Gradient Regularized V-Learning for Dynamic Treatment Regimes

Yao Zhang
University of Cambridge
yz555@cam.ac.uk

Mihaela van der Schaar
University of Cambridge
University of California, Los Angeles
The Alan Turing Institute
mv472@cam.ac.uk

Abstract

Deciding how to optimally treat a patient, including how to select treatments over time among the multiple available treatments, represents one of the most important issues that need to be addressed in medicine today. A dynamic treatment regime (DTR) is a sequence of treatment rules indicating how to individualize treatments for a patient based on the previously assigned treatments and the evolving covariate history. However, DTR evaluation and learning based on offline data remain challenging problems due to the bias introduced by time-varying confounders that affect treatment assignment over time; this may lead to suboptimal treatment rules being used in practice. In this paper, we introduce Gradient Regularized V-learning (GRV), a novel method for estimating the value function of a DTR. GRV regularizes the underlying outcome and propensity score models with respect to the optimality condition in semiparametric estimation theory. On the basis of this design, we construct estimators that are efficient and stable in finite samples regime. Using multiple simulation studies and one real-world medical dataset, we demonstrate that our method is superior in DTR evaluation and learning, thereby providing improved treatment options over time for patients.

1 Introduction

Clinical decision-makers regularly face the daunting challenge of choosing from multiple treatment options and treatment timings. While clinical trials represent the gold standard for causal inference, clinical trials for longitudinal studies are expensive to conduct. They have few patients and narrow inclusion criteria, and usually do not follow patients for long periods of time. Leveraging increasingly available observational data about patients, such as electronic health records, represents a more viable alternative for developing individualized treatment plans over time. Across multiple communities, including machine learning, statistics and economics, increasing attention has been paid to the need to understand how decision-makers decide which treatments to give patients, and how to construct treatment rules based on their static information [1, 2, 3, 4, 5, 6, 7].

Treatment individualization and adaptation over time are crucial for managing chronic diseases. For example, the time-varying patient information, such as side-effect severity and treatment adherence, drives the treatment for major depressive disorder [8]; clinicians routinely adjust therapy based on the risk of toxicity and antibiotics resistance in treating cystic fibrosis [9]. Unlike in static settings, developing time-varying treatment rules in longitudinal settings poses unique opportunities to understand how diseases evolve under different treatment plans, how individual patients respond to medication over time, and which timings are optimal for assigning treatments. Dynamic treatment regimes (DTRs) [10, 11] offer an attractive causal inference framework that serves this purpose. A DTR is a sequence of time-varying treatment rules that determine which treatment to provide

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

at each time step, given the patient's ongoing observation and evolving treatment history. These time-varying treatment rules are also known as adaptive treatment strategies [12, 13, 14, 15, 16] or treatment policies [17, 18, 19]. DTRs provide an effective vehicle for several application areas that have recently been discussed in biomedical literature, including personalized management of chronic conditions for cancer, diabetes, and mental illnesses [20]. Moreover, DTRs can be used to determine when to use chemotherapy or radiotherapy on lung cancer patients for controlling the growth of their tumour volume, when to stop using antibiotics treatment to decrease the white blood cell count (indicative of severe illness and poor outcome) for ICU patients, and when to use ventilation on patients with COVID-19 to maximize their survival outcomes.

To estimate the value function of a DTR, there are three classes of estimators to consider in the literature. Each of these classes has limitations in practice. It is well known that under some circumstances importance sampling-based estimators have high variance, while regression-based estimators have large biases. Doubly robust (DR) estimators have desirable asymptotic properties, but they can still suffer from high variance in finite samples, due to the unstable inverse propensity score (IPS) product. More in-depth discussion of existing methods is provided in Section 3.

In this paper, we introduce Gradient Regularized V-learning (GRV) as a new method for estimating the value function of a DTR. The GRV estimator is constructed to have stable finite sample behaviour while achieving the asymptotic efficiency. GRV achieves this by regularizing the underlying machine learning models to satisfy the estimating equation in semiparametric estimation theory. We prove theoretically that the estimating equation is satisfied when our proposed regularizer is minimized. Our work can be viewed as an extension of the targeted regularization (TR) method [21] to DTR settings. TR is a debiased method for estimating the average treatment effect (ATE) in static settings. Our regularizer is designed to solve the complex estimating equation for DTR evaluation, which involves the outcome and propensity score models across different time steps. Our regularizer can also be applied to estimate the average long-term effect of multiple treatments in longitudinal studies. With experiments on multiple simulation studies, we demonstrate that our method is superior to baseline methods in terms of accuracy in estimating value functions and learning better DTRs. We also demonstrate the application of our method on a real-world dataset extracted from the Medical Information Mart for Intensive Care (MIMIC) database [22].

2 DTR Problem

Under the standard assumptions of sequential ignorability, consistency and overlap [10, 23, 24] in the potential outcome framework for causal inference, we consider a dataset consisting of treatment trajectories from a population of N units. For each $i \in [N] = \{1, \dots, N\}$, unit i has a baseline covariate vector in $Z_i \in \mathcal{Z}$. At each time $t \in [T] = \{1, \dots, T\}$, unit i has a covariate vector $X_{i,t} \in \mathcal{X}_t$, outcome variable $Y_{i,t} \in \mathcal{Y}_t$, and a one-hot treatment vector $A_{i,t} \in \mathcal{A} = \{0,1\}^K$ which indicates which of the K treatment options is assigned to the unit i. In the case of two treatments, we have four options: no treatments, treatment 1, treatment 2, and both treatments.

We collect all the observations about unit i by $O_i = (Z_i, X_{i,[T]}, A_{i,[T]}, Y_{i,[T]})$. The vectors $O_{[N]}$ are N i.i.d sampled copies of $O \in (Z, X_{[T]}, A_{[T]}, Y_{[T]}) \sim P(O) = P$. We denote the history up to time t by $H_t = (X_{[t]}, A_{[t]}, Z) \in \mathcal{H}_t$, $\tilde{H}_t = (X_{[t]}, A_{[t-1]}, Z) \in \tilde{\mathcal{H}}_t$ and the baseline covariates Z by H_0. Then we can factorize the joint distribution P as

$$P(O) = P(H_0) \prod_{t=1}^{T} P(X_t \mid H_{t-1}) P(A_t \mid \tilde{H}_t) P(Y_t \mid H_t) = g(O) f(O) h(O), \tag{1}$$

where $g(O) = \prod_{t=1}^{T} P(A_t \mid \tilde{H}_t)$ is the intervention distribution, $f(O) = \prod_{t=1}^{T} P(Y_t \mid H_t)$ is the outcome distribution, and $h(O) = P(H_0) \prod_{t=1}^{T} P(X_t \mid H_{t-1})$ is the covariate distribution.

We denote the time t treatment rule by $d_t : \tilde{\mathcal{H}}_t \to \mathcal{A}$, which assigns the treatment A_t as a function of the history $\tilde{H}_t$. A dynamical treatment regime (DTR), $d := d_{1:T} = \{d_1, \dots, d_T\}$, is a set that collects all the treatment rules over time. We let $P_{d_{t:T}}$ denote the joint distribution under the intervention of $d_{t:T}$, i.e., replacing $P(A_s \mid \tilde{H}_s)$ in $P(O)$ with the indicator function $\mathbb{1}\big(A_s = d_s(\tilde{H}_s)\big)$, for

$s = t, \ldots, T$. The time t value function under $d_{t:T}$ is given as

$$V_t(d_{t:T}) = \mathbb{E}_{P_{d_{t:T}}}\left[\sum_{r=t}^{T} Y_r\right] = \int \sum_{r=t}^{T} Y_r \frac{dP_{d_{t:T}}}{dP} dP = \mathbb{E}_P\left[\frac{\prod_{s=t}^{T} \mathbb{1}\big(A_s = d_s(\tilde{H}_s))}{\prod_{s=t}^{T} P(A_s|\tilde{H}_s)} \sum_{r=t}^{T} Y_r\right]. \tag{2}$$

The optimal DTR is given as the optimizer $d^* = \arg\max_d V(d)$ where $V(d)$ is the value function over the entire treatment trajectory. To learn the optimal DTR d^*, one can first consider developing a sample-efficient and low variance estimator to evaluate $V(d)$, as will be discussed in Section 3.

3 Related works

In the literature, DTR learning in causal inference and off policy evaluation (OPE) in batch reinforcement learning (BRL) are the two branches of methods that attempt to solve the estimation problem of $V(d)$. Here, we summarize the existing value function estimators in three classes.

Importance Sampling estimators. The importance sampling (IS) estimator of $V_t(d_{t:T})$ is given by the empirical version of the R.H.S of Equation (2), with $P(A_s \mid \tilde{H}_s)$ in the denominator approximated by a propensity score model $\hat{g}_s(A_s, \tilde{H}_s)$. In the DTR literature, backward outcome weighted learning (BOWL) [6] is a method that derives the treatment rule $\hat{d}_t$ by optimizing the IS estimator of $V_t(d_t, \hat{d}_{t+1:T})$ backwardly through time given the previously optimized rules $\hat{d}_{t+1:T}$. Simultaneous outcome weighted learning (SOWL) [6] optimizes the treatment rules in d jointly based on the IS estimator of $V(d)$. In BRL, IS estimators [25, 26, 27] are used to evaluate the value function under a target policy by reweighting the rewards in the historical data with the probability ratio of the target policy and the policy that generates the data. IS estimators are known to be consistent and unbiased but suffer high variance due to the inverse propensity score product.

Regression-based estimators. Batch Q-learning (BQL) [28] is a classic machine learning method that uses regression models to directly estimate the value function of a DTR. The term "batch" is used to emphasize an important difference from standard Q-learning [29, 30], in which learning occurs only after the collection of the training set. Alternatively, A-Learning [10, 31] models regret functions which measure the loss incurred by not following the optimal DTR. More discussion on the relationship between Q- and A-learning can be found in [32]. The variants of BQL in OPE are known as directed methods (DM) [33, 34, 35]. Here, we briefly describe BQL and its limitations.

In BQL, we define the time T Q-function as $Q_T(a_T, \tilde{h}_T) = \mathbb{E}\big[Y_T \mid A_T = a_T, \tilde{H}_T = \tilde{h}_T\big]$. When the underlying data generating distribution P is known, dynamic programming (DP) [36, 37] shows that $d^*_T(h_T) \in \arg\max_{a_T \in \mathcal{A}} Q_T(a_T, \tilde{h}_T)$ and recursively $d^*_t(h_t) \in \arg\max_{a_t \in \mathcal{A}} Q_t(a_t, \tilde{h}_t)$ where $Q_t(a_t, \tilde{h}_t) = \mathbb{E}\big[Y_t(a_t) + \max_{a_{t+1} \in \mathcal{A}} Q_{t+1}(a_{t+1}, \tilde{H}_{t+1}) \mid A_t = a_t, \tilde{H}_t = \tilde{h}_t\big]$. BQL approximates each Q_t using some machine learning models $\hat{Q}_t$, and updates them backwards through time $t = T, T-1, \ldots, 1$ by solving the optimization problem:

$$\hat{Q}_t \in \underset{Q_t \in \mathcal{F}}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \left[Y_{i,t} + \max_{a_{t+1} \in \mathcal{A}} \hat{Q}_{t+1}(a_{t+1}, \tilde{H}_{i,t+1}) - Q_t(A_{i,t}, \tilde{H}_{i,t})\right]^2 . \tag{3}$$

The optimal time t treatment rule is derived as $\hat{d}_t(\tilde{H}_t) = \arg\max_{a_t \in \mathcal{A}} \hat{Q}_t(a_t, \tilde{H}_t)$. Despite BQL being able to produce low-variance estimators, the bias in these estimators is hard to quantify. Learning the Q-functions with Equation (3) implicitly restricts the solution to be in a postulated function class $\tilde{\mathcal{F}}$ and induces a corresponding class of DTRs given by the arg max of the postulated Q-functions. The work [38] shows that even when the optimal DTR resides in the class of induced DTRs, Q-learning can still fail to be consistent when $\tilde{\mathcal{F}}$ is misspecified. Hence, in Q-learning, the quality of DTR optimization will rely on the correctness of postulated regression models which we can not quantify easily. This limitation of Q-learning shows the advantage of simultaneous learning of $V(d)$ in some importance sampling based methods.

Doubly robust estimators. Developing doubly robust (DR) estimators is a well-studied extension in the literature [39, 40, 41, 42, 43] to improve value function estimation in terms of sample efficiency and robustness to model misspecification. These DR estimators have a similar expression as the well-known augmented inverse probability weighted (AIPW) estimator [44, 45] for the average treatment effect (ATE) estimation in static settings. Despite their asymptotic optimality, they can be unstable in finite samples due to the IPS product. For example, the AIPW estimator in [42] is given as

$$\hat{V}_{\text{AIPW}}(d) = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{C_\delta}{\hat{g}_{i,1:T}}\Big(\sum_{t=1}^{T} Y_{i,t}\Big) + \sum_{t=1}^{T}\frac{C_{\delta,t}}{\hat{g}_{i,1:t}}\hat{Q}_t(A_{i,t}, \tilde{H}_{i,t})\right\}, \tag{4}$$

where C_δ and $C_{\delta,t}$ are some indicator functions of whether DTR treatments align with observed data, and $\hat{g}_{i,1:t}$ is a function of $\prod_{s=1}^{t}\hat{g}_t(A_{i,s}, \tilde{H}_{i,s})$. By adaptively using a DR estimator for the value function in part of a trajectory and then using a regression-based estimator for the remainder, the weighted doubly-robust (WDR) and MAGIC estimator in [40] offer more accurate evaluation than the standard DR estimators. This result sheds the light on the need for an estimator that has a stable expression against high variance due to the IPS products while achieving the asymptotic optimality. Next, we introduce our method, Gradient Regularized V-Learning (GRV), that fills this need.

4 Gradient Regularized V-Learning

This section is organized as follows: (1) we first introduce a set of the outcome and propensity models, $\mathcal{M}_{t:T}$, required for estimating the value function $V_t(d_{t:T})$; (2) we discuss under what condition $\hat{\mathcal{M}}_{t:T}$ can construct an efficient estimator of $V_t(d_{t:T})$ in semiparametric theory; (3) we describe our neural network architecture that parameterizes the models in $\hat{\mathcal{M}}_{t:T}$, and a theory that demonstrates that our proposed regularizer can encourage the models to satisfy the optimality condition of efficient estimators; and (4) we briefly discuss the two GRV based DTR learning algorithms.

4.1 Set up

The time t value function $V_t(d_{t:T}) = \mathbb{E}_{P_{d_{t:T}}}\big[\sum_{r=t}^{T} Y_r\big]$ is defined as the sum of the mean outcomes after time $t-1$ under the intervention of $d_{t:T}$, in which

$$Y_r = Y_r(d_{t:r}) = \sum_{a_{t:r}\in\mathcal{A}^{r-t}}\prod_{s=t}^{r}\mathbb{1}(d_s(\tilde{H}_s) = a_s)Y_r(a_{t:r}),$$

for $r = t, \ldots, T$. To model these intervened outcomes using the data generated by the distribution $P(O)$ defined in Equation (1), we make no parametric assumptions on the joint covariate distribution $h(O)$, but make parametric assumptions on the joint outcome distribution $f(O)$ and intervention distribution $g(O)$. We let $\mu_{t,t}(A_t, \tilde{H}_t)$ denote the mean of Y_t under the distribution $P(Y_t \mid H_t)$, and $g_t(A_t, \tilde{H}_t)$ denote the propensity score $P(A_t \mid \tilde{H}_t)$. We let $\mu_{t,r}(d_{t:r}, \tilde{H}_t)$ denote the mean of Y_r given the intervention of $d_{t:r}$ and the history $\tilde{H}_t$. In our notation of $\mu_{t,r}$, "t" indicates the given history is up to time t and "r" is the time of the future outcome Y_r. This long-term expectation can be computed via sequential regression. Starting backward from time r, given $\mu_{r,r}\big(A_r, \tilde{H}_r\big)$, we can compute $\mu_{r,r}(d_r, \tilde{H}_r)$ as $\mu_{r,r}(d_r(\tilde{H}_r), \tilde{H}_r) = \mathbb{E}\big[Y_r \mid A_r = d_r(\tilde{H}_r), \tilde{H}_r\big]$. Then by the law of iterated expectations, for $s = t, \ldots, r-1$, we can compute

$$\mu_{s,r}(d_{s+1:r}, H_s) = \mathbb{E}\big[\mu_{s+1,r}(d_{s+1:r}, \tilde{H}_{s+1}) \mid H_s\big]. \tag{5}$$

In practice, $\mu_{s,r}(d_{s+1:r}, H_s)$ is computed by regressing $\mu_{s+1,r}(d_{s+1:r}, \tilde{H}_{s+1})$ onto H_s. After $r-t$ regressions, we obtain the wanted expectation $\mu_{t,r}(d_{t:r}, \tilde{H}_t)$. We let $\mathcal{M}_{t:T}$ collect all the mean outcome and propensity score models that we need to compute in this sequential procedure for estimating the time t value function $V_t(d_{t:T})$. It is worth noting that the simplest treatment rules are the ones that assign the same treatment to all the units. The method we propose later also works for estimating the average long-term outcome $\mathbb{E}\big[Y_r(a_{t:r})\big]$ given some specific treatment plan $A_{t:r} = a_{t:r}$. We can estimate the effect of a target treatment plan by contrasting its outcome Y_r with the one under some simple treatment plans, such as the plan that never assigns any treatments.

4.2 Efficiency

We now turn to discuss how to construct a consistent and efficient estimator of $V_t(d_{t:T})$ based on semiparametric theory. For a complete textbook presentation of the theory, we refer the reader to

[46, 47, 48, 49, 50]. We let $\mathcal{P}_{t:T}$ denote the space of the data distribution $P_{t:T}$ of $\tilde{H}_t$, $X_{t+1:T}$, $A_{t:T}$ and $Y_{t:T}$, and $\hat{P}_{t:T}$ denote a estimate of $P_{t:T}$ with a set of nuisance models $\hat{\mathcal{M}}_{t:T}$. We define the target parameter mapping as $\Psi_t : \mathcal{P}_{t:T} \to \mathbb{R}$ such that $\Psi_t(P_{t:T}; d_{t:T}) = V_t(d_{t:T})$.

For parametric models with finite-dimensional parameters, we can define the nuisance tangent space as the subspace spanned by the nuisance score vector which is the gradient of $\log P_{t:T}$ w.r.t the parameters of the nuisance models. Similarly, we can define the target score vector for the target parameters. The efficient influence curve (EIC) is a scaled residual obtained by projecting the target score vector onto the nuisance tangent space. The EIC is the influence curve with the smallest variance equal to the inverse of the Fisher Information, which is well-known to be the Cramer-Rao lower bound on the variance of unbiased estimators of the target parameters and attainable via maximum likelihood estimation (MLE) under regularity assumptions.

In semiparametric models, we cannot define scores for the infinite-dimensional nuisance parameters. However, if we assume that the target parameter mapping Ψ_t is a pathwise differentiable functional at each distribution in $\mathcal{P}_{t:T}$, the EIC is given as the canonical gradient of Ψ_t, $D_t^*(\hat{P}_{t:T}; d_{t:T})$. This defines an estimator that is asymptotically efficient and linear with the EIC. To see this, we use the fact that the pathwise differentiable $\Psi_t(\hat{P}_{t:T}; d_{t:T}) = \hat{V}_t(d_{t:T})$ admits a von Mises Expansion [47]:

$$\hat{V}_t(d_{t:T}) - V_t(d_{t:T}) = (\hat{\mathbb{P}}_{t:T} - \mathbb{P}_{t:T})D_t^*(\hat{P}_{t:T}; d_{t:T}) + R_2(\hat{P}_{t:T}, P_{t:T}),$$

where $R_2(\hat{P}_{t:T}, P_{t:T})$ is the second-order remainder. Using the fact that EIC has zero mean such that $\hat{\mathbb{P}}_{t:T}D_t^*(\hat{P}_{t:T}; d_{t:T}) = 0$, we can decompose the first-order bias $-\mathbb{P}_{t:T}D_t^*(\hat{P}_{t:T}; d_{t:T})$ as

$$(\mathbb{P}_N - \mathbb{P}_{t:T})\big[D_t^*(\hat{P}_{t:T}; d_{t:T}) - D_t^*(P_{t:T}; d_{t:T})\big] + (\mathbb{P}_N - \mathbb{P}_{t:T})D_t^*(P_{t:T}; d_{t:T}) - \mathbb{P}_N D_t^*(\hat{P}_{t:T}; d_{t:T}).$$

In the decomposition, the first term will be $O_{P_{t:T}}(1/\sqrt{N})$ under empirical process conditions, e.g., if the nuisance models in $\hat{\mathcal{M}}_{t:T}$ are regular enough so that $D_t^*(\hat{P}_{t:T}; d_{t:T})$ lies in a Donsker class of functions, or if sample splitting is used so that the nuisance models are constructed on separate data. The second term is a sample-average of a fixed function $D_t^*(P_{t:T}; d_{t:T})$, and thus converges to a normal distribution after $\sqrt{N}$-scaling, by the central limit theorem. The remainder $R_2(\hat{P}_{t:T}, P_{t:T})$ can be $O_{P_{t:T}}(1/\sqrt{N})$ even when nuisance models converge at slower rates. Overall, we can have

$$\hat{V}_t(d_{t:T}) - V_t(d_{t:T}) = -\mathbb{P}_N D_t^*(\hat{P}_{t:T}; d_{t:T}) + O_{P_{t:T}}(1/\sqrt{N}), \tag{6}$$

which implies that $\hat{V}_t(d_{t:T})$ is a $\sqrt{N}$-consistent estimator of $V_t(d_{t:T})$ if we can set the sample-average EIC $\mathbb{P}_N D_t^*(\hat{P}_{t:T}; d_{t:T})$ to zero. The variance of the EIC gives the generalized Cramer-Rao lower bound of the estimator variance [50]. The estimator $\hat{V}_t(d_{t:T})$ reaches this lower bound asymptotically and hence efficient. A general introduction of deriving the EIC for large models can be found in [51]. In our case, the EIC takes the form

$$D_t^*(P_{t:T}; d_{t:T}) = \sum_{r=t}^{T}\Big(\mathbb{E}\big[Y_r(d_{t:r}) \mid \tilde{H}_t\big] - \mathbb{E}\big[Y_r(d_{t:r})\big]\Big) + \sum_{r=t}^{T}\sum_{s=t}^{r} D_{s,r}^*(P_{t:T}; d_{t:T}). \tag{7}$$

Let $\delta_{t:k}/g_{t:k}$ denote the IPS product $\prod_{m=t}^{k} \mathbb{1}(A_m = d_m(\tilde{H}_m))/\prod_{m=t}^{k} P(A_m \mid \tilde{H}_m)$. In the second term of Equation (7), as $s = t, \dots, r-1$,

$$D_{s,r}^*(P_{t:T}; d_{t:T}) = \frac{\delta_{t:s}}{g_{t:s}}\Big(\mathbb{E}\big[Y_r(d_{s+1:r}) \mid \tilde{H}_{s+1}\big] - \mathbb{E}\big[Y_r(d_{s:r}) \mid \tilde{H}_s\big]\Big), \tag{8}$$

and as $s = r$,

$$D_{r,r}^*(P_{t:T}; d_{t:T}) = \frac{\delta_{t:r}}{g_{t:r}}\Big(Y_r - \mathbb{E}\big[Y_r(d_r) \mid \tilde{H}_r\big]\Big). \tag{9}$$

Suppose we define a fluctuated estimate $\hat{P}_{t:T,\epsilon}$ of $P_{t:T}$ so that a small change in ϵ corresponds with a maximal small change in the estimator $\Psi_t(\hat{P}_{t:T,\epsilon}; d_{t:T})$ on the data, the local maximum ϵ^* is given by solving the estimating equation: $\mathbb{P}_N D_t^*(\hat{P}_{t:T,\epsilon*}; d_{t:T}) \approx 0$. Targeted minimum loss estimation (TMLE) [52, 53, 54, 55] is an estimation procedure that relies on an extra model parameter ϵ to fine-tune the initial nuisance models μ and g for solving the estimating equation. For a long EIC expression in Equations (7) to (9), TMLE would need to be applied to the models at each time step separately. It has been found even in the simple static setting of the average treatment effect (ATE) estimation (with only one μ and g model) [21] that the performance of TMLE is dependent on the quality of the initial models, and tuning the initial models and ϵ jointly can lead to more robust performance. To improve the quality of the initial models, one can also consider using cross-validation based methods such as super learning [50] that constructs an initial model by combining multiple machine learning models. However, the problem of TMLE is more severe in DTR settings which involves a relatively large number of initial models over time. Also, it is computationally prohibitive to use cross-validation repeatedly when we are optimizing a DTR.

4.3 Model

Next, we present our method, GRV, that regularizes all the nuisance models with a fluctuation parameter ϵ_t to solve the estimating equation defined by the sophisticated EIC in Equations (7) to (9) in the training time of the nuisance models. We prove theoretically that the equation is solved when the regularizer is minimized. Such a strategy simplifies the process of solving multiple estimating equations and avoids using bad initial nuisance models to construct the value function estimator.

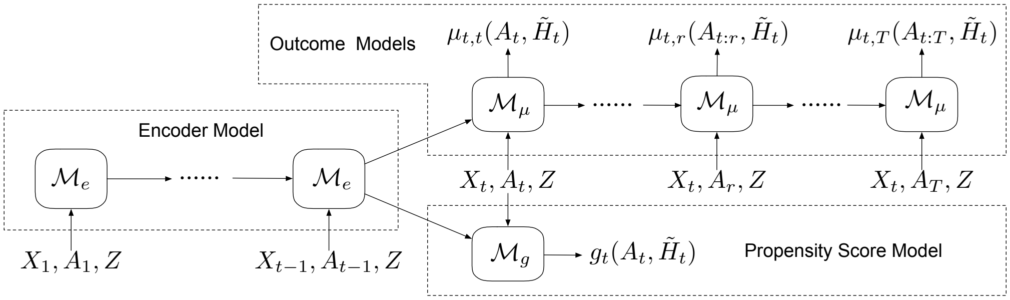

Figure 1: The network architecture of outcome and propensity score models

We start by introducing the network architecture that parameterizes the nuisance models. Our network is given as a composition of standard RNN models and feedforward neural network (i.e. multilayer perceptron) models. We opt for neural network models because they enable end-to-end training of the nuisance models. Figure 1 gives a pictorial description of the network architecture. We use a sequence encoder model $\mathcal{M}_e$ to extract information from the history H_{t-1}. At each time step s, the encoder $\mathcal{M}_e$ takes the covariates X_s, treatment variables A_s, and baseline covariates Z concatenated together as its input. We feed in the baseline covariates repeatedly for learning its interaction with the covariate and treatment variables at each time step. At time t, we use a sequence decoder model $\mathcal{M}_\mu$ to parametrize the outcome models $\mu_{t,r}(A_{t:r}, \tilde{H}_t), r = t, \ldots, T$. The decoder $\mathcal{M}_\mu$ takes the hidden states of $\mathcal{M}_e$ at time $t-1$ as its initial state. The time r input of $\mathcal{M}_\mu$ consists of X_t, A_r and Z. The propensity score models $g_s(A_s, \tilde{H}_s), s = t, \ldots, T$, are parameterized by a shared multilayer perceptron. The model $g_t(A_t, \tilde{H}_t)$ takes the hidden states of $\mathcal{M}_e$ at time $t-1$, X_t and Z as input, and use a softmax output layer to estimate the probability of A_t. We note that using a shared encoder is a common strategy in the recent neural network literature for treatment effect estimation. However, such a strategy is only beneficial when the outcomes and propensity scores are some similar functions, e.g., functions with the same dependence on the input covariates. On the contrary, if the outcomes and propensity scores are some different functions, we should model them with two separate encoders. For example, the treatment assignment is randomized while the outcomes have some complex dependence on the input covariates.

Objective function. To simplify the notation in our objective function, we define $\hat{\delta}_{i,s:r} = \prod_{m=s}^{r} \mathbb{1}\big(A_{i,m} = \hat{d}_m(\tilde{H}_{i,m})\big)$, $g_{i,s:r} = \prod_{m=s}^{r} g_m(A_{i,m}, \tilde{H}_{i,m})$, and $\hat{d}_{i,s:r} = \big[\hat{d}_m(\tilde{H}_{i,m})\big]_{m=s}^{r}$. Given the treatment rules $\hat{d}_{t:T}$, we optimize $\mathcal{M}_{\text{NN}} = \{\mathcal{M}_e, \mathcal{M}_\mu, \mathcal{M}_g\}$ with the objective function

$$\hat{\mathcal{M}}_{\text{NN}}, \hat{\epsilon}_t = \underset{\mathcal{M}_{\text{NN}}, \epsilon_t}{\arg\min} \Big\{ \mathcal{L}_t\big(\mathcal{M}_{\text{NN}}; O_{[N]}\big) + \mathcal{R}_t\big(\mathcal{M}_{\text{NN}}, \epsilon_t; \hat{d}_{t:T}, O_{[N]}\big) \Big\}, \tag{10}$$

where $\mathcal{L}_t\big(\mathcal{M}_{\text{NN}}; O_{[N]}\big)$ is given as

$$\mathcal{L}_t\big(\mathcal{M}_{\text{NN}}; O_{[N]}\big) = \sum_{s=t}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \left[\mu_{s,s}(A_{i,s}, \tilde{H}_{i,s}) - Y_{i,s}\right]^2 + \frac{1}{N} \sum_{i=1}^{N} \text{CrossEntropy}\big[g_{i,s}, A_{i,s}\big] \right),$$

where the first term is the loss for the outcome models, and the second term is the loss for the propensity score models. The regularizer $\mathcal{R}_t\big(\mathcal{M}_{\text{NN}}, \epsilon_t; \hat{d}_{t:T}, O_{[N]}\big)$ is factorized as $\mathcal{R}_t\big(\mathcal{M}_{\text{NN}}, \epsilon_t; \hat{d}_{t:T}, O_{[N]}\big) = \sum_{r=t}^{T} \sum_{s=t}^{r} \zeta^t_{s,r}$, where each factor $\zeta^t_{s,r}$ is defined as

$$\zeta^t_{s,r} = \frac{1}{N} \sum_{i=1}^{N} \left[\mu^{\epsilon_t}_{s+1:r}\big(\hat{d}_{i,s+1:r}, \tilde{H}_{i,s+1}\big) - \mu^{\epsilon_t}_{s:r}\big(\hat{d}_{i,s:r}, \tilde{H}_{i,s}\big)\right]^2, \text{ as } s = t, \ldots, r-1,$$

and $\zeta^t_{r,r} = \frac{1}{N}\sum_{i=1}^{N} \left[Y_{i,r} - \mu^{\epsilon_t}_{r,r}(\hat{d}_{i,r}, \tilde{H}_{i,r})\right]^2$, where $\mu^{\epsilon_t}_{k,r}$ is a fluctuated outcome model,

$$\mu^{\epsilon_t}_{k,r}\big(\hat{d}_{i,k:r}, \tilde{H}_{i,k}\big) = \mu_{k,r}\big(\hat{d}_{i,k:r}, \tilde{H}_{i,k}\big) + \epsilon_t \sum_{m=k}^{r} \frac{\hat{\delta}_{i,t:m}}{g_{i,t:m}}, \text{ for } r = t, \ldots, T,\ k = t, .., r.$$

The following theorem shows that minimizing $\mathcal{R}_t\big(\mathcal{M}_{\text{NN}}, \epsilon_t; \hat{d}_{t:T}, O_{[N]}\big)$ in Equation (10) encourages these fluctuated outcome models to solve the estimating equation, hence the resulting estimator is tailored to be efficient for estimating the value function $V_t(\hat{d}_{t:T})$ as discussed in Section 4.2. Since we already adapt all the nuisance models to solve the estimating equation, we have the freedom to scale or penalize the magnitude of ϵ_t to be small in training, which can partially relieve the high variance issue caused by the unstable IPS product. We note that minimizing $\mathcal{L}_t\big(\mathcal{M}_{\text{NN}}; O_{[N]}\big)$ is still important for the convergence of the remainder in Equation (6).

Theorem 1. *Given some treatment rules $\hat{d}_{t:T} = \{\hat{d}_t, \ldots, \hat{d}_T\}$, the nuisance models $\hat{\mu}$ and $\hat{g}$ parameterized by $\hat{\mathcal{M}}_{NN}$ and $\hat{\epsilon}_t$ satisfy the estimating equation when $\mathcal{R}_t$ is minimized such that*

$$0 = \partial_{\epsilon_t}\Big[\mathcal{R}_t\big(\hat{\mathcal{M}}_{NN}, \epsilon_t; \hat{d}_{t:T}, O_{[N]}\big)\Big]\Big|_{\epsilon_t = \hat{\epsilon}_t} = \mathbb{P}_N D^*_t(\hat{P}_{t:T}; \hat{d}_{t:T}). \tag{11}$$

Proof. Appendix A. □

Recall that $\hat{\mu}^{\hat{\epsilon}_t}_{t,r}\big(d_{t:r}, \tilde{H}_t\big)$ estimates $\mathbb{E}\big[Y_r(d_{t:r}) \mid \tilde{H}_t\big]$, the mean of Y_r given the intervention of $d_{t:r}$ and the history $\tilde{H}_t$. We can construct an empirical estimate $\hat{V}_t(d_{t:T})$ of the time t value function $\hat{V}_t(d_{t:T})$ by summing up the estimated outcomes over time. We can learn each d_t backwardly in a dynamic programming fashion, by optimizing the estimator of $\hat{V}_t(d_t, \hat{d}_{t+1:T})$ given the previously optimized rules $\hat{d}_{t+1:T}$. We also attempt to optimize all the treatment rules in d jointly. We denote the former DTR optimization algorithm as GRV-B and the latter as GRV-S. Both algorithms are computationally expensive because we need to update the nuisance models after each update of the DTR. Our hope is that training the DTR for one epoch would not change the DTR dramatically, then we only need to retrain the nuisance models for a few epochs. GRV-S is cheaper than GRV-B but unstable in practice. Finally, we note that optimizing efficient value function estimates is a popular strategy for learning treatment rules in the literature. However, even if the estimates are efficient, it does not imply that we have an efficient learning algorithm of the optimal DTR. It has been found recently that efficient gradient estimation is crucial for achieving efficient DTR learning [56, 57, 58].

5 Experiments

We evaluate the performance of GRV[1] in DTR evaluation and optimization. The first set of our experiments is based on two non-Markovian simulation studies adapted from [43]. We call them treatment cost trade-off and survival rate maximization respectively. In the first simulation, the DTR objective is to minimize the total treatment cost in the treatment trajectory while keeping the health metric above the threshold at the end of the trajectory. In the second simulation, the DTR objective is to maximize the survival rate of cancer patients using an invasive and non-invasive treatment in

[1] The code is provided at: https://bitbucket.org/mvdschaar/mlforhealthlabpub.

Table 1: Performance of benchmarks, our estimator and algorithms: the MSEs of the value function estimators (lower is better) and the value of the learned DTRs (higher is better). The mean and standard deviation are averaged across 20 independent runs on a testing set with 20,000 individuals.

Metrics	MSE				Value			
Dataset	Treatment cost trade-off							
Samples	IPW	BQL	AIPW	GRV	IPW	BQL	GRV-S	GRV-B
1000	.375 ± .018	.362 ± .003	.139 ± .006	**.134** ± **.004**	.741 ± .002	.748 ± .002	.753 ± .004	**.761** ± **.003**
5000	.223 ± .011	.348 ± .003	.035 ± .004	**.031** ± **.003**	.760 ± .002	.752 ± .001	.759 ± .003	**.767** ± **.002**
10000	.132 ± .013	.291 ± .002	**.023** ± **.003**	**.022** ± **.002**	.765 ± .001	.754 ± .001	.762 ± .002	**.768** ± **.002**
Dataset	Survival rate maximization							
Samples	IPW	BQL	AIPW	GRV	IPW	BQL	GRV-S	GRV-B
1000	.332 ± .035	.034 ± .001	.047 ± .005	**.032** ± **.002**	.154 ± .009	.137 ± .006	.146 ± .005	**.182** ± **.004**
5000	.065 ± .003	.017 ± .001	.006 ± .002	**.005** ± **.001**	.175 ± .008	.198 ± .003	.201 ± .002	**.224** ± **.002**
10000	.031 ± .004	.011 ± .001	**.003** ± **.001**	**.003** ± **.001**	.188 ± .006	.215 ± .002	.213 ± .002	**.226** ± **.002**

the treatment trajectories. We benchmark GRV against three methods that cover the three estimator classes summarized in Section 3, including IPW/IS estimator [26, 27], batch Q-learning (BQL) [28, 33], and AIPW estimator [40, 42]. The simulation and implementation details of GRV and benchmark methods can be found in Appendix B.

Given some offline data from a simulation environment and a target DTR, we want to test how accurate the methods can estimate the value of the target DTR. We compare the methods in terms of their mean squared errors (MSEs) in estimating the value of the optimal DTR. For DTR learning, we compare GRV-S and GRV-B against the benchmarks in terms of the value achieved by their learned DTRs. The DTR value is defined as the cumulative outcome under the learned DTR. Specifically, we generate a large testing dataset with 20,000 individuals. For each learned DTR, we let each individual in the testing set follow the treatments recommended by the DTR. At the end of the treatment trajectory, we obtain the cumulative outcome for each individual. We then compute the value of the DTR by averaging the cumulative outcome over all the individuals.

In Table 1, we report the MSEs of the value function estimators and the value of the learned DTRs. We repeat the same experiments with training sample size at 1000, 5000 and 10000. When the sample size is 1000 or 5000, our GRV estimator outperforms the competing estimators in estimating the value of the optimal DTR. When the sample size is 10000, AIPW and GRV achieve similar MSE performance. Additionally, the DTRs learned by GRV-B achieve larger value than the DTRs learned by the other methods. GRV-S performs relatively poorly because optimizing all the treatment rules together is unstable when the trajectory is long. Overall, the performance gain highlights the effectiveness of our regularizer, which constructs an efficient value function estimator by adapting the nuisance models to solve the estimating equation during the training process of the nuisance models.

In Appendix C, we show our method performs well on three additional simulation studies from [6] where the covariate vector is relatively high-dimensional while the training set is small with hundreds of samples. We also provide an ablation study which shows that ablating the regularizer leads to worse performance in our method. In Appendix D, we demonstrate the application of our method to antibiotics administration on a real-world dataset extracted from the MIMIC database [22]. We note that this experiment is only used to illustrate that the learned DTR will lead to gradually decrease the usage of the antibiotics treatment, which corresponds to the clinical literature. The real DTR application can only be demonstrated well through real-world experiments and simulation. For example, in a real-world test-bed, the clinician can check if his/her treatment decision is correct or not by comparing it with the DTR recommendation before prescribing it.

6 Conclusion

We have introduced Gradient Regularized V-Learning (GRV), a novel regularization method that enables recurrent neural network models to estimate the value function of a target DTR accurately and learn better DTRs. We prove theoretically that the nuisance models satisfy the estimating equation in semiparametric estimation theory when the proposed regularizer is minimized. We hope that GRV will become a useful regularization method when RNNs are deployed to tackle the challenges of treatment effect estimation and decision making in machine learning.

Broader Impact

Our work can help to develop accurate and individualized decision-making system in many real-world applications, such as treatment recommendation. Our method is easy to use for machine learning practitioners since it simply relies on regularization of the underlying recurrent network models. However, Offline DTR or policy evaluation is still a challenging problem because we often do not have sufficient samples to estimate the outcomes for all the possible treatment plans over time. Given some datasets that are high-dimensional or have long treatment trajectories, we would need to combine methods that solve different challenges of DTR evaluation, such as putting constraints on treatment assignment mechanisms over time or selecting variables that are most likely to affect the treatment decisions. Our work focuses on a particular aspect of DTR evaluation and does not cover other aspects that are also important for the real-world applications of DTRs.

Acknowledgments

This work was supported by GlaxoSmithKline (GSK), the US Office of Naval Research (ONR), and the National Science Foundation (NSF) 1722516. We thank all reviewers for their generous comments and suggestions.

References

[1] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

[2] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

[3] Maximilian Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 98(1):111–131, 2016.

[4] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. *Proceedings of machine learning research*, 97:4674, 2019.

[5] Jörg Stoye. Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156, 2012.

[6] Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

[7] Onur Atan, William R Zame, Qiaojun Feng, and Mihaela van der Schaar. Constructing effective personalized policies using counterfactual inference from biased data sets with many features. *Machine Learning*, 108(6):945–970, 2019.

[8] S. A. Murphy, D. Oslin, A. Rush, and J. Zhu. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology*, 32:257–262, 2007.

[9] P. Flume, B. O'sullivan, K. Robinson, C. Goss, P. Mogayzel, D. Willey-Courand, J. Bujan, J. Finder, M. Lester, L. Quittell, R. Rosenblatt, R. Vender, Leslie A Hazle, K. Sabadosa, and B. Marshall. Cystic fibrosis pulmonary guidelines: chronic medications for maintenance of lung health. *American journal of respiratory and critical care medicine*, 176 10:957–69, 2007.

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

[11] 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.

[12] Philip W Lavori and Ree Dawson. A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):29–38, 2000.

[13] Philip W Lavori and Ree Dawson. Adaptive treatment strategies in chronic disease. *Annu. Rev. Med.*, 59:443–453, 2008.

[14] Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005.

[15] Peter F Thall, Randall E Millikan, and Hsi-Guang Sung. Evaluating multiple treatment courses in clinical trials. *Statistics in medicine*, 19(8):1011–1028, 2000.

[16] Peter F Thall, Hsi-Guang Sung, and Elihu H Estey. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97(457):29–39, 2002.

[17] Jared K Lunceford, Marie Davidian, and Anastasios A Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58(1):48–57, 2002.

[18] Abdus S Wahed and Anastasios A Tsiatis. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 60(1):124–133, 2004.

[19] Abdus S Wahed and Anastasios A Tsiatis. Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika*, 93(1):163–177, 2006.

[20] Edward H Wagner, Brian T Austin, Connie Davis, Mike Hindmarsh, Judith Schaefer, and Amy Bonomi. Improving chronic illness care: translating evidence into action. *Health affairs*, 20(6):64–78, 2001.

[21] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2503–2513, 2019.

[22] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[23] Miguel A Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(454):440–448, 2001.

[24] James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

[25] Michael JD Powell and J Swann. Weighted uniform sampling—a monte carlo technique for reducing variance. *IMA Journal of Applied Mathematics*, 2(3):228–236, 1966.

[26] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[27] A Rupam Mahmood, Hado P van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 3014–3022, 2014.

[28] Susan A Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6(Jul):1073–1097, 2005.

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

[30] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[31] Doron Blatt, Susan A Murphy, and Ji Zhu. A-learning for approximate planning. *Ann Arbor*, 1001:48109–2122, 2004.

[32] Phillip J. Schulte, A. Tsiatis, E. Laber, and M. Davidian. Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 29 4:640–661, 2014.

[33] Hoang M Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

[34] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2016.

[35] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

[36] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[37] Richard E Bellman and Stuart E Dreyfus. *Applied dynamic programming*. Princeton university press, 2015.

[38] Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

[39] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

[40] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

[41] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

[42] Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

[43] Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.

[44] James M Robins, Andrea Rotnitzky, and Mark van der Laan. On profile likelihood: comment. *Journal of the American Statistical Association*, 95(450):477–482, 2000.

[45] James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

[46] Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

[47] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[48] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[49] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[50] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[51] Jonathan Levy. Tutorial: Deriving the efficient influence curve for large models. *arXiv preprint arXiv:1903.01706*, 2019.

[52] M. J. van der Laan and A. Luedtke. Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of Causal Inference*, 3:61 – 95, 2015.

[53] Mark J van der Laan and Alexander R Luedtke. Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of causal inference*, 3(1):61–95, 2015.

[54] Iván Díaz, Nicholas Williams, Katherine L Hoffman, and Edward J Schenck. Nonparametric causal effects based on longitudinal modified treatment policies. *arXiv preprint arXiv:2006.01366*, 2020.

[55] Aurélien F Bibaut, Ivana Malenica, Nikos Vlassis, and Mark J van der Laan. More efficient off-policy evaluation through regularized targeted learning. *arXiv preprint arXiv:1912.06292*, 2019.

[56] Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. *arXiv preprint arXiv:1910.09066*, 2019.

[57] Andrew Bennett and Nathan Kallus. Efficient policy learning from surrogate-loss classification reductions. *arXiv preprint arXiv:2002.05153*, 2020.

[58] Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy policy gradients. *arXiv preprint arXiv:2002.04014*, 2020.

[59] Liesbet De Bus, Bram Gadeyne, Johan Steen, Jerina Boelens, Geert Claeys, Dominique Benoit, Jan De Waele, Johan Decruyenaere, and Pieter Depuydt. A complete and multifaceted overview of antibiotic use and infection diagnosis in the intensive care unit: results from a prospective four-year registration. *Critical Care*, 22(1):241, 2018.

[60] Muhammad Ali, Humaira Naureen, Muhammad Haseeb Tariq, Muhammad Junaid Farrukh, Abubakar Usman, Shahana Khattak, and Hina Ahsan. Rational use of antibiotics in an intensive care unit: a retrospective study of the impact on clinical outcomes and mortality rate. *Infection and Drug Resistance*, 12:493, 2019.

[61] U Waheed, P Williams, S Brett, G Baldock, and N Soni. White cell count and intensive care unit outcome. *Anaesthesia*, 58(2):180–182, 2003.