Stateful Offline Contextual Policy Evaluation and Learning

Nathan Kallus
Cornell University and Cornell Tech
kallus@cornell.edu

Angela Zhou*
UC Berkeley
angela-zhou@berkeley.edu

Abstract

We study off-policy evaluation and learning from sequential data in a structured class of Markov decision processes that arise from repeated interactions with an exogenous sequence of arrivals with contexts, which generate unknown individual-level responses to agent actions that induce known transitions. This is a relevant model, for example, for dynamic personalized pricing and other operations management problems in the presence of potentially high-dimensional user types. The individual-level response is not causally affected by the state variable. In this setting, we adapt doubly-robust estimation in the single-timestep setting to the sequential setting so that a state-dependent policy can be learned even from a single timestep's worth of data. We introduce a *marginal MDP* model and study an algorithm for off-policy learning, which can be viewed as fitted value iteration in the marginal MDP. We also provide structural results on when errors in the response model leads to the persistence, rather than attenuation, of error over time. In simulations, we show that the advantages of doubly-robust estimation in the single time-step setting, via unbiased and lower-variance estimation, can directly translate to improved out-of-sample policy performance. This structure-specific analysis sheds light on the underlying structure on a class of problems, operations research/management problems, often heralded as a real-world domain for offline RL, which are in fact qualitatively easier.

1 Introduction

Offline reinforcement learning seeks to reuse existing data to evaluate and learn novel policies and is crucial in applications with limited freedom to experiment but plentiful logged data. In general Markov decision processes (MDPs), offline reinforcement learning can be very difficult, as we must understand the effect of actions in each state and time, whether in model-based (e.g., learn the transition kernel) or model-free methods (e.g., learn Q-functions). However, many practically-relevant problems fit in simpler, more tractable classes of MDPs with "sequential decision-making" but not "longitudinal data", for example because transitions arise in a stochastic system from exogenous arrivals. In this paper, we study off-policy evaluation and optimization from observational data in this special class. At each timestep the *same* contextual response model generates both transitions and rewards. The setting is a variant of offline contextual bandits with constraints, where the same randomness generates transitions in the system state (status of the constraints) and rewards in the system. We call this setting, common in operations management, "stateful" to emphasize the well-understood and usually simple system state, like inventory state, in contrast to the unknown potentially high-dimensional context-dependent response model, like an individual's propensity to purchase, that must be learned.

We first describe some stylized examples to illustrate how previously studied classical problems in fact share this broader structure: consider personalized dynamic pricing with inventory constraints,

Offline Reinforcement Learning Workshop at Neural Information Processing Systems, 2021

or managing a rideshare system and repositioning vehicles by making price offers to individuals. The system state includes capacities of each resource, or locations of cars in the system. Individuals with contexts (covariates) arrive exogenously. The system takes actions, such as personalized price or trip offers. Given a context and action, the individual response generates changes in system dynamics: the purchase of a product consumes resources, or accepting a price offer and ride from one location to another moves cars. But given that we can offer the resource at all, the state of the system does not further affect the response except by affecting our pricing decisions.

We focus on the evaluation and optimization of state-dependent policies from offline trajectories collected from these system dynamics. Confounding is of particular concern in such observational data. Naturally, system actions can be spuriously correlated with outcomes. For example, we expect observational operational data to bias toward higher-revenue actions: higher price offers are made to individuals deemed more likely to accept. However, in our setting, the underlying system state does not causally affect the individual response and therefore, unlike individual-level covariates, the system state is not a confounder, making it easier to learn the response model from observational data, and then use it to design sequential policies.

The contributions of this paper are as follows. We model the above structure and specialize fitted-value evaluation and iteration. We establish the sample complexity: because the special structure allows data pooling and a reduction to analysis of tabular MDPs, $\widetilde{O}\left(T^3/\epsilon^2\right)$ trajectories are required for off-policy optimization to achieve ϵ-suboptimal value, where T is the horizon (omitting logarithmic terms). Analyzing the problem structure, we show that bias from naive model-based approaches would generally persist when historical confounding is correlated with higher rewards. In the dynamic and capacitated pricing example, optimizing over threshold policies with our doubly-robust estimation approach improves upon naive model-based approaches. The key advantage of specializing to problem structure is reducing the number of required nuisance functions from $T+1$ to simply *two* (propensities and outcome), which we show empirically leads to finite-sample improvements while retaining standard improvements from doubly-robust estimators.

2 Problem setup: Stateful off-policy evaluation and learning

We first describe the generic full-information MDP that generates our data before describing the restrictions that characterize the stateful setting. For ease of reference, we partition the state space of the full-information MDP into a product space of the *discrete system state space* $\mathcal{S}$, potentially continuous *context/covariate space* $\mathcal{X}$, and discrete covariate-conditional response space $\mathcal{Y}$: $\mathcal{S} \times \mathcal{X} \times \mathcal{Y}$. The inclusion of Y in the state variable is purely informational; below we specify that it is the random variable that generates system transitions. Consider a finite-horizon setting with $T+1$ timesteps, and denote the initial system state s_0; timesteps are indexed $0, \ldots, T$. Uppercase (S, X) indicates random variables; lowercase (s, x) indicates fixed values; and prime (s', x') indicates next-timestep values. Let $\mathcal{A}(s, x)$ denote the discrete action space feasible from the state (s, x). A contextual policy $\pi_t \colon \mathcal{S} \times \mathcal{X} \mapsto \Delta^{\mathcal{A}}$ maps from system state/context to a distribution over actions, where $\Delta^{\mathcal{A}}$ is the set of distributions defined on $\mathcal{A}$, so that $\pi_t(a \mid s, x) = \mathbb{P}(A = a \mid S = s, X = x)$ gives the probability of taking action a given state and context information. Let $\pi = \{\pi_t\}_{t=0,\ldots,T}$ denote the MDP policy that is specified in a function class $\Pi_{0:T}$. Reward is a known deterministic function of next state transition, $R(s, a, s')$.

We next specify the restrictions on this MDP that give rise to our "stateful" setting. These are illustrated in Figure 1. Roughly: contexts arrive exogenously and contextual responses Y come from a stationary conditional distribution $\mathbb{P}(Y \mid X, A)$ and deterministically generate the system transitions. We henceforth use the shorthand $P(S' = s', X' = x', Y' = y \mid S = s, X = x, Y = y_{-1}, A = a) = P(s', x', y \mid s, x, y_{-1}, a)$ for the transition model.

The following assumptions characterize the "stateful" setting and identify general structure that appears commonly in more specialized problem contexts elsewhere. We state them as assumptions for ease of reference.

Assumption 1 (Exogenous context process)**.** The transition factorizes as

$$P(s', x', y \mid s, x, y_{-1}, a) = P(s', y \mid s, x, a) f(x') \;\; \forall s, s', x, x', y_{-1}, y.a$$

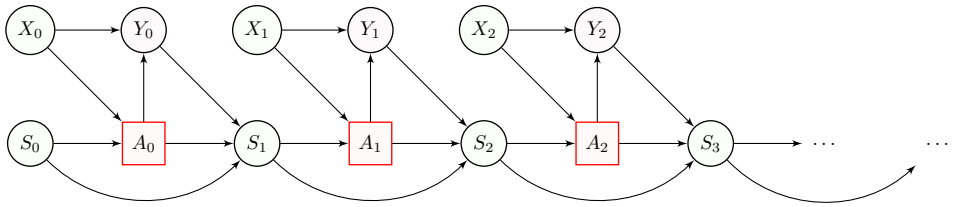


Figure 1: The "stateful" decision model we consider; rewards are functions of S_t, A_t, S_{t+1}.

Assumption 2 (Contextual-response transitions)**.** We know $s'(s,y) : \mathcal{S} \times \mathcal{Y} \mapsto \mathcal{S}$ such that when s' is not absorbing from s we have:

$$P(s', y \mid s, x, a) = \delta_{s'-s'(s,y)} \mathbb{P}(Y = y \mid x, a)$$

We can easily extend to random transitions given responses, but focus on deterministic for concreteness and as it captures the most relevant application settings. Assumption 1 arises from contextual bandits or uniformizing (with contexts) a stochastic system [20, 31]. Assumption 2 reflects the offline contextual bandit nature of the problem and encodes that Y_t is independent of the originating state.[1] [17] leverages a factorization with exogenous information, but not a contextual response model and notes that the "system transition function" construction is the norm in control/operations research [6, 38].

For ease of presentation we introduce $\{s', y \mid s, a\}$ as the pairs of next states and contextual responses reachable from s, a.

Definition 1 (Reachable state transition-potential outcomes)**.**

$$\{(s', y) \mid s\} := \{(s', y) \in \mathcal{S} \times \mathcal{Y} \colon \exists\, x, a \text{ s.t. } P(s', y \mid s, x, a) > 0\}$$

The corresponding full-information MDP is $\mathcal{M} = (\mathcal{S} \times \mathcal{X} \times \mathcal{Y}, \mathcal{A}, P, R, T)$ where P is the full-information transition kernel. Our *observational* data comprises of N trajectories; denote individual observations as S_t^i for t timestep of trajectory i: $\{S_t^i, X_t^i, A_t^i, Y_t^i, R_t^i\}_{i=1,\ldots,N; t=0,\ldots,T}$.

Without loss of generality we omit the Y_{t-1} information state from the full-information *value/reward to go* V- and state-action value Q-function, since $V_t(s, x, y_{-1}) = V_t(s, x)$:

$$V_t^\pi(s,x) = \mathbb{E}_\pi \left[\textstyle\sum_{t'=t}^T R_{t'} \mid S_t = s, X_t = x\right], \quad Q_t^\pi(s,x,a) = \mathbb{E}\left[\textstyle\sum_{t'=t}^T R_{t'} \mid S_t = s, X_t = x, A_t = a\right]$$

It is useful to define the *context-marginalized* value function $\tilde{V}_t^\pi(s) = \mathbb{E}[V^\pi(s, X)]$, the value function at system state S_t marginalized over context distribution X_t, and analogously $\tilde{Q}$. Under assumptions 1 and 2 and the notation of defn. 1, the Q-function in the full-information MDP is:

$$Q_t^\pi(s,x,a) = \textstyle\sum_{(s',y)|s} \mathbb{P}(Y = y \mid x, a)(R(s,a,s') + \tilde{V}_{t+1}^\pi(s')) \tag{1}$$

Examples of stateful problems. We discuss illustrative examples. The first example, single-item personalized and dynamic pricing with inventory constraints, is adapted from classical models for network revenue management ([20, 19]).[2]

Example 1 (Single item personalized and dynamic pricing)**.** $Y_t \in \{0, 1\}$ is purchase/no-purchase, respectively, and $A_t \in \{0, 1\}$ is whether a discount of value d is not or is offered. Let $p(a)$ be the price corresponding to taking action $A = a$. The reward is fixed given transition to s': $R(s, a, s') = p(a)\mathbb{I}[y = 1]\mathbb{I}[s > 0]$. For short let $R(a)$ denote price of product under a, e.g. reward only received if item is sold, and we can only sell if we have stock so

[1] This assumption would not be true in settings where customers had full observation of the system and responded to it, e.g. in queuing if customers can observe queue length and balk. Or, if customer arrivals are correlated with system state due to unobserved confounders, such as weather patterns that lead to higher propensity to accept a ride and higher customer demand at other locations.

[2] Instead of assuming arrivals would deterministically purchase and setting the control lever to be fare availability, we consider a stochastic demand response model.

$s'(s,y) = \mathbb{I}[s>0, y=1](s-1) + \mathbb{I}[s>0, y=0]s$. Denote the difference of value functions as $\Delta V_t^\pi(s) = \tilde{V}_t^\pi(s-1) - \tilde{V}_t^\pi(s)$, then the full-information Q_t function is 0 for $t = T$, $\mathbb{P}(Y = y \mid x, a)R(a)\mathbb{I}[s>0]$ for $t = T-1$, and for $t < T-1$: $Q_t(s,x,a) = \mathbb{P}(Y = y \mid x, a)(R(a)\mathbb{I}[s>0] + \Delta V_{t+1}^\pi(s)) + \tilde{V}_{t+1}^\pi(s)$.

Such structure also appears in *multi-item network revenue management* and *spatial pricing and repositioning*; we defer these descriptions to the appendix.

3 Related work

Off-policy policy learning for offline sequential decision-making. There has been extensive work on off-policy evaluation and learning in the sequential setting. We focus on work that builds on statistical model-free approaches, including doubly robust off-policy evaluation in incorporating value-function control variates [43, 24, 50], and study of the efficient influence function that specifically uses Markovian structure to obtain efficiency bounds for evaluation [26, 8, 27], as well as MIS or fitted-Q-evaluation [49, 15, 30, 21].

Our estimator does not require rejection sampling on entire trajectories. Therefore we show statefulness is in fact more closely related to single-timestep off-policy evaluation and learning [16, 29, 42, 46]; we are able to pool available data across states and timesteps. We do not claim novelty relative to the extensively-studied doubly-robust estimation in sequential OPE: rather we study the practical fact that specializing to structure admits estimation with 2 nuisance functions rather than $T+1$ nuisance functions.

Online contextual decision-making with constraints and algorithmic analysis under known distributions. In the main text we provide an abridged discussion highlighting contextual or stateful problems; see Section C.1. Contextual bandits with knapsack (CBwK [4]) does consider both contexts and statefulness, relative to an extensive literature (typically model-based) on *either* contextual [14, 23, 39, 41, 5, 12] or stateful [22, 7, 1] problems .[3] The closest work is [2], but it considers the Lagrangian relaxation of the resource constraints. Relative to CBwK and pricing bandits, we consider a general MDP embedding and our sample complexity analysis and algorithm do not require specific structure of the reward beyond assumptions 1 and 2.

4 Off-policy evaluation and learning in the *Marginal MDP*

Marginal MDP construction. Section 2 described the generating process of the data. We now marginalize over contexts and (policy-induced) outcomes in a lifted *marginal MDP* on a discrete state space and continuous action space where actions are given by policy parameters. This MDP is purely a *conceptual device* which is used in the analysis. Direct OPE methods cannot be used in the marginal MDP because observations in the dataset correspond to variation over different *actions*, but not necessarily different *policies* that are actions in the marginal MDP.

The marginal MDP state space is the system state space, $\mathcal{S}$,

action space is the set of parametrized policies, $\mathcal{A}(s) = \Pi(s)$, where $\Pi(s) = \{\pi(s,\cdot) \in \Pi\}$ is the set of policy functions given s,

transitions between s and $s'(s,y)$ occur with probability $P(Y = y \mid \pi)$, (eq. (2)), which is the *single-timestep batch policy value*.

It is defined with respect to the single-timestep expectation as: $P(Y = y \mid \pi)$ which provides the transition probabilities in the marginal MDP:

$$P(Y = y \mid \pi) = \textstyle\sum_{a \in \mathcal{A}} \mathbb{E}[\pi_e(a \mid X)\mathbb{P}(Y = y \mid a, X)] \tag{2}$$

and reward is $\tilde{R}(s,\pi) = \sum_a \mathbb{E}_X[\int_y \pi(a \mid X)R(s,a,s'(s,y))dP(y \mid X, a)]$, the expected reward induced by context-conditional policy actions and corresponding outcomes.

The marginal MDP is $\widetilde{\mathcal{M}} = (\mathcal{S}, \Pi, \tilde{P}, \tilde{R}, T)$. The key modeling insight is that expectations over individual exogenous arrivals may be estimated via a distribution of iid arrivals. It is immediate to

[3] We discuss CBwK for a full discussion of related work. But while our framework can readily handle unknown or multiple behavior policies, we do not consider data directly collected from a bandit algorithm (i.e. outcome-adapted data subject to adaptive sequential learning bias).

verify equivalence of policy values and optimal policies in the marginal and full-information MDPs (under the same policy-restricted class). (Note that higher-order moments are not equivalent.)

Proposition 1. Assume the policy class Π is a product space over $s \in \mathcal{S}, t \in [T]$. The marginal MDP $\widetilde{\mathcal{M}} = (\mathcal{S}, \Pi, \tilde{P}, \tilde{R}, T)$ has the same optimal policy, and policy value for tabular states, as the full-information MDP with policy class Π, $\mathcal{M} = (\mathcal{S} \times \mathcal{X} \times \mathcal{Y}, \mathcal{A}, P, R, T)$,.

Example: Marginal value function for single-item pricing. In the marginal MDP counterpart of Example 1, $\tilde{P}(s-1 \mid s, \pi) = P(Y = 1 \mid \pi)$, $s'(s, y) = \mathbb{I}[s > 0]\mathbb{I}[y = 1](s-1)$, and for $t < T-1$:

$$\tilde{V}_t^{\pi_{t:T}}(s) = \tilde{R}(s, \pi) + \tilde{V}_{t+1}^{\pi_{t+1:T}}(s) + P(Y = y \mid \pi)\Delta V_{t+1}^{\pi_{t+1:T}}(s)$$

Estimation via fitted value evaluation and iteration in the marginal MDP. We define the propensity score and outcome model as follows:

$$e_t(a, x) = \mathbb{P}(A_t = a \mid X = x), \qquad \mu(y \mid a, x) = \mathbb{P}(Y = y \mid A = a, X = x).$$

The propensity score only controls for X: while we allow the underlying behavior policy to be state-dependent, Assumption 2 implies that adjusting for X is sufficient to estimate the marginalized transition, eq. (2), because the state does not affect the outcome. To achieve the orthogonality and rate double-robustness benefits of the doubly-robust estimator we next introduce, we use two-fold sample splitting in trajectories and timesteps. We use cross-time fitting and introduce folds that partition trajectories and timesteps $k(i, t)$. For $K = 2$ we consider timesteps interleaved by parity (e.g. odd/even timesteps in the same fold). We let $k(i, t)$ denote that nuisance $\hat{\mu}^{-k(i,t)}$ is learned from $\{X_{t'}^{(i)}, Y_{t'}^{(i)}\}_{i \in \mathcal{I}_{k(i)}, t' \bmod 2 = t \bmod 2}$, e.g. from the $-k(i)$ trajectories and from timesteps of the same evenness or oddness but is only used for evaluation in the other fold. Interleaving between timesteps insures downstream policy evaluation errors are independent of errors in nuisance evaluation at time t.

We let $\hat{P}(Y = y \mid \pi)$ denote the empirical estimate: we verify that the standard doubly robust estimator, reweighting the empirical transitions in observational data, for single-timestep offline policy learning estimates the transition probabilities in the marginal MDP.

Proposition 2 (Single-time-step doubly robust estimator of transitions in the marginal MDP.)**.** Let

$$\Gamma_t^i(y \mid a) := \frac{\mathbb{I}[Y_t^i = y] - \hat{\mu}^{-k(i,t)}(y \mid A_t^i, X_t^i)}{\hat{e}_t^{-k(i,t)}\left(A_t^i, X_t^i\right)}\mathbb{I}[A_t^i = a] + \hat{\mu}^{-k(i,t)}(y \mid a, X_t^i)$$

$$\hat{P}(Y = y \mid \pi) := (NT)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{a \in \mathcal{A}} \pi(a \mid X_t^i)\Gamma_t^i(y \mid a).$$

$\hat{P}(Y = y \mid \pi)$ is an unbiased estimator of $P(Y = y \mid \pi)$ if at least one of $\hat{\mu}$ or $\hat{e}$ are unbiased.

Proposition 2 verifies *orthogonality*, that the estimator is doubly-robust against misspecification of one of μ or e. The estimator only adjusts for contexts since Assumption 2 specifies that the state variable does not affect the response (hence is not a confounder). Proposition 2 considers the stationary case; when X_t is time-varying but non-adversarial with fixed distributions, similar data-pooling is possible by estimating density ratios.[4]

Given a generic estimator $\hat{P}(Y = y \mid \pi)$ for the marginal transition probability $P(Y = y \mid \pi)$, we can construct a Q-estimate as follows: $\hat{P}(Y = y \mid \pi)$ can be the doubly robust estimator as in Proposition 2 or alternatively the IPW estimator (simply let $\hat{\mu}^{k(i,t)} = 0$ in Proposition 2) or direct method estimator (simply let $\hat{e}^{k(i,t)} = \infty$ in Proposition 2). We use backwards recursion to evaluate the off-policy value estimate $\widehat{V}_t^{\pi_{t:T}}(s)$ using model-based evaluation with $\hat{P}(Y = y \mid \pi)$ in the marginal MDP.

$$\widehat{Q}_t^{\pi, \pi_{t+1:T}}(s, \pi) = \sum_{(s', y) \mid s} \hat{P}(Y = y \mid \pi)\left(R(s, a, s') + \widehat{V}_{t+1}^{\pi_{t+1:T}}(s')\right). \tag{3}$$

Policy Learning. When the policy space is in fact a product set over the state space (i.e., the policy being optimized can vary independently for every value of the state), we study a policy learning proposal in Algorithm 1 which implements backwards-recursive policy learning to determine the optimal policy vector π.

[4] In revenue-management settings, it is common for X_t arrivals to be nonstationary. While online algorithms considers adversarial arrival distributions, relevant arrivals may also have highly structured nonstationarity, e.g., "business-class" arrivals arriving later on. To a limited extent, *adversarial* arrivals could also be modeled by robustness, e.g., using the approach of [28] over density ratios for each timestep's subproblem in Algorithm 1.

Algorithm 1 Backwards-Recursive Policy Learning

Input: estimate $\hat{P}(Y = y \mid \pi)$, policy class $\Pi_{0:T}$
for $t = T, \ldots, 0$ **do**:
  for $s \in \mathcal{S}$ **do**:
    Estimate off-policy value $\widehat{Q}_t^{\pi, \hat{\pi}^*_{t+1:T}}(s, \pi)$ via eq. (3)
    Optimize $\hat{\pi}^*_{t,s} \in \arg\max_{\pi \in \Pi_t(s)} \widehat{Q}_t^{\pi, \hat{\pi}^*_{t+1:T}}(s, \pi)$ and update $\widehat{V}^{\hat{\pi}^*_{t:T}}(s) \leftarrow \widehat{Q}_t^{\hat{\pi}^*_{t:T}}(s, \hat{\pi}^*_{t,s})$
return $\hat{\pi}^* = \{\hat{\pi}^*_{t,s} \colon t \in [T], s \in \mathcal{S}\}$

5 Analysis

Sample complexity. We first provide a generalization bound for Algorithm 1 on the out-of-sample regret $\tilde{V}_0^{\hat{\pi}^*}$, the true value achieved by the sample-optimal policy $\hat{\pi}^*$, relative to the best-in-class policy, $\tilde{V}_0^{\pi^*}$. We assume the policy class at a given s, t has restricted functional complexity in the sense of a finite entropy integral of the covering numbers [44, 47]. In the main text we state the bound for a binary-action setting with finite VC dimension d_{vc}. In the general multi-class setting we include in the appendix corresponding statements for multi-class notions such as Natarajan dimension [32].

Theorem 1 (Sample complexity and rate double-robustness)**.** *Suppose $\nu^{-1} \leq e(a \mid x) \leq 1 - \nu^{-1}$ uniformly over a, x, for $\nu > 0$, (overlap) and for some rates $0 < r_1, r_2 < 1$ and constants C_1, C_2, we have uniformly consistent estimation of nuisance estimates*

$$\mathbb{E}[(\mu(y \mid a, X) - \hat{\mu}(y \mid a, X))^2] = o_p(n^{-r_1}), \;\; \mathbb{E}[(e(a \mid X) - \hat{e}(a \mid X))^2] = o_p(n^{-r_2}),$$

where $r_1 + r_2 \geq 1$. Then there exists a random variable $\kappa = o_p((NT)^{-\frac{1}{2}})$ so that w.p. $\geq 1 - \delta$,

$$\tilde{V}_0^{\pi^*} - \tilde{V}_0^{\hat{\pi}^*} \leq \frac{{}^{9\nu R^{\max}}/_2(T^{1/2} + {}^1/_2 T^{3/2})\sqrt{d_{\text{vc}} \log({}^{5T|\mathcal{Y}|}/_\delta)}}{\sqrt{N}} + \kappa.$$

The main improvement in Theorem 1 is in specializing to statefulness so only two nuisance functions are required, rather than $T \times |\Pi_{0:T}|$ many as would arise in the case of Q-function nuisances. We also observe that the problem constants are independent of *concentratability coefficients* used in batch RL to quantify the degree of exploration in observational data but are clearly ill-suited for the setting here [33, 3, 34, 35, 30]. We discuss this in Section C with a stateful example where the *uniform* concentratability coefficient is exponential in the horizon, but the stagewise overlap bound is $O(T)$.

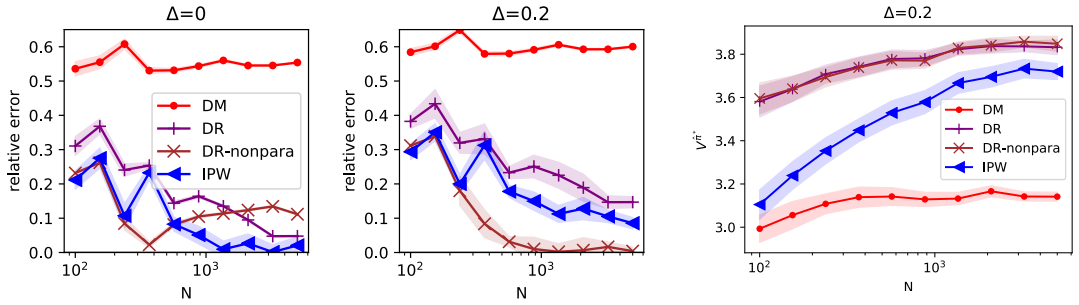
Structural analysis: error propagation in dynamic pricing. We study the possible drawbacks of naive plug-in approaches (DM) by analyzing error propagation in general. We provide structural conditions for error *persistence* in the sequential setting and study the context of practically anticipated demerits of naive model-based approaches with observational data: the combination of misspecification bias and endogeneity. Error could "persist" if model error in transitions and value functions continues to impact downstream learned policies or "attenuate" if these cancel each other out in the sequential setting; especially given the simple next-time-step-dependence on marginal V. We specialize to Example 1; similar results should hold for other settings with less interpretable sufficient conditions. Define the threshold θ^*, true conditional expectation ratio $\Delta\mu^*$, and optimal (threshold) policy π^*:

$$\theta_t^*(s, \Delta V_{t+1}(s)) = \tfrac{R(0) + \Delta V_{t+1}(s)}{R(1) + \Delta V_{t+1}(s)}, \;\; \Delta\mu^*(x) = \tfrac{\mu(1|1,x)}{\mu(1|0,x)}, \;\; \pi^*(1 \mid s, x) = \mathbb{I}\left[\Delta\mu^*(x) > \theta_t^*\right]. \quad (4)$$

The confounded-optimal policy incurs error from $\Delta\hat{\mu}$ or $\Delta V^{\hat{\pi}^*}$. We reparametrize the decision boundary on $\Delta\hat{\mu}$ relative to $\theta^*, \Delta\mu^*$. Then the biased optimal policy is a threshold $\hat{\theta}^*(s, \Delta V)$ with the following relationship to θ^*, where $\delta(a, x)$ is a pointwise error:

$$\delta(a, x) = \hat{\mu}(y \mid a, x) - \mu(y \mid a, x), \;\; \hat{\theta}^* = \theta_t^*(s) \cdot (1 + {}^{\delta(0,x)}/_{\mu(1|0,x)}) - \delta(1, x). \quad (5)$$

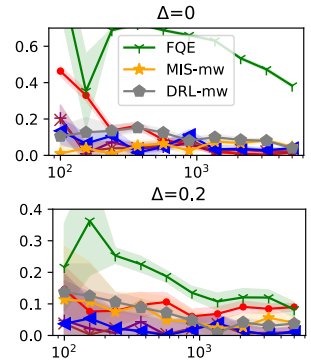
When self-evident we omit dependence of θ^* on arguments that remain fixed for brevity. Error "persists" if the error at different timesteps, including from value estimation, persists in the same direction relative to the optimal policy. We provide a sufficient condition for concluding the direction of error induced only from downstream errors in value estimation. In the main text we state a special case; the appendix includes the full theorem for $t < T - 2$ with less interpretable conditions.



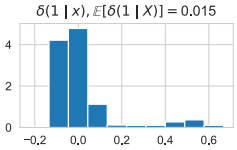
(a) Relative absolute error (y-axis) of off-policy evaluation with increasing model misspecification (Δ). Lower is better.

(b) Out-of-sample value $V^{\hat{\pi}^*}$ (y-axis) of policy optimization: threshold policies optimized via Algorithm 1. Higher is better.

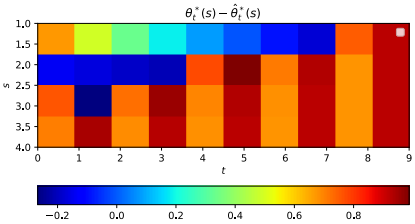
Figure 2: Policy evaluation and optimization as more trajectories (x-axis) are collected of T-horizon selling in contextual and capacitated dynamic pricing, example 1; specification of eq. (6).



(a) $\Delta = 0.2$; further OPE comparison on a different DGP.



(b) (Normalized) histogram of conditional bias $\delta(1, x)$.



(c) Heatmap of thresholds optimized via DR vs. (biased) thresholds, over states (y-axis) and timesteps (x-axis). Red is direction of single-timestep error (persistence).

Figure 3: Conditions of Theorem 2 for error persistence.

Theorem 2 (Conditions for error persistence)**.** *For $t = T - 2$, assume $R(1) > R(0)$, without loss of generality. Then, for $s > 2$,*

$$\mathbb{E}[-\tau(X)\mathbb{I}[\hat{\theta}^*_{T-1} \leq \Delta\mu^*(X) \leq \theta^*_{T-1}]] \geq 0 \implies \hat{\theta}^*_{T-2}(\Delta V^{\hat{\pi}^*_{T-1:T}}_{T-1}) < \hat{\theta}^*_{T-2}(\Delta V^{\pi^*_{T-1:T}}_{T-1}).$$

We discuss implications of Theorem 2 for bias persistence in the context of Example 1.

Example 2 (Error persistence in Example 1)**.** Suppose the cumulative effects of model misspecification and endogeneity is such that $\delta(1, x) > 0 > \delta(0, x)$ uniformly over x, for example if historical price increases were targeted towards those more likely to purchase them and discounts were targeted to those less likely to purchase overall. Then for any $s, \Delta V$, $\hat{\theta}^*(s, \Delta V) \leq \theta^*(s, \Delta V)$. By assumption on, e.g. price elasticity so that $\tau(x) \leq 0$, we expect the sufficient condition of Theorem 2 to be true so that bias persists; for $s > 2$,

$$\hat{\theta}^*_{T-2}(\Delta V^{\hat{\pi}^*_{T-1:T}}_{T-1}) < \hat{\theta}^*_{T-2}(\Delta V^{\pi^*_{T-1:T}}_{T-1}) < \theta^*_{T-2}(\Delta V^{\pi^*_{T-1:T}}_{T-1}).$$

6 Empirics

Data-generating process. We consider a simple example based on single-product dynamic pricing (example 1), with a response model that is a Δ-weighted mixture model of a logistic specification and a nonlinear specification, where $\sigma(\beta^\top x) = (1 + \exp(-\beta^\top x))^{-1}$.

$$\mu(1 \mid a, x) = (1 - \Delta)\sigma(\beta^\top x + \beta_0 \cdot p(a)) + \Delta\sigma(x_0^2 \cdot p(a)),\ e_t(1 \mid x) = \sigma(-0.8\beta^\top x) \quad (6)$$

We generate the data corresponding to the outcome specification for parameters $\beta = [-0.75, 0.75]$, $\beta_0 = -2$. We learn outcome models $\hat{\mu}$ by either logistic regression (for DM, direct method or DR, doubly robust) or a neural network for a nonparametric nuisance estimate (DR-nonpara), and the behavior policy by (well-specified) logistic regression. We consider a (time and state-stationary)

evaluation policy $\pi_e(1 \mid x) = \sigma(0.25(\beta^\top x))$. The time horizon is 10 timesteps, with initial state capacity $s_0 = 4$.

Policy evaluation and optimization. In Figure 2 we generate different outcome models with increasing levels of misspecification Δ, evaluate $V_0^{\pi_e}(s_0)$ by Monte Carlo rollouts with $N = 10000$ trajectories. We compare DM with logistic regression nuisance, DR doubly-robust with logistic regression, DR-nonpara with nonparametric nuisance, and IPW, inverse propensity weighting. (See Section D for further comparison including other baselines and policy optimization in the well-specified $\Delta = 0$ case, where the variance drawbacks of IPW do worse than model-based approaches.) Figure 2a considers off policy evaluation, with absolute relative error on the y-axis and trajectory size on the x-axis (log grid from $N = 50, \dots 5000$). When $\Delta = 0.2$ the logistic outcome model is misspecified, but orthogonality and the well-specified propensity score ensures estimates are asymptotically unbiased. Similar to other DR settings, although incorporating the outcome model reduces variance, incorporating a misspecified outcome model does worse than just using well-specified IPW, but we see faster convergence from the flexible, nonparametric nuisance which outperforms well-specified IPW. We also compared to nonparametric baselines FQE [30], and modified stateful versions of MIS [49] and DRL [26]. However, in this simple setting, the highly flexible nuisance estimators overfit and fail (incurring 40-50% absolute error). We discuss these baselines in greater detail in "OPE comparison" in a more favorable data-generating process.

We then consider policy optimization in Figure 2b, with a rich policy class to avoid misspecification error issues. Motivated by eq. (5), observe that the optimal threshold policy on the true $\Delta\mu$ is an affine transform relative to a threshold on the estimated $\Delta\hat{\mu}$ (possibly misspecified, hence biased), with an x-conditional term for the conditional error. We approximate optimizing over policies $\mathbb{I}[\Delta\mu > \theta]$ by ranging over all thresholds on $\mathbb{I}[\Delta\hat{\mu} > \theta']$; this approximates the x-conditional error term of eq. (5) by a constant. This is similar to a contextual version of "bid-price" policies [20]. We optimize over the class of threshold policies on $\Delta\hat{\mu}$ by enumerating thresholds and evaluating via the estimate from Proposition 2, so the functional specification does depend on the (unadjusted) nuisance estimation. The y-axis depicts out of sample value (higher is better) averaged over 48 replications. Both DR and IPW (inverse propensity-weighted) estimates translate to improvements in optimized policy value. We see dramatic benefits of DR when $\Delta = 0.2$. For small dataset sizes, IPW suffers from high variance as expected. Therefore, DR and its variance reduction/well-scaled estimates achieve sizeable improvements for small amounts of data. As the amount of data grows larger, the performance of IPW nears that of DR asymptotically. The DM plug-in approach remains biased and achieves worse performance, even asymptotically.

OPE comparison. We compare to state-of-the-art OPE: FQE of [30] which does not use the "stateful" structure, and we also derive "strong baselines" that leverage some of the structure (MIS-mw [49], DRL-mw [27]). (We reiterate our core contribution is not in off-policy evaluation but OPE with direct uniform convergence). For example, observe that since x is exogenously generated the finite-horizon *state-action density ratio* is *independent of* x. We endow MIS-mw and DRL-mw with this structural information by restricting functional specification of the density ratio $\mu_t(s, a)$ (see appendix Section D for details). As the general OPE literature prescribes, we use nonparametric nuisances, e.g. multi-layer perceptron with scikit-learn defaults for all nuisance predictors. We consider a more favorable DGP for OPE comparison in Figure 3a, using eq. (6) with $p = 5$ and $\beta = [-0.53, -0.56, -0.10, 0.40, 0.74], \beta_0 = -2.39$. Our ablated derivation of MIS-mw does well overall, although empirically we find other DGPs where MIS-mw underperforms FQE. In the misspecified setting, our doubly robust estimators outperform MIS-mw. FQE, which fits next timestep $Q(s, x, a)$, appears to converge but much slower than our approaches. The gap between FQE and DM for the (slightly misspecified) case precisely illustrates the benefits of encoding problem structure in Equation (1).

Assessing the structural conditions of Theorem 2 in practice. Figure 3b plots a histogram of the error δ: although it is symmetrically distributed for most x, there is overall marginal error in the direction of our anticipated confounding error. In the empirical example we optimize over marginal thresholds and so we expect, marginalizing over x, the directional error condition is satisfied. In Figure 3c, we show a heatmap of $\theta_t^*(s) - \hat{\theta}_t(s)$ over timesteps and state values. As the analysis suggests, for $s > 2$ for most timesteps the error persists: red indicates regions where naive thresholds are in the same direction, relative to the optimal threshold, and hence the error persists rather than attenuates over time.

References

[1] S. Agrawal and R. Jia. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 743–744, 2019.

[2] S. Agrawal, N. R. Devanur, and L. Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18, 2016.

[3] A. Antos, R. Munos, and C. Szepesvari. Fitted q-iteration in continuous action-space mdps. In *Neural Information Processing Systems*, 2007.

[4] A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[5] G.-Y. Ban and N. B. Keskin. Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Forthcoming, Management Science*, 2020.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[7] O. Besbes and A. Zeevi. Blind network revenue management. *Operations research*, 60(6): 1537–1550, 2012.

[8] A. F. Bibaut, I. Malenica, N. Vlassis, and M. J. van der Laan. More efficient off-policy evaluation through regularized targeted learning. *arXiv preprint arXiv:1912.06292*, 2019.

[9] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[10] P. Bumpensanti and H. Wang. A re-solving heuristic with uniformly bounded loss for network revenue management. *Management Science*, 2020.

[11] J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[12] X. Chen, Z. Owen, C. Pixton, and D. Simchi-Levi. A statistical learning approach to personalization in revenue management. *Management Science*, 2021.

[13] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[14] M. C. Cohen, I. Lobel, and R. P. Leme. Feature-based dynamic pricing. *EC*, 2016(10.1145): 2940716–2940728, 2016.

[15] Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[16] M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

[17] I. El Shar and D. Jiang. Lookahead-bounded q-learning. In *International Conference on Machine Learning*, pages 8665–8675. PMLR, 2020.

[18] Y. Emek, R. Lavi, R. Niazadeh, and Y. Shi. Stateful posted pricing with vanishing regret via dynamic deterministic markov decision processes. *Advances in Neural Information Processing Systems*, 33, 2020.

[19] G. Gallego and G. Van Ryzin. A multiproduct dynamic pricing problem and its applications to network yield management. *Operations research*, 45(1):24–41, 1997.

[20] G. Gallego, H. Topaloglu, et al. *Revenue management and pricing analytics*, volume 209. Springer, 2019.

[21] Y. Hu, N. Kallus, and M. Uehara. Fast rates for the regret of offline reinforcement learning. In *Conference on Learning Theory*, 2021.

[22] W. T. Huh, R. Levi, P. Rusmevichientong, and J. B. Orlin. Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research*, 59(4): 929–941, 2011.

[23] A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *arXiv preprint arXiv:1609.07574*, 2016.

[24] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[25] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

[26] N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*, 2019.

[27] N. Kallus and M. Uehara. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019.

[28] N. Kallus and A. Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 2020.

[29] T. Kitagawa and A. Tetenov. Empirical welfare maximization. 2015.

[30] H. Le, C. Voloshin, and Y. Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[31] S. Meyn and S. P. Meyn. *Control techniques for complex networks*. Cambridge University Press, 2008.

[32] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[33] R. Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567, 2003.

[34] R. Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

[35] R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[37] D. Pollard. *Empirical processes: theory and applications*. NSF-CBMS regional conference series in probability and statistics, 1990.

[38] W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2007.

[39] S. Qiang and M. Bayati. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

[40] B. Scherrer. Approximate policy iteration schemes: a comparison. In *International Conference on Machine Learning*, pages 1314–1322. PMLR, 2014.

[41] V. Shah, J. Blanchet, and R. Johari. Semi-parametric dynamic contextual pricing. *arXiv preprint arXiv:1901.02045*, 2019.

[42] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. *Proceedings of NIPS*, 2015.

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

[44] A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[45] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[46] S. Wager and S. Athey. Efficient policy learning. 2017.

[47] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[48] T. Xie, Y. Ma, and Y.-X. Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*, 2019.

[49] M. Yin, Y. Bai, and Y.-X. Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

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

[51] Y.-Q. Zhao, D. Zeng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.

[52] Z. Zhou, S. Athey, and S. Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.