More Efficient Off-Policy Evaluation through Regularized Targeted Learning

Aurélien F. Bibaut [*1] **Ivana Malenica** [*1] **Nikos Vlassis** [2] **Mark J. van der Laan** [1]

Abstract

We study the problem of off-policy evaluation (OPE) in Reinforcement Learning (RL), where the aim is to estimate the performance of a new policy given historical data that may have been generated by a different policy, or policies. In particular, we introduce a novel doubly-robust estimator for the OPE problem in RL, based on the Targeted Maximum Likelihood Estimation principle from the statistical causal inference literature. We also introduce several variance reduction techniques that lead to impressive performance gains in off-policy evaluation. We show empirically that our estimator uniformly wins over existing off-policy evaluation methods across multiple RL environments and various levels of model misspecification. Finally, we further the existing theoretical analysis of estimators for the RL off-policy estimation problem by showing their $O_P(1/\sqrt{n})$ rate of convergence and characterizing their asymptotic distribution.

1. Introduction

The study of *off-policy evaluation* is an increasingly important problem in reinforcement learning. *Off-policy evaluation* (OPE) addresses the pressing issue of evaluating the performance of a novel policy in a setting where actual enforcement might be too costly, infeasible, or even hazardous. This situation arises in many fields, including medicine, finance, advertising, and education, to name a few (Murphy et al., 2001; Petersen et al., 2014; Theocharous et al., 2015; Hoiles & Van Der Schaar, 2016). The OPE problem can be treated as a counterfactual quantity estimation problem, as we inquire about the mean reward we would have accrued, had we, contrary to fact, implemented the policy π_e at the time of data-collection. Estimating and inferring such counterfactual quantities is a well studied problem in statistical causal inference, and has led to many methodological developments. One of the things we aim to do in this work is to further earlier efforts (Dudik et al., 2011) in bridging the gap between the reinforcement learning and causal inference fields.

There are roughly two predominant classes of approaches to off-policy value evaluation in RL (Jiang & Li, 2015). The first is the *direct method* (DM), analogous to the *G-computation* procedure in causal inference (Robins et al., 1999; 2000). The direct method first fits a model of the system's dynamics and then uses the learned fit in order to estimate the mean reward of the target policy (evaluation policy). The estimators produced by this approach usually exhibit low variance, but suffer from high bias when the model fit is misspecified or the number of samples is low as compared to the complexity of the function class of the model (Mannor et al., 2007). The second major avenue for off-policy value evaluation is *importance sampling* methods, also termed *inverse propensity score* methods in statistical causal inference (Rosenbaum & Rubin, 1983). Importance sampling (IS) attempts to correct the mismatch between the distributions produced by the behavior and target policies (Precup et al., 2000; Precup, 2000). IS estimators are unbiased under mild conditions, but their variance tends to be large when the evaluation and behavior policies differ significantly (Farajtabar et al., 2018), and grows exponentially with the horizon, rendering them (Farajtabar et al., 2018) impractical for many RL settings. A third class of estimators, *Doubly Robust* (DR) estimators, obtained by combining a DM estimator and an IS estimator, are becoming standard in OPE (Farajtabar et al., 2018; Jiang & Li, 2015; Thomas & Brunskill, 2016). These originate from the statistics literature (Robins et al., 1994; Robins & Rotnitzky, 1995; Bang & Robins, 2005; van der Laan & Rubin, 2006; van der Laan & Rose, 2011; 2018), and were introduced in the RL literature by (Dudik et al., 2011). Combining a DM and an IS estimator under the form of a DR estimator leads to lower bias than DM alone, and lower variance than IS alone.

Our contribution to OPE in RL is multifold. First we adapt a doubly robust estimator from statistical causal inference, the Longitudinal Targeted Maximum Likelihood Estimator (LTMLE) to the OPE in RL setting. We show that our adapted estimator converges at rate $O_P(1/\sqrt{n})$ to the true

[*] Equal contribution [1] University of California, Berkeley, CA [2] Netflix, Los Gatos, CA. Correspondence to: Aurélien F. Bibaut <aurelien.bibaut@berkeley.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

policy value. Deriving the LTMLE requires us to identify a mathematical object known in semiparametric statistics as the *efficient influence function* (EIF) of the estimand (policy value). To the best of our knowledge, this article is the first one to explicitly derive the EIF of the policy value for the OPE problem in RL. Knowledge of the EIF allows us to prove that both our estimator (the LTMLE) and recently proposed DR estimators (Jiang & Li, 2015; Thomas & Brunskill, 2016) are optimal in the sense that they achieve the generalized Cramer-Rao lower bound.

Second, we introduce an idea from statistics to make better use of the data than prior OPE works (Jiang & Li, 2015; Thomas & Brunskill, 2016). We noticed that most OPE papers, at least in theory, use sample splitting: the Q-function is fitted on a split of the data, while the DR estimator is obtained by evaluating the fitted Q-function on another split. We propose a cross-validation-based technique that allows to essentially average the Q-function over the entire sample, leading to a constant-factor gain in risk.

Finally, and most importantly for practice, we propose several regularization techniques for the LTMLE estimators, out of which some, but not all, apply to other DR estimators. Using the MAGIC ensemble method from (Thomas & Brunskill, 2016), we construct an estimator that combines various regularized LTMLEs. We call our estimator RLTMLE (TMLE for RL). Our experiments demonstrate that RLTMLE outperforms all considered competing off-policy methods, uniformly across multiple RL environments and levels of model misspecification.

2. Statistical Formulation of the Problem

2.1. Markov Decision Process

Consider a Markov Decision Process (MDP) defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P_1, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, and $\gamma \in (0, 1]$ is a discount factor. A trajectory H is a succession of states S_t, actions A_t and rewards R_t, observed from $t = 1$ to the horizon $t = T$: $H = (S_1, A_1, R_1, ..., S_T, A_T, R_T)$. For all $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}$, $P(s', r|s, a)$ is the probability of collecting reward r and transitioning to state s', conditional on starting in state s and taking action a, and $P_1(s)$ is the probability that the initial is s. A policy π is a sequence of conditional distributions $(\pi_1, \pi_2, ...)$ that stochastically map a state to an action: for all t, $A_t|S_t \sim \pi_t$.

Suppose we are given n i.i.d. T-step trajectories of the MDP, $D = (H_1, ..., H_n)$, collected under the behavior policy $\pi_b = (\pi_{b,1}, ..., \pi_{b,T})$. We assume all trajectories have the same initial state s_1, allowing for the data-generating mechanism to be fully characterized by (P, π_b).

2.2. Estimation Target

The goal of OPE is to estimate the average cumulative discounted reward we would have obtained by carrying out the target policy π_e instead of policy π_b. That is, we want to estimate the following counterfactual quantity:

$$V_1^{\pi_e}(s_1) := E_{P,\pi_e}\left[\sum_{t=1}^{T} \gamma^t R_t | S_1 = s_1\right]. \tag{1}$$

Consider the following common assumption from the causal inference literature.

Assumption 1 (Absolute continuity)**.** *For all* $s, a \in \mathcal{S} \times \mathcal{A}$*, if* $\pi_b(a|s) = 0$*, then* $\pi_e(a|s) = 0$ *too.*

Under assumption 1 and the Markov assumption of the MDP model, $V_1^{\pi_e}(s_1)$ can be written as an expectation under the data-generating mechanism (P, π_b):

$$V_1^{\pi_e}(s_1) = E_{P,\pi_b}\left[\prod_{t=1}^{T} \frac{\pi_{e,t}(A_t|S_t)}{\pi_{b,t}(A_t|S_t)} \sum_{t=1}^{T} \gamma^t R_t \middle| S_1 = s_1\right]. \tag{2}$$

For $t = 1, ..., T$, define $\bar{R}_{t:T} := \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$as the total reward from step t to step T. For all $1 \leq t_1 \leq t_2 \leq T$, define $\rho_{t_1:t_2} := \prod_{\tau=t_1}^{t_2} \pi_{e,\tau}(A_\tau|S_\tau)/\pi_{b,\tau}(A_\tau|S_\tau)$. For all $t = 1, ..., T$, we will use the shortcut notation $\rho_t := \rho_{1:t}$. We use the convention that $\rho_0 = 0$. Denote $\bar{R}_{t:T}^{(i)}$, $\rho_t^{(i)}$, $\rho_{t_1:t_2}^{(i)}$ the corresponding quantities for a sample trajectory H_i. Consistently with (1) and (2), we define, for any $t = 1, ..., T$, and $s \in \mathcal{S}$, the value function (or reward-to-go) from time point t and state s, as

$$\begin{aligned} V_t^{\pi_e}(s) :&= E_{P,\pi_e}[\bar{R}_{t:T}|S_t = s] \\ &= E_{P,\pi_b}\left[\rho_{t:T}\bar{R}_{t:T}|S_t = s\right]. \end{aligned}$$

For every $t = 1, ..., T$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, we further define the action-value function from time step t as

$$\begin{aligned} Q_t^{\pi_e}(s, a) &:= E_{P,\pi_e}\left[\bar{R}_{t:T}|S_t = s, A_t = a\right] \\ &= E_{P,\pi_b}\left[\rho_{t:T}\bar{R}_{t:T}|S_t = s, A_t = a\right]. \end{aligned}$$

3. An existing state-of-the art approach

Our method can be seen as building upon and improving on (Thomas & Brunskill, 2016). We believe it helps understanding our contribution to first briefly describe their estimators. For a detailed review of OPE methods, we refer the interested reader to the vast and excellent literature on the topic (Precup et al., 2000; Thomas, 2015; Jiang & Li, 2015; Farajtabar et al., 2018).

3.1. Weighted Doubly Robust Estimator

(Jiang & Li, 2015) were the first authors to propose a doubly robust estimator for off-policy evaluation in the MDP setting.

(Thomas & Brunskill, 2016) propose a stabilized version of the DR estimator of (Jiang & Li, 2015), termed Weighted Doubly Robust (WDR) estimator, which they obtain by replacing the importance sampling weights by stabilized importance sampling weights. The stabilized importance sampling weight for observation i at time step t is defined as $w_t^{(i)} = \rho_t^{(i)} / \sum_{i=1}^n \rho_t^{(i)}$. The WDR estimator is thus defined as

$$WDR := \sum_{i=1}^{n} \Bigg\{ \frac{1}{n} V_1^{\pi_e}(S_1^{(i)}) + \sum_{t=1}^{T} \gamma^t w_t^{(i)} \left[R_t^{(i)} - Q_t^{\pi_e}(S_t^{(i)}, A_t^{(i)}) + \gamma V_{t+1}^{\pi_e}(S_{t+1}^{(i)}) \right] \Bigg\}. \tag{3}$$

3.2. MAGIC

While WDR has low bias and converges at rate $O_P(1/\sqrt{n})$ to the truth, its reliance on importance weights can make it highly variable. As a result, in some settings, especially if model misspecification isn't too strong, DM estimators can beat WDR (Thomas & Brunskill, 2016). This motivates the construction of an estimator that interpolates between DM and WDR, so as to benefit from the best of both worlds. (Thomas & Brunskill, 2016) propose the *partial importance sampling* estimators, which correspond to essentially cutting off the sum in (3) the terms with index $t \geq j$ for some $0 \leq j \leq T$. Formally, they define their partial importance sampling estimator as the average $g_j := \sum_{i=1}^n g_j^{(i)}$ of the so-called *off-policy j-step return*, that they define, for each trajectory i, as

$$g_i^{(j)} := \sum_{t=1}^{j} \underbrace{\gamma^t w_t^i R_t^{(i)}}_{\text{a}} + \underbrace{\gamma^{j+1} w_j^i V_{j+1}^{\pi_e}(S_{j+1}^i)}_{\text{b}} - \sum_{t=1}^{j} \underbrace{\gamma^t [w_t^i Q_t^{\pi_e}(S_t^{(i)}, A_t^{(i)}) - w_{t-1}^i V_t^{\pi_e}(S_t^{(i)})]}_{\text{c}},$$

Note that g_0 is equal to the DM estimator. Note that the last component, (c), represents the combined control variate for the importance sampling (a) and model based term (b). Hence, as j increases, we expect bias to decrease, at the expense of an increase in variance.

(Thomas & Brunskill, 2016)'s final estimator is a convex combination of the partial importance sampling estimators g_j. Ideally, we would like this convex combination to minimize mean squared error (MSE), that is we would like to use as estimator $(\mathbf{x}^*)^\top \boldsymbol{g}$, with $\boldsymbol{g} = (g_0, ..., g_T)$, where

$$\begin{aligned} \boldsymbol{x}^* &= \arg \min_{0 \leq \boldsymbol{x} \leq 1 : \sum_{j=0}^T x_j = 1} \text{MSE}(\boldsymbol{x}^\top \boldsymbol{g}, V_1^{\pi_e}) \\ &= \arg \min_{0 \leq \boldsymbol{x} \leq 1 : \sum_{j=0}^T x_j = 1} \Big\{ \text{Bias}^2(\boldsymbol{x}^\top \boldsymbol{g}, V_1^{\pi_e}) \\ &\qquad + \text{Var}(\boldsymbol{x}^\top \boldsymbol{g}) \Big\}. \end{aligned} \tag{4}$$

As we don't have access to the true variance and bias, (Thomas & Brunskill, 2016) propose to use as estimator $\hat{\boldsymbol{x}}^\top \boldsymbol{g}$, where $\hat{\boldsymbol{x}}$ is a minimizer, over the convex weights simplex, of an estimate of the MSE. The covariance matrix of $\boldsymbol{g}$ can be estimated as the empirical covariance matrix of the $\boldsymbol{g}^{(i)}$'s. Bias estimation is a more involved. For each $j = 1, ..., T$, (Thomas & Brunskill, 2016) estimate the bias of the partial importance sampling estimator g_j by its distance to a δ-confidence interval for g_T obtained by bootstrapping it, for some $\delta \in (0, 1)$. They named the resulting ensemble estimator MAGIC, standing for *model and guided importance sampling combining*. For further details, we refer the reader to the very clear presentation of their algorithm by (Thomas & Brunskill, 2016).

4. Longitudinal TMLE for MDPs

4.1. High level description

Our proposed estimator extends the longitudinal Targeted Maximum Likelihood Estimation (TMLE) methodology, initially developed in the statistics causal inference literature, to the MDP setting (van der Laan & Rubin, 2006; van der Laan & Gruber, 2011; van der Laan & Rose, 2011; 2018). In order to build intuition on our estimator, we start with a high-level description. Targeted Maximum Likelihood Estimation is a general framework that allows to construct efficient nonparametric estimators of low-dimensional characteristics of the data-generating distribution, given machine learning based estimators of high-dimensional characteristics. Let us illustrate on an example what these low-dimensional and high-dimensional characteristics can be. Suppose we want to estimate an average treatment effect (ATE), and that we have pre-treatment covariates X, a treatment T and an outcome Y, with $(X, T, Y) \sim P$. In this situation, the low-dimensional characteristic is the ATE $E_P[E_P[Y|T = 1, X] - E_P[Y|T = 0, X]]$, while the high-dimensional characteristics of P are the outcome regression function $x, a \mapsto E_P[Y|A = a, X = x]$ and the propensity score function $x \mapsto E_P[T|X = x]$.

4.2. Simplified sample-splitting based algorithm

In the following sections we present a simplified version of the algorithm that constructs our Longitudinal Targeted Maximum Likelihood Estimator. The full-blown version of the algorithm is presented in the appendix, with the corresponding theoretical justifications.

Suppose we are provided with n i.i.d. trajectories, $D = (H_1, ..., H_n)$. Make two splits of the sample: for some $0 < p < 1$, let $D^{(0)} = (H_1, ..., H_{(1-p)n})$ and $D^{(1)} = (H_{(1-p)n+1}, ..., H_n)$. Use $D^{(0)}$ to fit estimators $\hat{Q}_1^{\pi_e}, \cdots, \hat{Q}_T^{\pi_e}$ of the action value functions $Q_1^{\pi_e}, \cdots, Q_T^{\pi_e}$ We will call $\hat{Q}_1^{\pi_e}, \cdots, \hat{Q}_T^{\pi_e}$ the *initial estimators*. Such estimators can be obtained for instance by fitting a model of the dynamics of the MDP, or by SARSA, among other methods (Sutton & Barto, 1998). Estimators fitted in such a way tend to exhibit low variance but often suffer from misspecification bias. As mentioned in section 3, doubly-robust estimators take such initial estimators as input, and evaluate and then average a certain function on them on $D^{(1)}$ to produce an unbiased estimator or $V_1^{\pi_e}(s_1)$. These doubly-robust estimators rely on the addition of terms weighted by the importance sampling (IS) ratios $\rho_{i:t}^{(i)}$, $i = 1, \cdots, n$, $t = 1, \cdots, n$. The TMLE methodology takes another route: for each t, it defines, on top of the initial estimator fit, a parametric model, which we will call a *second-stage parametric model* $\hat{Q}_t^{\pi_e}$, and achieves bias reduction by fitting this parametric model by maximum likelihood, on the sample split $D^{(1)}$.

4.3. Formal presentation of the simplified algorithm

To formally describe our algorithm, it suffices to define the second-stage parametric models and describe the loss used for the fit. For all $x \in \mathbb{R}$, we define $\sigma(x) = 1/(1 + e^{-x})$ as the logistic function, and we denote σ^{-1} its inverse. Observe that bounding the range of rewards where $\forall t, R_t \in [r_{min}, r_{max}]$, implies that $\forall t$ and $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_t(s, a) \in [-\Delta_t, \Delta_t]$ with $\Delta_t := \sum_{\tau=t}^{T} \gamma^{\tau-t} \max(r_{max}, |r_{min}|)$. We further denote $\tilde{Q}_t^{\pi_e}(s, a) := (\hat{Q}_t^{\pi_e} + \Delta_t)/(2\Delta_t)$ as the normalized initial estimator. In addition, $\forall \delta \in (0, 1/2)$ and $\forall (s, a)$, we define the following thresholded version of $\tilde{Q}_t^{\pi_e}$:

$$\tilde{Q}_t^{\pi_e,\delta}(s, a) := \begin{cases} 1 - \delta & \text{if } \tilde{Q}_t^{\pi_e}(s, a) > 1 - \delta, \\ \tilde{Q}_t^{\pi_e}(s, a) & \text{if } \tilde{Q}_t^{\pi_e}(s, a) \in [\delta, 1 - \delta], \\ \delta & \text{if } \tilde{Q}_t^{\pi_e}(s, a) < \delta. \end{cases}$$

For all $\epsilon \in \mathbb{R}$, we can now define the normalized version of our second-stage parametric model as:

$$\tilde{Q}_t^{\pi_e,\delta}(\epsilon)(s, a) := \sigma(\sigma^{-1}(\tilde{Q}_t^{\pi_e,\delta}(s, a)) + \epsilon).$$

Finally, we denote $\hat{Q}_t^{\pi_e,\delta}(\epsilon) = 2\Delta_t(\tilde{Q}_t^{\pi_e,\delta}(\epsilon) - 1/2)$ as the rescaled version of $\tilde{Q}_t^{\pi_e,\delta}(\epsilon)$.

The normalization, thresholding and rescaling steps in the definition of the parametric second-stage model ensure that (1) $\tilde{Q}_t^{\pi_e,\delta}(\epsilon) \in [\delta, 1 - \delta] \subset (0, 1)$ for all ϵ, and that (2) $\hat{Q}_t^{\pi_e,\delta}(\epsilon)$ always stays in the allowed range of rewards $[-\Delta_t, \Delta_t]$. The definition of $\tilde{Q}_t^{\pi_e,\delta}(\epsilon)$ as a logistic transform of ϵ that lies in $(0, 1)$ makes the fitting of ϵ possible through maximum likelihood for a logistic likelihood. For $t = T$, since $Q_T^{\pi_e}(s, a) = E_{P,\pi_b}[\rho_{1:T} R_T | S_T = s, A_T = a]$, it is natural to consider the log likelihood,

$$\mathcal{R}_{n,T}^{\delta}(\epsilon) = \frac{1}{n} \sum_{i=1}^{n} \rho_{1:T}^{(i)} \Big(\tilde{U}_T^{(i)} \log(\tilde{Q}_T^{\pi_e,\delta}(\epsilon)(S_T^{(i)}, A_T^{(i)})) + (1 - \tilde{U}_T^{(i)}) \log(1 - \tilde{Q}_T^{\pi_e,\delta}(\epsilon)(S_T^{(i)}, A_T^{(i)})) \Big), \quad (5)$$

where $\tilde{U}_T^{(i)} := (R_T^{(i)} + \Delta_T)/(2\Delta_T)$ is the normalized reward at time T. Normalization of the reward is necessary since we are using logistic regression to optimize ϵ, and to keep the definition of $\tilde{U}_T^{(i)}$ and $\tilde{Q}_T^{\pi_e,\delta}(s, a)$ consistent. The thresholding step that defines $\tilde{Q}_t^{\text{e}lta}(s, a)$ prevents the log likelihood from taking on non-finite values. In order to make the bias introduced by thresholding vanish as the sample size grows, we use a vanishing sequence $\delta_n \downarrow 0$ of thresholding values.

Let $\epsilon_{n,T}$ be the minimizer over $\mathbb{R}$ of the log likelihood $\mathcal{R}_{n,t}^{\delta}$ for step t. We fit the second-stage models for $t = T-1, ..., 1$ by backward recursion, a procedure which we describe in details in this paragraph. Start with observing that for all $t = 1, ..., T$, and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_t^{\pi_e}(s, a) = E_{\pi_b}[\rho_{1:t}(R_t + \gamma V_{t+1}^{\pi_e}(S_{t+1}))|S_t = s, A_t = a]$. This motivates defining, as outcome of the rescaled logistic regression model for time step t, the normalized reward-to-go:

$$\tilde{U}_{t,n}^{(i)} := (R_t^{(i)} + \gamma \hat{V}_{t+1}^{\pi_e}(\epsilon_{n,t+1})(S_{t+1}^{(i)}) + \Delta_t)/(2\Delta_t).$$

Define $\hat{V}_t^{\pi_e}(\epsilon)$ as the value function corresponding to the action-value function $\hat{Q}_t^{\pi_e,\delta_n}(\epsilon)$, that is, for all $s \in \mathcal{S}$, set $\hat{V}_t^{\pi_e}(\epsilon)(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s) \hat{Q}^{\pi_e,\delta_n}(\epsilon)(s, a')$. We define the second-stage model log likelihood for each $t = T - 1, ..., 1$ as

$$\mathcal{R}_{t,n}^{\delta_n} = \frac{1}{n} \sum_{i=1}^{n} \rho_{1:t}^{(i)} \Big(\tilde{U}_t^{(i)} \log(\tilde{Q}_t^{\pi_e,\delta}(\epsilon)(S_t^{(i)}, A_t^{(i)})) + (1 - \tilde{U}_t^{(i)}) \log(1 - \tilde{Q}_t^{\pi_e,\delta}(\epsilon)(S_t^{(i)}, A_t^{(i)})) \Big). \quad (6)$$

The fact that the outcome in the second-stage logistic model at time step t depends on the second-stage model fit at time step $t + 1$ is why we have to proceed backwards in time. This is why we say this procedure is a *backward recursion*.

Finally, once all of the T second-stage models have been fitted, we define the LTMLE estimator of $V_1^{\pi_e}(s_1)$ as follows:

$$\hat{V}_1^{\pi_e,LTMLE}(s_1) := \hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1).$$

Algorithm 1 Longitudinal TMLE for MDPs

```
Input: Logged data split $D^{(1)}$, target policy $\pi_e$, initial estimators $\hat{Q}_1^{\pi_e}, ..., \hat{Q}_T^{\pi_e}$, discount factor $\gamma$.
Set $\Delta_T = 0$ and $\hat{V}_{T+1}^{\pi_e} = \mathbf{0}$.
for $t = T$ to $1$ do
    Set $\Delta_t = \max_{t,i} |R_t| + \gamma \Delta_t$.
    Set $\tilde{U}_t = (R_t + \gamma \hat{V}_{t+1}^{\pi_e} + \Delta_t)/2\Delta_t$.
    Set $\tilde{Q}_t^{\pi_e, \delta_n} =$ threshold$(\delta_n, (\hat{Q}_t^{\pi_e} + \Delta_t)/2\Delta_t)$.
    Compute $\epsilon_{n,t} = \arg\min_\epsilon \mathcal{R}_{n,t}^{\delta_n}(\epsilon)$.
    Set $\hat{Q}_t^{\pi_e, \delta_n} = 2\Delta_t(\tilde{Q}_t^{\pi_e, \delta_n} - 0.5)$.
    Set, for all $s \in \mathcal{S}$,
        $$\hat{V}_t^{\pi_e}(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s) \hat{Q}_t^{\pi_e, \delta_n}(s, a').$$
end for
return $\hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1)$.
```

This idea of backward recursion we just exposed was initially introduced in (Bang & Robins, 2005). They called it *sequential regression.*

We present the pseudo-code of the procedure as Algorithm 1.

4.4. Guarantees and benefits

It might at first appear surprising that fitting the second-stage models, which amounts to simply fitting the intercept of a logistic regression model, suffices to fully remove the bias. We nevertheless prove that it does so in theorem 1 under mild assumptions. Theorem 1 requires assumption 1 stated in section 2 and assumptions 2-4 stated below.

Assumption 2. *For all $t = 1, ..., T$, $r_t \in [r_{min}, r_{max}]$ almost surely.*

Assumption 3. *For all $t = 1, ..., T$, the initial estimator $\hat{Q}_{t,n}^{\pi_e}$ converges in probability to some limit $Q_{t,\infty} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, that is $\|\hat{Q}_{t,n}^{\pi_e} - Q_{t,\infty}\|_{P,2} = o_P(1)$.*

Assumption 4. *For all $t = 1, ..., T$, let $Q_{t,\infty}$ be the limit as defined in Assumption 3. Assume there exists a (small) positive constant $\eta \in (0, 1/2)$ such that $\forall t$ and $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_{t,\infty}(s, a) \in [\eta, 1 - \eta]$.*

Assumption 5. *Suppose there exists a finite positive constant M such that $\forall t$, $\rho_{1:t} \leq M$ almost surely.*

We can now state our main theoretical result, for the algorithm presented in section 4.3.

Theorem 1. *Suppose assumptions 2, 3, 4, and 5 hold. Then the LTMLE estimator has bias $o(1/\sqrt{n})$, that is*

$$E_{P,\pi_b}[\hat{V}_1^{\pi_e, LTMLE}(s_1)] - V_1^{\pi_e}(s_1) = o(1/\sqrt{n}).$$

In addition, the LTMLE estimator converges in probability at rate $\sqrt{n}$, such that:

$$\hat{V}_1^{\pi_e, LTMLE}(s_1) - V_1^{\pi_e}(s_1) = O_P(1/\sqrt{n}).$$

With a little extra work, we can also characterize the asymptotic distribution and the asymptotic variance of the LTMLE estimator. In particular, we show in the appendix that, provided that $\hat{Q}^{\pi_e}$ is consistent, our estimator attains the generalized Cramer-Rao bound and is therefore *locally efficient*. We also argue that it is asymptotically equivalent with the doubly robust estimator (Thomas & Brunskill, 2016; Jiang & Li, 2015).

5. RLTMLE

In this section, we introduce RLTMLE, a convex combination of regularized LTMLE estimators. The weighted average of base LTMLEs is obtained by minimizing an approximation of the MSE, with bias and variance decomposition as proposed for the MAGIC estimator described in Section 3. The RLTMLE performance stems from two sources: (1) it benefits from the stability of the LTMLE, (2) its has a rich pool of regularized base estimators. In addition, it accounts for several finite sample advantages over existing methods, including respecting the reward domain by design and avoiding directly summing over ρ_t. However, even more important than the stability of $\hat{V}_1^{\pi_e, RLTMLE}(s_1)$ itself, what contributes most to our finite sample gains is that our core estimator is amenable to various types of regularization. We emphasize that some regularizations employed by RLTMLE do not have a clear analogue in the DR and WDR realm, making LTMLE particularly serviceable.

5.1. Regularization and base estimators

We introduce three regularization techniques that allow to stabilize the variance of the LTMLE estimator. The first two have a clear WDR analogue, while the third one only applies to LTMLE.

1. **Weights softening.** For $\alpha \in [0, 1]$, $x \in \mathbb{R}^d$, define $\texttt{soften}(x) := (x_k^\alpha / \sum_{l=1}^d x_l^\alpha : k = 1, ..., d)$. The LTMLE algorithm corresponding to softening level α is obtained by replacing, in the second-stage log likelihoods (5) and (6), the IS ratios $(\rho_{1:t}^{(i)} : i = 1, ..., n)$ by $\texttt{soften}(\rho_{1:t}^{(i)} : i = 1, ..., n)$. The same operation can be applied as well to the importance weights of the WDR estimator.

2. **Partial horizon.** The LTMLE with partial horizon $j < T$ is obtained by setting to zero the coefficients $\epsilon_{n,j_1}, ..., \epsilon_{n,T}$ before fitting the other second-stage coefficients. This enforces that the importance sampling ratios $\rho_{1:t}$ for $t \geq j$ have no impact on the estimator. The WDR equivalent is to use the j-step return $g^{(j)}$.

3. **Penalization.** The penalized LTMLE is obtained by adding a penalty $\lambda|\epsilon_{n,t}|$ for some $\lambda \geq 0$ to the the log-likelihoods (5) and (6) of the second-stage models.

Our final estimator combines weight softened, penalized and partial LTMLE, creating a rich family of different regularized LTMLEs with an aim of minimizing MSE. In particular, we define a family of base estimators $\hat{V}_1^{\pi_e,j}$, $j = 1, ..., J$, where each $\hat{V}_1^{\pi_e,j}$ consists of a j^{th} combination of the above regularizations, defined by $(\alpha_j, \lambda_j, \tau_j)$. We consider a finite set of returns, such that $j \in \mathcal{J}$ where $\mathcal{J} < \infty$ and $J = |\mathcal{J}|$. Additionally, we define $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_J)$, $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_J)$ and $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_J)$.

5.2. Ensemble estimator

Our final estimator creates a convex combination of regularized LTMLE estimators that minimizes the MSE. First, we define a grid of possible combinations of regularization techniques to be incorporated in the LTMLE procedure described in Section 4.3, corresponding to the j^{th} return. Note however, that just arbitrarily combing different α, λ and τ is not optimal for all cases. For example, if the initial estimator is highly biased, penalized LTMLE might be sub-optimal in terms of MSE. On the other hand if $\hat{Q}_t^{\pi_e}$ does a good job, penalized LTMLE will lead to decreased variance. Similarly, if the second-stage model is not doing a good job estimating $\epsilon_{n,t}$, penalized TMLE will prevent highly variable perturbations of the previous (or initial) fit. With that in mind, we define our library of regularized LTMLEs such that we can obtain the benefits of regularization for each scenario. When $\lambda_j = 0$ and $\tau_j = T$, $\alpha_j > 0$. Similarly, when $\lambda_j > 0$ and $\tau_j < T$, $\alpha_j = 0$. With this distribution of $(\boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\tau})$, we are able to always reduce variance no matter the level of model misspecification.

We obtain estimates of $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$ for each $(\alpha_j, \lambda_j, \tau_j)_{j=1}^J$, with $\hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1) = \{\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)\}_{j=1}^J$. For simplicity sake, we omit the dependence of ϵ on j, even though each j will imply a potentially new set of ϵ. Analogous to Section 3, we aim to create a weighted average of j-length returns as in (4):

$$\boldsymbol{x}^* := \arg \min_{0 \leq x \leq 1 : \sum_i x_i = 1} \text{MSE}(\boldsymbol{x}^T \hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1), V_1^{\pi_e}). \tag{7}$$

For each $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$, we have that the stabilized efficient influence function is defined as:

$$\text{EIF}^j := \sum_{t=1}^{\tau_j} \gamma^t \rho_t (R^t + \gamma \hat{V}_{t+1}^{\pi_e,j}(\epsilon_{n,t+1})(S_{t+1}) - \\ - \hat{Q}_t^{\pi_e,j}(\epsilon_{n,t})(S_t, A_t)) \tag{8}$$

where dependence on j in $\hat{V}^{\pi_e}$ and $\hat{Q}^{\pi_e}$ comes from ϵ. Intuitively, the efficient influence function gives the weighted (stabilized) differences between the observed and estimated average reward, and it is at the core of our estimator analysis presented in the appendix. From the centered, stabilized EIF^j we can derive $\hat{\Omega}_n$ as the $J \times J$ sample covariance matrix, analogous to Section 3.

Algorithm 2 RLTMLE

```
**Input:** $D$, $\pi_e$, $\hat{Q}^{\pi_e}$, $\hat{V}^{\pi_e}$, $\gamma$, $n_{bootstrap}$
Set $J \in \mathbb{R}$.
Initialize $(\alpha_j, \lambda_j, \tau_j)$ for $j \in (1, \cdots J)$.
**for** $j = 1$ **to** $J$ **do**
  Run Algorithm (1), with $(\alpha_j, \lambda_j, \tau_j)$.
  Return $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$ and $\epsilon_{n,t}$, $\forall t$.
**end for**
$\forall j, j \in (1, \cdots, J)$, compute $\text{EIF}^j$ as in (8).
Compute $\hat{\Omega}_n$ as a sample covariance of $(\text{EIF}^j, \text{EIF}^k)$.
Compute $\hat{\boldsymbol{b}}_n$ using the percentile bootstrap and dist(a,b).
Compute $\boldsymbol{x}^*$ according to (7).
**return** $(\boldsymbol{x}^*)^T \hat{V}_1^{\pi_e}(\epsilon_{n,1})(s_1)$.
```

Recall that $\boldsymbol{b}_n = (b_n(1), \cdots, b_n(J))$, with $b_n(j)$ the bias for $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$. In order to estimate bias, we apply the percentile bootstrap as in (Thomas & Brunskill, 2016), with the estimate $b_n(j)$ defined as $\text{dist}(\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1), \text{CI}(WDR, 0.1))$. As distance metric, we also define $\text{dist}(a, b) := \min_{b \in \mathcal{B}} |a - b|$ as for MAGIC. Finally, we present a high-level pseudo code of RLTMLE as Algorithm 2.

6. Experiments

In this section, we demonstrate the effectiveness of RLTMLE by comparing it with other state-of-the-art methods used for OPE problem in various RL benchmark environments. We used three main domains, with detailed description of each allocated to the Appendix. We implement the same behavior and evaluation policies as in previous work (Thomas & Brunskill, 2016; Farajtabar et al., 2018).

1. **ModelFail**: a partially observable, deterministic domain with $T = 3$. Here the approximate model is incorrect, even asymptotically, due to three of the four states appearing identical to the agent.

2. **ModelWin**: a stochastic MDP with $T = 10$, where the approximate model can perfectly represent the MDP.

3. **GridWorld**: a 4×4 grid used for evaluating OPE methods, with an episode ending at $T = 100$ or when a final state ($s16$) is reached.

We omit benefits of RLTMLE over IS, PDIS (per-decision IS), WIS (weighted IS), CWPDIS (consistent weighted per-decision IS) and DR (doubly robust) estimators due to the

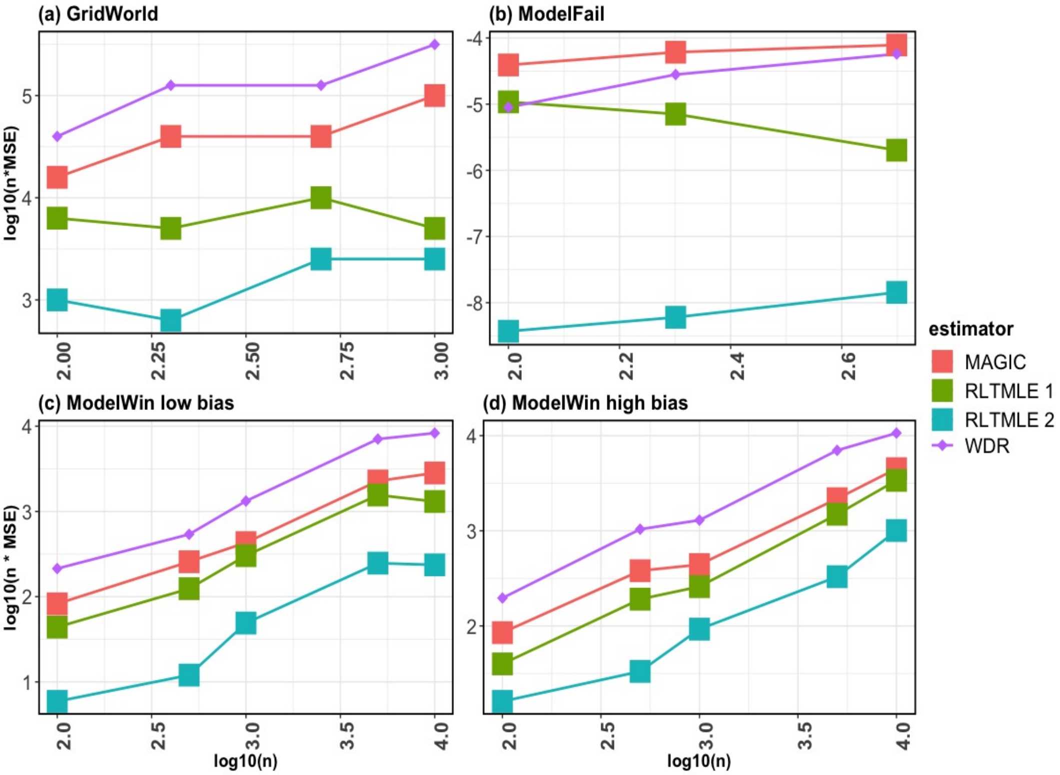

Figure 1. Empirical results for three different environments and varying level of model misspecification. **(a)** GridWorld MSE across varying sample size $n = (100, 200, 500, 1000)$ and bias equivalent to $b_0 = 0.005 * \text{Normal}(0, 1)$ over 71 trials; **(b)** ModelFail MSE across varying sample size $n = (100, 200, 500, 1000)$ and bias equivalent to $b_0 = 0.005 * \text{Normal}(0, 1)$ over 71 trials; **(c)** ModelWin MSE across varying sample size $n = (100, 500, 1000, 5000, 10000)$ and bias equivalent to $b_0 = 0.005 * \text{Normal}(0, 1)$ over 63 trials; **(d)** ModelWin MSE across varying sample size $n = (100, 500, 1000, 5000, 10000)$ and bias equivalent to $b_0 = 0.05 * \text{Normal}(0, 1)$ over 63 trials.

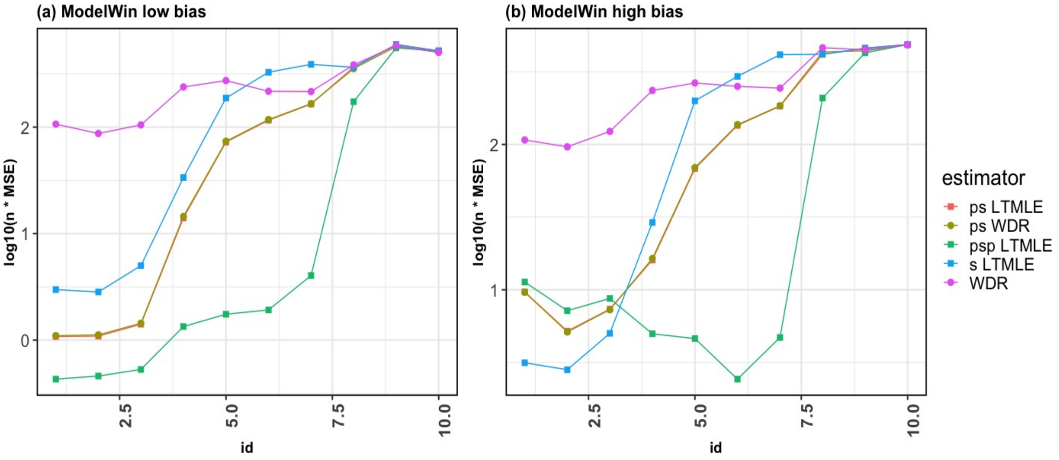

Figure 2. Comparison of WDR and LTMLE base estimators across various regularization methods in ModelWin at low ($b_0 = 0.005 * \text{Normal}(0, 1)$) and high ($b_0 = 0.05 * \text{Normal}(0, 1)$) model misspecification. Regularized base estimators include ps LTMLE (partial, softened LTMLE), ps WDR (partial, softened WDR), psp LTMLE (partial, softened, penalized LTMLE), s LTMLE (softened LTMLE) and WDR (no regularization). The x-axis indicates the id of the j^{th} estimator, corresponding to $(\alpha_j, \lambda_j, \tau_j)$. **(a)** ModelWin MSE for sample size $n = 1000$ and low bias over 315 trials; **(b)** ModelWin MSE for sample size $n = 1000$ and high bias over 315 trials.

extensive empirical studies performed by Thomas and Brunskill (Thomas & Brunskill, 2016). Instead, we compare our estimator to WDR and MAGIC, as they demonstrate improved performance over all simulations in benchmark RL environments considered (Thomas & Brunskill, 2016).

In evaluating our estimator, we also explore how various degree of model misspecification and sample size can affect the performance of considered methods. We start with small amount of bias, $b_0 = 0.005 * \text{Normal}(0, 1)$, where most estimators should do well. Consequently, we increase model misspecification to $b_0 = 0.05 * \text{Normal}(0, 1)$ at the same sample size, and consider the performance of all estimators. In addition, we test sensitivity to the number of episodes in D with $n = \{100, 200, 500, 1000)$ for GridWorld and ModelFail, and $n = \{100, 500, 1000, 5000, 10000)$ for ModelWin.

In addition, we consider the benefits of adding few regularization techniques as opposed to all three described in subsection 5.1. In particular, we concentrate on RLTMLE with only weight softening and partial LTMLE (RLTMLE 1) as opposed to using penalized LTMLE as well (RLTMLE 2). The goal of these experiments was to demonstrate the improved performance of our estimator when fully exploiting all the variance reduction techniques in a clever way. The MSE across varying sample size and model misspecification for GridWorld, ModelFail and ModelWin can be found in Figure 6. We can see that RLTMLE 2 outperforms all other estimators for all RL environments and varying levels of model misspecification.

Finally, we compare WDR and LTMLE base estimators augmented with various regularization methods before the ensemble step in Figure 6. In particular, for ModelWin, we look at the MSE of $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$ and $g^{(j)}$ for each j, where the j^{th} estimator corresponds to regularization $(\alpha_j, \lambda_j, \tau_j)$. Regularized base estimators considered include ps LTMLE (partial, softened LTMLE), ps WDR (partial, softened WDR), psp LTMLE (partial, softened, penalized LTMLE), s LTMLE (softened LTMLE) and WDR (no regularization). We note the vast improvement of WDR just by adding weight softening across all base estimators, evident for both low and high model misspecification setting. For the low bias environment of ModelWin, psp LTMLE (RLTMLE 2) uniformly outperforms all competitors for all j. High bias setting loses to s LTMLE for low j, but still outperforms majority of the time, including having the best ensemble MSE. While uniform win over all j is not necessary, we note that this behavior stems from the fact that for $j < 3$, $(\alpha_j, \lambda_j, \tau_j)$ used had very small τ_j and α_j. As such, with no strong debiasing effect of LTMLE, minimizing variance becomes more effective with respect to minimizing MSE.

7. Conclusion

In this paper, we proposed a new doubly robust estimator for off-policy value evaluation in reinforcement learning. In particular, we present a convex combination of regularized LTMLE estimators which aim at minimizing the MSE. We showed that our estimator is consistent and asymptotically optimal, achieving the Cramer-Rao lower bound. We prove the $O_P(1/\sqrt{n})$ rate of convergence of our estimator, and characterize its asymptotic distribution. The LTMLE is guaranteed to lie in the allowed rewards domain, both for discrete and continuous state, and is amenable to several regularization techniques. Finally, our experiments demonstrate uniform win of RLTMLE over all considered off-policy methods across multiple RL environments and various levels of model misspecification.

The RLTMLE enjoys multiple distinguishing features that contribute to its finite sample performance. First, its base estimator is a substitution estimator, therefore it inherently respects the reward domain for the RL problem. While this is true for DR if states and actions are discrete, our estimator by design produces estimates that lie in the allowed reward domain for both discrete and continuous state space. Our estimator also allows for clever usage of importance weights, instead of explicitly summing over IS terms. This property strives from using LTMLE as a base estimator, where stabilized IS ratios can be used as weights of the observations in the log likelihood of the second-stage models. This is an important feature of RLTMLE, that greatly contributes to its stability without introducing bias. Finally, LTMLE is amenable to many regularization methods, with RLTMLE enjoying a rich family of regularized base estimators. Our experiments show impressive performance gains from utilizing variance reduction techniques for both RLTMLE and WDR.

Finally, our method does not refit the entire reward-to-go model for each new target policy as the More Robust Doubly Robust estimator, demonstrating some practical advantages. Since refitting the reward-to-go model can be quite computationally expensive, our estimator might be beneficial in situations where one wants to scan through many candidate target policies.

References

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, Dec 2005.

Dudik, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1097–1104, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104620.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. *CoRR*, abs/1802.03493, 2018. URL http://arxiv.org/abs/1802.03493.

Hoiles, W. and Van Der Schaar, M. Bounded off-policy evaluation with missing data for course recommendation and curriculum design. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1596–1604. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045559.

Jiang, N. and Li, L. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. *arXiv e-prints*, art. arXiv:1511.03722, November 2015.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007. doi: 10.1287/mnsc.1060.0614. URL https://doi.org/10.1287/mnsc.1060.0614.

Murphy, S. A., van der Laan, M. J., and Robins, J. M. Marginal Mean Models for Dynamic Regimes. *J Am Stat Assoc*, 96(456):1410–1423, Dec 2001.

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., and M, v. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, 2(2):147–185, 2014. PMCID: PMC4405134.

Precup, D. *Temporal abstraction in reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 2000. https://scholarworks.umass.edu/dissertations/AAI9978540.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1-55860-707-2. URL http://dl.acm.org/citation.cfm?id=645529.658134.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291135.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290910.

Robins, J. M., Greenland, S., and Hu, F.-C. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association*, 94(447):687–700, 1999. doi: 10.1080/01621459.1999.10474168.

Robins, J. M., Hernan, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, Sep 2000.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41. URL http://dx.doi.org/10.1093/biomet/70.1.41.

Sutton, R. S. and Barto, A. G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 1806–1812. AAAI Press, 2015. ISBN 978-1-57735-738-4. URL http://dl.acm.org/citation.cfm?id=2832415.2832500.

Thomas, P. *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst, 2015.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2139–2148, New York, New York, USA, Jun 2016. PMLR.

van der Laan, M. J. and Gruber, S. Targeted minimum loss based estimation of an intervention specific mean outcome. Technical report, U.C. Berkeley Division of Biostatistics Working Paper Series, https://biostats.bepress.com/ucbbiostat/paper290/, 2011.

van der Laan, M. J. and Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data (Springer Series in Statistics)*. Springer, 2011.

van der Laan, M. J. and Rose, S. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media, 2018.

van der Laan, M. J. and Rubin, D. Targeted maximum likelihood learning. Technical Report Working Paper 213, U.C. Berkeley Division of Biostatistics Working Paper Series, 10 2006.