Performance Guarantees for Individualized Treatment Rules (Supplementary Material)

Min Qian and Susan A. Murphy

S.1 The overfitting problem

In this section, we discuss the problem with over-fitting due to the potentially large number of pretreatment variables (and/or complex approximation space for Q_0) mentioned in Section 4.

Consider the setting in which we know that Q_0 is linear in the $\{X, A\}$ variables and suppose that most coefficients are nonzero (some may be quite small). Then the least squares estimator using the best correct linear model (i.e. the model that contains and only contains variables with truly nonzero coefficients) may result in ITRs with poor Value as compared to the estimator from a more sparse model. Intuitively this occurs when the dimension of $\{X, A\}$ is too large for the size of the data set. This is similar to the case of stepwise model selection; a solution is to select the model that balances the approximation error with the estimation error instead of keeping all of the correct terms (Massart [3]). Indeed the l_1-PLS method aims to estimate a parameter possessing small approximation error (i.e. the excess prediction error) and controlled sparsity (which is directly related to the estimation error). As a result, the ITR produced by l_1-PLS will more reliably have higher Value than the rule produced by the OLS (ordinary least squares) estimator constructed when the correct model is known but is too non-sparse relative to the size of the data set.

In the following we use a simple simulation to support this argument. First we generate $X = (X_1, \ldots, X_{12})$, where $X_1, \ldots, X_{12}$ are mutually independent and each X_j is uniformly distributed on $[-1, 1]$. The treatment A is then generated independently of X from $\{-1, 1\}$ with probability 1/2 each. The response R is generated from a normal distribution with mean $Q_0(X, A) = (1, X_{-12}, A, X_{-12}A)\boldsymbol{\vartheta}$ and variance 1, where $X_{-12} = (X_1, \ldots, X_{11})$ and $\boldsymbol{\vartheta} \in \mathbb{R}^{24}$ is a vector parameter. We consider $\boldsymbol{\vartheta} = (1.458, -0.455, -0.311, -1.213, -1.600, 0.665, -0.431, -0.265, -0.113, -0.814, -0.128, 0.210, 0.442, 0.324, -0.090, 0.195, -0.047, 0.143, -0.008, 0.198 -0.389, 0.409, -0.085, -0.251)^T$. The effect size is 0.5. We approximate Q_0 using model $\mathcal{Q} = \{(1, X, A, XA)\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{26}\}$. Thus $Q_0 \in \mathcal{Q}$.

We simulate samples of sizes $n = 30$, 50 and 80. 1000 samples are generated for each n. For each sample, we apply the l_1-PLS based method, where the tuning parameter is selected using cross-validation with Value maximization as described in Section 5. In addition, we compute the OLS estimator over the best correct sub-model (i.e. $\hat{\boldsymbol{\theta}}_n^{OLS^*} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^{26},\theta_{13}=\theta_{26}=0} E_n[R-(1,X,A,XA)\boldsymbol{\theta}]^2$) and the associated ITR $\hat{d}_n^{OLS^*}(X) = \arg\max_{a\in\{-1,1\}}(1,X,\ a,Xa)\hat{\boldsymbol{\theta}}_n^{OLS^*}$. An independent test set of size 10000 is generated to evaluate the Value of each estimated ITR. Medians and median absolute deviations (MAD) of the Value and the number of variables in each ITR over 1000 samples are presented in Table S.1. Value of the optimal ITR is also evaluated and presented in the table. It is easy to see that in this case (i.e. the approximation model for Q_0 is sufficiently good and the best correct linear sub-model is too non-sparse for the sample size), the l_1-PLS estimator from the full model tends to have better performance in Value maximization and often yield much simpler decision rules than the OLS estimator from the best correct sub-model.

	Median (MAD) for	
Method	Value of the ITR	# variables used in the ITR
$n = 30$		
l_1-PLS	1.780 (0.138)	2 (2)
OLS*	1.636 (0.108)	12 (0)
$n = 50$		
l_1-PLS	1.889 (0.029)	3 (3)
OLS*	1.814 (0.058)	12 (0)
$n = 80$		
l_1-PLS	1.914 (0.016)	6 (5)
OLS*	1.887 (0.036)	12 (0)
$(V(d_0) = 1.9832)$		

Table S.1: Medians and MAD (in the parentheses) of the Value of each estimated ITR (left) and the number of variables in each estimated ITR (including the main treatment effect term, right) based on 1000 replications. (Value of the optimal ITR, $V(d_0)$, is given at the bottom. OLS* denote the OLS estimator from the best correct linear model.)

S.2 Some modifications of the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$

As demonstrated in van de Geer [4], sometimes it is natural not to penalize a subset of coefficients (e.g. coefficients corresponding to the constant term and/or to vari-

ables that are considered as definitely relevant). In this section, we discuss several modifications of the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ in this case.

Suppose one decides not to penalize coefficients indexed by $\mathcal{S} \subset \{1, \ldots, J_n\}$. A general modification is to exclude those terms from the penalty, i.e.

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} E_n(R - \Phi_n(X, A)\boldsymbol{\theta})^2 + \lambda_n \sum_{j \in \{1,\ldots,J_n\} \setminus \mathcal{S}} \hat{\sigma}_j |\theta_j|,$$

where $\hat{\sigma}_j = (E_n \phi_j^2)^{1/2}$. It is easy to see that with this modification, an analog of inequality (A.8) can be obtained after only slight adjustments in the proof.

Now suppose there are only two treatments $\mathcal{A} = \{1, -1\}$. A simple vector of basis functions that one may consider is $\Phi_n(X, A) = (1, X, A, XA)$, where X is a row vector of pretreatment variables. One may choose to leave the intercept term not penalized. Furthermore, if one believes that the main treatment effect exists, then the coefficient of A should not be penalized either (see the Nefazodone-CBASP data example in section 5.2). In both cases, one might want to change the weights $\hat{\sigma}_j$'s used in the penalty. In the following, we discuss these two special cases in a general framework.

1. When there is a constant term $\phi_1 \equiv 1$ and one decides not to penalize θ_1, it is natural to modify $\hat{\sigma}_j$ to $\hat{\sigma}_j \triangleq [E_n \phi_j^2 - (E_n \phi_j)^2]^{1/2}$ (so $\hat{\sigma}_1 = 0$). In this case, each $E\phi_j$ is estimated by $E_n \phi_j$. van de Geer [4] pointed out that "this additional source of randomness is in a sense of smaller order" and "the modification does not bring in new theoretical complications". The modified assumptions and outline of the proof for obtaining an analog of inequality (A.8) is provided below.

2. When Φ_n contains the main treatment effect terms and one decides not to penalize those terms, one may modify $\hat{\sigma}_j$ to an estimate of $\big(\sum_{a \in \mathcal{A}} var(\phi_j(X, A)|A = a)P(A = a)\big)^{1/2}$ (i.e. pooled standard deviation).

   For example, suppose $Q_0(X, a)$ is modeled by $\Psi_a(X)\boldsymbol{\theta}_a$ for each $a \in \mathcal{A}$, where the first term of each Ψ_a is $\psi_{a,1} \equiv 1$. Then the vector of basis functions is $\Phi_n(X, A) = (\Psi_a(X)1_{A=a})_{a \in \mathcal{A}}$ and $\{\psi_{a,1}1_{A=a} : a \in \mathcal{A}\}$ is the set of main treatment effect terms. Denote the index set of the main treatment effect terms in Φ_n by $\mathcal{S}$. If we use weights $\hat{\sigma}_j \triangleq \big(\sum_{a \in \mathcal{A}} \hat{var}(\phi_j(X, A)|A = a)E_n 1_{A=a}\big)^{1/2}$, where $\hat{var}(\phi_j(X, A)|A = a)$ is the sample variance of ϕ_j over the sub-sample that assigned treatment a, then $\hat{\sigma}_j = 0$ for all $j \in \mathcal{S}$. One can verify that

choosing $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$ to minimize $E_n(R - \Phi_n\boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j|\theta_j|$ is equivalent to choosing $\theta_j, j \in \{1, \ldots, J_n\} \setminus \mathcal{S}$, to minimize $E_n(R' - \sum_{j\in\{1,\ldots,J_n\}\setminus\mathcal{S}} \theta_j\phi_j')^2 + \lambda_n \sum_{j\in\{1,\ldots,J_n\}\setminus\mathcal{S}} \hat{\sigma}_j|\theta_j|$ and setting $\theta_j, j \in \mathcal{S}$ to be some appropriate quantities, where $R' = R - \sum_{a\in\mathcal{A}}(E_n 1_{A=a}R)1_{A=a}/E_n 1_{A=a}$ (so $E_nR' = 0$) and each ϕ_j' is a variation of ϕ_j (so that $E_n\phi_j' = 0$ and $E_n[(\phi_j')^2] = \hat{\sigma}_j^2$). This implies that the modification of $\hat{\sigma}_j$ is appropriate.

To obtain an analog of (A.8), we need to show the concentration of sample means (of quantities such as R and ϕ_j) around the true means within each treatment group and make some assumptions about the randomization probability $p(a|X)$. As we have discussed, these modifications only bring in further trivial technical complications rather than theoretical innovations.

In the rest of the section, we present modified assumptions and outline of the proof for obtaining an analog of (A.8) when $\phi_1 \equiv 1$ and θ_1 is not penalized.

In this case, $\hat{\sigma}_j$ and σ_j are modified to $\hat{\sigma}_j \triangleq [E_n\phi_j^2 - (E_n\phi_j)^2]^{1/2}$ and $\sigma_j \triangleq [E\phi_j^2 - (E\phi_j)^2]^{1/2}$, respectively, for $j = 1, \ldots, J_n$.

For any $0 \le \gamma < 1/2$ and $\eta_{2,n} \ge 0$, Θ_n^o is modified to

$$\Theta_n^{o\prime} = \Big\{\boldsymbol{\theta} \in \mathbb{R}^{J_n} : \exists\ \boldsymbol{\theta}^o \in [\boldsymbol{\theta}_n^*] \text{ s.t. } \|\Phi_n(\boldsymbol{\theta} - \boldsymbol{\theta}^o)\|_\infty \le \eta_{2,n} \text{ and } \max\Big\{|\theta_1 - \theta_1^o|, \max_{j\in\{2,\ldots,J_n\}} \Big|E\Big[\Phi_n(\boldsymbol{\theta} - \boldsymbol{\theta}^o)\frac{\phi_j}{\sigma_j}\Big]\Big|\Big\} \le \gamma\lambda_n\Big\}.$$

For any $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$ and $\rho \ge 0$, let

$$M_{\rho\lambda_n}(\boldsymbol{\theta})' \in \arg \min_{\{M\subseteq\{2,\ldots,J_n\}:\sum_{j\notin M}\sigma_j|\theta_j|\le\rho(|M|+1)\lambda_n\}} N_M.$$

Assumption A.2(a) is modified to

Assumption S.2(a) *There exists some $U_n > 0$ such that $\max_{j=2,\ldots,J_n} \|\phi_j\|_\infty/\sigma_j \le U_n$.*

Assumption A.3 is modified to

Assumption S.3 *There exists a positive number β_n such that*

$$\begin{aligned} &E[\Phi(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2(|M_{\rho\lambda_n}(\boldsymbol{\theta})'| + 1) \\ \ge&\beta_n\Big[\Big(|\tilde{\theta}_1 - \theta_1| + |\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})'} \sigma_j|\tilde{\theta}_j - \theta_j|\Big)^2 - \rho^2(|M_{\rho\lambda_n}(\boldsymbol{\theta})'| + 1)^2\lambda_n^2\Big] \end{aligned} \tag{S.1}$$

for all $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ satisfying conditions similar to those in Assumption A.3.

For any fixed $\boldsymbol{\theta} \in \Theta_n$, $\delta_1, \delta_2 \in (0,1)$, $\tau_1, \tau_2 > 0$, define the events

$$\begin{aligned}
\Omega_1' &= \cap_{j=2}^{J_n} \left\{(1-\delta_1)\sigma_j \le \hat{\sigma}_j \le (1+\delta_2)\sigma_j\right\}, \\
\Omega_2(\boldsymbol{\theta})' &= \Big\{ \max_{j=2,\ldots,J_n} \Big|(E-E_n)\frac{\phi_j}{\sigma_j}\Big| \le \tau_1 \frac{\beta_n}{|M_{\rho\lambda_n}(\boldsymbol{\theta})'|+1} \\
&\qquad \text{and} \quad \max_{j,k=2,\ldots,J_n} \Big|(E-E_n)\Big(\frac{\phi_j\phi_k}{\sigma_j\sigma_k}\Big)\Big| \le \tau_1 \frac{\beta_n}{|M_{\rho\lambda_n}(\boldsymbol{\theta})'|+1}\Big\}, \\
\Omega_3(\boldsymbol{\theta})' &= \Big\{ \big|E_n\big[(R-\Phi_n\boldsymbol{\theta})\phi_1\big]\big| \le \frac{2\tau_2+\delta_2+1}{2}\lambda_n \\
&\qquad \text{and} \quad \max_{j=2,\ldots,J_n} \Big|E_n\Big[(R-\Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| \le \tau_2\lambda_n\Big\}.
\end{aligned}$$

Using the same arguments as those in the proof of Theorem A.1, an analog of (A.8) can be obtained on the event $\Omega_1' \cap \Omega_2(\boldsymbol{\theta})' \cap \Omega_3(\boldsymbol{\theta})'$ with appropriate choices of $\delta_1, \delta_2, \tau_1$ and τ_2.

Next one can show that $\Omega_2(\boldsymbol{\theta})'$ and $\Omega_3(\boldsymbol{\theta})'$ occur with high probabilities under similar conditions as those in Lemmas A.5 and A.6. To show Ω_1' occurs with high probability, we define

$$\begin{aligned}
\Omega_{1,1}' &= \cap_{j=2}^{J_n} \Big\{|E\phi_j| - \nu_1\sqrt{E\phi_j^2} \le |E_n\phi_j| \le |E\phi_j| + \nu_2\sqrt{E\phi_j^2}\Big\} \\
\Omega_{1,2}' &= \cap_{j=2}^{J_n} \{(1-\kappa_1)E\phi_j^2 \le E_n\phi_j^2 \le (1+\kappa_2)E\phi_j^2\}
\end{aligned}$$

for some positive ν_1, ν_2, κ_1 and κ_2 to be chosen later. Under similar conditions as those in Lemma A.4, it is easy to see that $\Omega_{1,1}'$ and $\Omega_{1,2}'$ hold with high probabilities. In below we show that $\Omega_1' \subset \Omega_{1,1}' \cap \Omega_{1,2}'$ with appropriate choices of ν_1, ν_2, κ_1 and κ_2.

For $j = 2, \ldots, J_n$, on the event $\Omega_{1,1}' \cap \Omega_{1,2}'$, we have

$$\begin{aligned}
\hat{\sigma}_j^2 &= E_n\phi_j^2 - (E_n\phi_j)^2 \ge (1-\kappa_1)E\phi_j^2 - \Big(|E\phi_j| + \nu_2\sqrt{E\phi_j^2}\Big)^2 \\
&= (1-\delta_1)^2\sigma_j^2 + (2\delta_1 - \delta_1^2 - \kappa_1)E\phi_j^2 + (1-\delta_1)^2(E\phi_j)^2 - \Big(|E\phi_j| + \nu_2\sqrt{E\phi_j^2}\Big)^2 \\
&\ge (1-\delta_1)^2\sigma_j^2 + (2\delta_1 - \delta_1^2 - \kappa_1 - \nu_2^2 - 2\nu_2)E\phi_j^2 - (2\delta_1^2 - \delta_1^2)(E\phi_j)^2.
\end{aligned}$$

Taking κ_1 and ν_2 so that $\kappa_1 + \nu_2^2 + 2\nu_2 \le (2\delta_1 - \delta_1^2)\min_{j=1,\ldots,J_n}[1 - (E\phi_j)^2/E\phi_j^2]$, we obtain $\hat{\sigma}_j \ge (1-\delta_1)\sigma_j$.

Next we show that $\hat{\sigma}_j \le (1-\delta_1)\sigma_j$ on the event $\Omega_{1,1}' \cap \Omega_{1,2}'$. Note that when $E\phi_j = 0$,

$$\hat{\sigma}_j^2 = E_n\phi_j^2 - (E_n\phi_j)^2 \le (1+\kappa_2)E\phi_j^2 \le (1+\delta_2)^2\sigma_j^2$$

for any $\kappa_2 \leq \delta_2^2 + 2\delta_2$.

Consider $E\phi_j \neq 0$. When $\nu_1 \leq |E\phi_j|/\sqrt{E\phi_j^2}$,

$$\begin{aligned}\hat{\sigma}_j^2 &= E_n\phi_j^2 - (E_n\phi_j)^2 \leq (1+\kappa_2)E\phi_j^2 - \left(|E\phi_j| - \nu_1\sqrt{E\phi_j^2}\right)^2 \\ &= (1+\delta_2)^2\sigma_j^2 + (\kappa_2 - 2\delta_2 - \delta_2^2)E\phi_j^2 + (1+\delta_2)^2(E\phi_j)^2 - \left(|E\phi_j| - \nu_1\sqrt{E\phi_j^2}\right)^2 \\ &= (1+\delta_2)^2\sigma_j^2 + (\delta_2^2 + 2\delta_2)(E\phi_j)^2 - (\delta_2^2 + 2\delta_2 - \kappa_2 - 2\nu_1 + \nu_1^2)E\phi_j^2\end{aligned}$$

Taking κ_2 and ν_1 so that $\kappa_2 + 2\nu_1 - \nu_1^2 \leq (\delta_2^2 + 2\delta_2)\min_{j=1,\ldots,J_n}[1 - (E\phi_j)^2/E\phi_j^2]$, we obtain $\hat{\sigma}_j \leq (1+\delta_2)\sigma_j$.

S.3 Extra simulation examples

In this section, we consider extra four simulation examples (i.e. examples 5-8 below) in addition to the examples used in Section 5.1. To make the simulations more realistic, these examples are based on data from the Nefazodone-CBASP trial [1] (see Section 5.2 for description of the trial).

In the simulation study, we consider 50 pretreatment variables collected from the trial (i.e. $X \in \mathbb{R}^{50}$). Each variable is standardized using the sample mean and standard deviation. The Nefazodone-CBASP data provides an empirical distribution for the standardized pretreatment variables. This is the distribution we use to generate X. Treatment A is generated independently of X from $\{-1, 1\}$ with probability $1/2$ each. To generate R, the response HRSD score is reverse coded so that higher scores are desirable. We regress the reverse coded HRSD score on $(1, X)$ and denote the estimated regression coefficients by $\boldsymbol{\vartheta}^{(1)}$. Then the response R is generated from a normal distribution with mean $Q_0(X, A) = (1, X)\boldsymbol{\vartheta}^{(1)} + T_0(X, A)$ and variance 9. We consider 4 examples for T_0. There is no treatment effect in example 5. The covariates and parameters involved in examples 6-8 produce a medium effect size.

5. $T_0(X, A) = 0$.

6. $T_0(X, A) = (1, \widetilde{X})\boldsymbol{\vartheta}^{(2)}A$, where $\widetilde{X} = (X_6, X_{21}, X_{22}, X_{27}, X_{38})$ and $\boldsymbol{\theta}^{(2)} = (-1.222, -0.568, 0.416, -0.008, -0.776, 0.614)^T$. Note that the analysis model contains the the correct model for T_0.

7. $T_0(X, A) = |(1, \widetilde{X})\boldsymbol{\vartheta}^{(2)}|A$, where $\widetilde{X} = (X_8, X_9, X_{29}, X_{40}, X_{46})$ and $\boldsymbol{\vartheta}^{(2)} = (-0.875,$

-0.289, 0.121, 1.052, 0.344, $-0.424)^T$. In this case, treatment 1 is always better than -1.

8. $T_0(X, A) = sign((1, \widetilde{X}_{sub})\boldsymbol{\vartheta}^{(2),1})|(1, \widetilde{X})\boldsymbol{\vartheta}^{(2),2}|A$, where $\widetilde{X} = (X_{44}, X_{17}, X_{31}, X_{35}, X_{16})$, $\widetilde{X}_{sub}$ contains the first 3 covariates in $\widetilde{X}$, $\boldsymbol{\vartheta}^{(2),1} = (-0.841, 0.747, 0.141, 0.298)^T$ and $\boldsymbol{\vartheta}^{(2),2} = (-3.136, 0.793, -5.266, -1.787, -0.268, 2.324)^T$. Note that the analysis model does not contain the correct model for T_0.

We approximate Q_0 by model $\mathcal{Q} = \{(1, X, A, XA)\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{102}\}$.

For each example, we simulate 1000 data sets of size $n = 500$. The Value of each estimated ITR is evaluated via Monte Carlo using a test set of size $10,000$. The Value of the optimal ITR is also evaluated using the test set. Simulation results are presented in Table S.2. These simulations give us the same conclusion as those in Section 5.1.

S.4 Proofs of Lemmas A.1-A.5

In this section, we prove Lemmas A.1 - A.5 given in Appendix A.2.

Proof of Lemma A.1

First note that the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ satisfies the following first order condition:

$$-2E_n(R - \Phi_n\hat{\boldsymbol{\theta}}_n)\phi_j + \lambda_n\hat{\sigma}_j\text{sgn}(\hat{\theta}_{n,j}) = 0 \text{ for } j = 1, \ldots, J_n,$$

where $\text{sgn}(\theta_j) = 1$ if $\theta_j > 0$, $\text{sgn}(\theta_j) = -1$ if $\theta_j < 0$ and $\text{sgn}(\theta_j) \in [-1, 1]$ if $\theta_j = 0$ for any $\theta_j \in \mathbb{R}$. This implies

$$-2E_n[(R - \Phi_n\hat{\boldsymbol{\theta}}_n)\Phi_n\boldsymbol{\theta}] + \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j\text{sgn}(\hat{\theta}_{n,j})\theta_j = 0$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$. In particular, $-2E_n[(R - \Phi_n\hat{\boldsymbol{\theta}}_n)\Phi_n\hat{\boldsymbol{\theta}}_n] + \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\hat{\theta}_{n,j}| = 0$.

Therefore, for any $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$, we have

$$\begin{aligned} 0 &= 2E_n[(R - \Phi_n\hat{\boldsymbol{\theta}}_n)\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})] + \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j\text{sgn}(\hat{\theta}_{n,j})\theta_j - \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\hat{\theta}_j| \\ &\leq 2E_n[(R - \Phi_n\hat{\boldsymbol{\theta}}_n)\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})] + \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\theta_j| - \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\hat{\theta}_{n,j}|. \end{aligned} \tag{S.2}$$

Fix n. If $\boldsymbol{\theta} = \mathbf{0}$, on the event $\Omega_1 \cap \Omega_3(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
0 \leq & 2E_n[(R - \Phi_n\boldsymbol{\theta})\Phi_n\hat{\boldsymbol{\theta}}_n] - 2E_n(\Phi_n\hat{\boldsymbol{\theta}}_n)^2 - \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j |\hat{\theta}_{n,j}| \\
\leq & 2 \max_{j=1,\ldots,J_n} \Big| E_n\Big[(R - \Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| \Big(\sum_{j=1}^{J_n} \sigma_j |\hat{\theta}_{n,j}| \Big) - 2E_n(\Phi_n\hat{\boldsymbol{\theta}}_n)^2 - \frac{2(1+\gamma)}{3}\lambda_n \sum_{j=1}^{J_n} \sigma_j |\hat{\theta}_{n,j}| \\
\leq & \frac{2\gamma - 1}{3}\lambda_n \sum_{j=1}^{J_n} \sigma_j |\hat{\theta}_{n,j}| - 2E_n(\Phi_n\hat{\boldsymbol{\theta}}_n)^2 \leq 0.
\end{aligned}
$$

This implies $\hat{\boldsymbol{\theta}}_n = \mathbf{0}$. Thus (A.8) and (A.9) hold.

Otherwise, for any fixed $\boldsymbol{\theta} \in \Theta_n \setminus \{\mathbf{0}\}$, the index set $M_{\rho\lambda_n}(\boldsymbol{\theta})$ is non-empty. Following (S.2), on the event $\Omega_1 \cap \Omega_3(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
0 \leq & 2 \max_{j=1,\ldots,J_n} \Big| E_n\Big[(R - \Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| \Big(\sum_{j=1}^{J_n} \sigma_j |\hat{\theta}_{n,j} - \theta_j| \Big) - 2E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 \\
& + \lambda_n \sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \hat{\sigma}_j |\hat{\theta}_{n,j} - \theta_j| + \lambda_n \sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \hat{\sigma}_j (|\theta_j| - |\hat{\theta}_{n,j}|) \\
\leq & \frac{4\gamma + 1}{3}\lambda_n \Big(\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j} - \theta_j| + \rho |M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n + \sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j}| \Big) \\
& + \frac{2(2-\gamma)}{3}\lambda_n \Big(\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j} - \theta_j| + \rho |M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n \Big) \\
& - \frac{2(1+\gamma)}{3}\lambda_n \sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j}| - 2E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 \\
= & \frac{2\gamma + 5}{3}\lambda_n \Big(\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j} - \theta_j| + \rho |M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n \Big) \\
& - \frac{1 - 2\gamma}{3}\lambda_n \sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j}| - 2E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2.
\end{aligned}
$$

This implies

$$
\begin{aligned}
& \sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j}| \leq \frac{2\gamma + 5}{1 - 2\gamma} \Big(\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j} - \theta_j| + \rho |M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n \Big) \\
& \text{and } E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 \leq \frac{2\gamma + 5}{6}\lambda_n \Big(\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\hat{\theta}_{n,j} - \theta_j| + \rho |M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n \Big). \quad \text{(S.3)}
\end{aligned}
$$

Define the sets

$$\Theta_1(\boldsymbol{\theta}) = \Big\{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n} : \sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j| \leq \frac{2\gamma+5}{1-2\gamma}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)\Big\},$$

$$\Theta_2(\boldsymbol{\theta}) = \Big\{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n} : \sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| > \frac{[10(2\gamma+5)+3(21-2\gamma)\beta_n\rho]|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n}{3(19+2\gamma)\beta_n}\Big\},$$

$$\Theta_3(\boldsymbol{\theta}) = \Big\{\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n} : \sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n > \frac{10|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n}{3\beta_n}\Big\}$$

Note that $\hat{\boldsymbol{\theta}}_n \in \Theta_1(\boldsymbol{\theta})$ on the event $\Omega_1 \cap \Omega_3(\boldsymbol{\theta})$. In addition, on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$,

$$\begin{aligned}
&\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Big\{2E_n[(R-\Phi_n\tilde{\boldsymbol{\theta}})\Phi_n(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})] + \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\theta_j| - \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\tilde{\theta}_j|\Big\} \\
\leq &\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Big\{\frac{2\gamma+5}{3}\lambda_n\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big) - 2E[\Phi_n(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})]^2 \\
&+ 2\max_{j=1,\ldots,J_n}\Big|(E-E_n)\Big(\frac{\phi_j\phi_k}{\sigma_j\sigma_k}\Big)\Big|\Big(\sum_{j=1}^{J_n}\sigma_j|\tilde{\theta}_j-\theta_j|\Big)^2 - \frac{1-2\gamma}{3}\lambda_n\sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j|\Big\} \\
\leq &\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Big\{\frac{2\gamma+5}{3}\lambda_n\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big) \\
&\qquad + 2\beta_n\rho^2|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n^2 - \frac{2\beta_n}{|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|\Big)^2 \\
&+ \frac{(1-2\gamma)^2\beta_n}{60|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)\Big(\sum_{j=1}^{J_n}\sigma_j|\tilde{\theta}_j-\theta_j|\Big) \\
&+ \frac{1-2\gamma}{3}\Big[\frac{(1-2\gamma)\beta_n}{20|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\Big(\sum_{j=1}^{J_n}\sigma_j|\tilde{\theta}_j-\theta_j|\Big) - \lambda_n\Big]\Big(\sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j|\Big)\Big\} \\
\leq &\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Big\{\frac{2\gamma+5}{3}\lambda_n\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big) + 2\beta_n\rho^2|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n^2 \\
&- \frac{2\beta_n}{|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|\Big)^2 + \frac{(1-2\gamma)\beta_n}{10|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1-2\gamma}{3}\Big[\frac{3\beta_n}{10|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|+\rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)-\lambda_n\Big]\Big(\sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j|\Big)\Big\}\\
\leq&\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})\cap\Theta_3(\boldsymbol{\theta})^C}\Big\{\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|+\rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)\\
&\qquad\times\Big(\frac{2\gamma+5}{3}\lambda_n+\frac{21-2\gamma}{10}\beta_n\rho\lambda_n-\frac{(19+2\gamma)\beta_n}{10|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|\Big)\Big\}\\
&+\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})\cap\Theta_3(\boldsymbol{\theta})}\Big\{\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|+\rho|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)\\
&\qquad\times\Big(\frac{13}{5}\beta_n\rho\lambda_n-\frac{7\beta_n}{5|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j|\Big)\Big\}\\
<&0,
\end{aligned}
$$

where the second inequality follows from Assumption A.3 and the definition of $\Omega_2(\boldsymbol{\theta})$, the third inequality follows from the definition of $\Theta_1(\boldsymbol{\theta})$, the fourth equality follows from the definition of $\Theta_3(\boldsymbol{\theta})$ and simple algebra, and the last inequality follows from the definition of $\Theta_2(\boldsymbol{\theta})$, $\Theta_3(\boldsymbol{\theta})$ and the assumption that $\rho\beta_n\leq 1$.

Since $\hat{\boldsymbol{\theta}}_n$ satisfies inequality (S.2), we have $\hat{\boldsymbol{\theta}}_n\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})^C$ on the event $\Omega_1\cap\Omega_2(\boldsymbol{\theta})\cap\Omega_3(\boldsymbol{\theta})$. Algebra suffices to show (A.8).

Following (S.3) and the fact that $\hat{\boldsymbol{\theta}}_n\in\Theta_2(\boldsymbol{\theta})^C$, we have

$$
E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2\leq\frac{5(12\rho\beta_n+2\gamma+5)(2\gamma+5)}{9(19+2\gamma)\beta_n}|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n^2.
$$

on the event $\Omega_1\cap\Omega_2(\boldsymbol{\theta})\cap\Omega_3(\boldsymbol{\theta})$. Suppose (A.9) does not hold, i.e. $E[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2>130(12\rho\beta_n+2\gamma+5)^2|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n^2/[9(19+2\gamma)^2\beta_n]$. Then

$$
\frac{(E-E_n)[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2}{E[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2}\leq\frac{(1-2\gamma)^2\beta_n}{120|M_{\rho\lambda_n}(\boldsymbol{\theta})|}\cdot\frac{\Big(\sum_{j=1}^{J_n}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big)^2}{E[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2}\leq\frac{3}{13},
$$

where the first inequality follows from the definition of $\Omega_2(\boldsymbol{\theta})$ and the second inequality follows from (A.8). This implies

$$
E[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2\leq\frac{13}{10}E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2\leq\frac{13(12\rho\beta_n+2\gamma+5)(2\gamma+5)}{18(19+2\gamma)\beta_n}|M_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n^2,
$$

which contradicts the condition. Thus (A.9) holds on the event $\Omega_1\cap\Omega_2(\boldsymbol{\theta})\cap\Omega_3(\boldsymbol{\theta})$.□

Proof of Lemma A.2

Consider fixed n and fixed $\boldsymbol{\theta} \in \Theta_n$. Since $E[\Phi_n^{(2)}(X,A)^T|X] = \mathbf{0}$ a.s., we have $E(\phi_j\phi_{j'}) = 0$ for any $j \in \{1,\ldots,J_n^{(1)}\}$ and $j' \in \{J_n^{(1)}+1,\ldots,J_n\}$. On the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
&E_n\big[(\Phi_n\boldsymbol{\theta} - \Phi_n\hat{\boldsymbol{\theta}}_n)(\Phi_n^{(2)}\hat{\boldsymbol{\theta}}_n^{(2)} - \Phi_n^{(2)}\boldsymbol{\theta}^{(2)})\big] \\
\leq &\max_{j\in\{1,\ldots,J_n^{(1)}\},j'\in\{J_n^{(1)}+1,\ldots,J_n\}} \Big|E_n\Big(\frac{\phi_j\phi_{j'}}{\sigma_j\sigma_{j'}}\Big)\Big|\Big(\sum_{j=1}^{J_n^{(1)}}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big)\Big(\sum_{j=J_n^{(1)}+1}^{J_n}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big) \\
&+\max_{j,j'\in\{J_n^{(1)}+1,\ldots,J_n\}} \Big|(E-E_n)\Big(\frac{\phi_j\phi_{j'}}{\sigma_j\sigma_{j'}}\Big)\Big|\Big(\sum_{j=J_n^{(1)}+1}^{J_n}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big)^2 - E\big[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta}^{(2)})\big]^2 \\
\leq &\frac{(1-2\gamma)(12\beta_n\rho+2\gamma+5)}{6(2\gamma+19)}\lambda_n\Big(\sum_{j=J_n^{(1)}+1}^{J_n}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big) - E\big[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta}^{(2)})\big]^2,
\end{aligned}
$$

where the second inequality follows from the definition of $\Omega_2(\boldsymbol{\theta})$ and Lemma A.1.

Next, note that (S.2) holds for $(\hat{\boldsymbol{\theta}}_n^{(1)}, \boldsymbol{\theta}^{(2)})$. Thus on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
0 \leq &2E_n[(R-\Phi_n\hat{\boldsymbol{\theta}}_n)\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta}^{(2)})] + \lambda_n\sum_{j=J_n^{(1)}+1}^{J_n}\hat{\sigma}_j|\theta_j| - \lambda_n\sum_{j=J_n^{(1)}+1}^{J_n}\hat{\sigma}_j|\hat{\theta}_{n,j}| \\
\leq &2\max_{j=J_n^{(1)}+1,\ldots,J_n}\Big|E_n\Big[(R-\Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big|\Big(\sum_{j=J_n^{(1)}+1}^{J_n}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big) + \lambda_n\sum_{j\in M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}\hat{\sigma}_j|\hat{\theta}_{n,j}-\theta_j| \\
&+\lambda_n\sum_{j\in\{J_n^{(1)}+1,\ldots,J_n\}\setminus M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}\hat{\sigma}_j\big(|\theta_j|-|\hat{\theta}_{n,j}|\big) + 2E_n\big[(\Phi_n\boldsymbol{\theta}-\Phi_n\hat{\boldsymbol{\theta}}_n)(\Phi_n^{(2)}\hat{\boldsymbol{\theta}}_n^{(2)}-\Phi_n^{(2)}\boldsymbol{\theta}^{(2)})\big] \\
\leq &\frac{2\gamma+5}{3}\lambda_n\Big(\sum_{j\in M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho|M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})|\lambda_n\Big) - \frac{1-2\gamma}{3}\lambda_n\sum_{j\in\{J_n^{(1)}+1,\ldots,J_n\}\setminus M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}|\hat{\theta}_{n,j}| \\
&+\frac{(1-2\gamma)(12\beta_n\rho+2\gamma+5)}{3(2\gamma+19)}\lambda_n\Big(\sum_{j=J_n^{(1)}+1}^{J_n}\sigma_j|\hat{\theta}_{n,j}-\theta_j|\Big) - 2E\big[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta}^{(2)})\big]^2 \\
= &\frac{12(1-2\gamma)\beta_n\rho+20(2\gamma+5)}{3(2\gamma+19)}\lambda_n\Big(\sum_{j\in M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho|M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})|\lambda_n\Big) \\
&-\frac{2(1-2\gamma)(7-6\beta_n\rho)}{3(2\gamma+19)}\lambda_n\sum_{j\in\{J_n^{(1)}+1,\ldots,J_n\}\setminus M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}|\hat{\theta}_{n,j}| - 2E\big[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)}-\boldsymbol{\theta}^{(2)})\big]^2
\end{aligned}
$$

This implies

$$\frac{(1-2\gamma)(7-6\beta_n\rho)}{3(2\gamma+19)}\lambda_n \sum_{j\in\{J_n^{(1)}+1,\ldots,J_n\}\setminus M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})} |\hat{\theta}_{n,j}| + E\big[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}^{(2)})\big]^2$$
$$\leq \frac{6(1-2\gamma)\beta_n\rho + 10(2\gamma+5)}{3(2\gamma+19)}\lambda_n\Big(\sum_{j\in M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho|M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n\Big)$$

Using similar arguments as those in Lemma A.1, we obtain

$$\sum_{j\in M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho|M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n \leq \frac{10(12\beta_n\rho+2\gamma+5)}{3(2\gamma+19)\beta_n}|M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})|\lambda_n.$$

Algebra suffices to show (A.10) and (A.11). □

To proof Lemmas A.3, A.4 and A.5, we first introduce the following Bernstein's inequalities that will be repeatedly used.

Lemma S.1 *(**Bernstein's inequalities**; Massart [2]) Let $\zeta_1,\ldots,\zeta_n$ be independent and square integrable random variables such that $E[\zeta_i]=0$ for all $i=1,\ldots,n$.*

(a) Assume there exist some positive numbers b and ν such that $\zeta_i \leq b$ almost surely for all $i=1,\ldots,n$ and $\sum_{i=1}^n E\zeta_i^2 \leq \nu$. Then for any $s>0$,

$$\mathbf{P}\Big(\sum_{i=1}^n \zeta_i > s\Big) \leq \exp\Big(-\frac{s^2}{2(\nu+bs/3)}\Big).$$

(b) Assume there exist some positive numbers b and ν such that $\sum_{i=1}^n E[(\zeta_i^l)_+] \leq \frac{l!}{2}\nu b^{l-2}$ for all integers $l\geq 2$. Then for any $s>0$,

$$\mathbf{P}\Big(\sum_{i=1}^n \zeta_i > s\Big) \leq \exp\Big(-\frac{s^2}{2(\nu+bs)}\Big).$$

Proof of Lemma A.3.

For each $j=1,\ldots,J_n$, we apply Lemma S.1(a) with $\zeta_i = \phi_j(X_i,A_i)^2/\sigma_j^2 - 1$ and $s=(7-2\gamma)(1-2\gamma)n/9$. By Assumption A.2(a), we have $\zeta_i \leq U_n^2-1$ and $\sum_{i=1}^n E\zeta_i^2 \leq n(U_n^2-1)$. Thus

$$\mathbf{P}\Big(\hat{\sigma}_j \geq \frac{2(2-\gamma)}{3}\sigma_j\Big) \leq \exp\Big(-\frac{(7-2\gamma)^2(1-2\gamma)^2 n}{2(U_n^2-1)[81+3(7-2\gamma)(1-2\gamma)]}\Big)$$

$$\leq \exp\Big(-\frac{25(1-2\gamma)^2 n}{6(27U_n^2-10\gamma-22)}\Big).$$

Similarly, applying Lemma S.1(a) with $\zeta_i = 1-\phi_j(X_i,A_i)^2/\sigma_j^2$ and $s=(5+2\gamma)(1-2\gamma)n/9$, we have

$$\begin{aligned}\mathbf{P}\Big(\hat{\sigma}_j \leq \frac{2(1+\gamma)}{3}\sigma_j\Big) &\leq \exp\Big(-\frac{(5+2\gamma)^2(1-2\gamma)^2 n}{6[27(U_n^2-1)+(5+2\gamma)(1-2\gamma)]}\Big)\\ &\leq \exp\Big(-\frac{25(1-2\gamma)^2 n}{6(27U_n^2-10\gamma-22)}\Big).\end{aligned}$$

Using union bound argument and condition (A.6), we have

$$\mathbf{P}(\Omega_1^C) \leq 2J_n \exp\Big(-\frac{25(1-2\gamma)^2 n}{6(27U_n^2-10\gamma-22)}\Big) \leq \exp\Big(-\frac{13(1-2\gamma)^2 n}{6(27U_n^2-10\gamma-22)}\Big).\square$$

Proof of Lemma A.4.

Note that $\|\phi_j\phi_k/(\sigma_j\sigma_k) - E[\phi_j\phi_k/(\sigma_j\sigma_k)]\|_\infty \leq 2U_n^2$ and $E[\phi_j\phi_k/(\sigma_j\sigma_k)]^2 \leq U_n^2$ for all j,k. Applying Lemma S.1(a) with $\zeta_i = \pm[\phi_j(X_i,A_i)\phi_k(X_i,A_i)/(\sigma_j\sigma_k) - E(\phi_j\phi_k)/(\sigma_j\sigma_k)]$ and $s=(1-2\gamma)^2\beta_n n/[120|M_{\rho\lambda_n}(\boldsymbol{\theta})|]$ and using union bound argument, we obtain

$$\begin{aligned}\mathbf{P}(\{\Omega_2(\boldsymbol{\theta})\}^C) &\leq J_n(J_n+1)\exp\Big(-\frac{(1-2\gamma)^4\beta_n^2 n}{160U_n^2[180|M_{\rho\lambda_n}(\boldsymbol{\theta})|^2+(1-2\gamma)^2\beta_n|M_{\rho\lambda_n}(\boldsymbol{\theta})|]}\Big)\\ &\leq \frac{1}{3}\exp(-t),\end{aligned}$$

where the second inequality follows from the definition of Θ_n in (A.4). $\square$

Proof of Lemma A.5

For any $\boldsymbol{\theta}\in\Theta_n$, there is a $\boldsymbol{\theta}^o \in [\boldsymbol{\theta}_n^*]$ such that $\max_j |E[\Phi_n(\boldsymbol{\theta}^o-\boldsymbol{\theta})\phi_j/\sigma_j]| \leq \gamma\lambda_n$. Since $\boldsymbol{\theta}^o$ minimizes $E(R-\Phi_n\boldsymbol{\theta})^2$, we have $E\big[(R-\Phi_n\boldsymbol{\theta}^o)\phi_j\big]=0$ for $j=1,\ldots,J_n$. Thus

$$\max_j \Big|E\Big[(R-\Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| = \max_j \Big|E\Big[(\Phi_n\boldsymbol{\theta}^o-\Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| \leq \gamma\lambda_n.$$

This implies

$$\max_j \Big|E_n\Big[(R-\Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| \leq \max_j \Big|(E_n-E)\Big[\varepsilon\frac{\phi_j}{\sigma_j}\Big]\Big| + \max_j \Big|(E_n-E)\Big[(Q_0-\Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| + \gamma\lambda_n.$$

For $j=1,\ldots,J_n$, by Assumptions A.1 and A.2(a), we have $E(\varepsilon_i\phi_j(X_i,A_i)/\sigma_j)=0$ and $\sum_{i=1}^n E[(\varepsilon_i\phi_j(X_i,A_i)/\sigma_j)_+^l] \leq l!n\sigma^2(cU_n)^{l-2}/2$ for all integers $l\geq 2$. Applying

Lemma S.1(b) yields

$$\mathbf{P}\Big(\Big|(E_n - E)\Big[\varepsilon\frac{\phi_j}{\sigma_j}\Big]\Big| > \frac{1-2\gamma}{12}\lambda_n\Big) \leq 2\exp\Big(-\frac{(1-2\gamma)^2\lambda_n^2 n}{288\sigma^2 + 24c(1-2\gamma)U_n\lambda_n}\Big).$$

Similarly, the definition of Θ_n together with Assumption A.2 implies that, for any $\boldsymbol{\theta} \in \Theta_n$ and $j = 1, \ldots, J_n$, $\big\|(Q_0 - \Phi_n\boldsymbol{\theta})\phi_j/\sigma_j - E\big((Q_0 - \Phi_n\boldsymbol{\theta})\phi_j/\sigma_j\big)\big\|_\infty \leq 2(\eta_{n,1} + \eta_{n,2})U_n$ and $E\big[(Q_0 - \Phi_n\boldsymbol{\theta})\phi_j/\sigma_j\big]^2 \leq (\eta_{n,1} + \eta_{n,2})^2$. Applying Lemma S.1(a) yields

$$\begin{aligned}&\mathbf{P}\Big(\Big|(E_n - E)\Big[(Q_0 - \Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| > \frac{1-2\gamma}{12}\lambda_n\Big)\\ \leq& 2\exp\Big(-\frac{(1-2\gamma)^2\lambda_n^2 n}{288(\eta_{1,n} + \eta_{2,n})^2 + 16(1-2\gamma)(\eta_{1,n} + \eta_{2,n})U_n\lambda_n}\Big).\end{aligned}$$

The result follows from the union bound argument and condition (A.5). □

References

[1] Keller, M. B., McCullough, J. P., Klein, D. N., Arnow, B., Dunner, D. L., Gelenberg, A. J., Markowitz, J. C., Nemeroff. C. B., Russell, J. M., Thase, M. E., Trivedi, M. H. and Zajecka, J. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *The New England Journal of Medicine,* **342(20)**, 1462–1470.

[2] Massart. P. (2003). *Ecole d'Eté de Probabilités de Saint-Flour XXXIII, Concentration inequalities and model selection,* Springer.

[3] Massart, P. (2005). A non asymptotic theory for model selection. *Proceedings of the* 4^{th} *European Congress of Mathematicians (Ed. Ari Laptev),* European Mathematical Society, 309–323.

[4] van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics,* **36(2)**, 614–645.

Method	Median and MAD (in the parentheses) for	
	Value of the decision rules	# of variables needed for treatment assignment
Example 5 ($V(d_0) = 28.854$ = average of R over the test set)		
l_1-PLS	28.842 (0.023)	4 (4)
OLS	28.853 (0.035)	51 (0)
PP	28.855 (0.035)	49 (1)
Example 6 ($V(d_0) = 30.2354$)		
l_1-PLS	30.104 (0.034)	11 (7)
OLS	29.984 (0.050)	51 (0)
PP(CV)	30.008 (0.050)	50 (1)
Example 7 ($V(d_0) = 30.044$)		
l_1-PLS	30.042 (0.002)	7 (7)
OLS	29.797 (0.053)	51 (0)
PP	29.840 (0.051)	49 (1)
Example 8 ($V(d_0) = 33.275$)		
l_1-PLS	32.227 (0.426)	4 (2)
OLS	31.252 (0.219)	51 (0)
PP	31.359 (0.251)	42 (3)

Table S.2: Comparison of the l_1-PLS based method with the OLS method and the PP method (examples 5 - 8): Medians and MAD (in the parentheses) of the Value of the estimated decision rules (left) and the number of variables needed for treatment assignment (including the main treatment effect term, right) based on 1000 replications ($n = 500$). (The Value of the optimal treatment rule for each example is given as well. Note that in example 5, all decision rules should produce the same Value. The small differences in Value observed in example 5 are due only to Monte Carlo error.)