Markovian Interference in Experiments

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We consider experiments in dynamical systems where interventions on some experimental units impact other units through a limiting constraint (such as a limited supply of products). Despite outsize practical importance, the best estimators for this 'Markovian' interference problem are largely heuristic in nature, and their bias is not well understood. We formalize the problem of inference in such experiments as one of policy evaluation. Off-policy estimators, while unbiased, apparently incur a large penalty in variance relative to state-of-the-art heuristics. We introduce an on-policy estimator: the Differences-In-Q's (DQ) estimator. We show that the DQ estimator can in general have exponentially smaller variance than off-policy evaluation. At the same time, its bias is second order in the impact of the intervention. This yields a striking bias-variance tradeoff so that the DQ estimator effectively dominates state-of-the-art alternatives. Our empirical evaluation includes a set of experiments on a city-scale ride-hailing simulator.

1. Introduction

Experimentation is a broadly deployed learning tool in online commerce that is, in principle, simple: apply the treatment in question at random (e.g. an A/B test), and 'naively' infer the treatment effect by differencing the average outcomes under treatment and control. About a decade ago, Blake and Coey [8] pointed out a challenge in such experimentation on Ebay:

"Consider the example of testing a new search engine ranking algorithm which steers test buyers towards a particular class of items for sale. If test users buy up those items, the supply available to the control users declines."

The above violation of the so-called Stable Unit Treatment Value Assumption (SUTVA [13]), has been viewed as problematic in the context of online platforms at least as early as Reiley's seminal 'Magic on the Internet' work [40]; Blake and Coey [8] were simply pointing out that the resulting inferential biases were large, which is particularly problematic since treatment effects in this context are typically tiny. The *interference* problem above is germane to experimentation on commerce platforms where interventions on a given experimental unit impact other units since all units effectively share a common inventory of 'demand' or 'supply' depending on context.

Despite what appears to be the ubiquity of such interference, a practical solution is far from settled. The majority of approaches so far fall under the category of *experimental design*, the idea being that a more-careful assignment of treatment will render the bias of the 'naively'-derived inference negligible. This ongoing line of work has produced sophisticated experiment designs which, in the best cases, provably reduce bias under highly specialized

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

models. While this is promising in theory, experimentation on online platforms in particular still largely relies on the simplest designs, i.e. A/B tests. For reasons including cost and organizational frictions, sophisticated experimental designs are not be an ideal lever, and often infeasible.

Markovian Interference and Existing Approaches: We study a generic experimentation problem within a system represented as a Markov Decision Process (MDP), where treatment corresponds to an action which may interfere with state transitions. This form of interference, which we refer to as *Markovian*, naturally subsumes the platform examples above, as recently noted by others either implicitly [48] or explicitly [26, 52]. In that example, a user arrives at each time step, the platform chooses an action (whether to treat the user), and the user's purchase decision alters the system state (inventory levels).

Our goal is to estimate the Average Treatment Effect (ATE), defined as the difference in steady-state reward with and without applying the treatment. In light of the above discussion, we assume that experimentation is done under simple randomization (i.e. A/B testing). Now without design as a lever, there are perhaps two existing families of estimators:
1. Naive: We will explicitly define the *Naive* estimator in the next section, but the strategy amounts to simply ignoring the presence of interference. This is by and large what is done in practice. Of course it may suffer from high bias (we show this formally in Example 1), but it serves as more than just a strawman. In particular, bias is only one side of the estimation coin, and with respect to the other side, namely variance, the Naive estimator is effectively the best possible.
2. Off-Policy Evaluation (OPE): Another approach comes from viewing our problem as one of policy evaluation in reinforcement learning (RL). Succinctly, it can be viewed as estimating the average reward of two different policies (no treatment, or treatment) given observations from some *third* policy (simple randomization). This immediately suggests framing the problem as one of *Off-Policy Evaluation*, and borrowing one of many existing *unbiased* estimators, e.g. [59, 58, 39, 24, 31, 32]. This tack appears to be promising, e.g. [52], but we observe that the resulting variance is necessarily large (Theorem 3).

Our Contributions: Against the above backdrop, we propose a novel *on*-policy treatment-effect estimator, which we dub the 'Differences-In-Q's (DQ)' estimator, for experiments with Markovian interference. In a nutshell, we characterize our contribution as follows:

The DQ estimator has provably negligible bias relative to the treatment effect. Its variance can, in general be exponentially smaller than that of an efficent off-policy estimator. In both stylized and large-scale real-world models, it dominates state-of-the-art alternatives.

We next describe these relative merits in greater detail:
1. Second-order Bias: We show (Theorem 1) that when the impact of an intervention on transition probabilities is $O(\delta)$, the bias of the DQ estimator is $O(\delta^2)$. The DQ estimator thus leverages the one piece of structure we have relative to generic off-policy evaluation: treatment effects are typically small.
2. Variance: We show (Theorem 2) that the DQ estimator is asymptotically normal, and provide a non-trivial, explicit characterization of its variance. By comparison, we show (Theorem 4) that this variance can, in general, be exponentially (in the size of the state space) smaller than the variance of *any* unbiased off-policy estimator.

Summarizing the above points, we are the first (to our knowledge) to explicitly characterize the favorable bias-variance trade-off in using *on-policy* estimation to tackle off-policy evaluation. This new lens has broader implications for OPE and policy optimization in RL.
3. Practical Performance: We conduct experiments in both a caricatured one-dimensional environment proposed by others [26], as well as a city-scale simulator of a ride-sharing platform. We show that in both settings the DQ estimator has MSE that is substantially lower than (a) naive and off-policy estimators, and even (b) estimators given access to incumbent state-of-the-art experimental *designs*.

Related Literature: The largest portion of work in interference is in *experimental design*, with the design levers ranging from stopping times in A/B tests [34, 25, 66, 27], to any form of more-sophisticated 'clustering' of units [12, 18, 21, 15, 43, 62, 64, 17], to clustering specifically when interference is represented by a network [41, 63, 50, 2, 7, 45, 70], to the proportion of units treated [23, 57, 4], to the timing of treatment [53, 9, 19], and beyond [3, 33, 60, 41, 11, 6, 22, 50]. As alluded to earlier, these sophisticated designs can be powerful, but cost, user experience, and other implementation concerns restrict their application in practice [35, 36].

We view this paper as orthogonal to this literature, but will eventually compare against a recent state-of-the-art design, so-called *two-sided randomization* [26, 5], that is specific to the context of two-sided marketplaces (e.g. the one we simulate).

As stated earlier, the problem we study is one of *off-policy evaluation (OPE)* [46, 55]. The fundamental challenge in OPE is high variance, which can be attributed to the nature of the algorithmic tools used, e.g. sampling procedures [59, 58, 39]. Recent work on 'doubly-robust' estimators [24, 31, 32] has improved on variance (incidentally, our estimator is loosely tied to these, as we discuss in Section 6), but again we will show, via a formal lower bound, that unbiased estimators as a whole have prohibitively large variance. Finally, our motivation is close in spirit to a recent paper [52], which applies OPE directly in Markovian interference settings; we make a direct experimental comparison in Section 5.

In the policy optimization literature, 'trust-region' methods [51] and conservative policy iteration [30] use a related on-policy estimation approach to bound policy improvement. However, the explicit application of on-policy estimation in the context of OPE, and in particular the striking bias-variance tradeoff this enables, are novel to this paper.

2. Model

This section formalizes the inference problem that we tackle, casting it in the language of MDPs. Vis-à-vis the existing literature, this lens allows us to reason about the problem using a large, well-established toolkit, and makes obvious the fact that OPE provides unbiased estimation of the ATE. We then present what we call the 'Naive' estimator (alluded to in the introduction). This is the lowest-variance estimator one can hope for in this setting, but it can have significant bias, as we will see.

We begin by defining an MDP with state space $\mathcal{S}$. We denote by $s_t \in \mathcal{S}$ the state of the MDP at time $t \in \mathbb{N}$. Every state is associated with a set of available actions $\mathcal{A}$ which govern the transition probabilities between states via the (unknown) function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. We assume that $\mathcal{A} = \{0, 1\}$ irrespective of state; for descriptive purposes, we will associate the '1' action with the use of a prospective intervention, so that '0' is associated with not employing the intervention. We denote by $r(s, a)$ the reward earned in state s having employed action a. A policy $\pi : \mathcal{S} \to \mathcal{A}$ maps states to random actions. We define the average reward λ^π, under any (ergodic, unichain) policy π, according to:

$$\lambda^\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r(s_t, \pi(s_t)).$$

There are three policies we define explicitly:

The Incumbent Policy π_0: This policy never uses the intervention, so that $\pi_0(s) = 0$ for all s. This is 'business as usual'. Denote the associated transition matrix as P_0 (i.e. the entries of P_0 are exactly $p(\cdot, 0, \cdot)$)

The Intervention Policy π_1: This policy always uses the intervention, so that $\pi_1(s) = 1$ for all s. This reflects the system, should the intervention under consideration be 'rolled out'. Denote the associated transition matrix as P_1.

The Experimentation Policy π_p: This policy corresponds to the experiment design. Simple randomization would select $\pi(s) = 1$ with some fixed probability p, say $1/2$, independently at every period. This corresponds to the sort of search engine experiment alluded to in the introduction. The transition matrix associated with this design is then $P_{1/2} = \frac{1}{2}P_0 + \frac{1}{2}P_1$.

The Inference Problem: We are given a single sequence of T states, actions, and rewards, observed under the experimentation policy π_p (recall that cost and constraints [35, 36] prohibit us from running π_0 or π_1 separately until convergence). The data we have is the sequence $\{(s_t, a_t, r(s_t, a_t)) : t = 1, \ldots, T\}$, wherein $a_t \triangleq \pi_p(s_t)$. We must estimate the average treatment effect (ATE):

$$\text{ATE} \triangleq \lambda^{\pi_1} - \lambda^{\pi_0}.$$

2.1. The Naive Estimator and Bias

A natural approach to estimating the ATE is to use simple randomization (i.e. $P_{1/2}$) and the following *Naive* estimator:

$$\hat{\text{ATE}}_N = \frac{1}{|T_1|}\sum_{t\in T_1} r(s_t, a_t) - \frac{1}{|T_0|}\sum_{t\in T_0} r(s_t, a_t), \tag{1}$$

where $T_1 = \{t : a_t = 1\}$ and $T_0 = \{t : a_t = 0\}$. In the context of the search engine experiment, this corresponds to simply averaging some metric of interest (say, conversion) among the test users (T_1) and control users (T_0). What goes wrong is simply that the two empirical averages above, that seek to estimate λ^{π_1} and λ^{π_0} respectively, employ the wrong measure over states. This is sufficient to introduce bias that is on the order of the treatment effect being estimated:

Example 1. *Consider an MDP on two states, $\mathcal{S} = \{\mathbf{0}, \mathbf{1}\}$. We collect a reward of 0 in state $\mathbf{0}$ irrespective of the action taken in that state ($r(\mathbf{0}, 0) = r(\mathbf{0}, 1) = 0$), and a reward of 1 in state $\mathbf{1}$, again, irrespective of action ($r(\mathbf{1}, 0) = r(\mathbf{1}, 1) = 1$). On the other hand, transitions are impacted by our choice of action. Specifically, let $p(\mathbf{0}, 0, \mathbf{0}) = p(\mathbf{0}, 0, \mathbf{1}) = p(\mathbf{1}, 0, \mathbf{1}) = p(\mathbf{1}, 0, \mathbf{0}) = 1/2$. We maintain $p(\mathbf{0}, 1, \mathbf{1}) = p(\mathbf{0}, 1, \mathbf{0}) = 1/2$ so that the intervention has no effect at state $\mathbf{0}$. On the other hand, we let $p(\mathbf{1}, 1, \mathbf{1}) = 1/2 + \delta$, so that $p(\mathbf{1}, 1, \mathbf{0}) = 1/2 - \delta$, for some $\delta > 0$. In words, the intervention tends to discourage a transition to $\mathbf{0}$ from state $\mathbf{1}$.*

In the above example, it is easy to calculate that $\text{ATE} = (1/2)\delta/(1-\delta)$, reflecting the shift in the stationary distribution favoring state $\mathbf{1}$, induced under the intervention. On the other hand, we can calculate that $\lim_T \hat{\text{ATE}}_N = 0$, so that the bias induced by the 'experimentation' policy relative to the stationary distributions under the incumbent and intervention policies respectively, is comparable to the size of the treatment effect.

3. The Differences-In-Q's Estimator

We are now prepared to introduce our estimator for inference in the presence of Markovian interference. Before defining our estimator, which we will see is only slightly more complicated than the Naive estimator, we recall a few useful objects associated with MDPs. First, for a fixed policy π, define the Bellman operator $T_\pi : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R} \to \mathbb{R}^{|\mathcal{S}|}$ according to

$$T_\pi(V, \lambda) = r_\pi - \lambda\mathbf{1} + P_\pi V,$$

where $r_\pi : \mathcal{S} \to \mathbb{R}$ is defined according to $r_\pi(s) = \mathsf{E}[r(s, \pi(s))]$. The average cost of policy π, denoted λ^π, and the bias function corresponding to π, denoted V_π, are then a solution to the fixed point equation $T_\pi(V, \lambda) = V$. Finally, the Q-function associated with π, denoted $Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is defined according to

$$Q_\pi(s, a) = r(s, a) - \lambda^\pi + \mathsf{E}[V_\pi(s_1)|s_0 = s, a_0 = a]. \tag{2}$$

3.1. An Idealized First Step

In motivating our estimator, let us begin with the following idealization of the Naive estimator, where we denote by $\rho_{1/2}$ the steady state distribution under the randomization policy $\pi_{1/2}$:

$$\mathsf{E}_{\rho_{1/2}}\left[\hat{\text{ATE}}_N\right] = \sum_s \rho_{1/2}(s)\left[r(s, 1) - r(s, 0)\right].$$

It is not hard to see that in the context of Example 1, we continue to have $\mathsf{E}_{\rho_{1/2}}[\hat{\text{ATE}}_N] = 0$, so that this idealization of the Naive estimator continues to have bias on the order of the treatment effect. Consider then, the following alternative:

$$\mathsf{E}_{\rho_{1/2}}\left[\hat{\text{ATE}}_D\right] = \sum_s \rho_{1/2}(s)\left[Q_{\pi_{1/2}}(s, 1) - Q_{\pi_{1/2}}(s, 0)\right], \tag{3}$$

where the term $\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_D]$ can for now just be thought of as an idealized constant ($\hat{\mathrm{ATE}}_D$ is defined soon in (4)). Compared to $\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_N]$, we see that $\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_D]$ takes a remarkably similar form, except that as opposed to an average over differences in rewards, we compute an average of differences in Q-function values. The idea is that doing so will hopefully compensate for the shift in distribution induced by $\pi_{1/2}$. We return to our example to check:

Example 1 (Continued). *Continuing with our example, we can explicitly calculate $Q_{\pi_{1/2}}(\cdot,\cdot)$, the average reward $\lambda^{\pi_{1/2}}$, and the stationary distribution $\rho_{1/2}$ (see Appendix B). Doing so allows us to calculate that*

$$\mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right] = \frac{1}{2}\left(\frac{\delta}{(1-\delta/2)^2}\right).$$

That is, $|\mathrm{ATE} - \mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_D]| = O(\delta^2)$, so that the bias of this idealized estimator is second-order (i.e. negligible) relative to the ATE.

Is the dramatic mitigation of bias we see in Example 1 generic? If the experimentation policy mixes fast, our first set of results essentially answers this question in the affirmative. In particular, we make the following mixing time assumption:

Assumption 1 (Mixing time). *There exist constants C and λ such that for all $s \in \mathcal{S}$,*

$$d_{\mathrm{TV}}(P_{1/2}^k(s,\cdot), \rho_{1/2}) \le C\lambda^k,$$

where $d_{\mathrm{TV}}(\cdot,\cdot)$ denotes total variation distance.

We then have that the second order bias we saw in Example 1 is, in fact, generic:

Theorem 1 (Bias of DQ). *Assume that for any state $s \in \mathcal{S}$, $d_{\mathrm{TV}}(p(s,1,\cdot), p(s,0,\cdot)) \le \delta$. Then,*

$$\left|\mathrm{ATE} - \mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right]\right| \le C'\left(\frac{1}{1-\lambda}\right)^2 r_{\max}\cdot\delta^2$$

where $r_{\max} := \max_{s,a}|r(s,a)|$ and C' is a constant depending (polynomially) on $\log(C)$.

3.2. The Differences-In-Q's Estimator

Motivated by the development in the previous subsection, the *Differences-In-Q's (DQ)* estimator we propose to use is simply

$$\hat{\mathrm{ATE}}_D = \frac{1}{|T_1|}\sum_{t\in T_1}\hat{Q}_{\pi_{1/2}}(s_t,a_t) - \frac{1}{|T_0|}\sum_{t\in T_0}\hat{Q}_{\pi_{1/2}}(s_t,a_t), \tag{4}$$

where we take an empirical average over the state trajectory produced under the randomization policy, and $\hat{Q}_{\pi_{1/2}}$ is an estimator of the Q-function. For concreteness, we obtain $\hat{Q}_{\pi_{1/2}}$ by solving

$$\min_{\hat{V},\hat{\lambda}}\sum_{s\in\mathcal{S}}\left(\sum_{t,s_t=s} r(s_t,a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t)\right)^2. \tag{5}$$

Our main result characterizes the variance and asymptotic normality of $\hat{\mathrm{ATE}}_D$:

Theorem 2 (Variance and Asymptotic Normality of DQ). *The DQ estimator is asymptotically normal so that*

$$\sqrt{T}\left(\hat{\mathrm{ATE}}_D - \mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right]\right) \xrightarrow{d} \mathcal{N}(0,\sigma_D^2),$$

with standard deviation

$$\sigma_D \le C'\left(\frac{1}{1-\lambda}\right)^{5/2}\log\left(\frac{1}{\min_{s\in S}\rho_{1/2}(s)}\right) r_{\max}.$$

where C' is a constant depending (polynomially) on $\log(C)$.

One Extreme of the Bias-Variance Tradeoff: We may heuristically think of the Naive estimator as representing one extreme of the bias-variance tradeoff among reasonable estimators. For the sake of comparison, by the Markov Chain CLT, the Naive estimator is also asymptotically normal with standard deviation $\Theta(r_{\max}/(1-\lambda)^{1/2})$. This rate is efficient for the estimation of the mean of a Markov chain [20]. On the other hand, while the Naive estimator is effectively useless for the problem at hand given its bias is in general $\Theta(\delta)$, that of the DQ estimator is $O(\delta^2)$.

4. The Price of Being Unbiased

Thus far, we have seen that the DQ estimator provides a dramatic mitigation in bias (Theorem 1) at a relatively modest price in variance (Theorem 2). This suggests another question: could we hope to construct an *unbiased* estimator that has low variance (i.e. comparable to either the Naive or DQ estimators). We will see that the short answer is: no.

4.1. The Variance of an Optimal Unbiased Estimator

As noted earlier, a plethora of Off-policy evaluation (OPE) algorithms might be used to provide an unbiased estimate of the ATE. Rather than consider a particular OPE algorithm, here we produce a lower bound on the variance of *any* unbiased OPE algorithm. While such a bound is obviously of independent interest (since OPE is a far more general problem than what we seek to accomplish in this paper), we will primarily be interested in comparing this lower bound to the variance of the DQ estimator from Theorem 2.

Theorem 3 (Variance Lower Bound for Unbiased Estimators)**.** *Assume we are given a dataset* $\{(s_t, a_t, r(s_t, a_t)) : t = 0, \dots, T\}$ *generated under the experimentation policy* $\pi_{1/2}$*, with* s_0 *distributed according to* $\rho_{1/2}$*. Then for any unbiased estimator* $\hat{\tau}$ *of* ATE*, we have that*

$$T \cdot \mathrm{Var}(\hat{\tau}) \geq 2 \sum_s \frac{\rho_1(s)^2}{\rho_{1/2}(s)} \sum_{s'} p(s,1,s')(V_{\pi_1}(s') - V_{\pi_1}(s) + r(s,1) - \lambda^{\pi_1})^2$$
$$+ 2 \sum_s \frac{\rho_0(s)^2}{\rho_{1/2}(s)} \sum_{s'} p(s,0,s')(V_{\pi_0}(s') - V_{\pi_0}(s) + r(s,0) - \lambda^{\pi_0})^2 \triangleq \sigma_{\mathrm{off}}^2.$$

It is worth remarking that this lower bound is tight: in the appendix we show that an LSTD(0)-type OPE algorithm achieves this lower bound. While this is of independent interest vis-à-vis average cost OPE, we turn next to our ostensible goal here – evaluating the 'price' of unbiasedness. We can do so simply by comparing the variance of the DQ estimator with the lower bound above. In fact, we are able to exhibit a class of one-dimensional Markov chains (in essence the same model proposed by [26] as a caricature of the dynamic interference problem) for which we have:

Theorem 4 (Price of Unbiasedness)**.** *For any* $0 < \delta \leq \frac{1}{5}$*, there exists a class of MDPs parameterized by* $n \in \mathbb{N}$*, where* n *is the number of states, such that*

$$\frac{\sigma_D}{\sigma_{\mathrm{off}}} = O\left(\frac{n^8}{c^n}\right),$$

for some constant $c > 1$*. Furthermore,* $|(\mathrm{ATE} - \mathsf{E}[\hat{\mathrm{ATE}}_D])/\mathrm{ATE}| \leq \delta$*.*

Another Extreme of the Bias-Variance Tradeoff: Theorems 2, 3, and 4 together reveal the opposite extreme of the bias-variance tradeoff. Specifically, if we insisted on an unbiased estimator for our problem (of which there are many, thanks to our framing of the problem as one of OPE), we would pay a large price in terms of variance. In particular Theorem 4 illustrates that this price can grow exponentially in the size of the state space. This jibes with our empirical evaluation in both caricatured and large-scale MDPs in Section 5.

Taken together our results reveal that the DQ estimator accomplishes a striking bias-variance tradeoff: it has substantially smaller variance than any unbiased estimator (in fact, comparable to the Naive estimator), all while ensuring bias that is second order in the impact of the intervention.

5. Experiments

This section will empirically investigate the DQ estimator and a number of alternatives in two settings: a simple one-dimensional toy model proposed by [26], and more realistically, a city-scale simulator of a ride-hailing platform similar to what large ride-hailing operators use in production. The alternatives we consider include: 1) the Naive estimator; 2) TSRI-1 and TSRI-2, the "two-sided randomization" (TSR) designs/estimators from [26]; 3) a variety of OPE estimators. For the OPE estimators, we note that off-policy average reward estimation has only recently been addressed in [65, 69], and we implement their specific estimators which we simply denote as TD and GTD respectively. We also implement an extension to an LSTD type estimator proposed in [52].

5.1. A toy example

We first study all of our estimators in a simple setting that does not call for any sort of value function approximation. Our goal is to understand the relative merits of these estimators in terms of their bias and variance. To this end, we adopt precisely the toy MDP studied by [26]; a stylized model of a rental marketplace. This MDP is essentially a 1-D Markov chain on $N = 5000$ states parameterized by a 'customer arrival' rate λ and a 'rental duration' rate μ. At a given state n (so that n units of inventory are in the system), the probability that an arriving customer rents a unit is impacted by the intervention. As such if the intervention increases the probability of a customer renting, this reduces the inventory availability for customers that arrive later. Our MDP setup exactly replicates that of [26], with $N = 5000, \lambda = 1, \mu = 1$; see the appendix for further details. We run all estimators over

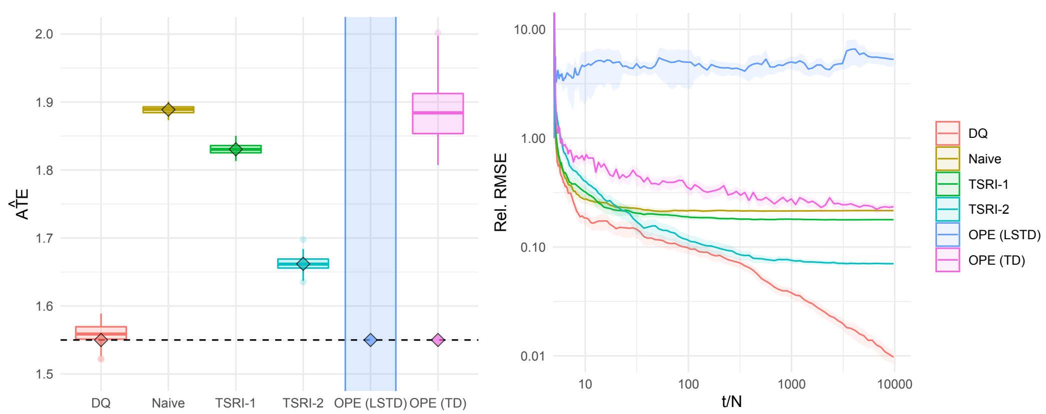

Figure 1: Toy-example from [26]. *Left*: Estimated ATE at time $t/N = 10^4$ across 100 trajectories. Dashed line indicates actual ATE. Diamonds indicate the asymptotic mean for each estimator. DQ shows compelling bias-variance tradeoff for this experimental budget. *Right*: Relative RMSE vs. Time; DQ dominates the alternatives at all timescales.

100 separate trajectories of length $t = 10^4 N$ of the above MDP initialized in its stationary distribution. Figure 1 summarizes the results of this experiment. Beginning with the left panel, which reports estimated quantities at $t = 10^4 N$, we immediately see:

TSR improves on Naive: The actual ATE in the experiment is 1.5%. Whereas it has the lowest variance of the estimators here, the Naive estimator has among the highest bias. The two TSR estimators reduce this bias substantially at a modest increase in variance. It is worth noting, as a sanity check, that these results precisely recreate those reported in [26].

OPE estimators are high variance: The OPE estimators have the highest variance of those considered here. The TD estimator has the lower variance but this is simply because it is implicitly regularized. Run long enough, both estimators will recover the treatment effect.

DQ shows a compelling bias-variance tradeoff: In contrast, the DQ estimator has the lowest bias at $t = 10^4 N$ and its variance is comparable to the TSR estimators (It is worth noting that run long enough, the DQ estimator had a bias of $\sim -5 \times 10^{-7}$).

Conclusions hold across experimental budgets: Turning our attention briefly to the right chart in Figure 1, we show the relative RMSE (i.e. RMSE normalized by the treatment effect) of the various estimators considered here *across all experimental budgets* t. RMSE effectively scalarizes bias and variance and we see that on this scalarization the DQ estimator dominates the other estimators considered here over all choice of t.

We note that specialized designs such as TSR can still be valuable in specific settings: when $\lambda \gg \mu$, for example, TSR is nearly unbiased (as shown in [26]), and can outperform DQ; see the appendix for such a study.

5.2. A Large-Scale Ridesharing Simulator

We next turn our attention to a city-scale ridesharing simulator similar to those used in production at large ride-hailing services. We will consider the problem of experimenting with changes to *dispatching* rules. Experimenting with these changes naturally creates Markovian interference by impacting the downstream supply/ positioning of drivers. Relative to the earlier toy example, the corresponding MDP here has an intractably large state-space, necessitating value function approximation for the DQ and OPE estimators.

The simulator: Ridesharing admits a natural MDP; see e.g. [48]. The state at the time of a request corresponds to that of all drivers at that time: position, assigned routes, riders, and the pickup/dropoff location of the request. Actions correspond to driver assignments and pricing decisions. The reward for a request is the price paid by the rider, less cost incurred to service the request. Our simulator models Manhattan. Riders and drivers are generated according to real world data, based on [1]; this yields $\sim 300k$ requests and $\sim 7k$ unique drivers per real day. An arriving request is served a menu of options generated by a price engine. The rider chooses an option based on a choice model calibrated on taxi prices (for the outside option) and implied delay disutility from typical match rates. A dispatch engine assigns a driver to the rider; the engine chooses the driver who can serve the rider at minimal marginal cost, subject to the product's constraints. Finally drivers proceed along their assigned routes until the next request is received. The simulator implements pooling. Users can switch out demand and supply generation, pricing and dispatch algorithms, driver repositioning, and the choice model via a simple API. Other simulators exist in the literature [48, 68], but lack either an open-source implementation, or implement a subset of the functionality here.

The experiment: We experiment with dispatch policies. Specifically, we consider assigning a request to an idle driver or a 'pool' driver, i.e. a driver who already has riders in their car. A dispatch algorithm might prefer the former, but only if the cost of the resulting trip is at most $\alpha\%$ higher than the cost of assigning to a pool driver. We consider three experiments, each of which changes α from a baseline of 0 to one of three distinct values: $30\%, 50\%$ or 70%, with ATEs of 0.5%, -0.9%, and -4.6% respectively. As we noted earlier, we would expect significant interference in this experiment (or indeed any experiment that experiments with pricing or dispatch) since an intervention changes the availability / position of drivers for subsequent requests.

Figure 2 summarizes the results of the above experiments, wherein each estimator was run over 50 independent simulator trajectories, each over 3×10^5 requests. The DQ and OPE estimators shared a common linear approximation architecture with basis functions that count the number of drivers at every occupancy level. We note that this approximation introduces its own bias which is not addressed by our theory. We immediately see:
Strong Impact of Interference: As we might expect, interference has a significant impact here as witnessed by the large bias in the Naive estimator.
Incumbent estimators do not improve on Naive: None of the incumbent estimators improve on Naive in this hard problem. This is also the case for the TSR designs, which in this large scale setting surprisingly appear to have significant variance. The OPE estimators have lower variance due to the regularization caused by value function approximation.
DQ works: In all three experiments, the bias in DQ (although in a relative sense higher than in the toy model) is *substantially* smaller than the alternatives, and also smaller than the ATE. This is evident in the left panel in Figure 2. Notice that in the rightmost experiment (ATE $= 0.5$), DQ is the only estimator to learn that the ATE is positive. Like in the toy model, the right panel shows that these results are robust over experimentation budgets.

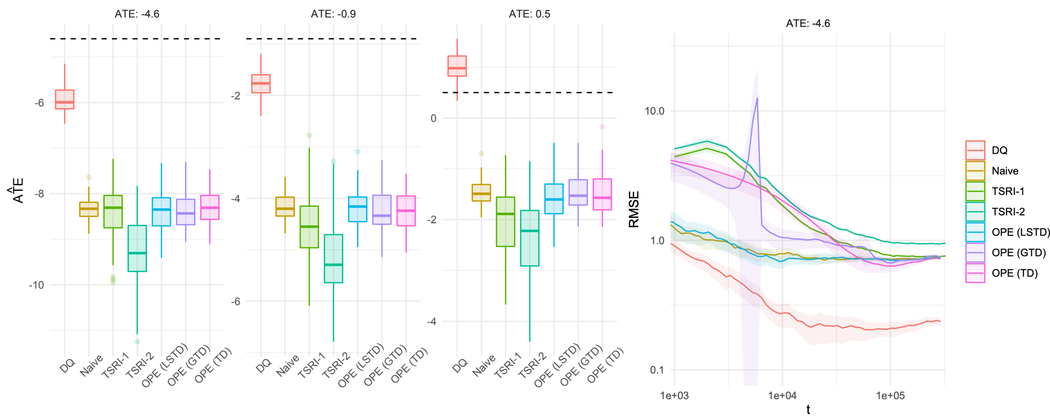

Figure 2: Ridesharing model *Left:* $\hat{\text{ATE}}$ at $t = 3 \times 10^5$ over 50 trajectories. Dashed line indicates actual ATE. DQ has lowest bias, and is only estimator to estimate correct sign of the treatment at all effect sizes. *Right:* RMSE vs. Time; DQ dominates at all time scales.

6. Discussion: refining the bias-variance tradeoff

To summarize, we have shown that the DQ estimator achieves a surprising bias-variance tradeoff by applying on-policy estimation to the Markovian interference problem, and more generally to OPE. Here we draw further connections between the Naive, DQ, and OPE estimators, and suggest how to interpolate between these estimators to realize other points along the bias-variance curve.

Dynkin's formula and an OPE meta-estimator. First, we situate the DQ estimator in the context of existing OPE techniques, using an identity referred to as Dynkin's formula in stochastic control, and re-derived several times in the RL literature:

$$\lambda_1 = \lambda_{1/2} + \mathsf{E}_{\rho_{1/2}}\left[\frac{\rho_1(s)}{\rho_{1/2}(s)}(Q_{\pi_{1/2}}(s,1) - V_{\pi_{1/2}}(s))\right]. \tag{6}$$

Taking $\lambda_1 - \lambda_0$, this translates into a familiar identity for the ATE:

$$\text{ATE} = \mathsf{E}_{\rho_{1/2}}\left[\zeta(s)(Q_{\pi_{1/2}}(s,1) - Q_{\pi_{1/2}}(s,0))\right] \tag{7}$$

where $\zeta(s) = \frac{1}{2}\frac{\rho_1(s)+\rho_0(s)}{\rho_{1/2}(s)}$ is the likelihood ratio of the stationary distributions. A variety of OPE estimators – including doubly-robust ([31, 61]) and primal-dual ([14, 56]) estimators – in fact estimate Equation (7) explicitly by plugging in estimates $\hat{\zeta}, \hat{Q}_{\pi_{1/2}}$ of the likelihood ratio and value functions (referred to as the "doubly-robust meta-estimator" in [31]):

$$\hat{\text{ATE}}_{\text{DR}} = \frac{1}{|T_1|}\sum_{t \in T_1}\hat{\zeta}(s_t)\hat{Q}_{\pi_{1/2}}(s_t,1) - \frac{1}{|T_0|}\sum_{t \in T_0}\hat{\zeta}(s_t)\hat{Q}_{\pi_{1/2}}(s_t,0)$$

Refining the bias-variance tradeoff. Immediately, we see that that by taking the likelihood ratio to be a constant $\hat{\zeta}(s) = 1\ \forall s$, we recover the DQ estimator $\hat{\text{ATE}}_{\text{DQ}}$. Furthermore, if we then take $\hat{V}_{\pi_{1/2}}$ to be any constant $\hat{V}_{\pi_{1/2}}(s) = c\ \forall s$, we recover the Naive estimator[1]. The DQ and Naive estimators' relationship to OPE then becomes clear: we obtain DQ by choosing a minimal variance (but highly biased) estimator of ζ; and we obtain the Naive estimator by subsequently choosing minimal variance (but highly biased) estimator of V.

[1]To see this, observe that $\hat{Q}_{\pi_{1/2}}(s,a) = r(s,a) + \mathsf{E}\left[\hat{V}_{\pi_{1/2}}(s')|s,a\right] = r(s,a) + c$

This suggests that we can interpolate between these extremes by making more refined bias-variance tradeoffs in estimating $\hat{\zeta}$ and $V_{\pi_{1/2}}$. It turns out that several natural approaches to variance reduction provide exactly such an interpolation:

- *Explicit regularization.* In estimating $\hat{\zeta}(s)$, one can directly penalize its deviation from one, where increasing the penalty interpolates from OPE to DQ. Given that estimation of $\hat{\zeta}(s)$ is the key difference between DQ and unbiased OPE – and therefore the source of the massive variance gap (Theorems **??** and **??**) – we would expect this to be a particularly powerful approach to OPE, and indeed some works have shown strong empirical performance using similar penalties [44].

  Similarly, one can directly penalize the deviation of $\hat{V}_{\pi_{1/2}}$ from zero (or any constant), as in regularized variants of LSTD (see e.g. [37]). As we increase the regularization penalty on $\hat{\zeta}(s)$, we interpolate from OPE to DQ; additionally increasing the regularization penalty on $\hat{V}_{\pi_{1/2}}$ then interpolates from DQ to Naive. Approaches combining both forms of regularization have been explored in [67].

- *Function approximation.* More generally, one can restrict $\hat{\zeta}(s)$ and $\hat{V}_{\pi_{1/2}}$ to lie in particular function classes, with one extreme being any mapping $\mathcal{S} \mapsto \mathbb{R}$, and the other extreme being the constant functions $\hat{V}_{\pi_{1/2}}(s) = c$ or $\hat{\zeta}(s) = 1$. As one example, when the state space is massive we may approximate it using state aggregation. At the extreme, aggregating all states into a single aggregate state implies that the value function (or likelihood ratio) must be a constant. As the aggregation for $\hat{\zeta}(s)$ goes from fine to coarse, we interpolate between OPE and DQ; subsequently increasing the coarseness of $\hat{V}_{\pi_{1/2}}(s)$ then interpolates between DQ and Naive.

- *Discounting.* A common technique to estimate the average reward value function is to instead estimate a discounted reward value function $Q_\gamma(s,a) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right]$, motivated by the fact that we obtain exactly the average-reward value function Q as the discount rate γ goes to one (under the proper scaling; precisely, $\lim_{\gamma \to 1}(1-\gamma)Q_\gamma(s,a) = Q(s,a)$ [47]). This approach is commonly applied to reduce variance in average reward RL (see e.g. [29]). Implementing DQ with $\hat{Q}_{\pi_{1/2}}(s,a) = (1-\gamma)Q_\gamma(s,a)$ yields the exact DQ estimator as $\gamma \to 1$, and the Naive estimator as $\gamma \to 0$.

References

[1] TLC Trip Record Data - TLC. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] Susan Athey, Dean Eckles, and Guido W Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

[3] Lars Backstrom and Jon Kleinberg. Network bucket testing. In *Proceedings of the 20th international conference on World wide web*, pages 615–624, 2011.

[4] Sarah Baird, J Aislinn Bohren, Craig McIntosh, and Berk Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018.

[5] Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas Richardson, and Ido M Rosen. Multiple randomization designs. *arXiv preprint arXiv:2112.13495*, 2021.

[6] Guillaume W Basse and Edoardo M Airoldi. Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858, 2018.

[7] Guillaume W Basse, Avi Feller, and Panos Toulis. Randomization tests of causal effects under interference. *Biometrika*, 106(2):487–494, 2019.

[8] Thomas Blake and Dominic Coey. Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 567–582, 2014.

[9] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020.

[10] Guan-Yu Chen and Laurent Saloff-Coste. On the mixing time and spectral gap for birth and death chains. *arXiv preprint arXiv:1304.4346*, 2013.

[11] David Choi. Estimation of monotone treatment effects in network experiments. *Journal of the American Statistical Association*, 112(519):1147–1155, 2017.

[12] Jerome Cornfield. Randomization by group: a formal analysis. *American journal of epidemiology*, 108(2):100–102, 1978.

[13] David Roxbee Cox. Planning of experiments. 1958.

[14] Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the Actor with Dual Critic. *arXiv:1712.10282 [cs]*, December 2017.

[15] Allan Donner and Neil Klar. Pitfalls of and controversies in cluster randomization trials. *American journal of public health*, 94(3):416–422, 2004.

[16] Joseph L Doob. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169, 1935.

[17] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1), 2017.

[18] John W Farquhar. The community-based model of life style intervention trials. *American journal of epidemiology*, 108(2):103–111, 1978.

[19] Peter W Glynn, Ramesh Johari, and Mohammad Rasouli. Adaptive Experimental Design with Temporal Interference: A Maximum Likelihood Approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 15054–15064. Curran Associates, Inc., 2020.

[20] Priscilla E Greenwood and Wolfgang Wefelmeyer. Efficiency of empirical estimators for markov chains. *The Annals of Statistics*, pages 132–143, 1995.

[21] COMMIT Research Group. Community intervention trial for smoking cessation (commit): summary of design and intervention. *JNCI: Journal of the National Cancer Institute*, 83(22):1620–1628, 1991.

[22] Huan Gui, Ya Xu, Anmol Bhasin, and Jiawei Han. Network a/b testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web*, pages 399–409, 2015.

[23] Michael G Hudgens and M Elizabeth Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.

[24] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[25] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[26] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 2022.

[27] Ramesh Johari, Leo Pekelis, and David Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research (To Appear)*, 2020.

[28] Galin L Jones. On the markov chain central limit theorem. *Probability surveys*, 1:299–320, 2004.

[29] Sham Kakade. Optimizing average reward using discounted rewards. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 605–615, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[30] Sham Kakade and John Langford. Approximately Optimal Approximate Reinforcement Learning. In *In Proc. 19th International Conference on Machine Learning*, pages 267–274, 2002.

[31] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res.*, 21(167):1–63, 2020.

[32] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 2022.

[33] Liran Katzir, Edo Liberty, and Oren Somekh. Framework and algorithms for network bucket testing. In *Proceedings of the 21st international conference on World Wide Web*, pages 1029–1036, 2012.

[34] Eugene Kharitonov, Aleksandr Vorobev, Craig Macdonald, Pavel Serdyukov, and Iadh Ounis. Sequential testing for early stopping of online experiments. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 473–482, 2015.

[35] John Kirn. Challenges in Experimentation. https://eng.lyft.com/challenges-in-experimentation-be9ab98a7ef4, April 2022.

[36] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

[37] J. Zico Kolter and Andrew Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press.

[38] Vijay R Konda. Actor-critic algorithms, 2002.

[39] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

[40] David Lucking-Reiley. Using Field Experiments to Test Equivalence between Auction Formats: Magic on the Internet. *American Economic Review*, 89(5):1063–1080, December 1999.

[41] Charles F Manski. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23, 2013.

[42] Carl D Meyer, Jr. The condition of a finite markov chain and perturbation bounds for the limiting probabilities. *SIAM Journal on Algebraic Discrete Methods*, 1(3):273–283, 1980.

[43] David M Murray et al. *Design and analysis of group-randomized trials*, volume 29. Monographs in Epidemiology and, 1998.

[44] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pages 2315–2325, 2019.

[45] Jean Pouget-Abadie, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

[46] Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pages 417–424, 2001.

[47] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, August 2014.

[48] Zhiwei Tony Qin, Hongtu Zhu, and Jieping Ye. Reinforcement Learning for Ridesharing: A Survey. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2447–2454, September 2021.

[49] Vladimir Rakočević. On continuity of the moore-penrose and drazin inverses. *Matematichki Vesnik*, 49(3-4):163–172, 1997.

[50] Martin Saveski, Jean Pouget-Abadie, Guillaume Saint-Jacques, Weitao Duan, Souvik Ghosh, Ya Xu, and Edoardo M Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1027–1035, 2017.

[51] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[52] Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework. *Journal of the American Statistical Association*, 0(ja):1–29, January 2022.

[53] Carla Sneider, Yixin Tang, and Y Tang. Experiment rigor for switchback experiment analysis. *URL: https://doordash. engineering/2019/02/20/experiment-rigor-for-switchbackexperiment-analysis*, 2018.

[54] Petre Stoica and Thomas L Marzetta. Parameter estimation problems with singular information matrices. *IEEE Transactions on Signal Processing*, 49(1):87–90, 2001.

[55] Richard S Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21):1609–1616, 2008.

[56] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly Robust Bias Reduction in Infinite Horizon Off-Policy Estimation, October 2019.

[57] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.

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

[59] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[60] Panos Toulis and Edward Kao. Estimation of causal peer influence effects. In *International conference on machine learning*, pages 1489–1497. PMLR, 2013.

[61] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9659–9668. PMLR, November 2020.

[62] Johan Ugander and Lars Backstrom. Balanced label propagation for partitioning massive graphs. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 507–516, 2013.

[63] Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 329–337, 2013.

[64] Dylan Walker and Lev Muchnik. Design of randomized experiments in networks. *Proceedings of the IEEE*, 102(12):1940–1951, 2014.

[65] Yi Wan, Abhishek Naik, and Richard S. Sutton. Learning and Planning in Average-Reward Markov Decision Processes. In *Proceedings of the 38th International Conference on Machine Learning*, pages 10653–10662. PMLR, July 2021.

[66] Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. A framework for multi-a (rmed)/b (andit) testing with online fdr control. *Advances in Neural Information Processing Systems*, 30, 2017.

[67] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-Policy Evaluation via the Regularized Lagrangian. In *Advances in Neural Information Processing Systems*, volume 33, pages 6551–6561. Curran Associates, Inc., 2020.

[68] Rui Yao and Shlomo Bekhor. A ridesharing simulation platform that considers dynamic supply-demand interactions. April 2021.

[69] Shangtong Zhang, Yi Wan, Richard S. Sutton, and Shimon Whiteson. Average-Reward Off-Policy Policy Evaluation with Function Approximation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12578–12588. PMLR, July 2021.

[70] Corwin M Zigler and Georgia Papadogeorgou. Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(1):109, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] No use of existing assets.

(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A. Notation

For a vector $a \in \mathbb{R}^n$, we use $\|a\|_1 = \sum_{i=1}^n |a_i|$ and $\|a\|_\infty = \max_{i=1}^n |a_i|$. For a matrix $M \in \mathbb{R}^{n\times m}$, we use $\|M\|_{1,\infty} = \max_{1\le i\le n}\sum_{j=1}^m |a_{ij}|$ to represent the maximal row-wise l_1-norms. We use $\mathbf{1}$ to represent the vectors with all ones. We use $A^{\#}$ to represent the group inverse of A. For an irreducible and aperiodic Markov chain with associated transition matrix P and the stationary distribution ρ, we have $(I-P)^{\#} = (I - P + \mathbf{1}\rho^\top)^{-1} - \mathbf{1}\rho^\top$.

B. Analysis of Example 1

To begin, let us derive the ATE. Under policy π_0, the transition matrix is

$$P_0 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

and the stationary distribution is $\rho_0 = [1/2, 1/2]^\top$ accordingly. Similarly, one can verify under policy π_1, the transition matrix is

$$P_1 = \begin{bmatrix} 1/2 & 1/2 \\ 1/2-\delta & 1/2+\delta \end{bmatrix}$$

and the stationary distribution is $\rho_1 = [\frac{1-2\delta}{2-2\delta}, \frac{1}{2-2\delta}]^\top$. Let $r_0 = [0,1]^\top, r_1 = [0,1]^\top$ be the reward vector under actions 0 or 1. Then, the ATE is

$$\begin{aligned} \text{ATE} &= r_1^\top \rho_1 - r_0^\top \rho_0 \\ &= \frac{1}{2-2\delta} - \frac{1}{2} \\ &= \frac{1-1+\delta}{2-2\delta} \\ &= \frac{\delta}{2}\frac{1}{1-\delta}. \end{aligned}$$

Next, we consider the computation of $\mathsf{E}_{\rho_{1/2}}[\hat{\text{ATE}}_D]$, which can be written as

$$\mathsf{E}_{\rho_{1/2}}[\hat{\text{ATE}}_D] = \rho_{1/2}^\top (Q_1 - Q_0)$$

where Q_a is the Q-value vector for the policy $\pi_{1/2}$ under the action a. To compute $\rho_{1/2}$, Q_0, and Q_1, consider the transition matrix P for the policy $\pi_{1/2}$:

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2-\delta/2 & 1/2+\delta/2 \end{bmatrix}.$$

Then one can verify that the stationary distribution $\rho_{1/2}$ is

$$\rho_{1/2} = \left[\frac{1-\delta}{2-\delta}, \frac{1}{2-\delta}\right]^\top$$

and the average reward $\lambda^{1/2} = \frac{1}{2-\delta}$.

Furthermore, consider the following Bellman equation for Q-value function:

$$Q(s,a) = r(s,a) - \lambda^{1/2} + \sum_{s',a'} P_a(s,s')\frac{1}{2}Q(s',a').$$

One can verify that one solution of the above equations is

$$Q(0,0) = 0, \quad Q(0,1) = 0$$
$$Q(1,0) = 1, \quad Q(1,1) = 1 + \frac{2\delta}{2-\delta}$$

Therefore,

$$\begin{aligned}\mathsf{E}_{\rho_{1/2}}[\hat{\text{ATE}}_D] &= \frac{1}{2-\delta}(Q(1,1) - Q(1,0)) \\ &= \frac{1}{2-\delta}\frac{2\delta}{2-\delta} \\ &= \frac{1}{2}\left(\frac{\delta}{(1-\delta/2)^2}\right).\end{aligned}$$

For the bias induced by the DQ estimator, we have

$$\begin{aligned}\text{ATE} - \mathsf{E}_{\rho_{1/2}}[\hat{\text{ATE}}_D] &= \frac{\delta}{2}\left(\frac{1}{1-\delta} - \frac{1}{(1-\delta/2)^2}\right) \\ &= \frac{\delta}{2}\left(\frac{1}{1-\delta} - \frac{1}{1-\delta+\delta^2/4}\right) \\ &= \frac{\delta}{2}\frac{\delta^2/4}{(1-\delta)(1-\delta+\delta^2/4)}.\end{aligned}$$

This completes the analysis.

C. Proof of Theorem 1

The proof of Theorem 1 is a simple proof built on a perturbation formula for stationary distributions of Markov chains. We in fact construct a novel Taylor series representation of the ATE parameterized by δ that controls the perturbation around $P_{1/2}$, which yields the Naive estimator as the zeroth-order truncation of the series; and the idealized DQ estimator as the natural first-order correction. Theorem 1 then proceeds by bounding the remainder. This strategy additionally allows us to generalize the DQ estimator to arbitrarily high-order bias corrections, by computing Q-functions iteratively. Here we present the proof (with some details omitted for simplicity).

We first define few pieces of useful notation. Let $\rho_0 \in \mathbb{R}^{|\mathcal{S}|}, \rho_{1/2} \in \mathbb{R}^{|\mathcal{S}|}, \rho_1 \in \mathbb{R}^{|\mathcal{S}|}$ be the vectors of the stationary distributions of $P_0, P_{1/2}, P_1$ accordingly. Let $r_0 \in \mathbb{R}^{|\mathcal{S}|}, r_{1/2} \in \mathbb{R}^{|\mathcal{S}|}, r_1 \in \mathbb{R}^{|\mathcal{S}|}$ be the reward vectors associated with policies $\pi_0, \pi_{1/2}, \pi_1$, i.e., $r_a(s) = r(s,a)$ and $r_{1/2} = \frac{1}{2}r_0 + \frac{1}{2}r_1$.

To begin, we parameterize $P_0 := P_{1/2} - \delta A$ and $P_1 := P_{1/2} + \delta A$ by δ with fixed $P_{1/2}$ and some fixed matrix $A \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with $\|A\|_{1,\infty} \le 1$ ($\|A\|_{1,\infty} = \max_i \sum_j |A_{ij}|$)[2]. Then, ρ_0 and ρ_1 can also be viewed as a function of δ. Also recall $\text{ATE} = \rho_1^\top r_1 - \rho_0^\top r_0$. Our goal is to represent ATE as a function of δ and then study the Taylor expansion of such a function. To do so, we use the following known perturbation formula of Markov chains.

Lemma 1 (Stationary Distribution Perturbation, Theorem 4.1 [42])**.** *Suppose $P \in \mathbb{R}^{n\times n}$ and $P' \in \mathbb{R}^{n\times n}$ are transitions matrices of two finite-state aperiodic and irreducible Markov Chains and $\rho \in \mathbb{R}^n, \rho' \in \mathbb{R}^n$ are the stationary distributions accordingly. Then $\rho'^\top = \rho^\top + \rho'^\top(P'-P)(I-P)^\#$ where $(I-P)^\#$ is the group inverse of $I-P$ given by $(I-P)^\# = (I-P+\mathbf{1}\rho^\top)^{-1} - \mathbf{1}\rho^\top$.*

Let us apply Lemma 1 to $\rho_1^\top r_1$ based on the perturbation between $\rho_{1/2}$ and ρ_1.

$$\begin{aligned}\rho_1^\top r_1 &= \rho_{1/2}^\top r_1 + \rho_1^\top (P_1 - P_{1/2})(I - P_{1/2})^\# r_1 \\ &= \rho_{1/2}^\top r_1 + \delta \cdot \rho_1^\top A (I-P_{1/2})^\# r_1 \end{aligned} \tag{8}$$

[2] This is always possible since $d_{\text{TV}}(p(s,1,\cdot), p(s,0,\cdot)) \le \delta$.

Note that we can apply Lemma 1 again to the ρ_1 in the RHS of Eq. (8) and then repeat this process,

$$(9) \qquad \rho_1^\top r_1 = \sum_{k=0}^{K} \delta^k \cdot \rho_{1/2}^\top \left(A(I-P_{1/2})^\#\right)^k r_1 + \delta^{K+1} \cdot \rho_1^\top \left(A(I-P_{1/2})^\#\right)^{K+1} r_1$$

for any $K = 0, 1, 2, \ldots$. Essentially Eq. (9) provides the K-th order Taylor expansion for $\rho_1^\top r_1$ with an explicit remainder. Furthermore, we can bound the remainder by

$$\begin{aligned} \left|\rho_1^\top \left(A(I-P_{1/2})^\#\right)^{K+1} r_1\right| &\overset{(i)}{\le} \|\rho_1\|_1 \left(\|A\|_{1,\infty} \|I - P_{1/2}^\#\|_{1,\infty}\right)^{K+1} \|r_1\|_{\max} \\ &\overset{(ii)}{\le} \|I - P_{1/2}^\#\|_{1,\infty}^{K+1} r_{\max} \\ &\overset{(iii)}{\le} \left(\frac{2\ln(C)+1}{1-\lambda}\right)^{K+1} r_{\max} \end{aligned}$$

Here in (i) we use that for any vector a, b and matrix B, we have $|a^\top b| \le \|a\|_1 \|b\|_{\max}$ and $\|a^\top B\|_1 \le \|a\|_1 \|B\|_{1,\infty}$. In (ii) we use that $\|\rho_1\|_1 = 1, \|A\|_{1,\infty} \le 1$. In (iii), we use the following lemma implied by the mixing time assumption and the series expansion of $(I-P)^\#$.

Lemma 2. *Suppose for any $s \in \mathcal{S}$, $d_{\mathrm{TV}}(P_{1/2}^k(s,\cdot), \rho_{1/2}) \le C\lambda^k$. Then $\|(I-P_{1/2})^\#\|_{1,\infty} \le \frac{2\ln(C)+1}{1-\lambda}$.*

Appplying a similar process to $\rho_0^\top r_0$, we obtain the Taylor expansion for the ATE.

$$(10) \quad \mathrm{ATE} = \sum_{k=0}^{K} \delta^k \cdot \left(\rho_{1/2}^\top \left(A(I-P_{1/2})^\#\right)^k r_1 - \rho_{1/2}^\top \left((-A)(I-P_{1/2})^\#\right)^k r_0\right) + \delta^{K+1} \cdot a_K$$

where $|a_K| \le 2\left(\frac{2\ln(C)+1}{1-\lambda}\right)^{K+1} r_{\max}$. It is easy to see that the Naive estimator $\rho_{1/2}^\top(r_1 - r_0)$ corresponds to the zeroth-order truncation. In fact, the DQ estimator, i.e., $\mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right]$, exactly matches the first-order truncation. To see this, by the definition of $\mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right]$ and Q-functions,

$$\begin{aligned} \mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right] &= \sum_s \rho_{1/2}(s)\left(Q_{\pi_{1/2}}(s,1) - Q_{\pi_{1/2}}(s,0)\right) \\ &= \sum_s \rho_{1/2}(s)\left(r_1(s) + \sum_{s'} V_{1/2}(s')P_1(s,s') - r_0(s) - \sum_{s'} V_{1/2}(s')P_0(s,s')\right) \\ &= \rho_{1/2}^\top\left(r_1 - r_0 + (P_1 - P_0)V_{1/2}\right) \end{aligned}$$

where $V_{1/2}$ is the induced vector of the V-function of policy $\pi_{1/2}$. By the well-known fact that $V_{1/2} = (I-P_{1/2})^\# r_{1/2}$ induced by the Bellman equation, we then have

$$\begin{aligned} \mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right] &= \rho_{1/2}^\top\left(r_1 - r_0 + (P_1 - P_0)(I-P_{1/2})^\# r_{1/2}\right) \\ &= \rho_{1/2}^\top r_1 - \rho_{1/2}^\top r_0 + \delta \rho_{1/2}^\top A(I-P_{1/2})^\#(r_1 + r_0). \end{aligned}$$

Then indeed $\mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_D\right]$ is the first-order Taylor truncation. Together, this completes the proof.

Generalization to Higher-Order Bias Correction. In fact, we can also use the K-th order Taylor expansion of ATE, to design estimators that can correct higher-order bias, in a similar way presented above.

D. Proof of Theorem 2

To begin, we present the outline of the proof. We aim to use Markov chain CLT ([28]) to provide the asymptotic normality of our estimator. Note that Markov chain CLT states that for a Markov chain $X_1, X_2, \ldots,$ and a bounded function u with the domain on the state space, there exists Σ_u such that

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} u(X_t) - u^*\right) \xrightarrow{d} N(0, \Sigma_u)$$

where u^* is the expected value of u under the stationary distribution of the Markov chain.

Delta method. Unfortunately, the estimator $\hat{\text{ATE}}_D$ can not be directly written as an empirical average of some function u. To address this issue, we use the delta method (see e.g. [16], Lemma 5). In particular, we write $\hat{\text{ATE}}_D = f(u_T)$ as a function of a random vector u_T given by $u_T := \frac{1}{T}\sum_{t=1}^{T} u(X_t)$. Under some minor conditions, the delta method states that

$$\sqrt{T}\left(f(u_T) - f(u^*)\right) \xrightarrow{d} N(0, \sigma_f^2)$$

where $\sigma_f^2 := \nabla f(u^*)^\top \Sigma_u \nabla f(u^*)$ and $\nabla f(u^*)$ is the gradient of f evaluating at the point u^*. This forms the basis of proving Theorem 2.

Linearization. To simplify the analysis for σ_f, instead of computing Σ_u explicitly, we "linearize" the function f by defining $\tilde{f}(X_t) := \nabla f(u^*)^\top (u(X_t) - u^*)$ and the delta method in fact implies (see Lemma 6)

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} \tilde{f}(X_t)\right) \xrightarrow{d} N(0, \sigma_f^2),$$

i.e., the linearized f converges with the same limiting variance as the original f. Therefore, we can focus on $\tilde{f}$ for analyzing σ_f.

Bounding σ_f with the Entry-wise Non-expansive Lemma. To bound σ_f, we will invoke Lemma 4, which states that

$$\sigma_f \leq \sqrt{\frac{1}{1-\lambda}}\tilde{f}_{\max}$$

where $\tilde{f}_{\max} := \max_s |\tilde{f}(s)|$. Then the problem boils down to bound $\tilde{f}_{\max}$. This bound is the key of Theorem 2 and requires us to bound

$$\max_k \rho_{1/2}^\top (P_1 - P_0)(I - P_{1/2})^{\#} D^{-1/2} e_k \tag{11}$$

where D is a diagonal matrix with entries $D_{ii} = \rho_{1/2}(i)$. It is not clear a priori that Eq. (11) is well-controlled. In fact, a loose analysis for Eq. (11) will give $\tilde{f}_{\max} = O(\frac{1}{(\rho_{\min})^{1/2}})$, which shows no advantage comparing to the off-policy estimators (the off-policy estimator requires a bound for $(\rho_0^\top + \rho_1^\top)(P_1 - P_0)(I - P_{1/2})^{\#} D^{-1/2})$.

Fortunately, we observe that based on the fact that $\rho_{1/2}$ is the stationary distribution of $P_{1/2}$ (on-policy estimator), there exists a non-expansive property (coined as the "Entry-wise Non-expansive Lemma", see Lemma 3), which states that

$$(\rho_{1/2}^\top (P_1 - P_0)(I - P_{1/2})^{\#})_k \leq c \cdot \rho_{1/2}(k)$$

for some c that depend only on λ and $\log(1/\rho_{\min})$. This is the key enabler for establishing the advantage of on-policy estimators rigorously, that leads to $\tilde{f}_{\max} = O(\log(\frac{1}{\rho_{\min}}))$. We believe this novel lemma is of independent interest for the field of OPE.

Next, we present the proof in full details.

D.1. Delta method and Linearization

To begin, consider the Markov chain $X_t = (s_t, a_t, s_{t+1})$. For $a \in \{0,1\}$, denote $F^{(a)}, h^{(a)}$ by

$$F^{(a)}(X_t) := 2E_{s_t}E_{s_{t+1}}^\top \cdot 1(a_t = a) \tag{12}$$

$$h^{(a)}(X_t) := 2r(s_t, a_t) \cdot E_{s_t} \cdot 1(a_t = a) \tag{13}$$

where E_s is a vector with all entries zero except that the s-th entry is one. Let $F_T^{(a)} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, h_T^{(a)} \in \mathbb{R}^{|\mathcal{S}|}$ be the empirical average of the function $F^{(a)}$ and $h^{(a)}$:

$$F_T^{(a)} := \frac{1}{T}\sum_{t=1}^T F^{(a)}(X_t)$$
$$h_T^{(a)} = \frac{1}{T}\sum_{t=1}^T h^{(a)}(X_t).$$

We aim to write $\hat{\mathrm{ATE}}_D := f(F_T^{(0)}, F_T^{(1)}, h_T^{(0)}, h_T^{(1)})$ as a function of $F_T^{(0)}, F_T^{(1)}, h_T^{(0)}, h_T^{(1)}$ for applying delta method. To do so, let $D_T^{(a)}$ be an diagonal matrix with entries $D_T^{(a)}(s,s) = \sum_{s'} F_T^{(a)}(s,s')$. One can verify that

$$\hat{V} = (D_T^{(0)} + D_T^{(1)} - F_T^{(0)} - F_F^{(1)})^\#(h_T^{(0)} + h_T^{(1)})$$

gives the estimation of V-function in Eq. (5). Further, one can verify that with a plugging-in estimator for Q, the DQ estimator is given by

$$\begin{aligned}\hat{\mathrm{ATE}}_D &= f(F_T^{(0)}, F_T^{(1)}, h_T^{(0)}, h_T^{(1)})\\ &=: \mathbf{1}^\top(F_T^{(1)} - F_T^{(0)})(D_T^{(0)} + D_T^{(1)} - F_T^{(0)} - F_F^{(1)})^\#(h_T^{(0)} + h_T^{(1)})\\ &\quad + \mathbf{1}^\top(h_T^{(1)} - h_T^{(0)}).\end{aligned}$$

By Markov chain CLT, we have when T goes to infinity

$$F_T^{(0)} \to F_0^* := DP_0, \quad F_T^{(1)} \to F_1^* := DP_1$$
$$h_T^{(0)} \to h_0^* := Dr_0, \quad h_T^{(1)} \to h_1^* := Dr_1$$

where D is a diagonal matrix with entries $D_{s,s} = \rho_{1/2}(s)$. Then by the delta method (see Lemma 5), we have[3]

$$\sqrt{T}(f(F_T^{(0)}, F_T^{(1)}, h_T^{(0)}, h_T^{(1)}) - f(F_0^*, F_1^*, h_0^*, h_1^*)) \xrightarrow{d} N(0, \sigma_f^2)$$

which is equivalent to

$$\sqrt{T}(\hat{\mathrm{ATE}}_D - \mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_D]) \xrightarrow{d} N(0, \sigma_f^2)$$

since $f(F_0^*, F_1^*, h_0^*, h_1^*) = \mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_D]$. To analyze σ_f, we consider the "linearization" of f around $u^* := (F_0^*, F_1^*, h_0^*, h_1^*)$. In particular, let $u(X_t) = (F^{(0)}(X_t), F^{(1)}(X_t), h^{(0)}(X_t), h^{(1)}(X_t))$. Let (λ, V) be the average reward and the "true" V-function under the policy $\pi_{1/2}$. One can verify that

$$\begin{aligned}\tilde{f}(s,a,s') &= \nabla f(u^*)^\top(u(s,a,s') - u^*)\\ &= (\mathbf{1}^\top D(P_1 - P_0)(I - P_{1/2})^\# D^{-1})E_s(r(s,a) - \lambda + V(s') - V(s))\\ &\quad + 2(1(a=1) - 1(a=0))(V(s') + r(s,a)) - c\end{aligned}$$

where $c := \mathsf{E}_{\rho_{1/2}}[2(1(a=1) - 1(a=0))(V(s') + r(s,a))]$. By Lemma 6, we have

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^T \tilde{f}(X_t)\right) \xrightarrow{d} N(0, \sigma_f^2).$$

[3] The group inverse is continuous if we consider the set of matrices with rank $|\mathcal{S}| - 1$ ([49]).

D.2. Bound σ_f

Next, we aim to provide a bound for σ_f. Note that that the mixing time of X_t is the same as s_t and by Lemma 4, we have

$$\sigma_f \leq \sqrt{2}\tilde{f}_{\max}\sqrt{\frac{2\ln(C)+1}{1-\lambda}}$$

where $\tilde{f}_{\max} = \max_{s,a,s'} |\tilde{f}(s,a,s')|$. Then the problem boils down to bound $\tilde{f}_{\max}$.

Let $z_s := (\mathbf{1}^\top D(P_1 - P_0)(I - P_{1/2})^{\#} D^{-1})E_s$. By the definition of $\tilde{f}$, we have

$$\tilde{f}_{\max} \leq 2(z_{\max} + 2)(V_{\max} + r_{\max})$$

where $z_{\max} := \max_s |z_s|, V_{\max} := \max_s |V(s)|$. For $V_{\max}$, we have

$$\begin{aligned}
\|V\|_\infty &= \|(I - P_{1/2})^{\#} r\|_\infty \\
&\leq \|(I - P_{1/2})\|_{1,\infty} r_{\max} \\
&\leq \frac{2\ln(C)+1}{1-\lambda} r_{\max}.
\end{aligned}$$

For $z_{\max}$, we have the following claim.

Lemma 3. *There exists a constant C' such that*

$$z_{\max} \leq C' \log\left(\frac{1}{\rho_{\min}}\right)\frac{1}{1-\lambda}.$$

Therefore, there exists a constant C'' such that

$$\sigma_f \leq C'' \log\left(\frac{1}{\rho_{\min}}\right)\left(\frac{1}{1-\lambda}\right)^{5/2} r_{\max}$$

which completes the proof.

Proof of Lemma 3. Let

$$v^\top := \mathbf{1}^\top D(P_1 - P_0) = \rho_{1/2}^\top (P_1 - P_0).$$

We claim that $|(P_1 - P_0)(s,s')| \leq 2P_{1/2}(s,s')$ for any s and s'. This is due to $2P_{1/2} = P_0 + P_1$ and for any $a \geq 0, b \geq 0$, we have

$$|a - b| \leq a + b.$$

Furthermore, note that $\rho_{1/2}^\top P_{1/2} = \rho_{1/2}^\top$. Then for any s',

$$\begin{aligned}
|v(s')| &= |\sum_s \rho_{1/2}(s)(P_1 - P_0)(s,s')| \\
&\leq \sum_s \rho_{1/2}(s)|(P_1 - P_0)(s,s')| \\
&\leq \sum_s \rho_{1/2}(s) 2P_{1/2}(s,s') \\
&\leq 2\rho_{1/2}(s').
\end{aligned}$$

This is to say, v is entry-wise bounded by $\rho_{1/2}$. Furthermore, this bound holds for any transformation under $P_{1/2}$.

$$\begin{aligned}
|(v^\top P_{1/2}^k)(s')| &= |\sum_s v(s) P_{1/2}^k(s,s')| \\
&\leq 2\sum_s \rho_{1/2}(s) P_{1/2}^k(s,s') \\
&\leq 2\rho_{1/2}(s').
\end{aligned}$$

Next, consider

$$v^\top (I - P_{1/2})^\# e_{s'} = \sum_{k=0}^{\infty} v^\top (P_{1/2}^k - \mathbf{1}\rho_{1/2}^\top) e_{s'}$$
$$=: \sum_{k=0}^{\infty} a_k.$$

Note that $|(v^\top P_{1/2}^k) e_{s'}| \leq 2\rho_{1/2}(s')$. Further, $|v^\top \mathbf{1}\rho_{1/2}^\top e_{s'}| \leq |v^\top \mathbf{1}|\rho_{1/2}(s') \leq 2\rho_{1/2}(s')$. Therefore, for any k,

$$|a_k| \leq 4\rho_{1/2}(s').$$

We also have

$$\begin{aligned} |a_k| &\leq \|v^\top\|_1 \|P^k - \mathbf{1}\rho^\top\|_{1,\infty} \|e_{s'}\|_{\max} \\ &\leq 2C\lambda^k. \end{aligned}$$

Let $\rho_{\min} := \min_s \rho_{1/2}(s)$. With $a := 4, b := 2C\frac{1}{\rho_{\min}}$, we have

$$\begin{aligned} \frac{1}{\rho_{1/2}(s')} \sum_{k=0}^{\infty} a_k &\leq \sum_{k=0}^{\infty} \min\left(4, 2C\lambda^k \frac{1}{\rho_{\min}}\right) \\ &\leq \sum_{k=0}^{\log_\lambda(1/b)-1} a + \sum_{k=\log_\lambda(1/b)} b\lambda^k \\ &= \log_\lambda(1/b) a + \frac{1}{1-\lambda} \\ &\leq \frac{\ln(b)}{1-\lambda} a + \frac{1}{1-\lambda} \\ &\lesssim \log\left(\frac{1}{\rho_{\min}}\right) \frac{1}{1-\lambda}. \end{aligned}$$

Then $v^\top (I - P_{1/2})^\# D^{-1} E_s \lesssim \log\left(\frac{1}{\rho_{\min}}\right)\frac{1}{1-\lambda}$, which completes the proof. ■

E. Proof of Theorem 3

The proof is based on a multi-variate Cramér-Rao bound. To begin, we assume $P_0(s, s') > 0, P_1(s, s') > 0$ for all (s, s').[4]

Consider the parameters $\theta = (F_0, F_1)$ which controls the transition matrices

$$P_0(s, s') = \frac{F_0(s, s')}{\sum_{s''} F_0(s, s'')}, \quad P_1(s, s') = \frac{F_1(s, s')}{\sum_{s''} F_1(s, s'')}.$$

Given the observations $X_t = (s_t, a_t), t = 0, 1, \ldots, T$ under the policy $\pi_{1/2}$. We can compute the log-likelihood

$$l(X_1, \ldots, X_T \mid \theta) = \left(\sum_{s,a,s'} n_{s,a,s'} \cdot \ln(P_a(s, s')) \right) - T\ln(2)$$

[4] The general case follows a similar proof and is omitted for simplicity.

where $n_{s,a,s'} = \sum_t 1(s_t = s, a_t = a, s_{t+1} = s')$. Then, the entry of the Fisher information matrix with $\theta^* = (P_0, P_1)$ is given by

$$
\begin{aligned}
I_{k,m} &= -\mathsf{E}_X \left[\frac{\partial l(X|\theta^*)}{\partial \theta_k \partial \theta_m} \right] \\
&= -\mathsf{E}_X \left[\sum_{s,a,s'} \frac{n_{s,a,s'}}{P_a(s,s')} \cdot \frac{\partial P_a(s,s')}{\partial \theta_k \partial \theta_m} \right] + \mathsf{E}_X \left[\sum_{s,a,s'} \frac{n_{s,a,s'}}{P_a(s,s')^2} \cdot \frac{\partial P_a(s,s')}{\partial \theta_k} \frac{\partial P_a(s,s')}{\partial \theta_m} \right] \\
&= -T \sum_{s,a,s'} \frac{1}{2} \rho_{1/2}(s) \cdot \frac{\partial P_a(s,s')}{\partial \theta_k \partial \theta_m} + T \sum_{s,a,s'} \frac{1}{2} \frac{\rho_{1/2}(s)}{P_a(s,s')} \cdot \frac{\partial P_a(s,s')}{\partial \theta_k} \frac{\partial P_a(s,s')}{\partial \theta_m} \\
&= -T \frac{\partial 1}{\partial \theta_k \partial \theta_m} + T \sum_{s,a,s'} \frac{1}{2} \frac{\rho_{1/2}(s)}{P_a(s,s')} \cdot \frac{\partial P_a(s,s')}{\partial \theta_k} \frac{\partial P_a(s,s')}{\partial \theta_m} \\
&= T \sum_{s,a,s'} \frac{1}{2} \frac{\rho_{1/2}(s)}{P_a(s,s')} \cdot \frac{\partial P_a(s,s')}{\partial \theta_k} \frac{\partial P_a(s,s')}{\partial \theta_m}.
\end{aligned}
$$

Consider $\theta_k = F_0(i,j), \theta_m = F_0(i,l)$, we have

$$\frac{1}{T} I_{k,m} = \frac{1}{2} \frac{\rho_{1/2}(i)}{P_0(i,j)} 1(j = l) - \frac{1}{2} \rho_{1/2}(i).$$

For $\theta_k = F_1(i,j), \theta_m = F_1(i,l)$, we have

$$\frac{1}{T} I_{k,m} = \frac{1}{2} \frac{\rho_{1/2}(i)}{P_1(i,j)} 1(j = l) - \frac{1}{2} \rho_{1/2}(i).$$

Otherwise it is easy to see that $I_{k,m} = 0$.

Next, consider an unbiased estimator $\hat{\tau}(X_1, \ldots, X_T)$ for ATE. We can write ATE $= f(F_0, F_1)$ as a function of F_0 and F_1. Further, one can verify that

$$
\begin{aligned}
\frac{\partial f(\theta^*)}{\partial F_0(i,j)} &= -\rho_0(i)(V_{\pi_0}(j) - V_{\pi_0}(i) + r_0(i) - \lambda^{\pi_0}) \\
\frac{\partial f(\theta^*)}{\partial F_1(i,j)} &= \rho_1(i)(V_{\pi_1}(j) - V_{\pi_1}(i) + r_1(i) - \lambda^{\pi_1}).
\end{aligned}
$$

Finally, we aim to use the multi-variate Cramér-Rao bound. To do so, let $v_i^{(1)}$ be an vector with the j-th element being $v_i^{(1)}(j) = \rho_1(i)(V_{\pi_1}(j) - V_{\pi_1}(i) + r_1(i) - \lambda^{\pi_1})$. Let

$$I_i^{(1)}(j,l) = \frac{T}{2} \frac{\rho_{1/2}(i)}{P_1(i,j)} 1(j = l) - \frac{T}{2} \rho_{1/2}(i)$$

be a matrix. Similarly, define $v_i^{(0)}$ and $I_i^{(0)}$ accordingly. Then, by the multi-variate Cramér-Rao bound for the singular Fisher information matrix [54], we have

$$
\begin{aligned}
T\mathrm{Var}(\hat{\tau}) &\geq \sum_i v_i^{(1)\top} (I_i^{(1)})^{-1} v_i^{(1)} + \sum_i v_i^{(0)\top} (I_i^{(0)})^{-1} v_i^{(0)} \\
&= 2 \sum_i \frac{\rho_0(i)^2}{\rho_{1/2}(i)} \sum_j P_0(i,j)(V_{\pi_0}(j) - V_{\pi_0}(i) + r_0(i) - \lambda^{\pi_0})^2 \\
&\quad + 2 \sum_i \frac{\rho_1(i)^2}{\rho_{1/2}(i)} \sum_j P_1(i,j)(V_{\pi_1}(j) - V_{\pi_1}(i) + r_1(i) - \lambda^{\pi_1})^2
\end{aligned}
$$

which completes the proof.

E.1. Unbiased Estimator that achieves the lower-bound

In this section, we construct an LSTD(0)-type OPE estimator that achieves the aforementioned Cramér-Rao lower bound. To do so, we solve the following least square optimization problems that are similar to Eq. (5),

$$(\hat{V}_1, \hat{\lambda}^{\pi_1}) = \arg\min_{\hat{V},\hat{\lambda}} \sum_{s\in\mathcal{S}} \left(\sum_{t, s_t = s, a_t = 1} r(s_t, a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t) \right)^2 \tag{14}$$

$$(\hat{V}_0, \hat{\lambda}^{\pi_0}) = \arg\min_{\hat{V},\hat{\lambda}} \sum_{s\in\mathcal{S}} \left(\sum_{t, s_t = s, a_t = 0} r(s_t, a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t) \right)^2 . \tag{15}$$

Then, the estimation for the average treatment effect is given by

$$\tau_{\text{off}} := \hat{\lambda}^{\pi_1} - \hat{\lambda}^{\pi_0}.$$

To analyze the variance of $\hat{\tau}$, we follow the similar analysis as in Theorem 2. To begin, one can verify that

$$\hat{\lambda}^{\pi_0} - \lambda^{\pi_0} = \left(\hat{\rho}_0^\top - \rho_0^\top\right) r_0$$

where $\hat{\rho}_0$ is the empirical stationary distribution for the empirical transition matrix $\hat{P}_0$ ($\hat{\rho}_1$ and $\hat{P}_1$ can be defined accordingly).

Next, by the perturbation bound of $\hat{\rho}_0$, we have

$$\hat{\rho}_0^\top - \rho_0^\top = \rho_0^\top (\hat{P}_0 - P_0)(I - \hat{P}_0)^{\#}.$$

Hence,

$$\begin{aligned} \hat{\lambda}_0 - \lambda^{\pi_0} &= (\hat{\rho}_0^\top - \rho_0^\top) r_0 \\ &= \rho_0^\top (\hat{P}_0 - P_0)(I - \hat{P}_0)^{\#} r_0. \end{aligned}$$

Note that $\hat{P}_0$ is a function of $F_T^{(0)}$ ($\hat{P}_0(i,j) = F_T^{(0)}(i,j)/\sum_k F_T^{(0)}(i,k)$, $F^{(0)}$ is defined in Eq. (12)). Therefore, we can define $f_0(F_T^{(0)}) := \hat{\lambda}_0 - \lambda^{\pi_0}$ as a function of $F_T^{(0)}$. Similarly, we can define

$$f_1(F_T^{(1)}) := \hat{\lambda}_1 - \lambda^{\pi_1} = \rho_1^\top (\hat{P}_1 - P_1)(I - \hat{P}_1)^{\#} r_1$$

Then by Lemma 6, we have the asymptotic normality for τ_{off}:

$$\sqrt{T}(\tau_{\text{off}} - \text{ATE}) = \sqrt{T}(f_1(F_T^{(1)}) - f_0(F_T^{(0)})) \xrightarrow{d} N(0, \sigma_{\text{off}}^2).$$

In order to compute σ_{off} by using Lemma 6, we will linearize $f_1 - f_0$ around (F_0^*, F_1^*). To do so, consider

$$\begin{aligned} \frac{\partial f_0(F_0)}{\partial (F_0)(i,j)} &= \rho_0^\top \frac{\partial (\hat{P}_0 - P_0)}{\partial F_0(i,j)} (I - P_0)^{-1} (r_0 - \lambda^\pi \mathbf{1}) \\ &\quad + \rho_0^\top (P_0 - P_0) \frac{\partial (I - P_0)^{-1}}{\partial (F_0)(i,j)} (r_0 - \lambda^\pi \mathbf{1}) \\ &= \rho_0^\top \frac{\partial \hat{P}_0}{\partial F_0(i,j)} V_0 \\ &= \sum_k \rho_0(i) V_0(k) \frac{\partial \hat{P}_0(i,k)}{\partial F_0(i,j)} \end{aligned}$$

Note that $\hat{P}(i,k) = \hat{F}_0(i,k)/\sum_l \hat{F}_0(i,l)$. Therefore,

$$\begin{aligned}
\frac{\partial f_0(F_0)}{\partial (F_0)(i,j)} &= \sum_k \rho_0(i) V_0(k) \frac{\partial \frac{F_0(i,k)}{\sum_l F_0(i,l)}}{\partial F_0(i,j)} \\
&= \sum_k \rho_0(i) V_0(k) \frac{1(j=k)\sum_l F_0(i,l) - F_0(i,k)}{(\sum_l F_0(i,l))^2} \\
&= \sum_k \rho_0(i) V_0(k) \frac{1(j=k)\rho(i) - \rho(i) P_0(k|i)}{\rho(i)^2} \\
&= \frac{\rho_0(i)}{\rho(i)} V_0(j) - \frac{\rho_0(i)}{\rho(i)} \sum_k P_0(k|i) V_0(k) \\
&= \frac{\rho_0(i)}{\rho(i)} (V_0(j) - V_0(i) + r_0(i) - \lambda^{\pi_0}).
\end{aligned}$$

Hence, the linearization of f_0 is

$$\begin{aligned}
&\sum_{ij} \frac{\partial f_0(F_0)}{\partial (F_0)(i,j)} \Big((F_0(s,s',a))_{ij} - F_0(i,j) \Big) \\
&= 2 \cdot 1(a=0) \frac{\rho_0(s)}{\rho(s)} (V_0(s') - V_0(s) + r_0(s) - \lambda^{\pi_0}) - \sum_{ij} \rho_0(i) (V_0(j) P_0(j|i) - V_0(i) + r_0(i) - \lambda^{\pi_0}) \\
&= 2 \cdot 1(a=0) \frac{\rho_0(s)}{\rho(s)} (V_0(s') - V_0(s) + r_0(s) - \lambda^{\pi_0}).
\end{aligned}$$

The similar linearization can be done for f_1. Then the linearization of $f_1 - f_0$ is

$$\begin{aligned}
g((s,s',a)) = &-2 \cdot 1(a=0) \frac{\rho_0(s)}{\rho(s)} (V_0(s') - V_0(s) + r_0(s) - \lambda^{\pi_0}) \\
&+ 2 \cdot 1(a=1) \frac{\rho_1(s)}{\rho(s)} (V_1(s') - V_1(s) + r_1(s) - \lambda^{\pi_1}).
\end{aligned}$$

Note that for any $E[g(X_k)|X_1 = (s,s',a)] = 0$ for any (s,s',a) and $k \geq 2$. Hence

$$\begin{aligned}
\sigma^2_{\text{off}} &= \text{Var}_\rho(g) + 2 \sum_{k=2}^{\infty} \text{Cov}_\rho[g(X_k) g(X_1)] \\
&= \text{Var}_\rho(g) \\
&= 2 \sum_{s,s'} \frac{\rho_0(s)^2 P_0(s'|s)}{\rho(s)} (V_0(s') - V_0(s) + r_0(s) - \lambda^{\pi_0})^2 \\
&\quad + 2 \sum_{s,s'} \frac{\rho_1(s)^2 P_1(s'|s)}{\rho(s)} (V_1(s') - V_1(s) + r_1(s) - \lambda^{\pi_1})^2
\end{aligned}$$

which completes the proof.

F. Proof of Theorem 4

We construct a birth-death Markov chain with n states. Let $P \in \mathbb{R}^{n \times n}$ be a transition matrix where $P(s,s+1) = \frac{1}{4} - \delta, P(s,s-1) = \frac{1}{4}$ and $P(s,s) = 1/2 + \delta$ (except at the two ends with $P(0,0) = 3/4 + \delta$ and $P(n-1,n-1) = 3/4$).

Let the stationary distribution of P be ρ. Then $\rho(s) = c\,(1-4\delta)^s$ for $0 \leq s \leq n-1$ and $c := \frac{1}{\sum_s (1-4\delta)^s}$ is a constant. By [10], we have that the spectral gap of the chain is on the order of $\gamma = O(1/n)$. Furthermore, the mixing time of the chain is bounded by

$$\|P^k - \mathbf{1}\rho^\top\|_{1,\infty} \leq \left(\frac{1}{\rho_{\min}}\right)(1-\gamma)^k$$
$$\|(I-P)^{\#}\|_{1,\infty} \leq \log\left(\frac{1}{\rho_{\min}}\right) O(n) = O(n^2).$$

Following the same proof in Theorem 2, we have that the on-policy variance is bounded by

$$\sigma_{\text{on}} = O(n^6).$$

On the other hand, consider the node k where $\sum_{s=k}^n \rho(s) \leq c'\delta/n^2$ and $\sum_{s=k-1}^n \rho(s) > c'\delta/n^2$ for some sufficiently small constant c'. Let P_1 be the same as P except $\forall s \geq k$

$$P_1(s, s+1) = \frac{1}{4}$$
$$P_1(s, s) = \frac{1}{2}.$$

Let ρ_1 be the stationary distribution of P_1. One can verify that $\rho_1(n) = O(1/n^2)$. We then construct r such that $r(n, 1) = 1$ and $\lambda^{\pi_1} = 0$. Then

$$\sigma_{\text{off}} \geq \sqrt{2\frac{\rho_1(n)^2}{\rho(n)}\frac{3}{4}}$$
$$= \Omega\left(\frac{e^{cn}}{n^2}\right)$$

for some constant c. Therefore,

$$\frac{\sigma_{\text{on}}}{\sigma_{\text{off}}} = O\left(\frac{n^8}{e^{cn}}\right).$$

Next, consider the bias of the DQ estimator. Suppose ATE $= \delta$ without loss (one can always achieve this by adding some constants to r). Let $P_0 = 2 \cdot P_1 - P$ and let ρ_0 be the stationary distribution of P_0. One can verify that

$$\|\rho_1 - \rho\|_1 = O(\delta/n^2), \|\rho_0 - \rho\|_1 = O(\delta/n^2).$$

Furthermore, following the proof in Theorem 1, we have

$$|(\text{ATE} - \mathsf{E}[\hat{\text{ATE}}_D])/\text{ATE}| \leq (\|\rho_1 - \rho\|_1 + \|\rho_0 - \rho\|_1)\|I - P\|_{1,\infty}^{\#}$$
$$\leq C \cdot c'\delta\frac{1}{n^2}n^2$$
$$\leq \delta$$

for sufficiently small constant c'. This completes the proof.

G. Technical Lemmas

Lemma 2. *Suppose $P \in \mathbb{R}^{n\times n}$ is the transition matrix of a finite-state aperiodic and irreducible Markov Chain and ρ is the stationary distribution. Suppose there exists C and λ such that for any $k = 0, 1, \ldots$*

$$\|P^k - \mathbf{1}\rho^\top\|_{1,\infty} \leq C\lambda^k.$$

Then

$$\|(I-P)^{\#}\|_{1,\infty} \leq \frac{2\ln(C) + 1}{1-\lambda}.$$

Proof. Note that

$$\begin{aligned}A &= (I - P + 1\rho^\top)^{-1} - 1\rho^\top \\ &= \sum_{k=0}^{\infty} \left(P^k - 1\rho^\top\right).\end{aligned}$$

Then

$$\begin{aligned}\|A\|_{1,\infty} &\le \sum_{k=0}^{\infty} \|P^k - 1\rho^\top\|_{1,\infty} \\ &\le \sum_{k=0}^{\infty} \min\left(2, C\lambda^k\right) \\ &\le \sum_{k=0}^{\log_\lambda(1/C)-1} 2 + \sum_{k=\log_\lambda(1/C)}^{\infty} C\lambda^k \\ &\le 2\log_\lambda(1/C) + \frac{1}{1-\lambda} \\ &= 2\frac{\ln(C)}{-\ln(\lambda)} + \frac{1}{1-\lambda} \\ &\overset{(i)}{\le} \frac{2\ln(C)+1}{1-\lambda}\end{aligned}$$

where (i) is due to $-\ln(x) \le 1 - x$ for $x > 0$. ■

Lemma 4. *For a finite-state aperiodic and irreducible Markov Chain $X_1, X_2, \ldots, X_t$. Let P be the transition matrix, ρ be the stationary distribution, and $\mathcal{S}$ be the state space. Suppose there exists C and λ such that for $k = 0, 1, \ldots,$*

$$\|P^k - \mathbf{1}\rho^\top\|_{1,\infty} \le C\lambda^k.$$

Then for any bounded function $f : \mathcal{S} \to [a, b]$, there exists σ such that when T goes to infinity,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(f(X_t) - f^*\right) \xrightarrow{d} N(0, \sigma^2) \tag{16}$$

where $f^* = \mathsf{E}_\rho(f)$ is the expected value of f under the stationary distribution and

$$\sigma \le \sqrt{2}(b-a)\sqrt{\frac{2\ln(C)+1}{1-\lambda}}. \tag{17}$$

Proof. Note that Eq. (16) is simply due to the Markov chain CLT ([28]). Let D be an diagonal matrix with entries $D_{ii} = \rho_i$. [28] further states that

$$\begin{aligned}\sigma^2 &= \mathrm{Var}_\rho(f) + 2\sum_{k=2}^{\infty} \mathsf{E}_\rho[(f(X_1) - f^*)(f(X_k) - f^*)] \\ &= (f - f^*)^\top D(f - f^*) + 2\sum_{k=1}^{\infty} (f - f^*)^\top D P^k (f - f^*) \\ &= 2\sum_{k=0}^{\infty} (f - f^*)^\top D(P^k - \mathbf{1}\rho^\top)(f - f^*) - (f - f^*)^\top D(f - f^*) \\ &\le 2\sum_{k=0}^{\infty} (f - f^*)^\top D(P^k - \mathbf{1}\rho^\top)(f - f^*) \\ &\le 2\|(f - f^*)^\top D\|_1 \|I - P\|_{1,\infty}^{\#} \|f - f^*\|_{\max} \\ &\le 2\|f - f^*\|_{\max}^2 \frac{2\ln(C)+1}{1-\lambda}.\end{aligned}$$

Therefore,

$$\sigma \leq \sqrt{2}(b-a)\sqrt{\frac{2\ln(C)+1}{1-\lambda}}.$$

■

Lemma 5 (Theorem 6.2 [38]). *Let U_k be a sequence of random variables in $\mathbb{R}^p$ converging in probability to u. Let a_k be a deterministic non-negative sequence increasing to ∞. Let $\sqrt{\alpha_k}(U_k - u)$ converge in distribution to $N(0,\Gamma)$. Let $f : R^p \to R^q$ be a function twice differentiable in a neighborhood of u. Then, denoting the Jacobian of f at u by $\nabla f(u)$, we have*

1. *$f(U_k)$ converges in probability to $f(u)$.*
2. *$\sqrt{\alpha_k}(f(U_k) - f(u))$ converges in distribution to $N(0, \nabla f(u^*)\Gamma\nabla f(u^*)^\top)$.*

Lemma 6. *Consider an irreducible and aperiodic finite-state space Markov Chain $X_1, X_2, \ldots, X_t$. Let S be the state space and ρ be the stationary distribution. Let $u : S \to \mathbb{R}^p$ be a function with each component $u_i, 1 \leq i \leq p$. Let $u^* = \sum_{s\in S}\rho(s)u(s)$ be the expected value of u under the stationary distribution ρ.*

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function twice differentiable in a neighbor of u^*. Then, there exists $\sigma \geq 0$ such that when $T \to \infty$,

$$\sqrt{T}\left(f\left(\frac{1}{T}\sum_{i=1}^{T} u(X_t)\right) - f(u^*)\right) \xrightarrow{d} N(0,\sigma^2)$$
$$\sqrt{T}\left(\sum_{i=1}^{p}(u_i(X_t) - u_i^*)\cdot\frac{\partial f(u^*)}{\partial u_i}\right) \xrightarrow{d} N(0,\sigma^2)$$

Proof. To begin, note that by the Markov Chain CLT (Corollary 5 [28]), we have

$$\sqrt{T}\left(\frac{1}{T}\sum_{i=1}^{T} u(X_t) - u^*\right) \xrightarrow{d} N(0,\Sigma)$$

for some covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. In particular,

$$\Sigma := E_\rho[(u(X_1) - u^*)(u(X_1) - u^*)^\top] + 2\sum_{k=2}^{\infty} E_\rho[(u(X_1) - u^*)(u(X_k) - u^*)^\top] \tag{18}$$

where E_ρ denotes the expectation when the initial distribution of the Markov chain is ρ.

Then, since f is twice differentiable in a neighborhood of u^*, we can invoke Lemma 5 to get

$$\sqrt{T}\left(f\left(\frac{1}{T}\sum_{i=1}^{T} u(X_t)\right) - f(u^*)\right) \xrightarrow{d} N(0,\sigma^2)$$

where $\sigma^2 := \nabla f(u^*)^\top\Sigma\nabla f(u^*)$.

Next, let $F(X) := \sum_{i=1}^{p}(u_i(X) - u_i^*)\cdot\frac{\partial f(u^*)}{\partial u_i} = (u(X) - u^*)^\top\nabla f(u^*)$. Then using the fact $\frac{1}{T}\sum_{t=1}^{T} u(X_t) \to u^*$ and invoking the Markov chain CLT again, we have

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T} F(X_t)\right) \xrightarrow{d} N(0,\sigma_F^2)$$

where

$$\sigma_F^2 := E_\rho[F(X_1)^2] + 2\sum_{k=2}^{\infty} E_\rho[F(X_1)F(X_k)].$$

Expanding $F(X)$ by $(u(X) - u^*)^\top \nabla f(u^*)$, we have

$$\begin{aligned}
\sigma_F^2 &= E_\rho[((u(X_1) - u^*)^\top \nabla f(u^*))^2] + 2\sum_{k=2}^{\infty} E_\rho[(u(X_1) - u^*)^\top \nabla f(u^*)(u(X_k) - u^*)^\top \nabla f(u^*)] \\
&= \nabla f(u^*)^\top E_\rho[(u(X_1) - u^*)(u(X_1) - u^*)^\top] \nabla f(u^*) \\
&\quad + \nabla f(u^*)^\top \sum_{k=2}^{\infty} E_\rho[(u(X_1) - u^*)(u(X_k) - u^*)^\top] \nabla f(u^*) \\
&\overset{(i)}{=} \nabla f(u^*)^\top \Sigma \nabla f(u^*) \\
&= \sigma^2
\end{aligned}$$

where (i) uses Eq. (18). This implies that F (the linearization of f at the point u^*) will converge with the same limiting variance as f. ■

H. Experiment details

H.1. Synthetic example

H.1.1. Environment

We replicate exactly the environment of [26]. We model a rental marketplace with $N = 5000$ homogeneous listings. Customers arrive according to a Poisson process with rate $N\lambda$, decide whether to rent a listing (with rental probability controlled by the intervention), and if they do rent, they occupy a listing for an exponentially distributed time with mean $\frac{1}{\mu}$.

Specifically, we define our MDP to be the discrete-time jump chain of this process, with events indexed by t and state $s_t \in \{0, 1 \ldots N\}$ representing the current inventory of listings. At the t^{th} event, the system chooses to apply control ($a_t = 0$) or treatment ($a_t = 1$). One of the following state transition and reward scenarios may then happen:

1. A previously occupied rental becomes available, i.e. $s_{t+1} = s_t + 1$ and $r_t = 0$; this occurs with probability $\frac{(N - s_t)\mu}{N\mu + N\lambda}$
2. A customer arrives, with probability $\frac{N\lambda}{N\mu + N\lambda}$, and subsequently:
   (a) Rents a listing, so $s_{t+1} = s_t - 1$ and $r_t = 1$; this occurs with probability $\frac{s_t v(a_t)}{N + s_t v(a_t)}$ conditional on a customer arrival, where $v(0) = 0.315$ and $v(1) = 0.3937$ are the average utility under control and treatment, respectively.
   (b) Does not rent a listing , so $s_{t+1} = s_t$ and $r_t = 0$; this occurs with probability $\frac{N}{N + s_t v(a_t)}$ conditional on a customer arrival.
3. No state change occurs; i.e. $s_{t+1} = s_t$ and $r_t = 0$.

[26] also describes a two-sided randomization scheme, where listings are also assigned to control or treatment, and the customer's purchase probability depends on both the customer's treatment assignment a_t, as well as the number of control listings and the number of treatment listings. This corresponds to a more complicated MDP with a two-dimensional state $s_t = (s_t^{\text{co}}, s_t^{\text{tr}})$, where s_t^{co} corresponds to the number of available control listings, and s_t^{tr} the number of available treatment listings. The average utility of a control listing is $v_{\text{co}}(0) = v_{\text{co}}(1) = v(0)$, while the average utility of a treatment listing is $v_{\text{tr}}(0) = v(0)$ and $v_{\text{tr}}(1) = v(1)$. We defer to [26] for further details of this scheme.

H.1.2. Implementation details

Here we list algorithms and hyperparameters tuned for this experiment. Hyperparameters were chosen to minimize MSE averaged over 10 held-out trajectories. As in [26], we also include a burn-in period of $T_0 = 5N$.

1. Naive. This has no hyperparameters.
2. TSRI. This has several hyperparameters, which affect both the experimental design (customer randomization probability p and listing randomization probability p_L), as well as the estimator (parameters k and β, as described in [26]). We set p, p_L, β assuming λ, μ are known, exactly as prescribed in [26]. Specifically, we compute the values reported in Table 1 as:

$$p = \left(1 - e^{-\lambda/\mu}\right) + 0.5e^{-\lambda/\mu} \qquad p_L = 0.5\left(1 - e^{-\lambda/\mu}\right) + e^{-\lambda/\mu} \qquad \beta = e^{-\lambda/\mu}$$

   We report results for both $k = 1$ and $k = 2$.
3. DQ with LSTD, which we estimate using a slight modification of Equation (5). Specifically, we directly estimate the state-action value function Q instead of separately estimating the state value function V and P_1, P_0, and we add an L_2 regularization term. In short, we approximate and solve for a fixed point to the regularized least-squares problem:

$$Q = \arg\min_{Q'} \|Q' - r - PQ + \lambda\|_2^2 + \xi\|Q'\|_2^2$$

   where $Q \in \mathbb{R}^{2(N+1)}$ is the vector of estimated $Q(s, a)$ values, and$P \in \mathbb{R}^{2(N+1)\times 2(N+1)}$ is the state-action transition matrix. We use sample means in each state to construct plug-in estimates of r, P and λ.
4. Off-Policy with LSTD, which we note is novel in the average reward literature. In Section **??** we describe this algorithm, provide convergence guarantees, and show that this algorithm is efficient. This can be construed as a direct analog of [52]'s off-policy estimator, which applies LSTD in the discounted-reward setting. It has no hyperparameters.
5. Off-Policy with TD, where Q -functions and off-policy average rewards are calculated according to the Differential TD algorithm of [65]. This approach has two hyperparameters: the learning rate for the Q -function $\gamma/\sqrt{t}$, and the learning rate for the mean reward estimate $\beta\gamma/\sqrt{t}$.

For these experiments, we exclude the Off-Policy GTD variant described in [69] as their convergence guarantees do not apply to the tabular setting.

Algorithm	Hyperparameters
TSRI	$p = \mathbf{0.816}, p_L = \mathbf{0.683}, k \in \{\mathbf{1}, \mathbf{2}\}, \beta = \mathbf{0.368}$
DQ (LSTD)	$\xi \in \{0.01, \mathbf{0.1}, 1, 10, 100\}$
Off-Policy (TD)	$\beta \in \{0.2, \mathbf{0.5}\}, \gamma \in \{0.001, \mathbf{0.01}, 0.1, 1.\}$

Table 1: Hyperparameters for the synthetic example of [26]. Parameter settings reported in the main text are in bold.

H.1.3. Additional results

We note that there are scenarios for which which specialized designs and estimators – specifically TSR, in this example – can provide a superior bias-variance tradeoff. [26] shows that the TSRI estimators become unbiased when $\lambda \gg \mu$. We ran the synthetic example setting $\lambda = 10, \mu = 1$ (also mirroring results from [26]), and indeed for this setting for reasonable horizons TSR achieves lower RMSE. Recall, however, that TSR is ill-defined for settings where there is no natural notion of two-sided randomization (i.e. in any MDP without a notion of two sides), and its bias properties are clearly highly instance-specific and depend on knowledge of λ, μ. DQ still outperforms all alternatives besides TSR in this setting, and even in this extremely unbalanced setting bachieves a much lower asymptotic bias than TSR (-5e-3 vs 1e-2, as a proportion of the treatment effect magnitude). .

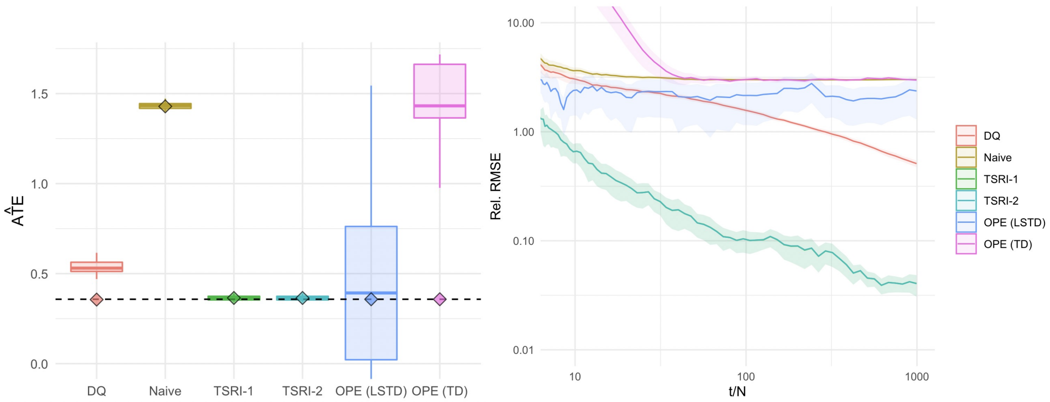

Figure 3: Toy-example from [26], with $\lambda = 10$. *Left*: Estimated ATE at time $t/N = 10^3$ across 100 trajectories. Dashed line indicates actual ATE. Diamonds indicate the asymptotic mean for each estimator. Over this horizon, TSRI-1 and TSRI-2 exhibit small bias and variance, although asymptotically DQ still has lower bias.

H.1.4. Computing environment

These experiments were performed on a personal desktop with a 24-core Intel Xeon X5670 CPU and 128 GB RAM. Total compute time per seed averaged less than two hours.

H.2. Ridesharing Simulator

H.2.1. Environment

We implement a ridesharing simulator, with code available on Github.

1. Riders are generated based on trips resampled from the NYC Taxi Dataset [1], with a random willingness-to-pay per second distributed as LogNormal$(\log(0.01), 1.)$. The the rider's outside option is assumed to be the trip they actually took in the dataset, and the cost (i.e., negative utility) the rider incurs for this option is the fare recorded in the dataset, plus the trip time times the rider's WTP per second.
2. Drivers enter the system at pickup locations in the same dataset, but at a lower arrival rate (tuned to achieve a utilization of $\sim 70\%$). Drivers stay in the system for an exponential time with a mean of two hours, and stop serving new requests once they exit the system.
3. When a request enters the system, the pricing engine computes the cost to serve that request with an idle driver (where cost is based on recent per-mile and per-minute fare rates), and discounts this by 10%; this is the price offered to the rider. The pricing engine also offers the rider a worst-case time-to-destination (ETD) guarantee, which is 1.5 times the time to serve the request with an idle driver. The rider then chooses to accept or reject the offer, based on whether their worst-case utility for the trip exceeds the utility of the outside option. If the rider rejects the offer they exit the system.
4. If the rider accepts, the request is submitted to the dispatch engine. The dispatcher searches for the nearest idle driver and the 10 nearest pool drivers to the request. This list of candidates is filtered to those who can serve the request while satisfying

the ETD guarantees of all riders. The pool candidates are then further filtered to those whose cost to service the request is at most $\frac{1}{1+\alpha_t}$ times the cost of the idle driver, where $\alpha_t = \alpha_{\text{co}} = 0$ in control ($a_t = 0$) and $\alpha_t = \alpha_{\text{tr}}$ in treatment ($a_t = 1$), where we vary $\alpha_{\text{tr}} \in \{0.3, 0.5, 0.7\}$. Finally, the minimum cost driver among this set is dispatched.

We can implement two-sided randomization in this market as follows. Each driver is also randomized into either treatment or control. The dispatcher then dispatches to the minimum cost driver among the following set:

- All idle drivers (i.e., drivers currently assigned no passengers).
- Control pool drivers, whose cost is at most $\frac{1}{1+\alpha_{\text{co}}}$ times the minimum cost idle driver.
- Treatment pool drivers, whose cost is at most $\frac{1}{1+a_t\alpha_{\text{tr}}+(1-a_t)\alpha_{\text{co}}}$ times the minimum cost idle driver.

H.2.2. Estimators

We use the same approximation architecture for each algorithm, where $Q(s,a) = \theta^\top \phi(s,a)$ is a linear function of features $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ with coefficients θ. We take features $\phi(s_t, a_t)$ to consist of the number of drivers in the system with each of 0, 1, 2, and 3 open seats remaining, as well as the price and cost of the current request, and an indicator variable for the action taken.

The estimators are then:

1. Naive, with no hyperparameters.
2. TSRI, again with hyperparameters p, p_L, k, β. We set these based on the relative supply and demand characteristics of the simulator. Specifically, with analogy to the synthetic problem, the system averages around 600 drivers active at any moment, with 3 passenger seats per driver, for a total of $N \approx 1800$ available units of capacity. The arrival rate is 4 passengers per second, yielding $\lambda \approx 4/1800$, while the average trip lasts 12 minutes, yielding $\mu \approx 720$. Ultimately we have $\lambda/\mu \approx 1.6$, and set the algorithm hyperparameters accordingly.
3. DQ with LSTD, with a single regularization hyperparameter ξ. Here we solve for θ by approximating and solving for a fixed point to the regularized least-squares problem:
$$\theta = \arg\min_{\theta'} \|\Phi\theta' - r - P\Phi\theta + \lambda\|_2^2 + \xi\|\theta'\|_2^2$$
where $\Phi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ is the matrix of state-action feature representations.
4. Off-Policy with LSTD, where we solve simultaneously for θ_1, λ_1 by solving for the unique fixed point of the projected Bellman equation $\Phi_1^\top \Phi_1\theta_1 = \Phi_1^\top(r_1 - \mathbf{1}\lambda_1) + \Phi_1^\top P_1\Phi_1\theta_1$, where $\Phi_1 \in \mathbb{R}^{|\mathcal{S}|}$ is the matrix of state-action features corresponding to action 1, and $r_1 \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of rewards for action 1. We solve an analogous equation for θ_0, λ_0. This effectively extends the algorithm of Section **??** to the setting of linear function approximation. This has no hyperparameters.
5. Off-Policy with TD, where Q -functions and off-policy average rewards are calculated according to the extension of [65] to linear function approximation, as provided in [69]. This approach has two hyperparameters: the learning rate for the Q -function $\gamma/\sqrt{t}$, and the learning rate for the mean reward estimate $\beta\gamma/\sqrt{t}$.
6. Off-Policy with Gradient TD (GTD), as in [69]. This has the same hyperparameters β, γ as TD.

A single hyperparameter was selected for each algorithm across all treatment effect settings, based on a scalarization of MSE across all settings, and tuned on 10 held-out trajectories for each setting.

Algorithm	Hyperparameters
TSRI	$p = \mathbf{0.9}, p_L = \mathbf{0.6}, k \in \{\mathbf{1}, \mathbf{2}\}, \beta = \mathbf{0.2}$
DQ (LSTD)	$\xi \in \{0.01, 0.1, \mathbf{1}, 10, 100\}$
Off-Policy (TD)	$\beta \in \{0.2, \mathbf{0.5}\}, \gamma \in \{0.001, \mathbf{0.01}, 0.1, 1.\}$
Off-Policy (GTD)	$\beta \in \{0.2, \mathbf{0.5}\}, \gamma \in \{0.001, \mathbf{0.01}, 0.1, 1.\}$

Table 2: Hyperparameters for the ridesharing setting. Parameter settings reported in the main text are in bold.

H.2.3. Computing environment

These experiments were performed on an internal cluster. Each run of the simulator took an average of around four hours, allocating a single CPU and 8GB of RAM.