HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Off Policy Reinforcement Learning for Real World Settings"

presented by Aarón Michael Sonabend Worthalter

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature .....

Typed name: Prof. Tianxi Cai

Signature .....

Typed name: Prof. Peter Szolovits

Signature Rajarshi Mukherjee
Rajarshi Mukherjee (Apr 29, 2021 18:40 EDT)

Typed name: Prof. Rajarshi Mukherjee

Signature .....

Typed name:

Date: April 29, 2021

Off Policy Reinforcement Learning for Real-World Settings

A dissertation presented

by

Aarón Michael Sonabend Worthalter

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

April 2021

Dissertation Advisor: Professor Tianxi Cai

Aarón Michael Sonabend Worthalter

Off Policy Reinforcement Learning for Real-World Settings

Abstract

In this dissertation, we aim to adapt reinforcement learning (RL) to real-world, high-risk settings. We study how to optimize sequential decision-making in complex settings with large observational data repositories where exploration is unfeasible. In particular, we are motivated by estimating optimal dynamic treatment regimes (DTR) with electronic health records (EHR). We address some of the challenges that differentiate off-policy RL in high-risk settings from other contexts. For example, we account for sampling bias and uncertainty to yield causally valid inference. Additionally, our resulting policy functions are interpretable by domain experts. We also provide measures of statistical efficiency, which are crucial in our settings of large but finite and noisy data.

In Chapter 1, we propose an offline policy and value function learning method based on Bayesian RL. Our estimated policy is optimal and safe as it handles uncertainty through hypothesis testing, allows for different levels of risk aversion, and is interpretable. We provide consistency results and a regret bound, which establishes sample efficiency. The theoretical results are independent of the risk aversion threshold and quality of the expert policy.

Chapter 2 develops a semi-supervised RL (SSRL) method for Q-learning and doubly-robust off-policy evaluation. SSRL is specifically relevant to EHR data where outcome information is often not well coded but rather embedded in clinical notes. Our approach leverages a small dataset with true outcomes observed and a large dataset with outcome surrogates. We provide theoretical results for our estimators to understand to what degree efficiency can be gained from SSRL. Our method is at least as efficient as the supervised approach and robust to the misspecification of the imputation models.

Chapter 3 seeks to find the optimal DTR, which maximizes the value function through non-parametric estimation. We frame this as a multi-stage classification problem. To address the discontinuity of the objective function, we use a smooth surrogate for the value function. In particular, we characterize a family of smooth surrogate functions that are Fisher consistent and provide a regret bound tailored to the non-parametric estimation method. In addition, smoothness in the surrogate value function makes the method scalable to large sample sizes.

Contents

List of Tables

List of Figures

Acknowledgments

I start by stating how fortunate I am to have Professor Tianxi Cai as my advisor. Tianxi motivated me to become the best researcher I have in me. She allowed me to learn how to find high-level motivation in real-world problems and translate them into challenging and relevant statistical problems. I am also deeply grateful for having Tianxi respect my own research and teaching interests and found ways to let me grow within these areas.

I also want to acknowledge my thesis committee members Professors Peter Szolovits and Rajarshi Mukherjee. Pete was always a friendly face that provided me with great research and life perspective, and of course, introduced me to reinforcement learning. Rajarshi shared with me his infinite mathematical wisdom in a patient and enjoyable way.

I would also like to thank Professor Junwei Lu for always being a part of my team and the endless hours of discussion, which helped me be strategic in my research and career. Jesse Gronsbell and Nilanjana Laha for being my friends and mentors who showed me the way forward. Finally, my acquired friends have been one of the most rewarding aspects of the PhD, especially Eric Dunipace and Greyson Liu, who made all this process very enjoyable and helped me get the often needed laughs distractions.

I want to dedicate this work to my parents Roberto and Fanny Sonabend, and my wife, Victoria. My marriage began the same summer that my PhD did. They have both been an intertwined adventure; all of the joys and challenges I have faced with her beside me. I sincerely thank you for your support, empathy, and for being a part of my life. This work is the result of our team. I am sincerely grateful to my parents, who have made me the person I am today, by encouraging me to pursue my curiosities. They have constantly made an effort to make me feel supported, to understand my interests, and be a part of my world.

To Roberto, Fanny, and Victoria

Introduction

Reinforcement learning (RL) is a family of machine learning and statistical methods that focus on optimizing sequential decision making. There is a vast realm of RL methods that aim to tackle different applications. Recent years have witnessed a remarkable success of RL methods in applications such as robotics, online marketing, video and board games, and physical simulators [1]. On the other hand, the digitization of electronic health records (EHR) has brought forth a vast amount of datasets rich in medical information. These data, originally collected for billing and insurance purposes, contain valuable clinical information in the form of large observational records. In this dissertation, we aim to develop RL methodology to learn from real observational data to estimate optimal policies. In particular, we motivate our methods using EHR data to find optimal treatment regimes for complex diseases.

RL has experienced a remarkable stream of successes in the last few years in the computer science field. Most of the literature has been primarily dedicated to self-contained "laboratory-like" environments. These environments include video games, board games, robotic manipulation, and physical simulations. These settings vary in complexity and often serve as a test-bed to develop powerful RL methods. There are several advantages of using these environments. Primarily, the state-space is always fully observed; for example, the video game screen at any given time point is always measurable. Another important advantage is that interactions with the environment are relatively cheap to collect or simulate, making the sample size of the training data virtually infinite. RL methods trained in these settings focus on computational efficiency and optimizing the exploitation vs. exploration strategies. In other words, when to follow the estimated optimal policy vs. when to keep exploring the

environment. There is a large amount of work that tackles both the practical and theoretical aspects of these types of problems, yielding exciting results.

Adapting RL to real-world settings such as healthcare has its own sets of challenges. At a high level, the problem is quite different from the one previously discussed. Research focuses on off-policy learning with a partially observed state and limited data. The off-policy learning setting implies that a different agent has already carried out all exploration of the environment. All that is left to do is learn an optimal policy based on the observed data. For example, in healthcare, the treatments and patient responses have already been carried out and logged in the medical records. An additional high-level challenge of dealing with real-world settings is that multiple sources of heterogeneity make the state high-dimensional and only partially observed. A treatment response might depend on demographic and socioeconomic status, physiological characteristics, genetic components, and medical history in healthcare settings. These treatment interactions make for a high-dimensional state and are often hard to measure. [2, 3, 4, 5, 6, 7, 8, 9]

These high-level difficulties that arise when working on real-world observational data sets such as EHR bring interesting statistical challenges when learning optimal treatments. Continuing with the healthcare setting example, EHR data are noisy, finite, and suffer from sampling bias. In this dissertation, we focus on developing methods that address these challenges. First, our algorithms yield results with associated measures for uncertainty. Such measures are important as patients are ultimately stochastic environments: based on what we can measure, we cannot expect two patients in the same state to respond to the same treatment in the same way. Second, we provide theoretical results that show our inference is statistically efficient, ensuring that we extract as much information as possible from our large but limited datasets. Third, observational data require off-policy RL methods; hence we need to be careful to account for treatment by indication bias. For example, sicker patients who are more likely to die are also often more likely to receive higher doses of medication. Failing to account for bias might lead the agent to interpret medicine as harmful to life. Fourth, we focus on providing methods that are interpretable in their application domain. We believe

this is important as it makes RL more likely to get adopted by users if the methods can be critiqued and the decisions can be understood through the practitioner's eyes.

Chapter 1

Expert-Supervised Reinforcement Learning for Offline Policy Learning and Evaluation

Aarón Sonabend[1], Junwei Lu[1], Leo A. Celi[2], Tianxi Cai[1] and Peter Szolovits[3]

[1]Department of Biostatistics

Harvard University

[2]Institute for Medical Engineering & Science

Massachusetts Institute of Technology

[3]Computer Science & Artificial Intelligence Laboratory

Massachusetts Institute of Technology

1.1 Summary

Offline Reinforcement Learning (RL) is a promising approach for learning optimal policies in environments where direct exploration is expensive or unfeasible. However, the adoption of such policies in practice is often challenging, as they are hard to interpret within the application

context, and lack measures of uncertainty for the learned policy value and its decisions. To overcome these issues, we propose an Expert-Supervised RL (ESRL) framework which uses uncertainty quantification for offline policy learning. In particular, we have three contributions: 1) the method can learn safe and optimal policies through hypothesis testing, 2) ESRL allows for different levels of risk averse implementations tailored to the application context, and finally, 3) we propose a way to interpret ESRL's policy at every state through posterior distributions, and use this framework to compute off-policy value function posteriors. We provide theoretical guarantees for our estimators and regret bounds consistent with Posterior Sampling for RL (PSRL). Sample efficiency of ESRL is independent of the chosen risk aversion threshold and quality of the behavior policy.

1.2 Introduction

With increasing success in reinforcement learning (RL), there is broad interest in applying these methods to real-world settings. This has brought exciting progress in offline RL and off-policy policy evaluation (OPPE). These methods allow one to leverage observed data sets collected by expert exploration of environments where, due to costs or ethical reasons, direct exploration is not feasible. Sample-efficiency, reliability, and ease of interpretation are characteristics that offline RL methods must have in order to be used for real-world applications with high risks, where a tendency is exhibited towards sampling bias. In particular there is a need for policies that shed light into the decision-making at all states and actions, and account for the uncertainty inherent in the environment and in the data collection process. In healthcare data for example, there is a common bias that arises: drugs are mostly prescribed only to sick patients; and so naive methods can lead agents to consider them harmful. Actions need to be limited to policies which are similar to the expert behavior and sample size should be taken into account for decision-making [8, 9].

To address these deficits we propose an Expert-Supervised RL (ESRL) approach for offline learning based on Bayesian RL. This method yields safe and optimal policies as it learns when to adopt the expert's behavior and when to pursue alternative actions. Risk aversion might

vary across applications as errors may entail a greater cost to human life or health, leading to variation in tolerance for the target policy to deviate from expert behavior. ESRL can accommodate different risk aversion levels. We provide theoretical guarantees in the form of a regret bound for ESRL, independent of the risk aversion level. Finally, we propose a way to interpret ESRL's policy at every state through posterior distributions, and use this framework to compute off-policy value function posteriors for any given policy.

While training a policy, ESRL considers the reliability of the observed data to assess whether there is substantial benefit and certainty in deviating from the behavior policy, an important task in a context of limited data. This is embedded in the method by learning a policy that chooses between the optimal action or the behavior policy based on statistical hypothesis testing. The posteriors are used to test the hypothesis that the seemingly optimal action is indeed better than the one from the behavior policy. Therefore, ESRL is robust to the quality of the behavior policy used to generate the data.

To understand the intuition for why hypothesis testing works for offline policy learning, we discuss an example. Consider a medical setting where we are interested in the best policy to treat a complex disease over time. We first assume there is a standardized treatment guideline that works well and that most physicians adopt it to treat their patients. The observed data will have very little exploration of the whole environment —in this case, meaning little use of alternative treatments. However, the state-action pairs observed will be near optimal. For any fixed state, those actions not recommended by the treatment guidelines will be rare in the data set and the posterior distributions will be dominated by the uninformative wide priors. The posteriors for the value associated with the optimal actions will incorporate more information from the data as they are commonly observed. Thus, testing for the null hypothesis that an alternative action is better than the treatment guideline will likely yield a failure to reject the null, and the agent will conclude the physician's action is best. Unless the alternative is substantially better for a given state, the learned policy will not deviate from the expert's behavior when there is a clear standard of care.

On the other hand, if there is no treatment guideline or consensus among physicians,

different doctors will try different strategies and state-action pairs will be more uniformly observed in the data. At any fixed state, some relatively good actions may have narrower posterior distributions associated with their value. Testing for the null hypothesis that a fixed action is better than what the majority of physicians chose is more likely to reject the null and point towards an alternative action in this case, as variance will be smaller across the sampled actions. Deviation from the (noisy) behavior policy will occur more frequently. Therefore, whether there is a clear care guideline or not, the method will have learned a suitable policy. A central point in Bayesian RL is that the posterior provides not just the expected value for each action, but also higher moments. We leverage this to produce interpretable policies which can be understood and analyzed within the context of the application. We illustrate this with posterior distributions and credible intervals (CI). We further propose a way to produce posterior distributions for OPPE with consistent and unbiased estimates.

Handling Uncertainty. To the best of our knowledge, there is no work that has incorporated hypothesis testing directly into the policy training process. However, accounting for the uncertainty in policy estimation is a successful idea which has been widely explored in other works. Methods range from confidence interval estimation using bootstrap, to model ensembles for guiding online exploration [10, 11, 12]. For example, a simple and effective way of incorporating uncertainty is through random ensembles (REM) [13]. These have shown promise on Atari games, significantly outperforming Deep Q networks (DQN) [14] and naive ensemble methods in the offline setting. We adopt the Bayesian framework, which has been proven successful in online RL [15, 16], as it provides a natural way to formalize uncertainty in finite samples. Bayesian model free methods such as temporal difference (TD) learning provide provably efficient ways to explore the dynamics of the MDP [17, 18, 19]. Gaussian Process TD can also be used to provide posterior distributions with mean value and CI for every state-action pair [20]. Although efficient for online exploration, TD methods require large data in high dimensional settings, which can be a challenge in complex offline applications such as healthcare. ESRL is model-based which makes it sample efficient [21]. Within model-based methods, the Bayesian framework allows for natural incorporation of uncertainty measures.

Posterior sampling RL proposed by Strens efficiently explores the environment by using a single MDP sample per episode [22]. ESRL fits within this line of methods, which are theoretically guaranteed to be efficient in terms of finite time regret bounds [23, 24].

Hypothesis Testing for Offline RL Naively applying model-based RL to offline, high dimensional tasks can degrade its performance, as the agent can be led to unexplored states where it fails to learn reliable policies. There are environments where simple approaches like behavior cloning (BC) on the offline data set is enough to ensure reliability. BC has actually been shown to perform quite well in offline benchmarks like RL Unplugged [25], D4RL [26] and Atari when the data is collected from a single noisy behavior policy [27]. The issue with these approaches is that interest lies in policy improvement with respect to the expert, and there is no guarantee that the learned policies are safe in all states, a necessary condition when treating patients. A common strategy is to regularize the learned policy towards the behavior policy whether directly in the state space or in the action space [25, 27, 28, 29, 30]. However, there are cases where the data logging policy is a noisy representation of the expert behavior, and regularization will lead to sub-optimal actions. ESRL can detect these cases through hypothesis testing [31] to check whether improvement upon the behavior policy is feasible and, if so, incorporate new actions into the policy in accordance with the user's risk tolerance. Additionally, as opposed to the regularization hyper-parameter that one must choose for methods like Batch Constrained deep Q-learning (BCQ) [25, 27], the risk-aversion parameter has a direct interpretation as the significance level that the user is comfortable with for the policy to deviate from the expert behavior. It allows the method to be tailored to different scientific and business applications where one might have different tolerance towards risk in search for higher rewards.

Off-Policy Policy Evaluation and Interpretation. Many of the aforementioned methods can be easily adapted for offline learning and often importance sampling is used to address the distribution shift between the behavior and target policies [1]. However, importance sampling can yield high variance estimates in finite samples, especially in long episodes. Doubly robust

estimation of the value function is proposed to address these issues. These methods will have low variance and consistent estimators if either the behavior policy or the model is correctly specified [6, 7]. Still, in finite samples or environments with high dimensional state-action spaces, these doubly robust estimators may still not be reliable, because only a few episodes end up contributing to the actual value estimate due to the product in the importance sampling weights [9]. Additionally, having point estimates without any measure of associated uncertainty can be dangerous, as it is hard to know whether the sample size is large enough for the estimate to be reliable. To this end, we use the ESRL framework to sample MDP models from the posterior and evaluate the policy value. Our estimates are unbiased and consistent, and are equipped with uncertainty measures.

1.3 Problem Set-up

We are interested in learning policies that can be used in real-world applications. To develop the framework we will use the clinical example discussed in Section 1.2. Consider a finite horizon MDP defined by the following tuple: $< \mathcal{S}, \mathcal{A}, R^M, P^M, P_0, \tau >$, where $\mathcal{S}$ is the state-space, $\mathcal{A}$ is the action space, M is the model over all rewards and state transition probabilities with prior $f(\cdot)$, $R^M(s,a) : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward distribution for fixed state-action pair (s,a) under model M, with mean $\bar{R}^M(s,a)$. $P_a^M(s'|s)$ is the probability distribution function for transitioning to state s' from state-action pair (s,a) under model M, $\tau \in \mathbb{N}$ is the fixed episode length, and P_0 is the initial state distribution. The true MDP model M^* has distribution f.

The behavior policy function is a noisy version of a deterministic policy. Going back to the clinical example there is generally a consensus of what the correct treatment is for a disease, but the data will be generated by different physicians who might adhere to the consensus to varying degrees. Thus, we model the standard of care as a deterministic policy function $\pi^0 : \mathcal{S} \times \{1, \ldots, \tau\} \mapsto \mathcal{A}$. The behavior policy is $\pi(s,t) = \pi^0(s,t)$ with probability (w.p.) $1-\epsilon$, and $\pi(s,t) = a$ sampled uniformly at random from $\mathcal{A}$ w.p. ϵ. For a fixed $\epsilon \in [0,1]$, π generates the observed data $\boldsymbol{D}_T = \{(s_{i1}, a_{i1}, r_{i1}, \ldots, s_{i\tau}, a_{i\tau}, r_{i\tau})\}_{i=1}^T$ which consists of T episodes (i.e. patient treatment histories), where $s_{i1} \sim P_0 \ \forall i = 1, \ldots, T$. Note that π^0 may generally yield

high rewards, however it is not necessarily optimal and can be improved upon.

We'll denote a policy function by $\mu : \mathcal{S} \times \{1, \ldots, \tau\} \to \mathcal{A}$. The associated value function for μ, model M is $V^M_{\mu,t}(s) = \mathbb{E}_{M,\mu}\left[\sum_{j=t}^{\tau} \bar{R}^M(s_j, a_j)|s_t = s\right]$, and the action-value function is $Q^M_{\mu,t}(s,a) = \bar{R}^M(s,a) + \sum_{s' \in \mathcal{S}} P^M_a(s'|s) V_{\mu,t+1}(s')$. At any fixed (s,t), $\mu(s,t) \equiv \arg\max_a Q_{\tilde{\mu},t}(s,a)$, note that we allow $\tilde{\mu}$ in the Q function to differ from μ. This distinction will be useful as $\tilde{\mu}$ can be μ, π (or the ESRL policy defined in Section 1.4). Finally, $\pi(a|s,t)$ is the probability of a given (s,t), under the behavior policy.

1.4 Expert-Supervised Reinforcement Learning

We are interested in finding a policy which improves upon π. Directly regularizing the target policy to the behavior might restrict the agent from finding optimal actions, especially when π has a high random component ϵ, or π^0 is not close to optimal. Thus we want to know when to use μ versus π. This motivates the use of posterior distributions to quantify how well each state has been explored in $\boldsymbol{D}_T$ and how close π is to π^0. At every state and time (s,t) in the episode we can sample K MDP models from $f(\cdot|\boldsymbol{D}_T)$. These samples are used to compare the quality of the behavior and target policy actions. We consider both the expected values of each action $Q_{\tilde{\mu},t}(s, \pi(s,t))$ versus $Q_{\tilde{\mu},t}(s, \mu(s,t))$, and their second moments for any fixed $\tilde{\mu}$. In particular, posterior distributions of $Q_{\tilde{\mu},t}(s,a)$, $a \in \mathcal{A}$ are used to test if the value for $\mu(s,t)$ is significantly better than π. This makes the learning process robust to the quality of the behavior policy. Next we formalize these arguments by a sampling scheme, define the ESRL policy, and state its theoretical properties.

Sampling Q functions. The distribution over the MDP model $f(\cdot|\boldsymbol{D}_T)$ implicitly defines a posterior distribution for any Q function: $Q_{\tilde{\mu},t}(s,a) \sim f_Q(\cdot|s,a,t,\boldsymbol{D}_T)$. As the true MDP model M^* is stochastic, we want to approximate the conditional mean Q value: $\mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)|s,a,t,\boldsymbol{D}_T\right]$. We do this by sampling K MDP models M_k, compute $Q^{(k)}_{\tilde{\mu},t}(s,a)$, $k = 1, \ldots, K$ and use $\hat{Q}_{\tilde{\mu},t}(s,a) \equiv \frac{1}{K}\sum_{k=1}^{K} Q^{(k)}_{\tilde{\mu},t}(s,a)$.

Lemma 1.4.1. *$\hat{Q}_{\tilde{\mu},t}(s,a)$ is consistent and unbiased for $Q^{M^*}_{\tilde{\mu},t}(s,a)$:*

$$\mathbb{E}\left[\hat{Q}_{\tilde{\mu},t}(s,a)|s,a,t,\boldsymbol{D}_T\right] = \mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)|s,a,t,\boldsymbol{D}_T\right],$$

$$\hat{Q}_{\tilde{\mu},t}(s,a) - \mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)|s,a,t,\boldsymbol{D}_T\right] = O_p\left(K^{-\frac{1}{2}}\right), \forall(t,s,a).$$

Lemma 1.4.1 establishes desirable properties for our Q function estimation. Choosing $K = 1$ yields an immediate result: every $Q^{(k)}_{\tilde{\mu},t}(s,a)$ from model M_k is unbiased.

The stochasticity of M^* and π suggests the mean Q values for π and μ are not enough to make a decision for whether it is beneficial to deviate from π. Next we discuss how to directly incorporate this uncertainty assessment into the policy training through Bayesian hypothesis testing.

ESRL Policy Learning Through Hypothesis Testing. For a fixed α-level, denote the ESRL policy by μ^α, we next describe the steps to learn this policy. By iterating backwards as in dynamic programming, assume we know $\mu^\alpha(s,j)\ \forall s \in \mathcal{S}, j \in \{t+1,\ldots,\tau\}$, and we have $V^M_{\mu^\alpha,\tau+1}(s) = 0, \forall s \in \mathcal{S}$. Intuitively, at any (s,t) we want to assess whether there is enough information in D_T to support choosing the seemingly best action μ over π. Denote $\mu(s,t) = \arg\max_a Q_{\mu^\alpha,t}(s,a)$ as the best action if we follow the ESRL policy μ^α onward, we formalize this with the following hypothesis:

$$H_0 : Q^M_{\mu^\alpha,t}(s,\mu(s,t)) \leq Q^M_{\mu^\alpha,t}(s,\pi(s,t)). \tag{1.1}$$

Note that in (1.1), both Q functions assume the agent proceeds with ESRL policy μ^α onward. If we can reject H_0, then it is safe to follow μ, if we fail to reject the null, it does not necessarily mean the behavior policy is better, but there is not enough information in the data to support following μ. To construct a safe ESRL policy we simply evaluate H_0 by computing the null probability $\mathbb{P}\left(H_0|t,s,\boldsymbol{D}_T\right)$, if this is below a pre-specified risk-aversion level α then we can safely choose μ. In other words if the learned policy does not yield a significantly better value estimate, then we fail to reject the null and proceed to use the behavior policy's action. The

ESRL policy at (s,t) is then

$$\mu^{\alpha}(s,t) = \begin{cases} \mu(s,t) & \text{if} \quad \mathbb{P}\left(H_0|t,s,\boldsymbol{D}_T\right) < \alpha, \\ \pi(s,t) & \text{else.} \end{cases}$$

To compute $\mu^{\alpha}(s,t)$, we start by sampling K MDP models from the posterior distribution, computing $\{Q^{(k)}_{\mu^{\alpha},t}(s,a)\}_{k=1}^{K}$ and splitting the samples into two disjoint sets $\mathcal{I}_1, \mathcal{I}_2$. We use $\mathcal{I}_1$ to draw the policy $\hat{\mu}(s,t)$ based on majority voting. Then we use $\mathcal{I}_2$ to assess the null hypothesis in (1.1), with estimator $\hat{\mathbb{P}}\left(H_0|t,s,\boldsymbol{D}_T\right) = \frac{1}{K}\sum_{k=1}^{K} I\left(Q^{(k)}_{\mu^{\alpha},t}(s,\hat{\mu}(s,t)) \leq Q^{(k)}_{\mu^{\alpha},t}(s,\pi(s,t))\right)$. We next discuss convergence of the null probability estimator, and how to choose $\hat{\mu}(s,t)\ \forall (s,t) \in \mathcal{S} \times \{1,\ldots,\tau\}$.

Lemma 1.4.2. *Let $\mathbb{P}^*\left(H_0|t,s,\boldsymbol{D}_T\right)$ be the null probability under true MDP M^* with policy μ^*,*

$$\hat{\mathbb{P}}\left(H_0|t,s,\boldsymbol{D}_T\right) - \mathbb{P}^*\left(H_0|t,s,\boldsymbol{D}_T\right) = O_p\left(K^{-\frac{1}{2}}\right).$$

Lemma 1.4.2 guarantees that we can construct a consistent policy μ^{α} by sampling from the MDP posterior. There are two factors that come into play in (1.1): the difference in mean Q values, and the second moments. If $Q^{M^*}_{\mu^{\alpha},t}(s,\mu(s,t))$ is much higher than $Q^{M^*}_{\mu^{\alpha},t}(s,\pi(s,t))$, but there are very few samples in $\boldsymbol{D}_T$ for $(s,\mu(s,t))$, the wide posterior will translate into a high $\hat{\mathbb{P}}\left(H_0|t,s,\boldsymbol{D}_T\right)$ leading ESRL to adopt $\pi(s,t)$. To choose $\mu(s,t)$ there needs to be both a substantial benefit for this new action and a high certainty of such gain. How averse the user is to deviating from π is controlled by parameter α. A small risk averse α will allow μ^{α} to deviate from π only with high certainty. When $\alpha = 1$, Algorithm 1 boils down to an offline version of PSRL after T episodes, which uses majority voting for a robust policy. Algorithm 1 collects these ideas in order to learn an ESRL policy μ^{α}. Disjoint sets $\mathcal{I}_1, \mathcal{I}_2$, ensure independence and keep theoretical guarantees under the Assumption 1.4.3.

Assumption 1.4.3. *Let $\mathbb{P}^*(H_0|s,t,\boldsymbol{D}_T)$ be defined as in (1.1) for the true M^*. The chosen risk-averse parameter $\alpha \in [0,1]$ satisfies $\mathbb{P}^*(H_0|s,t,\boldsymbol{D}_T) \neq \alpha\ \forall (s,t) \in \mathcal{S} \times \{1,\ldots,\tau\}$.*

As α is set by the user, Assumption 1.4.3 is easily satisfied as long as α is chosen carefully. Let $V^{M^*}_{\mu^{\alpha*},1}(s)$ be the value under the true MDP M^* and let $\mu^{\alpha*}$ be an ESRL policy which uses

Algorithm 1: Expert-Supervised RL

Sample $M_k \sim f(\cdot|\boldsymbol{D}_T)$ $k = 1, \ldots, K$, set $\mathcal{I}_1 = \{1, \ldots, \lceil \frac{K}{2} \rceil\}$, $\mathcal{I}_2 = \{\lceil \frac{K}{2} \rceil + 1, \ldots, K\}$;
Set $\hat{V}^{(k)}_{\tau+1}(s) \leftarrow 0$ $\forall s \in \mathcal{S}$, $k = 1, \ldots, K$;
Compute behavior distribution $\pi(a|s,t)$ from $\boldsymbol{D}_T$, set $\pi(s,t) = \arg\max_a \pi(a|s,t)$;
for $t = \tau, \ldots, 1$ **do**
  for $s \in \mathcal{S}$ **do**
    for $k = 1, \ldots, K$ **do**
      $\mu_k(s,t) \leftarrow \arg\max_a Q^{(k)}_{\mu^\alpha,t}(s,a)$;
    end
    $\hat{\mu}(s,t) \leftarrow$ maj. vote$\{\mu_k(s,t), k \in \mathcal{I}_1\}$;
    Compute $\hat{\mathbb{P}}(H_0|s,t,\boldsymbol{D}_T) = \frac{1}{|\mathcal{I}_2|} \sum_{k \in \mathcal{I}_2} I\left(Q^{(k)}_{\mu^\alpha,t}(s, \hat{\mu}(s,t)) < Q^{(k)}_{\mu^\alpha,t}(s, \pi(s,t))\right)$;
    for $k = 1, \ldots, K$ **do**
      $\mu^\alpha_k(s,t) \leftarrow I\left(\hat{\mathbb{P}}(H_0|s,t,\boldsymbol{D}_T) < \alpha\right) \mu_k(s,t) + I\left(\hat{\mathbb{P}}(H_0|s,t,\boldsymbol{D}_T) \geq \alpha\right) \pi(s,t)$;
      $\hat{V}^{(k)}_t(s) \leftarrow Q^{(k)}_{\mu^\alpha,t}(s, \mu^\alpha_k(s,t))$;
    end
    $\hat{\mu}^\alpha(s,t) \leftarrow$ maj. vote$\{\mu^\alpha_k(s,t), k \in \mathcal{I}_1\}$;
    $\mathcal{M}^\alpha(s,t) \leftarrow \{k | k \in \mathcal{I}_1, \mu^\alpha_k(s,t) = \hat{\mu}^\alpha(s,t)\}$;
  end
end
Define majority voting set: $\text{MV}^\alpha = \cap_{(s,t)} \mathcal{M}^\alpha(s,t)$;
if $\exists k \in MV^\alpha$ **then**
  choose $k \in \text{MV}^\alpha$ at random, set $k^{\text{MV}} \leftarrow k$
else
  Set k^{MV} to most common $k \in \mathcal{M}^\alpha(s,t), \forall (s,t)$
end
Set $\mu^\alpha = \mu_{k^{MV}}$

the null hypotheses in (1.1) defined under M^*. Then, for episode i we can define the regret for μ^α from Algorithm 1 as $\Delta_i = \sum_{s_i \in \mathcal{S}} P_0(s_i) \left(V^{M^*}_{\mu^{\alpha *},1}(s_i) - V^{M^*}_{\mu^\alpha,1}(s_i) \right)$, and the expected regret after T episodes as $\mathbb{E}[Regret(T)] = \mathbb{E}\left[\sum_{i=1}^{T} \Delta_i\right]$.

Theorem 1.4.4 (Regret Bound for ESRL)**.** *For any $\alpha \in [0,1]$ which satisfies Assumption 1.4.3, Algorithm 1 using $\boldsymbol{D}_T$ and choosing $K = \mathcal{O}(T)$ will yield*

$$\mathbb{E}\left[Regret\left(T\right)\right] = \mathcal{O}\left(\tau S \sqrt{AT \log(SAT)}\right).$$

Theorem 1.4.4 shows ESRL is sample efficient, flexible to risk aversion level α, and robust to the quality of behavior policy π. As the regret bound is true for any level of risk aversion α, Algorithm 1 universally converges to the oracle. This makes ESRL flexible for a wide range of applications. It also shows that ESRL is suitable to a large class of models, as the regret bound does not impose a specific form on f. Regarding access to $f(\cdot|\mathbf{D}_T)$ for sampling MDPs in real-world problems, as data increases, dependency of results on the prior decreases, so we can use any *working model* to approximate the MDP. Several models are computationally simple to sample from, and can be used for learning. For example, we use the Dirichlet/multinomial, and normal-gamma/normal conjugates for P^M and R^M respectively, which work well for all simulation and real data settings explored in Section 1.6. In fact, if a Dirichlet prior over the transitions is assumed, the regret bound in Theorem 1.4.4 can be improved. Chosen priors should be flexible enough to capture the dynamics and easy to sample from efficiently. Next we consider how to discern whether ESRL, or any other fixed policy, is an improvement on the behavior policy.

1.5 Off-Policy Policy Evaluation and Uncertainty Estimation

We now illustrate how the ESRL framework can be used to construct efficient point estimates of the value function, and their posterior distributions. Hypothesis testing can also be used to assess whether the difference in value of two policies is statistically significant (i.e. μ^α vs. π).

To compute the estimated value of a given policy $\tilde{\mu}$, we sample K models from the posterior

and navigate M_k using $\tilde{\mu}$. This yields samples $V_{\tilde{\mu},1}^{(k)} \sim f_V(\cdot|\boldsymbol{D}_T)$. We estimate $\mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right]$ with $\hat{V}_{\tilde{\mu}} = \frac{1}{K}\sum_{k=1}^{K} V_{\tilde{\mu},1}^{(k)}$. Note that we average over the initial states as well, as we are interested to know the marginal value of the policy. A conditional value of the policy function $V_{\tilde{\mu},1}^{M^*}(s_0)$ can also be computed simply by starting all samples at a fixed state s_0.

Theorem 1.5.1. *Let $\tilde{\mu} : \mathcal{S} \times \{1, \ldots, \tau\} \mapsto \mathcal{A}$ be a pre-specified policy,*

$$\mathbb{E}\left[\hat{V}_{\tilde{\mu}}\middle|\boldsymbol{D}_T\right] = \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)\middle|\boldsymbol{D}_T\right], \hat{V}_{\tilde{\mu}} - \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)\middle|\boldsymbol{D}_T\right] = O_p\left(K^{-\frac{1}{2}}\right).$$

Theorem 1.5.1 ensures that we are indeed estimating the quantity of interest. It establishes that $\hat{V}_{\tilde{\mu}}$ is consistent and unbiased for $\sum_{s\in\mathcal{S}} P_0(s)V_{\tilde{\mu},1}^{M^*}(s)$. As MDP M^* is stochastic, point estimates without measures of uncertainty are not sufficient to evaluate the quality of a policy. For example in an application such as healthcare, there might be policies for which the second best action (treatment) is not significantly different in terms of value, but has less associated secondary risks. Including a secondary risk directly into the method might force us to make strong modeling assumptions. Therefore, testing whether such policies yield a statistically significant difference in value is important. With this information, one can devise a policy that always chooses the safest action (e.g. in clinical terms) and if this yields an equivalent value to the optimal policy, then it is preferable.

Policy-level hypothesis testing. Define the value function null hypothesis for two fixed policies $\tilde{\mu}_1, \tilde{\mu}_2$ as the event in which policy $\tilde{\mu}_1$ has a higher expected value than $\tilde{\mu}_2$ conditional on $\boldsymbol{D}_T$: $H_0 : \mathbb{E}_{s\sim P_0,M^*}[V_{\tilde{\mu}_1,1}(s)|\boldsymbol{D}_T] > \mathbb{E}_{s\sim P_0,M^*}[V_{\tilde{\mu}_2,1}(s)|\boldsymbol{D}_T]$. The probability of the null under the true model M^* is

$$\mathbb{P}_\mu(H_0|\boldsymbol{D}_T) = \sum_{s\in\mathcal{S}} P_0(s)\mathbb{P}\left(V_{\tilde{\mu}_1,1}^{M^*}(s) > V_{\tilde{\mu}_2,1}^{M^*}(s)\middle|s, \boldsymbol{D}_T\right).$$

We use samples $V_{\tilde{\mu}_\ell}^{(k)}$, $\ell = 1, 2$ to estimate the probability of the null with $\hat{\mathbb{P}}_\mu(H_0|\boldsymbol{D}_T) = \frac{1}{K}\sum_{k=1}^{K} I\left(V_{\tilde{\mu}_1,1}^{(k)}(s) > V_{\tilde{\mu}_2,1}^{(k)}(s)\right)$. Consistency of this estimator is shown in the Appendix A.3.2.

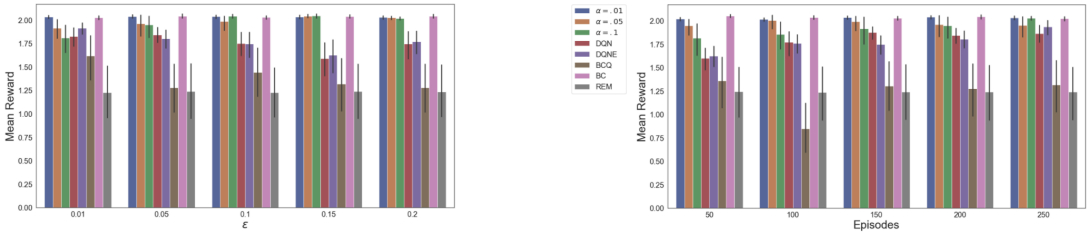
1.6 Experiments and Application

We perform several analyses to assess ESRL policy learning, sensitivity to the risk aversion parameter α, value function estimation, and finally illustrate how we can interpret the posteriors within the context of the application. The code for implementing ESRL with detailed comments is publicly available[1]. We use the Riverswim environment [32], and a Sepsis data set built from MIMIC-III data [33]. We compare ESRL to several methods: a) a *naive* baseline made from an ensemble of K DQN models (DQNE), where we simply use the mean for selecting actions, this benchmark is meant to shed light into the empirical benefit of the hypothesis testing in ESRL. b) We argue ESRL can deviate from the behavior policy when allowed by the hypothesis testing, for further investigating the benefit of hypothesis testing, we implement behavior cloning (BC). c) We explore Batch Constrained deep Q-learning (BCQ) which uses regularization towards the behavior policy for offline RL [25, 27, 34]. d) Finally, we also implement a strong benchmark which leverages ensembles and uncertainty estimation in the context of offline RL using random ensembles (REM) [13]. For Riverswim we use 2-128 unit layers, for Sepsis we use 128, 256 unit layers respectively [35]. For ESRL, we use conjugate Dirichlet/multinomial, and normal-gamma/normal for the prior and likelihood of the transition and reward functions respectively.

1.6.1 Riverswim

The Riverswim environment [32] requires deep exploration for achieving high rewards. There are 6 states and two actions: swim right or left. Only swimming left is always successful. There are only two ways to obtain rewards: swimming left while in the far left state will yield a small reward (5/1000) w.p. 1, swimming right in the far right state will yield a reward of 1 w.p. 0.6. The episode lasts 20 time points. We train policy π^0 using PSRL [23] for 10,000 episodes, we then generate data set $\boldsymbol{D}_T$ with π, varying both size T and noise ϵ. The offline trained policies are then tested on the environment for 10,000 episodes. This process is repeated 50 times.

[1] https://github.com/asonabend/ESRL



(a) *Mean reward for T=200 episodes, while varying ϵ-greedy behavior policy π.*

(b) *Mean reward for a $\epsilon = 0.05$ in the behavior policy, while varying number of episodes T in $\boldsymbol{D}_T$.*

Figure 1.1: *Mean test reward per episode for policies trained offline with ESRL ($\alpha = 0.01, 0.05, 0.1$), DQN, DQNE, BC, BCQ, and REM on Riverswim. Optimal policy expected reward is 2.*

Policy Learning. We first assess ESRL on Riverswim. The training set sample size T is kept low to make it hard to completely learn the dynamics of the environment. We train an offline policy using ESRL with different risk aversion parameters ($\alpha = 0.01, 0.05, 0.1$). Figure 1.1 (a) shows mean reward for $T = 200$ episodes while varying ϵ. ESRL proves to be robust to the behavior policy quality. This is expected as when ϵ is low the environment is not fully explored. This yields high variance in the Q posteriors, which leads ESRL to reject the null more often and favor the behavior policy. For low quality data generating policies there is greater exploration of the environment, which yields narrower posterior distributions for the Q function posteriors, leading ESRL to reject the null when it is indeed beneficial to do so. When behavior policy is almost deterministic, the smaller risk aversion parameter α seems to yield good results as ESRL almost always imitates the behavior policy. BC does well as it seems to estimate the expert behavior well enough regardless of the noise level. Overall Q-learning methods lack enough data to learn a good policy. Figure 1.1 (b) compares methods on an almost constant behavior policy ($\epsilon = 0.05$), so there is little exploration in $\boldsymbol{D}_T$. ESRL is robust as wide posteriors keep it from deviating from π. Methods other than BC generally fail likely to lack of exploration in $\boldsymbol{D}_T$. However note that in real world data π^0 is not necessarily optimal, in which case BC will likely not perform very well relative to ESRL or others if there is a high-noise expert policy, which yields a well explored MDP, this is the case in the Sepsis results shown in Figure 1.4 (c). Finally it's worth noting that REM does better than DQNE in Riverswim but not on Sepsis, we believe this is because the DQN neural networks are smaller,

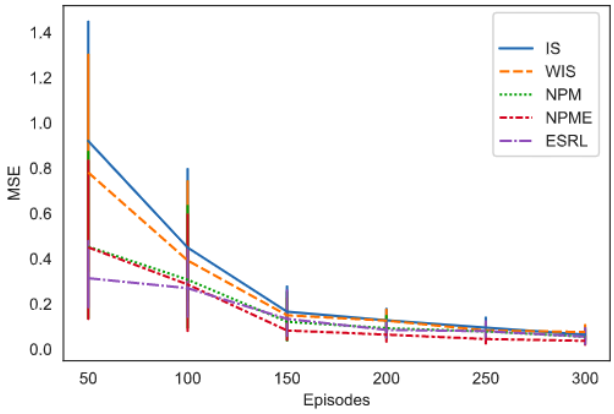
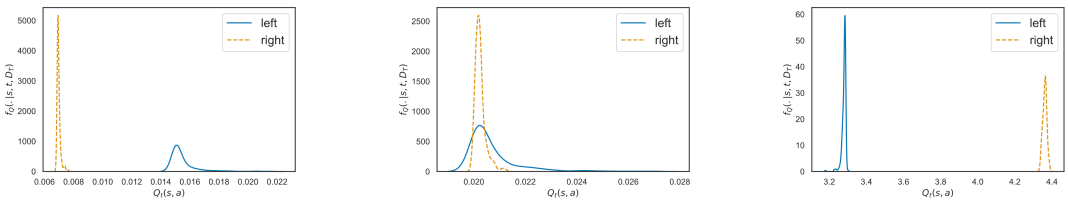


Figure 1.2: *Mean squared error and 95% confidence bands for OPPE of an ESRL policy. We compare step importance sampling (IS), step weighted IS (WIS), a non parametric model (NPM), an NPM ensemble (NPME) and ESRL estimation.*

REM outperforms DQNE in a more complex and higher variance setting with more training data such as the Sepsis setting in Section 1.6.2.

Figure 1.2 shows Mean Squared Error (MSE) and 95% confidence bands for value estimation of an ESRL policy using $\boldsymbol{D}_T$ while varying T. We compare it with sample-based estimates: step importance sampling (IS), and step weighted IS (WIS), and model based estimates which use a full non parametric model (NPM), and an NPM ensemble (NPME). The non parametric models compute the rewards and transition probability tables based on observed counts. The policy is evaluated by using the tables as an MDP model where states are drawn using the estimated transition probability matrix. NPM uses 1000 episodes to evaluate a policy, NPME is an average over 100 NPM estimates. In small data sets ESRL performs substantially better as it uses the model posteriors to overcome rarely visited states in $\boldsymbol{D}_T$. Eventually the priors (which are miss-specified for some state-action pairs) loose importance and ESRL converges to the non-parametric estimates. Sample based estimates are consistently less efficient but converge to the true policy with enough data.

Hypothesis testing and interpretability with Q function posterior distributions. We illustrate interpretability of the ESRL method in Riverswim as it is a simple, intuitive setting. Figure 1.3 shows 3 Q function posterior distributions $f_Q(\cdot|s,t,\boldsymbol{D}_T)$, each for a fixed



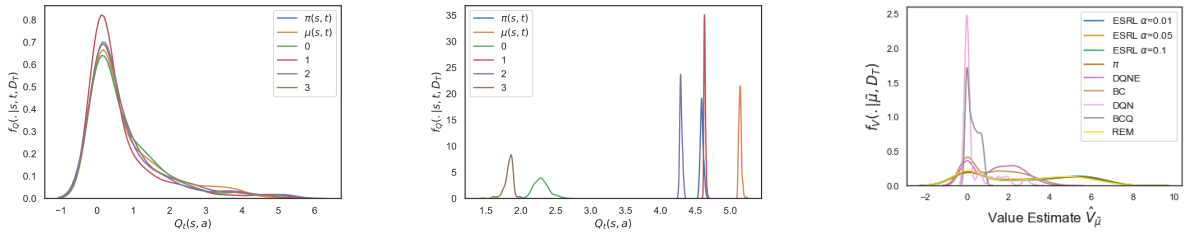
(a) $f_Q(\cdot|s=0,t=17,\boldsymbol{D}_T)$ **(b)** $f_Q(\cdot|s=2,t=2,\boldsymbol{D}_T)$ **(c)** $f_Q(\cdot|s=5,t=5,\boldsymbol{D}_T)$

Figure 1.3: *Posterior distributions of* $Q_t(s,a)$ *functions for fixed* (s,t)*. We use* $K = 250$ *MDP samples. Observed data,* $\boldsymbol{D}_T$ *has* $T = 1000$ *episodes, generated with* $\epsilon = 0.2$*.*

state-time pair (s,t). Display (a) shows Q functions for the far left state and an advanced time point $t = 17$. There is high certainty (no overlap in posteriors) that swimming left will yield a higher reward, as left is successful w.p. 1. $Q_{17}(0, \textit{left})$ has a wider posterior as this (s,a) is not common in $\boldsymbol{D}_T$. Display (b) is the most interesting, it sheds light into the utility of uncertainty measures. A naive RL method that only considers mean values, would choose the optimal action according to μ: swimming left. However, there is high uncertainty associated with such a choice. In fact, we know that the optimal strategy in Riverswim is $\pi(2,2) = \textit{right}$, hypothesis testing will fail to reject the null and use the behavior action which will lead to a higher expected reward. Display (c) shows Q posteriors for the state furthest to the right, at $t = 5$. Choosing right will be successful with high certainty: narrow $Q_5(5, \textit{left})$ posterior. Swimming left will still yield a relatively high reward as in the next time point the agent will proceed with the optimal policy (choosing right). As there is no overlap in (a) and (c), the best choice is clear as would be reflected with a hypothesis test.

1.6.2 Sepsis.

We further test ESRL on a Sepsis data set built from MIMIC-III [33]. Sepsis is a state of infection where the immune system gets overwhelmed and can cause tissue damage, organ failure, and death. Deciding treatments and medication dosage is a dynamic and highly challenging task for the clinicians. We consider an action space representing dosage of intravenous fluids for hypovolemia (IV fluids) and vasopressors to counteract vasodilation. The action space $\mathcal{A}$ is size 25: a 5×5 matrix over discretized dose of vasopressors and IV



(a) *Q function posterior distributions for* $(s,t) = (90,7)$, $a \in \{0,1,2,3,\pi(90,7),\mu(90,7)\}$.

(b) *Q function posterior distributions for* $(s,t) = (5,8)$, $a \in \{0,1,2,3,\pi(5,8),\mu(5,8)\}$.

(c) *Posterior distribution for* $\hat{V}_{\hat{\mu}}$, *for* π, *ESRL, BC, BCQ, DQN, DQNE, REM.*

Figure 1.4: *Display (a) & (b) show posterior distributions of Q functions at fixed* (s,t). *Display (c) shows posteriors* $\hat{V}$ *for policies:* π *and* μ^α *for* $\alpha = 0.01, 0.05, 0.1$, *DQN, DQNE, BC, BCQ and REM for* $K = 500$.

fluids. The state space is composed of 1,000 clusters estimated using K-means on a 46-feature space which contains measures of the patient's physiological state. We used negative SOFA score as a reward [35], we transform it to be between 0 and 1. The data set used has 12,991 episodes of 10 time steps- measurements every 4-hour interval. We used 80% of episodes for training and 20% for testing.

Figure 1.4 (a) & (b) show posterior distributions for two different (s,t) pairs in the Sepsis data set hand-picked to illustrate interpretability. For simplicity we restrict to show the best action: $\mu(s,t)$, physician's action $\pi(s,t)$, and four other low dose actions $a \in \{0,\ldots,3\}$. Display (a) shows posterior distributions over a state rarely observed in $\boldsymbol{D}_T$, hence the Q functions have relatively high standard errors. The expected cumulative inverse SOFA value for this state seems to be relatively stable no matter what action is taken. The Q posteriors for μ and π are practically overlapping so there's no reason to deviate from π, this is encoded into μ^α through hypothesis testing. Interpertability is useful in these cases, as a physician might see there is no difference in actions: all will yield similar SOFA scores. Therefore, an action can be chosen to lower risk of side effects. Display (b) on the other hand shows a common state in $\boldsymbol{D}_T$: the low standard errors allow the policy to deviate from π at any α level. Within this state, actions π and μ are usually selected so the posteriors for their Q functions are narrow, as opposed to those for $a = 0, 3$. These actions are not prevalent in $\boldsymbol{D}_T$ as they seem to be sub-optimal, so they are less often chosen by doctors and seen in $\boldsymbol{D}_T$.

Figure 1.4 (c) shows the posterior distribution of the Sepsis value function for different policies. There seems to be a bi-modal distribution: it is easier to control the SOFA scores for patients in the set of states shown in the right mode of the distribution. Physicians know how to do this well as shown by the posterior value function for π; and ESRL picks up on this. The other clusters of states in the left mode seem to be harder to control. We can appreciate how deviating from the physician's policy is strikingly damaging to the expected value on the test set. DQN and BCQ, DQNE and BC generalize better but under preform relative to ESRL and REM. The $\boldsymbol{D}_T$ is probably not enough to generalize to the test set due to the high dimensional state and action spaces. ESRL through hypothesis testing captures this and hardly deviates from the behavior policy. Thus, it is clear that we cannot do better than π given the information in the data, but the posterior suggest the need to learn safe policies as we can do substantially worse with methods that don't account for uncertainty and safety.

1.7 Conclusion

We propose an Expert-Supervised RL (ESRL) approach for offline learning based on Bayesian RL. This framework can learn safe policies from observed data sets. It accounts for uncertainty in the MDP and data logging process to assess when it is safe and beneficial to deviate from the behavior policy. ESRL allows for different levels of risk aversion, which are chosen within the application context. We show a $\tilde{\mathcal{O}}(\tau S\sqrt{AT})$ Bayesian regret bound that is independent of the risk aversion level tailored to the environment and noise level in the data set. The ESRL framework can be used to obtain interpretable posterior distributions for the Q functions and for OPPE. These posteriors are flexible to account for any possible policy function and are amenable to interpretation within the context of the application. An important limitation of ESRL is that it cannot readily handle continuous state spaces which are common in real world applications. Another extension we are interested in is in exploring the comparison of credible intervals as opposed to the null probability estimates. We believe ESRL is a step towards bridging the gap between RL research and real-world applications.

Chapter 2

Semi-Supervised Off Policy Reinforcement Learning

Aarón Sonabend[1], Nilanjana Laha[1], Ashwin N. Ananthakrishnan[2], Tianxi Cai[1] and Rajarshi Mukherjee[1]

[1] Department of Biostatistics

Harvard University

[2] Division of Gastroenterology

Massachusetts General Hospital

2.1 Summary

Reinforcement learning (RL) has shown great success in estimating sequential treatment strategies which take into account patient heterogeneity. However, health-outcome information, which is used as the reward for reinforcement learning methods, is often not well coded but rather embedded in clinical notes. Extracting precise outcome information is a resource intensive task, so most of the available well-annotated cohorts are small. To address this issue, we propose a semi-supervised learning (SSL) approach that efficiently leverages a small sized labeled data with true outcome observed, and a large unlabeled data with outcome

surrogates. In particular, we propose a semi-supervised, efficient approach to Q-learning and doubly robust off policy value estimation. Generalizing SSL to sequential treatment regimes brings interesting challenges: 1) Feature distribution for Q-learning is unknown as it includes previous outcomes. 2) The surrogate variables we leverage in the modified SSL framework are predictive of the outcome but not informative to the optimal policy or value function. We provide theoretical results for our Q-function and value function estimators to understand to what degree efficiency can be gained from SSL. Our method is at least as efficient as the supervised approach, and moreover safe as it robust to mis-specification of the imputation models.

2.2 Introduction

Finding optimal treatment strategies that can incorporate patient heterogeneity is a cornerstone of personalized medicine. When treatment options change over time, optimal sequential treatment rules (STR) can be learned using longitudinal patient data. With increasing availability of large-scale longitudinal data such as electronic health records (EHR) data in recent years, reinforcement learning (RL) has found much success in estimating such optimal STR [36]. Existing RL methods include G-estimation [37], Q-learning [2, 38], A-learning [39] and directly maximizing the value function [40]. Both G-estimation and A-learning attempt to model only the component of the outcome regression relevant to the treatment contrast, while Q-learning posits complete models for the outcome regression. Although G-estimation and A-learning models can be more efficient and robust to mis-specification, Q-learning is widely adopted due to its ease of implementation, flexibility and interpretability [2, 3, 4].

Learning STR with EHR data, however, often faces an additional challenge of whether outcome information is readily available. Outcome information, such as development of a clinical event or whether a patient is considered as a responder, is often not well coded but rather embedded in clinical notes. Proxy variables such as diagnostic codes or mentions of relevant clinical terms in clinical notes via natural language processing (NLP), while predictive of the true outcome, are often not sufficiently accurate to be used directly in place of the

outcome [41, 42]. On the other hand, extracting precise outcome information often requires manual chart review, which is resource intensive, particularly when the outcome needs to be annotated over time. This indicates the need for a semi-supervised learning (SSL) approach that can efficiently leverage a small sized labeled data $\mathcal{L}$ with true outcome observed and a large sized unlabeled data $\mathcal{U}$ for predictive modeling. It is worthwhile to note that the SSL setting differs from the standard missing data setting in that the probability of missing tends to 1 asymptotically, which violates the positivity assumption required by the classical missing data methods [43].

While SSL methods have been well developed for prediction, classification and regression tasks [e.g. 43, 44, 45, 46, 47, 48], there is a paucity of literature on SSL methods for estimating optimal treatment rules. Recently, [42] and [49] proposed SSL methods for estimating an average causal treatment effect. [50] proposed a semi-supervised RL method which achieves impressive empirical results and outperforms simple approaches such as direct imputation of the reward. However, there are no theoretical guarantees and the approach lacks causal validity and interpretability within a domain context. Additionally, this method does not leverage available surrogates. In this paper, we fill this gap by proposing a theoretically justified SSL approach to Q-learning using a large unlabeled data $\mathcal{U}$ which contains sequential observations on features $\mathbf{O}$, treatment assignment A, and surrogates $\mathbf{W}$ that are imperfect proxies of Y, as well as a small set of labeled data $\mathcal{L}$ which contains true outcome Y at multiple stages along with $\mathbf{O}$, A and $\mathbf{W}$. We will also develop robust and efficient SSL approach to estimating the value function of the derived optimal STR, defined as the expected counterfactual outcome under the derived STR.

To describe the main contributions of our proposed SSL approach to RL, we first note two important distinctions between the proposed framework and classical SSL methods. First, existing SSL literature often assumes that $\mathcal{U}$ is large enough that the feature distribution is known [51]. However, under the RL setting, the outcome of the stage $t-1$, denoted by Y_{t-1}, becomes a feature of stage t for predicting Y_t. As such, the feature distribution for predicting Y_t can not be viewed as known in the Q-learning procedure. Our methods for estimating an

optimal STR and its associated value function, carefully adapt to this sequentially missing data structure. Second, we modify the SSL framework to handle the use of surrogate variables $\mathbf{W}$ which are predictive of the outcome through the joint law $\mathbb{P}_{Y,\mathbf{O},A,\mathbf{W}}$, but are not part of the conditional distribution of interest $\mathbb{P}_{Y|\mathbf{O},A}$. To address these issues, we propose a two-step fitting procedure for finding an optimal STR and for estimating its value function in the SSL setting. Our method consists of using the outcome-surrogates ($\mathbf{W}$) and features ($\mathbf{O}, A$) for non-parametric estimation of the missing outcomes (Y). We subsequently use these imputations to estimate Q functions, learn the optimal treatment rule and estimate its associated value function. We provide theoretical results to understand when and to what degree efficiency can be gained from $\mathbf{W}$ and $\mathbf{O}, A$.

We further show that our approach is robust to mis-specification of the imputation models. To account for potential mis-specification in the models for the Q function, we provide a double robust value function estimator for the derived STR. If either the regression models for the Q functions or the propensity score functions are correctly specified, our value function estimators are consistent for the true value function.

We organize the rest of the paper as follows. In Section 2.3 we formalize the problem mathematically and provide some notation to be used in the development and analysis of the methods. In Section 2.4 we discuss traditional Q-learning and propose an SSL estimation procedure for the optimal STR. Section 2.5 details an SSL doubly robust estimator of the value function for the derived STR. In Section 2.6 we provide theoretical guarantees for our approach and discuss implications of our assumptions and results. Section 2.7 is devoted for numerical experiments as well as real data analysis with an inflammatory bowel disease (IBD) data-set. We end with a discussion of the methods and possible extensions in Section 3.6[1]. Finally all the technical proofs and supporting lemmas are collected in Appendices B.2 and B.3.

[1]The proposed method has been implemented in R and the code can be found at `github.com/asonabend/SSOPRL`.

2.3 Problem setup

We consider a longitudinal observational study with outcomes, confounders and treatment indices potentially available over multiple stages. Although our method is generalizable for any number of stages, for ease of presentation we will use two time points of (binary) treatment allocation as follows. For time point $t \in \{1, 2\}$, let $\mathbf{O}_t \in \mathbb{R}^{d_t^o}$ denote the vector of covariates measured prior to stage t of dimension d_t^o; $A_t \in \{0, 1\}$ a treatment indicator variable; and $Y_{t+1} \in \mathbb{R}$ the outcome observed at stage $t + 1$, for which higher values of Y_{t+1} are considered beneficial. Additionally we observe surrogates $\mathbf{W}_t \in \mathbb{R}^{d_t^\omega}$, a d_t^ω-dimensional vector of post-treatment covariates potentially predictive of Y_{t+1}. In the labeled data where $\mathbf{Y} = (Y_2, Y_3)^\intercal$ is annotated, we observe a random sample of n independent and identically distributed (iid) random vectors, denoted by

$$\mathcal{L} = \{\mathbf{L}_i = (\vec{\mathbf{U}}_i^\intercal, \mathbf{Y}_i^\intercal)^\intercal\}_{i=1}^n, \quad \text{where } \mathbf{U}_{ti} = (\mathbf{O}_{ti}^\intercal, A_{ti}, \mathbf{W}_{ti}^\intercal)^\intercal \text{ and } \vec{\mathbf{U}}_i = (\mathbf{U}_{1i}^\intercal, \mathbf{U}_{2i}^\intercal)^\intercal.$$

We additionally observe an unlabeled set consisting of N iid random vectors,

$$\mathcal{U} = \{\vec{\mathbf{U}}_j\}_{j=1}^N$$

with $N \gg n$. We denote the entire data as $\mathbb{S} = (\mathcal{L} \cup \mathcal{U})$. To operationalize our statistical arguments we denote the joint distribution of the observation vector $\mathbf{L}_i$ in $\mathcal{L}$ as $\mathbb{P}$. In order to connect to the unlabeled set, we assume that any observation vector $\vec{\mathbf{U}}_j$ in $\mathcal{U}$ has the distribution induced by $\mathbb{P}$.

We are interested in finding the optimal STR and estimating its *value function* to be defined as expected counterfactual outcomes under the derived regime. To this end, let $Y_{t+1}^{(a)}$ be the potential outcome for a patient at time $t + 1$ had the patient been assigned at time t to treatment $a \in \{0, 1\}$. A dynamic treatment regime is a set of functions $\mathcal{D} = (d_1, d_2)$, where $d_t(\cdot) \in \{0, 1\}$, $t = 1, 2$ map from the patient's history up to time t to the treatment choice $\{0, 1\}$. We define the patient's history as $\mathbf{H}_1 \equiv [\mathbf{H}_{10}^\intercal, \mathbf{H}_{11}^\intercal]^\intercal$ with $\mathbf{H}_{1k} = \boldsymbol{\phi}_{1k}(\mathbf{O}_1)$, $\mathbf{H}_2 = [\mathbf{H}_{20}^\intercal, \mathbf{H}_{21}^\intercal]^\intercal$ with $\mathbf{H}_{2k} = \boldsymbol{\phi}_{2k}(\mathbf{O}_1, A_1, \mathbf{O}_2)$, where $\{\boldsymbol{\phi}_{tk}(\cdot), t = 1, 2, k = 0, 1\}$ are pre-specified basis functions. We then define features derived from patient history for regression

modeling as $\bar{X}_1 \equiv [\mathbf{H}_{10}^\intercal, A_1\mathbf{H}_{11}^\intercal]^\intercal$ and $\bar{X}_2 \equiv [\mathbf{H}_{20}^\intercal, A_2\mathbf{H}_{21}^\intercal]^\intercal$. For ease of presentation, we also let $\check{\mathbf{H}}_1 = \mathbf{H}_1^\intercal$, $\check{\mathbf{H}}_2 = (Y_2, \mathbf{H}_2^\intercal)^\intercal$, $\check{\bar{X}}_1 = \bar{X}_1$, $\check{\bar{X}}_2 = (Y_2, \bar{X}_2^\intercal)^\intercal$, and $\mathbf{\Sigma}_t = \mathbb{E}[\check{\bar{X}}_t\check{\bar{X}}_t^\intercal]$.

Let $\mathbb{E}_{\mathcal{D}}$ be the expectation with respect to the measure that generated the data under regime $\mathcal{D}$. Then these sets of rules $\mathcal{D}$ have an associated value function which we can write as $V(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}\left[Y_2^{(d_1)} + Y_3^{(d_2)}\right]$. Thus, an optimal dynamic treatment regime is a rule $\bar{\mathcal{D}} = (\bar{d}_1, \bar{d}_2)$ such that $\bar{V} = V\left(\bar{\mathcal{D}}\right) \geq V\left(\mathcal{D}\right)$ for all $\mathcal{D}$ in a suitable class of admissible decisions [3]. To identify $\bar{\mathcal{D}}$ and $\bar{V}$ from the observed data we will require the following sets of standard assumptions [4, 52]: (i) consistency – $Y_{t+1} = Y_{t+1}^{(0)}I(A_t = 0) + Y_{t+1}^{(1)}I(A_t = 1)$ for $t = 1, 2$, (ii) no unmeasured confounding – $Y_{t+1}^{(0)}, Y_{t+1}^{(1)} \perp\!\!\!\perp A_t|\mathbf{H}_t$ for $t = 1, 2$ and (iii) positivity – $\mathbb{P}(A_t|\mathbf{H}_t) > \nu$, for $t = 1, 2$, $A_t \in \{0, 1\}$, for some fixed $\nu > 0$.

We will develop SSL inference methods to derive optimal STR $\bar{\mathcal{D}}$ as well the associated value function $\bar{V}$ by leveraging the richness of the unlabeled data and the predictive power of surrogate variables which allows us to gain crucial statistical efficiency. Our main contributions in this regard can be described as follows. First, we provide a systematic generalization of the Q-learning framework with theoretical guarantees to the semi-supervised setting with improved efficiency. Second, we provide a doubly robust estimator of the value function in the semi-supervised setup. Third, our Q-learning procedure and value function estimator are flexible enough to allow for standard off-the-shelf machine learning tools and are shown to perform well in finite-sample numerical examples.

2.4 Semi-Supervised Q-learning

In this section we propose a semi-supervised Q-learning approach to deriving an optimal STR. To this end, we first recall the basic mechanism of traditional linear parametric Q-learning [3] and then detail our proposed method. We defer the theoretical guarantees to Section 2.6.

2.4.1 Traditional Q-learning

Q-learning is a backward recursive algorithm that identifies optimal STR by optimizing two stage Q-functions defined as:

$$Q_2(\check{\mathbf{H}}_2, A_2) \equiv \mathbb{E}[Y_3|\check{\mathbf{H}}_2, A_2], \quad \text{and} \quad Q_1(\check{\mathbf{H}}_1, A_1) \equiv \mathbb{E}[Y_2 + \max_{a_2} Q_2(\check{\mathbf{H}}_2, a_2)|\check{\mathbf{H}}_1, A_1]$$

[38, 53]. In order to perform inference one typically proceeds by positing models for the Q functions. In its simplest form one assumes a (working) linear model for some parameters $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t^\intercal, \boldsymbol{\gamma}_t^\intercal)^\intercal$, $t = 1, 2$, as follows:

$$\begin{aligned} Q_1(\check{\mathbf{H}}_1, A_1; \boldsymbol{\theta}_1^0) =& \check{\tilde{X}}_1^\intercal \boldsymbol{\theta}_1^0 = \mathbf{H}_{10}^\intercal \boldsymbol{\beta}_1^0 + A_1(\mathbf{H}_{11}^\intercal \boldsymbol{\gamma}_1^0), \\ Q_2(\check{\mathbf{H}}_2, A_2; \boldsymbol{\theta}_2^0) =& \check{\tilde{X}}_2^\intercal \boldsymbol{\theta}_2^0 = Y_2 \beta_{21}^0 + \mathbf{H}_{20}^\intercal \boldsymbol{\beta}_{22}^0 + A_2(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2^0). \end{aligned} \tag{2.1}$$

Typical Q-learning consists of performing a least squares regression for the second stage to estimate $\widehat{\boldsymbol{\theta}}_2$ followed by defining the stage 1 pseudo-outcome for $i = 1, ..., n$ as

$$\widehat{Y}_{2i}^* = Y_{2i} + \max_{a_2} Q_2(\check{\mathbf{H}}_{2i}, a_2; \widehat{\boldsymbol{\theta}}_2) = Y_{2i}(1 + \hat{\beta}_{21}) + \mathbf{H}_{20i}^\intercal \widehat{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21i}^\intercal \widehat{\boldsymbol{\gamma}}_2]_+,$$

where $[x]_+ = xI(x > 0)$. One then proceeds to estimate $\widehat{\boldsymbol{\theta}}_1$ using least squares again, with $\widehat{Y}_2^*$ as the outcome variable. Indeed, valid inference on $\bar{\mathcal{D}}$ using the method described above crucially depends on the validity of the model assumed. However as we shall see, even without validity of this model we will be able to provide valid inference on suitable analogues of the Q-function working model parameters, and on the value function using a double robust type estimator. To that end it will be instructive to define the least square projections of Y_3 and Y_2^* onto $\check{\tilde{X}}_2$ and $\check{\tilde{X}}_1$ respectively. The linear regression working models given by (2.1) have $\boldsymbol{\theta}_1^0$, $\boldsymbol{\theta}_2^0$ as unknown regression parameters. To account for the potential mis-specification of the working models in (2.1), we define the target population parameters $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2$ as the population solutions to the expected normal equations

$$\mathbb{E}\left\{\check{\tilde{X}}_1(\bar{Y}_2^* - \check{\tilde{X}}_1^\intercal \bar{\boldsymbol{\theta}}_1)\right\} = \mathbf{0}, \quad \text{and} \quad \mathbb{E}\left\{\check{\tilde{X}}_2^\intercal \left(Y_3 - \check{\tilde{X}}_2^\intercal \bar{\boldsymbol{\theta}}_2\right)\right\} = \mathbf{0},$$

where $\bar{Y}_2^* = Y_2 + \max_{a_2} Q_2(\check{\mathbf{H}}_2, a_2; \bar{\boldsymbol{\theta}}_2)$. As these are linear in the parameters, uniqueness and existence for $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2$ are well defined. In fact, $Q_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\theta}}_1) = \check{\bar{X}}_1^\intercal \bar{\boldsymbol{\theta}}_1, Q_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\theta}}_2) = \check{\bar{X}}_2^\intercal \bar{\boldsymbol{\theta}}_2$ are the L_2 projection of $\mathbb{E}(Y_2^*|\check{\bar{X}}_1) \in \mathcal{L}_2\left(\mathbb{P}_{\check{\bar{X}}_1}\right)$, $\mathbb{E}(Y_3|\check{\bar{X}}_2) \in \mathcal{L}_2\left(\mathbb{P}_{\check{\bar{X}}_2}\right)$ onto the subspace of all linear functions of $\check{\bar{X}}_1, \check{\bar{X}}_2^\intercal$ respectively. Therefore, Q functions in (2.1) are the best linear predictors of $\bar{Y}_2^*$ conditional on $\check{\bar{X}}_1$ and Y_3 conditional on $\check{\bar{X}}_2^\intercal$.

Traditionally, one only has access to labeled data $\mathcal{L}$, and hence proceeds by estimating $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in (2.1) by solving the following sample version set of normal equations:

$$\begin{aligned} \mathbb{P}_n\left[\check{\bar{X}}_2(Y_3 - \check{\bar{X}}_2^\intercal \boldsymbol{\theta}_2)\right] \equiv \mathbb{P}_n \begin{bmatrix} Y_2\{Y_3 - (Y_2, \bar{X}_2^\intercal)\boldsymbol{\theta}_2\} \\ \bar{X}_2\{Y_3 - (Y_2, \bar{X}_2^\intercal)\boldsymbol{\theta}_2\} \end{bmatrix} &= \mathbf{0}, \\ \mathbb{P}_n\left[\bar{X}_1\{Y_2(1+\beta_{21}) + \mathbf{H}_{20}^\intercal \boldsymbol{\beta}_{22} + [\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2]_+ - \bar{X}_1^\intercal \boldsymbol{\theta}_1\}\right] &= \mathbf{0}. \end{aligned} \tag{2.2}$$

[3], where $\mathbb{P}_n$ denotes the empirical measure: i.e. for a measurable function $f : \mathbb{R}^p \mapsto \mathbb{R}$ and random sample $\{\mathbf{L}_i\}_{i=1}^n$, $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(\mathbf{L}_i)$. The asymptotic distribution for the Q function parameters in the fully-supervised setting has been well studied [see 54].

2.4.2 Semi-supervised Q-learning

We next detail our robust imputation-based semi-supervised Q-learning that leverages the unlabeled data $\mathcal{U}$ to replace the unobserved Y_t in (2.2) with their properly imputed values for subjects in $\mathcal{U}$. Our SSL procedure includes three key steps: (i) imputation, (ii) refitting, and (iii) projection to the unlabeled data. In step (i), we develop flexible imputation models for the conditional mean functions $\{\mu_t(\cdot), \mu_{2t}(\cdot), t = 2, 3\}$, where $\mu_t(\vec{\mathbf{U}}) = \mathbb{E}(Y_t|\vec{\mathbf{U}})$ and $\mu_{2t}(\vec{\mathbf{U}}) = \mathbb{E}(Y_2 Y_t|\vec{\mathbf{U}})$. The refitting in step (ii) will ensure the validity of the SSL estimators under potential mis-specifications of the imputation models.

Step I: Imputation.

Our first imputation step involves weakly parametric or non-parametric prediction modeling to approximate the conditional mean functions $\{\mu_t(\cdot), \mu_{2t}(\cdot), t = 2, 3\}$. Commonly used models such as non-parametric kernel smoothing, basis function expansion or kernel machine regression

can be used. We denote the corresponding estimated mean functions as $\{\widehat{m}_t(\cdot), \widehat{m}_{2t}(\cdot), t = 2, 3\}$ under the corresponding imputation models $\{m_t(\vec{\mathbf{U}}), m_{2t}(\vec{\mathbf{U}}), t = 2, 3\}$. Theoretical properties of our proposed SSL estimators on specific choices of the imputation models are provided in section 2.6. We also provide additional simulation results comparing different imputation models in section 2.7.

Step II: Refitting.

To overcome the potential bias in the fitting from the imputation model, especially under model mis-specification, we update the imputation model with an additional refitting step by expanding it to include linear effects of $\{\bar{X}_t, t = 1, 2\}$ with cross-fitting to control overfitting bias. Specifically, to ensure the validity of the SSL algorithm from the refitted imputation model, we note that the final imputation models for $\{Y_t, Y_{2t}, t = 2, 3\}$, denoted by $\{\bar{\mu}_t(\vec{\mathbf{U}}), \bar{\mu}_{2t}, t = 2, 3\}$, need to satisfy

$$\mathbb{E}\left[\vec{\bar{X}}\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\}\right] = \mathbf{0}, \qquad \mathbb{E}\left\{Y_2^2 - \bar{\mu}_{22}(\vec{\mathbf{U}})\right\} = 0,$$
$$\mathbb{E}\left[\bar{X}_2\{Y_3 - \bar{\mu}_3(\vec{\mathbf{U}})\}\right] = \mathbf{0}, \quad \mathbb{E}\left\{Y_2 Y_3 - \bar{\mu}_{23}(\vec{\mathbf{U}})\right\} = 0.$$

where $\vec{\bar{X}} = (1, \bar{X}_1^\intercal, \bar{X}_2^\intercal)^\intercal$. We thus propose a refitting step that expands $\{m_t(\vec{\mathbf{U}}), m_{2t}(\vec{\mathbf{U}}), t = 2, 3\}$ to additionally adjust for linear effects of $\bar{X}_1$ and/or $\bar{X}_2$ to ensure the subsequent projection step is unbiased. To this end, let $\{\mathcal{I}_k, k = 1, ..., K\}$ denote K random equal sized partitions of the labeled index set $\{1, ..., n\}$, and let $\{\widehat{m}_t^{(-k)}(\vec{\mathbf{U}}), \widehat{m}_{2t}^{(-k)}(\vec{\mathbf{U}}), t = 2, 3\}$ be the counterpart of $\{\widehat{m}_t(\vec{\mathbf{U}}), \widehat{m}_{2t}(\vec{\mathbf{U}}), t = 2, 3\}$ with labeled observations in $\{1, .., n\} \setminus \mathcal{I}_k$. We then obtain $\widehat{\boldsymbol{\eta}}_2$, $\widehat{\eta}_{22}$, $\widehat{\boldsymbol{\eta}}_3$, $\widehat{\eta}_{23}$ respectively as the solutions to

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \vec{\bar{X}}_i \left\{Y_{2i} - \widehat{m}_2^{(-k)}(\vec{\mathbf{U}}_i) - \boldsymbol{\eta}_2^\intercal \vec{\bar{X}}_i\right\} = \mathbf{0}, \qquad \sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \left\{Y_{2i}^2 - \widehat{m}_{22}^{(-k)}(\vec{\mathbf{U}}_i) - \eta_{22}\right\} = 0, \\
&\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \bar{X}_{2i} \left\{Y_{3i} - \widehat{m}_3^{(-k)}(\vec{\mathbf{U}}_i) - \boldsymbol{\eta}_3^\intercal \bar{X}_{2i}\right\} = \mathbf{0}, \sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \left\{Y_{2i} Y_{3i} - \widehat{m}_{23}^{(-k)}(\vec{\mathbf{U}}_i) - \eta_{23}\right\} = 0.
\end{aligned} \tag{2.3}$$

Finally, we impute Y_2, Y_3, Y_2^2 and $Y_2 Y_3$ respectively as $\widehat{\mu}_2(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^{K} \widehat{m}_2^{(-k)}(\vec{\mathbf{U}}) + \widehat{\boldsymbol{\eta}}_2^\intercal \vec{\bar{X}}$, $\widehat{\mu}_3(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^{K} \widehat{m}_3^{(-k)}(\vec{\mathbf{U}}) + \widehat{\boldsymbol{\eta}}_3^\intercal \bar{X}_2$, $\widehat{\mu}_{22}(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^{K} \widehat{m}_{22}^{(-k)}(\vec{\mathbf{U}}) + \widehat{\eta}_{22}$, and $\widehat{\mu}_{23}(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^{K} \widehat{m}_{23}^{(-k)}(\vec{\mathbf{U}}) + \widehat{\eta}_{23}$.

Step III: Projection

In the last step, we proceed to estimate $\widehat{\boldsymbol{\theta}}$ by replacing $\{Y_t, Y_2Y_t, t=2,3\}$ in (2.2) with their the imputed values $\{\widehat{\mu}_t(\vec{\mathbf{U}}), \widehat{\mu}_{2t}(\vec{\mathbf{U}}), t=2,3\}$ and project to the unlabeled data. Specifically, we obtain the final SSL estimators for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ via the following steps:

1. Stage 2 regression: we obtain the SSL estimator for $\boldsymbol{\theta}_2$ as

$$\widehat{\boldsymbol{\theta}}_2 = (\widehat{\boldsymbol{\beta}}_2^{\intercal}, \widehat{\boldsymbol{\gamma}}_2^{\intercal})^{\intercal} : \text{the solution to} \quad \mathbb{P}_N \begin{bmatrix} \widehat{\mu}_{23}(\vec{\mathbf{U}}) - [\widehat{\mu}_{22}(\vec{\mathbf{U}}), \widehat{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^{\intercal}]\boldsymbol{\theta}_2 \\ \bar{X}_2\{\widehat{\mu}_3(\vec{\mathbf{U}}) - [\widehat{\mu}_2(\vec{\mathbf{U}}), \bar{X}_2^{\intercal}]\boldsymbol{\theta}_2\} \end{bmatrix} = \mathbf{0}$$

2. We compute the imputed pseudo-outcome:

$$\widetilde{Y}_2^* = \widehat{\mu}_2(\vec{\mathbf{U}}) + \max_{a\in\{0,1\}} Q_2\left(\mathbf{H}_2, \widehat{\mu}_2(\vec{\mathbf{U}}), a; \widehat{\boldsymbol{\theta}}_2\right),$$

3. Stage 1 regression: we estimate $\widehat{\boldsymbol{\theta}}_1 = (\widehat{\boldsymbol{\beta}}_1^{\intercal}, \widehat{\boldsymbol{\gamma}}_1^{\intercal})^{\intercal}$ as the solution to:

$$\mathbb{P}_N\left\{\bar{X}_1(\widetilde{Y}_2^* - \bar{X}_1^{\intercal}\boldsymbol{\theta}_1)\right\} = \mathbf{0}.$$

Based on the SSL estimator for the Q-learning model parameters, we can then obtain an estimate for the optimal treatment protocol as:

$$\widehat{d}_t \equiv \widehat{d}_t(\mathbf{H}_t) \equiv d_t(\mathbf{H}_t; \widehat{\boldsymbol{\theta}}_t), \text{ where } d_t(\mathbf{H}_t, \boldsymbol{\theta}_t) = \operatorname*{argmax}_{a\in\{0,1\}} Q_t(\mathbf{H}_t, a; \boldsymbol{\theta}_t) = I\left(\mathbf{H}_{t1}^{\intercal}\boldsymbol{\gamma}_t > 0\right), \ t=1,2.$$

Theorems 2.6.4 and 2.6.6 of Section 2.6 demonstrate the consistency and asymptotic normality of the SSL estimators $\{\widehat{\boldsymbol{\theta}}_t, t=1,2\}$ for their respective population parameters $\{\bar{\boldsymbol{\theta}}_t, t=1,2\}$ even in the possible mis-specification of (2.1). As we explain next, this in turn yields desirable statistical results for evaluating the resulting policy $\bar{d}_t \equiv \bar{d}_t(\mathbf{H}_t) \equiv d_t(\mathbf{H}_t, \bar{\boldsymbol{\theta}}_t) = \operatorname{argmax}_{a\in\{0,1\}} Q_t(\check{\mathbf{H}}_t, a; \bar{\boldsymbol{\theta}}_t)$ for $t=1,2$.

2.5 Semi Supervised Off-Policy Evaluation of the Policy

To evaluate the performance of the optimal policy $\bar{\mathcal{D}} = \{\bar{d}_t(\mathbf{H}_t), t = 1, 2\}$, derived under the Q-learning framework, one may estimate the expected population outcome under the policy $\bar{\mathcal{D}}$:

$$\bar{V} \equiv \mathbb{E}\left[\mathbb{E}\{Y_2 + \mathbb{E}\{Y_3|\check{\mathbf{H}}_2, A_2 = \bar{d}_2(\mathbf{H}_2)\}|\mathbf{H}_1, A_1 = \bar{d}_1(\mathbf{H}_1)\}\right].$$

If models in (2.1) are correctly specified, then under standard causal assumptions (consistency, no unmeasured confounding, and positivity), an asymptotically consistent supervised estimator for the value function can be obtained as

$$\widehat{V}_Q = \mathbb{P}_n\left[Q_1^o(\check{\mathbf{H}}_1; \widehat{\boldsymbol{\theta}}_1)\right],$$

where $Q_t^o(\check{\mathbf{H}}_t; \boldsymbol{\theta}_t) \equiv Q_t\left(\check{\mathbf{H}}_t, d_t(\mathbf{H}_t; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t\right)$. However, $\widehat{V}_Q$ is likely to be biased when the outcome models in (2.1) are mis-specified. This occurs frequently in practice since $Q_1(\check{\mathbf{H}}_1, A_1)$ is especially difficult to specify.

To improve the robustness to model mis-specification, we augment $\widehat{V}_Q$ via propensity score weighting. This gives us an SSL doubly robust (SSL_{DR}) estimator for $\bar{V}$. To this end, we define propensity scores:

$$\pi_t(\check{\mathbf{H}}_t) = \mathbb{P}\{A_t = 1|\check{\mathbf{H}}_t\}, \quad t = 1, 2.$$

To estimate $\{\pi_t(\cdot), t = 1, 2\}$, we impose the following generalized linear models (GLM):

$$\pi_t(\check{\mathbf{H}}_t; \boldsymbol{\xi}_t) = \sigma\left(\check{\mathbf{H}}_t^{\mathsf{T}}\boldsymbol{\xi}_t\right), \quad \text{with} \quad \sigma(x) \equiv 1/(1 + e^{-x}) \quad \text{for} \quad t = 1, 2. \tag{2.4}$$

We use the logistic model with potentially non-linear basis functions $\check{\mathbf{H}}$ for simplicity of presentation but one may choose other GLM or alternative basis expansions to incorporate non-linear effects in the propensity model. We estimate $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^{\mathsf{T}}, \boldsymbol{\xi}_2^{\mathsf{T}})^{\mathsf{T}}$ based on the standard maximum likelihood estimators using labeled data, denoted by $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\xi}}_1^{\mathsf{T}}, \widehat{\boldsymbol{\xi}}_2^{\mathsf{T}})^{\mathsf{T}}$. We denote the limit of $\widehat{\boldsymbol{\xi}}$ as $\bar{\boldsymbol{\xi}} = (\bar{\boldsymbol{\xi}}_1^{\mathsf{T}}, \bar{\boldsymbol{\xi}}_2^{\mathsf{T}})^{\mathsf{T}}$. Note that this is not necessarily equal to the true model parameter under correct specification of (2.4), but corresponds to the population solution of the fitted models.

Our framework is flexible to allow an SSL approach to estimate the propensity scores. As these are nuisance parameters needed for estimation of the value function, and SSL for GLMs has been widely explored [See 55, Ch. 2], we proceed with the usual GLM estimation to keep the discussion focused. However, SSL for propensity scores can be beneficial in certain cases, as we show in Proposition 2.6.2.

2.5.1 SUP_{DR} Value Function Estimation

To derive a supervised doubly robust (SUP_{DR}) estimator for $\bar{V}$ overcoming confounding in the observed data, we let $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\intercal, \boldsymbol{\xi}^\intercal)^\intercal$ and define the inverse probability weights (IPW) using the propensity scores

$$\omega_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \equiv \frac{d_1(\mathbf{H}_1;\boldsymbol{\theta}_1)A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)} + \frac{\{1-d_1(\mathbf{H}_1;\boldsymbol{\theta}_1)\}\{1-A_1\}}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}, \quad \text{and}$$
$$\omega_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta}) \equiv \omega_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \left(\frac{d_2(\mathbf{H}_2;\boldsymbol{\theta}_2)A_2}{\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)} + \frac{\{1-d_2(\mathbf{H}_2;\boldsymbol{\theta}_2)\}\{1-A_2\}}{1-\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)} \right).$$

Then we augment $Q_1^o(\mathbf{H}_1;\widehat{\boldsymbol{\theta}}_1)$ based on the estimated propensity scores via

$$\mathcal{V}_{\text{SUP}_{\text{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}}) = Q_1^o(\mathbf{H}_1;\widehat{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \widehat{\boldsymbol{\Theta}}) \left[Y_2 - \left\{ Q_1^o(\mathbf{H}_1,\widehat{\boldsymbol{\theta}}_1) - Q_2^o(\check{\mathbf{H}}_2;\widehat{\boldsymbol{\theta}}_2) \right\} \right]$$
$$+ \omega_2(\check{\mathbf{H}}_2, A_2, \widehat{\boldsymbol{\Theta}}) \left\{ Y_3 - Q_2^o(\check{\mathbf{H}}_2;\widehat{\boldsymbol{\theta}}_2) \right\}$$

and estimate $\bar{V}$ as

$$\widehat{V}_{\text{SUP}_{\text{DR}}} = \mathbb{P}_n \left\{ \mathcal{V}_{\text{SUP}_{\text{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}}) \right\}. \tag{2.5}$$

The importance sampling estimators previously proposed in [7] and [6] for value function estimation employ similar augmentation strategies. However, they consider a fixed policy, and we account for the fact that the STR is estimated with the same data. The construction of augmentation in $\widehat{V}_{\text{SUP}_{\text{DR}}}$ also differs from the usual augmented IPW estimators [3]. As we are interested in the value had the population been treated with function $\bar{\mathcal{D}}$ and not a fixed sequence (A_1, A_2), we augment the weights for a fixed treatment (i.e. $A_t = 1$) with the propensity score weights for the estimated regime $I(A_t = \bar{d}_t)$. Finally, we note that this estimator can easily be extended to incorporate non-binary treatments.

The supervised value function estimator $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$ is doubly robust in the sense that if either the outcome models of the propensity score models are correctly specified, then $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}} \xrightarrow{\mathbb{P}} \bar{V}$ in probability. Moreover, under certain reasonable assumptions, $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$ is asymptotically normal. Theoretical guarantees and proofs for this procedure are shown in Appendix B.4.1.

2.5.2 $\mathsf{SSL}_{\mathsf{DR}}$ Value Function Estimation

Analogous to semi-supervised Q-learning, we propose a procedure for adapting the augmented value function estimator to leverage $\mathcal{U}$, by imputing suitable functions of the unobserved outcome in (2.5). Since $\check{\mathbf{H}}_2$ involves Y_2, both $\omega_2(\check{\mathbf{H}}_2, A_2; \boldsymbol{\Theta})$ and $Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2) = Y_2\beta_{21} + Q_{2-}^o(\mathbf{H}_2; \boldsymbol{\theta}_2)$ are not available in the unlabeled set, where $Q_{2-}^o(\mathbf{H}_2; \boldsymbol{\theta}_2) = \mathbf{H}_{20}^\intercal \boldsymbol{\beta}_{22} + [\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2]_+$. By writing $\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}})$ as

$$\begin{aligned}\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}) = Q_1^o(\mathbf{H}_1; \widehat{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \widehat{\boldsymbol{\Theta}}) \left\{(1+\widehat{\beta}_{21})Y_2 - Q_1^o(\mathbf{H}_1, \widehat{\boldsymbol{\theta}}_1) + Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\right\} \\ + \omega_2(\check{\mathbf{H}}_2, A_2, \widehat{\boldsymbol{\Theta}}) \left\{Y_3 - \widehat{\beta}_{21}Y_2 - Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\right\},\end{aligned}$$

we note that to impute $\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}})$ for subjects in $\mathcal{U}$, we need to impute Y_2, $\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}})$, and $Y_t\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}})$ for $t = 2, 3$. We define the conditional mean functions

$$\mu_2^v(\vec{\mathbf{U}}) \equiv \mathbb{E}[Y_2|\vec{\mathbf{U}}], \quad \mu_{\omega_2}^v(\vec{\mathbf{U}}) \equiv \mathbb{E}[\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}})|\vec{\mathbf{U}}], \quad \mu_{t\omega_2}^v(\vec{\mathbf{U}}) \equiv \mathbb{E}[Y_t\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}})|\vec{\mathbf{U}}],$$

for $t = 2, 3$, where $\bar{\boldsymbol{\Theta}} = (\bar{\boldsymbol{\theta}}^\intercal, \bar{\boldsymbol{\xi}}^\intercal)^\intercal$. As in Section 2.4.2 we approximate these expectations using a flexible imputation model followed by a refitting step for bias correction under possible mis-specification of the imputation models.

Step I: Imputation

We fit flexible weakly parametric or non-parametric models to the labeled data to approximate the functions $\{\mu_2^v(\vec{\mathbf{U}}), \mu_{\omega_2}^v(\vec{\mathbf{U}}), \mu_{t\omega_2}^v(\vec{\mathbf{U}}),\ t = 2, 3\}$ with unknown parameter $\boldsymbol{\Theta}$ estimated via the SSL Q-learning as in Section 2.4.2 and the propensity score modeling as discussed above. Denote the respective imputation models as $\{m_2(\vec{\mathbf{U}}), m_{\omega_2}(\vec{\mathbf{U}}), m_{t\omega_2}(\vec{\mathbf{U}}),\ t = 2, 3\}$ and their fitted values as $\{\widehat{m}_2(\vec{\mathbf{U}}), \widehat{m}_{\omega_2}(\vec{\mathbf{U}}), \widehat{m}_{t\omega_2}(\vec{\mathbf{U}}),\ t = 2, 3\}$.

Step II: Refitting

To correct for potential biases arising from finite sample estimation and model mis-specifications, we perform refitting to obtain final imputed models for $\{Y_2, \omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}),\ Y_t\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}),$ $t = 2, 3\}$ as $\{\bar{\mu}_2^v(\vec{\mathbf{U}}) = m_2(\vec{\mathbf{U}}) + \eta_2^v, \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}}) = m_{\omega_2}(\vec{\mathbf{U}}) + \eta_{\omega_2}^v, \bar{\mu}_{t\omega_2}^v(\vec{\mathbf{U}}) = m_{t\omega_2}(\vec{\mathbf{U}}) + \eta_{t\omega_2}^v, t = 2, 3\}$. As for the estimation of $\boldsymbol{\theta}$ for Q-learning training, these refitted models are not required to be correctly specified but need to satisfy the following constraints:

$$\begin{aligned}
\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}})\left\{Y_2 - \bar{\mu}_2^v(\vec{\mathbf{U}})\right\}\right] &= 0, \\
\mathbb{E}\left[Q_{2-}^o(\vec{\mathbf{U}}; \boldsymbol{\theta}_2)\left\{\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}) - \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}})\right\}\right] &= 0, \\
\mathbb{E}\left[\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}})Y_t - \bar{\mu}_{t\omega_2}^v(\vec{\mathbf{U}})\right] &= 0,\ t = 2, 3.
\end{aligned}$$

To estimate η_2^v $\eta_{\omega_2}^v$, and $\eta_{t\omega_2}^v$ under these constraints, we again employ cross-fitting and obtain $\widehat{\eta}_2^v$ $\widehat{\eta}_{\omega_2}^v$, and $\widehat{\eta}_{t\omega_2}^v$ as the solution to the following estimating equations

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \omega_1(\check{\mathbf{H}}_{1i}, A_{1i}; \widehat{\bar{\boldsymbol{\Theta}}})\left\{Y_2 - \widehat{m}_2^{(-k)}(\vec{\mathbf{U}}_i) - \widehat{\eta}_2^v\right\} &= 0, \\
\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} Q_{2-}^o(\vec{\mathbf{U}}_i; \widehat{\boldsymbol{\theta}}_2)\left\{\omega_2(\check{\mathbf{H}}_{2i}, A_{2i}; \widehat{\bar{\boldsymbol{\Theta}}}) - \widehat{m}_{\omega_2}^{(-k)}(\vec{\mathbf{U}}_i) - \widehat{\eta}_{\omega_2}^v\right\} &= 0, \\
\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \left\{\omega_2(\check{\mathbf{H}}_{2i}, A_{2i}; \widehat{\bar{\boldsymbol{\Theta}}})Y_{ti} - \widehat{m}_{t\omega_2}^{(-k)}(\vec{\mathbf{U}}_i) - \widehat{\eta}_{t\omega_2}^v\right\} &= 0,\ t = 2, 3.
\end{aligned} \tag{2.6}$$

The resulting imputation functions for $Y_2, \omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}})$ and $Y_t\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}})$ are respectively constructed as $\widehat{\mu}_2^v(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^K \widehat{m}_2^{(-k)}(\vec{\mathbf{U}}) + \widehat{\eta}_2^v$, $\widehat{\mu}_{\omega_2}^v(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^K \widehat{m}_{\omega_2}(\vec{\mathbf{U}}) + \widehat{\eta}_{\omega_2}^v$, and $\widehat{\mu}_{t\omega_2}^v(\vec{\mathbf{U}}) = K^{-1}\sum_{k=1}^K \widehat{m}_{t\omega_2}^{(-k)}(\vec{\mathbf{U}}) + \widehat{\eta}_{t\omega_2}^v$, for $t = 2, 3$.

Step III: Semi-supervised augmented value function estimator.

Finally, we proceed to estimate the value of the policy $\bar{V}$, using the following semi-supervised augmented estimator:

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} = \mathbb{P}_N\left\{\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \widehat{\bar{\boldsymbol{\Theta}}}, \widehat{\mu})\right\}, \tag{2.7}$$

where $\widehat{\mathcal{V}}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}})$ is the semi-supervised augmented estimator for observation $\vec{\mathbf{U}}$ defined as:

$$\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\widehat{\boldsymbol{\Theta}},\widehat{\mu}) = Q_1^o(\check{\mathbf{H}}_1;\widehat{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \widehat{\boldsymbol{\Theta}})\left[(1+\widehat{\beta}_{21})\widehat{\mu}_2^v(\vec{\mathbf{U}}) - Q_1^o(\check{\mathbf{H}}_1;\widehat{\boldsymbol{\theta}}_1) + Q_{2-}^o(\mathbf{H}_2;\widehat{\boldsymbol{\theta}}_2)\right]$$
$$+\widehat{\mu}_{3\omega_2}(\vec{\mathbf{U}}) - \widehat{\beta}_{21}\widehat{\mu}_{2\omega_2}(\vec{\mathbf{U}}) - Q_{2-}^o(\mathbf{H}_2;\widehat{\boldsymbol{\theta}}_2)\widehat{\mu}_{\omega_2}(\vec{\mathbf{U}}).$$

The above SSL estimator uses both labeled and unlabeled data along with outcome surrogates to estimate the value function, which yields a gain in efficiency as we show in Proposition 2.6.2. As its supervised counterpart, $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ is doubly robust in the sense that if either the Q functions or the propensity scores are correctly specified, the value function will converge in probability to the true value $\bar{V}$. Additionally, it does not assume that the estimated treatment regime was derived from a different sample. These properties are summarized in Theorem 2.6.9 and Proposition 2.6.2 of the following section.

2.6 Theoretical Results

In this section we discuss our assumptions and theoretical results for the semi-supervised Q-learning and value function estimators. Throughout, we define the norm $\|g(x)\|_{L_2(\mathbb{P})} \equiv \sqrt{\int g(x)^2 d\mathbb{P}(x)}$ for any real valued function $g(\cdot)$. Additionally, let $\{U_n\}$, and $\{V_n\}$ be two sequences of random variables. We will use $U_n = O_{\mathbb{P}}(V_n)$ to denote stochastic boundedness of the sequence $\{U_n/V_n\}$, that is, for any $\epsilon > 0$, $\exists M_\epsilon, n_\epsilon \in \mathbb{R}$ such that $\mathbb{P}\left(|U_n/V_n| > M_\epsilon\right) < \epsilon$ $\forall n > n_\epsilon$. We use $U_n = o_{\mathbb{P}}(V_n)$ to denote that $U_n/V_n \xrightarrow{\mathbb{P}} 0$.

2.6.1 Theoretical Results for SSL Q-learning

Assumption 2.6.1. *(a) Sample size for $\mathcal{U}$, and $\mathcal{L}$, are such that $n/N \longrightarrow 0$ as $N, n \longrightarrow \infty$, (b) $\check{\mathbf{H}}_t \in \mathcal{H}_t$, $\check{\check{X}}_t \in \mathcal{X}_t$ have finite second moments and compact support in $\mathcal{H}_t \subset \mathbb{R}^{q_t}$, $\mathcal{X}_t \subset \mathbb{R}^{p_t}$ $t = 1, 2$ respectively (c) $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ are nonsingular.*

Assumption 2.6.2. *Functions m_s, $s \in \{2, 3, 22, 23\}$ are such that (i) $\sup_{\vec{\mathbf{U}}} |m_s(\vec{\mathbf{U}})| < \infty$, and (ii) the estimated functions $\hat{m}_s$ satisfy (ii) $\sup_{\vec{\mathbf{U}}} |\widehat{m}_s(\vec{\mathbf{U}}) - m_s(\vec{\mathbf{U}})| = o_{\mathbb{P}}(1)$.*

Assumption 2.6.3. *Suppose Θ_1, Θ_2 are open bounded sets, and p_1, p_2 fixed under (2.1). We*

define the following class of functions:

$$\mathcal{Q}_t \equiv \{Q_t : \mathcal{X}_1 \mapsto \mathbb{R} | \boldsymbol{\theta}_1 \in \Theta_1 \subset \mathbb{R}^{p_t}\}, \ t = 1, 2.$$

Further suppose for $t = 1, 2$*, the solutions for* $\mathbb{E}[S_t^{\theta}(\boldsymbol{\theta}_t)] = \mathbf{0}$*, i.e.* $\bar{\boldsymbol{\theta}}_1$ *and* $\bar{\boldsymbol{\theta}}_2$ *satisfy*

$$S_2^{\theta}(\boldsymbol{\theta}_2) = \frac{\partial}{\partial \boldsymbol{\theta}_2^{\intercal}} \|Y_3 - Q_2(\check{\check{X}}_2; \boldsymbol{\theta}_2)\|_2^2, \ S_1^{\theta}(\boldsymbol{\theta}_1) = \frac{\partial}{\partial \boldsymbol{\theta}_1^{\intercal}} \|Y_2^* - Q_1(\check{\check{X}}_1; \boldsymbol{\theta}_1)\|_2^2.$$

The target parameters satisfy $\bar{\boldsymbol{\theta}}_t \in \Theta_t \, , t = 1, 2$*. We write* $\bar{\boldsymbol{\beta}}_t, \bar{\gamma}_t$ *as the components of* $\bar{\boldsymbol{\theta}}_t$*, according to equation* (2.2)*.*

Assumption 2.6.1 (a) distinguishes our setting from the standard missing data context. Theoretical results for the missing completely at random (MCAR) setting generally assume that the missingness probability is bounded away from zero [56], which enables the use of standard semiparametric theory. However, in our setting one can intuitively consider the probability of observing an outcome being $\frac{n}{n+N}$ which converges to 0.

Assumption 2.6.2 is fairly standard as it just requires boundedness of the imputation functions – which is natural to expect from the boundedness of the covariates. We also require uniform convergence of the estimated functions to their limit. This allows for the normal equations targeting the imputation residuals in (2.3) and (2.6) to be well defined. Moreover, several off-the-shelf flexible imputation models for estimation can satisfy these conditions. See for example, local polynomial estimators, basis expansion regression like natural cubic splines or wavelets [57]. In particular, it is worth noting that we do not require any specific rate of convergence. As a result, the required condition is typically much easier to verify for many off-the-shelf algorithms. It is likely that other classes of models such as random forests can satisfy Assumption 2.6.2. Recent work suggests that it is plausible to use the existing point-wise convergence results to show uniform convergence. [see 58, 59].

Assumption 2.6.3 is fairly standard in the literature and ensures well-defined population level solutions for Q-learning regressions $\bar{\boldsymbol{\theta}}$ exist, and belong to that parameter space. In this regard, we differentiate between population solutions $\bar{\boldsymbol{\theta}}$ and true model parameters $\boldsymbol{\theta}^0$

shown in equation (2.1). If the working models are mis-specified, Theorems 2.6.4 and 2.6.6 still guarantee the $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal centered at the population solution $\bar{\boldsymbol{\theta}}$. However, when equation (2.1) is correct, $\widehat{\boldsymbol{\theta}}$ is asymptotically normal and consistent for the true parameter $\boldsymbol{\theta}^0$. Now we are ready to state the theoretical properties of the semi-supervised Q-learning procedure described in Section 2.4.2.

Theorem 2.6.4 (Distribution of $\widehat{\boldsymbol{\theta}}_2$)**.** *Under Assumptions 2.6.1-2.6.3, $\widehat{\boldsymbol{\theta}}_2$ satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2) = \boldsymbol{\Sigma}_2^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}_2(\mathbf{L}_i; \bar{\boldsymbol{\theta}}_2) + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}, \mathbf{V}_{2\mathsf{SSL}}(\bar{\boldsymbol{\theta}}_2)\Big),$$

where $\boldsymbol{\Sigma}_2 = \mathbb{E}[\overset{\scriptscriptstyle\vee}{X}_2 \overset{\scriptscriptstyle\vee}{X}_2^{\intercal}]$ is defined in Section 2.3, the influence function $\boldsymbol{\psi}_2$ is given by

$$\boldsymbol{\psi}_2(\mathbf{L}; \bar{\boldsymbol{\theta}}_2) = \begin{bmatrix} \{Y_2 Y_3 - \bar{\mu}_{23}(\vec{\mathbf{U}})\} - \bar{\beta}_{21}\{Y_2^2 - \bar{\mu}_{22}(\vec{\mathbf{U}})\} - Q_{2-}(\mathbf{H}_2, A_2; \bar{\boldsymbol{\theta}}_2)\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\} \\ \bar{X}_2\{Y_3 - \bar{\mu}_3(\vec{\mathbf{U}})\} - \bar{\beta}_{21}\bar{X}_2\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\} \end{bmatrix},$$

and $\mathbf{V}_{2\mathsf{SSL}}(\bar{\boldsymbol{\theta}}_2) = \boldsymbol{\Sigma}_2^{-1} \mathbb{E}\left[\boldsymbol{\psi}_2(\mathbf{L}; \bar{\boldsymbol{\theta}}_2) \boldsymbol{\psi}_2(\mathbf{L}; \bar{\boldsymbol{\theta}}_2)^{\intercal}\right] \left(\boldsymbol{\Sigma}_2^{-1}\right)^{\intercal}$.

We hold off remarks until the end of the results for the Q-learning parameters. Since the first stage regression depends on the second stage regression through a non-smooth maximum function, we make the following standard assumption [54] in order to provide valid statistical inference.

Assumption 2.6.5. *Non-zero estimable population treatment effects $\bar{\boldsymbol{\gamma}}_t$, $t = 1, 2$: i.e. the population solution to* (2.2)*, is such that (a) $\mathbf{H}_{21}^{\intercal} \bar{\boldsymbol{\gamma}}_2 \neq 0$ for all $\mathbf{H}_{21} \neq \mathbf{0}$, and (b) $\bar{\boldsymbol{\gamma}}_1$ is such that $\mathbf{H}_{11}^{\intercal} \bar{\boldsymbol{\gamma}}_1 \neq 0$ for all $\mathbf{H}_{11} \neq \mathbf{0}$.*

Assumption 2.6.5 yields regular estimators for the stage one regression and the value function, which depend on non-smooth components of the form $[x]_+$. This is needed to achieve asymptotic normality of the Q-learning parameters for the first stage regression. Note that the estimating equation for the stage one regression in Section 2.4.2 includes $[\mathbf{H}_{21}^{\intercal} \widehat{\boldsymbol{\gamma}}_2]_+$. Thus, for the asymptotic normality of $\widehat{\boldsymbol{\theta}}_1$, we require $\sqrt{n}\mathbb{P}_n\left([\mathbf{H}_{21}^{\intercal} \widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal} \bar{\boldsymbol{\gamma}}_2]_+\right)$ to be asymptotically normal. The latter is automatically true if $\mathbf{H}_{11}$ contains continuous covariates as $\mathbb{P}(\mathbf{H}_{21}^{\intercal} \bar{\boldsymbol{\gamma}}_2 = 0) = 0$. Violation of Assumption 2.6.5 will yield non-regular estimates which

translate into poor coverage for the confidence intervals (see [54] for a thorough discussion on this topic).

Theorem 2.6.6 (Distribution of $\hat{\boldsymbol{\theta}}_1$)**.** *Under Assumptions 2.6.1-2.6.3, and 2.6.5 (a),* $\widehat{\boldsymbol{\theta}}_1$ *satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_1) = \boldsymbol{\Sigma}_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}_1(\mathbf{L}_i; \bar{\boldsymbol{\theta}}_1) + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}, \mathbf{V}_{1\mathsf{SSL}}(\bar{\boldsymbol{\theta}}_1)\Big)$$

where $\boldsymbol{\Sigma}_1^{-1} = \mathbb{E}[\breve{X}_1 \breve{X}_1^{\intercal}]$*, the influence function* $\boldsymbol{\psi}_1$ *is given by*

$$\begin{aligned}\boldsymbol{\psi}_1(\mathbf{L}; \bar{\boldsymbol{\theta}}_1) =& \bar{X}_1(1 + \bar{\beta}_{21})\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\} + \mathbb{E}\left[\bar{X}_1 \left(Y_2, \mathbf{H}_{20}^{\intercal}\right)\right] \boldsymbol{\psi}_{\beta_2}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2) \\ &+ \mathbb{E}\left[\bar{X}_1 \mathbf{H}_{21}^{\intercal} | \mathbf{H}_{21}^{\intercal} \bar{\boldsymbol{\gamma}}_2 > 0\right] \mathbb{P}\left(\mathbf{H}_{21}^{\intercal} \bar{\boldsymbol{\gamma}}_2 > 0\right) \boldsymbol{\psi}_{\gamma_2}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2),\end{aligned}$$

$\mathbf{V}_{1\mathsf{SSL}}(\bar{\boldsymbol{\theta}}_1) = \boldsymbol{\Sigma}_1^{-1} \mathbb{E}\left[\boldsymbol{\psi}_1(\mathbf{L}; \bar{\boldsymbol{\theta}}_1) \boldsymbol{\psi}_1(\mathbf{L}; \bar{\boldsymbol{\theta}}_1)^{\intercal}\right] \left(\boldsymbol{\Sigma}_1^{-1}\right)^{\intercal}$*, and* $\boldsymbol{\psi}_{\beta_2}$*,* $\boldsymbol{\psi}_{\gamma_2}$ *are the elements corresponding to* $\bar{\boldsymbol{\beta}}_2$*,* $\bar{\boldsymbol{\gamma}}_2$ *of the influence function* $\boldsymbol{\psi}_2$ *defined in Theorem 2.6.4.*

1) Theorems 2.6.4 and 2.6.6 establish the $\sqrt{n}$-consistency and asymptotic normality (CAN) of $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$ for any $K \geq 2$. Beyond asymptotic normality at $\sqrt{n}$ scale, these theorems also provide an asymptotic linear expansion of the estimators with influence functions $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ respectively.

2) $\mathbf{V}_{1\mathsf{SSL}}(\bar{\boldsymbol{\theta}})$, $\mathbf{V}_{2\mathsf{SSL}}(\bar{\boldsymbol{\theta}})$ reflect an efficiency gain over the fully supervised approach due to sample $\mathcal{U}$ and the surrogates contribution in prediction performance. This gain is formalized in Proposition 2.6.1 which quantifies how correlation between surrogates and outcome increases efficiency.

3) Let $\boldsymbol{\psi} = [\boldsymbol{\psi}_1^{\intercal}, \boldsymbol{\psi}_2^{\intercal}]^{\intercal}$, we collect the vector of estimated Q-learning parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\intercal}, \boldsymbol{\theta}_2^{\intercal})^{\intercal}$, then under Assumptions 2.6.1-2.6.3, 2.6.5 (a), we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = \boldsymbol{\Sigma}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{L}_i; \bar{\boldsymbol{\theta}}) + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}, \mathbf{V}_{\mathsf{SSL}}\left(\bar{\boldsymbol{\theta}}\right)\Big)$$

with $\mathbf{V}_{\mathsf{SSL}}(\bar{\boldsymbol{\theta}}) = \boldsymbol{\Sigma}^{-1} \mathbb{E}\left[\boldsymbol{\psi}(\mathbf{L}; \bar{\boldsymbol{\theta}}) \boldsymbol{\psi}(\mathbf{L}; \bar{\boldsymbol{\theta}})^{\intercal}\right] \left(\boldsymbol{\Sigma}^{-1}\right)^{\intercal}$.

4) Theorems 2.6.4 and 2.6.6 hold even when the Q functions are mis-specified, that is, $\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2$ are CAN for $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2$. Furthermore, if model (2.1) is correctly specified then we can simply

replace $\bar{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}^0$ in the above result.

3) We estimate $\mathbf{V}_{\mathsf{SSL}}(\bar{\boldsymbol{\theta}})$ via sample-splitting as

$$\begin{aligned}
\widehat{\mathbf{V}}_{\mathsf{SSL}}(\widehat{\boldsymbol{\theta}}) &= \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{A}}(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)^{\mathsf{T}}, \text{ where} \\
\widehat{\mathbf{A}}(\widehat{\boldsymbol{\theta}}) &= n^{-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\boldsymbol{\psi}^{(-k)}\left(\mathbf{L}_i;\widehat{\boldsymbol{\theta}}\right)\boldsymbol{\psi}^{(-k)}\left(\mathbf{L}_i;\widehat{\boldsymbol{\theta}}\right)^{\mathsf{T}}, \\
\widehat{\boldsymbol{\Sigma}}_t &= \mathbb{P}_n\left\{\bar{X}_t\bar{X}_t^{\mathsf{T}}\right\}, \ t=1,2.
\end{aligned}$$

Note that we can decompose $\boldsymbol{\psi}$ into the influence function for each set of parameters. For example, we have $\boldsymbol{\psi}_2 = \left(\boldsymbol{\psi}_{\beta_2}^{\mathsf{T}}, \boldsymbol{\psi}_{\gamma_2}^{\mathsf{T}}\right)^{\mathsf{T}}$ where $\boldsymbol{\psi}_{\gamma_2}(\mathbf{L};\bar{\boldsymbol{\theta}}_2) = \mathbf{H}_{21}A_2\left[\{Y_3-\bar{\mu}_3(\vec{\mathbf{U}})\}-\bar{\beta}_{21}\{Y_2-\bar{\mu}_2(\vec{\mathbf{U}})\}\right]$. Therefore we can decompose the variance-covariance matrix into a component for each parameter, the variance-covariance for the treatment effect for stage 2 regression $\boldsymbol{\gamma}_2$ is

$$\mathbb{E}\left[\boldsymbol{\psi}_{\gamma_2}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)\boldsymbol{\psi}_{\gamma_2}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)^{\mathsf{T}}\right] = \mathbb{E}\left[\mathbf{H}_{21}\mathbf{H}_{21}^{\mathsf{T}}A_2^2\left\{Y_3-\bar{\mu}_3(\vec{\mathbf{U}})-\beta_{21}\left(Y_2-\bar{\mu}_2(\vec{\mathbf{U}})\right)\right\}^2\right].$$

This gives us some insight into how the predictive power of $\vec{\mathbf{U}}$, which contains surrogates $\mathbf{W}_1, \mathbf{W}_2$, decreases parameter standard errors. This is the case for the influence functions for estimating $\bar{\boldsymbol{\theta}}_1$, $\bar{\boldsymbol{\theta}}_2$ as well. We formalize this result with the following proposition. Let $\widehat{\boldsymbol{\theta}}_{\mathsf{SUP}}$ be the estimator for the fully supervised Q-learning procedure (i.e. only using labeled data), with influence function and asymptotic variance denoted as $\boldsymbol{\psi}_{\mathsf{SUP}}$ and $\mathbf{V}_{\mathsf{SUP}}$ respectively (see Appendix B.2.1 for the exact form of $\boldsymbol{\psi}_{\mathsf{SUP}}$ and $\mathbf{V}_{\mathsf{SUP}}$).

For the following proposition we need the imputation models $\bar{\mu}_s$, $s \in \{2,3,22,23\}$ to satisfy additional constraints of the form $\mathbb{E}\left[\bar{X}_2\bar{X}_2^{\mathsf{T}}\{Y_2Y_3-\bar{\mu}_{23}(\vec{\mathbf{U}})\}\right] = \mathbf{0}$. We list them in Assumption B.2.1, Appendix B.2.1. One can construct estimators which satisfy such conditions by simply augmenting $\boldsymbol{\eta}_2$, η_{22}, $\boldsymbol{\eta}_3$, η_{23} in (2.3) with additional terms in the refitting step.

Under Assumptions 2.6.1-2.6.3, 2.6.5 (a), and B.2.1 then

$$\mathbf{V}_{\mathsf{SSL}}(\bar{\boldsymbol{\theta}}) = \mathbf{V}_{\mathsf{SUP}}(\bar{\boldsymbol{\theta}}) - \boldsymbol{\Sigma}^{-1}\mathrm{Var}\left[\boldsymbol{\psi}_{\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}) - \boldsymbol{\psi}_{\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\theta}})\right]\left(\boldsymbol{\Sigma}^{-1}\right)^{\mathsf{T}}.$$

Proposition 2.6.1 illustrates how the estimates for the semi-supervised Q-learning parame-

ters are at least as efficient, if not more so, than the supervised ones. Intuitively, the difference in efficiency is explained by how much information is gained by incorporating the surrogates $\mathbf{W}_1, \mathbf{W}_2$ into the estimation procedure. If there is no new information in the surrogate variables, then residuals found in $\boldsymbol{\psi}_{\mathsf{SSL}}(\mathbf{L};\boldsymbol{\theta})$ will be of similar magnitude to those in $\boldsymbol{\psi}_{\mathsf{SUP}}(\mathbf{L};\boldsymbol{\theta})$, and thus the difference in efficiency will be small: $\text{Var}\left[\boldsymbol{\psi}_{\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}) - \boldsymbol{\psi}_{\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\theta}})\right] \approx 0$. In this case both methods will yield equally efficient parameters. The gain in precision is especially relevant for the treatment interaction coefficients $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2$ used to learn the dynamic treatment rules. Finally, note that for Proposition 2.6.1, we do not need the correct specification of Q-functions or imputation models.

2.6.2 Theoretical Results for SSL Estimation of the Value Function

If model (2.1) is correct, one only needs to add Assumption 2.6.5 (b) for $\mathbb{P}_N\{Q_1^o(\mathbf{H}_1;\widehat{\boldsymbol{\theta}}_1)\}$ to be a consistent estimator of the value function $\bar{V}$ [60]. However, as we discussed earlier, (2.1) is likely mis-specified. Therefore, we show our semi-supervised value function estimator is doubly robust. We also show it is asymptotically normal and more efficient that its supervised counterpart. To that end, define the following class of functions:

$$\mathcal{W}_t \equiv \{\pi_t : \mathcal{H}_t \mapsto \mathbb{R} | \boldsymbol{\xi}_t \in \Omega_t\}\,,\ t = 1, 2,$$

under propensity score models $\pi_1,\ \pi_2$ in (2.4).

Assumption 2.6.7. *Let the population equations* $\mathbb{E}\left[S_t^{\xi}(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)\right] = \mathbf{0}, t = 1, 2$ *have solutions* $\bar{\boldsymbol{\xi}}_1, \bar{\boldsymbol{\xi}}_2$*, where*

$$S_t^{\xi}(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t) = \frac{\partial}{\partial \boldsymbol{\xi}_t} \log\left[\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)^{A_t}\{1-\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)\}^{(1-A_t)}\right],\ t = 1, 2,$$

(i) Ω_1, Ω_2 *are open, bounded sets and the population solutions satisfy* $\bar{\boldsymbol{\xi}}_t \in \Omega_t, t = 1, 2$*,*

(ii) for $\bar{\boldsymbol{\xi}}_t, t = 1, 2,\ \inf_{\check{\mathbf{H}}_t \in \mathcal{H}_1} \pi_1(\check{\mathbf{H}}_t;\bar{\boldsymbol{\xi}}_t) > 0$*,*

(iii) Finite second moment: $\mathbb{E}\left[S_t^{\xi}(\check{\mathbf{H}}_t;\boldsymbol{\Theta}_t)^2\right] \leq \infty$*, and Fisher information matrix:* $\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\xi}_t} S_t^{\xi}(\check{\mathbf{H}}_t;\boldsymbol{\Theta}_t)\right]$ *exists and is non singular,*

(iv) Second-order partial derivatives of $S_t^{\xi}(\check{\mathbf{H}}_t;\boldsymbol{\Theta}_t)$ *with respect to* $\boldsymbol{\xi}$ *exist and for every* $\check{\mathbf{H}}_t$*,*

and satisfy $|\partial^2 S_t^{\xi}(\check{\mathbf{H}}_t;\boldsymbol{\Theta}_t)/\partial\boldsymbol{\xi}_i\partial\boldsymbol{\xi}_j| \leq \tilde{S}_t(\check{\mathbf{H}}_t)$ *for some integrable measurable function* $\tilde{S}_t$ *in a neighborhood of* $\bar{\boldsymbol{\xi}}$.

Assumption 2.6.8. *Functions* $m_2, m_{\omega_2}, m_{t\omega_2}$ $t = 2,3$ *are such that (i)* $\sup_{\vec{\mathbf{U}}} |m_s(\vec{\mathbf{U}})| < \infty$, *and (ii) the estimated functions* $\hat{m}_s$ *satisfy (ii)* $\sup_{\vec{\mathbf{U}}} |\widehat{m}_s(\vec{\mathbf{U}}) - m_s(\vec{\mathbf{U}})| = o_{\mathbb{P}}(1)$, $s \in \{2, \omega_2, 2\omega_2, 3\omega_2\}$.

Assumption 2.6.7 is standard for Z-estimators [see 61, Ch. 5.6]. Assumption 2.6.8 is the propensity score equivalent version of Assumption 2.6.2. Finally, we use $\boldsymbol{\psi}^{\xi}$ and and $\boldsymbol{\psi}^{\theta}$ to denote the influence function for $\widehat{\boldsymbol{\xi}}$, and $\widehat{\boldsymbol{\theta}}$ respectively. We are now ready to state our theoretical results for the value function estimator in equation (2.7). The proof, and the exact form of $\boldsymbol{\psi}^{\xi}$ can be found in Appendix B.2.2.

Theorem 2.6.9 (Asymptotic Normality for $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$)**.** *Under Assumptions 2.6.1-2.6.8,* $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ *defined in* (2.7) *satisfies*

$$\sqrt{n}\left\{\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}},\bar{\mu})\right]\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}}) + o_{\mathbb{P}}(1),$$

where

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}}) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathsf{SSL}_{\mathsf{DR}}}\right).$$

Here

$$\begin{aligned}
\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) =& \nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) + \boldsymbol{\psi}^{\theta}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\theta}}\int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}} \\
&+ \boldsymbol{\psi}^{\xi}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}\int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}, \\
\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) =& \omega_1(\check{\mathbf{H}}_1, A_1;\bar{\boldsymbol{\Theta}}_1)(1+\bar{\beta}_{21})\left\{Y_2 - \bar{\mu}^{v}_2(\vec{\mathbf{U}})\right\} + \omega_2(\check{\mathbf{H}}_2, A_2, \bar{\boldsymbol{\Theta}}_2)Y_3 - \bar{\mu}_{3\omega_2}(\vec{\mathbf{U}}) \\
&-\bar{\beta}_{21}\left\{\omega_2(\check{\mathbf{H}}_2, A_2, \bar{\boldsymbol{\Theta}}_2)Y_2 - \bar{\mu}_{2\omega_2}(\vec{\mathbf{U}})\right\} - Q^{o}_{2-}(\mathbf{H}_2;\bar{\boldsymbol{\theta}}_2)\left\{\omega_2(\check{\mathbf{H}}_2, A_2, \bar{\boldsymbol{\Theta}}_2) - \bar{\mu}_{\omega_2}(\vec{\mathbf{U}})\right\},
\end{aligned}$$

$\sigma^2_{\mathsf{SSL}_{\mathsf{DR}}} = \mathbb{E}\left[\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})^2\right]$, *and* $\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})$ *is as defined in* (2.5).

[Double Robustness of $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ as an estimator of $\bar{V}$] (a) If either $\|Q_t(\check{\mathbf{H}}_t, A_t;\widehat{\boldsymbol{\theta}}_t) - Q_t(\check{\mathbf{H}}_t, A_t)\|_{L_2(\mathbb{P})} \to 0$, or $\|\pi_t(\check{\mathbf{H}}_t;\widehat{\boldsymbol{\xi}}_t) - \pi_t(\check{\mathbf{H}}_t)\|_{L_2(\mathbb{P})} \to 0$ for $t = 1,2$, then under Assumptions 2.6.1-2.6.8, $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ satisfies

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} \xrightarrow{\mathbb{P}} \bar{V}.$$

(b) If $\|Q_t(\check{\mathbf{H}}_t, A_t; \widehat{\boldsymbol{\theta}}_t) - Q_t(\check{\mathbf{H}}_t, A_t)\|_{L_2(\mathbb{P})} \|\pi_t(\check{\mathbf{H}}_t; \widehat{\boldsymbol{\xi}}_t) - \pi_t(\check{\mathbf{H}}_t)\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ for $t = 1, 2$, then under Assumptions 2.6.1-2.6.8, $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ satisfies

$$\sqrt{n}\left(\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \bar{V}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\mathsf{SSL}_{\mathsf{DR}}}\right).$$

Next we define the supervised influence function for estimator $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$. Let $\boldsymbol{\psi}^{\theta}_{\mathsf{SUP}}$, be the influence function for the supervised estimator $\widehat{\boldsymbol{\theta}}_{\mathsf{SUP}}$ for model (2.1). The influence function for $\mathrm{SUP}_{\mathsf{DR}}$ Value Function Estimation estimator (2.5) and its variance is (see Theorem B.4.2 in Appendix B.4.1):

$$\begin{aligned}
\psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}}) =& \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}}) - \mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right] \\
&+ \boldsymbol{\psi}^{\theta}_{\mathsf{SUP}}(\mathbf{L})^{\intercal} \frac{\partial}{\partial \boldsymbol{\theta}} \int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \boldsymbol{\Theta}) d\mathbb{P}_{\mathbf{L}} \Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}} + \boldsymbol{\psi}^{\xi}(\mathbf{L})^{\intercal} \frac{\partial}{\partial \boldsymbol{\xi}} \int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \boldsymbol{\Theta}) d\mathbb{P}_{\mathbf{L}} \Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}, \\
\sigma^2_{\mathsf{SUP}_{\mathsf{DR}}} =& \mathbb{E}\left[\psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})^2\right].
\end{aligned}$$

The flexibility of our SSL value function estimator $V_{\mathsf{SSL}_{\mathsf{DR}}}$, allows the use of either supervised or SSL approach for estimation of propensity score nuisance parameters $\boldsymbol{\xi}$. For SSL estimation, we can use an approach similar to Section 2.4.2, [see 43, Ch. 2] for details. This can be beneficial in that we can then quantify the efficiency gain of $V_{\mathsf{SSL}_{\mathsf{DR}}}$ vs. $V_{\mathsf{SUP}_{\mathsf{DR}}}$ by comparing the asymptotic variances. In light of this, we assume SSL is used for $\boldsymbol{\xi}$ when estimating $V_{\mathsf{SSL}_{\mathsf{DR}}}$.

Before stating the result we discuss an additional requirement for the imputation models. As for Proposition 2.6.1, models $\bar{\mu}^{v}_{2}(\vec{\mathbf{U}})$, $\bar{\mu}^{v}_{\omega_2}(\vec{\mathbf{U}})$, $\bar{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}})$, $t = 2, 3$ need to satisfy a few additional constraints of the form

$$\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1) Q^{o}_{2-}(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_1)\{Y_2 - \bar{\mu}^{v}_{2}(\vec{\mathbf{U}})\}\right] = \mathbf{0}.$$

As there are several constraints, we list them in Appendix B.2.2, and condense them in

Assumption B.2.5, Appendix B.2.2. Again, one can construct estimators which satisfy such conditions by simply augmenting η_2^v, $\eta_{\omega_2}^v$, $\eta_{t\omega_2}^v$, $t = 2, 3$ in (2.6) with additional terms in the refitting step.

Under Assumptions 2.6.1-2.6.8, and B.2.5, asymptotic variances $\sigma^2_{\mathsf{SSL}_{\mathsf{DR}}}$, $\sigma^2_{\mathsf{SUP}_{\mathsf{DR}}}$ satisfy

$$\sigma^2_{\mathsf{SSL}_{\mathsf{DR}}} = \sigma^2_{\mathsf{SUP}_{\mathsf{DR}}} - \mathrm{Var}\left[\psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}}) - \psi^v_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right].$$

1) Proposition 2.6.2 illustrates how $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ is asymptotically unbiased if either the Q functions or the propensity scores are correctly specified.

2) An immediate consequence of Proposition 2.6.2 is that the semi-supervised estimator is at least as efficient (or more) as its supervised counterpart, that is $\mathrm{Var}\left[\psi_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \boldsymbol{\Theta})\right] \leq \mathrm{Var}\left[\psi_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \boldsymbol{\Theta})\right]$. As with Proposition 2.6.1, the difference in efficiency is explained by the information gain from incorporating surrogates.

3) To estimate standard errors for $V_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \bar{\boldsymbol{\Theta}})$, we will approximate the derivatives of the expectation terms $\frac{\partial}{\partial \bar{\boldsymbol{\Theta}}} \int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}}) d\mathbb{P}_{\mathbf{L}}$ using kernel smoothing to replace the indicator functions. In particular, let $\mathbb{K}_h(x) = \frac{1}{h}\sigma(x/h)$, σ defined as in (2.4), we approximate $d_t(\mathbf{H}_t, \boldsymbol{\theta}_2) = I(\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t > 0)$ with $\mathbb{K}_h(\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t)$ $t = 1, 2$, and define the smoothed propensity score weights as

$$\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \equiv \frac{A_1 \mathbb{K}_h(\mathbf{H}_{11}^\intercal \boldsymbol{\gamma}_1)}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{\{1 - A_1\}\{1 - \mathbb{K}_h(\mathbf{H}_{11}^\intercal \boldsymbol{\gamma}_1)\}}{1 - \pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}, \quad \text{and}$$
$$\tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta}) \equiv \tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \left[\frac{A_2 \mathbb{K}_h(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2)}{\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)} + \frac{\{1 - A_2\}\{1 - \mathbb{K}_h(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2)\}}{1 - \pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)}\right].$$

We simply replace the propensity score functions with these smooth versions in $\psi^v_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})$, detail is given in Appendix B.2.2. To estimate the variance we use the sample-split estimators:

$$\hat{\sigma}^2_{\mathsf{SSL}_{\mathsf{DR}}} = n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \psi^{v(-k)}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}_i; \widehat{\boldsymbol{\Theta}})^2.$$

2.7 Simulations and application to EHR data:

We perform extensive simulations to evaluate the finite sample performance of our method. Additionally we apply our methods to an EHR study of treatment response for patients with inflammatory bowel disease to identify optimal treatment sequence. These data have treatment response outcomes available for a small subset of patients only.

2.7.1 Simulation results

We compare our SSL Q-learning methods to fully supervised Q-learning using labeled datasets of different sizes and settings. We focus on the efficiency gains of our approach. First we discuss our simulation settings, then go on to show results for the Q function parameters under correct and incorrect working models for (2.1). We then show value function summary statistics under correct models, and mis-specification for the Q models in (2.1) and the propensity score function π_2 in (2.4).

Following a similar set-up as in [4], we first consider a simple scenario with a single confounder variable at each stage with $\mathbf{H}_{10} = \mathbf{H}_{11} = (1, O_1)^\intercal$, $\check{\mathbf{H}}_{20} = (Y_2, 1, O_1, A_1, O_1A_1, O_2)^\intercal$, and $\mathbf{H}_{21} = (1, A_1, O_2)^\intercal$. Specifically, we sequentially generate

$$
\begin{aligned}
&O_1 \sim \mathsf{Bern}(0.5), \qquad A_1 \sim \mathsf{Bern}(\sigma\left\{\mathbf{H}_{10}^\intercal \boldsymbol{\xi}_1^0\right\}), \qquad\qquad Y_2 \sim \mathcal{N}(\breve{\breve{X}}_1^\intercal \boldsymbol{\theta}_1^0, 1),\\
&O_2 \sim \mathcal{N}(\check{\mathbf{H}}_{20}^\intercal \boldsymbol{\delta}^0, 2), \quad A_2 \sim \mathsf{Bern}\left(\sigma\left\{\mathbf{H}_{20}^\intercal \boldsymbol{\xi}_2^0 + \xi_{26}^0 O_2^2\right\}\right), \quad \text{and} \quad Y_3 \sim \mathcal{N}(m_3\left\{\check{\mathbf{H}}_{20}\right\}, 2).
\end{aligned}
$$

where $m_3\{\check{\mathbf{H}}_{20}\} = \mathbf{H}_{20}^\intercal \boldsymbol{\beta}_2^0 + A_2(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2^0) + \beta_{27}^0 O_2^2 Y_2 \sin\{[O_2^2(Y_2+1)]^{-1}\}$. Surrogates are generated as $W_t = \lfloor Y_{t+1} + Z_t \rfloor$, $Z_t \sim \mathcal{N}(0, \sigma_{z,t}^2)$, $t = 1, 2$ where $\lfloor x \rfloor$ corresponds to the integer part of $x \in \mathbb{R}$. Throughout, we let $\boldsymbol{\xi}_1^0 = (0.3, -0.5)^\intercal$, $\boldsymbol{\beta}_1^0 = (1, 1)^\intercal$, $\boldsymbol{\gamma}_1^0 = (1, -2)^\intercal$ $\boldsymbol{\delta}^0 = (0, 0.5, -0.75, 0.25)^\intercal$, $\boldsymbol{\xi}_2^0 = (0, 0.5, 0.1, -1, -0.1)^\intercal$ $\boldsymbol{\beta}_2^0 = (.1, 3, 0, 0.1, -0.5, -0.5)^\intercal$, $\boldsymbol{\gamma}_2^0 = (1, 0.25, 0.5)^\intercal$.

We consider an additional case to mimic the structure of the EHR data set used for the real-data application. Outcomes Y_t are binary, and we use a higher number of covariates for the Q functions and multivariate count surrogates $\mathbf{W}_t$ $t = 1, 2$. Data is simulated with $\mathbf{H}_{10} = (1, O_1, \ldots, O_6)^\intercal$, $\mathbf{H}_{11} = (1, O_2, \ldots, O_6)^\intercal$, $\check{\mathbf{H}}_{20} = (Y_2, 1, O_1, \ldots, O_6, A_1, Z_{21}, Z_{22})^\intercal$, and

$\mathbf{H}_{21} = (1, O_1, \ldots, O_4, A_1, Z_{21}, Z_{22})^\intercal$, generated according to

$$\begin{aligned}
&\mathbf{O}_1 \sim \mathcal{N}(\mathbf{0}, I_6), && A_1 \sim \mathsf{Bern}(\sigma\{\mathbf{H}_{10}^\intercal \boldsymbol{\xi}_1^0\}), && Y_2 \sim \mathsf{Bern}(\sigma\{\breve{X}_1^\intercal \boldsymbol{\theta}_1^0\}),\\
&\mathbf{O}_2 = [I\{Z_1 > 0\}, I\{Z_2 > 0\}]^\intercal && A_2 \sim \mathsf{Bern}\left(\tilde{m}_2\{\breve{\mathbf{H}}_{20}\}\right), \quad \text{and} && Y_3 \sim \mathsf{Bern}(\tilde{m}_3\left\{\breve{\mathbf{H}}_{20}\right\}),
\end{aligned}$$

with $\tilde{m}_2 = \sigma\left\{\mathbf{H}_{20}^\intercal \boldsymbol{\xi}_2^0 + \tilde{\boldsymbol{\xi}}_2^\intercal \mathbf{O}_2\right\}$, $\tilde{m}_3(\breve{\mathbf{H}}_{20}) = \mathbf{H}_{20}^\intercal \boldsymbol{\beta}_2^0 + A_2(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2^0) + \tilde{\boldsymbol{\beta}}_2^\intercal \mathbf{O}_2 Y_2 \sin\{\|\mathbf{O}_2\|_2^2/(Y_2+1)\}$ and $Z_l = O_{1l}\delta_l^0 + \epsilon_z$, $\epsilon_z \sim \mathcal{N}(0,1)$ $l = 1, 2$. The dimensions for the Q functions are 13 and 37 for the first and second stage respectively, which match with our IBD dataset discussed in Section 2.7.2. The surrogates are generated according to $\mathbf{W}_t = \lfloor \mathbf{Z}_t \rfloor$, with $\mathbf{Z}_t \sim \mathcal{N}\left(\boldsymbol{\alpha}^\intercal(1, \mathbf{O}_t, A_t, Y_t), I\right)$. Parameters are set to $\boldsymbol{\xi}_1^0 = (-0.1, 1, -1, 0.1)^\intercal$, $\boldsymbol{\beta}_1^0 = (0.5, 0.2, -1, -1, 0.1, -0.1, 0.1)^\intercal$, $\boldsymbol{\gamma}_1^0 = (1, -2, -2, -0.1, 0.1, -1.5)^\intercal$, $\boldsymbol{\xi}_2^0 = (0, 0.5, 0.1, -1, 1, -0.1)^\intercal$, $\boldsymbol{\beta}_2^0 = (1, \boldsymbol{\beta}_1^0, 0.25, -1, -0.5)^\intercal$, $\boldsymbol{\gamma}_2^0 = (1, 0.1, -0.1, 0.1, -0.1, 0.25, -1, -0.5)^\intercal$, and $\boldsymbol{\alpha} = (1, \mathbf{0}, 1)^\intercal$.

For all settings, we fit models $Q_1(\mathbf{H}_1, A_1) = \mathbf{H}_{10}^\intercal \boldsymbol{\beta}_1^0 + A_1(\mathbf{H}_{11}^\intercal \boldsymbol{\gamma}_1^0)$, $Q_2(\breve{\mathbf{H}}_2, A_2) = \breve{\mathbf{H}}_{20}^\intercal \boldsymbol{\beta}_2^0 + A_2(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2^0)$ for the Q functions, $\pi_1(\mathbf{H}_1) = \sigma\left(\mathbf{H}_{10}^\intercal \boldsymbol{\xi}_1\right)$ and $\pi_2(\breve{\mathbf{H}}_2) = \sigma\left(\mathbf{H}_{20}^\intercal \boldsymbol{\xi}_2\right)$ for the propensity scores. The parameters ξ_{26}^0 and β_{27}^0 and $\tilde{\boldsymbol{\xi}}_2, \tilde{\boldsymbol{\beta}}_2$ index mis-specification in the fitted Q-learning outcome models and the propensity score models with a value of 0 corresponding to a correct specification. In particular, we set $\xi_{26}^0 = 1$, $\tilde{\boldsymbol{\xi}}_2 = \frac{1}{\|(1,\ldots,1)\|_2}(1, \ldots, 1)^\intercal$, and $\beta_{27}^0 = 1$, $\tilde{\boldsymbol{\beta}}_2 = \frac{1}{\|(1,\ldots,1)\|_2}(1, \ldots, 1)^\intercal$ for mis-specification of propensity score π_2 and Q_1, Q_2 functions respectively. We set $\xi_{26}^0 = \beta_{27}^0 = 0$ and $\tilde{\boldsymbol{\xi}} = \mathbf{0}$, $\tilde{\boldsymbol{\beta}}_2 = \mathbf{0}$ for correct model specification. Under mis-specification of the outcome model or propensity score model, the term omitted by the working models is highly non-linear, in which case the imputation model will be mis-specified as well. We note that our method does not need correct specification of the imputation model. For the imputation models, we considered both random forest (RF) with 500 trees and basis expansion (BE) with piecewise-cubic splines with 2 equally spaced knots on the quantiles 33 and 67 [62]. Finally, we consider two choices of (n, N): $(135, 1272)$ which are similar to the sizes of our EHR study and larger sizes of $(500, 10000)$. For each configuration, we summarize results based on $1,000$ replications.

We start discussing results under correct specification of the Q functions. In Table 2.1, we present the results for the estimation of treatment interaction coefficients $\bar{\gamma}_1, \bar{\gamma}_2$, under the

(a) $n = 135$ and $N = 1272$

	Supervised		Semi-Supervised									
			Random Forests					Basis Expansion				
Parameter	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
γ_{11}=1.4	-0.03	0.41	0.00	0.26	0.24	0.93	1.57	0.00	0.24	0.23	0.93	1.68
γ_{12}=-2.6	0.04	0.58	-0.01	0.36	0.34	0.94	1.61	-0.02	0.35	0.31	0.90	1.69
γ_{21}=0.8	0.00	0.34	0.01	0.21	0.20	0.93	1.61	0.00	0.20	0.19	0.94	1.71
γ_{22}=0.2	-0.02	0.45	-0.01	0.28	0.28	0.95	1.60	-0.01	0.27	0.26	0.94	1.70
γ_{23}=0.5	0	0.18	0.01	0.11	0.11	0.94	1.59	0.00	0.11	0.11	0.94	1.68

(b) $n = 500$ and $N = 10,000$

	Supervised		Semi-Supervised									
			Random Forests					Basis Expansion				
Parameter	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
γ_{11}=1.4	0.01	0.22	0.01	0.12	0.11	0.92	1.76	0.01	0.12	0.11	0.92	1.80
γ_{12}=-2.6	0	0.29	0	0.17	0.16	0.93	1.73	-0.01	0.16	0.15	0.93	1.80
γ_{21}=0.8	0.00	0.17	0.00	0.10	0.09	0.93	1.80	0.00	0.09	0.09	0.93	1.86
γ_{22}=0.2	-0.01	0.23	0	0.13	0.12	0.93	1.81	0	0.13	0.12	0.94	1.83
γ_{23}=0.5	0.00	0.09	0.00	0.05	0.05	0.94	1.78	0.00	0.05	0.05	0.95	1.81

Table 2.1: *Bias, empirical standard error (ESE) of the supervised and the SSL estimators with either random forest imputation or basis expansion imputation strategies for $\bar{\boldsymbol{\gamma}}_1, \bar{\boldsymbol{\gamma}}_2$ when (a) $n = 135$ and $N = 1272$ and (b) $n = 500$ and $N = 10,000$. For the SSL estimators, we also obtain the average of the estimated standard errors (ASE) as well as the empirical coverage probabilities (CovP) of the 95% confidence intervals.*

correct model specification, continuous outcome setting with $\beta_{27}^0 = \xi_{26}^0 = 0$. The complete tables for all $\bar{\boldsymbol{\theta}}$ parameters for the continuous and EHR-like settings can be found in Appendix B.1. We report bias, empirical standard error (ESE), average standard error (ASE), 95% coverage probability (CovP) and relative efficiency (RE) defined as the ratio of supervised ESE over SSL estimate ESE. Overall, compared to the supervised approach, the proposed semi-supervised Q-learning approach has substantial gains in efficiency while maintaining comparable or even lower bias. This is likely due to the refitting step which helps take care of the finite sample bias, both from the missing outcome imputation and Q function parameter estimation. Imputation with BE yields slightly better estimates than when using RF, both in terms of efficiency and bias. Coverage probabilities are close to the nominal level due to the good performance of the standard error estimation.

We next turn to Q-learning parameters under mis-specification of (2.1). Figure 2.1 shows the bias and root mean square error (RMSE) for the treatment interaction coefficients in the

2-stage Q functions. We focus on the continuous setting, where we set $\beta_{27}^0 \in \{-1, 0, 1\}$. Note that $\beta_{27}^0 \neq 0$ implies that both Q functions are mis-specified as the fitting of Q_1 depends on formulation of Q_2 as seen in (2.2). Semi-supervised Q-learning is more efficient for any degree of mis-specification for both small and large finite sample settings. As the theory predicts, there is no real difference in efficiency gain of SSL across mis-specification of the Q function models. This is because asymptotic distribution of $\widehat{\bar{\gamma}}_{\text{SSL}}$ shown in Theorems 2.6.4 & 2.6.6 are centered on the target parameters $\bar{\gamma}$. Thus, both SSL and SUP have negligible bias regardless of the true value of β_{27}^0.

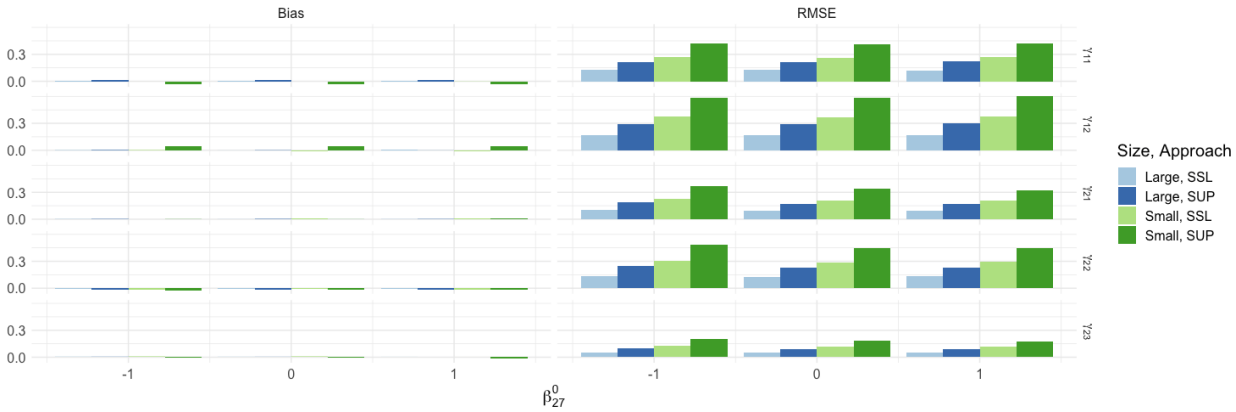


Figure 2.1: *Monte Carlo estimates of bias and RMSE ratios for estimation of* γ_{11}, γ_{12}, γ_{21}, γ_{22}, γ_{23} *under mis-specification of the Q-functions through* β_{27}^0. *Results are shown for the large* ($N = 10,000$, $n = 500$) *and small* ($N = 1,272$, $n = 135$) *data samples for the continuous setting over 1,000 simulated datasets.*

Next we analyze performance of the doubly robust value function estimators for both continuous and EHR-like settings. Table 2.2 shows bias and RMSE across different sample sizes, and comparing SSL vs. SUP estimators. Results are shown for the correct specification of the Q functions and propensity scores, and when either is mis-specified. Bias across simulation settings is relatively similar between $\widehat{V}_{\text{SSL}_{\text{DR}}}$ and $\widehat{V}_{\text{SUP}_{\text{DR}}}$, and appears to be small relative to RMSE. The low magnitude of bias suggests both estimators are robust to model mis-specification. There is an exception on the EHR setting with small sample size, for which the bias is non-negligible. This is likely due to the fact that the Q function parameters to estimate are 13+37, and the propensity score functions have 12 parameters which add up to

(a) $n = 135$ and $N = 1272$

			Supervised		Semi-Supervised									
					Random Forests					Basis Expansion				
Setting	Model	$\bar{V}$	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
Continuous	Correct	6.08	0.02	0.27	0.04	0.21	0.24	0.97	1.27	0.02	0.23	0.25	0.97	1.18
	Missp. Q	6.34	0.01	0.24	0.03	0.19	0.22	0.97	1.27	0.00	0.20	0.22	0.97	1.20
	Missp. π	6.08	0.01	0.28	0.02	0.22	0.24	0.97	1.24	0.01	0.25	0.25	0.97	1.12
EHR	Correct	1.38	0.09	0.15	0.05	0.12	0.12	0.94	1.24	0.04	0.13	0.12	0.95	1.12
	Missp. Q	1.43	0.09	0.14	0.04	0.12	0.12	0.96	1.12	0.03	0.14	0.12	0.95	1.02
	Missp. π	1.38	0.09	0.15	0.05	0.14	0.13	0.96	1.13	0.04	0.14	0.13	0.96	1.05

(b) $n = 500$ and $N = 10,000$

			Supervised		Semi-Supervised									
					Random Forests					Basis Expansion				
Setting	Model	$\bar{V}$	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
Continuous	Correct	6.08	0.02	0.15	0.03	0.11	0.12	0.96	1.32	0.02	0.13	0.13	0.95	1.16
	Missp. Q	6.34	0.01	0.13	0.03	0.10	0.10	0.96	1.31	0.01	0.11	0.11	0.96	1.16
	Missp. π	6.08	0.01	0.14	0.03	0.11	0.12	0.96	1.28	0.02	0.12	0.12	0.95	1.16
EHR	Correct	1.38	0.02	0.07	0.01	0.04	0.06	0.99	1.55	0.00	0.06	0.06	0.98	1.23
	Missp. Q	1.43	0.01	0.07	0.00	0.04	0.05	0.99	1.66	0.00	0.05	0.06	0.98	1.35
	Missp. π	1.38	0.02	0.08	0.01	0.06	0.07	0.99	1.22	0.00	0.07	0.07	0.97	1.03

Table 2.2: *Bias, empirical standard error (ESE) of the supervised estimator $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$ and bias, ESE, average standard error (ASE) and coverage probability (CovP) for $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ with either random forest imputation or basis expansion imputation strategies when (a) $n = 135$ and $N = 1272$ and (b) $n = 500$ and $N = 10,000$. We show performance and relative efficiency across both simulation settings for estimation under correct models, and mis-specification of Q function or propensity score function.*

a large number relative to the labeled sample size: $n = 135$. The SSL bias is lower in this case which could be due to the refitting step, which helped to reduce the finite sample bias. Efficiency gains of $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ are consistent across model specification. We next illustrate our approach using an IBD dataset.

2.7.2 Application to an EHR Study of Inflammatory Bowel Disease

Anti–tumor necrosis factor (anti-TNF) therapy has greatly changed the management and improved the outcomes of patients with inflammatory bowl disease (IBD) [63]. However, it remains unclear whether a specific anti-TNF agent has any advantage in efficacy over other agents, especially at the individual level. There have been few randomized clinical trials performed to directly compare anti-TNF agents for treating IBD patients [64]. Retrospective studies comparing infliximab and adalimumab for treating IBD have found limited and sometimes conflicting evidence of their relative effectiveness [65, 66, 67]. There is even less

evidence regarding optimal STR for choosing these treatments over time [68]. To explore this, we performed RL using data from a cohort of IBD patients previously identified via machine learning algorithms from the EHR systems of two tertiary referral academic centers in the Greater Boston metropolitan area [69]. We focused on the subset of $N = 1,272$ patients who initiated either Infliximab ($A_1 = 0$) or Adalimumab ($A_1 = 1$) and continued to be treated by either of these two therapies during the next 6 months. The observed treatment sequence distributions are shown in Table 2.3. The outcomes of interest are the binary indicator of treatment response at 6 months ($t = 2$) and at 12 months ($t = 3$), both of which were only available on a subset of $n = 135$ patients whose outcomes were manually annotated via chart review.

To derive the STR, we included gender, age, Charlson co-morbidity index [70], prior exposure to anti-TNF agents, as well as mentions of clinical terms associated with IBD such as bleeding complications extracted from the clinical notes via natural language processing (NLP) features as confounding variables at both time points. To improve the imputation of Y_t, we use 15 relevant NLP features such as mentions of rectal or bowel resection surgery as surrogates at $t = 1, 2$. We transformed all count variables using $x \mapsto \log(1 + x)$ to decrease skewness in the distributions, and centered continuous features. We used RF with 500 trees to carry out the imputation step, and 5-fold cross-validation (CV) to estimate the value function.

The supervised and semi-supervised estimates are shown in Table 2.4 for the Q-learning models and in Table 2.5 for the value functions associated with the estimated STR. Similar to those observed in the simulation studies, the semi-supervised Q-learning has more power to detect significant predictors of treatment response. Relative efficiency for almost all Q function estimates is near or over 2. The supervised Q-learning does not have the power to detect predictors such as prior use of anti-TNF agents, which are clearly relevant to treatment response [68]. Semi-supervised Q-learning is able to detect that the efficacy of Adalimumab wears off as patients get older, meaning younger patients in the first stage experienced a higher rate of treatment response to Adalimumab, a finding that cannot be detected with supervised Q-learning. Additionally, supervised Q-learning does not pick up that there is a higher rate of

response to Adalimumab among patients that are male or have experienced an abscess. This translates into a far from optimal treatment rule as seen in the cross-validated value function estimates. Table 2.5 reflects that using our semi-supervised approach to find the regime and to estimate the value function of such treatment rules yields a more efficient estimate, as the semi-supervised value function estimate $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$ yielded a smaller standard error than that of the supervised estimate $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$. However, the standard errors are large relative to the point estimates. On the upside, they both yield estimates very close in numerical value which is reassuring: both should be unbiased as predicted by theory and simulations.

		A_1	
		0	1
A_2	0	912	327
	1	27	183

Table 2.3: *Distribution of treatment trajectories for observed sample of size 1407.*

Stage 1 Regression							
	Supervised			Semi-Supervised			
Parameter	Estimate	SE	P-val	Estimate	SE	P-val	RE
Intercept	**0.424**	**0.082**	**0.00**	**0.518**	**0.028**	**0.00**	2.937
Female	-0.237	0.167	0.16	**-0.184**	**0.067**	**0.007**	2.514
Age	0.155	0.088	0.081	**0.18**	**0.034**	**0.00**	2.588
Charlson Score	0.006	0.072	0.929	-0.047	0.026	0.075	2.776
Prior anti-TNF	-0.038	0.06	0.524	**-0.085**	**0.019**	**0.00**	3.177
Perianal	**0.138**	**0.06**	**0.022**	**0.179**	**0.022**	**0.00**	2.688
Bleeding	0.049	0.08	0.54	0.058	0.03	0.055	2.675
A1	0.163	0.488	0.739	0.148	0.206	0.473	2.374
Female$\times A_1$	0.168	0.696	0.81	-0.042	0.287	0.886	2.424
Age$\times A_1$	-0.177	0.264	0.503	**-0.278**	**0.109**	**0.013**	2.418
Charlson Score$\times A_1$	0.136	0.391	0.728	0.195	0.178	0.276	2.194
Perianal$\times A_1$	-0.113	0.226	0.618	-0.019	0.08	0.808	2.838
Bleeding$\times A_1$	0.262	0.364	0.474	0.127	0.161	0.431	2.267

Stage 2 Regression							
	Supervised			Semi-Supervised			
Parameter	Estimate	SE	P-val	Estimate	SE	P-val	RE
Y_1	**0.37**	**0.11**	**0.00**	**0.55**	**0.05**	**0.00**	2.08
Intercept	0.08	0.06	0.17	0.04	0.02	0.14	2.40
Female	-0.01	0.10	0.92	-0.00	0.05	0.98	2.21
Age	0.05	0.06	0.35	**0.07**	**0.02**	**0.00**	2.33
Charlson Score	0.04	0.04	0.33	**0.06**	**0.02**	**0.01**	2.06
Prior anti-TNF	-0.05	0.05	0.29	**-0.09**	**0.02**	**0.00**	2.39
Perianal	-0.01	0.04	0.80	**-0.03**	**0.02**	**0.06**	2.31
Bleeding	-0.04	0.05	0.49	-0.03	0.03	0.29	2.14
A1	0.11	0.25	0.67	0.03	0.10	0.74	2.60
Abscess_2	0.06	0.04	0.16	**0.05**	**0.01**	**0.00**	2.68
Fistula_2	0.02	0.05	0.67	0.01	0.02	0.62	2.33
Female$\times A_1$	0.13	0.38	0.74	0.17	0.16	0.30	2.37
Age$\times A_1$	-0.02	0.12	0.88	-0.09	0.06	0.17	1.94
Charlson Score$\times A_1$	-0.02	0.16	0.89	0.04	0.07	0.55	2.19
Perianal$\times A_1$	-0.14	0.09	0.15	**-0.17**	**0.04**	**0.00**	2.34
Bleeding$\times A_1$	0.13	0.20	0.51	0.03	0.09	0.76	2.17
A2	0.07	0.17	0.69	**0.22**	**0.07**	**0.00**	2.55
Female$\times A_2$	-0.39	0.28	0.16	**-0.51**	**0.11**	**0.00**	2.53
Age$\times A_2$	0.09	0.10	0.40	**0.15**	**0.04**	**0.00**	2.27
Charlson Score$\times A_2$	0.01	0.07	0.84	-0.03	0.03	0.42	2.08
Perianal$\times A_2$	**0.20**	**0.09**	**0.04**	**0.23**	**0.04**	**0.00**	2.23
Bleeding$\times A_2$	0.03	0.08	0.77	0.02	0.04	0.49	2.34
$\text{Abscess}_2 \times A_2$	-0.13	0.07	0.06	**-0.09**	**0.03**	**0.00**	2.31
$\text{Fistula}_2 \times A_2$	-0.04	0.06	0.56	-0.03	0.03	0.36	2.17

Table 2.4: *Results of Inflammatory Bowel Disease data set, for first and second stage regressions. Fully supervised Q-learning is shown on the left and semi-supervised is shown on the right. Last columns in the panels show relative efficiency (RE) defined as the ratio of standard errors of the semi-supervised vs. supervised method, RE greater than one favors semi-supervised. Significant coefficients at the 0.05 level are in bold.*

	Estimate	SE
$\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$	0.851	0.486
$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$	0.871	0.397

Table 2.5: *Value function estimates for Inflammatory Bowel Disease data set, the first row has the estimate for treatment rule learned using $\mathcal{U}$ and its respective value function, the second row shows the same for a rule estimated using $\mathcal{L}$ and its estimated value.*

2.8 Discussion

We have proposed an efficient and robust strategy for estimating optimal dynamic treatment rules and their value function, in a setting where patient outcomes are scarce. In particular, we developed a two step estimation procedure amenable to non-parametric imputation of the missing outcomes. This helped us establish $\sqrt{n}$-consistency and asymptotic normality for both the Q function parameters $\widehat{\boldsymbol{\theta}}$ and the doubly robust value function estimator $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$. We additionally provided theoretical results which illustrate if and when the outcome-surrogates $\mathbf{W}$ contribute towards efficiency gain in estimation of $\widehat{\boldsymbol{\theta}}_{\mathsf{SSL}}$ and $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$. This lets us conclude that our procedure is always preferable to using the labeled data only: since estimation is robust to mis-specification of the imputation models, our approach is safe to use and will be at least as efficient as the supervised methods.

We focused on the 2-time point, binary action setting for simplicity but all our theoretical results and algorithms can be easily extended to a higher finite time horizon, and multiple actions with careful bookkeeping of notation. In practice, one would need to be careful with the variability of the IPW-value function which increases substantially with time. However, the SSL approach would come in handy to estimate propensity scores, providing an efficiency gain that would help stabilize the IPW in longer horizons.

We are interested in extending this framework to handle missing at random (MAR) sampling mechanisms. In the EHR setting, it is feasible to sample a subset of the data completely at random in order to annotate the records. Hence, we argue the MCAR assumption is true by design in our context. However, the MAR context allows us to leverage different data sources for $\mathcal{L}$ and $\mathcal{U}$. For example, we could use an annotated EHR data cohort and a large unlabeled

registry data repository for our inference, ultimately making the policies and value estimation more efficient and robust. We believe this line of work has the potential to leverage massive observational cohorts, which will help to improve personalized clinical care for a wide range of diseases.

Chapter 3

Estimating Optimal Dynamic Treatment Regimes with Smooth Surrogate Value Functions

Nilanjana Laha[1], Aarón Sonabend[1], Rajarshi Mukherjee and Tianxi Cai

Department of Biostatistics

Harvard University

3.1 Introduction

As electronic health records (EHR) become ubiquitous, healthcare data is increasingly available. These rich data capture heterogeneity in response to treatment over time and across patients, making it easier to adapt the treatment process to the evolving status of each person. Treatment decisions tailored to patient's individual characteristics are often referred to as dynamic treatment regimes (DTRs). These consist of a sequence of decisions informed by a patient's individual characteristics, as well as previous responses to treatment and clinical history. The

[1]Denotes equal contribution

objective is to maximize the counterfactual outcome, often referred to as the value function. EHR data are a vast, relatively untapped source of information from which optimal DTRs can be estimated.

There are several approaches for estimating DTRs [See 5, 54, for thorough discussions on DTR methodology and applications]. A popular paradigm consists of postulating outcome models and using them to optimize decisions sequentially. For example, Q-learning approaches focus on modeling the conditional outcome function at every timepont, then use approximate dynamic programming to find the optimal treatment rules [2, 38, 53]. Q-learning is a frequently employed and intuitive method, however it relies on correct specification of the conditional outcome models at every time point. This is highly unlikely to be the case in practice do to the sequential nature of the problem [71]. A family of methods, including A-learning and G-estimation address this by positing models for the propensity scores, thus offering doubly robust estimation and a higher efficiency in estimating the interaction effects [72**?**]. Although these approaches are less dependent on correct specification of the outcome regression models, they still require that the interaction effect, and either the propensity score or main effect models are correct. [See 4, for a detailed discussion on this]. This is an important limitation as the main quantity of interest lies in the value function, and these nuisance functions are often high-dimensional and easy to misspecify in the longitudinal context.

An alternative family of methods focus on directly maximizing the counterfactual value function to estimate the optimal treatment rule. One approach is to posit a model for the value function. If one is willing to assume a suitable stochastic class of treatment functions, the value function model can be differentiable with respect to the treatment rule and is thus directly optimizable. This is especially helpful in infinite horizon trajectories [73, 74]. However, a drawback of stochastic treatment rules is that their practical implementation is challenging in the clinical setting. Additionally, misspecification of the value function model can lead to sub-optimal DTRs. Working with a deterministic DTR class yields a discontinuous value function. To address this issue [75**?**] propose to use a classification approach to find the optimal DTR. Intuitively, given a patient's features at any given time, finding the optimal

treatment among a finite set which maximizes the outcome, is equivalent to predicting the correct treatment (class) for such features. In contrast to regression-based methods, this has the added benefit that there is no need for an outcome or value-function model. Thus, the approach does not rely on any model specification which is likely to be wrong in this setting. Additionally, [**?**] propose a simultaneous outcome weighted learning (SOWL) algorithm for the multi-stage problem. This method had the advantage that it can learn the optimal DTR simultaneously for all timepoints. SOWL can be understood as a solution for a multi-class classification problem.

To overcome the discontinuity of the objective function, [**?**] follow the classification literature by using a continuous surrogate to optimize the value function. They replace the discontinuous value function with a generalized surrogate hinge-loss which lends itself to standard optimization techniques. In their work, a slightly modified version of support vector machines (SVMs) are used to solve the classification problem for SOWL. They weigh the standard SVM dual objective function by a function of the outcomes. Then proceed by solving the problem with quadratic programming.

The paradigm shift of estimating DTRs by finding classification rules is a powerful idea. However, using the lack of smoothness in the hinge function raises some challenges in its use as a surrogate loss. SOWL through SVMs work best with small-sized sequential multiple assignment randomized trials (SMART). These sequential trials, small by nature, are where treatment assignment is randomized [76]. SOWL is suitable for such small data settings primarily due to the need for SVMs to optimize over the hinge-loss. As the optimization problem is solved in the dual space, SVMs are known to be slow and unable to scale well with large data. The computational complexity to find the solution in the dual space with n samples scales at $O(n^3)$ [77]. This is unfortunate as to best capture heterogeneity in patient response, there is a need for massive cohorts such as those found in large data repositories such as EHR.

In their work, [**?**] argue that this classification paradigm offers flexibility in the choice of loss functions. In line with their discussion and to address the practical challenges associated

with the generalized hinge loss, we discuss a general class of loss functions. In particular, this alternative surrogate class is composed of smooth, differentiable loss functions. The smoothness in our surrogate functions yields an algorithm which scales with sample size. We provide sufficient conditions for Fisher consistency and a rate on the regret bound consistent with classification literature. Our proposed class of surrogate loss functions make the optimization problem suitable to a large range of standard machine learning algorithms. We demonstrate this using neural networks, wavelet series and natural cubic splines. Coupled with different choices of surrogate functions within our proposed class, our approach offers flexibility for learning the DTRs so that practitioners can tailor their methods to the data and problem at hand.

The rest of the chapter is organized as follows. In Section 3.2 we outline the problem, discuss the mathematical formulation and implementation. In Section 3.3 we discuss calibration and Fisher consistency in the binary classification setting and generalize these concepts to the DTR setting. We continue to characterize the family of surrogate value functions and establish associated theoretical results. We continue to discuss the regret bound for the value function in Section 3.4. In particular we analyze the approximation and generalization error for the general setting. We follow by showing bounds for when estimating the DTRs using neural networks or wavelet series. Then in section 3.5 we illustrate our method's empirical performance with extensive simulations and an application to a Sepsis cohort. We continue with a discussion in Section 3.6. In Section 3.7 we discuss in detail our theoretical results for the the regret bound rate when using neural networks and wavelet series. Appendices C.1-C.4, collect the proofs of our results and some necessary technical lemmas.

3.2 Problem and Methodology

We consider a longitudinal setting where we denote by O_t the patient state at time t. This is followed by a binary treatment decision $A_t \in \{\pm 1\}$, and a response to such treatment Y_{t+1}. Throughout we assume higher values of response Y_{t+1} are desirable. For simplicity, we will focus on a two-stage setting: $(O_1, A_1, Y_2, O_2, A_2, Y_3)$. In the clinical setting it is usual to make

decisions based on all previous states, treatments and the associated responses. We therefore define the patient history by

$$H_1 = O_1, \text{ and } H_2 = (O_1, A_1, O_2, Y_2),$$

where H_1 and H_2 take values in sets $\mathcal{X}_1 \subset \mathbb{R}^{dim(H_1)}$ and $\mathcal{X}_2 \subset \mathbb{R}^{dim(H_2)}$, respectively.

The goal is to find the treatment sequence that maximizes the expected sum of rewards $Y_2 + Y_3$. To formalize this, consider the decision rules

$$d_1 : \mathcal{X}_1 \mapsto \{\pm 1\}, \quad \text{and} \quad d_2 : \mathcal{X}_2 \mapsto \{\pm 1\}.$$

We seek to find decision rule $d = (d_1, d_2)$ which maximizes the value function $\mathbb{E}_{d_1,d_2}(Y_2 + Y_3)$. Here $\mathbb{E}_d$ is the expectation under the measure P_d generated by observations under treatments determined by the rule d. The optimal decision rules are

$$(d_1^*, d_2^*) = \underset{d_1,d_2 \mathbb{E}_{d_1,d_2}(Y_2+Y_3),}{\operatorname{argmax}}$$

and in particular

$$d_1^*(H_1) = \underset{a_1 \in \{\pm 1\} \mathbb{E}\left[Y_2 + \max_{a_2 \in \{\pm 1\}} \mathbb{E}[Y_3 | A_1 = a_1, A_2 = a_2, H_1, Y_2, O_2] \middle| A_1 = a_1, H_1 \right]}{\operatorname{argmax}}$$

$$d_2^*(H_2) = \max_{a_2 \in \{\pm 1\}} \mathbb{E}[Y_3 | A_2 = a_2, H_2].$$

Since the optimal decision rules remain unchanged if a constant c is added to both Y_2 and Y_3, we can assume without loss of generality that $Y_2, Y_3 > 0$.

We use $\pi_1(a_1|H_1)$, $\pi_2(a_2|H_2)$ to denote the propensity scores $P(A_1 = a_1|H_1)$, $P(A_2 = a_2|H_2)$ respectively. Assuming (i) sequential randomization of $A_t|H_t$ $t = 1, 2$, and (ii) bounded known propensity scores: $P(A_t = 1|H_t) > \nu$ for $\nu > 0$, we can express the expected sum or rewards under treatment rule d, with the observed data, by using importance sampling [**?**]. Therefore we can re-write $\mathbb{E}_d(Y_2 + Y_3)$ as

$$\mathbb{E}_d(Y_2 + Y_3) = \mathbb{E}\left[\Big(Y_2 + Y_3 \Big) \frac{I\{A_1 = d_1(H_1)\} I\{A_2 = d_2(H_2)\}}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)} \right].$$

We use $\mathbb{E}$ without any subscript to denote expectation under the measure generated by observations under the propensity scores. As written, the objective function $\mathbb{E}_d(Y_2 + Y_3)$ is discontinuous and non-convex so unless H_1 and H_2 are discrete and low dimensional, this is an intractable optimization problem. Thus, directly searching for (d_1^*, d_2^*) is difficult in practice. However, by characterizing a class of functions:

$$\mathcal{F} = \left\{ (f_1, f_2) \;\middle|\; f_1 : \mathcal{H}_1 \mapsto \mathbb{R}; \quad f_2 : \mathcal{H}_2 \times \{\pm 1\} \mapsto \mathbb{R} \text{ are measurable} \right\}, \tag{3.1}$$

we can define a new objective in terms of functions in $\mathcal{F}$:

$$V(f_1, f_2) \equiv \mathbb{E}\left[\Big(Y_2 + Y_3\Big) \frac{I\{A_1 = sign(f_1(H_1))\} I\{A_2 = sign(f_2(H_2))\}}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)} \right].$$

It can be shown that solving for $(f_1^*, f_2^*) = \text{argmax}_{(f_1,f_2)\in\mathcal{F}} V(f_1, f_2)$, yields the optimal treatment rule:

$$d_t^*(H_t) = sign(f_t^*(H_t)), \quad t = 1, 2.$$

This maximization problem resembles a two-stage extension of the classical binary classification problem [78], or a multi-label classification problem. To find the optimal DTR, we can solve $\text{argmax}_{(f_1,f_2)\in\mathcal{F}} V(f_1, f_2)$. However, this is still a challenging problem in practice. The reason is that the non-convexity and discontinuity of $V(f_1, f_2)$ caused by the indicator functions deems the maximization problem NP-hard. This leads us to look for alternate solutions.

The traditional classification literature suggests using a concave surrogate to replace the indicator 0-1 loss functions in $V(f_1, f_2)$. Suitable choices of ψ make the problem tractable for optimization and can yield optimal DTRs and fast estimation rates. In light of this, for a pre-specified function $\psi : \mathbb{R}^2 \mapsto \mathbb{R}$, we define

$$V_\psi(f_1, f_2) = \mathbb{E}\left[\frac{(Y_2 + Y_3)\psi\Big(A_1 f_1(H_1), A_2 f_2(H_2)\Big)}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)} \right]. \tag{3.2}$$

Note, however, that we do not have access to the unknown distribution to solve the maximization

problem above, thus we consider the empirical version instead:

$$\widehat{V}_{\psi}(f_1, f_2) = \mathbb{P}_n \left[\frac{(Y_2 + Y_3)\psi\Big(A_1 f_1(H_1), A_2 f_2(H_2)\Big)}{\pi_1(A_1|H_1)\pi_2(A_2|H_2)} \right]. \tag{3.3}$$

We will focus on characterizing a family of smooth surrogate functions ψ which will yield desirable theoretical properties on the treatment rules found using the objective function in (3.3).

3.2.1 Gradient Descent for DTR

Optimizing the empirical surrogate value function $\widehat{V}_{\psi}(f_1, f_2)$ defined in (3.3) requires us to parametrize functions $f_t(H_t)$, $t = 1, 2$. As we restrict to using smooth surrogates ψ, our framework allows for a wide range of parametrizations $f_t(H_t; \theta_t)$. We only require that they are differentiable with respect to $\theta = (\theta_1^{\intercal}, \theta_2^{\intercal})$. For example, we can use linear functions $f(H_t; \theta_t) = \varphi(H_t)^{\intercal}\theta_t$ on any type of pre-specified basis expansion $\varphi(\cdot)$ such as splines or wavelets. We can alternatively use neural networks for $f(H_t; \theta_t)$. For any such choice, given an observed dataset, the optimization problem can be written as

$$\max_{\theta} \quad \widehat{V}_{\psi}(f_1(H_1; \theta_1), f_2(H_2; \theta_2)).$$

Using a smooth ψ ensures all derivatives are well defined, allowing us to solve the above optimization problem with gradient descent. Starting with an initial value $\theta^{(0)}$ and a step $\eta \in (0, 1)$, an update at the k^{th} iteration is given by

$$\theta^{(k+1)} \coloneqq \theta^{(k)} - \eta \frac{\partial \widehat{V}_{\psi}}{\partial \psi} \frac{\partial \psi(A_1 f_1, A_2 f_2)}{\partial (f_1(H_1; \theta_1), f_2(H_2; \theta_2))} \frac{\partial (f_1, f_2)}{\partial \theta}.$$

This framework is simple and can be easily implemented with standard libraries for machine learning or optimization. It can also be implemented with stochastic gradient descent and variations of this algorithm such as Nesterov accelerated gradient, Adaptive Gradient Algorithm, root mean square propagation (RMSprop), etc. [79, 80, 81].

3.3 Calibration and Fisher Consistency

We now have a tractable optimization problem defined by the objective function $V_\psi(f_1, f_2)$ and have shown how we can obtain a solution. However, we are actually interested in maximizing $V(f_1, f_2)$, thus we need to ensure our solutions from the surrogate objective function satisfy Fisher consistency. This means that the replacement of the indicator product by ψ yields the same solutions obtained by directly maximizing $V(f_1, f_2)$. To formalize this concept we denote

$$(\tilde{f}_1, \tilde{f}_2) = \underset{(f_1,f_2)\in\mathcal{F} V_\psi(f_1,f_2),}{\operatorname{argmax}}$$

where $\mathcal{F}$ is as defined in (3.1). Fisher consistency in this context implies $(\tilde{f}_1, \tilde{f}_2) = (f_1^*, f_2^*) = \operatorname{argmax}_{(f_1,f_2)\in\mathcal{F} V(f_1,f_2)}$. We define the t-stage decision rule by $\tilde{d}_t(H_t) = sign(\tilde{f}_t(H_t))$ and 1 when $\tilde{f}_t(H_t) = 0$, $t = 1, 2$. Fisher consistency guarantees that the DTR found using our surrogate value function $(\tilde{d}_1, \tilde{d}_2)$ is consistent with the optimal DTR (d_1^*, d_2^*) for the population value function V^*.

One possible choice for ψ is the generalized hinge-loss. [**?**] show that $\psi(x, y) = \min(x - 1, y - 1, 0) + 1$ is Fisher consistent. However, as previously discussed, this can cause issues in practice as it is not scalable with sample size.

In this section we propose an alternative family of smooth surrogate functions ψ. We motivate the class of smooth surrogates ψ by analyzing the properties of $\tilde{f}_1(H_1)$ and $\tilde{f}_2(H_2)$ needed to ensure consistency. We follow with some discussion on Fisher consistency and characterize the class of surrogates that we will consider. In order to motivate our proposed class, we first introduce the concept of calibration, and discuss what makes a loss function calibrated in the simpler binary classification regime. Since the problem at hand shares a deep connection with binary classification, it is natural to expect that the calibrated classes for these two regimens will share some similarities.

3.3.1 Binary Classification Calibration

In the binary setting it is useful to first understand the concept of classification-calibration. Consider the case where we have $Y \in \{-1, 1\}$, $X \in \mathcal{X}$, and define $\varphi(x) \equiv P(Y = 1|X = x)$. A

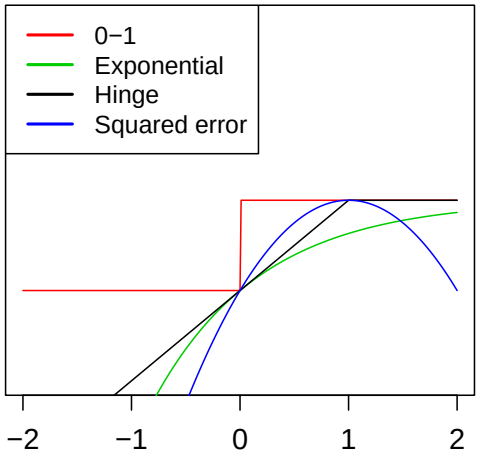


Figure 3.1: *Plots of $\phi(x)$ vs x for concave calibrated value function ϕ for binary decision rules. Here are the functions, 0-1: $\phi(x) = 1[x > 0]$, Exponential: $\phi(x) = 1 - e^{-x}$, Hinge: $\phi(x) = \min(x, 1)$, Squared error: $\phi(x) = 1 - (1 - x)^2$.*

function ϕ is calibrated if every solution $f^* = \text{argmin}_{f\mathbb{E}[\phi(Yf(X))|X=x]}$ has the same sign as the Bayes rule: $2\varphi(x) - 1$, for any x such that $\varphi(x) \neq 1/2$. Intuitively, calibration ensures that the classification rule given by $sign(f^*(x))$ is consistent with the one which targets the 0-1 loss: $\mathbb{E}[I\{Y = sign(f(X))\}|X = x]$ at every x. Thus, classification-calibration is a form of pointwise Fisher consistency [78, 82].

In the binary setting, a univariate concave value function ϕ is classification-calibrated if and only if it is differentiable at 0 with positive derivative [see 78, Theorem 6]. Many commonly used functions for classification satisfy these conditions. We show some of these in Figure 3.1. [See 83, for a study on using smooth surrogates for estimation of optimal single-stage DTRs.]

To develop further understanding regarding classification calibration, we direct the readers to an interesting geometric property of these functions. Any such function mimics the graph of the 0-1 loss function. After proper shifting and scaling, its image lies below that of the 0-1 loss function (see Figure 3.1). Of course, concavity is not necessary for classification-calibration, and this geometric property is shared by non-concave classification calibrated losses as well [see Lemma 9 of 78].

3.3.2 Fisher Consistency for the Value Function

To generalize the Fisher consistency to the multi-stage setting we will consider functions of form $\psi(x, y) = \phi_1(x)\phi_2(y)$, where ϕ_1 and ϕ_2 are smooth. Interestingly, the sufficient conditions to attain calibration in the binary case do not directly generalize to the multi-stage setting. In the following we derive the conditions on ϕ_1, ϕ_2 so that ψ is Fisher consistent.

It is easy to see that even if we impose concavity on functions ϕ_1, ϕ_2, their product ψ will not necessarily be concave. Therefore, we cannot use the binary calibration results directly. However, from the preceding section, we know that if, say ϕ_2, is positive on the real line and approximates the $0-1$ loss function, then ϕ_2 will be calibrated. An example of this is the sigmoid function: $\phi_2(x) = (1 + e^{-x})^{-1}$. This is a differentiable, increasing function with limits 0 and 1 respectively for $x \to -\infty$, $x \to \infty$. In fact, any function which behaves like the sigmoid function in this manner will satisfy calibration. The disadvantage of such functions that satisfy the above requirements is that they can not be concave, thus $\psi(x, y) = \phi_1(x)\phi_2(y)$ will necessarily be non-concave. However, sigmoid-type functions suffice to guarantee Fisher consistency, and thus we can sacrifice concavity of ψ.

We can summarize the characteristics of sigmoid-type functions as functions ϕ which satisfy

$$\sup_{x\in\mathbb{R}} \Big(\eta\phi_2(x) + (1-\eta)\phi_2(-x)\Big) = C\max\{\eta, 1-\eta\}.$$

This ensures calibration for ψ. We note that the above property is not exclusive to sigmoid-type functions. Other non-smooth functions like $\phi_2 = \max\{1-x, 0\}$ satisfy this property, namely, the hinge loss, which is in fact concave.

We formalize these concepts in the following conditions. $\phi(x)$ is calibrated for the 0-1 loss:

$$\sup_{x:x(2\eta-1)\le 0} \Big(\eta\phi(x) + (1-\eta)\phi(-x)\Big) < \sup_{x\in\mathbb{R}} \Big(\eta\phi(x) + (1-\eta)\phi(-x)\Big)$$

for all $\eta \in [0, 1]$. ϕ satisfies Condition 3.3.2. Moreover,

$$\sup_{x\in\mathbb{R}} \Big(\eta\phi(x) + (1-\eta)\phi(-x)\Big) = C\max\{\eta, 1-\eta\}$$

for all $\eta \in [0, 1]$.

Condition 3.3.2 requires ϕ to be calibrated in the sense of [78]. Thus, any concave function ϕ that is differentiable at 0 with $\phi'(0) > 0$, satisfies Condition 3.3.2 [78, Theorem 6]. These functions are commonly used in binary classification [83] as well as sequential classification procedures for dynamic treatment regimen [84]. Some common examples are exponential, $\phi(x) = -\exp(-x)$, logistic, etc.

As Condition 3.3.2 might be hard to verify for a given function, in the next lemma we provide a way of finding functions that satisfy Condition 3.3.2. The proof is found in Appendix C.1.

Lemma 3.3.1. *Suppose ϕ is a monotonous smooth function such that*

1. *$\phi(x) > 0$ for all $x \in \mathbb{R}$.*
2. *For all $x \in \mathbb{R}$, $\phi(x)$ satisfies $\phi(x) + \phi(-x) = C$ where $C > 0$.*
3. *$\lim_{x\to\infty} \phi(x) = C$.*
4. *$\lim_{x\to-\infty} \phi(x) = 0$.*

Then ϕ satisfies Condition 3.3.2.

We can easily construct a function ϕ which satisfies the requirements for Lemma 3.3.1, by applying a location shift of a bounded odd function g. In Corollary 3.3.1.1, we provide some examples of functions that are part of this surrogate family.

Corollary 3.3.1.1. *The following odd functions are non-decreasing and have range $[-1, 1]$. Thus $\phi(x) = 1 + g(x)$ satisfies Lemma 3.3.1 with $C = 2$.*

(i) $g(x) = \frac{x}{1+|x|}$.

(ii) $g(x) = \frac{2}{\pi} \arctan\left(\frac{\pi x}{2}\right)$.

(iii) $g(x) = \frac{x}{\sqrt{1+x^2}}$.

(iv) $g(x) = 2(1 + e^{-x})^{-1} - 1$.

Corollary 3.3.1.1 lists some useful examples which will determine the rates at which the optimal DTRs can be estimates. However there are evidently a wide range of functions which satisfy the conditions required by Lemma 3.3.1. Next we state a desirable results for any surrogate function characterized by the above conditions. We defer the proof to Appendix C.1.

Lemma 3.3.2. *Suppose $\psi(x,y) = \phi_1(x)\phi_2(y)$ where ϕ_1 satisfies Condition 3.3.2 and ϕ_2 satisfies Condition 3.3.2. Further suppose $(\tilde{f}_1, \tilde{f}_2)$ is the maximizer of $V_\psi(f_1, f_2)$ defined in* (3.2) *over all measurable functions $f_1 : \mathcal{X}_1 \mapsto \mathbb{R}$ and $f_2 : \mathcal{X}_2 \mapsto \mathbb{R}$. Then*

$$\begin{aligned} \tilde{d}_1(H_1) = sign(\tilde{f}_1(H_1)) = & \underset{a_1 \in \{\pm 1\}}{argmax} \mathbb{E}\left[Y_2 + U_2^*(H_2) \middle| a_1, H_1 \right], \\ \tilde{d}_2(H_2) = sign(\tilde{f}_2(H_2)) = & \underset{a_2 \in \{\pm 1\}}{argmax} \mathbb{E}[Y_3 | a_2, H_3], \end{aligned} \tag{3.4}$$

where $U_3^*(H_2) = argmax_{a_2 \in \{\pm 1\}} \mathbb{E}[Y_3 | a_2, H_2]$.

Note that using Lemma 3.3.2 we can simply plug the optimal treatment for the first stage: $\tilde{d}_1(H_1)$ as the first treatment in H_2, which yields

$$\text{sign}\Big(\tilde{f}_2(O_1, \tilde{d}_1(H_1), Y_2, O_2)\Big) = \underset{a_2 \in \{\pm 1\}}{\text{argmax}} \mathbb{E}[Y_3 | O_1, \tilde{d}_1(H_1), Y_2, O_2].$$

Thus, whenever they are unique for all $(H_1, H_2) \in \mathcal{X}_1 \times \mathcal{X}_2$,

$$\tilde{d}_1(H_1) = d_1^*(H_1) \quad \text{and} \quad \tilde{d}_2(H_2) = d_2^*(H_2).$$

Therefore, the population solution for the DTR associated with surrogate value function $V_\psi(f_1, f_2)$ is the same as the solution for the optimal DTR in the the 0-1 value function V^*. Notice that Lemma 3.3.2 is a result for $(\tilde{f}_1, \tilde{f}_2) = \text{argmax}_{(f_1,f_2) \in \mathcal{F} V_\psi(f_1,f_2)}$ where $\mathcal{F}$ consists of all measurable functions. We next discuss a feasible sub-class of estimable functions, define the regret bound, and analyze convergence rates over different function classes.

3.4 Regret Bound for the Value Function

In the previous section we have provided results which state that by optimizing over the surrogate value function we can recover the optimal DTR for the 0-1 value function. We next discuss the approximation and estimation of error rates associated with our value function of interest $V(f_1, f_2)$ and surrogate value function $V_\psi(f_1, f_2)$. In this context it is useful to define excess risk in line with the excess risk in the context of classification. Recalling the optimal value and surrogate functions are $V^* \equiv V(f_1^*, f_2^*)$, and $V_\psi^* \equiv V_\psi(\tilde{f}_1, \tilde{f}_2)$, we define the regret and excess ψ-regret of using $(f_1, f_2) \in \mathcal{F}$ respectively by

$$V^* - V(f_1, f_2), \quad \text{and} \quad V_\psi^* - V_\psi(f_1, f_2).$$

Note that the regret and ψ-regret are always non-negative. As discussed in Section 3.2, although attaining V_ψ^* is possible if one can recover $\tilde{f}_1$ and $\tilde{f}_2$, it is intractable owing to our inability to search over $\mathcal{F}$. In practice, one considers optimization over a nested class which can depend on sample size n:

$$\mathcal{H}_1 \subset \ldots \subset \mathcal{H}_n \subset \mathcal{F}.$$

Clearly, higher sample size provides a richer $\mathcal{H}_n$; examples include reproducible kernel Hilbert space with dimension n, and the wavelet series class defined as a subspace of a Hölder space. The latter will be discussed in detail in Section 3.7.2.

Before we state our result for the regret bound, we motivate it by discussing the binary setting. Consider data of the form $(Y, X) \in \{\pm 1\} \times \mathcal{X}$ as in Section 3.3.1. In line with the value function notation, we define the risk and the surrogate ψ-risk as $R(f) = \mathbb{E}[I\{Yf(X) < 0\}]$, $R_\phi(f) = \mathbb{E}[\phi(Yf(X))]$ respectively. Also denote the optimal values respectively as $R^* = \inf_f R(f)$, $R_\phi^* = \inf_f R_\phi(f)$. In the binary classification setting, [78] proved existence of a function $g : [0, 1] \mapsto \mathbb{R}^+$ so that

$$g(R(f) - R^*) \leq R_\phi(f) - R_\phi^* \tag{3.5}$$

for any loss function ϕ. This g is a non-negative concave function, and vanishes at zero.

Moreover, when ϕ is calibrated, [78] proved g to be strictly monotone on $[0, 1]$, which guarantees the existence of the inverse $g_{\mathcal{T}}^{-1}$. Also, it can be shown that ϕ is calibrated if and only if g is positive on $(0, 1]$.

[85] generalized the above result for a wide range of methods encompassing binary, multicategory, and multilabel classification. In this process, he modified the definition of $g_{\mathcal{T}}$ to adjust for the additional complexity arising in the multivariate settings.

The benefit of a bound as in (3.5) is two-fold. First, observe that, since $g_{\mathcal{T}}(0) = 0$ and $g_{\mathcal{T}}$ is positive on $(0, 1]$ for calibrated ψ, the bound readily implies Fisher consistency under calibration. [86] used the results of [85] to show classification calibration results in Fisher consistency in multilabel classification settings. Second, one can estimate the rate of convergence of $R(\hat{f}_n) - R^*$ from that of $R_\psi(\hat{f}_n) - R_\psi^*$ using this $g_{\mathcal{T}}$-transform. [78] initiated this topic for binary settings, although a more thorough treatment can be found in [87]. Similar results can be shown in multicategory classification as well [85].

In the next theorem, we develop bounds on the regret $V^* - V(f_1, f_2)$ using the construction in [85, 87]. We first state our required assumptions:

Assumption 3.4.1. *Outcomes Y_2, Y_3 and propensity score functions π_1, π_2 are such that $\max(Y_2 + Y_3) \leq C_y$ and $\pi_1, \pi_2 > C_\pi$ for some $C_y, C_\pi > 0$.*

Assumption 3.4.1 is standard in the inverse probability weighing literature as it ensures the optimal rule is identifiable. Additionally we assume outcomes are bounded. We now give a relationship between regrets, the proof is found in Appendix C.2.1.

Theorem 3.4.2. *Suppose $\psi(x, y) = \phi(x)\phi(y)$ where ϕ is increasing and satisfies the conditions of Lemma 3.3.1. Then under assumption 3.4.1, for $C > 0$*

$$V^* - V(f_1, f_2) \leq \frac{C}{\min(1, \phi(0))}\Big(V_\psi^* - V_\psi(f_1, f_2)\Big).$$

We have established that we can bound the value function regret $V(f_1^*, f_2^*) - V(f_1, f_2)$ by the surrogate value regret $V_\psi(f_1^*, f_2^*) - V_\psi(f_1, f_2)$. Therefore, it suffices to focus on solving for $(\tilde{f}_1, \tilde{f}_2) = \text{argmax}_{(f_1,f_2)\in\mathcal{F}} V_\psi(f_1, f_2)$ to find the optimal DTR. Theorem 3.4.2 implies Fisher consistency of the family of functions $\psi(x, y) = \phi_1(x)\phi_2(y)$ characterized by Conditions 3.3.2

& 3.3.2. We next restrict the class of functions, as $\mathcal{F}$ is a very large class of functions making the problem intractable.

In practice we will search over a smaller class of functions $\mathcal{H}_n \subset \mathcal{F}$. There are numerous statistical learning methods that can optimize over a smooth surrogate function. We will first develop general results for the regret bound using estimated functions $\widehat{f}_{n,1}, \widehat{f}_{n,2}$ such that $(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \in \mathcal{H}_n$. We then go on to analyze the rates for function estimation using function spaces $\mathcal{H}_n$, suitable for estimation with neural networks and wavelet basis. We will ensure this restricted class is rich enough so that $\mathcal{H}_n \to \mathcal{F}$, meaning our sub-class will tend to the class of all measurable functions $\mathcal{F}$ on $\mathcal{X}_1 \times \mathcal{X}_2$.

To analyze the ψ-regret, recall $V_\psi^* = \sup_{f_1,f_2} V_\psi^*(f_1, f_2)$. We can decompose it as follows:

$$
\begin{aligned}
&V_\psi^* - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \\
&\quad = V_\psi^* - \sup_{(f_1,f_2)\in\mathcal{H}_n} V_\psi(f_1, f_2) + \sup_{(f_1,f_2)\in\mathcal{H}_n} \Big(V_\psi(f_1, f_2) - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \Big) \\
&\quad = \underbrace{\inf_{(f_1,f_2)\in\mathcal{H}_n} \Big(V_\psi^* - V_\psi(f_1, f_2) \Big)}_{T_1} + \underbrace{\sup_{(f_1,f_2)\in\mathcal{H}_n} \Big(V_\psi(f_1, f_2) - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \Big)}_{T_2}.
\end{aligned}
$$

Term T_1 is known as the approximation error which comes from searching over the space of functions $\mathcal{H}_n$ instead of $\mathcal{F}$. Term T_2 is known as the generalization or estimation error due to estimating functions $(\hat{f}_{n,1}, \hat{f}_{n,2})$ with a finite dataset. We next discuss each of them separately for generic function class $\mathcal{H}_n$. We then show results for when $\mathcal{H}_n$ are suitable function classes for estimation with neural networks, and with wavelet series.

3.4.1 Approximation Error

We start by inspecting the approximation error

$$
T_1 = \inf_{(f_1,f_2)\in\mathcal{H}_n} \Big(V_\psi^* - V_\psi(f_1, f_2) \Big)
$$

with our surrogate family $\psi(x, y) = \phi(x)\phi(y)$, were ϕ satisfies Condition 3.3.2. To provide some guidance, we consider the classification problem for the second stage. For a patient with clinical history H_2, the optimal treatment will be $\tilde{d}_2(H_2) = 1$ if $\mathbb{E}[Y_2 + Y_3 | A_2 = 1, H_2] >$

$\mathbb{E}[Y_2 + Y_3|A_2 = -1, H_2]$ and $\tilde{d}_2(H_2) = -1$ otherwise. Building on this concept, it is useful to define functions $G_1 : \mathcal{X}_1 \mapsto \mathbb{R}$ and $G_2 : \mathcal{X}_2 \mapsto \mathbb{R}$ as

$$\begin{aligned} G_1(H_1) &= \frac{\mathbb{E}[Y_2 + U_3^*(H_2)|A_1 = 1, H_1]}{\mathbb{E}[Y_2 + U_3^*(H_2)|A_1 = 1, H_1] + \mathbb{E}[Y_2 + U_3^*(H_2)|A_1 = -1, H_1]} - \frac{1}{2}, \\ G_2(H_2) &= \frac{\mathbb{E}[Y_2 + Y_3|A_2 = 1, H_2]}{\mathbb{E}[Y_2 + Y_3|A_2 = 1, H_2] + \mathbb{E}[Y_2 + Y_3|A_2 = -1, H_2]} - \frac{1}{2}, \end{aligned} \tag{3.6}$$

with $U_3^*(H_2)$ defined as in Lemma 3.3.2. With this construction, we have that the optimal DTR is given by

$$(\tilde{d}_1(H_1), \tilde{d}_2(H_2)) = (sign(G_1(H_1)), sign(G_2(H_2))).$$

In light of this, let a_n be any increasing function of n, and let $(a_n\tilde{f}_{n,2}, a_n\tilde{f}_{n,1}) = \arg\inf_{(a_nf_1, a_nf_2)\in\mathcal{H}_n} \left(V_\psi^* - V_\psi(f_1, f_2) \right)$. It is helpful to decompose the approximation error as

$$T_1 = V_\psi^* - V_\psi(a_nG_1, a_nG_2) + V_\psi(a_nG_1, a_nG_2) - V_\psi(a_n\tilde{f}_{n,2}, a_n\tilde{f}_{n,1}).$$

This decomposition suggests that if we can approximate functions G_1, G_2 well with $\mathcal{H}_n$ the approximation error, induced by restricting our search to functions within this class, will be small. We summarize this in lemmas 3.4.4 and 3.4.5. Before stating the results we discuss a necessary assumption on the treatment effect, encoded in functions G_1, G_2.

Assumption 3.4.3. *There exists constant $C > 0$ and α so that*

$$P(0 < |G_1(H_1)| \leq t) + P(0 < |G_2(H_2)| \leq t) \leq Ce^{-\alpha/t}, \quad \textit{for all } t > 0.$$

Note that a larger value of α indicates a faster decay of the tail of the distribution of $G_1(H_1)$ and $G_2(H_1)$ near zero, which means the classification problem becomes easier as α increases. Assumption 3.4.3 is weaker than assuming strong separability: $\inf_{H_t\in\mathcal{X}_t} |G_t(H_t)| > 0$, $t = 1, 2$. However, it is a stronger version of the Margin Assumption [cf. Assumption (MA) of 88] or the low-noise assumption, which says there exist $\alpha > 0$ such that

$$P(0 < |G_1(H_1)| \leq t) + P(0 < |G_2(H_2)| \leq t) \leq Ct^{-\alpha}, \quad \text{for all } t > 0.$$

Assumption 3.4.3 ensures that the treatment effect is bounded away in probability from

the decision boundary. To see this, note that we can define the treatment effect for treatment 2 as

$$\Delta_2(H_2) \equiv \mathbb{E}[Y_3|H_2, A_2 = 1] - \mathbb{E}[Y_3|H_2, A_2 = -1].$$

This assumption implies that the treatment effect is bounded way from zero with high probability: $P(|\Delta(H_2)| > t) = 1 - e^{-\alpha/t}$. It suffices for H_2 to have a continuous covariate so that we have non-zero treatment effects: $P(\Delta(H_2) = 0) = 0$. Assumption 3.4.3 states that the treatment effect is strong with high-probability which makes the classification task easier.

The next two Lemmas help us control the approximation error. The proofs can be found in Appendix C.2.2.

Lemma 3.4.4. *Let G_1 and G_2 be defined as in (3.6), under assumptions 3.4.1 and 3.4.3, for any $a_n \to \infty$ any function ϕ in Corollary 3.3.1.1 (i)-(iii) satisfies,*

$$V_\psi^* - V_\psi(a_n G_1, a_n G_2) = O(a_n^{-1})$$

and ϕ as in Corollary 3.3.1.1 (iv) satisfies

$$V_\psi^* - V_\psi(a_n G_1, a_n G_2) = O(\exp(-a_n)).$$

We next state a result which lets us bound the difference between the surrogate value function evaluated at the true G_1, G_2, and at the approximations $\tilde{f}_{n,1}$, $\tilde{f}_{n,2}$.

Lemma 3.4.5. *Suppose ϕ is as in Corollary 3.3.1.1 (i)-(iii). Further suppose Assumptions 3.4.1, 3.4.3 hold, and there exists functions $(\tilde{f}_{n,1}, \tilde{f}_{n,2}) \in \mathcal{H}_n$ such that*

$$\|\tilde{f}_{n,1} - G_1/C_y\|_\infty \leq (\log a_n)^{-1},$$

and

$$\|\tilde{f}_{n,2} - G_2/C_y\|_\infty \leq (\log a_n)^{-1}.$$

Then

$$|V_\psi(a_n G_1, a_n G_2) - V_\psi\Big(a_n C_y \tilde{f}_{n,1}, C_y \tilde{f}_{n,2}\Big)| = O((\log a_n)^2/a_n).$$

If ϕ is as in Corollary 3.3.1.1 (iv), then under the same conditions,

$$|V_\psi(a_nG_1, a_nG_2) - V_\psi\Big(a_nC_y\tilde{f}_{n,1}, C_y\tilde{f}_{n,2}\Big)| = O(a_n^{-1}).$$

Lemma 3.4.4 indicates that if a_nG_1 and a_nG_2 can be well approximated by functions $(f_1, f_2) \in \mathcal{H}_n$, then the approximation error due to restricting our search to functions within these classes will be small. The uniform bound assumption for $\tilde{f}_{n,1}$, $\tilde{f}_{n,2}$ in Lemma 3.4.5, requires that $\mathcal{H}_n$ is rich enough to approximate G_1 G_2 well as $a_n \to \infty$. Next, we discuss the second term in the ψ-regret bound.

3.4.2 Generalization Error

We now turn to the generalization error:

$$T_2 = \sup_{(f_1,f_2)\in\mathcal{H}_n} \Big(V_\psi(f_1, f_2) - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2})\Big).$$

This term, also called estimation error, is caused by the fact that we are training $(\widehat{f}_{n,1}, \widehat{f}_{n,2})$ on a finite dataset. Intuitively, the generalization error measures how well $(\widehat{f}_{n,1}, \widehat{f}_{n,2})$ would perform on unseen data. Thus it naturally decreases as the training sample size increases. On the other hand, the generalization error increases with the complexity of the class $\mathcal{H}_n$. The error rate depends on the entropy integral for $\mathcal{H}_n$, which is defined by

$$J(\mathcal{H}_n) \equiv \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{H}_n, L_2(P))}d\epsilon.$$

The following result bounds the generalization error for estimated functions $(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \in \mathcal{H}_n$.

Theorem 3.4.6. *For any surrogate function of the form $\psi(x, y) = \phi(x)\phi(y)$ with ϕ Lipschitz continuous. We have*

$$\sup_{(f_1,f_2)\in\mathcal{H}_n} V_\psi(f_1, f_2) \;-\; V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) \;=\; O_p(J(\mathcal{H}_n)/\sqrt{n}).$$

The proof can be found in Appendix C.2.3. We note that for the bound we do not need ϕ to satisfy condition 3.3.2, we need only for it to be Lipschitz continuous. All examples listed

in Corollary 3.3.1.1 satisfy this. Additionally, Theorem 3.4.6 suggests that the generalization error rate will depend on the order of the entropy integral $J(\mathcal{H}_n)$. Thus, the radius and complexity of $\mathcal{H}_n$ play an important role in how well we can generalize. This is illustrated better using concrete examples for $\mathcal{H}_n$ in the next section.

3.4.3 Regret Bound Results

We have established rates for the approximation and generalization error in the previous sections. As an immediate result of Lemmas 3.4.4, 3.4.5 and Theorem 3.4.6 we have the following corollary which bounds the ψ-regret.

Corollary 3.4.6.1. *Under assumptions 3.4.1 and 3.4.3, if $a_n \to \infty$, and functions $(\tilde{f}_{n,1}, \tilde{f}_{n,2}) \in \mathcal{H}_n$ are such that*

$$\|\tilde{f}_{n,1} - G_1\|_\infty \leq (\log a_n)^{-1}, \quad \|\tilde{f}_{n,2} - G_2\|_\infty \leq (\log a_n)^{-1},$$

then any ϕ in Corollary 3.3.1.1 satisfies,

$$V_\psi^* - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) = O_p(J(\mathcal{H}_n)/\sqrt{n}) + O\left((\log a_n)^2/a_n\right).$$

Corollary 3.4.6.1 ensures consistency of $V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2})$ for V_ψ^* as long as $J(\mathcal{H}_n) = o_p(\sqrt{n})$. Next we discuss two particular examples of function spaces $\mathcal{H}_n$ which are well suited for neural networks and wavelet series estimation. We give the summary of our results in the current section and go over them in greater detail in Section 3.7.

We first define the neural network space in line with [89] construction. We set $\mathcal{H}_n = \bar{\mathcal{F}}(L_n, p, s_n, b_n)$ to be the class of rectified linear unit (ReLU) networks, which are uniformly bounded by b_n, with depth of $L_n \in \mathbb{N}$, layers of width vector p, and sparsity $s_n \in \mathbb{N}$ with weights bounded by 1.

For neural network estimation in Section 3.7.1 we further specify the smoothness degree of functions G_1, G_2 which we summarize in Assumption 3.7.1. With this, we next state in Corollary 3.4.6.2 (a) & (b) the approximation and generalization error respectively for using neural networks for estimation.

Corollary 3.4.6.2. *Suppose that the network depth is $L_n \geq C \log n$, sparsity $s_n \geq C \log a_n$, and $b_n \geq 2a_n$. If G_1, G_2 are Hölder classes with exponent $\beta \geq 1$, then under Assumptions 3.4.1, 3.4.3, and 3.7.1*

$$(a) \quad V_\psi^* - \sup_{f_1, f_2 \in \bar{\mathcal{F}}(L_n, p, s_n, b_n)} V_\psi(f_1, f_2) = O(\log^2(a_n)/a_n),$$

$$(b) \quad \textit{the entropy integral is } J_n = \sqrt{s_n L_n \log(s_n \wedge \max p)}.$$

Thus the regret bound for neural networks is

$$(c) \quad V_\psi^* - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) = O_p(\log n/\sqrt{n})$$

The proof is deferred to Appendix C.3.1. Next we give a similar corollary for wavelet series. Let $\mathcal{H}_n = \mathcal{H}_{1n} \times \mathcal{H}_{2n}$ be the class of S-regular, r-dimensional wavelet basis with $S \geq \beta$. Further assuming certain smoothness properties for G_1, G_2 summarized in Assumption 3.7.2 in Section 3.7.1, we have the following result.

Corollary 3.4.6.3. *If G_1, G_2 are Hölder classes with exponent $\beta \geq Cr^2$, then under Assumptions 3.4.1, 3.4.3, and 3.7.2*

$$(a) \quad V_\psi^* - \sup_{f_1, f_2 \in \mathcal{H}_n} V_\psi(f_1, f_2) = O(\log^2(a_n)/a_n).$$

$$(b) \quad \textit{the entropy integral is } J_n = \sqrt{a + K(\log n)^{r/\beta}},$$

where K depends on β, r. Thus the regret bound for wavelet series is

$$(c) \quad V_\psi^* - V_\psi(\widehat{f}_{n,1}, \widehat{f}_{n,2}) = O_p\left(\frac{(\log n)^{r/(2\beta)}}{\sqrt{n}}\right).$$

3.5 Empirical Analysis

This section contains results for extensive simulations performed to evaluate the finite sample properties of our methods under different generating data mechanisms. We then proceed to

use the method on a real data set of patients with Sepsis treated in the intensive care unit (ICU).

3.5.1 Simulations

Following are several simulation scenarios designed to compare the methods under discrete and continuous feature spaces, linear and non-linear decision boundaries, and continuous and binary outcomes.

Setting 1: Discrete space To understand the method's performance when the set of all measurable functions $\mathcal{F}$ is attainable, we designed a setting with discrete covariates only. We then estimate $\hat{f}_{n,1}, \hat{f}_{n,2}$ using saturated models. Specifically, suppose $O_1 \equiv (X_{11}, X_{12}, X_{13})$, $O_2 \equiv X_{21}$, where $X_1, X_2, X_3, A_1, A_2 \in \{-1, 1\}$, $P(X_j = 1) = P(A_k = 1) = 0.5$, $j = 1, 2, 3$, $k = 1, 2$, and $X_{21} = I\{-1.75X_{12}A_1 + Z_1 > 0\}$, with $Z_1 \sim N(0, 1)$. Further, we have

$$Y_2|O_1, A_1 \sim \text{Bern}\left[\sigma\left\{(X_{13} - 0.5X_{12})A_1\right\}\right], \text{ and}$$
$$Y_3|O_1, A_1, O_2, A_2 \sim \text{Bern}\left[\sigma\left\{(0.5X_{11} + X_{12} - 0.2X_{21} + Y_2)A_2\right\}\right].$$

Setting 2: Toy non-linear setting In this setting we explore how the methods perform with decision boundaries that are not linearly-separable. (See Figure 3.2). If we have treatments A_1,A_2 $\sim$Bern(0.5), and $O_1 \equiv X_1$, $O_2 \equiv X_2$ with $X_1, X_2, Z_1, Z_2 \overset{iid}{\sim} N(0, 1)$, then outcomes are

$$Y_2 = A_1(2I\{|X_1| < 1\} - 1) + Z_1, \text{ and}$$
$$Y_3 = A_2(2I\{X_2 > X_2^2\} - 1) + Z_2.$$

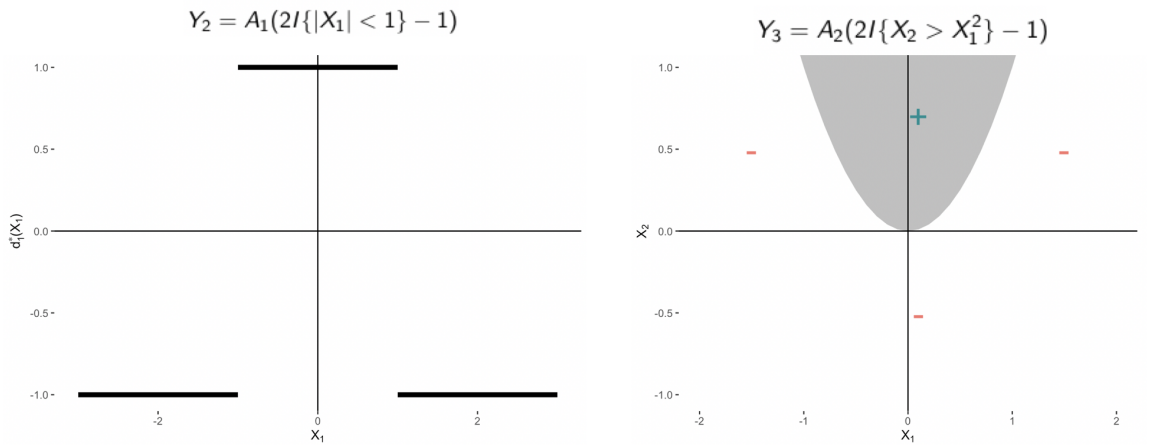


Figure 3.2: *Plots of the optimal DTR given the non linear decision boundaries in setting 2. The treatment rules are given by* $d_1^*(X_1) = sign(1 - |X_1|)$ *and* $d_2^*(X_1, A_1, X_2) = sign(X_2 - X_1^2)$ *respectively, shown in black and gray.*

Setting 3: Simple decision boundary We use a setting similar to [**?**] were there is a mix of continuous and discrete covariates. We define $O_1 \equiv (X_{11}, X_{12}, X_{13}) \sim MVN(0, I_3)$, and $O_2 \equiv (X_{21}, X_{22})$ with $X_{21} = I\{1.25X_{11}A_1 + Z_1 > 0\}$, $X_{22} = I\{-1.75X_{12}A_1 + Z_2 > 0]$. The treatments are $A_1, A_2 \in \{-1, 1\}$ with $P(A_t = 1) = 0.5$, $t = 1, 2$. Outcomes are set to

$$Y_2 = 10 + A_1(1 + 1.5X_{13}) + Z_3, \text{ and}$$
$$Y_3 = 10 + A_2(-5 + 0.5Y_2 + 0.5A_1 + 0.5X_{21} - 0.5X_{22}) + Z_4,$$

where $Z_1, \ldots, Z_4 \overset{iid}{\sim} N(0, 1)$.

Setting 4: Continuous, non-linear decision boundary This setting is designed to explore performance of the methods in non-linear outcome and decision functions, with a higher number of features relative to setting 2. We use O_1, treatments A_1,A_2 and noise Z_1, Z_2 as defined in Setting 2, we set $O_2 \equiv X_{21} \sim N(0, 1)$. Treatments A_1,A_2 ∼Bern(0.5). Outcomes are defined as

$$Y_2 = 2 + A_1(1 + 1.5I\{X_{13} > 0\}) + Z_1, \text{ and}$$
$$Y_3 = 2 + A_1(1 + 1.5I\{X_{13} > 0\}) + 10 * A_2 sign\Big(.01X_{12}^2 - .05X_{13}^2\Big) + X_{12} + Z_2.$$

Setting 5: Sepsis This last setting is designed to be similar to our Sepsis data set. This has a higher number of correlated covariates to assimilate our real data. In addition we differentiate between main effect and treatment effect covariates using H_{t0}, H_{t1} respectively for $t = 1, 2$. Data is simulated with $H_{10} = (1, X_1, \ldots, X_6)^\intercal$, $H_{11} = (1, X_2, \ldots, X_6)^\intercal$, $H_{20} = (Y_2, 1, X_1, \ldots, X_6, A_1, Z_{21}, Z_{22})^\intercal$, and $H_{21} = (1, X_1, \ldots, X_4, A_1, Z_{21}, Z_{22})^\intercal$, generated according to $X_1 = [W]$, $W \sim \mathcal{N}(0, \sigma_z I_6 + \rho_z)$, where $[W]$ denotes the floor function of vector W, and second stage covariates are $O_2 = [I\{Z_1 > 0\}, I\{Z_2 > 0\}]^\intercal$, $Z_l = O_{1l}\delta_l^0 + \epsilon_z$, $\epsilon_z \sim \mathcal{N}(0, 1)$ $l = 1, 2$. Treatments are no longer randomized, they are generated according to $A_1 \sim \text{Bern}(\sigma\{H_{10}^\intercal \xi_1^0\})$, $A_2 \sim \text{Bern}\left(\tilde{m}_2\{H_{20}\}\right)$, with $\tilde{m}_2 = \sigma\left\{H_{20}^\intercal \xi_2^0 + \tilde{\xi}_2^\intercal X_2\right\}$. Finally, outcomes are set to

$$
\begin{aligned}
Y_2 &= H_{10}^\intercal \beta_1^0 + A_1 H_{11}^\intercal \gamma_1^0 + Z_3, \text{and} \\
Y_3 &= H_{20}^\intercal \beta_2^0 + A_2 (H_{21}^\intercal \gamma_2^0) + \tilde{\beta}_2^\intercal X_2 Y_2 \sin\{\|X_2\|_2^2/(Y_2 + 1)\} + Z_4,
\end{aligned}
$$

with $Z_3, Z_4 \sim (0, 1)$. Parameters are set to $\xi_1^0 = (-0.1, 1, -1, 0.1)^\intercal$, $\beta_1^0 = (0.5, 0.2, -1, -1, 0.1, -0.1, 0.1)^\intercal$, $\gamma_1^0 = (1, -2, -2, -0.1, 0.1, -1.5)^\intercal$, $\sigma_x = .9$, and $\rho_x = 0.1$ for the first stage and $\xi_2^0 = (0, 0.5, 0.1, -1, 1, -0.1)^\intercal$, $\beta_2^0 = (1, \beta_1^0, 0.25, -1, -0.5)^\intercal$, $\gamma_2^0 = (1, 0.1, -0.1, 0.1, -0.1, 0.25, -1, -0.5)^\intercal$ for the second stage.

We implement our proposed method using several linear functions, basis expansion and neural networks. The linear functions are of the form $f_t(H_t; \theta_t) = H_t^\intercal \theta_t$, $t = 1, 2$. Similarly, we use a linear function on the basis. In particular, for our basis expansions we use natural cubic splines with three knots at the quantiles 0.25, 0.5, and 0.75, and Daubechies tensor wavelet basis with expansions up to $J = 5$. Finally, we implement a neural network (NN) for $f_t(H_t; \theta_t)$ which consists of 2-layers with 128, 64 units each and ReLU activation functions. We use 50% dropout as regularization. All models are trained using RMSprop, a variation of stochastic gradient descent which scales the learning rate by a moving average of the magnitudes of recent gradients [81].

Additionally, we compare the proposed method with SOWL. We use a linear kernel as in the original work, and a Gaussian kernel (RBF) to incorporate non-linearity. We implement

the weighted SVM procedure using CVXOPT, a Python package for convex optimization [90]. We use 5-fold cross validation for selecting the tuning parameter among a grid of 15 candidate values as in [**?**]. We also use Q-learning as a benchmark. In this non-Markovian context, as with dynamic programming, we estimate $\hat{\theta}_2$ as the minimizer of $\mathbb{P}_n(Y_3 - Q(H_2, A_2; \theta_2))^2$, then set the pseudo outcome to $\hat{Y}_{2i} = Y_{2i} + \max_a Q(H_{2i}, a; \hat{\theta}_2)$ $i = 1, \ldots, n$ and estimate $\hat{\theta}_1$ as the minimizer of $\mathbb{P}_n(\hat{Y}_2 - Q(H_1, A_1; \theta_1))^2$. The estimated DTR using Q-learning is $\hat{d}_t(H_t) = \text{argmax}_{aQ(H_t,a;\hat{\theta}_t)}$. We consider two models for the Q functions. The first one consists of Q-functions which are linear in the patient's history, with a main and treatment effect: $Q(H_t, A_t; \theta_t) = H_t^\intercal \theta_{t0} + A_t H_t^\intercal \theta_{t1}$ $t = 1, 2$. The second model uses a NN for each Q function. The network consists of 2-layers with 128, 64 units each and ReLU activation functions. We use 50% dropout as regularization and RMSprop for optimization.

We generate 500 different datasets of size $2n$ where $n = 250, 2500, 25000$ under each listed setting, then split the data in two equal sizes for a training and test set. We implement the described methods to estimate DTR $(\hat{d}_1(H_1), \hat{d}_2(H_2))$ using the n training samples. The derived treatment rule is evaluated using the true value function $V(\hat{d}_1(H_1), \hat{d}_2(H_2))$ on the test data, and compared with $V^* = V(d_1^*(H_1), d_2^*(H_2))$. We report these estimates as well as run-time (seconds) for each case.

Table 3.1 shows the value function for estimated rule for each method under different data size and settings. The optimal value V^* is shown for each setting. Setting 1 is a relatively simple setting as the space is small and discrete, so with the necessary number of parameters (no-less than the possible values in feature space) any possible DTR is estimable by any method. This is reflected on the performance of all surrogate methods and NN Q-learning with enough data. SOWL fails to scale with sample size and the linear Q-learning is misspecified in this setting as it is under-parameterized. The NN on the other hand could be under-performing due to the over-parameterization and relatively small training sample. Settings 2 and 4 have highly non-linear decision boundaries. Thus, all smooth surrogate methods except for the linear one perform well. This is because they can account for non-linear types of decision boundaries. NN Q-learning also handles these settings well, however on smaller data samples

Setting	n	Smooth Surr. (linear)	Smooth Surr. (wavelets)	Smooth Surr. (splines)	Smooth Surr. (NN)	SOWL (linear)	SOWL (RBF)	Q-learning (linear)	Q-learning (NN)
1	250	1.23 (0.07)	1.25 (0.07)	1.16 (0.09)	1.23 (0.06)	1.16 (0.12)	0.96 (0.16)	1.06 (0.08)	1.24 (0.03)
	2500	1.33 (0.02)	1.33 (0.02)	1.22 (0.07)	1.32 (0.02)	1.24 (0.06)	1.06 (0.13)	1.04 (0.01)	1.31 (0.01)
	25000	1.35 (0.01)	1.35 (0.01)	1.35 (0.08)	1.35 (0.01)			1.04 (0.02)	1.32 (0.01)
	V^*				1.36				
2	250	1.15 (0.15)	0.98 (0.29)	1.25 (0.25)	1.47 (0.16)	0.44 (0.1)	0.61 (0.42)	0.56 (0.4)	1.0 (0.09)
	2500	1.25 (0.03)	1.29 (0.28)	1.8 (0.14)	1.7 (0.13)	0.21 (0.17)	0.81 (0.33)	0.8 (0.02)	1.8 (0.04)
	25000	1.31 (0.03)	1.35 (0.25)	1.89 (0.05)	1.83 (0.14)			0.8 (0.01)	1.95 (0.01)
	V^*				2.0				
3	250	20.58 (0.68)	20.28 (0.44)	20.81 (0.63)	20.61 (0.65)	20.52 (0.63)	20.12 (0.65)	20.57 (0.83)	20.09 (0.31)
	2500	21.07 (0.57)	20.91 (0.42)	21.63 (0.44)	21.06 (0.45)	20.63 (0.62)	19.42 (0.7)	21.0 (0.04)	21.1 (0.06)
	25000	21.43 (0.22)	21.72 (0.25)	22.35 (0.09)	21.33 (0.16)			21.0 (0.01)	21.57 (0.05)
	V^*				22.61				
4	250	7.08 (0.87)	6.98 (1.65)	10.8 (1.55)	7.64 (0.95)	3.36 (1.65)	3.87 (2.31)	6.73 (0.82)	5.95 (0.72)
	2500	7.91 (0.29)	8.89 (1.7)	13.33 (0.61)	9.99 (0.9)	3.82 (1.74)	3.68 (2.59)	6.9 (0.19)	10.51 (0.57)
	25000	8.08 (0.24)	10.15 (1.43)	13.82 (0.49)	13.06 (0.73)			6.91 (0.06)	13.7 (0.15)
	V^*				14.75				
5	250	7.29 (0.48)	7.43 (0.46)	6.99 (0.52)	7.22 (0.51)	7.09 (0.62)	4.25 (2.42)	7.07 (0.53)	7.34 (0.48)
	2500	7.56 (0.16)	7.57 (0.16)	7.17 (0.26)	7.56 (0.16)	7.29 (0.15)	5.67 (1.6)	7.13 (0.16)	7.77 (0.14)
	25000	7.77 (0.08)	7.72 (0.08)	7.57 (0.26)	7.63 (0.09)			7.12 (0.05)	7.77 (0.05)
	V^*				7.96				

Table 3.1: *Value function $V(\hat{d}_1, \hat{d}_2)$ for estimated DTR for the surrogate loss method, SOWL and Q-learning across different data generating mechanisms.*

(n = 250, 2500) its performance is not ideal even with high drop-out regularization. As expected, linear functions in either smooth surrogates, Q-learning, or SOWL cannot correctly classify independent of their training sample size. This is because they can never closely approximate the decision boundary.

It is interesting to note that in Setting 2, SOWL with RBF kernel does better than its linear counterpart, however it fails to scale with n. In setting 4 on the other hand, RBF appears to hinder SOWL's performance, possibly due to the higher number of features. In settings 3 and 5 most methods do well due to the relatively simple decision boundary. Complexity does seem to hinder SOWL again in setting 5, as RBF is outperformed by a linear kernel. The smooth surrogate methods have comparable or better performance across all scenarios. This is because the smoothness and flexibility in choice of surrogates and in types of estimation, make them easily tailored to the data set at hand. We discuss this further in Section 3.6.

We show computation run-time in Table 3.2. All smooth surrogate and Q-learning methods are trained in a similar way. They all use stochastic gradient descent with RMSprop for

Setting, n		Smooth Surr. (linear)	Smooth Surr. (wavelets)	Smooth Surr. (splines)	Smooth Surr. (NN)	SOWL (linear)	SOWL (RBF)	Q-learning (linear)	Q-learning (NN)
1	250	0.04	0.04	0.05	0.15	0.13	0.15	0.07	0.21
	2500	0.42	0.54	0.44	1.17	66.24	94.77	0.72	2.22
	25000	5.97	4.32	4.48	11.66			6.66	22.69
2	250	0.05	0.04	0.05	0.14	1.27	1.39	0.08	0.23
	2500	0.44	0.44	0.46	1.23	847.96	828.77	0.79	2.14
	25000	4.62	3.98	4.64	11.8			7.56	22.14
3	250	0.05	0.04	0.05	0.12	0.1	0.18	0.07	0.35
	2500	0.51	0.5	0.47	1.48	54.57	113.39	1.07	3.52
	25000	6.34	4.22	4.78	12.71			6.98	33.06
4	250	0.05	0.05	0.05	0.12	1.24	1.36	0.07	0.23
	2500	0.45	0.47	0.46	1.29	826.9	788.03	0.8	2.13
	25000	4.68	5.83	4.31	10.65			8.04	23.85
5	250	0.04	0.04	0.05	0.17	1.39	1.18	0.08	0.23
	2500	0.42	0.41	0.46	1.6	837.5	870.7	0.73	2.31
	25000	4.38	4.66	4.66	11.46			7.52	20.61

Table 3.2: *Run-time (seconds) for estimating DTR for the surrogate loss method, SOWL, and Q-learning across different data generating mechanisms.*

optimization of the respective loss functions. All these methods are trained for 20 epochs and use a batch size of 128. Run time for smooth surrogate methods with linear and basis expansion functions is relatively similar. This doubles for NN functions. Nonetheless, these are all in the order of 10 seconds, even for samples of $n = 25000$. The Q-learning methods are slightly slower but the time increase is still negligible. As expected, SOWL methods are an order or two higher in run-time. This is because SVMs utilize the dual space for optimization. The time cost is especially high in settings 2 and 4 which have highly non-linear decision boundaries, and setting 5 which has over 40 features.

3.5.2 Data Application: Sepsis Cohort

We evaluate our methods and benchmarks on a Sepsis cohort from MIMIC-IV data [91]. Sepsis is a severe state of infection which overwhelms the body's immune system, potentially causing damage to tissue, organ failure, and in some cases death. This makes it an interesting condition from a DTR perspective [92, 93]. Physician's usually treat it with a high, constant dose of antibiotics. They also use intravenous (IV) fluids for hypovolemia. Deciding the optimal dose of IV fluids is highly challenging due to the heterogeneity in response of the septic body.

Interest lies in learning when to administer IV fluids, thus we code our actions as $A_t = -1$ if no dose is necessary and $A_t = 1$ otherwise. The state space is comprised of 46 covariates O_t at each time-step. Measured variables include age, body mass index, diastolic and systolic blood pressure, etc. We use an inverse transformation of lactate acid level as the outcome Y_t. In particular $Y_t = (LA_t + 5)^{-1} + 2$ where LA_t stands for lactic acid level at time t, this is done to ensure we have positive outcomes, as stated in Assumption 3.4.1. Each time-step consists of a four hour interval. The data includes $n = 9,872$ trajectories. We use a 50% split for training the methods and for estimating the respective value function.

We evaluate the smooth surrogate methods using linear, wavelets, splines and NN functions. Additionally we include SOWL with linear and RBF kernels, and Q-learning with linear and NN Q-functions. To evaluate the derived DTRs we estimate the value function using a doubly robust approach [6, 7, 94]. In particular, we estimate a Q-function for each time-step using an NN with the same specifications as our benchmark. We estimate treatment propensity score models $\pi_t(A_t|H_t;\gamma_t)$ for $P(A_t = a|H_t)$ using logistic regression with a ridge penalty. For any treatment rule $(\hat{d}_1(H_1), \hat{d}_2(H_2))$ we can estimate the value function on the test set with

$$\begin{aligned}\widehat{V}_{DR} =& \mathbb{P}_n \Bigg[Q_1(H_1, \hat{d}_1(H_1), \hat{\theta}_1) \\ &+ \frac{I\{A_1 = \hat{d}_1(H_1)\}}{\pi_1(A_1|H_1;\hat{\gamma}_1)} \left[Y_2 - \left\{ Q_1(H_1, \hat{d}_1(H_1), \hat{\theta}_1) - Q_2(H_2, \hat{d}_2(H_2), \hat{\theta}_2) \right\} \right] \\ &+ \frac{I\{A_1 = \hat{d}_1(H_1), A_2 = \hat{d}_2(H_2)\}}{\pi_1(A_1|H_1;\hat{\gamma}_1)\pi_2(A_2|H_2;\hat{\gamma}_2)} \left\{ Y_3 - Q_2(H_2, \hat{d}_2(H_2); \hat{\theta}_2) \right\} \Bigg].\end{aligned}$$

We compute the doubly robust value function estimator $\widehat{V}_{DR}$ for the different methods using 2-fold cross-validation. We repeat this for 100 different random splits and report the mean and SE.

	Smooth surr. (linear)	Smooth surr. (splines)	Smooth surr. (NN)	Smooth surr. (wavelets)	SOWL (linear)	SOWL (RBF)	Q-learning (linear)	Q-learning (NN)
$\hat{V}_{DR}(\hat{d}_1, \hat{d}_2)$	1.66	1.51	1.315	1.331	1.1	0.858	1.464	1.328
SE	0.031	0.017	0.054	.31	0.07	1.1	0.08	0.077

Table 3.3: *Estimated Value function and 90% CI for DTRs derived with different methods.*

Table 3.3 shows the estimated values using the doubly robust estimator for each method and their corresponding standard errors. There seems to be a better performance when using linear functions for both the decision boundary (as in our smooth surrogate method and SOWL) as well as for the Q-functions. As in simulation setting 5, this might be because the underlying decision function is near linear, therefore the complex, more flexible models might be suffering because of their complexity. The linear smooth surrogate method still outperforms linear Q-learning. This is possibly due to the fact that Q-learning is imposing a linear model on the outcome functions, which as we have previously discussed, are likely to be mis-specified. Especially in this multiple stage setting, the first stage outcome function might not be well approximated by a linear model. Using our smooth surrogates we bypass the need for outcome modeling. SOWL methods are probably suffering because the training data consists of ≈ 5000, which is already a large data set in the dual space.

3.6 Discussion

In this work, we propose an extension of smooth surrogate loss functions used for classification, into the multi-stage setting. We then illustrate how to use this extension to estimate the optimal DTR for a given data set. We provide results which guarantee that our surrogate value function is Fisher consistent for the value function of interest, generated by the 0-1 indicators. We additionally show a ψ-regret bound based on a general class of estimable functions $\mathcal{H}_n$, with complexity measured by the entropy integral. With these results, we provide regret bounds for when we estimate the decision functions with NNs, or wavelet series. We also have shown $1/\sqrt{n}$ rates for these methods, under weak separability conditions.

There are some immediate steps that we are interested in exploring in order to make our approach more robust to different data generating mechanisms. For example, our approximation error results require a noise condition of the order of $\exp\{-t/\alpha\}$ (see Assumption 3.4.3). Recall that this is for any $t > 0$ where α measures the decay around the decision boundary. This assumption implies that the treatment effect is relatively strong across the patient population. We are interested in relaxing this to the commonly used low-noise assumption, which is of the order of t^α. With this, we will generalize the rate of our approximation error by allowing it to be a function of α.

In addition, the $1/\sqrt{n}$ generalization error rate currently dominates the approximation error $(\log n)^2/n$ rate, which makes the ψ-regret bound of $O(1/\sqrt{n})$. This is due to the surrogate loss function ψ not being concave. However, we will explore local concavity, as this will likely improve the generalization error rate to $O(1/n)$, which in turn yields a preferable regret bound for our proposed surrogate family. We are also interested in extending this framework to multiple stages and multiple treatment options.

We have shown that the smooth-surrogate family brings an advantage in empirical performance over competing methods such as SOWL and Q-learning. This is due to a few reasons. First, the fact that the surrogates are differentiable make the optimization amenable to gradient descent. This lets us use state-of-the-art machine learning packages to implement stochastic gradient descent (SGD). In addition, we can also choose among the large number of SGD variants such as RMSprop, which is often an empirically superior modification to SGD. Moreover, smoothness in ψ allows us to use a wide range of methods for approximating the decision boundary. We show that our method is usually among the best for at least one of linear function, wavelets, splines or NN. Finally, the additional choice of ϕ-functions adds a degree of flexibility that lets us tailor the method to a particular data set. All this choice in hyper-parameter selection brings little increased cost. As the method takes only a few seconds to train even with large samples, there is very little run-time needed to search for the best choice of surrogate functions ϕ_1, ϕ_2, and approximation functions $f_t(H_t;\theta_t)$. We believe this approach makes simultaneous optimization of DTRs a feasible task, while offering strong

theoretical guarantees. All of which brings us one step closer to personalized medicine.

3.7 Technical Results for Neural Networks and Wavelets

In this section we develop the necessary. notation, definitions, and theoretical results for Corollaries 3.4.6.2 and 3.4.6.3. In particular, we detail the notation and construction of NN and wavelet series function classes, and show they satisfy the required assumptions for our results.

We start by defining a Hölder class. Suppose $\mathcal{C} \subset \mathbb{R}^r$. A function $f : \mathcal{C} \mapsto \mathbb{R}$ is said to have Hölder smoothness index β if for all $\alpha \in \mathbb{N}^r$ satisfying $|\alpha|_1 < \beta$, $\partial^\alpha f$ exists and there exists constant C so that

$$\frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|^{\beta - \lfloor * \rfloor \beta}} < C \quad \text{for all } x, y \in \mathcal{C}.$$

Here $\alpha = (\alpha_1, \ldots, \alpha_d)$ and $\partial^\alpha f = \partial^{\alpha_1} \partial^{\alpha_2} \ldots \partial^d$. We will denote by $\mathcal{C}_r^\beta(D, C_a)$ the class

$$\left\{ f : D \subset \mathbb{R}^r \mapsto \mathbb{R} \;\middle|\; \sum_{\alpha : |\alpha|_1 < \beta} \|\partial^\alpha f\|_\infty + \sum_{\alpha : |\alpha| = \lfloor * \rfloor \beta} \sup_{\substack{x,y \in \mathcal{C} \\ x \neq y}} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|^{\beta - \lfloor * \rfloor \beta}} \leq C_a \right\}.$$

Note that functions in $\mathcal{C}_r^\beta(C_a)$ are bounded by C_a.

3.7.1 Neural Networks

We next discuss the case where we use NNs to approximate G_1, G_2 defined in (3.6). Our construction follows that of [89]. We first go over the notation necessary to define the class of NNs. We use NNs with L layers, and summarize the width of each layer in vector $p = (p_{L+1}, p_L, \ldots, p_1, p_0)$. In our case, p_0 is the length of the feature dimension, i.e. a unit larger than the dimension of $\mathcal{X}_2$. We denote it by r_2. Therefore, $p_0 = r_2 + 1$. Let $v_0, \ldots, v_L$ be $L+1$ many shift vectors. We denote by $\sigma_v = \sigma(\cdot - v)$ the elementwise shifted ReLU activation function, its dimension is the same as v. Here $\sigma(x) = \max\{0, x\}$ is the elementwise ReLU function. For any $x \in \mathbb{R}^r$ consider the function of the form

$$x \mapsto f(x) = W_L \sigma_{v_L} \ldots W_1 \sigma_{v_1} W_0 \sigma_{v_0}(x), \tag{3.7}$$

where W_L is a vector and all other W_i's are weight matrices.

We use the term sparsity to denote the number of non-zero parameters in the network. Hence, s is the sparsity of the network if

$$s = \sum_{i=0}^{L} (\|W_i\|_0 + \|v_i\|_0).$$

Then by $\mathcal{F}(L, p, s)$ we denote the class of functions of the form (3.7) with sparsity bounded by s. Let us define

$$\mathcal{F}(L, p, s) = \Bigg\{ f : f(x) = W_L \sigma_{v_L} \dots W_1 \sigma_{v_1} W_0 \sigma_{v_0}(x), \\ s = \sum_{i=0}^{L} (\|W_i\|_0 + \|v_i\|_0),\ \|W_i\|_\infty, \|v_i\|_\infty < 1 \text{ for all } i = 0 : L \Bigg\}.$$

The output layer the networks in $\mathcal{F}(L, p, s)$ are linear because they have the form of (3.7). We consider another class of network whose final layer has a ReLU gate without any additional shift. In particular, we consider the form

$$x \mapsto f(x) = \sigma_0 W_L \sigma_{v_L} \dots W_1 \sigma_{v_1} W_0 \sigma_{v_0}(x). \tag{3.8}$$

We define

$$\overline{\mathcal{F}}(L, p, s) = \Bigg\{ f : f(x) = \sigma_0 W_L \sigma_{v_L} \dots W_1 \sigma_{v_1} W_0 \sigma_{v_0}(x), \\ s = \sum_{i=0}^{L} (\|W_i\|_0 + \|v_i\|_0),\ \|W_i\|_\infty, \|v_i\|_\infty < 1 \text{ for all } i = 0 : L \Bigg\}.$$

Note that the networks in $\mathcal{F}(L, p, s)$ are a composition of networks in $\overline{\mathcal{F}}(L', p', s')$, as the only difference is that the networks in $\overline{\mathcal{F}}(L', p', s')$ have the standard ReLU gate in the last layer. This also implies that composition of two $\overline{\mathcal{F}}(L, p, s)$ networks give another $\overline{\mathcal{F}}(L', p', s')$ network. We will denote $\mathcal{F}(L, p, \infty)$ and $\overline{\mathcal{F}}(L, p, \infty)$ by $\mathcal{F}(L, p)$ and $\overline{\mathcal{F}}(L, p)$, respectively.

We also consider the class

$$\mathcal{F}(L, p, s, b_n) = \Big\{ f \in \mathcal{F}(L, p, s, b_n) : \|f\|_\infty < b_n \Big\}$$

where b_n is some large class. The above classes are expected to approximate functions which

can be written as compositions of smooth functions reasonably well.

We will consider a composition of functions $g_i : U_i \subset \mathbb{R}^{r_i} \mapsto \mathbb{R}^{r_{i+1}}$ where $g_i = (g_{i1}, \ldots, g_{id_{i+1}})$. We let $i = 0, \ldots, q$ where q is a fixed positive integer. We take $\mathcal{X}_2 = U_0$ and $r_{q+1} = 1$. We will also consider $U_i = [a_i, b_i]_i^r \subset \mathbb{R}^{r_i}$ for $1 \leq i \leq q$ for some positive integer $q \in \mathbb{N}$ and $|b_i|, |c_i| < C_\alpha$. Suppose $\mathbf{r}$ is the vector of the dimensions $(r_0, r_1, \ldots, r_q)$. $\mathbf{t}$ is the vector of integers $(t_0, t_1, \ldots, t_q)$ and β is the vector $(\beta_1, \ldots, \beta_r)$. Suppose $[a_0, b_0]$ contains any one-dimensional projection of $\mathcal{X}_2$. Let us denote the class

$$\Omega(q, \mathbf{r}, \mathbf{t}, \beta, C_a) = \Big\{ f = g_q \circ g_{q-1} \circ \ldots \circ g_0 \;\Big|\; g_i = (g_{ij})_j : [a_i, b_i]^{r_i} \mapsto [a_{i+1}, b_{i+1}]^{r_{i+1}}, \\ a_i, b_i \in \mathbb{R} \;\text{ for } i \geq 1, g_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([a_i, b_i], C_a) \Big\}.$$

We assumed the supports of g_i's to be rectangular for simplicity of the proof. If the g_i's have bounded support, then also the proof goes through after some modification.

We will now impose further conditions on G_1 and G_2. First, let us separate the continuous and categorical variables of H_1 and H_2 by using H_1^c and H_2^c to denote the continuous variables of H_1 and H_2. We assume $H_1^c \in \mathcal{X}_1^c$ and $H_2^c \in \mathcal{X}_2^c$. Similarly, we denote the categorical parts $H_1^d \in \mathcal{X}_1^d$ and $H_2^d \in \mathcal{X}_2^d$. Note that $\mathcal{X}_1 = \mathcal{X}_1^c \cup \mathcal{X}_1^d$ and $\mathcal{X}_2 = \mathcal{X}_2^c \cup \mathcal{X}_2^d$. Also note that $\mathcal{X}_1^c \subset \mathcal{X}_2^c$ and $\mathcal{X}_1^d \subset \mathcal{X}_2^d$. For $j = 1, 2$, we let $r_{j,c}$ and $r_{j,d}$ denote the dimensions of $\mathcal{X}_j^c$ and $\mathcal{X}_j^d$, respectively. Therefore $r_{1,c} < r_{2,c}$ and $r_{1,d} < r_{2,d}$.

Assumption 3.7.1. *For any fixed $h_1^d \in \mathcal{X}_1^d$ and $h_2^d \in \mathcal{X}_2^d$, denote by $G_1(\cdot, h_1^d)$ and $G_2(\cdot, h_2^d)$ the maps $h_1^c \mapsto G_1(h_1^c, h_1^d)$. Also, for $j = 1, 2$, we assume $\mathcal{X}_j^c = \prod_{i=1}^{r_{j,c}} [b_i', c_i']$ where $b_i' < c_i' \in \mathbb{R}$.*

1. *There exist $q \in \mathbb{N}$, $\boldsymbol{r}, \boldsymbol{t} \in \mathbb{N}^{q+1}$ and $\beta \in [1, \infty)^{q+1}$ with $r_0 = r_{2,c}$ so that the map $G_2(\cdot, h_2^d) \in \Omega(q, \boldsymbol{r}, \boldsymbol{t}, \beta, C_a)$ for each h_2^d.*

2. *There exist $q' \in \mathbb{N}$, $\boldsymbol{r}', \boldsymbol{t}' \in \mathbb{N}^{q'+1}$ and $\beta' \in [1, \infty)^{q'+1}$ with $r_0' = r_{1,c}$ so that for each h_1^d, the map $G_1(\cdot, h_1^d) \in \Omega'(q', \boldsymbol{r}', \boldsymbol{t}', \beta', C_a)$.*

$G_2(\cdot, h_2^d)$'s may have different levels of smoothness and they may belong to different $\Omega(q, \mathbf{r}, \mathbf{t}, \beta, C_a)$'s for different h_2^d's. Since the identity map is an infinitely differentiable smooth map, we can show that the union over all such $\Omega(q, \mathbf{r}, \mathbf{t}, \beta, C_a)$'s is also of the form

$\Omega(q', \mathbf{r}', \mathbf{t}', \beta', C_a')$ where q' is the maximum of all q's. Therefore even when $G_2(\cdot, h_2^d)$'s have different smoothness levels, Assumption 3.7.1 1 stays valid provided each of them are in $\Omega(q, \mathbf{r}, \mathbf{t}, \beta, C_a)$ for some q, $\mathbf{r}$, $\mathbf{t}$, β and C_a. The same holds for $G_1(\cdot, h_1^d)$'s. We opted for the uniform scenario to simplify notational burden.

Under this detailed construction, Corollary 3.4.6.2 states that by using the loss function ϕ, and having the data generating mechanism satisfies Assumptions 3.4.1, 3.4.3, and. 3.7.1. Given $\epsilon > 0$ and sequence $a_n \to \infty$, there exists

$$\tilde{h}_{n,1} \in \mathcal{F}(O(\log a_n), \mathbf{p}, O(\log a_n), 2a_n)$$
$$\tilde{h}_{n,2} \in \mathcal{F}(O(\log(a_n)), \mathbf{p}, O(\log a_n), 2a_n),$$

so that the maximal width in $\mathbf{p}$ is $O(1)$ and the $n^{-1/2}$ ψ-regret bound follows.

3.7.2 Wavelets

This section considers estimation of f_1 and f_2 using wavelet series. Before going into any further detail, we introduce the notion of Besov spaces because of its intrinsic connection to the wavelet series estimation. As usual, we make some smoothness assumption regarding G_1 and G_2. In particular, we assume that they belong to Hölder spaces with coefficient β.

Assumption 3.7.2. *The functions G_1 and G_2 defined in* (3.6) *satisfy $G_1 \in \mathcal{C}_{r_1}^{\beta}(\mathcal{X}_1, C_a)$ and $G_2 \in \mathcal{C}_{r_2}^{\beta}(\mathcal{X}_2, C_a)$ where $C_a > 0$ and $\mathcal{X}_1 \subset \mathbb{R}^{r_1}$ and $\mathcal{X}_2 \subset \mathbb{R}^{r_2}$.*

Suppose $r \geq 1$. It can be shown that for any compact set $\mathcal{S} \subset \mathbb{R}^r$, the Hölder space $C_r^{\beta}(\mathcal{S}, C)$ is embedded into the Besov space $B_{\infty,\infty}^{\beta}(\mathcal{S})$ [cf. p.350-p.351 of 95, for a discussion when the dimension r=1]. $B_{\infty,\infty}^{\beta}(\mathcal{S})$ is a particular case of the more general Besov spaces $B_{p,q}^{\beta}(\mathcal{S})$ when $p = \infty$ and $q = \infty$ [see Section 4.3.1 of 95, for definition and characterization].

Let us consider father and mother wavelets ϕ and ψ. We will assume that these wavelets are S-regular wavelets [cf. Definition 4.2.14, 95] with $S > \beta$, whose examples include, but are not limited to Meyer and Daubechies wavelets. The corresponding r-dimensional wavelets are

constructed using tensor products

$$\Phi(x) = \phi(x_1) \cdots \phi(x_d), \quad x = (x_1, \ldots, x_d) \in [0,1]^r.$$

We let $\Phi_k(x) = \Phi(x-k)$, $k \in \mathcal{K}_0 \subset \mathbb{Z}^r$, where $\mathcal{K}_0$ is the set of possible translations for $k \in \mathbb{Z}$ such that $x-k \in [0,1]^r$. We define the vector of indices $\mathcal{I} = \{i = (i_1, \ldots, i_d) | i \in \{0,1\}^r,\ i_1 + \cdots + i_d \neq 0\}$, note that $|\mathcal{I}| = 2^r - 1$. Further define $\Psi^i(x) = \psi^{i_1}(x_1) \cdots \psi^{i_d}(x_d)$, $\Psi^i_{lk}(x) = 2^{lr/2}\Psi^i(2^l x - k)$, $l \in \mathbb{N} \cup \{0\}$, $k \in \mathcal{K}(l)$. Where $\mathcal{K}(l)$ are the possible translations for $k \in \mathbb{Z}$ such that $2^l x - k \in [0,1]^r$. It can be shown that any $f \in L_2(D)$ has the expansion

$$f = \sum_{k \in \mathcal{K}_0} \langle f, \Phi_k \rangle \Phi_k + \sum_{l=0}^{\infty} \sum_{k \in \mathcal{K}(l), i \in \mathcal{I}} \langle f, \Psi^i_{lk} \rangle \Psi^i_{lk}. \tag{3.9}$$

The inner products $\langle f, \Phi_k \rangle$ and $\langle f, \Psi^i_{lk} \rangle$ are generally referred as the wavelet coefficients. As the smoothness in f increases, the wavelet coefficients decay at a faster rate. In particular it can be shown that if $f \in \mathcal{C}^\beta_r(\mathcal{S}, C)$, then the wavelet coefficients satisfy $\sup_{k \in \mathcal{K}_0} \|\langle f, \Phi_k \rangle\|_\infty \leq M$ and

$$\sup_{k \in \mathcal{K}(l)} \|\langle f, \Psi_{lk} \rangle\|_\infty \leq M 2^{-l(\beta + r/2)} \quad \text{for all } l \in \{0, 1, \ldots, \infty\}$$

for some positive number $M > 0$ [cf. p. 351 95]. As a consequence, we have that

$$\|f\|^B_\beta = \|\langle f, \Phi_\cdot \rangle\|_\infty + \sup_l 2^{l(\beta + r/2)} \|\langle f, \Psi^i_{l\cdot} \rangle\|_\infty < \infty$$

for such an f. The class of all continuous functions with $\|f\|^B_\beta < \infty$ is the Besov space $B^\beta_{p,q}$ when $p = \infty$ and $q = \infty$, i.e.

$$B^\beta_{\infty,\infty}(\mathcal{S}) \equiv \{f : \mathcal{S} \mapsto \mathbb{R} : \|f\|^B_\beta < \infty, f \text{ is continuous}\}. \tag{3.10}$$

For all $\beta > 0$, therefore,

$$C^\beta_r(\mathcal{S}, C) \subset B^\beta_{\infty,\infty}(\mathcal{S}). \tag{3.11}$$

Moreover, as we mentioned earlier, when β is non-integer, it can be shown that $B^\beta_{\infty,\infty}(\mathcal{S}) = C^\beta_r(\mathcal{S}, C)$. In this case, the Besov norm $\|f\|^B_\beta$ is equivalent to the canonical norm of $C^\beta_r(\mathcal{S}, C)$ [cf. Proposition 4.3.23 of 95, for the one-dimensional case]. The above justifies the use of

wavelet series expansion to estimate a smooth function.

In light of the above discussion, to approximate an f with smoothness parameter β, it makes sense to use a truncated version of the wavelet series expansion in (3.9). Therefore, we consider $l \in \{0, \ldots, b_n\}$ where b_n is a sequence of integers diverging to ∞, and search for f in the class

$$\mathcal{H}_n(\mathcal{S}) = \Bigg\{ f_n : \mathcal{S} \mapsto \mathbb{R} \;\Bigg|\; f_n = \sum_{k \in \mathcal{K}_0} c_k \Phi_k + \sum_{l=0}^{b_n} \sum_{k \in \mathcal{K}(l), i \in \mathcal{I}} c_{lk} \Psi_{lk}^i \text{ where } c_k \in \mathbb{R} \text{ for } k \in \mathbb{N} \text{ and } c_{lk} \in \mathbb{R} \text{ for all } l \in \mathbb{Z} \text{ and } k \in \mathcal{K}(l) \Bigg\}. \tag{3.12}$$

Since the functions Φ_k's and Ψ_{lk}'s are orthogonal, the estimation procedure is not different from any other basis expansion method. Also because $c_k = \langle f_n, \Phi_k \rangle$ and $c_{lk} = \langle f, \Psi_{lk} \rangle$, $f_n \in \mathcal{H}_n(\mathcal{S})$ satisfies

$$\|f_n\|_\beta^B = \|\mathbf{c}_\cdot\|_\infty + \sup_{l \geq 0} 2^{l(\beta + r/2)} \sup_{k \in \mathcal{K}(l)} |\mathbf{c}_{l\cdot}|.$$

Define by

$$\mathcal{H}_n^M(\mathcal{S}) = \Bigg\{ f_n : \mathcal{S} \mapsto \mathbb{R} \;\Bigg|\; f_n = \sum_{k \in \mathcal{K}_0} c_k \Phi_k + \sum_{l=0}^{b_n} \sum_{k \in \mathcal{K}(l), i \in \mathcal{I}} c_{lk} \Psi_{lk}^i \text{ where } c_k \in \mathbb{R} \text{ } \|\mathbf{c}_\cdot\|_\infty + \sup_{l \geq 0} 2^{l(\beta + r/2)} \sup_{k \in \mathcal{K}(l)} |\mathbf{c}_{l\cdot}| < M \Bigg\}. \tag{3.13}$$

We consider the problem of maximizing $\widehat{V}_\psi(f_1, f_2)$ over $\mathcal{H}_n = \mathcal{H}_{1n} \times \mathcal{H}_{2n}$, where $f_1 \in \mathcal{H}_{1n} = \mathcal{H}_n^{Ca_n}(\mathcal{X}_1)$ and $f_2 \in \mathcal{H}_{2n} = \mathcal{H}_n^{Ca_n}(\mathcal{X}_2)$ for some $C > 0$. To that end, we first state a general result regarding approximation using $\mathcal{H}_n^M(\mathcal{S})$ for any compact set $\mathcal{S}$ and $M > 0$. The proof can be found in Appendix C.3.2.

Lemma 3.7.3. *Suppose the wavelet basis $(\Phi_\cdot, \Psi_{l\cdot},\ l \in \mathcal{Z})$ is S-regular for some $S > \beta \geq 1$ and $f_0 \in \mathcal{C}_r^\beta(\mathcal{S}, C)$ for some constant $C > 0$. Then for $\beta > r^2$, there exists a constant $M > 0$ depending on the compact $\mathcal{S}$ and the constant C such that*

$$\inf_{f \in \mathcal{H}_n^M(\mathcal{S})} \|f - f_0\|_\infty = \mathcal{O}\left(M 2^{-(\beta - d^2)(b_n + 1)} \right),$$

where $\mathcal{H}_n^M$ is as defined in (3.13).

Thus when f_0 is sufficiently smooth, i.e. if $\beta > r^2$, the tail of the expansion in (3.9) decays exponentially fast. The equivalence relation discussed earlier implies that if $f_0 \in \mathcal{C}_r^{\beta}(\mathcal{S}, C)$, then there exists $M > 0$, so that f_0 belongs to the norm-ball of radius M in $B^{\beta}_{\infty,\infty}([0,1]^r)$, to be denoted by $\mathcal{B}(d, \beta, M)$ from now on.

Note that Lemma 3.7.3 implies that given $\epsilon > 0$, there exists $f_n \in \mathcal{H}_n^M(\mathcal{S})$ so that $\|f_n - f_0\|_\infty \leq M\epsilon$ if we choose $b_n = \frac{-2\log_2 \epsilon}{\beta - d^2}$. Also, note that $a_n f_n, a_n f_0 \in \mathcal{H}^{M_n}(\mathcal{S})$ where $M_n = a_n M$.

The above, combined with Lemma 3.4.5 give us the following result.

Corollary 3.7.3.1. *Suppose the wavelet basis $(\Phi_\cdot, \Psi_{l\cdot},\ l \in \mathcal{Z})$ is as in Lemma 3.7.3, and G_1, G_2 satisfy Assumption 3.4.3. Then there exists a constant $M > 0$ and $C > 0$ so that taking $b_n \leq C \log a_n/(\beta - d^2)$, we can find $\tilde{f}_{n,1}, \tilde{f}_{n,2} \in \mathcal{H}_n^M(\mathcal{S})$ such that $a_n\tilde{f}_{n,1}, a_n\tilde{f}_{n,2} \in \mathcal{H}_n^{a_n M}(\mathcal{S})$ and*

$$V_\psi^* - V_\psi\Big(a_n C_y \tilde{f}_{n,1}, C_y \tilde{f}_{n,2}\Big) = O((\log a_n)^2/a_n),$$

where ϕ is as in Corollary 3.3.1.1 (i)-(iii). If ϕ is as in Corollary 3.3.1.1 (iv), then $\tilde{f}_{n,1}$ and $\tilde{f}_{n,2}$ satisfy

$$V_\psi^* - V_\psi\Big(a_n C_y \tilde{f}_{n,1}, C_y \tilde{f}_{n,2}\Big) = O(a_n^{-1}).$$

Therefore, using 3.7.3.1 along with Theorem 4.3.36 in [95] we get the results for the approximation, generalization error, and finally the $n^{-1/2}$ ψ-regret bound in Corollary 3.4.6.3.

Conclusion

In this dissertation, we have discussed and addressed some of the challenges that arise when RL is adapted to real-world settings. We are particularly motivated to learn optimal treatment regimes from EHR data. We develop three methods to address these challenges. ESRL aims to use Bayesian RL to build a safe policy. The semi-supervised learning approach for RL leverages data with unknown rewards to estimate an efficient policy function and a doubly-robust value function. Finally, our surrogate value function for direct policy learning, which is efficient and fully non-parametric.

We believe these methods are valuable off-policy learning tools to leverage observed data sets where direct exploration is not feasible (e.g., EHR data). These methods can be interpreted by clinicians and practitioners, making them more likely to get implemented in real-world settings. They account for uncertainty in the environment and the observed data either through posterior distribution credible intervals for ESRL, asymptotic standard errors for our SSL methods, and regret bounds for ESRL and the surrogate value methods. Moreover, the SSL approach and the regret bounds shown for ESRL and surrogate value function methods ensure our inference is sample-efficient, an important quality for limited data settings. Additionally, we control for sampling bias by inverse probability weighting and using a doubly robust estimator for the value function estimate. Finally, as this work is all based on observational data sets, we provide theoretical results for optimizing the policy functions estimated with our methods. These results ensure that the learned policies are safe to be implemented in the real world. We hope our work motivates the extension of these and other RL methods with similar goals to eventually contribute to improving patient care, ultimately advancing the field of RL

for real-world settings.

References

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press Cambridge, Massachusetts London, England, 2017.

[2] Christopher John Cornish Hellaby Watkins. Learning from delayed rewards, 1989.

[3] Bibhas Chakraborty and Erica E.M Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Statistics for Biology and Health. Springer New York, New York, NY, 2013 edition, 2013.

[4] Phillip J. Schulte, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. **Q**- and **A**-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.*, 29(4):640–661, 11 2014.

[5] Michael R. Kosorok and Eric B. Laber. Precision medicine. 6(1):263–286, 2019.

[6] Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. 2016.

[7] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv.org*, 2016.

[8] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16, 2019.

[9] Omer Gottesman, Fredrik D. Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li-Wei H. Lehman, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David A. Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *CoRR*, abs/1805.12298, 2018.

[10] Leslie Pack Kaelbling. *Learning in Embedded Systems.* A Bradford Book Ser. 1993.

[11] Martha White and Adam White. Interval estimation for reinforcement-learning algorithms in continuous-state domains. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2433–2441. Curran Associates, Inc., 2010.

[12] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. 2018.

[13] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning, 2019.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[15] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

[16] Brendan O'Donoghue, Ian Osband, Rémi Munos, and Volodymyr Mnih. The uncertainty bellman equation and exploration. *CoRR*, abs/1709.05380, 2017.

[17] Richard Dearden, Nir Friedman, and Stuart J. Russell. Bayesian q-learning. In *AAAI/IAAI*, 1998.

[18] John Asmuth, Lihong Li, Michael L. Littman, Ali Nouri, and David Wingate. A bayesian sampling approach to exploration in reinforcement learning. 2012.

[19] Alberto Maria Metelli, Amarildo Likmeta, and Marcello Restelli. Propagating uncertainty in reinforcement learning via wasserstein barycenters. In H. Wallach, H. Larochelle, A. Beygelzimer, Falche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4333–4345. Curran Associates, Inc., 2019.

[20] Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proceedings of the 22nd international conference on machine learning*, volume 119 of *ICML '05*, pages 201–208. ACM, 2005.

[21] Marc Deisenroth and Carl Rasmussen. Pilco: A model-based and data-efficient approach to policy search. pages 465–472, 01 2011.

[22] Malcolm Strens. A bayesian framework for reinforcement learning. In *In Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950. ICML, 2000.

[23] Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. 2013.

[24] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? 2016.

[25] Caglar Gulcehre, Ziyu Wang, Alexander Novikov, Tom Le Paine, Sergio Gomez Colmenarejo, Konrad Zolna, Rishabh Agarwal, Josh Merel, Daniel Mankowitz, Cosmin Paduraru, Gabriel Dulac-Arnold, Jerry Li, Mohammad Norouzi, Matt Hoffman, Ofir Nachum, George Tucker, Nicolas Heess, and Nando de Freitas. Rl unplugged: Benchmarks for offline reinforcement learning, 2020.

[26] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2020.

[27] Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. Benchmarking batch deep reinforcement learning algorithms, 2019.

[28] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *CoRR*, abs/1906.00949, 2019.

[29] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel : Model-based offline reinforcement learning, 2020.

[30] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning, 2019.

[31] Quentin F. Gronau, Alexander Ly, and Eric-Jan Wagenmakers. Informed bayesian t-tests. *The American Statistician*, 74(2):137–143, 2020.

[32] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[33] Johnson A. E. W, Pollard T. J., Shen L., Lehman L.-W. H., Feng M., Ghassemi M., Moody B., Szolovits P., Anthony Celi L., , and R. G. Mark. A freely accessible critical care database mimic-iii. *Scientific Data*, 4(160035), 2016.

[34] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062, 2019.

[35] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment - a deep reinforcement learning approach. 2017.

[36] Michael R. Kosorok and Eric B. Laber. Precision medicine. 6(1):263–286, 2019.

[37] J. M. Robins. *Optimal structural nested models for optimal sequential decisions.* Lin, D. Y. Proceedings of the Second Seattle Symposium in Biostatistics : Analysis of Correlated Data Lecture notes in statistics (Springer-Verlag) ; 179. Springer New York : Imprint: Springer, New York, NY, 2004.

[38] SA Murphy. A generalization error for q-learning. *Journal Of Machine Learning Research*, 6:1073–1097, 2005.

[39] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[40] Ying-Qi Zhao, Donglin Zeng, Eric B. Laber, and Michael R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110:583–598, 2015.

[41] Yichi Zhang, Tianrun Cai, Sheng Yu, Kelly Cho, Chuan Hong, Jiehuan Sun, Jie Huang, Yuk-Lam Ho, Ashwin N Ananthakrishnan, Zongqi Xia, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (phecap). *Nature Protocols*, 14(12):3426–3444, 2019.

[42] David Cheng, Ashwin N Ananthakrishnan, and Tianxi Cai. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, 2020.

[43] Abhishek Chakrabortty, Tianxi Cai, et al. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

[44] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 2006.

[45] Xiaojin Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2008.

[46] John Blitzer and Xiaojin Zhu. Semi-supervised learning for natural language processing. In *ACL (Tutorial Abstracts)*, page 3, 2008.

[47] Wang Zhixing and Chen Shaohong. Web page classification based on semi-supervised naïve bayesian em algorithm. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 242–245. IEEE, 2011.

[48] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition, 2018.

[49] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

[50] Chelsea Finn, Tianhe Yu, Justin Fu, Pieter Abbeel, and Sergey Levine. Generalizing skills with semi-supervised reinforcement learning. 2016.

[51] Larry Wasserman and John D. Lafferty. Statistical analysis of semi-supervised regression. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 801–808. Curran Associates, Inc., 2008.

[52] J. Robins. Causal inference from complex longitudinal data. *Latent Variable Modeling and Applications to Causality*, pages 69—117, 1997.

[53] Richard S. Sutton. *Reinforcement learning : an introduction*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts ; London, England, second edition. edition, 2018.

[54] Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. Dynamic treatment regimes: technical challenges and applications. *Electronic journal of statistics*, 8(1):1225–1272, 2014.

[55] Abhishek Chakrabortty. Robust semi-parametric inference in semi-supervised settings, 2016.

[56] Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, New York, NY, 2006.

[57] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[58] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716, 2015.

[59] G Biau, L Devroye, and G Lugosi. Consistency of random forests and other averaging classifiers. *Journal Of Machine Learning Research*, 9:2015–2033, 2008.

[60] Wensheng Zhu, Donglin Zeng, and Rui Song. Proper inference for value function in high-dimensional q-learning for dynamic treatment regimes. *Journal of the American Statistical Association*, 114(527):1404–1417, 2019.

[61] A. W. van der Vaart. *Asymptotic statistics*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, Cambridge, UK ; New York, NY, USA, 1998.

[62] T.J Hastie. *Statistical Models in S*. CRC Press, 1 edition, 1992.

[63] L Peyrin-Biroulet. Anti-tnf therapy in inflammatory bowel diseases: a huge review. *Minerva gastroenterologica e dietologica*, 56(2):233, 2010.

[64] Bruce E Sands, Laurent Peyrin-Biroulet, Edward V Loftus Jr, Silvio Danese, Jean-Frédéric Colombel, Murat Törüner, Laimas Jonaitis, Brihad Abhyankar, Jingjing Chen, Raquel Rogers, et al. Vedolizumab versus adalimumab for moderate-to-severe ulcerative colitis. *New England Journal of Medicine*, 381(13):1215–1226, 2019.

[65] Toshihiro Inokuchi, Sakuma Takahashi, Sakiko Hiraoka, Tatsuya Toyokawa, Shinjiro Takagi, Koji Takemoto, Jiro Miyaike, Tsuyoshi Fujimoto, Reiji Higashi, Yuki Morito, et al. Long-term outcomes of patients with crohn's disease who received infliximab or adalimumab as the first-line biologics. *Journal of gastroenterology and hepatology*, 34(8):1329–1336, 2019.

[66] Yongil Lee, Jae Hee Cheon, Yehyun Park, Soo Jung Park, Tae Il Kim, and Won Ho Kim. Comparison of long-term outcomes between infliximab and adalimumab in biologic-naive patients with ulcerative colitis. *Gut & Liver*, 13, 2019.

[67] Mark T Osterman and Gary R Lichtenstein. Infliximab vs adalimumab for uc: Is there a difference? *Clinical Gastroenterology and Hepatology*, 15(8):1197–1199, 2017.

[68] An Ananthakrishnan, A Cagan, Tianxi Cai, Vs Gainer, S Shaw, S Churchill, E Karlson, I Kohane, K Liao, and S Murphy. Comparative effectiveness of infliximab and adalimumab in crohn's disease and ulcerative colitis. *Gastroenterology*, 150(4):S979–S979, 2016.

[69] An Ananthakrishnan, Tianxi Cai, SC Cheng, Pj Chen, G Savova, RG Perez, Vs Gainer, Sn Murphy, P Szolovits, K Liao, Ew Karlson, S Churchill, I Kohane, and RM Plenge. Improving case definition of crohn's disease and ulcerative colitis in electronic medical

records using natural language processing - a novel informatics approach. *Gastroenterology*, 142(5):S791–S791, 2012.

[70] Mary E Charlson, Peter Pompei, Kathy L Ales, and C.Ronald Mackenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373–383, 1987.

[71] S A Murphy, M J van der Laan, and J M Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[72] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[73] Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020.

[74] Peng Liao, Zhengling Qi, and Susan Murphy. Batch policy learning in average reward markov decision processes, 2020.

[75] Zeng D. Rush A. J. Kosorok M. R. Zhao, Y. Q. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107:1106–1118, 2012.

[76] Susan A Murphy, David W Oslin, A. John Rush, and Ji Zhu. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology (New York, N.Y.)*, 32(2):257–262, 2007.

[77] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001.

[78] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[79] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. 1983.

[80] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[81] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning, overview of mini-batch gradient descent, 2012. Available at `https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf`.

[82] Yi Lin. A note on margin-based loss functions in classification. *Statistics probability letters*, 68(1):73–82, 2004.

[83] Shuai Chen, Lu Tian, Tianxi Cai, and Menggang Yu. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209, 2017.

[84] Binyan Jiang, Rui Song, Jialiang Li, and Donglin Zeng. Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 29(4):1633–1655, 2019.

[85] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

[86] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*, pages 341–358, 2011.

[87] Jingwei Zhang, Tongliang Liu, and Dacheng Tao. On the rates of convergence from surrogate risk minimizers to the bayes optimal classifier. *arXiv preprint arXiv:1802.03688*, 2018.

[88] Jean-Yves Audibert, Alexandre B Tsybakov, et al. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

[89] Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

[90] M. S. Andersen, J. Dahl, and L. Vandenberghe. Cvxopt: A python package for convex optimization, version 1.2, 2021. Available at https://cvxopt.org, 2021.

[91] A Johnson, L Bulgarelli, T Pollard, S Horng, L A Celi, and R Mark. Mimic-iv (version 0.4). physionet., 2020.

[92] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11), 2018.

[93] Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, and Peter Szolovits. Expert-supervised reinforcement learning for offline policy learning and evaluation. In *Advances in Neural Information Processing Systems*, volume 33, pages 18967–18977, 2020.

[94] Aaron Sonabend, Nilanjana Laha, Ashwin N. Ananthakrishnan, Tianxi Cai, and Rajarshi Mukherjee. Semi-supervised off policy reinforcement learning, 2021.

[95] Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge, 2015.

[96] R.M Dudley. Balls in rk do not cut all subsets of k + 2 points. *Advances in mathematics (New York. 1965)*, 31(3):306–308, 1979.

[97] Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, New York, 1996.

[98] Aad W. Van Der Vaart and Jon A. Wellner. Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*, 55:234–252, 2007.

[99] Vladimir Koltchinskii. 2008 saint flour lectures oracle inequalities in empirical risk minimization and sparse recovery problems. 2009.

[100] Aad Van Der Vaart and Jon A Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.

Appendix A

Appendix to Chapter 1

Supplementary Material for Expert-Supervised Reinforcement Learning for Offline Policy Learning and Evaluation

A.1 Off-Policy Policy Evaluation and Uncertainty Estimation

In this Section, we follow the lines of Section 4 in the main text with more discussion. We show an Algorithm that collects the ideas presently discussed and an additional Lemma regarding the convergence of the null probability estimator.

We leverage $f(\cdot|\boldsymbol{D}_T)$ to estimate the value function for any policy, and use hypothesis testing for whether there is a meaningful difference in two policy functions (i.e. μ^α vs. π). Recall, we compute the estimated value of a given policy $\tilde{\mu}$, by sampling K models from the posterior and navigating M_k using $\tilde{\mu}$ to obtain $V_{\tilde{\mu},1}^{M_k} \sim f_V(\cdot|\boldsymbol{D}_T)$. We estimate $\mathbb{E}\left[V_{\tilde{\mu}}^{M^*}|\boldsymbol{D}_T\right]$ with $\hat{V}_{\tilde{\mu}} = \frac{1}{K}\sum_{k=1}^{K} V_{\tilde{\mu},1}^{(k)}$. This process is shown in Algorithm 2.

Algorithm 2: Value function estimation

for $k = 1, \ldots, K$ **do**

Set $V_0^{(k)} \leftarrow 0$;

Sample $M_k \sim f(\cdot|\boldsymbol{D}_T)$, $k = 1, \ldots, K$;

Sample $s \sim P_0^{M_k}$;

for $t = 1, \ldots, \tau$ **do**

$a \leftarrow \tilde{\mu}(s, t)$;

$V_t^{(k)} \leftarrow V_{t-1}^{(k)} + \bar{R}^{M_k}(s, a)$;

Sample $s' \sim P_a^{M_k}(s'|s)$;

Set $s \leftarrow s'$;

end

Set $V_{\tilde{\mu},1}^{(k)} \leftarrow V_\tau^{(k)}$;

end

Note that we average over the initial states as well, as we are interested to know the marginal value of the policy. A conditional value of the policy function $V_{\tilde{\mu},1}^{M^*}(s)$ can also be computed simply by starting all samples at a fixed state. Analogous to Section 3, we use samples $\left\{V_{\tilde{\mu},1}^{(k)}\right\}_{k=1}^{K}$ to define a $(1-\alpha)$ CI using the α and $1-\alpha$ quantiles. Note that for policies which are very different from the behavior policy, the posterior distribution will have wider CIs due to the wide distribution shift. This signals that there is not enough information in $\boldsymbol{D}_T$ for the rarely visited state-action pairs (s, a). This happens with OPPE importance sampling estimators as well [9]. As opposed to only considering point estimators of the value function, these CI help to assess whether the estimated value is likely to be accurate or if the estimate is unreliable given the information in $\boldsymbol{D}_T$. Importance sampling based estimators reflect this large distribution shift in high variance estimators.

Policy-level hypothesis testing. We use Algorithm 2 to assess whether there is a statistically significant difference in value from two different policies. Define the value function null hypothesis for two fixed policies $\tilde{\mu}_1, \tilde{\mu}_2$ as the event in which policy $\tilde{\mu}_1$ has a higher expected

value than $\tilde{\mu}_2$ conditional on $\boldsymbol{D}_T$: $H_0 : \mathbb{E}_{s\sim P_0,M^*}\left[V_{\tilde{\mu}_1}(s)|\boldsymbol{D}_T\right] > \mathbb{E}_{s\sim P_0,M^*}\left[V_{\tilde{\mu}_2}(s)|\boldsymbol{D}_T\right]$. The probability of the null under the true model M^* is

$$\mathbb{P}^*_\mu\left(H_0|\boldsymbol{D}_T\right) = \mathbb{P}\left(V^{M^*}_{\tilde{\mu}_1}(s) > V^{M^*}_{\tilde{\mu}_2}(s)\middle|\boldsymbol{D}_T\right) = \sum_{s\in\mathcal{S}} P_0(s)\mathbb{P}\left(V_{\tilde{\mu}_1}(s) > V_{\tilde{\mu}_2}(s)\middle|s,\boldsymbol{D}_T\right).$$

We use the following estimator from samples generated from Algorithm 2:

$$\hat{\mathbb{P}}_\mu\left(H_0|\boldsymbol{D}_T\right) = \frac{1}{K}\sum_{k=1}^{K} I\left(V^{M_k}_{\tilde{\mu}_1}(s) - V^{M_k}_{\tilde{\mu}_2}(s) > 0\right). \tag{A.1}$$

Lemma A.1.1. *Let μ_1, $\mu_2 : \mathcal{S}\times\{1,\dots,\tau\}$ be two pre-specified policy functions, and let $\hat{P}_\mu(H_0|\boldsymbol{D}_T)$ be defined as in* (A.1)*,*

$$\hat{\mathbb{P}}_\mu\left(H_0|\boldsymbol{D}_T\right) - \mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right) = O_p\left(K^{-\frac{1}{2}}\right),$$

Lemma A.1.1 ensures consistency of the probability of the null-hypothesis for the value function testing.

A.2 Supporting Lemma

Lemma A.2.1. *(Lemma 1 in [23]) If f is the distribution of M^* then, for any $\sigma(\boldsymbol{D}_T)-$measurable function g, and model $M_k \sim f(\cdot|\boldsymbol{D}_T)$:*

$$\mathbb{E}\left[g(M^*)|\boldsymbol{D}_T\right] = \mathbb{E}\left[g(M_k)|\boldsymbol{D}_T\right].$$

A.3 Proof of results in main body

A.3.1 Theorem 3.4

In this Subsection we develop the necessary definitions and lemmas, and eventually go on to prove Theorem 3.4. To simplify notation let $\mathbb{P}^*(H_0) \equiv \mathbb{P}^*(H_0|s,t,\boldsymbol{D}_T)$ and $\hat{\mathbb{P}}(H_0) \equiv \hat{\mathbb{P}}(H_0|s,t,\boldsymbol{D}_T)$. Given the behavior policy as defined in Algorithm 1 and the optimal policy under the true MDP M^*, we can write the ESRL policy obtained from any M_k sample from

Algorithm 1, and it's equivalent version under M^* as:

$$\mu_k^{\alpha}(s,t) = I\left(\hat{\mathbb{P}}(H_0) < \alpha\right)\mu_k(s,t) + I\left(\hat{\mathbb{P}}(H_0) \geq \alpha\right)\pi(s,t),$$
$$\mu^{\alpha *}(s,t) = I\left(\mathbb{P}^*(H_0) < \alpha\right)\mu^*(s,t) + I\left(\mathbb{P}^*(H_0) \geq \alpha\right)\pi(s,t),$$

we show our result is true for any μ_k^{α} and thus it's true for the ESRL policy μ^{α}. Next we define the policy μ_k^{α} which uses the true null probabilities and μ_k as:

$$\mu_k^{\alpha *}(s,t) = I\left(\mathbb{P}^*(H_0) < \alpha\right)\mu_k(s,t) + I\left(\mathbb{P}^*(H_0) \geq \alpha\right)\pi(s,t).$$

finally let

$$\Delta_i^{\mu} = \sum_{s\in\mathcal{S}} P_0(s)\left(V_{\mu_k^{\alpha *},1}^{M^*}(s) - V_{\mu_k^{\alpha},1}^{M^*}(s)\right)$$
$$\Delta_i^{*} = \sum_{s\in\mathcal{S}} P_0(s)\left(V_{\mu_k^{\alpha *},1}^{M_k}(s) - V_{\mu_k^{\alpha *},1}^{M^*}(s)\right).$$

Consider function $g : M \mapsto V_{\mu^{\alpha}*,1}^{M}$, g is $\sigma(\boldsymbol{D}_T)$ measurable for a fixed $\alpha \in [0,1]$ as $\pi(s,t)$, $\mathbb{P}^*(H_0)$ are fixed $\forall (s,t) \in \mathcal{S}\times\{1,\ldots,\tau\}$, thus, by Lemma A.2.1 for any $M_k \sim f(\cdot|\boldsymbol{D}_T)$

$$\mathbb{E}\left[V_{\mu_k^{\alpha *},1}^{M_k}(s)|\boldsymbol{D}_T\right] = \mathbb{E}\left[V_{\mu^{\alpha *},1}^{M^*}(s)|\boldsymbol{D}_T\right],$$

now using iterated expectations we get $\mathbb{E}\left[V_{\mu_k^{\alpha *},1}^{M_k}(s)\right] = \mathbb{E}\left[V_{\mu^{\alpha *},1}^{M^*}(s)\right]$.

We use this to re-express the expected regret for episode i under model k computed with Algorithm 1 as

$$\begin{aligned}
\mathbb{E}\left[\Delta_i\right] &= \mathbb{E}\left[\sum_{s\in\mathcal{S}} P_0(s)\left(V_{\mu^{\alpha *},1}^{M^*}(s) - V_{\mu_k^{\alpha},1}^{M^*}(s)\right)\right] \\
&= \sum_{s\in\mathcal{S}} P_0(s)\left(\mathbb{E}\left[V_{\mu^{\alpha *},1}^{M^*}(s)\right] - \mathbb{E}\left[V_{\mu_k^{\alpha},1}^{M^*}(s)\right]\right) \\
&= \sum_{s\in\mathcal{S}} P_0(s)\left(\mathbb{E}\left[V_{\mu_k^{\alpha *},1}^{M_k}(s)\right] - \mathbb{E}\left[V_{\mu_k^{\alpha},1}^{M^*}(s)\right]\right) \\
&= \mathbb{E}\left[\Delta_i^{*}\right] + \mathbb{E}\left[\Delta_i^{\mu}\right],
\end{aligned}$$

where the last step follows from adding and subtracting $\mathbb{E}\left[V^{M^*}_{\mu_k^{\alpha*},1}(s)\right]$.

We first consider $\mathbb{E}\left[\Delta_i^*\right]$, we use a strategy similar to [23], but do not make an *iid* assumption for within-episode observations. Define the following Bellman operator $\mathcal{T}^M_{\mu^\alpha}$ for any MDP M, policy μ^α, and value function V to be

$$\mathcal{T}^M_{\mu^\alpha} V(s) = \bar{R}^M(s, \mu^\alpha(s,t)) + \sum_{s' \in \mathcal{S}} P^M_{\mu^\alpha(s,t)}(s'|s) V(s'), \tag{A.2}$$

this lets us write $V^M_{\mu^\alpha,t}(s) = \mathcal{T}^M_{\mu^\alpha} V^M_{\mu^\alpha,t+1}(s)$.
The next Lemma will let us express term $\mathbb{E}\left[\Delta_i^* \middle| M^*, M_k\right]$ in terms of the Bellman operator.

Lemma A.3.1. *If f is the distribution of M^*, then*

$$\mathbb{E}\left[\Delta_i^* \middle| M^*, M_k\right] = \mathbb{E}\left[\sum_{j=1}^{\tau} \left(\mathcal{T}^{M_k}_{\mu_k^{\alpha*}(\cdot,j)} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,j)}\right) V^{M_k}_{\mu_k^{\alpha*},j+1}(s_{j+1}) \middle| M^*, M_k\right].$$

We now define a confidence set for the reward and transition estimated probabilities.

Lemma A.3.2. *Let $\mathcal{I}$ denote the set of index i, j for episodes in $\boldsymbol{D}_T = \{(s_{i1}, a_{i1}, r_{i1}, \ldots, s_{i\tau}, a_{i\tau}, r_{i\tau})\}_{i=1}^T$, that is: $\mathcal{I} = \left\{(i,j) \middle| i \in \{1,\ldots,T\}, j \in \{1,\ldots,\tau\}\right\}$. Further let $N_T(s,a)$ be the number of times (s,a) was sampled in $\boldsymbol{D}_T$: $N_T(s,a) = \sum_{i,j \in \mathcal{I}} I(S_{ij} = s, A_{ij} = a)$, let $\hat{P}_a(\cdot|s)$ and $\hat{R}(s,a)$ be non-parametric estimators for the distribution of transitions and rewards observed after sampling T episodes:*

$$\hat{P}_a(s'|s) = \frac{\sum_{i,j \in \mathcal{I}} I(s_{i,j+1} = s') I(s_{ij} = s, a_{ij} = a)}{N_T(s,a)}, \;\; \hat{R}(s,a) = \frac{\sum_{ij \in \mathcal{I}} I(s_{ij} = s, a_{ij} = a) r_{ij}}{N_T(s,a)}.$$

Define the confidence set:

$$\mathcal{M}_T \equiv \left\{M : \left\|\hat{P}_a(\cdot|s) - P^M_a(\cdot|s)\right\|_1 \leq \beta_T(s,a), \left|\hat{R}(s,a) - R^M(s,a)\right|_1 \leq \beta_T(s,a) \; \forall (s,a)\right\},$$

where $\beta_T(s,a) \equiv \frac{\sqrt{8ST\log(2SAT)}}{\max\{1, N_T(s,a)\}}$, then $P\left(M^* \notin \mathcal{M}_T\right) < \frac{1}{T}$.

Proof of Theorem 3.4. We start by summing Δ_i^* over all episodes:

$$\mathbb{E}\left[\sum_{i=1}^{T}\Delta_i^*\right] \leq \mathbb{E}\left[\sum_{i=1}^{T}\Delta_i^* I(M_k, M^* \in \mathcal{M}_T)\right] + \tau\sum_{i=1}^{T}\left(\mathbb{P}(M_k \notin \mathcal{M}_T) + \mathbb{P}(M^* \notin \mathcal{M}_T)\right)$$
$$\leq \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{T}\Delta_i^* | M_k, M^*\right] I(M_k, M^* \in \mathcal{M}_k)\right] + 2\tau$$
$$\leq \mathbb{E}\left[\sum_{i=1}^{T}\sum_{j=1}^{\tau}\left|\left(\mathcal{T}^{M_k}_{\mu_k^{\alpha*}(\cdot,j)} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,j)}\right) V^{M_k}_{\mu_k^{\alpha*},j+1}(s_j)\right| I(M_k, M^* \in \mathcal{M}_k)\right] + 2\tau$$

where the first step follows by conditioning on event $I(M_k \in \mathcal{M}_T, M^* \in \mathcal{M}_T)$ and it's complement, and from the fact that $\Delta_i^* \leq \tau$ as all rewards $R(s,a) \in [0,1]$. The second step follows from iterated expectations and Lemma A.3.2 as $\mathbb{P}[I(M^* \notin \mathcal{M}_T)] \leq \frac{1}{T}$. Also since $\mathcal{M}_T$ is a $\sigma(D_T)$-measurable function by Lemma A.2.1 we have $\mathbb{E}\left[I\left(M_k \notin \mathcal{M}_T\right)|D_T\right] = \mathbb{E}\left[I\left(M^* \notin \mathcal{M}_T\right)|D_T\right]$, using iterated expectations we have $\mathbb{P}[I(M_k \notin \mathcal{M}_T)] \leq \frac{1}{T}$. The last step follows from Lemma A.3.1. Next using (A.2) the last equation can be re-written as

$$\mathbb{E}\left[\sum_{i=1}^{T}\sum_{j=1}^{\tau} I\left(\mathbb{P}^*(H_0) \geq \alpha\right)\left\{\bar{R}^{M_k}(s,\pi(s,j)) - \bar{R}^{M^*}(s,\pi(s,j))\right\} I(M_k, M^* \in \mathcal{M}_k)\right]$$
$$+ \mathbb{E}\left[\sum_{i=1}^{T}\sum_{j=1}^{\tau} I\left(\mathbb{P}^*(H_0) \geq \alpha\right)\left\{\sum_{s'\in\mathcal{S}}\left|P^{M_k}_{\pi(s,j)}(s'|s) - P^{M^*}_{\pi(s,j)}(s'|s)\right| V^{M_k}_{\mu_k^{\alpha*},j+1}(s_{j+1})\right\} I(M_k, M^* \in \mathcal{M}_k)\right]$$
$$+ \mathbb{E}\left[\sum_{i=1}^{T}\sum_{j=1}^{\tau} I\left(\mathbb{P}^*(H_0) < \alpha\right)\left\{\bar{R}^{M_k}(s,\mu_k(s,j)) - \bar{R}^{M^*}(s,\mu_k(s,j))\right\} I(M_k, M^* \in \mathcal{M}_k)\right]$$
$$+ \mathbb{E}\left[\sum_{i=1}^{T}\sum_{j=1}^{\tau} I\left(\mathbb{P}^*(H_0) < \alpha\right)\left\{\sum_{s'\in\mathcal{S}}\left|P^{M_k}_{\mu_k(s,j)}(s'|s) - P^{M^*}_{\mu_k(s,j)}(s'|s)\right| V^{M_k}_{\mu_k^{\alpha*},j+1}(s_{j+1})\right\} I(M_k, M^* \in \mathcal{M}_k)\right]$$
$$+ 2\tau$$
$$\leq \mathbb{E}\left[\tau\sum_{i=1}^{T}\sum_{j=1}^{\tau}\min\left\{\beta_T(s_{ij},\pi(s_{ij},j)),1\right\}\right] + \mathbb{E}\left[\tau\sum_{i=1}^{T}\sum_{j=1}^{\tau}\min\left\{\beta_T(s_{ij},\mu_k(s_{ij},j)),1\right\}\right] + 2\tau,$$

where the last step follows by Lemma A.3.2, next:

$$\leq \mathbb{E}\left[\tau \sum_{i=1}^{T}\sum_{j=1}^{\tau} \frac{\sqrt{8ST\log(2SAT)}}{\min\{N_T(s_{ij},\mu_k(s_{ij},j))\}}\right] + \mathbb{E}\left[\tau \sum_{i=1}^{T}\sum_{j=1}^{\tau} \frac{\sqrt{8ST\log(2SAT)}}{\min\{1, N_T(s_{ij},\pi(s_{ij},j))\}}\right] + 2\tau$$

$$\leq M_1\sqrt{\tau^2 SAT} + M_2\tau\sqrt{S^2AT\log(SAT)} + 2\tau < M_3\tau S\sqrt{AT\log(SAT)} + 2\tau,$$

where the last step follows by Appendix B in [23] with constants M_1, M_2, M_3.

We next analyze

$$\mathbb{E}\left[\Delta_i^{\mu}\right] = \sum_{s\in\mathcal{S}} P_0(s)\left(\mathbb{E}\left[V_{\mu_k^{\alpha*},1}^{M^*}(s)\right] - \mathbb{E}\left[V_{\mu_k^{\alpha},1}^{M^*}(s)\right]\right).$$

We can write the second term as

$$\mathbb{E}\left[V_{\mu_k^{\alpha},1}^{M^*}(s)\right] = \mathbb{E}\left[\sum_{j=1}^{\tau} I\left(\hat{\mathbb{P}}(H_0) < \alpha\right) R^{M^*}(s_j,\mu_k(s_j,j)) + I\left(\hat{\mathbb{P}}(H_0) \geq \alpha\right) R^{M^*}(s_j,\pi(s_j,j))\right],$$

we extend the null probability notation to be explicit on the time index: $\mathbb{P}_j^*(H_0) = \mathbb{P}^*(H_0|s_j, j, \boldsymbol{D}_T), \hat{\mathbb{P}}_j(H_0) = \hat{\mathbb{P}}(H_0|s_j, j, \boldsymbol{D}_T)$. By Lemma 3.2, $\exists\delta > 0$ such that $\hat{\mathbb{P}}_j(H_0) - \mathbb{P}_j^*(H_0) \leq \delta\ \forall s \in \mathcal{S}, j \in \{1,\ldots,\tau\}$ with high probability, therefore

$$\begin{aligned} \mathbb{P}_j^*(H_0) < \alpha - \delta &\implies \mathbb{P}\left(\hat{\mathbb{P}}_j(H_0) < \alpha\right) = 1 - O_p\left(K^{-\frac{1}{2}}\right), \\ \mathbb{P}_j^*(H_0) \geq \alpha + \delta &\implies \mathbb{P}\left(\hat{\mathbb{P}}_j(H_0) \geq \alpha\right) = 1 - O_p\left(K^{-\frac{1}{2}}\right). \end{aligned} \tag{A.3}$$

As $\mathcal{I}_1$, $\mathcal{I}_2$ in Algorithm 1 are mutually exclusive, $\hat{\mathbb{P}}_j(H_0)$ are independent to $\mu_k(s,j)\ \forall s \in \mathcal{S}, j \in \{1,\ldots,\tau\}$, therefore starting with $V_{\mu_k^{\alpha},\tau}^{M^*}(s)$ we have

$$
\begin{aligned}
&\mathbb{E}\left[V^{M^*}_{\mu_k^{\alpha},\tau}(s)\right] \\
=&I\left(\mathbb{P}^*_\tau(H_0) < \alpha-\delta\right)\left\{\mathbb{E}\left[I\left(\hat{\mathbb{P}}_\tau(H_0)<\alpha\right)\right]\bar{R}^{M^*}(s_\tau,\mu_k(s_\tau,\tau)) + \mathbb{E}\left[I\left(\hat{\mathbb{P}}_\tau(H_0)\geq\alpha\right)\right]\bar{R}^{M^*}(s_\tau,\pi(s_\tau,\tau))\right\} \\
+&I\left(\mathbb{P}^*_\tau(H_0) \geq \alpha-\delta\right)\left\{\mathbb{E}\left[I\left(\hat{\mathbb{P}}_\tau(H_0)<\alpha\right)\right]\bar{R}^{M^*}(s_\tau,\mu_k(s_\tau,\tau)) + \mathbb{E}\left[I\left(\hat{\mathbb{P}}_\tau(H_0)\geq\alpha\right)\right]\bar{R}^{M^*}(s_\tau,\pi(s_\tau,\tau)\right\} \\
+&I\left(\mathbb{P}^*_\tau(H_0) \in [\alpha-\delta,\alpha+\delta)\right)\left\{\mathbb{E}\left[I\left(\hat{\mathbb{P}}_\tau(H_0)<\alpha\right)\right]\bar{R}^{M^*}(s_\tau,\mu_k(s_\tau,\tau)) + \mathbb{E}\left[I\left(\hat{\mathbb{P}}_\tau(H_0)\geq\alpha\right)\right]\bar{R}^{M^*}(s_\tau,\pi(s_\tau,\tau))\right\} \\
=&I\left(\mathbb{P}^*_\tau(H_0) < \alpha-\delta\right)\bar{R}^{M^*}(s_\tau,\mu_k(s_\tau,\tau)) + O_p\left(K^{-\frac{1}{2}}\right) \\
+&I\left(\mathbb{P}^*_\tau(H_0) \geq \alpha-\delta\right)\bar{R}^{M^*}(s_\tau,\mu_k(s_\tau,\tau)) + O_p\left(K^{-\frac{1}{2}}\right) \\
+&I\left(\mathbb{P}^*_\tau(H_0) \in [\alpha-\delta,\alpha+\delta)\right)O_p\left(K^{-\frac{1}{2}}\right) \\
=&\mathbb{E}\left[V^{M^*}_{\mu_k^{\alpha*},\tau}(s)\right] + O_p\left(K^{-\frac{1}{2}}\right),
\end{aligned}
$$

where the first step follows from $\mathcal{I}_1$, $\mathcal{I}_2$ being independent, the second step follows from (A.3) and last step from definition of $V^{M^*}_{\mu_k^{\alpha*},\tau}(s)$. Iterating backards from $\tau-1\ldots,1$ and applying the same steps as above we get

$$
\mathbb{E}\left[V^{M^*}_{\mu_k^{\alpha},1}(s)\right] = \mathbb{E}\left[V^{M^*}_{\mu_k^{\alpha*},1}(s)\right] + O_p\left(\tau K^{-\frac{1}{2}}\right).
$$

therefore we have $\mathbb{E}\left[\sum_{i=1}^T \Delta_i^{\mu}\right] = O_p\left(T\tau K^{-\frac{1}{2}}\right)$, choosing $K=\mathcal{O}(T)$ we get $\mathbb{E}\left[\sum_{i=1}^T \Delta_i^{\mu}\right] = O_p\left(\sqrt{T}\tau\right)$ which is dominated by $\mathbb{E}\left[\sum_{i=1}^T \Delta_i^{*}\right]$.

Putting both terms together we have

$$
\mathbb{E}\left[\sum_{i=1}^T \Delta_i\right] = \mathbb{E}\left[\sum_{i=1}^T \Delta_i^{*}\right] + \mathbb{E}\left[\sum_{i=1}^T \Delta_i^{\mu}\right] = \mathcal{O}\left(\tau S\sqrt{AT\log(SAT)}\right).
$$

□

A.3.2 Proofs for other results in main body

Proof of Lemma 3.1. To establish $\hat{Q}_{\tilde{\mu},t}(s,a)$ is unbiased, note that for any fixed (t,s,a), $M_k \sim f(\cdot|\boldsymbol{D}_T)$ are *iid*, now for a given policy function $\tilde{\mu}$:

$$\begin{aligned}&\mathbb{E}\left[\hat{Q}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right]=\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}Q^{(k)}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right]\\ =&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[Q^{(k)}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right]=\mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right]\end{aligned}$$

where the last step follows from Lemma A.2.1 with $g: M \mapsto Q^{M}_{\tilde{\mu},t}(s,a)$ which is $\sigma(\boldsymbol{D}_T)-$ measurable.

To establish the rate, we have that $R^M(s,a) \in [0,1]\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}$, $t=1,\ldots,\tau$ thus $Q^{(k)}_t(s,a) \leq \tau$. By definition $\hat{Q}_t(s,a) - \mathbb{E}\left[Q^{M^*}_{\mu,t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right] = O_p\left(K^{-\frac{1}{2}}\right)$ if and only if for any $\epsilon>0$, $\exists M_\epsilon>0$ such that

$$\mathbb{P}\left(\hat{Q}_{\tilde{\mu},t}(s,a) - \mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right] > K^{-\frac{1}{2}}M_\epsilon\middle|s,a,t,\boldsymbol{D}_T\right)\leq\epsilon\ \ \forall K.$$

Note that for any $M>0$,

$$\begin{aligned}&\mathbb{P}\left(\hat{Q}_{\tilde{\mu},t}(s,a) - \mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)\middle|t,s,a,\boldsymbol{D}_T\right] > K^{-\frac{1}{2}}M\middle|t,s,a,\boldsymbol{D}_T\right)\\ =&\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}Q^{(k)}_{\tilde{\mu},t}(s,a) - \mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right] > K^{-\frac{1}{2}}M\middle|s,a,t,\boldsymbol{D}_T\right)\\ \leq&\exp\left\{-\frac{2M^2K^{-1}K^2}{K\tau^2}\right\}=\exp\left\{-\frac{2M^2}{\tau^2}\right\},\end{aligned}$$

which follows from Hoeffding's inequality as conditional on $s,a,t,\tilde{\mu}$ and $\boldsymbol{D}_T$, $\left\{Q^{(k)}_{\tilde{\mu},t}(s,a)\right\}_{k=1}^{K}$ are *iid* with mean $\mathbb{E}\left[Q^{M^*}_{\tilde{\mu},t}(s,a)\middle|s,a,t,\boldsymbol{D}_T\right]$. The result follows from choosing $M_\epsilon>0$ large enough such that $\exp\left\{-\frac{2M_\epsilon^2}{\tau^2}\right\}<\epsilon$.

□

Proof of Lemma 3.2. To simplify notation, let $Z^{(k)} \equiv I\left(Q^{(k)}_{\mu^\alpha_k,t}(s,\mu_k(s,t)) - Q^{(k)}_{\mu^\alpha_k,t}(s,\pi(s,t)) \leq 0\right)$,

then by definition $Z^{(k)} - \mathbb{E}\left[Z^{(k)}\right] = O_p\left(K^{-\frac{1}{2}}\right)$ if and only if for any $\epsilon > 0$, $\exists M_\epsilon > 0$ such that

$$\mathbb{P}\left(Z^{(k)} - \mathbb{E}\left[Z^{(k)}\right] > K^{-\frac{1}{2}} M_\epsilon \middle| t, s, \boldsymbol{D}_T\right) \leq \epsilon \ \ \forall K.$$

Note that for any $M > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\hat{\mathbb{P}}(H_0|t,s,\boldsymbol{D}_T) - \mathbb{E}\left[Z^{(k)}|t,s,\boldsymbol{D}_T\right] > K^{-\frac{1}{2}} M|t,s,\boldsymbol{D}_T\right) \\
=&\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} Z^{(k)} - \mathbb{E}\left[Z^{(k)}|t,s,\boldsymbol{D}_T\right] > MK^{-\frac{1}{2}} \middle| t,s,\boldsymbol{D}_T\right) \\
\leq &\exp\left\{-\frac{2M^2K^{-1}K^2}{K\tau^2}\right\} = \exp\left\{-\frac{2M^2}{\tau^2}\right\},
\end{aligned}$$

where the inequality follows from Hoeffding's inequality as $\left\{Z^{(k)}\right\}_{k=1}^{K}$ are iid with mean $\mathbb{E}\left[Z^{(k)} \middle| t, s, \boldsymbol{D}_T\right]$, since $\mathcal{I}_1$, $\mathcal{I}_2$ in Algorithm 1 are disjoint. We can choose $M_\epsilon > 0$ large enough such that $\exp\left\{-\frac{2M^2}{\tau^2}\right\} < \epsilon$. Next note that as π is fixed, by Lemma A.2.1, with $g : M \mapsto I\left(Q^M_{\mu^\alpha,t}(s,\mu(s,t)) - Q^M_{\mu^\alpha,t}(s,\pi(s,t)) \leq 0\right)$ for any $M_k \sim f(\cdot|\boldsymbol{D}_T)$

$$\begin{aligned}
&\mathbb{E}\left[I\left(Q^{(k)}_{\mu^\alpha_k,t}(s,\mu_k(s,t)) - Q^{(k)}_{\mu^\alpha_k,t}(s,\pi(s,t)) \leq 0\right) \middle| t,s,\boldsymbol{D}_T\right] \\
=&\mathbb{E}\left[I\left(Q^{M^*}_{\mu^{\alpha*},t}(s,\mu^*(s,t)) - Q^{M^*}_{\mu^{\alpha*},t}(s,\pi(s,t)) \leq 0\right) \middle| t,s,\boldsymbol{D}_T\right] \\
=&\mathbb{P}(H_0|t,s,\boldsymbol{D}_T)
\end{aligned}$$

which follows from using disjoint sets $\mathcal{I}_1, \mathcal{I}_2$ in Algorithm 1. Substituting this in the probability statement gives us

$$\hat{\mathbb{P}}\left(H_0|t,s,\boldsymbol{D}_T\right) - \mathbb{P}\left(H_0|t,s,\boldsymbol{D}_T\right) = O_p\left(K^{-\frac{1}{2}}\right),$$

which is our required result. □

Proof of Theorem 4.1. We start by showing $\hat{V}_{\tilde{\mu}}$ is unbiased:

$$\mathbb{E}\left[\hat{V}_{\tilde{\mu}}(s)|\boldsymbol{D}_T,\tilde{\mu}\right] = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[V^{(k)}_{\tilde{\mu},1}(s)\middle|\boldsymbol{D}_T\right].$$

where the first step follows from definition, and the $M_k \sim f(\cdot|\boldsymbol{D}_T)$ being iid, now by Lemma

A.2.1 with $g : M \mapsto V_{\mu,1}^{M}$ we have

$$\mathbb{E}\left[\hat{V}_{\tilde{\mu}}|\boldsymbol{D}_T\right] = \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right].$$

To establish the rate, we have that $V_{\tilde{\mu},1}^{(k)} \leq \tau$ as all rewards are between $[0,1]$ by definition $\hat{V}_{\tilde{\mu}} - \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right] = O_p\left(K^{-\frac{1}{2}}\right)$ if and only if for any $\epsilon > 0$, $\exists M_\epsilon > 0$ such that

$$\mathbb{P}\left(\hat{V}_{\tilde{\mu}} - \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right] > K^{-\frac{1}{2}}M_\epsilon\right) \leq \epsilon \ \ \forall K.$$

Note that for any $M > 0$,

$$\begin{aligned}\mathbb{P}\left(\hat{V}_{\tilde{\mu}} - \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right] > K^{-\frac{1}{2}}M\right) =& \mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} V_{\tilde{\mu},1}^{(k)} - \mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right] > K^{-\frac{1}{2}}M\right) \\ \leq& \exp\left\{-\frac{2M^2K^{-1}K^2}{K\tau^2}\right\} = \exp\left\{-\frac{2M^2}{\tau^2}\right\},\end{aligned}$$

where the inequality follows from Hoeffding's inequality as $\left\{V_{\tilde{\mu},1}^{(k)}\right\}_{k=1}^{K}$ are *iid* with mean $\mathbb{E}\left[V_{\tilde{\mu},1}^{M^*}(s)|\boldsymbol{D}_T\right]$. The result follows from choosing $M_\epsilon > 0$ large enough such that $\exp\left\{-\frac{2M^2}{\tau^2}\right\} < \epsilon$. □

A.4 Proofs for Supplementary results

Proof of Lemma A.1.1. First note that conditional on $\boldsymbol{D}_T$ with $g : M \mapsto I\left(V_{\mu_1}^{M}(s) - V_{\mu_2}^{M}(s) > 0\right)$, by Lemma A.2.1

$$\mathbb{E}\left[I\left(V_{\mu_1}^{M_k}(s) - V_{\mu_2}^{M_k}(s) > 0\right)\middle|\boldsymbol{D}_T\right] = \mathbb{E}\left[I\left(V_{\mu_1}^{M^*}(s) - V_{\mu_2}^{M^*}(s) > 0\right)\middle|\boldsymbol{D}_T\right] = \mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right)$$

By definition $\hat{\mathbb{P}}_\mu(H_0|\boldsymbol{D}_T) - \mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right) = O_p\left(K^{-\frac{1}{2}}\right)$ if and only if for any $\epsilon > 0$, $\exists M_\epsilon > 0$ such that

$$\mathbb{P}\left(\hat{\mathbb{P}}_\mu(H_0|\boldsymbol{D}_T) - \mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right) > K^{-\frac{1}{2}}M_\epsilon\middle|\boldsymbol{D}_T\right) \leq \epsilon \ \ \forall K.$$

Now, for any $M > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\hat{\mathbb{P}}_\mu(H_0|\boldsymbol{D}_T) - \mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right) > K^{-\frac{1}{2}} M_\epsilon \middle| \boldsymbol{D}_T\right) \\
=&\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} I\left(V^{(k)}_{\mu_1,1} - V^{(k)}_{\mu_2,1} > 0\right) - \mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right) > MK^{-\frac{1}{2}} \middle| \boldsymbol{D}_T\right) \\
\leq& \exp\left\{-\frac{2M^2K^{-1}K^2}{K\tau^2}\right\} = \exp\left\{-\frac{2M^2}{\tau^2}\right\},
\end{aligned}$$

where the inequality follows from Hoeffding's inequality as the indicators $\left\{I\left(V^{(k)}_{\mu_1,1} - V^{(k)}_{\mu_2,1} > 0\right)\right\}_{k=1}^{K}$ are *iid* with mean $\mathbb{P}_\mu\left(H_0|\boldsymbol{D}_T\right)$. We can choose $M_\epsilon > 0$ large enough such that $\exp\left\{-\frac{2M^2}{\tau^2}\right\} < \epsilon$. □

Proof of Lemma A.3.1. We first write the estimated regret as a sum of difference in value functions and a Bellman error.

I) We'll denote the sequence of states for an episode as $s_1, s_2, \ldots, s_\tau$, define

$$\begin{aligned}
\mathcal{W}_j &= \left(\mathcal{T}^{M_k}_{\mu_k^{\alpha*}(\cdot,j)} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,j)}\right) V^{M_k}_{\mu_k^{\alpha*},j+1}\left(s_{j+1}\right) \\
\mathbb{T}_j &= \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,j)}\left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1}\right)\left(s_{j+1}\right)
\end{aligned}$$

using (A.2) we can write

$$\begin{aligned}
\left(V^{M_k}_{\mu_k^{\alpha*},1} - V^{M^*}_{\mu_k^{\alpha*},1}\right)(s_1) &= \left(\mathcal{T}^{M_k}_{\mu_k^{\alpha*}(\cdot,1)} V^{M_k}_{\mu_k^{\alpha*},2} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,1)} V^{M^*}_{\mu_k^{\alpha*},2}\right)(s_2) \\
&= \left(\mathcal{T}^{M_k}_{\mu_k^{\alpha*}(\cdot,1)} V^{M_k}_{\mu_k^{\alpha*},2} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,1)} V^{M_k}_{\mu_k^{\alpha*},2} + \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,1)} V^{M_k}_{\mu_k^{\alpha*},2} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,1)} V^{M^*}_{\mu_k^{\alpha*},2}\right)(s_2) \\
&= \mathcal{W}_1 + \mathbb{T}_1,
\end{aligned}$$

with the same steps we can generalize this to

$$\left(V^{M_k}_{\mu_k^{\alpha*},j} - V^{M^*}_{\mu_k^{\alpha*},j}\right)(s_j) = \mathcal{W}_j + \mathbb{T}_j. \tag{A.4}$$

Next let

$$e_j = \left(I\left(\mathbb{P}^*(H_0) < \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\mu_k(s,j)}(s'|s) + I\left(\mathbb{P}^*(H_0) \geq \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\pi(s,j)}(s'|s) \right)$$
$$\times \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1} \right)(s') - \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1} \right)(s_{j+1}),$$

using the Bellman operator we get

$$\mathbb{T}_j = \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1} \right)(s_{j+1}) + e_j,$$

then we can write $\mathbb{T}_1 = \left(V^{M_k}_{\mu_k^{\alpha*},2} - V^{M^*}_{\mu_k^{\alpha*},2} \right)(s_2) + e_1$, with the above definitions and repeated use of (A.4):

$$\begin{aligned}
\left(V^{M_k}_{\mu_k^{\alpha*},1} - V^{M^*}_{\mu_k^{\alpha*},1} \right)(s_1) &= \mathcal{W}_1 + \mathbb{T}_1 \\
&= \mathcal{W}_1 + \left(V^{M_k}_{\mu_k^{\alpha*},2} - V^{M^*}_{\mu_k^{\alpha*},2} \right)(s_2) + e_1 \\
&= \mathcal{W}_1 + \mathcal{W}_2 + \left(V^{M_k}_{\mu_k^{\alpha*},3} - V^{M^*}_{\mu_k^{\alpha*},3} \right)(s_3) + e_1 + e_2 \\
&\vdots \\
&= \sum_{j=1}^{\tau} \mathcal{W}_j + e_j.
\end{aligned}$$

II) Next we consider $\mathbb{E}\left[e_j | M_k, M^*\right]$:

$$
\begin{aligned}
&\mathbb{E}\left[e_j \middle| M_k, M^*\right] \\
&= \mathbb{E}\left[I\left(\mathbb{P}^*(H_0) < \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\mu_k(s,j)}(s'|s) \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1}\right)(s') \middle| M_k, M^*\right] \\
&+ \mathbb{E}\left[I\left(\mathbb{P}^*(H_0) \geq \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\pi(s,j)}(s'|s) \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1}\right)(s') \middle| M_k, M^*\right] \\
&- \mathbb{E}\left[\left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1}\right)(s_{j+1}) \middle| M_k, M^*\right] \\
&= \left(I\left(\mathbb{P}^*(H_0) < \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\mu_k(s,j)}(s'|s) + I\left(\mathbb{P}^*(H_0) \geq \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\pi(s,j)}(s'|s)\right) \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1}\right)(s') \\
&- \left(I\left(\mathbb{P}^*(H_0) < \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\mu_k(s,j)}(s'|s) + I\left(\mathbb{P}^*(H_0) \geq \alpha\right) \sum_{s' \in \mathcal{S}} P^{M^*}_{\pi(s,j)}(s'|s)\right) \left(V^{M_k}_{\mu_k^{\alpha*},j+1} - V^{M^*}_{\mu_k^{\alpha*},j+1}\right)(s') \\
&= 0,
\end{aligned}
$$

which follows by the expectation conditional on M_k, M^* and definition of policy $\mu_k^{\alpha*}$.

Putting I) and II) together we get

$$
\begin{aligned}
\mathbb{E}\left[\left(V^{M_k}_{\mu_k^{\alpha*},1} - V^{M^*}_{\mu_k^{\alpha*},1}\right)(s_1) \middle| M^*, M_k\right] &= \mathbb{E}\left[\sum_{j=1}^{\tau} \mathcal{W}_j + e_j \middle| M^*, M_k\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{\tau} \left(\mathcal{T}^{M_k}_{\mu_k^{\alpha*}(\cdot,j)} - \mathcal{T}^{M^*}_{\mu_k^{\alpha*}(\cdot,j)}\right) V^{M_k}_{\mu_k^{\alpha*},j+1}(s_j) \middle| M^*, M_k\right]
\end{aligned}
$$

□

Proof of Lemma A.3.2. First consider Azuma-Hoeffding's Inequality: Let $Z_1, Z_2, \ldots$ be a martingale sequence difference with $|Z_j| \leq c \ \forall j$. Then $\forall \epsilon > 0$ and $n \in \mathbb{N}$ $P\left[\sum_{i=1}^{n} Z_i > \epsilon\right] \leq \exp\left\{-\frac{\epsilon^2}{2nc^2}\right\}$.

By definition the difference between the estimated transition and reward functions and their

true respective functions are:

$$\hat{P}_a(s'|s) - P_a^M(s'|s) = \frac{\sum_{i,j\in\mathcal{I}} \left(I(s_{i,j+1} = s') - P_a^M(s'|s)\right) I(s_{ij} = s, a_{ij} = a)}{N_T(s,a)},$$
$$\hat{R}(s,a) - R^M(s,a) = \frac{\sum_{i,j\in\mathcal{I}} (r_{ij} - R^M(s,a)) I(s_{ij} = s, a_{ij} = a)}{N_T(s,a)},$$

now let $\tilde{\beta}_T(s,a) \equiv \sqrt{8ST\log(2TSA)}$, and consider the transition probability function, for a fixed state action pair (s,a), let $\boldsymbol{\xi} = (\xi(s_1), \ldots, \xi(s_S)) \in \{-1,1\}^S$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{s'\in\mathcal{S}} \left| \frac{\sum_{i,j\in\mathcal{I}} \left(I(s_{i,j+1} = s') - P_a^M(s'|s)\right) I(s_{ij} = s, a_{ij} = a)}{N_T(s,a)} \right| \geq \frac{\tilde{\beta}_T(s,a)}{N_T(s,a)} \right) \\
\leq &\mathbb{P}\left(\max_{\boldsymbol{\xi}\in\{-1,1\}^s} \sum_{s'\in\mathcal{S}} \xi(s') \sum_{i,j\in\mathcal{I}} \left(I(s_{i,j+1} = s') - P_a^M(s'|s)\right) I(s_{ij} = s, a_{ij} = a) \geq \tilde{\beta}_T(s,a) \right) \\
\leq &2^S \mathbb{P}\left(\sum_{s'\in\mathcal{S}} \sum_{i,j\in\mathcal{I}} \xi(s') \Big(I(s_{i,j+1} = s') - P_a^M(s'|s) \Big) I(s_{ij} = s, a_{ij} = a) \geq \tilde{\beta}_T(s,a) \right)
\end{aligned}$$

where the first step follows from multiplying by $N_T(s,a)$, and eliminating the absolute value with $\boldsymbol{\xi}$, we use a union bound for the second step as there are 2^S possible $\boldsymbol{\xi}$ for a fixed (s,a) pair. Next we use Azuma-Hoeffding's inequality to bound the 2^S probability terms, note that within the probability function we are summing over T terms:

$$\begin{aligned}
&2^S \mathbb{P}\left(\sum_{s'\in\mathcal{S}} \sum_{i,j\in\mathcal{I}} \xi(s') \Big(I(s_{i,j+1} = s') - P_a^M(s'|s) \Big) I(s_{ij} = s, a_{ij} = a) \geq \tilde{\beta}_T(s,a) \right) \\
&\leq 2^S \exp\left\{ -\frac{8ST\log(2TSA)}{2\times 2^2 T} \right\} \\
&\leq 2^S \exp\left\{ \log\left((2TSA)^{-S}\right) \right\} = 2^S \frac{1}{(2TSA)^S} < \frac{1}{TSA},
\end{aligned}$$

next we sum over all (s,a) pairs and get

$$\begin{aligned}
&\mathbb{P}\left(\left\| \hat{P}_a(s'|s) - P_a^M(s'|s) \right\|_1 \geq \beta_T(s,a) \right) \\
&\leq \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mathbb{P}\left(\sum_{s'\in\mathcal{S}} \left| \frac{\sum_{i,j\in\mathcal{I}} \left(I(s_{i,j+1} = s') - P_a^M(s'|s)\right) I(s_{ij} = s, a_{ij} = a)}{N_T(s,a)} \right| \geq \frac{\tilde{\beta}_T(s,a)}{N_T(s,a)} \right) \\
&\leq SA \frac{1}{TSA} = \frac{1}{T},
\end{aligned}$$

which follows from using a union bound again. Analogous we can show that $\mathbb{P}\left(\left|\hat{R}(s,a) - R^M(s,a)\right| \geq \beta_T(s,a)\right) \leq \frac{1}{T}$, thus

$$\mathbb{P}\left(M^* \notin \mathcal{M}_T\right), \mathbb{P}\left(M_T \notin \mathcal{M}_T\right) < \frac{1}{T}.$$

□

Appendix B

Appendix to Chapter 2

B.1 Simulation Results for Alternative Settings

In this Section we provide additional results for data generating scenarios described in Section 2.7. Tables B.1 and B.1 contain results for estimation of Q function parameters for the EHR simulation setting for small and large sample sizes respectively. Table B.3 contains the complete parameter results for the continuous data generating setting for both small and large samples.

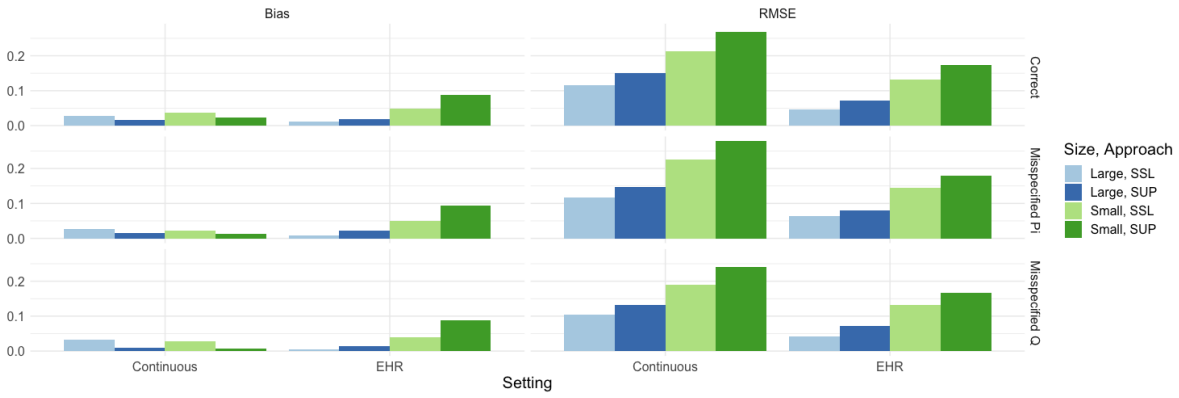


Figure B.1: *Monte Carlo estimates for doubly-robust value function estimation:* $\widehat{V}_{\text{SSL}_{\text{DR}}}$, $\widehat{V}_{\text{SUP}_{\text{DR}}}$ *under continuous, and EHR settings. Columns show bias and RMSE respectively, rows show different misspecification scenarios. Results are shown for the large* ($N = 10,000$, $n = 500$) *and small data samples* ($N = 1,272$, $n = 135$) *for the continuous setting over 1,000 simulated datasets.*

B.2 Proof of Main Results

B.2.1 Semi-supervised Q-learning asymptotics

In this section we first show the proofs for the theoretical results on the generalized semi-supervised Q-learning shown in section 2.6.

Proofs for theoretical results for Q-learning in section 2.6

We first define $\boldsymbol{\theta}_{2-} \equiv (\boldsymbol{\beta}_{22}^{\intercal}, \boldsymbol{\gamma}_2^{\intercal})^{\intercal}$, and $\hat{\Delta}_s^{(\text{-}k)}(\vec{\mathbf{U}}) \equiv \widehat{m}_s^{(\text{-}k)}(\vec{\mathbf{U}}) - m_s(\vec{\mathbf{U}})$, $s \in \{2, 3, 22, 23\}$, and note that from Assumptions 2.6.1, 2.6.2 & 2.6.3 it follows that:

$$\begin{aligned}
&\sum_{k=1}^{K} \sup_{\vec{\mathbf{U}}} \left| \hat{\Delta}_{2t}^{(\text{-}k)}(\vec{\mathbf{U}}) \right| = o_{\mathbb{P}}(1) \text{ for } t = 2, 3, \\
&\sum_{k=1}^{K} \sup_{\vec{\bar{X}}, \vec{\mathbf{U}}} \| \vec{\bar{X}} \hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}) \| = o_{\mathbb{P}}(1), \\
&\sum_{k=1}^{K} \sup_{, \bar{X}_2 \vec{\mathbf{U}}} \| \bar{X}_2 \hat{\Delta}_3^{(\text{-}k)}(\vec{\mathbf{U}}) \| = o_{\mathbb{P}}(1),
\end{aligned} \tag{B.1}$$

Next we remind that, to ensure the validity of the SSL algorithm from the refitted imputation model, the final imputation models for $\{Y_t, Y_{2t}, t = 2, 3\}$, denoted by $\{\bar{\mu}_t(\vec{\mathbf{U}}), \bar{\mu}_{2t}, t = 2, 3\}$, need to satisfy the constraints shown in Section 2.4.2:

$$\begin{aligned}
\mathbb{E}\left[\vec{\bar{X}} \{ Y_2 - \bar{\mu}_2(\vec{\mathbf{U}}) \} \right] = \mathbf{0}, \quad \mathbb{E}\left\{ Y_2^2 - \bar{\mu}_{22}(\vec{\mathbf{U}}) \right\} = 0, \\
\mathbb{E}\left[\bar{X}_2 \{ Y_3 - \bar{\mu}_3(\vec{\mathbf{U}}) \} \right] = \mathbf{0}, \quad \mathbb{E}\left\{ Y_2 Y_3 - \bar{\mu}_{23}(\vec{\mathbf{U}}) \right\} = 0.
\end{aligned} \tag{B.2}$$

where $\vec{\bar{X}} = (1, \bar{X}_1^{\intercal}, \bar{X}_2^{\intercal})^{\intercal}$.

Proof of Theorem 2.6.4. Recall the estimating equation for stage 2 regression in Section 2.4.2 is

$$\mathbb{P}_N \begin{bmatrix} \widehat{\mu}_{23}(\vec{\mathbf{U}}) - \hat{\beta}_{21} \widehat{\mu}_{22}(\vec{\mathbf{U}}) - \widehat{\mu}_2(\vec{\mathbf{U}}) \bar{X}_2^{\intercal} \widehat{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2 \left\{ \widehat{\mu}_3(\vec{\mathbf{U}}) - \hat{\beta}_{21} \widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{X}_2^{\intercal} \widehat{\boldsymbol{\theta}}_{2-} \right\} \end{bmatrix} = \mathbf{0}.$$

Centering the above at $\bar{\boldsymbol{\theta}}_2$ we get

$$\mathbb{P}_N \begin{bmatrix} \widehat{\mu}_{22}(\vec{\mathbf{U}}), \widehat{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^\intercal \\ \bar{X}_2\widehat{\mu}_2(\vec{\mathbf{U}}), \bar{X}_2\bar{X}_2^\intercal \end{bmatrix} (\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2) = \mathbb{P}_N \begin{bmatrix} \widehat{\mu}_{23}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\widehat{\mu}_{22}(\vec{\mathbf{U}}) - \widehat{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2 \left\{ \widehat{\mu}_3(\vec{\mathbf{U}}) - \bar{\beta}_{21}\widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \right\} \end{bmatrix}. \tag{B.3}$$

Define

$$\begin{aligned}
\mathcal{R}_{\mathcal{U}} =& \mathbb{P}_N \begin{bmatrix} \bar{\mu}_{23}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\bar{\mu}_{22}(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2 \left\{ \bar{\mu}_3(\vec{\mathbf{U}}) - \bar{\beta}_{21}\bar{\mu}_2(\vec{\mathbf{U}}) - \bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \right\} \end{bmatrix}, \\
\hat{\mathcal{R}}_{\mathbb{S}}^{(K)} =& \mathbb{P}_N \begin{bmatrix} \left\{ \widehat{\mu}_{23}(\vec{\mathbf{U}}) - \bar{\mu}_{23}(\vec{\mathbf{U}}) \right\} - \bar{\beta}_{21} \left\{ \widehat{\mu}_{22}(\vec{\mathbf{U}}) - \bar{\mu}_{22}(\vec{\mathbf{U}}) \right\} - \left\{ \widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}}) \right\} \bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2 \left\{ \widehat{\mu}_3(\vec{\mathbf{U}}) - \bar{\mu}_3(\vec{\mathbf{U}}) \right\} - \bar{\beta}_{21}\bar{X}_2 \left\{ \widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}}) \right\} \end{bmatrix}, \\
\Gamma_{\mathcal{U}} =& \mathbb{P}_N \begin{bmatrix} \bar{\mu}_{22}(\vec{\mathbf{U}}) & \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^\intercal \\ \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2 & \bar{X}_2\bar{X}_2^\intercal \end{bmatrix}, \\
\hat{\Gamma}_{\mathbb{S}}^{(K)} =& \mathbb{P}_N \begin{bmatrix} \widehat{\mu}_{22}(\vec{\mathbf{U}}) - \bar{\mu}_{22}(\vec{\mathbf{U}}) & \left\{ \widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}}) \right\} \bar{X}_2^\intercal \\ \left\{ \widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}}) \right\} \bar{X}_2 & \mathbf{0} \end{bmatrix},
\end{aligned}$$

with these we can re-write equation (B.3) as $\left(\Gamma_{\mathcal{U}} + \hat{\Gamma}_{\mathbb{S}}^{(K)}\right)(\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2) = \mathcal{R}_{\mathcal{U}} + \hat{\mathcal{R}}_{\mathbb{S}}^{(K)}$. We next deal with each term.

(I) We first consider $\hat{\mathcal{R}}_{\mathbb{S}}^{(K)}$, let

$$\begin{aligned}
\hat{\mathcal{S}}_{\mathbb{S}}^{\eta} =& \mathbb{P}_N \begin{bmatrix} (\widehat{\eta}_{23} - \eta_{23}) - \bar{\beta}_{21}(\widehat{\eta}_{22} - \eta_{22}) - (\widehat{\boldsymbol{\eta}}_2 - \boldsymbol{\eta}_2)^\intercal\bar{X}_2\bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2\bar{X}_2^\intercal \left\{ (\widehat{\boldsymbol{\eta}}_3 - \boldsymbol{\eta}_3) - \bar{\beta}_{21} (\widehat{\boldsymbol{\eta}}_2 - \boldsymbol{\eta}_2) \right\} \end{bmatrix} \\
\hat{\mathcal{S}}_{\mathbb{S}}^{(K)} =& \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}_N \begin{bmatrix} \hat{\Delta}_{23}^{(\text{-}k)}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\hat{\Delta}_{22}^{(\text{-}k)}(\vec{\mathbf{U}}) - \hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2 \left\{ \hat{\Delta}_3^{(\text{-}k)}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}) \right\} \end{bmatrix} \\
\bar{\mathcal{S}}_k =& \mathbb{E}_{\mathcal{L}} \begin{bmatrix} \hat{\Delta}_{23}^{(\text{-}k)}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\hat{\Delta}_{22}^{(\text{-}k)}(\vec{\mathbf{U}}) - \hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\bar{X}_2^\intercal\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2 \left\{ \hat{\Delta}_3^{(\text{-}k)}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}) \right\} \end{bmatrix} \quad \text{for } k \in \{1, \ldots, K\}.
\end{aligned}$$

From (2.3) it follows that $\hat{\mathcal{R}}_{\mathbb{S}}^{(K)} = \hat{\mathcal{S}}_{\mathbb{S}}^{\eta} + \hat{\mathcal{S}}_{\mathbb{S}}^{(K)}$. Next using (B.1), Assumption 2.6.2, and Lemma B.3.2 it follows that $\hat{\mathcal{S}}_{\mathbb{S}}^{(K)} = \frac{1}{K}\sum_k \bar{\mathcal{S}}_k + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)$, which lets us write $\hat{\mathcal{R}}_{\mathbb{S}}^{(K)} = \hat{\mathcal{S}}_{\mathbb{S}}^{\eta} + \frac{1}{K}\sum_k \bar{\mathcal{S}}_k + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)$.

Now consider $\hat{\mathcal{S}}^{\eta}_{\mathbb{S}}$, note that by the central limit theorem (CLT) $\mathbb{P}_n\bar{X}_2\bar{X}_2 = \mathbb{E}\bar{X}_2\bar{X}_2 + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. Thus using this, Slutsky's theorem and Assumption 2.6.1

$$(\mathbb{P}_n\bar{X}_2\bar{X}_2)^{-1}(\mathbb{P}_N\bar{X}_2\bar{X}_2) = I + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),$$

then using (B.2), (2.3) and Assumption 2.6.2 we can write

$$\begin{aligned}
&\mathbb{P}_N\left\{(\widehat{\boldsymbol{\eta}}_2 - \boldsymbol{\eta}_2)^{\intercal}\bar{X}_2\bar{X}_2^{\intercal}\bar{\boldsymbol{\theta}}_{2-}\right\} \\
=&\left[\left(\mathbb{P}_n\bar{X}_2\bar{X}_2^{\intercal}\right)^{-1}\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{2i}\left\{Y_{2i} - \bar{\mu}_2(\vec{\mathbf{U}}_i) + m_2(\vec{\mathbf{U}}_i) - m_2^{(-k)}(\vec{\mathbf{U}}_i)\right\}\right]^{\intercal}\mathbb{P}_N(\bar{X}_2\bar{X}_2^{\intercal})\boldsymbol{\theta}_{2-} \\
=&\left[\mathbb{P}_n\bar{X}_2^{\intercal}\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{2i}^{\intercal}\hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}}_i)\right]\left(\mathbb{P}_n\bar{X}_2\bar{X}_2^{\intercal}\right)^{-1}\mathbb{P}_N(\bar{X}_2\bar{X}_2^{\intercal})\boldsymbol{\theta}_{2-} \\
=&\mathbb{P}_n\bar{X}_2\boldsymbol{\theta}_{2-}^{\intercal}\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{2i}^{\intercal}\boldsymbol{\theta}_{2-}\hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}}_i) \\
&+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\left[\mathbb{P}_n\bar{X}_2\boldsymbol{\theta}_{2-}^{\intercal}\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{2i}^{\intercal}\boldsymbol{\theta}_{2-}\hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}}_i)\right] \\
=&\mathbb{P}_n\bar{X}_2\boldsymbol{\theta}_{2-}^{\intercal}\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} + \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{2i}^{\intercal}\boldsymbol{\theta}_{2-}\hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}}_i) + O_{\mathbb{P}}\left(n^{-1}\right) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)o_{\mathbb{P}}(1).
\end{aligned}$$

Analogous derivations for all terms in $\hat{\mathcal{S}}^{\eta}_{\mathbb{S}}$ gives us

$$\hat{\mathcal{S}}^{\eta}_{\mathbb{S}} = \mathbb{T}_{\mathcal{L}} - \mathbb{T}^{(K)}_{\mathcal{L}} + O_{\mathbb{P}}\left(n^{-1}\right) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)o_{\mathbb{P}}(1),$$

where

$$\begin{aligned}
\mathbb{T}_{\mathcal{L}} =&\mathbb{P}_n\begin{bmatrix}\left\{Y_2Y_3 - \bar{\mu}_{23}(\vec{\mathbf{U}})\right\} - \bar{\beta}_{21}\left\{Y_2^2 - \bar{\mu}_{22}(\vec{\mathbf{U}})\right\} - \left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\}\bar{X}_2^{\intercal}\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2\left\{Y_3 - \bar{\mu}_3(\vec{\mathbf{U}})\right\} - \bar{\beta}_{21}\bar{X}_2\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\}\end{bmatrix}, \\
\mathbb{T}^{(K)}_{\mathcal{L}} =&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\begin{bmatrix}\hat{\Delta}_{23}^{(-k)}(\vec{\mathbf{U}}_i) - \bar{\beta}_{21}\hat{\Delta}_{22}^{(-k)}(\vec{\mathbf{U}}_i) - \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}}_i)\bar{X}_{2i}^{\intercal}\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_{2i}\left\{\hat{\Delta}_3^{(-k)}(\vec{\mathbf{U}}_i) - \bar{\beta}_{21}\hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}}_i)\right\}\end{bmatrix}.
\end{aligned}$$

From the above it follows that $\hat{\mathcal{R}}^{(K)}_{\mathbb{S}} = \mathbb{T}_{\mathcal{L}} - \mathbb{T}^{(K)}_{\mathcal{L}} + \frac{1}{K}\sum_k \bar{\mathcal{S}}_k + O_{\mathbb{P}}\left(n^{-1}\right) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)o_{\mathbb{P}}(1)$. Next by Assumption 2.6.2 and using Lemma B.3.4 with $\hat{C}_{n,N} = 1$, and setting functions

$\hat{l}_n(\cdot)$, $\hat{\pi}_n(\cdot)$ to be the constant 1, and $f(\bar{X}_2) = \bar{X}_2$ to be the identity function, we have $\sqrt{n}\left(\mathbb{T}_{\mathcal{L}}^{(K)} - \frac{1}{K}\sum_k \bar{\mathcal{S}}_k\right) = O_{\mathbb{P}}\left(c_{n_K^-}\right)$. Therefore $\hat{\mathcal{R}}_{\mathbb{S}}^{(K)} = \mathbb{T}_{\mathcal{L}} + O_{\mathbb{P}}\left(n^{-\frac{1}{2}} c_{n_K^-}\right)$.

(II) Now we consider $\mathcal{R}_{\mathcal{U}}$, from the CLT, assuming working model (2.1), as constraints (B.2) are satisfied it follows that

$$\mathcal{R}_{\mathcal{U}} = \mathbb{E}\begin{bmatrix} \bar{\mu}_{23}(\vec{\mathbf{U}}) - \bar{\beta}_{21}\bar{\mu}_{22}(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^{\intercal}\bar{\boldsymbol{\theta}}_{2-} \\ \bar{X}_2\{\bar{\mu}_3(\vec{\mathbf{U}}) - \bar{\beta}_{21}\bar{\mu}_2(\vec{\mathbf{U}}) - \bar{X}_2^{\intercal}\bar{\boldsymbol{\theta}}_{2-}\} \end{bmatrix} + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right) = \mathbf{1}O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right).$$

(III) Next we focus on $\hat{\Gamma}_{\mathbb{S}}^{(K)}$, we use a similar expansion to (I) and define

$$\begin{aligned}
\hat{\mathcal{F}}_{\mathbb{S}}^{\eta} =& \begin{bmatrix} \widehat{\eta}_{22} - \eta_{22} & (\widehat{\eta}_2 - \eta_2)\bar{X}_2^{\intercal} \\ (\widehat{\eta}_2 - \eta_2)\,\bar{X}_2 & \mathbf{0} \end{bmatrix}, \\
\hat{\mathcal{F}}_{\mathbb{S}}^{(K)} =& \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}_N \begin{bmatrix} \hat{\Delta}_{22}^{(-k)}(\vec{\mathbf{U}}) & \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}})\bar{X}_2^{\intercal} \\ \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}})\bar{X}_2 & \mathbf{0} \end{bmatrix}, \\
\bar{\mathcal{F}}_k =& \mathbb{E}_{\mathcal{L}} \begin{bmatrix} \hat{\Delta}_{22}^{(-k)}(\vec{\mathbf{U}}) & \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}})\bar{X}_2^{\intercal} \\ \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}})\bar{X}_2 & \mathbf{0} \end{bmatrix} \quad \forall k \in \{1, \ldots, K\},
\end{aligned}$$

We argue as in (I), that from (2.3) it follows that $\hat{\Gamma}_{\mathbb{S}}^{(K)} = \hat{\mathcal{F}}_{\mathbb{S}}^{\eta} + \hat{\mathcal{F}}_{\mathbb{S}}^{(K)}$. Using (B.1), Assumptions 2.6.2 and Lemma B.3.2 $\hat{\mathcal{F}}_{\mathbb{S}}^{(K)} - \frac{1}{K}\sum_k \bar{\mathcal{F}}_k = O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)$, therefore $\hat{\Gamma}_{\mathbb{S}}^{(K)} = \hat{\mathcal{F}}_{\mathbb{S}}^{\eta} + \frac{1}{K}\sum_k \bar{\mathcal{F}}_k + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)$. Next we follow the same decomposition for $\hat{\mathcal{F}}_{\mathbb{S}}^{\eta}$ as we did in (I) for $\hat{\mathcal{S}}_{\mathbb{S}}^{\eta}$, it follows that

$$\begin{aligned}
\hat{\Gamma}_{\mathbb{S}}^{(K)} =& \mathbb{P}_n \begin{bmatrix} Y_2^2 - \bar{\mu}_{22}(\vec{\mathbf{U}}) & \left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\}\bar{X}_2^{\intercal} \\ \left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\}\bar{X}_2 & \mathbf{0} \end{bmatrix} \\
& - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k} \begin{bmatrix} \hat{\Delta}_{22}^{(-k)}(\vec{\mathbf{U}}) & \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}})\bar{X}_2^{\intercal} \\ \hat{\Delta}_2^{(-k)}(\vec{\mathbf{U}})\bar{X}_2 & \mathbf{0} \end{bmatrix} + \frac{1}{K}\sum_k \bar{\mathcal{F}}_k + O_{\mathbb{P}}\left(n^{-1}\right) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) o_{\mathbb{P}}(1).
\end{aligned}$$

The first term in the right hand side is $O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ by he CLT, the next two terms together are $O_{\mathbb{P}}\left(n^{-\frac{1}{2}} c_{n_K^-}\right)$ by Lemma B.3.4, thus $\hat{\Gamma}_{\mathbb{S}}^{(K)} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}} c_{n_K^-}\right)$.

(IV) Finally we consider $\Gamma_{\mathcal{U}}$. By central limit theorem and (B.2) it follows that

$$\Gamma_{\mathcal{U}} = \mathbb{E}\begin{bmatrix} \bar{\mu}_{22}(\vec{\mathbf{U}}) & \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^\intercal \\ \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2 & \bar{X}_2\bar{X}_2^\intercal \end{bmatrix} + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right) = \mathbb{E}[\breve{\bar{X}}_2\breve{\bar{X}}_2^\intercal] + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right).$$

From (I)-(IV) we can write (B.3) as $(\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2) = \mathbb{E}\left[\breve{\bar{X}}_2\breve{\bar{X}}_2^\intercal\right]^{-1}\mathbb{T}_{\mathcal{L}} + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}c_{n_K^-}\right)$, it follows that

$$\begin{aligned}
&\sqrt{n}(\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2) \\
=&\mathbb{E}\left[\breve{\bar{X}}_2\breve{\bar{X}}_2^\intercal\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} \left\{Y_{2i}Y_{3i} - \bar{\mu}_{23}(\vec{\mathbf{U}}_i)\right\} - \bar{\beta}_{21}\left\{Y_{2i}^2 - \bar{\mu}_{22}(\vec{\mathbf{U}}_i)\right\} - \breve{\bar{X}}_{2i}^\intercal\bar{\boldsymbol{\theta}}_{2-}\left\{Y_{2i} - \bar{\mu}_2(\vec{\mathbf{U}}_i)\right\} \\ \bar{X}_{2i}\left\{Y_{3i} - \bar{\mu}_3(\vec{\mathbf{U}}_i)\right\} - \bar{\beta}_{21}\bar{X}_{2i}\left\{Y_{2i} - \bar{\mu}_2(\vec{\mathbf{U}}_i)\right\} \end{bmatrix} \\
&+o_{\mathbb{P}}(1).
\end{aligned}$$

□

Proof of Theorem 2.6.6. The solution to stage 1 estimating equation $\boldsymbol{\theta}_1$ in Section 2.4.2 satisfies

$$\mathbb{P}_N\left[\bar{X}_1\left\{\widehat{\mu}_2(\vec{\mathbf{U}}) + \hat{\beta}_{21}\widehat{\mu}_2(\vec{\mathbf{U}}) + \mathbf{H}_{20}^\intercal\widehat{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+ - \bar{X}_1^\intercal\widehat{\boldsymbol{\theta}}_1\right\}\right] = \mathbf{0}.$$

We center the above at $\bar{\boldsymbol{\theta}}_1$ and get

$$\mathbb{P}_N\left[\bar{X}_1\bar{X}_1^\intercal\right]\left(\widehat{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_1\right) = \mathbb{P}_N\left[\bar{X}_1\left\{\bar{\mu}_2(\vec{\mathbf{U}}) + \hat{\beta}_{21}\bar{\mu}_2(\vec{\mathbf{U}}) + \mathbf{H}_{20}^\intercal\widehat{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+ - \bar{X}_1^\intercal\bar{\boldsymbol{\theta}}_1\right\}\right]. \tag{B.4}$$

Next, with the following definitions

$$\begin{aligned}
\hat{\Sigma}_{\mathcal{U}} =&\mathbb{P}_N\left[\bar{X}_1\bar{X}_1^\intercal\right], \quad \hat{\Sigma}_{\mathcal{L}} = \mathbb{P}_n\left[\bar{X}_1\bar{X}_1^\intercal\right], \\
\mathcal{R}^{(1)} =&\mathbb{P}_N\left[\bar{X}_1\left\{\bar{\mu}_2(\vec{\mathbf{U}}) + \hat{\beta}_{21}\bar{\mu}_2(\vec{\mathbf{U}}) + \mathbf{H}_{20}^\intercal\widehat{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+ - \bar{X}_1^\intercal\bar{\boldsymbol{\theta}}_1\right\}\right], \\
\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)} =&\mathbb{P}_N\left[\bar{X}_1\left\{\widehat{\mu}_2(\vec{\mathbf{U}}) - \bar{\mu}_2(\vec{\mathbf{U}})\right\}\right],
\end{aligned}$$

we can write (B.4) as $\hat{\Sigma}_{\mathcal{U}}(\widehat{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_1) = \mathcal{R}^{(1)} + (1 + \hat{\beta}_{21})\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)}$. We now analyze both terms $\mathcal{R}^{(1)}$, and $(1 + \hat{\beta}_{21})\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)}$.

I) First we consider $(1+\hat{\beta}_{21})\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)}$, define

$$\begin{aligned}
\hat{\mathcal{S}}_{\mathbb{S}}^{(1\eta)} =& \hat{\Sigma}_{\mathcal{U}}\left(\widehat{\boldsymbol{\eta}}_2 - \boldsymbol{\eta}_2\right), \\
\hat{\mathcal{S}}_{\mathbb{S}}^{(1\mathbb{K})} =& \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_N\left[\bar{X}_1\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\right], \\
\bar{\mathcal{S}}_k^{(1)} =& \mathbb{E}\left[\bar{X}_1\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\right],
\end{aligned}$$

from (2.3) it follows that $\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)} = \hat{\mathcal{S}}_{\mathbb{S}}^{(1\eta)} + \hat{\mathcal{S}}_{\mathbb{S}}^{(1\mathbb{K})}$, next from Assumptions 2.6.1, 2.6.2, we get $\sum_{k=1}^{K}\sup_{\bar{X}_1,\vec{\mathbf{U}}}\|\bar{X}_1\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\| = o_{\mathbb{P}}(1)$, thus by Lemma B.3.2 $\hat{\mathcal{S}}_{\mathbb{S}}^{(1\mathbb{K})} = \bar{\mathcal{S}}_k^{(1)} + \left(N^{-\frac{1}{2}}\right)$. Using (2.3) again, and recalling $\bar{\mu}_2(\vec{\mathbf{U}}) = m_2(\vec{\mathbf{U}}) + \bar{X}_1^{\intercal}\boldsymbol{\eta}_2$ we have

$$\begin{aligned}
\hat{\mathcal{S}}_{\mathbb{S}}^{(1\eta)} =& \hat{\Sigma}_{\mathcal{U}}\hat{\Sigma}_{\mathcal{L}}^{-1}\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{1i}\left\{Y_{2i} - \bar{\mu}_2(\vec{\mathbf{U}}_i) - \widehat{m}_2^{(\text{-}k)}(\vec{\mathbf{U}}_i) + m_2(\vec{\mathbf{U}}_i)\right\} \\
=& \frac{1}{n}\sum_{i=1}^{n}\bar{X}_{1i}\left\{Y_{2i} - \bar{\mu}_2(\vec{\mathbf{U}}_i)\right\} - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{X}_{1i}\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}_i) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),
\end{aligned}$$

where the last line follows by the CLT and Assumptions 2.6.1 and 2.6.2 as

$$\hat{\Sigma}_{\mathcal{U}}\hat{\Sigma}_{\mathcal{L}}^{-1} = I + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$$

Now using Lemma B.3.2 and Assumptions 2.6.1, 2.6.2 again, it follows that

$$\hat{\mathcal{S}}_{\mathbb{S}}^{(1\mathbb{K})} = \bar{\mathcal{S}}_k^{(1)} + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right),$$

combining the above we can write

$$\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)} = \mathbb{P}_n\bar{X}_1\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} - \frac{1}{n}\sum_{k=1}^{K}\left\{\sum_{i\in\mathcal{I}_k}\bar{X}_{1i}\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}_i) - \mathbb{E}\left[\bar{X}_1\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\right]\right\} + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

Next by Assumption 2.6.2 and Lemma B.3.4 we have

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\left\{\sum_{i\in\mathcal{I}_k}\bar{X}_{1i}\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}_i) - \mathbb{E}\left[\bar{X}_1\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\right]\right\} = O_{\mathbb{P}}\left(c_{n_K^-}\right),$$

therefore $\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)} = \mathbb{P}_n\bar{X}_1\left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}c_{n_K^-}\right)$. Finally using Theorem 2.6.4 we have

$\hat{\beta}_{21} - \bar{\beta}_{21} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, and by CLT $\mathbb{P}_n \bar{X}_1 \left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, thus we can write

$$(1+\hat{\beta}_{21})\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)} = (1+\bar{\beta}_{21})\mathbb{P}_n \bar{X}_1 \left\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\right\} + O_{\mathbb{P}}\left(n^{-\frac{1}{2}} c_{n_K^-}\right).$$

II) Next we consider $\mathcal{R}^{(1)}$ by writing

$$\begin{aligned}\mathcal{R}^{(1)} =& \mathbb{P}_N \left[\bar{X}_1 \left\{\bar{\mu}_2(\vec{\mathbf{U}}) + \bar{\beta}_{21}\bar{\mu}_2(\vec{\mathbf{U}}) + \mathbf{H}_{20}^{\intercal}\bar{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+ - \bar{X}_1^{\intercal}\bar{\boldsymbol{\theta}}_1\right\}\right] \\ &+ \mathbb{P}_N \left[\bar{X}_1 \left\{\bar{\mu}_2(\vec{\mathbf{U}})(\hat{\beta}_{21} - \bar{\beta}_{21}) + \mathbf{H}_{20}^{\intercal}(\widehat{\boldsymbol{\beta}}_{22} - \bar{\boldsymbol{\beta}}_{22}) + [\mathbf{H}_{21}^{\intercal}\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+\right\}\right],\end{aligned}$$

note that under (B.2) using model (2.1) the first term in the right hand side is mean zero, therefore from Assumption 2.6.1 and CLT

$$\mathbb{P}_N \left\{\bar{X}_1 \left(\bar{\mu}_2(\vec{\mathbf{U}}) + \bar{\boldsymbol{\beta}}_{21}\bar{\mu}_2(\vec{\mathbf{U}}) + \mathbf{H}_{20}^{\intercal}\bar{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+ - \bar{X}_1^{\intercal}\bar{\boldsymbol{\theta}}_1\right)\right\} = O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right).$$

Hence, we have

$$\begin{aligned}\sqrt{n}\mathcal{R}^{(1)} =& \sqrt{n}\mathbb{P}_N \left[\bar{X}_1 \left\{\bar{\mu}_2(\vec{\mathbf{U}})(\hat{\beta}_{21} - \bar{\beta}_{21}) + \mathbf{H}_{20}^{\intercal}(\widehat{\boldsymbol{\beta}}_{22} - \bar{\boldsymbol{\beta}}_{22}) + [\mathbf{H}_{21}^{\intercal}\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+\right\}\right] + O_{\mathbb{P}}\left(\sqrt{\frac{n}{N}}\right) \\ =& \mathbb{P}_N \left[\bar{X}_1 \left(\bar{\mu}_2(\vec{\mathbf{U}}), \mathbf{H}_{20}^{\intercal}\right)\right] \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_2 - \bar{\boldsymbol{\beta}}_2\right) + \sqrt{n}\mathbb{P}_N \left[\bar{X}_1 \left([\mathbf{H}_{21}^{\intercal}\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+\right)\right] + O_{\mathbb{P}}\left(\sqrt{\frac{n}{N}}\right) \\ =& \mathbb{E}\left[\bar{X}_1 \left(\bar{\mu}_2(\vec{\mathbf{U}}), \mathbf{H}_{20}^{\intercal}\right)\right] n^{-\frac{1}{2}} \sum_{i=1}^{n} \boldsymbol{\psi}_{2i\beta} + \sqrt{n}\mathbb{P}_N \left[\bar{X}_1 \left([\mathbf{H}_{21}^{\intercal}\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+\right)\right] + O_{\mathbb{P}}\left(\sqrt{\frac{n}{N}}\right),\end{aligned}$$

where the last inequality follows from the CLT, where $\boldsymbol{\psi}_{2i\beta}$ is the element corresponding to $\widehat{\boldsymbol{\beta}}_2$ of the influence function $\boldsymbol{\psi}_{2i}$ defined in Theorem 2.6.4.

Next by Theorem 2.6.4 we know that

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}}_2 - \bar{\boldsymbol{\gamma}}_2) = O_{\mathbb{P}}(1),$$

using Lemma B.3.5 (a) we have

$$\mathbb{P}\left[\sqrt{n}\mathbb{P}_N \left\{\bar{X}_1 \left([\mathbf{H}_{21}^{\intercal}\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+\right)\right\} = \mathbb{P}_N\left\{\bar{X}_1\mathbf{H}_{21}^{\intercal} I\left(\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2 > 0\right)\right\}\sqrt{n}\left(\widehat{\boldsymbol{\gamma}}_2 - \bar{\boldsymbol{\gamma}}_2\right)\right] \to 1.$$

Therefore, letting $\boldsymbol{\psi}_{2i\gamma}$ be the element corresponding to $\widehat{\boldsymbol{\gamma}}_2$ of the influence function $\boldsymbol{\psi}_{2i}$

defined in Theorem 2.6.4,

$$
\begin{aligned}
&\sqrt{n}\mathbb{P}_N\left\{\bar{X}_1\left([\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2]_+\right)\right\}\\
&=\mathbb{P}_N\left\{\bar{X}_1\mathbf{H}_{21}^\intercal I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0)I\left(\widehat{\boldsymbol{\gamma}}_2\in\mathcal{A}\right)\right\}\sqrt{n}\left(\widehat{\boldsymbol{\gamma}}_2-\bar{\boldsymbol{\gamma}}_2\right)\\
&+\sqrt{n}\mathbb{P}_N\left\{\bar{X}_1\left([\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+ - [\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2]_+\right)\right\}I_{\{\widehat{\boldsymbol{\gamma}}_2\notin\mathcal{A}\}}\\
&=\mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^\intercal|\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0,\widehat{\boldsymbol{\gamma}}_2\in\mathcal{A}\right]\mathbb{P}\left(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right)\mathbb{P}\left(\widehat{\boldsymbol{\gamma}}_2\in\mathcal{A}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{2\gamma_2 i}+O_{\mathbb{P}}\left(c_{n_K^-}\right)+o_{\mathbb{P}}(1)\\
&=\mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^\intercal|\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right]\mathbb{P}\left(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{2\gamma_2 i}+o_{\mathbb{P}}(1),
\end{aligned}
$$

combining all terms

$$
\begin{aligned}
\sqrt{n}\mathcal{R}^{(1)}=&\mathbb{E}\left[\bar{X}_1\left(\bar{\mu}_2(\vec{\mathbf{U}}),\mathbf{H}_{20}^\intercal\right)\right]n^{-\frac{1}{2}}\sum_{i=1}^{n}\boldsymbol{\psi}_{2i(\beta)}\\
&+\mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^\intercal|\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right]\mathbb{P}\left(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_{2i(\gamma)}\\
&+O_{\mathbb{P}}\left(c_{n_K^-}\right).
\end{aligned}
$$

Finally, from I), II), and since $\hat{\Sigma}_{\mathcal{U}}^{-1}=\mathbb{E}\left[\bar{X}_1\bar{X}_1^\intercal\right]^{-1}+o_{\mathbb{P}}(1)$ by the LLN, we have

$$
\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\theta}}_1-\bar{\boldsymbol{\theta}}_1)=&\mathbb{E}\left[\bar{X}_1\bar{X}_1^\intercal\right]^{-1}\hat{\Sigma}_{\mathcal{U}}^{-1}\mathcal{R}^{(1)}+\mathbb{E}\left[\bar{X}_1\bar{X}_1^\intercal\right]^{-1}(1+\hat{\beta}_{21})\hat{\mathcal{R}}_{\mathbb{S}}^{(1K)}+o_{\mathbb{P}}(1)\\
=&\mathbb{E}\left[\bar{X}_1\bar{X}_1^\intercal\right]^{-1}(1+\bar{\beta}_{21})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{Y_{2i}-\bar{\mu}_2(\vec{\mathbf{U}}_i)\}\\
&+\mathbb{E}\left[\bar{X}_1\bar{X}_1^\intercal\right]^{-1}\mathbb{E}\left[\bar{X}_1\left(\bar{\mu}_2(\vec{\mathbf{U}}),\mathbf{H}_{20}^\intercal\right)\right]\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_{2i(\beta)}\\
&+\mathbb{E}\left[\bar{X}_1\bar{X}_1^\intercal\right]^{-1}\mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^\intercal|\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right]\mathbb{P}\left(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_{2i\gamma}\\
&+o_{\mathbb{P}}(1),
\end{aligned}
$$

using (B.2) we have $\mathbb{E}\left[\bar{X}_1\left(\bar{\mu}_2(\vec{\mathbf{U}}),\mathbf{H}_{20}^\intercal\right)\right]=\mathbb{E}\left[\bar{X}_1\left(Y_2,\mathbf{H}_{20}^\intercal\right)\right]$ which yields our required results

□

Next we discuss some results and assumptions needed for Proposition 2.6.1. First we show

the asymptotic results for the supervised estimation of the Q-function parameters. Recall $\widehat{\boldsymbol{\theta}}_{1\mathsf{SUP}}$, $\widehat{\boldsymbol{\theta}}_{2\mathsf{SUP}}$ are the estimators for the Q-function parameters, when using the labeled data $\mathcal{L}$ only. From [54] we have that the following results for $\widehat{\boldsymbol{\theta}}_{2\mathsf{SUP}}$:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{2\mathsf{SUP}}-\bar{\boldsymbol{\theta}}_2\right)=\Sigma_2^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)\to\mathcal{N}\left(\mathbf{0},\mathbf{V}_{2\mathsf{SUP}}\left[\bar{\boldsymbol{\theta}}_2\right]\right),$$

with

$$\begin{aligned}\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)&=\check{\bar{X}}_2\{Y_{3i}-\check{\bar{X}}_{2i}^{\mathsf{T}}\bar{\boldsymbol{\theta}}_2\},\\ \mathbf{V}_{2\mathsf{SUP}}\left[\bar{\boldsymbol{\theta}}_2\right]&=\boldsymbol{\Sigma}_2^{-1}\mathbb{E}\left[\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)^{\mathsf{T}}\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^{\mathsf{T}},\end{aligned}$$

and for $\widehat{\boldsymbol{\theta}}_{1\mathsf{SUP}}$:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{1\mathsf{SUP}}-\bar{\boldsymbol{\theta}}_1\right)=\Sigma_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_{1\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_1)\to\mathcal{N}\left(\mathbf{0},\mathbf{V}_{1\mathsf{SUP}}\left[\bar{\boldsymbol{\theta}}_1\right]\right),$$

with

$$\begin{aligned}\boldsymbol{\psi}_{1\mathsf{SUP}}(\mathbf{L}_i;\bar{\boldsymbol{\theta}}_2)=&\bar{X}_{i1}\{Y_{2i}+Y_{2i}\bar{\beta}_{21}+\mathbf{H}_{20i}^{\mathsf{T}}\bar{\boldsymbol{\beta}}_{22}+[\mathbf{H}_{21i}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}_2]_+-\bar{X}_{1i}^{\mathsf{T}}\bar{\boldsymbol{\theta}}_1\}\\ &+\mathbb{E}\left[\bar{X}_1\left(Y_2,\mathbf{H}_{20}^{\mathsf{T}}\right)\right]\boldsymbol{\psi}_{2\mathsf{SUP},(\beta)}(\mathbf{L}_i)\\ &+\mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^{\mathsf{T}}|\mathbf{H}_{21}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}_2>0\right]\mathbb{P}\left(\mathbf{H}_{21}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}_2>0\right)\boldsymbol{\psi}_{2\mathsf{SUP},(\gamma)}(\mathbf{L}_i),\\ \mathbf{V}_{1\mathsf{SUP}}\left[\bar{\boldsymbol{\theta}}_1\right]=&\boldsymbol{\Sigma}_1^{-1}\mathbb{E}\left[\boldsymbol{\psi}_{1\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_1)\boldsymbol{\psi}_{1\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_1)^{\mathsf{T}}\right]\left(\boldsymbol{\Sigma}_1^{-1}\right)^{\mathsf{T}}.\end{aligned}$$

Next we discuss the assumption required for Proposition 2.6.1. We need the imputation models $\bar{\mu}_s(\vec{\mathbf{U}})$, $s\in\{2,3,22,23\}$ to satisfy several additional constraints. For example, for the stage two Q-function parameters, recall $\boldsymbol{\theta}_{2-}=(\boldsymbol{\beta}_{22}^{\mathsf{T}},\boldsymbol{\gamma}_2^{\mathsf{T}})^{\mathsf{T}}$, the imputation models should satisfy:

$$\mathbb{E}\left[\vec{\bar{X}}^{\mathsf{T}}\bar{\mu}_j(\vec{\mathbf{U}})\{g_s(\mathbf{Y})-\bar{\mu}_s(\vec{\mathbf{U}})\}\right]=\mathbf{0}\quad\mathbb{E}\left[\vec{\bar{X}}^{\mathsf{T}}\bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^{\mathsf{T}}\bar{\boldsymbol{\theta}}_{2-}\{g_s(\mathbf{Y})-\bar{\mu}_s(\vec{\mathbf{U}})\}\right]=\mathbf{0},s,j\in\{2,3,22,23\}$$

$$\mathbb{E}\left[\vec{\bar{X}}\vec{\bar{X}}^{\mathsf{T}}\bar{\mu}_j(\vec{\mathbf{U}})\{g_s(\mathbf{Y})-\bar{\mu}_s(\vec{\mathbf{U}})\}\right]=\mathbf{0},\qquad\mathbb{E}\left[\vec{\bar{X}}\vec{\bar{X}}^{\mathsf{T}}\bar{X}_2^{\mathsf{T}}\bar{\boldsymbol{\theta}}_{2-}\{g_s(\mathbf{Y})-\bar{\mu}_s(\vec{\mathbf{U}})\}\right]=\mathbf{0},s,j\in\{2,3\},$$

where $\bar{X}=(1,\bar{X}_1^{\mathsf{T}},\bar{X}_2^{\mathsf{T}})^{\mathsf{T}}$, $g_2(\mathbf{Y})=Y_2$, $g_3(\mathbf{Y})=Y_3$, $g_{22}(\mathbf{Y})=Y_2^2$, $g_{23}(\mathbf{Y})=Y_2Y_3$.

To summarize all the assumptions needed, we define the following functions:

$$\begin{aligned}
\mathcal{E}^{\theta}(\vec{\mathbf{U}}) \equiv & \left\{\mathcal{E}_1(\vec{\mathbf{U}})^{\intercal}, \mathcal{E}_2(\vec{\mathbf{U}})^{\intercal}\right\}^{\intercal}, \\
\mathcal{E}_2(\vec{\mathbf{U}}) \equiv & \begin{bmatrix} \bar{\mu}_{23}(\vec{\mathbf{U}}) - [\bar{\mu}_{22}(\vec{\mathbf{U}}), \bar{\mu}_2(\vec{\mathbf{U}})\bar{X}_2^{\intercal}]\bar{\boldsymbol{\theta}}_2 \\ \bar{X}_2\{\bar{\mu}_3(\vec{\mathbf{U}}) - [\bar{\mu}_2(\vec{\mathbf{U}}), \bar{X}_2^{\intercal}]\bar{\boldsymbol{\theta}}_2\} \end{bmatrix}, \\
\mathcal{E}_1(\vec{\mathbf{U}}) \equiv & \bar{X}_1\{\bar{\mu}_2(\vec{\mathbf{U}})(1+\bar{\beta}_{21}) + Q^o_{2-}(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2) - \bar{X}_1^{\intercal}\bar{\boldsymbol{\theta}}_1\} \\
& + \mathbb{E}\left[\bar{X}_1\left(Y_2, \mathbf{H}_{20}^{\intercal}\right)\right]\mathcal{E}_{2\beta}(\vec{\mathbf{U}}) \\
& + \mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^{\intercal} | \mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2 > 0\right]\mathbb{P}\left(\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2 > 0\right)\mathcal{E}_{2\gamma}(\vec{\mathbf{U}}),
\end{aligned} \tag{B.5}$$

where $\mathcal{E}_{2\beta}(\vec{\mathbf{U}})$, $\mathcal{E}_{2\gamma}(\vec{\mathbf{U}})$ are the elements corresponding to $\bar{\boldsymbol{\beta}}_2$, $\bar{\boldsymbol{\gamma}}_2$ of $\mathcal{E}_2(\vec{\mathbf{U}})$. Now we can succinctly summarize the constraints, by having $\bar{\mu}_s(\vec{\mathbf{U}})$, $s \in \{2, 3, 22, 23\}$ satisfy

$$\mathbb{E}\left[\left\{\boldsymbol{\psi}_{\mathsf{2SUP}}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2) - \mathcal{E}_2(\vec{\mathbf{U}})\right\}\mathcal{E}_2(\vec{\mathbf{U}})^{\intercal}\right] = \mathbf{0}.$$

This is condensed in the following assumption.

Assumption B.2.1. *Let* $\mathcal{E}^{\theta}(\vec{\mathbf{U}})$ *be as defined in* (B.5)*, and*

$$\boldsymbol{\psi}_{\mathsf{SUP}}(\mathbf{L}; \bar{\boldsymbol{\theta}}) = \left[\boldsymbol{\psi}_{\mathsf{1SUP}}(\mathbf{L}; \bar{\boldsymbol{\theta}}_1)^{\intercal}, \boldsymbol{\psi}_{\mathsf{2SUP}}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2)^{\intercal}\right]^{\intercal},$$

the imputation models $\bar{\mu}_s(\vec{\mathbf{U}})$*,* $s \in \{2, 3, 22, 23\}$ *satisfy*

$$\mathbb{E}\left[\left\{\boldsymbol{\psi}_{\mathsf{SUP}}(\mathbf{L}; \bar{\boldsymbol{\theta}}) - \mathcal{E}^{\theta}(\vec{\mathbf{U}})\right\}\mathcal{E}^{\theta}(\vec{\mathbf{U}})^{\intercal}\right] = \mathbf{0}.$$

Proof of Proposition 2.6.1. We first show the result is true for $\mathbf{V}_{\mathsf{2SSL}}\left[\bar{\boldsymbol{\theta}}_2\right]$. To simplify algebra, we denote the influence function from Theorem 2.6.4 as $\boldsymbol{\psi}_{\mathsf{2SSL}}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2)$. Using the influence function of $\widehat{\boldsymbol{\theta}}_{\mathsf{2SUP}}$ and Theorem 2.6.4 we have the following relationship:

$$\boldsymbol{\psi}_{\mathsf{2SSL}}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2) = \boldsymbol{\psi}_{\mathsf{2SUP}}(\mathbf{L}; \bar{\boldsymbol{\theta}}_2) - \mathcal{E}_2(\vec{\mathbf{U}}).$$

Therefore

$$\begin{aligned}
\mathbf{V}_{2\mathsf{SSL}}\left(\bar{\boldsymbol{\theta}}_2\right) =& \boldsymbol{\Sigma}_2^{-1}\mathbb{E}\left[\boldsymbol{\psi}_{2\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)\boldsymbol{\psi}_{2\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)^\intercal\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^\intercal \\
=& \boldsymbol{\Sigma}_2^{-1}\mathbb{E}\left[\left\{\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)-\mathcal{E}_2(\vec{\mathbf{U}})\right\}\left\{\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)-\mathcal{E}_2(\vec{\mathbf{U}})\right\}^\intercal\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^\intercal \\
=& \boldsymbol{\Sigma}_2^{-1}\mathbb{E}\left[\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)^\intercal\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^\intercal \\
&+\boldsymbol{\Sigma}_2^{-1}\mathbb{E}\left[\mathcal{E}_2(\vec{\mathbf{U}})\mathcal{E}_2(\vec{\mathbf{U}})^\intercal\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^\intercal \\
&-2\boldsymbol{\Sigma}_2^{-1}\mathbb{E}\left[\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)\mathcal{E}_2(\vec{\mathbf{U}})^\intercal\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^\intercal
\end{aligned}$$

Now, since our imputation models satisfy Assumption B.2.1, it follows that

$$\mathbb{E}\left[\left\{\boldsymbol{\psi}_{2\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_2)-\mathcal{E}_2(\vec{\mathbf{U}})\right\}\mathcal{E}_2(\vec{\mathbf{U}})^\intercal\right]=\mathbf{0}.$$

Therefore we have

$$\mathbf{V}_{2\mathsf{SSL}}\left(\bar{\boldsymbol{\theta}}_2\right) =\mathbf{V}_{2\mathsf{SUP}}\left(\bar{\boldsymbol{\theta}}_2\right)-\boldsymbol{\Sigma}_2^{-1}\mathrm{Var}\left[\mathcal{E}_2(\vec{\mathbf{U}})\right]\left(\boldsymbol{\Sigma}_2^{-1}\right)^\intercal.$$

To show the result is true for $\mathbf{V}_{1\mathsf{SSL}}\left[\bar{\boldsymbol{\theta}}_1\right]$, We denote by $\mathcal{E}_{2\beta}(\vec{\mathbf{U}})$ and $\mathcal{E}_{2\gamma}(\vec{\mathbf{U}})$ the vectors corresponding to $\bar{\boldsymbol{\beta}}_2$, $\bar{\boldsymbol{\gamma}}_2$ in $\mathcal{E}_2(\vec{\mathbf{U}})$ respectively, and further recall the definition of $\mathcal{E}_1(\vec{\mathbf{U}})$:

$$\begin{aligned}
\mathcal{E}_1(\vec{\mathbf{U}}) =& \bar{X}_1\{\bar{\mu}_2(\vec{\mathbf{U}})+\bar{\mu}_2(\vec{\mathbf{U}})\bar{\beta}_{21}+\mathbf{H}_{20}^\intercal\bar{\boldsymbol{\beta}}_{22}+[\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2]_+-\bar{X}_1^\intercal\bar{\boldsymbol{\theta}}_1\} \\
&+\mathbb{E}\left[\bar{X}_1\left(Y_2,\mathbf{H}_{20}^\intercal\right)\right]\mathcal{E}_{2\beta}(\vec{\mathbf{U}}) \\
&+\mathbb{E}\left[\bar{X}_1\mathbf{H}_{21}^\intercal|\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right]\mathbb{P}\left(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2>0\right)\mathcal{E}_{2\gamma}(\vec{\mathbf{U}}).
\end{aligned}$$

From the form of the influence function of $\widehat{\boldsymbol{\theta}}_{1\mathsf{SUP}}$, and Theorems 2.6.4 & 2.6.6 we have that:

$$\boldsymbol{\psi}_{1\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\theta}}_1)=\boldsymbol{\psi}_{1\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}_1)-\mathcal{E}_1(\vec{\mathbf{U}}).$$

Analogous steps for the proof of $\bar{\boldsymbol{\theta}}_2$ can then be used to show

$$\mathbf{V}_{1\mathsf{SSL}}\left(\bar{\boldsymbol{\theta}}_1\right) =\mathbf{V}_{1\mathsf{SUP}}\left(\bar{\boldsymbol{\theta}}_1\right)-\boldsymbol{\Sigma}_1^{-1}\mathrm{Var}\left[\mathcal{E}_1(\vec{\mathbf{U}})\right]\left(\boldsymbol{\Sigma}_1^{-1}\right)^\intercal.$$

The required result is obtained by stacking the influence functions for $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2$ for the

supervised and semi-supervised versions, noting that

$$\boldsymbol{\psi}_{\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\theta}}) = \boldsymbol{\psi}_{\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\theta}}) - \mathcal{E}^{\theta}(\vec{\mathbf{U}}).$$

and repeating the steps above. □

B.2.2 Value Function Results

In this Section we prove the main results for our SSL value function estimator. Before the proofs we go over some useful definitions, notation and lemmas. First recall that, in order to correct for potential biases arising from finite sample estimation and model mis-specifications, the final imputed models for $\{Y_2, \omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}),\ Y_t\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}),\ t = 2, 3\}$ satisfy the following constraints:

$$\begin{aligned}
\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}})\left\{Y_2 - \bar{\mu}_2^v(\vec{\mathbf{U}})\right\}\right] &= 0, \\
\mathbb{E}\left[Q_{2-}^o(\vec{\mathbf{U}}; \boldsymbol{\theta}_2)\left\{\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}) - \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}})\right\}\right] &= 0, \\
\mathbb{E}\left[\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}})Y_t - \bar{\mu}_{t\omega_2}^v(\vec{\mathbf{U}})\right] &= 0,\ t = 2, 3.
\end{aligned} \tag{B.6}$$

Next, define the set

$$\begin{aligned}
\mathcal{S}(\delta) = \Bigg\{(\boldsymbol{\theta}, \boldsymbol{\xi}) \Bigg| \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 < \delta, \|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}\|_2^2 < \delta, \boldsymbol{\theta}_t \in \Theta_t, \boldsymbol{\xi}_t \in \Omega_t, t = 1, 2, \\
\pi_1(\mathbf{H}_1; \boldsymbol{\xi}_1) > 0,\ \pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2) > 0,\ \forall\, \mathbf{H} \in \mathcal{H}\Bigg\}.
\end{aligned}$$

We will be using the influence functions for our model parameters $\boldsymbol{\Theta}$. In this regard let $\boldsymbol{\psi}^{\theta} = (\boldsymbol{\psi}_1^{\mathsf{T}}, \boldsymbol{\psi}_2^{\mathsf{T}})^{\mathsf{T}}$. By Theorems 2.6.4 & 2.6.6 $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}^{\theta}(\vec{\mathbf{U}}_i) + o_{\mathbb{P}}(1)$. Next, from Assumption 2.6.7, it can be shown that $\widehat{\boldsymbol{\xi}}$ has the following expansion: $\sqrt{n}(\widehat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\psi}^{\xi}\left(\mathbf{L}_i; \bar{\boldsymbol{\xi}}\right) + o_{\mathbb{P}}(1)$, where

$$\boldsymbol{\psi}_t^{\xi}\left(\mathbf{L}; \bar{\boldsymbol{\xi}}\right) = \mathbb{E}\left\{\check{\mathbf{H}}_t^{\mathsf{T}}\check{\mathbf{H}}_t\sigma\left(\check{\mathbf{H}}_t^{\mathsf{T}}\bar{\boldsymbol{\xi}}_t\right)\left[1 - \sigma\left(\check{\mathbf{H}}_t^{\mathsf{T}}\bar{\boldsymbol{\xi}}_t\right)\right]\right\}^{-1}\check{\mathbf{H}}_t\left\{A_t - \sigma\left(\check{\mathbf{H}}_t^{\mathsf{T}}\bar{\boldsymbol{\xi}}_t\right)\right\},\ \ t = 1, 2,$$

$\boldsymbol{\psi}^{\xi}\left(\mathbf{L}; \bar{\boldsymbol{\xi}}\right) = \left[\boldsymbol{\psi}_1^{\xi}\left(\mathbf{L}; \bar{\boldsymbol{\xi}}\right), \boldsymbol{\psi}_2^{\xi}\left(\mathbf{L}; \bar{\boldsymbol{\xi}}\right)\right]$ and $\mathbb{E}[\boldsymbol{\psi}^{\xi}] = 0$, $\mathbb{E}[(\boldsymbol{\psi}^{\xi})^{\mathsf{T}}\boldsymbol{\psi}^{\xi}] < \infty$.

We now introduce a set of definitions used in this section to make the proofs easier to read. Recall from (2.7) we have

$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} = \mathbb{P}_N\left\{\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \widehat{\boldsymbol{\Theta}}, \widehat{\mu})\right\}$, where $\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \widehat{\boldsymbol{\Theta}}, \widehat{\mu})$ is the semi-supervised augmented

estimator for observation $\vec{\mathbf{U}}$, we re-write $\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \widehat{\boldsymbol{\Theta}}, \widehat{\mu})$ as $\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\widehat{\mu}}(\vec{\mathbf{U}})$ recall its definition, and define the following functions:

$$\begin{aligned}
\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\widehat{\mu}}(\vec{\mathbf{U}}) \equiv & Q_1^o(\check{\mathbf{H}}_1; \widehat{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \widehat{\boldsymbol{\Theta}}) \left[(1+\widehat{\beta}_{21})\widehat{\mu}_2^v(\vec{\mathbf{U}}) - Q_1^o(\check{\mathbf{H}}_1; \widehat{\boldsymbol{\theta}}_1) + Q_{2-}^o\{\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2\}\right] \\
& + \widehat{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \widehat{\beta}_{21}\widehat{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\widehat{\mu}_{\omega_2}^v(\vec{\mathbf{U}}), \\
\mathcal{V}_{\bar{\boldsymbol{\Theta}},\widehat{\mu}}(\vec{\mathbf{U}}) \equiv & Q_1^o(\check{\mathbf{H}}_1; \bar{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \bar{\boldsymbol{\Theta}}) \left[(1+\bar{\beta}_{21})\widehat{\mu}_2^v(\vec{\mathbf{U}}) - Q_1^o(\check{\mathbf{H}}_1; \bar{\boldsymbol{\theta}}_1) + Q_{2-}^o\{\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2\}\right] \\
& + \widehat{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \bar{\beta}_{21}\widehat{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - Q_{2-}^o(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\widehat{\mu}_{\omega_2}^v(\vec{\mathbf{U}}).
\end{aligned} \tag{B.7}$$

We next replace the estimated imputation functions with their limits $\bar{\mu}_2^v$, $\bar{\mu}_{2\omega_2}^v$, $\bar{\mu}_{3\omega_2}^v$ and $\bar{\mu}_{\omega_2}^v$, and define:

$$\begin{aligned}
\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}}(\vec{\mathbf{U}}) \equiv & Q_1^o(\check{\mathbf{H}}_1; \widehat{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \widehat{\boldsymbol{\Theta}}) \left[(1+\widehat{\beta}_{21})\bar{\mu}_2^v(\vec{\mathbf{U}}) - Q_1^o(\check{\mathbf{H}}_1; \widehat{\boldsymbol{\theta}}_1) + Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\right] \\
& + \bar{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \hat{\beta}_{21}\bar{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}}), \\
\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}(\vec{\mathbf{U}}) \equiv & Q_1^o(\check{\mathbf{H}}_1; \bar{\boldsymbol{\theta}}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \bar{\boldsymbol{\Theta}}) \left[(1+\bar{\beta}_{21})\bar{\mu}_2^v(\vec{\mathbf{U}}) - Q_1^o(\check{\mathbf{H}}_1; \bar{\boldsymbol{\theta}}_1) + Q_{2-}^o(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\right] \\
& + \bar{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \bar{\beta}_{21}\bar{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - Q_{2-}^o(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}}).
\end{aligned} \tag{B.8}$$

Finally we define the following functions which are weighted sums of the imputation function errors:

$$\begin{aligned}
\mathcal{E}_{\widehat{\boldsymbol{\Theta}}}(\vec{\mathbf{U}}) \equiv & \omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}})(1+\hat{\beta}_{21}) \left\{\widehat{\mu}_2^v(\vec{\mathbf{U}}) - \bar{\mu}_2^v(\vec{\mathbf{U}})\right\} + \widehat{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) \\
& - \hat{\beta}_{21}\left\{\widehat{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{2\omega_2}^v(\vec{\mathbf{U}})\right\} - Q_{2-}^o\left(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2\right)\left\{\widehat{\mu}_{\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}})\right\}, \\
\mathcal{E}_{\bar{\boldsymbol{\Theta}}}(\vec{\mathbf{U}}) \equiv & \omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}})(1+\bar{\beta}_{21}) \left\{\widehat{\mu}_2^v(\vec{\mathbf{U}}) - \bar{\mu}_2^v(\vec{\mathbf{U}})\right\} + \widehat{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) \\
& - \bar{\beta}_{21}\left\{\widehat{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{2\omega_2}^v(\vec{\mathbf{U}})\right\} - Q_{2-}^o\left(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2\right)\left\{\widehat{\mu}_{\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}})\right\}.
\end{aligned} \tag{B.9}$$

These definitions will come in handy in the following proofs as we can use them to write $\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\widehat{\mu}}(\vec{\mathbf{U}}) = \mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}}(\vec{\mathbf{U}}) + \mathcal{E}_{\widehat{\boldsymbol{\Theta}}}(\vec{\mathbf{U}})$, $\mathcal{V}_{\bar{\boldsymbol{\Theta}},\widehat{\mu}}(\vec{\mathbf{U}}) = \mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}(\vec{\mathbf{U}}) + \mathcal{E}_{\bar{\boldsymbol{\Theta}}}(\vec{\mathbf{U}})$. Finally, recalling that $\mathbb{P}_{\vec{\mathbf{U}}}$ is the underlying distribution of the data, we define function $g_1 : \boldsymbol{\Theta} \mapsto \mathbb{R}$ as

$$g_1(\boldsymbol{\Theta}) = \int \mathcal{V}_{\boldsymbol{\Theta},\bar{\mu}}(\vec{\mathbf{U}}) d\mathbb{P}_{\vec{\mathbf{U}}}.$$

With the above definitions we proceed by stating three lemmas that will be used to prove Theorem 2.6.9. We defer the proofs of these lemmas for after proving the main Theorem in this section.

Lemma B.2.2. *Under Assumptions 2.6.1-2.6.8, we have*

I) $\sqrt{n}\left\{\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1\left(\bar{\boldsymbol{\Theta}}\right)\right\} = o_{\mathbb{P}}(1),$

II) $\sqrt{n}\left\{g_1(\widehat{\boldsymbol{\Theta}}) - g_1\left(\bar{\boldsymbol{\Theta}}\right)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left(\frac{\partial}{\partial\boldsymbol{\theta}}g_1\left(\bar{\boldsymbol{\Theta}}\right)\right)^{\intercal}\boldsymbol{\psi}^{\theta}(\vec{\mathbf{U}}_i) + \left(\frac{\partial}{\partial\boldsymbol{\xi}}g_1\left(\bar{\boldsymbol{\Theta}}\right)\right)^{\intercal}\boldsymbol{\psi}^{\xi}(\vec{\mathbf{U}}_i)\right\} + o_{\mathbb{P}}(1).$

Lemma B.2.3. *Under Assumptions 2.6.1-2.6.8, the following holds:*

$$\sqrt{n}\left\{\left(\mathbb{P}_N\left[\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\widehat{\boldsymbol{\Theta}})\right) - \left(\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\bar{\boldsymbol{\Theta}})\right)\right\} = o_{\mathbb{P}}\left(1\right).$$

Lemma B.2.4. *Under Assumptions 2.6.1-2.6.8, the following assertions hold:*

I) $\sqrt{n}\mathbb{P}_N\left\{\mathcal{E}_{\widehat{\boldsymbol{\Theta}}} - \mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right\} = o_{\mathbb{P}}(1),$

II) $$\begin{aligned}\sqrt{n}\mathbb{P}_N\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right] = \mathbb{G}_n\left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right\} \\ + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\boldsymbol{\psi}^{\theta}(\mathbf{L}_i)^{\intercal}\frac{\partial}{\partial\boldsymbol{\theta}}\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\boldsymbol{\Theta})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}} + \boldsymbol{\psi}^{\xi}(\mathbf{L}_i)^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\boldsymbol{\Theta})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}\right\} \\ + o_{\mathbb{P}}(1).\end{aligned}$$

Proof of Theorem 2.6.9. We start by expanding the expression in (2.7) and using definitions (B.7), (B.8), (B.9):

$$
\begin{aligned}
&\sqrt{n}\left\{\mathbb{P}_N\left[\mathcal{V}_{\widehat{\bar{\boldsymbol{\Theta}}},\widehat{\bar{\mu}}}\right]-\mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]\right\}\\
=&\sqrt{n}\left\{\underbrace{\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]+\mathbb{P}_N\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right]}_{(I)}-\underbrace{g_1(\bar{\boldsymbol{\Theta}})-\mathbb{E}_{\mathbb{S}}\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right]}_{(II)}\right\}\\
+&\sqrt{n}\left\{\left(\mathbb{P}_N\left[\mathcal{V}_{\widehat{\bar{\boldsymbol{\Theta}}},\bar{\mu}}\right]+\mathbb{P}_N\left[\mathcal{E}_{\widehat{\bar{\boldsymbol{\Theta}}}}\right]-\underbrace{g_1(\widehat{\bar{\boldsymbol{\Theta}}})}_{(III)}\right)-\mathbb{E}_{\mathbb{S}}\left[\mathcal{E}_{\widehat{\bar{\boldsymbol{\Theta}}}}\right]\right\}-\left\{\underbrace{\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]+\mathbb{P}_N\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right]}_{(I)}-\underbrace{g_1(\bar{\boldsymbol{\Theta}})-\mathbb{E}_{\mathbb{S}}\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right]}_{(II)}\right\}\\
+&\sqrt{n}\left\{\underbrace{g_1(\widehat{\bar{\boldsymbol{\Theta}}})+\mathbb{E}_{\mathbb{S}}\left[\mathcal{E}_{\widehat{\bar{\boldsymbol{\Theta}}}}\right]}_{(III)}-\mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]\right\}\\
=&\sqrt{n}\left\{\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]-g_1(\bar{\boldsymbol{\Theta}})\right\}\\
+&\sqrt{n}\left\{g_1(\widehat{\bar{\boldsymbol{\Theta}}})-g_1(\bar{\boldsymbol{\Theta}})\right\}\\
+&\sqrt{n}\left\{\left(\mathbb{P}_N\left[\mathcal{V}_{\widehat{\bar{\boldsymbol{\Theta}}},\bar{\mu}}\right]-g_1(\widehat{\bar{\boldsymbol{\Theta}}})\right)-\left(\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]-g_1(\bar{\boldsymbol{\Theta}})\right)\right\}\\
+&\sqrt{n}\mathbb{P}_N\left[\mathcal{E}_{\widehat{\bar{\boldsymbol{\Theta}}}}-\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right]\\
+&\sqrt{n}\mathbb{P}_N\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right]\\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}})+o_{\mathbb{P}}(1).
\end{aligned}
$$

which follows from Lemmas B.2.2, B.2.3 & B.2.4 with the influence function $\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}$ defined as

$$
\begin{aligned}
\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})=&\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})+\boldsymbol{\psi}^{\theta}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\theta}}\int\left\{\mathcal{V}_{\boldsymbol{\Theta},\bar{\mu}}(\mathbf{L})+\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right\}d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}\\
&+\boldsymbol{\psi}^{\xi}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}\int\left\{\mathcal{V}_{\boldsymbol{\Theta},\bar{\mu}}(\vec{\mathbf{U}})+\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right\}d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}},\\
\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})=&\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)(1+\bar{\beta}_{21})\left\{Y_2-\bar{\mu}^{v}_2(\vec{\mathbf{U}})\right\}+\omega_2(\check{\mathbf{H}}_2,A_2,\bar{\boldsymbol{\Theta}}_2)Y_3-\bar{\mu}_{3\omega_2}(\vec{\mathbf{U}})\\
&-\bar{\beta}_{21}\left\{\omega_2(\check{\mathbf{H}}_2,A_2,\bar{\boldsymbol{\Theta}}_2)Y_2-\bar{\mu}_{2\omega_2}(\vec{\mathbf{U}})\right\}-Q^{o}_{2-}(\mathbf{H}_2;\bar{\boldsymbol{\theta}}_2)\left\{\omega_2(\check{\mathbf{H}}_2,A_2,\bar{\boldsymbol{\Theta}}_2)-\bar{\mu}_{\omega_2}(\vec{\mathbf{U}})\right\}
\end{aligned}
$$

Next note that

$$\int \left(\mathcal{V}_{\boldsymbol{\Theta},\bar{\mu}}(\vec{\mathbf{U}}) + \nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right) d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}} = \int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta}) d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}},$$

where $\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})$ is defined in (2.5). Finally, all random variables in the expression of $\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})$ are bounded by Assumptions 2.6.1 and 2.6.7 we have $\mathbb{E}\left[\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})^2\right] < \infty$, the central limit theorem yields that

$$\sqrt{n}\left\{\mathbb{P}_N\left[\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\widehat{\mu}}\right] - g_1(\bar{\boldsymbol{\Theta}})\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{v}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}}) + o_{\mathbb{P}}(1) \xrightarrow{d} N\left(0, \sigma^2_{\mathsf{SSL}_{\mathsf{DR}}}\right).$$

□

Proof of Lemma B.2.2. I) We start with $\sqrt{n}\left\{\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\bar{\boldsymbol{\Theta}})\right\}$. Note that $\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}(\vec{\mathbf{U}})$ is a deterministic function of random variable $\vec{\mathbf{U}}$ as parameters and imputation functions are fixed. We have that $\mathbb{E}\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}(\vec{\mathbf{U}})^2\right] < \infty$ holds by Assumption 2.6.1 & 2.6.7. Thus the central limit theorem yields $\mathbb{G}_N\left\{\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right\} \xrightarrow{d} \mathcal{N}\left(0, Var\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right]\right)$, therefore

$$\sqrt{n}\left\{\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\bar{\boldsymbol{\Theta}})\right\} = \sqrt{\frac{n}{N}}\mathbb{G}_N\left\{\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right\} = O_{\mathbb{P}}\left(\frac{\sqrt{n}}{N}\right) = o_{\mathbb{P}}(1).$$

II) We next consider $\sqrt{n}\left\{g_1(\widehat{\boldsymbol{\Theta}}) - g_1(\bar{\boldsymbol{\Theta}})\right\}$. Using a Taylor series expansion

$$g_1(\widehat{\boldsymbol{\Theta}}) = g_1(\bar{\boldsymbol{\Theta}}) + (\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^{\intercal}\frac{\partial}{\partial \boldsymbol{\theta}}g_1(\bar{\boldsymbol{\Theta}}) + (\widehat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}})^{\intercal}\frac{\partial}{\partial \boldsymbol{\xi}}g_1(\bar{\boldsymbol{\Theta}}) + O_{\mathbb{P}}\left(n^{-1}\right),$$

as both $\|\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\|_2^2 = O_{\mathbb{P}}\left(n^{-1}\right)$ and $\|\widehat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}}\|_2^2 = O_{\mathbb{P}}\left(n^{-1}\right)$ by Theorems 2.6.4, 2.6.6 and Assumption 2.6.7, therefore

$$\sqrt{n}\left\{g_1(\widehat{\boldsymbol{\Theta}}) - g_1(\bar{\boldsymbol{\Theta}})\right\} = \sqrt{n}(\widehat{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}})^{\intercal}\frac{\partial}{\partial \boldsymbol{\theta}}g_1(\bar{\boldsymbol{\Theta}}) + \sqrt{n}(\widehat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}})^{\intercal}\frac{\partial}{\partial \boldsymbol{\xi}}g_1(\bar{\boldsymbol{\Theta}}) + o_{\mathbb{P}}(1).$$

We can write

$$\sqrt{n}\left\{g_1(\widehat{\boldsymbol{\Theta}}) - g_1(\bar{\boldsymbol{\Theta}})\right\} = \frac{\partial}{\partial \boldsymbol{\theta}}g_1(\bar{\boldsymbol{\Theta}})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}^{\theta}(\vec{\mathbf{U}}_i) + \frac{\partial}{\partial \boldsymbol{\xi}}g_1(\bar{\boldsymbol{\Theta}})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}^{\xi}(\vec{\mathbf{U}}_i) + o_{\mathbb{P}}(1).$$

□

Proof of Lemma B.2.3. We consider $\sqrt{n}\left\{\left(\mathbb{P}_N\left[\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\widehat{\boldsymbol{\Theta}})\right) - \left(\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\bar{\boldsymbol{\Theta}})\right)\right\}$, re-

call that $d_t(\check{\mathbf{H}}_t, \boldsymbol{\theta}_t) = I(\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t > 0)$ $t = 1, 2$, thus the inverse probability weight functions are defined as

$$\omega_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \equiv \frac{I(\mathbf{H}_{11}^\intercal \boldsymbol{\gamma}_1 > 0) A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{\{1 - I(\mathbf{H}_{11}^\intercal \boldsymbol{\gamma}_1 > 0)\}\{1 - A_1\}}{1 - \pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}, \quad \text{and}$$
$$\omega_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta}) \equiv \omega_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \left(\frac{I(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2 > 0) A_2}{\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)} + \frac{\{1 - I(\mathbf{H}_{21}^\intercal \boldsymbol{\gamma}_2 > 0)\}\{1 - A_2\}}{1 - \pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)} \right).$$

Define the class

$$\ell_t = \{ I\left(\mathbf{H}_t^\intercal \boldsymbol{\gamma}_t \geq 0\right) : \mathcal{H}_{t1}, \boldsymbol{\gamma} \in \mathbb{R}^{q_t} \}, \; t = 1, 2$$

and the collection of half spaces $\mathcal{C}_\ell \equiv \{\mathbf{H}_t \in \mathbb{R}^{q_t} : \mathbf{H}_t^\intercal \boldsymbol{\gamma}_t \geq 0, \boldsymbol{\gamma} \in \mathbb{R}^{q_t}, t \in \{1, 2\}\}$. By **(author?)** [96] $\mathcal{C}_\ell$ is a VC class of VC dimension $q_t + 1$. Next by [97] we have that as $\mathcal{C}_\ell$ is a VC-class ℓ_t is a class of the same index. Finally, by Theorem 2.6.7 we have that ℓ_t is a $\mathbb{P}$-Donsker class. Next define the following function

$$\begin{aligned} f_{\boldsymbol{\Theta}}(\vec{\mathbf{U}}) =& Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1) + \omega_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \left[(1 + \beta_{21}) \bar{\mu}_2^v(\vec{\mathbf{U}}) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1) + Q_{2-}^o(\mathbf{H}_2; \boldsymbol{\theta}_2) \right] \\ &+ \bar{\mu}_{3\omega_2}^v(\vec{\mathbf{U}}) - \beta_{21} \bar{\mu}_{2\omega_2}^v(\vec{\mathbf{U}}) - Q_{2-}^o(\mathbf{H}_2; \boldsymbol{\theta}_2) \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}}). \end{aligned}$$

We define the associated class of functions $\mathcal{C}_1 = \left\{ f_{\boldsymbol{\Theta}}(\vec{\mathbf{U}}) | \vec{\mathbf{U}}, \boldsymbol{\Theta} \in \mathcal{S}(\delta) \right\}$.

i) By Assumptions 2.6.3, 2.6.7 and Theorem 19.5 in [61], ℓ_t, $\mathcal{W}_t$, $\mathcal{Q}_t, t = 1, 2$ are $\mathbb{P}$-Donsker classes. Thus it follows that $\mathcal{C}_1$ is a Donsker class.

ii) We estimate $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ for (2.4) with their maximum likelihood estimators, $\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2$, solving $\mathbb{P}_n \left[S_t(\boldsymbol{\xi}_t) \right] = \mathbf{0}, t = 1, 2$. By Assumption (2.6.7) and Theorem 5.9 in [61] $\widehat{\boldsymbol{\xi}}_t \xrightarrow{p} \bar{\boldsymbol{\xi}}_t, t = 1, 2$. Next, by Theorems 2.6.4, 2.6.6, under Assumptions 2.6.1, 2.6.2, $\widehat{\boldsymbol{\theta}}_t \xrightarrow{p} \bar{\boldsymbol{\theta}}_t, t = 1, 2$. Thus $\mathbb{P}\left(\widehat{\boldsymbol{\Theta}} \in \mathcal{S}(\delta) \right) \to 1$, $\forall \delta$.

iii) We next show $\int \left(\mathcal{V}_{\widehat{\boldsymbol{\Theta}}, \bar{\mu}} - \mathcal{V}_{\bar{\boldsymbol{\Theta}}, \bar{\mu}} \right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \longrightarrow 0$. By Assumptions 2.6.7 (ii), 2.6.8, and bounded covariates and there exists a constant $c \in \mathbb{R}$ such that we can write

$$
\begin{aligned}
&\int \left(\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}} - \mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \\
\leq &\int \left(Q_1^o(\mathbf{H}_1;\widehat{\boldsymbol{\theta}}_1) - Q_1^o(\mathbf{H}_1;\bar{\boldsymbol{\theta}}_1)\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \\
+c &\int \left(\frac{1}{1-\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)} - \frac{1}{1-\pi_1(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1)}\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \\
+c &\int \left(\frac{1}{\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)} - \frac{1}{\pi_1(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1)}\right)^2 \\
+c &\int \left\{Q_{2-}^o(\check{\mathbf{H}}_2;\widehat{\boldsymbol{\theta}}_2) - Q_{2-}^o(\mathbf{H}_2;\bar{\boldsymbol{\theta}}_2)\right\}^2 d\mathbb{P}_{\vec{\mathbf{U}}} \\
+c &\int \left\{I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 > 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 > 0)\right\}^2 d\mathbb{P}_{\vec{\mathbf{U}}} \\
+ &\left(\hat{\beta}_{21} - \bar{\beta}_{21}\right)^2 \\
+c &\int \left(\mathbf{H}_{20}^\intercal\bar{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{20}^\intercal\bar{\boldsymbol{\gamma}}_2]_+ - \mathbf{H}_{20}^\intercal\widehat{\boldsymbol{\beta}}_{22} - [\mathbf{H}_{20}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}}
\end{aligned}
$$

where we use $(a-b)^2, (a+b)^2 \leq 2a^2 + 2b^2\ \forall a, b \in \mathbb{R}$, $\widehat{d}_1, A_1 \leq 1$ for all $\mathbf{H} \in \mathcal{H}$, and boundedness of $\widehat{\boldsymbol{\theta}}_t, t = 1, 2$ by Assumptions 2.6.1-2.6.3. Next note that all terms outside integrals are bounded by Assumptions 2.6.1-2.6.3. Finally we consider terms within the integrals with the following example

$$
\begin{aligned}
\int \left(Q_{2-}^o(\mathbf{H}_2;\widehat{\boldsymbol{\theta}}_2) - Q_{2-}^o(\mathbf{H}_2;\bar{\boldsymbol{\theta}})\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} = &\int \left(\mathbf{H}_{20}^\intercal\widehat{\boldsymbol{\beta}}_{22} + [\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+ - \mathbf{H}_{20}^\intercal\bar{\boldsymbol{\beta}}_{22} - [\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2]_+\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \\
=&4\|\widehat{\boldsymbol{\beta}}_{22} - \bar{\boldsymbol{\beta}}_{22}\|_2^2 \int \mathbf{H}_{20}^\intercal\mathbf{H}_{20} d\mathbb{P}_{\vec{\mathbf{U}}} \\
&+4\|\widehat{\boldsymbol{\gamma}}_2 - \bar{\boldsymbol{\gamma}}_2\|_2^2 \int \mathbf{H}_{21}^\intercal\mathbf{H}_{21} d\mathbb{P}_{\vec{\mathbf{U}}} = O_{\mathbb{P}}(n^{-1}),
\end{aligned}
$$

which follows from Theorem 2.6.4 and Lemma B.3.5 (a). All similar terms can be handled accordingly. We get the convergence in probability to 0: $\int \left(\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}} - \mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \to 0$ as all other terms within expectation are $O_{\mathbb{P}}\left(n^{-1}\right)$ by the dominating convergence theorem, boundedness conditions as stated in Assumptions 2.6.2, 2.6.7, and the consistency of $\widehat{\boldsymbol{\xi}}$ and $\widehat{\boldsymbol{\theta}}$ as $\mathbb{P}\left(\widehat{\boldsymbol{\Theta}} \in \mathcal{S}(\delta)\right) \to 1$, $\forall \delta > 0$.

Finally, we have i) $\mathbb{P}\left(\widehat{\boldsymbol{\Theta}} \in \mathcal{S}(\delta)\right) \to 1$, ii) $\mathcal{C}_1$ is a Donsker class, and

iii) $\int \left(\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}} - \mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right)^2 d\mathbb{P}_{\vec{\mathbf{U}}} \longrightarrow 0$, then by Theorem 2.1 in [98],

$$\sqrt{\frac{n}{N}}\sqrt{n}\left\{\left(\mathbb{P}_N\left[\mathcal{V}_{\widehat{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\widehat{\boldsymbol{\Theta}})\right) - \left(\mathbb{P}_N\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - g_1(\bar{\boldsymbol{\Theta}})\right)\right\} = \sqrt{\frac{n}{N}}o_{\mathbb{P}}(1).$$

□

Proof of Lemma B.2.4. I) First note that from the empirical normal equations (2.6), we have that the solution $\widehat{\eta}_2^v$ satisfies $\widehat{\eta}_2^v - \eta_2^v = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. Therefore

$$\begin{aligned}
\sup_{\vec{\mathbf{U}}}\left|\widehat{\mu}_2^v(\vec{\mathbf{U}}) - \mu_2^v(\vec{\mathbf{U}})\right| &= \sup_{\vec{\mathbf{U}}}\left|\frac{1}{K}\widehat{m}_2^{(\text{-}k)}(\vec{\mathbf{U}}) + \widehat{\eta}_2^v - m_2(\vec{\mathbf{U}}) + \eta_2^v\right| \\
&\leq \frac{1}{K}\sup_{\vec{\mathbf{U}}}\left|\widehat{m}_2^{(\text{-}k)}(\vec{\mathbf{U}}) + m_2(\vec{\mathbf{U}})\right| + |\widehat{\eta}_2^v - \eta_2^v| \\
&= o_{\mathbb{P}}(1) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) = o_{\mathbb{P}}(1),
\end{aligned}$$

where we additionally use Assumption 2.6.8 for the difference of estimated and true imputation models $\hat{m}_2$, m_2. Similarly $\sup_{\vec{\mathbf{U}}}\left|\widehat{\mu}_{t\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{t\omega_2}^v(\vec{\mathbf{U}})\right| = o_{\mathbb{P}}(1)$, $\sup_{\vec{\mathbf{U}}}\left|\widehat{\mu}_{\omega_2}^v(\vec{\mathbf{U}}) - \bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}})\right| = o_{\mathbb{P}}(1)$, $t = 2, 3$. Next, using the triangle and Jensen's inequalities, we have

$$\begin{aligned}
&\mathbb{P}_N\left[\mathcal{E}_{\widehat{\boldsymbol{\Theta}}} - \mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right] \\
\leq&\mathbb{P}_N\left|\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1)(1+\hat{\beta}_{21}) - \omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)(1+\bar{\beta}_{21})\right|\sup_{\vec{\mathbf{U}}}\left|\widehat{\mu}_2^v(\vec{\mathbf{U}}) - \bar{\mu}_2^v(\vec{\mathbf{U}})\right| \\
+&\left|\hat{\beta}_{21} - \bar{\beta}_{21}\right|\sup_{\vec{\mathbf{U}}}\left|\widehat{\mu}_{2\omega_2}(\vec{\mathbf{U}}) - \mu_{2\omega_2}(\vec{\mathbf{U}})\right| \\
+&\mathbb{P}_N\left|Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2) - Q_{2-}^o(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\right|\sup_{\vec{\mathbf{U}}}\left|\widehat{\mu}_{\omega_2}(\vec{\mathbf{U}}) - \bar{\mu}_{\omega_2}(\vec{\mathbf{U}})\right| \\
\leq&\mathbb{P}_N\left|\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1) - \omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\right|o_{\mathbb{P}}(1) + \mathbb{P}_N\left|\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1)\hat{\beta}_{21} - \omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\bar{\beta}_{21}\right|o_{\mathbb{P}}(1) \\
+&\left|\hat{\beta}_{21} - \bar{\beta}_{21}\right|o_{\mathbb{P}}(1) + \mathbb{P}_N\left|\left(\widehat{\boldsymbol{\beta}}_{22} - \widehat{\boldsymbol{\beta}}_{22}\right)^{\mathsf{T}}\mathbf{H}_{20} + [\widehat{\boldsymbol{\gamma}}_2^{\mathsf{T}}\mathbf{H}_{21}]_+ - [\bar{\boldsymbol{\gamma}}_2^{\mathsf{T}}\mathbf{H}_{21}]_+\right|o_{\mathbb{P}}(1).
\end{aligned}$$

By Theorem 2.6.4 we have $\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, also from Lemma B.3.5 (a) it follows that

$\mathbb{P}_N\left([\mathbf{H}_{21}^{\intercal}\widehat{\bar{\gamma}}_2]_+ - [\mathbf{H}_{21}^{\intercal}\bar{\gamma}_2]_+\right) = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, hence as covariates are bounded we have

$$\begin{aligned}
&\left|\hat{\beta}_{21} - \bar{\beta}_{21}\right| o_{\mathbb{P}}(1) + \mathbb{P}_N\left|\left(\widehat{\boldsymbol{\beta}}_{22} - \widehat{\boldsymbol{\beta}}_{22}\right)^{\intercal}\mathbf{H}_{20} + [\widehat{\bar{\gamma}}_2^{\intercal}\mathbf{H}_{21}]_+ - [\bar{\gamma}_2^{\intercal}\mathbf{H}_{21}]_+\right| \\
&\leq \left\{o_{\mathbb{P}}(1) + \sup_{\mathbf{H}_{20}}\|\mathbf{H}_{20}\|_2\|\widehat{\boldsymbol{\beta}}_{22} - \bar{\boldsymbol{\beta}}_{22}\|_2 + \sup_{\mathbf{H}_{21}}\|\mathbf{H}_{21}\|_2\right\} O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Next, we can write

$$\omega_1(\mathbf{H}_1, A_1; \widehat{\boldsymbol{\Theta}}_1) = I\left\{A_1 = d_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)\right\}\left\{\frac{A_1}{\pi_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)} + \frac{1 - A_1}{1 - \pi_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)}\right\}.$$

By Lemma B.3.5 (b) it follows that

$$\begin{aligned}
\mathbb{P}_N\left[I\left\{A_1 = d_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)\right\} - I\left\{A_1 = d_1\left(\mathbf{H}_1; \bar{\boldsymbol{\xi}}_1\right)\right\}\right] &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right), \\
\mathbb{P}_N\left[\frac{A_1}{\pi_1(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1)} - \frac{A_1}{\pi_1\left(\mathbf{H}_1; \bar{\boldsymbol{\xi}}_1\right)}\right] &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right), \\
\mathbb{P}_N\left[\frac{1 - A_1}{1 - \pi_1(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1)} - \frac{1 - A_1}{1 - \pi_1\left(\mathbf{H}_1; \bar{\boldsymbol{\xi}}_1\right)}\right] &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Using the above and Lemma B.3.1 we get

$$\begin{aligned}
\hat{\beta}_{21}\mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1)\right\} - \bar{\beta}_{21}\mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\right\} &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right), \\
\mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1)\right\} - \mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\right\} &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

From the above we get

$$\mathbb{P}_N\left\{\mathcal{E}_{\widehat{\boldsymbol{\Theta}}} - \mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right\} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) o_{\mathbb{P}}(1).$$

II) To show the relevant result, we first recall the definition of $\nu_{\mathsf{SSL}_{\mathsf{DR}}}$ from Theorem 2.6.9 and show that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i; \widehat{\boldsymbol{\Theta}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i; \bar{\boldsymbol{\Theta}}) \\
+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\left[\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i; \bar{\boldsymbol{\Theta}})\right]\right)^{\intercal}\boldsymbol{\psi}^{\theta}(\mathbf{L}_i) \qquad \text{(B.10)} \\
+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\xi}}\mathbb{E}\left[\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i; \bar{\boldsymbol{\Theta}})\right]\right)^{\intercal}\boldsymbol{\psi}^{\xi}(\mathbf{L}_i) + o_{\mathbb{P}}(1).
\end{aligned}$$

We start expanding $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\widehat{\boldsymbol{\Theta}})$ as

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\widehat{\boldsymbol{\Theta}})\\&=\mathbb{G}_n\left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right\}+\mathbb{G}_n\left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}})-\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right\}+\sqrt{n}\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}},\end{aligned}$$

we next consider the limit of each term above.

1) Using a Taylor series expansion on $\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}$ we get

$$\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}=\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}+\left(\widehat{\boldsymbol{\Theta}}-\bar{\boldsymbol{\Theta}}\right)^{\intercal}\frac{\partial}{\partial\boldsymbol{\Theta}}\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}+O_{\mathbb{P}}\left(n^{-1}\right),$$

where the remaining terms are of order $O\left\{\left(\widehat{\boldsymbol{\Theta}}-\bar{\boldsymbol{\Theta}}\right)^2\right\}$ which by Theorems 2.6.4 & 2.6.6 are $O_{\mathbb{P}}\left(n^{-1}\right)$. Next note that from (B.6) it follows that $\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}=0$, and thus letting $g_2(\boldsymbol{\Theta})=\int\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})d\mathbb{P}_{\mathbf{L}}$ we have

$$\sqrt{n}g_2(\widehat{\boldsymbol{\Theta}})=\sqrt{n}(\widehat{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}})^{\intercal}\frac{\partial}{\partial\boldsymbol{\Theta}}g_2(\boldsymbol{\Theta})\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}+\sqrt{n}(\widehat{\boldsymbol{\xi}}-\bar{\boldsymbol{\xi}})^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}g_2(\boldsymbol{\Theta})\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}+o_{\mathbb{P}}(1).$$

We can write

$$\sqrt{n}g_2(\widehat{\boldsymbol{\Theta}})=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}^{\theta}(\mathbf{L}_i)^{\intercal}\frac{\partial}{\partial\boldsymbol{\theta}}g_2(\boldsymbol{\Theta})\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}^{\xi}(\mathbf{L}_i)^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}g_2(\boldsymbol{\Theta})\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}+o_{\mathbb{P}}(1).$$

2) We next show

$$\mathbb{G}_n\left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}})-\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right\}=o_{\mathbb{P}}(1),$$

define the class

$$\ell_t=\{I(\mathbf{H}^{\intercal}\boldsymbol{\gamma}_t\geq 0):\mathcal{H}_{t1},\boldsymbol{\gamma}\in\mathbb{R}^{q_t}\},\ t=1,2$$

and the collection of half spaces $\mathcal{C}_\ell\equiv\{\mathbf{H}_t\in\mathbb{R}^{q_t}:\mathbf{H}^{\intercal}\boldsymbol{\gamma}_t\geq 0,\boldsymbol{\gamma}\in\mathbb{R}^{q_t},t\in\{1,2\}\}$, by [96] $\mathcal{C}_\ell$ is a VC class of VC dimension q_t+1, next by [97] we have that as $\mathcal{C}_\ell$ is a VC-class ℓ_t is a class of the same index. Finally, by Theorem 2.6.7 we have that ℓ_t is a Donsker class.

$$\begin{aligned}f_{\boldsymbol{\Theta}}(\mathbf{L}_i)=&\omega_1(\check{\mathbf{H}}_{1i},A_{1i};\boldsymbol{\Theta}_1)(1+\beta_{21})\left\{Y_{2i}-\bar{\mu}_2^{v}(\vec{\mathbf{U}}_i)\right\}+\omega_2(\check{\mathbf{H}}_{2i},A_{2i};\boldsymbol{\Theta}_2)Y_{3i}-\bar{\mu}_{3\omega_2}(\vec{\mathbf{U}}_i)\\&-\beta_{21}\left\{\omega_2(\check{\mathbf{H}}_{2i},A_{2i};\boldsymbol{\Theta}_2)Y_{2i}-\bar{\mu}_{2\omega_2}(\vec{\mathbf{U}}_i)\right\}-Q_{2-}^{o}(\mathbf{H}_{2i};\boldsymbol{\theta}_2)\left\{\omega_2(\check{\mathbf{H}}_{2i},A_{2i};\boldsymbol{\Theta}_2)-\bar{\mu}_{\omega_2}(\vec{\mathbf{U}}_i)\right\},\end{aligned}$$

we define the class of functions $\mathcal{C}_2 = \{f_{\boldsymbol{\Theta}}(\mathbf{L}) | \boldsymbol{\Theta} \in \mathcal{S}(\delta)\}$.

i) By Assumptions 2.6.3, 2.6.7 and Theorem 19.5 in [61], $\mathcal{W}_t,\ \mathcal{Q}_t, t = 1, 2$ are a $\mathbb{P}$-Donsker class. Additionally, the terms in the $\omega_t(\mathbf{H}_t, A_t; \boldsymbol{\Theta}_t)$ functions of the form $\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t I(\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t > 0)$ constitute a $\mathbb{P}$-Donsker class, as $\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t$ is linear in $\boldsymbol{\gamma}_t$ and $I(\mathbf{H}_{t1}^\intercal \boldsymbol{\gamma}_t > 0)$ is $\mathbb{P}$-Donsker. Thus it follows that $\mathcal{C}_2$ is a $\mathbb{P}$-Donsker class.

ii) We estimate $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ for (2.4) with their maximum likelihood estimators, $\widehat{\boldsymbol{\xi}}_1, \widehat{\boldsymbol{\xi}}_2$, solving $\mathbb{P}_n [S_t(\boldsymbol{\xi}_t)] = \mathbf{0}, t = 1, 2$, by Assumption 2.6.7 and Theorem 5.9 in [61] $\widehat{\boldsymbol{\xi}}_t \xrightarrow{p} \bar{\boldsymbol{\xi}}_t, t = 1, 2$. Next, by Theorems 2.6.4, 2.6.6, under Assumptions 2.6.1, 2.6.2, $\widehat{\boldsymbol{\theta}}_t \xrightarrow{p} \bar{\boldsymbol{\theta}}_t, t = 1, 2$. Thus $\mathbb{P}\left(\widehat{\boldsymbol{\Theta}} \in \mathcal{S}(\delta)\right) \xrightarrow{p} 1$, $\forall \delta$. Therefore, we have $\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}) \in \mathcal{C}_2$ with high probability.

iii) We then show $\int \left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}) - \nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \longrightarrow 0$. Using simple algebra for a large enough constant c we have

$$
\begin{aligned}
&\int \left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}) - \nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \\
\leq & c \sup_{Y_2, \vec{\mathbf{U}}} \left\{Y_2 - \bar{\mu}_2^v(\vec{\mathbf{U}})\right\}^2 \\
& \times \sup_{\check{\mathbf{H}}_1, A_1} \left\{(1 + \hat{\beta}_{21})\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1) - (1 + \bar{\beta}_{21})\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\right\}^2 \\
& + c \sup_{Y_3} Y_3^2 \sup_{\check{\mathbf{H}}_2, A_2} \left\{\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_2) - \omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}_2)\right\}^2 \\
& + c \sup_{Y_2} Y_2^2 \sup_{\check{\mathbf{H}}_2, A_2} \left\{\hat{\beta}_{21}\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_2) - \bar{\beta}_{21}\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}_2)\right\}^2 \\
& + c \sup_{\vec{\mathbf{U}}} \bar{\mu}_{2\omega_3}(\vec{\mathbf{U}})^2 \left(\hat{\beta}_{21} - \bar{\beta}_{21}\right)^2 \\
& + c \sup_{\check{\mathbf{H}}_2, A_2} \left\{Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_2) - Q_{2-}^o(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}_2)\right\}^2 \\
& + c \sup_{\vec{\mathbf{U}}} \bar{\mu}_{2\omega_2}(\vec{\mathbf{U}})^2 \sup_{\mathbf{H}_2} \left\{Q_{2-}^o(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2) - Q_{2-}^o(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\right\}^2 \\
& \xrightarrow{p} 0
\end{aligned}
$$

where we use $(a-b)^2, (a+b)^2 \leq 2a^2 + 2b^2\ \forall a, b \in \mathbb{R}$, boundedness of $\bar{\boldsymbol{\Theta}}$ and covariates by Assumptions 2.6.1, 2.6.2 to bound all supremum quantities.

By Theorems 2.6.4 and 2.6.6 we have $\widehat{\boldsymbol{\theta}}_2 - \bar{\boldsymbol{\theta}}_2 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, $\widehat{\boldsymbol{\theta}}_1 - \bar{\boldsymbol{\theta}}_1 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, also from Lemma B.3.5 (a) it follows that

$$\begin{aligned}
&\sup_{\mathbf{H}_2}\left\{Q^o_{2-}(\mathbf{H}_2;\widehat{\boldsymbol{\theta}}_2) - Q^o_{2-}(\mathbf{H}_2;\bar{\boldsymbol{\theta}}_2)\right\}^2 \\
&\leq 2\sup_{\mathbf{H}_{20}}\|\mathbf{H}_{20}\|_2^2\|\widehat{\boldsymbol{\beta}}_{22} - \bar{\boldsymbol{\beta}}_{22}\|_2^2 + 2\sup_{\mathbf{H}_{21}}\|\mathbf{H}_{21}\|_2^2\|\widehat{\boldsymbol{\gamma}}_{22} - \bar{\boldsymbol{\gamma}}_{22}\|_2 \\
&= O_{\mathbb{P}}\left(n^{-1}\right).
\end{aligned}$$

Next, we can write

$$\begin{aligned}
\omega_1(\mathbf{H}_1, A_1; \widehat{\boldsymbol{\Theta}}_1) =& I\left\{A_1 = d_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)\right\}\left\{\frac{A_1}{\pi_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)} + \frac{1-A_1}{1-\pi_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1\right)}\right\} \\
\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_1) =& \omega_1(\mathbf{H}_1, A_1; \widehat{\boldsymbol{\Theta}}_1) I\left\{A_2 = d_2\left(\mathbf{H}_2; \widehat{\boldsymbol{\xi}}_2\right)\right\}\left\{\frac{A_2}{\pi_2\left(\check{\mathbf{H}}_2; \widehat{\boldsymbol{\xi}}_2\right)} + \frac{1-A_2}{2-\pi_2\left(\check{\mathbf{H}}_2; \widehat{\boldsymbol{\xi}}_2\right)}\right\}.
\end{aligned}$$

By Lemma B.3.5 (b) it follows that

$$\begin{aligned}
\sup_{\mathbf{H}_1, \mathbf{a}_1}\left|I(\widehat{d}_1 = A_1) - I(\bar{d}_1 = A_1)\right| =& o_{\mathbb{P}}(1), \\
\sup_{\mathbf{H}_2, \mathbf{a}_2}\left|I(\widehat{d}_1 = A_1)I(A_2 = \widehat{d}_2) - I(\bar{d}_1 = A_1)I(\bar{d}_2 = A_2)\right| =& o_{\mathbb{P}}(1), \\
\sup_{\mathbf{H}_1}\left|\frac{1}{\pi_1(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_1)} - \frac{1}{\pi_1(\mathbf{H}_1; \bar{\boldsymbol{\xi}}_1)}\right| =& O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Using the above and Lemma B.3.1 we get

$$\begin{aligned}
\sup_{\check{\mathbf{H}}_1, A_1}\left\{(1+\hat{\beta}_{21})\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1) - (1+\bar{\beta}_{21})\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\right\}^2 &= o_{\mathbb{P}}(1), \\
\sup_{\check{\mathbf{H}}_2, A_2}\left\{(1+\hat{\beta}_{21})\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_2) - (1+\bar{\beta}_{21})\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}_2)\right\}^2 &= o_{\mathbb{P}}(1), \\
\sup_{\check{\mathbf{H}}_2, A_2}\left\{Q^o_{2-}(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_2) - Q^o_{2-}(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_2)\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}_2)\right\}^2 &= o_{\mathbb{P}}(1), \\
\sup_{\check{\mathbf{H}}_2, A_2}\left\{\hat{\beta}_{21}\omega_2(\check{\mathbf{H}}_2, A_2; \widehat{\boldsymbol{\Theta}}_2) - \bar{\beta}_{21}\omega_2(\check{\mathbf{H}}_2, A_2; \bar{\boldsymbol{\Theta}}_2)\right\}^2 &= o_{\mathbb{P}}(1).
\end{aligned}$$

which gives us $\int\left\{\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}) - \nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \xrightarrow{p} 0$.

Therefore we have i) $\mathbb{P}\left(\widehat{\boldsymbol{\Theta}} \in \mathcal{S}(\delta)\right) \to 1$, $\forall\delta$, ii) $\mathcal{C}_2$ is a $\mathbb{P}$-Donsker class, and

iii) $\int \left(\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\bar{\mathbf{\Theta}}}) - \nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right)^2 d\mathbb{P}_{\mathbf{L}} \to 0$. By Theorem 2.1 in [98]

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\left(\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\widehat{\bar{\mathbf{\Theta}}}) - \mathbb{E}_{\mathbb{S}}[\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\widehat{\bar{\mathbf{\Theta}}})]\right) - \left(\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\mathbf{\Theta}}) - \mathbb{E}_{\mathbb{S}}[\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})]\right)\right\} = o_{\mathbb{P}}(1).$$

by 1), 2) and noting that $\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\mathbf{\Theta}})$ has mean zero we obtain the result in (B.10).

We next re-write $\sqrt{n}\mathbb{P}_N\left[\mathcal{E}_{\bar{\mathbf{\Theta}}}\right]$ by expressing the estimated imputation functions in $\mathcal{E}_{\bar{\mathbf{\Theta}}}$ in terms of the labeled sample $\mathcal{L}$. Letting

$$\hat{C}^{(1)}_{n,N} = \frac{(1+\bar{\beta}_{21})\mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1, A_1;\bar{\mathbf{\Theta}})\right\}}{(1+\hat{\beta}_{21})\mathbb{P}_n\left\{\omega_1(\check{\mathbf{H}}_1, A_1;\widehat{\bar{\mathbf{\Theta}}})\right\}},\ \hat{C}^{(2)}_{n,N} = \frac{\mathbb{P}_N\left\{Q^o_{2-}(\mathbf{H}_2;\bar{\boldsymbol{\theta}}_2)\right\}}{\mathbb{P}_n\left\{Q^o_{2-}(\mathbf{H}_2;\widehat{\boldsymbol{\theta}}_2)\right\}},$$

we can write:

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\omega_1(\check{\mathbf{H}}_{1j}, A_{1j}, \bar{\mathbf{\Theta}})(1+\bar{\beta}_{21})\left\{\widehat{\mu}^v_2(\vec{\mathbf{U}}_j) - \bar{\mu}^v_2(\vec{\mathbf{U}}_j)\right\}\\
=&\frac{1}{N}\sum_{j=1}^{N}\omega_1(\check{\mathbf{H}}_{1j}, A_{1j}, \bar{\mathbf{\Theta}})(1+\bar{\beta}_{21})\left\{\frac{1}{K}\sum_{k=1}^{K}\widehat{m}^{(-k)}_2(\vec{\mathbf{U}}_j) + \widehat{\eta}^v_2 - m_2(\vec{\mathbf{U}}_j) - \eta^v_2\right\}\\
=&(1+\bar{\beta}_{21})\frac{1}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}\omega_1(\check{\mathbf{H}}_{1j}, A_{1j}, \bar{\mathbf{\Theta}})\hat{\Delta}^{(-k)}_2(\vec{\mathbf{U}}_j) + (\widehat{\eta}^v_2 - \eta^v_2)\frac{1}{N}\sum_{j=1}^{N}\omega_1(\check{\mathbf{H}}_{1j}, A_{1j}, \bar{\mathbf{\Theta}})(1+\bar{\beta}_{21}),
\end{aligned}$$

where the first step follows from constrains shown in (2.6) and we simply regroup terms in the second step.

Next note that we can use Lemma B.3.2 to replace

$$\mathbb{P}_N\left[(1+\bar{\beta}_{21})\omega_1(\check{\mathbf{H}}_1, A_1, \bar{\mathbf{\Theta}})\hat{\Delta}^{(-k)}_2(\vec{\mathbf{U}}_j)\right] \quad \text{by} \quad \mathbb{E}_{\mathcal{L}}\left[(1+\bar{\beta}_{21})\omega_1(\check{\mathbf{H}}_1, A_1, \bar{\mathbf{\Theta}})\hat{\Delta}^{(-k)}_2(\vec{\mathbf{U}}_j)\right] + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right),$$

using $\mathbb{E}_{\mathcal{L}}[\cdot]$ to denote expectation with respect to $\mathcal{L}$. Additionally, using (2.6) and the definition

of $\bar{\mu}_2^v(\vec{\mathbf{U}})$ for the second term we get:

$$
\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}\omega_1(\check{\mathbf{H}}_{1j},A_{1j};\bar{\boldsymbol{\Theta}})(1+\bar{\beta}_{21})\left\{\widehat{\mu}_2^v(\vec{\mathbf{U}}_j)-\bar{\mu}_2^v(\vec{\mathbf{U}}_j)\right\}\\
=&\mathbb{E}_{\mathcal{L}}\left[\frac{1}{K}\sum_{k=1}^{K}(1+\bar{\beta}_{21})\omega_1(\check{\mathbf{H}}_1,A_1,\bar{\boldsymbol{\Theta}})\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})\right]+O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)\\
&-\hat{C}_{n,N}^{(1)}\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}(1+\hat{\beta}_{21})\omega_1(\check{\mathbf{H}}_{1i},A_{1i};\widehat{\boldsymbol{\Theta}})\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}}_i)\\
&+\hat{C}_{n,N}^{(1)}(1+\hat{\beta}_{21})\frac{1}{n}\sum_{i=1}^{n}\omega_1(\check{\mathbf{H}}_{1i},A_{1i};\widehat{\boldsymbol{\Theta}})\left\{Y_{2i}-\bar{\mu}_2^v(\vec{\mathbf{U}}_i)\right\}\\
=&\left\{1+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\right\}(1+\hat{\beta}_{21})\frac{1}{n}\sum_{i=1}^{n}\omega_1(\check{\mathbf{H}}_{1i},A_{1i};\widehat{\boldsymbol{\Theta}})\left\{Y_{2i}-\bar{\mu}_2^v(\vec{\mathbf{U}}_i)\right\}+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}c_{n_K^-}\right),
\end{aligned}
$$

where the last step follows from Assumption 2.6.8 and Lemma B.3.4 choosing f to be the constant function 1, setting $\hat{\Delta}_k(\vec{\mathbf{U}})=\hat{\Delta}_2^{(\text{-}k)}(\vec{\mathbf{U}})$, $\hat{l}(\check{\mathbf{H}}_1)=A_1 I(\mathbf{H}_{11}^{\intercal}\widehat{\boldsymbol{\gamma}}_1>0)$, and $\hat{\pi}(\check{\mathbf{H}}_1)=\pi_1(\check{\mathbf{H}}_1;\widehat{\boldsymbol{\xi}}_1)$ and with $\hat{C}_{n,N}=\hat{C}_{n,N}^{(1)}$ -which satisfies $\hat{C}_{n,N}^{(1)}=1+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ by Lemma B.3.5 (c).

Using similar arguments we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N}Q_{2-}^{o}(\mathbf{H}_{2j};\bar{\boldsymbol{\theta}}_2)\{\widehat{\mu}_{\omega_2}(\vec{\mathbf{U}}_j)-\bar{\mu}_{\omega_2}(\vec{\mathbf{U}}_j)\}\\
&=\frac{1}{N}\sum_{j=1}^{N}Q_{2-}^{o}(\mathbf{H}_{2j};\bar{\boldsymbol{\theta}}_2)\left\{\frac{1}{K}\sum_{k=1}^{K}\widehat{m}_{\omega_2}^{(\text{-}k)}(\vec{\mathbf{U}}_j)+\widehat{\eta}_{\omega_2}^v-m_{\omega_2}(\vec{\mathbf{U}}_j)-\eta_{\omega_2}^v\right\}\\
&=\frac{1}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}Q_{2-}^{o}(\mathbf{H}_{2j};\bar{\boldsymbol{\theta}}_2)\hat{\Delta}_{\omega_2 k}(\vec{\mathbf{U}}_j)+\left(\widehat{\eta}_{\omega_2}^v-\eta_{\omega_2}^v\right)\frac{1}{N}\sum_{j=1}^{N}Q_{2-}^{o}(\mathbf{H}_{2j};\bar{\boldsymbol{\theta}}_2)\\
&=\mathbb{E}_{\mathcal{L}}\left[\frac{1}{K}\sum_{k=1}^{K}Q_{2-}^{o}(\mathbf{H}_2;\bar{\boldsymbol{\theta}}_2)\hat{\Delta}_{\omega_2 k}(\vec{\mathbf{U}})\right]-\hat{C}_{n,N}^{(2)}\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}Q_{2-}^{o}(\mathbf{H}_{2i};\bar{\boldsymbol{\theta}}_2)\hat{\Delta}_{\omega_2 k}(\vec{\mathbf{U}}_i)\\
&+\hat{C}_{n,N}^{(2)}\frac{1}{n}\sum_{i=1}^{n}Q_{2-}^{o}(\mathbf{H}_{2i};\widehat{\boldsymbol{\theta}}_2)\{\omega_2(\check{\mathbf{H}}_{2i},A_{2i};\widehat{\boldsymbol{\Theta}})-\bar{\mu}_{\omega_2}(\vec{\mathbf{U}}_i)\}+O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)\\
&=\left\{1+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\right\}\frac{1}{n}\sum_{i=1}^{n}Q_{2-}^{o}(\mathbf{H}_{2i};\widehat{\boldsymbol{\theta}}_2)\{\omega_2(\check{\mathbf{H}}_{2i},A_{2i};\widehat{\boldsymbol{\Theta}})-\bar{\mu}_{\omega_2}(\vec{\mathbf{U}}_i)\}+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}c_{n_K^-}\right),
\end{aligned}
$$

and for $t = 2, 3$

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N}\left\{\widehat{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}}_j) - \bar{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}}_j)\right\} =& \frac{1}{KN}\sum_{j=1}^{N}\left\{\sum_{k=1}^{K}\widehat{m}^{(\text{-}k)}_{t\omega_2}(\vec{\mathbf{U}}_j) + \widehat{\eta}^{v}_{t\omega_2} - m_{t\omega_2}(\vec{\mathbf{U}}_j) - \eta^{v}_{t\omega_2}\right\} \\
=& \frac{1}{KN}\sum_{j=1}^{N}\sum_{k=1}^{K}\hat{\Delta}_{3\omega_2 k}(\vec{\mathbf{U}}_j) + (\widehat{\eta}^{v}_{t\omega_2} - \eta^{v}_{t\omega_2}) \\
=& \mathbb{E}_{\mathcal{L}}\left[\frac{1}{K}\sum_{k=1}^{K}\hat{\Delta}_{3\omega_2 k}(\vec{\mathbf{U}})\right] - \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\hat{\Delta}_{3\omega_2 k}(\vec{\mathbf{U}}_i) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\omega_2(\check{\mathbf{H}}_{2i}, A_{2i}; \widehat{\boldsymbol{\Theta}})Y_{ti} - \bar{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}}_i) + O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right) \\
=& \frac{1}{n}\sum_{i=1}^{n}\omega_2(\check{\mathbf{H}}_{2i}, A_{2i}; \widehat{\boldsymbol{\Theta}})Y_{ti} - \bar{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}}_i) + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}c_{n^{-}_{K}}\right),
\end{aligned}$$

finally by Theorem 2.6.4, $\hat{\beta}_{21} - \bar{\beta}_{21} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$.

Therefore, recalling the definition of $\nu_{\mathsf{SSL}_{\mathsf{DR}}}$ from Theorem 2.6.9, using the derivations above, we can write

$$\begin{aligned}
\frac{\sqrt{n}}{N}\sum_{j=1}^{N}\mathcal{E}_{\bar{\boldsymbol{\Theta}}}(\vec{\mathbf{U}}_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}&\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}_i; \widehat{\boldsymbol{\Theta}}) \\
&+ O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\boldsymbol{\Theta}}_1)(1+\hat{\beta}_{21})\left\{Y_2 - \bar{\mu}^{v}_{2}(\vec{\mathbf{U}})\right\} \\
&- O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\beta}_{21}\left\{\omega_2(\check{\mathbf{H}}_2, A_2, \widehat{\boldsymbol{\Theta}}_2)Y_2 - \bar{\mu}_{2\omega_2}(\vec{\mathbf{U}})\right\} \\
&- O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Q^{o}_{2-}(\mathbf{H}_2; \widehat{\boldsymbol{\theta}}_2)\left\{\omega_2(\check{\mathbf{H}}_2, A_2, \bar{\boldsymbol{\Theta}}_2) - \bar{\mu}_{\omega_2}(\vec{\mathbf{U}})\right\} \\
&+ O_{\mathbb{P}}\left(c_{n^{-}_{K}}\right).
\end{aligned} \tag{B.11}$$

Using result (B.10) we know $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}_i; \widehat{\boldsymbol{\Theta}}) = O_{\mathbb{P}}(1)$, therefore the second, third and fourth terms in (B.11) are $o_{\mathbb{P}}(1)$. Using (B.10) again for the first term in (B.11) we get our

required result:

$$\begin{aligned}\sqrt{n}\mathbb{P}_N\left[\mathcal{E}_{\bar{\boldsymbol{\Theta}}}\right] = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}}) \\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\left[\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}})\right]\right)^{\mathsf{T}}\boldsymbol{\psi}^{\theta}(\mathbf{L}_i) \\ &+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\boldsymbol{\xi}}\mathbb{E}\left[\nu_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\boldsymbol{\Theta}})\right]\right)^{\mathsf{T}}\boldsymbol{\psi}^{\xi}(\mathbf{L}_i)+o_{\mathbb{P}}(1).\end{aligned}$$

□

Proof of Proposition 2.6.2. Recall the definition of $\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})$ in (2.5), using (B.6) we have $\mathbb{E}\left[\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\bar{\boldsymbol{\Theta}},\bar{\mu})\right] = \mathbb{E}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right]$, therefore

$$\mathrm{Bias}\left\{\bar{V},\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}\left(\mathbf{L};\bar{\boldsymbol{\Theta}}\right)\right\} = \mathrm{Bias}\left\{\bar{V},\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\bar{\boldsymbol{\Theta}},\bar{\mu})\right\}.$$

Therefore, by Lemma B.3.6 we have

$$\begin{aligned}&\mathrm{Bias}\left\{\bar{V},\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\bar{\boldsymbol{\Theta}},\bar{\mu})\right\} \\ \leq&\sqrt{\sup_{\check{\mathbf{H}}_1}|\{1-\pi_1(\check{\mathbf{H}}_1;\bar{\boldsymbol{\xi}}_1)\}^{-1}|}\sqrt{\|\pi_1(\check{\mathbf{H}}_1;\bar{\boldsymbol{\xi}}_1)-\pi_1(\check{\mathbf{H}}_1)\|_{L_2(\mathbb{P})}}\sqrt{\|Q_1^o(\check{\mathbf{H}}_1;\bar{\boldsymbol{\theta}}_1)-Q_1^o(\check{\mathbf{H}}_1)\|_{L_2(\mathbb{P})}} \\ +&\sqrt{\sup_{\check{\mathbf{H}}_2}\left|\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1;\bar{\boldsymbol{\xi}}_1)}+\frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1;\bar{\boldsymbol{\xi}}_1)}\right\}\{1-\pi_1(\check{\mathbf{H}}_1;\bar{\boldsymbol{\xi}}_1)\}^{-1}\{1-\pi_2(\check{\mathbf{H}}_2;\bar{\boldsymbol{\xi}}_2)\}^{-1}\right|} \\ \times&\sqrt{\|\pi_2(\check{\mathbf{H}}_2;\bar{\boldsymbol{\xi}}_2)-\pi_2(\check{\mathbf{H}}_2)\|_{L_2(\mathbb{P})}}\sqrt{\|Q_2^o(\check{\mathbf{H}}_2;\bar{\boldsymbol{\theta}}_2)-Q_2^o(\check{\mathbf{H}}_2)\|_{L_2(\mathbb{P})}}.\end{aligned}$$

Next using Theorem 2.6.9

$$\sqrt{n}\left\{\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}-\bar{V}\right\}+\sqrt{n}\mathrm{Bias}\left\{\bar{V},\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\bar{\boldsymbol{\Theta}},\bar{\mu})\right\}\xrightarrow{d}N\left(0,\sigma^2_{\mathsf{SSL}_{\mathsf{DR}}}\right), \tag{B.12}$$

if either (2.1) or (2.4) are correct then $\mathrm{Bias}\left\{\bar{V},\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\bar{\boldsymbol{\Theta}},\bar{\mu})\right\} = o_{\mathbb{P}}(1)$, multiplying (B.12) by $n^{-\frac{1}{2}}$ we have

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}-\bar{V}\xrightarrow{\mathbb{P}}0,$$

which is the required result for Proposition 2.6.2 (a).

Next, if $\sqrt{\|\pi_t(\check{\mathbf{H}}_t;\widehat{\boldsymbol{\xi}}_t)-\pi_t(\check{\mathbf{H}}_t)\|_{L_2(\mathbb{P})}}\sqrt{\|Q_t^o(\check{\mathbf{H}}_t;\widehat{\boldsymbol{\theta}}_t)-Q_t^o(\check{\mathbf{H}}_t)\|_{L_2(\mathbb{P})}} = O_{\mathbb{P}}\left(n^{-1}\right)$ for $t=1,2$

then Bias $\left\{\bar{V}, \mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \bar{\boldsymbol{\Theta}}, \bar{\mu})\right\} = O_{\mathbb{P}}\left(n^{-1}\right)$ and from (B.12) we get

$$\sqrt{n}\left\{\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \bar{V}\right\} \xrightarrow{d} N\left(0, \sigma^2_{\mathsf{SSL}_{\mathsf{DR}}}\right),$$

which is the required result for Proposition 2.6.2 (b). □

Before proving Proposition 2.6.2, we introduce a useful definition and state the necessary assumption to prove the result. Let $\psi^{\xi}_{\mathsf{SUP}}(\mathbf{L}; \bar{\boldsymbol{\xi}})$ and $\psi^{\xi}_{\mathsf{SSL}}(\mathbf{L}; \bar{\boldsymbol{\xi}})$ be the supervised and SSL influence functions respectively for $\boldsymbol{\xi}$, then we define

$$\begin{aligned}
\mathcal{E}^{v}(\vec{\mathbf{U}}) =& \mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}; \bar{\boldsymbol{\Theta}}, \bar{\mu}) - \mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right] + \mathcal{E}^{\theta}(\vec{\mathbf{U}})^{\intercal} \frac{\partial}{\partial \boldsymbol{\theta}} \int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}}) d\mathbb{P}_{\mathbf{L}} \Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}} \\
&+ \mathcal{E}^{\xi}(\vec{\mathbf{U}})^{\intercal} \frac{\partial}{\partial \boldsymbol{\xi}} \int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}}) d\mathbb{P}_{\mathbf{L}} \Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}, \\
\mathcal{E}^{\xi}(\vec{\mathbf{U}}) =& \psi^{\xi}_{\mathsf{SUP}}(\mathbf{L}; \bar{\boldsymbol{\xi}}) - \psi^{\xi}_{\mathsf{SSL}}(\mathbf{L}; \bar{\boldsymbol{\xi}}).
\end{aligned}$$

We need to ensure that the imputation models $\bar{\mu}^{v}_{2}(\vec{\mathbf{U}})$, $\bar{\mu}^{v}_{\omega_2}(\vec{\mathbf{U}})$, $\bar{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}})$, $t = 2, 3$ used in the SSL value function estimator $V_{\mathsf{SSL}_{\mathsf{DR}}}$ are unbiased when multiplied by several functions. For example, we need additional constraints of the type:

$$\begin{aligned}
\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1) Q^{o}_{2-}(\mathbf{H}_2; \bar{\boldsymbol{\theta}}_1)\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\}\right] =& \mathbf{0}, \\
\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)^2\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\}\right] =& \mathbf{0}, \\
\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1; \bar{\boldsymbol{\Theta}}_1)^2\{Y_2 - \bar{\mu}_2(\vec{\mathbf{U}})\}\right] =& \mathbf{0},
\end{aligned}$$

so the imputation models are unbiased in expectation when multiplied by every term and cross-product of terms in $\psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})$, $\mathcal{E}^{v}(\vec{\mathbf{U}})$. These constraints can be summarized in the following Assumption.

Assumption B.2.5. *Imputation models* $\bar{\mu}^{v}_{2}(\vec{\mathbf{U}})$, $\bar{\mu}^{v}_{\omega_2}(\vec{\mathbf{U}})$, $\bar{\mu}^{v}_{t\omega_2}(\vec{\mathbf{U}})$, $t = 2, 3$ *satisfy*

$$\mathbb{E}\left[\left\{\mathcal{E}^{v}(\vec{\mathbf{U}}) - \psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right\} \mathcal{E}^{v}(\vec{\mathbf{U}})\right] = 0.$$

Proof of Proposition 2.6.2. From Theorem B.4.2 in Appendix B.4.1 we have that the influence

function for the fully-supervised value function estimator (2.5) is:

$$\begin{aligned}\psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) =&\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) - \mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right] + \boldsymbol{\psi}^{\theta}_{\mathsf{SUP}}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\theta}}\int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}\\ &+\boldsymbol{\psi}^{\xi}_{\mathsf{SUP}}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}\int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}\Big|_{\boldsymbol{\Theta}=\bar{\boldsymbol{\Theta}}}.\end{aligned}$$

Next, as we estimate $\boldsymbol{\xi}$ with a semi-supervised approach such that $\boldsymbol{\psi}^{\xi}_{\mathsf{SSL}}(\mathbf{L};\bar{\boldsymbol{\xi}}) = \boldsymbol{\psi}^{\xi}_{\mathsf{SUP}}(\mathbf{L};\bar{\boldsymbol{\xi}}) - \mathcal{E}^{\xi}(\vec{\mathbf{U}})$, simple algebra can be used to show that

$$\psi^v_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) = \psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) - \mathcal{E}^v(\vec{\mathbf{U}}).$$

Using the above we can write

$$\begin{aligned}\sigma^2_{\mathsf{SSL}_{\mathsf{DR}}} = \mathbb{E}\left[\psi^v_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})^2\right] =&\mathbb{E}\left[\left\{\psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}}) - \mathcal{E}^v(\vec{\mathbf{U}})\right\}^2\right]\\ =&\mathbb{E}\left[\psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})^2\right] + \mathbb{E}\left[\mathcal{E}^v(\vec{\mathbf{U}})^2\right]\\ &-2\mathbb{E}\left[\psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\mathcal{E}^v(\vec{\mathbf{U}})\right].\end{aligned}$$

By Assumption B.2.5, we have $\mathbb{E}\left[\left\{\mathcal{E}^v(\vec{\mathbf{U}}) - \psi^v_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})\right\}\mathcal{E}^v(\vec{\mathbf{U}})\right] = 0$, hence

$$\sigma^2_{\mathsf{SSL}_{\mathsf{DR}}} = \sigma^2_{\mathsf{SUP}_{\mathsf{DR}}} - \mathrm{Var}\left[\mathcal{E}^v(\vec{\mathbf{U}})\right].$$

□

Variance Estimation for $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$

As discussed in Remark 2.6.2, to estimate standard errors for $V_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}};\bar{\boldsymbol{\Theta}})$, we will approximate the derivatives of the expectation terms $\frac{\partial}{\partial\boldsymbol{\Theta}}\int \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\boldsymbol{\Theta}})d\mathbb{P}_{\mathbf{L}}$ using kernel smoothing to replace the indicator functions. In particular, let $\mathbb{K}_h(x) = \frac{1}{h}\sigma(x/h)$, with σ defined as in (2.4), we approximate $d_t(\mathbf{H}_t,\boldsymbol{\theta}_t) = I(\mathbf{H}_{t1}^{\intercal}\boldsymbol{\gamma}_t > 0)$ with $\mathbb{K}_h(\mathbf{H}_{t1}^{\intercal}\boldsymbol{\gamma}_t)$ $t = 1, 2$, and define the smoothed propensity score weights as

$$\begin{aligned}\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) \equiv& \frac{A_1\mathbb{K}_h(\mathbf{H}_{11}^{\intercal}\boldsymbol{\gamma}_1)}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)} + \frac{\{1-A_1\}\{1-\mathbb{K}_h(\mathbf{H}_{11}^{\intercal}\boldsymbol{\gamma}_1)\}}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}, \quad \text{and}\\ \tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta}) \equiv& \tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\left[\frac{A_2\mathbb{K}_h(\mathbf{H}_{21}^{\intercal}\boldsymbol{\gamma}_2)}{\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)} + \frac{\{1-A_2\}\{1-\mathbb{K}_h(\mathbf{H}_{21}^{\intercal}\boldsymbol{\gamma}_2)\}}{1-\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)}\right].\end{aligned}$$

For simplicity we'll set $h = 1$, the derivatives are as follows:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta}) = \frac{\partial}{\partial \boldsymbol{\theta}}Q_1^o(\mathbf{H}_1;\boldsymbol{\theta}_1) &+ \left\{\frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\right\}\left[Y_2 - \left\{Q_1^o(\mathbf{H}_1,\boldsymbol{\theta}_1) - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right] \\ &+\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\left[-\frac{\partial}{\partial \boldsymbol{\theta}}Q_1^o(\mathbf{H}_1,\boldsymbol{\theta}_1) + \frac{\partial}{\partial \boldsymbol{\theta}}Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right] \\ &+\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta})\right\}\left[Y_3 - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right] \\ &-\tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta})\frac{\partial}{\partial \boldsymbol{\theta}}Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2),\end{aligned}$$

where

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}}Q_1^o(\mathbf{H}_1;\boldsymbol{\theta}_1) &= [\mathbf{H}_{10}^\intercal, \mathbf{H}_{11}^\intercal I\left(\mathbf{H}_{11}^\intercal\boldsymbol{\gamma}_1 > 0\right), \mathbf{0}^\intercal]^\intercal, \\ \frac{\partial}{\partial \boldsymbol{\theta}}Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2) &= [\mathbf{0}^\intercal, \mathbf{H}_{20}^\intercal, \mathbf{H}_{21}^\intercal I\left(\mathbf{H}_{21}^\intercal\boldsymbol{\gamma}_2 > 0\right)]^\intercal, \\ \frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) &= \left[\mathbf{0}^\intercal, \mathbf{H}_{11}^\intercal\mathbb{K}_h(\mathbf{H}_{11}^\intercal\boldsymbol{\gamma}_1)\{1-\mathbb{K}_h(\mathbf{H}_{11}^\intercal\boldsymbol{\gamma}_1)\}\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)} - \frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}, \mathbf{0}^\intercal\right]^\intercal \\ \frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\left\{\frac{A_2 d_2(\mathbf{H}_2;\boldsymbol{\theta}_2)}{\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)} + \frac{\{1-A_2\}\{1-d_2(\mathbf{H}_2;\boldsymbol{\theta}_2)\}}{1-\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)}\right\} \\ &\quad + \tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\left[\mathbf{0}^\intercal, \mathbf{H}_{21}^\intercal\mathbb{K}_h(\mathbf{H}_{21}^\intercal\boldsymbol{\gamma}_2)(1-\mathbb{K}_h(\mathbf{H}_{21}^\intercal\boldsymbol{\gamma}_2))\left\{\frac{A_2}{\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)} - \frac{1-A_2}{1-\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)}\right\}\right]^\intercal.\end{aligned}$$

Next we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\xi}}\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta}) = &\left\{\frac{\partial}{\partial \boldsymbol{\xi}}\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\right\}\left[Y_2 - \left\{Q_1^o(\mathbf{H}_1,\boldsymbol{\theta}_1) - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right] \\ &+\left\{\frac{\partial}{\partial \boldsymbol{\xi}}\tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta})\right\}\left[Y_3 - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right],\end{aligned}$$

where

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\xi}}\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta}) &= \left[\varpi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)^\intercal, \mathbf{0}^\intercal\right]^\intercal, \\ \frac{\partial}{\partial \boldsymbol{\xi}}\tilde{\omega}_2(\check{\mathbf{H}}_2, A_2, \boldsymbol{\Theta}) &= \left[\tilde{\omega}_1(\check{\mathbf{H}}_1, A_1, \boldsymbol{\Theta})\varpi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)^\intercal, \mathbf{0}^\intercal\right]^\intercal, \\ \varpi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t) &\equiv \check{\mathbf{H}}_{t1}\left\{-d_t(\check{\mathbf{H}}_t,\boldsymbol{\theta}_t)A_t\frac{1-\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)}{\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)} + \{1-d_t(\mathbf{H}_t,\boldsymbol{\theta}_t)\}\{1-A_t\}\frac{\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)}{1-\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)}\right\}.\end{aligned}$$

B.3 Technical Lemmas

We start with a simple Lemma that will save us some algebra:

Lemma B.3.1. *For a fixed ℓ, let $\bar{X} \in \mathbb{R}^\ell$ be a random bounded vector and functions*

$g_1(\bar{X})$, $g_2(\bar{X})$ be measurable functions of $\bar{X}$. Let $\mathbb{S}_n = \{\bar{X}\}_{i=1}^n$ be an i.i.d. sample, and $\hat{g}_1(\cdot)$, $\hat{g}_2(\cdot)$ be the estimators for functions g_1, $g_2 \in \mathbb{R}$ respectively with $\sup_{\bar{X}} |g_1(\bar{X})|$, $\sup_{\bar{X}} |g_2(\bar{X})|$, $\sup_{\bar{X}} |\hat{g}_1(\bar{X})|$, $\sup_{\bar{X}} |\hat{g}_2(\bar{X})| < \kappa$ for fixed $\kappa \in \mathbb{R}$. If $\mathbb{P}_n\{\hat{g}_k - g_k\} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, for $k = 1, 2$, then $\mathbb{P}_n\{\hat{g}_1\hat{g}_2 - g_1g_2\} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$.

Proof of Lemma B.3.1. By definition, $\mathbb{P}_n\{\hat{g}_1\hat{g}_2 - g_1g_2\} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ if and only if for a given any $\epsilon > 0$, $\exists M_\epsilon > 0$ such that
$\mathbb{P}\left(|\mathbb{P}_n\{\hat{g}_1\hat{g}_2 - g_1g_2\}| > M_\epsilon n^{-\frac{1}{2}}\right) \leq \epsilon \; \forall n$. Let $M_\epsilon > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(|\mathbb{P}_n\{g_1g_2 - g_1g_2\}| > M_\epsilon n^{-\frac{1}{2}}\right) \\
=&\mathbb{P}\left(|\mathbb{P}_n\{\hat{g}_1\hat{g}_2 - \hat{g}_1g_2 + \hat{g}_1g_2 - g_1g_2\}| > M_\epsilon n^{-\frac{1}{2}}\right) \\
&\leq \mathbb{P}\left(|\mathbb{P}_n\{\hat{g}_1(\hat{g}_2 - g_2)\}| + |\mathbb{P}_n\{g_2(\hat{g}_1 - g_1)\}| > M_\epsilon n^{-\frac{1}{2}}\right) \\
&\leq \mathbb{P}\left(\sup_{\bar{X}} |\hat{g}_1(\bar{X})||\mathbb{P}_n\{\hat{g}_2 - g_2\}| + \sup_{\bar{X}} |g_2(\bar{X})||\mathbb{P}_n\{\hat{g}_1 - g_1\}| > M_\epsilon n^{-\frac{1}{2}}\right)
\end{aligned}
$$

which follows from bounded functions, the union bound, now since $\mathbb{P}_n\{\hat{g}_k(\bar{X}) - g_k(\bar{X})\} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, $k = 1, 2$, there exists $M_\epsilon > 0$ such that

$$
\mathbb{P}\left(|\mathbb{P}_n\{\hat{g}_2 - g_2\}| > M_\epsilon n^{-\frac{1}{2}}\frac{1}{\kappa}\right) + \mathbb{P}\left(|\mathbb{P}_n\{\hat{g}_1 - g_1\}| > M_\epsilon n^{-\frac{1}{2}}\frac{1}{\kappa}\right) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
$$

□

Lemma B.3.2. *(Lemma (A.1) (a) in [43])*

Let $\bar{X} \in \mathbb{R}^\ell$ be any random vector and $g(\bar{X}) \in \mathbb{R}^\ell$ be any measurable function of $\bar{X}$, with ℓ and d fixed. Let $\mathbb{S}_n = \{\bar{X}\}_{i=1}^n, \mathbb{S}_N = \{\bar{X}\}_{j=1}^N$ be two random samples of n and N i.i.d observations of $\bar{X}$ respectively, such that $\mathbb{S}_n \perp\!\!\!\perp \mathbb{S}_N$. Let $\hat{g}_n(\cdot)$ be any estimator of $g(\cdot)$ estimated with $\mathbb{S}_n$ such that the random sequence: $\hat{T}_n = \sup_{x \in \mathcal{X}} \|\hat{g}_n(\cdot)\| = O_{\mathbb{P}}(1)$, where $\bar{X} \in \mathcal{X} \subseteq \mathbb{R}^\ell$. Further define the following random sequences: $\hat{\boldsymbol{G}}_{n,N} \equiv \frac{1}{N}\sum_{j=1}^N \hat{g}_n(\bar{X}_j)$, and $\bar{\boldsymbol{G}}_n \equiv \mathbb{E}_{\mathbb{S}_N}\left[\hat{G}_{n,N}\right] = \mathbb{E}_{\bar{X}}\left[\hat{g}_n(\bar{X})\right]$, where $\mathbb{E}_{\bar{X}}$ is the expectation with respect to $\bar{X} \in \mathbb{S}_N$. We assume all expectations involved are finite almost surely (a.s.) $\mathbb{S}_n \; \forall n$. Then $G_{n,N} - \bar{G}_n = O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)$.

Proof of lemma B.3.2. The following proof follows similar arguments to [43]. Let $\mathcal{G}_{n,N}$, $\bar{\mathcal{G}}_n$ be the j^{th} element of $\hat{\boldsymbol{G}}_{n,N}$ and $\bar{\boldsymbol{G}}_n$ respectively, with $j \in \{1, \dots, \ell\}$. We show that $\mathcal{G}_{n,N} - \bar{\mathcal{G}}_n = O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right)$, which implies Lemma B.3.2 for any ℓ dimensional $\hat{\boldsymbol{G}}_{n,N}$, $\bar{\boldsymbol{G}}_n$. Denote by $\mathbb{P}_{\mathbb{S}_n}$, $\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_N}$ denote the joint probability distributions of samples $\mathbb{S}_n$ and $\mathbb{S}_n, \mathbb{S}_N$ respectively. Further let $\mathbb{E}_{\mathbb{S}_n}[\cdot]$ denote the expectation with respect to $\mathbb{S}_n$. Since $\mathbb{S}_n \perp\!\!\!\perp \mathbb{S}_N$ using Hoeffding's inequality

$$\mathbb{P}_{\mathbb{S}_N}\left(\left|\hat{\mathcal{G}}_{n,N} - \hat{\mathcal{G}}_n\right| > N^{-\frac{1}{2}}t \middle| \mathbb{S}_n\right) \leq 2\exp\left(-\frac{2N^2t^2}{4N^2\hat{T}_n^2}\right) \text{ a.s. } \mathbb{P}_{\mathbb{S}_n}.$$

Also, as $\mathbb{S}_n \perp\!\!\!\perp \mathbb{S}_N$ we have

$$\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_N}\left[\left|\hat{\mathcal{G}}_{n,N} - \hat{\mathcal{G}}_n\right| > N^{-\frac{1}{2}}t\right] = \mathbb{E}_{\mathbb{S}_n}\left[\mathbb{P}_{\mathbb{S}_N}\left\{\left|\hat{\mathcal{G}}_{n,N} - \hat{\mathcal{G}}_n\right| > N^{-\frac{1}{2}}t \middle| \mathbb{S}_n\right\}\right].$$

Next, we have that $\hat{T}_n = \sup_{x\in\mathcal{X}} \|\hat{g}_n(\cdot)\| = O_{\mathbb{P}}(1)$ and is non-negative, thus $\forall \epsilon > 0$ $\exists\, \delta(\epsilon) > 0$ such that
$\mathbb{P}_{\mathbb{S}_n}\left(\hat{T}_n > \delta(\epsilon)\right) < \epsilon/4$, using the above we have that $\forall\ n, N$:

$$\begin{aligned}
&\mathbb{P}_{\mathbb{S}_n,\mathbb{S}_N}\left(\left|\hat{\mathcal{G}}_{n,N} - \hat{\mathcal{G}}_n\right| > N^{-\frac{1}{2}}t\right) \leq \mathbb{E}_{\mathbb{S}_n}\left[2\exp\left(-\frac{2N^2t^2}{4N^2\hat{T}_n^2}\right)\right] \\
=&\mathbb{E}_{\mathbb{S}_n}\left[2\exp\left(-\frac{t^2}{2\hat{T}_n^2}\right)\right] = \mathbb{E}_{\mathbb{S}_n}\left[2\exp\left(-\frac{t^2}{2\hat{T}_n^2}\right)\left(I\{\hat{T}_n > \delta(\epsilon)\} + I\{\hat{T}_n \leq \delta(\epsilon)\}\right)\right] \\
&\leq 2\mathbb{P}_{\mathbb{S}_n}\left(\hat{T}_n < \delta(\epsilon)\right) + 2\exp\left(-\frac{t^2}{2\delta^2(\epsilon)}\right)\mathbb{P}_{\mathbb{S}_n}\left(\hat{T}_n > \delta(\epsilon)\right) \leq 2\exp\left(-\frac{t^2}{2\delta^2(\epsilon)}\right) + \frac{\epsilon}{2} \leq \frac{2\epsilon}{2} = \epsilon,
\end{aligned}$$

where the last step follows from choosing t large enough such that $\exp\left(-\frac{t^2}{2\delta^2(\epsilon)}\right) \leq \epsilon/4$. □

For Assumption B.3.3 and Lemma B.3.4 we first define some notation and set up the problem. Let $\bar{X} = (\bar{X}_1, \bar{X}_2) \in \mathbb{R}^{\ell_1+\ell_2}$ be any random vector and $g(\bar{X}_1) \in \mathbb{R}$ be any measurable function of $\bar{X}_1 \in \mathbb{R}^{\ell_1}$ with ℓ_1, ℓ_2 fixed. Suppose we're interested in estimating $m(\bar{X}_2) = \mathbb{E}[g(\bar{X}_1)|\bar{X}_2]$. Let $\mathbb{S}_n = \{\bar{X}\}_{i=1}^n$ be a random sample of n i.i.d. observations of $\bar{X}$, and $\mathbb{S}_{k=1}^K$ denote a random partition of $\mathbb{S}_n$ into K disjoint subsets of size $n_K = \frac{n}{K}$ with index sets $\{\mathcal{I}_k\}_{k=1}^K$. We will use cross-validation to estimate $\widehat{m}(\bar{X}_2)$, that is, we use subset $\mathcal{I}_k$ to train estimator $\widehat{m}_k$ and we estimate $m(\bar{X}_2)$ with: $\widehat{m}(\bar{X}_2) = K^{-1}\sum_{k=1}^K \sum_{i\in\mathcal{I}_k} \widehat{m}_k(\bar{X}_2)$, $K \geq 2$. Denote by $\hat{C}_{n,N} \in \mathbb{R}$ an estimator which depends on both samples $\mathbb{S}_n, \mathbb{S}_N$. Additionally, let function $\hat{\pi}_n(\cdot) : \mathbb{R}^{\ell_2} \to (0,1)$

be a random function with limit $\pi(\cdot)$, $\hat{l}_n(\bar{X}_2) : \mathbb{R}^{\ell_2} \to \{0,1\}$, be a random function with limit $l(\bar{X}_2)$, and finally function $f : \mathbb{R}^{\ell_2} \to \mathbb{R}^d$, $d \leq \ell_2$ be any deterministic function of $\bar{X}_2$.

Assumption B.3.3. *Let $\mathcal{X} \subset \mathbb{R}^p$ for an arbitrary $p \in \mathbb{N}$ i) function $w : \mathcal{X} \mapsto \mathbb{R}$ and estimator $\hat{\pi}_n$ are such that $\sup_{\bar{X}_2} \left|\hat{\pi}_n(\bar{X}_2)^{-1} - \pi(\bar{X}_2)^{-1}\right| = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, ii) function $l : \mathcal{X} \mapsto \{0,1\}$ and estimator $\hat{l}_n$ are such that $\sup_{\bar{X}_2} \left|\hat{l}_n(\bar{X}_2) - l(\bar{X}_2)\right| = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, and iii) function $f : \mathbb{R}^{\ell_2} \to \mathbb{R}^d$, $d \leq \ell_2$ is such that $\sup_{\bar{X}_2} \|f(\bar{X}_2)\| < \infty$.*

Lemma B.3.4. *Define $\hat{\boldsymbol{G}}_k^n(\bar{X}_2) = \hat{C}_{n,N} \frac{\hat{l}_n(\bar{X}_2)}{\hat{\pi}_n(\bar{X}_2)} f(\bar{X}_2)\hat{\Delta}_k(\bar{X}_2) - \mathbb{E}\left[\frac{l(\bar{X}_2)}{\pi(\bar{X}_2)} f(\boldsymbol{x}_2)\hat{\Delta}_k(\bar{X}_2)\right]$ for $\hat{\Delta}_k(\bar{X}_2) = \widehat{m}_k(\bar{X}_2) - m(\bar{X}_2)$, and $\hat{C}_{n,N} \in \mathbb{R}$ which satisfies $\hat{C} = 1 + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. Under Assumptions 2.6.8 and B.3.3, there is $c_{n_K^-} = o(1)$ such that $\mathbb{G}_{n,K} = n^{-\frac{1}{2}} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \hat{\boldsymbol{G}}_k^n(\bar{X}_2) = O_{\mathbb{P}}\left(c_{n_K^-}\right)$,*

Proof of Lemma B.3.4. First we define

$$\mathcal{G}_k^{(n)} = n^{-\frac{1}{2}} \sum_{i \in \mathcal{I}_k} \frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i})\hat{\Delta}_k(\bar{X}_{2i}) - \mathbb{E}\left[\frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i})\hat{\Delta}_k(\bar{X}_{2i})\right],$$

for any sample subset $\mathbb{S}_K \subseteq \mathcal{L}$, let $\mathbb{P}_{\mathbb{S}_K}$ denote the joint probability distribution of $\mathbb{S}_K$, and let $\mathbb{E}_{\mathbb{S}_K}[\cdot]$ denote expectation with respect to $\mathbb{P}_{\mathbb{S}_K}$, and $\mathbb{G}_{n,K} = K^{-\frac{1}{2}} \sum_{k=1}^K \mathcal{G}_k^{(n)}$, Next by Assumption 2.6.8 we have $\widehat{d}_k \equiv \sup_{\bar{X}_2} \hat{\Delta}(\bar{X}_2) = o_{\mathbb{P}}(1)$. Finally let $B_1 = \sup_{\bar{X}_2} \|f(\bar{X}_2)\|_2 < \infty$, $B_2 < \infty$ be the upperbound to $\sup_{\bar{X}_2} |\pi(\bar{X}_2)^{-1}|, \sup_{\bar{X}_2} |\hat{l}_n(\bar{X}_2)| \sup_{\bar{X}_2} |\frac{\hat{l}_n(\bar{X}_2)}{\hat{\pi}_n(\bar{X}_2)}|$.

First note that

$$
\begin{aligned}
&\left\|\mathbb{G}_{n,K}\right\|_2 \\
&= \left\| n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \hat{C}_{n,N} \frac{\hat{l}_n(\bar{X}_{2i})}{\hat{\pi}_n(\bar{X}_{2i})} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) - \mathbb{E}\left[\frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) \right] \right\|_2 \\
&\leq \left\| \left(\hat{C}_{n,N} - 1 \right) n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) \frac{\hat{l}_n(\bar{X}_{2i})}{\hat{\pi}_n(\bar{X}_{2i})} \right\|_2 \\
&+ \left\| n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) \hat{l}_n(\bar{X}_{2i}) \left(\frac{1}{\hat{\pi}_n(\bar{X}_{2i})} - \frac{1}{\pi(\bar{X}_{2i})} \right) \right\|_2 \\
&+ \left\| n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) \frac{1}{\pi(\bar{X}_{2i})} \left(\hat{l}_n(\bar{X}_{2i}) - l(\bar{X}_{2i}) \right) \right\|_2 \\
&+ \left\| n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) - \mathbb{E}\left[\frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) \right] \right\|_2,
\end{aligned}
$$

which follows from the triangle inequality, next as $f(\cdot), \hat{\pi}_n(\cdot)^{-1}, \pi(\cdot)^{-1}, \hat{l}_n(\cdot)$ are bounded $\forall \bar{X}_2 \in \mathcal{X}$, and using uniform bounds of $O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ for the difference terms we have

$$
\begin{aligned}
\left\|\mathbb{G}_{n,K}\right\|_2 \leq & O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) n^{\frac{1}{2}} B_1 B_2 \left| \sum_{k=1}^{K} \widehat{d}_k \right| + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) n^{\frac{1}{2}} B_1 B_2 \left| \sum_{k=1}^{K} \widehat{d}_k \right| \\
& + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) n^{\frac{1}{2}} B_1 B_2 \left| \sum_{k=1}^{K} \widehat{d}_k \right| + \left\| \frac{1}{K} \sum_{k=1}^{K} \mathcal{G}_k^{(n)} \right\|_2, \\
\leq & \left\| n^{-\frac{1}{2}} \frac{1}{K} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) - \mathbb{E}\left[\frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})} f(\bar{X}_{2i}) \hat{\Delta}_k(\bar{X}_{2i}) \right] \right\|_2 + o_{\mathbb{P}}(1).
\end{aligned}
$$

where the last step follows from $\widehat{d}_k = o_{\mathbb{P}}(1)$. Next we want to bound the first term above by

$c_{n_K^-}$ in probability, note that $\forall \epsilon \; \exists M > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{k=1}^{K}\mathcal{G}_k^{(n)}\right\|_2 > Mc_{n_K^-}\right) \leq \mathbb{P}\left(K^{-\frac{1}{2}}\left\|\sum_{k=1}^{K}\mathcal{G}_k^{(n)}\right\|_2 > Mc_{n_K^-}\right) \\
&\leq \sum_{k=1}^{K}\mathbb{P}\left(\left\|\mathcal{G}_k^{(n)}\right\|_2 > \frac{Mc_{n_K^-}}{K^{\frac{1}{2}}}\right) \leq \sum_{k=1}^{K}\sum_{j=1}^{d}\mathbb{P}\left(\left|\mathcal{G}_{k[j]}^{(n)}\right| > \frac{Mc_{n_K^-}}{(Kd)^{\frac{1}{2}}}\right) \\
&\leq \sum_{k=1}^{K}\sum_{j=1}^{d}\mathbb{E}_{\mathcal{L}_k^-}\left[\mathbb{P}_{\mathcal{L}_k}\left(\left|\mathcal{G}_{k[j]}^{(n)}\right| > \frac{Mc_{n_K^-}}{(Kd)^{\frac{1}{2}}}\middle|\mathcal{L}_k^-\right)\right],
\end{aligned}$$

where the first 3 steps follow from applying Boole's inequality and the triangle inequality, the fourth step follows from iterated expectations for the the event $\left\{\left|\mathcal{G}_{k[j]}^{(n)}\right| > \frac{Mc_{n_K^-}}{(Kd)^{\frac{1}{2}}}\right\}$.

Next, we have $\mathcal{L}_k^- \perp\!\!\!\perp \mathcal{L}_k$, $\forall\, k \in \{1,\ldots,K\}$, thus conditional on $\mathcal{L}_k^-$, $n^{\frac{1}{2}}\mathcal{G}_k^{(n)}$ is a sum of iid centered random vectors $\left\{\frac{l(\bar{X}_{2i})}{\pi(\bar{X}_{2i})}f(\bar{X}_{2i})\hat{\Delta}_k(\bar{X}_{2i})\right\}_{i\in\mathcal{I}_k}$ which are bounded a.s. $\mathbb{P}_{\mathcal{L}}^-$, $\forall k, n$. Thus we can apply Hoeffding's inequality to $\mathcal{G}_{k[j]}^{(n)} \forall j$:

$$\mathbb{P}_{\mathcal{L}_k}\left(\left|\mathcal{G}_{k[j]}^{(n)}\right| > \frac{Mc_{n_K^-}}{(Kd)^{\frac{1}{2}}}\middle|\mathcal{L}_k^-\right) \leq 2\exp\left\{-\frac{M^2c_{n_K^-}^2}{2KdB^2\widehat{d}_k^2}\right\} \tag{B.13}$$

a.s. $\mathbb{P}_{\mathcal{L}_k^-}\forall n$; and for each $k \in \{1,\ldots,K\}, j \in \{1\ldots,d\}$. Note that $\frac{c_{n_K^-}}{D_k} \geq 0$ is stochastically bounded away from zero as $\widehat{d}_k = o_{\mathbb{P}}(1)$, therefore $\forall k$ and given $\epsilon > 0$, $\exists \delta(\epsilon,k) > 0$ such that $\mathbb{P}_{\mathcal{L}_k^-}\left(\frac{c_{n_K^-}}{D_k} \leq \delta(\epsilon,k)\right) \leq \frac{\epsilon}{4Kd}$, let $\delta^*(\epsilon,k) = \min_k\{\delta(\epsilon,k)\}$, we have that $\mathbb{P}_{\mathcal{L}_k^-}\left(\frac{c_{n_K^-}}{D_k} \leq \delta^*(\epsilon,k)\right) \leq \frac{\epsilon}{4Kd}$.

Therefore using the bound in (B.13) and event $\left\{\frac{c_{n_K^-}}{D_k} \leq \delta^*(\epsilon,k)\right\}$:

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{k=1}^{K}\mathcal{G}_k^{(n)}\right\|_2 > Mc_{n_K^-}\right)\\
&\leq \sum_{k=1}^{K}\sum_{j=1}^{d}\mathbb{E}_{\mathcal{L}_k^-}\left[\mathbb{P}_{\mathcal{L}_k}\left(\left|\mathcal{G}_{k[j]}^{(n)}\right| > \frac{Mc_{n_K^-}}{(Kd)^{\frac{1}{2}}}\middle|\mathcal{L}_k^-\right)\right]\\
&\leq \sum_{k=1}^{K}\sum_{j=1}^{d}\mathbb{E}_{\mathcal{L}_k^-}\left[2\exp\left\{-\frac{M^2c^2_{n_K^-}}{2KdB^2\widehat{d}_k^2}\right\}\left(I\left\{\frac{c_{n_K^-}}{D_k}\leq\delta^*(\epsilon,k)\right\}+I\left\{\frac{c_{n_K^-}}{D_k}\delta^*(\epsilon,k)\right\}\right)\right]\\
&\leq 2Kd\exp\left\{-\frac{M^2\delta^*(\epsilon,k)^2}{2KdB^2}\right\}\mathbb{P}_{\mathcal{L}_k^-}\left(\frac{c_{n_K^-}}{D_k}\leq\delta^*(\epsilon,k)\right)+2Kd\mathbb{P}_{\mathcal{L}_k^-}\left(\frac{c_{n_K^-}}{D_k}>\delta^*(\epsilon,k)\right)\\
&\leq 2Kd\frac{\epsilon}{4Kd}+2Kd\exp\left\{-\frac{M^2\delta^*(\epsilon,k)^2}{2KdB^2}\right\}\mathbb{P}_{\mathcal{L}_k^-}\left(\frac{c_{n_K^-}}{D_k}>\delta^*(\epsilon,k)\right),
\end{aligned}
$$

next note that choosing a large enough M such that $\exp\left\{-\frac{M^2\delta^*(\epsilon,k)^2}{2KdB^2}\right\} < \frac{\epsilon}{4Kd}$, since $\mathbb{P}_{\mathcal{L}_k^-}\left(\frac{c_{n_K^-}}{D_k}>\delta^*(\epsilon,k)\leq 1\right)$ we get $\mathbb{P}\left(\left\|\sum_{k=1}^{K}\mathcal{G}_k^{(n)}\right\|_2 > Mc_{n_K^-}\right)\leq\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon$.
Finally we have

$$\mathbb{G}_{n,K} = O_{\mathbb{P}}\left(c_{n_K^-}\right)+o_{\mathbb{P}}(1) = O_{\mathbb{P}}\left(c_{n_K^-}\right).$$

□

Lemma B.3.5. *Let $\widehat{\boldsymbol{\gamma}}\in\mathbb{R}^d$ be a random variable such that $\sqrt{n}\left(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}}\right)=O_{\mathbb{P}}(1)$, then for any fixed vector $\mathbf{a}\in\mathbb{R}^d$ we have that (a) $\sqrt{n}\left([\mathbf{a}^\intercal\widehat{\boldsymbol{\gamma}}]_+-[\mathbf{a}^\intercal\bar{\boldsymbol{\gamma}}]_+\right)=\sqrt{n}\left(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}}\right)I(\mathbf{a}^\intercal\bar{\boldsymbol{\gamma}}>0)+o_{\mathbb{P}}(1)$, (b) Functions $\widehat{d}_t$ $t=1,2$, defined in Section 2.5 and propensity scores π_1 in (2.4) satisfy*

$$
\begin{aligned}
\sup_{\mathbf{H}_1,\mathbf{a}_1}\left|I(\widehat{d}_1=A_1)-I(\bar{d}_1=A_1)\right| &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),\\
\sup_{\mathbf{H}_2,\mathbf{a}_2}\left|I(\widehat{d}_1=A_1)I(A_2=\widehat{d}_2)-I(\bar{d}_1=A_1)I(\bar{d}_2=A_2)\right| &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),\\
\sup_{\mathbf{H}_1}\left|\frac{1}{\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)}-\frac{1}{\pi_1(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1)}\right| &= O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}
$$

(c) For $\widehat{\boldsymbol{\theta}}$, $\widehat{\boldsymbol{\xi}}$ estimated via our semi-supervised approach, and limits $\bar{\boldsymbol{\theta}}$, $\bar{\boldsymbol{\xi}}$ defined in Assumptions 2.6.3 and 2.6.7 respectively

$$\hat{C}_{n,N}^{(1)}=\frac{(1+\hat{\beta}_{21})\mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)\right\}}{(1+\hat{\beta}_{21})\mathbb{P}_n\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\widehat{\boldsymbol{\Theta}}_1)\right\}},\ \hat{C}_{n,N}^{(2)}=\frac{\mathbb{P}_N\left\{Q_{2-}^o(\mathbf{H}_2,A_2;\bar{\boldsymbol{\theta}}_2)\right\}}{\mathbb{P}_n\left\{Q_{2-}^o(\mathbf{H}_2,A_2;\widehat{\boldsymbol{\theta}}_2)\right\}},$$

satisfy $\hat{C}^{(1)}_{n,N} = 1 + O_{\mathbb{P}}(n^{-\frac{1}{2}})$, $\hat{C}^{(2)}_{n,N} = 1 + O_{\mathbb{P}}(n^{-\frac{1}{2}})$.

Proof of Lemma B.3.5. Define set $\mathcal{A}_q$ for any q dimensional vector $\hat{\boldsymbol{\gamma}}$ as

$$\mathcal{A}_q = \left\{ \widehat{\gamma} \in \mathbb{R}^q \;\middle|\; \frac{1}{2}\mathbf{a}^\intercal \bar{\gamma} < \mathbf{a}^\intercal \widehat{\gamma} < 2\mathbf{a}^\intercal \bar{\gamma},\ \forall \mathbf{a} \in \mathbb{R}^q \right\}.$$

Now consider $\widehat{\gamma} \in \mathcal{A}_q$:

- if $\text{sign}(\mathbf{a}^\intercal \bar{\gamma}) = 1$, then $0 < \frac{1}{2}\mathbf{a}^\intercal \bar{\gamma} < \mathbf{a}^\intercal \widehat{\gamma} \implies \text{sign}(\mathbf{a}^\intercal \widehat{\gamma}) = 1$,
- if $\text{sign}(\mathbf{a}^\intercal \bar{\gamma}) = -1$, then $\mathbf{a}^\intercal \widehat{\gamma} < 2\mathbf{a}^\intercal \bar{\gamma} < 0 \implies \text{sign}(\mathbf{a}^\intercal \widehat{\gamma}) = -1$.

Assuming $\sqrt{n}(\widehat{\gamma} - \bar{\gamma}) = O_{\mathbb{P}}(1)$, $\mathcal{A}_q$ exists and in fact it is such that $\mathbb{P}\left(\widehat{\gamma} \in \mathcal{A}_q\right) \xrightarrow{p} 1$.

(a) Using the above:

$$\begin{aligned}
\sqrt{n}\left([\mathbf{a}^\intercal \widehat{\gamma}]_+ - [\mathbf{a}^\intercal \bar{\gamma}]_+\right) =& \sqrt{n}\left(\widehat{\gamma} - \bar{\gamma}\right) I(\mathbf{a}^\intercal \bar{\gamma} > 0) I\left(\widehat{\gamma} \in \mathcal{A}_q\right) + \sqrt{n}\left([\mathbf{a}^\intercal \widehat{\gamma}]_+ - [\mathbf{a}^\intercal \bar{\gamma}]_+\right) I\left(\widehat{\gamma} \notin \mathcal{A}_q\right) \\
=& \sqrt{n}\left(\widehat{\gamma} - \bar{\gamma}\right) I(\mathbf{a}^\intercal \bar{\gamma} > 0) + o_{\mathbb{P}}(1).
\end{aligned}$$

(b) As $A_{ti} \in \{0, 1\}$, $t = 1, 2$, we can write

$$\begin{aligned}
I(\widehat{d}_1 = A_1) I(\widehat{d}_2 = A_2) =& I\left\{A_1 = I(\mathbf{H}_{11}^\intercal \widehat{\gamma}_1 > 0)\right\} I\left\{A_2 = I(\mathbf{H}_{21}^\intercal \widehat{\gamma}_2 > 0)\right\} \\
=& I\{A_1 = I(\mathbf{H}_{11}^\intercal \widehat{\gamma}_1 > 0)\} I\{A_2 = I(\mathbf{H}_{21}^\intercal \widehat{\gamma}_2 > 0)\} \\
=& A_1 A_2 I(\mathbf{H}_{11}^\intercal \widehat{\gamma}_1 > 0) I(\mathbf{H}_{21}^\intercal \widehat{\gamma}_2 > 0) \\
&+ (1 - A_1)(1 - A_2) I(\mathbf{H}_{11}^\intercal \widehat{\gamma}_1 < 0) I(\mathbf{H}_{21}^\intercal \widehat{\gamma}_2 < 0) \\
&+ A_1 (1 - A_2) I(\mathbf{H}_{11}^\intercal \widehat{\gamma}_1 > 0) I(\mathbf{H}_{21}^\intercal \widehat{\gamma}_2 < 0) \\
&+ (1 - A_1) A_2 I(\mathbf{H}_{11}^\intercal \widehat{\gamma}_1 < 0) I(\mathbf{H}_{21}^\intercal \widehat{\gamma}_2 > 0),
\end{aligned}$$

therefore

$$
\begin{aligned}
&\left|I(\widehat{d}_1 = A_1)I(\widehat{d}_2 = A_2) - I(\bar{d}_1 = A_1)I(\bar{d}_2 = A_2)\right| \\
=&\Big|A_1A_2\left\{I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 > 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 > 0)\right\} \\
&+(1-A_1)(1-A_2)\left\{I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 < 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 < 0)\right\} \\
&+A_1(1-A_2)\left\{I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 < 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 < 0)\right\} \\
&+(1-A_1)A_2\left\{I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 > 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 > 0)\right\}\Big| \\
\leq&A_1A_2\left|I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 > 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 > 0)\right| \\
&+(1-A_1)(1-A_2)\left|I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 < 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 < 0)\right| \\
&+A_1(1-A_2)\left|I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 < 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 > 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 < 0)\right| \\
&+(1-A_1)A_2\left|I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 > 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 > 0)\right|
\end{aligned}
$$

where the first step follows from above, the second step from the triangle inequality, now as $\widehat{\boldsymbol{\gamma}}_1$, $\widehat{\boldsymbol{\gamma}}_2$ have dimensions q_{12}, q_{22} respectively, we use sets $\mathcal{A}_{q_{12}}$, $\mathcal{A}_{q_{22}}$ and have

$$
\begin{aligned}
&\left|I(\widehat{d}_1 = A_1)I(\widehat{d}_2 = A_2) - I(\bar{d}_1 = A_1)I(\bar{d}_2 = A_2)\right| \\
\leq&A_1A_2I(\widehat{\boldsymbol{\gamma}}_1 \notin \mathcal{A}_{q_{12}})I(\widehat{\boldsymbol{\gamma}}_2 \notin \mathcal{A}_{q_{22}}) + (1-A_1)(1-A_2)I(\widehat{\boldsymbol{\gamma}}_1 \notin \mathcal{A}_{q_{12}})I(\widehat{\boldsymbol{\gamma}}_2 \notin \mathcal{A}_{q_{22}}) \\
&+A_1(1-A_2)I(\widehat{\boldsymbol{\gamma}}_1 \notin \mathcal{A}_{q_{12}})I(\widehat{\boldsymbol{\gamma}}_2 \notin \mathcal{A}_{q_{22}}) + (1-A_1)A_2I(\widehat{\boldsymbol{\gamma}}_1 \notin \mathcal{A}_{q_{12}})I(\widehat{\boldsymbol{\gamma}}_2 \notin \mathcal{A}_{q_{22}}) \\
=&I(\widehat{\boldsymbol{\gamma}}_1 \notin \mathcal{A}_{q_{12}})I(\widehat{\boldsymbol{\gamma}}_2 \notin \mathcal{A}_{q_{22}})
\end{aligned}
$$

which follows from the fact that for any term within absolute value:

$$
\left|I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 > 0) - I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 > 0)\right| = I(\widehat{\boldsymbol{\gamma}}_1 \notin \mathcal{A}_{q_{12}})I(\widehat{\boldsymbol{\gamma}}_2 \notin \mathcal{A}_{q_{22}})
$$

since for $I(\mathbf{H}_{11}^\intercal\bar{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2 > 0) \neq I(\mathbf{H}_{11}^\intercal\widehat{\boldsymbol{\gamma}}_1 < 0)I(\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2 > 0)$ both $\widehat{\boldsymbol{\gamma}}_1, \widehat{\boldsymbol{\gamma}}_2$ have to be outside sets $\mathcal{A}_{q_{12}}, \mathcal{A}_{q_{22}}$ respectively. Thus $\left|I(\widehat{d}_1 = A_1)I(\widehat{d}_2 = A_2) - I(\bar{d}_1 = A_1)I(\bar{d}_2 = A_2)\right| = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, we can analogous show that $\left|I(\widehat{d}_1 = A_1) - I(\bar{d}_1 = A_1)\right| = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)\ \forall i$.
Next to see $\sup_{\mathbf{H}_1}\left|\frac{1}{\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)} - \frac{1}{\pi_1(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1)}\right| = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, note that as $\mathcal{H}_1$, Ω_1 are bounded sets

we have

$$\begin{aligned}
&\sup_{\mathbf{H}_1}\left|\frac{1}{\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)}-\frac{1}{\pi_1(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1)}\right| = \sup_{\mathbf{H}_1\in\mathcal{H}_1}\left|e^{-\mathbf{H}_1^\intercal\hat{\boldsymbol{\xi}}_1}-e^{-\mathbf{H}_1^\intercal\bar{\boldsymbol{\xi}}_1}\right| \\
&\leq \sup_{\mathbf{H}_1\in\mathcal{H}_1,\boldsymbol{\xi}_1\in\Omega_1}\left|\frac{d}{dx}e^{-x}\big|_{x=\mathbf{H}_1^\intercal\boldsymbol{\xi}_1}\right| \sup_{\mathbf{H}_1\in\mathcal{H}_1}\left|\mathbf{H}_1^\intercal\hat{\boldsymbol{\xi}}_1-\mathbf{H}_1^\intercal\bar{\boldsymbol{\xi}}\right| \\
&\leq \sup_{\mathbf{H}_1\in\mathcal{H}_1,\boldsymbol{\xi}_1\in\Omega_1}\left|\frac{d}{dx}e^{-x}\big|_{x=\mathbf{H}_1^\intercal\boldsymbol{\xi}_1}\right| \sup_{\mathbf{H}_1\in\mathcal{H}_1}\|\mathbf{H}_1\|\left\|\hat{\boldsymbol{\xi}}_1-\bar{\boldsymbol{\xi}}_1\right\|_2 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),
\end{aligned}$$

where we use the definition of π_1 in (2.4), Lipschitz and $\left\|\hat{\boldsymbol{\xi}}_1-\bar{\boldsymbol{\xi}}_1\right\|_2 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$ from Assumptions (2.6.7) and Theorem 5.21 in [61] as we are using Z-estimation for $\boldsymbol{\xi}_1$.

(c) By Theorem 2.6.4 we have $\hat{\beta}_{21}-\bar{\beta}_{21}=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. Next, we can write

$$\omega_1(\mathbf{H}_1,A_1;\widehat{\boldsymbol{\Theta}}_1)=I\left\{A_1=d_1\left(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1\right)\right\}\left\{\frac{A_1}{\pi_1\left(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1\right)}+\frac{1-A_1}{1-\pi_1\left(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1\right)}\right\}.$$

By Lemma B.3.5 (b) it follows that

$$\begin{aligned}
\mathbb{P}_n\left[I\left\{A_1=d_1\left(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1\right)\right\}-I\left\{A_1=d_1\left(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1\right)\right\}\right]&=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),\\
\mathbb{P}_n\left[\frac{A_1}{\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)}-\frac{A_1}{\pi_1\left(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1\right)}\right]&=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),\\
\mathbb{P}_n\left[\frac{1-A_1}{1-\pi_1(\mathbf{H}_1;\widehat{\boldsymbol{\xi}}_1)}-\frac{1-A_1}{1-\pi_1\left(\mathbf{H}_1;\bar{\boldsymbol{\xi}}_1\right)}\right]&=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Using the above and Lemma B.3.1 we get

$$(1+\hat{\beta}_{21})\mathbb{P}_n\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\widehat{\boldsymbol{\Theta}}_1)\right\}=(1+\bar{\beta}_{21})\mathbb{P}_n\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)\right\}+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$$

Also by CLT we have

$$\begin{aligned}
(1+\bar{\beta}_{21})\mathbb{P}_n\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)\right\}&=(1+\bar{\beta}_{21})\mathbb{E}\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)\right\}+O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right),\\
(1+\bar{\beta}_{21})\mathbb{P}_N\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)\right\}&=(1+\bar{\beta}_{21})\mathbb{E}\left\{\omega_1(\check{\mathbf{H}}_1,A_1;\bar{\boldsymbol{\Theta}}_1)\right\}+O_{\mathbb{P}}\left(N^{-\frac{1}{2}}\right),
\end{aligned}$$

finally by Slutsky's theorem $\hat{C}^{(1)}_{n,N}-1=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$. With similar arguments, and using Lemma B.3.5 (a) to see $\mathbb{P}_n\left([\mathbf{H}_{21}^\intercal\widehat{\boldsymbol{\gamma}}_2]_+-[\mathbf{H}_{21}^\intercal\bar{\boldsymbol{\gamma}}_2]_+\right)=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, we can show $\hat{C}^{(2)}_{n,N}-1=O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$.

□

Lemma B.3.6. *Let $Q_t(\check{\mathbf{H}}_t;\boldsymbol{\theta}_t)$, $\pi_t(\check{\mathbf{H}}_t;\boldsymbol{\xi}_t)$ $t=1,2$ be estimator functions of (2.1) & (2.4) respectively and define the bias as Bias $\left(\bar{V},\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right) \equiv \bar{V}-\mathbb{E}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right]$, then*

$$
\begin{aligned}
&Bias\left(\bar{V},\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right)\\
=&\mathbb{E}\left[\left\{1-\frac{\pi_1(\check{\mathbf{H}}_1)}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1)-Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\right]\\
+&\mathbb{E}\left[\left\{1-\frac{1-\pi_1(\check{\mathbf{H}}_1)}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1)-Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\right]\\
+&\mathbb{E}\left[\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}+\frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}\left\{1-\frac{\pi_2(\check{\mathbf{H}}_2)}{\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2)-Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right]\\
+&\mathbb{E}\left[\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}+\frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}\left\{1-\frac{1-\pi_2(\check{\mathbf{H}}_2)}{1-\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2)-Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right].
\end{aligned}
$$

where $\bar{V}=\mathbb{E}[\mathbb{E}[Y_2+\mathbb{E}[Y_3|\mathbf{H}_2,Y_2,A_2=\bar{d}_2(\check{\mathbf{H}}_2)]|\mathbf{H}_1,A_1=\bar{d}_1(\check{\mathbf{H}}_1)]]$ is the mean population value under the optimal treatment rule.

Proof of Lemma B.3.6.

$$
\begin{aligned}
\text{Bias}\left(\bar{V},\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right) =&\mathbb{E}[\mathbb{E}[Y_2+\mathbb{E}[Y_3|\mathbf{H}_2,Y_2,A_2=\bar{d}_2]|\mathbf{H}_1,A_1=\bar{d}_1]]-\mathbb{E}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right]\\
=&\mathbb{E}\left[Q_1^o(\mathbf{H}_1)-Q_1^o(\mathbf{H}_1;\boldsymbol{\theta}_1)\right]\\
-&\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1,A_1;\boldsymbol{\Theta}_1)\left\{Y_2-Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\right]\\
-&\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1,A_1;\boldsymbol{\Theta}_1)Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right]\\
-&\mathbb{E}\left[\omega_2(\check{\mathbf{H}}_2,A_2;\boldsymbol{\Theta}_2)\left\{Y_3-Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right].
\end{aligned}
$$

Adding and subtracting $\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1,A_1;\boldsymbol{\Theta}_1)Q_2^o(\check{\mathbf{H}}_2)\right]=\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1,A_1;\boldsymbol{\Theta}_1)\mathbb{E}[Y_3|\mathbf{H}_2,\bar{d}_2(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2),Y_2]\right]$,

$$
\begin{aligned}
&\text{Bias}\left(\bar{V},\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\boldsymbol{\Theta})\right)\\
&=\mathbb{E}\left[Q_1^o(\mathbf{H}_1)-Q_1^o(\mathbf{H}_1;\boldsymbol{\theta}_1)\right]\\
&-\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1,A_1;\boldsymbol{\Theta}_1)\left\{Y_2+\mathbb{E}[Y_3|\mathbf{H}_2\bar{d}_2(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2),Y_2]-Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\right]\\
&-\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1,A_1;\boldsymbol{\Theta}_1)\left\{Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)-Q_2^o(\check{\mathbf{H}}_2)\right\}\right]\\
&-\mathbb{E}\left[\omega_2(\check{\mathbf{H}}_2,A_2;\boldsymbol{\Theta}_2)\left\{Y_3-Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right],
\end{aligned}
$$

using iterated expectations in the second and fourth terms:

$$
\begin{aligned}
&\text{Bias}\left(\bar{V}, \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}\left(\mathbf{L};\boldsymbol{\Theta}\right)\right)\\
&= \mathbb{E}\left[Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right]\\
&- \mathbb{E}\left[\mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1;\boldsymbol{\Theta}_1)\left\{Y_2 + \mathbb{E}[Y_3|\check{\mathbf{H}}_2, \bar{d}_2(\check{\mathbf{H}}_2), Y_2] - Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\middle|\check{\mathbf{H}}_1, A_1\right]\right]\\
&- \mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1;\boldsymbol{\Theta}_1)\left\{Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2) - Q_2^o(\check{\mathbf{H}}_2)\right\}\right]\\
&- \mathbb{E}\left[\mathbb{E}\left[\omega_2(\check{\mathbf{H}}_2, A_2;\boldsymbol{\Theta}_2)\left\{Y_3 - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\middle|\check{\mathbf{H}}_2, A_2, Y_2\right]\right]\\
&= \mathbb{E}\left[Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right]\\
&- \mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1;\boldsymbol{\Theta}_1)\left\{\mathbb{E}\left[Y_2 + \mathbb{E}[Y_3|\check{\mathbf{H}}_2, \bar{d}_2(\check{\mathbf{H}}_2), Y_2]\middle|\check{\mathbf{H}}_1, A_1\right] - Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\right]\\
&- \mathbb{E}\left[\omega_1(\check{\mathbf{H}}_1, A_1;\boldsymbol{\Theta}_1)\left\{Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2) - Q_2^o(\check{\mathbf{H}}_2)\right\}\right]\\
&- \mathbb{E}\left[\omega_2(\check{\mathbf{H}}_2, A_2;\boldsymbol{\Theta}_2)\left\{\mathbb{E}\left[Y_3|\check{\mathbf{H}}_2, A_2, Y_2\right] - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right].
\end{aligned}
$$

using definitions of $\omega_t(\check{\mathbf{H}}_t, A_t;\boldsymbol{\Theta}_t)$ $t = 1, 2$ we can write:

$$
\begin{aligned}
&\text{Bias}\left(\bar{V}, \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}\left(\mathbf{L};\boldsymbol{\Theta}\right)\right)\\
=&\mathbb{E}\left[Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right]\\
-&\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)\right\}\right]\\
-&\mathbb{E}\left[\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right] - \mathbb{E}\left[\frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right]\\
-&\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1;\boldsymbol{\xi}_1)}\right\}\left\{\frac{\bar{d}_2 A_2}{\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)} + \frac{(1-\bar{d}_2)(1-A_2)}{1-\pi_2(\check{\mathbf{H}}_2;\boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\right]
\end{aligned}
$$

assuming $A_1 \perp A_2 | \mathbf{H}_2, Y_2$, we use iterated expectations:

$$
\begin{aligned}
&\operatorname{Bias}\left(\bar{V}, \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \boldsymbol{\Theta})\right) \\
=&\mathbb{E}\left[Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1)\right] \\
-&\mathbb{E}\left[\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1)\right\} \middle| \check{\mathbf{H}}_1\right]\right] \\
-&\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2)\right\}\right] \\
-&\mathbb{E}\left[\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{\frac{\bar{d}_2 A_2}{\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)} + \frac{(1-\bar{d}_2)(1-A_2)}{1-\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2)\right\} \middle| \check{\mathbf{H}}_2\right]\right] \\
=&\mathbb{E}\left[Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1)\right] \\
-&\mathbb{E}\left[\left\{\frac{\bar{d}_1 \pi_1(\check{\mathbf{H}}_1)}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{\{1-\bar{d}_1\}\{1-\pi_1(\check{\mathbf{H}}_1)\}}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1)\right\}\right] \\
-&\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2)\right\}\right] \\
-&\mathbb{E}\left[\left\{\frac{\bar{d}_1 A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{(1-\bar{d}_1)(1-A_1)}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{\frac{\bar{d}_2 \pi_2(\check{\mathbf{H}}_2)}{\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)} + \frac{\{1-\bar{d}_2\}\{1-\pi_2(\check{\mathbf{H}}_2)\}}{1-\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2)\right\}\right]
\end{aligned}
$$

finally, factorizing common terms:

$$
\begin{aligned}
&\operatorname{Bias}\left(\bar{V}, \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \boldsymbol{\Theta})\right) \\
=&\mathbb{E}\left[\bar{d}_1\left\{1 - \frac{\pi_1(\check{\mathbf{H}}_1)}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1)\right\}\right] \\
+&\mathbb{E}\left[\{1-\bar{d}_1\}\left\{1 - \frac{1-\pi_1(\check{\mathbf{H}}_1)}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{Q_1^o(\check{\mathbf{H}}_1) - Q_1^o(\check{\mathbf{H}}_1; \boldsymbol{\theta}_1)\right\}\right] \\
+&\mathbb{E}\left[\bar{d}_2\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{1 - \frac{\pi_2(\check{\mathbf{H}}_2)}{\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2)\right\}\right] \\
+&\mathbb{E}\left[\{1-\bar{d}_2\}\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\left\{1 - \frac{1-\pi_2(\check{\mathbf{H}}_2)}{1-\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)}\right\}\left\{Q_2^o(\check{\mathbf{H}}_2) - Q_2^o(\check{\mathbf{H}}_2; \boldsymbol{\theta}_2)\right\}\right] \\
\leq&\sqrt{\sup_{\check{\mathbf{H}}_1}|\{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)\}^{-1}|}\sqrt{\|\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1) - \pi_1(\check{\mathbf{H}}_1)\|_{L_2(\mathbb{P})}}\sqrt{\|Q_1^o(\check{\mathbf{H}}_1; \widehat{\boldsymbol{\theta}}_1) - Q_1^o(\check{\mathbf{H}}_1)\|_{L_2(\mathbb{P})}} \\
+&\sqrt{\sup_{\check{\mathbf{H}}_2}\left|\left\{\frac{A_1}{\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)} + \frac{1-A_1}{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)}\right\}\{1-\pi_1(\check{\mathbf{H}}_1; \boldsymbol{\xi}_1)\}^{-1}\{1-\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)\}^{-1}\right|} \\
\times&\sqrt{\|\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2) - \pi_2(\check{\mathbf{H}}_2)\|_{L_2(\mathbb{P})}}\sqrt{\|Q_2^o(\check{\mathbf{H}}_2; \widehat{\boldsymbol{\theta}}_2) - Q_2^o(\check{\mathbf{H}}_2)\|_{L_2(\mathbb{P})}},
\end{aligned}
$$

which follows by Cauchy–Schwarz Inequality. □

B.4 Additional Theoretical Results

B.4.1 Augmented value function estimation

We first re-write Assumption 2.6.7 to account for only using sample $\mathcal{L}$ in estimation of the Q functions and propensity scores.

Assumption B.4.1. *Define the following class of functions:*

$$\begin{aligned}
\mathcal{Q}_1 &\equiv \{Q_1(\mathbf{H}_1, A_1; \boldsymbol{\theta}_1) | \boldsymbol{\theta}_1 \in \Theta_1 \subset \mathbb{R}^{q_1}\}, \\
\mathcal{Q}_2 &\equiv \{Q_2(\mathbf{H}_2, A_2, Y_2; \boldsymbol{\theta}_2) | \boldsymbol{\theta}_2 \in \Theta_2 \subset \mathbb{R}^{q_2}\}, \\
\mathcal{W}_1 &\equiv \{\pi_1(\mathbf{H}_1; \boldsymbol{\xi}_1) | \boldsymbol{\xi}_1 \in \Omega_1 \subset \mathbb{R}^{p_1}\}, \\
\mathcal{W}_2 &\equiv \left\{\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2) | \boldsymbol{\xi}_2 \in \Omega_2 \subset \mathbb{R}^{p_2}\right\},
\end{aligned} \tag{B.14}$$

with p_1, p_2, q_1, q_2 fixed under model definitions (2.1) & (2.4). Let the population equations $\mathbb{E}\left[S_t^{\xi}(\boldsymbol{\xi}_t)\right] = \mathbf{0}, t = 1, 2$ have solutions $\bar{\boldsymbol{\xi}}_1, \bar{\boldsymbol{\xi}}_2$, where

$$\begin{aligned}
S_1^{\xi}(\boldsymbol{\xi}_1) =& \frac{\partial}{\partial \boldsymbol{\xi}_1} \log\left[\pi_1(\mathbf{H}_1; \boldsymbol{\xi}_1)^{A_1}(1 - \pi_1(\mathbf{H}_1; \boldsymbol{\xi}_1))^{(1-A_1)}\right], \\
S_2^{\xi}(\boldsymbol{\xi}_2) =& \frac{\partial}{\partial \boldsymbol{\xi}_2} \log\left[\pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2)^{A_2}(1 - \pi_2(\check{\mathbf{H}}_2; \boldsymbol{\xi}_2))^{(1-A_2)}\right],
\end{aligned}$$

and the population equations for the Q functions $\mathbb{E}[S_t^{\theta}(\boldsymbol{\theta}_t)] = \mathbf{0}, t = 1, 2$ have solutions $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2$, where

$$\begin{aligned}
S_2^{\theta}(\boldsymbol{\theta}_2) =& \frac{\partial}{\partial \boldsymbol{\theta}_2^{\intercal}} \|Y_3 - Q_2(\check{\mathbf{H}}_2, A_2; \boldsymbol{\theta}_2)\|_2^2, \\
S_1^{\theta}(\boldsymbol{\theta}_1) =& \frac{\partial}{\partial \boldsymbol{\theta}_1^{\intercal}} \|Y_2 + \bar{Q}_2(\check{\mathbf{H}}_2; \bar{\boldsymbol{\theta}}_2) - Q_1(\mathbf{H}_1, A_1; \boldsymbol{\theta}_1)\|_2^2,
\end{aligned}$$

(i) ξ_1, ξ_2 are bounded sets. (ii) Θ_1, Θ_2 are open bounded sets and for some $r > 0$ and $g_t(\cdot)$

$$\left|Q_t(\cdot; \boldsymbol{\theta}_t) - Q_t(\cdot; \boldsymbol{\theta}_t')\right| \leq g_t(\cdot)\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t'\| \;\; \forall \boldsymbol{\theta}_t, \boldsymbol{\theta}_t' \in \Theta_t, \; \mathbb{E}\left[|g_t(\cdot)|^r\right] < \infty, \; t = 1, 2. \tag{B.15}$$

(iii) The population minimizers satisfy $\bar{\boldsymbol{\theta}}_t \in \Theta_t, \bar{\boldsymbol{\xi}}_t \in \Omega_t, t = 1, 2$. (iv) For $\bar{\boldsymbol{\xi}}_t, t = 1, 2$, $\bar{\pi}_1(\mathbf{H}_1; \bar{\boldsymbol{\xi}}_1) > 0, \bar{\pi}_2(\check{\mathbf{H}}_2; \bar{\boldsymbol{\xi}}_2) > 0 \; \forall \mathbf{H} \in \mathcal{H}$.

Existence of solutions $\bar{\boldsymbol{\theta}}_t \in \Theta_t$, $t = 1, 2$ is clear as Θ_1, Θ_2 are open and bounded.

Theorem B.4.2 (Asymptotic Normality for $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$)**.** *Under Assumptions 2.6.1, 2.6.5, and B.4.1,* $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$ *as defined in* (2.5) *is such that*

$$\sqrt{n}\left\{\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}-\mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right]\right\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}_i;\bar{\mathbf{\Theta}})+o_{\mathbb{P}}(1)\xrightarrow{d}N\left(0,\sigma^2_{\mathsf{SUP}_{\mathsf{DR}}}\right).$$

where

$$\begin{aligned}\psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})=&\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})-\mathbb{E}_{\mathbb{S}}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right]+\boldsymbol{\psi}^{\theta}_{\mathsf{SUP}}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\theta}}\int\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})d\mathbb{P}_{\mathbf{L}}\Big|_{\mathbf{\Theta}=\bar{\mathbf{\Theta}}}\\&+\boldsymbol{\psi}^{\xi}_{\mathsf{SUP}}(\mathbf{L})^{\intercal}\frac{\partial}{\partial\boldsymbol{\xi}}\int\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})d\mathbb{P}_{\mathbf{L}}\Big|_{\mathbf{\Theta}=\bar{\mathbf{\Theta}}},\\\sigma^2_{\mathsf{SUP}_{\mathsf{DR}}}=&\mathbb{E}\left[\psi^{v}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})^2\right].\end{aligned}$$

proof of theorem B.4.2. Letting $g(\mathbf{\Theta})=\int\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\mathbf{\Theta})d\mathbb{P}_{\mathbf{L}}$, we start by centering (2.5) and scaling by $\sqrt{n}$:

$$\begin{aligned}&\sqrt{n}\left\{\mathbb{P}_n\left(\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\widehat{\mathbf{\Theta}}_{\mathsf{SUP}})\right)-\mathbb{E}\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right]\right\}\\=&\mathbb{G}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right\}+\mathbb{G}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\widehat{\mathbf{\Theta}}_{\mathsf{SUP}})-\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right\}+\sqrt{n}\left\{g(\widehat{\mathbf{\Theta}}_{\mathsf{SUP}})-g(\bar{\mathbf{\Theta}})\right\}\end{aligned}$$

I) Empirical Process Term

We first show that under Assumption B.4.1, $\mathbb{G}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\widehat{\mathbf{\Theta}}_{\mathsf{SUP}})-\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\bar{\mathbf{\Theta}})\right\}=o_{\mathbb{P}}(1)$, let

$$\begin{aligned}f_{\mathbf{\Theta}}(\vec{\mathbf{U}})=Q^{o}_1(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)+&\omega_1(\check{\mathbf{H}}_1,A_1,\mathbf{\Theta})\left\{Y_2-Q^{o}_1(\check{\mathbf{H}}_1;\boldsymbol{\theta}_1)+Q^{o}_2(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\}\\&+\omega_2(\check{\mathbf{H}}_1,A_1;\mathbf{\Theta})\left\{Y_3-Q^{o}_2(\check{\mathbf{H}}_2;\boldsymbol{\theta}_2)\right\},\end{aligned}$$

we define the class of functions $\mathcal{C}_3=\left\{f_{\mathbf{\Theta}}(\vec{\mathbf{U}})|\vec{\mathbf{U}},\mathbf{\Theta}\in\mathcal{S}(\delta)\right\}$, and

$$\ell=\{l:\{0,1\}^2\mapsto\{0,1\}\}.$$

i) By Assumptions B.4.1 and Theorem 19.5 in [61], ℓ, $\mathcal{W}_t$, $\mathcal{Q}_t, t=1,2$ are a $\mathbb{P}$-Donsker

class, thus it follows that $\mathcal{C}_3$ is a Donsker class.

ii) We estimate $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$ from (B.14) with their maximum likelihood estimator $\widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}}, \widehat{\boldsymbol{\xi}}_{2\mathsf{SUP}}$, solving $\mathbb{P}_n\left[S_t(\boldsymbol{\xi}_t)\right] = \mathbf{0}, t = 1,2$ and estimate functions $\pi_1(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}})$, $\pi_2(\check{\mathbf{H}}_2; \widehat{\boldsymbol{\xi}}_{2_{\mathsf{SUP}}})$ with $\widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}}, \widehat{\boldsymbol{\xi}}_{2\mathsf{SUP}}$. By Assumption B.4.1 and weak law of large numbers $\widehat{\boldsymbol{\xi}}_{t\mathsf{SUP}} \xrightarrow{p} \bar{\boldsymbol{\xi}}_t, t = 1,2$.

Analogous, under regularity conditions (**??**) and (**??**) have unique solutions $\widehat{\boldsymbol{\theta}}_{t\mathsf{SUP}}$ for which $\widehat{\boldsymbol{\theta}}_{t\mathsf{SUP}} \xrightarrow{p} \bar{\boldsymbol{\theta}}_t, t = 1,2$ by Assumption B.4.1 and weak law of large numbers. Both regardless of whether models (2.1) & (2.4) are correct. Thus $\mathbb{P}\left(\widehat{\boldsymbol{\Theta}}_{\mathsf{SUP}} \in \mathcal{S}(\delta)\right) \to 1$, $\forall \delta$.

iii) We next show $\int \left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}_{\mathsf{SUP}}) - \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \longrightarrow 0$. Using (2.7), for a large enough constant c we can write

$$
\begin{aligned}
&\int \left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\boldsymbol{\Theta}}_{\mathsf{SUP}}) - \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\boldsymbol{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \\
\leq & c \sup_{\mathbf{H}_1} \left(\mathbf{H}_{10}^{\intercal}\bar{\boldsymbol{\beta}}_1 + [\mathbf{H}_{11}^{\intercal}\bar{\boldsymbol{\gamma}}_1]_+ - \mathbf{H}_{10}^{\intercal}\widehat{\boldsymbol{\beta}}_{1\mathsf{SUP}} - [\mathbf{H}_{11}^{\intercal}\widehat{\boldsymbol{\gamma}}_{1\mathsf{SUP}}]_+\right)^2 \\
+ & c \sup_{\check{\mathbf{H}}_2} \left(\check{\mathbf{H}}_{20}^{\intercal}\bar{\boldsymbol{\beta}}_2 + [\mathbf{H}_{21}^{\intercal}\bar{\boldsymbol{\gamma}}_2]_+ - \check{\mathbf{H}}_{20}^{\intercal}\widehat{\boldsymbol{\beta}}_{2\mathsf{SUP}} - [\mathbf{H}_{21}^{\intercal}\widehat{\boldsymbol{\gamma}}_{2\mathsf{SUP}}]_+\right)^2 \\
+ & c \sup_{\mathbf{H}_1, A_1} \left\{\omega_1(\mathbf{H}_1, A_1; \widehat{\boldsymbol{\Theta}}_{1\mathsf{SUP}}) - \omega_1(\mathbf{H}_1, A_1; \bar{\boldsymbol{\Theta}}_1)\right\}^2 + \left(\hat{\beta}_{21\mathsf{SUP}} - \bar{\beta}_{21}\right)^2 \\
& \longrightarrow 0
\end{aligned}
$$

where we use $(a-b)^2, (a+b)^2 \leq 2a^2 + 2b^2 \; \forall a, b \in \mathbb{R}$, boundedness of $\bar{\boldsymbol{\Theta}}$ and covariates by Assumptions 2.6.1, 2.6.2, and B.4.1. Next,

from assumption (2.7) it can be shown that $\widehat{\boldsymbol{\theta}}_{2\mathsf{SUP}} - \bar{\boldsymbol{\theta}}_2 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, $\widehat{\boldsymbol{\theta}}_{1\mathsf{SUP}} - \bar{\boldsymbol{\theta}}_1 = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$, also from Lemma B.3.5 (a) it follows that for $t = 1,2$

$$
\begin{aligned}
& \sup_{\check{\mathbf{H}}_t} \left(\mathbf{H}_{t0}^{\intercal}\bar{\boldsymbol{\beta}}_t + [\mathbf{H}_{t1}^{\intercal}\bar{\boldsymbol{\gamma}}_t]_+ - \mathbf{H}_{t0}^{\intercal}\widehat{\boldsymbol{\beta}}_{t\mathsf{SUP}} - [\mathbf{H}_{t1}^{\intercal}\widehat{\boldsymbol{\gamma}}_{t\mathsf{SUP}}]_+\right)^2 \\
& \leq 2 \sup_{\check{\mathbf{H}}_{t0}} \|\check{\mathbf{H}}_{t0}\|_2^2 \|\widehat{\boldsymbol{\beta}}_t - \bar{\boldsymbol{\beta}}_t\|_2^2 + 2 \sup_{\mathbf{H}_{t1}} \|\mathbf{H}_{t1}\|_2^2 \|\widehat{\boldsymbol{\gamma}}_t - \bar{\boldsymbol{\gamma}}_t\|_2 \\
& = O_{\mathbb{P}}\left(n^{-1}\right).
\end{aligned}
$$

Next, we can write

$$\omega_1(\mathbf{H}_1, A_1; \widehat{\mathbf{\Theta}}_{1\mathsf{SUP}}) = I\left\{A_1 = d_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}}\right)\right\}\left\{\frac{A_1}{\pi_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}}\right)} + \frac{1-A_1}{1-\pi_1\left(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}}\right)}\right\}.$$

By Lemma B.3.5 (b) it follows that

$$\sup_{\mathbf{H}_1}\left|\frac{1}{\pi_1(\mathbf{H}_1; \widehat{\boldsymbol{\xi}}_{1\mathsf{SUP}})} - \frac{1}{\pi_1(\mathbf{H}_1; \bar{\boldsymbol{\xi}}_1)}\right| = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right).$$

Using the above and Lemma B.3.1 we get

$$\sup_{\check{\mathbf{H}}_1, A_1}\left\{\omega_1(\check{\mathbf{H}}_1, A_1; \widehat{\mathbf{\Theta}}_{1\mathsf{SUP}}) - \omega_1(\check{\mathbf{H}}_1, A_1; \bar{\mathbf{\Theta}}_1)\right\}^2 = o_{\mathbb{P}}(1),$$

which gives us $\int\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\mathbf{\Theta}}_{\mathsf{SUP}}) - \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \longrightarrow 0$.

Hence, we have i) $\mathbb{P}\left(\widehat{\mathbf{\Theta}}_{\mathsf{SUP}} \in \mathcal{S}(\delta)\right) \to 1$, $\forall\delta$, ii) $\mathcal{C}_1$ is a Donsker class, and iii) $\int\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\mathbf{\Theta}}_{\mathsf{SUP}}) - \mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}})\right\}^2 d\mathbb{P}_{\mathbf{L}} \longrightarrow 0$, then by Theorem 2.1 in [98]

$$\sqrt{n}\left[\mathbb{P}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \widehat{\mathbf{\Theta}}_{\mathsf{SUP}}) - g(\widehat{\mathbf{\Theta}}_{\mathsf{SUP}})\right\} - \mathbb{P}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}}) - g(\bar{\mathbf{\Theta}})\right\}\right] = o_{\mathbb{P}}(1).$$

Centered Sample Average

Next we consider $\mathbb{G}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}})\right\}$. Note that $\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}})$ is a deterministic function of random variable $\mathbf{L}$ as parameters are fixed. We have that $\mathbb{E}\left[\left(\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}}\right)^2\right] < \infty$ holds by Assumption 2.6.1 & B.4.1. Thus the central limit theorem yields

$$\mathbb{G}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}})\right\} \xrightarrow{d} \mathcal{N}\left(0, Var\left[\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L}; \bar{\mathbf{\Theta}})\right]\right).$$

Bias Term

We finally analyze the bias: $\sqrt{n}\left\{g(\widehat{\mathbf{\Theta}}_{\mathsf{SUP}}) - g(\bar{\mathbf{\Theta}})\right\}$. Using a Taylor series expansion

$$g(\widehat{\mathbf{\Theta}}_{\mathsf{SUP}}) = g(\bar{\mathbf{\Theta}}) + (\widehat{\boldsymbol{\theta}}_{\mathsf{SUP}} - \bar{\boldsymbol{\theta}})^{\mathsf{T}}\frac{\partial}{\partial\boldsymbol{\theta}_{\mathsf{SUP}}}g(\bar{\mathbf{\Theta}}) + (\widehat{\boldsymbol{\xi}}_{\mathsf{SUP}} - \bar{\boldsymbol{\xi}})^{\mathsf{T}}\frac{\partial}{\partial\boldsymbol{\xi}_{\mathsf{SUP}}}g(\bar{\mathbf{\Theta}}) + O_{\mathbb{P}}\left(n^{-1}\right),$$

therefore

$$\sqrt{n}\left\{g(\widehat{\mathbf{\Theta}}_{\mathsf{SUP}}) - g(\bar{\mathbf{\Theta}})\right\} = \sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathsf{SUP}} - \bar{\boldsymbol{\theta}})^{\mathsf{T}}\frac{\partial}{\partial\boldsymbol{\theta}_{\mathsf{SUP}}}g(\bar{\mathbf{\Theta}}) + \sqrt{n}(\widehat{\boldsymbol{\xi}}_{\mathsf{SUP}} - \bar{\boldsymbol{\xi}})^{\mathsf{T}}\frac{\partial}{\partial\boldsymbol{\xi}_{\mathsf{SUP}}}g(\bar{\mathbf{\Theta}}) + o_{\mathbb{P}}(1).$$

Using the Q-function and propensity score function influence functions we can write

$$\sqrt{n}\left\{g(\widehat{\boldsymbol{\Theta}}_{\mathsf{SUP}}) - g(\bar{\boldsymbol{\Theta}})\right\} = \frac{\partial}{\partial \boldsymbol{\theta}_{\mathsf{SUP}}} g(\bar{\boldsymbol{\Theta}}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}^{\theta}_{\mathsf{SUP}}(\mathbf{L}_i) + \frac{\partial}{\partial \boldsymbol{\xi}} g(\bar{\boldsymbol{\Theta}}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}^{\xi}_{\mathsf{SUP}}(\mathbf{L}_i) + o_{\mathbb{P}}(1)$$

□

(a) $n = 135$ and $N = 1272$

	Supervised		Semi-Supervised									
			Random Forests					Basis Expansion				
Parameter	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
β_{11}=1.2	0.05	0.09	0.03	0.06	0.05	0.88	1.65	0.03	0.06	0.05	0.89	1.60
β_{12}=0	0.00	0.06	0.00	0.04	0.04	0.90	1.57	0.00	0.04	0.04	0.91	1.62
β_{13}=-0.4	0	0.07	-0.01	0.05	0.04	0.92	1.53	0	0.05	0.05	0.93	1.56
β_{14}=-0.3	0.00	0.07	-0.01	0.04	0.04	0.93	1.67	0	0.04	0.04	0.93	1.64
β_{15}=0	0.00	0.08	0.00	0.04	0.04	0.93	1.69	0.00	0.04	0.04	0.92	1.69
β_{16}=0	0	0.07	0.00	0.04	0.04	0.93	1.67	0.00	0.04	0.04	0.93	1.74
β_{17}=0	0.00	0.08	0.00	0.05	0.04	0.92	1.62	0.00	0.05	0.04	0.92	1.62
γ_{11}=0.1	-0.01	0.14	0.00	0.09	0.08	0.91	1.55	0	0.09	0.07	0.89	1.55
γ_{12}=0	-0.01	0.09	-0.01	0.06	0.05	0.92	1.53	-0.01	0.06	0.06	0.93	1.51
γ_{13}=0	0	0.08	0	0.05	0.05	0.93	1.58	0	0.05	0.05	0.94	1.58
γ_{14}=0	0	0.08	0.00	0.05	0.05	0.93	1.58	0	0.05	0.05	0.93	1.58
γ_{15}=0	0.00	0.09	0.00	0.05	0.05	0.92	1.59	0	0.05	0.05	0.95	1.65
γ_{16}=-0.1	0	0.09	0	0.06	0.05	0.92	1.52	0	0.06	0.05	0.93	1.49
β_{21}=0.1	0.00	0.10	-0.01	0.15	0.13	0.91	0.71	0	0.14	0.13	0.93	0.75
β_{22}=0.6	0	0.13	0.01	0.11	0.10	0.91	1.16	0	0.11	0.11	0.94	1.18
β_{23}=0	0.00	0.06	0.00	0.04	0.04	0.93	1.44	0.00	0.04	0.04	0.93	1.47
β_{24}=-0.2	0.00	0.06	0	0.05	0.04	0.89	1.16	0	0.05	0.05	0.93	1.20
β_{25}=-0.2	0.00	0.05	0	0.05	0.04	0.90	1.13	0	0.04	0.04	0.92	1.18
β_{26}=0	0.00	0.04	0.00	0.02	0.02	0.94	1.50	0.00	0.02	0.02	0.94	1.50
β_{27}=0	0.00	0.04	0.00	0.03	0.02	0.94	1.52	0.00	0.02	0.02	0.94	1.58
β_{28}=0	0.00	0.05	0.00	0.04	0.03	0.92	1.49	0.00	0.04	0.03	0.92	1.49
β_{29}=0	0	0.12	0.00	0.08	0.07	0.91	1.49	0.00	0.08	0.08	0.93	1.52
β_{210}=-0.2	0.00	0.11	0	0.07	0.07	0.94	1.54	0.00	0.07	0.07	0.94	1.57
β_{211}=-0.1	0.01	0.11	0.00	0.07	0.07	0.94	1.54	0.00	0.07	0.07	0.93	1.56
γ_{21}=0.1	0.01	0.16	0.01	0.11	0.10	0.92	1.47	0.01	0.11	0.10	0.94	1.51
γ_{22}=0	0.00	0.08	0.00	0.06	0.05	0.94	1.47	0.00	0.06	0.06	0.93	1.50
γ_{23}=0	0	0.08	0.00	0.06	0.05	0.94	1.45	0.00	0.05	0.05	0.94	1.48
γ_{24}=0	0	0.07	0.00	0.05	0.05	0.93	1.43	0.00	0.05	0.05	0.94	1.46
γ_{25}=0	0	0.07	0	0.05	0.05	0.94	1.48	0	0.05	0.05	0.94	1.48
γ_{26}=0	0	0.18	0	0.12	0.11	0.92	1.45	0	0.12	0.11	0.94	1.52
γ_{27}=-0.2	-0.01	0.16	-0.01	0.11	0.10	0.93	1.47	-0.01	0.11	0.10	0.94	1.48
γ_{28}=-0.1	-0.01	0.15	-0.01	0.10	0.10	0.94	1.54	-0.01	0.10	0.10	0.94	1.57

Table B.1: *Bias, empirical standard error (ESE) of the supervised and the SSL estimators with either random forest imputation or basis expansion imputation strategies for $\bar{\boldsymbol{\theta}}$ when (a) $n = 135$ and $N = 1272$ under the EHR simulation setting. For the SSL estimators, we also obtain the average of the estimated standard errors (ASE) as well as the empirical coverage probabilities (CovP) of the 95% confidence intervals.*

(b) $n = 500$ and $N = 10,000$

	Supervised		Semi-Supervised									
			Random Forests					Basis Expansion				
Parameter	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
β_{11}=1.2	0.01	0.05	0.00	0.02	0.02	0.91	2.09	0.00	0.02	0.02	0.92	2.00
β_{12}=0	0.00	0.03	0.00	0.01	0.01	0.91	2.07	0.00	0.01	0.01	0.92	2.07
β_{13}=-0.4	0.00	0.04	0	0.02	0.02	0.92	2.05	0	0.02	0.02	0.92	2.05
β_{14}=-0.3	0	0.04	0	0.02	0.01	0.92	2.06	0	0.02	0.02	0.92	2.06
β_{15}=0	0.00	0.04	0	0.02	0.02	0.94	2.18	0	0.02	0.02	0.94	2.06
β_{16}=0	0	0.04	0.00	0.02	0.02	0.94	2.18	0.00	0.02	0.02	0.94	2.18
β_{17}=0	0.00	0.04	0.00	0.02	0.02	0.93	2.06	0.00	0.02	0.02	0.94	2.06
γ_{11}=0.1	0	0.07	0	0.03	0.03	0.91	2.00	0	0.03	0.03	0.91	2.00
γ_{12}=0	-0.01	0.05	0	0.02	0.02	0.90	2.00	0	0.02	0.02	0.89	2.00
γ_{13}=0	0.00	0.04	0.00	0.02	0.02	0.92	2.00	0.00	0.02	0.02	0.91	1.90
γ_{14}=0	0	0.04	0.00	0.02	0.02	0.94	2.00	0.00	0.02	0.02	0.94	1.90
γ_{15}=0	0.00	0.04	0.00	0.02	0.02	0.94	2.16	0.00	0.02	0.02	0.94	2.05
γ_{16}=-0.1	0	0.04	0	0.02	0.02	0.93	2.05	0	0.02	0.02	0.92	1.95
β_{21}=0.1	0.00	0.05	0.00	0.04	0.04	0.95	1.16	0.00	0.04	0.05	0.96	1.13
β_{22}=0.6	0	0.07	0	0.04	0.04	0.95	1.74	0	0.04	0.04	0.96	1.69
β_{23}=0	0.00	0.03	0.00	0.01	0.01	0.94	1.87	0.00	0.01	0.01	0.94	1.87
β_{24}=-0.2	0.00	0.03	0.00	0.02	0.02	0.94	1.71	0.00	0.02	0.02	0.95	1.71
β_{25}=-0.2	0.00	0.02	0	0.01	0.01	0.94	1.60	0	0.01	0.01	0.95	1.60
β_{26}=0	0.00	0.02	0.00	0.01	0.01	0.92	1.90	0.00	0.01	0.01	0.93	1.90
β_{27}=0	0.00	0.02	0.00	0.01	0.01	0.94	1.89	0.00	0.01	0.01	0.94	1.89
β_{28}=0	0.00	0.03	0.00	0.01	0.01	0.94	1.92	0.00	0.01	0.01	0.94	1.92
β_{29}=0	0.00	0.06	0.00	0.03	0.03	0.92	1.94	0.00	0.03	0.03	0.93	1.88
β_{210}=-0.2	0	0.05	0	0.03	0.03	0.94	2.00	0.00	0.03	0.03	0.94	2.00
β_{211}=-0.1	0.00	0.06	0.00	0.03	0.03	0.94	2.00	0.00	0.03	0.03	0.94	2.00
γ_{21}=0.1	0	0.08	0.00	0.04	0.04	0.94	1.98	0.00	0.04	0.04	0.94	1.98
γ_{22}=0	0.00	0.04	0.00	0.02	0.02	0.93	1.95	0.00	0.02	0.02	0.93	1.86
γ_{23}=0	0	0.04	0	0.02	0.02	0.94	1.81	0	0.02	0.02	0.93	1.90
γ_{24}=0	0	0.03	0.00	0.02	0.02	0.94	1.83	0.00	0.02	0.02	0.95	1.83
γ_{25}=0	0	0.04	0	0.02	0.02	0.94	1.84	0	0.02	0.02	0.94	1.84
γ_{26}=0	-0.01	0.09	0	0.04	0.04	0.93	2.00	0	0.04	0.04	0.93	2.00
γ_{27}=-0.2	0.01	0.08	0.00	0.04	0.04	0.94	1.98	0.00	0.04	0.04	0.94	1.98
γ_{28}=-0.1	0.00	0.08	0.00	0.04	0.04	0.94	1.95	0.00	0.04	0.04	0.94	1.95

Table B.2: *Bias, empirical standard error (ESE) of the supervised and the SSL estimators with either random forest imputation or basis expansion imputation strategies for $\bar{\boldsymbol{\theta}}$ when (b) $n = 500$ and $N = 10,000$ under the EHR simulation setting. For the SSL estimators, we also obtain the average of the estimated standard errors (ASE) as well as the empirical coverage probabilities (CovP) of the 95% confidence intervals.*

(a) $n = 135$ and $N = 1272$

	Supervised		Semi-Supervised									
			Random Forests					Basis Expansion				
Parameter	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
β_{11}=4.9	0.04	0.34	0.01	0.22	0.18	0.91	1.58	0.01	0.20	0.17	0.90	1.70
β_{12}=1.1	-0.03	0.42	0.00	0.26	0.24	0.94	1.61	0.01	0.25	0.23	0.92	1.68
γ_{11}=1.4	-0.03	0.41	0.00	0.26	0.24	0.93	1.57	0.00	0.24	0.23	0.93	1.68
γ_{12}=-2.6	0.04	0.58	-0.01	0.36	0.34	0.94	1.61	-0.02	0.35	0.31	0.90	1.69
β_{21}=0.1	0.00	0.10	0.00	0.13	0.12	0.94	0.82	0.00	0.16	0.17	0.94	0.64
β_{22}=3	0.00	0.33	0.00	0.24	0.23	0.93	1.39	0	0.26	0.25	0.93	1.30
β_{23}=0	-0.01	0.34	-0.01	0.24	0.22	0.93	1.43	-0.01	0.24	0.24	0.94	1.39
β_{24}=0.1	0	0.43	0	0.29	0.28	0.94	1.49	0	0.30	0.29	0.94	1.46
β_{25}=-0.5	0.01	0.15	0	0.09	0.09	0.93	1.62	0.00	0.09	0.09	0.93	1.71
β_{26}=-0.4	0.03	0.48	0.01	0.37	0.35	0.93	1.29	0.01	0.41	0.40	0.94	1.16
γ_{21}=0.8	0.00	0.34	0.01	0.21	0.20	0.93	1.61	0.00	0.20	0.19	0.94	1.71
γ_{22}=0.2	-0.02	0.45	-0.01	0.28	0.28	0.95	1.60	-0.01	0.27	0.26	0.94	1.70
γ_{23}=0.5	0	0.18	0.01	0.11	0.11	0.94	1.59	0.00	0.11	0.11	0.94	1.68

(b) $n = 500$ and $N = 10,000$

	Supervised		Semi-Supervised									
			Random Forests					Basis Expansion				
Parameter	Bias	ESE	Bias	ESE	ASE	CovP	RE	Bias	ESE	ASE	CovP	RE
β_{11}=4.9	0.00	0.17	0	0.10	0.09	0.91	1.72	0	0.10	0.08	0.92	1.79
β_{12}=1.1	0	0.22	0.00	0.12	0.11	0.93	1.80	0.00	0.12	0.11	0.93	1.86
γ_{11}=1.4	0.01	0.22	0.01	0.12	0.11	0.92	1.76	0.01	0.12	0.11	0.92	1.80
γ_{12}=-2.6	0	0.29	0	0.17	0.16	0.93	1.73	-0.01	0.16	0.15	0.93	1.80
β_{21}=0.1	-0.01	0.05	0	0.05	0.05	0.94	1.06	0	0.07	0.08	0.95	0.74
β_{22}=3	0.00	0.17	0.00	0.11	0.10	0.93	1.60	0.00	0.12	0.11	0.94	1.45
β_{23}=0	0.00	0.17	0.00	0.10	0.10	0.95	1.66	0.00	0.11	0.11	0.95	1.54
β_{24}=0.1	0.02	0.23	0.01	0.13	0.12	0.94	1.77	0.01	0.14	0.13	0.94	1.68
β_{25}=-0.5	0.00	0.07	0.00	0.04	0.04	0.93	1.74	0.00	0.04	0.04	0.94	1.78
β_{26}=-0.4	-0.01	0.25	-0.01	0.17	0.15	0.93	1.51	-0.01	0.19	0.18	0.94	1.31
γ_{21}=0.8	0.00	0.17	0.00	0.10	0.09	0.93	1.80	0.00	0.09	0.09	0.93	1.86
γ_{22}=0.2	-0.01	0.23	0	0.13	0.12	0.93	1.81	0	0.13	0.12	0.94	1.83
γ_{23}=0.5	0.00	0.09	0.00	0.05	0.05	0.94	1.78	0.00	0.05	0.05	0.95	1.81

Table B.3: *Bias, empirical standard error (ESE) of the supervised and the SSL estimators with either random forest imputation or basis expansion imputation strategies for $\bar{\boldsymbol{\theta}}$ when (a) $n = 135$ and $N = 1272$ and (b) $n = 500$ and $N = 10,000$ under the continuous outcome simulation setting. For the SSL estimators, we also obtain the average of the estimated standard errors (ASE) as well as the empirical coverage probabilities (CovP) of the 95% confidence intervals.*

Appendix C

Appendix to Chapter 3

C.1 Fisher Consistency Results

In this Section we write the proofs for the theoretical results shown in Section 3.3.

Proof of Lemma 3.3.1. Our assumptions imply that

$$\phi'(x) - \phi'(-x) = 0 \quad \text{for all } x \in \mathbb{R}.$$

Let us denote

$$C_\eta(x) = \eta\phi(x) + (1-\eta)\phi(-x).$$

Then

$$C'_\eta(x) = \eta\phi'(x) - (1-\eta)\phi'(-x) = (2\eta - 1)\phi'(x)$$

When $\eta = 1/2$, $C_\eta(x) = C/2$ does not depend on x. For $\eta > 1/2$, $C'_\eta(x) \geq 0$ because $\phi'(x) \geq 0$. Therefore, $C_\eta(x)$ is non-decreasing. we see that $\lim_{x\to\infty} C_\eta(x) = C\eta > C/2 = C_\eta(0)$. Therefore, either $\sup_{x\in\mathbb{R}} C_\eta(x)$ is never attained or it is attained by an interval of the form $[x, \infty)$ for some $x > 0$. In either case,

$$\sup_{x\in\mathbb{R}} C_\eta(x) = C\eta.$$

Clearly,

$$\sup_{x\leq 0} C_\eta(x) = C_\eta(0) = C\eta/2 < \sup_{x\in\mathbb{R}} C_\eta(x).$$

Similarly, when $0 \leq \eta < 1/2$, we can show that $C'_\eta(x) < 0$ and

$$\sup_{x\in\mathbb{R}} C_\eta(x) = \lim_{x\to-\infty} C_\eta(x) = (1-\eta)C.$$

In this case also, either $\sup_{x\in\mathbb{R}} C_\eta(x)$ is never attained or it is attained by an interval of the form $(\infty, -x]$ for some $x > 0$. Also,

$$\sup_{x\geq 0} C_\eta(x) = C_\eta(0) = C\eta/2 < \sup_{x\in\mathbb{R}} C_\eta(x).$$

Thus, Condition 3.3.2 is satisfied. Also,

$$\sup_{x\in\mathbb{R}} C_\eta(x) = C\max\{\eta, 1-\eta\},$$

which implies that Condition 3.3.2 is also satisfied. □

Proof of Corollary 3.3.1.1. This follows straight forward from definition of the functions $\phi(x)$. We show it for $\phi(x) = 1 + \frac{x}{1+|x|}$.

1) Note that $\frac{x}{1+|x|} \in (0,1)$ for all $x \in \mathbb{R}$, thus $\phi(x) > 0$.

2) $\phi(x) + \phi(-x) = 1 + \frac{x}{1+|x|} + 1 + \frac{-x}{1+|x|} = 2$.

3) $\lim_{x\to\infty} \phi(x) = 2$, and 4) $\lim_{x\to-\infty} \phi(x) = 0$.

The rest of the example function $\phi(\cdot)$ can be shown in a similar fashion straight from definitions.

□

Proof of Lemma 3.3.2. Writing $V_\psi(f_1, f_2)$ as in (3.2) yields

$$
\begin{aligned}
&V_\psi(f_1, f_2) \\
&\quad = \mathbb{E}\Big[\frac{(Y_2+Y_3)\psi\Big(A_1 f_1(H_1), A_2 f_2(H_2)\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\Big\{\frac{(Y_2+Y_3)\phi_1\Big(A_1 f_1(H_1)\Big)\phi_2\Big(A_2 f_2(H_2)\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big| A_1, A_2, H_1, O_2, Y_2\Big\}\Big] \\
&\quad = \mathbb{E}\Big[\pi_2(1 \mid H_2)\mathbb{E}\Big\{\frac{(Y_2+Y_3)\phi_1\Big(A_1 f_1(H_1)\Big)\phi_2\Big(f_2(H_2)\Big)}{\pi_1(A_1 \mid H_1)\pi_2(1 \mid H_2)}\Big| A_1, A_2 = 1, H_1, O_2, Y_2\Big\}\Big] \\
&\qquad + \mathbb{E}\Big[\pi_2(-1 \mid H_2)\mathbb{E}\Big\{\frac{(Y_2+Y_3)\phi_1\Big(A_1 f_1(H_1)\Big)\phi_2\Big(-f_2(H_2)\Big)}{\pi_1(A_1 \mid H_1)\pi_2(-1 \mid H_2)}\Big| A_1, A_2 = -1, H_1, O_2, Y_2\Big\}\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\Big\{\frac{(Y_2+Y_3)\phi_1\Big(A_1 f_1(H_1)\Big)\phi_2\Big(f_2(H_2)\Big)}{\pi_1(A_1 \mid H_1)}\Big| A_1, A_2 = 1, H_1, O_2, Y_2\Big\}\Big] \\
&\qquad + \mathbb{E}\Big[\mathbb{E}\Big\{\frac{(Y_2+Y_3)\phi_1\Big(A_1 f_1(H_1)\Big)\phi_2\Big(-f_2(H_2)\Big)}{\pi_1(A_1 \mid H_1)}\Big| A_1, A_2 = -1, H_1, O_2, Y_2\Big\}\Big] \\
&\quad = \mathbb{E}\Big[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\Big\{E[Y_2+Y_3|A_1, A_2 = 1, H_1, O_2, Y_2]\phi_2\Big(f_2(H_2)\Big) \\
&\qquad + \mathbb{E}[Y_2+Y_3|A_1, A_2 = -1, H_1, O_2, Y_2]\phi_2\Big(-f_2(H_2)\Big)\Big\}\Big] \\
&\quad = \mathbb{E}\Big[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\Big\{T(A_1, 1, H_1, O_2, Y_2)\phi_2\Big(f_2(H_2)\Big) \\
&\qquad + T(A_1, -1, H_1, O_2, Y_2)\phi_2\Big(-f_2(H_2)\Big)\Big\}\Big]
\end{aligned}
$$

where for a_1, a_2,

$$T(a_1, a_2, H_1, O_2, Y_2) = \mathbb{E}[Y_2+Y_3|a_1, a_2, H_1, O_2, Y_2] = Y_2 + \mathbb{E}[Y_3 \mid a_1, a_2, H_1, O_2, Y_2].$$

For $a_1 \in \{\pm 1\}$, and $H_2 \in \mathcal{X}_2$, define the function

$$\eta(a_1, H_1, O_2, Y_2) = \frac{T(a_1, 1, H_1, O_2, Y_2)}{T(a_1, 1, H_1, O_2, Y_2) + T(a_1, -1, H_1, O_2, Y_2)}.$$

We see that the last expression equals

$$\mathbb{E}\Big[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\Big(T(A_1,1,H_1,O_2,Y_2)+T(A_1,1,H_1,O_2,Y_2)\Big)\Big\{\eta(A_1,H_1,O_2,Y_2)\phi_2\Big(f_2(H_2)\Big) + \Big(1-\eta(A_1,H_1,O_2,Y_2)\Big)\phi_2\Big(-f_2(H_2)\Big)\Big\}\Big].$$

Therefore, for fixed a_1 and h_2, $z = f_2(H_2)$ maximizes

$$\eta(a_1,h_1,o_2,y_2)\phi_2(z) + \Big(1-\eta(a_1,h_1,o_2,y_2)\Big)\phi_2(-z)$$

By our assumption on ϕ_2, ϕ_2 satisfies Condition 3.3.2. Therefore, if $\eta(a_1,h_1,o_2,y_2) \neq 0$, then

$$sign(z)\Big(2\eta(a_1,h_1,o_2,y_2)-1\Big) > 0.$$

Hence, if $\mathbb{E}[Y_2+Y_3|a_1,1,H_1,O_2,Y_2] \neq \mathbb{E}[Y_2+Y_3|a_1,-1,H_1,O_2,Y_2]$, then we have

$$sign\Big(\tilde{f}_2(a_1,H_1,O_2,Y_2)\Big) = \underset{a_2\mathbb{E}[Y_2+Y_3|a_1,a_2,H_1,O_2,Y_2]=\text{argmax}_{a_2\mathbb{E}[Y_2|a_1,a_2,H_1,O_2,Y_2]},}{\text{argmax}}$$

or

$$\tilde{d}_2(a_1,H_1,O_2,Y_2) \in \underset{a_2\mathbb{E}[Y_2|a_1,a_2,H_1,O_2,Y_2].}{\text{argmax}}$$

Also noting that Condition 3.3.2 implies

$$\begin{aligned}&\max_x\Big\{T(A_1,1,H_1,O_2,Y_2)\phi_2(x)+T(A_1,-1,H_1,O_2,Y_2)\phi_2(-x)\Big\}\\ &= C\max\Big\{T(A_1,1,H_1,O_2,Y_2),T(A_1,-1,H_1,O_2,Y_2)\Big\}\end{aligned}$$

and using $\phi_1 > 0$, we also deduce

$$
\begin{aligned}
&\sup_{f_2} \mathbb{E}\Bigg[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\Big\{T(A_1,1,H_1,O_2,Y_2)\phi_1\Big(f_2(H_1,O_2,Y_2,A_1)\Big) \\
&\quad + T(A_1,-1,H_1,O_2,Y_2)\phi_1\Big(-f_2(H_1,O_2,Y_2,A_1)\Big)\Big\}\Bigg] \\
&= \mathbb{E}\Bigg[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)} \sup_z \Big\{T(A_1,1,H_1,O_2,Y_2)\phi(z) + T(A_1,-1,H_1,O_2,Y_2)\phi(-z)\Big\}\Bigg] \\
&= C\mathbb{E}\Bigg[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)} \max\Big\{Y_2 + \mathbb{E}[Y_3|A_1,1,H_1,O_2,Y_2], Y_2 + \mathbb{E}[Y_3|A_1,-1,H_1,O_2,Y_2]\Big\}\Bigg] \\
&= C\mathbb{E}\Bigg[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\Big(Y_2 + \max\Big\{\mathbb{E}[Y_3|A_1,1,H_1,O_2,Y_2], \mathbb{E}[Y_2|A_1,-1,H_1,O_2,Y_2]\Big\}\Big)\Bigg] \\
&= C\mathbb{E}\Bigg[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\mathbb{E}\Big[Y_2 + \max\Big\{\mathbb{E}[Y_3|A_1,1,H_1,O_2,Y_2], \mathbb{E}[Y_2|A_1,-1,H_1,O_2,Y_2]\Big\}\Big|A_1,O_1\Big]\Bigg] \\
&= C\mathbb{E}\Bigg[\frac{\phi_1\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\mathbb{E}\Big\{Y_2 + U_3^*(H_2)\Big|A_1,H_1\Big\}\Bigg] \\
&= C\mathbb{E}\Bigg[\sum_{a_1=\pm 1}\phi_1(a_1 f_1(o_1))\mathbb{E}\Big\{Y_2 + U_3^*(H_2)\Big|a_1,H_1\Big\}\Bigg].
\end{aligned}
$$

Since ϕ_1 satisfies Condition 3.3.2, we obtain that

$$
\tilde{f}_1(H_1) = \underset{x \sum_{a_1=\pm 1}\phi_1(a_1 x)\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|a_1,H_1\Big]}{\operatorname{argmax}} \tag{C.1}
$$

and it holds that

$$
\tilde{d}_1(H_1) \in \underset{a_1\in\{\pm 1\}\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|a_1,H_1\Big]}{\operatorname{argmax}}.
$$

□

C.2 Regret Bound Results

In this Section we write the proofs for the theoretical results shown in Section 3.4.

C.2.1 Relating Regret and ψ-Regret

Before we go on to prove Theorem 3.4.2, we introduce a useful Lemma and some definitions.

Lemma C.2.1. *Suppose ϕ satisfies the conditions of Lemma 3.3.1 and is strictly increasing. Then for an $\eta \in (0,1)$ such that $\eta \neq 1/2$,*

$$C_\eta(x) = \eta\phi(x) + (1-\eta)\phi(-x) \tag{C.2}$$

is maximized at $x = \infty$ if $2\eta > 1$ and at $x = -\infty$ if $2\eta < 1$.

The above result tells us that $\tilde{f}_1$ and $\tilde{f}_2$ do not take finite values if ϕ is a function following the conditions in Lemma 3.3.1. The proof can be found in Appendix C.4. Now suppose ϕ is as in Lemma C.4.1. Let us define

$$H(\eta) = \sup_{x\in\mathbb{R}} C_\eta(x), \quad \eta \in [0,1].$$

where C_η is as in (C.2) depends on ϕ. Since ϕ satisfies Condition 3.3.2, $H(\eta) = C\max\{\eta, 1-\eta\}$. Also define

$$H^-(\eta) = \sup_{\substack{x\in\mathbb{R}:\\ x(2\eta-1)\leq 0}} C_\eta(x).$$

Because $C'_\eta(x) = (2\eta-1)\phi'(x)$, it follows that

$$H^-(\eta) = C_\eta(0) = \phi(0) = C/2.$$

The quantity

$$\tilde{\psi}(\eta) := H\left(\frac{1+\eta}{2}\right) - H^-\left(\frac{1+\eta}{2}\right)$$

plays an important role in the approximation error for binary classification [78]. In our case,

$$\tilde{\psi}(\eta) = C/2\left(\max\{1+\eta, 1-\eta\} - 1\right) = C\eta/2$$

is a closed convex function.

Proof of Theorem 3.4.2. Note that for $\psi(x,y) = \phi(x)\phi(y)$,

$$
\begin{aligned}
V^{\psi}(\tilde{f}_1, \tilde{f}_2) - V^{\psi}(f_1, f_2) &= \mathbb{E}\Big[\frac{(Y_2+Y_3)\Big(\phi(A_1\tilde{f}_1(H_1))\phi(A_2\tilde{f}_2(H_2)) - \phi(A_1 f_1(H_1))\phi(A_2 f_2(H_2))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&= \mathbb{E}\Big[\frac{(Y_2+Y_3)\Big(\phi(A_1\tilde{f}_1(H_1))\phi(A_2\tilde{f}_2(H_2)) - \phi(A_1 f_1(H_1))\phi(A_2\tilde{f}_2(H_2))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&\quad + \mathbb{E}\Big[\frac{(Y_2+Y_3)\Big(\phi(A_1 f_1(H_1))\phi(A_2\tilde{f}_2(H_2)) - \phi(A_1 f_1(H_1))\phi(A_2 f_2(H_2))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&= \mathbb{E}\Big[\frac{(Y_2+Y_3)\phi(A_1 f_1(H_1))\Big(\phi(A_2\tilde{f}_2(H_2)) - \phi(A_2 f_2(H_2))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&\quad + \mathbb{E}\Big[\frac{(Y_2+Y_3)\phi(A_2\tilde{f}_2(H_2))\Big(\phi(A_1\tilde{f}_1(H_1)) - \phi(A_1 f_1(H_1))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[\frac{(Y_2+Y_3)\phi\Big(A_1 f_1(H_1)\Big)\Big(\phi(A_2\tilde{f}_2(H_2)) - \phi(A_2 f_2(H_2))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big] \\
&\quad + \mathbb{E}\Big[\frac{(Y_2+Y_3)1[A_2 d_2^*(H_2) > 0]\Big(\phi(A_1\tilde{f}_1(H_1)) - \phi(A_1 f_1(H_1))\Big)}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big]
\end{aligned}
$$

where (a) follows from Lemma C.4.1 and the fact that $sign(\tilde{f}) = \tilde{d} = d^*$. Note that,

$$
\begin{aligned}
&V^{\psi}(\tilde{f}_1, \tilde{f}_2) - V^{\psi}(f_1, f_2) \\
&= \mathbb{E}\Big[\frac{\phi\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)} \underbrace{\mathbb{E}\Big[\frac{(Y_2+Y_3)\Big(\phi(A_2\tilde{f}_2(H_2)) - \phi(A_2 f_2(H_2))\Big)}{\pi_2(A_2 \mid H_2)}\Big| H_2\Big]}_{T_1}\Big] \\
&+ \mathbb{E}\Big[\underbrace{\mathbb{E}\Big[\mathbb{E}\Big[\frac{(Y_2+Y_3)1[A_2 d_2^*(H_2) > 0]}{\pi_2(A_2 \mid H_2)}\Big| A_1, H_1\Big]\frac{\Big(\phi(A_1\tilde{f}_1(H_1)) - \phi(A_1 f_1(H_1))\Big)}{\pi_1(A_1 \mid H_1)}\Big| H_1\Big]}_{T_2}\Big]
\end{aligned}
$$

Noting $H_2 = (O_1, A_1, Y_2, O_2)$, we calculate

$$\begin{aligned}
&\mathbb{E}\left[\frac{(Y_2+Y_3)1[A_2 d_2^*(H_2)>0]}{\pi_2(A_2 \mid H_2)}\middle| A_1, H_1\right] \\
&\quad= \mathbb{E}\left[\frac{(Y_2+\mathbb{E}[Y_3|H_2,A_2])1[A_2 d_2^*(H_2)>0]}{\pi_2(A_2 \mid H_2)}\middle| A_1, H_1\right] \\
&\quad= \mathbb{E}\Big[(Y_2+\mathbb{E}[Y_3|H_2,A_2=1])1[d_2^*(H_2)>0]\Big| A_1, H_1\Big] \\
&\quad+ \mathbb{E}\Big[(Y_2+\mathbb{E}[Y_3|H_2,A_2=-1])1[d_2^*(H_2)<0]\Big| A_1, H_1\Big] \\
&\quad= \mathbb{E}\Big[Y_2+\mathbb{E}[Y_3|H_2,A_2=d_2^*(H_2)]\Big| A_1, H_1\Big] \\
&\quad= \mathbb{E}\Big[Y_2+U_3^*(H_2)|H_2\Big]
\end{aligned}$$

Let us denote

$$\eta(H_1) = \frac{\underset{a=\pm 1}{\operatorname{argmax}}\, \mathbb{E}\left[Y_2+U_3^*(H_2)\middle| A_1=a, H_1\right]}{\mathbb{E}\left[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\middle| A_1=1, H_1\right]+\mathbb{E}\left[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\middle| A_1=-1, H_1\right]}.$$

Note that $\eta(H_1) > 1/2$. Therefore, $\tilde{\psi}\Big(2\eta(H_1)-1\Big)$ is well defined, and

$$H(\eta(H_1)) - H^-(\eta(H_1)) = \frac{C\Big(2\eta(H_1)-1\Big)}{2}. \tag{C.3}$$

Without loss of generality we assume that

$$\underset{a=\pm 1}{\operatorname{argmax}}\, \mathbb{E}\left[Y_2+U_3^*(A_1=a,H_1,O_2,Y_2)\middle| A_1=a, H_1\right]=1$$

because the proof for the other case will be similar. Then for T_2, we calculate that

$$
\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1,H_1\Big]\frac{\Big(\phi(A_1\tilde{f}_1(H_1))-\phi(A_1f_1(H_1))\Big)}{\pi_1(A_1\mid H_1)}\Big|H_1\Big]\\
&\quad=\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=1,H_1\Big]\Big(\phi(\tilde{f}_1(H_1))-\phi(f_1(H_1))\Big)\\
&\qquad+\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=-1,H_1\Big]\Big(\phi(-\tilde{f}_1(H_1))-\phi(-f_1(H_1))\Big)\\
&\quad=\Big(\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=1,H_1\Big]+\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=-1,H_1\Big]\Big)\\
&\qquad\Big(\Big\{\eta(H_1)\phi\Big(\tilde{f}_1(H_1)\Big)+\Big(1-\eta(H_1)\Big)\phi\Big(-\tilde{f}_1(H_1)\Big)\Big\}\\
&\qquad-\Big\{\eta(H_1)\phi\Big(f_1(H_1)\Big)+\Big(1-\eta(H_1)\Big)\phi\Big(-f_1(H_1)\Big)\Big\}\Big)\\
&\overset{(a)}{\geq}\Big(\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=1,H_1\Big]+\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=-1,H_1\Big]\Big)\\
&\qquad\Big(\Big\{\eta(H_1)\phi\Big(\tilde{f}_1(H_1)\Big)+\Big(1-\eta(H_1)\Big)\phi\Big(-\tilde{f}_1(H_1)\Big)\Big\}\\
&\qquad-\Big\{\eta(H_1)\phi\Big(f_1(H_1)\Big)+\Big(1-\eta(H_1)\Big)\phi\Big(-f_1(H_1)\Big)\Big\}\Big)1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\\
&\overset{(b)}{\geq}\Big(\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=1,H_1\Big]+\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=-1,H_1\Big]\Big)\\
&\qquad\Big(H(\eta(H_1))-H^-(\eta(H_1))\Big)1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\\
&\overset{(c)}{=}\frac{C|2\eta(H_1)-1|}{2}\Big(\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=1,H_1\Big]\\
&\qquad+\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=-1,H_1\Big]\Big)1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\\
&\quad=\frac{C}{2}\Big|\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=1,H_1\Big]-\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1=-1,H_1\Big]\Big|\\
&\qquad1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big].
\end{aligned}
$$

Here (a) follows from (C.1) and the fact that ϕ, Y_2, Y_3 are non-negative, (b) follows from the definition of H, H^-, and (c) follows from (C.3). Now it can be shown very easily that in this

case,

$$
\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1,H_1\Big]\frac{1[(A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)}\Big|H_1\Big]\\
&\quad=\mathbb{E}\Big[\mathbb{E}\Big[Y_2+U_3^*(H_2)\Big|A_1,H_1\Big]\frac{1[(A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)}\Big|H_1\Big]\\
&\qquad 1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\\
&\quad=\Big(\mathbb{E}\Big[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\Big|1,H_1\Big]1[\tilde{f}_1(H_1)>0]\\
&\qquad+\mathbb{E}\Big[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\Big|-1,H_1\Big]1[\tilde{f}_1(H_1)<0]\Big)\\
&\qquad-\Big(\mathbb{E}\Big[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\Big|1,H_1\Big]1[f_1(H_1)>0]\\
&\qquad+\mathbb{E}\Big[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\Big|-1,H_1\Big]1[f_1(H_1)<0]\Big)\\
&\qquad 1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\\
&\qquad\overset{(a)}{=}\Big(\mathbb{E}\Big[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\Big|1,H_1\Big]1[\tilde{f}_1(H_1)>0]\\
&\qquad+\mathbb{E}\Big[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\Big|-1,H_1\Big]1[\tilde{f}_1(H_1)<0]\Big)\\
&\qquad-\Big(\mathbb{E}\Big[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\Big|1,H_1\Big]1[\tilde{f}_1(H_1)<0]\\
&\qquad+\mathbb{E}\Big[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\Big|-1,H_1\Big]1[\tilde{f}_1(H_1)>0]\Big)\\
&\qquad 1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\\
&\quad=\Big|\mathbb{E}\Big[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\Big|1,H_1\Big]-\mathbb{E}\Big[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\Big|-1,H_1\Big]\Big|\\
&\qquad\times\Big(1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\Big)
\end{aligned}
$$

where (a) follows because $\tilde{d}_1=sign(\tilde{f}_1)=d_1^*$. Thus, it follows that

$$
\begin{aligned}
T_2\geq\frac{C}{2}&\Big|\mathbb{E}\Big[Y_2+U_3^*(A_1=1,H_1,O_2,Y_2)\Big|1,H_1\Big]-\mathbb{E}\Big[Y_2+U_3^*(A_1=-1,H_1,O_2,Y_2)\Big|-1,H_1\Big]\Big|\\
&\times\Big(1\Big[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\Big]\Big).
\end{aligned}
$$

Now,

$$
\begin{aligned}
&\left|\mathbb{E}\left[Y_2+U_3^*(H_2)\middle|1,H_1\right]-\mathbb{E}\left[Y_2+U_3^*(H_2)\middle|-1,H_1\right]\right|\left(1\left[sign\Big(f_1(H_1)\Big)\neq d_1^*(H_1)\right]\right)\\
&\quad=\mathbb{E}\left[Y_2+U_3^*(H_2)\middle|A_1=1,H_1\right]\Big(1[\tilde{f}_1(H_1)>0]-1[f_1(H_1)>0]\Big)\\
&\qquad+\mathbb{E}\left[Y_2+U_3^*(H_2)\middle|A_1=-1,H_1\right]\Big(1[\tilde{f}_1(H_1)<0]-1[f_1(H_1)<0]\Big)\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[Y_2+U_3^*(H_2)\middle|A_1,H_1\right]\frac{1[A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)}\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[Y_2+Y_3\middle|H_2,A_2=d_2^*(H_2)\right]\frac{1[A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)}\right]\\
&\quad\overset{(a)}{=}\mathbb{E}\left[1\Big[A_2\tilde{f}_2(H_2)>0\Big]\mathbb{E}\left[Y_2+Y_3\middle|H_2,A_2\right]\frac{1[A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)\pi_2(A_2\mid H_2)}\right]\\
&\quad=\mathbb{E}\left[(Y_2+Y_3)1\Big[A_2\tilde{f}_2(H_2)>0\Big]\frac{1[A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)\pi_2(A_2\mid H_2)}\right],
\end{aligned}
$$

where (a) follows because $sign(\tilde{f}_2(H_2))=d_2^*(H_2)$. Therefore,

$$
T_2\geq\frac{C}{2}\mathbb{E}\left[(Y_2+Y_3)1\Big[A_2\tilde{f}_2(H_2)>0\Big]\frac{1[A_1\tilde{f}_1(H_1)>0]-1[A_1f_1(H_1)>0]}{\pi_1(A_1\mid H_1)\pi_2(A_2\mid H_2)}\right]. \tag{C.4}
$$

On the other hand,

$$
\begin{aligned}
T_1 &= \mathbb{E}\Bigg[\frac{(Y_2+Y_3)\Big(\phi(A_2\tilde{f}_2(H_2))-\phi(A_2 f_2(H_2))\Big)}{\pi(A_2,H_2)}\Bigg|H_2\Bigg] \\
&= \mathbb{E}\Bigg[(Y_2+\mathbb{E}[Y_3|H_2,A_2=1])\Big(\phi(\tilde{f}_2(H_2))-\phi(f_2(H_2))\Big) \\
&\quad +(Y_2+\mathbb{E}[Y_3|H_2,A_2=-1])\Big(\phi(-\tilde{f}_2(H_2))-\phi(-f_2(H_2))\Big)\Bigg|H_2,A_1\Bigg] \\
&\overset{(a)}{\geq} \mathbb{E}\Bigg[(Y_2+\mathbb{E}[Y_3|H_2,A_2=1])\Big(\phi(\tilde{f}_2(H_2))-\phi(f_2(H_2))\Big) \\
&\quad +(Y_2+\mathbb{E}[Y_3|H_2,A_2=-1])\Big(\phi(-\tilde{f}_2(H_2))-\phi(-f_2(H_2))\Big)\Bigg|H_2,A_1\Bigg] \\
&\quad 1\Big[sign(f_2(H_2))\neq d_2^*(H_2)\Big] \\
&\overset{(b)}{\geq} \frac{C}{2}\Big|\mathbb{E}[Y_2|H_2,A_2=1]-\mathbb{E}[Y_2|H_2,A_2=-1]\Big|1\Big[sign(f_2(H_2))\neq d_2^*(H_2)\Big] \\
&= \frac{C}{2}\mathbb{E}\Bigg[\frac{(Y_2+Y_3)\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi(A_2,H_2)}\Bigg|H_2\Bigg]
\end{aligned}
$$

where (a) follows from the proof of Lemma 3.3.2 and (b) follows in the same way as the proof for T_2 using the properties of ϕ. Hence,

$$
\begin{aligned}
&\mathbb{E}\Bigg[\frac{\phi\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1\mid H_1)}T_1\Bigg] \\
&\quad\geq \frac{C}{2}\mathbb{E}\Bigg[\frac{\phi\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1\mid H_1)}\mathbb{E}\Bigg[\frac{(Y_2+Y_3)\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi(A_2,H_2)}\Bigg|H_2\Bigg]\Bigg].
\end{aligned}
$$

Also, since

$$
\mathbb{E}\Bigg[\frac{(Y_2+Y_3)\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi(A_2,H_2)}\Bigg|H_2\Bigg]
$$

is non-negative, it follows that

$$\begin{aligned}
&\mathbb{E}\Bigg[\frac{\phi\Big(A_1 f_1(H_1)\Big)}{\pi_1(A_1 \mid H_1)}\mathbb{E}\Bigg[\frac{(Y_2+Y_3)\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi(A_2,H_2)}\Bigg|H_2\Bigg]1\Big[A_1 f_1(H_1)>0\Big]\Bigg] \\
&\quad\geq \phi(0)\mathbb{E}\left[\frac{1\Big[A_1 f_1(H_1)>0\Big]}{\pi_1(A_1 \mid H_1)}\mathbb{E}\left[\frac{(Y_2+Y_3)\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi(A_2,H_2)}\Bigg|H_2\right]\right] \\
&\quad= \phi(0)\mathbb{E}\left[(Y_2+Y_3)\frac{1[A_1 f_1(H_1)>0]\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi_1(A_1 \mid H_1)\pi(A_2,H_2)}\right]
\end{aligned}$$

because ϕ is strictly increasing. Combining with (C.4), we obtain that

$$\begin{aligned}
&V^{\psi}(f_1,f_2)-V^{\psi}(f_1,f_2) \\
&\quad\geq \frac{C\phi(0)}{2}\mathbb{E}\left[(Y_2+Y_3)\frac{1[A_1 f_1(H_1)>0]\Big(1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_2(H_2)>0]\Big)}{\pi_1(A_1 \mid H_1)\pi(A_2,H_2)}\right] \\
&\qquad+\frac{C}{2}\mathbb{E}\left[(Y_2+Y_3)1\Big[A_2\tilde{f}_2(H_2)>0\Big]\frac{1[A_1\tilde{f}_1(H_1)>0]-1[A_1 f_1(H_1)>0]}{\pi_1(A_1 \mid H_1)\pi(A_2,H_2)}\right] \\
&\quad\geq \frac{C}{2}\min\Big(\phi(0),1\Big) \\
&\qquad\mathbb{E}\left[(Y_2+Y_3)\frac{1[A_2\tilde{f}_1(H_2)>0]1[A_2\tilde{f}_2(H_2)>0]-1[A_2 f_1(H_2)>0]1[A_2 f_2(H_2)>0]}{\pi_2(A_2 \mid H_2)\pi(A_2 \mid H_2)}\right] \\
&\quad= \frac{C}{2}\min\Big(\phi(0),1\Big)\Big(V(\tilde{f}_1,\tilde{f}_2)-V(f_1,f_2)\Big),
\end{aligned}$$

which completes the proof.

□

C.2.2 Proof of Approximation Error Results

Proof of Lemma 3.4.4. Lemma C.2.1 implies that

$$\begin{aligned}
V_{\psi}^{*} &= \sup_{f_1,f_2}\mathbb{E}\left[(Y_2+Y_3)\frac{\phi(A_1 f_1(H_1))\phi(A_2 f_2(H_2))}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\right] \\
&= 4\mathbb{E}\left[(Y_2+Y_3)\frac{1[A_1 G_1(H_1)>0]1[A_2 G_2(H_2)>0]}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\right].
\end{aligned}$$

Suppose f_1 and f_2 are such that $sign(f_1) = d_1 = sign(G_1)$ and $sign(f_2) = d_2 = sign(G_2)$. Then

$$
\begin{aligned}
&V_\psi^* - V_\psi(f_1, f_2) \\
&= \left| \mathbb{E}\left[(Y_2 + Y_3) \frac{\left\{ 4 \times 1[A_1 G_1(H_1) > 0, A_2 G_2(H_2) > 0] - \phi(A_1 f_1(H_1)) \phi(A_2 f_2(H_2)) \right\}}{\pi_1(A_1 \mid H_1) \pi_2(A_2 \mid H_2)} \right] \right| \\
&= \left| \mathbb{E}\left[(Y_2 + Y_3) \frac{4 \times 1[A_1 G_1(H_1) > 0] \Big(1[A_2 G_2(H_2) > 0] - \phi(A_2 f_2(H_2)) \Big)}{\pi_1(A_1 \mid H_1) \pi_2(A_2 \mid H_2)} \right] \right. \\
&\quad \left. + \mathbb{E}\left[(Y_2 + Y_3) \frac{\phi(A_2 G_2(H_2)) \Big(1[A_1 G_1(H_1) > 0] - \phi(A_1 f_1(H_1)) \Big)}{\pi_1(A_1 \mid H_1) \pi_2(A_2 \mid H_2)} \right] \right| \\
&\leq C\mathbb{E}\left[\left| 4 \times 1[A_1 G_1(H_1) > 0] - \phi(A_1 f_1(H_1)) \right| \right] + C\mathbb{E}\left[\left| 2 \times 1[A_2 G_2(H_2) > 0] - \phi(A_2 f_2(H_2)) \right| \right] \\
&\overset{(a)}{=} C\mathbb{E}\left[\left| 2 \times 1[A_1 f_1(H_1) > 0] - \phi(A_1 f_1(H_1)) \right| \right] + C\mathbb{E}\left[\left| 2 \times 1[A_2 f_2(H_2) > 0] - \phi(A_2 f_2(H_2)) \right| \right]
\end{aligned}
$$

where

$$C \geq \frac{\max(Y_2 + Y_3)}{\min \pi_1(A_1 \mid H_1) \min \pi_2(A_2 \mid H_2)} = \frac{2B_Y}{c}$$

is a constant and (a) follow because the pairs G_1 and f_1, and G_2 and f_2 have the same signs. Next we use the above to show the results follows for each function listed in Corollary 3.3.1.1.

Corollary 3.3.1.1 (i):

We have $\phi(x) = 1 + x/(1 + |x|)$. Then

$$
\begin{aligned}
\Big| 2 \times 1[x > 0] - \phi(x) \Big| =& \frac{1[x > 0]}{1 + x} + \frac{1[x < 0]}{1 - x} \\
=& \frac{1}{1 + |x|} \\
<& |x|^{-1}.
\end{aligned}
$$

Therefore,

$$V_\psi^* - V_\psi(f_1, f_2) \leq C\mathbb{E}\Big(|f_1(H_1)|^{-1} + |f_2(H_2)|^{-1} \Big).$$

Corollary 3.3.1.1 (ii):

Now $\phi(x) = 1 + \frac{2}{\pi}\arctan\left(\frac{\pi x}{2}\right)$. Therefore

$$\begin{aligned}
&\Big|2 \times 1[x > 0] - \phi(x)\Big| \\
&= 1[x > 0]\left(1 - \frac{2}{\pi}\arctan\left(\frac{\pi x}{2}\right)\right) + 1[x < 0]\left(1 + \frac{2}{\pi}\arctan\left(\frac{\pi x}{2}\right)\right) \\
&= 1[x > 0]\left(1 - \frac{2}{\pi}\arctan\left(\frac{\pi|x|}{2}\right)\right) + 1[x < 0]\left(1 - \frac{2}{\pi}\arctan\left(\frac{\pi|x|}{2}\right)\right) \\
&= 1 - \frac{2}{\pi}\arctan\left(\frac{\pi|x|}{2}\right) \\
&\overset{(b)}{=} \frac{2}{\pi}\int_0^\infty \frac{dy}{1+y^2} - \frac{2}{\pi}\int_0^{\pi|x|/2} \frac{dy}{1+y^2} \\
&= \frac{2}{\pi}\int_{\pi|x|/2}^\infty \frac{dy}{1+y^2} \\
&\leq \frac{2}{\pi}\int_{\pi|x|/2}^\infty \frac{dy}{y^2} \\
&\leq |x|^{-1}.
\end{aligned}$$

where (b) follows from $\int_{\mathbb{R}}(1+x^2)^{-1}dx = \pi/2$. This again yields,

$$V_\psi^* - V_\psi(f_1, f_2) \leq C\mathbb{E}\Big(|f_1(H_1)|^{-1} + |f_2(H_2)|^{-1}\Big).$$

Corollary 3.3.1.1 (iii):

Now with $\phi(x) = 1 + \frac{x}{\sqrt{1+x^2}}$, we have

$$\begin{aligned}
&\Big|2 \times 1[x > 0] - \phi(x)\Big| \\
&= 1[x > 0]\left(\frac{\sqrt{1+x^2} - x}{\sqrt{1+x^2}}\right) + 1[x < 0]\left(\frac{\sqrt{1+x^2} + x}{\sqrt{1+x^2}}\right) \\
&= \frac{\sqrt{1+x^2} - |x|}{\sqrt{1+x^2}}
\end{aligned}$$

which is not greater than

$$\frac{\sqrt{1+2|x|+x^2} - |x|}{\sqrt{1+x^2}} \leq \frac{1}{\sqrt{1+x^2}} \leq |x|^{-1}.$$

Thus,

$$V_\psi^* - V_\psi(f_1, f_2) \leq C\mathbb{E}\Big(|f_1(H_1)|^{-1} + |f_2(H_2)|^{-1}\Big)$$

Corollary 3.3.1.1 (iv):

Suppose $\phi(x) = 2(1 + \exp(-x))^{-1}$. Then

$$\begin{aligned}
&\Big|2 \times 1[x > 0] - \phi(x)\Big| \\
&\quad = 2 \times 1[x > 0]\Big(\frac{1}{1 + \exp(x)}\Big) + 2 \times 1[x < 0]\Big(\frac{1}{1 + \exp(-x)}\Big) \\
&\quad = \frac{2}{1 + \exp(|x|)}
\end{aligned}$$

which is bounded by $\exp(-|x|)$. Therefore,

$$V_\psi^* - V_\psi(f_1, f_2) \leq C\mathbb{E}\Big[\exp(-|f_1(H_1)|) + \exp(-|f_2(H_2)|)\Big].$$

Suppose $a_n \to \infty$ and G_1 and G_2 where G_1 and G_2 are as in (3.6). Then for ϕ as in Corollary 3.3.1.1 (i)-(iii),

$$V_\psi^* - V_\psi(a_n G_1, a_n G_2) \leq \frac{C}{a_n}\mathbb{E}\Big(|G_1(H_1)|^{-1} + |G_2(H_2)|^{-1}\Big)$$

next we have

$$\begin{aligned}
\mathbb{E}\Big(|G_1(H_1)|^{-1} + |G_2(H_2)|^{-1}\Big) &= \int_0^\infty \mathbb{P}\Big(|G_1(H_1)|^{-1} + |G_2(H_2)|^{-1} > x\Big)dx \\
&\leq \int_0^\infty \mathbb{P}\Big(|G_1(H_1)| < \frac{1}{x}\Big) + \mathbb{P}\Big(|G_2(H_2)| < \frac{1}{x}\Big)dx, \\
&\leq \int_0^\infty Ce^{-x\alpha}dx \\
&= C/\alpha,
\end{aligned}$$

which follows from using indentity $\mathbb{E}[X] = \int_0^\infty P(X > x)dx$ for $X \geq 0$, the union bound and Assumption 3.4.3. Therefore

$$V_\psi^* - V_\psi(a_n G_1, a_n G_2) = O(n^{-1}).$$

Next, for Corollary 3.3.1.1 (iv)

$$\begin{aligned}
V_\psi^* - V_\psi(a_nG_1, a_nG_2) \leq & C\mathbb{E}\Big[\exp(-a_n|G_1(H_1)|) + \exp(-a_n|G_2(H_2)|)\Big] \\
= & C\mathbb{E}\Big[\left(\exp(-a_n|G_1(H_1)|) + \exp(-a_n|G_2(H_2)|)\right) I\{G_1(H_1) < \delta\}\Big] \\
& + C\mathbb{E}\Big[\left(\exp(-a_n|G_1(H_1)|) + \exp(-a_n|G_2(H_2)|\right) I\{G_1(H_1) \geq \delta\})\Big] \qquad false \\
\leq & 2C\exp(-\alpha/\delta) + 2C\exp(-a_n)
\end{aligned}$$

where the last step follows from Assumption 3.4.3, choosing $\delta = O\left(a_n^{-1}\right)$ we get

$$V_\psi^* - V_\psi(a_nG_1, a_nG_2) = O(\exp(-a_n)),$$

which proves our result. □

Proof of Lemma 3.4.5. First notice that

$$\begin{aligned}
& |V_\psi(a_nG_1, a_nG_2) - V_\psi\Big(a_nC_y\tilde{f}_{n,1}, C_y\tilde{f}_{n,2}\Big)| \\
& \leq \mathbb{E}\left[\frac{Y_2 + Y_3}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big|\phi(A_1a_nG_1(H_1))\Big(\phi(a_nA_2G_2(H_2)) - \phi(a_nA_2C_y\tilde{f}_{n,2}(H_2))\Big)\Big|\right] \\
& + \mathbb{E}\left[\frac{Y_2 + Y_3}{\pi_1(A_1 \mid H_1)\pi_2(A_2 \mid H_2)}\Big|\phi(a_nA_2C_y\tilde{f}_{n,2}(H_2))\Big(\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\Big)\Big|\right] \\
& \leq C\Big\{\mathbb{E}\Big|\phi(a_nA_2G_2(H_2)) - \phi(a_nA_2C_y\tilde{f}_{n,2}(H_2))\Big| + \mathbb{E}\Big|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\Big|\Big\}
\end{aligned}$$

Which follows as ϕ is bounded, Y_i's are bounded above and π's are bounded below by Assumption 3.4.1.

Next, we focus on the second term as the analysis of the first term will be identical. For any $\delta > 0$ we have

$$\begin{aligned}
& \mathbb{E}\Big|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\Big| \\
& \quad = \mathbb{E}\Big\{\Big|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\Big|1[0 < |G_1(H_1)| < \delta]\Big\} \\
& \quad + \mathbb{E}\Big\{\Big|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\Big|1[\delta < |G_1(H_1)|]\Big\}.
\end{aligned}$$

By Assumption 3.4.3,

$$\mathbb{E}\left\{\left|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\right|1[0 < |G_1(H_1)| < \delta]\right\} \leq 2Ce^{-\alpha/\delta}, \tag{C.5}$$

where we used the fact that any ϕ in Corollary 3.3.1.1 is such that $0 < \phi < 2$. Note ϕ in Corollary 3.3.1.1(i)-(iii) is $0 < \phi < 1$.

We consider the case where ϕ is as in Corollary 3.3.1.1 (i)-(iii) first. If $G_1(H_1) > \delta$ and $\|\tilde{f}_{n,1} - G_1/C_y\|_\infty \leq \delta/2$, then $|\tilde{f}_{n,1}(H_1)| > \delta/2$, and $\tilde{f}_{n,1}(H_1))$ and $G_1(H_1)$ have the same sign. Therefore from Lemma C.4.2 and (C.5) it follows that

$$\begin{aligned}
&\mathbb{E}\left|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\right| \\
&= \mathbb{E}\left\{\left|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\right|1[0 < |G_1(H_1)| < \delta]\right\} \\
&+\mathbb{E}\left\{\left|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))\right|1[\delta < |G_1(H_1)|]\right\} \\
&\leq C\exp(-\alpha/\delta) + \frac{C\delta}{2a_n\min(\delta^3/8, 1)}
\end{aligned}$$

When $\delta = \alpha(\log a_n)^{-1} < 8$, the above bound becomes

$$\frac{C}{a_n} + \frac{4C\alpha(\log a_n)^2}{a_n}.$$

Next we consider the case of ϕ as in Corollary 3.3.1.1(iv), we have

$$\phi'(x) = \frac{2e^{-x}}{(1+e^{-x})^2} < 2e^{-x}.$$

Therefore, by Taylor's theorem,

$$|\phi(a_nA_1G_1(H_1)) - \phi(a_nA_1C_y\tilde{f}_{n,1}(H_1))| \leq a_n\phi'(\zeta)C_y\|\tilde{f}_{n,1} - G_1/C_y\|_\infty$$

for ζ is between $A_1a_nG_1(H_1)$ and $A_1a_nC_y\tilde{f}_{n,1}(H_1)$. If $G_1(H_1) > \delta$ and $\|\tilde{f}_{n,1} - G_1/C_y\|_\infty \leq \delta/2$, then $|\tilde{f}_{n,1}(H_1)| > \delta/2$, and $\tilde{f}_{n,1}(H_1))$ and $G_1(H_1)$ have the same sign. Because $\zeta/(A_1a_nC_y)$ is between $G_1(H_1)/C_y$ and $\tilde{f}_{n,1}(H_1)$,

$$|\zeta/(A_1a_nC_y)| \geq \delta/2.$$

As a result, if $G_1(H_1) > \delta$, then

$$\begin{aligned}
&|\phi(a_n A_1 G_1(H_1)) - \phi(a_n A_1 C_y \tilde{f}_{n,1}(H_1))| \\
&\leq a_n 2\exp(-a_n C_y \delta/2) C_y \|\tilde{f}_{n,1} - G_1/C_y\|_\infty \\
&\leq a_n \exp(-a_n C_y \delta/2) C_y \delta.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\Big|\phi(a_n A_1 G_1(H_1)) - \phi(a_n A_1 C_y \tilde{f}_{n,1}(H_1))\Big| \\
&\quad \leq 2C\exp(-\alpha/\delta) + a_n C_y \delta \exp(-a_n C_y \delta/2).
\end{aligned}$$

If we take $\delta = \alpha(\log a_n)^{-1}$, then the above bound can be written as

$$2Ca_n^{-1} + \alpha \frac{a_n C_y}{\log a_n} \exp\Big(-\frac{a_n C_y \alpha}{2\log a_n}\Big),$$

whose dominating term is $O(a_n^{-1})$.

□

C.2.3 Proofs of Generalization Error Results

Proof of Theorem 3.4.6. First, we introduce some new notations and terminologies. For fixed f_1, f_2, f_3, and f_4, the function $\zeta_{f_1,f_2,g_1,g_2} : \mathcal{X}_1 \times \mathcal{X}_2 \times \{\pm 1\}^2 \times \mathbb{R}^2 \mapsto \mathbb{R}$ is defined by

$$\begin{aligned}
\zeta_{f_1,f_2,g_1,g_2}(h_1,h_2,a_1,a_2,y) = \frac{y_1+y_2}{\pi_1(a_1|h_1)\pi_2(a_2|h_2)}\Big(&\psi\Big(y_1 f_1(h_1), y_2 f_2(h_2,a_2)\Big) \\
&- \psi\Big(y_1 g_1(h_1), y_2 g_2(h_2,a_2)\Big)\Big).
\end{aligned}$$

Let us denote

$$V_\psi^*(n) = \sup_{\substack{f_1 \in \mathcal{H}_{1n}, \\ f_2 \in \mathcal{H}_{2n}}} \mathbb{E}\Big[\psi\Big(Y_2 f_1(H_1), Y_3 f_2(H_2)\Big)\Big].$$

Finally, we define the following class for functions ζ_{f_1,f_2,g_1,g_2}:

$$\mathcal{G}_n = \Big\{\zeta_{f_1,f_2,g_1,g_2} : (f_1,f_2),(g_1,g_2) \in \mathcal{H}_n,\ \psi \text{ is Lipschitz with constant } L_\psi\Big\}.$$

Note that as $(\widehat{f}_{n,1}, \widehat{f}_{n,2})$ are the empirical value function maximizers, we can write

$$
\begin{aligned}
T_2 &\leq \sup_{\substack{f_1\in\mathcal{H}_{1n},\\ f_2\in\mathcal{H}_{2n}}} (P-\mathbb{P}_n)\Bigg[\frac{Y_1+Y_2}{\pi_1(A_1\mid H_1)\pi_2(A_2\mid H_2)}\Bigg(\psi\Big(Y_2 f_1(H_1), Y_3 f_2(H_2)\Big) \\
&\qquad\qquad - \psi\Big(Y_2\widehat{f}_{n,1}(H_1), Y_3\widehat{f}_{n,2}(H_2)\Big)\Bigg)\Bigg] \\
&\leq \sup_{\zeta\in\mathcal{G}_n} (P-\mathbb{P}_n)\zeta \\
&\leq \sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta|.
\end{aligned}
$$

Denote by D_n the diameter of the set $\mathcal{G}_n$:

$$
D_n = \sup_{\zeta\in\mathcal{G}_n} P\zeta^2.
$$

Talagrand's inequality [p.10, 99] implies that

$$
P\Bigg(\sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta| \geq \mathbb{E}\sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta| + D_n\sqrt{\frac{t}{n}} + \frac{t}{n}\Bigg) \leq e^{-t}. \tag{C.6}
$$

Then using the fact that ϕ, Y_1, Y_2, π_1^{-1}, and π_2^{-1} are bounded, we derive that $D_n \leq C^2$ for some $C > 0$. This yields

$$
\sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta| = \mathbb{E}\sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta| + O_p(n^{-1/2}).
$$

Due to a result by Pollard [cf. (1.2) of 100], the expected supremum of the empirical process can be bounded by

$$
\mathbb{E}\sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta| \lesssim n^{-1/2} J_{[]}(\mathcal{G}_n, L_2(P))\sqrt{\mathbb{E}Env(\mathcal{G}_n)^2} \tag{C.7}
$$

where $Env(\mathcal{F})$ denotes the envelope function for the class $\mathcal{F}$, and

$$
J_{[]}(\mathcal{F}, \|\cdot\|_{P,2}) = \int_0^1 \sqrt{1 + \log N(\epsilon\|Env(\mathcal{F})\|_{P,2}, \mathcal{F}, \|\cdot\|)}d\epsilon
$$

is the entropy integral of $\mathcal{F}$ corresponding to metric $\|\cdot\|_{P,2}$. Since ϕ is bounded, $\|Env(\mathcal{F})\|_{P,2}$ is bounded by a constant, from which, it follows that

$$
\mathbb{E}\sup_{\zeta\in\mathcal{G}_n} |(\mathbb{P}_n - P)\zeta| \lesssim n^{-1/2} J_{[]}(\mathcal{G}_n, L_2(P)).
$$

Because Y_1, Y_2 are in some compact sets, and π_1 and π_2 are known functions bounded awy from zero, standard bracketing entropy results show that

$$N(\epsilon, \mathcal{G}_n, \|\cdot\|) = N(\epsilon, \mathcal{H}_n, \|\cdot\|) + 1/\epsilon.$$

Typically $\epsilon^{-1} \lesssim N(\epsilon, \mathcal{H}_n, \|\cdot\|)$, which yields

$$J_{[]}(\mathcal{G}_n, L_2(P)) \lesssim J_{[]}(\mathcal{H}_n, L_2(P)),$$

hence the result follows. □

C.3 Proofs for Results in Section 3.4.1

C.3.1 Proof for Neural network Results, Section 3.7.1

Proof of Corollary 3.4.6.2. First of all, note that G_1 and G_2 are bounded above by $1/2$ and below by $-3/2$. Therefore, $\|G_1\|_\infty, \|G_2\|_\infty \leq 3/2$. Let us consider $C_y = 4$.

We prove the theorem showing the following steps:

A. For each $h_1^d \in \mathcal{H}_1^d$, and $h_2^d \in \mathcal{H}_2^d$, there exist $h_1 \in \mathcal{F}(L, \mathbf{p}, s, 1)$ and $h_2 \in \mathcal{F}(L, \mathbf{p}, s, 1)$ so that $|h_1 - G_1(\cdot, h_1^d)/C_y|_\infty < \epsilon$ and $|h_2 - G_2(\cdot, h_2^d)/C_y|_\infty < \epsilon$ where $\max \mathbf{p} = O(\epsilon^{-|r/\beta|_\infty})$, $L = O(\log(1/\epsilon))$, and $s = O(\log(1/\epsilon)\epsilon^{-|r/\beta|_\infty})$. Here by $|r/\beta|_\infty$, we denote $\max_{0 \leq i \leq q}(r_i/\beta_i)$.

B. There are $\tilde{f}_{n,1} \in \mathcal{F}(l, \mathbf{p}, s, 1)$ and $\tilde{f}_{n,2} \in \mathcal{F}(l, \mathbf{p}, s, 1)$ so that $|\tilde{f}_{n,1} - G_1/C_y|_\infty < \epsilon$ and $|\tilde{f}_{n,2} - G_2/C_y|_\infty < \epsilon$.

C. $\tilde{h}_{n,1} = a_n C_y \tilde{f}_{n,1} \in \mathcal{F}(l, \mathbf{p}, s, 1)$ and $\tilde{h}_{n,2} = a_n C_y \tilde{f}_{n,1} \in \mathcal{F}(l, \mathbf{p}, s, 1)$.

D.
$$|V_\psi(a_n G_1, a_n G_2) - V_\psi(\tilde{h}_{n,1}, \tilde{h}_{n,2})| \leq \frac{C\epsilon}{a_n \min(c^3, 1)}$$
for some absolute constant C.

Proof of Step A

This part follows from [89]. First, from (21) and the paragraph before (21) of [89], we derive that any $f \in \Omega(q, \mathbf{r}, \mathbf{t}, \beta, C_a)$ with the form

$$f = g_q \circ \ldots \circ g_1 \circ g_0$$

can be written as

$$f = h_q \circ \ldots \circ h_1 \circ h_0,$$

where h_{0j} takes value in $[0,1]$, $h_{0j} \in \mathcal{C}_{t_0}^{\beta}([a_0,b_0]^{t_0},1)$, $h_{ij} \in \mathcal{C}_{t_i}^{\beta_i}([0,1]^{t_i},(2C_a)^{\beta_i})$ for $i = 1,\ldots,q-1$ and $h_{qj} \in \mathcal{C}_{t_q}^{\beta_q}([0,1]^{t_q},2^{\beta_q}C_a^{\beta_q+1})$. By Lemma **??**, for any $i = 0,\ldots,q$, and $j = 1,\ldots,r_{i+1}$, for h_{ij}, we can find a network $\tilde{h}_{ij} \in \mathcal{F}(L_i, \mathbf{p}_i, s_i)$ satisfying $\|\tilde{h}_{ij} - h_{ij}\|_\infty \leq \epsilon$. Here $L_i = O(\log(1/\epsilon))$, $s_i = O(\log(1/\epsilon)\epsilon^{-r/\beta})$, and $\max p_i = O(\epsilon^{-r/\beta})$. By adjoining the parallel networks $\{\tilde{h}_{ij}\}_{1\leq j\leq r_{i+1}}$, we have a network $\tilde{h}_i \in \mathcal{F}(L_i,(r+1)\mathbf{p},(r+1)s_i)$, which satisfies

$$\||\tilde{h}_i - h_i|_\infty\|_\infty \leq \epsilon.$$

Consider the composite network

$$\tilde{h} = \tilde{h}_q \circ \ldots \circ \tilde{h}_1 \circ \tilde{h}_0.$$

Clearly,

$$\tilde{h} \in \mathcal{F}(\sum_{i=0}^{q} L_i, \mathbf{p}, \sum_{i=0}^{q}(r_i+1)s_i)$$

where $\max \mathbf{p} = \max_{0\leq i\leq q} \max p_i$, which is $O(\epsilon^{-\max_{0\leq i\leq q}(r_i/\beta_i)})$. Clearly, $L = O(q\log(1/\epsilon)) = O(\log(1/\epsilon))$ because q is finite. Also, $s = O(\log(1/\epsilon)\epsilon^{-\max_{0\leq i\leq q}(r_i/\beta_i)})$. From Lemma 3 of [89] it also follows that

$$\|\tilde{h} - f\|_\infty \leq C\epsilon$$

where the constant C depends only on q, r, β, and C_a. If we want $C\epsilon = a_n^{-1}$, then we have

$$L = O(\log a_n), s = O(a_n^{|r/\beta|_\infty}\log a_n), \text{and}\, \max p = O(a_n^{|r/\beta|_\infty}).$$

Now since $C_y > \|G_1\|_\infty, \|G_2\|_\infty + 1/2$, for small enough ϵ, if $f = G_1/C_y$ or G_2/C_y, such

an $\tilde{h}$ should take value in $[-1, 1]$, which completes the proof of this step.

Proof of Step B:

We show the proof only for $\tilde{f}_{n,2}$ because the proof for $\tilde{f}_{n,1}$ will follow similarly. We first build a network *multi* which takes h_2^c as input and approximates the vector $\{G_2(h_2^c, z)/C_y\}_{z \in \mathcal{H}_d^2}$. Since we have finitely many categorical variables with finitely many categories, using Step A we can show that this operation can be done finitely many steps. Then we show that we can obtain $G_2(h_2^c, h_2^d)$ from the output of the network multi.

Let $h_2^d = (v_1, \ldots, v_{r_{2,d}})$ be the typical categorical variable. We suppose $v_i \in D_i$ where D_i is represented by the set $\{1, \ldots, l_i\}$ for each $1 \leq i \leq r_{2,d}$. We also set $D = \prod_{1 \leq i \leq r_{2,d}} D_i$. Therefore, $\mathcal{H}_2^d = D$. Suppose $m = |D|$. Let us consider the m-valued function

$$f_1 : h_2^c \mapsto (f(h_2^c, v))_{v \in D}.$$

Note that Step A implies that for each $v \in D$, the function

$$f_v : h_2^c \mapsto (f(h_2^c, v)$$

can be approximated by a function in $\mathcal{F}(l, p, s, 1)$ where l, p, s depend only on G_1 and G_2. Here we used the fact that $\|G\|_\infty / C_y < 1$. Then using m parallel networks, we can create a network $f_2 \in \mathcal{F}(l, p_2, ms, 1)$ where $p_2 = (r_2^c, 6zmN, \ldots, 6zmN, m)$ so that

$$\| |f_1 - f_2|_\infty \|_\infty \leq \epsilon.$$

Noting the map $x \mapsto (x_+, x_-)$ is given by a single layer $(\sigma(x), \sigma(-x))$. The shift vector corresponding to this transformation has value zero and the weight matrix has $2m$ non-zero entries. Therefore, we can have a network $multi \in \mathcal{F}(l+1, (p, 2m), s+2p, 1)$ mapping $\mathbb{R}^{r_{2,c}}$ to $[-1, 1]^{2m}$ so that

$$\sup_{h_2^c} |multi(h_2^c) - (f_1(h_2^c)_+, f_1(h_2^c)_-)| < \epsilon.$$

Now we can build a parallel network of depth $l+1$ with identity weight matrices so that when

appended with multi, the resulting network $multi_1$ maps

$$multi_1 : h_2 \mapsto (multi(h_2^c), h_2^d).$$

The new network $multi_1$ is in $\mathcal{F}(l+1, (p + r_{2,d}, 2m + r_{2,d}), s + 2p + (l+1)r_{2,d}, 1)$. We write $l_{multi_1} = l + 1$, $p_{multi_1} = (p + r_{2,d}, 2m + r_{2,d})$, and $s_{multi_1} = s + 2p + (l+1)r_{2,d}$. Hence, $multi_1 \in (l_{multi_1}, p_{multi_1}, s_{multi_1}, 1)$. Let us write

$$multi(h_2^c) \equiv x = ((x_v)_+, (x_v)_-)_{v \in \mathcal{H}_2^d}$$

where for each $v \in \mathcal{H}_2^d$, $x_v \in [-1, 1]$ and $|\mathcal{H}_2^d| = |D| = m$. Lemma C.4.4 implies that there exists a network $ind \in \overline{\mathcal{F}}(l_{ind}, p_{ind}, s_{ind})$, where $l_{ind} = O\Big((\log d_{2,r})^3 \log(1/\epsilon)\Big)$, $s_{ind} = O(\log(1/\epsilon))$, and the maximal width in p_{ind} is of the order $mr_{2,d}$. which can take as input $((x_v)_{v \in \mathcal{H}_2^d}, h_2^d)$ and outputs a 2 dimensional vector (z_+, z_-), where

$$\sup_{x \in [-1,1]^m, h_2^d \in \mathcal{H}_2^d} |(z_+, z_-) - \Big((x_{h_2^d})_+, (x_{h_2^d})_-\Big)| < \epsilon.$$

Composing $multi_1$ with ind gives a network $init$ so that

$$init : h_2 \in \mathcal{H}_2 \mapsto ind \circ multi_1(h_2).$$

Note that since the output layer of init is the output layer of ind, it is of the form (z_+, z_-) for $z \in [-1, 1]$, implying

$$\sup_{h_2 \in \mathcal{H}_2} \Big| init(h_2) - \Big(G_2(h_2)_+, G_2(h_2)_-\Big)/C_y \Big| < 2\epsilon.$$

Note that $init \in \overline{\mathcal{F}}(l_{multi_1} + l_{ind}, (p_{multi_1}, p_{ind}), s_{multi_1} + s_{ind}, 1)$. Our $\tilde{f}_{n,2}$ is given by adding the elements given by the $init$ gate, i.e.

$$\tilde{f}_{n,2}(h_2) = (init(h_2))_1 + (init(h_2))_2. \tag{C.8}$$

Therefore, $\|\tilde{f}_{n,2} - G_2/C_y\|_\infty < 4\epsilon$. Similarly we can construct a $\tilde{f}_{n,1} : H_1 \mapsto \mathbb{R}$ so that

$$\|\tilde{f}_{n,1} - G_1/C_y\|_\infty < 4\epsilon \tag{C.9}$$

However, we will not carry out this operation in this step, because the next step will scale $\tilde{f}_{n,2}$ by a_n. Keeping the form (z_+, z_-) will help with this step.

Proof of step 3:

Suppose $\log_2(a_n C_y)$ is an integer. If not, then replace a_n by the sequence $2^{\log_2\lceil a_n C_y\rceil}/C_y$. Since the new sequence has the same order as a_n, all the results remain unchanged. By Lemma C.4.3, there exists a network $scale \in \mathcal{F}(\log_2(a_n C_y)-1, p, 4k(\lceil\log_2(a_n C_y)\rceil+2)$ where $p = (2,\ldots,2,1)$ which maps $(x_+, x_-) \mapsto a_n C_y x$. Letting net to be the combined network $scale \circ init$, we see that

$$net : h_2 \mapsto a_n C_y \tilde{f}_{n,2}(h_2).$$

When the dimension of the variables are $O(1)$, $net \in \mathcal{F}(l, \mathbf{p}, s, a_n C_y)$ where $l = O(\log a_n)$, $s = O(\log a_n)$, and the maximal width is $O(1)$.

Proof of step 4: It follows directly from the result of Step 2 and Lemma 3.4.5. □

Next for the generalization error, we use the following result from [89], Lemma 5.

Lemma C.3.1. *Suppose $\mathcal{H}_n = \mathcal{H}(L(n), p(n), s(n)) \equiv \mathcal{H}(L, p, s)$. Then*

$$\log \mathcal{N}(\epsilon, \mathcal{H}_n, \|\cdot\|_\infty) \lesssim (s+1)\log\big(\epsilon^{-1} C_L\big)$$

where $C_L = 2(L+1)\prod_{i=0}^{L}(p_i+1)^2$.

Note that Lemma C.3.1 does not require the functions in $\mathcal{H}_n$ to be bounded.

Lemma C.3.2. *Suppose $f_1, f_2 \in \mathcal{H}$, a class of neural networks with ReLU activation functions. Suppose the networks in $\mathcal{H}_n$ has maximum layer number L, maximal width W, and sparsity s, i.e. the number of non-zero parameters in each network is less than or equal to s. Then*

$$V_\psi^* - V(\widehat{f}_{n,1}, \widehat{f}_{n,2}) = O_p\left(\frac{\sqrt{sL\log(s \wedge (W \wedge d_2))}}{\sqrt{n}}\right)$$

where d_2 is the total number of covariates after second stage.

Proof. We will use the covering number bound in Lemma C.3.1. Since L_∞ norm is larger than

$L_2(P)$ norm, we have

$$\log \mathcal{N}(\epsilon, \mathcal{H}_n, L_2(P)) \lesssim (s+1)\log\big(\delta^{-1}C_L\big)$$

This bound on covering number yields the following entropy integral bound:

$$\begin{aligned}
J_{[]}(1, \mathcal{G}_n, L_2(P)) =& \int_0^1 \sqrt{1 + 4\log \mathcal{N}(\epsilon \|Env(\mathcal{G}_n)\|_2, \mathcal{H}_{2n}, L_2(P))}d\epsilon \\
\leq& \int_0^1 \sqrt{1 + 4\log \mathcal{N}(\epsilon L_\phi^{-1} C_\phi, \mathcal{H}_{2n}, L_2(P))}d\epsilon \\
\leq& \int_0^1 \sqrt{1 + 4(s+1)\Big(-\log\epsilon + \log\Big(L_\phi C_\phi^{-1} C_L\Big)\Big)}d\epsilon \\
=& \sqrt{\underbrace{1 + 4(s+1)\log\Big(L_\phi C_\phi^{-1} C_L\Big)}_{C_H}} \int_0^1 \sqrt{1 - t'\log\epsilon}d\epsilon,
\end{aligned}$$

where $t' = 4(s+1)/C_H$. The above integral equals $\sqrt{C}_H + \sqrt{C}_H\sqrt{t}'c$ where c is an absolute constant. Therefore, (C.7) implies

$$\mathbb{E}\sup_{h\in\mathcal{G}_n(\delta_n)} |(\mathbb{P}_n - P)h| \lesssim (\sqrt{C}_H + 2\sqrt{s+1})n^{-1/2}.$$

Note that

$$C_H \lesssim (s+1)\log C_L \leq (s+1)\Big\{ \log L + L\log(\max p + 1)\Big\} = O(sL\max p + 1)).$$

Therefore,

$$\mathbb{E}\sup_{h\in\mathcal{G}_n(\delta_n)} |(\mathbb{P}_n - P)h| = O\bigg(\frac{\sqrt{sL\log(\max p \wedge s)}}{\sqrt{n}}\bigg).$$

Hence, from (C.6) we obtain that

$$\sup_{h\in\mathcal{G}_n(\delta_n)} |(\mathbb{P}_n - P)h| = O_p\bigg(\frac{\sqrt{sL\log(s \wedge \max p)}}{\sqrt{n}}\bigg).$$

Therefore,

$$V_\psi^* - V(\widehat{f}_{n,1}, \widehat{f}_{n,2}) = O_p\bigg(\frac{\sqrt{sL\log(s \wedge (W \wedge d_2))}}{\sqrt{n}}\bigg)$$

where W is the maximal width of any intermediate layer. □

C.3.2 Proof for Wavelets Series Results, Section 3.7.2

Proof of Lemma 3.7.3. Since $f_0 \in \mathcal{C}_r^\beta(\mathcal{S}, C)$, by (3.11), there exists $M > 0$, so that f_0 belongs to the norm-ball of radius M in $B_{\infty,\infty}^\beta([0,1]^r)$. Therefore

$$\begin{aligned}
&\inf_{f\in\mathcal{H}_n} \|f - f_0\|_\infty \\
&\leq \left\| \sum_{l=0}^{b_n} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \langle f, \Psi_{lk}^i\rangle \Psi_{lk}^i - \sum_{l=0}^{\infty} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \langle f, \Psi_{lk}^i\rangle \Psi_{lk}^i \right\|_\infty \\
&= \left\| \sum_{l>b_n} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \langle f_0, \Psi_{lk}^i\rangle \Psi_{lk}^i \right\|_\infty \\
&\leq \sum_{l>b_n} \left\| \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \langle f_0, \Psi_{lk}^i\rangle \Psi_{lk}^i \right\|_\infty \\
&\lesssim \sum_{l>b_n} 2^{lr(1/2+r)} \left\| \langle f_0, \Psi_{l\cdot}^i\rangle \right\|_\infty
\end{aligned}$$

where the last step uses $\left\|\sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \langle f_0, \Psi_{lk}^i\rangle \Psi_{lk}^i\right\|_\infty \lesssim 2^{lr(1/2+r)} \left\|\langle f_0, \Psi_{l\cdot}^i\rangle\right\|_\infty$ which follows by Proposition C.3.2. Next as $\|f_0\| \in \mathcal{B}$, we have it follows that for any l:

$$2^{l(\beta+\frac{r}{2})} \|\langle f_0, \Psi_{l\cdot}^i\rangle\|_\infty \leq M,$$

therefore $\|\langle f_0, \Psi_{l\cdot}^i\rangle\|_\infty \leq M 2^{-l(\beta+\frac{r}{2})}$. Using this and the above we get

$$\sum_{l>b_n} 2^{lr(1/2+r)} \left\| \langle f_0, \Psi_{l\cdot}^i\rangle \right\|_\infty \leq M \sum_{l>b_n} 2^{-l(\beta-r^2)} = \mathcal{O}\left(a_n M 2^{-(\beta-r^2)(b_n+1)}\right),$$

which completes the proof.

□

Suppose ψ is S-regular for $S \geq 1$. Let $c \equiv \{c_{lk} : k \in \mathbb{Z}^r\} \subset \mathbb{R}$. Then for any fixed $l \in \mathbb{Z}$ and $\mathcal{K}(l) \subset \mathbb{Z}^r$, we have

$$\left\| \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} c_{lk} \Psi_{lk}^i(x) \right\|_\infty \lesssim 2^{lr(1/2+r)} \|c\|_\infty$$

Proof of Proposition C.3.2. First, since ψ is S-regular, by Definition 4.2.14 of [95], $\psi \in L^1(\mathbb{R})$ and there exists $c > 0$ so that

$$\sup_{x\in[0,1]} \sum_{k\in\mathbb{Z}} |\psi(x-k)| < c. \tag{C.10}$$

Noting we defined $\Psi^i_{lk}(x) = 2^{lr/2}\Psi^i(2^l x - k)$, we derive

$$\begin{aligned}
\left\| \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} c_{lk}\Psi^i_{lk} \right\|_\infty &= \sup_{x\in[0,1]^r} \left| \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} c_{lk} 2^{lr/2}\Psi^i(2^l x - k) \right| \\
&\leq 2^{lr/2} \sup_{x\in[0,1]^r} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} |c_{lk}| \left|\Psi^i(2^l x - k)\right| \\
&\leq 2^{lr/2} \sup_k |c_{lk}| \sup_{x\in[0,1]^r} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \left|\Psi^i(2^l x - k)\right|.
\end{aligned}$$

Next using $\Psi^i(2^l x - k) = \prod_{j=1}^r \psi^{i_j}(2^l x_j - k_j)$ we get

$$\begin{aligned}
\sup_{x\in[0,1]^r} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \left|\Psi^i(2^l x - k)\right| &= \sup_{x\in[0,1]^r} \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \prod_{j=1}^r \left|\psi^{i_j}(x_j - k_j)\right| \\
&\leq \sup_{x\in[0,1]^r} \prod_{j=1}^r \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} \left|\psi^{i_j}(x_j - k_j)\right| \\
&\leq \prod_{j=1}^r \sup_{x\in[0,1]^r} \sum_{k\in\mathcal{K}(l)} (|\psi(x_j - k_j)| + 1) \\
&\leq (c + |\mathcal{K}(l)|)^r
\end{aligned}$$

where the last step follows from (C.10). Here $|\mathcal{K}(l)|$ is the cardinality of set $\mathcal{K}(l)$ which is $\mathcal{O}\left(2^{lr}\right)$. Using the Binomial Theorem we have $(\kappa + |\mathcal{K}(l)|)^r = \mathcal{O}\left(2^{lr^2}\right)$. Therefore,

$$\left\| \sum_{k\in\mathcal{K}(l),i\in\mathcal{I}} c_{lk}\Psi^i_{lk} \right\|_\infty \lesssim 2^{lr(1/2+r)} \|c\|_\infty.$$

□

Next we show the proof for the wavelet generalization error bound.

Proof of Corollary 3.4.6.3 (b). We first note that from Theorem 4.3.36 in [95] we have

$$\log N_{[]}(\epsilon, \mathcal{H}_n, L_2(P)) \leq K \left(\frac{M}{\epsilon}\right)^{r/\beta},$$

for all $\epsilon > 0$ and for $K, M > 0$ where K depends on r and β, and M is the radius of the function space. This gives us that

$$\begin{aligned} J(\mathcal{H}_n) \equiv & \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{H}_n, L_2(P))} d\epsilon \\ \leq & \int_0^1 \sqrt{1 + K \left(\frac{M}{\epsilon}\right)^{r/\beta}} d\epsilon \\ = & \int_0^1 \epsilon^{-r/(2\beta)} \sqrt{\epsilon^{r/\beta} + K M^{r/\beta}} d\epsilon \\ \leq & \sqrt{1 + K M^{(r/\beta)}} \int_0^1 \epsilon^{-r/(2\beta)} d\epsilon. \end{aligned}$$

The above is equal to $\sqrt{1 + KM^{(r/\beta)}} 2\beta/(2\beta - r)$. Thus choosing $M = \log(n)$ we get the required result. □

C.4 Technical Lemmas

C.4.1 Relating Regret and ψ-Regret

Our next Lemma sheds some light on $\tilde{f}_1$ and $\tilde{f}_2$, which helps in calculating the regret.

Lemma C.4.1. *Suppose $\psi(x, y) = \phi(x)\phi(y)$, where ϕ satisfies the conditions in Lemma 3.3.2 and is strictly increasing. Then*

1. *$\tilde{f}_2(H_2)$ takes the value ∞ if*

$$\mathbb{E}[Y_2 + Y_3 | A_1 = 1, A_2 = 1, H_2] > \mathbb{E}[Y_2 + Y_3 | A_1 = 1, A_2 = -1, H_2]$$

and takes the value $-\infty$ if

$$\mathbb{E}[Y_2 + Y_3 | A_1 = 1, A_2 = 1, H_2] < \mathbb{E}[Y_2 + Y_3 | A_1 = 1, A_2 = -1, H_2].$$

2. $\tilde{f}_1(H_1)$ *takes the value* $\tilde{d}_1(H_1)\infty$ *where*

$$\tilde{d}_1(H_1) = \underset{a\in\{\pm 1\}\mathbb{E}[Y_2+U_3^*(a,H_1,O_2,Y_2)|A_1=a,H_1].}{argmax}$$

Since the result follows from the proof of Lemma 3.3.2, we skip the formal proof.

Proof of Lemma C.2.1. From the proof of Lemma 3.3.1, it follows that $C'_\eta(x) = (2\eta - 1)\phi'(x)$. If $2\eta > 1$, it follows that $C'_\eta(x) > 0$ since ϕ is strictly increasing. Therefore, $C_\eta(x)$ is maximized at ∞. Similarly, if $2\eta < 1$, we can show that $C_\eta(x)$ is maximized at $-\infty$. □

Lemma C.4.2. *Suppose* ϕ *is as in Corollary 3.3.1.1 (i)-(iii). Suppose* $|x| > c$ *for some* $c > 0$ *and* $\epsilon \in (0, c/2)$*. Let* a_n *be a large positive integer. Then*

$$|\phi(a_n x) - \phi(a_n(x-\epsilon))| < \frac{C\epsilon}{a_n \min(c^3, 1)}.$$

for some absolute constant $C > 0$*. For* ϕ *in Corollary 3.3.1.1(iv), we have*

$$|\phi(a_n x) - \phi(a_n(x-\epsilon))| \leq a_n e^{-ca_n/2}\epsilon.$$

Proof. First note that x and $x-\epsilon$ have the same sign. For the rest of the proof, we will assume that they are positive. When they are negative, the proof follows in an identical way. Suppose ϕ is as in Corollary 3.3.1.1(i). Then

$$\begin{aligned}
|\phi(a_n x) - \phi(a_n(x-\epsilon))| &= \left| \frac{x}{a_n^{-1} + x} - \frac{x-\epsilon}{a_n^{-1} + x - \epsilon} \right| \\
&= \frac{|-x\epsilon + \epsilon(a_n^{-1} + x)|}{(a_n^{-1} + x)(a_n^{-1} + x - \epsilon)} \\
&= \frac{a_n^{-1}}{(a_n^{-1} + x)(a_n^{-1} + x - \epsilon)}
\end{aligned}$$

which is not larger than $4a_n^{-1}\epsilon/c^2$.

Now suppose ϕ is as in Corollary 3.3.1.1(ii). Then

$$\begin{aligned}
|\phi(a_n x) - \phi(a_n(x-\epsilon))| &= \left| \frac{x}{\sqrt{a_n^{-2} + x^2}} - \frac{x-\epsilon}{\sqrt{a_n^{-2} + (x-\epsilon)^2}} \right| \\
&= \frac{\left| x\sqrt{a_n^{-2} + (x-\epsilon)^2} - (x-\epsilon)\sqrt{a_n^{-2} + x^2} \right|}{\sqrt{a_n^{-2} + x^2}\sqrt{a_n^{-2} + (x-\epsilon)^2}} \\
&= \frac{|x^2 a_n^{-2} + x^2(x-\epsilon)^2 - a_n^{-2}(x-\epsilon)^2 - x^2(x-\epsilon)^2|}{\sqrt{a_n^{-2} + x^2}\sqrt{a_n^{-2} + (x-\epsilon)^2}\left(x\sqrt{a_n^{-2} + (x-\epsilon)^2} + (x-\epsilon)\sqrt{a_n^{-2} + x^2} \right)} \\
&\leq a_n^{-2} \frac{x\epsilon + \epsilon^2}{\sqrt{a_n^{-2} + x^2}(c/2)c^2} \\
&\leq 2a_n^{-2} \frac{2x\epsilon}{\sqrt{a_n^{-2} + x^2}c^3},
\end{aligned}$$

which is bounded by $4a_n^{-2}\epsilon c^{-3}$.

Suppose ϕ is as in Corollary 3.3.1.1(iii). Then using the fact that $\arctan(x) = \int_{-\infty}^{x} \frac{dy}{1+y^2}$, we obtain that

$$\begin{aligned}
&|\phi(a_n x) - \phi(a_n(x-\epsilon))| \\
&= \frac{2}{\pi} \left| \int_{a_n\pi(x-\epsilon)/2}^{a_n\pi x/2} \frac{dy}{1+y^2} \right| \\
&\leq \frac{2}{\pi} \int_{a_n\pi(x-\epsilon)/2}^{a_n\pi x/2} \frac{dy}{y^2} \\
&= \frac{4}{\pi a_n} \left(\frac{1}{x-\epsilon} - \frac{1}{x} \right) \\
&\leq \frac{8\pi\epsilon}{c^2 a_n}.
\end{aligned}$$

Finally, we have for the ϕ in Corollary 3.3.1.1(iv),

$$\phi'(x) = \frac{\partial(1+e^{-x})^{-1}}{\partial x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{2+e^x+e^{-x}} \leq e^{-x}$$

By first order Taylor series expansion,

$$|\phi(a_n x) - \phi(a_n(x-\epsilon))| = a_n \phi'(a_n \xi)\epsilon$$

where $\xi \in (x-\epsilon, x)$. Noting $\phi'(a_n\xi) \leq e^{-ca_n/2}$, we have

$$|\phi(a_n x) - \phi(a_n(x-\epsilon))| \leq a_n e^{-ca_n/2}\epsilon.$$

□

Lemma C.4.3. *Suppose* $\log_k a_n$ *is an integer for some* $k \geq 2$. *Then there exists a network* $f \in \mathcal{F}(\log a_n/\log k - 1, p, s, a_n x)$ *so that* $f : (x_+, x_-) \mapsto \mathbb{R}$ *outputs* $a_n x$ *for any* $x \in \mathbb{R}$,

$$s \leq 4k(\log a_n/\log k + 2),$$

and

$$p = (2, k, \ldots, k, 2, 1).$$

Proof. We consider a network with the first layer outputs $(x_+, \ldots, x_+, x_-, \ldots, x_-)$, which is a vector of length $2k$. Therefore the associated weight matrix

$$W_0 = \begin{bmatrix} 1_{k\times 1} & 0_{k\times 1} \\ 0_{k\times 1} & 1_{k\times 1} \end{bmatrix}$$

has $2k$ parameters, where $1_{k\times 1}$ and $0_{k\times 1}$ are k-length vectors of ones and zeros, respectively. The associated shift vector is zero in this layer, and all consecutive layers. The second layer outputs k times the first layer using a weight matrix W_1

$$W_1 = \begin{bmatrix} J_{k\times k} & O_{k\times k} \\ O_{k\times k} & J_{k\times k} \end{bmatrix}$$

and this process is repeated for r layer. The r-th hidden layer outputs $k^r(x_+, \ldots, x_+, x_-, \ldots, x_-)$. These values are consolidated in the $r+1$-th layer, which outputs $k^{r+1}x_+$ and $k^{r+1}x_-$ as a result of using $W_r = W_1^T$ as the weight matrix. The final output is $k^{r+1}x_+ - k^{r+1}x_- = k^{r+1}x$ which needs only 4 network parameters. Note also that the total number of network parameters are $4k(r+3)$. Since we want $k^{r+1} = a_n$, $r = \log a_n/\log k - 1$. Therefore, the result follows. □

Lemma C.4.4. *Suppose we have* k *categories* $D_1, \ldots, D_k$. *Each* D_i *is presented by the set* $\{1, \ldots, l_i\}$ *where* l_i *is a positive integer. Suppose* v_i *is any element from* D_i *and* $v =$

$(v_1, \ldots, v_k) \in D = \prod_{i=1}^{k} D_i$. *$v$ is the categorical input vector to the network. Suppose we have* $x = (x_v)_{v \in D}$ *is a vector in* $[-1,1]^m$ *where* $m = |D|$. *Let us denote* $m' = \sum_{i=1}^{k} l_i$. *If* $\sup\{l_i : 1 \leq i \leq k\} \leq c$ *for a constant* $c > 0$, *then given any* $\epsilon > 0$, *there exists a network* $ind \in \overline{\mathcal{F}}(l, p, s)$ *so that*

$$\sup_{x \in [-1,1]^m, v \in D} |f(x_+, x_-, v) - ((x_v)_+, (x_v)_-)| < \epsilon$$

where

$$l \leq 2\lceil \log(k+1) \rceil^3 \log \epsilon^{-1},$$

$p = (2m + k, 6m' + 2m, 6m' + 2m, 2m' + 2m, 12m(k+1), 12m(k+1), \ldots 12m(k+1), 2m, 2)$,

and $s \leq c^{Ck} \log(\epsilon)^{-1}$ *where* $C > 0$ *is an absolute constant.*