangle| \\
&+ \frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)}+\check{\boldsymbol{v}}^{(k)}) - \sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)})\rangle| \\
&\leq \|\hat{\boldsymbol{u}}^{(0)}\|_1\frac{1}{N}\|\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)})\|_\infty + \frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=0}^{K}\nabla^2 L^{(k)}(\boldsymbol{w}^{(k)})\check{\boldsymbol{v}}^{(k)}\rangle| \\
&+ \frac{C}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{N}^{(k)}}\dddot{\psi}(\boldsymbol{x}_i^\intercal\boldsymbol{w}^{(k)})|\boldsymbol{x}_i^\intercal\check{\boldsymbol{v}}^{(k)}|^2|\boldsymbol{x}_i^\intercal\hat{\boldsymbol{u}}^{(0)}|,
\end{aligned}$$

where the last step is due to $\max_{1\leq k\leq K}\max_{i\in\mathcal{N}^{(k)}}|\boldsymbol{x}_i^\intercal\check{\boldsymbol{v}}^{(k)}| = O(h\sqrt{\log p/N^{(0)}}) = o(1)$ and Condition 2.

For the second term, using the sub-exponential property of $\widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{w}^{(k)})$, we have

$$\begin{aligned}
\frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=1}^{K}\nabla^2 L^{(k)}(\boldsymbol{w}^{(k)})\check{\boldsymbol{v}}^{(k)}\rangle| &\leq \frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=1}^{K}\frac{N^{(k)}}{N}\boldsymbol{H}^{(k)}(\boldsymbol{w}^{(k)})\check{\boldsymbol{v}}^{(k)}\rangle| \\
&+ \frac{1}{N}|\langle\hat{\boldsymbol{u}}^{(0)},\sum_{k=1}^{K}\frac{N^{(k)}}{N}\{\widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{w}^{(k)}) - \boldsymbol{H}^{(k)}(\boldsymbol{w}^{(k)})\}\check{\boldsymbol{v}}^{(k)}\rangle| \\
&\leq c_1\|\hat{\boldsymbol{u}}^{(0)}\|_2^2 + \frac{1}{c_1}\sum_{k=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2 + \|\hat{\boldsymbol{u}}^{(0)}\|_1\sum_{k=1}^{K}\frac{\sqrt{N^{(k)}\log p}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_1 \\
&\leq \|\hat{\boldsymbol{u}}^{(0)}\|_2^2 + \frac{1}{c_1}\sum_{k=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2 + c_2\|\hat{\boldsymbol{u}}^{(0)}\|_1\sqrt{\frac{\log p}{N}},
\end{aligned}$$

where the last step is due to $\max_{1\leq k\leq K}\|\check{\boldsymbol{v}}^{(k)}\|_1 = o(1)$ and K is finite.

For the last term, using the upper restricted eigenvalue condition on $\sum_{i\in N^{(k)}}\boldsymbol{x}_i\boldsymbol{x}_i^\intercal/N^{(k)}$ and $\max_{1\leq k\leq K}\|\check{\boldsymbol{v}}^{(k)}\|_0 \leq C\sqrt{N^{(0)}/\log p}$, we have for $\sqrt{N^{(0)}\log p} = o(\min_{1\leq k\leq K} N^{(k)})$,

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\mathcal{N}^{(k)}}\dddot{\psi}(\boldsymbol{x}_i^\intercal\boldsymbol{w}^{(k)})|\boldsymbol{x}_i^\intercal\check{\boldsymbol{v}}^{(k)}|^2|\boldsymbol{x}_i^\intercal\hat{\boldsymbol{u}}^{(0)}| \leq \|\hat{\boldsymbol{u}}^{(0)}\|_1\sum_{i=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2.$$

To summarize, for

$$\lambda_\beta \geq C_1\sqrt{\frac{\log p}{N}} + C_2\sum_{i=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2,$$

we have

$$c_1\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \leq \frac{3\lambda_\beta}{2}\|\hat{\boldsymbol{u}}_S^{(0)}\|_1 - \frac{\lambda_\beta}{2}\|\hat{\boldsymbol{u}}_{S^c}^{(0)}\|_1 + \sum_{k=0}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2.$$

As $\|\check{\boldsymbol{v}}^{(k)}\|_0 = O(\sqrt{N^{(0)}/\log p})$, we have $\|\check{\boldsymbol{v}}^{(k)}\|_0\log p/N^{(k)} = o(1)$ and $\frac{1}{N}\sum_{k=0}^{K}\|\boldsymbol{X}^{(k)}\check{\boldsymbol{v}}^{(k)}\|_2^2 \leq C\sum_{k=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2$. Standard analysis lead to

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \leq s\lambda_\beta^2 + C\sum_{k=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}^{(k)}\|_2^2$$

with probability at least $1-\exp(-c_1\log p)$. Notice that it suffices to take

$$\lambda_\beta \geq c_1\sqrt{\frac{\log p}{N}} + \frac{h\log p}{N^{(0)}}.$$

We arrive at

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2 \leq \frac{s\log p}{N} + \frac{h\log p}{N^{(0)}}(1+\frac{sh\log p}{N^{(0)}}).$$

As we assume $sh\log p = O(N^{(0)})$, the upper bound is established.

We are left to prove the RSC of $\Delta^{(k)}(\hat{\boldsymbol{\beta}}+\check{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta}+\check{\boldsymbol{\delta}}^{(k)})$, $k=1,\ldots,K$. Notice that

$$\begin{aligned}&\Delta^{(k)}(\hat{\boldsymbol{\beta}}+\check{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta}+\check{\boldsymbol{\delta}}^{(k)}) - \Delta^{(k)}(\hat{\boldsymbol{\beta}}+\boldsymbol{\delta}^{(k)},\boldsymbol{w}^{(k)})\\&= \frac{1}{N^{(k)}}\int_0^1\hat{\boldsymbol{u}}^\intercal[\widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{\beta}+\check{\boldsymbol{\delta}}^{(k)}+t\hat{\boldsymbol{u}}) - \widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{w}^{(k)}+t\hat{\boldsymbol{u}})]\hat{\boldsymbol{u}}(1-t)^2dt.\end{aligned}$$

As $\max_{1\leq k\leq K}\max_{i\in\mathcal{N}^{(k)}}|\boldsymbol{x}_i^\intercal\check{\boldsymbol{v}}^{(k)}| = o_P(1)$, it is easy to show that

$$\Delta^{(k)}(\hat{\boldsymbol{\beta}}+\check{\boldsymbol{\delta}}^{(k)},\boldsymbol{\beta}+\check{\boldsymbol{\delta}}^{(k)}) \geq \Delta^{(k)}(\hat{\boldsymbol{\beta}}+\boldsymbol{\delta}^{(k)},\boldsymbol{w}^{(k)}) - o_P(1)$$

and the RSC for the RHS follows from standard arguments.

□

C.2 Proof of Theorem 3.1

Lemma C.1. *Assume that $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \dots, N$ are independent sub-Gaussian random vectors with mean zero. Given that $\|\boldsymbol{u}\|_0 \le s_n$ for some $(s_n \vee \log p)^2 \le cN$,*

$$\mathbb{P}\left(\sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{u})^4 \le CN\|\boldsymbol{u}\|_2^4\right) \ge 1 - \exp\{-c\log p\}.$$

Proof of Lemma C.1. We use the concentration inequalities in Kuchibhotla and Chakrabortty (2018). By the sub-Gaussian property of $\boldsymbol{x}_i$, $(\boldsymbol{x}_i^\intercal \boldsymbol{u})^4$ is sub-Weibull(α) with 1/2. Let $\mathcal{T}$ be a fixed set with cardinality s_n. There are at most $\begin{pmatrix} p \\ s_n \end{pmatrix} \le C\exp\{\sqrt{s_n \log p}\}$ possible sets. Hence,

$$\mathbb{P}\left(\sup_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{u}\|_0 \le s_n} \sum_{i=1}^N (\boldsymbol{x}_i \boldsymbol{u})^4 \ge t\right) \le \exp\{\sqrt{s_n \log p}\} \max_{\mathcal{T}} \mathbb{P}\left(\sup_{\|\boldsymbol{u}\|_2=1, supp(\boldsymbol{u})=\mathcal{T}} \sum_{i=1}^N (\boldsymbol{x}_i \boldsymbol{u})^4 \ge t\right). \tag{C.6}$$

For a fixed $\mathcal{T}$, we consider an ϵ-net of $\{\|\boldsymbol{u}\|_2 = 1, supp(\boldsymbol{u}) = \mathcal{T}\}$ such that for any $u \in \{\|\boldsymbol{u}\|_2 = 1, supp(\boldsymbol{u}) = \mathcal{T}\}$, there is a vector $\boldsymbol{v} \in \mathcal{N}(\mathcal{T}, \epsilon)$ with $\|\boldsymbol{u} - \boldsymbol{v}\|_2 \le \epsilon$ for some constant $\epsilon > 0$. Hence,

$$\begin{aligned}
&|\sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{u})^4 - \sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{v})^4| \\
&\le |\sum_{i=1}^N \{\boldsymbol{x}_i(\boldsymbol{u}-\boldsymbol{v})\}^4| + 4|\sum_{i=1}^N \{\boldsymbol{x}_i(\boldsymbol{u}-\boldsymbol{v})\}^3\{\boldsymbol{x}_i\boldsymbol{v}\}| + 6|\sum_{i=1}^N \{\boldsymbol{x}_i(\boldsymbol{u}-\boldsymbol{v})\}^2\{\boldsymbol{x}_i\boldsymbol{v}\}^2| \\
&+ 4|\sum_{i=1}^N \{\boldsymbol{x}_i(\boldsymbol{u}-\boldsymbol{v})\}\{\boldsymbol{x}_i\boldsymbol{v}\}^3|.
\end{aligned}$$

Notice that $(\boldsymbol{u} - \boldsymbol{v})/\epsilon \in \{b : \|\boldsymbol{b}\|_2 = 1, supp(\boldsymbol{b}) = \mathcal{T}\}$ and hence,

$$|\sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{u})^4 - \sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{v})^4| \le (4\epsilon + 6\epsilon^2 + 4\epsilon^3 + \epsilon^4) \sup_{\|\boldsymbol{u}\|_2=1, supp(\boldsymbol{u})=\mathcal{T}} \sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{u})^4.$$

For small enough constant ϵ, we have

$$\sup_{\|\boldsymbol{u}\|_2=1, supp(u)=\mathcal{T}} \sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{u})^4 \le \max_{\boldsymbol{v} \in \mathcal{N}(\mathcal{T},\epsilon)} \sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{v})^4 + \frac{1}{2} \sup_{\|\boldsymbol{u}\|_2=1, supp(\boldsymbol{u})=\mathcal{T}} \sum_{i=1}^N (\boldsymbol{x}_i^\intercal \boldsymbol{u})^4.$$

Hence,

$$\max_{\mathcal{T}} \mathbb{P}\left(\sup_{\|u\|_2=1, supp(u)=\mathcal{T}} \sum_{i=1}^{N}((x_i u)^4 \geq t\right) \leq \max_{\mathcal{T}} \mathbb{P}\left(\max_{\boldsymbol{v}\in\mathcal{N}(\mathcal{T},\epsilon)} \sum_{i=1}^{N}((\boldsymbol{x}_i^\intercal \boldsymbol{v})^4 \geq t\right)$$
$$\leq |\mathcal{N}(\mathcal{T},\epsilon)| \max_{\mathcal{T}} \max_{\boldsymbol{v}\in\mathcal{N}(\mathcal{T},\epsilon)} \mathbb{P}\left(\sum_{i=1}^{N}((\boldsymbol{x}_i^\intercal \boldsymbol{v})^4 \geq t\right).$$

As $(\boldsymbol{x}_i^\intercal \boldsymbol{v})^4$ are sub-Weibull (α) with $\alpha = 1/2$,

$$\mathbb{P}\left(\sum_{i=1}^{N}((\boldsymbol{x}_i^\intercal \boldsymbol{v})^4 \geq N + \sqrt{Nt} + t^2\right) \leq \exp\{-t\}.$$

Using the fact that $|\mathcal{N}(\mathcal{T},\epsilon)| \leq \exp\{s_n\}$ and (C.6), we have

$$\mathbb{P}\left(\sup_{\|\boldsymbol{u}\|_2=1, \|\boldsymbol{u}\|_0 \leq \sqrt{N_1^{(0)}}/\log p} \sum_{i=1}^{N}(\boldsymbol{x}_i^\intercal \boldsymbol{u})^4 \geq N + \sqrt{Nt} + t^2\right) \leq \exp\{\sqrt{s_n \log p} + s_n - t\}. \quad \text{(C.7)}$$

Taking $t = s_n \vee \log p$, we arrive at desired results.

□

For $t = 1, \ldots, T$, define

$$E_t = \Bigg\{\max_{1\leq k\leq K} \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_1 = o(1),\ \lambda^{(k)} \geq \frac{2}{N^{(k)}}\|\nabla L^{(k)}(\boldsymbol{w}^{(k)})\|_\infty, k = 1, \ldots, K,$$
$$\text{RSC holds for } \widehat{\boldsymbol{H}}_t^{(k)}\Bigg\} \cap \{\text{events in Lemma C.1 hold}\}, \quad \text{(C.8)}$$

Lemma C.2 (Convergence rate of $\hat{\boldsymbol{w}}_t^{(k)}$)**.** *Assume Conditions 1 and 2 holds and* $s\log p/N^{(k)} = o(1)$ *for* $k = 1, \ldots, K$*. We take* $\lambda^{(k)} = c\sqrt{\log p/N^{(k)}}$*. Then (i) in event* E_t*,*

$$\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2 \lesssim (s+h)(\lambda^{(k)})^2 + \|\hat{\boldsymbol{u}}_{t-1}^{(k)}\|_2^4.$$

(ii) In event E_0 *defined in (C.16), then with probability at least* $1 - \exp(-c_1 \log p)$ *that*

$$\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2 \lesssim (s+h)(\lambda^{(k)})^2 + \|\hat{\boldsymbol{u}}_0^{(k)}\|_2^{4t},\ k = 1, \ldots, K.$$

Proof of Lemma C.2. The oracle inequality for $\hat{\boldsymbol{w}}_t^{(k)}$ is

$$\widehat{R}^{(k)}(\hat{\boldsymbol{w}}_t^{(k)}; \check{\boldsymbol{w}}_{t-1}^{(k)}) + \lambda^{(k)}\|\hat{\boldsymbol{w}}_t^{(k)}\|_1 \leq \widehat{R}^{(k)}(\boldsymbol{w}^{(k)}; \check{\boldsymbol{w}}_{t-1}^{(k)}) + \lambda^{(k)}\|\boldsymbol{w}^{(k)}\|_1.$$

It implies that

$$\frac{1}{2}\langle \hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(k)} \hat{\boldsymbol{u}}_t^{(k)}\rangle \leq \frac{1}{N^{(k)}}|\langle \hat{\boldsymbol{u}}_t^{(k)}, \nabla L(\check{\boldsymbol{w}}_{t-1}^{(k)}) - \widehat{\boldsymbol{H}}_t^{(k)}\check{\boldsymbol{u}}_{t-1}^{(k)}\rangle| + \lambda^{(k)}\|\boldsymbol{w}^{(k)}\|_1 - \lambda^{(k)}\|\hat{\boldsymbol{w}}_t^{(k)}\|_1$$
$$\leq \frac{\lambda^{(k)}}{2}\|\hat{\boldsymbol{u}}_t^{(k)}\|_1 + \lambda^{(k)}\|\boldsymbol{w}^{(k)}\|_1 - \lambda^{(k)}\|\hat{\boldsymbol{w}}_t^{(k)}\|_1 + |\langle \hat{\boldsymbol{u}}_t^{(k)}, \frac{1}{N^{(k)}}(\nabla L(\check{\boldsymbol{w}}_{t-1}^{(k)}) - \nabla L(\boldsymbol{w}^{(k)})) - \widehat{\boldsymbol{H}}_t^{(k)}\check{\boldsymbol{u}}_{t-1}^{(k)}\rangle|, \quad \text{(C.9)}$$

where the last step is due to the second statement in E_t.

For the RHS of (C.9), note that $supp(\boldsymbol{w}^{(k)}) \subseteq S \cup H_k$.

$$\begin{aligned}
&\text{RHS of (C.9)} \leq \frac{3\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{S\cup H_k}\|_1 - \frac{\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{\{S\cup H_k\}^c}\|_1 \\
&+ \langle \hat{\boldsymbol{u}}_t^{(k)}, \{\widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{w}^{(k)} + \rho_1 \check{\boldsymbol{u}}_{t-1}^{(k)}) - \widehat{\boldsymbol{H}}(\boldsymbol{w}^{(k)})\}\check{\boldsymbol{u}}_{t-1}^{(k)}\rangle \\
&\quad \leq \frac{3\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{S\cup H_k}\|_1 - \frac{\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{\{S\cup H_k\}^c}\|_1 + \frac{1}{N^{(k)}}\sum_{i\in\mathcal{N}^{(k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal \check{\boldsymbol{w}}_{t-1}^{(k)})|\boldsymbol{x}_i^\intercal \hat{\boldsymbol{u}}_t^{(k)}|\{\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(k)}\}^2,
\end{aligned}$$

where the last line is due to Condition 2 and $\|\boldsymbol{X}^{(k)}\|_{\infty,\infty}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_1 = o(1)$ in E_t.

Using the Cauchy-Schwartz on the last term and combining with the LHS of (C.9), we have for some small enough positive constant c_1,

$$c_1\langle \hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(k)}\hat{\boldsymbol{u}}_t^{(k)}\rangle \leq \frac{3\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{S\cup H_k}\|_1 - \frac{\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{\{S\cup H_k\}^c}\|_1 + \frac{c_2}{N^{(k)}}\sum_{i\in\mathcal{N}^{(k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal \check{\boldsymbol{w}}_{t-1}^{(k)})\{\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(k)}\}^4.$$

(i) If $\frac{3\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{S\cup H_k}\|_1 \geq c_2/N^{(k)}\sum_{i\in\mathcal{N}^{(k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal \check{\boldsymbol{w}}_{t-1}^{(k)})\{\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(k)}\}^4$, then we arrive at

$$c_1\langle \hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(k)}\hat{\boldsymbol{u}}_t^{(k)}\rangle \leq 3\lambda^{(k)}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{S\cup H_k}\|_1 - \frac{\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{\{S\cup H_k\}^c}\|_1.$$

Under the RSC condition of $\widehat{\boldsymbol{H}}_t^{(k)}$ in E_t, we arrive at for $(s+h)(\lambda^{(k)})^2 = o(1)$,

$$\langle \hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(k)}\hat{\boldsymbol{u}}_t^{(k)}\rangle \vee \|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2 \leq C(s+h)(\lambda^{(k)})^2. \tag{C.10}$$

(ii) If $\frac{3\lambda^{(k)}}{2}\|\{\hat{\boldsymbol{u}}_t^{(k)}\}_{S\cup H_k}\|_1 \leq c_2/N^{(k)}\sum_{i\in\mathcal{N}^{(k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal \check{\boldsymbol{w}}_{t-1}^{(k)})\{\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(k)}\}^4$, we leverage Lemma C.1 and the sparsity of $\check{\boldsymbol{u}}_{t-1}^{(k)}$. Specifically,

$$\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_0 \leq c_n + (s+h) \ll \sqrt{N^{(k)}}.$$

Hence, we arrive at

$$\frac{1}{N^{(k)}}\sum_{i\in\mathcal{N}^{(k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal \check{\boldsymbol{w}}_{t-1}^{(k)})|\{\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(k)}\}^4 \leq \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^4.$$

As $\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^4 = o(1)$, we have

$$\langle \hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(k)}\hat{\boldsymbol{u}}_t^{(k)}\rangle \vee \|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2 \lesssim \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^4. \tag{C.11}$$

Combining the arguments in (i) and (ii), we have in E_t,

$$\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2 \lesssim (s+h)(\lambda^{(k)})^2 + \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^4.$$

Since $c_n \geq s + h$, by Lemma 17 in Yuan et al. (2018), we have

$$\|\check{\boldsymbol{u}}_t^{(k)}\|_2^2 \lesssim \|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2.$$

Hence the result in (i) is proved.

It is left to verify that $\mathbb{P}(\cap_{t=1}^{\intercal} E_t) \geq \mathbb{P}(E_0) - \exp(-c_0 \log p)$ for any finite T with E_0 defined in (C.16). By our assumptions, it is easy to verify that $\mathbb{P}(E_1) \geq \mathbb{P}(E_0) - \exp(-c_0 \log p)$.

$$\begin{aligned}\mathbb{P}(E_1 \cap E_2) &\geq \mathbb{P}(E_1) - \mathbb{P}(\|\check{\boldsymbol{u}}_1^{(k)}\|_1 > c_1 | E_1) \\ &\geq \mathbb{P}(E_1) - \exp(-c_1 \log p)\end{aligned}$$

given that

$$s\lambda^{(k)} + \sqrt{s^2(\lambda^{(k)})^2 + s\|\check{\boldsymbol{u}}_0\|_2^4} \leq c_1.$$

As $s\lambda^{(k)} = O(1)$, it is easy to show that $\mathbb{P}(E_1 \cap E_2) \geq P(E_0) - \exp(-c_0 \log p)$. The rest of proofs follow by induction. □

For $t = 1, \ldots, T$, define

$$\begin{aligned}G_t = \Bigg\{ &\max_{0\leq k\leq K} \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_1 = o(1),\ \lambda^{(k)} \geq \frac{2}{N^{(k)}}\|\nabla L^{(k)}(\boldsymbol{w}^{(k)})\|_\infty, k = 1, \ldots, K, \lambda_\delta \geq \frac{2}{N^{(0)}}\|\nabla L^{(0)}(\boldsymbol{\beta})\|_\infty, \\ &\lambda_\beta \geq \frac{2}{N}\|\nabla L^{(0)}(\boldsymbol{\beta}) + \sum_{k=1}^K \nabla L^{(k)}(\boldsymbol{w}^{(k)})\|_\infty, \\ &\text{RSC holds for } \widehat{\boldsymbol{H}}_t^{(k)} \text{ and } \widehat{\boldsymbol{H}}_t^{(0)} \Bigg\} \cap \{\text{events in Lemma C.1 hold}\},\end{aligned} \tag{C.12}$$

Lemma C.3 (Convergence rate of $\hat{\boldsymbol{\delta}}_t^{(k)}$). *Assume Conditions 1 and 2 holds. We take $\lambda_\delta = c\sqrt{\log p/N^{(0)}}$. Then in event G_t,*

$$\|\check{\boldsymbol{v}}_t^{(k)}\|_2^2 \lesssim h\lambda_\delta^2 + \|\hat{\boldsymbol{u}}_{t-1}^{(0)}\|_2^4.$$

Proof of Lemma C.3. By the optimality of $\hat{\boldsymbol{\delta}}_t^{(k)}$,

$$\begin{aligned}&\frac{1}{2}(\hat{\boldsymbol{v}}_t^{(k)})^\intercal \widehat{\boldsymbol{H}}_t^{(0)} \hat{\boldsymbol{v}}_t^{(k)} \leq |\langle \hat{\boldsymbol{v}}_t^{(k)}, \frac{1}{N^{(0)}}\nabla L^{(0)}(\check{\boldsymbol{\beta}}_{t-1}) - \widehat{\boldsymbol{H}}^{(0)}(\check{\boldsymbol{\beta}}_{t-1})(\hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{\delta}^{(k)} - \check{\boldsymbol{\beta}}_{t-1})\rangle| \\ &\quad + \lambda_\delta\|\boldsymbol{\delta}^{(k)}\|_1 - \lambda_\delta\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1 \\ \Longrightarrow\ &\frac{1}{2}\langle \hat{\boldsymbol{v}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(0)} \hat{\boldsymbol{v}}_t^{(k)}\rangle \leq |\langle \hat{\boldsymbol{v}}_t^{(k)}, \frac{1}{N^{(0)}}\nabla L^{(0)}(\boldsymbol{\beta})\rangle| + \lambda_\delta\|\boldsymbol{\delta}^{(k)}\|_1 - \lambda_\delta\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1 \\ &\quad + \underbrace{|(\hat{\boldsymbol{v}}_t^{(k)})^\intercal\{\nabla L^{(0)}(\check{\boldsymbol{\beta}}_{t-1}) - \nabla L^{(0)}(\boldsymbol{\beta}) - \widehat{\boldsymbol{H}}_t^{(0)}(\hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{\delta}^{(k)} - \check{\boldsymbol{\beta}}_{t-1})\}|}_{F_t}.\end{aligned} \tag{C.13}$$

In event G_t,

$$\text{RHS of (C.13)} \leq F_t + \frac{\lambda_\delta}{4}\|\hat{\boldsymbol{v}}_t^{(k)}\|_1 + \lambda_\delta\|\boldsymbol{\delta}^{(k)}\|_1 - \lambda_\delta\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1,$$

where for some $\rho \in [0,1]$,

$$\begin{aligned} F_t &\le |(\hat{\boldsymbol{v}}_t^{(k)})^\intercal \{\widehat{\boldsymbol{H}}^{(0)}(\rho\check{\boldsymbol{\beta}}_{t-1} + (1-\rho)\boldsymbol{\beta})(\check{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}) - \widehat{\boldsymbol{H}}^{(0)}(\check{\boldsymbol{\beta}}_{t-1})(\hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{\delta}^{(k)} - \check{\boldsymbol{\beta}}_{t-1})\}| \\ &\le \underbrace{|(\hat{\boldsymbol{v}}_t^{(k)})^\intercal \widehat{\boldsymbol{H}}^{(0)}(\check{\boldsymbol{\beta}}_{t-1})\hat{\boldsymbol{u}}_t^{(k)}|}_{F_{1,t}} + \underbrace{|(\hat{\boldsymbol{v}}_t^{(k)})^\intercal \{\widehat{\boldsymbol{H}}^{(0)}(\rho\check{\boldsymbol{\beta}}_{t-1} + (1-\rho)\boldsymbol{\beta}) - \widehat{\boldsymbol{H}}^{(0)}(\check{\boldsymbol{\beta}}_{t-1})\}(\check{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta})|}_{F_{2,t}}. \end{aligned}$$

For $F_{1,t}$, we use the independence of $\hat{\boldsymbol{u}}_t^{(k)}$ and $\boldsymbol{X}^{(0)}$ to arrive at

$$F_{1,t} \le \|\hat{\boldsymbol{v}}_t^{(k)}\|_1 \frac{1}{N^{(0)}} \sum_{i \in \mathcal{N}^{(0)}} |\boldsymbol{x}_i^\intercal \hat{\boldsymbol{u}}^{(k)}| \le C\sqrt{\frac{\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2}{N_0}} = o(1)\|\hat{\boldsymbol{v}}_t^{(k)}\|_1 \lambda_\delta$$

with probability at least $1 - \exp(-c_1 N_0)$ as long as $\|\hat{\boldsymbol{u}}_t^{(k)}\|_2 = o(1)$.

For $F_{2,t}$, we use the Lipschitz property of $\dddot{\psi}$ and $\max_{i \in \mathcal{N}^{(0)}} |\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(0)}| \le C$ in G_t such that

$$\begin{aligned} F_{2,t} &\le \frac{1}{N^{(0)}} \sum_{i \in \mathcal{N}^{(0)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal \check{\boldsymbol{\beta}}_{t-1}) |\boldsymbol{x}_i^\intercal \hat{\boldsymbol{v}}_t^{(k)}| \{\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(0)}\}^2 \\ &\le \langle \check{\boldsymbol{v}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(0)} \check{\boldsymbol{v}}_t^{(k)} \rangle^{1/2} \sqrt{\frac{1}{N^{(0)}} \|\boldsymbol{X}^{(0)} \check{\boldsymbol{u}}_{t-1}^{(0)}\|_4^4}. \end{aligned}$$

Notice that

$$\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_0 \le c_n + s \le Cs \ll \sqrt{N^{(0)}}.$$

Using Young's inequality and Lemma C.1, we have

$$F_{2,t} \le c_1 \langle \check{\boldsymbol{v}}_{t-1}^{(k)}, \widehat{\boldsymbol{H}}_t^{(0)} \check{\boldsymbol{u}}_{t-1}^{(k)} \rangle + c_2 \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_4^4.$$

Using the RSC condition of $\widehat{\boldsymbol{H}}_t^{(0)}$ in G_t, we have the following oracle inequality

$$\|\boldsymbol{v}_t^{(k)}\|_2^2 \le c_1 \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^4 + \|\boldsymbol{\delta}^{(k)}\|\lambda_\delta - \|\hat{\boldsymbol{\delta}}_t^{(k)}\|\lambda_\delta. \tag{C.14}$$

Using the sparsity of $\boldsymbol{\delta}^{(k)}$, $h \log p = o(\sqrt{N^{(0)}})$, it is easy to show that

$$\|\hat{\boldsymbol{v}}_t\|_2^2 \le h\lambda_\delta^2 + c_1 \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^4.$$

As $\check{\boldsymbol{\beta}}_{t-1} = \mathcal{H}_{cs}(\hat{\boldsymbol{\beta}}_{t-1})$, we arrive at

$$\|\check{\boldsymbol{v}}_t\|_2^2 \le h\lambda_\delta^2 + c_1 \|\hat{\boldsymbol{u}}_{t-1}^{(0)}\|_2^4.$$

□

Proof of Theorem 3.1. Oracle inequality:

$$\begin{aligned}&\widehat{R}^{(0)}(\hat{\boldsymbol{\beta}}_t;\check{\boldsymbol{\beta}}_{t-1})+\sum_{k=1}^{K}\widehat{R}^{(k)}(\hat{\boldsymbol{\beta}}_t+\check{\boldsymbol{\delta}}_t^{(k)};\check{\boldsymbol{w}}_{t-1}^{(k)})+\lambda_\beta\|\hat{\boldsymbol{\beta}}_t\|_1\\ &\quad\le \widehat{R}^{(0)}(\boldsymbol{\beta};\check{\boldsymbol{\beta}}_{t-1})+\sum_{k=1}^{K}\widehat{R}^{(k)}(\boldsymbol{\beta}+\check{\boldsymbol{\delta}}_t^{(k)};\check{\boldsymbol{w}}_{t-1}^{(k)})+\lambda_\beta\|\boldsymbol{\beta}\|_1.\end{aligned}$$

Reorganizing the terms, we arrive at

$$\begin{aligned}&\frac{1}{2}(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})^\intercal\{\sum_{k=0}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(k)}\}(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})\\ &\le \underbrace{\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\frac{1}{N}\sum_{k=0}^{K}\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})-\sum_{k=1}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(k)}\{\boldsymbol{\beta}+\check{\boldsymbol{\delta}}_t^{(k)}-\check{\boldsymbol{w}}_{t-1}^{(k)}\}\rangle\big|}_{U_t}+\lambda_\beta\|\boldsymbol{\beta}\|_1-\lambda_\beta\|\hat{\boldsymbol{\beta}}_t\|_1.\end{aligned}$$

For U_t, we have

$$\begin{aligned}U_t&\le\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\frac{1}{N}\sum_{k=0}^{K}\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})-N^{(k)}\widehat{\boldsymbol{H}}_t^{(k)}\{\boldsymbol{w}^{(k)}-\check{\boldsymbol{w}}_{t-1}^{(k)}\}\rangle\big|+\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\sum_{k=1}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(k)}\{\check{\boldsymbol{\delta}}_t^{(k)}-\boldsymbol{\delta}^{(k)}\}\rangle\big|\\ &\le\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\frac{1}{N}\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)})\rangle\big|+\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\sum_{k=1}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(k)}\{\check{\boldsymbol{\delta}}_t^{(k)}-\boldsymbol{\delta}^{(k)}\}\rangle\big|\\ &\quad+\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\frac{1}{N}\sum_{k=0}^{K}\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})-\nabla L^{(k)}(\boldsymbol{w}^{(k)})-N^{(k)}\widehat{\boldsymbol{H}}_t^{(k)}\{\boldsymbol{w}^{(k)}-\check{\boldsymbol{w}}_{t-1}^{(k)}\}\rangle\big|\\ &\le\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}\|_1\frac{\lambda_\beta}{2}+\big|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\sum_{k=1}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(k)}\check{\boldsymbol{v}}_t^{(k)}\rangle\big|+\frac{1}{N}\left\{\sum_{k=0}^{K}\sum_{i\in\mathcal{N}^{(k)}}\ddot{\psi}(\boldsymbol{x}_i^\intercal\check{\boldsymbol{w}}_{t-1}^{(k)})|\boldsymbol{x}_i^\intercal(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})||\boldsymbol{x}_i^\intercal\check{\boldsymbol{u}}_{t-1}^{(k)}|^2\right\},\end{aligned}$$

where we use the Lipschitz condition and the first statement of G_t in the last step. Using Cauchy-Schwartz on the last two terms, we have

$$\begin{aligned}c_1(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})^\intercal\{\sum_{k=0}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(k)}\}(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})&\le\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}\|_1\frac{\lambda_\beta}{2}+\lambda_\beta\|\boldsymbol{\beta}\|_1-\lambda_\beta\|\hat{\boldsymbol{\beta}}_t\|_1\\ &\quad+4\sum_{k=1}^{K}\frac{N^{(k)}}{N}(\check{\boldsymbol{v}}_t^{(k)})^\intercal\widehat{\boldsymbol{H}}_t^{(k)}\check{\boldsymbol{v}}_t^{(k)}+c_1\frac{1}{N}\sum_{k=0}^{K}\sum_{i\in\mathcal{N}^{(k)}}\ddot{\psi}(\boldsymbol{x}_i^\intercal\check{\boldsymbol{w}}_{t-1}^{(k)})|\boldsymbol{x}_i^\intercal\check{\boldsymbol{u}}_{t-1}^{(k)}|^4\end{aligned}$$

By Lemma C.1 and the RSC condition on $\widehat{\boldsymbol{H}}_t^{(k)}$, we have

$$\begin{aligned}c_1\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}\|_2^2&\le\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}\|_1\frac{\lambda_\beta}{2}+\lambda_\beta\|\boldsymbol{\beta}\|_1-\lambda_\beta\|\hat{\boldsymbol{\beta}}_t\|_1+C\sum_{k=1}^{K}\frac{N_k}{N}\|\check{\boldsymbol{v}}_t^{(k)}\|_2^2\\ &\quad+C\sum_{k=0}^{K}\frac{N_k}{N}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^4.\end{aligned}$$

By Lemma C.3, we have for $N \geq KN^{(0)}$,

$$\begin{aligned}\sum_{k=1}^{K} \frac{N_k}{N} \|\check{\boldsymbol{v}}_t^{(k)}\|_2^2 &\leq \sum_{k=1}^{K} \frac{N_k}{N} \{h\lambda_\delta^2 + \|\hat{\boldsymbol{u}}_{t-1}^{(0)}\|_2^4\} \\ &\leq \frac{h \log p}{N^{(0)}} + \|\hat{\boldsymbol{u}}_{t-1}^{(0)}\|_2^4.\end{aligned}$$

Since K is finite, we arrive at

$$c_1 \|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|_2^2 \leq s\lambda_\beta^2 + \frac{h \log p}{N^{(0)}} + \|\hat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}\|_2^4. \tag{C.15}$$

By Lemma 17 in Yuan et al. (2018), since $c_n \geq s$,

$$\|\check{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|_2^2 \leq (1 + o(1)) \|\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|_2^2.$$

This shows that in event G_t, the results of Theorem 3.1 holds.

Finally, we show that $\mathbb{P}(\cap_{t=1}^{\mathrm{T}} G_t) \geq \mathbb{P}(E_0) - \exp(-c_1 \log p)$ for any fixed T. First notice that RSC conditions in each G_t is guaranteed by the first statement in G_t with probability at least $1 - \exp(-c_1 \log p)$. Specifically, for $T = 1$,

$$\mathbb{P}(G_1) \geq 1 - \exp(-c_1 \log p) - \mathbb{P}(\max_{0 \leq k \leq K} \|\check{\boldsymbol{u}}_0^{(k)}\|_1 \geq c_1) \geq \mathbb{P}(E_0) - \exp(-c_1 \log p)$$

by assumption. For $T = 2$,

$$\mathbb{P}(G_1 \cap G_2) \geq \mathbb{P}(G_1) - \mathbb{P}(\max_{0 \leq k \leq K} \|\check{\boldsymbol{u}}_1^{(k)}\|_1 \geq c_1 | G_1).$$

By the thresholding step, we have

$$\|\check{\boldsymbol{u}}_t^{(k)}\|_1 \leq \|\check{\boldsymbol{u}}_t^{(k)}\|_0^{1/2} \|\check{\boldsymbol{u}}_t^{(k)}\|_2 \leq \|\check{\boldsymbol{u}}_t^{(k)}\|_0^{1/2} \|\hat{\boldsymbol{u}}_t^{(k)}\|_2.$$

For $t = 1$,

$$\begin{aligned}\max_{1 \leq k \leq K} \|\check{\boldsymbol{u}}_1^{(k)}\|_1 &\leq \max_{1 \leq k \leq K} \|\check{\boldsymbol{u}}_1^{(k)}\|_0^{1/2} \|\hat{\boldsymbol{u}}_1^{(k)}\|_2 \\ &\leq \max_{1 \leq k \leq K} \sqrt{\{s + h + c_n\}\{\frac{(s+h)\log p}{N^{(k)}} + \|\check{\boldsymbol{u}}_0^{(k)}\|_2^4\}} = o(1)\end{aligned}$$

given that $c_n \geq s + h$ and

$$(s + h)\sqrt{\log p / N^{(k)}} = o(1) \text{ and } \max_{1 \leq k \leq K} \{s + c_n\} \|\check{\boldsymbol{u}}_0^{(k)}\|_2^4 = o(1).$$

For $t = 1$ and $k = 0$,

$$\|\check{\boldsymbol{u}}_1^{(0)}\|_1 \leq \sqrt{\{s + c_n\}\{\frac{s \log p}{N} + h\frac{\log p}{N^{(0)}} + \max_{0 \leq k \leq K} \|\check{\boldsymbol{u}}_0^{(k)}\|_2^4\}} = o(1)$$

given that

$$sh \log p = o(N^{(0)}) \text{ and } \{s + c_n\} \|\check{\boldsymbol{u}}_0^{(0)}\|_2^4 = o(1).$$

To summarize, we define an event

$$E_0 = \left\{ \max_{0 \le k \le K} s \|\hat{\boldsymbol{u}}_0^{(k)}\|_2^2 \le c_1, \max_{0 \le k \le K} s \|\hat{\boldsymbol{u}}_0^{(k)}\|_2^4 = o(1) \right\}. \tag{C.16}$$

The proof is complete when event E_0 holds. □

C.3 Proofs of Corollaries 1 and 2

Let $\check{\boldsymbol{w}}_0^{(k)} = \mathcal{H}_{c_n}(\hat{\boldsymbol{w}}_0^{(k)})$ for $\hat{\boldsymbol{w}}_0$ defined via (3.2).

Proof of Corollary 1. Under the conditions of Corollary 1, it is easy to show that

$$\|\hat{\boldsymbol{u}}_0^{(0)}\|_2^2 \lesssim \frac{s \log p}{N^{(m^*)}} + \frac{h \log p}{n^{(m^*,0)}}$$
$$\|\hat{\boldsymbol{u}}_0^{(k)}\|_2^2 \lesssim \max_{1 \le k \le K} \frac{s \log p}{n^{(m^*,k)}}.$$

Hence, E_0 holds as long as

$$\max_{1 \le k \le K} \frac{s^2 \log p}{n^{(m^*,k)}} + \frac{h \log p}{n^{(m^*,0)}} = o(1).$$

For prediction error, note that

$$\mathbb{E}_{x_*}[\{x_*^\intercal(\check{\boldsymbol{\beta}}_t - \boldsymbol{\beta})\}^2] \le \Lambda_{\max}(\mathbb{E}[x_* x_*^\intercal]) \|\check{\boldsymbol{\beta}}_t - \boldsymbol{\beta}\|_2^2.$$

As $\Lambda_{\max}(\mathbb{E}[x_* x_*^\intercal])$ is upper bounded by a constant, the proof is complete now.

□

Proof of Corollary 2. Under the conditions of Corollary 2, it is easy to show that

$$\max_{0 \le k \le K} \|\check{\boldsymbol{u}}_0^{(k)}\|_2^2 \lesssim \max_{0 \le k \le K} \frac{s \log p}{n^{(I_k,k)}}.$$

Hence, E_0 holds as long as

$$\max_{0 \le k \le K} \frac{s^2 \log p}{n^{(I_k,k)}} = o(1).$$

□

Proof of Remark 4.

$$\mathbb{P}(sgn(x_*^\intercal\hat{\boldsymbol{\beta}}_T) \neq sgn(x_*^\intercal\boldsymbol{\beta}^*)) \leq \mathbb{P}(|x_*^\intercal\boldsymbol{\beta}| \leq \epsilon) + \mathbb{P}(|x_*^\intercal\boldsymbol{\beta}| \geq \epsilon, sgn(x_*^\intercal\hat{\boldsymbol{\beta}}_T) \neq sgn(x_*^\intercal\boldsymbol{\beta}))$$
$$\leq \mathbb{P}(|x_*^\intercal\boldsymbol{\beta}| \leq \epsilon) + \mathbb{P}(|x_*^\intercal(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})| \geq \epsilon).$$

Conditioning on $\hat{\boldsymbol{\beta}}_T$, $x_*^\intercal(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta})$ is sub-Gaussian with sub-Gaussian norm $C\|\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}\|_2$.Taking

$$\epsilon = C_1\sqrt{\frac{s\log p}{N} + \frac{h\log p}{N^{(0)}} + \left\{\frac{s\log p}{N^{(m^*)}} + \frac{h\log p}{n^{(m^*,0)}}\right\}^{2T}}$$

for large enough constant C_1, we arrive at desired results.

□

D Proofs of convergence rate of algorithm B.1

Condition D.1 (Homogeneous covariates)**.** *Assume that* $\{\boldsymbol{x}_i\}_{i\in\mathcal{N}^{(m,k)}}$ *and* $\{\boldsymbol{x}_i\}_{i\in\mathcal{N}^{(m',k)}}$ *are identically distributed for any* $0 \leq k \leq K$ *and* $1 \leq m, m' \leq M$*.*

D.1 Convergence rate analysis

Define an event

$$E_0' = \left\{\max_{0\leq k\leq K}\|\check{\boldsymbol{u}}_0^{(k)}\|_1 = o(1), \|\check{\boldsymbol{u}}_0^{(k)}\|_2 \leq c\sqrt{s}\lambda_0^{(k)} = o(1), k = 1, \ldots, K, \ \|\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\|_2 = o(1)\right\}. \tag{D.1}$$

For the tuning parameters in Algorithm B.1, we take

$$\lambda_t^{(k)} \geq \sqrt{\frac{\log p}{N^{(k)}}} + s^t(\frac{\log p}{n^{(m^*,k)}})^{t/2}\lambda_0^{(k)}, \quad \lambda_{\delta,t}^{(k)} \geq \sqrt{\frac{\log p}{N^{(0)}}} + \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2\sqrt{\frac{s\log p}{n^{(m^*,0)}}} + \sqrt{\frac{\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2}{n^{(m^*,0)}}},$$
$$\lambda_{\beta,t} \geq \sqrt{\frac{\log p}{N}} + \max_{0\leq k\leq K}\sqrt{\frac{s\log p}{n^{(m^*,k)}}}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2. \tag{D.2}$$

Theorem D.1 (Convergence rate of Algorithm B.1)**.** *Assume Conditions 1, 2, and Condition D.1. Assume that* $n^{(m^*,1)} \geq n^{(m^*,0)}$ *and* $N^{(1)} \geq N^{(0)}$, $s^2\log p/N^{(m^*)}+hs\log p/n^{(m^*,0)} = o(1)$*. If event* E_0' *in (D.1) holds, then with probability at least* $1 - \exp(-c_1\log p)$*, it holds that*

$$\|\hat{\boldsymbol{u}}_t^{(0)}\|_2^2 \lesssim \frac{s\log p}{N} + \frac{h\log p}{N^{(0)}} + s^{2t+1}(\frac{\log p}{N^{(m^*)}})^t(\lambda_0^{(1)})^2 + s(\frac{s\log p}{N^{(m^*)}} + \frac{h\log p}{n^{(m^*,0)}})\|\hat{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2.$$

By induction, we have

$$\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}\|_2^2 \lesssim \frac{s\log p}{N} + \frac{h\log p}{N^{(0)}} + \max\{s(\lambda_0^{(1)})^2, \|\hat{\boldsymbol{u}}_0^{(0)}\|_2^2\}s^T(\frac{s\log p}{N^{(m^*)}} + \frac{h\log p}{n^{(m^*,0)}})^T.$$

D.2 Proof of Theorem B.1

Under the conditions of Theorem B.1, with probability at least $1-\exp(-c_1\log p)$, the initial values satisfy

$$\|\check{\boldsymbol{u}}_0^{(k)}\|_2^2 \lesssim \frac{s\log p}{n^{(m^*,k)}},\ \ \|\check{\boldsymbol{u}}_0^{(k)}\|_1 \lesssim s\sqrt{\frac{\log p}{n^{(m^*,k)}}},\ \ k=1,\ldots,K$$
$$\|\check{\boldsymbol{u}}_0^{(0)}\|_2^2 \lesssim \frac{s\log p}{N^{(m^*)}} + \frac{h\log p}{n^{(m^*,0)}},\ \ \|\check{\boldsymbol{u}}_0^{(0)}\|_1 \lesssim s\sqrt{\frac{\log p}{N^{(m^*)}}} + \sqrt{\frac{sh\log p}{n^{(m^*,0)}}}.$$

For $t=1,\ldots,T$, define

Lemma D.1. *Assume Conditions 1, 2, and Condition D.1. We take*

$$\lambda_t^{(k)} \geq \sqrt{\frac{\log p}{N^{(k)}}} + (s+h)\lambda_{t-1}^{(k)}\sqrt{\frac{\log p}{n^{(m^*,k)}}},\ \ k=1,\ldots,K.$$

If $s\sqrt{\log p/n^{(m^*,k)}} = o(1)$, then with

$$\lambda_t^{(k)} \geq \sqrt{\frac{\log p}{N^{(k)}}} + s^t(\frac{\log p}{n^{(m^*,k)}})^{t/2}\lambda_0^{(k)},$$

it holds that

$$\|\hat{\boldsymbol{u}}^{(k)}\|_2^2 \vee \|\check{\boldsymbol{u}}_t^{(k)}\|_2^2 \lesssim (s+h)(\lambda_t^{(k)})^2,\ \ \ \|\hat{\boldsymbol{u}}^{(k)}\|_1 \vee \|\check{\boldsymbol{u}}_t^{(k)}\|_1 \lesssim (s+h)\lambda_t^{(k)}$$

with probability at least $1-\exp(-c_1\log p)$.

Proof of Lemma D.1. It follows from (C.9) in Lemma C.2 that

$$\frac{1}{2}\langle\hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(m^*,k)}\hat{\boldsymbol{u}}_t^{(k)}\rangle \leq \frac{\lambda_t^{(k)}}{4}\|\hat{\boldsymbol{u}}_t^{(k)}\|_1 + \lambda_t^{(k)}\|\boldsymbol{w}^{(k)}\|_1 - \lambda_t^{(k)}\|\hat{\boldsymbol{w}}_t^{(k)}\|_1$$
$$+ |\langle\hat{\boldsymbol{u}}_t^{(k)}, \nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)}) - \nabla L^{(k)}(\boldsymbol{w}^{(k)}) - \widehat{\boldsymbol{H}}_t^{(m^*,k)}\check{\boldsymbol{u}}_{t-1}^{(k)}\rangle|. \tag{D.3}$$

The last term on the RHS of (D.3) can be upper bounded by $\|\hat{\boldsymbol{u}}_t^{(k)}\|_1\lambda_t^{(k)}$ for $\lambda_t^{(k)} \geq \|\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)}) - \nabla L^{(k)}(\boldsymbol{w}^{(k)}) - \widehat{\boldsymbol{H}}_t^{(m^*,k)}\check{\boldsymbol{u}}_{t-1}^{(k)}\|_\infty$. By (D.3), we have

$$\frac{1}{2}\langle\hat{\boldsymbol{u}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(m^*,k)}\hat{\boldsymbol{u}}_t^{(k)}\rangle \leq \frac{\lambda_t^{(k)}}{2}\|\hat{\boldsymbol{u}}_t^{(k)}\|_1 + \lambda_t^{(k)}\|\boldsymbol{w}^{(k)}\|_1 - \lambda_t^{(k)}\|\hat{\boldsymbol{w}}_t^{(k)}\|_1.$$

We now verify the RSC for $\widehat{\boldsymbol{H}}_t^{(m^*,k)}$ given that $(s+h)\log p = o(n^{(m^*,k)})$, $k=1,\ldots,K$. For any $\boldsymbol{u}\in\mathbb{R}^p$,

$$\boldsymbol{u}^\intercal\widehat{\boldsymbol{H}}_t^{(m^*,k)}\boldsymbol{u} \geq \boldsymbol{u}^\intercal\widehat{\boldsymbol{H}}^{(m^*,k)}(\boldsymbol{w}^{(k)})\boldsymbol{u} - |\boldsymbol{u}^\intercal\{\widehat{\boldsymbol{H}}^{(m^*,k)}(\check{\boldsymbol{w}}_{t-1}^{(k)}) - \widehat{\boldsymbol{H}}^{(m^*,k)}(\boldsymbol{w}^{(k)})\}\boldsymbol{u}|.$$

For $\check{\boldsymbol{u}}_{t-1}^{(k)}$ such that $\max_{i\in\mathcal{N}^{(m^*,k)}} |\boldsymbol{x}_i^\intercal \check{\boldsymbol{u}}_{t-1}^{(k)}| \lesssim \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_1 = o(1)$,

$$\boldsymbol{u}^\intercal \widehat{\boldsymbol{H}}_t^{(m^*,k)} \boldsymbol{u} \geq \boldsymbol{u}^\intercal \widehat{\boldsymbol{H}}^{(m^*,k)}(\boldsymbol{w}^{(k)})\boldsymbol{u}(1-o(1)).$$

Using the sub-Gaussian property of $\boldsymbol{x}_i$ and the positive definiteness of $\boldsymbol{H}^{(k)}(\boldsymbol{w}^{(k)})$, it is easy to show that with probability at least $1-\exp(-c_1\log p)$,

$$\sup_{\|\boldsymbol{u}_{S\cup H_k}\|_1 \geq 3\|\boldsymbol{u}_{(S\cup H_k)^c}\|_1} \boldsymbol{u}^\intercal \widehat{\boldsymbol{H}}^{(m^*,k)}(\boldsymbol{w}^{(k)})\boldsymbol{u} \geq C\|\boldsymbol{u}\|_2^2$$

given that $(s+h)\log p = o(n^{(m^*,k)})$. Standard analysis lead to

$$\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2 \leq C(s+h)(\lambda_t^{(k)})^2 \;\text{ and } \|\hat{\boldsymbol{u}}_t^{(k)}\|_1 \leq C(s+h)\lambda_t^{(k)}. \tag{D.4}$$

It is left to find $\lambda_t^{(k)}$.

$$\begin{aligned}
&\|\{\widehat{\boldsymbol{H}}^{(k)}(c\check{\boldsymbol{w}}_{t-1}^{(k)} + (1-c)\boldsymbol{w}^{(k)}) - \widehat{\boldsymbol{H}}^{(m^*,k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})\}\check{\boldsymbol{u}}_{t-1}^{(k)}\|_\infty \\
&\leq \|\boldsymbol{X}^{(k)}\|_{\infty,\infty}\frac{1}{n^{(m^*,k)}}\sum_{i\in\mathcal{N}^{(m^*,k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal\check{\boldsymbol{w}}_{t-1}^{(k)})\{\boldsymbol{x}_i^\intercal\check{\boldsymbol{u}}_{t-1}^{(k)}\}^2 + \|\boldsymbol{X}^{(k)}\|_{\infty,\infty}\frac{1}{N^{(k)}}\sum_{i\in\mathcal{N}^{(k)}} \dddot{\psi}(\boldsymbol{x}_i^\intercal\check{\boldsymbol{w}}_{t-1}^{(k)})\{\boldsymbol{x}_i^\intercal\check{\boldsymbol{u}}_{t-1}^{(k)}\}^2 \\
&\quad + \|\{\widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{w}^{(k)}) - \widehat{\boldsymbol{H}}^{(m^*,k)}(\boldsymbol{w}^{(k)})\}\check{\boldsymbol{u}}_{t-1}^{(k)}\|_\infty \\
&\lesssim \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2 + \|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_1\sqrt{\frac{\log p}{n^{(m^*,k)}}} \\
&\lesssim (s+h)(\lambda_{t-1}^{(k)})^2 + (s+h)\lambda_{t-1}^{(k)}\sqrt{\frac{\log p}{n^{(m^*,k)}}},
\end{aligned}$$

where the last step is due to $\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_0 \asymp s+h$ and the upper restricted eigenvalue condition holds.

Hence, if $(s+h)\sqrt{\frac{\log p}{n^{(m^*,k)}}} = o(1)$, then it suffices to take

$$\lambda_t^{(k)} \gtrsim \sqrt{\frac{\log p}{N^{(k)}}} + (s+h)\lambda_{t-1}^{(k)}\sqrt{\frac{\log p}{n^{(m^*,k)}}}.$$

□

Lemma D.2. *Assume Conditions 1, 2, and D.1. We take*

$$\lambda_{\delta,t}^{(k)} \geq \sqrt{\frac{\log p}{N^{(0)}}} + \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2\sqrt{\frac{s\log p}{n^{(m^*,0)}}} + \sqrt{\frac{\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2}{n^{(m^*,0)}}}.$$

If $s\log p/n^{(m^*,k)} + h\log p/n^{(m^*,0)} = o(1)$, *then with probability at least* $1-\exp(-c_1\log p)$,

$$\|\check{\boldsymbol{v}}_t^{(k)}\|_2^2 \lesssim h(\lambda_{\delta,t}^{(k)})^2.$$

Proof of Lemma D.2. Oracle inequality: By the optimality of $\hat{\boldsymbol{\delta}}_t^{(k)}$,

$$\begin{aligned}
&\frac{1}{2}(\hat{\boldsymbol{v}}_t^{(k)})^{\intercal}\widehat{\boldsymbol{H}}_t^{(m^*,0)}\hat{\boldsymbol{v}}_t^{(k)} \leq |\langle \hat{\boldsymbol{v}}_t^{(k)}, (\nabla L^{(0)}(\check{\boldsymbol{\beta}}_{t-1}) - \widehat{\boldsymbol{H}}^{(m^*,0)}(\check{\boldsymbol{\beta}}_{t-1})(\hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{\delta}^{(k)} - \check{\boldsymbol{\beta}}_{t-1})\rangle| \\
&\quad + \lambda_{\delta,t}^{(k)}\|\boldsymbol{\delta}^{(k)}\|_1 - \lambda_{\delta,t}^{(k)}\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1 \\
\Longrightarrow &\frac{1}{2}\langle \hat{\boldsymbol{v}}_t^{(k)}, \widehat{\boldsymbol{H}}_t^{(m^*,0)}\hat{\boldsymbol{v}}_t^{(k)}\rangle \leq |\langle \hat{\boldsymbol{v}}_t^{(k)}, \nabla L^{(0)}(\boldsymbol{\beta})\rangle| + \lambda_{\delta,t}^{(k)}\|\boldsymbol{\delta}^{(k)}\|_1 - \lambda_{\delta,t}^{(k)}\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1 \\
&\quad + \underbrace{|(\hat{\boldsymbol{v}}_t^{(k)})^{\intercal}\{\nabla L^{(0)}(\check{\boldsymbol{\beta}}_{t-1}) - \nabla L^{(0)}(\boldsymbol{\beta}) - \widehat{\boldsymbol{H}}_t^{(m^*,0)}(\hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{\delta}^{(k)} - \check{\boldsymbol{\beta}}_{t-1})\}|}_{F_t}. \qquad \text{(D.5)}
\end{aligned}$$

In event G_t,

$$\text{RHS of (C.13)} \leq F_t + \frac{\lambda_{\delta,t}^{(k)}}{4}\|\hat{\boldsymbol{v}}_t^{(k)}\|_1 + \lambda_{\delta,t}\|\boldsymbol{\delta}^{(k)}\|_1 - \lambda_{\delta,t}^{(k)}\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1,$$

where

$$\begin{aligned}
F_t &\leq |(\hat{\boldsymbol{v}}_t^{(k)})^{\intercal}\{\widehat{\boldsymbol{H}}^{(0)}(\rho\check{\boldsymbol{\beta}}_{t-1} + (1-\rho)\boldsymbol{\beta})(\check{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}) - \widehat{\boldsymbol{H}}^{(m^*,0)}(\check{\boldsymbol{\beta}}_{t-1})(\hat{\boldsymbol{w}}_t^{(k)} - \boldsymbol{\delta}^{(k)} - \check{\boldsymbol{\beta}}_{t-1})\}| \\
&\leq \underbrace{|(\hat{\boldsymbol{v}}_t^{(k)})^{\intercal}\widehat{\boldsymbol{H}}^{(m^*,0)}(\check{\boldsymbol{\beta}}_{t-1})\hat{\boldsymbol{u}}_t^{(k)}|}_{F_{1,t}} + \underbrace{|(\hat{\boldsymbol{v}}_t^{(k)})^{\intercal}\{\widehat{\boldsymbol{H}}^{(0)}(\rho\check{\boldsymbol{\beta}}_{t-1} + (1-\rho)\boldsymbol{\beta}) - \widehat{\boldsymbol{H}}^{(m^*,0)}(\check{\boldsymbol{\beta}}_{t-1})\}(\check{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta})|}_{F_{2,t}}.
\end{aligned}$$

Similar analysis of $F_{1,t}$ as in Lemma C.3,

$$F_{1,t} \lesssim c_1\|\hat{\boldsymbol{v}}_t^{(k)}\|_1\sqrt{\frac{\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2}{n^{(m^*,0)}}}.$$

For $F_{2,t}$, we have

$$\begin{aligned}
F_{2,t} &\leq \|\hat{\boldsymbol{v}}_t^{(k)}\|_1\|\{\widehat{\boldsymbol{H}}^{(0)}(\rho\hat{\boldsymbol{\beta}}_{t-1} + (1-\rho)\boldsymbol{\beta}) - \widehat{\boldsymbol{H}}^{(m^*,0)}(\check{\boldsymbol{\beta}}_{t-1})\}\check{\boldsymbol{u}}_{t-1}^{(0)}\|_\infty \\
&\leq \|\hat{\boldsymbol{v}}_t^{(k)}\|_1\left\{\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_1\sqrt{\frac{\log p}{n^{(m^*,0)}}} + \frac{1}{n^{(m^*,0)}}\|\boldsymbol{X}^{(m^*,0)}\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2 + \frac{1}{N^{(0)}}\|\boldsymbol{X}^{(0)}\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2\right\}.
\end{aligned}$$

Using the sparsity of $\check{\boldsymbol{\beta}}_{t-1}^{(0)}$, we have for

$$\begin{aligned}
\lambda_{\delta,t}^{(k)} &\geq \sqrt{\frac{\log p}{N^{(0)}}} + \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_1\sqrt{\frac{\log p}{n^{(m^*,0)}}} + \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2 + \sqrt{\frac{\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2}{n^{(m^*,0)}}} \\
&\geq \sqrt{\frac{\log p}{N^{(0)}}} + \|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2\sqrt{\frac{s\log p}{n^{(m^*,0)}}} + \sqrt{\frac{\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2}{n^{(m^*,0)}}},
\end{aligned}$$

$$\frac{1}{2}(\hat{\boldsymbol{v}}_t^{(k)})^{\intercal}\widehat{\boldsymbol{H}}_t^{(m,0)}\hat{\boldsymbol{v}}_t^{(k)} \leq \frac{\lambda_{\delta,t}}{2}\|\hat{\boldsymbol{v}}_t^{(k)}\|_1 + \lambda_{\delta,t}\|\boldsymbol{\delta}_t^{(k)}\|_1 - \lambda_{\delta,t}\|\hat{\boldsymbol{\delta}}_t^{(k)}\|_1 + \|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2.$$

Standard analysis leads to

$$\|\hat{\boldsymbol{v}}_t^{(k)}\|_2^2 \lesssim (h\lambda_{\delta,t}^{(k)})^2.$$

For the thresholded $\check{\boldsymbol{v}}_t^{(k)}$, since $\sqrt{n^{(m^*,0)}} \geq h$, by Lemma 17 of Yuan et al. (2018) we have

$$\|\check{\boldsymbol{v}}_t^{(k)}\|_2^2 \lesssim h(\lambda_{\delta,t}^{(k)})^2.$$

□

Proof of Theorem B.1. Oracle inequality:

$$\begin{aligned}
&\frac{N^{(0)}}{N}\widehat{R}^{(0,local)}(\hat{\boldsymbol{\beta}}_t;\check{\boldsymbol{\beta}}_{t-1}) + \sum_{k=1}^{K}\frac{N^{(k)}}{N}\widehat{R}^{(k,local)}(\hat{\boldsymbol{\beta}}_t+\check{\boldsymbol{\delta}}_t^{(k)};\check{\boldsymbol{w}}_{t-1}^{(k)}) + \lambda_{\beta,t}\|\hat{\boldsymbol{\beta}}_t\|_1 \\
&\leq \frac{N^{(0)}}{N}\widehat{R}^{(0,local)}(\boldsymbol{\beta};\check{\boldsymbol{\beta}}_{t-1}) + \sum_{k=1}^{K}\frac{N^{(k)}}{N}\widehat{R}^{(k,local)}(\boldsymbol{\beta}+\check{\boldsymbol{\delta}}_t^{(k)};\check{\boldsymbol{w}}_{t-1}^{(k)}) + \lambda_{\beta,t}\|\boldsymbol{\beta}\|_1.
\end{aligned}$$

Reorganizing the terms, we arrive at

$$\begin{aligned}
&\frac{1}{2}(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})^\intercal\{\sum_{k=0}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(m^*,k)}\}(\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}) \\
&\leq |\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\frac{1}{N}\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)})\rangle| \\
&\quad+\underbrace{\left|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\frac{1}{N}\sum_{k=0}^{K}\{\nabla L^{(k)}(\check{\boldsymbol{w}}_{t-1}^{(k)})-\nabla L^{(k)}(\boldsymbol{w}^{(k)})-N^{(k)}\widehat{\boldsymbol{H}}_t^{(m^*,k)}(\boldsymbol{\beta}+\check{\boldsymbol{\delta}}_t^{(k)}-\check{\boldsymbol{w}}_{t-1}^{(k)})\}\right|}_{J_t} \\
&\quad+\lambda_{\beta,t}\|\boldsymbol{\beta}\|_1-\lambda_{\beta,t}\|\hat{\boldsymbol{\beta}}_t\|_1,
\end{aligned}$$

where

$$\begin{aligned}
J_t &\leq |\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\sum_{k=0}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}^{(k)}(\boldsymbol{w}^{(k)}+\rho\check{\boldsymbol{u}}_{t-1}^{(k)})\check{\boldsymbol{u}}_{t-1}^{(k)}\rangle-\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\sum_{k=0}^{K}\frac{N^{(k)}}{N}\widehat{\boldsymbol{H}}_t^{(m^*,k)}(\check{\boldsymbol{u}}_{t-1}^{(k)}+\check{\boldsymbol{v}}_t^{(k)})\rangle| \\
&\leq \sum_{k=1}^{K}\frac{N^{(k)}}{N}|\langle\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta},\widehat{\boldsymbol{H}}_t^{(m^*,k)}\check{\boldsymbol{v}}_t^{(k)}\rangle|+\|\hat{\boldsymbol{u}}_t^{(0)}\|_1\max_{0\leq k\leq K}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2 \\
&\quad+\|\hat{\boldsymbol{u}}_t^{(0)}\|_1\|\sum_{k=0}^{K}\frac{N^{(k)}}{N}(\widehat{\boldsymbol{H}}^{(k)}-\widehat{\boldsymbol{H}}^{(m^*,k)})\check{\boldsymbol{u}}_{t-1}^{(k)}\|_\infty
\end{aligned}$$

Hence, for

$$\begin{aligned}
\lambda_{\beta,t} &\geq \sqrt{\frac{\log p}{N}}+\max_{0\leq k\leq K}\sqrt{\frac{s\log p}{n^{(m^*,k)}}}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2 \\
&\geq \frac{2}{N}\|\sum_{k=0}^{K}\nabla L^{(k)}(\boldsymbol{w}^{(k)})\|_\infty+K\max_{0\leq k\leq K}\sqrt{\frac{\log p}{n^{(m^*,k)}}}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_1+\max_{0\leq k\leq K}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2,
\end{aligned}$$

we have

$$
\begin{aligned}
\|\hat{\boldsymbol{\beta}}_t-\boldsymbol{\beta}\|_2^2 &\lesssim s(\lambda_{\beta,t})^2+\sum_{k=1}^{K}\frac{N^{(k)}}{N}\|\check{\boldsymbol{v}}_t^{(1)}\|_2^2\\
&\lesssim s(\lambda_{\beta,t})^2+\sum_{k=1}^{K}\frac{N^{(k)}}{N}h(\lambda_{\delta,t}^{(k)})^2\\
&\lesssim s(\lambda_{\beta,t})^2+\frac{h\log p}{N^{(0)}}+\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2\frac{hs\log p}{n^{(m^*,0)}}+\frac{h}{n^{(m^*,0)}}\sum_{k=1}^{K}\frac{N^{(k)}}{N}\|\hat{\boldsymbol{u}}_t^{(k)}\|_2^2\\
&\lesssim s(\lambda_{\beta,t})^2+\frac{h\log p}{N^{(0)}}+\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2\frac{hs\log p}{n^{(m^*,0)}}+\frac{h}{n^{(m^*,0)}}(\frac{s\log p}{N}+\max_{1\le k\le K}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2\frac{s\log p}{n^{(m^*,k)}}).
\end{aligned}
$$

To summarize,

$$
\begin{aligned}
\|\hat{\boldsymbol{u}}_t^{(0)}\|_2^2 \lesssim{}& \frac{s\log p}{N}+\max_{0\le k\le K}\frac{s^2\log p}{n^{(m^*,k)}}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2+\frac{h\log p}{N^{(0)}}+\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2\frac{hs\log p}{n^{(m^*,0)}}\\
&+\frac{h}{n^{(m^*,0)}}(\frac{s\log p}{N}+\max_{1\le k\le K}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2\frac{s\log p}{n^{(m^*,k)}}).\\
\le{}& \frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+\frac{s^2\log p}{n^{(m^*,0)}}\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2+\max_{1\le k\le K}\frac{s^2\log p}{n^{(m^*,k)}}\|\check{\boldsymbol{u}}_{t-1}^{(k)}\|_2^2.
\end{aligned}
$$

By Lemma D.1, we arrive at

$$
\|\hat{\boldsymbol{u}}_t^{(0)}\|_2^2 \lesssim \frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+\max_{1\le k\le K}s^{2t+1}(\frac{\log p}{n^{(m^*,k)}})^t(\lambda_0^{(k)})^2+\frac{s^2\log p}{n^{(m^*,0)}}\|\check{\boldsymbol{u}}_{t-1}^{(0)}\|_2^2.
$$

Let $b_n=\frac{s^2\log p}{n^{(m^*,0)}}$. Given the results of Theorem B.1, it is easy to show

$$
\begin{aligned}
\|\hat{\boldsymbol{u}}_T^{(0)}\|_2^2 &\lesssim (\frac{s\log p}{N}+\frac{h\log p}{N^{(0)}})\sum_{t=1}^{\intercal}b_n^{T-t}+\max_{1\le k\le K}(\lambda_0^{(k)})^2\sum_{t=1}^{\intercal}s^{2t+1}(\frac{\log p}{n^{(m^*,k)}})^t b_n^{T-t}+b_n^{\intercal}\|\check{\boldsymbol{u}}_0^{(0)}\|_2^2\\
&\lesssim \frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+\max_{1\le k\le K}s(\lambda_0^{(k)})^2Tb_n^{\intercal}+b_n^{\intercal}\|\check{\boldsymbol{u}}_0^{(0)}\|_2^2\\
&\lesssim \frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+(\max_{1\le k\le K}s(\lambda_0^{(k)})^2+\|\check{\boldsymbol{u}}_0^{(0)}\|_2^2)(\frac{s^2\log p}{n^{(m^*,0)}})^T
\end{aligned}
$$

□

D.3 Proofs of Corollary B.1 and Corollary B.2

Proof of Corollary B.1. Notice that $\lambda_0^{(k)}\asymp\sqrt{\log p/n^{(m^*,k)}}$ for $k=1,\dots,K$. and $\|\check{\boldsymbol{u}}_0^{(0)}\|_2^2=O_P(s\log p/N^{(m^*)}+h\log p/n^{(m^*,0)})$. It is easy to show

$$
\|\hat{\boldsymbol{u}}_T^{(0)}\|_2^2 \lesssim \frac{s\log p}{N}+\frac{h\log p}{N^{(0)}}+(\frac{s\log p}{\min_{1\le k\le K}n^{(m^*,k)}}+\frac{h\log p}{n^{(m^*,0)}})(\frac{s^2\log p}{n^{(m^*,0)}})^T.
$$

□

Proof of Corollary B.2. Notice that $\lambda_0^{(k)} \asymp \sqrt{\log p / n^{(I_k,k)}}$ and $\|\check{\boldsymbol{u}}_0^{(0)}\|_2^2 = O_P(s \log p / n^{(I_0,0)})$.

Given the results of Theorem B.1, it is easy to show

$$\|\hat{\boldsymbol{u}}_T^{(0)}\|_2^2 \lesssim \frac{s\log p}{N} + \frac{h\log p}{N^{(0)}} + \frac{s\log p}{\min_{0\le k\le K} n^{(I_k,k)}} (\frac{s^2\log p}{n^{(m^*,0)}})^T.$$

□

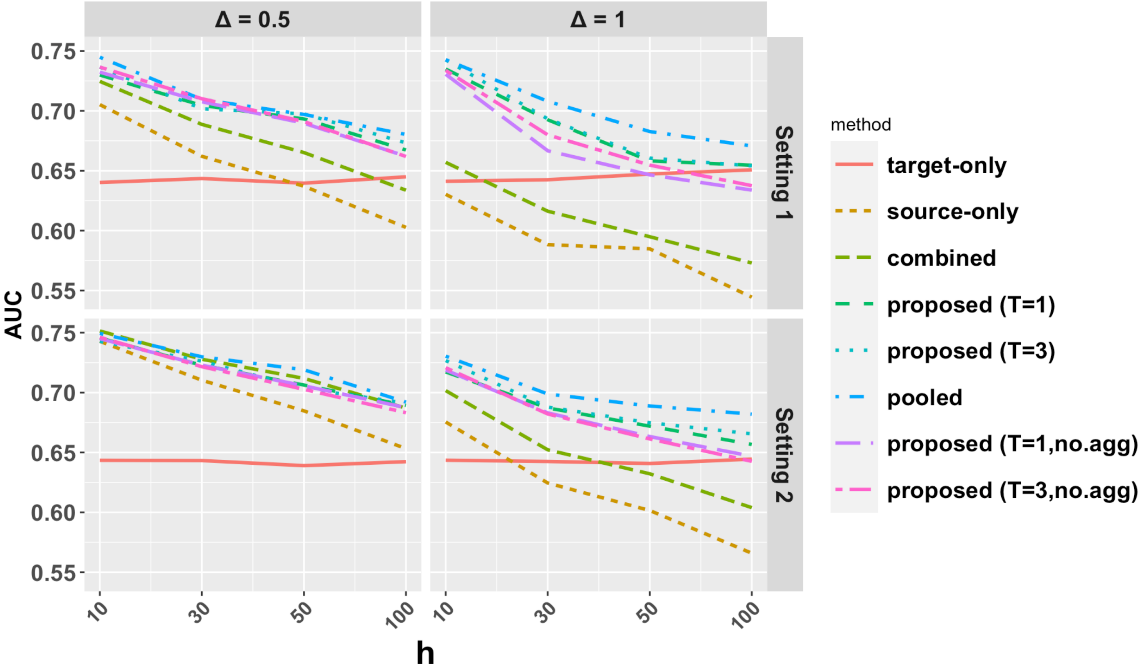
E Additional simulation results

To generate genotypes for data in the source population, we first generate p-dimensional multivariate Gaussian vector $\boldsymbol{z}_i$ with mean $\mathbf{0}$ and covariance matrix Σ_1. We choose Σ_1 to be a block-wise matrix with 20 blocks each has dimension 100×100. We set the all the 20 blocks to be the same, denoted by B_1, where $B_{1,ij} = 0.5^{|i-j|}$. We then randomly generate minor allele frequencies for the p genetic variants from $U(0, 0.5)$. Then we obtain $\boldsymbol{x}_i$ by categorize each $\boldsymbol{z}_i$ into $0, 1$ and 2 based on the corresponding minor allele frequencies. For the target data, we follow the same procedure with Σ_1 replaced by Σ_0, which has 100 blocks each with dimension 50×50. We set the block to be $B_{0,ij} = 0.3^{|i-j|}$.

In this section, we include more simulation results. In addition to methods compared in the main paper, we added three methods, which are (1) the proposed approach with $T = 1$ without aggregation (*FETA (T = 1, no agg)*); (5) the proposed approach with $T = 3$ without aggregation (*FETA (T = 3, no agg)*); (6) the pooled transfer learning method without aggregation (*pooled (no.agg)*).

Compared with the proposed methods considered in the main paper, these three methods are performed without the aggregation step. The methods are evaluated based on their mean squared error (MSE) and the out-sample area under the receiver operating characteristic curve (AUC) based on a randomly generated testing sample with sample size $n = 1000$. In this simulation, we used 10% of the total sample size of the target population in the leading site as a validation dataset to learn the weights introduced in equation (2.9) in the main paper. From Figures (E.1) - (E.2), we observed that FETA with aggregation perform no worse than the corresponding methods without aggregation. When the level of heterogeneity is low, FETA without aggregation has comparable performance as FETA with aggregation. However, as we expected, as the level of heterogeneity increases, the robustness of the proposed methods is shown to be better than FETA without aggregation. When the heterogeneity is large, the proposed federated transfer learning algorithms without aggregation have much poor performance and a large variation than the pooled transfer learning methods without aggregation. The variability reduces when the number of iteration increases.

Figure E.1: Comparison of AUC over 200 replications.



In sum, the additional simulation results demonstrated that the aggregation step is necessary unless there is strong prior knowledge supporting that the level of heterogeneity is low.

F Additional details in the eMERGE data application

F.1 Data processing

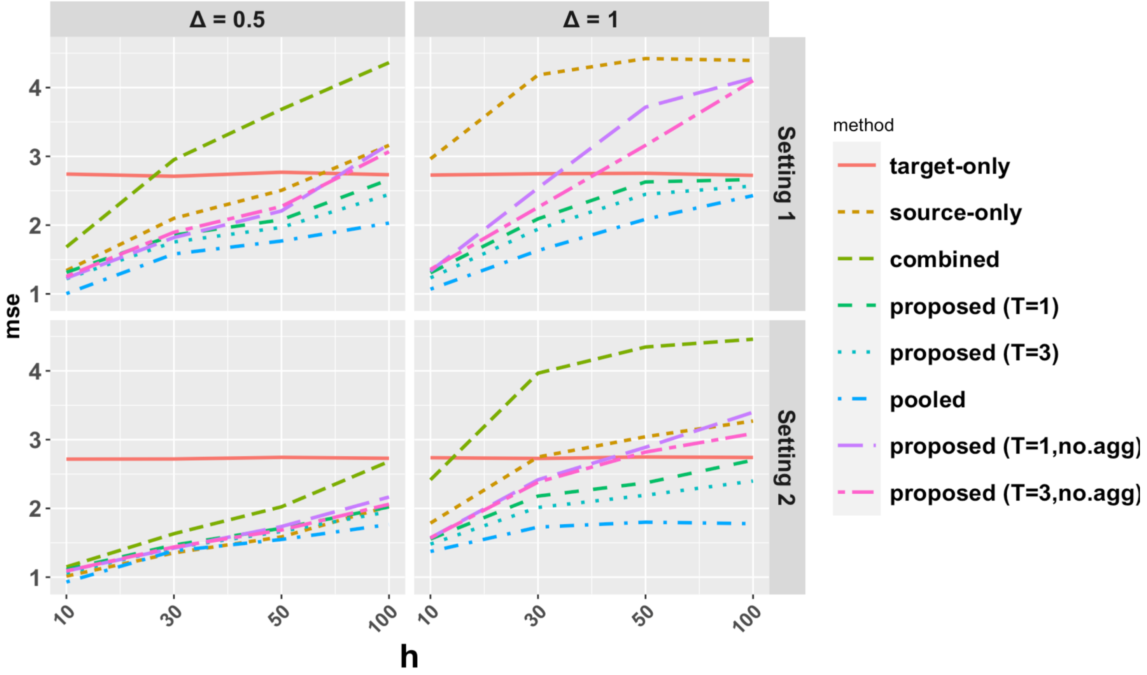
The data used in this application was applied is public available upon request from the database of Genotypes and Phenotypes (dbGaP) with accession phs000888.v1.p1. Here we introduce more details in selecting the genetic variations used in this paper:

1. We performed standard quality controls to remove SNPs with minor allele frequencies less than 0.05, and missing rates higher than 0.05. Missing SNPs are imputed by 0.

2. We then perform a GWAS study controlling for age, gender and top principal components.

3. GWAS p-values are used to clump the SNPs with a p-value threshold 5×10^{-5} and R^2 threshold 0.5, and physical distance threshold for clumping to be 1000 kb. Af-

Figure E.2: Comparison of MSE over 200 replications.



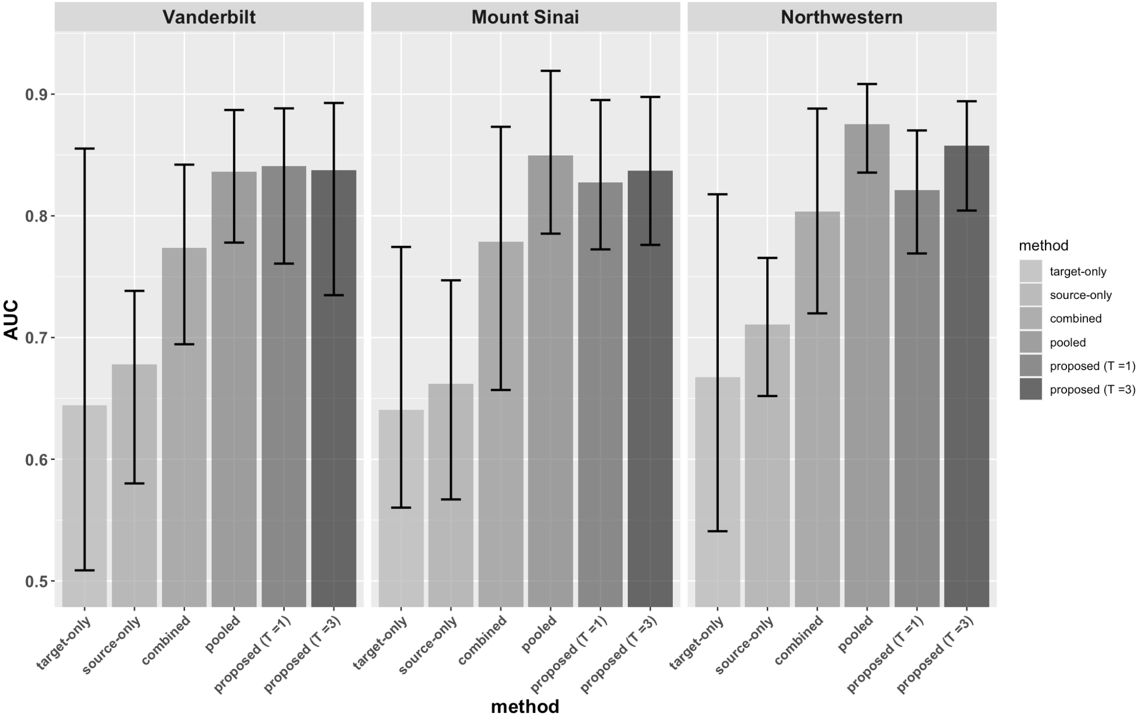
ter clumping, we obtain 2048 selected SNPsm which can be found in attached files "obesity_SNPs.txt".

F.2 Sensitivity analysis

We conducted sensitivity analysis by only including patients who are White (source) and Black or African Americans (target). The evaluation of performance of our method compared with other benchmark methods can be found in the following Figure F.1, where the definitions of training and testing datasets are the same as described in the main paper.

As we can see from Figure F.1, the proposed methods reach higher AUC than the target-only, source-only, and combined approaches, and are comparable with the pooled transfer learning method. However, by comparing with Figure 5 in the main paper, we see that the source-only model has decreased performance, which might be due to that the Unknown race group are more similar to the target population. So when excluding Unknown race from source, the performance of the all the approaches that involve source data drops.

Figure F.1: Sensitivity Study excluding participants who are not White, Black or African Americans.



References

Jordan, M. I., J. D. Lee, and Y. Yang (2018). Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc*, 1–14.

Kuchibhotla, A. K. and A. Chakrabortty (2018). Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*.

Wang, J., M. Kolar, N. Srebro, and T. Zhang (2017). Efficient distributed learning with sparsity. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3636–3645. JMLR. org.

Yuan, X.-T., P. Li, and T. Zhang (2018). Gradient hard thresholding pursuit. *Journal of Machine Learning Research 18*, 1–43.