Inferring the Long-Term Causal Effects of Long-Term Treatments from Short-Term Experiments

Allen Tran [1] **Aurélien Bibaut** [2] **Nathan Kallus** [2 3]

Abstract

We study inference on the long-term causal effect of a continual exposure to a novel intervention, which we term a long-term treatment, based on an experiment involving only short-term observations. Key examples include the long-term health effects of regularly-taken medicine or of environmental hazards and the long-term effects on users of changes to an online platform. This stands in contrast to short-term treatments or "shocks," whose long-term effect can reasonably be mediated by short-term observations, enabling the use of surrogate methods. Long-term treatments by definition have direct effects on long-term outcomes via continual exposure, so surrogacy conditions cannot reasonably hold. We connect the problem with offline reinforcement learning, leveraging doubly-robust estimators to estimate long-term causal effects for long-term treatments and construct confidence intervals.

1. Introduction

Long-term effects of interventions are often of primary importance yet their direct measurement is hampered by the difficulty of performing long-term randomized control trials. For example, both medical and policy trials are often interested in long-term health or welfare impact, but following subjects for prolonged periods is difficult. Similarly, businesses in digital settings, constrained by operational considerations and motivated by fast-paced innovation, often use short-run A/B tests to inform decisions that ultimately aim to improve long-term outcomes.

Surrogate methods offer a route to connect short term tests to their longer term outcomes (Athey et al., 2019; Prentice, 1989). These methods rely on the existence of intermediate-term surrogate variables and/or an observational dataset that associates surrogate variables with their eventual long-term outcomes. The key requirements are that the surrogate(s) fully mediate the effect of the treatment on the outcome of interest and that we can identify the effect of the surrogates.

However, the treatment of interest may be explicitly *long-term*, that is, involving a continuous exposure to a novel intervention that extends beyond the length of the experiment. For example, persistent environmental hazards, regular medication, or a change to the user experience in a digital setting. This stands in contrast to short-term treatments, such as a training course or a pharmacological regimen confined in time, whose consequences could reasonably be captured within a short time frame. For long-term treatments, unless the experiment itself (or the measurement of the surrogates) is long-term, surrogate methods are incapable of reliably capturing their effect.

In this paper, we develop a method that is capable of estimating the long-term effects of long-term treatments from short-term experiments, provided the short-term observations sufficiently characterize the long-term trajectory, even if they do not mediate the effect on it. The method learns long-term temporal dynamics directly from the short-run experimental dataset, which eliminates the need both for the surrogate assumption and for an observational dataset linking surrogates to long-term outcomes. Provided these dynamics persist, this enables the estimation of long-term effects of arbitrary-length treatments, both short and long. In contrast, we show that surrogate methods, even when their assumptions hold, implicitly estimate a truncated effect in our setting, that of a treatment that persists up to the point that surrogates are measured.

In place of the two the key assumptions of surrogate methods (perfect mediation and identification of mediated effect), we make two novel assumptions which connect the problem of estimating long-term effects from experiments with offline reinforcement learning (ORL), which broadly considers the problem of evaluating "policies" on their expected cumulative reward, with evaluation policies potentially differing from the policy generating the data. We make use of the connection with ORL by leveraging recent literature that develops efficient doubly-robust estimators for off-policy

[1]Netflix Inc., Los Angeles, USA [2]Netflix Inc., Los Gatos, USA [3]Cornell University, New York, USA. Correspondence to: Allen Tran <allent@netflix.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

evaluation. In particular, we show how long-term causal effects can be estimated from the outcomes of two types of policies: a null treatment policy and a set of policies indexed by T, where T denotes the duration of treatment.

The paper is organized as follows. The next section sets up the methodology and provides conditions for identification. Section 3 introduces the estimator and conditions for root-N consistency and asymptotic normality. Section 4 uses simulated data to evaluate our method against a range of alternatives, as well as exploring robustness to real world complications. We conclude in Section 5.

1.1. Related Literature

There exist a long history in Biostatistics of using the response of short-term proxy variables to interventions to infer longer-term effects on a primary outcome of interest. These short-term proxies are referred to as surrogate endpoints and their validity relies on various surrogacy assumptions that share the requirement that the surrogate mediates the treatment effect (Prentice, 1989; VanderWeele, 2013).

However, surrogate assumptions are unlikely to hold for a single surrogate and can potentially lead to sign-reversing bias (Chen et al., 2007). The surrogate index literature extends the surrogate method to allow for multiple surrogate variables and the use of observational datasets to infer the relationship between short term surrogates and longer term outcomes (Athey et al., 2019). Further extensions to this line of work include learning optimal policies (Yang et al., 2023), and combining long and short-term data to tackle confounding (Imbens et al., 2023; Athey et al., 2020) and improve efficiency (Kallus & Mao, 2020).

The most related extension similarly focuses on inferring the effects of long-term treatments with short-term experimental measurements (Huang et al., 2023). The approach uses a discrete-time sequential environment with an underlying surrogate-space. To overcome the curse of dimensionality, (Huang et al., 2023) assume linearity in both surrogate transitions and surrogate-reward mappings.

The dynamic treatment effects literature similarly seeks to estimate effects for a sequence of treatments (Murphy, 2003; Lewis & Syrgkanis, 2021; Chernozhukov et al., 2023). However, the key difference in our setting is that we aim to estimate treatment effects that extrapolate beyond the horizon of the observed short experiment, whereas dynamic treatment effects methods estimate effects for a horizon that matches the observed data. A related literature exists which aims to undo confounding in a dynamic setting (Battocchi et al., 2021; Bica et al., 2020). Here, confounding is not a concern since our data come from an experiment where treatment is at worst randomly assigned conditional on the initial state (see Assumption 2).

We lean heavily on the reinforcement learning literature, which estimates long-term outcomes from the perspective of quantifying the value of different "policies" (Sutton & Barto, 1998). We make direct use of an estimator from (Kallus & Uehara, 2022) that combines two functions: the Q function, which has a long history in reinforcement learning and the density ratio function (Liu et al., 2018; Uehara et al., 2020).

2. Methodology

Let Y denote the long-term outcome of interest and define a treatment policy, π^T, as a sequence of treatments for T periods and null treatment thereafter.[1] For example, the control policy is π^0 and a permanent treatment policy is π^∞. The potential long-term outcome associated with a particular treatment policy, π, is denoted as $Y(\pi)$.

Our estimand of interest is the average treatment effect of a particular treatment policy: the expected difference in potential long-term outcomes between a T-duration treatment policy and the control policy.

$$\varphi^T = \mathbb{E}\left[Y(\pi^T) - Y(\pi^0)\right] \tag{2.1}$$

We assume that we can decompose long-term outcomes into the discounted sum of per period outcomes, normalized so that Y can be interpreted as the weighted average per period potential outcome, weighted towards the present. Let γ denote the discount rate, Y_t the per period outcome and $Y_t(a)$ with $a \in \mathcal{A} = \{0, 1\}$ the per period potential outcome.

Assumption 1 (Additive rewards)**.**

$$Y(\pi^T) \equiv (1-\gamma)\sum_{t=0}^{\infty} \gamma^t Y_t(\mathbb{1}_{t<T}) \tag{2.2}$$

The experiment that generates our data is described in Figure 1. There exists an initial distribution of "states", p_b, from which initial states, S_0, are drawn and in turn which treatment, A_0 is assigned. We observe an outcome for the first period, Y_0, which depends on both the initial state and treatment assignment. Finally, we observe a transition to a subsequent state, S_1, which similarly depends both on the initial state and treatment assignment.

We want to evaluate the ATE with a different treatment policy and potentially a different distribution of initial states, p_e, than the experimental distribution. Figure 2 depicts the treatment policy and outcomes that we are interested in estimating. Figure 1 depicts the short experiment we observe, where treatment is assigned potentially depending

[1] The more general case of non-contiguous treatment policies easily fits within our framework, with the addition of more complex notation and a less elegant mapping to stationary state-independent treatment policies (see Section 2.1.1).

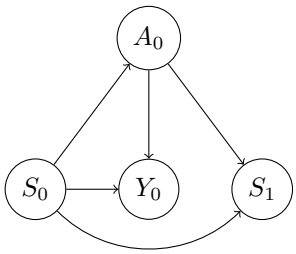

Figure 1. DAG of Observed Experiment

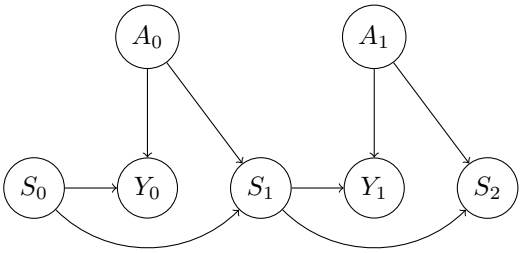

Figure 2. DAG of Treatment Policy of Interest

on the initial state. In contrast, Figure 2 depicts the target policy we want to evaluate where treatment is assigned according to our target policy.

The following two assumptions are assumptions on the experimental design. They are standard assumptions in the causal inference literature and allow us to "fill in" the missing counterfactual outcomes with observed outcomes.

Assumption 2 (Unconfoundedness)**.**

$$(Y_t(0), Y_t(1)) \perp\!\!\!\perp A_0 \mid S_0$$

Given a set of initial states S_0, we assume treatment assignment in the experiment is independent of potential outcomes conditional on the initial state, which should be satisfied with experimental data.

Assumption 3 (Overlap)**.**

$$\forall s \in \mathcal{S}, a \in \mathcal{A}\colon\ 0 < p_b(s,a) < 1$$

Note that our overlap condition is stronger than in traditional causal inference settings since it technically applies to the entire state-space and not just the initial states. For instance, it requires that treatment is rolled out to all types of users as opposed to being rolled out only to new users.

The next two assumptions depart from existing methods and allow us to extrapolate beyond the short term, using only data from the experiment. The states are assumed to satisfy the Markov property. The Markov assumption is implicitly a requirement that the state-space is sufficiently rich.

Assumption 4 (Markov property in states and actions)**.**

$$\forall s \in \mathcal{S}^t, a \in \mathcal{A}^{t-1}:$$
$$p\left(s_t | s_{t-1}, a_{t-1}, \ldots s_0, a_0\right) = p\left(s_t | s_{t-1}, a_{t-1}\right)$$

Define $p(y, s|\pi^T; t)$ as the marginal distribution of the states s and outcomes y "induced" by projecting the transition probabilities t periods from the initial distribution of states, $p_0(s_0)$, under the policy π^T.

$$p(y, s|\pi^T; t) = \int_{s_0,\ldots,s_{t-1}} p_0(s_0) \prod_{k=1}^{t} p(s_k|s_{k-1}, \mathbb{1}_{k-1<T}) \times p(y|s, \mathbb{1}_{t<T}) \tag{2.3}$$

Assumption 5 (Stationarity)**.**

$$\forall t, y, s, \pi^T:$$
$$p_t(y, s|\pi^T) - p_t(y, s|\pi^0) = p(y, s|\pi^T; t) - p(y, s|\pi^0; t)$$

Stationarity assumes that the *difference* in the marginal distributions of s and y with respect to any treatment policy π^T and the control policy π^0 at each period match those induced by the Markov transition probabilities. The assumption allows *levels* of these distributions to change, as long as the changes apply equally to treatment and control populations.

2.1. Identification

A necessary step in estimating the average treatment effect of a treatment policy depicted in Figure 2 is to express the estimand as a function of observable data available from an experiment. In particular, we assume the observable data consists of N i.i.d. tuples (a, s, y, s') generated from the process illustrated in Figure 1.

To do so, we will exploit the fact that the environment and assumptions above describe a Markov Decision Problem (MDP). Our setup is an MDP with a binary action space, $\mathcal{A}$, state-space $\mathcal{S}$, expected reward emission function, $p(y|s,a)$, state-transition kernel $p(s'|s,a)$ and a *non-stationary* policy, $\pi_t^T : \mathcal{A} \times \mathcal{S} \times \mathbb{N}^+ \to [0,1]$.

Leaning on the framing as a MDP, we can summarize the cumulative discounted outcomes recursively using the state-action value function (the Q function) defined as follows.

$$q_t^T(s,a) \equiv \mathbb{E}_y\left[y|s,a\right] + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}\left[q_{t+1}^T(s', \mathbb{1}_{t+1<T})\right] \tag{2.4}$$

The superscript of the Q function denotes the associated policy is the T-duration treatment policy π^T and the subscript denotes the dependence on time. The Q function as described in (2.4) is non-stationary because the T-duration treatment policy is non-stationary. Note that for fixed policies, either persistent treatment or control, the Q function is stationary since the underlying policy is constant over time.

Theorem 1 (Identification by Q function with non-stationary policy)**.** *Suppose Assumptions 1-4 hold. Then the average treatment effect of a T-duration treatment policy is composed of the following function of observable data.*

$$\varphi^T = (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot)}\left[q_0^T(s, \mathbb{1}_{0<T}) - q^0(s,0)\right] \quad (2.5)$$

Theorem 1 transforms the complex task of estimating infinite horizon per-period outcomes into a straightforward computation of the Q function, weighted appropriately and evaluated at the initial state and action. The proof is provided in the Appendix in Section A.1.

2.1.1. Stationary Policies

Theorem 1 is challenging to use directly since the Q function in Equation (2.5) is difficult to estimate due to it inheriting non-stationarity from the underlying T-duration treatment policy.[2] Instead, we prove the existence of an equivalent stationary stochastic policy and construct a computationally efficient approximation. With such a stationary policy, we construct a practical version of Theorem 2 in Corollary 4 which uses a stationary policy making the Q function tractable.

A key concept we need is that of an occupancy measure, the discounted fraction of time an agent spends in state s and action a.

$$\rho_{\pi,\gamma}(s,a) \equiv (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \rho_{\pi,t}(s,a) \quad (2.6)$$

Similarly, there exists the state occupancy measure which is Equation (2.6) marginalizing out actions. The occupancy measures provide a way to express the cumulative discounted outcomes, a summation across time, instead as a single point-in-time weighted average of outcomes.

$$(1-\gamma)\mathbb{E}_{p_0,\pi}\left[\sum_{t=0}^{\infty}\gamma^t y_t\right] = \mathbb{E}_{s,a\sim\rho_{\pi,\gamma}(\cdot),y\sim p(\cdot|s,a)}\left[y\right] \quad (2.7)$$

An implication of Equation (2.7) is that two policies that lead to the same occupancy measures will have the same cumulative discounted rewards.

Lemma 2 (Stationary equivalents of non-stationary policies)**.** *For any non-stationary policy $\pi = \pi_0, \pi_1, \ldots$, there exists a stationary policy $\bar{\pi}$ that generates the same occupancy measure. In particular construct a stationary policy as follows:*

$$\bar{\pi}(a|s) = \frac{\rho_{\pi,\gamma}(s,a)}{\rho_{\pi,\gamma}(s)}, \quad (2.8)$$

then

$$\rho_{\pi,\gamma} = \rho_{\bar{\pi},\gamma}. \quad (2.9)$$

See (Bertsekas, 2001) for a proof.

Theorem 3 (Stationary T-Duration Treatments)**.** *For a non-stationary policy π^T that sets $a=1$ for T periods and $a=0$ thereafter, (i) there exists an equivalent stationary stochastic policy $\bar{\pi}^T$ that yields the same cumulative discounted reward and (ii) the average of that stationary stochastic policy across states is $1-\gamma^T$.*

Proof. Let π^T be an arbitrary non-stationary policy. That non-stationary policy leads to associated occupancy measures, $\rho_{\pi^T,\gamma}$. Construct a candidate stationary policy:

$$\bar{\pi}^T(s,a) = \frac{\rho_{\pi^T,\gamma}(s,a)}{\rho_{\pi^T,\gamma}(s)} \quad (2.10)$$

Lemma 2 shows that $\bar{\pi}^T$ leads to an equivalent occupancy measure, and hence will result in the same expected cumulative discounted reward as under π^T.

For (ii), the weighted average treatment policy across states is:

$$\begin{aligned}\int_s \bar{\pi}^T(a|s)\rho_{\bar{\pi}^T,\gamma}(s)ds &= \int_s \rho_{\pi^T,\gamma}(s,a)ds \\ &= \int_s (1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t\rho_{\pi^T,t}(s,a)ds \\ &= (1-\gamma)\textstyle\sum_{t=0}^{\infty}\gamma^t\int_s\rho_{\pi^T,t}(s,a)ds \\ &= (1-\gamma)\textstyle\sum_{t=0}^{T}\gamma^t = 1-\gamma^T. \quad \square\end{aligned}$$

Intuitively, a stationary policy with a constant treatment probability of 0 corresponds to a control policy indexed by $T=0$. As T increases, so does this probability, and as $T\to\infty$, it approaches 1. In general, constructing the exact state-dependent equivalent stationary policy is intractable since it requires estimating the occupancy measures under the T-duration treatment policy. Instead, we suggest using the state-independent policy, $\forall s\colon\ \bar{\pi}^T(a|s) = 1-\gamma^T$, which offers a practical and computationally efficient approximation.

Constructing a stationary policy from a T-duration policy via Equation (2.10) leads to a stationary Q function, equivalent in occupancy measures and expected outcomes to the non-stationary Q function in Equation (2.4), when starting from the same initial distribution of states.

$$q^{T}(s,a) \equiv \mathbb{E}_y\left[y|s,a\right] + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a),a'\sim\bar{\pi}^T(\cdot|s')}\left[q^T(s',a')\right] \quad (2.11)$$

Hence we can state a stationary version of Theorem 1, with a stationary and hence learnable Q function.

Corollary 4 (Identification by Stationary-policy Q)**.** *Suppose Assumptions 1-4 hold. Then the expected average treatment effect of a T-duration treatment policy is equal to*

[2] For our specific form of non-stationarity, one would need to first estimate the Q function at T under the deterministic control policy, then the $T-1,\ldots,0$ Q functions in order under the deterministic treatment policy.

expectation over the difference of Q functions, associated with the equivalent stationary policy, $\bar{\pi}^T$ and the control policy.

$$\varphi^T = (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot),a\sim\bar{\pi}^T(\cdot|s)}\left[q^T(s,a) - q^0(s,0)\right] \tag{2.12}$$

2.2. Comparison to Surrogate Index Method

Within the current environment, it is instructive to pinpoint the key difference between our method and the surrogate index method (Athey et al., 2019). Assume the setup above where we observe everything up to some period t, where we observe only the transition to the tth period state. In other words, we observe N tuples of $(s_0, y_0, a_0, \ldots, s_t)$.

We focus on the difference in outcomes for a permanent treatment policy for periods past t since the covariate adjusted difference in means will recover the treatment effect prior to t. Following the assumptions in Section 2.1, the true potential outcome for period $t+k$ where $k>0$, can be expressed as

$$\begin{aligned}\mathbb{E}\left[Y_{t+k}(1)\right] = \int_{s_t,s,y} & yp(y \mid s, a=1) \\ & \times p(s \mid s_t, a=1; k)p_t(s_t \mid a=1). \end{aligned} \tag{2.13}$$

where $p(s'|s,a;k)$ is the transition kernel projected k periods ahead starting from s.

A surrogate method that relies on an observational dataset will instead calculate the expectation of the $t+k$ period outcome conditional on the distribution of s_t from the experiment but also in part on a probability model, p^o learned from an observational dataset.

$$\begin{aligned}&\mathbb{E}\left[Y_{t+k}(1)\right]] \leftarrow \mathbb{E}_{s_t}\left[\mathbb{E}^o_Y\left[Y_{t+k}|s_t, a=0\right]|a=1\right] \\ &= \int_{s_t,y_{t+k}} y_{t+k}p^o(y_{t+k} \mid s_t, a=0)p_t(s_t|a=1) \\ &= \int_{s_t,s_{t+k},y_{t+k}} y_{t+k}p^o(y_{t+k} \mid s_{t+k}, a=0) \\ &\quad \times p^o(s_{t+k}|s_t, a=0)p_t(s_t|a=1) \end{aligned} \tag{2.14}$$

Note that when the observational model is used, it conditions on null treatment since our novel treatment doesn't exist in the observational dataset. The comparability assumption ensures that the observational and experimental probabilities are equal, $p^o = p$.

Equation (2.14) makes it clear that the surrogate estimate only captures the partial treatment effect that is mediated through the surrogate, s_t. For periods beyond the measurement period of the surrogate, it misses that *permanent* interventions may alter (i) state transitions and hence affect the distribution of future states, $p(s_{t+k}|s_t, a=1) \neq p(s_{t+k}|s_t, a=0)$ and (ii) the contemporaneous relationship between state and outcome, $p(y|s,a=1) \neq p(y|s,a=0)$.[3]

Hence surrogate index methods capture long-term effects, but only for treatment durations up to the time when the surrogate is measured. The effects captured are indirect long-term effects due to the persistence of initial treatment effects. We verify this experimentally in Section 4.

3. Estimation

We want to estimate the long-term average treatment effect via the Q function in Equation (2.12). With discrete states, we can solve for Q exactly via dynamic programming methods subject to computational constraints (Sutton & Barto, 1998). But when the state-space is large or continuous, we need to rely on machine learning techniques to approximate the Q function.

It is well known that relying on ML-based estimators in a statistical estimand may lead to bias due to overfitting and regularization techniques used in training (Chernozhukov et al., 2018). Hence we develop a double ML based estimator centered around the efficient influence function (Kallus & Uehara, 2022). The estimator is $N^{-\frac{1}{2}}$ consistent and doubly robust with respect to ML-learned Q and density ratio functions, which are only required to converge at slow rates.

3.1. Efficient Influence Function Based Estimator

The estimator we propose is the naive plug-in estimator with a bias correction term based on the efficient influence function. The efficient influence function for one half of the estimand (the potential outcome under the policy π) is a function of the observed tuple (s,a,y,s'), a stationary policy π and the nuisance functions q and w, representing the Q and density ratio functions.

$$\begin{aligned}&\phi^\pi(s,a,y,s';q,w) = -\varphi^\pi + (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot)}\left[q^\pi(s,a)\right] \\ &+\frac{\pi(a|s)}{p_b(a|s)}w(s)\left(y+\gamma\sum_{a'\in\mathcal{A}}\pi(a'|s')q^\pi(s',a') - q^\pi(s,a)\right)\end{aligned} \tag{3.15}$$

The efficient influence function for the estimand is simply the difference of the respective efficient influence functions for treatment and control.

$$\begin{aligned}\phi(s,a,y,s';q,w,\pi,\pi^0) = \phi^\pi(s,a,y,s';q,w) \\ - \phi^{\pi^0}(s,a,y,s';q,w)\end{aligned} \tag{3.16}$$

[3] Of course, these effects diminish as the period of surrogate measurement increases. But this point is moot as the problem at hand is to estimate long-term effects on short term experimental measurements.

The density ratio function is defined as

$$w(s) \equiv \frac{\rho_{\pi,\gamma}(s)}{p_b(s)}. \tag{3.17}$$

An intuitive description of the density ratio function is the occupancy measure under the policy relative to the probability density function under the experiment.

The bias-corrected estimator is

$$\varphi^{\pi}_{BC} \equiv (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot),a\sim\pi(\cdot|s)}\left[q^{\pi}(s,a) - q^{\pi^0}(s,0)\right] + \mathbb{E}\left[\phi(s,a,y,s';q,w,\pi,\pi^0)\right] \tag{3.18}$$

where the final term is a bias correction term that undoes any asymptotic bias from using ML-based estimators of the nuisance and value functions (Kennedy, 2023).

3.2. Asymptotic Properties

By design, we specified the ATE as a function of a well studied objective in reinforcement learning: the normalized discounted outcomes associated with a policy. Hence we can make use of results from (Kallus & Uehara, 2022) who propose efficient, doubly robust estimators for off-policy evaluation using semiparametric methods. This section summarizes the relevant results.

The difference between the bias-corrected estimator and the true estimand can be decomposed into three components: a central limit theorem term, an empirical process term and a second order remainder term. The key is to show that the empirical process term and the second order remainder term converge to zero faster than the central limit theorem term.

The empirical process term is $o_p(N^{-\frac{1}{2}})$ if we use cross fitting in estimation. Cross fitting involves splitting the data into K partitions, estimating nuisance functions on the $K-1$ held out partitions, evaluating the estimator for each single partition and finally averaging over the K estimators to get the final estimate.

For brevity, we refer to (Kallus & Uehara, 2022) for details on controlling the second order remainder term.

Theorem 5 (Double Robustness)**.** *If either one of $\|\hat{q} - q\|_2 = o_p(1)$ or $\|\hat{w} - w\|_2 = o_p(1)$ holds, then $\hat{\varphi}_{BC} - \varphi = o_p(1)$.*

Double robustness implies that we only need to "correctly" estimate one of either the Q or the density ratio functions to ensure our bias corrected estimator is consistent.

Theorem 6 (Asymptotic Normality and Efficiency)**.** *Suppose that (i) $\hat{q}$ and $\hat{w}$ converge to q and w in probability at rates such that the product of those rates is $o_p(N^{-\frac{1}{2}})$ and (ii) the propensity score $p_0(a|x)$ is known. Then the bias-corrected estimator is asymptotically normal and efficient.*

$$\sqrt{N}\left(\hat{\varphi}^{\pi}_{BC} - \varphi^{\pi}\right) \xrightarrow{d} \mathcal{N}\left(0, \phi^2\right)$$

Crucially, the convergence rate requirements on the Q and density ratio function estimates are each slower than square-root which enables the use of a range of ML algorithms along with techniques such as regularization. If the propensity score needs to be estimated, then the rate requirement on the density ratio function instead applies to the product of the density ratio function and the propensity score.

3.3. Q Function Estimation

If states are finite, then dynamic programming techniques can solve for the Q function exactly given the availability of transition probabilities. However, since we potentially have continuous states or a large state space which is subject to the curse of dimensionality, we need to use ML techniques that parameterize the Q function.

An obvious choice is the family of Temporal Difference (TD) algorithms used for policy evaluation. TD algorithms estimate the Q function on a dataset of state transitions, actions and rewards, as available in an experiment. In particular, it requires a dataset of (s_i, a_i, y_i, s'_i) tuples for $i = 1, \ldots, N$ units and parameterizes the Q function with a vector of parameters, θ_q. Hence

$$q^{\pi}(s,a;\theta_q) \approx q^{\pi}(s,a).$$

Using the definition of the Q function from Equation (2.11), we can form the TD error term

$$L_Q(s,a,y,s') = y + \gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[q^{\pi}(s',a';\theta_q)\right] - q^{\pi}(s,a;\theta_q) \tag{3.19}$$

whose expectation is zero for the true Q function.

Within the family of TD methods, various approaches have been proposed which center on minimizing the TD error (Sutton & Barto, 1998). Framing the TD error in Equation 3.19 as an estimating equation, one can use techniques from M-estimation to derive asymptotic properties such as asymptotic normality and consistency. For example, (Kallus & Uehara, 2022) derive the asymptotic lower bound for an M-estimator that seeks to minimize a weighted form of Equation (3.19).

3.4. Density Ratio Estimation

The use of density ratio functions in reinforcement learning is relatively new, finding recent use in methods for efficient off-policy evaluation (Liu et al., 2018; Uehara et al., 2020; Kallus & Uehara, 2022). Their estimation has centered on the following relationship:

$$L_W(f,w) = \mathbb{E}_{s,a,y,s'\sim p_0,a'\sim\pi(\cdot|s')}\left[w(s,a)\left(\gamma f(s',a') - f(s,a)\right)\right] - (1-\gamma)\mathbb{E}_{s\sim p_0,a\sim\pi(\cdot|s)}\left[f(s,a)\right], \tag{3.20}$$

which equals zero for the true density ratio function and can be derived from the definition of the Q function.

Under some mild technical conditions, (Uehara et al., 2020) show that if $L_W(f, \hat{w}) = 0$ for all f, then $\hat{w} = w$. Moreover, the reverse also holds, that the true density ratio function is the only function for which the statement is true.

This leads to a Minimax-style estimator, with two function classes, $\mathcal{F}$ and $\mathcal{W}$, each encompassing the discriminator and the density ratio functions.

$$\hat{w}(s, a) = \arg\min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} L_W(f, w)^2 \tag{3.21}$$

Minimax estimators can be challenging to implement due to the inner maximization. Fortunately, the Minimax objective can be reduced to avoid the inner maximization in two cases (Uehara et al., 2020). First, when the function classes of the density ratio functions and the discriminator are linear under the same feature maps for state and actions. Second, when the discriminator function class corresponds to a reproducing kernel Hilbert space.

4. Experiments

To demonstrate the effectiveness of the ORL method, we perform experiments on simulated data.[4] First, we empirically validate some of the theoretical results on a simple MDP which allow us to precisely control aspects of the data generating process so that we can see the effects of gradually relaxing the adherence to our assumptions. Second, we evaluate the method using a simulator of sepsis where effect sizes and randomness reflect real-world settings (Oberst & Sontag, 2019).

4.1. Simple MDP

Simulation Details We generate data with characteristics that mirror a streaming video on demand service. We construct a single state Markov chain mirroring the dynamics of on-service tenure (state) and subscription revenue (reward). On-service tenure increases if members do not churn, which we model as a drift-diffusion process with positive drift:

$$s_t = s_{t-1} + \mu + \tau_s a_t + \sigma_s w_t \tag{4.22}$$

where $w_t \sim \mathcal{N}(0, 1)$ and we restrict $s_t \in [0, 1]$ by clipping.

Since longer tenured members often have higher yet capped per period revenue, we set the reward to be a diminishing scalar multiple of the state.

$$y_t = \alpha s_t^{\theta} + \tau_y a_t + \sigma_r e_t \tag{4.23}$$

where $e_t \sim \mathcal{N}(0, 1)$.

[4]Code and data for the experiments is available at: https://github.com/allentran/long-term-ate-orl .

Treatment effects both increase the probability of a transition to a higher state and increase average per-period rewards conditional on the state. Table 1 in the Appendix lists the parameter values used in the experiments.

Within this setting, we (i) benchmark a variety of estimators across a range of underlying treatment durations, (ii) determine the effect of a hidden state on estimates while varying the "relevance" of the hidden state and (iii) assess the impact of decreased support in the experimental data for states whose treatment effects we wish to evaluate.

Treatment duration We simulate trajectories over a long horizon to mimic an infinite horizon where treatments vary in duration from 1, 6, 12, 24, 48 and 120 periods.[5] Note that the estimand is infinite horizon, so that even for short duration treatments, we are interested in the long-term ATE.

In addition to the ORL method, we evaluate a naive method and two variants of the surrogate index method. The naive method assumes the 1-period experimental ATE is constant throughout the duration of treatment and zero thereafter. The first surrogate index variant sees two periods of the experiment, resembling access to data from a short experiment whereas the second variant sees the experiment for the duration treatment is applied.

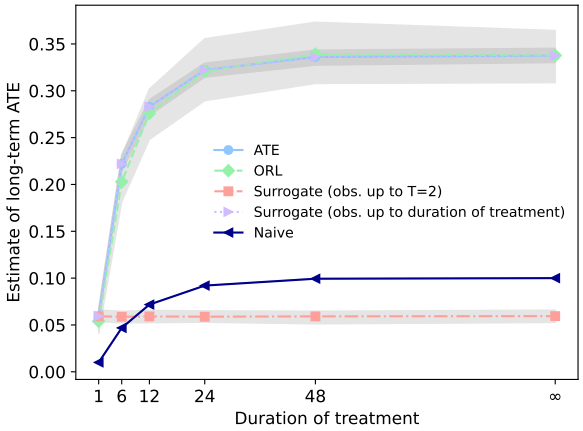

Figure 3. Estimates of the ATE for differing treatment durations

Figure 3 shows estimates across the estimators for varying durations of treatment. Since the treatment increases both the rewards per state and the likelihood of a positive state transition, the true ATE (light blue) increases with the duration of the treatment. Estimates from the ORL method (green) match the true ATE, as the duration of treatment extends all the way to a permanent intervention.

[5]We label the 120 period duration as ∞ as the two durations are quantitatively equivalent given our choice of the discount rate.

Conversely, the surrogate index method estimates outcomes as if treatment is applied at most, up to the point at which the last surrogate is measured. Hence the surrogate index method with the last surrogate observed in the second period estimates the ATE for treatment active for a single period. The estimate is constant and therefore underestimates the true ATE for $T > 1$. The surrogate method (purple) matches the true ATE but crucially, requires an infeasible duration of measurement, with surrogates measured up to the target duration of treatment.

Lack of coverage in experimental data The ORL method uses data from an experiment with a particular distribution over states and actions to estimate causal effects for a potentially different distribution of initial states. The two distributions can differ vastly in situations such as medical trials where it may be difficult to enrol some segments of the population of interest. This difference is summarized by w, the density ratio function in Equation (3.17) and explodes as the coverage in the experimental distribution goes to zero. Accordingly, the asymptotic variance of the ORL estimator increases with the square of the density ratio function.

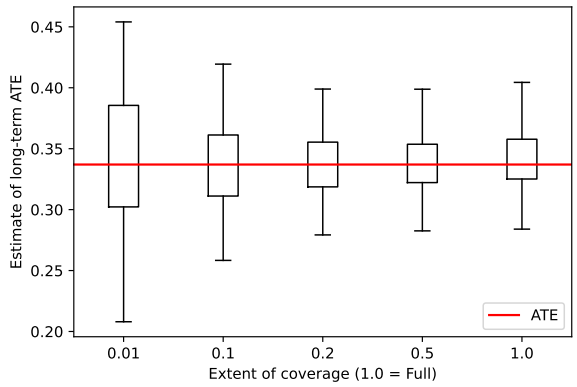

Figure 4. ORL estimates for differing state coverage in experimental data

To assess the effect of a lack of "coverage", we downweight observations with states within an interval of 0.2 and 0.8 by various degrees and resample with replacement to keep the number of observations fixed.[6] Results in Figure 4 show that as expected, the variance of estimates increases as coverage decreases, although only having an effect when the missing data is downweighted 1:100.

Effects of a hidden state A key assumption is that we observe all the states underlying the MDP. The existence of a hidden state may lead to bias and increased variance in estimates of the ATE since the Q and density ratio functions are underspecified.

[6]The evaluation state distribution is uniform between 0 and 1.

The impact of the hidden state depends largely on how independent the hidden state is and whether it affects state transitions that govern reward emissions. To demonstrate this, we add an additional state and alter state transitions via Equation (4.24). When $\rho = 0$, the hidden state is decoupled from the MDP and functions as noise. When $\rho = 1$, the state which governs reward emissions depends entirely on the hidden state and hence the hidden state needs to be observed for accurate Q function estimation.

$$\mathbf{s_t} = P\mathbf{s_{t-1}} + \mu + \tau_s a + \sigma_s \mathbf{w_t} \tag{4.24}$$

where bolded variables represent column vectors and

$$P = \begin{bmatrix} 1-\rho & \rho \\ 0 & 1 \end{bmatrix}. \tag{4.25}$$

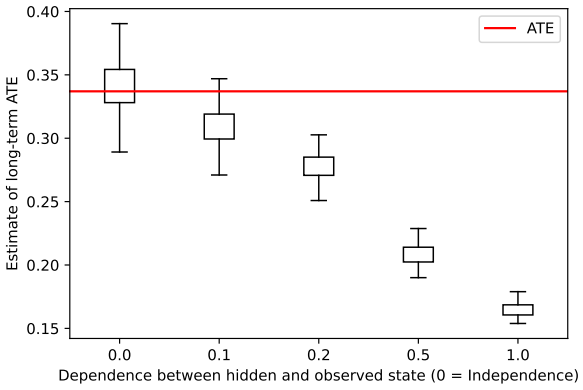

Figure 5. ORL estimates with unobserved hidden state of varying relevance

Figure 5 shows that increasing the relevance of a hidden state introduces severe bias. Although not depicted, the cause for the bias is downwards bias in the estimate of rewards from the treatment. As $\rho \to 1$, the model has difficulty predicting state transitions which leads to predicting transitions closer to the mean of the state distribution, understating treatment effects. This highlights the importance of using a large state-space and methods capable of overcoming the curse of dimensionality.

4.2. Sepsis Simulator

To evaluate our method in a more realistic setting, we generate data from a simulator of sepsis (Oberst & Sontag, 2019). The simulator is based on an underlying MDP with measurements of vital signs and previous treatment as states and rewards of $+1$ and -1 respectively upon patient discharge and death. Using a more realistic environment allows us to assess more quantitative dimensions of performance, such as the size of confidence intervals compared to effect

sizes. To maintain the constraint of realism, we use feature representations that are human-interpretable (e.g each vital sign and its measurement is a feature) as opposed to using one-hot encoded representations of each discrete state.

We mimic experiments by using each of the available medical treatments, the cartesian product of {antibiotics, vasopressors, ventilation}, as binary treatments to evaluate against the control of null treatment. As before, we simulate ground truth ATEs by applying the treatment for varying durations (1, 6, 12, 24, 48 and ∞) and then generate an experimental dataset of treatment and control. Results for two of the treatments, ventilation and ventilation in addition to antibiotics, spanning the spectrum of behavior of all 7 treatments, is shown in Figure 6.

For both treatments, estimates for the ORL method largely center on the true ATE. Estimates of the ATE from ventilation are close to zero and correspondingly confidence intervals for all treatment durations span zero. On the other hand, treatment effects from ventilation and antibiotics together are estimated to be large and increasing with the duration of treatment, matching the increase in the true ATE with treatment duration. Interestingly, estimates for both treatments appear to degrade as the treatment duration shortens. This is due to the fact that the stationary policy used to mimic a short term experiment is an approximation that gets worse the shorter the target duration of treatment.

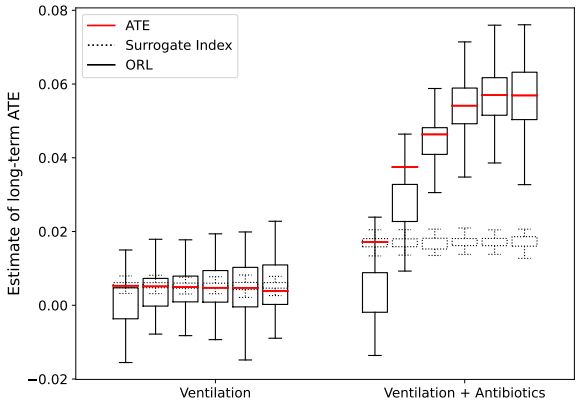

Figure 6. Estimated ATEs of two treatments for Sepsis, for treatment durations of 1, 6, 12, 24, 48 and ∞

In addition, Figure 6 shows estimates for the surrogate index method where the surrogate variable is the state at $t = 1$. As in the case for the simple MDP, the surrogate index estimate is only accurate for the treatment duration of 1 period. In the case of the ventilation treatment, the surrogate index estimate is correct but purely due to the fact that treatment duration has no effect on the long-term ATE. On the other hand, the ventilation and antibiotics estimates make it clear that the surrogate index method is biased when treatment duration has a sizeable effect on the long-term ATE.

5. Conclusion

We develop a method for inferring the ATE of continual exposure to a long-term treatment, when only data from a short-term experiment is available. The key difficulty is that the treatment we consider is both novel and long-term. Together, this means that surrogate methods are unsuitable since surrogacy assumptions do not hold. Instead, we proceed by making a connection to offline reinforcement learning and embed our problem within a Markov decision process.

We reframe the problem of estimating the long-term ATE as evaluating the difference in the long-term outcomes of two different policies: a treatment and control policy. By constructing stationary policies equivalent to arbitrary-duration treatment regimes, we are able to make use of tools from off-policy reinforcement learning. In particular, we use an estimator which depends on the Q and density ratio functions. Importantly, the estimator is doubly-robust with respect to these nuisance functions and asymptotically efficient.

Experiments on simulated data demonstrate the effectiveness of the ORL method over alternative methods. Estimates from the ORL method generally match the true ATE for the full spectrum of treatment durations, from single-period to permanent, using only two periods of data from the an experiment.

In conclusion, our proposed method provides a robust and efficient solution for estimating the long-term ATE of continual exposure to a novel long-term treatment, when only short-term experimental data are available. We do so without the need for long-term data, thereby bridging a significant gap in the study of long-term treatments. This opens up new possibilities for research and interventions in various fields where understanding the long-term effects of treatments is crucial.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. NBER Working Papers 26463, National Bureau of Economic Research, Inc, November 2019. URL https://ideas.repec.org/p/nbr/nberwo/26463.html.

Athey, S., Chetty, R., and Imbens, G. Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*, 2020.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oprescu, M., and Syrgkanis, V. Estimating the long-term effects of novel treatments. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2925–2935. Curran Associates, Inc., 2021.

Bertsekas, D. *Dynamic Programming and Optimal Control*, volume 1 and 2. Athena Scientific, 2 edition, 2001.

Bica, I., Alaa, A. M., and Van Der Schaar, M. Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Chen, H., Geng, Z., and Jia, J. Criteria for surrogate end points. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 69(5):919–932, 2007. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/4623303.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. Automatic debiased machine learning for dynamic treatment effects and general nested functionals, 2023.

Huang, S., Wang, C., Yuan, Y., Zhao, J., and Zhang, J. Estimating effects of long-term treatments. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, pp. 907, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597701. URL https://doi.org/10.1145/3580507.3597701.

Imbens, G., Kallus, N., Mao, X., and Wang, Y. Long-term causal inference under persistent confounding via data combination, 2023.

Kallus, N. and Mao, X. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022. doi: 10.1287/opre.2021.2249. URL https://doi.org/10.1287/opre.2021.2249.

Kennedy, E. H. Semiparametric doubly robust targeted double machine learning: a review, 2023.

Lewis, G. and Syrgkanis, V. Double/debiased machine learning for dynamic treatment effects. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22695–22707. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bf65417dcecc7f2b0006e1f5793b7143-Paper.pdf.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/dda04f9d634145a9c68d5dfe53b21272-Paper.pdf.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

Murphy, S. A. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003. doi: https://doi.org/10.1111/1467-9868.00389. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00389.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4881–4890. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/oberst19a.html.

Prentice, R. L. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, 8(4):431–440, 1989. doi: https://doi.org/10.1002/sim.4780080407. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780080407.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

VanderWeele, T. J. Surrogate measures and consistent surrogates. *Biometrics*, 69(3):561–569, 2013. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/24538119.

Yang, J., Eckles, D., Dhillon, P., and Aral, S. Targeting for long-term outcomes. *Management Science*, 0(0), 2023.

A. Proofs

A.1. Proof of Theorem 1

Since expectations are linear, it suffices to show that each per period outcome of the long-term outcome (each term in Equation (2.2)) can be expressed as a function of observable data. For periods beyond the first:

$$\begin{aligned}
\mathbb{E}\left[Y_t(\pi^T)|S_0, A_0\right] &= \mathbb{E}_Y\left[Y_t(\pi^T) \mid S_0, A_0\right] \\
&= \mathbb{E}_Y\left[Y_t(\pi^T) \mid \pi^T, S_0, A_0\right] \\
&= \mathbb{E}_Y\left[Y_t \mid A_t = 1_{t<T}, \pi^T, S_0, A_0\right] \\
&= \mathbb{E}_{S_t}\left[\mathbb{E}_Y\left[Y_t \mid A_t = 1_{t<T}, S_t\right]|\pi^T, S_0, A_0\right] \\
&= \mathbb{E}_S\left[\mathbb{E}_Y\left[Y \mid A = 1_{t<T}, S\right] \mid \pi^T, S_0, A_0; t\right] \qquad \text{(A.26)}
\end{aligned}$$

The first and fourth equalities rely on the law of iterated expectations, the second is justified via unconfoundedness and the third uses the definition of a potential outcome. The final equality relies on stationarity where the notation $\mathbb{E}\left[\cdot; t\right]$ denotes the expectation induced by projecting t periods ahead under the Markov model. The same derivation can be done for the first period where the action is A_0.

Applying this for all periods

$$\mathbb{E}\left[\sum_t \gamma^t Y_t(\pi^T)|S_0, A_0\right] = \mathbb{E}_Y\left[Y \mid A_0, S_0\right]\sum_{t=1}\gamma^t \sum_S \mathbb{E}_Y\left[Y \mid A = 1_{t<T}, S\right] p(S|\pi^T, S_0, A_0; t). \qquad \text{(A.27)}$$

From here, one can use the definition of the Q function and use the standard proof to show the equivalence of expected discounted rewards from an initial state-action to the Q function (Sutton & Barto, 1998).

B. Details of Experiments

B.1. Q function and density ration model implementation details

The ORL method requires the estimation of two nuisance functions, the Q function and the density ratio functions. For the Q function, we use a feed-forward neural network parameterized separately for each of treatment and control. Each network consists of two hidden layers with 128 and 64 features respectively with sigmoid activation functions and a linear final layer with no activation function. Additionally, we maintain separate "target" networks by freezing the parameters of each network for 64 epochs, which proved invaluable in stabilizing training (Mnih et al., 2015).

For the density ratio functions, we use the Minimax weight estimator from (Uehara et al., 2020) where we restrict both the discriminator and density ratio function classes to be linear with the feature maps $\phi(s, a) = \begin{bmatrix} s & s^2 & sa & s^2 & s^2a & a & 1 \end{bmatrix}$.

B.2. Sepsis simulator model details

As the sepsis simulator relies on discrete states, with enough data, a one-hot encoded representation of the states identifying each discrete state would recover the ATE exactly. To ensure a realistic setting where the true states are unknown, we keep states in the original human-interpretable format, with each vital sign and its measurement as a dimension of the state-space. In addition, the simulator generates trajectories that are episodic since death and discharge are terminal events whereas our method applies to tasks that are continuing. To handle this, we force outputs from the learned Q function to be zero for terminal states. Alternatively, we could have introduced another dimension of the state-space to be a boolean terminal state indicator and added terminal state transitions with zero reward to the training data.

Table 1. Simulation Parameter Values

Notation	Description	Value
γ	Discount rate	0.9
μ	Drift in state transition	0.05
σ_s	Std. dev. in state transition in 5	0.1
σ_r	Std. dev. in state-outcome mapping	0.1
θ	Curvature in state-outcome mapping	0.8
τ_s	Treatment effect on state transition	0.05
τ_y	Treatment effect on per-period outcome	0.01