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

Allen Tran
Netflix Inc.
Los Angeles, USA
allent@netflix.com

Aurélien Bibaut
Netflix Inc.
Los Gatos, USA
abibaut@netflixcontractors.com

Nathan Kallus
Netflix Inc. & Cornell University
New York, NY, USA
nkallus@netflix.com

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 cannot reasonably hold. Our approach instead learns long-term temporal dynamics directly from short-term experimental data, assuming that the initial dynamics observed persist but avoiding the need for both surrogacy assumptions and auxiliary data with long-term observations. 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. Finally, we demonstrate the method in simulated experiments.

KEYWORDS

Experimentation, Long-term Causal Inference, Markov Decision Processes

ACM Reference Format:
Allen Tran, Aurélien Bibaut, and Nathan Kallus. 2023. Inferring the Long-Term Causal Effects of Long-Term Treatments from Short-Term Experiments. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 of short-run A/B tests to inform decisions that ultimately aim to improve long-term outcomes.

Surrogate methods appear to offer a route to connect short term tests to their longer term outcomes [1, 12]. These methods rely on the existence of intermediate-term surrogate variables and/or an observational dataset that associates surrogate variables to 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 a 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 persits, 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. First, that dynamics of the underlying environment satisfy the Markov property. Second, the experiment is designed such that the treatment is applied widely, providing observations that cover the entire state-space of the population. The Markov assumption allows us to project long term effects from the experiment while the coverage assumptions ensures common support exists in the experimental data.

These assumptions 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 differing from the policy generating the data. We make use of the connection

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

with ORL by leveraging recent literature that develops efficient doubly-robust estimators for off-policy 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 provides an overview of related literature. Section 2 sets up the methodology, provides conditions for the identifiability of the estimand and demonstrates the potential bias of surrogate methods under permanent interventions. Section 3 introduces the estimator and conditions for root-n consistency and asymptotic normality. Section 4 simulates experimental data and compares the proposed method to a surrogate method baseline for varying treatment durations. 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 various surrogacy assumptions that share the requirement that the surrogate mediates the treatment effect [12, 15].

However, surrogate assumptions are unlikely to hold for a single surrogate and can potentially lead to sign-reversing bias [4]. Hence, more recent research 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. In particular, Athey et al. [1] use semiparametric methods to perform inference on an estimator using many surrogates that combines experimental and observational data. Extensions to this line of work include extending the environment to a dynamic setting [2], learning optimal policies [16], and combining long- and short-term data to tackle confounding [6**?**] and improve efficiency [7].

The reinforecement learning literature also estimates long term outcomes, albeit from the perspective of quantifying the value of different "policies" [13]. We make direct use of an estimator from Kallus and Uehara [8] that combines two functions: the Q function, which has a long history in reinforcement learning and the density ratio function [10, 14].

Our contribution is to map the problem of estimating long term effects from interventions to estimating the difference in outcomes from two specific types of policies. Moreover, we do so by avoiding the use of the surrogacy assumptions which allows us to generalize the estimation of long term effects to treatment regimes of any duration.

2 METHODOLOGY

We're interested in estimating long term effects from an intervention of some duration, where experimental evidence is only available for a shorter duration. For example, estimating the effect on customer lifetime value from a "permanent" change in a recommendation algorithm with evidence from a short-run experiment.

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 $Y(\pi)$.

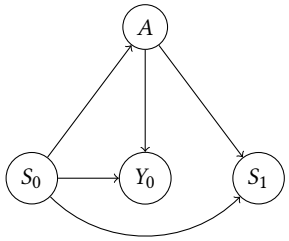

Figure 1: DAG of Experiment

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.

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

The experiment that generates our data is described in Figure 1. There exists an initial distribution of "states", S_0, from 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. The distribution of these variables under the experiment will be denoted by p_0.

The traditional causal inference challenge is to estimate treatment effects for the experiment, which entails overcoming the missing counterfactual outcomes. Here, we face an additional problem in that we are interested in long outcomes for various treatment policies of interest. Figure 2 depicts the treatment policy and outcomes that we are interested in estimating. The first 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 1. *Unconfoundedness*

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

Given a set of initial states S_0, we assume treatment assignment in the experiment is independent of potential outcomes conditional

[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).

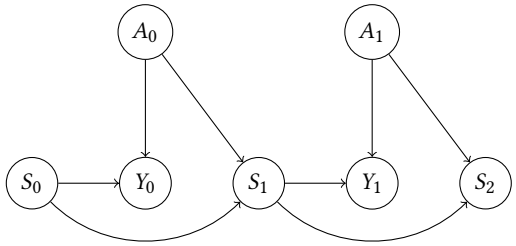

Figure 2: DAG of Treatment Policy of Interest

on the initial state, which should be satisfied with experimental data.

Assumption 2. *Overlap*

$$\forall s \in \mathcal{S}, a \in \mathcal{A}\colon\ 0 < p_0(s,a) < 1$$

As standard in the causal inference literature, overlap requires that we see both treated and control units for every state. Here, our overlap condition is stronger than in traditional settings since it technically applies to the entire state-space and not just the initial states. For instance, an algorithm change that is rolled out to all types of users of the platform as opposed to being rolled out for only 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 3. *Markov property*

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

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.

$$p(y,s|\pi^T;t) = \int_{s_0,\ldots,s_{t-1}} p_0(s_0)p(s_1|s_0,\mathbb{1}_{0<T})\ldots p(s|s_{t-1},\mathbb{1}_{t-1<T})p(y|s,\mathbb{1}_{t<T}) \tag{2.3}$$

Assumption 4. *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.

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

$$\rho_{\pi,\gamma}(s,a) \equiv (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \rho_{\pi,t}(s,a) \tag{2.5}$$

Similarly, there exists the state occupancy measure:

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

The occupancy measures provide a way to express the cumulative discounted outcomes, a summation across time, instead as a static weighted average of single period outcomes across states and actions.

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

Theorem 1 (Identification by Non-stationary policy Q). *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.*

$$\varphi^T = (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot)}\left[q_0^T(s,\mathbb{1}_{0<T}) - q^0(s,0)\right] \tag{2.8}$$

Theorem 1 simplifies the estimation of long-term treatment effects. It 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 key insight is that if we observe state transitions for all pairs of states for each of treatment and control (the actions), then we can identify the probability of a unit being in a state at any time period. Since we also observe the expected outcome conditional on the state, we can characterize the distribution of expected outcomes across time. Stationarity guarantees that these conditional probabilities are constant over time (up to differencing) and hence we can estimate these distributions on our short-term experimental data. The proof is provided in the Appendix in Section A.1.

As an illustrative example, consider estimating the effect of a permanent intervention policy against the control policy. These policies are simpler to evaluate since these policies are stationary and are not state-dependent. By assumption, the experiment allocates both treatment and control for each state. Hence we can simply estimate state *value* functions separately for each of the treatment and control arms and difference the two making sure to account for any initial imbalance in the treatment assignment.

2.1.1 Stationary Policies. Theorem 1 is challenging to use directly since the Q function in Equation (2.8) 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 can state a practical version of Theorem 2 which uses a stationary and hence more tractable Q function.

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:*

$$\bar{\pi}(a|s) = \frac{\rho_{\pi,\gamma}(s,a)}{\rho_{\pi,\gamma}(s)}, \tag{2.9}$$

then

$$\rho_{\pi,\gamma} = \rho_{\bar{\pi},\gamma}. \tag{2.10}$$

See Bertsekas [3] 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:

$$\bar{\pi}^T(s,a) = \frac{\rho_{\pi^T,\gamma}(s,a)}{\rho_{\pi^T,\gamma}(s)} \tag{2.11}$$

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:

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

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 (2.11) 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.

$$q^T(s,a) \equiv \mathbb{E}_y\left[y|s,a\right] + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a),a'\sim\bar{\pi}^T(\cdot|s')}\left[q^T(s',a')\right] \tag{2.12}$$

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

$$\varphi^T = (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot),a\sim\bar{\pi}^T(\cdot|s)}\left[q^T(s,a) - q^0(s,0)\right] \tag{2.13}$$

2.2 Comparison to Surrogate Index Method

Similar to the derivations in the Athey et al. [1], we can pinpoint the key difference between the method above and the surrogate index method. 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 can focus on the difference in potential outcomes for a permanent treatment policy for periods past t since a simple 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

$$\mathbb{E}\left[Y_{t+k}(1)\right] = \int_{s_t,s,y} yp(y \mid s, a=1)p(s \mid s_t, a=1; k)p_t(s_t|a=1), \tag{2.14}$$

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.

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

[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.

$$= \int_{s_t, s_{t+k}, y_{t+k}} y_{t+k} p^o(y_{t+k} \mid s_{t+k}, a=0) p^o(s_{t+k}|s_t, a=0) p_t(s_t|a=1) \tag{2.15}$$

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

Equation (2.15) 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, $p(s_{t+k}|s_t, a=1) \neq p(s_{t+k}|s_t, a=1)$ 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. To accurately estimate the effects of longer-term treatments, another method is needed.

3 ESTIMATION

We want to estimate the long term average treatment effect via the Q function in Equation (2.13). With discrete states, we can solve for Q exactly via dynamic programming methods subject to computational constraints [13]. 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 [5]. Hence we develop a Double ML based estimator centered around the efficient influence function [8]. 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.

$$\phi^\pi(s, a, y, s'; q, w) = -\varphi^\pi + (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot)}\left[q^\pi(s, a)\right] + \frac{\pi(a|s)}{p_0(a|s)} w(s) \left(y + \gamma \sum_{a' \in \mathcal{A}} \pi(a'|s') q^\pi(s', a') - q^\pi(s, a) \right) \tag{3.16}$$

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

[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.

$$\phi(s, a, y, s'; q, w, \pi, \pi^0) = \phi^\pi(s, a, y, s'; q, w) - \phi^{\pi^0}(s, a, y, s'; q, w) \tag{3.17}$$

The density ratio function is defined as

$$w(s) \equiv \frac{\rho_{\pi,\gamma}(s)}{p_0(s)}. \tag{3.18}$$

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

$$\varphi^\pi_{BC} \equiv (1-\gamma)\mathbb{E}_{s\sim p_0(\cdot), a\sim\pi(\cdot|s)}\left[q^\pi(s, a) - q^{\pi^0}(s, 0)\right] + \mathbb{E}\left[\phi(s, a, y, s'; q, w, \pi, \pi^0)\right] \tag{3.19}$$

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 [9].

3.2 Asymptotic Properties

By design, we specified the long term 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 and Uehara [8] who propose efficient, doubly robust estimators for off-policy evaluation using semiparametric methods. This section summarizes the relevant results.

Without making distributional assumptions, it is possible to show that 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.

The second order remainder term is:

$$R = \hat{\varphi}^\pi - \varphi^\pi + P\hat{\phi}^\pi$$

where P indicates the true underlying probability distribution.

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. Intuitively, if the estimate of the Q function is correct, $\hat{q} = q$, then $R = 0$ since simple algrebra shows that $P\hat{\phi}^\pi(s, y, s', x, a'; q, \hat{w}) = -\hat{\varphi}^\pi + \varphi^\pi$. On the other hand, when only the density ratio functions are correct, Lemma 1 (a characterization of $w(s)$) is required to show that $R = 0$ when $\hat{w} = w$.

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. The assumption on the propensity score holds naturally for experiments where treatment assignment is controlled. 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

The Q function is central to the long term average treatment effect. 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 computationally infeasible given the need to construct $\mathbb{S} \times \mathbb{S}$ transition probabilities), 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, without the need for long term sequences of rewards.

Moreover, TD methods are practical since we do not need to construct $\mathbb{S} \times \mathbb{S}$ transition probabilities and are free to use a large state-space, which helps to ensure the Markov property holds. The TD algorithm requires a dataset of $\left(s_i, a_i, y_i, s_i'\right)$ 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 (2.12), we can form the TD error term

$$L_Q(s, a, y, s') = y + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}\left[q^{\pi}(s', a'; \theta_q\right] - q^{\pi}(s, a; \theta_q) \quad (3.20)$$

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 [13]. Framing the TD error in Equation 3.20 as an estimating equation, one can use techniques from M-estimation to derive asymptotic properties such as asymptotic normality and consistency. For example, Kallus and Uehara [8] derive the asymptotic lower bound for an M-estimator that seeks to minimize a weighted form of Equation (3.20).

3.4 Density Ratio Estimation

Unlike Q functions which have a long history in reinforcement learning, the practical utility of density ratio functions is newer, finding recent use in methods for efficient off-policy evaluation [8, 10, 14].

Estimating density ratio functions has centered on the following relationship

$$L_W(f, w) = \mathbb{E}_{s,a,y,s' \sim p_0, a' \sim \pi(\cdot|s')}\left[w(s, a)\left(\gamma f(s', a') - f(s, a)\right)\right] - (1 - \gamma)\mathbb{E}_{s \sim p_0, a \sim \pi(\cdot|s)}\left[f(s, a)\right], \quad (3.21)$$

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

Uehara et al. [14] show that if $L_W(f, \hat{w}) = 0$ for all square-integrable f, then $\hat{w} = w$. Moreover, the reverse also holds under some mild technical conditions, 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.

$$\hat{w}(s, a) = \arg\min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} L_W(f, w)^2 \quad (3.22)$$

Loosely speaking, finding a $\hat{w}$ that sets the Minimax objective close to 0 guarantees that $\hat{w} \approx w$ since the inner maximization bounds the error across all $f \in \mathcal{F}$ [14].

Minimax estimators can be challenging to implement due to the inner maximization. Fortunately, the Minimax objective can be reduced to a simpler form in two cases [14]. First, if the function classes of the density ratio functions and the discriminator are linear under the same feature maps for state and actions, then there exists a closed form solution for w that sets $L_W(w, f) = 0 \ \forall f \in \mathcal{F}$. Second, if the discriminator function class corresponds to a reproducing kernel Hilbert space, then the inner maximization over $\mathcal{F}$ has a closed form solution which reduces the search space to functions within $\mathcal{W}$.

4 EXPERIMENTS

To demonstrate the effectiveness of the ORL method, we perform experiments on simulated data. In particular, we compare the method against a baseline surrogate index method in estimating a range of known long-term ATEs, each associated with a different treatment duration regime. Code and data for the experiments is available at: https://github.com/allentran/long-term-ate-orl.

The results confirm that the surrogate index method recovers the true long-term ATE, albeit only for a treatment of duration of a single period. When applied to data from experiments with treatment durations longer than a single period, the estimate remains static and hence its bias grows with the underlying treatment duration. On the other hand, the ORL method produces estimates that match the true ATE, growing as the duration of treatment extends from a single period to permanent.

4.1 Simulation Details

We generate ground truth data for the experiments with features that mirror those typically found within personalization algorithm tests for a global streaming video on demand service. In particular, we construct a single state Markov chain that mimics the dynamics of on-service tenure. Over time, on-service tenure increases if members do not churn, which we model as a drift-diffusion process with positive drift:

$$s_t = s_{t-1} + \mu + \sigma_s w_t \quad (4.23)$$

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

To reflect that longer tenured members often have higher per period revenue, we set the reward to be a diminishing scalar multiple of the state.

$$y_t = \alpha s_t^{\theta} + \sigma_r e_t \quad (4.24)$$

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

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

We assume that treatment effects both increase the probability of a transition to a higher state and increase average per-period rewards conditional on the state. These are implemented as positive constants in Equations (4.23) and (4.24) when the treatment is active in that period. This mirrors the decomposition of long term treatment effects in Equation (2.14). Table 1 lists the parameter values used in the experiments. For simplicity, we assume the experiment lasts a single period, with all variables observed in the first period as well as the transition to the second period state.

4.2 Model Implementation Details

The baseline surrogate index method estimates the potential outcome means for the first period as empirical means for each of treatment and control. To project long term outcomes, we first construct a synthetic long term observational dataset by simulating trajectories over a long horizon where treatment effects are set to zero. We then fit a regression model using the Random Forests algorithm where the input features is just the surrogate, the state from the second period, and the target is the normalized discounted sum of outcomes from the second period onwards.[4] The long-term estimand is obtained by combining the two by adding the first period potential outcome means to the discounted sum of outcomes from the second period onwards, discounted and normalized appropriately.

To demonstrate the effectiveness of the surrogate index method when its assumptions hold, we estimate a variant where the surrogacy assumption holds even with a long treatment: the surrogate mediates the long-term treatment effect. To do, we must include data from the experiment up to the last treatment period. In doing so, we effectively create a intermediate-term experiment, which moves us away from the central constraint of the paper, that of a *short* experiment.[5]

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 a single hidden layer with 48 features and a sigmoid activation function and a linear final layer with no activation function. Additionally, we maintain separate "target" networks by freezing the parameters of each network for 32 epochs, which proved invaluable in stabilizing training [11].

For the density ratio functions, we use the Minimax weight estimator from Uehara et al. [14] where we restrict both the discriminator and density ratio function classes to be linear with the feature maps $\phi(s) = \begin{bmatrix} s & s^2 & 1 \end{bmatrix}$.

4.3 Results

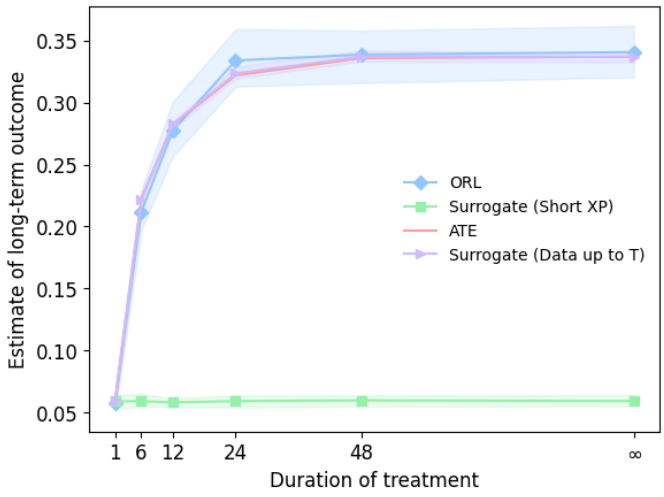

Figure 3: Estimates of the ATE for differing treatment durations

Figure 3 shows the results of simulations of A/B tests where the treatment arm has treatment active for the first 1, 6, 12, 24, 48 and ∞ periods and no treatment thereafter. From the entire history of the each experiment, we calculate the true long-term ATE. In contrast, to mimic real-world scenarios, the estimators evaluated only see a fraction of the underlying data from the experiment.

Since the treatment increases both the state-outcome mapping and the likelihood of a positive state transition, the true ATE (red) increases with the duration of the treatment. Estimates from the ORL method (blue) match the true ATE, as the duration of treatment extends all the way to a permanent intervention.

As Section 2.2 previewed, the surrogate index method estimates long term outcomes as if treatment is applied up to the point at which the last surrogate is measured regardless of the underlying treatment duration. Hence the baseline method (green), where the surrogate is the second period state, estimates the long-term ATE for when treatment is active only for a single period. The estimate is constant and therefore underestimates the true ATE for $T > 1$.

On the other hand, if one is explicit about the treatment duration of interest and measures surrogates up to that point, the intermediate-term surrogate method (purple) matches the true ATE. While both methods match the true ATE, their requirements on measurement are vastly different. The key benefit of the ORL method is that one can estimate the long term ATE from an arbitrary duration treatment regime with a short experiment. With the ORL method, the requirements on the duration of measurement remain static

[4]We use the defaults from the scikit-learn package. Although we could tune hyperparameters, the estimates using the defaults match the single period treatment ATE almost perfectly.

[5]Note that the intermediate-term experiment is not the same as observing the entire experiment since the surrogate method only sees data up to the last treatment period.

independent of the treatment duration of interest. However, one downside is that the variance of the ORL estimates appear to be larger than the surrogate methods.

5 CONCLUSION

We develop a method for inferring the long-term 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. Both of these features mean that surrogate methods are unsuitable since surrogacy assumptions do not hold. Instead, we proceed by making a connection to reinforcement learning and embed our problem within a Markov decision process.

By doing so, we frame 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, while having a different data-generating policy. We construct stationary policies equivalent to arbitrary-duration treatment regimes and hence can 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.

We demonstrate the effectiveness of the ORL method with experiments based on simulated data. Estimates from the ORL method match the true long-term ATE for the full spectrum of treatment durations, from single-period to permanent, using only two periods of data from the an experiment. In contrast, the surrogate index method matches the true ATE only when surrogates are observed up to or beyond the point where the treatment is last active. Hence for long-term treatments, the measurement requirements of the surrogate index method are severe.

In conclusion, our proposed method provides a robust and efficient solution for estimating the long-term ATE of continual exposure to a long-term treatment, when only short-term experimental data are available. The approach can be seen as a complement to surrogate methods, which are well suited to short-term treatments. Our method allows for the estimation of long-term effects 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.

REFERENCES

[1] Susan Athey, Raj Chetty, Guido W. Imbens, and Hyunseung Kang. 2019. *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. https://ideas.repec.org/p/nbr/nberwo/26463.html

[2] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Miruna Oprescu, and Vasilis Syrgkanis. 2021. Estimating the Long-Term Effects of Novel Treatments. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 2925–2935. https://proceedings.neurips.cc/paper_files/paper/2021/file/16fa2b0294e410b2551c3bf6965c0853-Paper.pdf

[3] D.P. Bertsekas. 2001. *Dynamic Programming and Optimal Control* (2 ed.). Vol. 1 and 2. Athena Scientific.

[4] Hua Chen, Zhi Geng, and Jinzhu Jia. 2007. Criteria for Surrogate End Points. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69, 5 (2007), 919–932. http://www.jstor.org/stable/4623303

[5] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (01 2018), C1–C68. https://doi.org/10.1111/ectj.12097 arXiv:https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf

[6] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2023. Long-term Causal Inference Under Persistent Confounding via Data Combination. (2023). arXiv:stat.ME/2202.07234

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

[8] Nathan Kallus and Masatoshi Uehara. 2022. Efficiently Breaking the Curse of Horizon in Off-Policy Evaluation with Double Reinforcement Learning. *Operations Research* 70, 6 (2022), 3282–3302. https://doi.org/10.1287/opre.2021.2249

[9] Edward H. Kennedy. 2023. Semiparametric doubly robust targeted double machine learning: a review. (2023). arXiv:stat.ME/2203.06469

[10] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. 2018. Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/dda04f9d634145a9c68d5dfe53b21272-Paper.pdf

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (Feb. 2015), 529–533. http://dx.doi.org/10.1038/nature14236

[12] Ross L. Prentice. 1989. Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine* 8, 4 (1989), 431–440. https://doi.org/10.1002/sim.4780080407 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780080407

[13] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA.

[14] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. 2020. Minimax Weight and Q-Function Learning for off-Policy Evaluation. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20).* JMLR.org, Article 895, 10 pages.

[15] Tyler J. VanderWeele. 2013. Surrogate Measures and Consistent Surrogates. *Biometrics* 69, 3 (2013), 561–569. http://www.jstor.org/stable/24538119

[16] Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. 2023. 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:

$$\begin{aligned}
\mathbb{E}\left[Y_t(\pi^T)|S_0, A_0\right] &= \mathbb{E}_Y\left[Y_t(\pi^T) \mid S_0, A_0\right] \\
&= \mathbb{E}_Y\left[Y_t(\pi^T) \mid \pi^T, S_0, A_0\right] \\
&= \mathbb{E}_Y\left[Y_t \mid A_t = 1_{t<T}, \pi^T, S_0, A_0\right] \\
&= \mathbb{E}_{S_t}\left[\mathbb{E}_Y\left[Y_t \mid A_t = 1_{t<T}, S_t\right] |\pi^T, S_0, A_0\right] \\
&= \mathbb{E}_S\left[\mathbb{E}_Y\left[Y \mid A = 1_{t<T}, S\right] \mid \pi^T, S_0, A_0; t\right]
\end{aligned} \tag{A.25}$$

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

$$\begin{aligned}
\mathbb{E}\left[\sum_t \gamma^t Y_t(\pi^T)|S_0, A_0\right] &= \mathbb{E}_Y\left[Y \mid A_0, S_0\right] + \\
&\sum_{t=1} \gamma^t \sum_S \mathbb{E}_Y\left[Y \mid A = 1_{t<T}, S\right] p(S|\pi^T, S_0, A_0; t).
\end{aligned} \tag{A.26}$$

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 [13].