Semi-Supervised Learning for Doubly Robust Offline Policy Evaluation

Aaron Sonabend-W
asonabend@g.harvard.edu
Department of Biostatistics
Harvard University

Nilanjana Laha
nilanjanaaa.laha@gmail.com
Department of Biostatistics
Harvard University

Rajarshi Mukherjee
ram521@mail.harvard.edu
Department of Biostatistics
Harvard University

Tianxi Cai
tcai@hsph.harvard.edu
Department of Biostatistics
Harvard University

1 Introduction

Finding optimal treatment strategies that can incorporate patient heterogeneity is a cornerstone of personalized medicine. When treatment options can 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 [1, 2, 3, 4, 5] and directly maximizing the value function [6]. Although G-estimation and A-learning models can be more efficient and robust to miss-specification, Q-learning is widely adopted due to its ease of implement, flexible and interpretable [3, 7, 8].

Learning STR with EHR data, however, often faces an additional challenge that 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 code 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 replace of the outcome [9, 10, 11]. On the other hand, extracting precise outcome information often requires manual chart review, which is resource intensive particularly when aiming to learn STR since the outcome needs to be annotated over time. This signifies the need for semi-supervised learning (SSL) 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 [12].

While SSL methods have been well developed for prediction, classification and regression tasks [13, 14, 15, 16, 17, 12, e.g.]. Recently, [11, 18] proposed SSL methods for estimating an average causal treatment effect. [19] proposed a semi-supervised RL method which achieves impressive empirical results and outperforms simple approaches such as direct imputation of the reward. However, to the best of our knowledge there is no previous work on SSL methods for off-policy policy evaluation (OPE). OPE differs from estimating the average treatment effect as the estimand depends on a dynamic conditional treatment rule (policy function) and not a pre-specified treatment. In this paper, we fill this gap by proposing a theoretically justified SSL approach to OPE 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 reward (or health outcome) R as well as a small set of labeled data $\mathcal{L}$ which contains true outcome R at multiple stages along with $\mathbf{O}$, A and $\mathbf{W}$. The policy evaluation

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2020.

we use is defined as the expected counterfactual outcome under the optimal STR. We further show that our approach is robust to miss-specification of the imputation models. Additionally, our OPE is doubly robust, meaning if either the regression models for the Q functions or the propensity score functions are correctly specified, our estimator is consistent for the true value function. Finally, our value function estimator is flexible enough to allow for standard off-the-shelf machine learning tools and is shown to perform well in finite-sample numerical examples. An extended version with additional theoretical and empirical results are shown in [20]. In Section 2 we go over the problem set-up and notation, in Section 3 we give a brief overview of Q-learning used to derive optimal STR, in Section 4 we derive our proposed Semi Supervised Off-Policy Evaluation of the Policy. Finally in Section 5 we show empirical results and conclude in Section 6.

2 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 the ease of presentation we'll use two time points of (binary) treatment allocation as follows. For time point $t \in \{1, 2\}$, let $\mathbf{O}_t \in \mathbb{R}^{d_t^S}$ denote the vector of covariates measured prior at stage t of dimension d_t^S; $A_t \in \{0, 1\}$ a treatment indicator variable; and $R_{t+1} \in \mathbb{R}$ the outcome observed at stage $t + 1$, for which higher values of R_{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 R_{t+1}. In the labeled data where $\mathbf{R} = (R_2, R_3)^\intercal$ is annotated, we observe a random sample of n independent and identically distributed (iid) random vectors, denoted by

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

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 estimating the optimal STR *value function*, defined as expected counterfactual outcomes under the optimal regime. To this end, let $R_t^{(a)}$ be the potential outcome for a patient at time t had the patient been treated at time $t - 1$ with 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 $\mathbf{S}_1 \equiv [\mathbf{S}_{10}^\intercal, \mathbf{S}_{11}^\intercal]^\intercal$ with $\mathbf{S}_{1k} = \boldsymbol{\phi}_{1k}(\mathbf{O}_1)$, $\mathbf{S}_2 = [\mathbf{S}_{20}^\intercal, \mathbf{S}_{21}^\intercal]^\intercal$ with $\mathbf{S}_{2k} = \boldsymbol{\phi}_{2k}(\mathbf{O}_1, A_1, \mathbf{O}_2)$, where $\{\boldsymbol{\phi}_{tk}(\cdot), t = 1, 2, k = 0, 1\}$ are pre-specified basis functions, for example wavelets basis, natural cubic splines, etc. We use $\mathbf{S}_{t0}$, $\mathbf{S}_{t1}$ $t = 1, 2$ to denote baseline features, and treatment effect features respectively for the Q functions, more detail will follow in Section 3. For ease of presentation, we also let $\check{\mathbf{S}}_1 = \mathbf{S}_1^\intercal$, $\check{\mathbf{S}}_2 = (R_2, \mathbf{S}_2^\intercal)^\intercal$.

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 $V(\mathcal{D}) = \mathbb{E}_\mathcal{D}\left[R_2^{(d_1)} + R_3^{(d_2)}\right]$. Thus, an optimal dynamic treatment regime is a rule $\bar{\mathcal{D}} = (\bar{d}_1, \bar{d}_2)$ such that $\bar{V} = V\left(\bar{\mathcal{D}}\right) \geq V\left(\mathcal{D}\right) \forall \mathcal{D}$, where $\mathcal{D}$ belongs to a suitable class of admissible decisions [7]. To identify $\bar{\mathcal{D}}$ and $\bar{V}$ from the observed data we will require the following sets of standard assumptions [21, 8]: (i) consistency – $R_{t+1} = R_{t+1}^{(0)} I_{\{A_t=0\}} + R_{t+1}^{(1)} I_{\{A_t=1\}}$ for $t = 1, 2$, (ii) no unmeasured confounding – $R_{t+1}^{(0)}, R_{t+1}^{(1)} \perp\!\!\!\perp A_t | \mathbf{S}_t$ for $t = 1, 2$ and (iii) positivity – $\mathbb{P}(A_t|\mathbf{S}_t) > \nu,$ for $t = 1, 2,$ $A_t \in \{0, 1\}$, for some fixed $\nu > 0$.

3 Q-learning for STR

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

$$Q_2(\check{\mathbf{S}}_2, A_2) \equiv \mathbb{E}[R_3|\check{\mathbf{S}}_2, A_2], \quad \text{and} \quad Q_1(\check{\mathbf{S}}_1, A_1) \equiv \mathbb{E}[R_2 + \max_{a_2} Q_2(\check{\mathbf{S}}_2, a_2)|\mathbf{S}_1, A_1]$$

[4, 22]. See [20] for discussion of a semi-supervised Q learning approach. 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:

$$\begin{aligned} Q_1(\check{\mathbf{S}}_1, A_1; \boldsymbol{\theta}_1^0) =& \mathbf{S}_{10}^\intercal \boldsymbol{\beta}_1^0 + A_1(\mathbf{S}_{11}^\intercal \boldsymbol{\gamma}_1^0), \\ Q_2(\check{\mathbf{S}}_2, A_2; \boldsymbol{\theta}_2^0) =& R_2 \beta_{21}^0 + \mathbf{S}_{20}^\intercal \boldsymbol{\beta}_{22}^0 + A_2(\mathbf{S}_{21}^\intercal \boldsymbol{\gamma}_2^0). \end{aligned} \tag{1}$$

Note that in model (1), parameters $\boldsymbol{\gamma}_1^0, \boldsymbol{\gamma}_2^0$ correspond to the treatment effect as $A_t \in \{0, 1\}$, $t = 1, 2$. 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 as $\widehat{Y}_{2i}^* = R_{2i} + \max_{a_2} Q_2(\check{\mathbf{S}}_{2i}, a_2; \widehat{\boldsymbol{\theta}}_2)$, for $i = 1, ..., n$. One then proceeds to estimate $\widehat{\boldsymbol{\theta}}_1$ using least squares again, using $\widehat{Y}_2^*$ 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 the value function using a doubly robust type estimator. Based on the Q-learning models, we can then obtain an estimate for the optimal treatment protocol as:

$$\widehat{d}_t \equiv d_t(\mathbf{S}_t; \widehat{\boldsymbol{\theta}}_t), \qquad \text{where} \quad d_t(\mathbf{S}_t; \boldsymbol{\theta}_t) = \underset{a \in \{0,1\}}{\operatorname{argmax}} Q_t(\check{\mathbf{S}}_t, a; \boldsymbol{\theta}_t), \quad t = 1, 2.$$

As we explain next, this in turn yields desirable statistical results for evaluating the resulting policy $\bar{d}_t \equiv d_t(\mathbf{S}_t; \bar{\boldsymbol{\theta}}_t) = \operatorname{argmax}_{a \in \{0,1\}} Q_t(\check{\mathbf{S}}_t, a; \bar{\boldsymbol{\theta}}_t)$, for $t = 1, 2$.

4 Semi Supervised Off-Policy Evaluation of the Policy

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

$$\bar{V} \equiv \bar{V}(\bar{\boldsymbol{\theta}}) = \mathbb{E}\left(\mathbb{E}[R_2 + \mathbb{E}\{R_3 | \check{\mathbf{S}}_2, A_2 = \bar{d}_2(\mathbf{S}_2; \bar{\boldsymbol{\theta}}_2)\} | \mathbf{S}_1, A_1 = \bar{d}_1(\mathbf{S}_1; \bar{\boldsymbol{\theta}}_1)]\right),$$

where $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\theta}}_1^\intercal, \bar{\boldsymbol{\theta}}_2^\intercal)^\intercal$. Recall that both $\mathcal{U}$ and $\mathcal{L}$ have distribution $\mathbb{P}$, thus we can equivalently define $\bar{V} = \mathbb{E}_{\bar{\mathcal{D}}}\left[R_2^{(\bar{d}_1)} + R_3^{(\bar{d}_2)}\right]$, with expectation over $\mathbb{P}$ holding $\bar{\mathcal{D}}$ fixed. If models in (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 $\widehat{V}_Q = \mathbb{P}_n\left[Q_1^o(\check{\mathbf{S}}_1; \widehat{\boldsymbol{\theta}}_1)\right]$, where $Q_t^o(\check{\mathbf{S}}_t; \boldsymbol{\theta}_t) \equiv Q_t\{\check{\mathbf{S}}_t, d_t(\mathbf{S}_t; \boldsymbol{\theta}_t); \boldsymbol{\theta}_t\}$. However, $\widehat{V}_Q$ is likely to be biased when the outcome models in (1) are miss-specified which arise frequently in practice since $Q_1(\check{\mathbf{S}}_1, A_1)$ is especially difficult to specify.

To improve the robustness to model miss-specification, we propose an SSL doubly robust (SSL_{DR}) estimator for $\bar{V}$ by augmenting $\widehat{V}_Q$ via propensity score weighting. To this end, we define propensity scores $\pi_t(\check{\mathbf{S}}_t) = \mathbb{P}\{A_t = \bar{d}_t | \check{\mathbf{S}}_t\}$, $t = 1, 2$. To estimate $\{\pi_t(\cdot), t = 1, 2\}$, we impose the following generalized linear models (GLM):

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

We use the logistic model with potentially non-linear basis functions $\check{\mathbf{S}}$ 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 $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\intercal, \boldsymbol{\xi}_2^\intercal)^\intercal$ based on the standard maximum likelihood estimators using labeled data, denoted by $\widehat{\boldsymbol{\xi}} = (\widehat{\boldsymbol{\xi}}_1^\intercal, \widehat{\boldsymbol{\xi}}_2^\intercal)^\intercal$. Denote the limit of $\widehat{\boldsymbol{\xi}}$ as $\bar{\boldsymbol{\xi}} = (\bar{\boldsymbol{\xi}}_1^\intercal, \bar{\boldsymbol{\xi}}_2^\intercal)^\intercal$, which is equal to the true model parameter under correct specification of (2).

4.1 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

$$\omega_1(\check{\mathbf{S}}_1, A_1, \boldsymbol{\Theta}) \equiv \frac{I\{A_1 = d_1(\check{\mathbf{S}}_1, \boldsymbol{\theta}_1)\}}{\pi_1(\check{\mathbf{S}}_1; \boldsymbol{\xi}_1)}, \text{ and } \omega_2(\check{\mathbf{S}}_2, A_2, \boldsymbol{\Theta}) \equiv \omega_1(\check{\mathbf{S}}_1, A_1, \boldsymbol{\Theta}) \frac{I\{A_2 = d_2(\check{\mathbf{S}}_2, \boldsymbol{\theta}_2)\}}{\pi_2(\check{\mathbf{S}}_2; \boldsymbol{\xi}_2)}.$$

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

$$\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}}) = Q_1^o(\mathbf{S}_1;\widehat{\boldsymbol{\theta}}_1)+\omega_1(\check{\mathbf{S}}_1,A_1,\widehat{\boldsymbol{\Theta}})\left[R_2-\left\{Q_1^o(\mathbf{S}_1,\widehat{\boldsymbol{\theta}}_1)-Q_2^o(\check{\mathbf{S}}_2;\widehat{\boldsymbol{\theta}}_2)\right\}\right] \\ +\omega_2(\check{\mathbf{S}}_2,A_2,\widehat{\boldsymbol{\Theta}})\left\{R_3-Q_2^o(\check{\mathbf{S}}_2;\widehat{\boldsymbol{\theta}}_2)\right\}$$

and estimate $\bar{V}$ as

$$\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}} = \mathbb{P}_n\left\{\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}})\right\}. \tag{3}$$

The augmented importance sampling estimator in (3) is similar to those previously proposed in [23] and [24]. The construction of augmentation in $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$ also differs from the usual augmented IPW estimators [7]. As we are interested in the value had the population been treated with $\bar{\mathcal{D}}$ and not a fixed sequence (A_1,A_2), we replace the weights for a fixed treatment (i.e. $A_t=1$) with the propensity score weights for the optimal regime $I(A_t=\bar{d}_t)$. Finally, we note that this estimator can easily be extended to incorporate non-binary treatments.

4.2 $\mathsf{SSL}_{\mathsf{DR}}$ Value Function Estimation

We next detail our robust imputation based semi-supervised procedure that leverages the unlabeled data $\mathcal{U}$ to replace the unobserved R_t in (3) 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 outcome mean functions. The refitting in step (ii) will ensure the validity of the SSL estimators under potential mis-specifications of the imputation models. Since $\check{\mathbf{S}}_2$ involves R_2, both $\omega_2(\check{\mathbf{S}}_2,A_2;\boldsymbol{\Theta})$ and $Q_2^o(\check{\mathbf{S}}_2;\boldsymbol{\theta}_2)=R_2\beta_{21}+Q_{2-}^o(\mathbf{S}_2;\boldsymbol{\theta}_2)$ are not available in the unlabeled set, where $Q_{2-}^o(\mathbf{S}_2;\boldsymbol{\theta}_2)=\mathbf{S}_{20}^{\mathsf{T}}\boldsymbol{\beta}_{22}+[\mathbf{S}_{21}^{\mathsf{T}}\boldsymbol{\gamma}_2]_+$. Writing

$$\mathcal{V}_{\mathsf{SUP}_{\mathsf{DR}}}(\mathbf{L};\widehat{\boldsymbol{\Theta}}) = Q_1^o(\mathbf{S}_1;\widehat{\boldsymbol{\theta}}_1)+\omega_1(\check{\mathbf{S}}_1,A_1,\widehat{\boldsymbol{\Theta}})\left\{(1+\widehat{\beta}_{21})R_2-Q_1^o(\mathbf{S}_1,\widehat{\boldsymbol{\theta}}_1)+Q_{2-}^o(\mathbf{S}_2;\widehat{\boldsymbol{\theta}}_2)\right\} \\ +\omega_2(\check{\mathbf{S}}_2,A_2,\widehat{\boldsymbol{\Theta}})\left\{R_3-\widehat{\beta}_{21}R_2-Q_{2-}^o(\mathbf{S}_2;\widehat{\boldsymbol{\theta}}_2)\right\},$$

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 R_2, $\omega_2(\check{\mathbf{S}}_2,A_2;\widehat{\boldsymbol{\Theta}})$, and $R_t\omega_2(\check{\mathbf{S}}_2,A_2;\widehat{\boldsymbol{\Theta}})$ for $t=2,3$. Define conditional mean functions

$$\mu_2(\vec{\mathbf{U}})\equiv\mathbb{E}[R_2|\vec{\mathbf{U}}],\quad \mu_{\omega_2}(\vec{\mathbf{U}})\equiv\mathbb{E}[\omega_2(\check{\mathbf{S}}_2,A_2;\bar{\boldsymbol{\Theta}})|\vec{\mathbf{U}}],\quad \mu_{t\omega_2}(\vec{\mathbf{U}})\equiv\mathbb{E}[R_t\omega_2(\check{\mathbf{S}}_2,A_2;\bar{\boldsymbol{\Theta}})|\vec{\mathbf{U}}],$$

for $t=2,3$, where $\bar{\boldsymbol{\Theta}}=(\bar{\boldsymbol{\theta}}^{\mathsf{T}},\bar{\boldsymbol{\xi}}^{\mathsf{T}})^{\mathsf{T}}$ We approximate these expectations by a flexible imputation model followed by a refitting step for bias correction under possible miss-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 $\{\mu_2(\vec{\mathbf{U}}),\mu_{\omega_2}(\vec{\mathbf{U}}),\mu_{t\omega_2}(\vec{\mathbf{U}}),\ t=2,3\}$ with unknown estimated parameter $\boldsymbol{\theta}$ 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 $\{R_2,\omega_2(\check{\mathbf{S}}_2,A_2;\bar{\boldsymbol{\Theta}}),R_t\omega_2(\check{\mathbf{S}}_2,A_2;\bar{\boldsymbol{\Theta}}),t=2,3\}$ as $\{\bar{\mu}_2^v(\vec{\mathbf{U}})=m_2(\vec{\mathbf{U}})+\eta_2^v,\bar{\mu}_{\omega_2}^v(\vec{\mathbf{U}})=m_{\omega_2}(\vec{\mathbf{U}})+\eta_{\omega_2}^v,\bar{\mu}_{t\omega_2}^v(\vec{\mathbf{U}})=m_{t\omega_2}(\vec{\mathbf{U}})+\eta_{t\omega_2}^v,\ t=2,3\}$. These refitted models are not required to be correctly specified but need to satisfy the following constraints are satisfied:

$$\begin{aligned}\mathbb{E}\left[\omega_1(\check{\mathbf{S}}_1,A_1;\bar{\boldsymbol{\Theta}})\left\{R_2-\bar{\mu}_2^v(\vec{\mathbf{U}})\right\}\right]&=0,\\ \mathbb{E}\left[Q_{2-}^o(\vec{\mathbf{U}};\boldsymbol{\theta}_2)\left\{\omega_2(\check{\mathbf{S}}_2,A_2;\bar{\boldsymbol{\Theta}})-\bar{\mu}_{\omega_2}(\vec{\mathbf{U}})\right\}\right]&=0,\\ \mathbb{E}\left[\omega_2(\check{\mathbf{S}}_2,A_2;\bar{\boldsymbol{\Theta}})R_t-\bar{\mu}_{t\omega_2}(\vec{\mathbf{U}})\right]&=0,\ t=2,3.\end{aligned} \tag{4}$$

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 $\{\widehat{m}_t^{(-k)}(\vec{\mathbf{U}}), \widehat{m}_{\omega_2}^{(-k)}(\vec{\mathbf{U}}), \widehat{m}_{t\omega_2}^{(-k)}(\vec{\mathbf{U}}), t = 2, 3\}$ be the counterpart of $\{\widehat{m}_t(\vec{\mathbf{U}}), \widehat{m}_{\omega_2}(\vec{\mathbf{U}}), \widehat{m}_{2t}(\vec{\mathbf{U}}), t = 2, 3\}$ with labeled observations in $\{1, .., n\} \setminus \mathcal{I}_k$. 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

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \omega_1(\check{\mathbf{S}}_{1i}, A_{1i}; \widehat{\mathbf{\Theta}}) \left\{ R_2 - \widehat{m}_2^{(-k)}(\vec{\mathbf{U}}_i) - \widehat{\eta}_2^v \right\} = 0, \\
\sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} Q_{2-}^o(\vec{\mathbf{U}}_i; \widehat{\boldsymbol{\theta}}_2) \left\{ \omega_2(\check{\mathbf{S}}_{2i}, A_{2i}; \widehat{\mathbf{\Theta}}) - \widehat{m}_{\omega_2}^{(-k)}(\vec{\mathbf{U}}_i) - \widehat{\eta}_{\pi_2}^v \right\} = 0, \\
\sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \left\{ \omega_2(\check{\mathbf{S}}_{2i}, A_{2i}; \widehat{\mathbf{\Theta}}) R_{ti} - \widehat{m}_{t\omega_2}^{(-k)}(\vec{\mathbf{U}}_i) - \widehat{\eta}_{t\omega_2}^v \right\} = 0, \ t = 2, 3.
\end{aligned} \tag{5}$$

The resulting imputation functions for $R_2, \omega_2(\check{\mathbf{S}}_2, A_2; \bar{\mathbf{\Theta}})$ and $R_t\omega_2(\check{\mathbf{S}}_2, A_2; \bar{\mathbf{\Theta}})$ 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 $\widehat{\mu}_{t\omega_2}(\vec{\mathbf{U}}) = K^{-1} \sum_{k=1}^{K} \widehat{m}_{t\omega_2}^{(-k)}(\vec{\mathbf{U}}) + \widehat{\eta}_{t\omega_2}^v$, 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:

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} = \mathbb{P}_N \left\{ \widehat{\mathcal{V}}_{\mathsf{SSL}_{\mathsf{DR}}}(\vec{\mathbf{U}}) \right\}, \tag{6}$$

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

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

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. $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ 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, we can use the same data to estimate both the optimal STR and it's value function. We next give a proposition which states our semi-supervised estimator is doubly robust, for further theoretical results see [20].

Let $g(\cdot)$ be a real valued function, we define the norm $\|g(x)\|_{L_2(\mathbb{P})} \equiv \sqrt{\int g(x)^2 d\mathbb{P}(x)}$. Additionally, let $\{U_n\}$, and $\{V_n\}$ be two sequences of random variables, we use $U_n = o_{\mathbb{P}}(V_n)$ to denote that $U_n / V_n \xrightarrow{\mathbb{P}} 0$. Assumptions for the following Proposition which include class functions for the Q functions and propensity scores are listed in the Appendix B.

Proposition 4.1 *If either* $\|\hat{\pi}_t - \pi_t\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$ *or* $\left\| \hat{Q}_t - Q_t \right\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$, $t = 1, 2$ *then under Assumptions B.1-B.4*

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} \xrightarrow{p} \bar{V},$$

where $\bar{V} = \mathbb{E}[\mathbb{E}[R_2 + \mathbb{E}[R_3 | \mathbf{S}_2, A_2 = \bar{d}_2, R_2] | \mathbf{S}_1, A_1 = \bar{d}_1]]$.

5 Simulation results

We performed extensive simulations to evaluate the finite sample performance of our method. We compare our semi-supervised estimator to a fully supervised approach using labeled data sets of size

n. In particular, we evaluate the value function comparing supervised vs. semi-supervised estimators while varying degree of miss-specification for the Q models in (1) and the propensity score functions w in (2). We detail the simulation settings in Appendix A. In essence, to evaluate robustness to misspecification in the Q function, we generate continuous R_3 as

$$R_3|\check{\mathbf{S}}_2 \sim \mathcal{N}\left(\check{\mathbf{S}}_{20}^{\intercal}\boldsymbol{\beta}_2^0 + \beta_{27}^0 O_2^2 R_2 \sin\left(\frac{1}{O_2^2 R_2}\right) + A_2(\check{\mathbf{S}}_{21}^{\intercal}\boldsymbol{\gamma}_2^0), 2\right)$$

and use the following model for the second stage Q function: $Q_2(\check{\mathbf{S}}_2, A_2; \boldsymbol{\theta}_2) = \check{\mathbf{S}}_{20}^{\intercal}\boldsymbol{\beta}_2^0 + A_2(\mathbf{S}_{21}^{\intercal}\boldsymbol{\gamma}_2^0)$. The binary reward version uses a sigmoid transformation of the above mean function for R_3. Similarly, to evaluate robustness to miss-specification of the propensity score, we use

$$A_2|O_1, O_2, A_1, R_2 \sim \text{Bern}\left(\sigma\left(\mathbf{S}_2^{\intercal}\boldsymbol{\xi}_2^0 + \xi_{26}^0 O_2^2\right)\right)$$

and fit model $\pi_2(\check{\mathbf{S}}_2; \boldsymbol{\xi}_2) = \sigma\left(\mathbf{S}_2^{\intercal}\boldsymbol{\xi}_2^0\right)$. Both base scenarios use $\beta_{27}^0 = \xi_{26}^0 = 0$, for correct model specification. For imputation we use random forests with 500 trees (RF). All results are based on 500 replications.

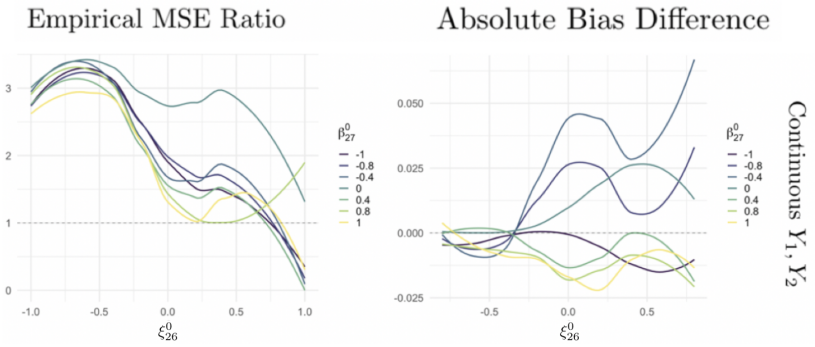

Figure 1: Monte Carlo estimates for 500 datasets for simulations of continuous rewards R_2, R_3. We show empirical MSE ratio and difference of the absolute value of bias for value function estimation. Comparison is done across degree of miss-specificaction of the Q function by ranging β_{27} in different colored lines and on the propensity score for A_2 by varying ξ_{26} in the x-axis. MSE ratios > 1 & absolute bias difference > 0 favor semi supervised value function estimation.

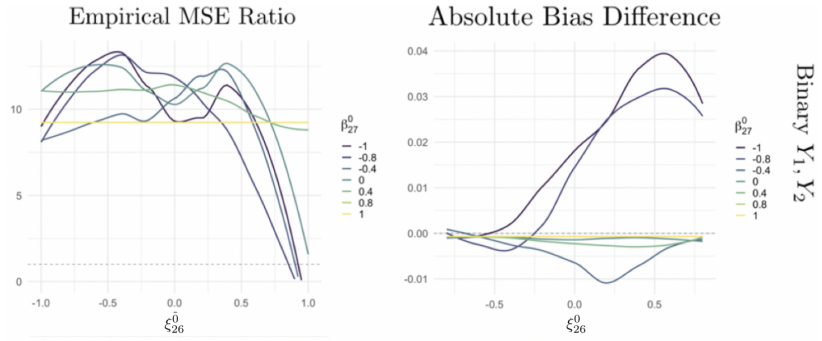

Figure 2: Monte Carlo estimates for 500 datasets for simulations of binary rewards R_2, R_3. We show empirical MSE ratio and difference of the absolute value of bias for value function estimation. Comparison is done across degree of miss-specificaction of the Q function by ranging β_{27} in different colored lines and on the propensity score for A_2 by varying ξ_{26} in the x-axis. MSE ratios > 1 & absolute bias difference > 0 favor semi supervised value function estimation.

Through simulations, we analyze performance of the doubly robust value function estimators for both binary and continuous outcomes. We vary degree of miss-specification of the Q functions and propensity scores through β_{27} and ξ_{26} respectively. Figure 1 & 2 show empirical MSE ratio and bias difference across varying levels of miss-specification of the Q functions as different lines, and

propensity scores across the x-axis. Bias across simulation settings are relatively similar between $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ and $\widehat{V}_{\mathsf{SUP}_{\mathsf{DR}}}$. The low magnitude of bias suggests both estimators are robust to model misspecification. Negative levels of β_{27}^0 favors $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$ for both binary and continuous outcomes. The semi-supervised value function estimator, however, is generally more efficient across levels of β_{27}^0 and ξ_{26}^0. Specially for negative ξ_{26}^0 values. On the other hand, large ξ_{26}^0 values result in efficiency loss for $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}$: there is a price to pay due to the over-fitting bias in the refitting step. For binary rewards, there is a better efficiency gain as seen in the empirical MSE ratio of Figure 2, however the continuous setting shows a similar trend in terms of efficiency gain, but a more modest one.

6 Discussion

We have proposed an efficient strategy for the semi-supervised setting for estimation of dynamic treatment rules and their value function. In particular we develop a two step estimation procedure which is amenable to non-parametric imputation of the missing outcomes. This helped us leverage the unlabeled data $\mathcal{U}$ to gain efficiency of our value function estimation as shown empirically through simulations. We additionally show our value function estimator is doubly robust.

There are several ways to extend this work. The most natural one is to extend this framework to more than two treatment time points. We focus on the 2-time point setting for simplicity but there is nothing in our theoretical result which cannot be extended to a higher time horizon. In implementation one would need to be careful with the variability of the IPW-value function which increases substantially with time. Additionally, this strategy can be used to estimate DTRs using A-learning which can yield efficiency gain under certain scenarios [8].

References

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

[2] James M. Robins. *Optimal Structural Nested Models for Optimal Sequential Decisions*, pages 189–326. Springer New York, New York, NY, 2004.

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

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

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

[6] 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(510):583–598, 2015.

[7] 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.

[8] 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.

[9] Chuan Hong, Katherine P Liao, and Tianxi Cai. Semi-supervised validation of multiple surrogate outcomes with application to electronic medical records phenotyping. *Biometrics*, 75(1):78–89, 2019.

[10] 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.

[11] 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.

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

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

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

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

[16] 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.

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

[18] 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.

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

[20] Aaron Sonabend-W, Nilanjana Laha, Rajarshi Mukherjee, and Tianxi Cai. Semi-supervised off policy reinforcement learning, 2020.

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

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

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

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

[25] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer series in statistics. Springer, New York, 2008.