Supplimentary Material for Variable Selection via Penalized Neural Network: a Drop-Out-One Loss Approach

Mao Ye [*1] **Yan Sun** [*1]

1. Proof of Theorem 1

1.1. Irrelevant Weights

We use $\ell_{\boldsymbol{\eta}}$ to represent $\ell_{\boldsymbol{\eta}}(y, \boldsymbol{x})$ to simplify. $\forall j \in S^c$, by Taylor Theorem,

$$\begin{aligned}
& \mathbb{P}_n(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) \\
= \; & \nabla^T_{\boldsymbol{w}_{j,*}} \mathbb{P}_n \ell_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}} (-\hat{\boldsymbol{w}}_{j,*}) + \frac{1}{2} \hat{\boldsymbol{w}}^T_{j,*} \nabla^2_{\boldsymbol{w}_{j,*}} \mathbb{P}_n \ell_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*} \hat{\boldsymbol{w}}_{j,*} \\
= \; & \nabla^T_{\boldsymbol{w}_{j,*}} \left\{ -\lambda_1 \sum_{j=1}^{p} \Omega_\alpha(\boldsymbol{w}_{j,*}) \right\} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}} (-\hat{\boldsymbol{w}}_{j,*}) + \frac{1}{2} \hat{\boldsymbol{w}}^T_{j,*} \nabla^2_{\boldsymbol{w}_{j,*}} \mathbb{P}_n \ell_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*} \hat{\boldsymbol{w}}_{j,*} \\
= \; & \left\{ sign\left(\hat{\boldsymbol{w}}_{j,*}\right) \lambda_1 (1-\alpha) + \lambda_1 \alpha \frac{\hat{\boldsymbol{w}}_{j,*}}{\sqrt{\sum_{i=1}^m \hat{\boldsymbol{w}}^2_{j,i}}} \right\}^T \hat{\boldsymbol{w}}_{j,*} \\
& + \frac{1}{2} \hat{\boldsymbol{w}}^T_{j,*} \nabla^2_{\boldsymbol{w}_{j,*}} \nabla^T_{\boldsymbol{w}_{j,*}} \mathbb{P}_n \ell_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*} \hat{\boldsymbol{w}}_{j,*},
\end{aligned}$$

where $\hat{\boldsymbol{\eta}}^* = \xi \hat{\boldsymbol{\eta}} + (1-\xi)\hat{\boldsymbol{\eta}}^{-j}$ for some $\xi \in [0,1]$. By Assumption 2 and Assumption 3 and Theorem 1 and Theorem 2 in (Feng & Simon, 2017), we have

$$\left\{ sign\left(\hat{\boldsymbol{w}}_{j,*}\right) \lambda_1 (1-\alpha) + \lambda_1 \alpha \frac{\hat{\boldsymbol{w}}_{j,*}}{\sqrt{\sum_{i=1}^m \hat{\boldsymbol{w}}^2_{j,i}}} \right\}^T \hat{\boldsymbol{w}}_{j,*} = O_p\left(\lambda_1 \frac{\sqrt{\log p} (\log n)^{\frac{3}{2}}}{\sqrt{n}} \right),$$
$$\tfrac{1}{2} \hat{\boldsymbol{w}}^T_{j,*} \nabla^2_{\boldsymbol{w}_{j,*}} \mathbb{P}_n \ell_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*} \hat{\boldsymbol{w}}_{j,*} = O_p(\tfrac{\log p (\log n)^3}{n}).$$

Thus we conclude that $\left| \Delta_n \mathcal{L}(\hat{\boldsymbol{\eta}}^{-j}, \hat{\boldsymbol{\eta}}) \right| = O_p(\frac{\log^3 n \log p}{n}), \; \forall j \in S^c$.

1.2. Relevant Weights

We first introduce the following claim and lemma.

Claim1: Let $S^{-j} = S - \{j\}$, define

$$Eq_0(S^{-j}) = \{\boldsymbol{\eta} : \; \boldsymbol{\eta} = \arg\min_{\boldsymbol{\eta}} \mathbb{P}\ell_{\boldsymbol{\eta}}, \; s.t. \; supp(\boldsymbol{\eta}) = S^{-j}\}.$$

[*] Equal contribution [1] Department of Statistics, Purdue University, West Lafayette, IN, USA. Correspondence to: Mao Ye <ye207@purdue.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

Then $\min_{j\in S}\mathbb{P}(\ell_{\boldsymbol{\eta}_0(S^{-j})} - \ell_{\boldsymbol{\eta}_0}) > 0$, where $\boldsymbol{\eta}_0(S^{-j}) \in Eq_0(S^{-j})$.

This claim can be easily obtained by Assumption 4. We denote

$$C \equiv \min_{j\in S}\mathbb{P}(\ell_{\boldsymbol{\eta}_0(S^{-j})} - \ell_{\boldsymbol{\eta}_0}).$$

Lemma1: Let $\boldsymbol{X}$ be a random variable such that $P(|\boldsymbol{X}| > t) \leq 2\exp^{(-\frac{t^2}{2\sigma^2})}$, then for any positive integer $K \geq 1$ we have $E(|\boldsymbol{X}|^K) \leq (2\sigma^2)^{\frac{K}{2}} K\Gamma(\frac{K}{2})$.

Use the property of sub-gaussian random variable, this lemma can be easily obtained.

$\forall j \in S$, we have,

$$\begin{aligned}
& \mathbb{P}_n(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) \\
= \; & \mathbb{P}_n\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \mathbb{P}_n\ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}} + \mathbb{P}_n\ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}} - \mathbb{P}_n\ell_{\hat{\boldsymbol{\eta}}} \\
= \; & (\mathbb{P}_n - \mathbb{P})(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) + \mathbb{P}(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}}) + \mathbb{P}(\ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}} - \ell_{\hat{\boldsymbol{\eta}}}) \\
\geq \; & (\mathbb{P}_n - \mathbb{P})(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) + \mathbb{P}(\ell_{\boldsymbol{\eta}_0(S-\{j\})} - \ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}}) + \mathbb{P}(\ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}} - \ell_{\hat{\boldsymbol{\eta}}}) \\
\geq \; & (\mathbb{P}_n - \mathbb{P})(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) + \mathbb{P}(\ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}} - \ell_{\hat{\boldsymbol{\eta}}}) + C.
\end{aligned}$$

Since $\left|\mathbb{P}(\ell_{\boldsymbol{\eta}_0^{(\hat{\eta})}} - \ell_{\hat{\boldsymbol{\eta}}})\right| = O_p(\frac{\log p(\log n)^{\frac{3}{2}}}{n})$ by Theorem 1 and Theorem 2 in (Feng & Simon, 2017), we then study the converge rate of $(\mathbb{P}_n - \mathbb{P})(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}})$. By the definition of the loss function, for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \boldsymbol{\Theta}$, we have

$$\begin{aligned}
& \left|(\mathbb{P}_n - \mathbb{P})(\ell_{\boldsymbol{\eta}_1} - \ell_{\boldsymbol{\eta}_2})\right| \\
= \; & \left|(\mathbb{P}_n - \mathbb{P})\left\{-2y(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2}) + (f_{\boldsymbol{\eta}_1}^2 - f_{\boldsymbol{\eta}_2}^2)\right\}\right| \\
= \; & \left|(\mathbb{P}_n - \mathbb{P})\left\{-2(f^* + \varepsilon)(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2}) + (f_{\boldsymbol{\eta}_1}^2 - f_{\boldsymbol{\eta}_2}^2)\right\}\right| \\
= \; & \left|(\mathbb{P}_n - \mathbb{P})\left\{-2\varepsilon(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2}) + \Big(f_{\boldsymbol{\eta}_1}^2 - f_{\boldsymbol{\eta}_2}^2 - 2f^*(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\Big)\right\}\right| \\
\leq \; & \left|(\mathbb{P}_n - \mathbb{P})\left\{-2\varepsilon(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\}\right| + \left|(\mathbb{P}_n - \mathbb{P})\left\{f_{\boldsymbol{\eta}_1}^2 - f_{\boldsymbol{\eta}_2}^2 - 2f^*(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\}\right|.
\end{aligned}$$

Using Chebyshev inequality and by the boundedness of $f_{\boldsymbol{\eta}}$ on $\mathcal{X}$ we have,

$$\begin{aligned}
& \sup_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \boldsymbol{\Theta}} P\left((\mathbb{P}_n - \mathbb{P})\left\{-2\varepsilon(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\} \geq a\right) \\
\leq \; & \sup_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \boldsymbol{\Theta}} \frac{Var\left\{-2\varepsilon(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\}}{na^2} \\
\leq \; & \sup_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \boldsymbol{\Theta}} \frac{E\left\{4\varepsilon^2(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})^2\right\}}{na^2} \\
\leq \; & \frac{4(mK_{\lambda_0}c_1)^2 E(\varepsilon^2)}{na^2} \\
\leq \; & \frac{16(mK_{\lambda_0}c_1)^2\sigma^2}{na^2},
\end{aligned}$$

where the last inequality is obtained by Lemma1. Thus we have

$$(\mathbb{P}_n - \mathbb{P})\left\{-2\varepsilon(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\} = O_p(\frac{1}{\sqrt{n}}).$$

Similarly, use Chebyshev inequality again, we also have

$$(\mathbb{P}_n - \mathbb{P})\left\{f^2_{\boldsymbol{\eta}_1} - f^2_{\boldsymbol{\eta}_2} - 2f^*(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\} = O_p(\frac{1}{\sqrt{n}}).$$

We conclude that $\Delta_n\mathcal{L}(\hat{\boldsymbol{\eta}}^{-j}, \hat{\boldsymbol{\eta}}) \geq C + O_p(\frac{1}{\sqrt{n}} \vee \frac{\log^3 n \log p}{n}),\ \forall j \in S$.

2. Proof of Theorem 2

2.1. Irrelevant Weights

$\forall j \in S^c$, By mean value theorem

$$\begin{aligned}
& (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) \\
= & (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)\nabla^T_{\boldsymbol{w}_{j,*}}\ell_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*} (-\hat{\boldsymbol{w}}_{j,*}) \\
= & (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(-2(y - f_{\boldsymbol{\eta}})\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*} (-\hat{\boldsymbol{w}}_{j,*})) \\
= & (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(-2(y - f_{\boldsymbol{\eta}})\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*})(-\hat{\boldsymbol{w}}_{j,*}) \\
= & (\mathbb{P}_{\tilde{n}} - \mathbb{P})(-2(y - f_{\boldsymbol{\eta}})\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*})(-\hat{\boldsymbol{w}}_{j,*}) \\
& +(\mathbb{P} - \mathbb{P}_n)(-2(y - f_{\boldsymbol{\eta}})\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*})(-\hat{\boldsymbol{w}}_{j,*})
\end{aligned}$$

where $\hat{\boldsymbol{\eta}}^* = \xi\hat{\boldsymbol{\eta}} + (1-\xi)\hat{\boldsymbol{\eta}}^{-j}$ for some $\xi \in [0, 1]$. Recall that $f_{\boldsymbol{\eta}}(\boldsymbol{x}) = \boldsymbol{\beta}^T\psi(\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{t}) + b$, $\|\boldsymbol{t}\|_2^2 + ||\boldsymbol{\beta}||_2^2 + b^2 \leq K_{\lambda_0}$, by Assumption 2 and chain rule, it is easy to show that $\sup_{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \boldsymbol{\Theta}} \left|\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}}\right| \leq c_2$ for some constant c_2. Using the condition $\tilde{n} = O(n)$, Chebyshev inequality and the same procedure in proof of Theorem 1, we have

$$\begin{aligned}
(\mathbb{P}_{\tilde{n}} - \mathbb{P})(-2(y - f_{\boldsymbol{\eta}})\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*}) = O_p(\frac{1}{\sqrt{n}}), \\
(\mathbb{P} - \mathbb{P}_n)(-2(y - f_{\boldsymbol{\eta}})\nabla^T_{\boldsymbol{w}_{j,*}} f_{\boldsymbol{\eta}} \mid_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}^*}) = O_p(\frac{1}{\sqrt{n}}).
\end{aligned}$$

By Theorem 1 and Theorem 2 in (Feng & Simon, 2017), we have $\|\hat{\boldsymbol{w}}_{j,*}\|_1 = O_p(\frac{\sqrt{\log p}(\log n)^{\frac{3}{2}}}{\sqrt{n}})$. Hence

$$(\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) = O_p(\frac{1}{\sqrt{n}}) \times O_p(\frac{\sqrt{\log p}(\log n)^{\frac{3}{2}}}{\sqrt{n}}) = O_p(\frac{\sqrt{\log p}(\log n)^{\frac{3}{2}}}{n}).$$

By Theorem 1, we have $\mathbb{P}_n(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) = O_p(\frac{\log^3 n \log p}{n})$. Hence

$$\begin{aligned}
\left|\Delta_{\tilde{n}}\mathcal{L}(\hat{\boldsymbol{\eta}}^{-j}, \hat{\boldsymbol{\eta}})\right| &= |\mathbb{P}_{\tilde{n}}(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}})| \\
&= |(\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) + \mathbb{P}_n(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}})| \\
&= O_p(\frac{\log^3 n \log p}{n}),\ \forall j \in S^c.
\end{aligned}$$

2.1.1. Relevant Weights

$\forall j \in S$

$$\begin{aligned}
& (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) \\
= & (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(-2y(f_{\hat{\boldsymbol{\eta}}^{-j}} - f_{\hat{\boldsymbol{\eta}}}) + (f^2_{\hat{\boldsymbol{\eta}}^{-j}} - f^2_{\hat{\boldsymbol{\eta}}})) \\
= & (\mathbb{P}_{\tilde{n}} - \mathbb{P})(-2y(f_{\hat{\boldsymbol{\eta}}^{-j}} - f_{\hat{\boldsymbol{\eta}}}) + (f^2_{\hat{\boldsymbol{\eta}}^{-j}} - f^2_{\hat{\boldsymbol{\eta}}})) \\
& +(\mathbb{P} - \mathbb{P}_n)(-2y(f_{\hat{\boldsymbol{\eta}}^{-j}} - f_{\hat{\boldsymbol{\eta}}}) + (f^2_{\hat{\boldsymbol{\eta}}^{-j}} - f^2_{\hat{\boldsymbol{\eta}}}))
\end{aligned}$$

Similar to the proof of Theorem 1, we have

$$(\mathbb{P}_{\tilde{n}} - \mathbb{P})(-2y(f_{\hat{\boldsymbol{\eta}}^{-j}} - f_{\hat{\boldsymbol{\eta}}}) + (f^2_{\hat{\boldsymbol{\eta}}^{-j}} - f^2_{\hat{\boldsymbol{\eta}}})) = O_p(\frac{1}{\sqrt{n}}),$$

$$(\mathbb{P} - \mathbb{P}_n)(-2y(f_{\hat{\boldsymbol{\eta}}^{-j}} - f_{\hat{\boldsymbol{\eta}}}) + (f^2_{\hat{\boldsymbol{\eta}}^{-j}} - f^2_{\hat{\boldsymbol{\eta}}})) = O_p(\frac{1}{\sqrt{n}}).$$

By Theorem 1,

$$\mathbb{P}_n(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) = \Delta_n \mathcal{L}(\hat{\boldsymbol{\eta}}^{-j}, \hat{\boldsymbol{\eta}}) \geq C + O_p(\frac{1}{\sqrt{n}} \vee \frac{\log^3 n \log p}{n}),\ \forall j \in S.$$

Hence

$$\Delta_{\tilde{n}} \mathcal{L}(\hat{\boldsymbol{\eta}}^{-j}, \hat{\boldsymbol{\eta}}) = (\mathbb{P}_{\tilde{n}} - \mathbb{P}_n)(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) + \mathbb{P}_n(\ell_{\hat{\boldsymbol{\eta}}^{-j}} - \ell_{\hat{\boldsymbol{\eta}}}) \geq C + O_p(\frac{1}{\sqrt{n}} \vee \frac{\log^3 n \log p}{n}),\ \forall j \in S.$$

As mentioned in the paper, Theorem 2 directly implies Corollary 3, thus we omit the proof of Corollary 3.

3. Proof of Corollary 4

3.1. Irrelevant groups

Recall that $S_g = \{i,\ g_i \cap S \neq \emptyset\}$. Therefore $\forall j \in S_g^c,\ i \in g_j$, we have $i \notin S$. By Theorem 1 and Theorem2 in (Feng & Simon, 2017), we have $\sum_{i \in g_j} \|\hat{\boldsymbol{w}}_{i,*}\|_1 = O_p(\frac{\sqrt{\log p}(\log n)^{\frac{3}{2}}}{\sqrt{n}})$. In the proof of Theorem 1 and Theorem 2, by replacing partial derivative with respect to $\hat{\boldsymbol{w}}_{j,*}$ with partial derivative with respect to $\hat{\boldsymbol{w}}_{g_j,*}$(here we use $\hat{\boldsymbol{w}}_{g_j,*}$to represent $\{\hat{\boldsymbol{w}}_{i,*},\ i \in g_j\}$), we can use the same procedure to prove that $\left|\Delta_{\tilde{n}} \mathcal{L}(\hat{\boldsymbol{\eta}}^{-g_j}, \hat{\boldsymbol{\eta}})\right| = O_p(\frac{\log^3 n \log p}{n}),\ \forall j \in S_g^c$.

3.2. Relevant groups

Recall that $supp_g(\boldsymbol{\eta}) = \{j :\ \sum_{i \in g_j} \|\boldsymbol{w}_{i,*}\|_2 \neq 0\}$. $\forall j \in S_g$, by Assumption 4*, we have $\forall \boldsymbol{\eta}_0 \in Eq_0$, $\sum_{i \in g_j} \|\boldsymbol{w}_{i,*}\|_2 \neq 0$. Similar to the proof of Theorem 1, we can have the following claim:

Claim 2: Let $S_g^{-g_j} = S_g - \{j\}$, define

$$Eq_0(S_g^{-g_j}) = \{\boldsymbol{\eta} :\ \boldsymbol{\eta} = \arg\min_{\boldsymbol{\eta}} \mathbb{P}\ell_{\boldsymbol{\eta}},\ s.t.\ supp_g(\boldsymbol{\eta}) = S_g^{-g_j}\}.$$

Then $\min_{j \in S_g} \mathbb{P}(\ell_{\boldsymbol{\eta}_0(S_g^{-g_j})} - \ell_{\boldsymbol{\eta}_0}) > 0$, where $\boldsymbol{\eta}_0(S_g^{-g_j}) \in Eq_0(S_g^{-g_j})$.

We denote $C_g \equiv \min_{j \in S} \mathbb{P}(\ell_{\boldsymbol{\eta}_0(S^{-j})} - \ell_{\boldsymbol{\eta}_0})$, C_g is a positive constant.

Note that in the proof of Theorem 1, our proof for convergence properties like $(\mathbb{P}_n - \mathbb{P})\left\{-2\varepsilon(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\} = O_p(\frac{1}{\sqrt{n}})$, $(\mathbb{P}_n - \mathbb{P})\left\{f^2_{\boldsymbol{\eta}_1} - f^2_{\boldsymbol{\eta}_2} - 2f^*(f_{\boldsymbol{\eta}_1} - f_{\boldsymbol{\eta}_2})\right\} = O_p(\frac{1}{\sqrt{n}})$, etc. is for all $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \boldsymbol{\Theta}$. By replacing $\hat{\boldsymbol{\eta}}^{-j}$ with $\hat{\boldsymbol{\eta}}^{-g_j}$, we can use the same procedure to prove that $\Delta_{\tilde{n}} \mathcal{L}(\hat{\boldsymbol{\eta}}^{-g_j}, \hat{\boldsymbol{\eta}}) \geq C_g + O_p(\frac{1}{\sqrt{n}} \vee \frac{\log^3 n \log p}{n}),\ \forall j \in S_g$.

Therefore, when $thre^{(t)}$ is properly tuned, we have $\hat{S}_g = S_g$ with probability 1 as $n \to \infty$.

4. Implement and tuning

GAM, RF, BART, SIS-SCAD, LROGL, SGL and Knockoffs are implemented using R packages $gamsel$, $randomForest$, $bart$, SIS, $grpregOverlap$, SGL and $knockoff$. We use default setting recommended by the authors for tuning and running.

For GAM, SIS-SCAD, LROGL, SGL, we tune the regularization parameters by grid search in a sufficiently large set generated by the packages, which is the recommended way. The chosen parameters fall into the inner part of the set (not boundary). The sets contain 50, 100, 100 and 20 values respectively. e.g, in simulation1, for GAM, the set is {0.00345, 0.00380...3.303, 3.542} (uniformly distributed in log sense). Similar to (Bleich et al., 2014; Liang et al., 2017), for BART and RF, we select variable (group) with variable (group) importance greater than 1%. See (Bleich et al., 2014; Liang et al., 2017) for the definition of variable importance. The definition of group importance is simply adding up the variable importance in one group and normalizing to make the sum of group importances equal to 1. We use a 500-tree RF (default setting in the package) and try BART with 25, 35 and 50 trees. The best structure of BART is chosen based on validation set. In knockoff, we use FDR=0.1, 0.2 and 0.3. In simulations, we give it correct distribution of variables. For the real data, we use default method to construct knockoff of variables. The W stat is based on random forest. We implement the Spinn and l_1-NN following (Feng & Simon, 2017). λ_1 of Spinn, l_1-NN and our method is chosen from the same set in every experiment. All the other parameters (including the ones for optimization algorithm) for the 3 methods are set to be the same in every experiment. In all the experiments, we set $\alpha = 0.8$ and initial $lr = 0.0005$. The tuning strategy for the proposed method is as follows: we set $thre^{(t)}$ to be

$$
\begin{aligned}
thre^{(t)} &= thre_{reg} \wedge \left(the\ \delta - th\ percentile\ of\ \Delta_{\tilde{n}}\mathcal{L}(\hat{\boldsymbol{\eta}}^{-g_j}, \hat{\boldsymbol{\eta}})\right), t \in \{1, ..., \bar{t}\}, \\
thre^{(t)} &= \left(\frac{\vartheta_1}{\sum_{j=1}^{d}\left(\Delta_{\tilde{n}}\mathcal{L}(\hat{\boldsymbol{\eta}}^{-g_j}, \hat{\boldsymbol{\eta}})\right)_+} \vee thre_{reg}\right) \wedge \frac{\vartheta_2}{\sum_{j=1}^{d}\left(\Delta_{\tilde{n}}\mathcal{L}(\hat{\boldsymbol{\eta}}^{-g_j}, \hat{\boldsymbol{\eta}})\right)_+}, t > \bar{t}.
\end{aligned}
$$

Here, we use $thre_{reg}$, ϑ_1 and ϑ_2 to regularize the threshold to make the algorithm more robust. The choose of $thre_{reg}$, ϑ_1 and ϑ_2 is not sensitive. A recommended setting is $\vartheta_1 = 0.01$ and $\vartheta_2 = 0.025$. We use this setting in all the experiments in this article. The choice of δ and $thre_{reg}$ depends on the dataset but the choice of the two tuning parameters is not sensitive. This proposed tuning strategy, which is alternative to the one mentioned in the main text, allows us to avoid tuning δ and $thre_{reg}$. And as mentioned in the main text, we can also set $thre^{(t)} = \frac{\vartheta}{\sum_{j=1}^{d}\left(\Delta_{\tilde{n}}\mathcal{L}(\hat{\boldsymbol{\eta}}^{-g_j}, \hat{\boldsymbol{\eta}})\right)_+}, t > \bar{t}$ and tune ϑ. Some more advanced technique such as (Liang & Zhang, 2008) can also be easily applied to determine the $thre^{(t)}$.

In simulation case 1 and case 2, the number of hidden units is set to be 6, $\lambda_0 = 0.0001$, $\bar{t} = 2$ and $\delta = 95$. In simulation 1, λ_1 is chosen from $\{0.025, 0.05, 0.1, 0.15, 0.2, 0.25\}$ and $thre_{reg} = 0.01$. In simulation 2, λ_1 is chosen from $\{0.035, 0.05, 0.1, 0.15, 0.20, 0.25\}$ and $thre_{reg} = 0.025$. In CCLE, Airfoil, CCPP and Boston $\lambda_0 = 0.00001$, $\bar{t} = 3$, $\delta = 30$. In CCLE, Airfoil and CCPP, $thre_{reg} = 0.001$ and the number of hidden units is 3. In Boston, $thre_{reg} = 0.1$ and the he number of hidden units is set to be 2 (since we introduce nonlinear features). In CCLE, λ_1 is chosen from $\{0.05, 0.06, 0.07, 0.08, 0.09, 0.1\}$. In Airfoil, λ_1 is chosen from $\{0.06, 0.07...0.34, 0.35\}$. In CCPP, λ_1 is chosen from $\{0.01, 0.02, ..., 0.15\}$. In Boston, λ_1 is chosen from $\{0.8, 0.09, ..., 0.2\}$. The chosen parameters fall inside the boundary of the candidate set in all experiments.

References

Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. Variable selection for bart: An application to gene regulation. *The Annals of Applied Statistics*, pp. 1750–1781, 2014.

Feng, J. and Simon, N. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*, 2017.

Liang, F. and Zhang, J. Estimating fdr under general dependence using stochastic approximation. *Biometrika*, 95(4): 961–977, 2008.

Liang, F., Li, Q., and Zhou, L. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, (just-accepted), 2017.