MODEL SELECTION AND l_1 PENALIZATION FOR INDIVIDUALIZED TREATMENT RULES

by

Min Qian

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2010

Doctoral Committee:

Professor Susan A. Murphy, Chair
Associate Professor Moulinath Banerjee
Associate Professor Bin Nan
Professor Runze Li, Pennsylvania State University

To my family

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my advisor, Dr. Susan Murphy. She taught me much more than just Statistics. Her passion for research has been a constant source of inspiration for me. Her commitment, encouragement, guidance and support helped me grow as a student and as a researcher. Working with her was an invaluable experience.

I would also like to give special thanks to my committee members: Dr. Runze Li, Dr. Moulinath Banerjee and Dr. Bin Nan for their advice and support through these valuable years at the University of Michigan.

I want to thank: faculties and staffs from the Department of Statistics and the Institute for Social Research - Dr. George Michailidis, Dr. Yu Xie, Rhonda Moats, Amy Rundquist, Lu Ann Custer, Mary Ann King and many others for providing a comfortable and supportive working environment; lab mates - Daniel Almirall, Bibhas Chakraborty, Lacey Gunter, Alena Oetting, Eric Laber, Daniel Lizotte, Zhiguo Li, Peng Zhang and Inbal Shani for helpful discussions and stimulating conversations; many friends for their encouragement through this exciting and challenging journey.

I acknowledge the support from NIH grants R01 MH080015 and P50 DA10075 (Dr. Susan Murphy). And I thank Martin Keller and the investigators of the Nefazodone-CBASP trial for the use of their data, John Rush, MD, for the technical support, and Bristol-Myers Squibb for helping fund the trial.

Last but not the least, I want to thank my family for their endless love and support in my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Many illnesses show heterogeneous response to treatment. For example, a study on schizophrenia (Ishigooka et al., 2001) found that patients who take the same antipsychotic (olanzapine) may have very different responses. Some may have to discontinue the treatment due to serious adverse events and/or acutely worsened symptoms, while others may experience few if any adverse events and have improved clinical outcomes. These types of results have motivated researchers to advocate the individualization of treatment to each patient (Lesko, 2007; Piquette-Miller and Grant, 2007; Insel, 2009). One step in this direction is to estimate each patient's risk level and then match treatment to risk category (Cai et al., 2008a,b). However, this approach is best used to decide whether to treat; otherwise it assumes the knowledge of the best treatment for each risk category. Alternatively, one can directly estimate a decision rule that recommends treatment according to individual characteristics. Such a decision rule is sometimes called an individualized treatment rule. In this dissertation, we consider the latter approach. Our goal is to develop a high quality individualized treatment rule using data from a randomized trial. We investigate model selection and penalization techniques aiming to improve the quality of the estimated individualized treatment rule.

1.1 Individualized treatment rules

We use upper case letters to denote random variables and lower case letters to denote values of the random variables. Consider data from a randomized trial. On each subject we have the pretreatment variables $X \in \mathcal{X}$, treatment A taking values in a finite, discrete treatment space $\mathcal{A}$, and a real-valued primary outcome Y (assuming large values are desirable). An *individualized treatment rule* d is a deterministic decision rule from space $\mathcal{X}$ into the treatment space $\mathcal{A}$.

Denote the distribution of (X, A, Y) by P. This is the distribution of the clinical trial data; in particular, denote the known randomization distribution of A given X by $p(\cdot|X)$. The likelihood of (X, A, Y) under P is then $f_0(x)p(a|x)f_1(y|x,a)$, where f_0 is the unknown density of X and f_1 is the unknown density of Y conditional on (X, A). Denote the expectations with respect to the distribution P by an E. For any individualized treatment rule $d : \mathcal{X} \to \mathcal{A}$, let P^d denote the distribution of (X, A, Y) in which d is used to assign treatments. Then the likelihood of (X, A, Y) under P^d is $f_0(x)1_{a=d(x)}f_1(y|x,a)$. Denote expectations with respect to the distribution P^d by an E^d. The *Value* of d is defined as $V(d) = E^d(Y)$. An *optimal individualized treatment rule* is a rule that has the maximal Value, i.e.

$$d^{opt} \in \arg\max_{d} V(d),$$

where the argmax is over all possible treatment rules. The Value of d^{opt}, $V(d^{opt})$, is the *optimal Value*.

Assume $P[p(a|X) > 0] = 1$ for all $a \in \mathcal{A}$ (i.e. all treatments in $\mathcal{A}$ are possible for all values of X a.s.). Then P^d is absolutely continuous with respect to P and a version of the Radon-Nikodym derivative is $dP^d/dP = 1_{a=d(x)}/p(a|x)$. Thus the Value of d satisfies

$$V(d) = E^d(Y) = \int Y dP^d = \int Y \frac{dP^d}{dP} dP = E\Big[\frac{1_{A=d(X)}}{p(A|X)} Y\Big]. \tag{1.1}$$

Our goal is to estimate d^{opt}, i.e. the individualized treatment rule that maximizes (1.1) using data from distribution P.

1.2 Comparison with classification

The decision making problem stated in the previous section is similar to a weighted classification problem. In binary classification, the goal is to estimate the classifier $\pi : \mathcal{X} \to \{-1, 1\}$ that minimizes the classification error $E[1_{Y \neq \pi(X)}]$, where $Y \in \{-1, 1\}$ is the correct label of X. In the decision making problem, the goal is to estimate the decision rule $d : \mathcal{X} \to \mathcal{A}$ that maximizes (1.1). One can think of d as a classifier which, given the observation $X = x$ as input, predicts the optimal treatment. Notice that $V(d)$ can be written as

$$V(d) = E\Big[\frac{1_{A=d(X)}}{p(A|X)} Y\Big] = E\Big[\frac{Y}{p(A|X)}\Big] - E\Big[\frac{Y}{p(A|X)} 1_{A \neq d(X)}\Big]. \tag{1.2}$$

The first term on the RHS of (1.2) is a fixed number and the second term can be viewed as a weighted classification error. Consequently, from an algorithmic view point, estimating the optimal individualized treatment rule is similar to learning the classifier with the minimal weighted classification error.

Thus ideas in classification can be used to estimate the optimal individualized treatment rule. When X is low dimensional and the best rule within a simple class of treatment rules is desired, empirical versions of the Value can be used to construct estimators (Murphy et al., 2001; Robins et al., 2008). However if the best rule within a larger class of treatment rules is of interest, these approaches are no longer feasible due to the non-smoothness and non-concavity of $1_{A=d(X)}$. In this dissertation, we consider a surrogate convex minimization method to regularize the non-concavity problem.

1.3 Estimating optimal rules based on convex minimization

Denote $Q^{opt}(X, A) := E(Y|X, A)$ (here notation "Q" stands for "quality" since $Q^{opt}(x, a)$ measures the quality of treatment a at observation $X = x$). It follows from (1.1) that for any individualized treatment rule d,

$$V(d) = E\Big[\frac{1_{A=d(X)}}{p(A|X)} Q^{opt}(X, A)\Big] = E\Big[\sum_{a \in \mathcal{A}} 1_{d(X)=a} Q^{opt}(X, a)\Big] = E[Q^{opt}(X, d(X))].$$

Thus $V(d^{opt}) = E[Q^{opt}(X, d^{opt}(X))] \leq E[\max_{a\in\mathcal{A}} Q^{opt}(X, a)]$. On the other hand, by the definition of d^{opt},

$$V(d^{opt}) \geq \max_{d:\ d(X) \in \arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} V(d) = E\Big[\max_{a\in\mathcal{A}} Q^{opt}(X, a)\Big].$$

Hence an optimal individualized treatment rule satisfies $d^{opt}(X) \in \arg\max_{a\in\mathcal{A}} Q^{opt}(X, a)$ a.s.

This suggests that we may estimate the conditional mean function Q^{opt} first and then estimate d^{opt} by maximizing the estimated Q^{opt} over $\mathcal{A}$. We propose to estimate Q^{opt} based on minimization of the quadratic loss $(Y-Q)^2$ over a function class for Q^{opt}. In particular, if the function class is linear in the parameter space, then we have a convex minimization problem. Compared with directly maximizing a Value estimator, this approach reduces the computational difficulty and allows the consideration of a large space of individualized treatment rules.

1.4 Contribution and Outline of the dissertation

In the present Chapter, we have formulated the decision making problem and compared it with classification. This comparison has motivated us to estimate the conditional mean function Q^{opt} first over a function class and then estimate the treatment rule by maximizing the estimated Q^{opt}. For clarity, we include in Table 1.1 symbols used throughout the dissertation, and in Tables 1.2 and 1.3 extra symbols used in Chapter III and Chapter IV, respectively.

In Chapter II, we relate the Value of an individualized treatment rule d to the prediction quality of Q for any square integrable function Q on $\mathcal{X} \times \mathcal{A}$ such that $d(X) \in \arg\max_{a\in\mathcal{A}} Q(X, a)$. This relationship implies that the estimated individualized treatment rule will be of high quality if Q^{opt} is well estimated. We also demonstrate that although the convex minimization approach reduces the computational difficulty, it may however deviate from the goal of estimating the best treatment rule if the approximation space for Q^{opt} is poor. This will motivate us to improve the performance of the convex minimization approach by using penalization and model selection techniques.

In Chapter III, we consider a sufficiently rich linear approximation space for Q^{opt}. l_1 penalty is employed to regularize possible overfitting problem and produce a simple individualized treatment rule. To justify this approach, we provide a finite sample upper bound on the difference between the Value of the estimated individualized treatment rule and the Value of the optimal individualized treatment rule. In practical implementation, we consider a data dependent criterion for selecting tuning parameter involved in the l_1 penalty that is targeted for Value maximization. This method is evaluated using simulation studies and illustrated with data from the Nefazodone-CBASP trial (Keller et al., 2000)

In Chapter IV, we use model selection techniques to deal with possible deviation of the convex minimization method from the goal of maximizing the Value. We consider a sequence of models for Q^{opt}. Within each model, an individualized treatment rule is estimated by minimizing the quadratic loss $(Y-Q)^2$. And the rule that maximizes a penalized Value estimator is selected. This approach is also justified by a finite sample upper bound on difference between the Value of the estimated individualized treatment rule and the Value of the optimal individualized treatment rule.

In Chapter V, we discuss possible extensions and future work. This includes the extension of current one-stage decision making problem to sequential decisions and issues related to efficient estimation.

In Chapter VI, we list mathematical tools that are useful in deriving theorems presented in Chapters III and IV.

X	patient pretreatment variables ($X \in \mathcal{X}$)
A	treatment ($A \in \mathcal{A}$, where $\mathcal{A}$ is a finite space)
Y	a one-dimensional summary of primary outcome (larger is better)
(x, a, y)	particular instances of random variables (X, A, Y)
$\mathbb{R}$	real line
$p(A\|X)$	randomization distribution of A given X in the clinical trial data
P	distribution of (X, A, Y) where A is assigned according to $p(A\|X)$
E	expectation with respect to the distribution P
n	sample size of the clinical trial data
(X_i, A_i, Y_i)	data collected from the i-th patient
E_n	empirical expectation with respect to the clinical trial data
d	an individualized treatment rule (mapping from $\mathcal{X}$ to $\mathcal{A}$)
P^d	distribution of (X, A, Y) where A is assigned according to rule d
E^d	expectation with respect to the distribution P^d
$V(d)$	the Value of d, $V(d) = E[1_{A=d(X)}Y/p(A\|X)]$
d^{opt}	an optimal individualized treatment rule, $d^{opt} \in \arg\max_d V(d)$
Q^{opt}	the conditional mean function, $Q^{opt}(X, A) = E(Y\|X, A)$
Q	a square integrable function on $\mathcal{X} \times \mathcal{A}$
$L(Q)$	prediction error of Q, $L(Q) = E[Y - Q(X, A)]^2$
$\mathcal{Q}$	approximation space for Q^{opt}
α	margin parameter defined in the Margin condition (2.2)
$\mathbb{P}$	probability with respect to all random variables
$\mathbb{E}$	expectation with respect to all random variables

Table 1.1: List of symbols used throughout the dissertation.

Φ (or Φ_n)	a row vector of basis functions on $\mathcal{X} \times \mathcal{A}$
J (or J_n)	dimension of Φ (or Φ_n)
ϕ_j	the j-th component of Φ (or Φ_n)
$\boldsymbol{\theta}$	a J (or J_n) by 1 parameter vector, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)^T$
$[\boldsymbol{\theta}^*]$ (or $[\boldsymbol{\theta}^*_n]$)	set of prediction error minimizers in the linear model
$\hat{\boldsymbol{\theta}}_n$	l_1 penalized least squares estimator of $\boldsymbol{\theta}$
λ_n	tuning parameter used in the l_1-PLS
σ_j	the L_2 norm of basis ϕ_j, $\sigma_j = \left(E\phi_j^2\right)^{1/2}$
$\hat{\sigma}_j$	the l_2 norm of basis ϕ_j, $\sigma_j = \left(E_n\phi_j^2\right)^{1/2}$
$\hat{d}_n$	the estimated l_1-PLS individualized treatment rule
$M_0(\boldsymbol{\theta})$	the index set of nonzero components in $\boldsymbol{\theta}$, $M_0(\boldsymbol{\theta}) = \{j = 1, \dots J : \theta_j \neq 0\}$
$M_{\rho\lambda_n}(\boldsymbol{\theta})$	the smallest index set of "large" components in $\boldsymbol{\theta}$ (defined in Section 3.4.1)
N_M	cardinality of the index set M
Θ_n^o	set of parameters with controlled prediction error (define in (3.4))
Θ_n	a subset of Θ_n^o with controlled sparsity (define in (3.5))
T^{opt}	the interaction term in Q^{opt}, $T^{opt}(X, A) = Q^{opt}(X, A) - E[Q^{opt}(X, A)\|X]$
$\Phi^{(1)}$ (or $\Phi_n^{(1)}$)	all components in Φ (or Φ_n) that do not contain A, $\Phi^{(1)} = (\phi_1, \dots, \phi_{J^{(1)}})$
$J^{(1)}$ (or $J_n^{(1)}$)	dimension of $\Phi^{(1)}$ (or $\Phi_n^{(1)}$)
$\Phi^{(2)}$ (or $\Phi_n^{(2)}$)	all components in Φ (or Φ_n) that contain A, $\Phi^{(2)} = (\phi_{J^{(1)}+1}, \dots, \phi_J)$
$\boldsymbol{\theta}^{(1)}$	parameter vector corresponding to $\Phi^{(1)}$, $\boldsymbol{\theta}^{(1)} = (\theta_1, \dots, \theta_{J^{(1)}})^T$ for a given $\boldsymbol{\theta} \in \mathbb{R}^J$
$\boldsymbol{\theta}^{(2)}$	parameter vector corresponding to $\Phi^{(2)}$, $\boldsymbol{\theta}^{(2)} = (\theta_{J^{(1)}+1, \dots, \theta_J})^T$ for a given $\boldsymbol{\theta} \in \mathbb{R}^J$
$M_0^{(1)}(\boldsymbol{\theta})$	the index set of nonzero components in $(\theta_1, \dots, \theta_{J^{(1)}})$
$M_0^{(2)}(\boldsymbol{\theta})$	the index set of nonzero components in $(\theta_{J^{(1)}+1}, \dots, \theta_J)$
$M_{\rho\lambda_n}^{(1)}(\boldsymbol{\theta})$	the smallest index set of "large" components in $(\theta_1, \dots, \theta_{J^{(1)}})$
$M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})$	the smallest index set of "large" components in $(\theta_{J^{(1)}+1}, \dots, \theta_J)$

Table 1.2: List of extra symbols used in Chapter III.

M_n	number of models for Q^{opt}
$\mathcal{Q}_m$	the m-th model for Q^{opt}
$\hat{Q}_{n,m}$	the least square estimator of Q^{opt} in the m-th model
Q^*_m	prediction error minimizer in the m-th model
$\mathcal{D}_m$	class of individualized treatment rules associated with $\mathcal{Q}_m$, $\mathcal{D}_m = \{d(X) \in \arg\max_a Q(X,a) : Q \in \mathcal{Q}_m\}$
$\hat{d}_{n,m}$	the treatment rule associated with $\hat{Q}_{n,m}$, $\hat{d}_{n,m}(X) \in \arg\max_a \hat{Q}_{n,m}(X,a)$
d^*_m	the treatment rule associated with Q^*_m
$\widetilde{d}_m$	the best treatment rule in $\mathcal{D}_m$, $\widetilde{d}_m \in \arg\max_{d \in \mathcal{D}_m} V(d)$
m^*	the model that has the maximal Value $V(\hat{d}_{n,m})$
$\hat{m}$	model selection by maximizing the penalized empirical Value (defined in (4.3))
$\gamma(n,m)$	a quantity that increases as the model complexity increases and decreases to zero as the sample size $n \to \infty$
$f(d)$	a function of (X,A,Y) and treatment rule d, $f(d) = \frac{1_{A=d(X)}}{p(A\|X)}Y$
$\mathcal{F}_m$	class of functions $f(d)$ for $d \in \mathcal{D}_m$
$N(\epsilon, \mathcal{G}, L_1(Pn))$	the ϵ-covering number of $\mathcal{G}$ relative to the $L_1(P_n)$ norm for any function class $\mathcal{G}$ on $\mathcal{X} \times \mathcal{A} \times \mathbb{R}$
$u_n(\mathcal{G})$	$u_n(\mathcal{G}) = \mathbb{E} \log[N(1/n, \mathcal{G}, L_1(P_n)) + 1]/n$ for any function class $\mathcal{G}$ on $\mathcal{X} \times \mathcal{A} \times \mathbb{R}$
$\xi_1, \ldots, \xi_n$	i.i.d. Rademacher random variables, $\xi_i = 1$ or -1 with probability 1/2 each

Table 1.3: List of extra symbols used in Chapter IV.

CHAPTER II

Quadratic loss minimization

In this chapter, we relate the Value of an individualized treatment rule d to the prediction quality of the associated square integrable function Q on $\mathcal{X} \times \mathcal{A}$. We also demonstrate possible deviation of the quadratic loss minimization method from the goal of estimating the best rule in the class of treatment rules under consideration.

2.1 Relating Value to quadratic loss

For any square integrable function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, let $L(Q)$ denote the prediction error of Q (i.e. the expected quadratic loss, $L(Q) := E[Y - Q(X, A)]^2$) and d be an individualized treatment rule associated with Q (i.e. $d(X) \in \arg\max_{a \in \mathcal{A}} Q(X, a)$ a.s.). In this section, we provide an upper bound on the excess Value of d, $V(d^{opt}) - V(d)$, in terms of the excess prediction error of Q, $L(Q) - L(Q^{opt})$.

Suppose there exists a positive constant S such that $p(a|x) \geq S^{-1}$ for all (x, a) pairs. Murphy (2005) showed that

$$V(d^{opt}) - V(d) \leq 2S^{1/2}\big[L(Q) - L(Q^{opt})\big]^{1/2}. \tag{2.1}$$

Intuitively, this bound tells us that if the excess prediction error of Q is small, then the Value of the associated individualized treatment rule will be close to the optimal

Value. Furthermore, the exponent 1/2 on the right hand side of (2.1) implicitly gives a rate of convergence. For example, suppose we approximate the conditional mean function Q^{opt} by a linear model and we estimate it by least squares. In addition, suppose Q^{opt} is in the linear model. Then the excess prediction error of the estimated Q-function will converge to zero at rate $1/n$. This bound implies that the Value of the estimated individualized treatment rule will converge to the optimal Value at rate at least $1/\sqrt{n}$. To guarantee a fast rate of convergence, an upper bound with exponent greater than 1/2 is needed. Such an improved bound can be obtained under a margin condition as follows.

Margin condition. *There exists some constants $C > 0$ and $\alpha \geq 0$ such that*

$$\mathbb{P}\Big(0 < \max_{a \in \mathcal{A}} Q^{opt}(X, a) - \max_{a \in \mathcal{A} \setminus \arg\max_{a \in \mathcal{A}} Q^{opt}(X,a)} Q^{opt}(X, a) \leq \epsilon\Big) \leq C\epsilon^{\alpha} \qquad (2.2)$$

for all positive ϵ satisfying $C\epsilon^{\alpha} \leq 1$.

The above condition is similar to the margin condition in classification (Polonik, 1995; Mammen and Tsybakov, 1999; Tsybakov, 2004); in classification this assumption is often used to obtain sharp upper bounds on the excess $0-1$ risk in terms of other surrogate risks (Bartlett et al., 2006). Here $\max_{a \in \mathcal{A}} Q^{opt}(x, a) - \max_{a \in \mathcal{A} \setminus \arg\max_{a \in \mathcal{A}} Q^{opt}(x,a)} Q^{opt}(x, a)$ can be viewed as the "margin" of Q^{opt} at observation $X = x$. It measures the difference in mean outcomes between the optimal treatment(s) and the best suboptimal treatment(s) at x.

Note that the margin condition (2.2) always holds with $C = 1, \alpha = 0$. In addition, the margin condition does not exclude multiple optimal treatments for any observation x. However, when $\alpha > 0$, it does exclude suboptimal treatments that yield a conditional mean outcome very close to the optimal conditional mean outcome for a set of x with nonzero probability.

Theorem II.1. *Suppose $p(a|x) \geq S^{-1}$ for a positive constant S for all (x, a) pairs.*

Assume the margin condition (2.2) holds with some $C > 0$ and $\alpha \geq 0$. Then for any individualized treatment rule $d : \mathcal{X} \to \mathcal{A}$ and square integrable function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that $d(X) \in \arg\max_{a \in \mathcal{A}} Q(X, a)$ a.s., we have

$$V(d^{opt}) - V(d) \leq C_1 \left[L(Q) - L(Q^{opt})\right]^{(1+\alpha)/(2+\alpha)}, \tag{2.3}$$

where $C_1 = (2^{2+3\alpha} S^{1+\alpha} C)^{1/(2+\alpha)}$. Furthermore, for any decomposition of $Q^{opt}(X, A)$ into $W^{opt}(X) + T^{opt}(X, A)$ and $Q(X, A)$ into $W(X) + T(X, A)$,

$$V(d^{opt}) - V(d) \leq C_1 \left[E\big(T(X, A) - T^{opt}(X, A)\big)^2\right]^{(1+\alpha)/(2+\alpha)}. \tag{2.4}$$

The proof of Theorem II.1 is in Section 2.3.

Remarks:

1. Inequality (2.3) is adaptive in the sense that if the margin condition (2.2) holds with some $\alpha > 0$, the exponent on the RHS (right hand side) of (2.3) is larger than 1/2; otherwise (2.3) is equivalent to (2.1) (since $C' = 2S^{1/2}$ when $C = 1$ and $\alpha = 0$).

2. $T^{opt}(X, A)$ in the inequality (2.4) need only contain terms in Q^{opt} that involve A. The consequence is that the quality of an estimated individualized treatment rule only depends on how well we estimate T^{opt}. In some cases, the estimation of T^{opt} will not be effected by the estimation of W^{opt}. See Chapter III for further discussion.

3. The exponent on the RHS of (2.3) and (2.4) approaches 1 as $\alpha \to \infty$. In this case, the margin condition requires that the LHS of (2.2) equals 0 for all $\epsilon \in (0, 1)$, which is unlikely to be true. However, the following holds.

Suppose $p(a|x) \geq S^{-1}$ *for all* (x,a) *pairs. Assume there is an* $\epsilon > 0$, *such that*

$$\mathbb{P}\Big(0 < \max_{a\in\mathcal{A}} Q^{opt}(X,a) - \max_{a\in\mathcal{A}\backslash \arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} Q^{opt}(X,a) < \epsilon\Big) = 0.$$

Then $V(d^{opt}) - V(d) \leq 4S[L(Q) - L(Q^{opt})]/\epsilon$ *and*
$V(d^{opt}) - V(d) \leq 4SE(T - T^{opt})^2/\epsilon$.

The proof is essentially the same as that of Theorem II.1 and is omitted.

Theorem II.1 implies that estimation based on minimization of the quadratic loss will yield a high quality individualized treatment rule if the associated estimator of Q^{opt} has prediction error close to $L(Q^{opt})$ (or more precisely, the part in Q^{opt} involving A is well estimated). In particular, if the excess prediction error converges to zero at a certain rate, then the estimated rule will have Value converges to the optimal Value at this rate to a power no smaller than 1/2.

2.2 Possible deviation from the goal

Recall that the non-concavity of the indicator function $1_{A=d(X)}$ has motivated us to use quadratic loss minimization method instead of directly maximizing an estimate of the Value. Below we demonstrate that although the convex minimization approach reduces the computational difficulty, it may, however, deviate from the goal of estimating the best individualized treatment rule.

In the previous section, we provided a quantitative relationship between the Value of an individualized treatment rule and the prediction error of the associated Q-function. This relationship is built through the optimal treatment rule d^{opt} and the true conditional mean function Q^{opt}. In practical implementation, we propose to estimate Q^{opt} over a function class, say $\mathcal{Q}$. The approximation space $\mathcal{Q}$ together with the definition of the estimated individualized treatment rule as the argmax of the estimated Q^{opt} places an implicit restriction on the space of decision rules that will be considered. In effect the space of decision rules is $\mathcal{D}_{\mathcal{Q}} = \{d(X) \in \arg\max_a Q(X,a) : Q \in \mathcal{Q}\}$. Thus it appears that the goal is to find the treatment rule in $\mathcal{D}_{\mathcal{Q}}$ that maximizes the Value. However, asymptotically the quadratic loss minimization method tries to estimate $d^*(X) \in \arg\max_a Q^*(X,a)$, where $Q^* = \arg\min_{Q \in \mathcal{Q}} L(Q)$. As one can see in the following example, d^* may not be the treatment rule in $\mathcal{D}_{\mathcal{Q}}$ when $Q^{opt} \notin \mathcal{Q}$.

A toy example

Suppose X is uniformly distributed in $[-1,1]$, A is binary $\{-1,1\}$ with probability $1/2$ each and is independent of X, and Y is normally distributed with mean $Q^{opt}(X,A) = (X-1/3)^2 A$ and variance 1. It is easy to see that the optimal individualized treatment rule satisfies $d^{opt}(X) = 1$ a.s. and $V(d^{opt}) = 4/9$. Now we consider model $\mathcal{Q} = \{Q(X,A;\boldsymbol{\theta}) = (1,X,A,XA)\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^4\}$ for Q^{opt}. Thus the space of decision rules under consideration is $\mathcal{D}_{\mathcal{Q}} = \{d(X) = sign(\theta_3 + \theta_4 X) : \theta_3, \theta_4 \in \mathbb{R}\}$. Note that $d^{opt} \in \mathcal{D}_{\mathcal{Q}}$ since $d^{opt}(X)$ can be written as $sign(\theta_3 + \theta_4 X)$ for any $\theta_3 > 0$ and $\theta_4 = 0$. d^{opt} is the best treatment rule in $\mathcal{D}_{\mathcal{Q}}$. However, minimizing the prediction error over $\mathcal{Q}$ yields the individualized treatment rule $d^*(X) = sign(2/3 - X)$, which has lower Value than d^{opt} ($V(d^*) = 29/81 < V(d^{opt})$). □

From this toy example we see that, if the approximation space for Q^{opt} is poor, estimation based on quadratic loss minimization may not (even asymptotically) reach the goal of maximizing the Value. In the rest of the dissertation, we consider two approaches to deal with this deviation. In the first approach, we consider a rich linear model for Q^{opt}, and we use l_1 penalization to avoid possible overfitting problem and produce a simple treatment rule. In the second approach, we consider a set of different models for Q^{opt}. An individualized treatment rule is estimated from each model using the quadratic loss minimization based method. The final rule will be chosen by maximizing a penalized Value estimator.

2.3 Proof of Theorem II.1

For any decision rule $d : \mathcal{X} \to \mathcal{A}$, denote $\triangle Q_d := \max_{a\in\mathcal{A}} Q^{opt}(X,a) - Q^{opt}(X, d(X))$. Following the arguments in Section 1.3, we have $V(d^{opt}) - V(d) = E(\triangle Q_d)$.

If $V(d^{opt}) - V(d) = 0$, then (2.3) and (2.4) automatically hold. Otherwise, $E(\triangle Q_d)^2 \geq (E\triangle Q_d)^2 > 0$. In this case, for any $\epsilon > 0$, define the event

$$\Omega_\epsilon = \Big\{0 < \max_{a\in\mathcal{A}} Q^{opt}(X,a) - \max_{a\in\mathcal{A}\backslash \arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} Q^{opt}(X,a) \leq \epsilon\Big\}.$$

Then on the event Ω_ϵ^C, $\triangle Q_d = 0$ or $\triangle Q_d > \epsilon$, which implies $\triangle Q_d \leq (\triangle Q_d)^2/\epsilon$. Thus

$$V(d^{opt}) - V(d) = E\big(1_{\Omega_\epsilon^C} \triangle Q_d\big) + E\big(1_{\Omega_\epsilon} \triangle Q_d\big) \leq \frac{1}{\epsilon} E\big[1_{\Omega_\epsilon^C} (\triangle Q_d)^2\big] + E\big(1_{\Omega_\epsilon} \triangle Q_d\big).$$

Since $\triangle Q_d \leq (\triangle Q_d)^2/\epsilon + \epsilon/4$, we have

$$V(d^{opt}) - V(d) \leq \frac{1}{\epsilon} E\big[1_{\Omega_\epsilon^C} (\triangle Q_d)^2\big] + E\Big[1_{\Omega_\epsilon}\Big(\frac{(\triangle Q_d)^2}{\epsilon} + \frac{\epsilon}{4}\Big)\Big] \leq \frac{1}{\epsilon} E\big[(\triangle Q_d)^2\big] + \frac{C}{4}\epsilon^{1+\alpha},$$

where the second inequality follows from the margin condition (2.2). Choosing $\epsilon = \big(4E(\triangle Q_d)^2/C\big)^{1/(2+\alpha)}$ to minimize the above upper bound yields

$$V(d^{opt}) - V(d) \leq 2^{\alpha/(2+\alpha)} C^{1/(2+\alpha)} \big[E(\triangle Q_d)^2\big]^{(1+\alpha)/(2+\alpha)}. \tag{2.5}$$

Next, for any d and Q such that $d(X) \in \max_{a\in\mathcal{A}} Q(X,a)$,

$$\begin{aligned}
&E(\triangle Q_d)^2\\
=&E\Big[\Big(\max_{a\in\mathcal{A}} Q^{opt}(X,a) - \max_{a\in\mathcal{A}} Q(X,a) + Q(X,d(X)) - Q^{opt}(X,d(X))\Big)^2\Big]\\
\leq&2E\Big[\Big(\max_{a\in\mathcal{A}} Q^{opt}(X,a) - \max_{a\in\mathcal{A}} Q(X,a)\Big)^2 + \Big(Q^{opt}(X,d(X)) - Q(X,d(X))\Big)^2\Big]
\end{aligned}$$

$$\leq 4E\Big[\max_{a\in\mathcal{A}}\big(Q^{opt}(X,a)-Q(X,a)\big)^2\Big],$$

where the last inequality follows from the fact that neither $|\max_{a\in\mathcal{A}}Q^{opt}(X,a)-\max_{a\in\mathcal{A}}Q(X,a)|$ nor $|Q(X,d(X))-Q^{opt}(X,d(X))|$ is larger than $\max_{a\in\mathcal{A}}|Q^{opt}(X,a)-Q(X,a)|$. Since $p(a|x)\geq S^{-1}$ for all (x,a) pairs, we have

$$E(\triangle Q_d)^2\leq 4SE\Big[\sum_{a\in\mathcal{A}}\big(Q^{opt}(X,a)-Q(X,a)\big)^2p(a|X)\Big]=4S\big[L(Q)-L(Q^{opt})\big]. \quad (2.6)$$

Inequality (2.3) follows by substituting (2.6) into (2.5).

In addition, note that $\triangle Q_d=\max_{a\in\mathcal{A}}T^{opt}(X,a)-T^{opt}(X,d(X))$ for any decomposition of $Q^{opt}(X,A)$ into $W^{opt}(X)+T^{opt}(X,A)$. Following the same procedure, we have that

$$E(\triangle Q_d)^2\leq 4E\Big[\max_{a\in\mathcal{A}}\big(T(X,a)-T^{opt}(X,a)\big)^2\Big]\leq 4SE[T(X,A)-T^{opt}(X,A)]^2$$

for any decomposition of $Q(X,A)$ into $W(X)+T(X,A)$. Inequality (2.4) follows immediately. □

CHAPTER III

Least squares with l_1 penalization

In this chapter, we consider an estimation procedure based on l_1 penalized least squares. And we provide a performance guarantee for the quality of the estimated individualized treatment rule.

3.1 l_1 penalized least squares

Let $\{(X_i, A_i, Y_i)\}_{i=1}^n$ represent i.i.d. observations on n subjects in a trial. For convenience, we use E_n to denote the associated empirical expectation (i.e. $E_n f = \sum_{i=1}^n f(X_i, A_i, Y_i)/n$ for any real-valued function f on $\mathcal{X} \times \mathcal{A} \times \mathbb{R}$). From the previous chapter, we see that if the interaction term (i.e. term involving A) in an estimated $\hat{Q}^{opt}$ is a high quality estimator of the interaction term in Q^{opt}, then the individualized treatment rule, $\hat{d}_n(X) = \arg\max_{a \in \mathcal{A}} \hat{Q}^{opt}(X, a)$, will have Value near optimal Value. Thus we focus on the estimation of Q^{opt}.

We estimate Q^{opt} via l_1-PLS (l_1 penalized least squares, Tibshirani 1996) over a linear approximation space $\mathcal{Q}$ for Q^{opt}. Because this is a convex optimization problem, the computational difficulty is reduced as compared to directly maximizing an empirical version of the Value.

We use penalization for two reasons. The first reason is that the use of least squares, while reducing computational difficulty, may, however, deviate from the goal of estimating an optimal individualized treatment rule if the interaction term in Q^{opt} is poorly modeled. As a result we consider complex models for Q^{opt}. The second reason is to deal with over-fitting due to the potentially large number of pretreatment variables (and/or complex approximation space for Q^{opt}). To illustrate this issue, consider the setting in which we know the form of Q^{opt} 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 correct linear model (i.e. the model that only contains variables with true nonzero coefficients) may result in decision rules with poor Value as well as estimated $\hat{Q}^{opt}$ with large prediction error. 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 nonzero coefficients (Massart, 2005). Indeed we will see in Theorem III.3 that the l_1-PLS method estimates a parameter with balanced prediction error and sparsity. As a result, the individualized treatment rule produced by l_1-PLS will more reliably have higher Value than the rule produced by the least squares estimator constructed when the correct model is known but is too complex relative to the size of the data set.

We selected l_1 penalization as this penalty does some variable selection and as a result will help us to construct individualized treatment rules that are cheaper to implement (fewer variables to collect per patient) and easier to interpret. See Section 3.3 for the discussion of other potential penalization methods.

Let $\mathcal{Q} := \{Q(X, A; \boldsymbol{\theta}) = \Phi(X, A)\boldsymbol{\theta}, \boldsymbol{\theta} \in \mathbb{R}^J\}$ be the approximation space for Q^{opt}, where $\Phi(X, A) = (\phi_1(X, A), \ldots, \phi_J(X, A))$ is a 1 by J vector composed of basis functions on $\mathcal{X} \times \mathcal{A}$, $\boldsymbol{\theta}$ is a J by 1 parameter vector, and J is the number of basis functions (for clarity here J will be fixed in n, see Section 3.4.1 for results with J

increasing as n increases). The l_1-PLS estimator of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^J} \Big\{ E_n[Y - \Phi(X,A)\boldsymbol{\theta}]^2 + \lambda_n \sum_{j=1}^{J} \hat{\sigma}_j |\theta_j| \Big\}, \tag{3.1}$$

where $\hat{\sigma}_j = \big[E_n \phi_j(X,A)^2\big]^{1/2}$, θ_j is the j^{th} component of $\boldsymbol{\theta}$ and λ_n is a tuning parameter that controls the amount of penalization. The weights $\hat{\sigma}_j$'s are used to balance the scale of different basis functions; these weights were used in Bunea et al. (2007b) and van de Geer (2008). In some situations, it is natural to penalize only a subset of coefficients and/or use different weights in the penalty; see Section 3.4.3 for modifications of $\hat{\boldsymbol{\theta}}_n$ to this case. The resulting estimated individualized treatment rule satisfies

$$\hat{d}_n(X) \in \arg\max_{a\in\mathcal{A}} \Phi(X,a)\hat{\boldsymbol{\theta}}_n. \tag{3.2}$$

3.1.1 Performance guarantees for the l_1-PLS

In this section we prove that the Value of the individualized treatment rule produced by the l_1-PLS method is larger than the optimal Value minus a quantity with high probability. As the sample size goes to infinity, this quantity converges to a constant which will be small if the interaction term in Q^{opt} is approximated sufficiently well. Even though we hope to have a good approximation model, the results below do not require this condition to hold.

Define $M_0(\boldsymbol{\theta}) = \{j = 1, \ldots, J : \theta_j \neq 0\}$. For a set M, let N_M denote the cardinality of M.

Let $\boldsymbol{\theta}^* \in \mathbb{R}^J$ be the prediction error minimizer in the linear model, i.e.

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^J} L(\Phi\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^J} E(Y - \Phi\boldsymbol{\theta})^2. \tag{3.3}$$

Note that the minimizer of $L(\Phi\boldsymbol{\theta})$ may not be unique. In that case, we use $[\boldsymbol{\theta}^*]$ to denote the equivalence class of $\boldsymbol{\theta}^*$ that contains all $\boldsymbol{\theta}$s having the same prediction error as $\boldsymbol{\theta}^*$. Let $\|f\|_\infty = \sup_{x\in\mathcal{X},a\in\mathcal{A}} |f(x,a)|$ for any bounded function f on $\mathcal{X} \times \mathcal{A}$. For any $\gamma \in [0, 1/2)$, $\eta_1 \geq 0$, $t > 0$ and tuning parameter $\lambda_n > 0$, define the sets

$$\Theta_n^o = \Big\{\boldsymbol{\theta} \in \mathbb{R}^J : \exists\ \boldsymbol{\theta}^o \in [\boldsymbol{\theta}^*] \text{ s.t. } \|\Phi(\boldsymbol{\theta} - \boldsymbol{\theta}^o)\|_\infty \leq \eta_1 \\ \text{and } \max_{j=1,\ldots,J} \left|\frac{E[\phi_j \Phi(\boldsymbol{\theta} - \boldsymbol{\theta}^o)]}{\sigma_j}\right| \leq \gamma\lambda_n\Big\} \quad (3.4)$$

$$\Theta_n = \Big\{\boldsymbol{\theta} \in \Theta_n^o : N_{M_0(\boldsymbol{\theta})} \leq \frac{(1-2\gamma)^2\beta}{120}\Big(\sqrt{\frac{1}{9} + \frac{n}{2U^2[\log(3J(J+1)) + t]}} - \frac{1}{3}\Big)\Big\}, \quad (3.5)$$

where $\sigma_j = (E\phi_j^2)^{1/2}$, and β and U are positive constants that will be defined in Theorem III.1. For small η_1 and γ, Θ_n^o contains $\boldsymbol{\theta}$s that are close to the elements in $[\boldsymbol{\theta}^*]$ (note that γ controls the closeness between $\boldsymbol{\theta} \in \Theta_n^o$ and $[\boldsymbol{\theta}^*]$ via first order derivatives of the prediction errors since $|E[\phi_j\Phi(\boldsymbol{\theta} - \boldsymbol{\theta}^o)]/\sigma_j| = \left|\partial L(\Phi\boldsymbol{\theta})/\partial\theta_j - \partial L(\Phi\boldsymbol{\theta}^o)/\partial\theta_j^o\right|/2\sigma_j)$. Θ_n^o is non-empty. Θ_n contains all $\boldsymbol{\theta} \in \Theta_n^o$ that have the required sparsity (note that Θ_n is always non-empty for large n since $N_{M_0(\boldsymbol{\theta})} \leq J$). In the following, we fist provide an upper bound for the Value of $\hat{d}_n$ in terms of the prediction error.

Theorem III.1. *Suppose $p(a|x) \geq S^{-1}$ for a positive constant S for all (x,a) pairs and the margin condition (2.2) holds for some $C > 0$, $\alpha \geq 0$ and all positive ϵ satisfying $C\epsilon^\alpha \leq 1$. Assume*

1. *the error terms $\varepsilon_i = Y_i - Q^{opt}(X_i, A_i), i = 1, \ldots, n$, are i.i.d. with $E(\varepsilon_i|X_i, A_i) = 0$ and $E[|\varepsilon_i|^l] \leq \frac{l!}{2}c^{l-2}\sigma^2$ for some $c, \sigma^2 > 0$ for all $l \geq 2$; and*

2. *there exist constants $0 < U < \infty$ and $0 \leq \eta_2 < \infty$ such that $\max_{j=1,\ldots,J}$ $\|\phi_j\|_\infty/\sigma_j \leq U$ and $\sup_{\boldsymbol{\theta}\in[\boldsymbol{\theta}^*]} \|Q^{opt} - \Phi\boldsymbol{\theta}\|_\infty \leq \eta_2$.*

For any $\eta_1 \geq 0$, $0 \leq \gamma < 1/2$ and $t > 0$, consider the estimated individualized *treatment rule* $\hat{d}_n$ *defined by (3.2) with tuning parameter*

$$\lambda_n \geq k \max\Big\{\frac{\log 6J + t}{n}, \sqrt{\frac{\log 6J + t}{n}}\Big\}, \tag{3.6}$$

where $k = (8\max\{3c, 2(\eta_1+\eta_2)\}U + 12\sqrt{2}\max\{\sigma, \eta_1+\eta_2\})/(1-2\gamma)$. *Let* Θ_n^o *be the set of parameters defined in (3.4). Assume*

3. *there exists a constant* $\beta > 0$ *such that, for all* $\boldsymbol{\theta} \in \Theta_n^o \setminus \{\mathbf{0}\}$ *and* $\tilde{\boldsymbol{\theta}} \in \{\mathbb{R}^J : \sum_{j\in\{1,\ldots,J_n\}\setminus M_0(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j| \leq (2\gamma+5)\sum_{j\in M_0(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j - \theta_j|/(1-2\gamma)\}$,

$$E[\Phi(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2 N_{M_0(\boldsymbol{\theta})} \geq \beta\Big(\sum_{j\in M_0(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j - \theta_j|\Big)^2.$$

Let Θ_n *be the set of parameters defined in (3.5). Then for any* $n \geq (27U^2 - 10\gamma - 22)\log 2J/[2(1-2\gamma)^2]$ *and for which* Θ_n *is non-empty, we have, with probability at least* $1 - \exp\big(-k'n) - \exp(-t)$, *that*

$$V(d^{opt}) - V(\hat{d}_n) \leq C_1\Big[\min_{\boldsymbol{\theta}\in\Theta_n}\Big(L(\Phi\boldsymbol{\theta}) + KN_{M_0(\boldsymbol{\theta})}\lambda_n^2/\beta\Big)\Big]^{(1+\alpha)/(2+\alpha)}$$

where $k' = 13(1-2\gamma)^2/[6(27U^2 - 10\gamma - 22)]$, $K = 50(2\gamma+5)(4\gamma^2+116\gamma+13)/[9(1-2\gamma)(2\gamma+19)^2]$ *and* C_1 *is defined in Theorem II.1.*

The result follows from Theorem II.1 and Theorem III.3 in Section 3.1.2. Similar results for J increasing in n can be obtained by combining Theorem II.1 with Theorem III.4 in Section 3.4.1

Remarks

Assumptions 1-3 in Theorem III.1 are employed to derive the finite sample prediction error bound for the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ defined in (3.1). Below we briefly discuss those assumptions.

1. Assumption 1 implicitly implies that the error terms do not have heavy tails. It is easy to verify that this assumption holds if each ε_i is bounded. Moreover, it also holds for some commonly used error distributions that have unbounded support, such as the normal or double exponential. This condition is often assumed to show that the sample mean of a variable is concentrated around its true mean with a high probability.

2. Assumption 2 implies that Q^{opt} and all basis functions are bounded. Note that we do not assume $\mathcal{Q}$ to be a good approximation space for Q^{opt}. However, if $\Phi\boldsymbol{\theta}^*$ approximates Q^{opt} well, η_2 will be small, which will result in a smaller upper bound in (3.8). This assumption is also used to show the concentration of the sample mean around the true mean. It is possible to replace the boundedness condition by conditions on moments similar to those in Assumption 1.

3. Assumption 3 employed to avoid collinearity. It is easy to verify that when $E[\phi_j\phi_k/(\sigma_j\sigma_k)]_{j,k\in\{1,\ldots,J\}}$ is positive definite, this condition trivially holds with β to be the smallest eigenvalue of $E[\phi_j\phi_k/(\sigma_j\sigma_k)]_{j,k\in\{1,\ldots,J\}}$. Similar conditions have been used in van de Geer (2008), where the minimum is taken over all $\boldsymbol{\theta} \in \mathbb{R}^J$. Assumption 3 is also similar to the restricted eigenvalue assumptions in Bickel et al. (2009)) in which E is replaced by E_n, and a fixed design matrix is considered. It is satisfied if the "mutual coherence" assumption in Bunea et al. (2007b) ($N_{M_0(\boldsymbol{\theta})} \max_{j\neq k, j\in M_0(\boldsymbol{\theta})} |E\phi_j\phi_k|/(\sigma_j\sigma_k) \leq$ a small constant) holds for all $\boldsymbol{\theta} \in \Theta_n^o$ (similar results in the fixed design setting have been proved in Bickel et al. (2009)). See Bickel et al. (2009) for other sufficient conditions for Assumption 3.

Define $T^{opt}(X, A) := Q^{opt}(X, A) - E[Q^{opt}(X, A)|X]$. Then T^{opt} is the interaction term in Q^{opt}. In particular, the vector of basis functions can be written as $\Phi(X, A) = (\Phi^{(1)}(X), \Phi^{(2)}(X, A))$, where $\Phi^{(1)} = (\phi_1(X), \ldots, \phi_{J^{(1)}}(X))$ is composed of all components in Φ that do not contain A and $\Phi^{(2)} = (\phi_{J^{(1)}+1}(X,A), \ldots, \phi_J(X,A))$ is composed of all components in Φ that contain A. Since A takes only finite values and the randomization distribution $p(a|x)$ is known, we can code A so that $E[\Phi^{(2)}(X,A)^T|X] = \mathbf{0}$ a.s. (see Section 3.2.1 for examples). For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T \in \mathbb{R}^J$, denote $\boldsymbol{\theta}^{(1)} = (\theta_1, \ldots, \theta_{J^{(1)}})^T$ and $\boldsymbol{\theta}^{(2)} = (\theta_{J^{(1)}+1}, \ldots, \theta_J)^T$. Then $\Phi^{(1)}\boldsymbol{\theta}^{(1)}$ approximates $E(Q^{opt}(X,A)|X)$ and $\Phi^{(2)}\boldsymbol{\theta}^{(2)}$ approximates T^{opt}. Define $M_0^{(1)}(\boldsymbol{\theta}) = \{j = 1, \ldots, J^{(1)} : \theta_j \neq 0\}$ and $M_0^{(2)}(\boldsymbol{\theta}) = \{j = J^{(1)}+1, \ldots, J : \theta_j \neq 0\}$

In the following we relate the Value of $\hat{d}_n$ to the estimator T^{opt}. Note that the conclusion of Theorem III.1 and the following theorem hold for all choices of λ_n that satisfy (3.6). Suppose $\lambda_n = o(1)$. The following theorem implies that if T^{opt} can be well approximated by a sparse representation (i.e. $E(\Phi^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2$ and $N_{M_0^{(2)}(\boldsymbol{\theta})}$ are small for some $\boldsymbol{\theta} \in \Theta_n$), then $\hat{d}_n$ will have Value close to the optimal Value.

Theorem III.2. *Suppose $p(a|x) \geq S^{-1}$ for a positive constant S for all (x,a) pairs and the margin condition (2.2) holds for some $C > 0$, $\alpha \geq 0$ and all positive ϵ satisfying $C\epsilon^\alpha \leq 1$. Suppose $E[\Phi^{(2)}(X,A)^T|X] = 0$ a.s. and Assumptions 1 and 2 in Theorem III.1 hold. For any $0 \leq \gamma < 1/2$, $\eta_1 \geq 0$ and $t > 0$, let $\hat{\boldsymbol{\theta}}_n$ be the l_1-PLS estimator defined in (3.1) with λ_n satisfying condition (3.6) and Θ_n^o be defined in (3.4). Assume*

4. there exists a constant $\beta > 0$, such that for all $\boldsymbol{\theta} \in \Theta_n^o \setminus \{\mathbf{0}\}$ and $\tilde{\boldsymbol{\theta}} \in \{\mathbb{R}^J : \sum_{j \in \{1,\ldots,J_n\} \setminus M_0(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j| \leq (2\gamma+5) \sum_{j \in M_0(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j - \theta_j| / (1-2\gamma)\}$,

$$E[\Phi^{(1)}(\tilde{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^{(1)})]^2 N_{M_0^{(1)}(\boldsymbol{\theta})} \geq \beta\Big(\sum_{j \in M_0^{(1)}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j - \theta_j|\Big)^2 \qquad (3.7)$$

$$\textit{and} \quad E[\Phi^{(2)}(\tilde{\boldsymbol{\theta}}^{(2)} - \boldsymbol{\theta}^{(2)})]^2 N_{M_0^{(2)}(\boldsymbol{\theta})} \geq \beta\Big(\sum_{j \in M_0^{(2)}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j - \theta_j|\Big)^2.$$

Let Θ_n be defined in (3.5). Then for any $n \geq (27U^2 - 10\gamma - 22)\log 2J/[2(1-2\gamma)^2]$ and *for which* Θ_n *is non-empty, we have, with probability at least* $1-\exp\big(-k'n)-\exp(-t)$, *that*

$$V(d^{opt}) - V(\hat{d}_n) \leq C_1 \Big[\min_{\boldsymbol{\theta} \in \Theta_n} \Big(E(\Phi^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2 + K' N_{M_0^{(2)}(\boldsymbol{\theta})} \lambda_n^2 / \beta \Big) \Big]^{\frac{1+\alpha}{2+\alpha}}, \qquad (3.8)$$

where $K' = 20\gamma(2\gamma+5)/[7(1-2\gamma)] + 200(2\gamma+5)^2/[9(2\gamma+19)^2]$, C_1 *is defined in Theorem II.1, and* k' *is defined in Theorem III.1.*

The result follows from Theorem II.1 and Corollary III.1 in Section 3.1.2.

Remark

Assumption 4 is a sufficient condition for Assumption 3 in Theorem III.1. We need (3.7) to show that the cross product term $E_n[(\Phi^{(1)}\hat{\boldsymbol{\theta}}_n^{(1)} - \Phi^{(1)}\boldsymbol{\theta}_n^{(1)})(\Phi^{(2)}\hat{\boldsymbol{\theta}}_n^{(2)} - \Phi^{(2)}\boldsymbol{\theta}_n^{(2)})]$ converging to 0 at the desired rate. We may use a really poor model for $E(Q^{opt}(X,A)|X)$ (e.g. $\Phi^{(1)} \equiv 1$, and (3.7) holds with $\beta = 1$). When the sample size is large (so that λ_n is small), the estimated treatment rule will be of high quality as long as T^{opt} is well approximated.

3.1.2 Prediction error bound for the l_1-PLS estimator

l_1-penalization in regression has been extensively studied in recent years. Much literature focused on variable selection/parameter estimation accuracy (see Meinshausen and Buhlmann 2006; Zhao and Yu 2006; Zhang and Huang 2008; Meinshausen and Buhlmann 2009; Zhang 2009 for examples). Others studied the behavior of the prediction loss (see Greenshtein 2006; Bunea et al. 2007a,b; van de Geer 2008; Bickel et al. 2009; Koltchinskii 2009 for examples). We are mainly interested in the latter.

In this section we provide a finite sample upper bound for the prediction error of the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$. We present the result here for the following two reasons. First, the result is needed to prove Theorem III.1. Second, the result itself strengthens existing literature on l_1-PLS method in prediction in the following way. Finite sample prediction error bounds for the l_1-PLS estimator in the random design setting have been provided in Bunea et al. (2007b) for quadratic loss, van de Geer (2008) mainly for general Lipschitz loss functions and Koltchinskii (2009) for loss functions satisfying some conditions. With the quadratic loss, and permitting J to increase with n (so Φ depends on n as well), both Bunea et al. (2007b) and van de Geer (2008) assumed the existence of some sparse $\boldsymbol{\theta} \in \mathbb{R}^J$ such that $E(\Phi\boldsymbol{\theta} - Q^{opt})^2$ is upper bounded by a quantity that decreases to 0 at a certain rate as $n \to \infty$; while Koltchinskii (2009) requires the primary outcome Y to be bounded. We improve the results in the sense that we do not make any of these assumptions (see Section 3.4.1 for results when Φ, J are indexed by n and J diverges).

In this section we consider the case where the sparsity of $\boldsymbol{\theta}$ is measured by the number of nonzero components (see Section 3.4.1 for proofs with a laxer definition of sparsity). The l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ estimates a parameter with balanced prediction error and sparsity. This target parameter lies in Θ_n defined in (3.5). By definition, elements in Θ_n have prediction error close to $\boldsymbol{\theta}^*$ and have the required sparsity. When Θ_n is non-empty, we define

$$\boldsymbol{\theta}^{**}(u) = \arg\min_{\boldsymbol{\theta}\in\Theta_n} \left[L(\Phi\boldsymbol{\theta}) + uN_{M_0(\boldsymbol{\theta})}\right] \tag{3.9}$$

for any $u > 0$. Note that $\boldsymbol{\theta}^{**}(u)$ is at least as sparse as $\boldsymbol{\theta}^*$ since by (3.3), $L(\Phi\boldsymbol{\theta}) + uN_{M_0(\boldsymbol{\theta})} > L(\Phi_n\boldsymbol{\theta}^*) + uN_{M_0(\boldsymbol{\theta}^*)}$ for any $\boldsymbol{\theta}$ such that $N_{M_0(\boldsymbol{\theta})} > N_{M_0(\boldsymbol{\theta}^*)}$. Thus the individualized treatment rule produced by $\boldsymbol{\theta}^{**}(u)$ could be simpler than the rule produced by $\boldsymbol{\theta}^*$. The l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ estimates $\boldsymbol{\theta}^{**}(u_n)$ for a particular "u_n" that gives a balanced prediction error and sparsity. Under appropriate conditions, $u_n \to 0$ as $n \to \infty$ (see remark 1 after Theorem III.3); in this case the prediction error of $\boldsymbol{\theta}^{**}(u_n)$ converges to $L(\Phi\boldsymbol{\theta}^*)$ as $n \to \infty$ (since $N_{M_0(\boldsymbol{\theta})} \leq J$).

In the following theorem, we show that $L(\Phi\hat{\boldsymbol{\theta}}_n) \leq L(\Phi\boldsymbol{\theta}^{**}(u_n)) + u_n N_{M_0(\boldsymbol{\theta}^{**}(u_n))}$ with high probability. That is, up to the $u_n N_{M_0(\boldsymbol{\theta}^{**}(u_n))}$ term, $\hat{\boldsymbol{\theta}}_n$ will have prediction error roughly as if the sparseness of $\boldsymbol{\theta}^{**}(u_n)$ were known.

Theorem III.3. *Suppose Assumptions 1 and 2 in Theorem III.1 hold. For any $0 \leq \gamma < 1/2$, $\eta_1 \geq 0$ and $t > 0$, let $\hat{\boldsymbol{\theta}}_n$ be the l_1-PLS estimator defined in (3.1) with λ_n satisfying condition (3.6) and Θ_n^o be defined in (3.4). Suppose Assumption 3 in Theorem III.1 holds. Let Θ_n and $\boldsymbol{\theta}^{**}(u)$ be defined in (3.5) and (3.9), respectively. Then for any $n \geq (27U^2-10\gamma-22)\log 2J/[2(1-2\gamma)^2]$ and for which Θ_n is non-empty, we have, with probability at least $1 - \exp\big(-k'n\big) - \exp(-t)$, that*

$$L(\Phi\hat{\boldsymbol{\theta}}_n) \leq \min_{\boldsymbol{\theta}\in\Theta_n} \Big(L(\Phi\boldsymbol{\theta}) + u_n N_{M_0(\boldsymbol{\theta})}\Big) = L\big(\Phi\boldsymbol{\theta}^{**}(u_n)\big) + u_n N_{M_0(\boldsymbol{\theta}^{**}(u_n))}, \tag{3.10}$$

where $u_n = K\lambda_n^2/\beta$, k' and K are defined in Theorem III.1.

The results follow directly from Theorem III.4 in Section 3.4.1 with $\rho = 0$.

Remarks:

1. The conclusion of Theorem III.3 holds for all choices of λ_n that satisfy (3.6). Suppose $\lambda_n = o(1)$, then $L(\Phi\boldsymbol{\theta}^{**}(u_n)) - L(\Phi\boldsymbol{\theta}^*) \to 0$ as $n \to \infty$ (since $N_{M_0(\boldsymbol{\theta})}$ is bounded). Then Theorem III.3 implies that $L(\Phi\hat{\boldsymbol{\theta}}_n) - L(\Phi\boldsymbol{\theta}^*) \to 0$ in probability. To achieve the best rate of convergence, equal sign should be taken in (3.6).

2. Note that $\boldsymbol{\theta}^{**}(u_n)$ defined by (3.9) is the parameter in Θ_n that minimizes $L(\Phi\boldsymbol{\theta}) - L(Q^{opt}) + u_n N_{M_0(\boldsymbol{\theta})}$. Intuitively, the minimum of $L(\Phi\boldsymbol{\theta}) - L(Q^{opt}) + u_n N_{M_0(\boldsymbol{\theta})}$ can be viewed as the approximation error plus a "tight" upper bound of the estimation error of an "oracle" in the stepwise model selection framework (when "=" is taken in (3.6)). Here "tight" means the convergence rate in the bound is the best known rate, and "oracle" is defined as follows.

Let m denote a non-empty subset of the index set $\{1, \ldots, J\}$. Then each m represents a model which uses a non-empty subset of $\{\phi_1, \ldots, \phi_J\}$ as basis functions. Define $\hat{\boldsymbol{\theta}}_{n,m} = \arg\min_{\{\boldsymbol{\theta}\in\mathbb{R}^J:\theta_j=0,j\notin m\}} E_n(Y-\Phi\boldsymbol{\theta})^2$ and $\boldsymbol{\theta}^*_m = \arg\min_{\{\boldsymbol{\theta}\in\mathbb{R}^J:\theta_j=0,j\notin m\}} E(Y-\Phi\boldsymbol{\theta})^2$. In this setting, an ideal model selection criterion will pick model m^* such that $L(\Phi\hat{\boldsymbol{\theta}}_{n,m^*}) = \inf_m L(\Phi\hat{\boldsymbol{\theta}}_{n,m})$. $\hat{\boldsymbol{\theta}}_{n,m^*}$ is referred as an "oracle" in Massart (2005). Note that the excess prediction error of each $\hat{\boldsymbol{\theta}}_{n,m}$ can be written as

$$L(\Phi\hat{\boldsymbol{\theta}}_{n,m}) - L(Q^{opt}) = \big[L(\Phi\boldsymbol{\theta}^*_m) - L(Q^{opt})\big] + \big[L(\Phi\hat{\boldsymbol{\theta}}_{n,m}) - L(\Phi\boldsymbol{\theta}^*_m)\big],$$

where the first term is called the approximation error of model m and the second term is the estimation error. It can be shown that (Bartlett, 2008) for each model m and $x_m > 0$, with probability at least $1 - \exp(-x_m)$,

$$L(\Phi\hat{\boldsymbol{\theta}}_{n,m}) - L(\Phi\boldsymbol{\theta}^*_m) \leq \mathit{constant} \times \Big(\frac{x_m + N_m \log(n/N_m)}{n}\Big)$$

under appropriate technical conditions, where N_m is the cardinality of the index set m. To our knowledge this is the best rate known so far. Taking $x_m = \log n + N_m \log J$ and using the union bound argument, we have with probability at least $1 - O(1/n)$,

$$\begin{aligned}
& L(\Phi_n\hat{\boldsymbol{\theta}}_{n,m^*}) - L(Q^{opt}) \\
=& \min_m \Big(\big[L(\Phi\boldsymbol{\theta}^*_m) - L(Q^{opt})\big] + L(\Phi\hat{\boldsymbol{\theta}}_{n,m}) - L(\Phi\boldsymbol{\theta}^*_m)\Big) \\
\leq& \min_m \Big(\big[L(\Phi\boldsymbol{\theta}^*_m) - L(Q^{opt})\big] + \mathit{constant} \times \frac{N_m(\log J + \log n)}{n}\Big) \\
=& \min_{\boldsymbol{\theta}} \Big(\big[L(\Phi\boldsymbol{\theta}) - L(Q^{opt})\big] + \mathit{constant} \times \frac{N_{M_0(\boldsymbol{\theta})}(\log J + \log n)}{n}\Big). \qquad (3.11)
\end{aligned}$$

On the other hand, take $t = \log(n/6)$ in (11) and select λ_n so that condition (3.6) holds with "=". We have $\lambda_n = \mathit{constant} \times \sqrt{(\log J + \log n)/n}$ for large n.

(3.10) implies that, with probability at least $1 - O(1/n)$,

$$L(\Phi\hat{\boldsymbol{\theta}}_n) - L(Q^{opt}) \leq \min_{\boldsymbol{\theta}\in\Theta_n} \Big(\big[L(\Phi\boldsymbol{\theta}) - L(Q^{opt})\big] + constant \times \frac{N_{M_0(\boldsymbol{\theta})}(\log J + \log n)}{n} \Big),$$

which is essentially (3.11) with the constraint of $\boldsymbol{\theta} \in \Theta_n$. (The "*constant*" in the above inequalities may take different values.) Since the minimum is achieved at $\boldsymbol{\theta} = \boldsymbol{\theta}^{**}(u_n)$, we refer to $\boldsymbol{\theta}^{**}(u_n)$ as an oracle.

3. Note that $\hat{\boldsymbol{\theta}}_n$ minimizes $E_n(R - \Phi\boldsymbol{\theta})^2$ plus an l_1 penalty whereas $\boldsymbol{\theta}^{**}(u_n)$ minimizes the prediction error $L(\Phi\boldsymbol{\theta})$ plus an l_0 penalty. We provide an intuitive connection between these two quantities. First note that $E_n(Y - \Phi\boldsymbol{\theta})^2$ estimates $L(\Phi\boldsymbol{\theta})$ and $\hat{\sigma}_j$ estimates σ_j. We use "$\approx$" to denote this relationship. Thus

$$\begin{aligned} &E_n(Y - \Phi\boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^{J} \hat{\sigma}_j |\theta_j| \qquad (3.12)\\ \approx &L(\Phi\boldsymbol{\theta}) + \lambda_n \sum_{j=1}^{J} \sigma_j |\theta_j| \\ \leq &L(\Phi\boldsymbol{\theta}) + \lambda_n \sum_{j=1}^{J} \sigma_j |\hat{\theta}_{n,j} - \theta_j| + \lambda_n \sum_{j=1}^{J} \sigma_j |\hat{\theta}_{n,j}|, \end{aligned}$$

where $\hat{\theta}_{n,j}$ is the j^{th} component of $\hat{\boldsymbol{\theta}}_n$. In Section 3.4.1 we show that for any $\boldsymbol{\theta} \in \Theta_n$, $\lambda_n \sum_{j=1}^{J} \sigma_j |\hat{\theta}_{n,j} - \theta_j|$ is upper bounded by $N_{M_0(\boldsymbol{\theta})}\lambda_n^2/\beta$ up to a constant with a high probability. Thus $\hat{\boldsymbol{\theta}}_n$ minimizes (3.12) and $\boldsymbol{\theta}^{**}(u_n)$ roughly minimizes an upper bound of (3.12).

4. The constants involved in the theorem can be improved; we focused on a readable result rather than providing best constants.

Following Theorem III.3 and Theorem III.4 in Section 3.4.1, we obtain a finite sample upper bound on the mean square difference between T^{opt} and its estimator. This upper bound implies that l_1-PLS estimator of T^{opt} is of high quality as long as

T^{opt} can be approximated sufficiently well by a sparse linear representation in the approximation model.

Corollary III.1. *Suppose $E[\Phi^{(2)}(X,A)^T|X] = 0$ a.s. and Assumptions 1 and 2 in Theorem III.1 hold. For any $0 \leq \gamma < 1/2$, $\eta_1 \geq 0$ and $t > 0$, let $\hat{\boldsymbol{\theta}}_n$ be the l_1-PLS estimator defined in (3.1) with λ_n satisfying condition (3.6) and Θ_n^o be defined in (3.4). Suppose Assumption 4 in Theorem III.1 holds. Let Θ_n be defined in (3.5). Then for any $n \geq (27U^2-10\gamma-22)\log 2J/[2(1-2\gamma)^2]$ and for which Θ_n is non-empty, we have, with probability at least $1-\exp\big(-k'n)-\exp(-t)$, we have*

$$E(\Phi^{(2)}\hat{\boldsymbol{\theta}}_n^{(2)} - T^{opt})^2 \leq \min_{\boldsymbol{\theta}\in\Theta_n} \Big(E(\Phi^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2 + K'N_{M_0^{(2)}(\boldsymbol{\theta})}\lambda_n^2/\beta\Big), \qquad (3.13)$$

where k' and K' are defined in Theorem III.1.

3.2 Numerical Studies

The proof of the theorems in the previous section requires a non-stochastic tuning parameter. However in practical implementation, it is more realistic to use data-dependent methods to select λ_n. Since our primary goal is to maximize the Value, we select λ_n to maximize the cross-validated Value. For any individualized treatment rule d, it is easy to verify that $E[1_{A=d(X)}/p(A|X)] = 1$. Thus an unbiased estimator of $V(d)$ is $E_n[1_{A=d(X)}Y/p(A|X)]/E_n[1_{A=d(X)}/p(A|X)]$ (Murphy et al., 2001). We split the data into 10 roughly equal-sized parts; then we apply the l_1-PLS based method on each 9 parts of the data to obtain a treatment rule, and estimate the Value of this rule using the remaining part (i.e. the average of $1_{A=d(X)}Y/p(A|X)$ divided by the average of $1_{A=d(X)}/p(A|X)$ over the remaining part); this method will select λ_n that maximizes the average of the 10 estimated Values. Since the Value of an individualized treatment rule is noncontinuous in the parameters, the resulting λ_n is usually non-unique. If necessary, we select the λ_n that produces the simplest decision rule (the rule using the least number of variables), from the set of λ_n's that maximize the average estimated Value. In the simulation below this second criterion effectively reduced the number of candidate λ_n around 25% of the time, and multiple λ_n still remained around 90% of the time. This is not surprising since the Value of a decision rule only depends on the relative magnitudes of parameters in the decision rule. In this case, we select the one among the remaining λ_n that minimizes the 10-fold cross validated prediction error estimator; that is, minimization of the prediction error is used as a final tie breaker.

In Section 3.2.1, we evaluate the l_1-PLS based method. In Section 3.2.2, we use data collected from the Nefazodone-CBASP trial (Keller et al., 2000) to illustrate the application of the l_1-PLS based method.

3.2.1 Simulations

In this section we evaluate the l_1-PLS based method by comparing it with treatment assignment via separate prognosis prediction for each treatment.

Prognosis prediction is the prediction of the outcome of a disease following a treatment. Usually this method is used on multiple data sets, each of which involves one active treatment. A natural approach to individualizing treatment is then to compare the predicted prognosis of a patient for each treatment and recommend the treatment that is associated with the best predicted prognosis (Kent et al., 2002). As a comparison to the l_1-PLS method, we estimate the prognosis $E(Y|X, A = a)$ via least squares with l_1 penalization separately for each treatment $a \in \mathcal{A}$. We call this method the "prognosis prediction" approach. In this approach the individualized treatment rule results in the treatment that yields the best predicted prognosis (i.e. the estimated individualized treatment rule satisfies $\hat{d}_{PP}(X) \in \arg\max_{a\in\mathcal{A}} \hat{E}(Y|X, A = a)$). The tuning parameters involved in this approach will be selected by minimizing the 10-fold cross-validated prediction error estimator. In the following examples, the approximation model we use for prognosis prediction under each treatment is consistent with the model we use in l_1-PLS (e.g. if Q^{opt} is approximated by $(1, X, A, XA)\boldsymbol{\theta}$ in l_1-PLS, then we approximate $E(Y|X, A = a)$ by $(1, X)\boldsymbol{\theta}_{PP}$ for each treatment group in the prognosis prediction approach). The intercept is penalized in neither method.

We consider 9 examples. In all the examples, treatment A is generated from $\{-1, 1\}$ independent of X with probability 1/2 each, and the outcome Y given X and A is normally distributed with mean Q^{opt}. In examples 1-3, we consider $X \in \mathbb{R}^5$ and three simple examples for Q^{opt}. In example 4, we consider $X \sim U[0, 1]$ and a complex Q^{opt}, which mimics the blocks function used in Donoho and Johnstone (Donoho and Johnstone, 1994). To make the simulations more realistic, examples 5-9 are based on data from the Nefazodone-CBASP trial (Keller et al. (2000), see Section 3.2.2 for description of the trial). We consider 50 pretreatment variables collected from the trial (i.e. $X \in \mathbb{R}^{50}$) and five examples for Q^{opt}. Detailed simulation design for the examples are presented in Section 3.4.2.

For example 4, we approximate Q^{opt} by Haar wavelets. The number of basis functions may increase as n increases (we index J, Φ and $\boldsymbol{\theta}^*$ by n in this case). Example plots for $Q^{opt}(X, A)$ and the associated best wavelet fits $\Phi_n(X, A)\boldsymbol{\theta}_n^*$ are provided in Figure 3.1. For all other examples, we approximate Q^{opt} by $(1, X, A, XA)\boldsymbol{\theta}$.

In examples 1, 2, 5 and 6, the interaction term $T^{opt}(= Q^{opt} - E(Q^{opt}|X))$ is contained in the analysis model. In particular, there is no treatment effect in example 1 and 5 (i.e. $T^{opt} \equiv 0$). In other examples, the analysis model does not contain T^{opt}. However, in example 4 the Haar wavelets approximate Q^{opt} (and thus T^{opt}) sufficiently well when J_n is large.

For each of the examples 1 - 4, we simulate data sets of sizes n between 40 and 1000. For each of the examples 5 - 9, we simulate data sets of size $n = 500$. 500

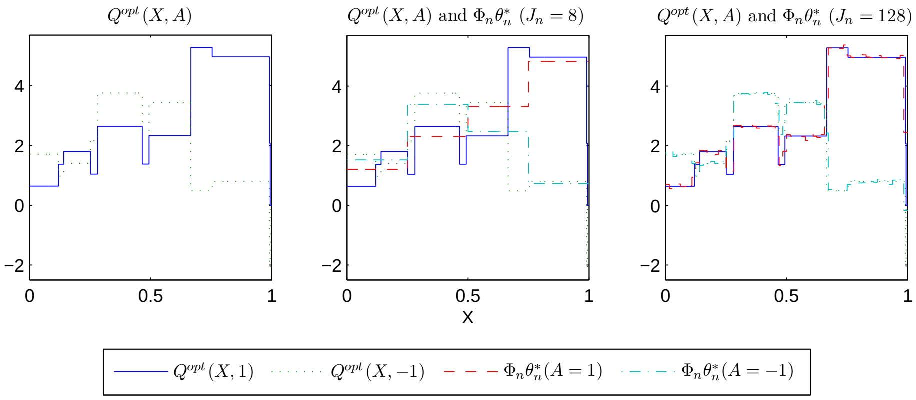

Figure 3.1: Plots for: the conditional mean function $Q^{opt}(X, A)$ (left), $Q^{opt}(X, A)$ and the associated best wavelet fit when $J_n = 8$ (middle), and $Q^{opt}(X, A)$ and the associated best wavelet fit when $J_n = 128$ (right) (example 4).

data sets are generated for each sample size. We apply the l_1-PLS based method (denoted by l_1-PLS) and the method based on separate prognosis prediction for each treatment (denoted by PP) on each data set. The Value of each estimated decision rule is evaluated via Monte Carlo using a test set of size $10,000$.

Simulation results are presented in Figure 3.2 and Table 3.1. When the approximation model for the interaction term in Q^{opt} is of high quality, both methods produce decision rules with similar Value. However, when the approximation model for the interaction term in Q^{opt} is poor (example 4 for small J_n and examples 3, 7, 8 and 9), the l_1-PLS method generally produces higher Value than PP (see examples 3, 8 and 9). Note that in example 3 the Value of the decision rule produced by l_1-PLS method has larger median absolute deviation (MAD) than that from PP when the sample size is small. One possible reason is that the Value estimator used in cross-validation is a non-smooth function of the data. Nonetheless, the l_1-PLS method is still preferred after taking the variation into account (l_1-PLS produces treatment rules with higher Value than PP 59.4%, 64.4%, 70.2% and 79.4% of the times when $n = 40, 64, 101$ and 160). Furthermore, in general the l_1-PLS method uses much fewer variables for treatment assignment than PP. This is expected since many variables may be useful in predicting the primary outcome under each treatment but only a few of them are helpful in selecting the best treatment.

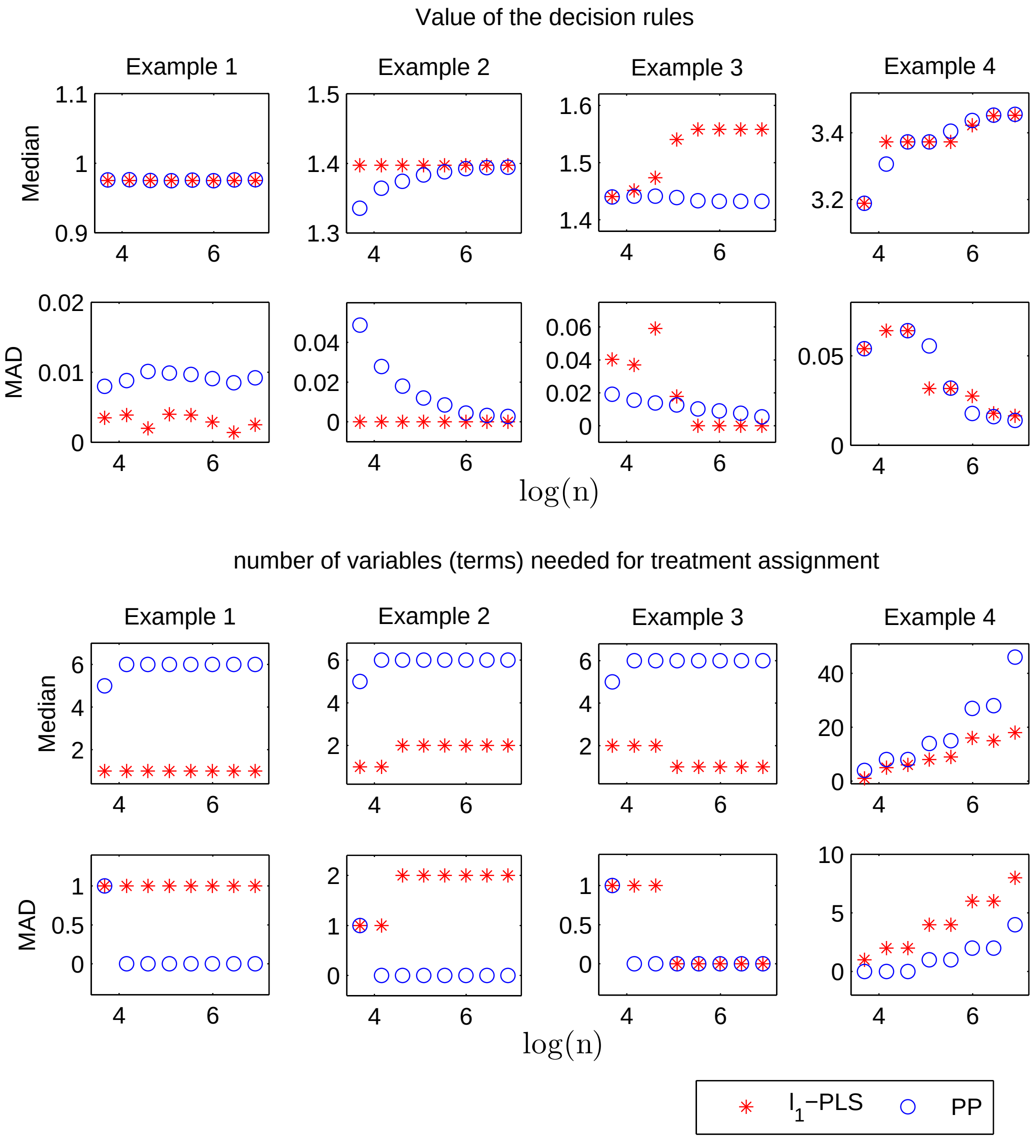

Figure 3.2: Comparison of the l_1-PLS based method with separate prognosis prediction for each treatment (examples 1 - 4): Plots for medians and median absolute deviations (MAD) of the Value of the estimated decision rules (top panels) and the number of variables (terms) needed for treatment assignment (including the main treatment effect term, bottom panels) over 500 samples versus sample size on the log scale ($n = 40, 64, 101, 160, 253, 401, 633, 1000$. The corresponding numbers of basis functions in example 4 are $J_n = 8, 16, 16, 32, 32, 64, 64, 128$).

Method	Median and MAD (in the parentheses) for	
	Value of the decision rules	# of variables needed for treatment assignment
Example 5		
l_1-PLS	28.8515 (0.0947)	4 (4)
PP	28.8300 (0.1659)	49 (1)
Example 6		
l_1-PLS	30.0865 (0.0239)	10.5 (6.5)
PP(CV)	30.0011 (0.0404)	50 (1)
Example 7		
l_1-PLS	30.0798 (0.0007)	6 (6)
PP	29.8959 (0.0533)	49 (1)
Example 8		
l_1-PLS	32.1382 (0.3611)	4 (2)
PP	31.3168 (0.2512)	42 (3)
Example 9		
l_1-PLS	30.1579 (0.0064)	7 (5)
PP	29.9262 (0.0592)	50 (1)

Table 3.1: Comparison of the l_1-PLS based method with separate prognosis prediction based method: Medians and MAD (in the parentheses) of the Value of each estimated decision rule (left) and the number of variables needed for treatment assignment (including the main treatment effect term, right) based on 500 replications (examples 5 - 9) ($n = 500$).

3.2.2 Nefazodone-CBASP trial example

The Nefazodone-CBASP trial was conducted to compare the efficacy of several alternate treatments for patients with chronic depression. The study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. Various assessments were taken throughout the study, among which the score on the 24-item Hamilton Rating Scale for Depression (HRSD) was the primary outcome. Low HRSD scores are desirable. See Keller et al. (2000) for more details of the study design and the primary analysis.

In the data analysis, we use a subset of the Nefazodone-CBASP data consisting of 656 patients for whom the outcome HRSD score was observed. Pairwise comparisons show that the combination treatment resulted in significantly lower HRSD scores than either of the single treatments, and there was no overall difference between the single treatments.

We use l_1-PLS to develop an individualized treatment rule. In the analysis HRSD score is reverse coded so that higher is better. There are 50 pretreatment variables $X = (X_1, \ldots, X_{50})$. Treatments are coded using contrast coding of dummy variables $A = (A_1, A_2)$, where $A_1 = 2$ if the combination treatment is assigned and -1 otherwise and $A_2 = 1$ if CBASP is assigned, -1 if nefazodone and 0 otherwise. The vector of basis functions, $\Phi(X, A)$, is of the form $(1, X, A_1, XA_1, A_2, XA_2)$. So the number of basis functions is $J = 153$. As a contrast, we also consider treatment assignment via separate prognosis prediction for each treatment (PP). The vector of basis functions used in PP is $(1, X)$ for each treatment group. Neither the intercept term nor the main treatment effect terms in these methods will be penalized (see Section 3.4.3 for the modification of the weights $\hat{\sigma}_j$'s used in (3.1)).

The individualized treatment rule given by the l_1-PLS method recommends the combination treatment to all (so none of the pretreatment variables enter the rule). On the other hand, the PP method produces a treatment rule that uses 29 variables. If the individualized treatment rule produced by PP were used to assign treatment for the 656 patients in the trial, it would recommend the combination treatment for 614 patients and nefazodone for the other 42 patients.

We have found that, in general, if one treatment is overwhelmingly better than the other treatments, the treatment rules produced by both methods are likely to recommend the same treatments for most patients; however as the difference in treatments decreases these two treatment rules will recommend different treatments for more and more patients. To see this, we consider the following 5 examples. The first example uses the original data, in which the combination treatment is overwhelmingly better. Cohen's f effect size index is around 0.25 (Cohen's f index is the square root of the between-group variance divided by the square root of the within-group variance; 0.25 is considered as a medium effect size; Cohen 1988). In each of the examples 2 to 5, we subtract a constant from the reverse coded HRSD scores for the combination treatment group so that the Cohen's f index is around 0.2, 0.15, 0.1 and 0.05, respectively. Both methods are used on each example. The l_1-PLS method produces treatment rules that use $0, 0, 0, 30$ and 23 variables and the PP method produces treatment rules that always use 29 variables for treatment assignment for examples 1 to 5, respectively. If the treatment rules produced by the two methods were used to assign treatment for the 656 patients in the trial, they would recommend different treatments on 42, 81, 132, 264 and 331 patients for examples 1 to 5, respectively.

3.3 Discussion

Our goal is to construct an individualized treatment rule that can be employed to benefit future patients. In this chapter, we considered l_1-PLS based estimation method and provided a finite sample upper bound for $V(d^{opt}) - V(\hat{d}_n)$, the excess Value of the estimated treatment rule.

The use of an l_1 penalty allows us to consider a large model for the conditional mean function Q^{opt} yet permits a sparse estimated individualized treatment rule. In fact, many other penalization methods such as SCAD (Fan and Li, 2001) and l_1 penalty with adaptive weights (adaptive Lasso; Zou 2006) also have this property. We choose the non-adaptive l_1 penalty to represent these methods. Interested readers may justify other PLS methods using similar proof techniques.

An important issue is how to select the sequence of basis functions so that the mean square error $\min_{\boldsymbol{\theta}\in\Theta_n} E(\Phi^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2$ converges to 0 as $n \to \infty$, where T^{opt} is the term in Q^{opt} containing A and $\Phi^{(2)}\boldsymbol{\theta}^{(2)}$ approximates T^{opt}. Although our theoretical result does not require this condition, if this condition does hold then our result implies that $V(\hat{d}_n)$ converges to the optimal Value. We refer to Barron et al. (1999) for general results on the construction of approximation spaces that guarantee this condition. In addition, note that the obtained high probability bound (3.8) cannot be used to construct a prediction interval for $V(d^{opt}) - V(\hat{d}_n)$ due to the unknown quantities in the upper bound. How to develop a high probability computable upper bound to assess the quality of $\hat{d}_n$ is an open question.

We used cross validation with Value maximization to select the tuning parameter involved in the l_1-PLS method. As compared to treatment assignment via separate prognosis prediction, this method yields individualized treatment rules that use less variables. However, since only the Value is used to select the tuning parameter, this method may produce a complex individualized treatment rule for which the Value is only slightly higher than that of a much simpler treatment rule. In that case, the simple treatment rule may be preferred due to the interpretability and cost of collecting the variables. Investigation of a tuning parameter selection criterion that trades off the Value with the number of variables in an individualized treatment rule is needed.

3.4 Appendices

3.4.1 Generalization of Theorem III.3

In this section, we present a generalization of Theorem III.3 where J may depend on n and the sparsity of any $\boldsymbol{\theta} \in \mathbb{R}^J$ is measured by the number of "large" components in $\boldsymbol{\theta}$ as described in Zhang and Huang (2008). In this case, J, Φ and the prediction error minimizer $\boldsymbol{\theta}^*$ from (3.3) are denoted as J_n, Φ_n and $\boldsymbol{\theta}^*_n$, respectively. We allow some constants used in Θ^o_n defined in (3.4), Θ_n defined in (3.5) and Assumptions 1-3 used in Theorem III.3 to depend on n. Those sets and assumptions are re-stated below.

Let N_M denote the cardinality of any index set $M \subseteq \{1, \ldots, J_n\}$. For any $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$ and constant $\rho \geq 0$, define

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

Then $M_{\rho\lambda_n}(\boldsymbol{\theta})$ is the smallest index set that contains only "large" components in $\boldsymbol{\theta}$. It is easy to see that when $\rho = 0$, $M_0(\boldsymbol{\theta})$ is the index set of nonzero components in $\boldsymbol{\theta}$. Moreover, $M_{\rho\lambda_n}(\boldsymbol{\theta})$ is an empty set if and only if $\boldsymbol{\theta} = \mathbf{0}$.

When $E(\Phi_n^{(2)}(X, A)^T | X) = \mathbf{0}$ a.s. ($\Phi_n^{(2)}$ is defined in Section 3.1.1), we define

$$M^{(1)}_{\rho\lambda_n}(\boldsymbol{\theta}) \in \arg \min_{\{M \subseteq \{1,\ldots,J_n^{(1)}\}: \sum_{j \in \{1,\ldots,J_n^{(1)}\} \setminus M} \sigma_j |\theta_j| \leq \rho N_M \lambda_n\}} N_M.$$

$$\text{and} \quad M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta}) \in \arg \min_{\{M \subseteq \{J_n^{(1)}+1,\ldots,J_n\}: \sum_{j \in \{J_n^{(1)}+1,\ldots,J_n\} \setminus M} \sigma_j |\theta_j| \leq \rho N_M \lambda_n\}} N_M.$$

Assumption III.1. *The error terms $\varepsilon_i, i = 1, \ldots, n$, are independently distributed with $E(\varepsilon_i | X_i, A_i) = 0$ and $E[|\varepsilon_i|^l] \leq \frac{l!}{2} c^{l-2} \sigma^2$ for some $c, \sigma^2 > 0$ for all $l \geq 2$.*

Assumption III.2. *For all $n \geq 1$,*

(a) there exists an $0 < U_n < \infty$ such that $\max_{j=1,\ldots,J_n} \|\phi_j\|_\infty / \sigma_j \leq U_n$, where $\sigma_j := (E\phi_j^2)^{1/2}$.

(b) there exists an $0 \leq \eta_{2,n} < \infty$, such that $\sup_{\boldsymbol{\theta} \in [\boldsymbol{\theta}_n^*]} \|Q^{opt} - \Phi_n \boldsymbol{\theta}\|_\infty \leq \eta_{2,n}$.

For any $0 \leq \gamma < 1/2$, positive number $\eta_{1,n}$ (which may depend on n) and tuning parameter λ_n, define

$$\Theta_n^o = \Big\{ \boldsymbol{\theta} \in \mathbb{R}^{J_n} : \exists\ \boldsymbol{\theta}^o \in [\boldsymbol{\theta}_n^*] \text{ s.t. } \|\Phi_n(\boldsymbol{\theta} - \boldsymbol{\theta}^o)\|_\infty \leq \eta_{1,n}$$
$$\text{and } \max_{j=1,\ldots,J_n} \Big| E\Big[\Phi_n(\boldsymbol{\theta} - \boldsymbol{\theta}^o) \frac{\phi_j}{\sigma_j}\Big] \Big| \leq \gamma \lambda_n \Big\}.$$

Assumption III.3. *For any $n \geq 1$, there exists a $\beta_n > 0$ such that*

$$E[\Phi_n(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})} \geq \beta_n \left[\left(\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j - \theta_j| \right)^2 - \rho^2 N^2_{M_{\rho\lambda_n}(\boldsymbol{\theta})} \lambda_n^2 \right] \tag{3.14}$$

for all $\boldsymbol{\theta} \in \Theta_n^o \setminus \{\mathbf{0}\}$, $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n}$ and $\sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j| \leq \frac{2\gamma+5}{1-2\gamma} (\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} |\tilde{\theta}_j - \theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})} \lambda_n)$.

When $E(\Phi_n^{(2)}(X, A)^T | X) = \mathbf{0}$ a.s., we consider the following assumption instead of Assumption III.3.

Assumption III.4. *For any $n \geq 1$, there exists a $\beta_n > 0$ such that*

$$\begin{aligned} & E[\Phi_n^{(1)}(\tilde{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}^{(1)})]^2 N_{M^{(1)}_{\rho\lambda_n}(\boldsymbol{\theta})} \geq \beta_n \left[\left(\sum_{j \in M^{(1)}_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j - \theta_j| \right)^2 - \rho^2 N^2_{M^{(1)}_{\rho\lambda_n}(\boldsymbol{\theta})} \lambda_n^2 \right] \\ \textit{and} \quad & E[\Phi_n^{(2)}(\tilde{\boldsymbol{\theta}}^{(2)} - \boldsymbol{\theta}^{(2)})]^2 N_{M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})} \geq \beta_n \left[\left(\sum_{j \in M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j - \theta_j| \right)^2 - \rho^2 N^2_{M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})} \lambda_n^2 \right] \end{aligned}$$

for all $\boldsymbol{\theta} \in \Theta_n^o \setminus \{\mathbf{0}\}$, $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n}$ and $\sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j |\tilde{\theta}_j| \leq \frac{2\gamma+5}{1-2\gamma} (\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})} |\tilde{\theta}_j - \theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})} \lambda_n)$.

Without loss of generality, we can assume $\rho\beta_n \leq 1$.

For any $t > 0$, define

$$\Theta_n = \left\{ \boldsymbol{\theta} \in \Theta_n^o : N_{M_{\rho\lambda_n}(\boldsymbol{\theta})} \leq \frac{(1-2\gamma)^2 \beta_n}{120} \left(\sqrt{\frac{1}{9} + \frac{n}{2U_n^2[\log(3J_n(J_n+1)) + t]}} - \frac{1}{3} \right) \right\}. \tag{3.15}$$

Note that we allow $U_n, \eta_{1,n}, \eta_{2,n}$ and β_n^{-1} to increase as n increases. However, if those quantities are small, the upper bound in (3.18) will be tighter.

Theorem III.4. *Suppose Assumptions III.1 and III.2 hold. For any given $0 \leq \gamma < 1/2$, $\eta_{1,n} \geq 0$, $\rho \geq 0$ and $t > 0$, let $\hat{\boldsymbol{\theta}}_n$ be the l_1-PLS estimator defined in (3.1) with* *tuning parameter*

$$\lambda_n \geq \frac{8\max\{3c, 2(\eta_{1,n}+\eta_{2,n})\}U_n(\log 6J_n + t)}{(1-2\gamma)n} + \frac{12\max\{\sigma, (\eta_{1,n}+\eta_{2,n})\}}{(1-2\gamma)}\sqrt{\frac{2(\log 6J_n + t)}{n}}. \tag{3.16}$$

Suppose Assumption III.3 holds with $\rho\beta_n \leq 1$*. Let* Θ_n *be the set defined in (3.15) and assume* Θ_n *is non-empty. If*

$$\frac{\log 2J_n}{n} \leq \frac{2(1-2\gamma)^2}{27U_n^2 - 10\gamma - 22}, \tag{3.17}$$

then with probability at least $1 - \exp(-k_n' n) - \exp(-t)$*, we have*

$$L(\Phi_n \hat{\boldsymbol{\theta}}_n) \leq \min_{\boldsymbol{\theta}\in\Theta_n}\left[L(\Phi_n\boldsymbol{\theta}) + K_n \frac{N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}{\beta_n}\lambda_n^2\right], \tag{3.18}$$

where $k_n' = 13(1-2\gamma)^2/[6(27U_n^2 - 10\gamma - 22)]$ *and* $K_n = [40\gamma(12\beta_n\rho + 2\gamma + 5)]/[(1-2\gamma)(2\gamma+19)] + 130(12\beta_n\rho + 2\gamma + 5)^2/[9(2\gamma+19)^2]$.

Furthermore, suppose $E(\Phi_n^{(2)}(X,A)^T|X) = \mathbf{0}$ *a.s. Let* $T^{opt} := Q^{opt} - E(Q^{opt}|X)$*. Instead of Assumption III.3, suppose Assumption III.4 holds with* $\rho\beta_n \leq 1$*. Then with probability at least* $1 - \exp(-k_n' n) - \exp(-t)$*, we have*

$$E(\Phi_n^{(2)}\hat{\boldsymbol{\theta}}_n^{(2)} - T^{opt})^2 \leq \min_{\boldsymbol{\theta}\in\Theta_n}\left[E(\Phi_n^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2 + K_n' \frac{N_{M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})}}{\beta_n}\lambda_n^2\right].$$

where $K_n' = 20(12\beta_n\rho + 2\gamma + 5)\{\gamma/[(1-2\gamma)(7-6\beta_n\rho)] + [3(1-2\gamma)\beta_n\rho + 10(2\gamma+5)]/[9(2\gamma+19)^2]\}$.

Remark

Note that K_n is upper bounded by a constant under the assumption $\beta_n\rho \leq 1$. In the asymptotic setting when $n \to \infty$ and $J_n \to \infty$, (3.18) implies that with probability tending to 1, $L(\Phi_n\hat{\boldsymbol{\theta}}_n) - L(\Phi_n\boldsymbol{\theta}_n^*) \to 0$ if (i) $N_{M_{\rho\lambda_n}(\boldsymbol{\theta}_n^*)}\lambda_n^2/\beta_n = o(1)$, (ii) $U_n^2 \log J_n/n \leq k_1$ and $N_{M_{\rho\lambda_n}(\boldsymbol{\theta}_n^*)} \leq k_2\beta_n\sqrt{n/(U_n^2\log J_n)}$ for some sufficiently small positive constants k_1 and k_2, and (iii) $\lambda_n \geq k_3\sqrt{\log J_n/n}$ when $\eta_{1,n} + \eta_{2,n} = O(1)$ or $\lambda_n \geq k_3(\eta_{1,n} + \eta_{2,n})\sqrt{\log J_n/n}$ when $(\eta_{1,n} + \eta_{2,n})^{-1} = o(1)$ for a sufficiently large constant k_3 (take $t = \log J_n$).

Proof. For any $\boldsymbol{\theta} \in \Theta_n$, define the events

$$\Omega_1 = \bigcap_{j=1}^{J_n} \Big\{ \frac{2(1+\gamma)}{3}\sigma_j \leq \hat{\sigma}_j \leq \frac{2(2-\gamma)}{3}\sigma_j \Big\} \text{ (where } \hat{\sigma}_j = (E_n\phi_j^2)^{1/2}),$$
$$\Omega_2(\boldsymbol{\theta}) = \left\{ \max_{j,k=1,\ldots,J_n} \Big|(E - E_n)\Big(\frac{\phi_j\phi_k}{\sigma_j\sigma_k}\Big)\Big| \leq \frac{(1-2\gamma)^2\beta_n}{120 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}} \right\},$$
$$\Omega_3(\boldsymbol{\theta}) = \left\{ \max_{j=1,\ldots,J_n} \Big|E_n\Big[(Y - \Phi_n\boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big| \leq \frac{4\gamma+1}{6}\lambda_n \right\}.$$

Then there exists a $\boldsymbol{\theta}^o \in [\boldsymbol{\theta}_n^*]$ such that

$$\begin{aligned} L(\Phi_n\hat{\boldsymbol{\theta}}_n) =& L(\Phi_n\boldsymbol{\theta}) + 2E[(\Phi_n\boldsymbol{\theta}^o - \Phi_n\boldsymbol{\theta})\Phi_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)] + E[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 \\ \leq& L(\Phi_n\boldsymbol{\theta}) + 2\max_{j=1,\ldots,J_n}\Big|E\Big[\Phi_n(\boldsymbol{\theta}^o - \boldsymbol{\theta})\frac{\phi_j}{\sigma_j}\Big]\Big|\left(\sum_{j=1}^{J_n}\sigma_j|\hat{\theta}_{n,j} - \theta_j|\right) + E[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 \\ \leq& L(\Phi_n\boldsymbol{\theta}) + 2\gamma\lambda_n\left(\sum_{j=1}^{J_n}\sigma_j|\hat{\theta}_{n,j} - \theta_j|\right) + E[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2, \end{aligned}$$

where the first equality follows from the fact that $E[(Y - \Phi_n\boldsymbol{\theta}^o)\phi_j] = 0$ for any $\boldsymbol{\theta}^o \in [\boldsymbol{\theta}_n^*]$ for $j = 1, \ldots, J_n$ and the last inequality follows from the definition of Θ_n^o.

Based on Lemma III.1 below, we have that on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$,

$$L(\Phi_n\hat{\boldsymbol{\theta}}_n) \leq L(\Phi_n\boldsymbol{\theta}) + K_n\frac{N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}{\beta_n}\lambda_n^2.$$

Similarly, by Lemma III.2, we have that on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$,

$$E(\Phi_n^{(2)}\hat{\boldsymbol{\theta}}_n^{(2)} - T^{opt})^2$$

$$\begin{aligned}&\leq E(\Phi_n^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2 + 2\gamma\lambda_n\left(\sum_{j=J_n^{(1)}+1}^{J_n} \sigma_j|\hat{\theta}_{n,j} - \theta_j|\right) + E[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}^{(2)})]^2\\ &\leq E(\Phi_n^{(2)}\boldsymbol{\theta}^{(2)} - T^{opt})^2 + K_n' \frac{N_{M_{\rho\lambda_n}^{(2)}(\boldsymbol{\theta})}}{\beta_n}\lambda_n^2.\end{aligned}$$

The conclusion of the theorem follows from the union probability bounds of the events Ω_1, $\Omega_2(\boldsymbol{\theta})$ and $\Omega_3(\boldsymbol{\theta})$ provided in Lemmas III.3, III.4 and III.5. □

Lemma III.1. *Suppose Assumption III.3 holds with $\rho\beta_n \leq 1$. Then for any $\boldsymbol{\theta} \in \Theta_n$, on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$, we have*

$$\sum_{j=1}^{J_n} \sigma_j|\hat{\theta}_{n,j} - \theta_j| \leq \frac{20(12\beta_n\rho + 2\gamma + 5)}{(1-2\gamma)(19+2\gamma)\beta_n} N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n \tag{3.19}$$

$$\textit{and} \quad E[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 \leq \frac{130(12\rho\beta_n + 2\gamma + 5)^2}{9(19+2\gamma)^2\beta_n} N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n^2 \tag{3.20}$$

Remark

This lemma implies that $\hat{\boldsymbol{\theta}}_n$ is close to each $\boldsymbol{\theta} \in \Theta_n$ on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$. The intuition is as follows. Since $\hat{\boldsymbol{\theta}}_n$ minimizes (3.1), the first order conditions imply that $\max_j |E_n(Y - \Phi_n\hat{\boldsymbol{\theta}}_n)\phi_j/\hat{\sigma}_j| \leq \lambda_n/2$. Similar property holds for $\boldsymbol{\theta}$ on the event $\Omega_1 \cap \Omega_3(\boldsymbol{\theta})$. Assumption III.3 together with event $\Omega_2(\boldsymbol{\theta})$ ensures that there is no collinearity in the $n \times J_n$ design matrix $\big(\Phi_n(X_i, A_i)\big)_{i=1}^n$. These two aspects guarantee the closeness of $\hat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}$.

Proof. First note that $\hat{\boldsymbol{\theta}}_n$ (defined in (3.1)) satisfies the following first order condition:

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

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

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

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

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

$$\begin{aligned} 0 &= 2E_n[(Y - \Phi_n \hat{\boldsymbol{\theta}}_n)\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})] + \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j \mathrm{sgn}(\hat{\theta}_{n,j})\theta_j - \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j |\hat{\theta}_{n,j}| \\ &\leq 2E_n[(Y - \Phi_n \hat{\boldsymbol{\theta}}_n)\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})] + \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j |\theta_j| - \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j |\hat{\theta}_{n,j}|. \end{aligned} \tag{3.21}$$

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

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

This implies $\hat{\boldsymbol{\theta}}_n = \mathbf{0}$. Thus (3.19) and 3.20) 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 (3.21), on the event $\Omega_1 \cap \Omega_3(\boldsymbol{\theta})$, we have

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

$$
\begin{aligned}
&+ \frac{2(2-\gamma)}{3}\lambda_n\left(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\right)\\
&- \frac{2(1+\gamma)}{3}\lambda_n\sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}| - 2E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2\\
=\ & \frac{2\gamma+5}{3}\lambda_n\left(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\right)\\
&- \frac{1-2\gamma}{3}\lambda_n\sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}| - 2E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2.
\end{aligned}
$$

This implies

$$
\begin{aligned}
\sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}| &\le \frac{2\gamma+5}{1-2\gamma}\left(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\right)\\
\text{and } E_n[\Phi_n(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})]^2 &\le \frac{2\gamma+5}{6}\lambda_n\left(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\hat{\theta}_{n,j}-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\right). \quad (3.22)
\end{aligned}
$$

Define the sets

$$
\begin{aligned}
\Theta_1(\boldsymbol{\theta}) &= \Bigg\{\tilde{\boldsymbol{\theta}}\in\mathbb{R}^{J_n} : \sum_{j\in\{1,\ldots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j|\\
&\qquad \le \frac{2\gamma+5}{1-2\gamma}\left(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\right)\Bigg\},\\
\Theta_2(\boldsymbol{\theta}) &= \left\{\tilde{\boldsymbol{\theta}}\in\mathbb{R}^{J_n} : \sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| > \frac{[10(2\gamma+5)+3(21-2\gamma)\beta_n\rho]N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n}{3(19+2\gamma)\beta_n}\right\},\\
\Theta_3(\boldsymbol{\theta}) &= \left\{\tilde{\boldsymbol{\theta}}\in\mathbb{R}^{J_n} : \sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n > \frac{10N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n}{3\beta_n}\right\}.
\end{aligned}
$$

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 $\Omega_1\cap\Omega_2(\boldsymbol{\theta})\cap\Omega_3(\boldsymbol{\theta})$,

$$
\sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})}\left\{2E_n[(Y-\Phi_n\tilde{\boldsymbol{\theta}})\Phi_n(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})] + \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\theta_j| - \lambda_n\sum_{j=1}^{J_n}\hat{\sigma}_j|\tilde{\theta}_j|\right\}
$$

$$
\begin{aligned}
&\le \sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Bigg\{ \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 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big) \\
&\qquad - 2E[\Phi_n(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})]^2 + 2\max_{j=1,\dots,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 \\
&\qquad - \frac{1-2\gamma}{3}\lambda_n \sum_{j\in\{1,\dots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j|\Bigg\} \\
&\le \sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Bigg\{ \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 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big) \\
&\qquad + 2\beta_n\rho^2 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n^2 - \frac{2\beta_n}{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 \\
&\qquad + \frac{(1-2\gamma)^2\beta_n}{60 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big)\Big(\sum_{j=1}^{J_n}\sigma_j|\tilde{\theta}_j-\theta_j|\Big) \\
&\qquad + \frac{1-2\gamma}{3}\Big[\frac{(1-2\gamma)\beta_n}{20 N_{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,\dots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j|\Big)\Bigg\} \\
&\le \sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})} \Bigg\{ \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 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big) \\
&\qquad + 2\beta_n\rho^2 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n^2 - \frac{2\beta_n}{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 \\
&\qquad + \frac{(1-2\gamma)\beta_n}{10 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big)^2 \\
&\qquad + \frac{1-2\gamma}{3}\Big[\frac{3\beta_n}{10 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big) - \lambda_n\Big] \\
&\qquad\quad \times\Big(\sum_{j\in\{1,\dots,J_n\}\setminus M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j|\Big)\Bigg\} \\
&\le \sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})\cap\Theta_3(\boldsymbol{\theta})^C} \Bigg\{\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big) \\
&\qquad\quad \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 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j|\Big)\Bigg\} \\
&\quad + \sup_{\tilde{\boldsymbol{\theta}}\in\Theta_1(\boldsymbol{\theta})\cap\Theta_2(\boldsymbol{\theta})\cap\Theta_3(\boldsymbol{\theta})} \Bigg\{\Big(\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})} \sigma_j|\tilde{\theta}_j-\theta_j| + \rho N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n\Big)
\end{aligned}
$$

$$\times \Big(\frac{13}{5}\beta_n\rho\lambda_n - \frac{7\beta_n}{5N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}}\sum_{j\in M_{\rho\lambda_n}(\boldsymbol{\theta})}\sigma_j|\tilde{\theta}_j - \theta_j|\Big)\Bigg\}$$

$<$0,

where the second inequality follows from Assumption III.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 (3.21), 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 (3.19).

Following (3.22) and the fact that $\hat{\boldsymbol{\theta}}_n \in \Theta_2(\boldsymbol{\theta})^C$, we have

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

on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$. Suppose (3.20) does not hold, i.e. $E[\Phi_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})]^2 > \frac{130(12\rho\beta_n+2\gamma+5)^2}{9(19+2\gamma)^2\beta_n} N_{M_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n^2$. Then

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

where the first inequality follows from the definition of $\Omega_2(\boldsymbol{\theta})$ and the second inequality follows from (3.19). This implies

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

which contradicts the condition. Thus (3.20) holds on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$. □

Lemma III.2. *Suppose* $E\big[\Phi_n^{(2)}(X, A)^T|X\big] = \mathbf{0}$ *a.s. and Assumption III.4 holds with*

$\rho\beta_n \leq 1$. *Then for any* $\boldsymbol{\theta} \in \Theta_n$*, on the event* $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$*, we have*

$$\sum_{j=J_n^{(1)}+1}^{J_n} \sigma_j |\hat{\theta}_{n,j} - \theta_j| \leq \frac{10(12\beta_n\rho + 2\gamma + 5)}{(1-2\gamma)(7-6\beta_n\rho)\beta_n} N_{M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n \tag{3.23}$$

and

$$E[\Phi_n^{(2)}(\hat{\boldsymbol{\theta}}_n^{(2)} - \boldsymbol{\theta}^{(2)})]^2 \leq \frac{20(12\rho\beta_n + 2\gamma + 5)[3(1-2\gamma)\beta_n\rho + 10(2\gamma+5)]}{9(2\gamma+19)^2\beta_n} N_{M^{(2)}_{\rho\lambda_n}(\boldsymbol{\theta})}\lambda_n^2. \tag{3.24}$$

Proof. Consider fixed n and fixed $\boldsymbol{\theta} \in \Theta_n$. Since $E(\Phi_n^{(2)}|X) = \mathbf{0}$ 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

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

where the second inequality follows from the definition of $\Omega_2(\boldsymbol{\theta})$ and Lemma III.1 (note that Assumption III.4 implies Assumption III.3).

Next, note that (3.21) holds for $(\boldsymbol{\theta}^{(1)}, \hat{\boldsymbol{\theta}}_n^{(2)})$. Thus on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$, we have

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

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

This implies

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

Using similar argument as that in lemma III.1, we obtain

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

Algebra suffices to show (3.23) and (3.24). □

Lemma III.3. *Suppose Assumption III.2(a) and inequality (3.17) hold. Then*

$$
\mathbb{P}(\Omega_1^C) \le \exp\left(-\frac{13(1-2\gamma)^2 n}{6(27U_n^2-10\gamma-22)}\right).
$$

Proof. For each $j = 1, \dots, J_n$, we apply Lemma VI.8(a) with $O_i = (\phi_j(X_i, A_i)^2/\sigma_j^2 - 1)/(U_n^2 - 1)$ and $t = (7-2\gamma)(1-2\gamma)n/9(U_n^2-1)$. By Assumption III.2(a), we have $O_i \leq 1$ and $\sum_{i=1}^n EO^2 \leq n/(U_n^2 - 1)$. Thus

$$\begin{aligned}\mathbb{P}\Big(\hat{\sigma}_j \geq \frac{2(2-\gamma)}{3}\sigma_j\Big) &\leq \exp\Big(-\frac{(7-2\gamma)^2(1-2\gamma)^2 n}{2(U_n^2-1)[81+3(7-2\gamma)(1-2\gamma)]}\Big)\\ &\leq \exp\Big(-\frac{25(1-2\gamma)^2 n}{6(27U_n^2-10\gamma-22)}\Big).\end{aligned}$$

Similarly, applying Lemma VI.8(a) with $-O_i$, we have

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

Using union bound argument and condition (3.17), we have

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

□

Lemma III.4. *Suppose Assumption III.2(a) holds. Then for any $\boldsymbol{\theta} \in \Theta_n$ and $t > 0$, $\mathbb{P}(\{\Omega_2(\boldsymbol{\theta})\}^C) \leq \exp(-t)/3$.*

Proof. 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 VI.8(a) with $O_i = \pm[\phi_j(X_i, A_i)\phi_k(X_i, A_i)/(\sigma_j\sigma_k) - E(\phi_j\phi_k)/(\sigma_j\sigma_k)]/U_n^2$ and $t = (1-2\gamma)^2\beta_n n/[120 N_{M_{\rho\lambda_n}(\boldsymbol{\theta})} U_n^2]$ and using union bound argument, we obtain

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

where the second inequality follows from the definition of Θ_n in (3.15). □

Lemma III.5. *Suppose Assumptions III.1 and III.2 hold. For any $t > 0$ and $\eta_{1,n} \geq 0$, take λ_n so that it satisfies condition (3.16). Then for any $\boldsymbol{\theta} \in \Theta_n$, we have $\mathbb{P}(\{\Omega_3(\boldsymbol{\theta})\}^C) \leq 2\exp(-t)/3$.*

Proof. 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 $E(Y - \Phi_n\boldsymbol{\theta})^2$, we have $E\big[(Y - \Phi_n\boldsymbol{\theta}^o)\phi_j\big] = 0$ for $j = 1, \ldots, J_n$. Thus

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

This implies

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

By Assumptions III.1 and III.2(a), we have $E(\varepsilon_i\phi_j(X_i, A_i)/\sigma_j) = 0$ and $\sum_{i=1}^n E[(\varepsilon_i\phi_j(X_i, A_i)/\sigma_j)_+^l] \leq \frac{l!}{2}n\sigma^2(cU_n)^{l-2}$ for all integers $l \geq 2$. Applying lemma VI.8(b) yields

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

Similarly, the definition of Θ_n^o together with Assumption III.2 implies that, for any $\boldsymbol{\theta} \in \Theta_n$ and $j = 1, \ldots, J_n$, $\big\|(Q^{opt} - \Phi_n\boldsymbol{\theta})\phi_j/\sigma_j - E\big((Q^{opt} - \Phi_n\boldsymbol{\theta})\phi_j/\sigma_j\big)\big\|_\infty \leq 2(\eta_{n,1} + \eta_{n,2})U_n$ and $E\big[(Q^{opt} - \Phi_n\boldsymbol{\theta})\phi_j/\sigma_j\big]^2 \leq (\eta_{n,1} + \eta_{n,2})^2$. Applying Lemma VI.8(a) yields

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

The result follows from the union bounds argument and condition (3.16). □

3.4.2 Design of simulations in section 3.2.1

In this section, we present the detailed simulation design of the examples used in Section 3.2.1.

In examples 1 - 3, we generate $X = (X_1, \ldots, X_5)$, where $X_1, \ldots, X_5$ are mutually independent and each $X_j, j = 1, \ldots, 5$, is uniformly distributed on $[-1, 1]$. The treatment A is then generated independently of X from $\{-1, 1\}$ with probability 1/2 each. Given X and A, the outcome Y is generated from a normal distribution with mean $Q^{opt}(X, A) = 1 + 2X_1 + X_2 + 0.5X_3 + T^{opt}(X, A)$ and variance 1 (recall that $T^{opt}(X, A) := Q^{opt}(X, A) - E[Q^{opt}(X, A)|X]$). We consider the following three examples for T^{opt}.

1. $T^{opt}(X, A) = 0$ (i.e. there is no treatment effect).

2. $T^{opt}(X, A) = 0.4190(1 - X_1)A$.

3. $T^{opt}(X, A) = 0.4464sign(X_1)(1 - X_1)^2 A$.

We approximate Q^{opt} by $\mathcal{Q} = \{(1, X, A, XA)\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{11}\}$. Thus in example 2 the correct model is contained in the approximation space, while in example 3 the correct model is not in the approximation space.

The effect sizes in examples 2 and 3 are medium according to Cohen's d index. When there are two treatments, the Cohen's d effect size index is defined as the standardized difference in mean outcomes between two treatment groups. Cohen (1988) tentatively defined the effect size as "small" if the Cohen's d index is 0.2, "medium" if the index is 0.5 and "large" if the index is 0.8.

In example 4, we consider a complex Q^{opt}. We generate X from $U[0, 1]$. Treatment A is generated independently of X from $\{-1, 1\}$ with probability 1/2 each. The outcome R is generated from a normal distribution with mean $Q^{opt}(X, A) = \sum_{j=1}^{5} \vartheta_{(0),j} 1_{X < u_{(0),j}} + [\sum_{j=1}^{5} \vartheta_{(1),j} 1_{X < u_{(1),j}}]A$ and variance 1, where $\vartheta_{(s),j}$ and $u_{(s),j}$

($\in[0,1]$) for $s=0,1, j=1,\dots,5$ are parameters specified in (3.25). The effect size is medium.

$$\begin{aligned}&\vartheta_{(0),1}=-0.4260, \vartheta_{(0),2}=2.8856, \vartheta_{(0),3}=-1.6010, \vartheta_{(0),4}=-0.9513, \vartheta_{(0),5}=1.2680;\\&\vartheta_{(1),1}=2.0822, \vartheta_{(1),2}=-0.7318, \vartheta_{(1),3}=0.7559, \vartheta_{(1),4}=0.3185, \vartheta_{(1),5}=-2.9579;\\&u_{(0),1}=0.1408, u_{(0),2}=0.9902, u_{(0),3}=0.2807, u_{(0),4}=0.4929, u_{(0),5}=0.4651;\\&u_{(1),1}=0.9934, u_{(1),2}=0.1191, u_{(1),3}=0.2509, u_{(1),4}=0.7541, u_{(1),5}=0.6660.\end{aligned} \tag{3.25}$$

We approximate Q^{opt} by Haar wavelets

$$\mathcal{Q}=\Big\{\theta_{(0),0}h_0(X)+\sum_{lk}\theta_{(0),lk}h_{lk}(X)+\Big(\theta_{(0),1}h_0(X)+\sum_{lk}\theta_{(1),lk}h_{lk}(X)\Big)A:\theta_{\cdot,\cdot}\in\mathbb{R}\Big\},$$

where $h_0(x)=1_{x\in[0,1]}$ and $h_{lk}(x)=2^{l/2}\big(1_{2^lx\in[k+1/2,k+1)}-1_{2^lx\in[k,k+1/2)}\big)$ for $l=0,\dots,\bar{l}_n$. We choose $\bar{l}_n=\lfloor 3\log_2 n/4\rfloor-2$. For a given l and sample $(X_i,A_i,R_i)_{i=1}^n$, k takes integer values from $\lfloor 2^l\min_i X_i\rfloor$ to $\lceil 2^l\max_i X_i\rceil-1$. Then $J_n=2^{\lfloor 3\log_2 n/4\rfloor}\le n^{3/4}$.

Examples 5-9 are based on data from the Nefazodone-CBASP trial (Keller et al., 2000). In the simulation study, we consider 50 pretreatment variables collected from the trial. 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 Y, the outcome 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 outcome R is generated from a normal distribution with mean $Q^{opt}(X,A)=(1,X)\boldsymbol{\vartheta}^{(1)}+T^{opt}(X,A)$ and variance 9. We consider 5 examples for T^{opt}. There is no treatment effect in example 5. The covariates and parameters involved in examples 6 - 9 produce a medium effect size.

5. $T^{opt}(X, A) = 0$.

6. $T^{opt}(X, A) = (1, \widetilde{X})\boldsymbol{\vartheta}^{(2)}A$, where $\widetilde{X} = (X_{38}, X_{27}, X_{22}, X_{21}, X_6)$ and $\boldsymbol{\theta}^{(2)} = (-1.2223,\ 0.6141, -0.7756, -0.0079, 0.4163, -0.5676)^T$. Note that the analysis model contains the the correct model for T^{opt}.

7. $T^{opt}(X, A) = |(1, \widetilde{X})\boldsymbol{\vartheta}^{(2)}|A$, where $\widetilde{X} = (X_{40}, X_8, X_{46}, X_9, X_{29})$ and $\boldsymbol{\vartheta}^{(2)} = (-0.8745,\ 0.3439,\ -0.2885,\ -0.4241,\ 0.1214,\ 1.0515)^T$. In this case, treatment 1 is always better than -1.

8. $T^{opt}(X, A) = sign((1, \widetilde{X}_{sub})\boldsymbol{\vartheta}^{(2),2})|(1, \widetilde{X})\boldsymbol{\vartheta}^{(2),2}|A$, where $\widetilde{X} = (X_{44},\ X_{17},\ X_{31},\ X_{35},\ X_{16})$, $\widetilde{X}_{sub}$ contains the first 3 covariates in $\widetilde{X}$, $\boldsymbol{\vartheta}^{(2),1} = (-0.8410,\ 0.7471,\ 0.1411,\ 0.2981)^T$ and $\boldsymbol{\vartheta}^{(2),2} = (-3.1364,\ 0.7930,\ -5.2663,\ -1.7865,\ -0.2682,\ 2.3239)^T$. Note that the analysis model does not contain the correct model for T^{opt}.

9. Same as example 8, but with a different set of covariates and parameters. $\widetilde{X} = (X_{27}, X_{30}, X_{12}, X_{50}, X_{32})$, $\widetilde{X}_{sub}$ contains the first 3 covariates in $\widetilde{X}$, $\boldsymbol{\vartheta}^{(2),1} = (-1.7428, -0.0478, 1.6312, -0.1969)^T$ and $\boldsymbol{\vartheta}^{(2),2} = (-0.3859,\ 0.5457,\ 0.7019,\ 0.6935,\ 1.0135,\ -1.1039)^T$.

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

3.4.3 Some modifications of the l_1-PLS estimator

As demonstrated in van de Geer (2008), sometimes it is natural not to penalize a subset of coefficients (e.g. coefficients corresponding to the constant term and/or to variables 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.

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

where $\hat{\sigma}_j = (E_n \phi_j^2)^{1/2}$. It is easy to see that with this modification, an analog of inequality (3.18) 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 3.2.2). 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 $\hat{\sigma}_j := [E_n \phi_j^2 - (E_n \phi_j)^2]^{1/2}$ (so $\hat{\sigma}_1 = 0$). In this case, each $E\phi_j$ is estimated by $E_n \phi_j$. van de Geer (2008) 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 (3.18) 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 $\big(\sum_{a \in \mathcal{A}} var(\phi_j(X, A)|A = a) E 1_{A=a}\big)^{1/2}$ (i.e. pooled standard deviation).

   For example, suppose $Q^{opt}(X, a)$ 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 $\hat{\sigma}_j := \big(\sum_{a\in\mathcal{A}} v\hat{a}r(\phi_j(X, A)|A = a)E_n 1_{A=a}\big)^{1/2}$, where $v\hat{a}r(\phi_j(X, A)|A = a)$ 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 $E_n(Y - \Phi_n\boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^{J_n} \hat{\sigma}_j|\theta_j|$ is equivalent to choosing $\theta_j, j \in \{1, \dots, J_n\} \setminus \mathcal{S}$, to minimize $E_n(Y' - \sum_{j\in\{1,\dots,J_n\}\setminus\mathcal{S}} \theta_j\phi'_j)^2 + \lambda_n \sum_{j\in\{1,\dots,J_n\}\setminus\mathcal{S}} \hat{\sigma}_j|\theta_j|$ and setting $\theta_j, j \in \mathcal{S}$ to be some appropriate quantities, where $R' = R - \sum_{a\in\mathcal{A}}(E_n1_{A=a}R)1_{A=a}/E_n1_{A=a}$ (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 (3.18), 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 (3.18) when $\phi_1 \equiv 1$ and θ_1 is not penalized.

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

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

$$\begin{aligned}\Theta_n^{o\prime} = \Big\{&\boldsymbol{\theta} \in \mathbb{R}^{J_n} : \exists\ \boldsymbol{\theta}^o \in [\boldsymbol{\theta}_n^*] \text{ s.t. } \|\Phi_n(\boldsymbol{\theta} - \boldsymbol{\theta}^o)\|_\infty \le \eta_{1,n} \\ &\text{and } \max\Big\{|\theta_1 - \theta_1^o|, \max_{j\in\{2,\dots,J_n\}} \Big|E\Big[\Phi_n(\boldsymbol{\theta} - \boldsymbol{\theta}^o)\frac{\phi_j}{\sigma_j}\Big]\Big|\Big\} \le \gamma\lambda_n\Big\}.\end{aligned}$$

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

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

Assumption III.2(a) is modified to

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

Assumption III.3 is modified to

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

$$\begin{aligned} &E[\Phi(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})]^2(N_{M_{\rho\lambda_n}(\boldsymbol{\theta})'} + 1) \\ \geq &\beta_n\Big[\Big(|\tilde{\theta}_1 - \theta_1| + |\sum_{j \in M_{\rho\lambda_n}(\boldsymbol{\theta})'} \sigma_j|\tilde{\theta}_j - \theta_j|\Big)^2 - \rho^2(N_{M_{\rho\lambda_n}(\boldsymbol{\theta})'} + 1)^2\lambda_n^2\Big] \end{aligned} \tag{3.26}$$

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

For any fixed $\boldsymbol{\theta} \in \Theta_n$, define the events

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

Using the same arguments as those in the proof of Theorem A.1, an analog of (3.18) 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 III.4 and III.5. To show Ω_1' occurs with high probability, we define

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

for some ν_1, ν_2, κ_1 and κ_2 to be chosen later. Under similar conditions as those in Lemma III.3, 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.

Note that Assumption A.2(a)' implies $E\phi_j^2 \ge (1+c_0)(E\phi_j)^2$ for $j = 2, \ldots, J_n$ for some $c_0 > 0$. Thus on the event $\Omega_{1,1}' \cap \Omega_{1,2}'$,

$$\begin{aligned}
\hat{\sigma}_j^2 &= E_n\phi_j^2 - (E_n\phi_j)^2 \ge (1-\kappa_1)E\phi_j^2 - \left(|E\phi_j| + \nu_2\sqrt{E\phi_j^2}\right)^2 \\
&\ge (1-\delta_1)^2\sigma_j^2 + (2\delta_1 - \delta_1^2 - \kappa_1)E\phi_j^2 + (1-\delta_1)^2(E\phi_j)^2 - \left(|E\phi_j| + \nu_2\sqrt{E\phi_j^2}\right)^2 \\
&\ge (1-\delta_1)^2\sigma_j^2 + \left[c_0(2\delta_1 - \delta_1^2) - (1+c_0)\kappa_1 - 2\sqrt{1+c_0}\nu_2 - (1+c_0)\nu_2^2\right](E\phi_j)^2 \\
&\ge (1-\delta_1)^2\sigma_j^2
\end{aligned}$$

for $j = 2, \ldots, J_n$ for some small enough ν_2 and κ_1 depending on c_0 and δ_1.

On the other hand, for any $j = 2, \ldots, J_n$ and $\kappa_2 < \delta_2^2 + 2\delta_2$, if $(E\phi_j)^2 \le (\delta_2^2 + 2\delta_2 - \kappa_2)E\phi_j^2/(1+\delta_2)^2$, then

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

Otherwise, for any $0 < \nu_1 \le \sqrt{\delta_2^2 + 2\delta_2 - \kappa_2}/(1+\delta_2)$, we have

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

$$\le (1+\delta_2)^2\sigma_j^2 + \left[2\nu_1\sqrt{\frac{(1+\delta_2)^2}{\delta_2^2+2\delta_2-\kappa_2}} + (1+c_0)\kappa_2 - c_0(\delta_2^2+2\delta_2)\right](E\phi_j)^2$$

$$\le (1+\delta_2)^2\sigma_j^2$$

for some small enough ν_1 and κ_2 depending on c_0 and δ_2.

CHAPTER IV

Model Selection

From the toy example in Section 2.2, we see that by using the quadratic loss minimization based method, we may deviate from the goal of estimating the best individualized treatment rule under consideration if the conditional mean function Q^{opt} is poorly approximated. In fact, it is also easy to verify that asymptotically the treatment rule estimated from a poor model for Q^{opt} may have higher Value than that from a better but still wrong model for Q^{opt} (e.g. in the toy example in Section 2.2, the rule estimated from the constant space $\mathcal{Q} = \{\theta_1 + \theta_2 A : \theta_1, \theta_2 \in \mathbb{R}\}$ would be better than the rule estimated from the linear space). In this chapter, we propose to deal with the deviation using step-wise model selection techniques. We will consider different models for Q^{opt}. For each model, we estimate an individualized treatment rule by minimizing the empirical quadratic loss. Model selection techniques will then be used to select a treatment rule with the highest Value.

Throughout this chapter, we will always assume suprema of empirical processes (i.e. quantities of the form $\sup_{g \in \mathcal{G}}(E_n - E)g$) are measurable. In other words, we assume that the class $\mathcal{G}$ and the distribution P satisfy appropriate (mild) conditions for measurability of this supremum (see Pollard (1984) and Massart (2003) for the conditions).

4.1 Model selection procedure

We still use $\{(X_i, A_i, Y_i)\}_{i=1}^n$ to represent i.i.d. observations on n subjects in a trial. Let $\{\mathcal{Q}_m : m = 1 \ldots, M_n\}$ be a collection of models for Q^{opt}, where the number of models M_n may increase as n increases. For each model $m \in \{1, \ldots, M_n\}$, we estimate Q^{opt} using least squares,

$$\hat{Q}_{n,m} = \arg \min_{Q \in \mathcal{Q}_m} E_n[Y - Q(X, A)]^2.$$

And the estimated individualized treatment rule is

$$\hat{d}_{n,m}(X) \in \arg \max_{a \in \mathcal{A}} \hat{Q}_{n,m}(X, a).$$

Now with the M_n candidate treatment rules $\{\hat{d}_{n,m} : m = 1, \ldots, M_n\}$, we want to select the one that gives the highest Value $V(\hat{d}_{n,m})$. Hence, the oracle selector satisfies

$$m^* \in \arg \max_{m \in \{1, \ldots, M_n\}} V(\hat{d}_{n,m}). \tag{4.1}$$

Note that m^* is a random variable since the estimated treatment rules $\hat{d}_{n,m}$'s vary from data sets to data sets.

For any individualized treatment rule d, denote

$$f(d) = f(X, A, Y; d) := \frac{1_{A=d(X)}}{p(A|X)} Y. \tag{4.2}$$

Then $V(d) = Ef(d)$ and $E_n f(d)$ is an unbiased estimator of $V(d)$ for any fixed d (see Section 1.1). However, $E_n f(\hat{d}_{n,m})$ may not be a good estimator of $V(\hat{d}_{n,m})$ since we use the same data set to estimate and evaluate $\hat{d}_{n,m}$. There might be over-fitting effect. We propose to conduct model selection via penalization to compensate for the possible over-fitting effect.

Our model selection criterion is

$$\hat{m} = \arg \max_{m \in \{1, \cdots, M_n\}} \Big[E_n f(\hat{d}_{n,m}) - pen_n(m) \Big], \tag{4.3}$$

where $pen_n(m)$ is the penalty, which depends on the sample size n, the model complexity and possibly the data.

We will discuss possible methods for constructing the penalty in Section 4.4. In the following section, we give a literature review on step-wise model selection with penalization.

4.2 Literature review

The vast majority of penalty based model selection literatures focuses on prediction (e.g. regression or classification). Despite the differences between prediction and decision making, much insight could be gained by investigating penalization methods developed for prediction.

In regression or classification, one observes i.i.d. copies $\{(X_i, Y_i) : i = 1, \ldots, n\}$, where each X_i takes values in a measurable space $\mathcal{X}$ and Y_i is real-valued ($Y_i \in \{-1, 1\}$ in classification). Define $\psi^*(x) = E(Y|X = x)$ for every $x \in \mathcal{X}$. In the regression case, one is interested in the estimation of ψ^*, and the most commonly used method is to minimize the empirical quadratic risk $E_n(Y - \psi(X))^2$ over a function class Ψ. While in the classification case, one wants to estimate the Bayes classifier $\pi^b(x) = sign(\psi^*(x))$, and one approach is to minimize the empirical classification error $E_n 1_{Y \neq \pi(x)}$ over a class of classifiers Π. This is the so called *empirical risk minimization (ERM) principle* (Vapnik, 1999). On one hand, one may choose a sufficiently large Ψ or Π so as to approximate well any function, but then the estimator get from the ERM principle may fit the data too well and cannot be generalized. This is the so called over-fitting problem. On the other hand, using a small Ψ or Π would make it hard to approximate the truth well. Therefore, selecting the right model Ψ or Π is the key to success.

For this reason, various penalty based model selection methods have been proposed. Basically, one considers a sequence of models with different complexities. For each model, the ERM principle is used to get an estimator. And one then chooses the estimator with minimal penalized empirical risk. Since the empirical risk of the estimators consistently decreases with model complexity, it is natural to incorporate model complexity into the penalty to compensate for possible overfitting effect.

Historically, penalty based model selection began with the work of Mallows (1973) (C_p) and Akaike (1974) (AIC) in the context of linear regression. Schwarz (1978) introduced the BIC criterion under Bayesian considerations. Rissanen (1978, 1983) proposed MDL criterion. Those classical methods are motivated by asymptotic (large-sample) properties of the linear estimators. For practical situations where the sample size is finite, they suffer from large variability of finite data and are often not optimal (Cherkassky et al., 1992). Another disadvantage is that all of the above penalties depend on the number of parameters in each model. This works well with linear model, but for models nonlinear in parameters, the number of parameters is not a good measurement of model complexity.

To deal with the above difficulties, Vapnik and Chervonenkis (1974) and Vapnik (1982, 1995) proposed the *structural risk minimization (SRM)* approach to model selection with finite sample sizes. In this approach, one considers a hierarchy of model classes with increasing complexities measured by VC-indices (defined in Chapter VI). For each model, the empirical risk minimizer is selected. One then chooses the estimator whose sum of empirical risk and VC confidence is minimal. According to the SRM principle, the hierarchy of model classes is defined before the training data appear (Vapnik, 1995). An extension of SRM in the case of classification can be found in Shawe-Taylor et al. (1998). They proved that one can get better risk bounds if the hierarchy of models is chosen according to data.

In literature of late 1990's on nonparametric inference, a general approach quite similar to SRM was developed by Barron et al. (1999) and Birgé and Massart (1997, 1998). They used sieve theory to define a sequence of nested models characterized by some dimensions, where the dimension of the model grows as the sample size increases. In particular, they pioneered the construction of penalties based on the upper bounds of the maximal deviation between the empirical risk and true risk in each model, and obtaining some oracle inequalities. In the context of classification, the oracle inequalities are of the form

$$E[1_{Y\neq\hat{\pi}_{n,\hat{m}}(X)} - 1_{Y\neq\pi^b(X)}] \leq K \inf_m \Big(\inf_{\pi\in\Pi_m} E[1_{Y\neq\pi(X)} - 1_{Y\neq\pi^b(X)}] + \gamma(n,m)\Big), \quad (4.4)$$

where $\{\Pi_m : m = 1, 2, \ldots\}$ is a collection of classes of classifiers, $\hat{\pi}_{n,m}$ is the empirical $0-1$ risk minimizer in the m-th model, K is a constant that is at least as large as 1, and $\gamma(n,m)$ is a quantity that increases with model complexity and decreases to zero as $n \to \infty$. Here, concentration inequalities for empirical processes (van der Vaart and Wellner, 1996; Massart, 2000, 2003) play an important role in bounding the maximal deviation. This approach has become a popular way to prove optimality in nonparametric estimation. Illustration of this method in regression can be found in Baraud (2000).

So far, all penalization methods mentioned above are based on dimension of the competing models. That is, they choose models by balancing the empirical risk with dimensionality. These approaches work well in situations where they apply. However, the dimension of each model is often hard to compute in some situations. Even if the dimension is computable, the obtained estimator may not work well for all distributions since the penalties are chosen independently of the data. Indeed, Kearns et al. (1995) compared (hold-out) cross-validation with some data-independent penalization methods (Rissanen's MDL and Vapnik's SRM). Their overall analysis showed that cross-validation is favored in most common circumstances. This has motivated people to investigate data-dependent penalties (Lugosi and Nobel, 1999).

Using symmetrization techniques in empirical processes (van der Vaart and Wellner, 1996), Koltchinskii (2001) and Bartlett et al. (2002) suggested penalties based on Rademacher averages. For a given function class $\mathcal{G}$, the Rademacher average of $\mathcal{G}$ is defined as $\mathbb{E}_\xi \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \xi_i g(X_i, Y_i)$, where $\xi_1, \ldots, \xi_n$ are i.i.d binary $\{-1, 1\}$ random variables with probability 1/2 each, and the expectation is taken over the distribution of ξ_i's. Note that this quantity measures the model complexity since a large function class is associated with a large Rademacher average. Lozano (2000) gave experimental evidence that the Rademacher penalization outperforms Vapnik's SRM and cross-validation for the interval selection problem in classification. Fromont (2007) further related Rademacher penalization to other resampling techniques. She proposed a penalty based on i.i.d. weighted bootstrap samples of the data (Efron, 1979, 1982), and proved that the Rademacher averages are actually special examples of bootstrap type penalties.

The above SRM and Rademacher penalization methods are based on upper bounds of the maximal deviation between the empirical risk and risk in the entire function class (e.g. $\sup_{\pi \in \Pi_m} (E - E_n) 1_{Y \neq \pi(X)}$ in classification), and ignore the fact that the empirical risk minimizer will likely have small risk and thus only a small subset of each function class should be used. For example, Bartlett (2008) showed that for a nontrivial class Π, the expectation of the maximal deviation would converge to zero at rate no faster than $1/\sqrt{n}$. Thus these penalties may lead to an oracle inequality of the form (4.4) with $\gamma(n, m)$ converging to 0 at rate no faster than $1/\sqrt{n}$ for each m. It is possible to get a faster rate of convergence if the penalty only measures the complexity of a small subset in each function class. This approach was first proposed by Massart (2000, 2003) in the general prediction setting, where the penalty is a data-independent upper bound on the deviation between the empirical excess risk and the excess risk in a subset of functions with small risk in each model. Bartlett et al. (2005) extended Massart's idea by considering local Rademacher complexities. Their work can be used to construct a data dependent computable penalty with a fast rate of convergence if the risk function satisfies some conditions (e.g. bounded regression). In the classification case, Lugosi and Wegkamp (2004) constructed a local Rademacher penalty, which will give a fast rate if the minimal risk is zero. Koltchinskii (2006) and Arlot and Bartlett (2008) further investigated the use of local Rademacher complexities in classification and proposed some sharp data-dependent penalties.

The key condition that allows one to obtain a fast rate of convergence in classification is the fact that the variance of the the excess loss is upper bounded by the power (≤ 1) of the expectation of the excess loss up to a constant. That is,

$$Var[1_{Y=\pi(X)} - 1_{Y=\pi^b(X)}] \leq C_2(E[1_{Y=\pi(X)} - 1_{Y=\pi^b(X)}])^{\beta} \tag{4.5}$$

for any $\pi \in \Pi$ for some $C_2 > 0$ and $\beta \in (0, 1]$. Intuitively, (4.5) implies that the variance of $1_{Y=\pi(X)} - 1_{Y=\pi^b(X)}$ decreases as π approaches π^b. So the risk of the empirical risk minimizer converges to the minimal risk more quickly than the uniform convergence results (Boucheron et al., 2005; Bartlett et al., 2006). Thus penalties based on local complexity measurement should be sharper than those based on maximal deviation within the entire function class.

In fact, since our goal is to minimize the risk, it is obvious that an ideal penalty measures the deviation between the risk and the empirical risk of the empirical risk minimizer (in classification, an ideal penalty is $(E - E_n)1_{Y \neq \hat{\pi}_{n,m}}$). While almost all of the above methods considered upper bounds of the ideal penalty. In a recent paper by Arlot (2009), he provided a bootstrap estimator of the ideal penalty in the histogram selection case. This is a regression setting with indicator predictors and orthogonal design matrix. Thus the ideal penalty has a close form and the bootstrapped version can be shown to concentrate around the truth with high probability. How to obtain such a penalty in general regression and/or classification is still a challenge.

4.3 Oracle inequalities

As demonstrated in the previous section, in the decision making problem, an ideal penalty is $pen_{id}(m) = (E_n - E)f(X, A, Y; \hat{d}_{n,m})$. This penalty will guarantee that

$$V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \leq \inf_{m=1,\ldots,M_n} \left[V(d^{opt}) - V(\hat{d}_{n,m})\right],$$

which is known as the benevolent oracle. However, this penalty depends on the unknown distribution P. There is no hope to perfectly mimic the behavior of the benevolent oracle. It is more realistic to incorporate a factor K and an additive term of the form $\gamma(n, m)$ in the oracle inequality, so that

$$V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \leq \inf_{m=1,\ldots,M_n} \Big\{K[V(d^{opt}) - V(\hat{d}_{n,m})] + \gamma(n, m)\Big\}, \tag{4.6}$$

where the constant K is at least as large as 1 and $\gamma(n, m)$ is increasing in model complexity and decreasing to 0 as $n \to \infty$.

However, since the RHS of (4.6) involves $V(\hat{d}_{n,m})$, which is random and the asymptotic behavior is unclear, (4.6) is not appropriate to serve as an oracle inequality. Below we briefly discuss model selection in classification and propose a reasonable oracle inequality.

Existing theoretical model selection literature for classification can be classified into the following two categories. The first category is empirical risk minimization (ERM): choosing the classifier that minimizes the empirical $0-1$ risk within each model, and then selecting the model with minimal penalized empirical $0-1$ risk. The second category is empirical surrogate risk minimization: choosing the classifier that minimizes an empirical surrogate risk (e.g. hinge loss), and then selecting the model with minimal penalized empirical surrogate risk. In both cases, people estimate the classifiers and perform model selection based on the same loss function. In the context of the decision making problem, this is similar to the following two scenarios.

1. Estimate $\hat{d}_{n,m}$ by maximizing $E_n f(d)$ over a class of rules $\mathcal{D}_m$, and then select the model that maximizes the penalized Value (4.3). In this case, $V(d^{opt}) - V(\hat{d}_{n,m})$ can be decomposed into $[V(d^{opt}) - V(\widetilde{d}_m)] + [V(\widetilde{d}_m) - V(\hat{d}_{n,m})]$, where $\widetilde{d}_m$ is the individualized treatment rule in $\mathcal{D}_m$ that maximizes the Value. Note that $[V(d^{opt}) - V(\widetilde{d}_m)]$ is the approximation error (irreducible) of model m, and $[V(\widetilde{d}_m) - V(\hat{d}_{n,m})]$ is the estimation error $(V(\widetilde{d}_m) - V(\hat{d}_{n,m}) \leq \sup_{d \in \mathcal{D}_m}(E - E_n)[f(\widetilde{d}_m) - f(d)]$, which converges to 0 as $n \to \infty$ under regular conditions). Thus a desirable oracle inequality is that, with a high probability,

$$V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \leq \inf_m \Big\{ K[V(d^{opt}) - V(\widetilde{d}_m)] + \gamma(n,m) \Big\}, \qquad (4.7)$$

   where K is at least as large as 1 and $\gamma(n,m)$ converges to 0 at the same rate as $[V(\widetilde{d}_m) - V(\hat{d}_{n,m})]$.

2. Estimate $\hat{Q}_{n,m}$ by minimizing $E_n(Y - Q)^2$ over $\mathcal{Q}_m$ and then select the model $\hat{m}$ that minimizes the penalized empirical quadratic risk $E_n[Y - \hat{Q}_{n,m}(X,A)]^2 + pen_n(m)$. The final estimated treatment rule $\hat{d}_{n,\hat{m}}$ chooses the treatment that maximizes $\hat{Q}_{n,\hat{m}}(X,a)$. In this approach, one usually relates the Value to the prediction error first (see Theorem II.1), and then constructs a high probability upper bound for the excess prediction error:

$$L(\hat{Q}_{n,m}) - L(Q^{opt}) \leq KL[(Q^*_m) - L(Q^{opt}) + \gamma(n,m).$$

where $L(Q) = E(Y - Q)^2$, $Q_m^* = \arg\min_{Q \in \mathcal{Q}_m} L(Q)$, K is at least as large as 1, and $\gamma_{n,m}$ converges to 0 at the same rate as $L(\hat{Q}_{n,m}) - L(Q_m^*)$. This together with Theorem II.1 implies an oracle inequality of the form

$$V(d^{opt}) - V(\hat{d}_{n,m}) \leq K' \inf_m \left\{ E[(Y - Q_m^*)^2 - (Y - Q^{opt})^2] + \gamma(n,m) \right\}^{(1+\alpha)/(2+\alpha)}. \tag{4.8}$$

Now let us go back to our problem. For each model $m = 1, \ldots, M_n$, define

$$Q_m^* = \arg\min_{Q \in \mathcal{Q}_m} L(Q)$$
$$\text{and } d_m^* \in \arg\max_{a \in \mathcal{A}} Q_m^*(X, a).$$

In this chapter, we use the quadratic loss minimization based method to estimate $\hat{d}_{n,m}$ and choose model by maximizing the penalized empirical Value (4.3). Ideally, we would hope the “estimation error” part $V(d_m^*) - V(\hat{d}_{n,m})$ converges to 0 for every m. And an oracle inequality similar to (4.7) is of the form

$$V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \leq \inf_{m=1,\ldots,M_n} \Big\{ K[V(d^{opt}) - V(d_m^*)] + \gamma(n,m) \Big\}. \tag{4.9}$$

However, under certain circumstances, $V(d_m^*) - V(\hat{d}_{n,m})$ may not converge to 0. Consider the following example.

Example IV.1. Suppose there are two treatments $\mathcal{A} = \{-1, 1\}$ and the randomization probability $p(a|x)$ is $1/2$ for all (x, a) combinations. We consider linear models for Q^{opt}, i.e. for each model m,

$$\mathcal{Q}_m = \Big\{ \Phi_{1,m}(X)\theta_1 + A\Phi_{2,m}(X)\theta_2 : \theta_1 \in \mathbb{R}^{dim(\Phi_{1,m})}, \theta_2 \in \mathbb{R}^{dim(\Phi_{2,m})} \Big\},$$

where $\Phi_{1,m}$ and $\Phi_{2,m}$ are row vectors of basis functions on $\mathcal{X}$. Then $d_m^*(X) =$

$sign(\Phi_{2,m}(X)\theta^*_{2,m})$ (define $sign(0) = 1$), where $(\theta^*_{1,m}, \theta^*_{2,m}) = \arg\min_{\theta_1,\theta_2} E(Y - \Phi_{1,m}(X)\theta_1 - A\Phi_{2,m}(X)\theta_2)^2$. If $\theta^*_{2,m} = 0$, then

$$V(d^*_m) = 2E[1_{A=1}Y] = E\big[Q^{opt}(X,1)\big].$$

However in this case, it is easy to verify that

$$V(\hat{d}_{n,m}) = E\left[1_{\Phi_{2,m}(X)\sqrt{n}(\hat{\theta}_{2,m}-\theta^*_{2,m})\geq 0}Q^{opt}(X,1) + 1_{\Phi_{2,m}(X)\sqrt{n}(\hat{\theta}_{2,m}-\theta^*_{2,m})<0}Q^{opt}(X,-1)\right].$$

As $n \to \infty$, $\sqrt{n}(\hat{\theta}_{2,m} - \theta^*_{2,m})$ converges to a normal random vector. Thus $\Phi_{2,m}(X)\sqrt{n}(\hat{\theta}_{2,m} - \theta^*_{2,m})$ may be positive or negative. If an optimal rule d^{opt} is indecisive (i.e. $Q^{opt}(x,1) = Q^{opt}(x,-1)$ for all $x \in \mathcal{X}$), then $V(\hat{d}_{n,m}) = V(d^*_m)$. Otherwise $V(\hat{d}_{n,m})$ may not converge to $V(d^*_m)$.

Thus in order to obtain (4.9), we need to assume that either $\theta^*_{2,m} \neq 0$ for all m (i.e. all d^*_m are decisive) or d^{opt} is indecisive. This condition is strong in the sense that it depends on both the unknown system dynamics and all the models under consideration. □

To avoid making assumptions stated in the previous example, we adopt oracle inequality of the form (4.8). Note that by the weighted AM-GM inequality, (4.8) is equivalent to

$$V(d^{opt}) - V(\hat{d}_{n,m}) \leq \inf_m \left\{\bar{K}\big[L(Q^*_m) - L(Q^{opt})\big]^{(1+\alpha)/(2+\alpha)} + \bar{\gamma}(n,m)\right\}, \tag{4.10}$$

where $\bar{K}$ is a constant and $\bar{\gamma}(n,m)$ converges to 0 as $n \to \infty$ for every m.

4.4 Penalization methods

In this section, we will discuss penalization methods that give us oracle inequality of form (4.10) with high probability.

4.4.1 Penalty based on maximal deviation

Note that by the definition of $\hat{m}$,

$$
\begin{aligned}
&V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \\
=&\Big[V(d^{opt}) - V(\hat{d}_{n,m})\Big] + Ef(\hat{d}_{n,m}) - Ef(\hat{d}_{n,\hat{m}}) \\
\leq&\Big[V(d^{opt}) - V(\hat{d}_{n,m})\Big] + (E - E_n)f(\hat{d}_{n,m}) + pen_n(m) + (E_n - E)f(\hat{d}_{n,\hat{m}}) - pen_n(\hat{m}).
\end{aligned}
$$

By Theorem II.1, we have that $V(d^{opt}) - V(\hat{d}_{n,m}) \leq C_1[L(\hat{Q}_{n,m}) - L(Q^{opt})]^{(1+\alpha)/(2+\alpha)}$ under the margin condition (2.2), which can be further upper bounded by $K(L(Q^*_m) - L(Q^{opt}))^{(1+\alpha)/(2+\alpha)} + \gamma(n,m)$ under appropriate conditions. Thus to achieve an oracle inequality of form (4.10), it is sufficient to choose the penalty as a nontrivial upper bound on the maximal deviation $\sup_{d\in\mathcal{D}_m}(E_n - E)f(d)$ (the upper bound is nontrivial in the sense that it converges to 0 as $n \to \infty$). Such an upper bound can be obtained by using concentration inequalities in the theory of empirical processes (Bartlett et al., 2002; Massart, 2000, 2003; Bartlett, 2008; Fromont, 2007). This upper bound could be either distribution-free, such as quantities depending on the dimension of the parameter space in each model class (Barron et al., 1999), or data-dependent, such as quantities based on Rademacher averages (Koltchinskii, 2001; Bartlett et al., 2002) or bootstrap estimators (Fromont, 2007).

As we have discussed in Section 4.2, penalties based on the upper bound of the maximal deviation in the entire function class measures the complexity of the entire function class and the resulting $\bar{\gamma}(n,m)$ in (4.10) may approach 0 at a rate no faster than $1/\sqrt{n}$. This makes us think of penalties that estimate the maximal deviation in a small ball around $\hat{d}_{n,m}$. In the following section, we will consider such a penalization based on local Rademacher complexities.

4.4.2 Margin adaptive model selection

In this section, we propose a penalization method so that $\bar{\gamma}(n,m)$ in (4.10) converges to zero at a rate adapting to the Margin condition (2.2). First note that if $Q^{opt}(X,A)$ is constant in A, then $V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) = 0$. The oracle inequality (4.10) trivially holds. In the following, we consider the case where Q^{opt} is not a constant function in A. We start with the following assumptions.

Assumption IV.1. *There exist some constants $S \geq 1$ and $b > 0$ such that*

(a) $p(a|x) > S^{-1}$ *for all combinations of* (x,a)*;*

(b) $|Y| \leq b$*; and*

(c) $\sup_{Q \in \cup_m \mathcal{Q}_m} \|Q\|_\infty \leq b$*.*

This assumption requires that all relevant quantities are bounded. Later we will see that it is directly relevant to Assumption IV.3 below. In addition, this technical condition is often employed in concentration inequalities for empirical processes. It is possible to replace Assumption IV.1(b) with a moment assumption on error terms and a boundedness assumption on Q^{opt}.

Assumption IV.2. *Each approximation space $\mathcal{Q}_m$ is convex for $m = 1, \ldots, M$.*

This assumption has two functionalities. First, this together with assumption IV.1(b) and (c) implies a bernstein condition on the quadratic loss function (see Lemma IV.9), which is the key to obtain a fast rate of convergence in prediction error within each model (Bartlett and Mendelson, 2006). Second, this condition ensures that the class $\mathcal{Q}_m$ is star-shaped (see Definition VI.4), which allows us to construct a data-dependent penalty with fast rate of convergence. It is easy to verify that this condition holds if each function class $\mathcal{Q}_m$ is linear.

Assumption IV.3. *For any optimal individualized treatment rule d^{opt} such that $d^{opt}(X) \in \arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)$, square integrable function Q on $\mathcal{X} \times \mathcal{A}$ and individualized treatment rule d such that $d(X) \in \arg\max_{a\in\mathcal{A}} Q(X,a)$, there exist some $C_1, C_2 > 0$ and $\alpha \geq 0$ such that*

(a) $V(d^{opt}) - V(d) \leq C_1[L(Q) - L(Q^{opt})]^{(1+\alpha)/(2+\alpha)}$; and

(b) $E[f(d^{opt}) - f(d)]^2 \leq C_2[Ef(d^{opt}) - Ef(d)]^{\alpha/(1+\alpha)}$, where $f(d)$ is defined in (4.2).

This assumption is the key to show a rate of convergence faster than $1/\sqrt{n}$ when $\alpha > 0$. Note that this assumption always holds with $\alpha = 0$ under Assumption IV.1(a) and (b). When $\alpha > 0$, condition (a) implies a tighter upper bound between the excess Value and the excess prediction error, and condition (b) implies the variance of $f(d^{opt}) - f(d)$ is upper bounded by its expectation to a power between 0 and 1. Thus the variance of $f(d^{opt}) - f(d)$ is small if the Value of d is close to the optimal Value, which gives a fast rate of convergence.

In fact, Assumption IV.3 is closely related to the margin condition (2.2). In Chapter II, we have showed that Assumption IV.3(a) holds if Assumption IV.1(a) and the margin condition hold (see Theorem II.1). The following proposition explains the origin of Assumption IV.3(b) and its relation to the margin condition when $\alpha > 0$.

Proposition IV.1. *Assume the margin condition (2.2) holds with some $C > 1$ and $\alpha > 0$. Suppose Assumptions IV.1(a), (b) hold and $\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)$ is unique a.s. Then for any square integrable function Q on $\mathcal{X}\times\mathcal{A}$ and individualized treatment rule d such that $d(X) \in \arg\max_{a\in\mathcal{A}} Q(X,a)$, we have*

$$E[f(d^{opt}) - f(d)]^2 \leq C_2[Ef(d^{opt}) - Ef(d)]^{\alpha/(1+\alpha)}, \tag{4.11}$$

where $C_2 = 2b^2S(1+\alpha)C^{1/(1+\alpha)}\alpha^{-\alpha/(1+\alpha)}$.

The proof is given in Section 4.6.1.

Remark

The exponent on the RHS of (4.11) approaches 1 as $\alpha \to \infty$. In this case, the margin condition requires that the LHS of (2.2) equals 0 for all $\epsilon \in (0,1)$, which is unlikely to be true. However, we can replace the margin condition (2.2) by the following condition. *There exists an* $\epsilon > 0$ *such that*

$$\mathbb{P}\Big(0 < \max_{a\in\mathcal{A}} Q^{opt}(X,a) - \max_{a\in\mathcal{A}\setminus\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} Q^{opt}(X,a) \le \epsilon\Big) = 0.$$

Then (4.11) holds with $C_2 = 2b^2S/\epsilon$ and $\alpha = \infty$.

After the list of assumptions, below we provide a sufficient condition for the penalty term to attain margin adaptivity.

For any function class $\mathcal{G}$ on $\mathcal{X}\times\mathcal{A}\times\mathbb{R}$, let $N(\epsilon,\mathcal{G},L_1(P_n))$ denote the ϵ-covering number of $\mathcal{G}$ relative to the $L_1(P_n)$ norm and denote $u_n(\mathcal{G}) = \mathbb{E}\log[N(1/n,\mathcal{G},L_1(P_n))+1]/n$.

For each model $m = 1,\ldots,M_n$ and any $t > 0$, define the sets of functions

$$\begin{aligned}\mathcal{D}_m =& \Big\{d(X) \in \arg\max_{a\in\mathcal{A}} Q(X,a) : Q \in \mathcal{Q}_m\Big\},\\ \mathcal{F}_m =& \{f(d) : d \in \mathcal{D}_m\},\ \text{where } f(d) = \frac{1_{A=d(X)}}{p(A|X)}Y.\end{aligned}$$

Theorem IV.1. *Suppose Assumptions IV.1(a),(b) and IV.3(b) hold. For any* $t > 0$ *and* $0 < \delta < 1$, *assume there exists a positive constant* c *such that the penalty term satisfies*

$$pen_n(m) + \frac{ct}{n} \ge [(1-\delta)E - E_n](f(d_m^*) - f(\hat{d}_{n,m})). \tag{4.12}$$

Let $\hat{m}$ be the selected model defined in (4.3). Then with probability at least $1-\exp(-t)$, *we have*

$$\begin{aligned}
&(1-\delta)\big[V(d^{opt})-V(\hat{d}_{n,\hat{m}})\big] \\
\le &\inf_{m=1,\dots,M_n}\Big\{(1+\delta)\Big[V(d^{opt})-V(\hat{d}_{n,m})\Big]+pen_n(m) \\
&\quad + K_1\Big[u_n(\mathcal{F}_m)^{(1+\alpha)/(2+\alpha)}+\Big(\frac{t+\log(2M_n)}{n}\Big)^{(1+\alpha)/(2+\alpha)}+\frac{t+\log(2M_n)}{n}\Big]\Big\}
\end{aligned}$$

for a sufficiently large constant K_1 *depending on* $b, S, \alpha, C_2, \delta$ *and* c.

The proof is given in Section 4.6.2.

Note that $V(d^{opt})-V(\hat{d}_{n,m})$ is bounded above by $C_1[L(\hat{Q}_{n,m})-L(Q^{opt})]^{(1+\alpha)/(2+\alpha)}$ under Assumptions IV.1(a) and IV.3(a). It also can be shown that with a high probability $L(\hat{Q}_{n,m}-L(Q^*_m)\le O(u_n(\mathcal{Q}_m))$ under Assumptions IV.1(b), (c) and IV.2 (Bartlett et al., 2005). This, together with the weighted AM-GM inequality, will give an oracle inequality of form (4.10) with the desired rate of convergence as long as the penalty is of the right order.

The above theorem applies to any penalization procedure that satisfies condition (4.12). Now we propose a data-dependent penalty based on local Rademacher complexities. Below we define precisely these complexities.

For any $s \ge 0$, define the set of individualized treatment rules

$$\hat{B}_m(s)=\Big\{d(X)\in\arg\max_a Q(X,a): E_n[(Y-Q)^2-(Y-\hat{Q}_{n,m})^2]\le s, Q\in\mathcal{Q}_m\Big\}.$$

Let $\xi_1,\dots,\xi_n$ be i.i.d. Rademacher random variables (i.e. $\mathbb{P}(\xi_i=1)=\mathbb{P}(\xi_i=-1)=1/2$). For any $t>0$, $r>0$, define

$$\hat{\eta}_m(r)=256b^3\Big(\frac{4}{3}\mathbb{E}_\xi \sup_{Q\in\mathcal{Q}_m:b^2E_n(Q-\hat{Q}_m)^2\le 170r}\frac{1}{n}\sum_{i=1}^n\xi_i Q(X_i,A_i)+\frac{6bt}{n}\Big), \tag{4.13}$$

where the expectation $\mathbb{E}_\xi$ is taken with respect to $(\xi_1, \ldots, \xi_n)$ and can be empirically approximated by repeatedly sampling $(\xi_1, \ldots, \xi_n)$.

By Lemmas VI.11 and VI.12 in Section 4.6.3, $\hat{\eta}_m(r)$ has a unique positive fixed point under Assumption IV.2. Let $\hat{r}_{n,m}$ denote the positive fixed point of $\hat{\eta}_m(r)$. Define

$$\hat{s}_{n,m} = \frac{132}{b^2}\hat{r}_{n,m} + \frac{5300b^2 t}{n}. \tag{4.14}$$

We have the following theorem.

Theorem IV.2. *Suppose Assumptions IV.1, IV.2 and IV.3 hold. For any given* $t > 0$ *and* $\delta \in (0,1)$*, let* $\hat{m}$ *be the model selected according to (4.3) with*

$$\begin{aligned} pen_n(m) = \delta & \sup_{d_1,d_2 \in \hat{B}_m(\hat{s}_{n,m})} E_n[f(d_1) - f(d_2)] \\ & + 10(1+\delta)\mathbb{E}_\xi \sup_{d_1,d_2 \in \hat{B}_m(\hat{s}_{n,m})} \frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i, A_i, Y_i; d_1) - f(X_i, A_i, Y_i; d_2)] \\ & + 2(1+\delta)\sqrt{\sup_{d_1,d_2 \in \hat{B}_m(\hat{s}_{n,m})} E_n[f(d_1) - f(d_2)]^2 \frac{t}{n}}. \end{aligned} \tag{4.15}$$

Then with probability at least $1 - \exp(-t)$*,*

$$\begin{aligned} & V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \\ \le & K_2 \inf_{m=1,\ldots,M_n} \Big\{ \big[L(Q_m^*) - L(Q^{opt})\big]^{(1+\alpha)/(2+\alpha)} + u_n(\mathcal{F}_m) + u_n(\mathcal{F}_m)^{(1+\alpha)/(2+\alpha)} \\ & \quad + u_n(\mathcal{Q}_m)^{(1+\alpha)/(2+\alpha)} + \Big(\frac{t + \log(15M_n)}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{t+\log(15M_n)}{n} \Big\} \end{aligned}$$

for a sufficiently large constant K_2 *depending on* b, S, α, C_1, C_2 *and* δ*.*

The proof is given in Section 4.6.3.

Note that penalty (4.15) is a data-dependent penalty. It measures the deviation in the ball $\hat{B}_m(\hat{s}_{n,m})$. If the models are nested, then the ball is large for large models.

Thus the penalty reflects the model complexity. The rate of convergence in the final result depends on the margin parameter α. To better illustrate the rate of convergence, we give the following corollary.

Corollary IV.1. *Suppose there are only two treatments $\mathcal{A} = \{-1, 1\}$. Assume Assumptions IV.1, IV.2 and IV.3 hold. For any given $t > 0$ and $\delta \in (0, 1)$, let $\hat{m}$ be the model selected according to (4.3) with penalty satisfying (4.15). If each model $\mathcal{Q}_m$ is a VC-class, then with probability at least $1 - \exp(-t)$,*

$$
\begin{aligned}
&V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \\
\leq &K_3 \inf_{m=1,\ldots,M_n} \Big\{ \big[L(Q_m^*) - L(Q^{opt})\big]^{(1+\alpha)/(2+\alpha)} + \Big(\frac{vc(\mathcal{Q}_m)\log n}{n}\Big)^{(1+\alpha)/(2+\alpha)} \\
&\quad + \frac{vc(\mathcal{Q}_m)\log n}{n} + \Big(\frac{t + \log(15M_n)}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{t + \log(15M_n)}{n}\Big\}
\end{aligned}
$$

for a sufficiently large constant K_3 depending on b, S, α, C_1, C_2 and δ, where $vc(\mathcal{Q}_m)$ is the VC-index of the set of subgraphs of functions in $\mathcal{Q}_m$.

The result follows from Lemma VI.4 (which provides a connection between the covering number and VC-index) and the preservation properties of VC class (Lemma VI.5). And the proof is omitted.

Note that if each model $\mathcal{Q}_m$ is a convex subset of a l_m-dimensional vector space, then $vc(\mathcal{Q}_m) \leq l_m + 2$. For example, suppose each model class is of the form $\mathcal{Q}_m = \{\Phi_m \boldsymbol{\theta} : |\boldsymbol{\theta}| \leq b\}$, where Φ_m is a $1 \times l_m$ vector of basis functions, and the sup-norm of each component in Φ_m is bounded above by 1. Then it is easy to verify that Assumptions IV.1(c) and IV.2 hold and $\mathcal{Q}_m$ is a subset of a finite dimensional vector space. In general, one can take $t = \log n$. Intuitively, this oracle inequality means that if there is a simple model that approximates Q^{opt} sufficiently well (so that both $vc(\mathcal{Q}_m)$ and $L(Q_m^*) - L(Q^{opt})$) are small, then the estimated individualized treatment rule will have Value close to the optimal Value.

4.5 Discussion

In this chapter, we considered a step-wise model selection procedure to improve the ability of the quadratic loss minimization based method. Unlike other theoretical model selection work, this approach is novel in the sense that the estimation and model selection are based on different loss functions. We justified this approach by providing a high probability upper bound on the quality of the estimated individualized treatment rule.

Although the proposed penalization method gives us a theoretical margin adaptive rate of convergence, it is practically hard to implement. First, the upper bound b in Assumption IV.1 is required in order to compute the penalty. In addition, we need to compute the fixed point of every $\hat{\eta}_m$ (defined in (4.13)), take sup over each sequence of sampled Rademacher variables ξ_i's and then average over sampled sequence of Rademacher variables. A possible future direction is to develop an easy-to-compute penalty with a fast rate of convergence.

As discussed in Section 4.2, an ideal penalty is $(E_n - E)f(\hat{d}_{n,m})$. In the simple histogram selection setting, Arlot (2009) provided a bootstrap penalty, which is concentrated around the idea penalty with high probability. This is a regression setting with piecewise constant predictors and orthogonal design matrix. Thus the ideal penalty has a closed form. How to obtain such a penalty in other general settings is an open problem.

In this chapter, we provided an oracle inequality in which the RHS contains a measure of approximation error $E[(Y - Q^*_m)^2 - (Y - Q^{opt})^2]$. This oracle inequality implies that the estimated rule $\hat{d}_{n,\hat{m}}$ is of high quality if $\inf_{m=1,\ldots,M_n}(E[(Y-Q^*_m)^2-(Y-Q^{opt})^2])$ is small and the sample size is large. However, this oracle inequality is still not ideal since $V(d^{opt}) - V(\hat{d}_{n,\hat{m}})$ could be small and $E[(Y-Q^*_m)^2-(Y-Q^{opt})^2]$ could be very large when all models are poor. Further exploration of an ideal oracle inequality that could better justify the quality of the estimated individualized treatment rule is needed.

4.6 Appendices

4.6.1 Proof of Proposition IV.1

First note that for any $d^{opt}(X) \in \arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)$ and individualized treatment rule d,

$$\begin{aligned} E[f(d^{opt}) - f(d)]^2 =& E\Big[\frac{1_{A=d^{opt}(X)} - 1_{A=d(X)}}{p(A|X)}Y\Big]^2 \\ \leq& b^2 SE\left[\frac{(1_{A=d^{opt}(X)} - 1_{A=d(X)})^2}{p(A|X)}\right] \\ =& b^2 SE\Big[\sum_{a\in\mathcal{A}}(1_{d^{opt}(X)=a} - 1_{d(X)=a})^2\Big] \\ \leq& 2b^2 SE1_{d(X)\neq d^{opt}(X)} \end{aligned}$$

This together with the assumption that $\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)$ is unique a.s. implies

$$E[f(d^{opt}) - f(d)]^2 \leq 2b^2 SE1_{d(X)\in\mathcal{A}\setminus\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)}. \tag{4.16}$$

For any $\epsilon > 0$, define the event

$$\Omega_\epsilon = \Big\{0 < \max_{a\in\mathcal{A}} Q^{opt}(X,a) - \max_{a\in\mathcal{A}\setminus\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} Q^{opt}(X,a) \leq \epsilon\Big\}.$$

If the margin condition (2.2) holds with some $\alpha > 0$, then following the arguments in Section 1.3, we have

$$E[f(d^{opt}) - f(d)] = E\left[\big(\max_{a\in\mathcal{A}} Q^{opt}(X,a) - Q^{opt}(X,d(X))\big)1_{d(X)\in\mathcal{A}\setminus\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)}\right]$$

$$\begin{aligned}
&\geq E\left[\big(\max_{a\in\mathcal{A}} Q^{opt}(X,a) - Q^{opt}(X,d(X))\big) 1_{d(X)\in\mathcal{A}\backslash\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} 1_{\Omega_\epsilon^C}\right] \\
&\geq \epsilon\big[E1_{d(X)\in\mathcal{A}\backslash\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} - E1_{\Omega_\epsilon}\big] \\
&\geq \epsilon E1_{d(X)\in\mathcal{A}\backslash\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)} - C\epsilon^{\alpha+1},
\end{aligned}$$

Choosing $\epsilon = \big(E1_{d(X)\in\mathcal{A}\backslash\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)}/[(1+\alpha)C]\big)^{1/\alpha}$ to maximize the above lower bound yields

$$E[f(d^{opt}) - f(d)] \geq C^{-1/\alpha}\alpha(1+\alpha)^{-(1+\alpha)/\alpha}\big(E1_{d(X)\in\mathcal{A}\backslash\arg\max_{a\in\mathcal{A}} Q^{opt}(X,a)}\big)^{(1+\alpha)/\alpha}.$$

The result follows by combining the above result with inequality (4.16). □

4.6.2 Proof of Theorem IV.1

Fix an optimal individualized treatment rule d^{opt}. For each model $m = 1, \ldots, M_m$, define the quantity

$$H_m = \sup_{d\in\mathcal{D}_m}\left|\frac{(E_n - E)[f(d^{opt}) - f(d)]}{[E(f(d^{opt}) - f(d))]^{\alpha/(1+\alpha)} + h_m}\right|,$$

where $h_m = k\big(u_n(\mathcal{F}_m)^{\alpha/(2+\alpha)}\big[1 \vee u_n(\mathcal{F}_m)^{1/(2+\alpha)}\big] + (t/n)^{\alpha/(2+\alpha)}\big[1 \vee (t/n)^{1/(2+\alpha)}\big]\big)$, k is a large enough constant depending on b, S and C_2 (see Lemma IV.1); and the events

$$\begin{aligned}
\Omega_{1.m} =&\Big\{H_m \leq u_n(\mathcal{F}_m)^{1/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{1/(2+\alpha)}\Big\} \\
\Omega_{2,m} =&\Big\{(E - E_n)[f(d^{opt}) - f(d_m^*)] \leq \sqrt{\frac{2tE[f(d^{opt}) - f(d_m^*)]^2}{n}} + \frac{2bSt}{3n}\Big\}.
\end{aligned}$$

By the definition of $\hat{m}$, $E_n f(\hat{d}_{n,m}) - pen_n(m) \leq E_n f(\hat{d}_{n,\hat{m}}) - pen_n(\hat{m})$. Thus for any $m = 1, \ldots, M_n$, on the event $\Omega_{1,m}$,

$$\begin{aligned}
&V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \\
=&\Big[V(d^{opt}) - V(\hat{d}_{n,m})\Big] + Ef(\hat{d}_{n,m}) - Ef(\hat{d}_{n,\hat{m}}) \\
\leq&\Big[V(d^{opt}) - V(\hat{d}_{n,m})\Big] + (E_n - E)[f(d^{opt}) - f(\hat{d}_{n,m})] + pen_n(m) \\
&+ (E - E_n)[f(d^{opt}) - f(\hat{d}_{n,\hat{m}})] - pen_n(\hat{m}) \\
\leq&\big[V(d^{opt}) - V(\hat{d}_{n,m})\big] + h_m\Big[u_n(\mathcal{F}_m)^{1/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{1/(2+\alpha)}\Big] \\
&+ \big[V(d^{opt}) - V(\hat{d}_{n,m})\big]^{\alpha/(1+\alpha)}\Big[u_n(\mathcal{F}_m)^{1/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{1/(2+\alpha)}\Big] \\
&+ pen_n(m) + (E - E_n)[f(d^{opt}) - f(\hat{d}_{n,\hat{m}})] - pen_n(\hat{m}) \\
\leq&(1+\delta)\big[V(d^{opt}) - V(\hat{d}_{n,m})\big] + h_m\Big[u_n(\mathcal{F}_m)^{1/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{1/(2+\alpha)}\Big] \\
&+ \frac{1}{(1+\alpha)\delta^\alpha}\Big[u_n(\mathcal{F}_m)^{(1+\alpha)/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{(1+\alpha)/(2+\alpha)}\Big] \\
&+ pen_n(m) + (E - E_n)[f(d^{opt}) - f(\hat{d}_{n,\hat{m}})] - pen_n(\hat{m}), \qquad (4.17)
\end{aligned}$$

where the last inequality follows from the weighted AM-GM inequality and the fact that $\alpha/(1+\alpha) \leq 1$.

Next by Assumption IV.3(b), we have on the event $\Omega_{2,m}$,

$$\begin{aligned}
&(E - E_n)[f(d^{opt}) - f(d^*_m)] \\
\leq&\sqrt{\frac{2tC_2[E(f(d^{opt}) - f(d^*_m))]^{\alpha/(1+\alpha)}}{n}} + \frac{2bSt}{3n} \\
\leq&\delta[V(d^{opt}) - V(d^*_m)] + \Big(\frac{1}{2\delta}\Big)^{\alpha/(2+\alpha)}\Big(\frac{2C_2t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{2bSt}{3n}.
\end{aligned}$$

Thus on the event $\cap_m \Omega_{2,m}$,

$$\begin{aligned}
&(E - E_n)[f(d^{opt}) - f(\hat{d}_{n,\hat{m}})] \\
\leq&\delta[V(d^{opt}) - V(d^*_{\hat{m}})] + \Big(\frac{1}{2\delta}\Big)^{\alpha/(2+\alpha)}\Big(\frac{2C_2t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{2bSt}{3n}
\end{aligned}$$

$$
\begin{aligned}
&+ (E - E_n)[f(d^*_{\hat{m}}) - f(\hat{d}_{n,\hat{m}})] \\
\le &\delta[V(d^{opt}) - V(\hat{d}_{n,\hat{m}})] + \Big(\frac{1}{2\delta}\Big)^{\alpha/(2+\alpha)}\Big(\frac{2C_2 t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{2bSt}{3n} \\
&+ ((1-\delta)E - E_n)[f(d^*_{\hat{m}}) - f(\hat{d}_{n,\hat{m}})]. \qquad (4.18)
\end{aligned}
$$

Substituting (4.18) into (4.17) and using the penalty condition (4.12), we obtain

$$
\begin{aligned}
(1-\delta)V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \le \inf_{m=1,\ldots,M_n} \Big\{ &(1+\delta)\Big[V(d^{opt}) - V(\hat{d}_{n,m})\Big] + pen_n(m) \\
&+ K_1\Big[u_n(\mathcal{F}_m)^{(1+\alpha)/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{t}{n}\Big]\Big\}
\end{aligned}
$$

on the event $\cap_m(\Omega_{1,m} \cap \Omega_{2,m})$, where K_1 is a sufficiently large constant depending on $b, S, \alpha, C_2, \delta$ and c.

Taking $t = t + \log(2M_n)$, the result follows from Lemmas IV.1, IV.2 and the union bound argument.

Lemma IV.1. *Assume assumptions IV.1(a),(b) and IV.3(b) hold. Then* $\mathbb{P}(\Omega_{1,m}) \ge 1 - \exp(-t)$ *for a sufficiently large constant* k *depending on* b, S *and* C_2.

Proof. Under Assumptions IV.1(a), (b) and IV.3(b), for any $d \in \mathcal{D}_m$

$$
\begin{aligned}
&\Big\|\frac{f(d^{opt}) - f(d) - P[f(d^{opt}) - f(d)]}{(E[f(d^{opt}) - f(d)])^{\alpha/(1+\alpha)} + h_m}\Big\|_\infty \le \frac{2bS}{h_m}; \quad \text{and} \\
&Var\Big(\frac{f(d^{opt}) - f(d) - E[f(d^{opt}) - f(d)]}{(E[f(d^{opt}) - f(d)])^{\alpha/(1+\alpha)} + h_m}\Big) \le \frac{E[f(d^{opt}) - f(d)]^2}{4h_m(E[f(d^{opt}) - f(d)])^{\alpha/(1+\alpha)}} \le \frac{C_2}{4h_m}.
\end{aligned}
$$

By Lemma VI.9, we have with probability at least $1 - \exp(-t)$,

$$
H_m \le \mathbb{E}H_m + \frac{1}{n}\sqrt{2t\Big(\frac{C_2 n}{4h_m} + \frac{4bSn}{h_m}\mathbb{E}H_m\Big)} + \frac{2bSt}{3nh_m} \le 2\mathbb{E}H_m + \sqrt{\frac{C_2 t}{2nh_m}} + \frac{8bSt}{3nh_m}.
$$

Let $\mathcal{G}_m = \{g = f(d^{opt}) - f(d) : d \in \mathcal{D}_m\}$. Then

$$H_m = \sup_{g \in \mathcal{G}_m} \Big| \frac{(E_n - E)g}{(Eg)^{\alpha/(1+\alpha)} + h_m} \Big|.$$

It is easy to verify that $u_n(\mathcal{G}_m) = u_n(\mathcal{F}_m)$.

Let $\mathcal{G}_{m,0}$ be a $2/n$-net in $L_1(P_n)$ over $\mathcal{G}_m$. The cardinality of $\mathcal{G}_{m,0}$ can be chosen equal to $N(1/n, \mathcal{G}_m, L_1(P_n))$. Then by symmetrization inequality and the definition of $\mathcal{G}_{m,0}$, for any $r > 0$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{g \in \mathcal{G}_m : Eg^2 \le r} |(E_n - E)g|\right] \le& 2\mathbb{E}\left[\sup_{g \in \mathcal{G}_m : Eg^2 \le r} \Big|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big|\right] \\ \le& 2\mathbb{E}\left[\sup_{g \in \mathcal{G}_{m,0} : Eg^2 \le r} \Big|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big|\right] + \frac{4}{n}.\end{aligned}$$

In addition, note that

$$\begin{aligned}&\mathbb{E}\left[\sup_{g \in \mathcal{G}_{m,0} : Eg^2 \le r} \Big|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big|\right] \\ \le& \frac{1}{\sqrt{\log 2}n}\mathbb{E}\left[\Big\|\sup_{g \in \mathcal{G}_{m,0} : Eg^2 \le r} \Big|\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big|\Big\|_{\psi_2}\right] \\ &\qquad\qquad\text{(by the definition of Orlicz norm)} \\ \le& \frac{2\sqrt{2}}{n\log 2}\mathbb{E}\Big[\sqrt{\log(N(1/n, \mathcal{G}_m, L_1(P_n)) + 1)} \sup_{g \in \mathcal{G}_{m,0} : Eg^2 \le r} \Big\|\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big\|_{\psi_2}\Big] \\ &\qquad\qquad\text{(by Maximal inequality)} \\ \le& \frac{2\sqrt{2}}{n\log 2}\mathbb{E}\Big[\sqrt{\log(N(1/n, \mathcal{G}_m, L_1(P_n)) + 1)} \sup_{g \in \mathcal{G}_{m,0} : Eg^2 \le r} \sqrt{6nE_n g^2}\Big] \\ &\qquad\qquad\text{(by Hoeffding's inequality)} \\ \le& \frac{4\sqrt{3}}{\log 2}\sqrt{u_n(\mathcal{G}_m)\mathbb{E}\sup_{g \in \mathcal{G}_{m,0} : Eg^2 \le r} E_n g^2} \quad \text{(by Cauchy-Schwarz inequality)}\end{aligned}$$

$$\leq \frac{4\sqrt{3}}{\log 2}\sqrt{u_n(\mathcal{G}_m)\Big[r + 4bS\mathbb{E}\sup_{g\in\mathcal{G}_{m,0}:Eg^2\leq r}\Big|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big|\Big]}$$

(by assumption IV.1(a), (b) and lemma VI.16),

which implies

$$\mathbb{E}\sup_{g\in\mathcal{G}_{m,0}:Eg^2\leq r}\Big|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i, A_i, Y_i)\Big| \leq \frac{192bSu_n(\mathcal{G}_m)}{(\log 2)^2} + \frac{4\sqrt{3u_n(\mathcal{G}_m)r}}{\log 2}.$$

Since $u_n(\mathcal{G}_m) = u_n(\mathcal{F}_m)$, we have for any $r > 0$,

$$\begin{aligned}\mathbb{E}\left[\sup_{g\in\mathcal{G}_m:Eg^2\leq r}|(E_n - E)g|\right] \leq& \frac{384bSu_n(\mathcal{F}_m)}{(\log 2)^2} + \frac{8\sqrt{3u_n(\mathcal{F}_m)r}}{\log 2} + \frac{4}{n}\\ \leq& \frac{(384bS + 4\log 2)u_n(\mathcal{F}_m)}{(\log 2)^2} + \frac{8\sqrt{3u_n(\mathcal{F}_m)r}}{\log 2},\end{aligned}$$

where the second inequality follows from the fact that $u_n(\mathcal{F}_m) \geq \log 2/n$.

By simple algebra, for any $r \geq (96bS/\log 2+1)^2 u_n(\mathcal{F}_m)^{\alpha/(2+\alpha)}[1\vee u_n(\mathcal{F}_m)^{1/(2+\alpha)}]/12$,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_m:Eg^2\leq r}|(E_n - E)g|\right] \leq \frac{16\sqrt{3}}{\log 2}\sqrt{u_n(\mathcal{F}_m)r[1\vee u_n(\mathcal{F}_m)^{1/(2+\alpha)}]}.$$

For any $s > 1$, let j' be the smallest integer such that $s^{2(j'+1)}h_m \geq b^2S^2$. Then under Assumption IV.3(b),

$$\begin{aligned}H_m \leq& \sup_{g\in\mathcal{G}_m:Eg^2\leq h_m}\Big|\frac{C_2(E_n - E)g}{Eg^2 + C_2h_m}\Big| + \sum_{j=0}^{j'}\sup_{g\in\mathcal{G}_m:s^{2j}h_m\leq Eg^2\leq s^{2(j+1)}h_m}\Big|\frac{C_2(E_n - E)g}{Eg^2 + C_2h_m}\Big|\\ \leq& \frac{1}{h_m}\sup_{g\in\mathcal{G}_m:Eg^2\leq h_m}\big|(E_n - E)g\big|\\ &+ \sum_{j=0}^{j'}\frac{C_2}{(s^{2j} + C_2)h_m}\sup_{g\in\mathcal{G}_m:s^{2j}h_m\leq Eg^2\leq s^{2(j+1)}h_m}\big|(E_n - E)g\big|.\end{aligned}$$

If $h_m \geq (96bS/\log 2+1)^2 u_n(\mathcal{F}_m)^{\alpha/(2+\alpha)}[1\vee u_n(\mathcal{F}_m)^{1/(2+\alpha)}]/12$, then

$$\begin{aligned}\mathbb{E}H_m \leq & \Big[1+\sum_{j=0}^{j'}\frac{C_2 s^{j+1}}{(s^{2j}+C_2)}\Big]\frac{16\sqrt{3}}{\log 2}\sqrt{\frac{u_n(\mathcal{F}_m)[1\vee u_n(\mathcal{F}_m)^{1/(2+\alpha)}]}{h_m}} \\ \leq & \frac{16\sqrt{3}(1+4C_2)}{\log 2}\sqrt{\frac{u_n(\mathcal{F}_m)[1\vee u_n(\mathcal{F}_m)^{1/(2+\alpha)}]}{h_m}}\end{aligned}$$

by taking $s=2$.

By the definition of h_m, for any $k \geq (96bS/\log 2+1)^2/12$, we have w.p. $\geq 1-\exp(-t)$,

$$H_m \leq \frac{32\sqrt{3}(1+4C_2)}{\sqrt{k}\log 2}u_n(\mathcal{F}_m)^{1/(2+\alpha)} + \Big(\sqrt{\frac{C_2}{2k}}+\frac{8bS}{3k}\Big)\Big(\frac{t}{n}\Big)^{1/(2+\alpha)},$$

which is no larger than $u_n(\mathcal{F}_m)^{1/(2+\alpha)}+(t/n)^{1/(2+\alpha)}$ if k is sufficiently large. □

Lemma IV.2. *Assume assumption IV.1(a),(b) holds. Then* $\mathbb{P}(\Omega_{2,m}) \geq 1-\exp(-t)$.

This directly follows from the bernstein inequality (Lemma VI.8). □

4.6.3 Proof of Theorem IV.2

We first define some quantities and events that will be used in the proof.

For any class of individualized treatment rules B, define the quantities

$$\begin{aligned}I_n(B) = & \ \delta \sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)] + 2\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)] \\ & + \sup_{d_1,d_2\in B}\sqrt{E(f(d_1)-f(d_2))^2\frac{2t}{n}}+\frac{8bSt}{3n}, \\ \hat{I}_n(B) = & \ \delta \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)] \\ & + 10(1+\delta)\mathbb{E}_\xi \sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,R_i;d_1)-f(X_i,A_i,R_i;d_2)]\end{aligned}$$

$$
\begin{aligned}
&+ 2(1+\delta)\sqrt{\sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2\frac{t}{n}} + \frac{83(1+\delta)bSt}{3n},\\
\bar{I}_n(B) =& \Big[22(1+\delta)\sqrt{\frac{1}{n}}+\delta\Big]\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]\\
&+ (42+44\delta)\mathbb{E}\sup_{d_1,d_2\in B}|(E_n-E)[f(d_1)-f(d_2)]|\\
&+ [4+5\delta]\sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^2\frac{t}{n}} + \frac{(52+55\delta)bSt}{n},
\end{aligned}
$$

and the events

$$
\begin{aligned}
\Omega_3(B) =& \Big\{ \sup_{d_1,d_2\in B}(E-E_n)(f(d_1)-f(d_2)) \le 2\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)(f(d_1)-f(d_2))\\
&\quad + \sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^2\frac{2t}{n}} + \frac{8bSt}{3n}\Big\},\\
\Omega_4(B) =& \Big\{ \quad \mathbb{E}\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)]\\
&\le 2\mathbb{E}_\xi\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)] + \frac{2bSt}{n}\Big\}\\
\Omega_5(B) =& \Big\{ \sup_{d_1,d_2\in B}(E-E_n)(f(d_1)-f(d_2))^2 \le 2\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)(f(d_1)-f(d_2))^2\\
&\quad + \sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^4\frac{2t}{n}} + \frac{8b^2S^2t}{3n}\Big\},\\
\Omega_6(B) =& \Big\{ \sup_{d_1,d_2\in B}(E_n-E)(f(d_1)-f(d_2)) \le 2\mathbb{E}\sup_{d_1,d_2\in B}(E_n-E)(f(d_1)-f(d_2))\\
&\quad + \sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^2\frac{2t}{n}} + \frac{8bSt}{3n}\Big\}\\
\Omega_7(B) =& \Big\{ \quad \mathbb{E}_\xi\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)]\\
&\le 2\mathbb{E}\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)] + \frac{5bSt}{6n}\Big\},\\
\Omega_8(B) =& \Big\{ \sup_{d_1,d_2\in B}(E_n-E)(f(d_1)-f(d_2))^2 \le 2\mathbb{E}\sup_{d_1,d_2\in B}(E_n-E)(f(d_1)-f(d_2))^2
\end{aligned}
$$

$$+\sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^4\frac{2t}{n}}+\frac{8b^2S^2t}{3n}\Big\}.$$

For each model $m = 1,\ldots,M_n$ and any $s \geq 0$ and $r > 0$, define the set of individualized treatment rules

$$B_m(s) =\{d(X)\in \arg\max_a Q(X,a) : L(Q)-L(Q_m^*)] \leq s, Q\in\mathcal{Q}_m\}$$

and the quantities

$$\begin{aligned}\eta_m(r) =&256b^3\Big(\mathbb{E}\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r}\frac{1}{n}\sum_{i=1}^n\xi_iQ(X_i,A_i)+\frac{10bt}{3n}\Big),\\ \bar{\eta}_m(r) =&256b^3\Big(\frac{5}{3}\mathbb{E}\sup_{Q\in\mathcal{Q}_m:b^2E(Q-Q_m^*)^2\leq 2650r}\frac{1}{n}\sum_{i=1}^n\xi_iQ(X_i,A_i)+\frac{82bt}{9n}\Big).\end{aligned}$$

By Assumption IV.2, lemma VI.11 and lemma VI.12, each of η_m and $\bar{\eta}_m(r)$ has unique positive fixed point. Let $r_{n,m}^*$ and $\bar{r}_{n,m}$ be the positive fixed points of $\eta_m(r)$ and $\bar{\eta}_m(r)$, respectively. Denote

$$s_{n,m}^* = \frac{44}{b^2}r_{n,m}^*+\frac{1752b^2t}{n} \text{ and } \bar{s}_{n,m} = \frac{308}{b^2}\bar{r}_{n,m}+\frac{12352b^2t}{n}.$$

Define the events

$$\begin{aligned}\Omega_{9,m} =&\Big\{\sup_{Q\in\mathcal{Q}_m}\big([L(Q)-L(Q_m^*)]-2E_n[(Y-Q)^2-(Y-Q_m^*)^2]\big)\\ &\leq\frac{44}{b^2}r_{n,m}^*+\frac{1752b^2t}{n}\Big\}\\ \Omega_{10,m} =&\Big\{\sup_{Q\in\mathcal{Q}_m}\big(E_n[(Y-Q)^2-(Y-Q_m^*)^2]-2[L(Q)-L(Q_m^*)]\big)\\ &\leq\frac{22}{b^2}r_{n,m}^*+\frac{920b^2t}{n}\Big\}\\ \Omega_{11,m}(r) =&\Big\{\mathbb{E}\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r}\frac{1}{n}\sum_{i=1}^n\xi_iQ(X_i,A_i)\end{aligned}$$

$$\leq \frac{4}{3}\mathbb{E}_\xi \sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i) + \frac{8bt}{3n}\Big\}$$

$$\Omega_{12,m}(r) = \Big\{\mathbb{E}_\xi \sup_{Q\in\mathcal{Q}_m:b^2E(Q-Q_m^*)^2\leq 2650r} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i)$$

$$\leq \frac{5}{4}\mathbb{E} \sup_{Q\in\mathcal{Q}_m:b^2E(Q-Q_m^*)^2\leq 2650r} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i) + \frac{7bt}{3n}\Big\}$$

$$\Omega_{13,m}(r) = \Big\{\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r} (E_n-E)(Q-Q_m^*)^2 \leq \sqrt{\frac{rt}{2n}} + \frac{52b^2t}{3n}$$

$$+ \frac{5}{2}\mathbb{E} \sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r} \frac{1}{n}\sum_{i=1}^n \xi_i (Q(X_i,A_i) - Q_m^*(X_i,A_i))^2\Big\}$$

$$\Omega_{14,m}(r) = \Big\{\sup_{l\in\mathcal{L}_m(r)} (E-E_n)l \leq \frac{5}{2}\mathbb{E}\sup_{l\in\mathcal{L}_m(r)} \frac{1}{n}\sum_{i=1}^n \xi_i l(X_i,A_i,Y_i) + \sqrt{\frac{rt}{2n}} + \frac{52b^2t}{3n}\Big\},$$

$$\text{where } \mathcal{L}_m(r) = \Big\{l = \frac{r(Q-Q_m^*)^2}{\max\{16b^2E(Q-Q_m^*)^2, r\}} : Q\in\mathcal{Q}_m\Big\}.$$

Now we start the proof. First note that for $m = 1,\ldots,M_n$,

$$pen_n(m) = \hat{I}_n(\hat{B}_m(\hat{s}_{n,m})) - \frac{83(1+\delta)bSt}{3n}.$$

Since $E_n[(Y-\hat{Q}_{n,m})^2 - (Y-Q_m^*)^2] \leq 0$, we have $\hat{d}_{n,m} \in B_m(s_{n,m}^*)$ on the event $\Omega_{9,m}$. In addition, by the definition of d_m^*, $d_m^* \in B_m(s_{n,m}^*)$. Thus

$$\begin{aligned}
&[(1-\delta)E - E_n](f(d_m^*) - f(\hat{d}_{n,m})) \\
&\leq \sup_{d_1,d_2\in B_m(s_{n,m}^*)} [(1-\delta)E - E_n](f(d_1) - f(d_2)) \\
&\leq \sup_{d_1,d_2\in B_m(s_{n,m}^*)} (E-E_n)(f(d_1)-f(d_2)) + \delta \sup_{d_1,d_2\in B_m(s_{n,m}^*)} E(f(d_1)-f(d_2)),
\end{aligned}$$

which is no larger than $I_n(B_m(s_{n,m}^*))$ on the event $\Omega_{3,m}(s_{n,m}^*)$.

By lemmas IV.3 and IV.6, we have $I_n(B_m(s_{n,m}^*)) \leq \hat{I}_n(B_m(s_{n,m}^*)) \leq \hat{I}_n(\hat{B}_m(\hat{s}_{n,m}))$ on the event $\Omega_3(B_m(s_{n,m}^*)) \cap \Omega_4(B_m(s_{n,m}^*)) \cap \Omega_5(B_m(s_{n,m}^*)) \cap \Omega_{9,m} \cap \Omega_{10,m} \cap \Omega_{11,m}(r_{n,m}^*) \cap$

$\Omega_{13,m}(r^*_{n,m}) \cap \Omega_{13,m}((1408 + \frac{657}{10})r^*_{n,m})$. This implies

$$pen_n(m) + \frac{83(1+\delta)bSt}{3n} \geq [(1-\delta)E - E_n](f(d^*_m) - f(\hat{d}_{n,m})).$$

Following the proof of Theorem IV.1, there exists a positive constant K_1 depending on $b, S, \alpha, C_2, \delta$ such that

$$\begin{aligned}(1-\delta)V(d^{opt}) - V(\hat{d}_{n,\hat{m}}) \leq \inf_{m=1,\dots,M_n} \Big\{ &(1+\delta)\Big[V(d^{opt}) - V(\hat{d}_{n,m})\Big] + pen_n(m) \\ &+ K_1\Big[u_n(\mathcal{F}_m)^{(1+\alpha)/(2+\alpha)} + \Big(\frac{t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{t}{n}\Big]\Big\}\end{aligned} \tag{4.19}$$

on the event $\cap_{m=1}^{M_n}(\Omega_{1,m} \cap \Omega_{2,m} \cap \Omega_3(B_m(s^*_{n,m})) \cap \Omega_4(B_m(s^*_{n,m})) \cap \Omega_5(B_m(s^*_{n,m})) \cap \Omega_{9,m} \cap \Omega_{10,m} \cap \Omega_{11,m}(r^*_{n,m}) \cap \Omega_{13,m}(r^*_{n,m}) \cap \Omega_{13,m}((1408 + \frac{657}{10})r^*_{n,m}))$.

Next, note that $\{f(d) : d \in B_m(\bar{s}_{n,m})\} \subset \mathcal{F}_m$. By Lemmas IV.4, IV.6 and IV.5, we have that, on the event $\Omega_6(B_m(\bar{s}_{n,m})) \cap \Omega_7(B_m(\bar{s}_{n,m})) \cap \Omega_8(B_m(\bar{s}_{n,m})) \cap \Omega_{9,m} \cap \Omega_{10,m} \cap \Omega_{12,m}(\bar{r}_{n,m}) \cap \Omega_{13,m}((1408 + \frac{657}{10})r^*_{n,m}) \cap \Omega_{14,m}(\bar{r}_{n,m})$,

$$\begin{aligned}&pen_n(m) \\ &\leq \hat{I}_n(B_m(\bar{s}_{n,m})) - \frac{83(1+\delta)bSt}{3n} \leq \bar{I}_n(B_m(\bar{s}_{n,m})) - \frac{83(1+\delta)bSt}{3n} \\ &\leq \Big[\frac{46\alpha + 51\alpha\delta + 2\delta}{2(1+\alpha)} + 22(1+\delta)\sqrt{\frac{1}{n}}\Big] \sup_{d \in B_m(\bar{s}_{n,m})} E[f(d^{opt}) - f(d)] \\ &\frac{768(21+22\delta)bSu_n(\mathcal{F}_m)}{(\log 2)^2} + \frac{(21+22\delta)(2+\alpha)}{1+\alpha}\Big(\frac{768C_2u_n(\mathcal{F}_m)}{(\log 2)^2}\Big)^{(1+\alpha)/(2+\alpha)} \\ &+ \frac{(4+5\delta)(2+\alpha)}{2(1+\alpha)}\Big(\frac{4C_2t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{(73+82\delta)bSt}{3n} + \frac{16(21+22\delta)}{n}\end{aligned} \tag{4.20}$$

By Assumption IV.3(a), the definition of $\bar{s}_{n,m}$ and Lemma IV.7, there exists a constant

K_4 depending on b such that

$$\begin{aligned}
&\sup_{d\in B_m(\bar{s}_{n,m})}[V(d^{opt})-V(d)]\\
\leq& C_1 \sup_{Q\in\mathcal{Q}_m:L(Q)-L(Q^*_m)\leq\bar{s}_{n,m}}\Big[L(Q)-L(Q^{opt})\Big]^{(1+\alpha)/(2+\alpha)}\\
\leq& C_1\Big[L(Q^*_m)^2-L(Q^{opt})^2+\bar{s}_{n,m}\Big]^{(1+\alpha)/(2+\alpha)}\\
\leq& C_1\Big[L(Q^*_m)-L(Q^{opt})+K_4\Big(u_n(\mathcal{Q}_m)+\frac{t+1}{n}\Big)\Big]^{(1+\alpha)/(2+\alpha)}. \qquad (4.21)
\end{aligned}$$

In addition, note that $(1+\delta)[V(d^{opt})-V(\hat{d}_{n,m})]\leq(1+\delta)\sup_{d\in B_m(\bar{s}_{n,m})}[V(d^{opt})-V(d)]$ on the event $\Omega_{9,m}$. Substituting (4.20) and (4.21) into (4.19) and using Lemma VI.2, we have

$$\begin{aligned}
&V(d^{opt})-V(\hat{d}_{n,\hat{m}})\\
\leq& K_2\inf_{m=1,\ldots,M_n}\Big\{\big[L(Q^*_m)-L(Q^{opt})\big]^{(1+\alpha)/(2+\alpha)}\\
&+u_n(\mathcal{F}_m)+u_n(\mathcal{F}_m)^{(1+\alpha)/(2+\alpha)}+u_n(\mathcal{Q}_m)+\Big(\frac{t}{n}\Big)^{(1+\alpha)/(2+\alpha)}+\frac{t+1}{n}\Big\}
\end{aligned}$$

for a sufficiently large constant K_2 depending on b,S,α,C_1,C_2 and δ on the event $\cap_{m=1}^{M_n}(\Omega_{1,m}\cap\Omega_{2,m}\cap\Omega_3(B_m(s^*_{n,m}))\cap\Omega_4(B_m(s^*_{n,m}))\cap\Omega_5(B_m(s^*_{n,m}))\cap\Omega_6(B_m(\bar{s}_{n,m}))\cap\Omega_7(B_m(\bar{s}_{n,m}))\cap\Omega_8(B_m(\bar{s}_{n,m}))\cap\Omega_{9,m}\cap\Omega_{10,m}\cap\Omega_{11,m}(r^*_{n,m})\cap\Omega_{12,m}(\bar{r}_{n,m})\cap\Omega_{13,m}(r^*_{n,m})\cap\Omega_{13,m}((1408+\frac{657}{10})r^*_{n,m}))\cap\Omega_{14,m}(\bar{r}_{n,m})$.

Taking $t=t+\log(15M_n)$, the conclusion follows from Lemma IV.8 and the union bound argument. □

Lemma IV.3. *Suppose Assumption IV.1(a), (b) hold. Then for any class of individualized treatment rules B, $I_n(B)\leq\hat{I}_n(B)$ on the event $\Omega_3(B)\cap\Omega_4(B)\cap\Omega_5(B)$.*

Proof. First note that on the event $\Omega_3(B)$,

$$\begin{aligned}
&\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]\\
&\leq \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)] + \sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]\\
&\leq \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)] + 2\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)(f(d_1)-f(d_2))\\
&\qquad\qquad + \sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^2\frac{2t}{n}} + \frac{8bSt}{3n}.
\end{aligned}$$

Thus

$$\begin{aligned}
I_n(B) \leq \delta \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)] &+ 2(1+\delta)\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]\\
&+ (1+\delta)\sqrt{\sup_{d_1,d_2\in B} E(f(d_1)-f(d_2))^2\frac{2t}{n}} + \frac{8(1+\delta)bSt}{3n} \qquad (4.22)
\end{aligned}$$

In addition, on the event $\Omega_4(B)$, we have

$$\begin{aligned}
&\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]\\
\leq& 2\mathbb{E}\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)]\\
\leq& 4\mathbb{E}_\xi\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,R_i;d_1)-f(X_i,A_i,R_i;d_2)] + \frac{4bSt}{n}, \qquad (4.23)
\end{aligned}$$

where the first inequality follows from the symmetrization inequality (6.3).

Furthermore, on the event $\Omega_5(B)$

$$\begin{aligned}
&\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2\\
&\leq \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2 + \sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]^2\\
&\leq \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2 + 2\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]^2
\end{aligned}$$

$$
\begin{aligned}
&+\sqrt{\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^4 \frac{2t}{n}}+\frac{8b^2S^2t}{3n} \\
\leq& \sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2+2\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]^2 \\
&+\frac{1}{2}\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2+\frac{11b^2S^2t}{3n}, \qquad (4.24)
\end{aligned}
$$

where the last inequality follows from the fact that $E[f(d_1)-f(d_2)]^4 \leq b^2S^2E[f(d_1)-f(d_2)]^2$ (by assumption IV.1(a) and (b)). By symmetrization inequality (6.3) and contraction inequality (6.1), we have on the event $\Omega_4(B)$,

$$
\begin{aligned}
&\mathbb{E}\sup_{d_1,d_2\in B}(E-E_n)[f(d_1)-f(d_2)]^2 \\
\leq& 2\mathbb{E}\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)]^2 \\
\leq& 4bS\mathbb{E}\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)] \\
\leq& 8bS\mathbb{E}_\xi\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)]+\frac{8b^2S^2t}{n}.
\end{aligned}
$$

This together with (4.24) implies that

$$
\begin{aligned}
&\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2 \\
\leq& 2\sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2 \\
&+32bS\mathbb{E}_\xi\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{i=1}^n \xi_i[f(X_i,A_i,Y_i;d_1)-f(X_i,A_i,Y_i;d_2)]+\frac{118b^2S^2t}{3n}.
\end{aligned}
$$

By Lemma VI.2, we have on the event $\Omega_4(B)\cap\Omega_5(B)$,

$$
\begin{aligned}
&\sqrt{\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2 \frac{2t}{n}} \\
\leq& 2\sqrt{\sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2 \frac{t}{n}}
\end{aligned}
$$

$$+ 2\mathbb{E}_\xi \sup_{d_1,d_2\in B} \frac{1}{n}\sum_{1=i}^{n} \xi_i[f(X_i, A_i, Y_i; d_1) - f(X_i, A_i, Y_i; d_2)] + \frac{17bSt}{n}. \tag{4.25}$$

The result follows by substituting (4.23) and (4.25) into (4.22). □

Lemma IV.4. *Suppose Assumption IV.1(a), (b) holds. Then for any class of individualized treatment rules B, $\hat{I}(B) \leq \bar{I}(B)$ on the event $\Omega_6(B) \cap \Omega_7(B) \cap \Omega_8(B)$.*

Proof. First on the event $\Omega_6(B)$,

$$\begin{aligned}
&\sup_{d_1,d_2\in B} E_n[f(d_1) - f(d_2)] \\
\leq &\sup_{d_1,d_2\in B} E[f(d_1) - f(d_2)] + \sup_{d_1,d_2\in B} (E_n - E)[f(d_1) - f(d_2)] \\
\leq &\sup_{d_1,d_2\in B} E[f(d_1) - f(d_2)] + 2\mathbb{E} \sup_{d_1,d_2\in B} (E_n - E)(f(d_1) - f(d_2)) \\
&\qquad + \sqrt{\sup_{d_1,d_2\in B} E(f(d_1) - f(d_2))^2 \frac{2t}{n}} + \frac{8bSt}{3n}.
\end{aligned} \tag{4.26}$$

In addition, on the event $\Omega_7(B)$,

$$\begin{aligned}
&\mathbb{E}_\xi \sup_{d_1,d_2\in B} \frac{1}{n}\sum_{i=1}^{n} \xi_i[f(X_i, A_i, R_i; d_1) - f(X_i, A_i, R_i; d_2)] \\
\leq &2\mathbb{E} \sup_{d_1,d_2\in B} \frac{1}{n}\sum_{i=1}^{n} \xi_i[f(X_i, A_i, Y_i; d_1) - f(X_i, A_i, Y_i; d_2)] + \frac{5bSt}{6n} \\
\leq &2\mathbb{E} \sup_{d_1,d_2\in B} \frac{1}{n}\sum_{i=1}^{n} \xi_i[f(X_i, A_i, Y_i; d_1) - f(X_i, A_i, Y_i; d_2) - E(f(d_1) - f(d_2))] \\
&+ 2\mathbb{E} \sup_{d_1,d_2\in B} \frac{1}{n}\sum_{i=1}^{n} \xi_i E[f(d_1) - f(d_2)] + \frac{5bSt}{6n} \\
\leq &4\mathbb{E} \sup_{d_1,d_2\in B} |(E_n - E)(f(d_1) - f(d_2))]| \\
&+ 2\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n} \xi_i\Big| \sup_{d_1,d_2\in B} |E[f(d_1) - f(d_2)]| + \frac{5bSt}{6n}
\end{aligned}$$

$$\leq 4\mathbb{E}\sup_{d_1,d_2\in B}|(E_n-E)(f(d_1)-f(d_2))|+2\sqrt{\frac{1}{n}}\sup_{d_1,d_2\in B}E[f(d_1)-f(d_2)]+\frac{5bSt}{6n}, \tag{4.27}$$

where the third inequality follows from the desymmetrization inequality (6.2) and the last inequality follows from the fact that $\mathbb{E}\left|\sum_{i=1}^n \xi_i/n\right| \leq \left[\mathbb{E}(\sum_{i=1}^n \xi_i/n)^2\right]^{1/2} = (1/n)^{1/2}$.

By assumption IV.1(a) and (b), $|(f(d_1)-f(d_2))^2| \leq b^2S^2$. Thus on the event $\Omega_8(B)$

$$\begin{aligned}
&\sup_{d_1,d_2\in B} E_n[f(d_1)-f(d_2)]^2\\
&\leq \sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2+\sup_{d_1,d_2\in B}(E_n-E)[f(d_1)-f(d_2)]^2\\
&\leq \sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2+2\mathbb{E}\sup_{d_1,d_2\in B}(E_n-E)(f(d_1)-f(d_2))^2\\
&\qquad\quad+\sqrt{\sup_{d_1,d_2\in B}E(f(d_1)-f(d_2))^4\frac{2t}{n}}+\frac{8b^2S^2t}{3n}\\
&\leq \frac{5}{2}\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2+2\mathbb{E}\sup_{d_1,d_2\in B}(E_n-E)(f(d_1)-f(d_2))^2+\frac{3b^2S^2t}{n}\\
&\leq \frac{5}{2}\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2+\frac{3b^2S^2t}{n}\\
&\quad+8bS\mathbb{E}\sup_{d_1,d_2\in B}\frac{1}{n}\sum_{1=i}^{n}\xi_i[f(X_i,A_i,R_i;d_1)-f(X_i,A_i,R_i;d_2)]\\
&\leq \frac{5}{2}\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2+\frac{3b^2S^2t}{n}\\
&\quad+16bS\mathbb{E}\sup_{d_1,d_2\in B}|(E_n-E)[f(d_1)-f(d_2)]|+8bS\sqrt{\frac{1}{n}}\sup_{d_1,d_2\in B}E[f(d_1)-f(d_2)],
\end{aligned}$$

where the last two inequalities follows from the desymmetrization-symmetrization inequality (6.2), (6.3) and the contraction inequality (6.1). This implies that

$$\sqrt{\sup_{d_1,d_2\in B}E_n[f(d_1)-f(d_2)]^2\frac{t}{n}}$$

$$
\begin{aligned}
\leq & \mathbb{E} \sup_{d_1,d_2\in B} |(E_n - E)[f(d_1) - f(d_2)]| + \sqrt{\sup_{d_1,d_2\in B} E[f(d_1)-f(d_2)]^2 \frac{5t}{2n}} \\
& + \sqrt{\frac{1}{n}} \sup_{d_1,d_2\in B} E[f(d_1) - f(d_2)] + \frac{(6+\sqrt{3})bSt}{n} \qquad (4.28)
\end{aligned}
$$

The result follows by substituting (4.26), (4.27) and (4.28) into $\hat{I}_n(B)$. □

Lemma IV.5. *Suppose Assumptions IV.1(a), (b) and IV.3(b) hold. For any non-stochastic class of individualized treatment rules B, let $\mathcal{F} = \{f(d) : d \in B\}$. Then*

$$
\begin{aligned}
\bar{I}_n(B) \leq & \Big[\frac{46\alpha + 51\alpha\delta + 2\delta}{2(1+\alpha)} + 22(1+\delta)\sqrt{\frac{1}{n}}\Big] \sup_{d\in B} E[f(d^{opt}) - f(d)] \\
& \frac{768(21+22\delta)bSu_n(\mathcal{F})}{(\log 2)^2} + \frac{(21+22\delta)(2+\alpha)}{1+\alpha}\Big(\frac{768C_2u_n(\mathcal{F})}{(\log 2)^2}\Big)^{(1+\alpha)/(2+\alpha)} \\
& + \frac{(4+5\delta)(2+\alpha)}{2(1+\alpha)}\Big(\frac{4C_2t}{n}\Big)^{(1+\alpha)/(2+\alpha)} + \frac{(52+55\delta)bSt}{n} + \frac{16(21+22\delta)}{n}.
\end{aligned}
$$

Proof. Define the function class $\mathcal{G}(B) = \{g = f(d_1) - f(d_2) : d_1, d_2 \in B\}$. Let $\mathcal{G}_0(B)$ be a $4/n$-net in $L_1(P_n)$ over $\mathcal{G}(B)$. The cardinality of $\mathcal{G}_0(B)$ can be chosen equal to $N(2/n, \mathcal{G}(B), L_1(P_n))$. Then by symmetrization inequality (6.3) and the definition of $\mathcal{G}_{m,0}$,

$$
\begin{aligned}
\mathbb{E} \sup_{d_1,d_2\in B} |(E - E_n)[f(d_1) - f(d_2)]| \leq & 2\mathbb{E} \sup_{g\in\mathcal{G}(B)} \Big|\frac{1}{n}\sum_{i=1}^{n} \xi_i g(X_i, A_i, Y_i)\Big| \\
\leq & 2\mathbb{E} \sup_{g\in\mathcal{G}_0(B)} \Big|\frac{1}{n}\sum_{i=1}^{n} \xi_i g(X_i, A_i, Y_i)\Big| + \frac{8}{n}. \qquad (4.29)
\end{aligned}
$$

Next, we can verify that $N(2/n, \mathcal{G}(B), L_1(P_n)) = N(1/n, \mathcal{F}, L_1(P_n))$. Thus $\mathbb{E}\log[N(2/n, \mathcal{G}(B), L_1(P_n)) + 1]/n = u_n(\mathcal{F})$. Following the same argument as that in the proof of Lemma IV.1, we have

$$
\mathbb{E} \sup_{g\in\mathcal{G}_0(B)} \Big|\frac{1}{n}\sum_{i=1}^{n} \xi_i g(X_i, A_i, Y_i)\Big| \leq \frac{192bSu_n(\mathcal{F})}{(\log 2)^2} + \frac{4\sqrt{3}}{\log 2}\sqrt{u_n(\mathcal{F}) \sup_{g\in\mathcal{G}_0(B)} Eg^2}.
$$

This together with 4.29 implies that

$$
\begin{aligned}
&\mathbb{E}\sup_{d_1,d_2\in B}|(E-E_n)[f(d_1)-f(d_2)]| \\
\leq&\frac{384bSu_n(\mathcal{F})}{(\log 2)^2}+\frac{8\sqrt{3}}{\log 2}\sqrt{u_n(\mathcal{F})\sup_{d_1,d_2\in B}E[f(d_1)-f(d_2)]^2}+\frac{8}{n} \\
\leq&\frac{384bSu_n(\mathcal{F})}{(\log 2)^2}+\frac{16\sqrt{3}}{\log 2}\sqrt{C_2u_n(\mathcal{F})\sup_{d\in B}E[f(d^{opt})-f(d)]^{\alpha/(1+\alpha)}}+\frac{8}{n} \\
\leq&\frac{\alpha}{2(1+\alpha)}\sup_{d\in B}E[f(d^{opt})-f(d)]+\frac{384bSu_n(\mathcal{F})}{(\log 2)^2} \\
&+\frac{2+\alpha}{2(1+\alpha)}\Big(\frac{768C_2u_n(\mathcal{F})}{(\log 2)^2}\Big)^{(1+\alpha)/(2+\alpha)}+\frac{8}{n}
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\sqrt{\sup_{d_1,d_2\in B_m(s)}E(f(d_1)-f(d_2))^2\frac{t}{n}} \\
\leq&\frac{\alpha}{2(1+\alpha)}\sup_{d\in B_m(s)}E[f(d^{opt})-f(d)]+\frac{2+\alpha}{2(1+\alpha)}\Big(\frac{4C_2t}{n}\Big)^{(1+\alpha)/(2+\alpha)}.
\end{aligned}
$$

The result follows by substituting the above two inequalities into $\bar{I}_n(B)$. □

Lemma IV.6. *Suppose assumptions IV.1(b), (c) and IV.2 hold. Then* $\hat{d}_{n,m}\in B_m(s^*_{n,m})\subset\hat{B}_m(\hat{s}_{n,m})\subset\bar{B}_m(\bar{s}_{n,m})$ *on the event* $\Omega_{9,m}\cap\Omega_{10,m}\cap\Omega_{11,m}(r^*_{n,m})\cap\Omega_{12,m}(\bar{r}_{n,m})\cap\Omega_{13,m}(r^*_{n,m})\cap\Omega_{13,m}\big(\big(1408+\frac{657}{10}\big)r^*_{n,m}\big)\cap\Omega_{14,m}(\bar{r}_{n,m})$.

Proof. First, by the definition of $B_m(s)$ and $s^*_{n,m}$, $\hat{d}_{n,m}\in B_m(s^*_{n,m})$ on the event $\Omega_{9,m}$.

Next, suppose $r^*_{n,m}\leq\hat{r}_{n,m}\leq\bar{r}_{n,m}$.

For any $d\in B_m(s^*_{n,m})$, there is a $Q\in\mathcal{Q}_m$ such that $d(X)\in\arg\max_a Q(X,a)$ and $E[(Y-Q)^2-(Y-Q^*_m)^2]\leq s^*_{n,m}$. On the event $\Omega_{9,m}\cap\Omega_{10,m}$, we have

$$
\begin{aligned}
&E_n[(Y-Q)^2-(Y-\hat{Q}_m)^2] \\
=&E_n[(Y-Q)^2-(Y-Q^*_m)^2]+E_n[(Y-Q^*_m)^2-(Y-\hat{Q}_m)^2]
\end{aligned}
$$

$$\begin{aligned}
&\leq 2E[(Y-Q)^2-(Y-Q_m^*)^2]+\frac{22r_{n,m}^*}{b^2}+\frac{920b^2t}{n}+\frac{22r_{n,m}^*}{b^2}+\frac{876b^2t}{n}\\
&=\frac{132}{b^2}r_{n,m}^*+\frac{5300b^2t}{n}\leq \hat{s}_{n,m}.
\end{aligned}$$

Thus $B_m(s_{n,m}^*)\subset \hat{B}_m(\hat{s}_{n,m})$ if $r_{n,m}^*\leq \hat{r}_{n,m}$.

Similarly, for any $d\in \hat{B}_m(\hat{s}_{n,m})$, there is a $Q\in\mathcal{Q}_m$ such that $d(X)\in\arg\max_a Q(X,a)$ and $E[(Y-Q)^2-(Y-\hat{Q}_m)^2]\leq \hat{s}_{n,m}$. On the event $\Omega_{9,m}\cap\Omega_{10,m}$, we have

$$\begin{aligned}
E[(Y-Q)^2-(Y-Q_m^*)^2]\leq& 2E_n[(Y-Q)^2-(Y-Q_m^*)^2]+\frac{44r_{n,m}^*}{b^2}+\frac{1752b^2t}{n}\\
\leq& 2\hat{s}_{n,m}+\frac{44r_{n,m}^*}{b^2}+\frac{1752b^2t}{n}\leq \bar{s}_{n,m}.
\end{aligned}$$

Thus $\hat{B}_m(\hat{s}_{n,m})\subset \bar{B}_m(\bar{s}_{n,m})$ if $\hat{r}_{n,m}\leq \bar{r}_{n,m}$.

To show $r_{n,m}^*\leq \hat{r}_{n,m}\leq \bar{r}_{n,m}$, note that if

$$\{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r_{n,m}^*\}\subset\{Q\in\mathcal{Q}_m:b^2E_n(Q-\hat{Q}_{n,m})^2\leq 170r_{n,m}^*\}, \tag{4.30}$$

then on the event $\Omega_{11,m}(r_{n,m}^*)$,

$$\begin{aligned}
r_{n,m}^*=&\eta_m(r_{n,m}^*)\\
&\leq 256b^3\Big(\frac{4}{3}\mathbb{E}_\xi \sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r_{n,m}^*}\frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i)+\frac{6bt}{n}\Big)\\
&\leq \hat{\eta}_m(r_{n,m}^*).
\end{aligned}$$

By Lemmas VI.11 and VI.12, we have $r_{n,m}^*\leq \hat{r}_{n,m}$.

Similarly, if

$$\{Q\in\mathcal{Q}_m:b^2E_n(Q-\hat{Q}_{n,m})^2\leq 170\bar{r}_{n,m}\}\subset\{Q\in\mathcal{Q}_m:b^2E(Q-Q_m^*)\leq 2650\bar{r}_{n,m}\}, \tag{4.31}$$

then on the event $\Omega_{12,m}(\bar{r}_{n,m})$,

$$\begin{aligned}\hat{\eta}_m(\bar{r}_{n,m}) =&256b^3\Big(\frac{4}{3}\mathbb{E}_\xi \sup_{Q\in\mathcal{Q}_m: b^2E_n(Q-\hat{Q}_m)^2\leq 170\bar{r}_{n,m}} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i) + \frac{6bt}{n}\Big)\\ \leq&256b^3\Big(\frac{4}{3}\mathbb{E}_\xi \sup_{Q\in\mathcal{Q}_m: b^2E(Q-Q^*_m)^2\leq 2650\bar{r}_{n,m}} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i) + \frac{6bt}{n}\Big)\\ \leq&\bar{\eta}_m(\bar{r}_{n,m}) = \bar{r}_{n,m}\end{aligned}$$

By Lemmas VI.11 and VI.12, we have $\hat{r}_{n,m} \leq \bar{r}_{n,m}$.

In the following, we show that (4.30) and (4.31) hold on the event $\Omega_{9,m}\cap\Omega_{13,m}(r^*_{n,m})\cap \Omega_{13,m}\big((1408+\frac{657}{10})r^*_{n,m}\big)\cap\Omega_{14,m}(\bar{r}_{n,m})$.

Since $\eta_m(r)$ is a sub-root function (Lemma VI.12) and $r^*_{n,m}$ is the positive fixed point of $\eta_m(r)$, $r \geq \eta_m(r)$ if and only if $r \geq r^*_{n,m}$. Thus, for any $r \geq r^*_{n,m}$, on the event $\Omega_{13,m}(r)$, we have

$$\begin{aligned}&\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q^*_m)^2\leq r} E_n(Q-Q^*_m)^2\\ \leq&\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q^*_m)^2\leq r} E(Q-Q^*_m)^2 + \sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q^*_m)^2\leq r}(E_n-E)(Q-Q^*_m)^2\\ \leq&\frac{r}{16b^2}+\frac{5}{2}\mathbb{E}\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q^*_m)^2\leq r}\frac{1}{n}\sum_{i=1}^n\xi_i(Q(X_i,A_i)-Q^*_m(X_i,A_i))^2+\sqrt{\frac{rt}{2n}}+\frac{52b^2t}{3n}\\ \leq&\frac{r}{16b^2}+10b\mathbb{E}\sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q^*_m)^2\leq r}\frac{1}{n}\sum_{i=1}^n\xi_iQ(X_i,A_i)+\frac{r}{128b^2}+\frac{100b^2t}{3n}\\ \leq&\frac{7r}{64b^2},\end{aligned}$$

where the third inequality follows from the contraction inequality (6.1) and the assumption that $\sup_{Q\in\mathcal{Q}_m}\|Q\|_\infty \leq b$.

In addition, following the argument in the proof of lemma IV.9, on the event $\Omega_{9,m}$, we have

$$\begin{aligned}16b^2E(\hat{Q}_{n,m}-Q_m^*)^2 \leq & 32b^2\big[E(Y-\hat{Q}_{n,m})^2-E(Y-Q_m^*)^2\big] \\ \leq & 1408r_{n,m}^*+\frac{56064b^4t}{n} \leq \Big(1408+\frac{657}{10}\Big)r_{n,m}^*.\end{aligned}$$

This implies that on the event $\Omega_{9,m}\cap\Omega_{13,m}\big(\big(1408+\frac{657}{10}\big)r_{n,m}^*\big)$, for any $r \geq r_{n,m}^*$, we have

$$E_n(\hat{Q}_{n,m}-Q_m^*)^2 \leq \frac{7}{64}\Big(1408+\frac{657}{10}\Big)\frac{r_{n,m}^*}{b^2} \leq \Big(154+\frac{4599}{640}\Big)\frac{r}{b^2}.$$

Thus for any $r \geq r_{n,m}^*$ and $Q \in \mathcal{Q}_m$ such that $16b^2E(Q-Q_m^*)^2 \leq r$,

$$E_n(Q-\hat{Q}_{n,m})^2 \leq \left(\sqrt{E_n(Q-Q_m^*)^2}+\sqrt{E_n(Q_m^*-\hat{Q}_{n,m})^2}\right)^2 \leq \frac{170r}{b^2}$$

Setting $r = r_{n,m}^*$ in the above argument shows that (4.30) holds on the event $\Omega_{9,m}\cap\Omega_{13,m}(r_{n,m}^*)\cap\Omega_{13,m}\big(\big(1408+\frac{657}{10}\big)r_{n,m}^*\big)$.

Next, for any $r \geq r_{n,m}^*$ and $Q \in \mathcal{Q}_m$ such that $b^2E_n(Q-\hat{Q}_m)^2 \leq 170r$, on the event $\Omega_{9,m}\cap\Omega_{13,m}\big(\big(1408+\frac{657}{10}\big)r_{n,m}^*\big)$,

$$E_n(Q-Q_m^*)^2 \leq \left(\sqrt{E_n(Q-\hat{Q}_m)^2}+\sqrt{E_n(\hat{Q}_m-Q_m^*)^2}\right)^2 \leq \frac{1325r}{2b^2}.$$

Let $\mathcal{L}_m(r)$ be the class of functions defined in the event $\Omega_{14,m}(r)$. For each $l_0 \in \mathcal{L}_m(r)$, there is a $Q_0 \in \mathcal{Q}_m$, such that $l_0 = \frac{r(Q_0-Q_m^*)^2}{\max\{16b^2E(Q_0-Q_m^*)^2,r\}}$. If $16b^2E(Q_0-Q_m^*)^2 \leq r$, then $l_0 = (Q_0-Q_m^*)^2$. Otherwise,

$$l_0 = \alpha_{Q_0}^2(Q_0-Q_m^*)^2 = [\alpha_{Q_0}Q_0+(1-\alpha_{Q_0})Q_m^*-Q_m^*]^2 = (Q_1-Q_m^*)^2,$$

where $\alpha_{Q_0} = \sqrt{r/[16b^2E(Q_0-Q_m^*)^2]}$ and $Q_1 = \alpha_{Q_0}Q_0+(1-\alpha_{Q_0})Q_m^*$. Since $\mathcal{Q}_m$ is convex, $Q_1 \in \mathcal{Q}_m$. It is also easy to verify that $16b^2E(Q_1 - Q_m^*)^2 \leq r$.

Thus

$$
\begin{aligned}
&\mathbb{E} \sup_{l \in \mathcal{L}_m(r)} \frac{1}{n} \sum_{i=1}^{n} \xi_i l(X_i, A_i, Y_i) \\
\leq &\mathbb{E} \sup_{Q \in \mathcal{Q}_m : 16b^2 E(Q-Q_m^*)^2 \leq r} \frac{1}{n} \sum_{i=1}^{n} \xi_i [Q(X_i, A_i, Y_i) - Q_m^*(X_i, A_i, Y_i)]^2 \\
\leq &4b\mathbb{E} \sup_{Q \in \mathcal{Q}_m : 16b^2 E(Q-Q_m^*)^2 \leq r} \frac{1}{n} \sum_{i=1}^{n} \xi_i Q(X_i, A_i, Y_i),
\end{aligned}
$$

where the last inequality follows from the contraction inequality (6.1) and the symmetry of the Rademacher random variables.

Hence, for any $r \geq r_{n,m}^*$, on the event $\Omega_{14,m}(r)$,

$$
\begin{aligned}
&(E - E_n) \frac{r(Q - Q_m^*)^2}{\max\{16b^2 E(Q - Q_m^*)^2, r\}} \\
\leq &10b\mathbb{E} \sup_{Q \in \mathcal{Q}_m : 16b^2 E(Q-Q_m^*)^2 \leq r} \frac{1}{n} \sum_{i=1}^{n} \xi_i Q(X_i, A_i, Y_i) + \sqrt{\frac{rt}{2n}} + \frac{52b^2 t}{3n} \\
\leq &\frac{r}{128b^2} + \frac{5\eta_m(r)}{128b^2} \\
\leq &\frac{3r}{64b^2}.
\end{aligned}
$$

For any $Q \in \mathcal{Q}_m$ such that $b^2 E_n(Q - \hat{Q}_m)^2 \leq 170r$, if $16b^2 E(Q - Q_m^*)^2 > r$, the above inequality implies that

$$
E(Q - Q_m^*)^2 \leq 4E_n(Q - Q_m^*)^2 \leq \frac{2650r}{b^2}.
$$

Since $\eta_m(r) \leq \bar{\eta}_m(r)$ for any $r > 0$, $\bar{r}_{n,m} \geq r_{n,m}^*$. Setting $r = \bar{r}_{n,m}$ in the above argument shows that (4.31) holds on the event $\Omega_{9,m} \cap \Omega_{13,m}\big((1408 + \frac{657}{10})r_{n,m}^*\big) \cap \Omega_{14,m}(\bar{r}_{n,m})$.

This completes the proof. □

Lemma IV.7. *Assume assumptions IV.1(c) and IV.2 hold. Then there exists a positive constant K_4 depending on b, such that*

$$\bar{s}_{n,m} \leq K_4\Big(u_n(\mathcal{Q}_m) + \frac{t+1}{n}\Big).$$

Proof. Let $\mathcal{Q}_{m,0}$ be a $2/n$-net in $L_1(P_n)$ over $\mathcal{Q}_m$. The cardinality of $\mathcal{Q}_{m,0}$ can be chosen equal to $N(1/n, \mathcal{Q}_m, L_1(P_n))$. Then

$$\begin{aligned}
&\mathbb{E} \sup_{Q\in\mathcal{Q}_m : b^2E(Q-Q^*_m)^2\leq 2650r} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i, A_i)\\
=&\mathbb{E} \sup_{Q\in\mathcal{Q}_m : b^2E(Q-Q^*_m)^2\leq 2650r} \frac{1}{n}\sum_{i=1}^n \xi_i \big[Q(X_i, A_i) - Q^*_m(X_i, A_i)\big]\\
\leq&\mathbb{E} \sup_{Q\in\mathcal{Q}_{m,0} : b^2E(Q-Q^*_m)^2\leq 2650r} \Big|\frac{1}{n}\sum_{i=1}^n \xi_i \big[Q(X_i, A_i) - Q^*_m(X_i, A_i)\big]\Big| + \frac{2}{n}.
\end{aligned}$$

Following the same argument as that in the proof of Lemma IV.1, we can show that

$$\mathbb{E} \sup_{Q\in\mathcal{Q}_m : b^2E(Q-Q^*_m)^2\leq 2650r} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i, A_i) \leq \frac{200\sqrt{318u_n(\mathcal{Q}_m)r}}{(\log 2)b} + \frac{384bu_n(\mathcal{Q}_m)}{(\log 2)^2} + \frac{2}{n}.$$

Thus,

$$\bar{r}_{n,m} = \bar{\eta}_m(\bar{r}_{n,m}) \leq \frac{512}{3}\Big(\frac{500b^2\sqrt{318u_n(\mathcal{Q}_m)\bar{r}_{n,m}}}{\log 2} + \frac{960b^4u_n(\mathcal{Q}_m)}{(\log 2)^2} + \frac{5b^3}{n} + \frac{41b^4t}{3n}\Big).$$

The result follows by solving the above inequality for $\bar{r}_{n,m}$ and from the definition of $\bar{s}_{n,m}$. □

Lemma IV.8. *Suppose Assumptions IV.1 and IV.2 hold. Then for any nonstochastic class of individualized treatment rules B and nonstochastic positive quantity r,*

$$\mathbb{P}(\Omega_j(B)) \geq 1 - \exp(-t) \quad \textit{for } j = 3, \ldots, 8, \tag{4.32}$$

$$\mathbb{P}(\Omega_{9,m}) \geq 1 - \exp(-t), \quad \mathbb{P}(\Omega_{10,m}) \geq 1 - \exp(-t), \tag{4.33}$$

$$\textit{and}\quad \mathbb{P}(\Omega_{j,m}(r)) \geq 1 - \exp(-t) \quad \textit{for } j = 11, \ldots, 14. \tag{4.34}$$

Proof. (4.32) and (4.34) follow directly from Lemmas VI.9, VI.10 and the symmetrization inequality (6.3).

Now we prove (4.33). First by Lemma IV.9,

$$E[(Y-Q)^2 - (Y-Q_m^*)^2]^2 \leq 32b^2 E[(Y-Q)^2 - (Y-Q_m^*)^2].$$

In addition note that $|(Y-Q_1)+(Y-Q_2)| \leq 4b$ for any $Q_1, Q_2 \in \mathcal{Q}_m$. By lemma IV.9 and lemma VI.14,

$$\begin{aligned}
&32b^2\mathbb{E} \sup_{Q\in\mathcal{Q}_m:32b^2E[(Y-Q)^2-(Y-Q_m^*)^2]\leq r} \frac{1}{n}\sum_{i=1}^n \xi_i\big[(Y_i - Q(X_i,A_i))^2 - (Y_i - Q_m^*(X_i,A_i))^2\big] \\
=&32b^2\mathbb{E} \sup_{Q\in\mathcal{Q}_m:32b^2E[(Y-Q)^2-(Y-Q_m^*)^2]\leq r} \frac{1}{n}\sum_{i=1}^n \xi_i(Y_i - Q(X_i,A_i))^2 \\
\leq&32b^2\mathbb{E} \sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r} \frac{1}{n}\sum_{i=1}^n \xi_i(Y_i - Q(X_i,A_i))^2 \\
\leq&128b^3\mathbb{E} \sup_{Q\in\mathcal{Q}_m:16b^2E(Q-Q_m^*)^2\leq r} \frac{1}{n}\sum_{i=1}^n \xi_i Q(X_i,A_i) \\
\leq&\eta_m(r).
\end{aligned}$$

Since $\eta_m(r)$ is a subroot function, (4.33) follows from Lemma VI.13. □

Lemma IV.9. *Suppose Assumptions IV.1(b), (c) and IV.2 hold. Then*

$$E[(Y-Q)^2 - (Y-Q_m^*)^2]^2 \leq 32b^2 E[(Y-Q)^2 - (Y-Q_m^*)^2] \tag{4.35}$$

for any $Q \in \mathcal{Q}_m$.

(4.35) is also known as the bernstein condition. Similar to (4.11), this condition is the key to show that $L(\hat{Q}_{n,m}) - L(Q_m^*)$ converges to 0 at rate faster than $1/\sqrt{n}$.

Proof. On one hand, by assumption IV.1(b) and (c), we have

$$E[(Y-Q)^2-(Y-Q_m^*)^2]^2 = E[(2Y-Q-Q_m^*)^2(Q-Q_m^*)^2] \leq 16b^2E(Q-Q_m^*)^2.$$

On the other hand, assumption IV.2 implies that $(Q+Q_m^*)/2 \in \mathcal{Q}_m$ for any $Q \in \mathcal{Q}_m$. Thus

$$\begin{aligned} E(Y-Q)^2+E(Y-Q_m^*)^2 =& 2E\Big(Y-\frac{Q+Q_m^*}{2}\Big)^2+\frac{1}{2}E(Q-Q_m^*)^2 \\ \geq& 2E(Y-Q_m^*)^2+\frac{1}{2}E(Q-Q_m^*)^2, \end{aligned}$$

which implies $E(Q-Q_m^*)^2 \leq 2E[(Y-Q)^2-(Y-Q_m^*)^2]$. This completes the proof. $\square$

CHAPTER V

Future Work

This dissertation investigates l_1 penalization and model selection for decision making. In this chapter we briefly discuss several problems that worth further exploration in future.

5.1 Extension to multi-stage decision

So far we have investigated a one stage decision problem. However, it is evident that some diseases may require time-varying treatments. For example, individuals with a chronic disease often experience a waxing and waning course of illness. In these settings the goal is to construct a sequence of individualized treatment rules that tailor the type and dosage of treatment through time according to an individual's changing status. There is an abundance of statistical literature in this area (Thall et al., 2000, 2002; Murphy, 2003, 2005; Robins, 2004; Lunceford et al., 2002; Wahed and Tsiatis, 2006; van de Laan et al., 2005). In the following we briefly introduce the multistage decision problem.

Consider data from a sequentially randomized trial. The longitudinal data on each subject is of the form $\{X_1, A_1, \ldots, X_T, A_T, Y\}$, where T is the number of stages, $X_t \in \mathcal{X}_t$ includes patient's variables observed prior to the treatment at stage t, $A_t \in \mathcal{A}_t$

is the treatment at stage t for $t = 1, \ldots, T$, and Y is a one-dimensional primary outcome. Denote $\bar{X}_t = (X_1, \ldots, X_t)$ and $\bar{A}_t = (A_1, \ldots, A_t)$ for $t = 1, \ldots, T$. At stage t, an individualized treatment rule d_t will take patient's history prior to stage t treatment (i.e. $(\bar{X}_t, \bar{A}_{t-1})$) as input and output a treatment A_t. The sequence of rules, $\mathbf{d} = (d_1, \ldots, d_T)$, is called a *dynamic treatment regime* (Murphy, 2003).

We still use E to denote the expectations with respect to the distribution of $(X_1, A_1, \ldots, X_T, A_T, Y)$, where the stage t treatment is assigned according to randomization probability $p_t(\cdot|\bar{X}_t, \bar{A}_{t-1})$ for $t = 1, \ldots, T$. In addition, we use $E^{\mathbf{d}}$ to denote the distribution of $(X_1, A_1, \ldots, X_T, A_T, Y)$ where the sequence of treatments is assigned according to rule $\mathbf{d}$. The *Value* of $\mathbf{d}$, $V(\mathbf{d})$, is the expected primary outcome that would have been observed if the dynamic treatment regime $\mathbf{d}$ were used to recommend treatment sequence for the entire study population. It is easy to verify that

$$V(\mathbf{d}) = E^{\mathbf{d}}(Y) = E\left[\prod_{t=1}^{T} \frac{1_{A_t = d_t(\bar{X}_t, \bar{A}_{t-1})}}{p_t(A_t|\bar{X}_t, \bar{A}_{t-1})} Y\right].$$

Our goal is to construct a dynamic treatment regime that maximizes the Value. Such a regime is called an *optimal dynamic treatment regime* and is denoted by $\mathbf{d}^{opt}$.

Define the conditional mean functions

$$\begin{aligned} Q_T^{opt}(\bar{X}_T, \bar{A}_T) =& E(Y|\bar{X}_T, \bar{A}_T), \\ Q_t^{opt}(\bar{X}_t, \bar{A}_t) =& E\Big[\max_{a_{t+1}} Q_{t+1}^{opt}(\bar{X}_{t+1}, \bar{A}_t, a_{t+1})\big|\bar{X}_t, \bar{A}_t\Big] \text{ for } t = T-1, \ldots, 1. \end{aligned}$$

Using backwards induction (Murphy, 2003), an optimal dynamic treatment regime satisfies

$$d_t^{opt}(\bar{X}_t, \bar{A}_{t-1}) \in \arg\max_{a_t} Q_t^{opt}(\bar{X}_t, \bar{A}_t)$$

for $t = 1, \ldots, T$.

Based on the above argument, an intuitive approach is to estimate the conditional mean functions backwards (the stage t dependent variable is $\max_{a_{t+1}} \hat{Q}_{n,t+1}(\bar{X}_{t+1}, \bar{A}_{t+1})$ where $\hat{Q}_{n,t+1}$ is the estimator of Q_{t+1}^{opt}), and then estimate the decision rules by maximizing the estimated Q_t^{opt}'s. This approach is known as Q-learning and has been extensively studied in computer science literature (Watkins, 1989; Sutton and Barto, 1998; Ormoneit and Sen, 2002; Lagoudakis and Parr, 2003; Ernst et al., 2005; Murphy, 2005).

To justify the Q-learning approach, an inequality similar to (2.1), which measures the closeness between the excess Value and the associated excess prediction error, has been provided in Murphy (2005). More precisely, suppose there is some $S \geq 1$ such that $p_t(a_t|\bar{x}_t, \bar{a}_{t-1}) \geq S^{-1}$ for all $(\bar{x}_t, \bar{a}_t)$ pairs for $t = 1, \ldots, T$. For any dynamic treatment regime $\mathbf{d} = (d_1, \ldots, d_T)$ and any functions $\{Q_1, \ldots, Q_T\}$ such that $d_t(\bar{X}_t, \bar{A}_{t-1}) \in \arg\max_{a_t} Q_t(\bar{X}_t, \bar{A}_{t-1}, a_t)$, Murphy (2005) showed that

$$V(\mathbf{d}^{opt}) - V(\mathbf{d}) \leq \sum_{t=1}^{T} 2S^{t/2} \sqrt{E\big(\max_{a_{t+1}} Q_{t+1}^{opt} - Q_t\big)^2 - E\big(\max_{a_{t+1}} Q_{t+1}^{opt} - Q_t^{opt}\big)^2},$$

where $Q_{T+1}^{opt} \equiv Y$.

Following the same arguments as those in Chapter II, we can further improve the upper bound under a margin type condition.

Theorem V.1. *Suppose there is some $S \geq 1$ such that $p_t(a_t|\bar{x}_t, \bar{a}_{t-1}) \geq S^{-1}$ for all $(\bar{x}_t, \bar{a}_t)$ pairs for $t = 1, \ldots, T$. Assume there are some constants $C > 0$ and $\alpha \geq 0$ such that for any positive ϵ satisfying $C\epsilon^\alpha < 1$,*

$$\mathbb{P}\Big(\exists \bar{a}_{t-1} \in \bar{\mathcal{A}}_{t-1} \text{ s.t.} \\ \max_{a_t} Q_t^{opt}(\bar{X}_t, \bar{a}_{t-1}, a_t) - \max_{a_t \notin \arg\max_{a_t} Q_t^{opt}(\bar{X}_t, \bar{a}_{t-1}, a_t)} Q_t^{opt}(\bar{X}_t, \bar{a}_{t-1}, a_t) \leq \epsilon\Big) \leq C\epsilon^\alpha$$

for $t = 1, \ldots, T$. Then for any dynamic treatment regime $\mathbf{d} = (d_1, \ldots, d_T)$ and any *functions* $\{Q_1, \ldots, Q_T\}$ *such that* $d_t(\bar{X}_t, \bar{A}_{t-1}) \in \arg\max_{a_t} Q_t(\bar{X}_t, \bar{A}_{t-1}, a_t)$,

$$V(\mathbf{d}^{opt}) - V(\mathbf{d}) \leq \sum_{t=1}^{T} C_{1,t} \left[E\big(\max_{a_{t+1}} Q_{t+1}^{opt} - Q_t\big)^2 - E\big(\max_{a_{t+1}} Q_{t+1}^{opt} - Q_t^{opt}\big)^2 \right]^{(1+\alpha)/(2+\alpha)},$$

where $C_{1,t} = [2^{2+3\alpha} S^{(1+\alpha)t} C]^{1/(2+\alpha)}$ *and* $Q_{T+1}^{opt} \equiv Y$.

The proof is similar to that of Theorem II.1 and thus is omitted.

To extend the l_1-PLS based method described in Chapter III to the multi-stage setting, we can approximate each conditional mean function Q_t^{opt} by a linear model, and then estimate Q_t^{opt} using least squares with an l_1 penalty. We can obtain a high probability bound for the excess prediction error at the last stage using the same techniques as that in the one-stage decision problem. However, the performance of the l_1-PLS at stages prior to the last stage is not clear and worth further investigation.

Now we consider the extension of the step-wise model selection as described in Chapter IV to the multi-stage setting. To develop a penalty with margin adaptive rate of convergence, we need to show that the difference between the prediction error of the empirical quadratic risk minimizer and the prediction error of the quadratic risk minimizer is upper bounded by a quantity which converges to zero at the desired rate of convergence at each stage. However, since the dependent variable at each stage prior to the last is estimated from the whole training set, existing methods can not be directly applied to construct such an upper bound. An interesting future research direction is to develop new techniques so as to obtain a fast rate of convergence of the prediction error in this setting.

5.2 Efficient estimation

In the previous chapters, we considered the use of model selection and penalization techniques to improve the quality of the estimated individualized treatment rule (dynamic treatment regime). In fact, those methods could be further improved by considering efficient estimation. In this section, we will discuss issues about improving estimation efficiency. We will illustrate this problem in the multi-stage setting.

5.2.1 An efficient estimator of Value

For any dynamic treatment regime $\mathbf{d} = (d_1, \ldots, d_T)$, define the weights

$$W_t(\bar{X}_t, \bar{A}_t; \mathbf{d}) = \prod_{s=1}^{t} \frac{1_{A_s = d_s(\bar{X}_s, \bar{A}_{s-1})}}{p_s(A_s | \bar{X}_s, \bar{A}_{s-1})} \quad \text{for } t = 1, \ldots, T.$$

It is easy to see that $E_n[W_T(\bar{X}_T, \bar{A}_T; \mathbf{d})Y]$ is an unbiased estimator of $V(\mathbf{d})$. However, this estimator is not efficient. To achieve the goal of estimating the optimal dynamic treatment regime, we wish to use a more efficient estimator of $V(\mathbf{d})$ in the future.

A doubly robust estimator of $V(\mathbf{d})$ has been provided in Murphy et al. (2001) by solving the following estimating equation for μ.

$$\begin{aligned} 0 = E_n\Big[& W_T\Big(Y - \mu\Big) - \sum_{t=1}^{T} W_t(\bar{X}_t, \bar{A}_t)\left(z_t(\bar{X}_t, \bar{A}_t) - \mu\right) \\ & + \sum_{t=1}^{T} \sum_{a_t \in \mathcal{A}_t} p_t(a_t | \bar{X}_t, \bar{A}_{t-1}) W_t(\bar{X}_t, \bar{A}_{t-1}, a_t)\left(z_t(\bar{X}_t, \bar{A}_{t-1}, a_t) - \mu\right)\Big] \end{aligned} \tag{5.1}$$

where $z_T(\bar{X}_T, \bar{A}_T) = E[Y | \bar{X}_T, \bar{A}_T]$ and
$z_t(\bar{X}_t, \bar{A}_t) = E\big[\sum_{a_{t+1} \in \mathcal{A}_{t+1}} 1_{a_{t+1} = d_{t+1}(\bar{X}_{t+1}, \bar{A}_t)} z_{t+1}(\bar{X}_{t+1}, \bar{A}_t, a_{t+1}) | \bar{X}_t, \bar{A}_t\big]$ for $t = T - 1, \ldots, 1$.

Note that z_t's are unknown functions and need to estimated. One can parameterize $\{z_t : t = 1, \ldots, T\}$ with vector parameter γ and set

$$E_n\Big[\Big(Y - z_T(\bar{X}_T, \bar{A}_T; \gamma)\Big)\frac{\partial z_T(\bar{X}_T, \bar{A}_T; \gamma)}{\partial \gamma} + \sum_{t=1}^{T-1} \frac{\partial z_t(\bar{X}_t, \bar{A}_t; \gamma)}{\partial \gamma}$$

$$\times \Big(\sum_{a_{t+1} \in \mathcal{A}_{t+1}} 1_{a_{t+1}=d_{t+1}(\bar{X}_{t+1},\bar{A}_t)} z_{t+1}(\bar{X}_{t+1}, \bar{A}_t, a_{t+1}; \gamma) - z_t(\bar{X}_t, \bar{A}_t; \gamma)\Big)\Big]$$

to 0 to get $\hat{\gamma}$.

The use of (5.1) leads to a consistent estimator of $V(\mathbf{d})$ even if models for z_t's are incorrect. This property holds because the randomization probabilities $p_1, \ldots, p_T$ are known. In addition, the resulting estimator is also efficient if z_t's are parameterized correctly.

5.2.2 Efficient regression

In this dissertation, we estimated the entire conditional mean functions Q_t^{opt}'s. however, it turns out that only part of each condition mean function is relevant to the construction of decision rules. To see this point, define the time-t advantage

$$\upsilon_t(\bar{X}_t, \bar{A}_t) = Q_t^{opt}(\bar{X}_t, \bar{A}_t) - \max_{a_t} Q_t^{opt}(\bar{X}_t, \bar{A}_{t-1}, a_t).$$

Then

$$Q_t^{opt}(\bar{X}_t, \bar{A}_t) = \upsilon_t(\bar{X}_t, \bar{A}_t) + \max_{a_t} Q_t^{opt}(\bar{X}_t, \bar{A}_{t-1}, a_t),$$

where only the first term υ_t contains A_t. Thus we only need to model the advantage functions υ_t's instead of modeling Q_t^{opt}'s.

This approach was first proposed in Murphy (2003), where an estimation procedure based on the least square characterization of the advantage functions is provided. Robins (2004) gave a refined estimation equation to gain efficiency. This is the so-called efficient A-learning in Almirall et al. (August, 2005). In the following, we will describe the efficient estimation procedure.

We parameterize $\{\upsilon_t, t = 1, \ldots, T\}$ with vector parameter θ and denote the parameterization by $\{\upsilon_t(\bar{X}_t, \bar{A}_t; \theta), t = 1, \ldots, T\}$. Define $H_t(\bar{X}_T, \bar{A}_T, Y) = Y -$

$\sum_{s=t}^{T} \upsilon_s(\bar{X}_t, \bar{A}_t)$ for $t = 1, \ldots, T$. Assume that

$$Var(H_t|\bar{X}_t, \bar{A}_t) = Var(H_t|\bar{X}_t, \bar{A}_{t-1}) \quad \text{denoted as } \sigma_t(\bar{X}_t, \bar{A}_{t-1})$$

for $t = 1, \ldots, T$.

A doubly robust estimator of θ can be found by solving the following estimating equations

$$\begin{aligned} 0 = E_n\Big(\sum_{t=1}^{T} & \big[H_t(\theta) - E(H_t(\theta)|\bar{X}_t, \bar{A}_{t-1})\big]\, \sigma_t(\bar{X}_t, \bar{A}_{t-1})^{-1} \\ & \times \big[E(\nabla_\theta H_t(\theta)|\bar{X}_t, \bar{A}_t) - E(\nabla_\theta H_t(\theta)|\bar{X}_t, \bar{A}_{t-1})\big] \Big). \end{aligned}$$

Again, this estimator is efficient if $E(H_t(\theta)|\bar{X}_t, \bar{A}_{t-1})$'s, $E(\nabla_\theta H_t(\theta)|\bar{X}_t, \bar{A}_t)$'s and $\sigma_t(\bar{X}_t, \bar{A}_{t-1})$'s are modeled correctly.

CHAPTER VI

Tools

This chapter contains a collection of results that is needed in the proofs in Chapters III and IV.

We start with several basic inequalities.

Lemma VI.1. ***(Weighted AM-GM inequality)***

For any $x, y > 0$ and $p, q \geq 0$ such that $p + q = 1$,

$$x^p y^q \leq px + qy.$$

Lemma VI.2. *For any $x, y \geq 0$ and $\beta \in [1/2, 1]$,*

$$2^{1-2\beta}(x+y)^\beta \leq x^\beta + y^\beta \leq 2^{1-\beta}(x+y)^\beta$$

and for any $\alpha > 0$,

$$2\sqrt{xy} \leq \alpha x + \frac{y}{\alpha}.$$

Lemma VI.3. ***(Cauchy-Schwarz inequality)***

For random variables X and Y,

$$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2).$$

Let $O \in \mathcal{O}$ be a random variable distributed according to P ($O = (X, A, Y)$ in the main body of the dissertation). Let $O_1, \ldots, O_n$ be independent copies of O and P_n be the empirical measure supported on $(O_1, \ldots, O_n)$. We use E and E_n to denote the expectation with respect to P and P_n, respectively.

Let $(\mathcal{F}, \|\cdot\|)$ be a subset of a normed space of real-valued functions on $\mathcal{O}$. In the following we will always assume suprema of empirical processes (i.e. quantities of the form $\sup_{f \in \mathcal{F}}(E_n - E)f$) are measurable. In other words, we assume that the class $\mathcal{F}$ and the distribution P satisfy appropriate (mild) conditions for measurability of this supremum (see Pollard (1984) and Massart (2003) for the conditions).

Definition VI.1. (**Covering number;** van der Vaart and Wellner 1996)

The ϵ-covering number of $(\mathcal{F}, \|\cdot\|)$, denoted by $N(\epsilon, \mathcal{F}, \|\cdot\|)$, is the minimal number of balls $\{f : \|f - f_0\| < \epsilon\}$ of radius ϵ needed to cover the set $\mathcal{F}$. The centers of the balls need not belong to $\mathcal{F}$, but they should have finite norms.

Let $\mathcal{C}$ a collection of subsets of $\mathcal{O}$. For any collection of points $\{o_1, ..., o_n\}$ in a set $\mathcal{O}$, we say that $\mathcal{C}$ picks out a certain subset of $\{o_1, ..., o_n\}$ if this subset can be formed as $C \cap \{o_1, ..., o_n\}$ for a $C \in \mathcal{C}$. The collection $\mathcal{C}$ is said to *shatter* $\{o_1, ..., o_n\}$ if all of the 2^n possible subsets of $\{o_1, ..., o_n\}$ can be picked out in this manner. The *VC-index* of the class $\mathcal{C}$, $vc(\mathcal{C})$, is the smallest n for which no set of size n is shattered by $\mathcal{C}$. $\mathcal{C}$ is called a *VC-class* if its index is finite.

The subgraph of a function $f : \mathcal{O} \to \mathbb{R}$ is the subset of $\mathcal{O} \times \mathbb{R}$ given by

$$\{(o, t) : t < f(o)\}.$$

A collection $\mathcal{F}$ of measurable real-valued functions on the sample space $\mathcal{O}$ is called a *VC-subgraph class* or *VC-class*, if the collection of all subgraphs of functions in $\mathcal{F}$ forms a VC-class of sets in $\mathcal{O} \times \mathbb{R}$. We use $vc(\mathcal{F})$ to denote the VC-index of the set of subgraphs of functions in $\mathcal{F}$.

A connection between covering numbers and VC-index is given in the following lemma.

Lemma VI.4. *(Haussler, 1992)*

Let $\mathcal{F}$ be a VC-class of functions on $\mathcal{O}$ with $\|f\|_\infty \le b$ for all $f \in \mathcal{F}$ for some $b > 0$. Then for any $\epsilon > 0$,

$$N(\epsilon, \mathcal{F}, L_1(P_n)) \le 2\Big(\frac{2eb}{\epsilon}\Big)^{2(vc(\mathcal{F})-1)}.$$

Lemma VI.5. *(**Preservation properties of VC-class;** Kosorok 2008)*

Let $\mathcal{F}$ and $\mathcal{G}$ be VC classes of functions on $\mathcal{O}$ with VC-indices $vc(\mathcal{F})$ and $vc(\mathcal{G})$, respectively. Let $g : \mathcal{O} \to \mathbb{R}$, $\phi : \mathbb{R} \to \mathbb{R}$ be fixed functions. Then

(a) $\mathcal{F} \wedge \mathcal{G} \equiv \{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is a VC class with index $\le vc(\mathcal{F}) + vc(\mathcal{G}) - 1$;

(b) $\mathcal{F} \vee \mathcal{G}$ is a VC class with index $\le vc(\mathcal{F}) + vc(\mathcal{G}) - 1$;

(c) $\{\mathcal{F} > 0\} \equiv \big\{\{f > 0\} : f \in \mathcal{F}\big\}$ is a VC-class of sets with index $vc(\mathcal{F})$;

(d) $-\mathcal{F}$ is a VC class with index $vc(\mathcal{F})$;

(e) $\mathcal{F} + g \equiv \{f + g : f \in \mathcal{F}\}$ is a VC class with index $vc(\mathcal{F})$;

(f) $\mathcal{F} \cdot g \equiv \{f \cdot g : f \in \mathcal{F}\}$ is a VC class with index $\le 2vc(\mathcal{F}) - 1$;

(g) $\phi \circ \mathcal{F} \equiv \{\phi \circ f : f \in \mathcal{F}\}$ is a VC class with index $\le vc(\mathcal{F})$ for monotone ϕ.

Definition VI.2. (**Orlicz norm;** van der Vaart and Wellner 1996)

Let ψ be a nondecreasing, convex function with $\psi(0) = 0$. Then the Orlicz norm $\|O\|_\psi$ is defined as

$$\|O\|_\psi = \inf\{c > 0 : E\psi(|O|/c) \le 1\}.$$

Let $\psi_p(o) = \exp(o^p) - 1$ for $p \ge 1$. Then

$$E|O|^p \le p!\|O\|_{\psi_1},$$

$$\|O\|_{\psi_p} \leq \|O\|_{\psi_q}(\log 2)^{1/q-1/p} \quad \text{for} \quad p \leq q.$$

Lemma VI.6. *(**Maximal inequality;** van der Vaart and Wellner 1996)*

For any random variables $O_1, \dots, O_m$, there exists a constant K depending only on ψ_p such that

$$|| \max_{1\leq i \leq m} O_i||_{\psi_p} \leq K\psi_p^{-1}(m) \max_i ||X_i||_{\psi_p}.$$

Lemma VI.7. *(**Hoeffding's inequality;** van der Vaart and Wellner 1996)*

Let $a_1, \dots, a_n$ be constants and $\xi_1, \dots, \xi_n$ be independent Rademacher random variables. Then

$$P\left(\left|\sum_{i=1}^n \xi_i a_i\right| > t\right) \leq 2\exp\left(-\frac{1}{2}t^2/\|a\|^2\right),$$

for the Euclidean norm $\|\cdot\|$ for any $t > 0$. Consequently, $\|\sum_{i=1}^n \xi_i a_i\|_{\psi_2} \leq \sqrt{6}\|a\|$.

Lemma VI.8. *(**Bernstein's inequalities**; Massart 2003)*

Let $O_1, \dots, O_n$ be independent and square integrable random variables such that $E[O_i] = 0$ for all $i = 1, \dots, n$.

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

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

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

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

Lemma VI.9. *(**Concentration inequality;** Bartlett et al. 2005)*

Consider n independent random variables $O_1, \dots, O_n$ with values in some measurable space $\mathcal{O}$. Let $\mathcal{F}$ be a class of real valued functions on $\mathcal{O}$. Assume that all *functions f in $\mathcal{F}$ satisfy $Ef = 0$ and $\|f\|_\infty \leq b$ for some $b > 0$. Let*

$$Z = \sup_{f \in \mathcal{F}} \Big| \sum_{i=1}^n f(O_i) \Big| \quad \textit{or} \quad Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^n f(O_i).$$

Let σ be a positive constant such that $\sigma^2 \geq \sup_{f \in \mathcal{F}} \Big[\sum_{i=1}^n Var(f(O_i)) \Big]$. Then for any $t \geq 0$, with probability at least $1 - \exp(-t)$,

$$Z \leq \mathbb{E}Z + \sqrt{2t(n\sigma^2 + 2b\mathbb{E}Z)} + \frac{bt}{3}$$

Lemma VI.10. *(Bartlett et al. 2005)*

Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{O}$ such that $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq b$ for a positive constant b. Let

$$Z = \mathbb{E}_\xi \sup_{f \in \mathcal{F}} \sum_{i=1}^n \xi_i f(O_i).$$

Then for any $t > 0$,

$$\mathbb{P}\Big(Z \geq \mathbb{E}Z + \sqrt{2bt\mathbb{E}Z} + \frac{bt}{3} \Big) \leq \exp(-t)$$

and

$$\mathbb{P}\big(Z \leq \mathbb{E}Z - \sqrt{2bt\mathbb{E}Z} \big) \leq \exp(-t).$$

Definition VI.3. (**Sub-root function;** Bartlett et al. 2005)

A function $\eta : [0, \infty) \to [0, \infty)$ is sub-root if it is nonnegative, nondecreasing and if $r \to \eta(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Lemma VI.11. *(Bartlett et al. 2005)*

If $\eta : [0, \infty) \to [0, \infty)$ is a nontrivial sub-root function (i.e. η is not the constant zero function), then it is continuous on $[0, \infty)$ and the equation $\eta(r) = r$ has a unique positive solution. Moreover, if we denote the solution by r^*, then for all $r > 0$,

$r \geq \eta(r)$ if and only if $r \geq r^*$.

Definition VI.4. (**Star-shape;** Bartlett et al. 2005)

A class of functions $\mathcal{F}$ is star-shaped around f_0 if $f_0 + \alpha(f - f_0) \in \mathcal{F}$ for all $f \in \mathcal{F}$ and $\alpha \in [0, 1]$.

Lemma VI.12. *(Bartlett et al. 2005)*

If the class $\mathcal{F}$ is star-shaped around $\hat{f}$ (which may depend on the data), and $T : \mathcal{F} \to \mathbb{R}^+$ is a (possibly random) function that satisfies $T(\alpha f) \leq \alpha^2 T(f)$ for any $f \in \mathcal{F}$ and any $\alpha \in [0, 1]$, then the (random) function η defined for $r \geq 0$ by r

$$\eta(r) = \mathbb{E}_\xi \sup_{f \in \mathcal{F}: T(f - \hat{f}) \leq r} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(O_i)$$

is sub-root and $r \to \mathbb{E}\eta(r)$ is also sub-root.

Lemma VI.13. *(Bartlett et al. 2005)*

Let $\mathcal{F}$ be a class of functions with ranges in $[b_1, b_2]$ and assume that there are some functional $T : \mathcal{F} \to \mathbb{R}^+$ some constant B such that for every $f \in \mathcal{F}$, $Var[f] \leq T(f) \leq BEf$. Assume there is a sub-root function η such that

$$\eta(r) \geq B\mathbb{E} \sup_{f \in \mathcal{F}: T(f) \leq r} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(O_i)$$

for any $r > r^*$, where r^* is the fixed point of η. Then, with $c_1 = 704$ and $c_2 = 26$, for any $K > 1$ and every $t > 0$, with probability at least $1 - \exp(-t)$,

$$\forall f \in \mathcal{F}, \quad Ef \leq \frac{K}{K-1} E_n f + \frac{c_1 K}{B} r^* + \frac{(11(b_2 - b_1) + c_2 BK)t}{n}.$$

Also, with probability at least $1 - \exp(-t)$,

$$\forall f \in \mathcal{F}, \quad E_n f \leq \frac{K+1}{K} Ef + \frac{c_1 K}{B} r^* + \frac{(11(b_2 - b_1) + c_2 BK)t}{n}.$$

Lemma VI.14. *(**Contraction inequality;** Bartlett et al. 2005)*

Let $l(\cdot)$ be a contraction, that is, $|l(x) - l(y)| \leq |x - y|$. Then, for every class $\mathcal{F}$,

$$\mathbb{E}_\xi \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i l \circ f(O_i) \leq \mathbb{E}_\xi \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(O_i). \tag{6.1}$$

Lemma VI.15. *(**Desymmetrization - Symmetrization inequality;** Koltchinskii 2006)*

For any class of functions $\mathcal{F}$,

$$\frac{1}{2}\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i [f(O_i) - Ef(O_i)] \leq \max\Big(\mathbb{E} \sup_{f \in \mathcal{F}} (E - E_n)f, \mathbb{E} \sup_{f \in \mathcal{F}} (E_n - E)f\Big) \tag{6.2}$$

$$\max\Big(\mathbb{E} \sup_{f \in \mathcal{F}} (E - E_n)f, \mathbb{E} \sup_{f \in \mathcal{F}} (E_n - E)f\Big) \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(O_i). \tag{6.3}$$

The lower bound is often referred to as a desymmetrization inequality and the upper bound as a symmetrization inequality.

Lemma VI.16. *Suppose $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$. Then*

$$\mathbb{E}\Big[\sup_{f \in \mathcal{F}} E_n f^2(O)\Big] \leq \sup_{f \in \mathcal{F}} Ef^2 + 4b\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(O_i)\Big]$$

Proof. By the symmetrization inequality (6.3), we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{f \in \mathcal{F}} E_n f^2(O)\Big] &\leq \sup_{f \in \mathcal{F}} Ef^2(O) + \mathbb{E}\Big[\sup_{f \in \mathcal{F}} (E_n - E) f^2(O)\Big] \\
&\leq \sup_{f \in \mathcal{F}} Ef^2(O) + 2\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f^2(O_i)\Big] \\
&\leq \sup_{f \in \mathcal{F}} Ef^2(O) + 4b\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \xi_i f(O_i)\Big],
\end{aligned}$$

where the second inequality follows from the contraction inequality (6.1). □

BIBLIOGRAPHY

BIBLIOGRAPHY

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

D. Almirall, L.L. Gunter, and S.A. Murphy. Efficient a-learning for dynamic treatment regimes: a handout. *RAND Workshop on Dynamic Treatment Regimes, Santa Monica*, August, 2005.

S. Arlot. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624, 2009.

S. Arlot and P. Bartlett. Margin adaptive model selection in statistical learning. *Submitted*, 2008.

Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Ralated Fields*, 117:467–493, 2000.

A. Barron, L Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

P.L. Bartlett. Fast rates for estimation error and oracle inequalities for model selection. *Econometric Theory*, 24(2):545–552, 2008.

P.L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135:311–334, 2006.

P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

P.L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, 33:1497–1537, 2005.

P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101:138–156, 2006.

P.J. Bickel, Y. Ritov, and A.B. Tsybakow. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

L. Birgé and P. Massart. From model selection to adaptive estimation. In E. Torgersen D. Pollard and G. Yang, editors, *Festschrift for Lucien Le Cam: Research papers in Probability and Statistics*, pages 55–87. Springer, New York, 1997.

L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007a.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007b.

T. Cai, L. Tian, D.M. Lloyd-Jones, and L.J. Wei. Evaluating subject-level incremental values of new markers for risk classification rule. *Harvard University Biostatistics Working Paper Series, Working paper*, page 91, 2008a.

T. Cai, L. Tian, H. Uno, S.D. Solomon, and L.J. Wei. Calibrating parametric subject-specific risk estimation. *Harvard University Biostatistics Working Paper Series, Working paper*, page 92, 2008b.

V. Cherkassky, X. Shao, F. Mulier, and V. Vaplik. Model complexity control for regression using vc generalization bounds. *IEEE Transactions on Neural Networks*, 10:1075–1089, 1992.

J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1988.

D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 1994.

B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7: 1–26, 1979.

B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, Philadelphia, 1982.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and it oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

M. Fromont. Model selection by bootstrap penalization for classification. *Machine Learning*, 66:165–207, 2007.

E. Greenshtein. Best subset selection, persistence in high dimensional statistical learning and optimization under l_1 constraint. *The Annals of Statistics*, 34:2367–2386, 2006.

D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

T.R. Insel. Translating scientific opportunity into public health impact: a strategic plan for research on mental illness. *Archives of General Psychiatry*, 66(2):128–133, 2009.

J. Ishigooka, M. Murasaki, S. Miura, and The Olanzapine Late-Phase II Study Group. Olanzapine optimal dose: Results of an open-label multicenter study in schizophrenic patients. *Psychiatry and Clinical Neurosciences*, 54(4):467–478, 2001.

M. Kearns, Y. Mansour, A.Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. In *Proceedings of the Eighth Annual ACM Workshop on Computational Learning Theory*, pages 21–30, New York, 1995.

M.B. Keller, J.P. McCullough, D.N. Klein, B. Arnow, D.L. Dunner, A.J. Gelenberg, J.C. Markowitz, Nemeroff. C.B., J.M. Russell, M.E. Thase, M.H. Trivedi, and J. Zajecka. 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, 2000.

D.M. Kent, R.A. Hayward, J.L. Griffith, S. Vijan, J.R. Beshansky, R.M. Califf, and H.P. Selker. An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. *The American Journal of Medicine*, 113(2):104–111, 2002.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 4:1902–1914, 2001.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré Probabilitiés et Statistiques*, 45(1):7–57, 2009.

M Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

M.G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

L.J. Lesko. Personalized medicine: Elusive dream or imminent reality? *Clinical Pharmacology and Therapeutics*, 2007.

F. Lozano. Model selection using rademacher penalization. In *Proceedings of the Second ICSC Symposium on Neural Networks*, Berlin, 2000.

G. Lugosi and A. Nobel. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27:1830–1864, 1999.

G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, 32:1679–1697, 2004.

J.K. Lunceford, M. Davidian, and A.A. Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58:48–57, 2002.

C. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27:1808–1829, 1999.

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

P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciencies de Toulouse*, IX:245–303, 2000.

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

N. Meinshausen and P. Buhlmann. High-dimensional graphs and *The Annals of Statistics*, 34:1436–1462, 2006.

N. Meinshausen and P. Buhlmann. Lasso-type recovery of sparse data. *The Annals of Statistics*, 37(1):246–270, 2009.

S.A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B (with discussion)*, 65(2):331–366, 2003.

S.A. Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005.

S.A. Murphy, M.J. van der Laan, J.M. Robins, and CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96:1410–1423, 2001.

D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49:161–178, 2002.

P. Piquette-Miller and D.M. Grant. The art and science of personalized medicine. *Clinical Pharmacology and Therapeutics*, 81:311–315, 2007.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

W. Polonik. Measuring mass concentrations and estimating density contour clusters - an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

J.M. Robins. Optimal structural nested models for optimal sequential decisions. In D.Y. Lin and P. Heagerty, editors, *Proceedings of the Second Seattle Symposium on Biostatistics*, New York, Springer, 2004.

J.M. Robins, L. Orellana, and A. Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721, 2008.

G. Schwarz. Estimating the demension of a model. *The Annals of Statistics*, 6: 461–464, 1978.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

P.F. Thall, R.E. Millikan, and H.G. Sung. Evaluating multiple treatment courses in clinical trials. *Statistics in Medicine*, 19:1011–1028, 2000.

P.F. Thall, H.G. Sung, and E.H. Estey. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *Journal of the American Statistical Association*, 97:29–39, 2002.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 32:135–166, 1996.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

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

M.J. van de Laan, M.L. Petersen, and M.M. Joffe. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *The International Journal of Biostatistics*, page Article 4, 2005.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

V.N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10:988–999, 1999.

V.N. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition.* (in Russian), Nauka, Moscow, 1974.

A.S. Wahed and A.A. Tsiatis. Semiparametric efficient estimation of survival distribution for treatment policies in two-stage randomization designs in clinical trials with censored data. *Biometrika*, 93:163–177, 2006.

C.J.C.H. Watkins. *Learning from delayed rewards.* Ph.D. Thesis, Cambridge University, 1989.

C-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.