Title
Essays on Treatment Effect Estimation and Treatment Choice Learning

Author
Shi, Liqiang

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Treatment Effect Estimation and Treatment Choice Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Economics

by

Liqiang Shi

2022

ABSTRACT OF THE DISSERTATION

Essays on Treatment Effect Estimation
and Treatment Choice Learning

by

Liqiang Shi
Doctor of Philosophy in Economics
University of California, Los Angeles, 2022
Professor Andres Santos, Chair

This dissertation consists of three chapters that study treatment effect estimation and treatment choice learning under the potential outcome framework (Neyman, 1923; Rubin, 1974). The first two chapters study how to efficiently combine an experimental sample with an auxiliary observational sample when estimating treatment effects. In chapter 1, I derive a new semiparametric efficiency bound under the two-sample setup for estimating ATE and other functions of the average potential outcomes. The efficiency bound for estimating ATE with an experimental sample alone is derived in Hahn (1998), and has since become an important reference point for studies that aim at improving the ATE estimation. This chapter answers how an auxiliary sample containing only observable characteristics (covariates, or features) can lower this efficiency bound. The newly obtained bound has an intuitive expression and shows that the (maximum possible) amount of variance reduction depends positively on two factors: 1) the size of the auxiliary sample, and 2) how well the covariates predict the individual treatment effect. The latter naturally motivates having high dimensional covariates and the adoption of modern machine learning methods to avoid over-fitting.

In chapter 2, under the same setup, I propose a two-stage machine learning (ML) imputation estimator that achieves the efficiency bound derived in chapter 1, so that no other regular estimators for ATE can have lower asymptotic variance in the same setting. This estimator involves two steps. In the first step, conditional average potential outcome functions are estimated nonparametrically via ML, which are then used to impute the unobserved potential outcomes for every units in both samples. In the second step, the imputed potential outcomes are aggregated together in a robust way to produce the final estimate. Adopting the cross-fitting technique proposed in Chernozhukov et al. (2018), our two-step estimator can use a wide range of supervised ML tools in its first step, while maintaining valid inference to construct confidence intervals and perform hypothesis tests. In fact, any method that estimates the relevant conditional mean functions consistently in $L_2(P)$ norm, with no rate requirement, will lead to efficiency through the proposed two-step procedure. I also show that cross-fitting is not necessary when the first step is implemented via LASSO or post-LASSO. Furthermore, our estimator is robust in the sense that it remains consistent and $\sqrt{n}$ normal (no longer efficient) even if the first step estimators are inconsistent.

Chapter 3 (coauthored with Kirill Ponomarev) studies model selection in treatment choice learning. When treatment effects are heterogeneous, a decision maker, given either experiment or quasi-experiment data, can attempt to find a policy function that maps observable characteristics to treatment choices, aiming at maximizing utilitarian welfare. When doing so, one often has to pick a constrained class of functions as candidates for the policy function. The choice of this function class poses a model selection problem. Following Mbakop and Tabord-Meehan (2021) we propose a policy learning algorithm that incorporates data-driven model selection. Our method also leverages doubly robust estimation (Athey and Wager, 2021) so that it could retain the optimal $n^{-1/2}$ rate in expected regret in general setups including quasi-experiments where propensity scores are unknown. We also refined some related results in the literature and derived a new finite sample lower bound on expected regret to show that the $n^{-1/2}$ rate is indeed optimal.

The dissertation of Liqiang Shi is approved.

Jinyong Hahn

Denis Nikolaye Chetverikov

Chad J. Hazlett

Andres Santos, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I owe special thanks to my advisor Andres Santos. His conscientious attitude to work and kindness to people inspire me everyday. I could always look up to him as a great role model. I am also grateful to Denis Chetverikov and Jinyong Hahn for their valuable guidance and continuous support, Chad Hazlett for graciously serving on my committee, Zhipeng Liao, Rosa Matzkin and Shuyang Sheng for many helpful comments and discussions throughout the years. I also thank my supportive friend and classmate Kirill Ponomarev, with whom I had the pleasure of coauthoring the last chapter of this dissertation. All errors are mine.

VITA

2014	B.A. Economics, UIBE, Beijing, China
2014–2016	Graduate Study at CEMFI, Madrid, Spain
2017	M.A. Economics, UCLA, Los Angeles, California.
2018	Ph.D. Candidate, UCLA, Los Angeles, California.
2016–2022	Graduate Teaching Fellow, UCLA, Los Angeles, California.

CHAPTER 1

Efficiency Bound for ATE under Sample Combination

1.1 Introduction

Randomized controlled trials (RCTs) have long been a common tool in biomedical sciences and pharmaceutical industry, and are increasingly popular in social sciences and business as well (Mason et al., 2003; Box et al., 2005; Imbens and Rubin, 2015; Rosenberger and Lachin, 2015; Kohavi and Longbotham, 2017). For example, in the field of development economics, the research center J-PAL alone has conducted more than 1000 randomized experiments in more than 90 countries to date.[1] In business, data-driven decision-making is becoming a culture and experimentation (often referred as A/B tests) is practically a mantra (Tang et al., 2010). Large technology companies like Amazon and Microsoft conduct tens of thousands of RCTs each year (Kohavi and Thomke, 2017). This wide popularity of RCTs is mainly due to their ability to provide clean identification and unbiased estimates on causal effects.

In RCTs, one of the most important parameter to estimate is the average treatment effect (ATE). Although the identification is no problem, the accuracy of its estimation and the power of the relevant statistical tests are limited by the experimental sample size, which itself is typically subject to a number of practical constraints. For example, in certain field experiments, administering the treatment can be costly, as can be the extensive follow-ups on the test subjects for measuring outcomes. In such cases, improving estimation efficiency becomes important as it allows us to work with smaller sized samples, reducing the costs

[1] https://www.povertyactionlab.org/evaluations

of running the experiment. One way to improve estimation efficiency is to combine the experimental sample with an auxiliary observational sample. The latter is often readily available or much cheaper to collect. In this chapter, I will study how the additional auxiliary sample contributes to estimation efficiency by deriving a semiparametric effciency bound on the ATE parameter under the two-sample setup. In the next chapter, I will propose an estimator that acheives such efficiency bound. Before more details, let me provide some further motivation for improving estimation efficiency.

As more efficient estimators have lower asymptotic variances, the problem of improving estimation efficiency is also sometimes called variance reduction, especially in the context of online A/B tests. In this context, even though experiments seem relatively cheap and easy to carry out at large scales, variance reduction can still be very important. This is due to multiple reasons. First, the treatment effect could be simply small and hard to distinguish from zero. Lewis and Rao (2015), for example, document that some advertising experiments require more than 10 million person-weeks to accurately measure the return to advertising. Second, businesses may want to do more experiments with more treatment arms at the same time, which splits up their traffic and reduces sample size. Third, as is often the case, the sample is only collected as users visit the website over time. Therefore, more efficient estimator allows decisions to be made faster as less units are required for the experiment to yield statistical significance results. Fourth, some experiments are disruptive to user experience and are better limited at a small scale. Google has made clear that they are not satisfied with the amount of traffic they have (Tang et al., 2010), and researchers at Microsoft, Facebook and Netflix all have been developing ways to improve estimation efficiency for their online experiments (Deng et al., 2013; Xie and Aurisset, 2016; Liou and Taylor, 2020).

As mentioned earlier, this chapter and the following one study how to improve estimation efficiency by combining the experimental sample with an auxiliary observational sample. I assume that the auxiliary sample only contains observable characteristics/features and is

sampled from the same target population as the experimental sample. Depending on the application, such auxiliary samples could be readily available in large size. For instance, a common assumption for A/B tests (and also online bandit problems) is that the vistors coming to the website are i.i.d. draws from a large stable population. Under this assumption, traffic prior and after the experiment period, as well as traffic split into other experiments, are all sources of our auxiliary sample. In field RCTs, an experiment may have its participants drawn from a large database (e.g. administrative record) of individuals. This database would then likely also contain observable characteristics of people who are not participants of the experiment (Gagnon-Bartsch et al., 2021). Furthermore, researchers could potentially collect the auxiliary samples on purpose. Since no intervention or extensive follow-ups are required (essentially a baseline survey), collecting the auxiliary sample could be a cheaper and easier way to increase power than including more subjects into the experiment.

Deriving a new semiparametric efficiency bound[2] under the two-sample setup for estimating average treatment effects (ATE) is an important first step to understand how the auxiliary sample can be best utilized to improved estimation efficiency. The efficiency bound for estimating ATE with an experimental sample alone is derived in Hahn (1998), and has since become an important reference point for methods and techniques that aim at improving the ATE estimation. This chapter answers how an auxiliary sample can lower this efficiency bound. The obtained bound has an intuitive expression and shows that the (maximum possible) amount of variance reduction through incorporating the auxiliary sample depends positively on two factors: 1) the size of the auxiliary sample, and 2) how well the observable characteristics (covariates) predict the individual treatment effects.

A vast literature explores efficient estimation and covariate adjustments in RCT related setups without auxiliary data. Results are abundant under both low dimensional asymp-

[2]Semiparametric efficiency bounds are of fundamental importance for semiparametric models (Newey, 1990). The bounds provide a guide to estimation methods. They give a standard against which the asymptotic efficiency of any particular estimator can be measured. Intuitively, they characterizes the lowest asymptotic variance that regular estimators can attain.

totics, where the model is fixed and sample size goes to infinity (Hahn, 1998; Rosenbaum, 2002; Hirano et al., 2003; Freedman, 2008; Imbens and Wooldridge, 2009; Berk et al., 2013; Lin, 2013; Athey and Imbens, 2017; Ding et al., 2019), and more recently high-dimensional asymptotics, where number of covariates in the model grows with sample size (Belloni et al., 2014, 2015; Farrell, 2015; Athey et al., 2016; Wager et al., 2016; Bloniarz et al., 2016; Chernozhukov et al., 2018). My study extend these results by including the auxiliary sample into consideration.

Some recent efforts also study how to improve estimation efficiency in RCTs by incorporating additional data, but differ from this paper in various ways. Deng et al. (2013) propose a method (CUPED) that uses pre-experiment outcomes as an additional covariate in the covariate adjustments. Their method requires repeatedly observing the same units before and during the experiment. Gui (2020) considers an auxiliary sample sampled from the same target population as the experimental sample (same as this paper). He assumes that an endogenous outcome is observed in the auxiliary sample and adopts a linear causal model with IV-like structure to study efficiency gain. Gagnon-Bartsch et al. (2021) propose a method named reLOOP that combines two former methods LOOP (Wu and Gagnon-Bartsch, 2021) and rebar (Sales et al., 2018). The reLOOP method trains a predictor for outcome from the auxiliary data, and then use it to impute outcomes for the experimental units as an additional covariate. The inference in their paper is design-based, as opposed to under the large population model.

Studies on combining experimental data with observational data for reasons other than variance reduction (e.g. external validity, estimating long-run effects) include Hartman et al. (2015), Peysakhovich and Lada (2016) Rosenman et al. (2018), Athey et al. (2020), among others.

The rest of this chapter proceeds as follows. Section 1.2 formally states the treatment effect model and the statistical inference problem. Section 1.3 derives a result on semiparametric efficiency bound in a more general two-sample problem that encompasses our

treatment effect model as a special case. Section 1.4 presents efficiency bound on ATE and other functionals of the average potential outcomes, and Section 1.5 concludes. Proofs are all collected in the appendix that forms Section 1.6.

1.2 Model

We consider two i.i.d. random samples $\{(Y_i, T_i, Z_i)\}_{i=1}^n$ and $\{X_i\}_{i=n+1}^m$ sampled separately from the same large population. The first sample is an experimental sample from the canonical potential outcome model (Neyman, 1923; Rubin, 1974) where $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the outcome variable of interest, $Z_i \in \mathcal{Z} \subset \mathbb{R}^{d_Z}$ is a d_Z-dimensional vector of covariates (features), and $T_i \in \{0, 1, ..., \mathtt{T}\}$ is a treatment indicator, with $T_i = t$ if unit i receives treatment t. We do not impose any order among the treatment arms and $T = 0$ is reserved for the control group. Each unit has a set of potential outcomes $\{Y_i(t)\}_{t=0}^{\mathtt{T}}$ so that

$$Y_i = Y_i(T_i) = \sum_{t=1}^{\mathtt{T}} Y_i(t) \mathbb{1}\{T_i = t\}, \quad 1 \leq i \leq n. \tag{1.1}$$

Treatment T_i is randomly assigned to each unit i by the experimenter. The assignment probabilities can depend on the vector of covariates. Denote $P(T = t|Z = z) = \pi_t(z)$, these functions (propensity scores) are chosen by the researcher and are hence known functions in the statistical problem.[3] The second sample $\{X_i\}_{i=n+1}^m$ is an auxiliary sample on the covariates from the same target population. We assume X_i is a subvector of Z_i, hence we can write $Z_i = (X_i^E, X_i)$, where X_i^E is the subvector of the covariates in Z_i that are only observed in the experimental sample. Since the auxiliary sample is sampled from the same target population as the experimental sample, all X_i, $1 \leq i \leq m$, follow the same distribution. The auxiliary sample can be collected before, after or simultaneously with the experimental sample. Since no experiment is conducted on units in the auxiliary sample, we

[3]To focus on studying the efficiency gain from auxiliary data, we assume known propensity scores here for simplicity. An extension to unknown propensity score, e.g. observational data under conditional ignorability, is possible.

do not observe T_i nor Y_i for units $i > n$. All modeling assumptions are summarized below.

Assumption 1.2.1. *i) Independent random samples:* $\{(Y_i, T_i, Z_i)\}_{i=1}^n$ *and* $\{X_i\}_{i=n+1}^m$ *are both i.i.d. samples and are also independent from each other. ii) Conditional random assignment:* T_i *is independent from* $\{Y_i(t)\}_{t=0}^{\mathtt{T}}$ *conditional on* $Z_i = (X_i^E, X_i)$ *for all* $i \leq n$. *iii) Propensity score: For all* $t \in \{0, 1, ..., \mathtt{T}\}$, $P(T_i = t|Z_i) = \pi_t(Z_i)$ *is known and* $\pi_t(z) \geq \pi_{\min} > 0$ *almost surely. iv) Common population:* X_i *in both samples follow the same distribution.*

The parameters of interests are the average potential outcomes $\mu_t = E[Y(t)]$ or any differentiable functions of them including the average treatment effects $\tau_{t,t'} = \mu_t - \mu_{t'}, t \neq t'$. These parameters are identified in the first sample alone by the standard identification Assumption 1.2.1 (ii) and (iii). In particular,

$$\mu_t = E[E[Y|Z, T = t]]. \tag{1.2}$$

From now on, we focus on efficient estimation of the response vector $\boldsymbol{\mu} = (\mu_0, ..., \mu_{\mathtt{T}})'$ as the efficiency bound for any known differentiable function of $\boldsymbol{\mu}$, including the average treatment effects, would easily follow if we find the efficiency bound for $\boldsymbol{\mu}$.

Remark 1.2.1. The assumption that X_i is a subvector of Z_i is not substantial. A more general set up is to consider two samples $\{(Y_i, T_i, X_i, X_i^E)\}_{i=1}^n$ and $\{(X_i, X_i^A)\}_{i=n+1}^m$, where X^A is only observed in the auxiliary sample. One can show that the efficiency bounds for μ_t and $\tau_{t,t'}$ remain unchanged after adding X^A. In other words, $\{X_i^A\}_{i=n+1}^m$ does not provide any extra useful information.

Remark 1.2.2. The efficiency bound result in this paper can be easily extended to the problem of estimating parameter $\phi(E[g(W, \tilde{W})])$ from two independent i.i.d. samples $\{(W_i, \tilde{W}_i)\}_{i=1}^n$ and $\{\tilde{W}_i\}_{i=n+1}^m$, collected separately from the same large population, for any known vector-valued function g and continuously differentiable function $\phi : \mathbb{R}^{d_g} \to \mathbb{R}^{d_\phi}$.

Remark 1.2.3. An alternative way to study sample combination problem is to assume multinomial sampling/missing data model. For our problem, this requires to model the data

as one single i.i.d. sample $\{Y_i\Delta_i, T_i\Delta_i, X_i^E\Delta_i, \Delta_i, X_i\}_{i=1}^m$, where Δ_i is a "sample indicator" that follows a known Bernoulli distribution with $P(\Delta = 1) = \frac{n}{m}$, independent from all other variables.[4] This is a somewhat unnatural model for the sampling process that might arise in applications, especially when the two samples are collected separately with predetermined sample sizes. Certain statistics share the same asymptotic distributions under this set up and the two-sample set up considered in this paper. However, to what extent the two models are "equivalent" remains unclear. In particular, variable Δ_i does not exist in the two-sample set up. Whether the extra randomness it brings affects semiparametric efficiency bounds needs to be examined.

Next, we derive the asymptotic variance bounds for all regular estimators of $\boldsymbol{\mu}$ under the two-sample set up specified in Assumption 1.2.1. To this aim, we first derive a result on semiparametric efficiency bound in a more general two-sample problem that encompasses our treatment effect model as a special case, then we calculate the bounds for $\boldsymbol{\mu}$ and $\tau_{t,t'}$ from there. We follow the semiparametric efficiency bound theory. For a detailed introduction to this topic, please refer to textbooks Van der Vaart (2000) and Bickel et al. (1993).

1.3 General Two-sample Problem

In this section, we study the semiparametric efficiency theory in a general two-sample set up. Suppose we have two independent i.i.d. samples $\{W_i\}_{i=1}^n$ and $\{\tilde{W}_i\}_{i=n+1}^m$. For the asymptotic analysis, we think of m as a sequence m_n indexed by n and that $\frac{m_n}{n} \to \gamma$ as $n \to \infty$. This allows us to think of the statistical experiment (Van der Vaart, 2000) as only indexed by n. All definitions and theorems hereon are implicitly under this asymptotic framework. We choose n as the index for asymptotic analysis because we consider $\{W_i\}_{i=1}^n$ as the primary

[4]Under this alternative model, the size of the experimental sample, i.e. $\sum_{i=1}^m \Delta_i$, is random and only equals to n in expectation. One way to address this is to assume Δ_i are block randomized so that $\sum_{i=1}^m \Delta_i = n$ with probability one, but the sample would no longer be i.i.d., creating difficulties for studying semiparametric efficiency.

sample and $\{\tilde{W}_i\}_{i=n+1}^m$ as an auxiliary sample. The asymptotic distributions of estimators will also be obtained by scaling with $\sqrt{n}$, so that the asymptotic variances are directly comparable with those of estimators that don't utilize the auxiliary sample. In appendix 1.6.2, we provide a simple example to show that the $\frac{m_n}{n} \to \gamma$ asymptotics is sensible in the two-sample set up when γ is naturally replaced by $\frac{m}{n}$ in the variance estimation.

We assume $W \sim P \in \mathbf{P}$ and $\tilde{W} \sim Q \in \mathbf{Q}$. In other words, P and Q are the true distributions for random variables W and $\tilde{W}$, and they respectively belong to sets of distributions $\mathbf{P}$ and $\mathbf{Q}$. The sets $\mathbf{P}$ and $\mathbf{Q}$ are known and are restricted by modeling assumptions. $\{W_i\}_{i=1}^n$ and $\{\tilde{W}_i\}_{i=n+1}^m$ are two independent i.i.d. samples that follow the product measure $P^n \otimes Q^{m-n}$. The two distributions P and Q are potentially related, which means that the pair (P, Q) could belong to a restricted subset $\mathbf{M} \subset \mathbf{P} \times \mathbf{Q}$, instead of the whole Cartesian product $\mathbf{P} \times \mathbf{Q}$. The restrictions that characterize $\mathbf{M}$ are also known modeling assumptions. The set $\mathbf{M}$ can be interpreted as the statistical model.

Remark 1.3.1. To map our treatment effect model into here as an example, $W = (Y, T, X^E, X) \sim P \in \mathbf{P}$ and $\tilde{W} = \tilde{X} \sim Q \in \mathbf{Q}$. [5] $\mathbf{P}$ is restricted by Assumption 1.2.1 (iii), the known propensity score[6], and $\mathbf{Q}$ is not restricted. Moreover, $\mathbf{M}$ is further restricted by Assumption 1.2.1 (iv), which requires that the distribution Q of $\tilde{X}$ is the same as the marginal distribution of X under P. Interestingly, the crucial identification Assumption 1.2.1 (ii) does not put any restriction on the observable distribution. The role it plays here is to give the otherwise rather arbitrary parameter $E_P[E_P[Y|T = t, Z]]$ a causal interpretation through the latent potential outcomes, namely $E_P[Y(t)]$.

The semiparametric efficiency bound can be heuristically understood as the lowest achiev-

[5] We added the "tilde" here to distinguish $\tilde{X}$ from X. $\tilde{X}$ represents the random variables from the second sample. We were able to avoid using this notation in previous sections because $\tilde{X}$ and X has the same distribution under our model and the second sample is indexed from $n+1$ instead of 1 (which we still do in this section for consistency).

[6] $\mathbf{P}$ can be understood as a conditional moment restriction model, as the set of distributions such that $E_P[\mathbb{1}\{T = t\} - \pi_t(Z)|Z] = 0$ for all $t \in (0, ..., \mathtt{T})$ and $Z = (X^E, X)$.

able asymptotic variance under the hardest parametric specification of the nonparametric model. Following this literature to characterize efficiency, we introduce local one-dimentional smooth parametric submodels, defined as follow.

Definition 1.3.1. *A local smooth submodel $\iota \mapsto (P_{g,\iota}, Q_{h,\iota})$ is a function defined on $[0,1]$ such that for every $\iota \in [0,1]$, $(P_{g,\iota}, Q_{h,\iota})$ belongs to the restricted set $\mathbf{M}$ and $P_{g,\iota} \otimes Q_{h,\iota}$ is a measure for $(W, \tilde{W})$. Moreover, $(P_{g,0}, Q_{h,0}) = (P, Q)$ and as $\iota \to 0$,*

$$\int (\frac{dP_{g,\iota}^{\frac{1}{2}} - dP^{\frac{1}{2}}}{\iota} - \frac{1}{2} g dP^{\frac{1}{2}})^2 \to 0, \tag{1.3}$$

$$\int (\frac{dQ_{h,\iota}^{\frac{1}{2}} - dQ^{\frac{1}{2}}}{\iota} - \frac{1}{2} h dQ^{\frac{1}{2}})^2 \to 0, \tag{1.4}$$

for functions $g(W) \in L_2(P)$ and $h(\tilde{W}) \in L_2(Q)$.

Under this definition, we can show that the likelihood ratio process ((1.5) below) of this model converges to a Gaussian limit as is stated by the following lemma.

Lemma 1.3.1. *If the submodel $\iota \mapsto (P_{g,\iota}, Q_{h,\iota})$ in $\mathbf{M}$ satisfies (1.3) and (1.4), then $E_P[g] = 0$, $E_P[g^2] < \infty$, $E_Q[h] = 0$, $E_Q[h^2] < \infty$ and as $n \to \infty$, $\frac{m}{n} \to \gamma$,*

$$\log \prod_{i=1}^{n} \frac{dP_{g,\frac{1}{\sqrt{n}}}}{dP}(W_i) \prod_{i=n+1}^{m} \frac{dQ_{h,\frac{1}{\sqrt{n}}}}{dQ}(\tilde{W}_i) \tag{1.5}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(W_i) - \frac{1}{2} E_P[g^2] + \frac{1}{\sqrt{m-n}} \sum_{i=n+1}^{m} \sqrt{\gamma - 1} h(\tilde{W}_i) - \frac{1}{2}(\gamma - 1) E_Q[h^2] + o_P(1). \tag{1.6}$$

The proof is collected in appendix 1.6.1. The display in line (1.6) converges in distribution to

$$N(-\frac{1}{2} E[\tilde{g}^2], E[\tilde{g}^2]), \tag{1.7}$$

where $\tilde{g}(W, \tilde{W}) = g(W) + \sqrt{\gamma - 1} h(\tilde{W})$ and the expectation is taken under the product measure $P \otimes Q$. This Gaussian limit is the key to apply the asymptotic representation theorem, through which we can derive an asymptotic variance bound that depends on (in addition to the parameter of interest itself) the set of such $\tilde{g}$ functions that could arise with

different submodels. As the submodel $\iota \mapsto (P_{g,\iota}, Q_{h,\iota}) \in \mathbf{M}$ varies, the function $\tilde{g}$ varies and we can get a collection of such functions. We call this collection of functions the tangent set $\mathcal{T}(P,Q)$ and its elements score functions of the submodels.

$$\mathcal{T}(P,Q) = \{\tilde{g} = g + \sqrt{\gamma - 1}h : (1.3) \text{ and } (1.4) \text{ holds for some } \iota \mapsto (P_{g,\iota}, Q_{h,\iota}) \in \mathbf{M}\}. \quad (1.8)$$

Similar tangent set is considered in the original one-sample problem. Suppose we only observe one i.i.d. sample $\{W_i\}_{i=1}^n$ of $W \sim P \in \mathbf{P}$ and that the model is $\mathbf{P}$, then we have the following tangent set,

$$\mathcal{T}(P) = \{g \in L_2(P) : (1.3) \text{ holds for some } \iota \mapsto P_{g,\iota} \in \mathbf{P}\}.$$

We assume that the closure of these tangent sets under L_2 norms are linear subspaces. In particular, $\overline{\mathcal{T}}(P,Q)$ is the closure of $\mathcal{T}(P,Q)$ under $\|\cdot\|_{L_2(P\otimes Q)}$ and is a subspace of $L_2(P\otimes Q)$.

Next, we impose the following assumptions on the parameter of interest ϕ, which encompasses the treatment effect applications we see in this paper.

Assumption 1.3.1. *The parameter $\phi \in \mathbb{R}^{d_\phi}$ is i) identified in the first sample, i.e. $\phi(P,Q) = \phi(P)$, and ii) differentiable relative to the tangent set $\mathcal{T}(P,Q)$, that is for every smooth submodel $\iota \mapsto (P_{g,\iota}, Q_{h,\iota}) \in \mathbf{M}$ with score function $\tilde{g} = g + \sqrt{\gamma - 1}h \in \mathcal{T}(P,Q)$,*

$$\left.\frac{d}{d\iota}\right|_{\iota=0} \phi(P_{g,\iota}) =: \dot{\phi}_P(g), \quad (1.9)$$

where $\dot{\phi}_P$ is a continuous linear map from $L_2(P)$ to $\mathbb{R}^{d_\phi}$.

Assumption 1.3.1 requires that ϕ is a differentiable parameter that only depends on P. Since $\mathbf{M}$ is a subset of $\mathbf{P} \times \mathbf{Q}$, and ϕ only depends on P, differentiability relative to $\mathcal{T}(P)$ is sufficient for differnetiability relative to $\mathcal{T}(P,Q)$. Indeed, the first component of every submodel $\iota \mapsto (P_{g,\iota}, Q_{h,\iota}) \in \mathbf{M}$ is a submodel $\iota \mapsto P_{g,\iota} \in \mathbf{P}$. Moreover, let $\tilde{\phi}$ be the influence function under the one-sample problem, so that $\tilde{\phi} \in \overline{\mathcal{T}}(P)$ and[7]

$$\left.\frac{d}{d\iota}\right|_{\iota=0} \phi(P_{g,\iota}) = E_P[\tilde{\phi}(W)g(W)], \text{ for every } \iota \mapsto P_{g,\iota} \in \mathbf{P} \text{ with } g \in \mathcal{T}(P), \quad (1.10)$$

[7]The influence function $\tilde{\phi}$ is a $d_\phi \times 1$ vector-valued function, same with the $\bar{\phi}$ defined later. When $d_\phi > 1$, we abuse the notation $\tilde{\phi} \in \overline{\mathcal{T}}(P)$ and $\bar{\phi} \in \overline{\mathcal{T}}(P,Q)$ to mean element-wise belonging.

we can immediately get that for every $\iota \mapsto (P_{g,\iota}, Q_{h,\iota}) \in \mathbf{M}$ with $\tilde{g} = g + \sqrt{\gamma - 1}h$,

$$\begin{aligned}\frac{d}{d\iota}\bigg|_{\iota=0} \phi(P_{g,\iota}) =& E_P[\tilde{\phi}(W)g(W)] \\ =& E[\tilde{\phi}(W)(g(W) + \sqrt{\gamma - 1}h(\tilde{W}))] = E[\tilde{\phi}(W)\tilde{g}(W, \tilde{W})],\end{aligned}$$

where the plain expectation E is taken with the product measure $P \otimes Q$. The second equality follows because W and $\tilde{W}$ are independent under the product measure and $E[h(\tilde{W})] = 0$. Define a vector-valued function $\bar{\phi}$ as the projection (element-wise) of $\tilde{\phi}$ onto $\overline{\mathcal{T}}(P, Q)$, i.e.

$$\bar{\phi} \in \overline{\mathcal{T}}(P, Q) \text{ such that } E[(\tilde{\phi} - \bar{\phi})\tilde{g}] = 0 \text{ for any } \tilde{g} \in \overline{\mathcal{T}}(P, Q). \tag{1.11}$$

Then we have for every $\iota \mapsto (P_{g,\iota}, Q_{h,\iota}) \in \mathbf{M}$ with $\tilde{g} \in \mathcal{T}(P, Q)$,

$$\frac{d}{d\iota}\bigg|_{\iota=0} \phi(P_{g,\iota}) = E[\bar{\phi}(W, \tilde{W})\tilde{g}(W, \tilde{W})]. \tag{1.12}$$

Expression (1.12) is analogous to (1.10), with g replaced by the score function $\tilde{g}$ of the two-sample problem. This suggests that $\bar{\phi}$ is the new efficient influence function. At last, we restrict the set of estimators we consider to regular estimators. An estimator $\hat{\phi}_n$ is regular if along any local submodel $(P_{g,\frac{1}{\sqrt{n}}}, Q_{h,\frac{1}{\sqrt{n}}})$, we have

$$\sqrt{n}(\hat{\phi}_n - \phi(P_{g,\frac{1}{\sqrt{n}}})) \overset{\tilde{g}}{\rightsquigarrow} L, \tag{1.13}$$

where the converge in distribution is along the particular submodel $(P_{g,\frac{1}{\sqrt{n}}}, Q_{h,\frac{1}{\sqrt{n}}})$ but the limiting distribution L is invariant to the choice of the submodel. This means that the estimator is robust to local perturbations of the model. In other words, the scaled and centered limiting distribution of the estimator is invariant to the local model $(P_{g,\frac{1}{\sqrt{n}}}, Q_{h,\frac{1}{\sqrt{n}}})$. The motivation of such restriction is to rule out super-efficient estimators like the Hodges-Le Cam estimator, although it also rule out useful estimators like certain shrinkage estimators. Best regular is also local asym minimax over all estimators if the parameter is difirentiable

The following lemma states that the lower bound on asymptotic variance of all regular estimators of parameter ϕ is $E[\bar{\phi}\bar{\phi}']$.

Lemma 1.3.2. *Suppose $\overline{\mathcal{T}}(P,Q) \subseteq L_2(P \times Q)$ is a linear space, parameter ϕ satisfies Assumption 1.3.1, then the asymptotic covariance matrix of every regular sequence of estimators is bounded below by $E[\bar{\phi}\bar{\phi}']$, where $\bar{\phi}$ is defined by (1.11).*

The proof is collected in Section 1.6.1.

Remark 1.3.2. Suppose ϕ is a scalar parameter, since the new efficient influence function $\bar{\phi}$ is a projection of the original influence function $\tilde{\phi}$, by pythagorean theorem, we have $E[\bar{\phi}\bar{\phi}'] = \|\bar{\phi}\|^2 \leq \|\tilde{\phi}\|^2 = E[\tilde{\phi}\tilde{\phi}']$. This says that the new efficiency bound is always smaller or equal to the original bound. This is intuitive as the second sample can only provide more information for estimating ϕ and hence improving efficiency. Similarly, if ϕ is a vector, we can show that $E[\tilde{\phi}\tilde{\phi}'] - E[\bar{\phi}\bar{\phi}']$ is always positive semidefinite.

Remark 1.3.3. We can also see that for the efficiency bound to improve, the two distributions P and Q must be related. Otherwise, we have $(P,Q) \in \mathbf{M} = \mathbf{P} \times \mathbf{Q}$ and $\mathcal{T}(P,Q) = \{g + \sqrt{\gamma - 1}h : g \in \mathcal{T}(P), h \in \mathcal{T}(Q)\}$, where $\mathcal{T}(Q)$ here is a collection of score functions h with corresponding submodels $\iota \to Q_{h,\iota} \in \mathbf{Q}$ that satisfies (1.4). This means that g and h are not related. In particular, h can be the zero function, so that $\tilde{\phi} = \tilde{\phi} + \sqrt{\gamma - 1} \cdot 0 \in \mathcal{T}(P,Q)$. As a result, the projection of $\tilde{\phi}$ on $\mathcal{T}(P,Q)$ is $\tilde{\phi}$ itself, i.e. the efficient influence function remains the same. This is expected. If the second distribution Q has no relation at all with the first distribution P, the second sample generated by Q will not provide any information in estimating the parameter $\phi(P)$ that only depends on the first distribution.

If the parameter of interest is instead $\theta(\phi)$, where $\theta : \mathbb{R}^{d_\phi} \to \mathbb{R}^{d_\theta}$ is a known fully differentiable function with Jacobian matrix J, then the efficiency bound for $\theta(\phi)$ is $JE[\bar{\phi}\bar{\phi}']J'$, as stated in the following corollary.

Corollary 1.3.1. *Under the conditions of Lemma 1.3.2. If $\theta : \mathbb{R}^{d_\phi} \to \mathbb{R}^{d_\theta}$ is a known fully differentiable function with Jacobian matrix J, then the asymptotic covariance matrix of every regular sequence of estimators $\hat{\theta}_n$ for parameter $\theta(\phi(P))$ is bounded below by $JE[\bar{\phi}\bar{\phi}']J'$.*

1.4 Efficiency Bound for the Treatment Effect Parameters

In this section, we present the semiparametric efficiency bound for the causal parameters $\boldsymbol{\mu}$ in the two-sample set up. As for $\tau_{t,t'}$, we view it as a differentiable function of $\boldsymbol{\mu}$ so that the efficiency bound for it follows from corollary 1.3.1. Following our discussion in the previous section, the calculation of the efficiency bounds consists of two steps: i) characterize the two-sample tangent set $\mathcal{T}(P,Q)$ and ii) project the original efficient influence functions onto the new tangent space. The original efficient influence functions refer to the ones when only the first sample is available. In our case, the first sample is the canonical experimental sample as studied in Hahn (1998), so we know these influence functions are

$$\tilde{\mu}_t(Y,T,Z) = \frac{\mathbb{1}\{T=t\}(Y-\eta_t(Z))}{\pi_t(Z)} + \eta_t(Z) - \mu_t \tag{1.14}$$

for μ_t, $t \in \{0,...,\mathtt{T}\}$, where $\eta_t(z)$ denotes a conditional mean function

$$\eta_t(z) = E[Y(t)|Z=z].$$

We also use $\zeta_t(x)$ to denote a similar conditional mean function

$$\zeta_t(x) = E[Y(t)|X=x].$$

The characterization of the two-sample tangent space and the projection of (1.14) onto it are included in the proof of Theorem 1.4.1, which states the semiparametric efficiency bounds for $\boldsymbol{\mu}$.

Theorem 1.4.1. *Under Assumption 1.2.1, the asymptoticvariance ($n \to \infty$, $m/n \to \gamma$) of any regular sequence of estimators for $\boldsymbol{\mu}$ is bounded below by V, which is a $(\mathtt{T}+1)$ dimensional square matrix defined by*

$$V_{t,t'} = E[(\eta_t(Z)-\zeta_t(X))(\eta_{t'}(Z)-\zeta_{t'}(X))] + \frac{1}{\gamma}E[(\zeta_t(X)-\mu_t)(\zeta_{t'}(X)-\mu_{t'})], \quad \text{for } t \neq t', \tag{1.15}$$

$$V_{t,t} = E[\frac{\sigma_t^2(Z)}{\pi_t(Z)}] + E[(\eta_t(Z)-\zeta_t(X))^2] + \frac{1}{\gamma}E[(\zeta_t(X)-\mu_t)^2], \tag{1.16}$$

where $\sigma_t^2(Z) = Var(Y(t)|Z)$.

To see the variance reduction, we compare (1.16) with the semiparametric efficiency bound for μ_t when only the experimental sample is used. This bound is derived in Hahn (1998) which equals to

$$E[\frac{\sigma_t^2(Z)}{\pi_t(Z)}] + E[(\eta_t(Z) - \mu_t)^2] = E[\frac{\sigma_t^2(Z)}{\pi_t(Z)}] + E[(\eta_t(Z) - \zeta_t(X))^2] + E[(\zeta_t(X) - \mu_t)^2].$$

We see that (1.16) reduces this variance by multiplying the last term with the factor $\frac{1}{\gamma} < 1$, where γ is the limit of $\frac{m}{n}$. The amount of variance reduced depends on two factors. First, $E[(\zeta_t(X) - \mu_t)^2]$, which is directly related to the explanation power of X on $Y(t)$, as $Var(Y(t)) = E[(\zeta_t(X) - \mu_t)^2] + E[Var(Y(t)|X)]$. The more covariates X explain the variation of $Y(t)$, the larger the reduction.[8] Put differently, the variance reduction is most pronounced when X consists of good predictors of $Y(t)$. In fact, $E[(\zeta_t(X) - \mu_t)^2]$ is weakly increasing in the dimension of X. This motivates using high-dimensional covariates, hence calling for ML methods to avoid over-fitting/curse of dimensionality. Second, provided that $E[(\zeta_t(X) - \mu_t)^2] \neq 0$, the larger the size of the auxiliary sample, the larger the variance reduction, as $\frac{1}{\gamma}$ is close to zero if $\frac{m}{n}$ is large.

Intuitively speaking, we can learn $\mu_t = E[Y(t)]$ by learning the conditional distribution $Y(t)|X$ and the marginal distribution of X. The former can only be learned from the experimental data while the latter can also be learned from the auxiliary data. Our efficiency bound results suggests that there should be a way to make use of the information on marginal distribution of X contained in the auxiliary data to improve estimation efficiency. That is indeed true as we will show in the next chapter.

One can also show that if the auxiliary sample takes the form of $\{(X_i, X_i^A)\}_{i=n+1}^m$, i.e. each observation contains an additional vector of covariates X_i^A not observed in the experimental sample, the efficiency bound remains unchanged.

Proposition 1.4.1. *Replace the sample $\{X_i\}_{i=n+1}^m$ with $\{(X_i, X_i^A)\}_{i=n+1}^m$ in Assumption 1.2.1 (i), where the conditional distribution of X_i^A on X_i is unrestricted, keep the rest of*

[8]To be precise, holding $Var(Y(t))$ constant, larger $E[(\zeta_t(X) - \mu_t)^2]/Var(Y(t))$, which is a measure analogues to the usual R^2 in regressions, means larger $E[(\zeta_t(X)-\mu_t)^2]$ and therefore larger variance reduction.

Assumption 1.2.1, the asymptotic variance of any regular sequence of estimators for $\boldsymbol{\mu}$ *is bounded below by* V*.*

This is to say that our assumption of X being a subvector of Z in the baseline model is not substantial.

1.5 Conclusion

In this chapter, we derived new semiparametric efficiency bounds of estimating the average treatment effect and other functions of the average potential outcomes under the two sample setup. We focused on the case where the auxiliary sample only contains observable characteristics and is sampled from the same target population as the experimental sample. The result shed light on how to efficiently combine the two samples for efficient estimation. The newly obtained bound on ATE has an intuitive expression and shows that the (maximum possible) amount of variance reduction depends positively on two factors: 1) the size of the auxiliary sample, and 2) how well the covariates predict the individual treatment effects. The latter naturally motivates having high dimensional covariates.

1.6 Appendix

1.6.1 Proofs

1.6.1.1 Proof of Lemma 1.3.1

This proof closely follows the proof of Theorem 7.2 in Van der Vaart (2000). Throughout this proof, we denote the densities $dP_{g,\frac{1}{\sqrt{n}}}$ and $dQ_{h,\frac{1}{\sqrt{n}}}$ by p_n and q_n, and their dominating measure by $d\mu$. Condition (1.4) implies $\sqrt{n}(\sqrt{q_n} - \sqrt{q})$ converges to $\frac{1}{2}h\sqrt{q}$ in $L_2(\mu)$, which

in turn implies that $\sqrt{q_n} - \sqrt{q}$ converges to zero in $L_2(\mu)$. By continuity of inner product,

$$\begin{aligned} E_Q[h] &= \int hq d\mu = \int \frac{1}{2} h\sqrt{q} \cdot 2\sqrt{q} d\mu \\ &= \lim_{n\to\infty} \int \sqrt{n}(\sqrt{q_n} - \sqrt{q})(\sqrt{q_n} + \sqrt{q}) d\mu \\ &= \lim_{n\to\infty} \int \sqrt{n}(q_n - q) d\mu \\ &= \lim_{n\to\infty} \sqrt{n}(1-1) = 0. \end{aligned}$$

Similarly, condition (1.3) implies that $E_P[g] = 0$. Note that in $L_2(\mu)$,

$$\sqrt{m-n}(\sqrt{q_n} - \sqrt{q}) = \sqrt{\frac{m-n}{n}} \sqrt{n}(\sqrt{q_n} - \sqrt{q}) \to \frac{1}{2}\sqrt{\gamma - 1} h\sqrt{q}. \tag{1.17}$$

Now we apply the Taylor expansion $\log(1+x) = x - \frac{1}{2}x^2 + x^2 R(2x)$ to the log likelihood process, where $R(x) \to 0$ as $x \to 0$. Define $H_{n,i} = 2(\frac{\sqrt{q_n}}{\sqrt{q}} - 1)$,

$$\begin{aligned} \log \prod_{i=n+1}^{m} \frac{q_n}{q}(\tilde{W}_i) =& 2 \sum_{i=n+1}^{m} \log(1 + \frac{1}{2} \cdot 2(\frac{\sqrt{q_n}}{\sqrt{q}} - 1)) \\ =& 2 \sum_{i=n+1}^{m} \log(1 + \frac{1}{2} H_{n,i}) \\ =& \sum_{i=n+1}^{m} H_{n,i} - \frac{1}{4} \sum_{i=n+1}^{m} H_{n,i}^2 + \frac{1}{2} \sum_{i=n+1}^{m} H_{n,i}^2 R(H_{n,i}). \end{aligned} \tag{1.18}$$

We proceed with analyzing the three terms in (1.18) one by one, starting with $\sum_{i=n+1}^{m} H_{n,i}$.

$$\begin{aligned} Var(\sum_{i=n+1}^{m} H_{n,i} - \frac{1}{\sqrt{m-n}} \sum_{i=n+1}^{m} \sqrt{\gamma - 1} h(\tilde{W}_i)) \leq& E[(\sqrt{m-n} H_{n,i} - \sqrt{\gamma - 1} h(\tilde{W}_i))^2] \\ =& \sqrt{2} \int \sqrt{m-n}(\sqrt{q_n} - \sqrt{q}) \qquad (1.19) \\ & - \frac{1}{2}\sqrt{\gamma - 1} h\sqrt{q})^2 d\mu \\ \to& 0, \qquad (1.20) \end{aligned}$$

where the convergence follows from (1.17). Next,

$$\begin{aligned}E[\sum_{i=n+1}^{m} H_{n,i}] =& 2(m-n)\int(\frac{\sqrt{q_n}}{\sqrt{q}}-1)qd\mu = 2(m-n)(\int\sqrt{q_n}\sqrt{q}d\mu - 1)\\ =& -(m-n)\int(\sqrt{q_n}-\sqrt{q})^2 d\mu = -\int(\sqrt{m-n}(\sqrt{q_n}-\sqrt{q}))^2 d\mu\\ \to& -\frac{1}{4}(\gamma-1)E_Q[h^2]. \end{aligned} \tag{1.21}$$

(1.20), (1.21) and Chebyshev's inequality implies

$$\sum_{i=n+1}^{m} H_{n,i} = \frac{1}{\sqrt{m-n}}\sum_{i=n+1}^{m}\sqrt{\gamma-1}h(\tilde{W}_i) - \frac{1}{4}(\gamma-1)E_Q[h^2] + o_p(1) \tag{1.22}$$

Next we analyze the second term in (1.18). Define $A_{n,i} = (m-n)H_{n,i}^2 - (\gamma-1)h^2(\tilde{W}_i)$. As shown in (1.20), $E[(\sqrt{m-n}H_{n,i} - \sqrt{\gamma-1}h(\tilde{W}_i))^2] \to 0$, we can prove[9] that $E[|A_{n,i}|] \to 0$. Which implies $\frac{1}{m-n}\sum_{i=n+1}^{m} A_{n,i} = o_p(1)$ by Markov inequality. Then apply law of large number,

$$\begin{aligned}\sum_{i=n+1}^{m} H_{n,i}^2 =& \frac{1}{m-n}\sum_{i=n+1}^{m}(\gamma-1)h^2(\tilde{W}_i) + \frac{1}{m-n}\sum_{i=n+1}^{m} A_{n,i}\\ =& (\gamma-1)E_Q[h^2] + o_p(1). \end{aligned} \tag{1.23}$$

Now for the last term in (1.18), by union bound and Markov inequality, we have for any $\varepsilon > 0$,

$$\begin{aligned}P(\max_{n+1\le i\le m}|H_{n,i}| > \sqrt{2}\varepsilon) \le& (m-n)P(|H_{n,i}| > \sqrt{2}\varepsilon)\\ =& (m-n)P(|(\gamma-1)h^2(\tilde{W}_i) + A_{n,i}| > 2(m-n)\varepsilon^2)\\ \le& (m-n)P((\gamma-1)h^2(\tilde{W}_i) > (m-n)\varepsilon^2)\\ &+ (m-n)P(|A_{n,i}| > (m-n)\varepsilon^2)\\ \le& \frac{1}{\varepsilon^2}E[(\gamma-1)h^2(\tilde{W}_i)\mathbb{1}\{(\gamma-1)h^2(\tilde{W}_i) > (m-n)\varepsilon^2\}] + \frac{1}{\varepsilon^2}E[|A_{n,i}|]\\ \to& 0\end{aligned}$$

[9]If X_n converge to X in L_2 (i.e. $E[(X_n-X)^2]^{\frac{1}{2}} \to 0$) then $E[|X_n^2 - X^2|] \to 0$. Indeed, $E[|X_n^2 - X^2|] = E[|(X_n+X)(X_n-X)|] \le E[(X_n+X)^2]^{\frac{1}{2}}E[(X_n-X)^2]^{\frac{1}{2}}$, where $E[(X_n+X)^2]^{\frac{1}{2}}$ is bounded by $2E[X^2]^{\frac{1}{2}} + E[(X_n-X)^2]^{\frac{1}{2}}$ through triangular inequality.

Since $R(x) \to 0$ as $x \to 0$, we have

$$\sum_{i=n+1}^{m} H_{n,i}^2 R(H_{n,i}) \leq \max_{n+1 \leq i \leq m} |R(H_{n,i})| \sum_{i=n+1}^{m} H_{n,i}^2 = o_p(1) \cdot O_p(1) = o_p(1). \tag{1.24}$$

Plug (1.22), (1.23) and (1.24) into (1.18) we get

$$\log \prod_{i=n+1}^{m} \frac{q_n}{q}(\tilde{W}_i) = \frac{1}{\sqrt{m-n}} \sum_{i=n+1}^{m} \sqrt{\gamma - 1} h(\tilde{W}_i) - \frac{1}{2}(\gamma - 1) E_Q[h^2] + o_p(1).$$

Similarly, we can prove that

$$\log \prod_{i=1}^{n} \frac{p_n}{p}(W_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(W_i) - \frac{1}{2} E_P[g^2] + o_p(1).$$

The lemma follows as logarithm of the product equals the sum of logarithms.

1.6.1.2 Proof of Lemma 1.3.2

Before we prove Lemma 1.3.2, we state a version of the asymptotic representation theorem as below.

Theorem 1.6.1. *(Asymptotic Representation Theorem) Let $\{Q_{h,n} : h \in \mathbb{R}^k\}$ and P_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$, and let $T_n : \Omega_n \to \mathbb{R}^d$ be a sequence of random vectors that converge in distribution to S_h under $Q_{h,n}$. Suppose*

$$(T_n, \frac{dQ_{h,n}}{dP_n}) \overset{P_n}{\rightsquigarrow} (S_0, \exp(h'\Delta - \frac{1}{2} h' J h)) \tag{1.25}$$

where $\Delta \sim N(0, J)$, then there exists a randomized statistic T in the experiment $(N(h, J^{-1}) : h \in \mathbb{R}^k)$ such that $S_h \sim T$ for every h.

The above theorem can be proved using the general version of Le Cam's third lemma (Theorem 6.6 in Van der Vaart (2000)) and follow the steps in the proof of theorem 7.10 in Van der Vaart (2000). By saying T is a randomized statistic in the experiment $(N(h, J^{-1}) : h \in \mathbb{R}^k)$, we mean that T is a measurable function of (X, U) where $X \sim N(h, J^{-1})$, and U independent of X. h is understood as the unknown parameter.

Now we prove Lemma 1.3.2. This proof follows the steps in the proof of Theorem 25.20 (Convolution) in Van der Vaart (2000). Pick an orthonormal base $\mathbf{g}_p = (g_1, g_2, ..., g_p)$ in $\overline{\mathcal{T}}(P, Q)$, for any $p \times 1$ vector a we find a score function $a'\mathbf{g}_p$. Denote the submodel with score function $a'\mathbf{g}_p$ as $\iota \to (P_{a,\iota}, Q_{a,\iota})$, By Lemma 1.3.1 and (1.7), we have

$$\log \prod_{i=1}^{n} \frac{dP_{a,\frac{1}{\sqrt{n}}}}{dP}(W_i) \prod_{i=n+1}^{m} \frac{dQ_{a,\frac{1}{\sqrt{n}}}}{dQ}(\tilde{W}_i) \overset{0}{\rightsquigarrow} N(-\frac{1}{2}E[a'\mathbf{g}_p\mathbf{g}_p'a], E[a'\mathbf{g}_p\mathbf{g}_p'a]). \tag{1.26}$$

Since $E[a'\mathbf{g}_p\mathbf{g}_p'a] = a'I_pa$, the right-hand side of (1.26) can be written as $a'\Delta - \frac{1}{2}a'I_pa$, where Δ follows standard normal distribution $N(0, I_p)$. Next, take any regular estimator $\hat{\phi}_n$, by (1.12),

$$\begin{aligned}
\sqrt{n}(\hat{\phi}_n - \phi(P_{a,\frac{1}{\sqrt{n}}})) =& \sqrt{n}(\hat{\phi}_n - \phi(P)) - \sqrt{n}(\phi(P_{a,\frac{1}{\sqrt{n}}}) - \phi(P)) \\
=& \sqrt{n}(\hat{\phi}_n - \phi(P)) - E[\bar{\phi}\mathbf{g}_p'a] + o(1) \\
=& \sqrt{n}(\hat{\phi}_n - \phi(P)) - Aa + o(1),
\end{aligned} \tag{1.27}$$

where $A := E[\bar{\phi}\mathbf{g}_p']$. Denote the limiting distribution of $\sqrt{n}(\hat{\phi}_n - \phi(P))$ under (P_0, Q_0) by S_0 and its limiting distribution under $(P_{a,\frac{1}{\sqrt{n}}}, Q_{a,\frac{1}{\sqrt{n}}})$ by S_a. S_0 and S_a exist by the regularity condition (1.13) and (1.27). We have

$$(\sqrt{n}(\hat{\phi}_n - \phi(P)), \prod_{i=1}^{n} \frac{dP_{a,\frac{1}{\sqrt{n}}}}{dP}(W_i) \prod_{i=n+1}^{m} \frac{dQ_{a,\frac{1}{\sqrt{n}}}}{dQ}(\tilde{W}_i)) \overset{0}{\rightsquigarrow} (S_0, \exp(a'\Delta - \frac{1}{2}a'I_pa)). \tag{1.28}$$

This gives us condition (1.25) in theorem 1.6.1. Hence, there exists a randomized statistic T in the experiment $(N(a, I_p) : a \in \mathbb{R}^p)$ such that $T \sim S_a$ for every a. This is to say that the limiting distribution of $\sqrt{n}(\hat{\phi}_n - \phi(P_{a,\frac{1}{\sqrt{n}}}))$ under $(P_{a,\frac{1}{\sqrt{n}}}, Q_{a,\frac{1}{\sqrt{n}}})$ is matched by the distribution of $T - Aa$. Since $\hat{\phi}_n$ is regular, the distribution of $T - Aa$ doesn't depend on a, meaning T is an equivariant estimator for Aa in the simple normal experiment. We can write

$$\sqrt{n}(\hat{\phi}_n - \phi(P_{a,\frac{1}{\sqrt{n}}})) \overset{a}{\rightsquigarrow} (T - Aa) \sim L, \text{ for any } a \in \mathbb{R}^p. \tag{1.29}$$

By Proposition 8.4 in Van der Vaart (2000), the lower bound on the variance for such equivariant estimators is AA'. As a result, the lower bound for the asymptotic variance of $\hat{\phi}_n$ is also AA'. Note that

$$AA' = AE[\mathbf{g}_p\mathbf{g}_p']A' = E[A\mathbf{g}_p\mathbf{g}_p'A'],$$

where $A\mathbf{g}_p = E[\bar{\phi}\mathbf{g}_p']\mathbf{g}_p = E[\bar{\phi}\mathbf{g}_p']E[\mathbf{g}_p\mathbf{g}_p']^{-1}\mathbf{g}_p$ is the projection (component-wise) of $\bar{\phi}$ onto the linear span of $\mathbf{g}_p$. Since $\bar{\phi}$ is by definition in the closed linear span of $\mathcal{T}(P,Q)$, we can choose $\mathbf{g}_p$ to make $A\mathbf{g}_p$ arbitrarily close to $\bar{\phi}$ and hence AA' is arbitrarily close to $E[\bar{\phi}\bar{\phi}']$. Therefore, the asymptotic variance of $\hat{\phi}_n$ is bounded below by $E[\bar{\phi}\bar{\phi}']$.

1.6.1.3 Proof of Theorem 1.4.1

The proof consists of two steps. In the first step, we characterize the two-sample tangent space described in Section 1.3. In the second step, we project the original influence functions onto the tangent space. Assumption 1.3.1 are satisfied as shown in Hahn (1998), the result would then follow from Lemma 1.3.2.

STEP ONE. We follow the same method as in Hahn (1998) to characterize the tangent space. We need to consider smooth parametric submodels $\iota \mapsto (P_\iota, Q_\iota)$, where P is the distribution for (Y_i, T_i, X_i, X_i^E), $i \leq n$, in the first sample and Q is the distribution for X_i, $n < i \leq m$ in the second sample. We use (Y, T, X, X^E) to denote the random variables under distribution P and $\tilde{X}$ under Q.

Start with distribution P. Denote $\pi_t(x, x^E) = P(T = t|X = x, X^E = x^E)$ and $f_X(x)$ the density function of X. Let $f_t(y|x, x^E)$ denote the conditional density of Y on (X, X^E) and $T = t$, $f_E(x^E|x)$ the conditional density of X^E on X. The density of (Y, T, X, X^E) under distribution P is then equal to

$$\prod_{j=0}^{\mathtt{T}} (f_j(y|x, x^E)\pi_j(x, x^E))^{\mathbb{1}\{t=j\}} \cdot f_E(x^E|x)f_X(x). \tag{1.30}$$

For distribution Q, by Assumption 1.2.1 iv), the density function for $\tilde{X}$ should be the same

as X, hence it is $f_X(\tilde{x})$. Then a regular one-dimensional parametric submodel $\iota \mapsto (P_\iota, Q_\iota)$ is

$$(\prod_{j=0}^{\mathtt{T}}(f_j(y|x,x^E;\iota)\pi_j(x,x^E))^{\mathbb{1}\{t=j\}} \cdot f_E(x^E|x;\iota)f_X(x;\iota), \quad f_X(\tilde{x};\iota)),$$

which equals to

$$(\prod_{j=0}^{\mathtt{T}}(f_j(y|x,x^E)\pi_j(x,x^E))^{\mathbb{1}\{t=j\}} \cdot f_E(x^E|x)f_X(x), \quad f_X(\tilde{x}))$$

when $\iota = 0$. Note that the propensity score $\pi_j(x, x^E)$ is not parametrized as it is known in the statistical problem. Also note that the density functions of X and $\tilde{X}$ are the same for every ι. Taking derivatives to the log densities gives us the two components of our score function that satisfies (1.3) and (1.4) respectively.

$$\sum_{j=0}^{\mathtt{T}} \mathbb{1}\{t=j\}s_j(y|x,x^E) + t_E(x^E|x) + t_X(x), \quad \text{and} \quad t_X(\tilde{x}), \tag{1.31}$$

where

$$\begin{aligned} s_j(y|x,x^E) =& \frac{\partial}{\partial \iota}\Big|_{\iota=0} \log f_j(y|x,x^E;\iota), \\ t_E(x^E|x) =& \frac{\partial}{\partial \iota}\Big|_{\iota=0} \log f_E(x^E|x;\iota), \\ t_X(x) =& \frac{\partial}{\partial \iota}\Big|_{\iota=0} \log f_X(x;\iota). \end{aligned}$$

We note that the first component in (1.31) is the same as the original score function in the one-sample problem studied in Hahn (1998). Next, by (1.8), we obtain a tangent space for the two-sample problem as

$$\mathcal{T}(P,Q) = \{\sum_{j=0}^{\mathtt{T}} \mathbb{1}\{t=j\}s_j(y|x,x^E) + t_E(x^E|x) + t_X(x) + \sqrt{\gamma-1}t_X(\tilde{x}) : \text{“conditions”}\}$$

where the “conditions” are

$$\int s_j(y|x,x^E)f_j(y|x,x^E)dy = 0, \quad \forall (x,x^E), j \in \{0,..,\mathtt{T}\}, \tag{1.32}$$

$$\int t_E(x^E|x)f_E(x^E|x)dx^E = 0, \quad \forall x, \tag{1.33}$$

$$\int t_X(x)f_X(x)dx = 0. \tag{1.34}$$

STEP TWO. The original influence function is (1.14), we now find its projection on $\mathcal{T}(P,Q)$. We can check that the following function is indeed the projection,

$$\bar{\mu}_t = \frac{\mathbb{1}\{T=t\}(Y-\eta_t(X,X^E))}{\pi_t(X,X^E)} + (\eta_t(X,X^E)-\zeta_t(X)) \\ + \frac{1}{\gamma}(\zeta_t(X)-\mu_t) + \frac{\sqrt{\gamma-1}}{\gamma}(\zeta_t(\tilde{X})-\mu_t).$$

To see this, we check that the projection error is orthogonal to all the elements in $\mathcal{T}(P,Q)$, namely $E[(\tilde{\mu}-\bar{\mu})\tilde{g}]=0$ for any $\tilde{g}\in\mathcal{T}(P,Q)$. Note that the expectation here and onward in this proof are all taken with respect to the product measure $P\otimes Q$. We have

$$\tilde{\mu}-\bar{\mu} = \frac{\gamma-1}{\gamma}(\zeta_t(X)-\mu_t) - \frac{\sqrt{\gamma-1}}{\gamma}(\zeta_t(\tilde{X})-\mu_t),$$

and consider any generic score function

$$\tilde{g} = \sum_{j=0}^{\mathtt{T}} \mathbb{1}\{T=j\}s_j(Y|X,X^E) + t_E(X^E|X) + t_X(X) + \sqrt{\gamma-1}t_X(\tilde{X}).$$

By condition (1.32) to (1.34) and that $\tilde{X}$ are independent from (Y,T,X,X^E), we get that

$$E[(\tilde{\mu}-\bar{\mu})\tilde{g}] = \frac{\gamma-1}{\gamma}E[(\zeta_t(X)-\mu_t)t_X(X)] - \frac{\gamma-1}{\gamma}E[(\zeta_t(\tilde{X})-\mu_t)t_X(\tilde{X})] = 0,$$

the second equality is due to X and $\tilde{X}$ having the same marginal distribution under $P\otimes Q$. The vector-valued efficient influence function $\bar{\boldsymbol{\mu}}$ is obtained by stacking all the projections $\bar{\mu}_t$, $t\in\{0,...,\mathtt{T}\}$ together. Then we get efficient bound $E[\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}']$ by Lemma 1.3.2, which equals to V as defined in theorem 1.4.1.

1.6.2 Limit of the Two Sample Sizes

In this subsection we provide a simple example to show that the $\frac{m}{n}\to\gamma$ asymptotic framework is sensible in the two-sample set up, when γ is naturally replaced by $\frac{m}{n}$ in the variance estimation. Recall that the goal of asymptotic analysis is primarily to provide approximations to the finite-sample distributions of statistics.

Consider two i.i.d. samples $\{W_i\}_{i=1}^n$ and $\{\tilde{W}_i\}_{i=n+1}^m$. Note that the size of the second sample is $m-n$. Assume central limit theorem applies, we have

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} W_i - E[W]) \rightsquigarrow N(0, Var(W)), \quad \text{as } n \to \infty, \tag{1.35}$$

$$\sqrt{m-n}(\frac{1}{m-n}\sum_{i=n+1}^{m} \tilde{W}_i - E[\tilde{W}]) \rightsquigarrow N(0, Var(\tilde{W})), \quad \text{as } m-n \to \infty. \tag{1.36}$$

These results justify the following asymptotic approximations for the finite sample distributions of the sample means

$$\frac{1}{n}\sum_{i=1}^{n} W_i \sim N(E[W], \frac{Var(W)}{n}), \tag{1.37}$$

$$\frac{1}{m-n}\sum_{i=n+1}^{m} \tilde{W}_i \sim N(E[\tilde{W}], \frac{Var(\tilde{W})}{m-n}). \tag{1.38}$$

So far, we have let the two sample sizes go to infinity separately without specifying the limiting ratio between them. Now suppose we choose to index the statistical experiment by n only, hence thinking of m as a sequence m_n indexed by n. Let $\frac{m}{n} \to \gamma$, we have similar to (1.36)

$$\begin{aligned}
&\sqrt{n}(\frac{1}{m-n}\sum_{i=n+1}^{m} \tilde{W}_i - E[\tilde{W}]) \\
=&\frac{\sqrt{n}}{\sqrt{m-n}}\sqrt{m-n}(\frac{1}{m-n}\sum_{i=n+1}^{m} \tilde{W}_i - E[\tilde{W}]) \\
\rightsquigarrow& N(0, \frac{1}{\gamma-1}Var(\tilde{W})), \quad \text{as } n \to \infty.
\end{aligned}$$

This limiting distribution would provide the following approximation for the finite sample distribution of the sample mean

$$\frac{1}{m-n}\sum_{i=n+1}^{m} \tilde{W}_i \sim N(E[\tilde{W}], \frac{Var(\tilde{W})}{(\gamma-1)n}).$$

If we set $\gamma = \frac{m}{n}$, as we would when we estimate the asymptotic variance, then this approximated distribution is identical to the one in (1.38). Also note that the variances of these approximations are actually exact.

CHAPTER 2

Efficient Machine Learning Imputation for Estimating ATE with a Combined Sample

2.1 Introduction

In this chapter, I study how to use modern debiased machine learning techniques (Chernozhukov et al., 2018) to efficiently estimate the average treatment effect (ATE) with a combined sample. Specifically, I continue with the setup discussed in the previous chapter, where a canonical experimental sample is combined with an auxiliary observational sample that only contains observable characteristics (features or covariates). As suggested by the semiparametric efficiency bound derived in the previous chapter, the (maximum possible) amount of variance reduction through incorporating the auxiliary sample depends positively on two factors: 1) the size of the auxiliary sample, and 2) how well the observable characteristics predict the individual treatment effects. The latter naturally motivates having high dimensional covariates. Therefore, ML tools are called for to avoid over-fitting/curse of dimensionality.

I propose a two-stage machine learning (ML) imputation estimator that achieves the efficiency bound derived in chapter 1, so that no other regular estimators for ATE can have lower asymptotic variance in the same setting. This estimator involves two steps. In the first step, conditional average potential outcome functions are estimated nonparametrically via ML, which are then used to impute the unobserved potential outcomes for every unit in both samples. In the second step, the imputed potential outcomes are aggregated together

in a robust way to produce the final estimate. Adopting the cross-fitting technique proposed in Chernozhukov et al. (2018), our two-step estimator can use a wide range of supervised ML tools as the first-step, while maintaining valid inference to construct confidence intervals and perform hypothesis tests. These ML tools include LASSO and post-LASSO, elastic nets, neural nets, regression trees and random forests, or ensemble/aggregated methods of the above. In fact, any method that estimates the relevant conditional mean functions consistently in $L_2(P)$ norm, with no rate requirement[1], will lead to efficiency through our two-step procedure. I also show that cross-fitting is not necessary when the first-step is done via LASSO or post-LASSO. Furthermore, our estimator is robust in the sense that it remains consistent and $\sqrt{n}$ normal (no longer efficient) even if the first step ML estimator is inconsistent.

A brief discussion on related literature and the motivation for improving estimation efficiency can be found in Section 1.1 in the previous chapter. The rest of this chapter proceeds as follows. In Section 2.2, I briefly restate the model for completeness. In the next two sections, I introduce two efficient imputation estimators and derive their asymptotic distributions: Section 2.3 adopts the cross-fitting technique proposed in Chernozhukov et al. (2018), which allows us to use general first-step ML methods that satisfy weak high-level conditions; Section 2.4 focuses on using LASSO/post-LASSO in the first-step, in which case cross-fitting is not necessary. Section 2.5 presents simulation results and Section 2.6 concludes. All proofs are collected in the appendix which forms Section 2.7.

2.2 Model

We consider two i.i.d. random samples $\{(Y_i, T_i, Z_i)\}_{i=1}^n$ and $\{X_i\}_{i=n+1}^m$ sampled separately from the same large population. The first sample is an experimental sample from the canonical potential outcome model (Neyman, 1923; Rubin, 1974) and the second sample is an

[1]Due to propensity score being a known function in the experimental setup.

auxiliary sample on the covariates from the same target population. X_i is a subvector of Z_i and all X_i, $1 \leq i \leq m$, follow the same distribution. More details on the model can be found in Section 1.2, where all modeling assumptions are summarized in Assumption 1.2.1.

We are interested in estimating the average potential outcomes $\mu_t = E[Y(t)]$ or any differentiable functions of them including the average treatment effects $\tau_{t,t'} = \mu_t - \mu_{t'}, t \neq t'$. These parameters are identified in the first sample alone by the standard identification Assumptions 1.2.1 (ii) and (iii). In particular,

$$\mu_t = E[E[Y|Z, T = t]]. \tag{2.1}$$

This chapter shows an efficient way of combining the two samples to estimate these parameters. We allow both Z and X to be high-dimensional to incorporate the cases when the number of covariates in the data is very large (even larger than n) or that these vectors contain many flexible transformations (series base functions, e.g. polynomials, regression splines) of the original covariates.[2] Using high-dimensional covariates here is well-motivated as we have shown earlier that the asymptotic variances (and efficiency bounds) are most reduced when X consists of good predictors of the potential outcomes.

As in the previous chapter, we focus on estimating the response vector $\boldsymbol{\mu} = (\mu_0, ..., \mu_{\mathrm{T}})'$ with $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathrm{T}})'$. The efficient estimator for any known differentiable function of $\boldsymbol{\mu}$ can be obtained by plugging in $\hat{\boldsymbol{\mu}}$. For example, treatment effect $\tau_{t,t'}$ can be efficiently estimated by $\hat{\tau}_{t,t'} = \hat{\mu}_t - \hat{\mu}_{t'}$.

Recall that we denote the conditional mean functions by

$$\zeta_t(x) = E[Y(t)|X = x],$$
$$\eta_t(z) = E[Y(t)|Z = z].$$

The first step our proposed estimator is to estimate these condtional mean functions.

[2]Many estimation results in high-dimensional literature are established under the asymptotic framework where the dimension of the covariates increases with sample size n. Our estimation results in this chapter remains valid under this asymptotic framework. However, for discussions on semiparametric efficiency and results in the previous chapter, the dimension of the (original) covariates should be fixed.

2.3 General ML First Step with Cross-fitting

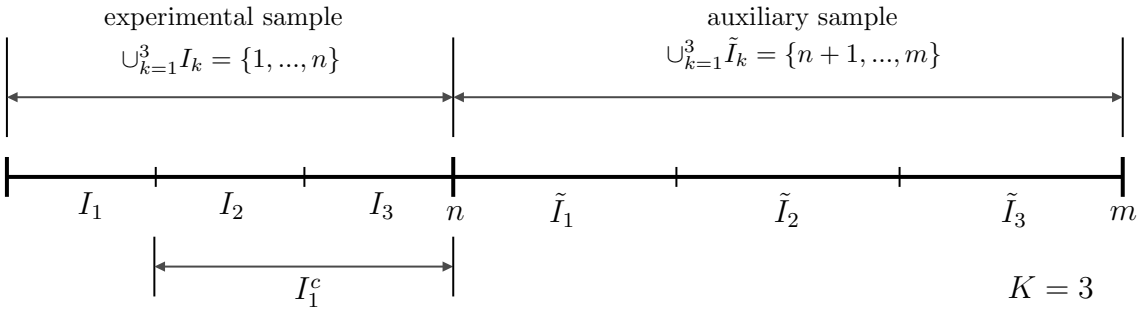


Figure 2.1: Illustration of the sample partitions for cross-fitting when $K = 3$.

Assume sample size n and m are divisible by K in order to simplify the notation. Let $\{I_k\}_{k=1}^K$ be a K-fold equal-sized random[3] partition of the indices $\{1, ..., n\}$ and $\{\tilde{I}_k\}_{k=1}^K$ of $\{n+1, ..., m\}$. Define $I_k^c = \{1, ..., n\} \setminus I_k$. See figure 2.1 for an illustration when $K = 3$. For each fold $k \in \{1, ..., K\}$, construct ML estimator $\hat{\zeta}_{t,k}$ for $\zeta_t(x) = E[Y(t)|X = x]$ and $\hat{\eta}_{t,k}$ for $\eta_t(z) = E[Y(t)|Z = z]$, using only the data from I_k^c (which is a subset of the experimental sample). Under Assumption 1.2.1, we have

$$\eta_t(z) = E[Y(t)|Z = z] = E[Y|T = t, Z = z], \tag{2.2}$$

$$\zeta_t(x) = E[Y(t)|X = x] = E[\frac{Y\mathbb{1}\{T = t\}}{\pi_t(Z)}|X = x]. \tag{2.3}$$

Hence, to obtain $\hat{\eta}_{t,k}$, we use the $T = t$ subsample in I_k^c, namely $\{i \in I_k^c : T_i = t\}$, and adopt supervised ML methods that predicts Y with Z. For $\hat{\zeta}_{t,k}$, we use the complete I_k^c and adopt supervised ML methods that predicts $\frac{Y\mathbb{1}\{T=t\}}{\pi_t(Z)}$ with X.[4] Different structured assumptions on ζ_t and η_t calls for the use of different machine learning tools. These tools include LASSO, neural nets, regression trees and random forests, or ensemble/aggregated

[3]We have assumed i.i.d. sample, so randomization in this partition process is technically unnecessary. The statistical model we consider does not reflect this randomization. Random partition is often advised in practice in case the data has been sorted by value.

[4]If we have $T_i \perp\!\!\!\perp \{Y_i(t)\}_{t=0}^{\intercal}|X$, then ζ_t can be estimated similarly as η_t from the $T = t$ subsample in I_k^c. This is recommended over working with $\frac{Y\mathbb{1}\{T=t\}}{\pi_t(Z)}$ as the latter adds noise to the estimation.

methods of the above. There are performance guarantees for most of these ML methods that make it possible to satisfy the following condition.

Assumption 2.3.1. *For every* $t \in \{0, ..., \mathtt{T}\}$ *and* $k \in \{1, ..., K\}$, *as* $n \to \infty$, $E_Z[(\hat{\eta}_{t,k}(Z) - \eta_t(Z))^2] = o_p(1)$ *and* $E_X[(\hat{\zeta}_{t,k}(X) - \zeta_t(X))^2] = o_p(1)$.

This assumption requires that the first step estimators are consistent for the conditional mean functions η_t and ζ_t in L_2-norm, with arbitrary rate of convergence. The consistency of the first-step estimators are needed for efficiency. We will also show that if the limiting functions η_t and ζ_t in Assumption 2.3.1 are replaced by any other square integrable functions, the proposed two-step estimator is still consistent for $\boldsymbol{\mu}$ and $\sqrt{n}$-normal, but without achieving the efficiency bound.

To construct the two-step estimator, for each $k \in \{1, ..., K\}$, we use $\{\hat{\zeta}_{t,k}\}_{t=0}^{\mathtt{T}}$ and $\{\hat{\eta}_{t,k}\}_{t=0}^{\mathtt{T}}$ to impute $\{Y_i(t)\}_{t=0}^{\mathtt{T}}$ for observations $i \in I_k \cup \tilde{I}_k$. Use the notation $k[i]$ to denote the data fold in which observation i belongs. The two-step estimator $\hat{\mu}_t$ is then given by

$$\hat{\mu}_t = \frac{1}{m}\sum_{i=1}^{m} \hat{\zeta}_{t,k[i]}(X_i) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}_{t,k[i]}(Z_i))}{\pi_t(Z_i)} + \hat{\eta}_{t,k[i]}(Z_i) - \hat{\zeta}_{t,k[i]}(X_i)\right). \tag{2.4}$$

The estimator for the vector $\boldsymbol{\mu} = (\mu_0, ..., \mu_{\mathtt{T}})'$ is then obtained by stacking $\hat{\mu}_t$, $t \in \{0, ..., \mathtt{T}\}$, together, i.e. $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathtt{T}})'$. The structure in (2.4) makes the estimator robust to the bias from the first-step ML estimation. To heuristically illustrate this, we can write

$$\begin{aligned}
\hat{\mu}_t =& E[\zeta_t(X_i)] + E[\frac{\mathbb{1}\{T_i = t\}(Y_i - \eta_t(Z_i))}{\pi_t(Z_i)} + \eta_t(Z_i) - \zeta_t(X_i)] + o_p(1) && (2.5)\\
=& E[\frac{\mathbb{1}\{T_i = t\}(Y_i - \eta_t(Z_i))}{\pi_t(Z_i)} + \eta_t(Z_i)] + o_p(1) && (2.6)\\
=& E[Y(t)] + o_p(1). && (2.7)
\end{aligned}$$

The moment in (2.6) is the well-known doubly robust score. (2.7) is obtained by law of iterated expectation on Z_i and Assumption 1.2.1 (ii). We see that (2.5) to (2.7) would still follow through for any other integrable functions in place of ζ_t and η_t. This suggests that the

estimator $\hat{\mu}_t$ is consistent even when the first-step estimators are not consistent but converge to some other limiting functions than ζ_t and η_t. (e.g. estimated using simple OLS regressions or under other mis-specified functional forms.) Furthermore, as shown in Chernozhukov et al. (2018), this type of robustness property (more generally, Neyman orthogonality) makes the second-step estimator insensitive to the bias of the first-step non-parametric estimators, allowing the adoption of machine learning methods.

The following theorem state the asymptotic distribution of $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathtt{T}})'$.

Theorem 2.3.1 (Cross-fitting with General ML First Step)**.** *Under Assumptions 1.2.1 and 2.3.1, consider the estimator $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathtt{T}})'$, where $\hat{\mu}_t$, $t \in \{0, ..., \mathtt{T}\}$, are defined by (2.4). Furthermore, assume that for every t, $Y(t)$, $\eta_t(Z)$ and $\zeta_t(X)$ all have finite second moment, then as $n \to \infty$ and $\frac{m}{n} \to \gamma$,*

$$V^{-\frac{1}{2}}\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to N(0, I_{\mathtt{T}+1}).$$

V is a $(\mathtt{T}+1)$ dimensional square matrix defined by

$$V_{t,t'} = E[(\eta_t(Z) - \zeta_t(X))(\eta_{t'}(Z) - \zeta_{t'}(X))] + \frac{1}{\gamma}E[(\zeta_t(X) - \mu_t)(\zeta_{t'}(X) - \mu_{t'})], \quad \text{for } t \neq t', \tag{2.8}$$

$$V_{t,t} = E[\frac{\sigma_t^2(Z)}{\pi_t(Z)}] + E[(\eta_t(Z) - \zeta_t(X))^2] + \frac{1}{\gamma}E[(\zeta_t(X) - \mu_t)^2], \tag{2.9}$$

where $\sigma_t^2(Z) = Var(Y(t)|Z)$.

Proof is collected in appendix 2.7.2.

We see that the asymptotic variance is identical to the semiparametric efficiency bound in theorem 1.4.1, hence our proposed estimator is efficient. In other words, there is no regular estimator that can have lower[5] asymptotic variance than V.

Corollary 2.3.1. *Under Assumptions 1.2.1 and 2.3.1, the estimator $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathtt{T}})'$, where $\hat{\mu}_t$, $t \in \{0, ..., \mathtt{T}\}$ are defined by (2.4), is semiparametric efficient.*

[5]As V is a matrix, we say $\tilde{V}$ is lower than V if $V - \tilde{V}$ is positive definite.

To estimate the variance matrix, we use the sample counterparts of (2.8) and (2.9). Define the $(\mathtt{T}+1)\times(\mathtt{T}+1)$ dimensional variance estimator $\hat{V}$ element-wise as the following. For $t \neq t'$,

$$\begin{aligned}\hat{V}_{t,t'} =& \frac{1}{n}\sum_{i=1}^{n}[(\hat{\eta}_{t,k[i]}(Z_i) - \hat{\zeta}_{t,k[i]}(X_i))(\hat{\eta}_{t',k[i]}(Z_i) - \hat{\zeta}_{t',k[i]}(X_i))] \\ &+ \frac{n}{m^2}\sum_{i=1}^{m}[(\hat{\zeta}_{t,k[i]}(X_i) - \hat{\mu}_t)(\hat{\zeta}_{t',k[i]}(X_i) - \hat{\mu}_{t'})], \end{aligned} \tag{2.10}$$

and for the diagonal elements,

$$\begin{aligned}\hat{V}_{t,t} =& \frac{1}{n}\sum_{i=1}^{n}[\frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}_{t,k[i]}(Z_i))^2}{\pi_t(Z_i)^2}] + \frac{1}{n}\sum_{i=1}^{n}[(\hat{\eta}_{t,k[i]}(Z_i) - \hat{\zeta}_{t,k[i]}(X_i))^2] \\ &+ \frac{n}{m^2}\sum_{i=1}^{m}[(\hat{\zeta}_{t,k[i]}(X_i) - \hat{\mu}_t)^2]. \end{aligned} \tag{2.11}$$

The following theorem states that the variance estimator is consistent.

Theorem 2.3.2 (Variance Estimation)**.** *Under the conditions of theorem 2.3.1, further assumes that for some $\delta > 0$ and every $t \in \{0, ..., \mathtt{T}\}$, $E[(Y(t)-\eta_t(Z))^{2+\delta}]$, $E[(\eta_t(Z)-\zeta_t(X))^{2+\delta}]$ and $E[(\zeta_t(X) - \mu_t)^{2+\delta}]$ are all bounded, then $\hat{V} \xrightarrow{p} V$.*

The proof is collected in appendix 2.7.2.

Theorems 2.3.1 and 2.3.2 can be used for standard construction of confidence intervals for treatment effects $\tau_{t,t'} = \mu_t - \mu_{t'}$.

Corollary 2.3.2. *Under the conditions of theorems 2.3.1 and 2.3.2, suppose we are interested in the parameter $\ell'\boldsymbol{\mu}$ for some $(\mathtt{T}+1)\times 1$ vector ℓ, then*

$$|P(\ell'\boldsymbol{\mu} \in [\ell'\hat{\boldsymbol{\mu}} \pm \Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\ell'\hat{V}\ell/n}]) - (1-\alpha)| \to 0.$$

At last, we state that the estimator $\boldsymbol{\mu}$ remains consistent and asymptotically normal when the first-step estimators are inconsistent. We relax Assumption 2.3.1 by changing the limiting functions ζ_t and η_t to some arbitrary square integrable functions $\tilde{\zeta}_t$ and $\tilde{\eta}_t$.

Assumption 2.3.2. *For every $t \in \{0, ..., \mathtt{T}\}$ and $k \in \{1, ..., K\}$, as $n \to \infty$, $E_Z[(\hat{\eta}_{t,k}(Z) - \tilde{\eta}_t(Z))^2] = o_p(1)$ and $E_X[(\hat{\zeta}_{t,k}(X) - \tilde{\zeta}_t(X))^2] = o_p(1)$, where $\tilde{\zeta}_t$ and $\tilde{\eta}_t$ are some arbitrary square integrable functions.*

Proposition 2.3.1. *Under Assumptions 1.2.1 and 2.3.2, consider the estimator $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathtt{T}})'$, where $\hat{\mu}_t$, $t \in \{0, ..., \mathtt{T}\}$, are defined by (2.4). Furthermore, assume that for every t, $E[Y(t)^2]$ is bounded, then as $n \to \infty$ and $\frac{m}{n} \to \gamma$,*

$$\tilde{V}^{-\frac{1}{2}}\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \to N(0, I_{\mathtt{T}+1}),$$

where $\tilde{V}$ is a $(\mathtt{T}+1)$ dimensional square matrix defined by (2.28) in appendix 2.7.2.4.

The proof is collected in appendix 2.7.2.

2.4 LASSO/post-LASSO First Step

In this subsection, we focus on using LASSO/post-LASSO in the first-step estimation of η_t and ζ_t. In this case, cross-fitting is not necessary under a couple of additional assumptions. LASSO/post-LASSO methods are suited when the conditional mean functions are believed to be approximately sparse, which is made precise by the following assumption. For a p dimensional vector δ, define the norm $\|\delta\|_0 = \sum_{j=1}^{p} \mathbb{1}\{|\delta_j| > 0\}$.

Assumption 2.4.1 (Approximated Sparsity)**.** *For every $t \in \{0, 1, ..., \mathtt{T}\}$, $\zeta_t(x)$ and $\eta_t(z)$ are well-approximated by sparse linear (in coefficients) functions, i.e. for some p_ζ and p_η dimensional vector coefficients $\beta_{\zeta,t}$ and $\beta_{\eta,t}$,*

$$\zeta_t(x) = f_\zeta(x)'\beta_{\zeta,t} + r_{\zeta,t}(x), \quad \eta_t(z) = f_\eta(z)'\beta_{\eta,t} + r_{\eta,t}(z),$$
$$\max\{\max_{0\le t\le \mathtt{T}} \|\beta_{\zeta,t}\|_0, \max_{0\le t\le \mathtt{T}} \|\beta_{\eta,t}\|_0\} \le s = o(n),$$
$$\max\{\max_{0\le t\le \mathtt{T}} E[r_{\zeta,t}(X)^2]^{\frac{1}{2}}, \max_{0\le t\le \mathtt{T}} E[r_{\eta,t}(Z)^2]^{\frac{1}{2}}\} = O(\sqrt{\frac{s}{n}}),$$

where $f_\zeta(x) := (f_{\zeta,1}(x), ..., f_{\zeta,p_\zeta}(x))'$ and $f_\eta(z) := (f_{\eta,1}(z), ..., f_{\eta,p_\eta}(z))'$ are the vectors of regressors.

The vectors of regressors f_η and f_ζ could be composed of polynomials, dummies, B-splines and various other series terms, or only of the original covariate themselves in high-dimensional settings. This assumption requires that at most s among all p_η (or p_ζ) of the regressor terms should be enough to approximate the conditional mean functions well, while the identities of these s terms can remain unknown. For more discussion on the sparsity assumption, please see Belloni and Chernozhukov (2011); Belloni et al. (2012).

Recall that η_t and ζ_t are identified by Equations (2.2) and (2.3). Hence, to estimate η_t, we can run LASSO regression of Y on $f_\eta(Z)$ in the subsample $\{i : 1 \leq i \leq n, T_i = t\}$. Denote $n_t = \sum_{i=1}^n \mathbb{1}\{T = t\}$,

$$\hat{\beta}_{\eta,t}^{LS} = \arg\min_{b \in \mathbb{R}^{p_\eta}} \frac{1}{n_t} \sum_{i=1}^{n} \mathbb{1}\{T = t\}(Y_i - f_\eta(Z_i)'b)^2 + \frac{\lambda_\eta}{n_t} \|\hat{\Lambda}_{\eta,t} b\|_1, \tag{2.12}$$

where λ_η is the penalty level and $\hat{\Lambda}_{\eta,t}$ is a diagonal matrix specifying penalty loadings. Proposed in Belloni et al. (2012), the penalty loadings allow the regression error $U_t := Y(t) - \eta_t(Z)$ to be heteroskedastistic and non-Gaussian. Details on the choice of $\lambda_{\eta,t}$ and $\hat{\Lambda}_{\eta,t}$ are included in Appendix 2.7.1 for completeness, which closely follows Belloni et al. (2012). The post-LASSO is defined as the OLS regression applied to regressors selected by LASSO. Denote

$$\hat{B}_{\eta,t} = \{b \in \mathbb{R}^{p_\eta} : b_j = 0 \text{ if } |\hat{\beta}_{\eta,t,j}^{LS}| = 0, j = 1, ..., p_\eta.\}.$$

The post-LASSO estimator $\hat{\beta}_{\eta,t}^{PL}$ is

$$\hat{\beta}_{\eta,t}^{PL} = \arg\min_{b \in \hat{B}_{\eta,t}} \frac{1}{n_t} \sum_{i=1}^{n} \mathbb{1}\{T = t\}(Y_i - f_\eta(Z_i)'b)^2. \tag{2.13}$$

To estimate ζ_t, we run LASSO regression of $\frac{Y\mathbb{1}\{T=t\}}{\pi_t(Z)}$ on $f_\zeta(X)$ in the whole experimental

sample $\{i : 1 \leq i \leq n\}$,[6]

$$\hat{\beta}^{LS}_{\zeta,t} = \arg\min_{b\in\mathbb{R}^{p_\zeta}} \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)} - f_\zeta(X_i)'b)^2 + \frac{\lambda_\zeta}{n}\|\hat{\Lambda}_{\zeta,t}b\|_1.$$

The choice of penalty level λ_ζ and penalty loadings $\hat{\Lambda}_{\zeta,t}$ are also collected in appendix 2.7.1. The post-LASSO estimator $\hat{\beta}^{PL}_{\zeta,t}$ is defined as the OLS regression applied to regressors selected by LASSO similar to $\hat{\beta}^{PL}_{\eta,t}$ and we omit the details. For $L \in \{LS, PL\}$, denote $\hat{\eta}^L_t(z) = f_\eta(z)'\hat{\beta}^L_{\eta,t}$ and $\hat{\zeta}^L_t(x) = f_\zeta(x)'\hat{\beta}^L_{\zeta,t}$, we can then define the two-step estimator as

$$\hat{\mu}^L_t = \frac{1}{m}\sum_{i=1}^{m}\hat{\zeta}^L_t(X_i) + \frac{1}{n}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}^L_t(Z_i))}{\pi_t(Z_i)} + \hat{\eta}^L_t(Z_i) - \hat{\zeta}^L_t(X_i)), \tag{2.14}$$

which is very similar to (2.4) only without cross-fitting. To derive the asymptotic distribution, we use the LASSO/post-LASSO rate results from the literature. For a scalar random variable W_i, define the norms $\|W_i\|^2_{2,n} = \frac{1}{n}\sum_{i=1}^n W_i^2$ and $\|W_i\|^2_{2,n_t} = \frac{1}{n_t}\sum_{i=1}^n \mathbb{1}\{T_i = t\}W_i^2$, so that $\|f_\eta(Z_i)'(\hat{\beta}^L_{\eta,t} - \beta_{\eta,t})\|_{2,n_t}$ and $\|f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t})\|_{2,n}$ are the prediction norms.

Assumption 2.4.2. *For $L \in \{LS, PS\}$ and every $t \in \{0, ..., \mathtt{T}\}$, as $n \to \infty$,*

$$\|\hat{\beta}^L_{\eta,t} - \beta_{\eta,t}\|_1 = O_p(\sqrt{\frac{s^2\log(p_\eta \vee n)}{n}}), \quad \|f_\eta(Z_i)'(\hat{\beta}^L_{\eta,t} - \beta_{\eta,t})\|_{2,n_t} = O_p(\sqrt{\frac{s\log(p_\eta \vee n)}{n}}),$$
$$\|\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t}\|_1 = O_p(\sqrt{\frac{s^2\log(p_\zeta \vee n)}{n}}), \quad \|f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t})\|_{2,n} = O_p(\sqrt{\frac{s\log(p_\zeta \vee n)}{n}}).$$

These rates can be found in corollary 1 from Belloni et al. (2012), which also provides the sufficient primitive conditions. In our application, η_t is estimated from the n_t sized subsample instead of the size n sample, however, since n_t and n are at the same rate and the asymptotic rate results for LASSO/post-LASSO are derived from finite sample bounds, we could assume that the primitive conditions in Belloni et al. (2012) holds for the subsamples with probability approaching 1. At last, we also impose that all the regressors have bounded support, i.e. $\|f_\zeta\|_\infty \vee \|f_\eta\|_\infty \leq K_B$, and that $s\log(p_\eta \vee p_\zeta)/\sqrt{n} \to 0$.

[6]As noted in a previous footnote, if we have $T_i \perp\!\!\!\perp \{Y_i(t)\}_{t=0}^{\mathtt{T}}|X$, then ζ_t can be estimated similarly as η_t from the $T = t$ subsample in I^c_k. This is recommended over working with $\frac{Y\mathbb{1}\{T=t\}}{\pi_t(Z)}$ as the latter adds noise to the estimation.

Assumption 2.4.3. *i) The regressors are bounded, i.e.* $\|f_\zeta\|_\infty \vee \|f_\eta\|_\infty \leq K_B$ *with probability* 1, *uniformly in* n. *ii)* $s\log(p_\eta \vee p_\zeta) = o(\sqrt{n})$.

The following theorem states the asymptotic distribution of the two-step estimator with LASSO/post-LASSO first step and without cross-fitting. The asymptotic variance matrix V is the same as in theorem 2.3.1.

Theorem 2.4.1 (LASSO/post-LASSO First Step)**.** *Under Assumptions 1.2.1, 2.4.1, 2.4.2 and 2.4.3, for* $L \in \{LS, PL\}$, *consider the estimator* $\hat{\boldsymbol{\mu}}^L = (\hat{\mu}_0^L, ..., \hat{\mu}_{\mathtt{T}}^L)'$, *where* $\hat{\mu}_t^L$, $t \in \{0, ..., \mathtt{T}\}$, *are defined by (2.14). Furthermore, assume that for every* t, $Y(t)$, $\eta_t(Z)$ *and* $\zeta_t(X)$ *all have finite second moment, then as* $n \to \infty$ *and* $\frac{m}{n} \to \gamma$,

$$V^{-\frac{1}{2}}\sqrt{n}(\hat{\boldsymbol{\mu}}^L - \boldsymbol{\mu}) \to N(0, I_{\mathtt{T}+1}).$$

The proof is collected in appendix 2.7.2.

The asymptotic variance V can also be consistently estimated with the LASSO/post-LASSO first-step and without cross-fitting. For $L \in \{LS, PL\}$, define variance estimator $\hat{V}^L$ by changing $\hat{\eta}_{t,k[i]}(Z_i)$ and $\hat{\zeta}_{t,k[i]}(X_i)$ to $\hat{\eta}_t^L(Z_i)$ and $\hat{\zeta}_t^L(X_i)$ in Equation (2.10) and (2.11).

Theorem 2.4.2 (Variance Estimation with LASSO/post-LASSO First Step)**.** *Under the conditions of theorem 2.4.1, further assumes that for some* $\delta > 0$ *and every* $t \in \{0, ..., \mathtt{T}\}$, $E[(Y(t) - \eta_t(Z))^{2+\delta}]$, $E[(\eta_t(Z) - \zeta_t(X))^{2+\delta}]$ *and* $E[(\zeta_t(X) - \mu_t)^{2+\delta}]$ *are all bounded,* $s^2(\log p_\eta \vee p_\zeta) = o(n)$, *then for* $L \in \{LS, PS\}$, $\hat{V}^L \xrightarrow{p} V$.

The proof is collected in appendix 2.7.2.

2.5 Simulation Exercise

In this section, we analyze the finite sample behavior of our method via a simulation study, and in particular examine the variance reduction property and robustness under different first-stage performances.

We generate an i.i.d. experimental sample $\{Y_i, T_i, X_i\}_{i=1}^n$, where $T_i \in \{0, 1\}$, and an independent i.i.d. auxiliary sample $\{X_i\}_{i=n+1}^m$ with the data generating process specified as follow. The potential outcomes follow sparse linear models,

$$Y_i(0) = \alpha_0(X_i'\beta_0) + \epsilon_{0,i}, \tag{2.15}$$

$$Y_i(1) = 5 + \alpha_1(X_i'\beta_1) + \epsilon_{1,i}, \tag{2.16}$$

where X_i has dimension d_x and follows joint normal distribution $N(0, I_{d_x})$. $\epsilon_{0,i}$ and $\epsilon_{1,i}$ are i.i.d. $N(0, 1)$. The coefficients in β_0 and β_1 are set to decay at a polynomial rate controlled by sparsity parameter δ, with

$$\beta_0 = (-1, -1, -1, -1, 1, 1, (\frac{1}{2})^\delta, -(\frac{1}{3})^\delta, (\frac{1}{4})^\delta, ...),$$
$$\beta_1 = (1, 1, 1, 1, ...(\frac{1}{4})^\delta, -(\frac{1}{3})^\delta, (\frac{1}{2})^\delta, 1, 1).$$

Most elements in β_0 and β_1 are set to have opposite signs so that the covariates X_i can explain the individual treatment effect $[Y_i(1) - Y_i(0)]$, i.e. X_i explains the heterogeneous treatment effect. The parameter of interest is the average treatment effect $E[Y_i(1) - Y_i(0)]$ which equals 5. Treatment T_i are generated with a propensity score function $\pi(X) = E[T|X]$, set to depend on the leading covariates (those that do not decay based on sparsity), namely

$$\pi(X) = \frac{1}{3} + \frac{1}{3}\mathbb{1}\{\sum_{j=1}^{6} X_j + X_{d_x} + X_{d_x-1} > 0\}.$$

Finally, we get $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. Throughout the simulation, we will have $n = 500$, $m = 5000$, and $d_x = 200$. In the first-step estimations, the treated and the untreated subsamples are used separately, and we adopt a five fold cross-fitting. As a result, the number of observations used in estimating the conditional mean functions is equal (in expectation) to the dimension of covariates ($500 \times \frac{1}{2} \times \frac{4}{5} = 200$).

Beside the sparsity parameter δ, the parameters α_0 and α_1 are of particular interest as they determine how much the variation of the potential outcomes, as well as the individual treatment effect, can be explained by the covariates. In the baseline specification, we let

$\delta = 2$, so that the model (2.15) and (2.16) are approximately sparse. We then set the values of α_0 and α_1 to be such that the population R^2 of (2.15) and (2.16) equal 0.8. According to our theory, this is a very favorable specification and we should expect large variance reduction. In specification II, we set $\delta = 0$, so that the sparsity assumption fails (all coefficients in β_0 and β_1 are 1). In specification III, we reduce α_0 and α_1 such that the population R^2 of model (2.15) and (2.16) is 0.1. We compare four estimators: 1) the inverse propensity score weighting estimator, which only uses the experimental sample; 2) Our ML imputation estimator without incorporating the auxiliary sample. This estimator is similar to the ones appeared in Chernozhukov et al. (2018) and Farrell (2015); 3) An infeasible version of our imputation estimator where the first-step estimates are replaced by the true conditional mean function; 4) Our ML imputation estimator. We use the post-LASSO estimator from the `hdm` package in `R` for the first-step estimations, which implement Belloni et al. (2012). The results are summerized in the table 2.1.

We see that the variance is reduced by 66% due to the inclusion of the auxiliary data under the favorable baseline specification. Under specification II, where the sparsity assumption fails, the first-stage post-LASSO estimates have a goodness of fit at around 20%, as opposed to the 80% from the true conditional mean function. As a result, the variance reduction is much less pronounced at around 26%. In specification III, due to very low population R^2 in the data generating process (high noise to signal ratio), the post-LASSO first stage estimates have virtually zero goodness of fit, hence there is no variance reduction as we would have expected. Noticeably though, in the latter two specifications where the first-stage estimates behave poorly, the ML imputation estimator is still consistent, and the variance is not drastically increased compared to the infeasible estimator or the IPW estimator. On top of that, the coverage remained good. This demonstrates the robustness of our estimator with respect to the performance of the first-stage estimation.

Table 2.1: Monte Carlo Simulation Results.

	Estimate	MC Var	Asym. Var	Coverage	GoF
Baseline Specification					
IPW	4.995	52.014	50.744	0.948	N/A
Exp. Only	4.998	18.374	19.262	0.951	0.778
Infeasible	5.000	5.687	6.958	0.971	0.797
ML Imputation	5.000	6.113	7.492	0.971	0.778
Specification II (Sparsity fails)					
IPW	5.000	54.543	53.563	0.950	N/A
Exp. Only	5.002	25.562	29.773	0.968	0.201
Infeasible	5.000	6.224	7.223	0.972	0.799
ML Imputation	4.998	18.843	23.124	0.973	0.201
Specification III ($R^2 = 0.1$)					
IPW	4.997	35.798	35.730	0.943	N/A
Exp. Only	4.999	5.257	6.187	0.965	0.007
Infeasible	4.999	4.672	5.666	0.971	0.089
ML Imputation	4.999	5.281	6.230	0.965	0.007

Note: The columns are: point estimate, Monte Carlo variance, estimated asymptotic variance, coverage of 95% confidence interval, out of sample goodness of fit of the first-step estimation. The number of simulations is 1000.

2.6 Conclusion

In this chapter, I proposed a two-stage machine learning (ML) imputation estimator for average treatment effect (ATE) that achieves the efficiency bound derived in the previous

chapter. This method is efficient in the sense that no other regular estimators for ATE can have lower asymptotic variance in the same setting. The efficiency is gained by aggregating imputed potential outcomes for every unit in both samples, hence fully utilizing the information on the marginal distribution of covariates in the auxiliary sample. Adopting the cross-fitting technique proposed in Chernozhukov et al. (2018), our two-step estimator can use a wide range of supervised ML tools as the first-step, while maintaining valid inference to construct confidence intervals and perform hypothesis tests. In fact, any method that estimates the relevant conditional mean functions consistently in $L_2(P)$ norm, with no rate requirement, will lead to efficiency through the proposed two-step procedure. I also show that cross-fitting is not necessary when the first-step is done via LASSO or post-LASSO. Furthermore, our estimator is robust in the sense that it remains consistent and $\sqrt{n}$ normal (no longer efficient) even if the first step estimators are inconsistent.

2.7 Appendix

2.7.1 Details on LASSO/post-LASSO First Step

In this subsection, for completeness, we provide the details on the choice of penalty levels and penalty loadings for the LASSO/post-LASSO estimators, which completely follows Belloni et al. (2012).

2.7.1.1 Estimating η_t

Recall that our conditional mean models are

$$Y_i(t) = \eta_t(Z_i) + U_{t,i}, \quad E[U_{t,i}|Z_i] = 0, \quad t \in \{1, ..., \mathtt{T}\},$$

Since $Y_i(t)$ is only observed in the $T = t$ subsample, we treat each $t \in \{1, ..., \mathtt{T}\}$ separately and estimate η_t only using the $T = t$ subsample. This is viable because under Assumption

1.2.1 we have

$$E[Y_i|T=t,Z_i] = E[Y_i(t)|T=t,Z_i] = E[Y_i(t)|Z_i] = \eta_t(Z_i).$$

Denote $n_t = \sum_{i=1}^n \mathbb{1}\{T=t\}$, the LASSO estimator is given by

$$\hat{\beta}_{\eta,t}^{LS} = \arg\min_{b\in\mathbb{R}^{p_\eta}} \frac{1}{n_t}\sum_{i=1}^n \mathbb{1}\{T=t\}(Y_i - f_\eta(Z_i)'b)^2 + \frac{\lambda_\eta}{n_t}\|\hat{\Lambda}_{\eta,t}b\|_1,$$

For the penalty parameter λ_η, we set

$$\lambda_\eta = 2c\sqrt{n}\Phi^{-1}(1-\tau_\eta/2p_\eta), \tag{2.17}$$

where $c > 1$ is a constant, $\tau_\eta \in (0,1)$, $\tau_\eta \to 0$, and $\log(\frac{1}{\tau_\eta}) = O(\log(p_\eta \vee n))$. Φ is the CDF of the standard normal distribution. Belloni et al. (2012) recommends $c = 1.1$ and $\tau_\eta = 0.1/\log(p_\eta \vee n)$. The penalty loadings $\hat{\Lambda}_{\eta,t}$ is a diagonal matrix computed from the following iteration procedure.

Denote the diagonal elements of $\hat{\Lambda}_{\eta,t}$ by $\hat{\gamma}_{\eta,t,j}$, $j \in \{1,...,p_\eta\}$. Denote $\bar{Y}(t) = \frac{1}{n_t}\sum_{i=1}^n \mathbb{1}\{T=t\}Y_i$. (a) Set the value of λ_η according to (2.17), and set the initial value of $\hat{\Lambda}_{\eta,t}$ by

$$\hat{\gamma}_{\eta,t,j} = \sqrt{\frac{1}{n_t}\sum_{i=1}^n \mathbb{1}\{T_i=t\}f_{\eta,j}(Z_i)^2(Y_i - \bar{Y}(t))^2}.$$

Then compute the LASSO or post-LASSO estimator $\hat{\beta}_{\eta,t}^L$ and calculate the residues $\hat{U}_{t,i} = Y_i - f_\eta(Z_i)'\hat{\beta}_{\eta,t}^L$ for $i \in \{1 \le i \le n : T_i = t\}$. (b) Update the penalty loadings by

$$\hat{\gamma}_{\eta,t,j} = \sqrt{\frac{1}{n_t}\sum_{i=1}^n \mathbb{1}\{T_i=t\}f_{\eta,j}(Z_i)^2\hat{U}_{t,i}^2},$$

update the LASSO/post-LASSO estimator $\hat{\beta}_{\eta,t}^L$ and compute a new set of residues. (c) Repeat the previous step K times.

The preferred approach in Belloni et al. (2012) is to use post-LASSO in every step and set $K = 15$.

2.7.1.2 Estimating ζ_t

As mentioned in the footnotes earlier, if we have $Y_i \perp\!\!\!\perp \{Y_i(t)\}_{t=1}^{\mathtt{T}} | X_i$, we recommend estimating ζ_t separately for each t using only the $T = t$ subsample in the same way as estimating η_t. Otherwise, we can run LASSO/post-LASSO regression of $\frac{\mathbb{1}\{T_i=t\}Y_i}{\pi_t(Z_i)}$ on X_i in the whole experimental sample. This is viable because by Assumption 1.2.1,

$$E[\frac{\mathbb{1}\{T_i = t\}Y_i}{\pi_t(Z_i)}|X_i] = E[E[\frac{\mathbb{1}\{T_i = t\}Y_i}{\pi_t(Z_i)}|Z_i]|X_i] = E[Y_i(t)|X_i] = \zeta_t(X_i).$$

Hence, our conditional mean models for ζ_t are

$$\frac{\mathbb{1}\{T_i = t\}Y_i}{\pi_t(Z_i)} = \zeta_t(X_i) + v_{t,i}, \quad E[v_{t,i}|X_i] = 0, \quad t \in \{1, ..., \mathtt{T}\},$$

The LASSO estimator is given by

$$\hat{\beta}_{\zeta,t}^{LS} = \arg\min_{b\in\mathbb{R}^{p_\zeta}} \frac{1}{n}\sum_{i=1}^{n}(\frac{Y_i\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)} - f_\zeta(X_i)'b)^2 + \frac{\lambda_\zeta}{n}\|\hat{\Lambda}_{\zeta,t}b\|_1.$$

For the penalty parameter λ_ζ, we set

$$\lambda_\zeta = 2c\sqrt{n}\Phi^{-1}(1 - \tau_\zeta/(2p_\zeta \cdot \mathtt{T})), \tag{2.18}$$

where $c > 1$ is a constant, $\tau_\zeta \in (0, 1)$, $\tau_\zeta \to 0$, and $\log(\frac{1}{\tau_\zeta}) = O(\log(p_\zeta \vee n))$. Φ is the CDF of the standard normal distribution. Belloni et al. (2012) recommends $c = 1.1$ and $\tau_\zeta = 0.1/\log(p_\zeta \vee n)$. The penalty loadings $\hat{\Lambda}_{\zeta,t}$ is a diagonal matrix computed from the following iteration procedure.

Denote the diagonal elements of $\hat{\Lambda}_{\zeta,t}$ by $\hat{\gamma}_{\zeta,t,j}$, $j \in \{1, ..., p_\zeta\}$. Denote $\tilde{Y}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{T=t\}Y_i}{\pi_t(Z_i)}$. (a) Set the value of λ_ζ according to (2.18), and set the initial value of $\hat{\Lambda}_{\zeta,t}$ by

$$\hat{\gamma}_{\zeta,t,j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{\zeta,j}(Z_i)^2(\frac{\mathbb{1}\{T = t\}Y_i}{\pi_t(Z_i)} - \tilde{Y}(t))^2}.$$

Then compute the LASSO or post-LASSO estimator $\hat{\beta}_{\zeta,t}^{L}$ and calculate the residues $\hat{v}_{t,i} = \frac{\mathbb{1}\{T=t\}Y_i}{\pi_t(Z_i)} - f_\zeta(Z_i)'\hat{\beta}_{\zeta,t}^{L}$ for $i \in \{1, ..., n\}$. (b) Update the penalty loadings by

$$\hat{\gamma}_{\zeta,t,j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{\zeta,j}(Z_i)^2\hat{v}_{t,i}^2},$$

update the LASSO/post-LASSO estimator $\hat{\beta}^L_{\zeta,t}$ and compute a new set of residues. (c) Repeat the previous step K times.

The preferred approach in Belloni et al. (2012) is to use post-LASSO in every step and set $K = 15$.

2.7.2 Proofs

2.7.2.1 Additional Lemmas

The following lemma states that conditional convergence in probability implies unconditional convergence in probability, which becomes useful when the cross-fitting technique is applied.

Lemma 2.7.1. *Let X_n and Y_n be two sequences of random variables indexed by n, if $E[|X_n|^\delta|Y_n] = o_p(1)$ for some $\delta > 0$, then $X_n = o_p(1)$.*

Proof. For any $\epsilon > 0$, by Markov inequality $P(|X_n| > \epsilon|Y_n) \leq \frac{E[|X_n|^\delta|Y_n]}{\epsilon^\delta} = o_p(1)$. Hence we have for any $\tilde{\eta} > 0$,

$$P(P(|X_n| > \epsilon|Y_n) > \tilde{\eta}) = o(1). \tag{2.19}$$

We want to show that for n large enough, for any $\epsilon > 0$ and $\eta > 0$, $P(|X_n| > \epsilon) < \eta$. Note that for some $\tilde{\eta} < \eta/2$,

$$\begin{aligned}
P(|X_n| > \epsilon) =& E[\mathbb{1}\{|X_n| > \epsilon\}] \\
=& E[E[\mathbb{1}\{|X_n| > \epsilon\}|Y_n]] \\
=& E[P(|X_n| > \epsilon|Y_n)\mathbb{1}\{P(|X_n| > \epsilon|Y_n) > \tilde{\eta}\}] \\
&+ E[P(|X_n| > \epsilon|Y_n)\mathbb{1}\{P(|X_n| > \epsilon|Y_n) \leq \tilde{\eta}\}] \\
\leq& E[\mathbb{1}\{P(|X_n| > \epsilon|Y_n) > \tilde{\eta}\}] + \tilde{\eta}.
\end{aligned}$$

The first term in the last line is smaller than $\tilde{\eta}$ for large n by (2.19), hence we have $P(|X_n| > \epsilon) \leq 2\tilde{\eta} < \eta$. ■

The following lemma is useful for proving the consistency of variance estimators.

Lemma 2.7.2. *Let* $\{(\hat{\Psi}_{t,i}, \hat{\Psi}_{t',i}, \Psi_{t,i}, \Psi_{t',i})\}_{i=1}^n$ *be a sequence of random vectors, suppose that i)* $\{(\Psi_{t,i}, \Psi_{t',i})\}_{i=1}^n$ *are i.i.d., ii) for some* $\delta > 0$, $E[\Psi_{t,i}^{2+\delta}]$ *and* $E[\Psi_{t',i}^{2+\delta}]$ *are bounded, iii)* $\frac{1}{n}\sum_{i=1}^n(\hat{\Psi}_{t,i} - \Psi_{t,i})^2 = o_p(1)$ *and* $\frac{1}{n}\sum_{i=1}^n(\hat{\Psi}_{t',i} - \Psi_{t',i})^2 = o_p(1)$*, then*

$$\frac{1}{n}\sum_{i=1}^n(\hat{\Psi}_{t,i}\hat{\Psi}_{t',i} - E[\Psi_{t,i}\Psi_{t',i}]) = o_p(1). \tag{2.20}$$

Proof.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n(\hat{\Psi}_{t,i}\hat{\Psi}_{t',i} - E[\Psi_{t,i}\Psi_{t',i}]) \\
=&\frac{1}{n}\sum_{i=1}^n(\hat{\Psi}_{t,i}\hat{\Psi}_{t',i} - \Psi_{t,i}\Psi_{t',i}) + \frac{1}{n}\sum_{i=1}^n(\Psi_{t,i}\Psi_{t',i} - E[\Psi_{t,i}\Psi_{t',i}])
\end{aligned} \tag{2.21}$$

We will show each of the two terms in (2.21) is $o_p(1)$, starting with the first term. Note that for any numbers a, b, $\tilde{a}$ and $\tilde{b}$ such that $|a| \vee |b| \leq c$ and $|\tilde{a}| \vee |\tilde{b}| \leq r$, we have

$$|(a+\tilde{a})(b+\tilde{b}) - ab| \leq 2r(c+r). \tag{2.22}$$

Apply (2.22) below, with $a = \Psi_{t,i}$, $b = \Psi_{t',i}$, $\tilde{a} = \hat{\Psi}_{t,i} - \Psi_{t,i}$ and $\tilde{b} = \hat{\Psi}_{t',i} - \Psi_{t',i}$, we get

$$\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^n(\hat{\Psi}_{t,i}\hat{\Psi}_{t',i} - \Psi_{t,i}\Psi_{t',i})| \\
\leq&\frac{1}{n}\sum_{i=1}^n|(\hat{\Psi}_{t,i}\hat{\Psi}_{t',i} - \Psi_{t,i}\Psi_{t',i})| \\
\leq&\frac{1}{n}\sum_{i=1}^n 2|\hat{\Psi}_{t,i} - \Psi_{t,i}| \vee |\hat{\Psi}_{t',i} - \Psi_{t',i}| \times (|\Psi_{t,i}| \vee |\Psi_{t',i}| + |\hat{\Psi}_{t,i} - \Psi_{t,i}| \vee |\hat{\Psi}_{t',i} - \Psi_{t',i}|) \\
\leq&2 \cdot (\frac{1}{n}\sum_{i=1}^n(|\hat{\Psi}_{t,i} - \Psi_{t,i}| \vee |\hat{\Psi}_{t',i} - \Psi_{t',i}|)^2)^{\frac{1}{2}} \\
&\times ((\frac{1}{n}\sum_{i=1}^n(|\Psi_{t,i}| \vee |\Psi_{t',i}|)^2)^{\frac{1}{2}} + (\frac{1}{n}\sum_{i=1}^n(|\hat{\Psi}_{t,i} - \Psi_{t,i}| \vee |\hat{\Psi}_{t',i} - \Psi_{t',i}|)^2)^{\frac{1}{2}}),
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz and triangular inequality. Define

$$R_{1,n}^2 = \frac{1}{n}\sum_{i=1}^{n}(|\Psi_{t,i}| \vee |\Psi_{t',i}|)^2,$$
$$R_{2,n}^2 = \frac{1}{n}\sum_{i=1}^{n}(|\hat{\Psi}_{t,i} - \Psi_{t,i}| \vee |\hat{\Psi}_{t',i} - \Psi_{t',i}|)^2.$$

We have bounded the first term in (2.21) by $2R_{2,n}(R_{1,n} + R_{2,n})$. Next,

$$R_{1,n}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\Psi_{t,i}^2 + \frac{1}{n}\sum_{i=1}^{n}\Psi_{t',i}^2,$$

so that $E[R_1^2] = E[\Psi_{t,i}^2] + E[\Psi_{t',i}^2] \leq \infty$ under conditions (i) and (ii) of this lemma. Hence $R_1 = O_p(1)$ by Markov inequality. Next,

$$R_{2,n} \leq (\frac{1}{n}\sum_{i=1}^{n}(\hat{\Psi}_{t,i} - \Psi_{t,i})^2)^{\frac{1}{2}} + (\frac{1}{n}\sum_{i=1}^{n}(\hat{\Psi}_{t',i} - \Psi_{t',i})^2)^{\frac{1}{2}} = o_p(1),$$

where the equality is by condition (iii) of this lemma. Put together, $2R_{2,n}(R_{1,n}+R_{2,n}) = o_p(1)$ and hence the first term in (2.21) is $o_p(1)$.

Next we show that the second term in (2.21) is also $o_p(1)$. Under condition (ii) of this theorem, if $\delta \geq 2$, we have

$$\begin{aligned}
&E[(\frac{1}{n}\sum_{i=1}^{n}(\Psi_{t,i}\Psi_{t',i} - E[\Psi_{t,i}\Psi_{t',i}]))^2] \\
\leq &\frac{1}{n}E[\Psi_{t,i}^2\Psi_{t',i}^2] \\
\leq &\frac{1}{n}E[\Psi_{t,i}^4]^{\frac{1}{2}}E[\Psi_{t',i}^4]^{\frac{1}{2}} = O(n^{-1}).
\end{aligned}$$

If $\delta \in (0,2)$ we use the Bahr-Esseen inequality[7] with $p = (2+\delta)/2$, so that

$$E[|\frac{1}{n}\sum_{i=1}^{n}(\Psi_{t,i}\Psi_{t',i} - E[\Psi_{t,i}\Psi_{t',i}])|^{\frac{2+\delta}{2}}]$$
$$\lesssim n^{-\frac{2+\delta}{2}} \cdot n \cdot E[\Psi_{t,i}^{2+\delta}]^{\frac{1}{2}} E[\Psi_{t',i}^{2+\delta}]^{\frac{1}{2}} = O(n^{-\frac{\delta}{2}}).$$

Follow by Markov inequality, the second term in (2.21) is $o_p(1)$, which completes the proof of this lemma.

■

Next, we state two maximal inequalities.

Lemma 2.7.3. *Let X_i be bounded $p \times 1$ vectors with mean μ, $|X_{i,j} - \mu_j| \leq B$ almost surely for all i, j, then as $n \to \infty$ and $p \to \infty$,*

$$P(\max_{1\leq j\leq p} |\frac{1}{n}\sum_{i=1}^{n} X_{i,j} - \mu_j| \geq 2B\sqrt{\frac{\log p}{n}}) \leq \frac{2}{p} \to 0,$$

so that $\|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_i]\|_\infty = O_p(\sqrt{\frac{\log p}{n}})$.

The following version is slightly different and works for matrices.

Lemma 2.7.4. *Let X_i be bounded $p \times p$ matrices with mean μ, $|X_{i,j,k} - \mu_{j,k}| \leq B$ almost surely for all i, j, k, then as $n \to \infty$ and $p \to \infty$,*

$$P(\max_{1\leq j\leq p}\max_{1\leq k\leq p} |\frac{1}{n}\sum_{i=1}^{n} X_{i,j,k} - \mu_{j,k}| \geq 2B\sqrt{\frac{\log p}{n}}) \leq \frac{2}{p^2} \to 0,$$

so that $\|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_i]\|_\infty = O_p(\sqrt{\frac{\log p}{n}})$.

Both Lemma 2.7.3 and 2.7.4 can be proved by applying the union bound and Hoeffding's inequality. The proofs are standard and omitted here.

---

[7] Let $1 \leq p \leq 2$, and let X_i, $i = 1, 2, ...$, be a sequence of independent random variables with finite p-th moment and mean zero (i.e. $E[|X_i|^p] < \infty$, $E[X_i] = 0$ for all $i = 1, 2, ...$). Then

$$E[|\sum_{i=1}^{n} X_i|^p] \leq (2 - n^{-1})\sum_{i=1}^{n} E[|X_i|^p].$$

2.7.2.2 Proof of Theorem 2.3.1

Recall our estimator $\hat{\mu}_t$ is defined as

$$\hat{\mu}_t = \frac{1}{m}\sum_{i=1}^{m}\hat{\zeta}_{t,k[i]}(X_i) + \frac{1}{n}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}_{t,k[i]}(Z_i))}{\pi_t(Z_i)} + \hat{\eta}_{t,k[i]}(Z_i) - \hat{\zeta}_{t,k[i]}(X_i)).$$

Add and subtract the true conditional mean functions $\zeta_t(X_i)$ and $\eta_t(Z_i)$, the scaled and centered distribution of $\hat{\mu}_t$ can be written into the sum of two terms,

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_t - \mu_t) =& \mathcal{I}_1 + \mathcal{I}_2, \\
\mathcal{I}_1 =& \sqrt{n}(\frac{1}{m}\sum_{i=1}^{m}\zeta_t(X_i) - \mu_t + \frac{1}{n}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}U_i^t}{\pi_t(Z_i)} + \eta_t(Z_i) - \zeta_t(X_i))), \\
\mathcal{I}_2 =& \sqrt{n}\sum_{k=1}^{K}(\frac{1}{m}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i)) \\
&+ \frac{1}{n}\sum_{i\in I_k}((1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)})(\hat{\eta}_{t,k}(Z_i) - \eta_t(Z_i)) - (\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i)))),
\end{aligned}$$

where $U_i^t = Y_i(t) - \eta_t(Z_i)$. The first term will have the normal limiting distribution and the second term is an $o_p(1)$. To see the latter, we can further rearrange the second term into

$$\begin{aligned}
\mathcal{I}_2 =& \sum_{k=1}^{K}(\mathcal{I}_{3,k} + \mathcal{I}_{4,k}), \\
\mathcal{I}_{3,k} =& \sqrt{n}\frac{1}{m}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i) - E[\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i)|I_k^c]), \\
\mathcal{I}_{4,k} =& \frac{1}{\sqrt{n}}\sum_{i\in I_k}((1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)})(\hat{\eta}_{t,k}(Z_i) - \eta_t(X_i)) - (\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i)) \\
&+ E[\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i)|I_k^c]).
\end{aligned}$$

To get the above expressions, we added and subtracted $E[\hat{\zeta}_{t,k}(X_i) - \zeta_t(X_i)|I_k^c]$ inside the summations. They will cancel as $\frac{|I_k\cup\tilde{I}_k|}{m} = \frac{|I_k|}{n} = \frac{1}{K}$. Note that both $\mathcal{I}_{3,k}$ and $\mathcal{I}_{4,k}$ are mean zero conditional on I_k^c. Next, we make use of the cross-fitting technique and bound $E[\mathcal{I}_{3,k}^2|I_k^c]$ and $E[\mathcal{I}_{4,k}^2|I_k^c]$ by Assumption 2.3.1. Since all the randomness in $\hat{\zeta}_{t,k}(x)$ and $\hat{\eta}_{t,k}(z)$ is from I_k^c,

we have $E[(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2|I_k^c]=E_X[(\hat{\zeta}_{t,k}(X)-\zeta_t(X))^2]$ and $E[(\hat{\eta}_{t,k}(Z_i)-\eta_t(Z_i))^2|I_k^c]=E_Z[(\hat{\eta}_{t,k}(Z)-\eta_t(Z))^2]$. Therefore,

$$\begin{aligned}
E[\mathcal{I}_{3,k}^2|I_k^c]=&\frac{n}{m^2}E[(\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i)-E[\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i)|I_k^c]))^2|I_k^c]\\
=&\frac{n}{mK}E[(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i)-E[\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i)|I_k^c])^2|I_k^c]\\
\leq&\frac{n}{mK}E[(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2|I_k^c]=o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
E[\mathcal{I}_{4,k}^2|I_k^c]=&\frac{1}{K}E[(1-\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)})^2(\hat{\eta}_{t,k}(Z_i)-\eta_t(Z_i))^2|I_k^c]\\
&+\frac{1}{K}E[(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i)-E[\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i)|I_k^c])^2|I_k^c]\\
\leq&\frac{1}{K\pi_{\min}^2}E[(\hat{\eta}_{t,k}(Z_i)-\eta_t(Z_i))^2|I_k^c]\\
&+\frac{1}{K}E[(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2|I_k^c]=o_p(1),
\end{aligned}$$

by Assumption 1.2.1 (i), (iii) and Assumption 2.3.1. Therefore by Lemma 2.7.1, $\mathcal{I}_{3,k}$ and $\mathcal{I}_{4,k}$ are $o_p(1)$ for every k. Since K is fixed and finite, $\mathcal{I}_2=o_p(1)$. Next we analyze the first term $\mathcal{I}_1$. Since $Y(t)$, $\zeta(X_i)$ and $\eta(Z_i)$ all have finite second moment under the condition of this theorem and that $\frac{m}{n}\to\gamma$,

$$\begin{aligned}
\mathcal{I}_1=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{n}{m}(\zeta_t(X_i)-\mu_t)+\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)}U_i^t+\eta_t(Z_i)-\zeta_t(X_i))\\
&+\sqrt{\frac{n(m-n)}{m^2}}\frac{1}{\sqrt{m-n}}\sum_{i=n+1}^{m}(\zeta_t(X_i)-\mu_t)\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{1}{\gamma}(\zeta_t(X_i)-\mu_t)+\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)}U_i^t+\eta_t(Z_i)-\zeta_t(X_i))\\
&+\sqrt{\frac{\gamma-1}{\gamma^2}}\frac{1}{\sqrt{m-n}}\sum_{i=n+1}^{m}(\zeta_t(X_i)-\mu_t)+o_p(1).
\end{aligned}$$

Let $\boldsymbol{\mu}=(\mu_0,...,\mu_{\mathrm{T}})'$ and $\hat{\boldsymbol{\mu}}=(\hat{\mu}_0,...,\hat{\mu}_{\mathrm{T}})'$, denote

$$\varphi_t(Y_i,T_i,Z_i)=\frac{1}{\gamma}(\zeta_t(X_i)-\mu_t)+\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)}U_i^t+\eta_t(Z_i)-\zeta_t(X_i),$$

we then have

$$\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \varphi_0(Y_i, T_i, Z_i) \\ ... \\ \varphi_{\mathtt{T}}(Y_i, T_i, Z_i) \end{pmatrix} + \sqrt{\frac{\gamma - 1}{\gamma^2}} \frac{1}{\sqrt{m-n}} \sum_{i=n+1}^{m} \begin{pmatrix} \zeta_0(X_i) - \mu_0 \\ ... \\ \zeta_{\mathtt{T}}(X_i) - \mu_{\mathtt{T}} \end{pmatrix} + o_p(1).$$

Since $Y(t)$, $\zeta(X_i)$ and $\eta(Z_i)$ all have finite second moment, and that the two samples are independent, by central limit theorem, the above expression converges in distribution to $N(0, V)$ where

$$\begin{aligned} V_{t,t'} =& \mathbb{1}\{t = t'\} E[\frac{\sigma_t^2(Z_i)}{\pi_t(Z_i)}] + E[(\eta_t(Z_i) - \zeta_t(X_i))(\eta_{t'}(Z_i) - \zeta_{t'}(X_i))] \\ &+ \frac{1}{\gamma} E[(\zeta_t(X_i) - \mu_t)(\zeta_{t'}(X_i) - \mu_{t'})]. \end{aligned}$$

This completes the proof of theorem 2.3.1.

2.7.2.3 Proof of Theorem 2.3.2

to prove this theorem, it suffices to show that for any $t, t' \in \{0, ..., \mathtt{T}\}$,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} (\hat{\eta}_{t,k[i]}(Z_i) - \hat{\zeta}_{t,k[i]}(X_i))(\hat{\eta}_{t',k[i]}(Z_i) - \hat{\zeta}_{t',k[i]}(X_i)) \\ &\qquad - E[(\eta_t(Z_i) - \zeta_t(X_i))(\eta_{t'}(Z_i) - \zeta_{t'}(X_i))] = o_p(1), \quad (2.23) \\ &\frac{1}{m} \sum_{i=1}^{m} (\hat{\zeta}_{t,k[i]}(X_i) - \hat{\mu}_t)(\hat{\zeta}_{t',k[i]}(X_i) - \hat{\mu}_{t'}) - E[(\zeta_t(X_i) - \mu_t)(\zeta_{t'}(X_i) - \mu_{t'})] = o_p(1), \quad (2.24) \\ &\frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}_{t,k[i]}(Z_i))^2}{\pi_t(Z_i)^2} - E[\frac{\sigma_t^2(Z_i)}{\pi_t(Z_i)}] = o_p(1). \quad (2.25) \end{aligned}$$

Starting with (2.23),

$$\frac{1}{n}\sum_{i=1}^{n}[(\hat{\eta}_{t,k[i]}(Z_i)-\hat{\zeta}_{t,k[i]}(X_i))(\hat{\eta}_{t',k[i]}(Z_i)-\hat{\zeta}_{t',k[i]}(X_i))] \tag{2.26}$$

$$-E[(\eta_t(Z_i)-\zeta_t(X_i))(\eta_{t'}(Z_i)-\zeta_{t'}(X_i))]$$

$$=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}((\hat{\eta}_{t,k}(Z_i)-\hat{\zeta}_{t,k}(X_i))(\hat{\eta}_{t',k}(Z_i)-\hat{\zeta}_{t',k}(X_i)) \tag{2.27}$$

$$-E[(\eta_t(Z_i)-\zeta_t(X_i))(\eta_{t'}(Z_i)-\zeta_{t'}(X_i))]).$$

We apply Lemma 2.7.2 to each fold k. Here, $\Psi_{t,i}=\eta_t(Z_i)-\zeta_t(X_i)$, $\hat{\Psi}_{t,i}=\hat{\eta}_{t,k}(Z_i)-\hat{\zeta}_{t,k}(X_i)$, $\Psi_{t',i}=\eta_{t'}(Z_i)-\zeta_{t'}(X_i)$, and $\hat{\Psi}_{t',i}=\hat{\eta}_{t',k}(Z_i)-\hat{\zeta}_{t',k}(X_i)$. Condition (i) and (ii) in Lemma 2.7.2 are directly satisfied by Assumption 1.2.1 and the conditions of this theorem. To see (iii), for any t, by triangular inequality,

$$\begin{aligned}(\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\Psi}_{t,i}-\Psi_{t,i})^2)^{\frac{1}{2}} =&(\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\eta}_{t,k}(Z_i)-\hat{\zeta}_{t,k}(X_i)-\eta_t(Z_i)+\zeta_t(X_i))^2)^{\frac{1}{2}}\\ \leq&(\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\eta}_{t,k}(Z_i)-\eta_t(Z_i))^2)^{\frac{1}{2}}+(\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2)^{\frac{1}{2}}.\end{aligned}$$

As a result of cross-fitting, by Assumption 2.3.1, we have

$$E[\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\eta}_{t,k}(Z_i)-\eta_t(Z_i))^2|I_k^c]=E_Z[(\hat{\eta}_{t,k}(Z_i)-\eta_t(Z_i))^2]=o_p(1),$$

$$E[\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2|I_k^c]=E_Z[(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2]=o_p(1).$$

Hence, by Lemma 2.7.1 (conditional convergence implies unconditional), condition (iii) for Lemma 2.7.2 is satisfied for both t and t'. Apply Lemma 2.7.2, we get for every $k\in\{1,...,K\}$,

$$\frac{1}{|I_k|}\sum_{i\in I_k}((\hat{\eta}_{t,k}(Z_i)-\hat{\zeta}_{t,k}(X_i))(\hat{\eta}_{t',k}(Z_i)-\hat{\zeta}_{t',k}(X_i))-E[(\eta_t(Z_i)-\zeta_t(X_i))(\eta_{t'}(Z_i)-\zeta_{t'}(X_i))])$$

$$=o_p(1)$$

Since K is finite and fixed, we have finished the proof of (2.23).

Next we move on to (2.24).

$$\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}_{t,k[i]}(X_i)-\hat{\mu}_t)(\hat{\zeta}_{t',k[i]}(X_i)-\hat{\mu}_{t'})-E[(\zeta_t(X_i)-\mu_t)(\zeta_{t'}(X_i)-\mu_{t'})]$$
$$=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k\cup\tilde{I}_k|}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i)-\hat{\mu}_t)(\hat{\zeta}_{t',k}(X_i)-\hat{\mu}_{t'})-E[(\zeta_t(X_i)-\mu_t)(\zeta_{t'}(X_i)-\mu_{t'})]$$

We again apply Lemma 2.7.2 to each fold k. This time $\Psi_{t,i}=\zeta_t(X_i)-\mu_t$, $\hat{\Psi}_{t,i}=\hat{\zeta}_{t,k}(X_i)-\hat{\mu}_t$, $\Psi_{t',i}=\zeta_{t'}(X_i)-\mu_{t'}$, and $\hat{\Psi}_{t',i}=\hat{\zeta}_{t',k}(X_i)-\hat{\mu}_{t'}$. Condition (i) and (ii) in Lemma 2.7.2 are directly satisfied by Assumption 1.2.1 and the condition of this theorem. To check (iii), for any t, by triangular inequality,

$$\begin{aligned}(\frac{1}{|I_k\cup\tilde{I}_k|}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\Psi}_{t,i}-\Psi_{t,i})^2)^{\frac{1}{2}}=&(\frac{1}{|I_k\cup\tilde{I}_k|}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i)-\hat{\mu}_t-\zeta_t(X_i)-\mu_t)^2)^{\frac{1}{2}}\\ \leq&(\frac{1}{|I_k\cup\tilde{I}_k|}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i)-\zeta_t(X_i))^2)^{\frac{1}{2}}\\ &+(\frac{1}{|I_k\cup\tilde{I}_k|}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\mu}_t-\mu_t)^2)^{\frac{1}{2}}.\end{aligned}$$

The first term in the upper bound can be shown to be $o_p(1)$ via the cross-fitting technique in the same way as in proving (2.23), the second term equals $\hat{\mu}_t-\mu_t$ which is $O_p(n^{-\frac{1}{2}})$ by theorem 2.3.1. Then by Lemma 2.7.2 and the fact that K is finite and fixed, (2.24) is proved.

At last, we prove (2.25).

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{T_i=t\}(Y_i-\hat{\eta}_{t,k[i]}(Z_i))^2}{\pi_t(Z_i)^2}-E[\frac{\sigma_t^2(Z_i)}{\pi_t(Z_i)}]$$
$$=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}(\frac{\mathbb{1}\{T_i=t\}(Y_i-\hat{\eta}_{t,k}(Z_i))}{\pi_t(Z_i)})^2-E[(\frac{\mathbb{1}\{T_i=t\}(Y_i-\eta_t(Z_i))}{\pi_t(Z_i)})^2]$$

We again apply Lemma 2.7.2 to each fold k. This time

$$\begin{aligned}\Psi_{t,i}=&\Psi_{t',i}=\frac{\mathbb{1}\{T_i=t\}(Y_i-\eta_t(Z_i))}{\pi_t(Z_i)},\\ \hat{\Psi}_{t,i}=&\hat{\Psi}_{t',i}=\frac{\mathbb{1}\{T_i=t\}(Y_i-\hat{\eta}_{t,k}(Z_i))}{\pi_t(Z_i)}.\end{aligned}$$

Condition (i) of Lemma 2.7.2 directly follows from Assumption 1.2.1. To check condition (ii),

$$E[\Psi_{t,i}^{2+\delta}] \leq \frac{1}{\pi_{\min}^{2+\delta}} E[(Y_i(t) - \eta_t(Z_i))^{2+\delta}] \leq \infty,$$

by triangular inequality, Assumption 1.2.1 and conditions of this theorem. At last, we check condition (iii),

$$\begin{aligned}(\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\Psi}_{t,i} - \Psi_{t,i})^2)^{\frac{1}{2}} =&(\frac{1}{|I_k|}\sum_{i\in I_k}(\frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)}(\hat{\eta}_{t,k}(Z_i) - \eta_{t,k}(Z_i)))^2)^{\frac{1}{2}} \\ \leq&\frac{1}{\pi_{\min}}(\frac{1}{|I_k|}\sum_{i\in I_k}(\hat{\eta}_{t,k}(Z_i) - \eta_{t,k}(Z_i))^2)^{\frac{1}{2}}.\end{aligned}$$

This upper bound is shown earlier to be $o_p(1)$ due to cross-fitting. The rest of (2.25) follows by Lemma 2.7.2 and K being finite and fixed.

2.7.2.4 Proof of Proposition 2.3.1

Recall the estimator $\hat{\mu}_t$ is defined as

$$\hat{\mu}_t = \frac{1}{m}\sum_{i=1}^{m}\hat{\zeta}_{t,k[i]}(X_i) + \frac{1}{n}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}_{t,k[i]}(Z_i))}{\pi_t(Z_i)} + \hat{\eta}_{t,k[i]}(Z_i) - \hat{\zeta}_{t,k[i]}(X_i)).$$

Add and subtract the limiting functions $\tilde{\zeta}_t(X_i)$ and $\tilde{\eta}_t(Z_i)$, the scaled and centered distribution of $\hat{\mu}_t$ can be written into the sum of two terms,

$$\begin{aligned}\sqrt{n}(\hat{\mu}_t - \mu_t) =&\mathcal{I}_1 + \mathcal{I}_2, \\ \mathcal{I}_1 =&\sqrt{n}(\frac{1}{m}\sum_{i=1}^{m}\tilde{\zeta}_t(X_i) - \mu_t + \frac{1}{n}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}(Y_i - \tilde{\eta}_t(Z_i))}{\pi_t(Z_i)} + \tilde{\eta}_t(Z_i) - \tilde{\zeta}_t(X_i))), \\ \mathcal{I}_2 =&\sqrt{n}\sum_{k=1}^{K}(\frac{1}{m}\sum_{i\in I_k\cup\tilde{I}_k}(\hat{\zeta}_{t,k}(X_i) - \tilde{\zeta}_t(X_i)) \\ &+ \frac{1}{n}\sum_{i\in I_k}((1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)})(\hat{\eta}_{t,k}(Z_i) - \tilde{\eta}_t(Z_i)) - (\hat{\zeta}_{t,k}(X_i) - \tilde{\zeta}_t(X_i)))),\end{aligned}$$

The second term $\mathcal{I}_2$ can be shown to be $o_p(1)$ following the same steps as in the proof of theorem 2.3.1. Hence we only discuss term $\mathcal{I}_1$ here.

$$\begin{aligned}\mathcal{I}_1 =&\sqrt{n}(\frac{1}{m}\sum_{i=1}^{m}(\tilde{\zeta}_t(X_i) - E[\tilde{\zeta}_t(X_i)]) \\ &+ \frac{1}{n}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)}Y_i - \mu_t + (1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)})\tilde{\eta}_t(Z_i) - \tilde{\zeta}_t(X_i) + E[\tilde{\zeta}_t(X_i)]))\end{aligned}$$

We can see that $E[\mathcal{I}_1] = 0$. Since $Y_i(t)$, $\tilde{\eta}_t(Z_i)$ and $\tilde{\zeta}_t(X_i)$ all have finite second moments, we can further write

$$\begin{aligned}\mathcal{I}_1 =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)}Y_i - \mu_t + (1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)})\tilde{\eta}_t(Z_i) + (\frac{1}{\gamma} - 1)(\tilde{\zeta}_t(X_i) - E[\tilde{\zeta}_t(X_i)])) \\ &+ \sqrt{\frac{\gamma - 1}{\gamma^2}}\frac{1}{\sqrt{m - n}}\sum_{i=n+1}^{m}(\tilde{\zeta}_t(X_i) - E[\tilde{\zeta}_t(X_i)]) + o_p(1).\end{aligned}$$

Define

$$\tilde{\varphi}_t(Y_i, T_i, Z_i) = \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)}Y_i - \mu_t + (1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)})\tilde{\eta}_t(Z_i) + (\frac{1}{\gamma} - 1)(\tilde{\zeta}_t(X_i) - E[\tilde{\zeta}_t(X_i)])$$

Let $\boldsymbol{\mu} = (\mu_0, ..., \mu_{\mathtt{T}})'$ and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_0, ..., \hat{\mu}_{\mathtt{T}})'$, we then have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\tilde{\varphi}_0(Y_i, T_i, Z_i) \\ ... \\ \tilde{\varphi}_{\mathtt{T}}(Y_i, T_i, Z_i)\end{pmatrix} \\ &+ \sqrt{\frac{\gamma - 1}{\gamma^2}}\frac{1}{\sqrt{m - n}}\sum_{i=n+1}^{m}\begin{pmatrix}\tilde{\zeta}_0(X_i) - E[\tilde{\zeta}_0(X_i)] \\ ... \\ \tilde{\zeta}_{\mathtt{T}}(X_i) - E[\tilde{\zeta}_{\mathtt{T}}(X_i)]\end{pmatrix} + o_p(1).\end{aligned}$$

Denote $\tilde{\boldsymbol{\varphi}} = (\tilde{\varphi}_0, ..., \tilde{\varphi}_{\mathtt{T}})'$ and $\boldsymbol{\zeta} = (\tilde{\zeta}_0(X_i) - E[\tilde{\zeta}_0(X_i)], ..., \tilde{\zeta}_{\mathtt{T}}(X_i) - E[\tilde{\zeta}_{\mathtt{T}}(X_i)])'$, by the conditions of this theorem and Cauchy-Schwarz inequality, $E[\tilde{\boldsymbol{\varphi}}\tilde{\boldsymbol{\varphi}}']$ and $E[\boldsymbol{\zeta}\boldsymbol{\zeta}']$ exists. Then by central limit theorem and the fact that the two samples are independent, we have

$$\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \rightsquigarrow N(0, \tilde{V}),$$

where

$$\tilde{V} = E[\tilde{\boldsymbol{\varphi}}\tilde{\boldsymbol{\varphi}}'] + \frac{\gamma - 1}{\gamma^2} E[\boldsymbol{\zeta}\boldsymbol{\zeta}']. \tag{2.28}$$

2.7.2.5 Proof of Theorem 2.4.1

Recall that for $L \in \{LS, PS\}$ our estimator $\hat{\mu}_t^L$ is defined as

$$\hat{\mu}_t^L = \frac{1}{m}\sum_{i=1}^{m} \hat{\zeta}_t^L(X_i) + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{T_i = t\}(Y_i - \hat{\eta}_t^L(Z_i))}{\pi_t(Z_i)} + \hat{\eta}_t^L(Z_i) - \hat{\zeta}_t^L(X_i)\right).$$

Add and subtract the true conditional mean functions $\zeta_t(X_i)$ and $\eta_t(Z_i)$, the scaled and centered distribution of $\hat{\mu}_t^L$ can be written into the sum of two terms,

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_t^L - \mu_t) =& \mathcal{I}_1 + \mathcal{I}_2, \\
\mathcal{I}_1 =& \sqrt{n}\left(\frac{1}{m}\sum_{i=1}^{m}\zeta_t(X_i) - \mu_t + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{T_i = t\}U_i^t}{\pi_t(Z_i)} + \eta_t(Z_i) - \zeta_t(X_i)\right)\right), \\
\mathcal{I}_2 =& \sqrt{n}\left(\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}_t^L(X_i) - \zeta_t(X_i))\right. \\
&\left. + \frac{1}{n}\sum_{i=1}^{n}\left(\left(1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)}\right)(\hat{\eta}_t^L(Z_i) - \eta_t(Z_i)) - (\hat{\zeta}_t^L(X_i) - \zeta_t(X_i))\right)\right),
\end{aligned}$$

where $U_i^t = Y_i(t) - \eta_t(Z_i)$. The first term $\mathcal{I}_1$ is identical to what we had in the proof of theorem 2.3.1 and the second term is an $o_p(1)$. As a result, the limiting distribution for $\hat{\boldsymbol{\mu}}$ is the same as in theorem 2.3.1. In the remaining of this proof, we show that the second term $\mathcal{I}_2$ is indeed an $o_p(1)$.

$$\begin{aligned}
\mathcal{I}_2 =& \mathcal{I}_3 + \mathcal{I}_4, \\
\mathcal{I}_3 =& \sqrt{n}\left(\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}_t^L(X_i) - \zeta_t(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(\hat{\zeta}_t^L(X_i) - \zeta_t(X_i))\right), \\
\mathcal{I}_4 =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1 - \frac{\mathbb{1}\{T_i = t\}}{\pi_t(Z_i)}\right)(\hat{\eta}_t^L(Z_i) - \eta_t(Z_i)).
\end{aligned}$$

To see $\mathcal{I}_3 = o_p(1)$, we plug in $\hat{\zeta}_t^L(X_i) = f_\zeta' \hat{\beta}_{\zeta,t}^L$ and $\zeta_t^L(X_i) = f_\zeta' \beta_{\zeta,t} + r_{\zeta,t}(X_i)$,

$$\begin{aligned}
\mathcal{I}_3 =& \sqrt{n}\left(\frac{1}{m}\sum_{i=1}^{m}(f_\zeta(X_i)'(\hat{\beta}_{\zeta,t} - \beta_{\zeta,t}^L) + r_{\zeta,t}(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(f_\zeta(X_i)'(\hat{\beta}_{\zeta,t}^L - \beta_{\zeta,t}) + r_{\zeta,t}(X_i))\right) \\
=& \sqrt{n}\left(\frac{1}{m}\sum_{i=1}^{m} f_\zeta(X_i) - E[f_\zeta(X_i)]\right)'(\hat{\beta}_{\zeta,t}^L - \beta_{\zeta,t}) \qquad (2.29) \\
&+ \sqrt{n}\left(E[f_\zeta(X_i)] - \frac{1}{n}\sum_{i=1}^{n} f_\zeta(X_i)\right)'(\hat{\beta}_{\zeta,t}^L - \beta_{\zeta,t}) \qquad (2.30) \\
&+ \sqrt{n}\frac{1}{m}\sum_{i=1}^{m} r_{\zeta,t}(X_i) - \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} r_{\zeta,t}(X_i). \qquad (2.31)
\end{aligned}$$

By Assumption 2.4.1,

$$Var(\sqrt{n}\frac{1}{m}\sum_{i=1}^{m} r_{\zeta,t}(X_i)) = \frac{n}{m}Var(r_{\zeta,t}(X_i)) \leq \frac{n}{m}E[r_{\zeta,t}(X_i)^2] = o(1),$$

hence the two terms in line (2.31) are $o_p(1)$. Line (2.29) and (2.30) can be bounded by applying the asymmetric Hölder's inequality,

$$\begin{aligned}
&\sqrt{n}\left(\frac{1}{m}\sum_{i=1}^{m} f_\zeta(X_i) - E[f_\zeta(X_i)]\right)'(\hat{\beta}_{\zeta,t}^L - \beta_{\zeta,t}) \\
\leq& \sqrt{n}\left\|\frac{1}{m}\sum_{i=1}^{m} f_\zeta(X_i) - E[f_\zeta(X_i)]\right\|_\infty \|\hat{\beta}_{\zeta,t}^L - \beta_{\zeta,t}\|_1 \\
=& \sqrt{n} \cdot O_p\left(\sqrt{\frac{\log p_\zeta}{n}}\right) \cdot O_p\left(\sqrt{\frac{s^2 \log(p_\zeta \vee n)}{n}}\right),
\end{aligned}$$

where the last equality is by Assumption 2.4.2 and the maximal inequality (Lemma 2.7.3), for the latter we have $f_\zeta(X_i)$ bounded under Assumption 2.4.3. At last, under the rate condition in Assumption 2.4.3, line (2.29) is an $o_p(1)$. (2.30) is also an $o_p(1)$ following the same argument.

Next we discuss term $\mathcal{I}_4$.

$$\begin{aligned}\mathcal{I}_4 =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}((1-\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)})r_{\eta,t}(Z_i)+(1-\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)})f_\eta(Z_i)'(\hat{\beta}^L_{\eta,t}-\beta_{\eta,t}))\\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)})r_{\eta,t}(Z_i) \qquad (2.32)\\ &+\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n}f_\eta(Z_i)-E[f_\eta(Z_i)])'(\hat{\beta}^L_{\eta,t}-\beta_{\eta,t}) \qquad (2.33)\\ &+\sqrt{n}(E[f_\eta(Z_i)]-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)}f_\eta(Z_i))'(\hat{\beta}^L_{\eta,t}-\beta_{\eta,t}). \qquad (2.34)\end{aligned}$$

Since $\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)}$ is bounded under Assumption 1.2.1 and $E[\frac{\mathbb{1}\{T_i=t\}}{\pi_t(Z_i)}f_\eta(Z_i)]=E[f_\eta(Z_i)]$, line (2.32) can be bounded by Markov inequality, and (2.33), (2.34) can be bounded by asymmetric Höelder's inequality in the same way as (2.29) - (2.31). So we have $\mathcal{I}_2=\mathcal{I}_3+\mathcal{I}_4=o_p(1)$.

2.7.2.6 Proof of Theorem 2.4.2

To prove this theorem, it suffices to show that for any $t,t'\in\{0,...,\mathtt{T}\}$ and $L\in\{LS,PS\}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}^L_t(Z_i)-\hat{\zeta}^L_t(X_i))(\hat{\eta}^L_{t'}(Z_i)-\hat{\zeta}^L_{t'}(X_i))\\ &\qquad -E[(\eta_t(Z_i)-\zeta_t(X_i))(\eta_{t'}(Z_i)-\zeta_{t'}(X_i))]=o_p(1), \qquad (2.35)\\ &\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}^L_t(X_i)-\hat{\mu}_t)(\hat{\zeta}^L_{t'}(X_i)-\hat{\mu}_{t'})-E[(\zeta_t(X_i)-\mu_t)(\zeta_{t'}(X_i)-\mu_{t'})]=o_p(1), \qquad (2.36)\\ &\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\{T_i=t\}(Y_i-\hat{\eta}^L_t(Z_i))^2}{\pi_t(Z_i)^2}-E[\frac{\sigma^2_t(Z_i)}{\pi_t(Z_i)}]=o_p(1). \qquad (2.37)\end{aligned}$$

All of the above three results are proved by applying Lemma 2.7.2.

Starting with (2.35), we apply Lemma 2.7.2 with $\Psi_{t,i}=\eta_t(Z_i)-\zeta_t(X_i)$, $\hat{\Psi}_{t,i}=\hat{\eta}^L_t(Z_i)-\hat{\zeta}^L_t(X_i)$, $\Psi_{t',i}=\eta_{t'}(Z_i)-\zeta_{t'}(X_i)$, and $\hat{\Psi}_{t',i}=\hat{\eta}^L_{t'}(Z_i)-\hat{\zeta}^L_t(X_i)$. Condition (i) and (ii) in Lemma 2.7.2 are directly satisfied by Assumption 1.2.1 and the conditions of this theorem. To see

(iii), for any t, by triangular inequality,

$$\begin{aligned}(\frac{1}{n}\sum_{i=1}^{n}(\hat{\Psi}_{t,i}-\Psi_{t,i})^2)^{\frac{1}{2}} =&(\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}_t^L(Z_i)-\hat{\zeta}_t^L(X_i)-\eta_t(Z_i)+\zeta_t(X_i))^2)^{\frac{1}{2}}\\ \leq&(\frac{1}{n}\sum_{i=1}^{n}(f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t}))^2)^{\frac{1}{2}}+(\frac{1}{n}\sum_{i=1}^{n}(f_\zeta(X_i)'(\hat{\beta}_{\zeta,t}^L-\beta_{\zeta,t}))^2)^{\frac{1}{2}} \qquad (2.38)\\ &+(\frac{1}{n}\sum_{i=1}^{n}r_\eta(Z_i)^2)^{\frac{1}{2}}+(\frac{1}{n}\sum_{i=1}^{n}r_\zeta(X_i)^2)^{\frac{1}{2}}. \qquad (2.39)\end{aligned}$$

The two terms in line (2.39) are $o_p(1)$ by Assumption 2.4.1 and Markov inequality, as

$$\begin{aligned}Var(\frac{1}{n}\sum_{i=1}^{n}r_\eta(Z_i)^2)\leq\frac{1}{n}E[r_\eta(Z_i)^2]=o(1),\\ Var(\frac{1}{n}\sum_{i=1}^{n}r_\zeta(X_i)^2)\leq\frac{1}{n}E[r_\zeta(X_i)^2]=o(1).\end{aligned}$$

Line (2.38) equals

$$\|f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t})\|_{2,n}+\|f_\zeta(X_i)'(\hat{\beta}_{\zeta,t}^L-\beta_{\zeta,t})\|_{2,n}, \qquad (2.40)$$

where the second term is the prediction norm of $\hat{\beta}_{\zeta,t}$ and is $o_p(1)$ by Assumption 2.4.2. For the first term in (2.40),

$$\begin{aligned}\|f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t})\|_{2,n}^2=&\frac{n_t}{n}\|f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t})\|_{2,n_t}^2+\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{T_i\neq t\}(f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t}))^2\\ =&\frac{n-n_t}{n}\frac{1}{n-n_t}\sum_{i=1}^{n}\mathbb{1}\{T_i\neq t\}(f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t}))^2+o_p(1),\end{aligned}$$

where the last equality follows from Assumption 2.4.2 and $n_t=O_p(n)$. Next, denote

$$\mathcal{W}_n=\{(\mathbb{1}\{T_i=t\},\mathbb{1}\{T_i=t\}Z_i)\}_{i=1}^n,$$

by Assumption 2.4.2 and the condition of this theorem, we have

$$\begin{aligned}&E[\frac{1}{n-n_t}\sum_{i=1}^{n}\mathbb{1}\{T_i\neq t\}(f_\eta(Z_i)'(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t}))^2|\mathcal{W}_n]\\ =&(\hat{\beta}_{\eta,t}^L-\beta_{\eta,t})'E[f_\eta(Z_i)f_\eta(Z_i)'|T_i\neq t](\hat{\beta}_{\eta,t}^L-\beta_{\eta,t})\\ \leq&K_B^2\|\hat{\beta}_{\eta,t}^L-\beta_{\eta,t}\|_1^2=O_p(\frac{s^2\log(p_\eta\vee n)}{n})=o_p(1).\end{aligned}$$

Then applying Lemma 2.7.1, we have

$$\frac{1}{n-n_t}\sum_{i=1}^{n}\mathbb{1}\{T_i\neq t\}(f_\eta(Z_i)'(\hat{\beta}^L_{\eta,t}-\beta_{\eta,t}))^2=o_p(1).$$

Hence condition (iii) of Lemma 2.7.2 which gives (2.35). Next, we prove (2.36) by applying Lemma 2.7.2 again. This time $\Psi_{t,i}=\zeta_t(X_i)-\mu_t$, $\hat{\Psi}_{t,i}=\hat{\zeta}^L_t(X_i)-\hat{\mu}_t$, $\Psi_{t',i}=\zeta_{t'}(X_i)-\mu_{t'}$, and $\hat{\Psi}_{t',i}=\hat{\zeta}^L_{t'}(X_i)-\hat{\mu}_{t'}$. Condition (i) and (ii) in Lemma 2.7.2 are directly satisfied by Assumption 1.2.1 and the condition of this theorem. To check (iii), for any t, by triangular inequality,

$$\begin{aligned}(\frac{1}{m}\sum_{i=1}^{m}(\hat{\Psi}_{t,i}-\Psi_{t,i})^2)^{\frac{1}{2}}=&(\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}^L_t(X_i)-\hat{\mu}_t-\zeta_t(X_i)-\mu_t)^2)^{\frac{1}{2}}\\ \leq&(\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}^L_t(X_i)-\zeta_t(X_i))^2)^{\frac{1}{2}}+(\frac{1}{m}\sum_{i=1}^{m}(\hat{\mu}_t-\mu_t)^2)^{\frac{1}{2}}.\end{aligned}$$

The second term in this upper bound equals $\hat{\mu}_t-\mu_t$ which is $O_p(n^{-\frac{1}{2}})$ by theorem 2.4.1 and for the first term, by triangular inequality,

$$\begin{aligned}&(\frac{1}{m}\sum_{i=1}^{m}(\hat{\zeta}^L_t(X_i)-\zeta_t(X_i))^2)^{\frac{1}{2}}\\ \leq&(\frac{1}{m}\sum_{i=1}^{m}(f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t}-\beta_{\zeta,t}))^2)^{\frac{1}{2}}+(\frac{1}{m}\sum_{i=1}^{m}r_\zeta(X_i)^2)^{\frac{1}{2}}.\end{aligned}\tag{2.41}$$

The second term in (2.41) is $o_p(1)$ by Markov inequality and Assumption 2.4.1, and for the first term, by Assumption 2.4.2, we have

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}(f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t}-\beta_{\zeta,t}))^2\\ =&\frac{n}{m}\|f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t}-\beta_{\zeta,t})\|^2_{2,n}+\frac{1}{n}\sum_{i=n+1}^{m}f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t}-\beta_{\zeta,t})^2\\ =&\frac{m-n}{n}\frac{1}{m-n}\sum_{i=n+1}^{m}f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t}-\beta_{\zeta,t})^2+o_p(1).\end{aligned}$$

Furthermore, denote the experimental sample by $\mathcal{C}_n = \{(Y_i, Z_i, T_i)\}_{i=1}^n$, we have

$$\begin{aligned}
&E[\frac{1}{m-n}\sum_{i=n+1}^{m} f_\zeta(X_i)'(\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t})^2|\mathcal{C}_n] \\
=&(\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t})'E[f_\zeta(X_i)f_\zeta(X_i)'](\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t}) \\
\leq& K_B^2\|\hat{\beta}^L_{\zeta,t} - \beta_{\zeta,t}\|_1^2 = o_p(1),
\end{aligned}$$

by Assumption 2.4.2, 2.4.3 and the condition of this theorem. Then by Lemma 2.7.1, condition (iii) of Lemma 2.7.2 is satisfied, which in turn gives (2.36). The proof of (2.37) is very similar to the proof of (2.25) for theorem 2.3.2 and the proof of (2.35) so we choose to omit it here.

CHAPTER 3

Model Selection in Doubly Robust Policy Learning

3.1 Introduction

When treatment effects are heterogeneous, an important question is to find out how to best assign treatment to individuals based on their observable characteristics, i.e. find a good policy rule $\pi(x)$ that maps a vector of characteristics x to a treatment. For example, a job training program might only benefit workers of certain education level; some drugs may only work on patients of certain age or with certain medical history; variant advertisement styles lift sales differently depending on customer demographics. In these scenarios, the decision maker might want to look beyond the average treatment effect and search for a good policy rule. Given either experiment or quasi-experiment data, researchers can formulate a statistical decision problem and evaluate policy rules by their expected regret (Manski, 2004; Dehejia, 2005; Stoye, 2009; Bhattacharya and Dupas, 2012; Armstrong and Shen, 2015; Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Mbakop and Tabord-Meehan, 2021).

The problem could be described as follows. A population of agents with observed characteristics $X \in \mathcal{X}$ is to be treated according to a rule $\pi : \mathcal{X} \to \mathcal{D}$ selected from a class of available rules (or interventions) $\pi \in \Pi$. Each treatment rule will result in an outcome $Y \in \mathbb{R}$ (interpreted as utility). Our goal is to learn a treatment rule that maximizes the expected value of Y, denoted $V(\pi)$. For that purpose, we attempt to estimate $V(\pi)$ with $\hat{V}_n(\pi)$ using available data on the outcomes Y_i, treatments T_i, covariates X_i, and optional auxiliary variables Z_i. Specifically, we have access to a collection $W_1^n = \{W_i\}_{i=1}^n$ of i.i.d.

samples $W_i = (Y_i, T_i, X_i, Z_i)$ distributed according to $P \in \mathbf{P}$. We note that the space of observed treatments $T_i \in \mathcal{T}$ may not be the same as the space of interventions $\mathcal{D}$.

The literature has pointed out that in many situations, the set of rules that a policy maker can choose from is constrained by various practical concerns such as budget, fairness, or simplicity. We notice that this constrained set of rules, call it Π, could nevertheless be ambiguous to a practitioner. For example, regulations may dictate that only certain variables could be included in the determination of treatment assignment and a decision tree up to depth four should be employed, but whether to use all of the variables and what exact depth of trees to consider is still up to the practitioner to decide. A better policy π would very likely exist in a larger class Π, but a too complex Π might not work well with the limited amount of data. Just like in many statistical estimation problems, there is a trade-off between bias and variance. Hence, picking a right class Π is a model selection problem for the practitioner.

In this chapter, we focus on the following question: if a practitioner can choose between several different classes of policy rules, denoted Π_k for $k \geqslant 1$, which class should they choose? To answer this question, we need a criterion to compare different data-dependent treatment rules. In line with the literature on statistical treatment rules, we evaluate the performance of treatment rules in terms of their expected regret, $\mathbb{E}[R(\hat{\pi}_n)]$, where regret is defined as

$$R(\pi) = \max_{\pi' \in \Pi^*} V(\pi') - V(\pi).$$

relative to some ideal policy class Π^* that may be infeasible, unknown or arbitrarily set. In the aforementioned example, Π^* could be thought as the largest set of rules allowed under the regulation. Now, to see the trade-off in picking the class, let $\hat{\pi}_{n,k}$ denote the optimal treatment rule chosen from a class Π_k, the regret can be written as:

$$R(\hat{\pi}_{n,k}) = \underbrace{\max_{\pi \in \Pi^*} V(\pi) - \max_{\pi \in \Pi_k} V(\pi)}_{\text{Approximation Error}} + \underbrace{\max_{\pi \in \Pi_k} V(\pi) - V(\hat{\pi}_{n,k})}_{\text{Estimation Error}}.$$

Intuitively, we see that: more complex rules have a better chance of reducing the approximation error, but, for a given sample size, might have larger estimation error.

We adapts and extends two recent methods proposed in Mbakop and Tabord-Meehan (2021) and Athey and Wager (2021). Mbakop and Tabord-Meehan (2021) introduces the penalized welfare maximization (PWM) rule, which itself is an extension to the empirical welfare maximization (EWM) rule proposed in Kitagawa and Tetenov (2018). The PWM rule adds penalization to the EWM rule to achieve model selection. The authors establish a finite-sample upper bound on the expected regret of the PWM rule. The bound converge to zero at $n^{-1/2}$ rate, which is proved to be optimal. A limitation to both EMW and PWM is that when the propensity score is unknown and has to be estimated, these methods would no longer be rate-optimal. Athey and Wager (2021) propose a method that could retain the $n^{-1/2}$ rate even with estimated propensity scores by leveraging doubly robust estimation, but their method does not incorporate model selection. In this chapter, we propose a method that could achieve both.

Following the aforementioned two papers, we propose the following procedure to select the best class. Define the penalized empirical welfare function:

$$Q_{n,k}(\pi) = \hat{V}_n(\pi) - \hat{C}_{n,k}(\pi),$$

where $\hat{V}_n(\pi)$ is a doubly robust estimate of $V(\pi)$ and $\hat{C}_{n,k}(\pi)$ represents a penalty for model complexity, which, informally speaking, estimates how much the model overfits the data. For each k, solve for

$$\hat{\pi}_{n,k} = \underset{\pi \in \Pi_k}{\operatorname{argmax}} \hat{V}_n(\pi),$$

choose

$$\hat{k} = \underset{k}{\operatorname{argmax}} Q_{n,k}(\hat{\pi}_{n,k}),$$

and set

$$\hat{\pi}_n \equiv \hat{\pi}_{n,\hat{k}}.$$

Our main result is to show that such $\hat{\pi}_n$ is adaptive in a sense that it automatically picks up the "right" class and has the optimal rate of convergence in terms of expected regret. Our regret bounds hold in finite samples, are tighter than the bounds available in the literature

and easily generalize to arbitrary discrete policy rules. Moreover, since the welfare estimation $\hat{V}_n(\pi)$ is based on doubly robust scores, our method retains the optimal $n^{-1/2}$ rate in general setups including quasi experiments where the propensity scores have to be estimated.

In Section 3.2, we further describe the setup and introduce our assumptions. In Section 3.3, we revisit known results from the literature and present modified and refined versions of them. Our main results are in Section 3.4, where we formally introduce our new algorithm, the robust penalized welfare maximization (RPWM) rule. We present bound on expect regret of the RPWM rule and prove that it is rate-optimal. Section 3.5 presents a simulation study and Section 3.6 concludes. Proofs are collected in the appendix which forms Section 3.7.

3.2 Setup

We consider the standard potential outcomes framework (Neyman, 1923; Rubin, 1974). Specifically, let $Y_i(t)$ denote an outcome that we would have observed if the treatment had been set to $T_i = t$, and $Y = Y(T)$ denote the observed outcome. Let $\theta = \mathbb{E}[\tau(X)]$ denote the average treatment effect. Our main assumption, following Athey and Wager (2021) and Chernozhukov et al. (2016), is that we can identify θ via a doubly-robust moment condition.

Assumption 3.2.1 (Identification)**.** *Let $m(x,t) = \mathbb{E}_P[Y(t)|X = x] \in \mathcal{M}$. Assume that $m(x,t)$ induces a treatment effect function $\tau_m(x,t)$ such that:*

1. *The welfare function can be expressed as $V(\pi) = \mathbb{E}_P[\pi(X)\tau(X)]$, where $\tau(X) = \mathbb{E}_P[\tau_m(X,T)|X]$.*

2. *The map $m \mapsto \tau_m$ is linear and there is a weighting function $g(x,z)$ such that for any $\tilde{m}(x,t) \in \mathcal{M}$*

$$\mathbb{E}_P[\tau_{\tilde{m}}(X,T) - g(X,Z)\tilde{m}(X,T)|X] = 0.$$

The auxiliary variable Z could be an instrumental variable, or equals to X when X is

exogeneous. We illustrate this setting with three important examples borrowed from (Athey and Wager, 2021).

Example 3.2.1 (Binary Treatments with Selection on Observables)**.** Under conditional ignorability assumption $T \perp (Y(1), Y(0))|X$, condition 2 in Assumption 3.2.1 is satisfied with

$$g(x,t) = \frac{t - e(x)}{e(x)(1-e(x))}, \quad \tau_m(x) = m(x,1) - m(x,0),$$

where $e(x) = P(T = 1|X = x)$ is the propensity score. Then the welfare function is

$$V(\pi) = \mathbb{E}_P[\pi(X)\tau(X)] = \mathbb{E}_P[Y(\pi(X))] - \mathbb{E}_P[Y(0)],$$

which corresponds to our utilitarian welfare objective.

Example 3.2.2 (Endogenous Binary Treatments with Binary Instruments)**.** Assume that Z is a valid instrument conditional on X in the sense of Assumption 2.1 of Abadie (2003), and further assume that conditional average treatment effect equals conditional local average treatment effect, then we can have

$$\tau_m(x) = m(x,1) - m(x,0) = \frac{\mathbb{C}ov[Y, Z|X = x]}{\mathbb{C}ov[T, Z|X = x]}.$$

Then condition 2 in Assumption 3.2.1 is satisfied with

$$\begin{aligned} g(x,z) &= \frac{1}{\Delta(x)} \frac{z - \Xi(x)}{\Xi(x)(1-\Xi(x))}, \\ \Xi(x) &= P[Z = 1|X = x], \\ \Delta(x) &= P[W = 1|Z = 1, X = x] - P[W = 0|Z = 1, X = x] \end{aligned}$$

Since $\tau_m(x)$ is the same as in the Example 3.2.1, the resulting welfare function is the same.

Example 3.2.3 (Continuous Treatments)**.** Suppose the treatment variable T is continuous and exogenous, i.e. $\{Y(t)\} \perp T|X$, then we let

$$\tau_m(x,t) = \frac{d}{dv} m(x, t+v)\Big|_{v=0}.$$

Under regularity conditions, condition 2 of Assumption 3.2.1 is then satisfied with a function $g(X,T)$ derived via integration by parts (Powell et al., 1989)

$$\int\int \frac{d}{dt} m(X,T)\Big|_{t=T} dF_{T|X} dF_X = \int\int \frac{d}{dt} g(X,T) m(X,T) dF_{T|X} dF_X,$$
$$g(X,T) = -\frac{d}{dt}\log(f(t|X))\Big|_{t=T}.$$

In this case the welfare function is

$$V(\pi) = \frac{d}{dv}\mathbb{E}[Y(T + v\pi(X))]\Big|_{v=0},$$

which is the average effect of a nudge following policy $\pi(x)$.

In the above settings, Chernozhukov et al. (2016) proposed estimating θ by

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1} \hat{\Gamma}_i, \quad \hat{\Gamma}_i = \tau_{\hat{m}}(X_i, T_i) + \hat{g}(X_i, Z_i)(Y_i - \hat{m}(X_i, T_i)),$$

where $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ are preliminary estimates of the nuisance functions $g(\cdot)$ and $m(\cdot)$. Using cross-fitting and Neyman orthogonality of the moment condition $\theta = \mathbb{E}[\Gamma(W; m, g)]$, the authors show that $\hat{\theta}_n$ is $\sqrt{n}$-consistent and asymptotically Normal, provided that $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ converge sufficiently fast, and may also be semiparametrically efficient (Newey, 1994).

Athey and Wager (2021) proposed using the orthogonal scores $\hat{\Gamma}_i$ for policy learning. Specifically, under Assumption 3.2.1, $V(\pi) = \mathbb{E}(\pi(X)\Gamma(W))$, so that a feasible sample analog can be constructed as $\hat{V}_n(\pi) = n^{-1}\sum_{i=1}^n \pi(X_i)\hat{\Gamma}_i$. Then, by establishing that $\hat{V}_n(\pi)$ approximates $V(\pi)$ uniformly well over $\pi \in \Pi$, Athey and Wager (2021) show that $\hat{\pi}_n = \text{argmax}_{\pi\in\Pi}\hat{V}_n(\pi)$ is rate-optimal in terms of expected regret.[1]

In section 3.4, we propose our method that complements their results with model selection. Specifically, we propose a procedure that selects the "best" class of treatment rules to choose from in a data-driven fashion. It resolves the trade-off between approximation and

[1]Athey and Wager (2021) work with $A(\pi) = 2V(\pi) - \mathbb{E}(\tau(X)) = \mathbb{E}((2\pi(X) - 1)\tau(X))$ and its feasible analog, but the modification here changes neither the problem nor the solution.

estimation error described earlier and can also be extended to handle policy classes of infinite VC-dimension as in Mbakop and Tabord-Meehan (2021). Another difference between our results to theirs is that our bounds hold in finite sample while they derived asymptotic bounds.

Now, we state the high-level assumptions on the first stage estimation that provides us with $\hat{\Gamma}_i$.

Assumption 3.2.2 (DGP and First-stage Estimators)**.** *In the setting of Assumption 3.2.1, assume that* $\mathbb{E}_P[m^2(X,T)] \vee \mathbb{E}_P[\tau_m^2(X,T)] \vee \mathbb{E}[g^2(X,Z)] < \infty$*, and we have access to estimators* $\hat{m}(x,d)$*,* $\tau_{\hat{m}}(x,d)$*, and* $\hat{g}(x,z)$ *depending on the data* W_1^n *and satisfying the following conditions. For some* $0 < \zeta_m, \zeta_g < 1$*, with* $\zeta_m + \zeta_g \geqslant 1$*, and a positive sequence* $a(n) \to 0$ *as* $n \to 0$*,*

$$\mathbb{E}_P[(\hat{m}(X,T) - m(X,T))^2] \vee \mathbb{E}_P[(\tau_{\hat{m}}(X,T) - \tau_m(X,T))^2] \leqslant \frac{a(n)}{n^{\zeta_m}},$$

$$\mathbb{E}_P[(\hat{g}(X,Z) - g(X,Z))^2] \leqslant \frac{a(n)}{n^{\zeta_g}},$$

where (X,D,Z) *is an independent test sample drawn from* P*, for all* $P \in \mathbf{P}$*.*

The above assumptions on first stage estimation is weaker than the equivalent in Athey and Wager (2021) as we do not assume uniform consistency. Next, we assume that the policy classes have finite VC dimensions.

Assumption 3.2.3 (Policy Rules)**.** *The class of available policy rules is* $\Pi = \bigcup_{k=1}^{K} \Pi_k$*, for some finite* K*, and each* Π_k *has a finite VC dimension denoted* $VC(\Pi_k)$*. The no-treatment rule,* $\pi(x) = 0$ *for all* $x \in \mathcal{X}$*, is included in each* Π_k*.*

At last, we assume that the function $g(x,z)$ is bounded away from zero.

Assumption 3.2.4 (Overlap Condition)**.** *There is an* $\eta > 0$ *such that the weighting function satisfies* $\sup_{x,z} |g(x,z)| \leqslant \eta^{-1}$ *for all* $P \in \mathbf{P}$*.*

3.3 Related Results

In this section, we revisit some closely related known results in the literature and present modified and improved version of them.

First, we revisit regret bounds of Kitagawa and Tetenov (2018). Assume that we are in the setting of Example 3.2.1 and the propensity score is known. Then, the welfare can be expressed as

$$V(\pi) = \mathbb{E}\left[\pi(X)\left(\frac{YT}{e(X)} - \frac{Y(1-T)}{1-e(X)}\right)\right],$$

where $e(X) = P(T = 1|X)$ denotes the propensity score, with a sample counterpart

$$\hat{V}_n^E(\pi) = \frac{1}{n}\sum_{i=1}^{n}\pi(X_i)\left(\frac{Y_iT_i}{e(X_i)} - \frac{Y_i(1-T_i)}{1-e(X_i)}\right). \tag{3.1}$$

Kitagawa and Tetenov (2018) consider the Empirical Welfare Maximization (EWM) rule, defined as

$$\hat{\pi}_n^{EWM} = \underset{\pi\in\Pi}{\operatorname{argmax}}\,\hat{V}_n^E(\pi).$$

They derive the upper bound on the worst-case expected regret of this rule over all distributions with bounded outcomes and propensity scores. In the following theorem, we extend and sharpen their result allowing for unbounded outcomes.[2] Define a set of distributions:

$$\mathcal{P}_{B,\eta} = \{P \in \mathbf{P} : \eta \leqslant P(T = 1|X) \leqslant 1 - \eta \text{ a.s.}, \mathbb{E}_P[Y^2] \leqslant B^2\}.$$

Theorem 3.3.1 (EWM Revisited)**.** *Assume that treatments are binary, $T = \{0, 1\}$, unconfoundedness holds, $(Y(0), Y(1)) \perp T|X$, and the propensity score $e(X)$ is known. Let $\hat{\pi}_n^{EWM} = \operatorname{argmax}_{\pi\in\Pi}\hat{V}_n^E(\pi)$, with $\hat{V}_n^E(\pi)$ defined in (3.1), denote the EWM rule. Then,*

$$\sup_{P\in\mathcal{P}_{B,\eta}}\mathbb{E}_P[R(\hat{\pi}_n^{EWM})] \leqslant C\frac{B}{\eta}\sqrt{\frac{VC(\Pi)}{n}},$$

[2]In addition to allowing unbounded outcomes, we obtain a substantially smaller constant. Kitagawa and Tetenov (2018) assume that $Y \in [-M/2, M/2]$ and derive an upper bound of the form $KM\eta^{-1}\sqrt{VC(\Pi)/n}$. A careful examination of the proof of their Theorem 1 suggests that the result holds with $K \approx 68$. To compare, note that for any distribution P such that $Y \in [-M/2, M/2]$, we have $(\mathbb{E}_P[Y^2])^{1/2} \leqslant M/2$. Then, our Theorem 3.3.1 implies that the expected regret bound holds with $C/2M\eta^{-1}\sqrt{VC(\Pi)/n}$, where $C/2 = 29$.

where $C \leqslant 58$ is a universal constant.

We complement this result with a tight lower bound to show that the EWM rule is rate-optimal.

Theorem 3.3.2 (Regret Lower Bound)**.** *Under the assumptions of Theorem 3.3.1, for all fixed $n \geqslant 4VC(\Pi)/\eta$,*

$$\inf_{\hat{\pi}_n} \sup_{P \in \mathcal{P}_{B,\eta}} \mathbb{E}_P[R(\hat{\pi}_n)] \geqslant 0.07 \cdot \frac{B}{\eta} \sqrt{\frac{VC(\Pi) - 1}{n}} - \kappa_n,$$

where $\kappa_n = 0.14B/\eta \cdot (VC(\Pi) - 1)/n$, and the right-hand side is positive.

Remark 3.3.1 (Unknown Propensity Score)**.** One important limitation of the above result is the assumption that the propensity score is known. If the propensity score is unknown and has to be estimated, one can plug the estimator in (3.1) and maximize the corresponding objective function. Then, a result similar to Theorem 3.3.1 holds with an additional $O(\phi_n^{-1})$ term, where ϕ_n is the rate of convergence of the propensity score estimator, which is generally slower than $\sqrt{n}$. In such cases, $\hat{\pi}_n^{EWM}$ is no longer rate-optimal.

In the same setting, Mbakop and Tabord-Meehan (2021) propose a treatment rule that accounts for model selection, called Penalized Welfare Maximization (PWM). Here, for simplicity, we only revisit the so-called holdout procedure defined as follows:

1. Let $l = \lceil (1-s)n \rceil$ and $r = n - l$ for some $s \in (0, 1)$, and call $W_1, \ldots, W_l$ the estimating sample, and $W_{l+1}, \ldots, W_n$ the test sample. We use subscripts l, r, and n for quantities that depend on the estimating sample only, on the test sample only, and on the entire sample.

2. Compute the EWM rules $\hat{\pi}_{l,k} \equiv \text{argmax}_{\pi \in \Pi_k} \hat{V}_l^E(\pi)$ for each Π_k using the estimating sample. Evaluate each $\hat{\pi}_{l,k}$ by computing the penalized welfare $Q_{n,k}(\hat{\pi}_{l,k}) = \hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{C}_{n,k}$ where the penalty is $\hat{C}_{n,k} = \hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{V}_r^E(\hat{\pi}_{l,k})$.

3. Select $\hat{k} = \text{argmax}_k \, Q_{n,k}(\hat{\pi}_{l,k})$, and define[3] $\hat{\pi}_n^{PWM} \equiv \hat{\pi}_{n,\hat{k}}$.

This procedure is natural: we estimate each $\hat{\pi}_{l,k}$ using the estimating sample, evaluate their performance by computing the empirical welfare on the test sample, $Q_{n,k} = \hat{V}_r(\hat{\pi}_{l,k})$, and select the best estimator. The following result shows that such estimator automatically selects the best class and attains the optimal rate of convergence.[4]

Theorem 3.3.3 (PWM Revisited)**.** *Assume that treatments are binary, $T = \{0, 1\}$, unconfoundedness holds, $(Y(0), Y(1)) \perp T|X$, and the propensity scores are known. Let $\hat{\pi}_n$ denote the PWM rule computed with the holdout penalty as described above. Then, for any $P \in \mathcal{P}_{B,\eta}$,*

$$\mathbb{E}_P[R(\hat{\pi}_n^{PWM})] \leqslant \inf_{k \leqslant K} \left\{ V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}_P[\hat{C}_{n,k}] \right\} + R_n$$

where V_Π^* and $V_{\Pi_k}^*$ denote the maximum welfare attainable within the corresponding classes (both depend on P), and $R_n = B/\eta \cdot K/\sqrt{sn}$.

Moreover, letting $\mathcal{P}_{B,\eta}^k \subset \mathcal{P}_{B,\eta}$ be a set of distributions such that $V_\Pi^* = V_{\Pi_k}^*$,

$$\sup_{P \in \mathcal{P}_{B,\eta}^k} \mathbb{E}_P[R(\hat{\pi}_n^{PWM})] \leqslant \frac{B}{\eta} \left(C \sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K \sqrt{\frac{1}{sn}} \right)$$

where $C \leqslant 58$ is a universal constant.

To gain interpretation, recall that selecting the best Π_k amounts to balancing the approximation error $V_\Pi^* - V_{\Pi_k}^*$ and the estimation error $V_{\Pi_k}^* - V(\hat{\pi}_{l,k})$. The estimation error is at the same rate as $\mathbb{E}[\hat{C}_{n,k}]$ under the hold-out penalty (Mbakop and Tabord-Meehan, 2021). Also, intuitively, one could think that the term $\hat{C}_{n,k} = \hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{V}_r^E(\hat{\pi}_{l,k})$ as an estimator for $V_{\Pi_k}^* - V(\hat{\pi}_{l,k})$, or at least a measure of over-fitting. Therefore, the above result shows

[3]A slight abuse of notation here: $\hat{\pi}_{n,\hat{k}}$ is obtained by plugging in $k = \hat{k}$ into $\hat{\pi}_{l,k}$. However, we replace the l by n here (didn't write $\hat{\pi}_{l,\hat{k}}$) to stress that the rule now depends on the whole sample as $\hat{k}$ depends on the whole sample.

[4]Our result refines Theorem 3.1. and Corollaries 3.2 and 3.3. of Mbakop and Tabord-Meehan (2021) for holdout penalty and a finite number of policy classes.

the oracle property of $\hat{\pi}_n^{PWM}$: it behaves as if we knew the right class ex ante and used it to compute the optimal treatment rule.

The difference in $V^*_{\Pi_k} - V(\hat{\pi}_{l,k})$ is in π but the difference in $\hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{V}_r^E(\hat{\pi}_{l,k})$ is in the way how V is estimated, so I don't see why $\hat{V}_l^E(\hat{\pi}_{l,k})$ would be estimating $V^*_{\Pi_k}$ but not $\hat{V}_r^E(\hat{\pi}_{l,k})$. Ideally we should have $\hat{V}_l^E(\pi^*_k)$, then you would say $\hat{\pi}_{l,k}$ is estimating π^*_k, but then what about $\hat{V}_r^E(\hat{\pi}_{l,k})$, why is this not estimating $\hat{V}_r^E(\pi^*_k)$ and then in turn also estimating $V^*_{\Pi_k}$

The goal of this chapter is to construct an estimator with a similar oracle property in a more general setting of Section 3.2 by combining doubly-robust welfare estimator and model selection.

3.4 Main Results

We return to the general setting introduced in Section 3.2. Under Assumption 3.2.1, the welfare can be written as

$$V(\pi) = \mathbb{E}[\pi(X)\Gamma(W)],$$

and the feasible sample analog is given by

$$\hat{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\hat{\Gamma}_i.$$

We further require that the estimated orthogonal scores $\hat{\Gamma}_i$ are computed using J-fold cross-fitting, defined as follows. Split the sample into J evenly sized folds of size $\lfloor n/J \rfloor$ distributing the remaining observations uniformly, and let $j : \{1, \dots, n\} \to \{1, \dots, J\}$ be a function that identifies the fold $j(i)$ to which observation i belongs. Then, let $\hat{g}^{(-j(i))}$, $\hat{m}^{(-j(i))}$, and $\tau_{\hat{m}}^{(-j(i))}$ denote the first-stage estimators computed using $(1 - J^{-1})n$ observations excluding the fold $j(i)$, and compute

$$\hat{\Gamma}_i = \tau_{\hat{m}}^{(-j(i))}(X_i, T_i) + \hat{g}^{(-j(i))}(X_i, Z_i)(Y_i - \hat{m}^{(-j(i))}(X_i, T_i)).$$

Following Athey and Wager (2021), we define a Doubly-Robust EWM estimator as

$$\hat{\pi}_n^{REWM} = \underset{\pi \in \Pi}{\operatorname{argmax}} \hat{V}_n(\pi).$$

Our first goal is to bound its expected regret in finite samples. To this end, we define:

$$\tilde{V}_n(\pi) = \frac{1}{n} \sum_{i=1}^{n} \pi(X_i) \Gamma_i,$$

and show that, under Assumption 3.2.1-2 and appropriate moment conditions,

$$\mathbb{E}\left[\sup_{\pi \in \Pi} |\tilde{V}_n(\pi) - V(\pi)|\right] \leqslant \tilde{C} \sqrt{\frac{VC(\Pi)}{n}} \qquad \mathbb{E}\left[\sup_{\pi \in \Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|\right] = o(n^{-1/2}).$$

That is, not knowing the propensity scores (and other nuisance parameters) only comes at a $o(n^{-1/2})$ price, meaning that $\hat{\pi}_n^{REWM}$ has the optimal rate of convergence.

Here, we impose more explicit restrictions on the distributions of the data, in line with our Assumptions 3.2.2 and 3.2.4. Specifically, we define:[5]

$$\mathcal{P}_{B_\tau, B, \eta} = \left\{ P \in \mathbf{P} : \begin{array}{c} \mathbb{E}_P[\tau_m^2(X,T)] \leqslant B_\tau^2 \\ \mathbb{E}_P[(Y - m(X,T))^2 | X, T] \overset{a.s.}{\leqslant} B^2 \\ \sup_{x,z} |g(x,z)| \leqslant \eta^{-1} \end{array} \right\}, \tag{3.2}$$

and prove the following result.

Theorem 3.4.1 (Doubly-Robust EWM)**.** *Let Assumptions 3.2.1 – 3.2.4 hold and $\hat{\pi}_n^{REWM}$ denote the Doubly-Robust EWM estimator defined above, with the first stage estimators for the nuisance parameters constructed using a J-fold sample splitting. Then,*

$$\sup_{P \in \mathcal{P}_{B_\tau, B, \eta}} \mathbb{E}_P[R(\hat{\pi}_n^{REWM})] \leqslant C \frac{\sqrt{B_\tau^2 \eta^2 + B^2}}{\eta} \sqrt{\frac{VC(\Pi)}{n}} + R_n,$$

[5]To relate this with the set $\mathcal{P}_{B,\eta}$ defined prior to Theorem 3.3.1, recall from Example 3.2.1 that $\tau_m(X,T) = m(X,1) - m(X,0)$ so that $\mathbb{E}(\tau_m^2) \leqslant 4B^2$ provided that $\mathbb{E}_P[(Y - m(X,T))^2 | X, T] \leqslant B^2$. The latter neither implies nor is implied by $\mathbb{E}_P[Y^2] \leqslant B^2$.

where $C \leqslant 58$ *is a universal constant, and* $R_n = 2(R_{1,n} + R_{2,n} + R_{3,n})$ *with*

$$R_{1,n} = C\sqrt{(J+2)B^2 \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}},$$

$$R_{2,n} = C\sqrt{(J+2)\frac{2(\eta^2+1)}{\eta^2} \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}},$$

$$R_{3,n} = \sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}}.$$

It is instructive to compare this result with Theorem 3.3.1 in the context of binary treatments under unconfoundedness (see Example 3.2.1). Recall that when the propensity scores are unknown, the analog of Theorem 3.3.1 holds with an additional $O(\phi_n^{-1})$ term, where ϕ_n is a rate of convergence of the propensity score estimator. The latter is generally slower than root-n, meaning that $\hat{\pi}_n^{EWM}$ is not rate-optimal. On the other hand, under the assumptions of Theorem 3.4.1, the extra term in the upper bound is $R_n = o(n^{-1/2})$. Therefore, $\hat{\pi}_n^{REWM}$ is rate-optimal, whether the propensity score is known or not, which illustrates the main advantage of using robust welfare estimates.

Next, we present our main result which adds model selection. We propose using a Robust Penalized Welfare Maximization (RPWM) treatment rule, defined as follows.

1. Let $l = \lceil (1-s)n \rceil$ and $r = n - l$ for some $s \in (0,1)$, and call $W_1, \ldots, W_l$ the estimating sample, and $W_{l+1}, \ldots, W_{l+r}$ the test sample. We use subscripts l, r, and n for quantities that depend on the estimating sample only, on the test sample only, and on the entire sample.

2. Compute the RWM rules $\hat{\pi}_{l,k} \equiv \operatorname{argmax}_{\pi \in \Pi_k} \hat{V}_l(\pi)$ for each Π_k using the estimating sample with $\hat{\Gamma}_i$ computed using a J-fold cross-fitting. Evaluate each $\hat{\pi}_{l,k}$ by computing the penalized welfare $Q_{n,k}(\hat{\pi}_{l,k}) = \hat{V}_l(\hat{\pi}_{l,k}) - \hat{C}_{n,k}$ where the penalty is $\hat{C}_{n,k} = \hat{V}_l(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})$.

3. Select $\hat{k} = \operatorname{argmax}_k Q_{n,k}(\hat{\pi}_{l,k})$, and define $\hat{\pi}_n^{RPWM} \equiv \hat{\pi}_{n,\hat{k}}$.

The following result shows that such estimator automatically selects the best class and attains the optimal rate of convergence.

Theorem 3.4.2. *Let Assumptions 3.2.1 – 3.2.4 hold and $\hat{\pi}_n^{RPWM}$ denote the Doubly-Robust PWM estimator defined above, with the first stage estimators for the nuisance parameters constructed using a J-fold sample splitting. Then:*

$$\mathbb{E}_P\left[R(\hat{\pi}_n^{RPWM})\right] \leqslant \inf_{k\leqslant K}\{V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}]\} + S_n,$$

where V_Π^* and $V_{\Pi_k}^*$ denote the maximum welfare attainable within the corresponding policy classes (both depend on P), and $S_n = O(\sqrt{B_\tau^2\eta^2 + B^2}/\eta \cdot 1/\sqrt{sn})$.

Moreover, letting $\mathcal{P}^k_{B_\tau,B,\eta} \subset \mathcal{P}_{B_\tau,B,\eta}$ be a set of distributions such that $V_\Pi^* = V_{\Pi_k}^*$,

$$\sup_{P\in\mathcal{P}^k_{B_\tau,B,\eta}} \mathbb{E}_P[R(\hat{\pi}_n^{RPWM})] \leqslant \frac{\sqrt{B_\tau^2\eta^2 + B^2}}{\eta}\left(C\sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K\sqrt{\frac{1}{sn}}\right) + S_{1,n}^k + S_{2,n}$$

where $C \leqslant 58$ is a universal constant, $S_{1,n}^k = R_{1,(1-s)n}^k + R_{2,(1-s)n}^k + R_{3,(1-s)n}$, where $R_{1,n}^k$, $R_{2,n}^k$, and $R_{3,n}$ are given in Theorem 3.4.1 with Π_k instead of Π, and $S_{2,n} = o(n^{-1/2})$.

Note that this theorem is comparable to Theorem 3.3.3. It shows the same oracle property as PWM discussed in Section 3.3. Moreover, by incorporating the doubly robust score, RPWM can retain the $n^{-1/2}$ rate in more general settings as the REWM rule. Hence, we are able to get the benefit of both worlds.

Or final result is a lower bound on expected regret that shows the $n^{-1/2}$ rate is indeed optimal.

Theorem 3.4.3. *Under Assumption 3.2.1 and 3.2.4, with $\mathcal{P}_{B,\eta}$ defined in 3.2, for any policy rule $\hat{\pi}_n$ as a function of W_1^n, we have*

$$\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant 0.07 \cdot \frac{B}{\eta}\sqrt{\frac{d}{n}} - \frac{0.14}{\sqrt{\eta}} \cdot \frac{B}{\eta}\frac{d}{n}.$$

3.5 Simulation

In this section, we conduct a simple simulation to demonstrate how RPWM rule balances between approximation error and estimation error.

We generate a random sample of size n with the following DGP.

$$\begin{aligned} Y(0) &= 0.7(X_3 + X_4 + \epsilon_0), \\ Y(1) &= X_2 - X_1 + 0.7(X_3 + X_4 + \epsilon_1), \\ P(T = 1|X) &= \Lambda(\log(0.5) + (X_1 + X_2 + X_3 + X_4)(\log(2) - \log(0.5))/4). \end{aligned}$$

where all covariates follow $U[0, 1]$ and errors follow $N(0, 1)$ independently. The $\Lambda(\cdot)$ denotes the logistic function so the propensity score is in between $\frac{1}{3}$ to $\frac{2}{3}$. Under this DGP, the average treatment effect is zero. However, There is heterogeneous treatment effect and

$$\mathbb{E}[Y(1) - Y(0)|X] = X_2 - X_1,$$

which suggest that the first best treatment policy is $\mathbb{1}\{X_2 \geq X_1\}$. Consider the $[0, 1]\times[0, 1] \subset R^2$ square where (X_1, X_2) belongs, the 45 degree diagonal line across this square would be the boundary of the first best treatment policy.

Now, for policy rule learning, suppose we arbitrarily decided to focus on decision trees that only splits on X_1 and X_2 and up to depth 4. We compare three different algorithms, the first one only considers depth 2 trees, the second one only considers depth 4 trees, and then an adaptive one which chooses across depths 2, 3 and 4 using the hold-out penalty. The last algorithm corresponds to RPWM and the first two REWM. We run Monte Carlo simulations with $n \in \{200, 400, 800, 1200, 1600, 2000\}$ and plot the regrets in Figure 3.1. 200 simulations were run for each sample size.

We see that when sample size is small, the estimation error would dominate, hence focusing on depth 2 trees leads to less regret. When the sample size is large, the approximation error would dominate so depth 4 trees become more favorable. The adaptive RPWM rule

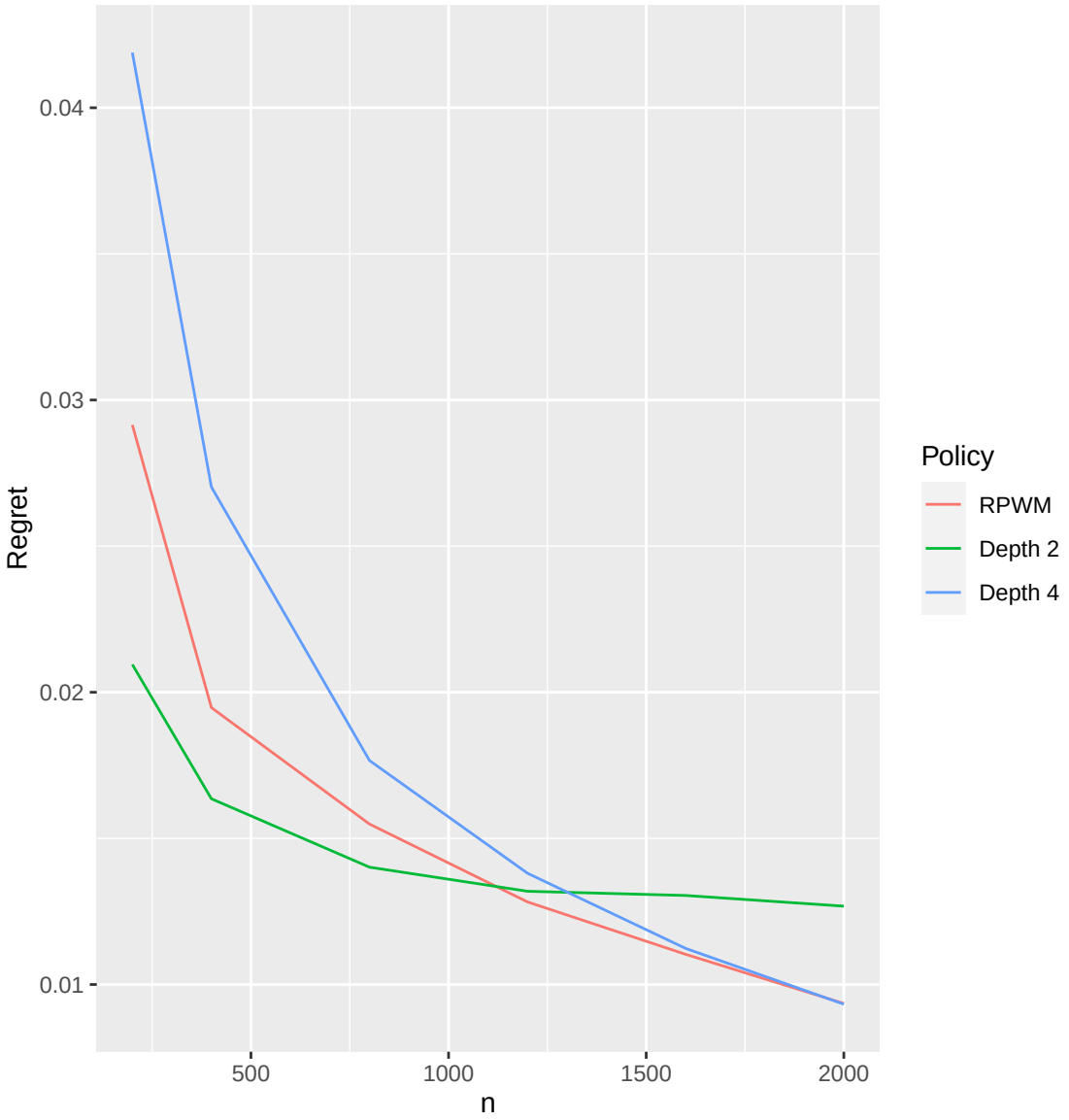
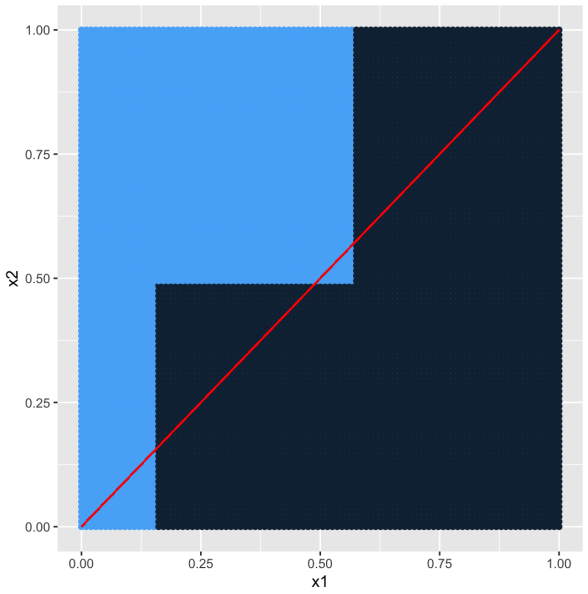


Figure 3.1: Regrets of 3 Algorithms with Different Sample Sizes

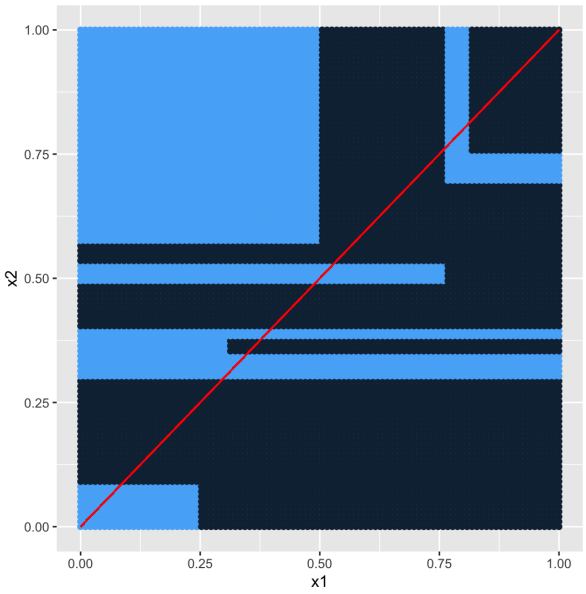
should ideally trace the lower envelope of the other two curves. That is similar to what it behaves in this simulation. We do notice a relatively poorer performance when sample size is small. This might be due to the fact that hold-out penalty effectively reduce sample size.

At last, we show some policy rules learned from the depth 2 and 4 trees at $n = 200$ and 2000 in Figure 3.2. We can see that the depth 4 tree behaves poorly at $n = 200$ due to

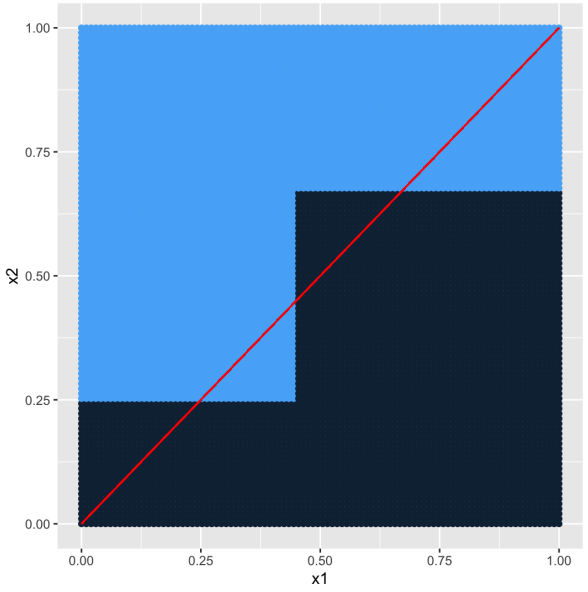
over-fitting while does a good job approximating the first best policy rule when $n = 2000$.



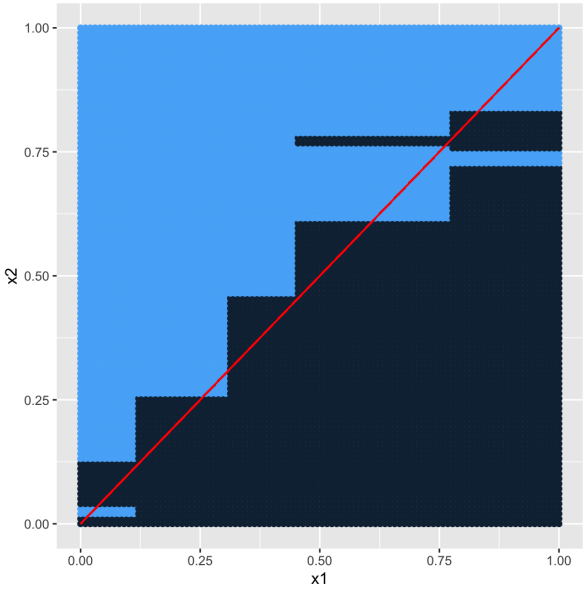
(a) A depth 2 tree with $n = 200$.



(b) A depth 4 tree with $n = 200$.



(c) A depth 2 tree with $n = 2000$.



(d) A depth 4 tree with $n = 2000$.

Figure 3.2: Examples of Policy Trees Learned with $n = 200$ and 2000.

3.6 Conclusion

In this chapter, we studied model selection in doubly robust policy learning. Following Mbakop and Tabord-Meehan (2021) and Athey and Wager (2021), we added hold-out penalty to the doubly robust policy learning algorithm. The resulting method could achieve data-driven model selection while retaining optimal $n^{-1/2}$ rate under general setups including quasi-experiments where propensity scores are unknown. By deriving finite sample upper bounds on expected regret, we show that the algorithm can automatically balance approximation error with estimation error. We also refined some related results in the literature and derived a new finite sample lower bound to show that the $n^{-1/2}$ rate is indeed optimal.

3.7 Appendix

3.7.1 Known Results for Reference and Some Refinements

First, we recite a well-known symmetrization inequality. See, e.g., Lemma 2.3.1. in van der Vaart and Wellner (1996).

Lemma 3.7.1 (Symmetrization)**.** *Let $W_1, \ldots, W_n$ be an i.i.d. sample. Then for any class of measurable functions $\mathcal{F}$,*

$$\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}(f(W_i))\right|\right] \leqslant 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(W_i)\right|\right]$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables independent from $W_1, \ldots, W_n$.

Let ψ be a strictly increasing, convex function with $\psi(0) = 0$ and X be a random variable. Then the Orlisz norm $||X||_\psi$ is defined as

$$||X||_\psi = \inf\left\{C > 0 : \mathbb{E}\left(\psi\left(\frac{|X|}{C}\right)\right) \leqslant 1\right\}.$$

Then, the following maximal inequality holds.

Lemma 3.7.2 (Maximal Inequality with Orlisz Norms). *For any random variables $X_1, \ldots, X_n$ and any strictly increasing, convex function ψ,*

$$\mathbb{E}\left[\max_{j\leqslant m}|X_j|\right] \leqslant \psi^{-1}(m)\max_{j\leqslant m}||X_j||_\psi$$

Proof. For any $C > 0$,

$$\begin{aligned}\psi\left(\mathbb{E}\left[\max_{j\leqslant m}\tfrac{|X_j|}{C}\right]\right) &\leqslant \mathbb{E}\left[\max_{j\leqslant m}\psi\left(\tfrac{|X_j|}{C}\right)\right]\\ &\leqslant m\max_{j\leqslant m}\mathbb{E}\left[\psi\left(\tfrac{|X_j|}{C}\right)\right],\end{aligned}$$

where the first inequality holds because ψ is convex and non-decreasing. Therefore, for any C such that $\max_{j\leqslant m}\mathbb{E}\left[\psi\left(|X_j|/C\right)\right] \leqslant 1$, we have

$$\mathbb{E}\left[\max_{j\leqslant m}|X_j|\right] \leqslant C\psi^{-1}(m).$$

Choosing $C = \max_{j\leqslant m}||X_j||_\psi$ concludes the proof. ■

The following result is Theorem 2.6.4. from Van der Vaart and Wellner (1996) with a precisely pineed down universal constant.

Lemma 3.7.3 (Covering Numbers for VC classes). *For any VC-class $\mathcal{C}$ of sets, any probability measure Q, any $r \geqslant 1$, and $0 < \varepsilon < 1$,*

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}}V(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{C})-1)}.$$

Proof. We closely follow the proof of Theorem 2.6.4. in van der Vaart and Wellner (1996). We start by referencing the main steps and introducing the necessary notation. First, note that $||\mathbf{1}_C - \mathbf{1}_D||_{Q,r} = Q^{1/r}(C\triangle D)$, so an ε^r-cover under $L_1(Q)$ produces an ε-cover under $L_r(Q)$. Therefore, the result for $r > 1$ follows immediately from the result for $r = 1$. Second, one can argue that it suffices to consider empirical type measures Q supported on a large enough finite set of distinct points $\{x_1, \ldots, x_n\}$. Third, it is more convenient to bound the packing number $D(\varepsilon, \mathcal{C}, L_1(Q))$ first and use the fact that $N(\varepsilon, \mathcal{C}, L_1)(Q)) \leqslant D(\varepsilon/2, \mathcal{C}, L_1(Q))$.

Each set $C \in \mathcal{C}$ can be identified with a binary vector $\mathbf{1}_C = (\mathbf{1}(x_i \in C))_{i=1}^n$, and the collection $\mathcal{C}$ can be identified with a binary matrix $\mathcal{Z}$ of size $n \times \#\mathcal{Z}$. Define $d(\mathbf{1}_{C_1}, \mathbf{1}_{C_2}) = n^{-1} \sum_{i=1}^n |\mathbf{1}_{C_1} - \mathbf{1}_{C_2}|$. Then, recalling that Q places probability $1/n$ on each x_i, $Q(C_1 \triangle C_2) = d(\mathbf{1}_{C_1}, \mathbf{1}_{C_2})$, so that $D(\varepsilon, \mathcal{C}, L_1(Q)) = D(\varepsilon, \mathcal{Z}, d)$. For simplicity of notation, assume that $\mathcal{Z}$ is ε-separated with respect to d, so the goal is to bound its size $\#\mathcal{Z}$ in terms of the VC dimension $V(\mathcal{C})$.

Denote $S = V(\mathcal{C}) - 1$ and fix an integer m such that $S \leqslant m < n$. For a subset $J \subset \{1, \ldots, n\}$ of size $\#J = m$, let $\mathcal{Z}_J$ denote the projection of $\mathcal{Z}$ onto $\{0,1\}^J$, and $\overline{\#\mathcal{Z}_J}$ denote the average size of $\mathcal{Z}_J$ over all subsets J or size m. Then, following the proof on Page 138 of van der Vaart and Wellner (1996) we arrive to the bound

$$\#\mathcal{Z} \leqslant \frac{\overline{\#\mathcal{Z}_J} n\varepsilon(m+1)}{\varepsilon n(m+1) - 2(n-m)S} \leqslant \frac{\varepsilon(m+1)\overline{\#\mathcal{Z}_J}}{\varepsilon(m+1) - 2S} \leqslant \frac{\varepsilon m \overline{\#\mathcal{Z}_J}}{\varepsilon m - 2S},$$

which holds without any extra constants. The number of points in any $\mathcal{Z}_J$ is equal to the number of subsets picked out by $\mathcal{C}$ from the points $\{x_i : i \in J\}$. By the Sauer-Shelah Lemma, this is bounded by $\sum_{j=0}^S \binom{m}{j}$, which is smaller than $(em/S)^S$ for $m \geqslant S$.[6] Therefore,

$$\#\mathcal{Z} \leqslant \Big(\frac{e}{S}\Big)^S \frac{m^{S+1}\varepsilon}{m\varepsilon - 2S}$$

holds for all integers m such that $S \leqslant m < n$. Denote the right-hand side of the preceding display by $f(m)$. This function is strictly decreasing until $m^* = 2(S+1)/\varepsilon$ and strictly increasing afterwards. Therefore, the optimal unconstrained choice is $m = m^*$, for which $f(m^*) = (2e/\varepsilon)^S (S+1)(1 + S^{-1})^S$. However, the argument leading to the upper bound on $\#\mathcal{Z}$ only applies to integer m such that $S \leqslant m < n$. To ensure that a similar bound holds for an integer value of m, we can simply use $f(m^* - 1)$ since somewhere between $m^* - 1$ and

[6] Indeed, for $t \in (0,1)$, $\sum_{j=0}^S \binom{m}{j} \leqslant \sum_{j=0}^S \binom{m}{j} \frac{t^j}{t^S} \leqslant \frac{(1+t)^m}{t^s}$. Set $t = \frac{S}{m}$ and use $(1 + S/m)^m \leqslant e^S$.

m^* there must be an integer, and $f(m)$ is decreasing on this interval. We have

$$\begin{aligned} f(m^*-1) &= \left(\tfrac{e}{S}\right)^S \frac{(2(S+1)/\varepsilon-1)^{S+1}\varepsilon}{(2(S+1)/\varepsilon-1)\varepsilon-2S} \\ &= \left(\tfrac{2e}{\varepsilon}\right)^S \tfrac{1}{1-\varepsilon/2}(S+1-\varepsilon/2)\left(1+\tfrac{1-\varepsilon/2}{S}\right)^S \\ &\leqslant \left(\tfrac{2e}{\varepsilon}\right)^S (S+1)^{\frac{1}{1-\varepsilon/2}} \exp(1-\varepsilon/2) \\ &\leqslant \left(\tfrac{2e}{\varepsilon}\right)^S (S+1)\cdot 2\sqrt{e}, \end{aligned}$$

for all $\varepsilon \in (0,1)$ since the function $g(\varepsilon) = (1-\varepsilon/2)^{-1}\exp(1-\varepsilon/2)$ is monotonically increasing. Therefore, we obtain the bound

$$\#\mathcal{Z} \leqslant \left(\frac{2e}{\varepsilon}\right)^S (S+1)\cdot 2\sqrt{e},$$

and it remains to check that this bound still holds when $m^*-1 < S$ or $m^* \geqslant n$. Note that $m^*-1 \geqslant S$ for all $\varepsilon \in (0,1)$. If $m^* \geqslant n$, by the Sauer-Shelah Lemma

$$\#\mathcal{Z} \leqslant \sum_{j=0}^{S}\binom{n}{j} \leqslant \left(\frac{en}{S}\right)^S \leqslant \left(\frac{em^*}{S}\right)^S \leqslant e\left(\frac{2e}{\varepsilon}\right)^S,$$

which certainly implies the bound in the previous display. Therefore, recalling that $\#\mathcal{Z} = D(\varepsilon, \mathcal{C}, L_1(Q))$,

$$\begin{aligned} N(\varepsilon, \mathcal{C}, L_1(Q)) &\leqslant D(\varepsilon/2, \mathcal{C}, L_1(Q)) \\ &\leqslant \left(\tfrac{4e}{\varepsilon}\right)^S (S+1)\cdot 2\sqrt{e} \\ &= \left(\tfrac{4e}{\varepsilon}\right)^{V(\mathcal{C})-1} V(\mathcal{C})\cdot 2\sqrt{e} \\ &= \tfrac{1}{2\sqrt{e}} V(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\tfrac{1}{\varepsilon}\right)^{(V(\mathcal{C})-1)}, \end{aligned}$$

and the desired result follows.

■

Next, we state and prove two simple lemmas about a specific VC-subgraph class of functions. A subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$C_f = \{(t,x) \in \mathbb{R}\times\mathcal{X} : t < f(x)\}.$$

A class of functions $\mathcal{F}$ is VC-subgraph if the class of all subgraphs

$$\mathcal{C}_{\mathcal{F}} = \{C_f : f \in \mathcal{F}\}$$

has a finite VC dimension. In this case we denote $V(\mathcal{F}) = V(\mathcal{C}_{\mathcal{F}})$.

The next result is Theorem 2.6.7. from van der Vaart and Wellner (1996). It is a direct consequence of the result for sets and holds with the same universal constant.

Lemma 3.7.4 (Covering Number for VC-subgraph Classes)**.** *For a VC-class of functions with measurable envelope function F and $r \geqslant 1$, one has for any probability measure Q with $||F||_{Q,r} > 0$,*

$$N(\varepsilon ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)},$$

for $0 < \varepsilon < 1$.

Next, we refine the above result for a particular VC-subgraph class of functions.

Lemma 3.7.5 (A Simple VC-Subgraph Class)**.** *Let $\mathcal{G}$ denote a class of subsets of $\mathcal{X}$ with a finite VC dimension $V(\mathcal{G})$, and $F : \mathcal{X} \to \mathbb{R}$ be an arbitrary function. Define a class of functions:*

$$\mathcal{F} = \{\mathbf{1}(x \in G)F(x) : G \in \mathcal{G}\}.$$

Then, $\mathcal{F}$ is VC-subgraph with $V(\mathcal{F}) \leqslant V(\mathcal{G})$.

Proof. Let $VC(\mathcal{G}) = d$ and $D = \{(t_1, x_1), \ldots, (t_d, x_{d+1})\} \subset \mathbb{R} \times \mathcal{X}$ be an arbitrary set of points. By definition, D is shattered by $\mathcal{F}$ if for every subset $\{(t_j, x_j) : j \in J\}$ there is a function f with subgraph C_f such that $C_f \cap D = \{(t_j, x_j) : j \in J\}$. Equivalently, D is shattered by $\mathcal{F}$ if for every subset $J \subset \{1, \ldots, d+1\}$ there is a set $G \in \mathcal{G}$ satisfying

$$\begin{aligned} t_j &< \mathbf{1}(x_j \in G)F(x_j) \text{ for } j \in J \\ t_k &\geqslant \mathbf{1}(x_k \in G)F(x_k) \text{ for } k \notin J \end{aligned} \tag{3.3}$$

We will argue that D cannot be shattered by $\mathcal{F}$.

First, if there is (t_j, x_j) such that $t_j < 0$ and $t_j < F(x_j)$, then $t_j < \mathbf{1}(x_j \in G)F(x_j)$ holds for all $G \in \mathcal{G}$. In this case, any subset of D that does not include t_j, x_j cannot be picked out, so D cannot be shattered by $\mathcal{F}$. Similarly, if there is (t_k, x_k) such that $t_k \geqslant 0$ and $t_k \geqslant F(x_k)$, then $t_k \geqslant \mathbf{1}(x_k \in G)F(x_k)$ holds for all $G \in \mathcal{G}$. So, any subset of D that includes this point cannot be picked out and D cannot be shattered by $\mathcal{F}$. Therefore, we will assume that each (t_j, x_j) satisfies either $t_j < 0, F(x_j) \geqslant 0$ or $t_j \geqslant 0, F(x_j) < 0$ for $j = 1, \ldots, d+1$.

Recall that $\mathcal{G}$ does not shatter $\{x_1, \ldots, x_{d+1}\}$, meaning that there exist a subset $\{x_j\}_{j \in J}$ that $\mathcal{G}$ cannot pick out. Then, for every $G \in \mathcal{G}$ we have either $x_j \notin G$ for some $j \in J$ or $x_k \in G$ for some $k \notin J$. If the inequalities in (3.3) do not hold for this J for any G, then $\{(t_j, x_j)\}_{j \in J}$ cannot be picked out and D cannot be shattered by $\mathcal{F}$. Suppose the inequalities in (3.3) hold for some $G \in \mathcal{G}$. If $x_j \notin G$ for some $j \in J$, it must be that $t_j < 0$ and, according to the previous discussion, $F(x_j) \geqslant 0$. Then the set $J' = J \backslash (t_j, x_j)$ cannot be picked out. If $x_k \in G$ for some $k \notin J$, it must be that $t_k \geqslant 0$ and $F(x_k) < 0$, so the set $J'' = J \cup k$ cannot be picked out. Therefore, D cannot be shattered by $\mathcal{F}$ and $VC(\mathcal{F}) \leqslant VC(\mathcal{G})$.

■

Lemma 3.7.6 (Covering Numbers for Special VC-Subgraph Classes)**.** *Let $\mathcal{G}$ denote a class of subsets of $\mathcal{X}$ with a finite VC dimension $V(\mathcal{G})$, and $F : \mathcal{X} \to \mathbb{R}$ be an arbitrary function. Define a class of functions:*

$$\mathcal{F} = \{\mathbf{1}(x \in G)F(x) : G \in \mathcal{G}\}.$$

Then, for any $r \geqslant 1$, probability measure Q with $||F||_{Q,r} > 0$, and $0 < \varepsilon < 1$,

$$N(\varepsilon \, ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Proof. By Lemma 3.7.5, $\mathcal{F}$ is VC-subgraph. For $r = 1$, note that:

$$||f_1 - f_2||_{Q,1} = \mathbb{E}_Q[|\mathbf{1}_{G_1} - \mathbf{1}_{G_2}||F|] = P(C_{f_1} \triangle C_{f_2}) \, ||F||_{Q,1},$$

where $P = \lambda \times Q/\left|\left|F\right|\right|_{Q,1}$ is a probability measure on $\mathbb{R} \times \mathcal{X}$ and λ is a Lebesgue measure on $\mathbb{R}$. Then, by Lemma 3.7.3,

$$N(\varepsilon \left|\left|F\right|\right|_{Q,1}, \mathcal{F}, L_1(Q)) = N(\varepsilon, \mathcal{C}_{\mathcal{F}}, L_1(P)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{(V(\mathcal{F})-1)}.$$

For $r > 1$, note that:

$$\left|\left|f_1 - f_2\right|\right|_{Q,r}^r = \mathbb{E}_Q(|\mathbf{1}_{G_1} F - \mathbf{1}_{G_2} F||F|^{r-1}) = \frac{\left|\left|f_1 - f_2\right|\right|_{R,1}}{\left|\left|F\right|\right|_{R,1}} \mathbb{E}_Q(|F|^r),$$

for the probability measure R with density $|F|^{r-1}/\mathbb{E}_Q(|F|^{r-1})$ with respect to Q. Therefore,

$$\left|\left|f_1 - f_2\right|\right|_{Q,r} = \left(\frac{\left|\left|f_1 - f_2\right|\right|_{R,1}}{\left|\left|F\right|\right|_{R,1}}\right)^{1/r} \left|\left|F\right|\right|_{Q,r},$$

so that by the previous argument applied to R instead of Q

$$N(\varepsilon \left|\left|F\right|\right|_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant N(\varepsilon^r \left|\left|F\right|\right|_{R,1}, \mathcal{F}, L_1(R)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}$$

■

3.7.2 Auxiliary Lemmas

Now we are ready to state and prove three auxiliary lemmas that give our main results.

Lemma 3.7.7 (Finite-Sample Bound on Rademacher Complexity)**.** *Let $W_1, \ldots, W_n$ be an i.i.d. sample and $\xi_1, \ldots, \xi_n$ be i.i.d. Rademacher random variables independent of $W_1, \ldots, W_n$.*

1. *Let $\mathcal{F}$ be a VC-subgraph of functions with $f_0(w) = 0 \in \mathcal{F}$, a finite VC dimension $VC(\mathcal{F})$, and a measurable envelope F such that $S = \mathbb{E}(F^2) < \infty$. Then:*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n} \sum_{i=1}^n \xi_i f(W_i)\right|\right] \leqslant C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

   where $C = 4\sqrt{12} \int_0^1 \sqrt{1/(2e^{3/2}) + \log(16e) + 2\log(1/u)} du \leqslant 34$.

2. *In the special case when $\mathcal{F} = \{f(x) = 1(x \in G)F(x) : G \in \mathcal{G}\}$, for a VC-class of sets $\mathcal{G}$ and an arbitrary measurable function F with $S = \mathbb{E}(F^2) < \infty$, the above holds with $C = 4\sqrt{12} \int_0^1 \sqrt{1/(2e^{3/2}) + \log(4e) + 2\log(1/u)} du \leqslant 29$.*

Proof. Denote $\mathbb{G}_n^0(f) = n^{-1/2}\sum_{i=1}^n \xi_i f(W_i)$. By the Law of Iterated Expectations,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\mathbb{G}^0(f)\right|\right] = \frac{1}{\sqrt{n}}\mathbb{E}_{W_1^n}\left[\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{G}^0(f)\right|\right]\right] \tag{3.4}$$

We will use a simple chaining argument to bound the right hand side of (3.4). Let $\eta = 2\,||F||_{2,n}$, and define $\mathcal{F}_0 = \{f_0\}$ and $\mathcal{F}_j$ contain centers of the balls in the minimal $\eta 2^{-j}$-cover of $\mathcal{F}$ under $||\cdot||_{2,n}$, so that $|\mathcal{F}_j| = N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n})$. Let $\phi_j : \mathcal{F} \to \mathcal{F}_j$ be a map that for a given f finds the closest element of $\mathcal{F}_j$. For any $f_k \in \mathcal{F}_k$ define a chain $f_{k-l} = \phi_{k-l}(f_{k-l+1})$ for $l = 1, \ldots, k$. Then,

$$\mathbb{G}_n^0(f_k) = \sum_{j=1}^k (\mathbb{G}_n^0(f_j) - \mathbb{G}_n^0(f_{j-1})) \leqslant \sum_{j=1}^k \max_{g\in\mathcal{F}_j} |\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))|,$$

Let $\psi_2(x) = e^{x^2} - 1$ and $||\cdot||_{\psi_2}$ denote the corresponding Orlisz norm. By Lemma 2.2.7. in van der Vaart and Wellner (1996), conditional on W_1^n, the process $\mathbb{G}_n^0(f)$ is sub-Gaussian for the metric $d_n(f_1, f_2) = ||f_1 - f_2||_{2,n}$, and satisfies $||\mathbb{G}_n^0(f) - \mathbb{G}_n^0(g)||_{\psi_2} \leqslant \sqrt{6}\,||f - g||_{2,n}$. By Lemma 3.7.2 and the above discussion,

$$\begin{aligned}\mathbb{E}_{\xi_1^n}\left[\max_{g\in\mathcal{F}_j}|\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))|\right] &\leqslant \psi_2^{-1}(|\mathcal{F}_j|)\max_{g\in\mathcal{F}_j}||\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))||_{\psi_2} \\ &\leqslant \sqrt{6}\cdot\psi_2^{-1}(N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n}))\cdot\eta 2^{-(j-1)}\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}_k}|\mathbb{G}_n^0(f)|\right] &\leqslant \sqrt{6}\sum_{j=1}^k \psi_2^{-1}(N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n}))\eta 2^{-(j-1)} \\ &\overset{(a)}{\leqslant} 4\sqrt{6}\int_0^{\eta/2}\psi^{-1}(N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n}))d\varepsilon \\ &= 4\sqrt{6}\int_0^{||F||_{2,n}}\sqrt{\log(N(\varepsilon, \mathcal{F}, ||F||_{2,n}) + 1)}d\varepsilon \\ &\overset{(b)}{\leqslant} 4\sqrt{12}\int_0^{||\cdot||_{2,n}}\sqrt{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n})}d\varepsilon,\end{aligned}$$

where (a) follows from rearranging rectangles under the curve $\varepsilon \mapsto \psi_2^{-1}(N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n}))$, and (b) follows from $\log(x+1) \leqslant 2\log(x)$ for $x \geqslant 2$. Since, conditional on W_1^n, the process $\mathbb{G}_n^0$ is separable, by letting $k \to \infty$ in the previous display we conclude that

$$\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n^0(f)|\right] \leqslant 4\sqrt{12}\int_0^{||F||_{2,n}} \sqrt{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n})}d\varepsilon. \tag{3.5}$$

Denote $V \equiv VC(\mathcal{F})$ and $K = (2\sqrt{e})^{-1}$. Applying Lemma 3.7.4 (or Lemma 3.7.6 for the special case) with $r = 2$ and $Q = P_n$,

$$\begin{aligned}
\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n}) &\leqslant \log(KV) + V\log(16e) + 2(V-1)\log\left(\tfrac{||F||_{2,n}}{\varepsilon}\right) \\
&= V\left(K\tfrac{\log(KV)}{KV} + \log(16e) + 2\tfrac{V-1}{V}\log\left(\tfrac{||F||_{2,n}}{\varepsilon}\right)\right) \\
&\leqslant V\left(K/e + \log(16e) + 2\log\left(\tfrac{||F||_{2,n}}{\varepsilon}\right)\right),
\end{aligned}$$

where the last line uses the fact that $\log(t)/t \leqslant 1/e$ for all $t > 0$. Therefore,

$$\begin{aligned}
\int_0^{||F||_{2,n}} \sqrt{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n})}d\varepsilon &\leqslant \int_0^{||F||_{2,n}} \sqrt{K/e + \log(16e) + 2\log\left(||F||_{2,n}/\varepsilon\right)}d\varepsilon \cdot \sqrt{V} \\
&\leqslant \int_0^1 \sqrt{K/e + \log(16e) + 2\log(1/u)}du\sqrt{V\,||F||_{2,n}^2},
\end{aligned} \tag{3.6}$$

where the second line follows from a change of variables $u = \varepsilon/\,||F||_{2,n}$. Combining (3.5) and (3.6), we obtain

$$\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n^0(f)|\right] \leqslant C\sqrt{V\,||F||_{2,n}^2}$$

where $C = 4\sqrt{12}\int_0^1 \sqrt{K/e + \log(16e) + 2\log(1/u)}du$ (or the same expression with $4e$ instead of $16e$ in the special case). By (3.4) and Jensen's inequality,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(W_i)\right|\right] \leqslant C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

which concludes the proof.

■

3.7.3 Proofs of Theorems 3.3.1, 3.3.2, and 3.3.3

3.7.3.1 Proof of Theorem 3.3.1

To keep notation simple, we write $\hat{\pi}_n$ instead of $\hat{\pi}_n^{EWM}$. Let π^* denote a rule such that $V(\pi^*) = V_\Pi^* = \sup_{\pi\in\Pi} V(\pi)$. Note that

$$\begin{aligned} R(\hat{\pi}_n) &= V(\pi^*) - V(\hat{\pi}_n) \\ &= V(\pi^*) - \hat{V}_n(\hat{\pi}_n) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) \\ &\leqslant V(\pi^*) - \hat{V}_n(\pi^*) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n), \end{aligned}$$

and, therefore,

$$\mathbb{E}[R(\hat{\pi}_n)] = \mathbb{E}[\hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n)] \leqslant \mathbb{E}[\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - V(\pi)|].$$

Define a class of functions

$$\mathcal{F} = \left\{ f(w) = \pi(x)\left(\frac{yt}{e(x)} - \frac{y(1-t)}{1-e(x)}\right) : \pi \in \Pi \right\},$$

so that

$$\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - V(\pi)| = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(W_i) - \mathbb{E}[f(W_i)] \right|.$$

Applying Lemma 3.7.1 and part 2 of Lemma 3.7.7,

$$\mathbb{E}\left[\left| \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(W_i) - \mathbb{E}[f(W_i)] \right| \right] \leqslant 2C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

where $C \leqslant 29$ is a universal constant and $S = \mathbb{E}[f(W)^2]$. By Lemma 3.7.5, $VC(\mathcal{F}) \leqslant VC(\Pi)$, and for any $P \in \mathcal{P}_{B,\eta}$,

$$\mathbb{E}_P[f(W)^2] \leqslant \mathbb{E}_P\left[\frac{Y^2 T}{e(X)^2} + \frac{Y^2(1-T)}{(1-e(X))^2} \right] \leqslant \frac{B^2}{\eta^2},$$

so the desired result follows.

3.7.3.2 Proof of Theorem 3.3.3

To keep notation simple, we write $\hat{\pi}_n = \hat{\pi}_{n,\hat{k}}$ instead of $\hat{\pi}_n^{PWM}$, and $\hat{V}_n$ instead of $\hat{V}_n^E$. The subscripts n, l, and r, indicate the the corresponding objects depend on the entire sample, only the estimating sample, and only the test sample correspondingly. For example, while $\hat{\pi}_{l,k}$ depends only on the estimating sample, $\hat{\pi}_{n,\hat{k}}$ depends on the entire sample by the choice of $\hat{k}$. Let π_k^* denote a rule such that $V(\pi_k^*) = V_{\Pi_k}^* = \max_{\pi \in \Pi_k} V(\pi)$. Recall that, by definition,

$$Q_{n,k}(\hat{\pi}_{n,\hat{k}}) = \hat{V}_l(\hat{\pi}_{n,\hat{k}}) - \hat{C}_{n,\hat{k}} = \hat{V}_r(\hat{\pi}_{n,\hat{k}}).$$

Write

$$\begin{aligned} R(\hat{\pi}_n) &= V_{\Pi}^* - V_{\Pi_k}^* \\ &\quad + V(\pi_k^*) - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) \\ &\quad + Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}}) \end{aligned}$$

By the definitions of $\hat{k}$ and $\hat{\pi}_{l,k}$, for any k,

$$V(\pi_k^*) - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) \leqslant V(\pi_k^*) - Q_{n,k}(\hat{\pi}_{l,k}) \leqslant V(\pi_k^*) - \hat{V}_l(\pi_k^*) + \hat{C}_{n,k},$$

so that

$$\mathbb{E}[V(\hat{\pi}_k^*) - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}})] \leqslant \mathbb{E}[\hat{C}_{n,k}].$$

Next, write

$$\mathbb{E}[\hat{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})] \leqslant r^{-1/2}\mathbb{E}\left[\max_{k \leqslant K} \sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})|\right], \tag{3.7}$$

and, working conditional on the estimating sample W_1^l,

$$\begin{aligned} &\mathbb{E}\left[\max_{k \leqslant K} \sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})| \;\middle|\; W_1^l\right] \\ &\qquad\qquad \leqslant K \max_{k \leqslant K} \mathbb{E}\left[\sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})| \;\middle|\; W_1^l\right]. \end{aligned} \tag{3.8}$$

Denoting $f_k(w) = \hat{\pi}_{m,k}(x)(yt/e(x) - y(1-t)/(1-e(x)))$, we have:

$$\begin{aligned}
\mathbb{E}\left[\sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})| \;\middle|\; W_1^l\right] &= \mathbb{E}\left[\left|r^{-1/2}\textstyle\sum_j f_k(W_j) - \mathbb{E}[f_k(W_j)]\right| \;\middle|\; W_1^l\right] \\
&\leqslant \mathbb{E}\left[\left(r^{-1/2}\textstyle\sum_j f_k(W_j) - \mathbb{E}[f_k(W_j)]\right)^2 \;\middle|\; W_1^l\right]^{1/2} \\
&\leqslant \mathbb{E}[f_k(W_j)^2 \mid W_1^l]^{1/2} \\
&\leqslant \tfrac{B}{\eta},
\end{aligned}$$

where the last inequality follows in the same fashion as in Theorem 3.3.1. Since this bound does not depend on k, taking expectations on both sides of (3.8) and recalling that $r = ln$, we obtain:

$$\mathbb{E}[\hat{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})] \leqslant \frac{B}{\eta}\frac{K}{\sqrt{ln}}.$$

Combining the above results, we conclude that

$$\mathbb{E}[R(\hat{\pi}_n)] \leqslant V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}] + \frac{B}{\eta}\frac{K}{\sqrt{ln}}, \tag{3.9}$$

holds for all $k \leqslant K$, so that

$$\mathbb{E}[R(\hat{\pi}_n)] \leqslant \inf_{k \leqslant K}\{V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}]\} + \frac{B}{\eta}\frac{K}{\sqrt{ln}},$$

and the first part of the statement follows.

For the second part of the statement, note that by the Law of Iterated Expectations

$$\begin{aligned}
\mathbb{E}[\hat{C}_{n,k}] &= \mathbb{E}[\hat{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k}) + V(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})] \\
&= \mathbb{E}[\hat{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})].
\end{aligned}$$

Then, repeating the proof of Theorem 3.3.1 with Π_k instead of Π and m instead of n, we obtain

$$\mathbb{E}[\hat{C}_{n,k}] \leqslant C\frac{B}{\eta}\sqrt{\frac{VC(\Pi_k)}{(1-s)n}}.$$

Plugging this in Equation (3.9) and recalling that $V_\Pi^* = V_{\Pi_k}^*$ for all $P \in \mathcal{P}_{B,\eta}^k$, we conclude that

$$\sup_{P \in \mathcal{P}_{B,\eta}^k} \mathbb{E}_P[R(\hat{\pi}_n^{PWM})] \leqslant \frac{B}{\eta}\left(C\sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K\sqrt{\frac{1}{ln}}\right),$$

and the proof is complete.

3.7.3.3 Proof of Theorem 3.3.2

We consider a particular subclass of $\mathcal{P}_{B,\eta}$ for which the worst-case regret can be bounded from below by a term proportional to $B/\eta\sqrt{d/n}$. The construction proceeds as follows. Let $x_1, \ldots, x_d$, where $d = VC(\Pi) - 1$, be a set shattered by Π with the largest possible cardinality. Let

$$X \in \{x_1, \ldots, x_d\}, \;\; P(X = x_j) = \tfrac{1}{d};$$
$$T \in \{0, 1\}, \;\; P(T = 1) = p, \;\; T \perp (X, Y_0, Y_1);$$
$$Y_0 = 0,$$

and, given a parameter vector $c = (c_1, \ldots, c_d) \in \{-1, 1\}^d$,

$$Y_1 | X = x_j = \begin{cases} A & \text{w.p. } \frac{1}{2}(1 + c_j \frac{\gamma}{A}) \\ -A & \text{w.p. } \frac{1}{2}(1 - c_j \frac{\gamma}{A}) \end{cases},$$

where $\gamma/A \leqslant 1$. Then, for $Y = TY_1 + (1 - T)Y_0$,

$$\mathbb{E}(Y^2) = pA^2,$$
$$\tau(x_j) = \mathbb{E}[Y_1 - Y_0 | X = x_j] = \gamma c_j.$$

For every $c \in \{-1, 1\}^d$, the joint distribution of $W = (Y, X, T)$ constructed above belongs to $P_{B,\eta}$ as long as $p \in [\eta, 1 - \eta]$ and $pA^2 \leqslant B^2$. We will specify such p and A later.

Let $C = (C_1, \ldots, C_d)$ consist of i.i.d. random variables $C_j \in \{-1, 1\}$ such that $P(C_j = 1) = 1/2$. The joint distribution of $W = (Y, X, T)$ given $C = c$ is

$$P(Y = y, X = x_j, T = t | C = c) = \begin{cases} (1 - p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + c_j \frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - c_j \frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases}.$$

We shall also derive the posterior probability $P(C_j = 1 | W_1^n)$ which will play a crucial role in deriving the lower bound.

We have

$$P(Y=y, X=x_j, T=t) = \begin{cases} (1-p)\frac{1}{d} & y=0, t=0 \\ \frac{1}{2}\frac{p}{d} & y=A, t=1 \\ \frac{1}{2}\frac{p}{d} & y=-A, t=1 \end{cases},$$

and

$$\begin{aligned} P(Y=y, X=x_k, T=t|C_j=1) &= \mathbf{1}(k\neq j)P(Y=y, X=x_j, T=t) \\ &+ \mathbf{1}(k=j)\begin{cases} (1-p)\frac{1}{d} & y=0, t=0 \\ \frac{1}{2}(1+\frac{\gamma}{A})\frac{p}{d} & y=A, t=1 \\ \frac{1}{2}(1-\frac{\gamma}{A})\frac{p}{d} & y=-A, t=1 \end{cases}. \end{aligned}$$

Therefore,

$$\frac{P(W_i|C_j=1)}{P(W_i)} = \mathbf{1}(X_i\neq x_j) + \mathbf{1}(X_i=x_j)\begin{cases} 1 & Y_i=0, T_i=0 \\ 1+\frac{\gamma}{A} & Y_i=A, T_i=1 \\ 1-\frac{\gamma}{A} & Y_i=-A, T_i=1 \end{cases},$$

and

$$P(C_j=1|W_1^n) = \frac{P(W_1^n|C_j=1)P(C_j=1)}{P(W_1^n)} = \frac{1}{2}\left(1+\frac{\gamma}{A}\right)^{N_j^+}\left(1-\frac{\gamma}{A}\right)^{N_j^-}, \tag{3.10}$$

where

$$N_j^+ = \#\{i : X_i = x_j, Y_i = A, T_i = 1\}$$
$$N_j^- = \#\{i : X_i = x_j, Y_i = -A, T_i = 1\},$$

so that a tuple $(N_j^+, N_j^-, n-N_j^+-N_j^-)$ has a multinomial distribution:

$$\begin{aligned} &P(N_j^+=k_1, N_j^-=k_2|C_j=1) \\ &\quad = \binom{n}{k_1}\binom{n-k_1}{k_2}\left(\frac{1}{2}(1+\frac{\gamma}{B})\frac{p}{d}\right)^{k_1}\left(\frac{1}{2}(1-\frac{\gamma}{B})\frac{p}{d}\right)^{k_2}\left(1-\frac{p}{d}\right)^{n-k_1-k_2}. \end{aligned} \tag{3.11}$$

Now we turn to the main part of the proof. Let $\mathcal{P}_{\mathcal{C}} = \{P_{W|C=c} : c \in \{-1,1\}^d\} \subset \mathcal{P}_{B,\eta}$ denote the set of distributions of $W=(Y,X,T)$ constructed above, and μ denote the

distribution of C. Let π_P^* denote the first-best treatment rule when the distribution of the data is P, and write $\pi_c^* = \pi^*_{P_{W|C=c}}$ for brevity. By construction, $\pi_c^*(x_j) = \mathbf{1}(c_j = 1)$, and $\pi_c^* \in \Pi$ since the class Π shatters $\{x_1, \ldots, x_d\}$. Note that:

$$V(\pi_c^*) - V(\hat{\pi}_n) = \frac{\gamma}{d} \sum_{j=1}^{d} c_j(\pi_c^*(x_j) - \hat{\pi}_n(x_j)) = \frac{\gamma}{d} \sum_{j=1}^{d} \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)).$$

Then,

$$\begin{aligned}
\sup_{P \in \mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \max_{P \in \mathcal{P}_C} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n))] \\
&\geqslant \int \mathbb{E}_{P_{W_1^n|C=c}}[V(\pi_c^*) - V(\hat{\pi}_n))] d\mu(c) \\
&= \frac{\gamma}{d} \sum_{j=1}^{d} \int \int \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)) dP_{W_1^n|C=c} d\mu(c) \qquad (3.12) \\
&= \frac{\gamma}{d} \sum_{j=1}^{d} P_{W_1^n, C_j}(\mathbf{1}(C_j = 1) \neq \hat{\pi}_n(x_j)) \\
&\geqslant \gamma \cdot \inf_{\pi} P_{W_1^n, C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n)).
\end{aligned}$$

Note that $P_{W_1^n, C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n))$ is the probability of misclassification of $\mathbf{1}(C_j = 1)$ using W_1^n. By Theorem 2.1. in Devroye and Lugosi (1996), the infimum is attained by the Bayes Classifier, $\pi^*(W_1^n) = \mathbf{1}(P(C_j = 1|W_1^n) > 0.5)$, and is equal to

$$\begin{aligned}
P(\mathbf{1}(C_j = 1) \neq \pi^*(W_1^n)) &= \tfrac{1}{2} P(\, P(C_j = 1|W_1^n) \leqslant 0.5 \,|C_j = 1) \\
&\quad + \tfrac{1}{2} P(\, P(C_j = 1|W_1^n) > 0.5 \,|C_j = -1).
\end{aligned}$$

Denote $a = \gamma/A$, and work conditional on $C_j = 1$ from now on. Recalling (3.18),

$$\begin{aligned}
P(P(C_j = 1|W_1^n) \leqslant 0.5) &= P((1+a)^{N_j^+}(1-a)^{N_j^-} \leqslant 1) \\
&\geqslant P((1-a^2)^{N_j^+} \leqslant 1 | N_j^+ \leqslant N_j^-) \cdot P(N_j^+ \leqslant N_j^-) \\
&= P(N_j^+ \leqslant N_j^-).
\end{aligned}$$

Let $D_i^+ = \mathbf{1}(X_i = x_j, Y_i = A, T_i = 1)$ and $D_i^- = \mathbf{1}(X_i = x_j, Y_i = -A, T_i = 1)$. Then, $\mathbb{E}[D_i^+ - D_i^-] = ap/d$, $\mathbb{V}ar[D_i^+ - D_i^-] = p/d - (ap/d)^2$, and $\mathbb{E}[(D_i^+ - D_i^-)^3] = p/d$. Letting Z_n denote the studentized version of $n^{-1}\sum_{i=1}^n (D_i^+ - D_i^-)$ and Φ denote the Standard Normal

CDF, using Berry-Esseen inequality we obtain

$$\begin{aligned} P(N_j^+ \leqslant N_j^-) &= P(\tfrac{1}{n}\textstyle\sum_{i=1}^n (D_i^+ - D_i^-) \leqslant 0) \\ &= P\left(Z_n \leqslant \tfrac{-\sqrt{n}ap/d}{\sqrt{p/d-(ap/d)^2}}\right) \\ &\geqslant \Phi\left(\tfrac{-\sqrt{n}ap/d}{\sqrt{p/d-(ap/d)^2}}\right) - \tfrac{K}{\sqrt{n}}\tfrac{1}{(p/d)^{1/2}(1-a^2p/d)^{3/2}}, \end{aligned}$$

where $K < 0.469$ (Shevtsova, 2013). Choosing $a = \gamma/A \equiv c/\sqrt{n}\sqrt{d/p}$ for some $c \in (0,1)$, assuming n is large enough to satisfy $\gamma/A \leqslant 1$, we obtain

$$P(N_j^+ \leqslant N_j^-) \geqslant \Phi\left(-\frac{c}{\sqrt{1-c^2/n}}\right) - \frac{K}{\sqrt{n}}\frac{1}{\sqrt{p/d}(1-c^2/n)^{3/2}}.$$

Choosing $p = \eta$, $A = B/\sqrt{\eta}$ so that $\gamma = c \cdot B/\eta\sqrt{d/n}$, we have, for $n \geqslant 3$,

$$\begin{aligned} \sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \tfrac{\gamma}{2} \cdot P(N_j^+ \leqslant N_j^- | C_j = 1) \\ &\geqslant \tfrac{1}{2}\tfrac{B}{\eta}\sqrt{\tfrac{d}{n}} \cdot c \cdot \Phi\left(-\tfrac{c}{\sqrt{1-c^2}}\right) - \tfrac{K}{2\sqrt{\eta}} \cdot \tfrac{B}{\eta}\tfrac{d}{n}\tfrac{c}{(1-c^2/3)^{3/2}} \end{aligned}$$

Choosing $c = 0.5162$, and plugging in $K = 0.469$ gives the final result

$$\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant 0.07 \cdot \frac{B}{\eta}\sqrt{\frac{d}{n}} - \frac{0.14}{\sqrt{\eta}} \cdot \frac{B}{\eta}\frac{d}{n}.$$

For $n \geqslant 4d/\eta$, the right-hand-side in the preceding display is positive, and $\gamma/A \leqslant 1$ is also satisfied.

3.7.4 Proofs of Theorems 3.4.1, 3.4.2 and 3.4.3

The proof of Theorem 3.4.1 is based on the following two lemmas. The first Lemma gives a maximal inequality in terms of the VC dimension of the class of policy rules Π, the number of observations n, and the second moment of the orthogonal score Γ.

Lemma 3.7.8 (Uniform Concentration Bound for $\tilde{V}_n$)**.** *Suppose that the class* Π *has VC-dimension* $VC(\Pi)$ *and includes the no-treatment policy* $\pi_0(x) = 0$ *for all* x*. Then,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi)|\right] \leqslant C\sqrt{\frac{VC(\Pi)S^2}{n}},$$

where $C \leqslant 58$ *is a universal constant and* $S^2 = \mathbb{E}(\Gamma_i^2)$*.*

Proof. Define a class of functions $\mathcal{F} = \{f(w) = \pi(x)\Gamma(w) : \pi \in \Pi\}$, which is a VC-subgraph class with $VC(\mathcal{F}) \leqslant VC(\Pi)$ and envelope $|\Gamma|$. Then, by Lemmas 3.7.1, 3.7.5, and the second part of Lemma 3.7.7,

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi)|\right] \leqslant 2\mathbb{E}\left[\sup_{\pi\in\Pi} \left|\frac{1}{n}\sum_{i=1}^{n} \xi_i \pi(X_i)\Gamma_i\right|\right] \leqslant 2C\sqrt{\frac{VC(\Pi)S^2}{n}}.$$

where $C \leqslant 29$ is a constant from Lemma 3.7.7. ■

The second Lemma establishes that $\hat{V}_n$ and $\tilde{V}_n$ are uniformly close in $\pi \in \Pi$. It is a finite-sample version of Lemma 4 from Athey and Wager (2021) proven under slightly weaker assumptions.

Lemma 3.7.9 (Uniform Coupling)**.** *Let assumptions 3.2.1 – 3.2.4 hold, and assume that* $\mathbb{E}((Y - m(X, D))^2|X, D) \leqslant B^2$ *almost surely. Suppose that* $\hat{\Gamma}_i$ *are computed using a* J*-fold sample splitting. Then,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|\right] \leqslant R_{1,n} + R_{2,n} + R_{3,n},$$

where $C \leqslant 58$ *is a universal constant, and*

$$R_{1,n} = C\sqrt{(J+2)\cdot B^2 \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}}$$

$$R_{2,n} = C\sqrt{(J+2)\cdot \frac{2(\eta^2+1)}{\eta^2} \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}},$$

and

$$R_{3,n} = \sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}}.$$

Proof. Let $\hat{m}^{(-j)}$, $\tau_{\hat{m}^{(-j)}}$ and $\hat{g}^{(-j)}$ denote the estimators computed on observations excluding j-th fold. Denote the indices of the observations included in j-th fold by I_j. For an observation $i \in I_j$, write the difference $\hat{\Gamma}_i - \Gamma_i$ as a sum of three terms

$$\begin{aligned}
\hat{\Gamma}_i - \Gamma_i &= (Y_i - m(X_i, T_i))(\hat{g}^{(-j)}(X_i, T_i) - g(X_i, T_i)) \\
&+ \tau_{\hat{m}^{(-j)}}(X_i, T_i) - \tau_m(X_i, T_i) - g(X_i, Z_i)(\hat{m}^{(-j)}(X_i, T_i) - m(X_i, T_i)) \\
&- (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))(\hat{m}^{(-j)}(X_i, T_i) - m(X_i, T_i))
\end{aligned}$$

and denote the corresponding summands in $\hat{V}_n(\pi) - \tilde{V}_n(\pi)$ by $D_1(\pi), D_2(\pi)$, and $D_3(\pi)$. We will bound each term separately.

First Term. Write $D_1(\pi) = \sum_{j=1}^{J} D_1^{(j)}(\pi)$, where $\frac{n}{n_k} D_1^{(j)}(\pi)$ is equal to

$$\frac{1}{n_j} \sum_{i \in I_j} \pi(X_i)(Y_i - m(X_i, T_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)).$$

Note that, by the law of iterated expectations,

$$\mathbb{E}[\pi(X_i)(Y_i - m(X_i, T_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)) \mid \hat{g}^{(-j)}] = 0,$$

and denote the conditional second moment by

$$V_{1,n}(j) = \mathbb{E}\left[\pi(X_i)^2 \cdot \mathbb{E}[(Y_i - m(X_i, T_i))^2 | X_i, T_i] \cdot (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))^2 \mid g^{(-j)}\right].$$

Applying, conditional on $\hat{g}^{(-j)}$, Lemma 3.7.8 with $(Y_i - m(X_i, T_i)) \cdot (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))$ in place of Γ_i, we get:

$$\frac{n}{n_j} \mathbb{E}\left[\sup_{\pi \in \Pi} |D_1^{(j)}(\pi)| \;\middle|\; \hat{g}^{(-j)}\right] \leqslant 2C \sqrt{\frac{VC(\Pi) V_{1,n}(j)}{n_j}}$$

Using Assumption 3.2.2, $\pi(X_i)^2 \leqslant 1$, and the bound on the conditional variance of Y,

$$\mathbb{E}(V_{1,n}(j)) \leqslant B \frac{a((\frac{J-1}{J})n)}{n^{\zeta_g}}$$

By the last two displays, the law of iterated expectations, and Jensen's inequality,

$$\mathbb{E}\left[\sup_{\pi \in \Pi} |D_1^{(j)}(\pi)|\right] \leqslant 2C \sqrt{\frac{n_j}{n}} \sqrt{B \frac{VC(\Pi) a((1 - J^{-1})n)}{n^{1+\zeta_g}}}$$

Since $n_j/n \leqslant 1/(J-1)$ and supremum is sub-additive,

$$\mathbb{E}\left[\sup_{\pi \in \Pi} |D_1(\pi)|\right] \leqslant 2C \sqrt{(J+2)B \cdot \frac{VC(\Pi) a((1 - J^{-1})n)}{n^{1+\zeta_g}}}$$

Second Term. As before, write $D_2(\pi) = \sum_{j=1}^{J} D_2^{(j)}(\pi)$, where $\frac{n}{n_j} D_2^{(j)}(\pi)$ is equal to

$$\frac{1}{n_j} \sum_{i \in I_j} \pi(X_i)(\tau_{\hat{m}^{(-j)}}(X_i, T_i) - \tau_m(X_i, T_i) - g(X_i, Z_i)(\hat{m}^{(-j)}(X_i, T_i) - m(X_i, T_i)))$$

Denote the individual summands in the previous display by $f(W_i;\pi)$. Note that

$$\mathbb{E}(f(W_i;\pi)|\hat{m}^{(-j)},\tau_{\hat{m}^{(-j)}})=0$$

by part (2) of Assumption 3.2.1 and the law of iterated expectations. Denote $V_{2,n}(j)=\mathbb{E}(f(W_i;\pi)^2|\hat{m}^{(-j)},\tau_{\hat{m}^{(-j)}})$. Applying, conditional on $\hat{m}^{(-j)}$ and $\tau_{\hat{m}^{(-j)}}$, Lemma 3.7.8 with $(\tau_{\hat{m}^{(-j)}}(X_i,T_i)-\tau_m(X_i,T_i)-g(X_i,Z_i)(\hat{m}^{(-j)}(X_i,T_i)-m(X_i,T_i)))$ in place of Γ_i, we get:

$$\frac{n}{n_j}\mathbb{E}\left[\sup_{\pi\in\Pi}|D_2^{(j)}(\pi)|\;\middle|\;\hat{g}^{(-j)}\right]\leqslant 2C\sqrt{\frac{VC(\Pi)V_{2,n}(j)}{n_j}}$$

Using $(a+b)^2\leqslant 2(a^2+b^2)$, $\pi(X_i)^2\leqslant 1$, and Assumptions 3.2.2 and 3.2.4, we get:

$$\mathbb{E}(V_{2,n}(j))\leqslant 2\left(\frac{a((1-J^{-1})n)}{n^{\zeta_m}}+\frac{1}{\eta^2}\frac{a((1-J^{-1})n)}{n^{\zeta_m}}\right)=\frac{2(\eta^2+1)}{\eta^2}\frac{a((1-J^{-1})n)}{n^{\zeta_m}}.$$

By the last two displays, the law of iterated expectation, and Jensen's inequality:

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|D_2^{(j)}(\pi)|\right]\leqslant 2C\sqrt{\frac{n_j}{n}}\sqrt{\frac{2(\eta^2+1)}{\eta^2}\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}}$$

Since $n_j/n\leqslant 1/(J-1)$ and supremum is sub-additive,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|D_1(\pi)|\right]\leqslant 2C\sqrt{(J+2)\frac{2(\eta^2+1)}{\eta^2}\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}}$$

Third Term. Let $j(i)$ denote the fold in which observation i belongs. We have:

$$D_3(\pi)=-\frac{1}{n}\sum_{i=1}^n\pi(X_i)(\hat{g}^{(-j(i))}(X_i,Z_i)-g(X_i,Z_i))(\hat{m}^{(-j(i))}(X_i,T_i)-m(X_i,T_i))$$

By Cauchy-Schwartz inequality and $\pi(X_i)^2\leqslant 1$,

$$\begin{aligned}|D_3(\pi)|\;&\leqslant\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\hat{g}^{(-j(i))}(X_i,Z_i)-g(X_i,Z_i))^2}\\&\quad\times\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\hat{m}^{(-j(i))}(X_i,T_i)-m(X_i,T_i))^2},\end{aligned}$$

where we note that the right hand side does not depend on π. Taking expectations on both sides, using Cauchy-Schwartz inequality one more time, and recalling Assumption 3.2.2, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|D_3(\pi)|\right]\leqslant\sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}},$$

and the proof is complete. ■

3.7.4.1 Proof of Theorem 3.4.1

To keep the notation simple, we write $\hat{\pi}_n$ instead of $\hat{\pi}_n^{REWM}$ and write $\mathbb{E}$ instead of $\mathbb{E}_P$ for a fixed distribution $P \in \mathcal{P}_{B_\tau,B,\eta}$. Let $\pi^* \in \Pi$ be such that $V(\pi^*) = \max_{\pi\in\Pi} V(\pi)$. Note that:

$$\begin{aligned} R(\hat{\pi}_n) &= V(\pi^*) - V(\hat{\pi}_n) \\ &= V(\pi^*) - \hat{V}_n(\hat{\pi}_n) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) \\ &\leqslant V(\pi^*) - \hat{V}_n(\pi^*) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n). \end{aligned}$$

Then, writing

$$\begin{aligned} V(\pi^*) - \hat{V}_n(\pi^*) &= V(\pi^*) - \tilde{V}_n(\pi^*) + \tilde{V}_n(\pi^*) - \hat{V}_n(\pi^*) \\ \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) &= \hat{V}_n(\hat{\pi}_n) - \tilde{V}_n(\hat{\pi}_n) + \tilde{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n), \end{aligned}$$

and using $\mathbb{E}[V(\pi^*) - \tilde{V}(\pi^*)] = 0$, we obtain:

$$\mathbb{E}[R(\hat{\pi}_n)] \leqslant \mathbb{E}[\sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi)|] + 2\mathbb{E}[\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|]. \tag{3.13}$$

By Lemma 3.7.8, the first term is bounded by $2C\sqrt{VC(\Pi)S^2/n}$, where $S^2 = \mathbb{E}[\Gamma^2]$. By the Law of Iterated Expectations and $P \in \mathcal{P}_{B_\tau,B,\eta}$,

$$\begin{aligned} \mathbb{E}[\Gamma^2] &= \mathbb{E}[(\tau_m(X,T) + g(X,Z)(Y - m(X,T)))^2] \\ &= \mathbb{E}[\tau_m^2(X,T)] + \mathbb{E}[g(X,Z)^2(Y - m(X,T))^2] \\ &\leqslant B_\tau^2 + \eta^{-2}B^2. \end{aligned}$$

The second term in (3.13) is bounded by Lemma 3.7.9, so the desired result follows.

Before proving the main result of the paper, we include another technical lemma for easier reference.

Lemma 3.7.10 (Addendum to Lemma 3.7.9)**.** *Let W_1^l denote the estimating sample with $l = (1-s)n$. In the notation of Lemma 3.7.9:*

1. *For every fixed $\pi \in \Pi$:*

$$\mathbb{E}[\hat{V}_l(\pi) - \tilde{V}_l(\pi)] \leqslant R_{3,l}.$$

2. For any $\hat{\pi}_{l,k}$ computed using the estimated sample W_1^l,

$$\mathbb{E}[\hat{V}_r(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})] \leqslant R_{3,l}.$$

Proof. To prove the first claim, we apply the same argument as in Lemma 3.7.9. The expectations of the first two corresponding terms, denoted there by $D_1(\pi)$ and $D_2(\pi)$, are equal to zero, and the expectation of the third term is shown to be less than $R_{3,l}$.

The proof of the second claim is easier, since we do not need to separate the contributions of different folds. Replacing the arguments of the functions with the index of the observation (from the test sample) to which they are applied, we can expand $\hat{\Gamma}_i - \Gamma_i$ as a sum of three terms:

$$\hat{\Gamma}_i - \Gamma_i = (\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) + (Y_i - m_i)(\hat{g}_i - g_i) - (\hat{m}_i - m_i)(\hat{g}_i - g_i).$$

Let D_1, D_2 and D_3 denote the corresponding terms in $\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})$. Then, by Assumption 3.2.1-2 and the Law of Iterated Expectations,

$$\mathbb{E}[D_1|W_1^l] = \mathbb{E}\left[\, \hat{\pi}_{l,k}(X_i) \cdot \mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|X_i, W_1^l] \;\middle|\; W_1^l \,\right] = 0.$$

Further, by the Law of Iterated Expectations and the exclusion restriction on Z_i,

$$\mathbb{E}[D_2|W_1^l] = \mathbb{E}\left[\, \hat{\pi}_{l,k}(X_i) \cdot \mathbb{E}[Y_i - m_i|X_i, T_i, W_1^l] \cdot (\hat{g}_i - g_i) \;\middle|\; W_1^l \,\right] = 0.$$

Finally, by Cauchy-Schwartz inequality and $\hat{\pi}_{l,k}(X_i)^2 \leqslant 1$,

$$D_3 \leqslant \sqrt{\tfrac{1}{r}\textstyle\sum_i (\hat{m}_i - m_i)^2} \cdot \sqrt{\tfrac{1}{r}\textstyle\sum_i (\hat{g}_i - g_i)^2}.$$

Taking expectations on both sides, applying Cauchy-Schwartz inequality again, and using the Law of Iterated Expectations, we obtain

$$\mathbb{E}[D_3] \leqslant \sqrt{\mathbb{E}[(\hat{m}_i - m_i)^2] \cdot \mathbb{E}[(\hat{g}_i - g_i)^2]} \leqslant R_{3,l},$$

■

3.7.4.2 Proof of Theorem 3.4.2

To keep the notation simple, we write $\hat{\pi}_{n,\hat{k}}$ instead of $\hat{\pi}_n^{RPWM}$ and $\mathbb{E}$ instead of $\mathbb{E}_P$ for a fixed distribution $P \in \mathcal{P}_{B_\tau,B,\eta}$. The subscripts l, r, and n indicate that the corresponding object depends only on the estimating sample, only on the test sample, or on the entire sample. For example, while $\hat{\pi}_{l,k}$ only depends on the estimating sample, $\hat{\pi}_{n,\hat{k}}$ depends on the entire sample due to the choice of $\hat{k}$. Let $\pi_k^* \in \Pi_k$ be such that $V(\pi_k^*) = V_{\Pi_k}^*$. Write:

$$V_\Pi^* - V(\hat{\pi}_{n,\hat{k}}) = V_\Pi^* - V_{\Pi_k}^* + \underbrace{V_{\Pi_k} - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}})}_{(I)} + \underbrace{Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})}_{(II)}. \tag{3.14}$$

First, since $Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) \geqslant Q_{n,k}(\hat{\pi}_{l,k})$, and $\hat{V}_l(\hat{\pi}_{l,k}) \geqslant \hat{V}_l(\pi_k^*)$, we can bound:

$$\begin{aligned}(I) &\leqslant V(\pi_k^*) - Q_{n,k}(\hat{\pi}_{l,k}) \\ &\leqslant V(\pi_k^*) - \hat{V}_l(\pi_k^*) + \hat{C}_{n,k} \\ &= V(\pi_k^*) - \tilde{V}_l(\pi_k^*) + \tilde{V}_l(\pi_k^*) - \hat{V}_l(\pi_k^*) + \hat{C}_{n,k}.\end{aligned}$$

Here, $\mathbb{E}[V(\pi_k^*) - \tilde{V}_l(\pi_k^*)] = 0$ and, by Lemma 3.7.10, $\mathbb{E}[\tilde{V}_l(\pi_k^*) - \hat{V}_l(\pi_k^*)] \leqslant R_{3,l}$. Therefore,

$$\mathbb{E}[(I)] \leqslant \mathbb{E}[\hat{C}_{n,k}] + R_{3,l}.$$

Next, consider

$$(II) = \hat{V}_r(\hat{\pi}_{n,\hat{k}}) - \tilde{V}_r(\hat{\pi}_{n,\hat{k}}) + (\tilde{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})).$$

The first summand can be bounded by

$$\begin{aligned}\mathbb{E}\left[\hat{V}_r(\hat{\pi}_{n,\hat{k}}) - \tilde{V}_r(\hat{\pi}_{n,\hat{k}})\right] &\leqslant \mathbb{E}\left[\max_{k\leqslant K} |\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right] \\ &\leqslant K \max_{k\leqslant K} \mathbb{E}\left[|\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right].\end{aligned}$$

As in the proof of Lemma 3.7.9, we can expand:

$$\hat{\Gamma}_i - \Gamma_i = (\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) + (Y_i - m_i)(\hat{g}_i - g_i) - (\hat{m}_i - m_i)(\hat{g}_i - g_i),$$

so that:

$$\begin{aligned}\mathbb{E}\left[|\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right] &= \mathbb{E}\left[|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{\Gamma}_i - \Gamma_i)|\right] \\ &\leqslant \tfrac{1}{\sqrt{r}}\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|\right] \\ &+ \tfrac{1}{\sqrt{r}}\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)|\right] \\ &+ \mathbb{E}\left[|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)|\right]\end{aligned}$$

By Assumption 3.2.1-2 and the Law of Iterated Expectations,

$$\mathbb{E}[\hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|W_1^n, X_i] = 0.$$

Using $\mathbb{E}[|W|]^2 \leqslant \mathbb{E}[W^2]$, the Law of Iterated Expectations, $\hat{\pi}_{l,k}^2(X_i) \leqslant 1$, and Assumption 3.2.2, we obtain:

$$\begin{aligned}\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|\right]^2 &\leqslant \mathbb{E}\left[\tfrac{1}{r}\textstyle\sum_i(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))^2\right] \\ &= \mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))^2] \\ &\leqslant 2(\mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i})^2] + \mathbb{E}[g_i^2(\hat{m}_i - m_i)^2]) \\ &\leqslant 2\tfrac{\eta^2+1}{\eta^2}\tfrac{a((1-J^{-1})l)}{l^{\zeta_m}}.\end{aligned}$$

A similar argument and the bound $\mathbb{E}[(Y_i - m_i)^2|X_i, T_i] \leqslant B^2$ yield:

$$\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)|\right]^2 \leqslant \mathbb{E}[(Y_i - m_i)^2(\hat{g}_i - g_i)^2] \leqslant B^2 \cdot \tfrac{a((1-J^{-1})l)}{l^{\zeta_g}}.$$

Next, by Cauchy-Schwartz inequality and $\hat{\pi}_{l,k}^2(X_i) \leqslant 1$,

$$|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)| \leqslant \sqrt{\tfrac{1}{r}\textstyle\sum_i(\hat{m}_i - m_i)^2} \cdot \sqrt{\tfrac{1}{r}\textstyle\sum_i(\hat{g}_i - g_i)^2}$$

Taking expectations on both sides, applying Cauchy-Schwartz inequality and the Law of Iterated Expectations,

$$\mathbb{E}[|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)|] \leqslant \sqrt{\mathbb{E}[(\hat{m}_i - m_i)^2] \cdot \mathbb{E}[(\hat{g}_i - g_i)^2]} \leqslant \sqrt{\tfrac{a((1-J^{-1})l)}{l^{\zeta_m+\zeta_g}}}$$

Combining the above results, we obtain:

$$\mathbb{E}\left[|\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right] \leqslant \sqrt{\tfrac{1}{s(1-s)^{\zeta_m \wedge \zeta_g}}}\sqrt{\tfrac{2(\eta^2+1)}{\eta^2} \vee B^2}\sqrt{\tfrac{a((1-J^{-1})(1-s)n)}{n^{1+\zeta_m \wedge \zeta_g}}} + R_{3,(1-s)n}$$

For the second summand in (II), arguing as in the proof of Theorem 3.3.3 (see Equations (3.7), (3.8), and the following argument and recall that $\hat{V}_n$ in that proof plays the same role as $\tilde{V}_r$ in this one),

$$\mathbb{E}\left[\tilde{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})\right] \leqslant \sqrt{B_\tau^2 + \eta^{-2}B^2}\frac{K}{\sqrt{sn}} = K\frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\sqrt{\frac{1}{sn}}.$$

Let $R_{3,(1-s)n}$ denote the rate Lemma 3.7.9 with $(1-s)n$ instead of n. Defining

$$S_{2,n} \equiv \sqrt{\frac{1}{s(1-s)^{\zeta_m\wedge\zeta_g}}}\sqrt{\frac{2(\eta^2+1)}{\eta^2}\vee B^2}\sqrt{\frac{a((1-J^{-1})(1-s)n)}{n^{1+\zeta_m\wedge\zeta_g}}} + 2R_{3,(1-s)n} \tag{3.15}$$

and

$$S_n \equiv K\frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\sqrt{\frac{1}{sn}} + S_{2,n}, \tag{3.16}$$

we conclude that

$$\mathbb{E}[(I)+(II)] \leqslant \mathbb{E}[\hat{C}_{n,k}] + S_n.$$

Therefore, for any $k \leqslant K$,

$$\mathbb{E}[R(\hat{\pi}_{n,\hat{k}})] \leqslant V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}] + S_n, \tag{3.17}$$

and the first statement of the Theorem follows from taking an infimum over $k \leqslant K$.

To prove the second statement, it remains to bound $\mathbb{E}[\hat{C}_{n,k}]$. To this end, write:

$$\begin{aligned}\hat{C}_{n,k} &= \tilde{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k}) + \hat{V}_l(\hat{\pi}_{l,k}) - \tilde{V}_l(\hat{\pi}_{l,k}) \\ &+ \tilde{V}_r(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k}) \\ &+ V(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k}).\end{aligned}$$

By Lemmas 3.7.8 and 3.7.9, for any $P \in \mathcal{P}_{B_\tau,B,\eta}$,

$$\begin{aligned}\mathbb{E}[\tilde{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})] &\leqslant C\frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\sqrt{\frac{VC(\Pi_k)}{(1-s)n}}, \\ \mathbb{E}[\hat{V}_l(\hat{\pi}_{l,k}) - \tilde{V}_l(\hat{\pi}_{l,k})] &\leqslant R^k_{1,(1-s)n} + R^k_{2,(1-s)n} + R_{3,(1-s)n},\end{aligned}$$

where $R^k_{j,(1-s)n}$, for $j=1,2,3$, are defined in Lemma 3.7.9 with Π_k instead of Π and $(1-s)n$ instead of n. Finally, by Lemma 3.7.10, $\mathbb{E}[\tilde{V}_r(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})] \leqslant R_{3,(1-s)n}$, and by the Law of

Iterated Expectations, $\mathbb{E}[V(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})] = 0$. Plugging the above into (3.17), and noting that for every $P \in \mathcal{P}^k_{B_\tau,B,\eta}$, we have $V^*_\Pi = V^*_{\Pi_k}$,

$$\sup_{P \in \mathcal{P}^k_{B_\tau,B,\eta}} \mathbb{E}_P[R(\hat{\pi}_{n,\hat{k}})] \leqslant \frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\left(C\sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K\sqrt{\frac{1}{sn}}\right) + S^k_{1,n} + S_{2,n},$$

where $S^k_{1,n} = R^k_{1,(1-s)n} + R^k_{2,(1-s)n} + R_{3,(1-s)n}$, and $S_{2,n}$ is given in Equation 3.15.

3.7.4.3 Proof of Theorem 3.4.3

We consider a particular subclass of $\mathcal{P}_{B,\eta}$ for which the worst-case regret can be bounded from below by a term proportional to $B/\eta\sqrt{d/n}$. The construction proceeds as follows. Let $x_1, \ldots, x_d$, where $d = VC(\Pi) - 1$, be a set shattered by Π with the largest possible cardinality. Let

$$X \in \{x_1, \ldots, x_d\}, \;\; P(X = x_j) = \tfrac{1}{d};$$
$$T \in \{0, 1\}, \;\; P(T = 1) = p, \;\; T \perp (X, Y_0, Y_1);$$
$$Y_0 = 0,$$

and, given a parameter vector $c = (c_1, \ldots, c_d) \in \{-1, 1\}^d$,

$$Y_1|X = x_j = \begin{cases} A & \text{w.p. } \frac{1}{2}(1 + c_j\frac{\gamma}{A}) \\ -A & \text{w.p. } \frac{1}{2}(1 - c_j\frac{\gamma}{A}) \end{cases},$$

where $\gamma/A \leqslant 1$. Then, for $Y = TY_1 + (1 - T)Y_0$,

$$\mathbb{E}(Y^2) = pA^2,$$
$$\tau(x_j) = \mathbb{E}[Y_1 - Y_0|X = x_j] = \gamma c_j.$$

For every $c \in \{-1, 1\}^d$, the joint distribution of $W = (Y, X, T)$ constructed above belongs to $P_{B,\eta}$ as long as $p \in [\eta, 1 - \eta]$ and $pA^2 \leqslant B^2$. We will specify such p and A later.

Let $C = (C_1, \ldots, C_d)$ consist of i.i.d. random variables $C_j \in \{-1, 1\}$ such that $P(C_j =$

$1) = 1/2$. The joint distribution of $W = (Y, X, T)$ given $C = c$ is

$$P(Y = y, X = x_j, T = t|C = c) = \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + c_j\frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - c_j\frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases} .$$

We shall also derive the posterior probability $P(C_j = 1|W_1^n)$ which will play a crucial role in deriving the lower bound.

We have

$$P(Y = y, X = x_j, T = t) = \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}\frac{p}{d} & y = -A, t = 1 \end{cases} ,$$

and

$$\begin{aligned} P(Y = y, X = x_k, T = t|C_j = 1) &= \mathbf{1}(k \neq j)P(Y = y, X = x_j, T = t) \\ &+ \mathbf{1}(k = j) \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + \frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - \frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases} . \end{aligned}$$

Therefore,

$$\frac{P(W_i|C_j = 1)}{P(W_i)} = \mathbf{1}(X_i \neq x_j) + \mathbf{1}(X_i = x_j) \begin{cases} 1 & Y_i = 0, T_i = 0 \\ 1 + \frac{\gamma}{A} & Y_i = A, T_i = 1 \\ 1 - \frac{\gamma}{A} & Y_i = -A, T_i = 1 \end{cases} ,$$

and

$$P(C_j = 1|W_1^n) = \frac{P(W_1^n|C_j = 1)P(C_j = 1)}{P(W_1^n)} = \frac{1}{2}\left(1 + \frac{\gamma}{A}\right)^{N_j^+}\left(1 - \frac{\gamma}{A}\right)^{N_j^-}, \tag{3.18}$$

where

$$\begin{aligned} N_j^+ &= \#\{i : X_i = x_j, Y_i = A, T_i = 1\} \\ N_j^- &= \#\{i : X_i = x_j, Y_i = -A, T_i = 1\}, \end{aligned}$$

so that a tuple $(N_j^+, N_j^-, n - N_j^+ - N_j^-)$ has a multinomial distribution:

$$P(N_j^+ = k_1, N_j^- = k_2 | C_j = 1)$$
$$= \binom{n}{k_1}\binom{n-k_1}{k_2}\left(\frac{1}{2}(1+\frac{\gamma}{B})\frac{p}{d}\right)^{k_1}\left(\frac{1}{2}(1-\frac{\gamma}{B})\frac{p}{d}\right)^{k_2}\left(1-\frac{p}{d}\right)^{n-k_1-k_2}. \quad (3.19)$$

Let $\mathcal{P}_\mathcal{C} = \{P_{W|C=c} : c \in \{-1,1\}^d\} \subset \mathcal{P}_{B,\eta}$ be a set of distributions of $W = (Y, X, T)$ introduced above, and μ denote the distribution of C. Let π_P^* denote the first-best treatment rule when the distribution of the data is P, and write $\pi_c^* = \pi^*_{P_{W|C=c}}$ for brevity. By construction, $\pi_c^*(x_j) = \mathbf{1}(c_j = 1)$, and $\pi_c^* \in \Pi$ since the class Π shatters $\{x_1, \ldots, x_d\}$. Note that:

$$V(\pi_c^*) - V(\hat{\pi}_n) = \frac{\gamma}{d}\sum_{j=1}^d c_j(\pi_c^*(x_j) - \hat{\pi}_n(x_j)) = \frac{\gamma}{d}\sum_{j=1}^d \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)).$$

Then,

$$\begin{aligned}
\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \max_{P\in\mathcal{P}_\mathcal{C}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n))] \\
&\geqslant \int \mathbb{E}_{P_{W_1^n|C=c}}[V(\pi_c^*) - V(\hat{\pi}_n))]d\mu(c) \\
&= \frac{\gamma}{d}\sum_{j=1}^d \int\int \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j))dP_{W_1^n|C=c}d\mu(c) \qquad (3.20) \\
&= \frac{\gamma}{d}\sum_{j=1}^d P_{W_1^n,C_j}(\mathbf{1}(C_j = 1) \neq \hat{\pi}_n(x_j)) \\
&\geqslant \gamma \cdot \inf_\pi P_{W_1^n,C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n)).
\end{aligned}$$

Note that $P_{W_1^n,C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n))$ is the probability of misclassification of $\mathbf{1}(C_j = 1)$ using W_1^n. By Theorem 2.1. in Devroye and Lugosi (1996), the infimum is attained by the Bayes Classifier, $\pi^*(W_1^n) = \mathbf{1}(P(C_j = 1|W_1^n) > 0.5)$, and is equal to

$$\begin{aligned}
P(\mathbf{1}(C_j = 1) \neq \pi^*(W_1^n)) &= \tfrac{1}{2}P(\ P(C_j = 1|W_1^n) \leqslant 0.5\ |C_j = 1) \\
&\quad + \tfrac{1}{2}P(\ P(C_j = 1|W_1^n) > 0.5\ |C_j = -1).
\end{aligned}$$

Denote $a = \gamma/A$, and work conditional on $C_j = 1$ from now on. Recalling (3.18),

$$\begin{aligned} P(P(C_j = 1|W_1^n) \leqslant 0.5) &= P((1+a)^{N_j^+}(1-a)^{N_j^-} \leqslant 1) \\ &\geqslant P((1-a^2)^{N_j^+} \leqslant 1 | N_j^+ \leqslant N_j^-) \cdot P(N_j^+ \leqslant N_j^-) \\ &= P(N_j^+ \leqslant N_j^-). \end{aligned}$$

Let $D_i^+ = \mathbf{1}(X_i = x_j, Y_i = A, T_i = 1)$ and $D_i^- = \mathbf{1}(X_i = x_j, Y_i = -A, T_i = 1)$. Then, $\mathbb{E}[D_i^+ - D_i^-] = ap/d$, $\mathbb{V}ar[D_i^+ - D_i^-] = p/d - (ap/d)^2$, and $\mathbb{E}[(D_i^+ - D_i^-)^3] = p/d$. Letting Z_n denote the studentized version of $n^{-1}\sum_{i=1}^n (D_i^+ - D_i^-)$ and Φ denote the Standard Normal CDF, using Berry-Esseen inequality we obtain

$$\begin{aligned} P(N_j^+ \leqslant N_j^-) &= P(\tfrac{1}{n}\textstyle\sum_{i=1}^n (D_i^+ - D_i^-) \leqslant 0) \\ &= P\left(Z_n \leqslant \frac{-\sqrt{n}ap/d}{\sqrt{p/d - (ap/d)^2}}\right) \\ &\geqslant \Phi\left(\frac{-\sqrt{n}ap/d}{\sqrt{p/d-(ap/d)^2}}\right) - \frac{K}{\sqrt{n}} \frac{1}{(p/d)^{1/2}(1-a^2p/d)^{3/2}}, \end{aligned}$$

where $K < 0.469$ (Shevtsova, 2013). Choosing $a = \gamma/A \equiv c/\sqrt{n}\sqrt{d/p}$ for some $c \in (0,1)$, assuming n is large enough to satisfy $\gamma/A \leqslant 1$, we obtain

$$P(N_j^+ \leqslant N_j^-) \geqslant \Phi\left(-\frac{c}{\sqrt{1-c^2/n}}\right) - \frac{K}{\sqrt{n}} \frac{1}{\sqrt{p/d}(1-c^2/n)^{3/2}}.$$

Choosing $p = \eta$, $A = B/\sqrt{\eta}$ so that $\gamma = c \cdot B/\eta\sqrt{d/n}$, we have, for $n \geqslant 3$,

$$\begin{aligned} \sup_{P \in \mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \tfrac{\gamma}{2} \cdot P(N_j^+ \leqslant N_j^- | C_j = 1) \\ &\geqslant \tfrac{1}{2}\tfrac{B}{\eta}\sqrt{\tfrac{d}{n}} \cdot c \cdot \Phi\left(-\tfrac{c}{\sqrt{1-c^2}}\right) - \tfrac{K}{2\sqrt{\eta}} \cdot \tfrac{B}{\eta}\tfrac{d}{n}\tfrac{c}{(1-c^2/3)^{3/2}} \end{aligned}$$

Choosing $c = 0.5162$, and plugging in $K = 0.469$ gives the final result

$$\sup_{P \in \mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant 0.07 \cdot \frac{B}{\eta}\sqrt{\frac{d}{n}} - \frac{0.14}{\sqrt{\eta}} \cdot \frac{B}{\eta}\frac{d}{n}.$$

For $n \geqslant 4d/\eta$, the right-hand-side in the preceding display is positive, and $\gamma/A \leqslant 1$ is also satisfied.

Bibliography

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Armstrong, T. and Shen, S. (2015). Inference on optimal treatment assignments.

Athey, S., Chetty, R., and Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

Athey, S., Imbens, G. W., Wager, S., et al. (2016). Efficient inference of average treatment effects in high dimensions via approximate residual balancing. Technical report.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2011). $\ell 1$-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2015). Program evaluation with high-dimensional data. Technical report, cemmap working paper.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. (2013). Covariance adjustments for the analysis of randomized field experiments. *Evaluation review*, 37(3-4):170–196.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

Box, G. E., Hunter, J. S., Hunter, W. G., et al. (2005). *Statistics for experimenters: design, innovation, and discovery*, volume 2. Wiley-Interscience New York.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125(1-2):141–173.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 123–132.

Ding, P., Feller, A., and Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Freedman, D. A. (2008). On regression adjustments in experiments with several treatments. *The annals of applied statistics*, 2(1):176–196.

Gagnon-Bartsch, J. A., Sales, A. C., Wu, E., Botelho, A. F., Erickson, J. A., Miratrix, L. W., and Heffernan, N. T. (2021). Precise unbiased estimation in randomized experiments using auxiliary observational data. *arXiv preprint arXiv:2105.03529*.

Gui, G. (2020). Combining observational and experimental data using first-stage covariates. *arXiv preprint arXiv:2010.05117*.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J. S. (2015). From sample average treatment e ect to population average treatment e ect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society, Series A ()*.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Kohavi, R. and Longbotham, R. (2017). Online controlled experiments and a/b testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929.

Kohavi, R. and Thomke, S. (2017). The surprising power of online experiments. *Harvard business review*, 95(5):74–82.

Lewis, R. A. and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *The Quarterly Journal of Economics*, 130(4):1941–1973.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Liou, K. and Taylor, S. J. (2020). Variance-weighted estimators to improve sensitivity in online experiments. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 837–850.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

Mason, R. L., Gunst, R. F., and Hess, J. L. (2003). *Statistical design and analysis of experiments: with applications to engineering and science*, volume 474. John Wiley & Sons.

Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Peysakhovich, A. and Lada, A. (2016). Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385*.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sales, A. C., Hansen, B. B., and Rowan, B. (2018). Rebar: Reinforcing a matching estimator with predictions from high-dimensional covariates. *Journal of Educational and Behavioral Statistics*, 43(1):3–31.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.

Tang, D., Agarwal, A., O'Brien, D., and Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 17–26.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678.

Wu, E. and Gagnon-Bartsch, J. A. (2021). Design-based covariate adjustments in paired experiments. *Journal of Educational and Behavioral Statistics*, 46(1):109–132.

Xie, H. and Aurisset, J. (2016). Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 645–654.