Markovian Interference in Experiments

Vivek F. Farias

Operations Research Center, Massachusetts Institute of Technology

Andrew A. Li

Tepper School of Business, Carnegie Mellon University

Tianyi Peng

Department of Aeronautics and Astronautics, Massachusetts Institute of Technology

Andrew Zheng

Operations Research Center, Massachusetts Institute of Technology

Abstract

We consider experiments in dynamical systems where interventions on some experimental units impact other units through a limiting constraint (such as a limited inventory). 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. From a theoretical perspective, we introduce three separate novel techniques that are of independent interest in the theory of Reinforcement Learning (RL). 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 simple to execute, in principle: apply the treatment in question at random (e.g. an A/B test), and 'naively' infer the average effect of the treatment by differencing the average outcomes under treatment and control. About a decade ago, Blake and Coey [3] 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."

This violation of the so-called Stable Unit Treatment Value Assumption (SUTVA) [6] has been viewed as problematic in online platforms as early as Reiley's seminal 'Magic on the Internet' work [26]. Blake and Coey [3] 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 share a common inventory of 'demand' or 'supply' depending on context.

Despite the ubiquity of such interference, a practical solution is far from settled. An ongoing line of work addresses the problem via *experimental design*, assigning treatments carefully to mitigate the bias of 'naively'-derived estimators. In the best cases such designs provably reduce bias by exploiting certain application specific structures, but often it is unclear whether the problem at hand affords such structure (a case in point being the search-engine example above, as will be apparent later). As such, experimentation on online platforms still largely relies on simple randomization, i.e. A/B tests. Motivated by this fact, we focus instead on *designing effective estimators* assuming simple randomization. We demonstrate a novel estimator which, thanks to an effective bias-variance tradeoff, is a compelling alternative to both alternative state-of-the-art estimators as well as bespoke experimental designs when they apply.

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 [31] or explicitly [14, 35]. 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 momentarily in Section 1.1), 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. [41, 40, 25, 12, 19, 20]. This tack appears to be promising, e.g. [35], 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. Our analysis introduces a novel Taylor-like expansion of the ATE (Theorem 5) that in addition to the current setting, is of general interest in the theory of RL (for instance, in the context of Policy Optimization).

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. Our analysis introduces two new techniques. First, we prove what we dub an 'Entrywise Non-expansive Lemma', that we believe is crucial to elucidating the variance reduction afforded by on-policy methods. Second, we introduce a novel linearization trick which dramatically simplifies the analysis of variance in RL via the delta method.

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 (e.g., this leads to a new approach with a provably lower bias than some widely used methods in policy optimization, see Section 6.3).

3. Practical Performance: Despite the technical novelty described above, we view this as our most important contribution. We conduct experiments in both a caricatured one-dimensional environment proposed by others [14], 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 several state-of-the-art off-policy estimators, and even (b) estimators given access to incumbent state-of-the-art experimental *designs*.

1.1 An Illustrative Example

It will be useful at this point to consider a simple example which highlights (a) the model of interference that we address, (b) the shortcomings of existing approaches to inference under such interference, and (c) our own approach to the problem. *Importantly, all results presented in this simple example will extend to general MDPs by virtue of our analysis in Sections 3 and 4.*

Consider the continuous-time Markov chain depicted in Fig. 1; this is simply an $M/M/N/N$ queue (or the 'Erlang B' model). The state space, ranging from 0 to N, can be thought of as the quantity of some resource (e.g. rental homes of a similar type and geography) currently ‘occupied’. Customers arrive according to a Poisson process with rate λ, and independently with probability p, occupy a resource if available, for an exponentially-distributed duration with mean $1/\mu$. In spite of its simplicity, this model is closely related to one previously studied by [14] in the context of interference in commerce platforms.

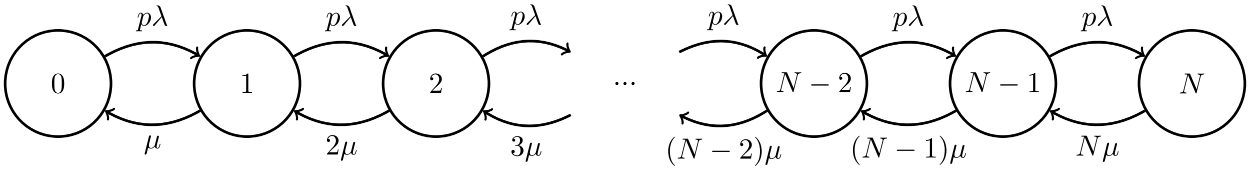

Figure 1: A continuous-time Markov chain. Arrows indicate *rates* of transition between states.

Now consider a treatment whose effect is to increase the probability p by some (unknown) quantity $\delta \geq 0$. Our goal is to measure the effect of the treatment on the steady-state rate of occupation (i.e. the steady-state rate of rightward transitions). We wish to estimate the treatment effect from a simple A/B test; i.e., an experiment that randomly applies (or does not apply) the treatment to each arriving customer. We now describe the various candidate estimators under this experimental design.

Existing Approach 1 – Naive: Given the observed trajectory during this experiment, the ‘naive’ approach measures the empirical rates at which customers with and without treatment occupy resources, and takes the difference – effectively ignoring interference. While this is largely what is done in practice, unfortunately the resulting estimator is *biased*. Specifically, its expected value overestimates the true treatment effect, loosely because it ignores the fact that an increase in p, while increasing the immediate likelihood of occupation, has the secondary effect of *decreasing* the availability of resources, and thus preventing new occupations in the future. This is *interference*.

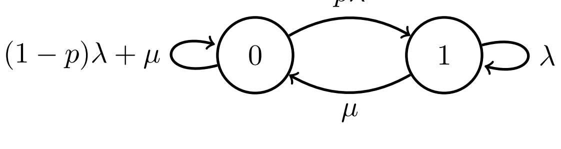

Figure 2: The discrete Markov chain analogous to the continuous-time chain depicted in Fig. 1, for the case $N = 1$. Arrows indicate transition *probabilities*, rather than rates. Without loss of generality, the parameters are normalized so that $\lambda + \mu = 1$.

To be concrete, consider the simplest case: $N = 1$. Fig. 2 depicts the equivalent *discrete* Markov chain, where we have assumed (without loss of generality) that $\lambda + \mu = 1$. A new occupation occurs whenever the chain transitions from state 0 to state 1. We are interested in the rate of such transitions, which can be worked out to be $p\lambda\mu/(p\lambda + \mu)$. The additive increase in this term when p is replaced with $p + \delta$, is the so-called *average treatment effect* (ATE) that we are after. For this example, it suffices to know that the ATE is $\Omega(\delta)$.

The chain we actually observe, e.g. resulting from an A/B test, applies the treatment with 1/2 probability at each time period, affecting the transition probabilities when in state 0. Given a single trajectory $\{s_t\}$ of length T from this chain, the *Naive* estimator then is

$$\hat{\text{ATE}}_{\text{NV}} = \frac{1}{|T_1|} \sum_{t \in T_1} \mathbb{I}_{\{s_t = 0, s_{t+1} = 1\}} - \frac{1}{|T_0|} \sum_{t \in T_0} \mathbb{I}_{\{s_t = 0, s_{t+1} = 1\}},$$

where T_1 and T_0 are, respectively, the sets of time periods in which the treatment was and was not applied. An explicit calculation then shows

$$\left| \lim_T \hat{\text{ATE}}_{\text{NV}} - \text{ATE} \right| \approx \frac{p\lambda}{\mu} \text{ATE}, \tag{1}$$

where the approximation ($\approx$) hides terms of size $O(\delta^2)$. In words, unless the system is extremely unoccupied ($\mu \gg p\lambda$), the Naive estimator *has bias that is on the order of the treatment effect.* It is worth noting that the variance of this Naive estimator is effectively $O(1)$ – i.e. it is as small as we can hope for.

Existing Approach 2 – Off-Policy Evaluation: We may view the problem at hand as one of *Off-Policy Evaluation* in reinforcement learning. To do this, we associate an MDP with our chain. The actions in this MDP correspond to treating or not treating an arrival at a given state in the chain. The reward associated with this action is 1 if the subsequent transition is to the right and 0 otherwise. The ATE then corresponds to the difference in average steady-state rewards between the two policies which always select, respectively, the treatment and non-treatment actions.

The task of estimating the ATE is now trivially viewed as one of OPE. This in turn, immediately suggests a whole host of existing OPE estimators that yield *unbiased* estimates of the ATE, e.g. [12, 19, 20]. This natural approach appears to be promising, e.g. [35], but outside of secondary issues (e.g. discounted vs. average reward), the primary issue is that being unbiased appears to come at a price: variance. Specifically, one may show that *any* unbiased OPE estimator has variance that grows *exponentially* with the number of states in our chain, as $e^{\Omega(N)}$.

This sets up two extremes of a bias-variance tradeoff in our simple example: the Naive estimator has $O(1)$ variance, but its bias is on the order of the treatment effect itself. Any unbiased OPE estimator on the other hand will have variance that scales like $e^{\Omega(N)}$.

Our Approach – The DQ Estimator: Continuing to keep in mind the MDP policy evaluation lens from above, observe that the Naive estimator effectively computes the average difference in instantaneous rewards, averaged over states visited under the policy corresponding to simple randomization. Our estimator makes one change to the Naive estimator: instead of computing the average difference in instantaneous rewards, we instead compute the average difference in *Q-functions* [1]. Intuitively, doing so allows us to partially account for the long-term effects of selecting the treatment over no-treatment at any given state, and consequently, we hope for a less biased estimate of the treatment effect.

It turns out that the DQ estimator (denoted by $\hat{\text{ATE}}_{\text{DQ}}$) provides a dramatic reduction in bias. Starting with bias, the DQ estimator's bias can be worked out explicitly here,

$$\left|\lim_{T} \hat{\text{ATE}}_{\text{DQ}} - \text{ATE}\right| \approx \frac{\delta}{2}\frac{\lambda}{(\mu+\lambda p)}\text{ATE}$$

from which we find that it is $O(\delta^2)$. See Appendix B for details. This is second order relative to the ATE, and, of course, a marked improvement over the Naive estimator's bias. It turns out that this reduction in bias is generic: one of our primary contributions (**Theorem 1**) is to prove the DQ estimator's bias is $O(\delta^2)$ in general MDPs.

Turning next to variance, we can show that the variance of the DQ estimator in our example is $O(N)$. In contrast, an optimal unbiased estimator has variance $e^{\Omega(N)}$, so that the DQ estimator provides an exponential reduction in variance for a relatively small increase in bias. In fact, these relative merits are also generic. Specifically, in **Theorem 2** we upper bound the variance of the DQ estimator for general MDPs. This upper bound scales as $\log(1/\rho_{\min})$, where $\rho_{\min}$ is probability of the least-visited state under the stationary distribution, which can be as large as $1/N$; in the given

[1] Q-functions are formally introduced in Section 3.

example $\rho_{\min} = e^{-\Omega(N)}$. In **Theorem 3** we prove a lower bound on the variance of *any* unbiased estimator in the context of general MDPs, which is exponentially larger – scaling as $\Omega(1/\rho_{\min})$. These two bounds show that the variance reduction relative to OPE is generic (**Theorem 4**).

In summary, this example illustrates precisely the bias-variance trade-offs embodied by each estimator in Table 1. In particular, the DQ estimator has bias second-order in the estimand, with variance exponentially smaller than any unbiased OPE estimator — capturing a particularly advantageous spot in the bias-variance curve. These results hold in generality for a large class of problems, which we formalize in the next section.

Estimator	**Bias**	**Variance**
Naive	$\Omega(\delta)$	$O(1)$
Off-Policy Evaluation	0	$e^{\Omega(N)}$
Differences-In-Q's (DQ)	$O(\delta^2)$	$O(N)$

Table 1: The bias-variance tradeoff of different estimators. Bias is parameterized by the additive impact δ of the intervention on transition probabilities – note that the ATE itself can be $\Omega(\delta)$. 'Variance' shows the limiting variance of each estimator on this example, as a function of the cardinality N of the state space. In full generality, variance is $O(\log(1/\rho_{\min}))$ for DQ, and $\Omega(1/\rho_{\min})$ for OPE, where $\rho_{\min}$ is the frequency of the least-visited state under the steady-state distribution. In this example $\rho_{\min} = e^{-\Omega(N)}$, but in general $\rho_{\min}$ can be up to $1/N$.

Aside – Alternative Experimental Designs: Whereas our focus is on estimation assuming simple-randomization, a more sophisticated *two-sided randomization* (TSR) design has also been studied for this specific system in [14]. In their scheme, both customers *and* resources are randomized independently into treatment and control, and the intervention is applied only if both the customer and the resource are treated. We provide empirical comparisons against this approach in Section 5, which show that DQ outperforms TSR in typical supply / demand regimes, despite a simpler design.

1.2 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 [13, 15], to any form of more-sophisticated 'clustering' of units [43, 10], to clustering specifically when interference is represented by a network [29, 48], to the proportion of units treated [39, 2], to the timing of treatment [4], and beyond [32]. As alluded to earlier, these sophisticated designs can be powerful, but cost, user experience, and other implementation concerns restrict their application in practice [21, 22].

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* [14, **?**], 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)* [30, 37]. The fundamental challenge in OPE is high variance, which can be attributed to the nature of the algorithmic tools used, e.g. sampling procedures [41, 40, 25]. Recent work on 'doubly-robust' estimators [12, 19, 20] 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 [35], 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 [33] and conservative policy iteration [17] use a related on-policy estimation approach to bound policy improvement. Relative to the existing literature, we develop an on-policy surrogate with provably lower bias than extant proposals; see Section 6.3. Furthermore, 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

Having discussed each estimator in a specific example, we now formalize the general 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 see in Eq. (1).

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 [21, 22] prohibit us from running π_0 or π_1 separately until convergence). We observe the sequence $\{(s_t, a_t, r(s_t, a_t)) : t = 1, \ldots, T\}$, wherein $a_t \triangleq \pi_p(s_t)$. Our goal is to estimate the average treatment effect (ATE): $\text{ATE} \triangleq \lambda^{\pi_1} - \lambda^{\pi_0}$.

The Naive Estimator and Bias. A natural approach to estimating the ATE is to use simple randomization (i.e. $P_{1/2}$) and the *Naive* estimator, which we define in the language of MDPs as: $\hat{\text{ATE}}_{\text{NV}} = \frac{1}{|T_1|}\sum_{t\in T_1} r(s_t, a_t) - \frac{1}{|T_0|}\sum_{t\in T_0} r(s_t, a_t)$, where $T_1 = \{t : a_t = 1\}$ and $T_0 = \{t : a_t = 0\}$. In the context of the example of Section 1.1, this corresponds to simply taking the difference between the probability of renting a resource among 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. As we saw, this is sufficient to introduce bias that is on the order of the treatment effect being estimated.

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 objectis in average-reward MDPs. Denote the average cost of a policy π by λ^π. The V-function of a policy π, V_π, characterizes the "reward-to-go" $V_\pi(s) := \mathsf{E}\Big[\sum_{t=0}^{\infty} r(s_t, a_t) - \lambda^\pi \mid s_0 = s\Big]$. It is also known that (V_π, λ^π) is the fixed point of the Bellman operator T_π with $T_\pi(V_\pi, \lambda^\pi) = V_\pi$. Here $T_\pi : \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R} \to \mathbb{R}^{|\mathcal{S}|}$ is given by $T_\pi(V, \lambda) = r_\pi - \lambda \mathbf{1} + P_\pi V$ where $r_\pi : \mathcal{S} \to \mathbb{R}$ is defined according to $r_\pi(s) = \mathsf{E}\left[r(s, \pi(s))\right]$. Finally, the Q-function associated with π, denoted $Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is defined according to $Q_\pi(s,a) := \mathsf{E}\Big[\sum_{t=0}^{\infty} r(s_t, a_t) - \lambda^\pi \mid s_0 = s, a_0 = a\Big]$. Put simply, the Q-function measures the 'excess' reward obtained starting from s with the action a relative to the average reward under π.

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 $\pi_{1/2}$: $\mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_{\mathrm{NV}}\right] = \sum_s \rho_{1/2}(s) \left[r(s,1) - r(s,0)\right]$. It is not hard to see that in the example of Section 1.1, we continue to have $|\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{NV}}] - \mathrm{ATE}| \approx \frac{p\lambda}{\mu}\mathrm{ATE}$, i.e. this idealization of the Naive estimator continues to have bias on the order of the treatment effect. Consider then, the following alternative:

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

where the term $\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}]$ can for now just be thought of as an idealized constant ($\hat{\mathrm{ATE}}_{\mathrm{DQ}}$ is defined soon in (2)). Compared to $\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{NV}}]$, we see that $\mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}]$ 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 $\pi_{1/2}$, as it does in the example of Section 1.1.

Is the dramatic mitigation of bias we see in the example 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 $s \in \mathcal{S}$, $d_{\mathrm{TV}}(P_{1/2}^k(s, \cdot), \rho_{1/2}) \leq$*

$C\lambda^k$ *where* $d_{\mathrm{TV}}(\cdot,\cdot)$ *denotes total variation distance.*

We then have that the second order bias we saw in Section 1.1 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,*

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

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

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

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

$$\min_{\hat{V},\hat{\lambda}}\sum_{s\in\mathcal{S}}\left(\sum_{t,s_t=s} r(s_t,a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t)\right)^2. \tag{3}$$

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

Theorem 2 (Variance and Asymptotic Normality of DQ). *The DQ estimator is asymptotically normal so that* $\sqrt{T}\left(\hat{\mathrm{ATE}}_{\mathrm{DQ}} - \mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_{\mathrm{DQ}}\right]\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2_{\mathrm{DQ}})$*, with limiting standard deviation*

$$\sigma_{\mathrm{DQ}} \le C'\left(\frac{1}{1-\lambda}\right)^{5/2}\log\left(\frac{1}{\rho_{\min}}\right) r_{\max}.$$

where $\rho_{\min} := \min_{s\in S}\rho_{1/2}(s)$ *and* C' *is a constant depending (polynomially) on* $\log(C)$*.*

The fact that σ_{DQ} in Theorem 2 only depends on $1/\rho_{\min}$ logarithmically is somewhat surprising. In fact, a coarse analysis will lead to $\sigma_D = \Omega\left(\frac{1}{\rho_{\min}}\right)$, which shows no advantage compared to the unbiased OPE estimators (which we will see momentarily). The key enabler for this striking result is a novel lemma that exploits an *entry-wise bound* for controlling the variance, even at states that are rarely visited (we dub this the "Entry-wise Non-expansive Lemma"; see Lemma 3). The lemma admits a simple form and may have broader implications for analyzing variance in OPE estimators (see Discussions in Section 6). In addition, our asymptotic normality analysis borrows the delta-method framework used in the context of on-policy LSTD [24], but with a novel linearization that dramatically simplifies the analysis. See Section 3.4 for more details.

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 [11]. 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 $O(\delta^2)$.

3.3 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 $\|A\|_{1,\infty} \leq 1$ ($\|A\|_{1,\infty} = \max_i \sum_j |A_{ij}|$)[2]. Then, ρ_0 and ρ_1 can also be viewed as a function of δ. Also recall 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 [27]). *Suppose $P \in \mathbb{R}^{n\times n}$ and $P' \in$*

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

$\mathbb{R}^{n\times n}$ *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.

$$\begin{aligned}\rho_1^\top r_1 &= \rho_{1/2}^\top r_1 + \rho_1^\top (P_1 - P_{1/2})(I - P_{1/2})^\# r_1 \\ &= \rho_{1/2}^\top r_1 + \delta \cdot \rho_1^\top A (I - P_{1/2})^\# r_1 \end{aligned} \tag{4}$$

Note that we can apply Lemma 1 again to the ρ_1 in the RHS of Eq. (4) and then repeat this process,

$$\rho_1^\top r_1 = \sum_{k=0}^{K} \delta^k \cdot \rho_{1/2}^\top \left(A(I-P_{1/2})^\#\right)^k r_1 + \delta^{K+1} \cdot \rho_1^\top \left(A(I-P_{1/2})^\#\right)^{K+1} r_1 \tag{5}$$

for any $K = 0, 1, 2, \ldots$. Essentially Eq. (5) 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.

$$\mathrm{ATE} = \sum_{k=0}^{K} \delta^k \cdot \left(\rho_{1/2}^\top \left(A(I-P_{1/2})^\#\right)^k r_1 - \rho_{1/2}^\top \left((-A)(I-P_{1/2})^\#\right)^k r_0\right) + \delta^{K+1} \cdot a_K \tag{6}$$

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., $\mathsf{E}_{\rho_{1/2}}\left[\hat{\mathrm{ATE}}_{\mathrm{DQ}}\right]$, exactly matches the first-order truncation. To see this, by the definition of $\mathsf{E}_{\rho_{1/2}}\left[\hat{\text{ATE}}_{\text{DQ}}\right]$ and Q-functions,

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

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

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

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

Generalization to Higher-Order Bias Correction. In fact, the K-th order Taylor expansion of ATE allows us to design estimators that can correct higher-order bias, based on computing difference-in-Q functions iteratively. See details in Section 6.1.

3.4 Proof Sketch of Theorem 2

We aim to use the Markov chain CLT ([16]) to show asymptotic normality of our estimator. The Markov chain CLT states that for a Markov chain $X_1, X_2, \ldots$, and a bounded function u with 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. See proof details in **??**.

Delta method. Unfortunately, the estimator $\hat{\text{ATE}}_{\text{DQ}}$ can not be directly written as an empirical average of some function u. To address this issue, we use the the delta method (traced back to [8], see Lemma 5). In particular, we write $\hat{\text{ATE}}_{\text{DQ}} = f(u_T)$ 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 for 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 Entry-wise Non-expansive Lemma. To bound σ_f, we will invoke Lemma 4, which states that $\sigma_f \leq \sqrt{2}\sqrt{\frac{2\ln(C)+1}{1-\lambda}}\tilde{f}_{\max}$ where $\tilde{f}_{\max} := \max_s |\tilde{f}(s)|$. Then the problem reduces to bounding $\tilde{f}_{\max}$, which will be controlled by the following key lemma.

Lemma 3 (Entry-wise non-expansive lemma)**.** *Let $W : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ be a map denoted by $W(\rho) := (I - P_{1/2})^{\#\top}(P_1 - P_0)^\top\rho$. Let $c := 4\frac{\ln(C)+\ln(1/\rho_{\min})+1}{1-\lambda}$. Then, for any $s \in \mathcal{S}$, $\frac{1}{c}\left|W(\rho_{1/2})(s)\right| \leq \rho_{1/2}(s)$.*

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 Cramér-rao 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 $\{(s_t, a_t, r(s_t, a_t)) : t = 0, \ldots, T\}$ 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*

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

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 model in Section 1.1) 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 $\frac{\sigma_{\mathrm{DQ}}}{\sigma_{\mathrm{off}}} = O\left(\frac{n}{c^n}\right)$, for some constant $c > 1$. Furthermore, $|(\mathrm{ATE} - \mathsf{E}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}])/\mathrm{ATE}| \leq \delta$.*

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: the simple example of Section 1.1, originally proposed by [14]; 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 [14]; and 3) a variety of OPE estimators. For the OPE estimators, we note that off-policy average reward estimation has only recently been addressed in [44, 47], 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 [35].

5.1 A Simple Example

We first study all of our estimators in the example of Section 1.1, a simple setting that does not call for any sort of value function approximation. Our goal now is to understand the relative merits of practical implementations of these estimators, in terms of their bias and variance.

To recap, this MDP is a stylized model of a rental marketplace, consisting of 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 and experimental setup exactly replicates that of [14], with $N = 5000, \lambda = 1, \mu = 1$. We run all estimators over 100 separate trajectories of length $t = 10^4 N$ of the above MDP initialized in its stationary distribution. Figure 3 summarizes the results of this experiment. Beginning with the left panel, which reports estimated quantities at $t = 10^4 N$, we immediately see:

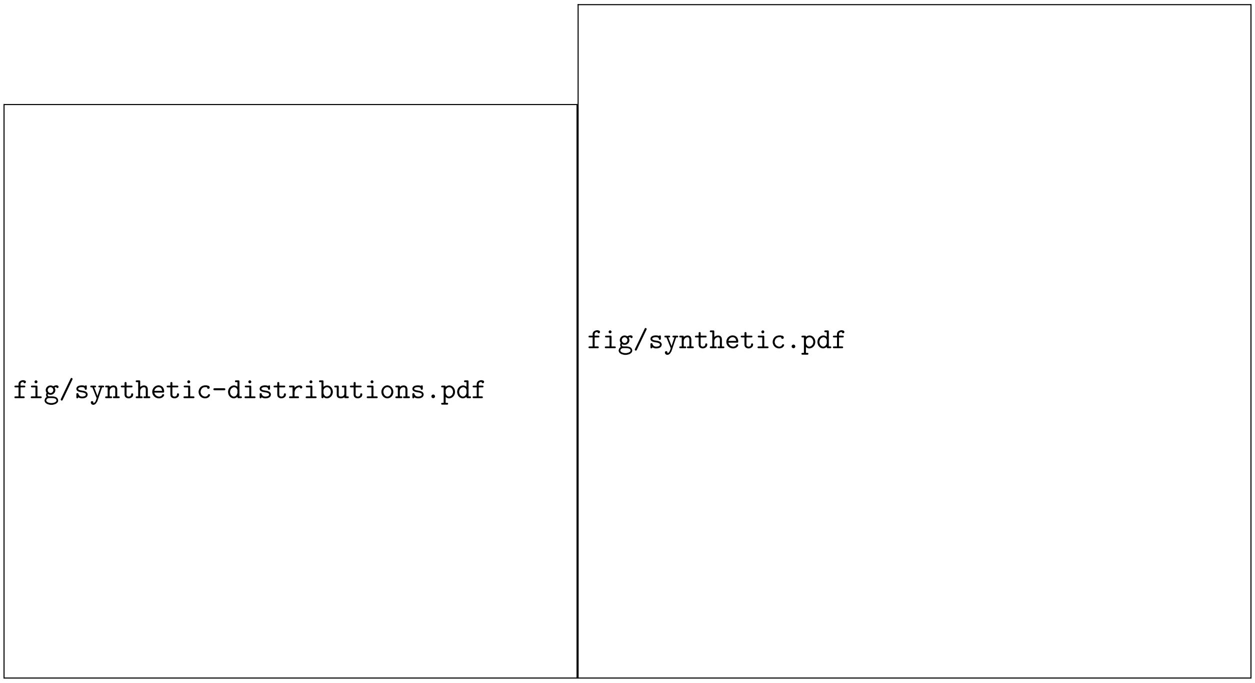

Figure 3: Toy-example from [14]. *Left*: Estimated ATE at time $t/N = 10^4$ across 100 trajectories. Dashed line indicates actual ATE. Diamonds indicate the asymptotic mean for each estimator. DQ shows compelling bias-variance tradeoff for this experimental budget. *Right*: Relative RMSE vs. Time; DQ dominates the alternatives at all timescales.

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 [14].

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 3, 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 (see [14]), 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. [31]. 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 delay disutility. 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 [31, 46], but either lack 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 4 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 fig/rideshare-distributions.pdf

fig/ridesharing.pdf

Figure 4: 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.

the toy model) is *substantially* smaller than the alternatives, and also smaller than the ATE. This is evident in the left panel in Figure 4. 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.

6 Discussion: Bias-Variance Tradeoffs, Policy Optimization

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 provide methods to realize other points on the bias-variance curve. Furthermore, we draw surprising connections between our estimator and trust-region methods in policy optimization, and show that DQ can serve as a drop-in replacement for these policy optimization surrogates with *provably* lower bias.

6.1 A k^{th}-order Bias Correction

As alluded to in Section 3, we can view the DQ estimator as a first-order correction to the Naive estimator, based on a Taylor series expansion of the ATE. This immediately motivates a k^{th}-order correction, with the goal of obtaining estimators with bias $O(\delta^{k+1})$ for arbitrary k.

This correction turns out to have a suprising and intuitive form[3]. In short, to obtain the k^{th} order correction term for k odd (the correction for k even is 0), we simply compute the DQ estimator, but replace rewards in the MDP with the $(k-1)^{\text{th}}$ order Difference-in-Q functions – effectively a Difference-in-Qs-of-Difference-in-Qs.

Precisely, for some reward function $f : \mathcal{S} \mapsto \mathbb{R}$, we can define an auxiliary MDP with the same transition probabilities, but rewards $f(s)$ at each state. Let $Q(s, a; f)$ be the corresponding Q function, under policy $\pi_{1/2}$. We now define the k^{th} order Q-function to be $Q^{(k)}(s, a) = Q(s, a; f^{(k-1)})$, where the rewards are defined as $f^{(k-1)}(s) = \frac{1}{2}\left(Q^{(k-1)}(s, 1) - Q^{(k-1)}(s, 0)\right)$; in other words, the previous-order Difference-in-Q functions. We take as a base case $f^{(0)}(s) = r(s)$. Finally, we can define the K^{th}-order Difference-in-Qs estimate of the ATE to be the sum of all lower-order correction terms: $\hat{\text{ATE}}^{(K)}_{\text{DQ}} = \sum_{k \text{ odd}, k \le K} \mathbb{E}_{\rho_{1/2}}\left[Q^{(k)}(s, 1) - Q^{(k)}(s, 0)\right]$.

In principle this approach enables *off-policy* evaluation with arbitarily low bias – entirely via

[3] For now, we assume for simplicity that rewards are only a function of state. Similar results can be derived for the more general case where r is a function of both state and action, although the resulting formulas are more complex

estimation of *on-policy* quantities. One can verify that $\mathrm{ATE}_{\mathrm{DQ}}^{(1)}$ is the expected value of the DQ estimator. We now generalize Theorem 1 to provide a bias bound for the k^{th} order correction[4]:

Theorem 5. *For any $K = 0, 1, 2, \ldots,$ we have $\left|\mathrm{ATE} - \mathrm{ATE}_{\mathrm{DQ}}^{(K)}\right| \le C' \left(\frac{1}{1-\lambda}\right)^{K+1} \delta^{K+1} r_{\max}$, where C' is a constant depending (polynomially) on $\log(C)$.*

6.2 Interpolating from OPE to Naive via Regularization

Here, we view the DQ estimator again as an intermediate point on the bias-variance curve between Naive and OPE. This time, however, we interpolate between these extremes by regularizing certain key nuisance parameters in estimating the ATE.

An OPE meta-estimator. First, we situate the DQ estimator in the context of existing OPE techniques. Consider the following exact identity for the ATE: $\mathrm{ATE} = \mathsf{E}_{\rho_{1/2}}[\zeta(s)(Q_{\pi_{1/2}}(s,1) - Q_{\pi_{1/2}}(s,0))]$ 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 ([19, 42]) and primal-dual ([7, 38]) estimators – in fact estimate ATE 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 [19]):

$$\hat{\mathrm{ATE}}_{\mathrm{DR}} = \frac{1}{|T_1|}\sum_{t \in T_1} \hat{\zeta}(s_t)\hat{Q}_{\pi_{1/2}}(s_t, 1) - \frac{1}{|T_0|}\sum_{t \in T_0} \hat{\zeta}(s_t)\hat{Q}_{\pi_{1/2}}(s_t, 0) \tag{7}$$

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 2 and 3) – we would expect this to be a particularly powerful approach to OPE, and indeed similar penalties have produced strong empirical performance [28]. Similarly, one can directly penalize the deviation of $\hat{V}_{\pi_{1/2}}$ from zero, as in regularized variants of LSTD (see e.g. [23]). 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 [45].

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

[4] The variance of such plug-in estimators can be bounded by iteratively applying Lemma 3, which is omitted for simplicity.

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; increasing the coarseness of $\hat{V}_{\pi_{1/2}}(s)$ then interpolates between DQ and Naive.

6.3 DQ as a Policy Optimization Objective

Surrogate objectives in trust-region methods DQ also has a suprising relationship to trust-region methods [33, 34, 18]. At each iteration, these methods essentially solve an *offline* policy optimization problem: using data collected under some "behavioral"policy π_{b}, they evaluate (and subsequently optimize) a candidate policy π. The policy evaluation step is exactly an OPE problem, and they construct a surrogate objective based on an identity sometimes referred to as Dynkin's identity [9]: $\lambda(\pi) = \mathsf{E}_{\pi_b}[r(s,a)] + \mathsf{E}_{s\sim\rho_{\pi_b},a\sim\pi}\left[\frac{\rho_\pi(s)}{\rho_{\pi_b}(s)}\left(Q^{\pi_b}(s,a) - V^{\pi_b}(s)\right)\right]$, where $\rho_{\pi_b}, V^{\pi_b}, Q^{\pi_b}$ is the stationary distribution, V-function, and Q-function of the policy π_b respectively and ρ_π is the stationary distribution of π. The referenced trust-region methods then effectively take the likelihood ratio $\rho_\pi(s)/\rho_{\pi_b}(s)$ to be identically one, yielding the (idealized) surrogate objective: $\hat{\lambda}_{\mathrm{TR}}(\pi) = \mathsf{E}_{\pi_b}[r(s,a)] + \mathsf{E}_{s\sim\rho_{\pi_b},a\sim\pi}\left[Q^{\pi_b}(s,a) - V^{\pi_b}(s)\right]$.

A DQ-based surrogate As it turns out, DQ derives from a very similar identity – with a small but critical difference which allows DQ to obtain *even lower bias.* Applying the peturbation bound in Lemma 1 to $\lambda(\pi)$, we obtain a slightly different identity for $\lambda(\pi)$: $\lambda(\pi) = \mathsf{E}_{\pi_b}[r_\pi(s)] + \mathsf{E}_{s\sim\rho_{\pi_b},a\sim\pi}\left[\frac{\rho_\pi(s)}{\rho_{\pi_b}(s)}\left(Q^{\pi_b}(s,a;r_\pi) - V^{\pi_b}(s;r_\pi)\right)\right]$ where $r_\pi(s) = \mathbb{E}_{a\sim\pi}\left[r(s,a)\right]$ is the expected reward under π, and recall that $Q^{\pi_b}(\cdot,\cdot;r_\pi), V^{\pi_b}(\cdot,\cdot;r_\pi)$ are the value functions for an auxiliary MDP with the same transition probabilities, but *taking r_π to be the reward function.* The biased form of this estimator, which forms the basis of the DQ estimator[5], is: $\hat{\lambda}_{\mathrm{DQ}}(\pi) = \mathsf{E}_{\pi_b}[r_\pi(s)] + \mathsf{E}_{s\sim\rho_{\pi_b},a\sim\pi}\left[Q^{\pi_b}(s,a;r_\pi) - V^{\pi_b}(s;r_\pi)\right]$ which is precisely $\hat{\lambda}_{\mathrm{TR}}(\pi)$, but computed on an MDP with rewards r_π.

[5] The DQ estimator in our original setting can actually be derived as either $\hat{\lambda}_{\mathrm{TR}}(\pi_1) - \hat{\lambda}_{\mathrm{TR}}(\pi_0)$ or $\hat{\lambda}_{\mathrm{DQ}}(\pi_1) - \hat{\lambda}_{\mathrm{DQ}}(\pi_0)$, but this results from a very suprising cancellation of terms in the subtraction; i.e. $\hat{\lambda}_{\mathrm{TR}}(\pi)$ and $\hat{\lambda}_{\mathrm{DQ}}(\pi)$ are individually different estimators with very different properties, as we will see; and the higher-order corrections must be derived from $\hat{\lambda}_{\mathrm{DQ}}$.

Lower-order bias This striking resemblance between the surrogatees $\hat{\lambda}_{\mathrm{DQ}}(\pi)$ and $\hat{\lambda}_{\mathrm{TR}}(\pi)$ naturally raises the question of how they compare. As it turns out, the simple act of replacing the rewards with r_π in $\hat{\lambda}_{\mathrm{DQ}}$ has significant consequences in terms of bias:

Theorem 6 (Bias of the DQ surrogate). *Suppose it holds that* $d_{\mathrm{TV}}(p(s,a,\cdot), p(s,a',\cdot)) \leq \delta$ *for all* s, a, a', *and that* $d_{\mathrm{TV}}(\pi(s,\cdot), \pi'(s,\cdot)) \leq \delta'$ *for all* s. *Then, the biases of* $\hat{\lambda}_{\mathrm{TR}}(\pi)$[6] *and* $\hat{\lambda}_{\mathrm{DQ}}(\pi)$ *satisfy*

$$|\hat{\lambda}_{\mathrm{TR}}(\pi) - \lambda^\pi| = O(\delta(\delta')^2) \qquad\qquad |\hat{\lambda}_{\mathrm{DQ}}(\pi) - \lambda^\pi| = O((\delta\delta')^2)$$

This characterization is sharp, in that there exist non-pathological examples where the exact bias of $\hat{\lambda}_{\mathrm{DQ}}$ is a factor δ smaller than that of $\hat{\lambda}_{\mathrm{TR}}$. Crucially, this means that even if the distance between *policies* has no non-vacuous upper bound (i.e. $\delta' = 2$), as long as the resulting *transition functions* are similar, then the bias of $\hat{\lambda}_{\mathrm{DQ}}$ will be small, whereas the bias of $\hat{\lambda}_{\mathrm{TR}}$ can be of the order of δ. This immediately suggests that optimizing $\hat{\lambda}_{\mathrm{DQ}}$ with respect to π should allow for both larger and more accurate policy improvement steps.

7 Conclusion

We propose a novel estimator, the DQ estimator, to solve the interference problem in experiments with simple randomized designs. The DQ estimator achieves second-order bias in estimating the average treatment effect, while its variance can be exponentially smaller than that of any unbiased estimator. We conducted a large scale ride-hailing experiment that demonstrated the superior performance of the DQ estimator over state-of-the-art approaches. The striking and rigorous bias-variance trade-offs induced by the DQ estimator and its generalizations provide a new lens for general off-policy evaluation and policy optimization in reinforcement learning.

References

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

[2] S. Baird, J. A. Bohren, C. McIntosh, and B. Özler. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018.

[3] T. Blake and D. Coey. Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proc. of the fifteenth ACM conf. on Economics and computation*, pages 567–582, 2014.

[4] I. Bojinov, D. Simchi-Levi, and J. Zhao. Design and analysis of switchback experiments. *Available at SSRN 3684168*, 2020.

[5] G.-Y. Chen and L. Saloff-Coste. On the mixing time and spectral gap for birth and death chains. *arXiv preprint arXiv:1304.4346*, 2013.

[6] D. R. Cox. Planning of experiments. 1958.

[7] B. Dai, A. Shaw, N. He, L. Li, and L. Song. Boosting the Actor with Dual Critic. *arXiv:1712.10282 [cs]*,

[6] This result for $\hat{\lambda}_{\mathrm{TR}}$ is in fact a slightly refined version of the key perturbation bound in [33].

Dec. 2017.

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

[9] E. Dynkin. Markov processes.

[10] D. Eckles, B. Karrer, and J. Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *J. of Causal Inference*, 5(1), 2017.

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

[12] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Intl. Conf. on Machine Learning*, pages 652–661. PMLR, 2016.

[13] R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proc. of the 23rd ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[14] R. Johari, H. Li, I. Liskovich, and G. Y. Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 2022.

[15] R. Johari, L. Pekelis, and D. Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research (To Appear)*, 2020.

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

[17] S. Kakade and J. Langford. Approximately Optimal Approximate Reinforcement Learning. In *In Proc. 19th Intl. Conf. on Machine Learning*, pages 267–274, 2002.

[18] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th Intl. Conf. on Machine Learning*. Citeseer, 2002.

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

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

[21] J. Kirn. Challenges in Experimentation. https://eng.lyft.com/challenges-in-experimentation-be9ab98a7ef4, Apr. 2022.

[22] R. Kohavi, D. Tang, and Y. Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

[23] J. Z. Kolter and A. Y. Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proc. of the 26th Annual Intl. Conf. on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada, 2009. ACM Press.

[24] V. R. Konda. Actor-critic algorithms, 2002.

[25] Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Adv. in Neural Information Processing Systems*, 31, 2018.

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

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

[28] O. Nachum, Y. Chow, B. Dai, and L. Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Adv. in Neural Information Processing Systems*, pages 2315–2325, 2019.

[29] J. Pouget-Abadie, K. Aydin, W. Schudy, K. Brodersen, and V. Mirrokni. Variance reduction in bipartite experiments through correlation clustering. *Adv. in Neural Information Processing Systems*, 32, 2019.

[30] D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pages 417–424, 2001.

[31] Z. T. Qin, H. Zhu, and J. Ye. Reinforcement Learning for Ridesharing: A Survey. In *2021 IEEE Intl. Intelligent Transportation Systems Conf. (ITSC)*, pages 2447–2454, Sept. 2021.

[32] M. Saveski, J. Pouget-Abadie, G. Saint-Jacques, W. Duan, S. Ghosh, Y. Xu, and E. M. Airoldi. Detecting network effects: Randomizing over randomized experiments. In *Proc. of the 23rd ACM SIGKDD intl. conf. on knowledge discovery and data mining*, pages 1027–1035, 2017.

[33] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *Intl. conf. on machine learning*, pages 1889–1897. PMLR, 2015.

[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[35] C. Shi, X. Wang, S. Luo, H. Zhu, J. Ye, and R. Song. Dynamic Causal Effects Evaluation in A/B Testing with a Reinforcement Learning Framework. *J. of the American Statistical Association*, 0(ja):1–29, Jan. 2022.

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

[37] R. S. Sutton, C. Szepesvári, and H. R. Maei. A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Adv. in neural information processing systems*, 21(21):1609–1616, 2008.

[38] Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu. Doubly Robust Bias Reduction in Infinite Horizon Off-Policy Estimation, Oct. 2019.

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

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

[41] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 29, 2015.

[42] M. Uehara, J. Huang, and N. Jiang. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. In *Proc. of the 37th Intl. Conf. on Machine Learning*, pages 9659–9668. PMLR, Nov. 2020.

[43] D. Walker and L. Muchnik. Design of randomized experiments in networks. *Proc. of the IEEE*, 102(12):1940–1951, 2014.

[44] Y. Wan, A. Naik, and R. S. Sutton. Learning and Planning in Average-Reward Markov Decision Processes. In *Proc. of the 38th Intl. Conf. on Machine Learning*, pages 10653–10662. PMLR, July 2021.

[45] M. Yang, O. Nachum, B. Dai, L. Li, and D. Schuurmans. Off-Policy Evaluation via the Regularized Lagrangian. In *Adv. in Neural Information Processing Systems*, volume 33, pages 6551–6561. Curran Associates, Inc., 2020.

[46] R. Yao and S. Bekhor. A ridesharing simulation platform that considers dynamic supply-demand interactions. Apr. 2021.

[47] S. Zhang, Y. Wan, R. S. Sutton, and S. Whiteson. Average-Reward Off-Policy Policy Evaluation with Function Approximation. In *Proc. of the 38th Intl. Conf. on Machine Learning*, pages 12578–12588. PMLR, July 2021.

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

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, there is $(I-P)^\# = (I - P + \mathbf{1}\rho^\top)^{-1} - \mathbf{1}\rho^\top$.

B Analysis of the Example

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

$$P_0 = \begin{bmatrix} (1-p)\lambda + \mu & p\lambda \\ \mu & \lambda \end{bmatrix}$$

and the stationary distribution is $\rho_0 = [\frac{\mu}{\mu+\lambda p}, \frac{\lambda p}{\mu+\lambda p}]^\top$ accordingly. Similarly, one can verify under policy π_1, the transition matrix is

$$P_1 = \begin{bmatrix} (1-p-\delta)\lambda + \mu & (p+\delta)\lambda \\ \mu & \lambda \end{bmatrix}$$

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

$$\begin{aligned}\text{ATE} &= r_1^\top \rho_1 - r_0^\top \rho_0 \\ &= \frac{\mu\lambda(p+\delta)}{\mu+\lambda(p+\delta)} - \frac{\mu\lambda p}{\mu+\lambda p} \\ &= \frac{\delta\mu^2\lambda}{(\mu+\lambda(p+\delta))(\mu+\lambda p)}.\end{aligned}$$

Consider the transition matrix for $\pi_{1/2}$,

$$P = \begin{bmatrix} (1-p-\delta/2)\lambda + \mu & (p+\delta/2)\lambda \\ \mu & \lambda \end{bmatrix}.$$

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

$$\rho_{1/2} = \left[\frac{\mu}{\mu+\lambda(p+\delta/2)}, \frac{\lambda(p+\delta/2)}{\mu+\lambda(p+\delta/2)}\right]^\top.$$

The naive estimator is

$$\mathsf{E}[\hat{\mathrm{ATE}}_{\mathrm{NV}}] = \frac{\delta\lambda\mu}{\mu + \lambda(p + \delta/2)}.$$

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

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

where Q_a is the Q-value vector for the policy $\pi_{1/2}$ under the action a. 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

$$\begin{aligned} Q(0,0) &= \frac{\mu\lambda p}{\mu + \lambda p}, \quad Q(0,1) = 0 \\ Q(1,0) &= \frac{\mu\lambda(p+\delta)}{\mu + \lambda p}, \quad Q(1,1) = 0 \end{aligned}$$

Therefore,

$$\begin{aligned} \mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}] &= \frac{\mu}{\mu + \lambda(p+\delta/2)}(Q(0,1) - Q(0,0)) \\ &= \frac{\mu}{\mu + \lambda(p+\delta/2)}\frac{\mu\lambda\delta}{\mu + \lambda p}. \end{aligned}$$

For the bias induced by the DQ estimator, we have

$$\begin{aligned} \mathrm{ATE} - \mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}] &= \frac{\delta\mu^2\lambda}{(\mu + \lambda(p+\delta))(\mu + \lambda p)} - \frac{\mu}{\mu + \lambda(p+\delta/2)}\frac{\mu\lambda\delta}{\mu + \lambda p} \\ &= \delta\frac{\mu^2\lambda}{\mu + \lambda p}\left(\frac{1}{\mu + \lambda(p+\delta)} - \frac{1}{\mu + \lambda(p+\delta/2)}\right) \\ &\approx \frac{\delta}{2}\frac{\lambda}{(\mu + \lambda p)}\mathrm{ATE} \end{aligned}$$

where $\approx$ ignore the terms $O(\delta^3)$. This completes the analysis.

C Proof of Theorem 2

C.1 Entry-wise Non-Expansive Lemma

To begin, we present a lemma that is the key enabler of establishing the striking variance improvement of the DQ estimator over the un-biased estimator. The lemma simply states

Lemma 3 (Entry-wise non-expansive lemma). *Let $W : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ be a map denoted by $W(\rho) := (I - P_{1/2})^{\#\top}(P_1 - P_0)^{\top}\rho$. Then, for any $s \in \mathcal{S}$,*

$$\frac{1}{c}\left|W(\rho_{1/2})(s)\right| \le \rho_{1/2}(s)$$

where $c := 4\frac{\ln(C)+\ln(1/\rho_{\min})+1}{1-\lambda}$.

This is to say, the mapping $\frac{1}{c}W$ does not expand $\rho_{1/2}$ in terms of entry-wise values. To see the necessity of this lemma and gain some intuition, consider a special case where $P_0, P_1, P_{1/2}, \rho_0, \rho_1, \rho_{1/2}$ are all known, while only the rewards are unknown and can only be sampled under the distribution $\rho_{1/2}$. For simplicity, assume $r_0 = r_1 = r$ and the sample for $r_t = r(s_t) + \epsilon_t$ is i.i.d from $\rho_{1/2}$ with some exogenous noise $\epsilon_t \sim \mathcal{N}(0, 1)$. Let us denote the empirical average estimator for r be $\hat{r}$. By CLT, we have

$$\sqrt{T}(\hat{r} - r) \xrightarrow{d} \mathcal{N}(0, D^{-1})$$

where D is a diagonal matrix with entries $D_{s,s} = \rho_{1/2}(s)$. This limiting variance captures the intuition that, for the state that is rarely visited, the variance for $\hat{r}(s) - r(s)$ can blow up. In fact, consider an un-biased estimator $(\rho_1^\top - \rho_0^\top)\hat{r}$ for the ATE, $(\rho_1^\top - \rho_0^\top)r$, (this is the un-biased estimator that achieves the optimal variance), we have

$$\sqrt{T}\left((\rho_1^\top - \rho_0^\top)\hat{r} - \text{ATE}\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2)$$

where

$$\begin{aligned}\sigma_0^2 &:= (\rho_1^\top - \rho_0^\top)D^{-1}(\rho_1 - \rho_0)\\ &= \sum_s \frac{(\rho_1(s) - \rho_0(s))^2}{\rho_{1/2}(s)}.\end{aligned}$$

Note that there is no guarantee for the likelihood ratio $\frac{\rho_0(s)}{\rho(s)}$ and $\frac{\rho_1(s)}{\rho(s)}$ and in general $\sigma_0^2 = \Omega\left(\frac{1}{\rho_{\min}}\right)$ and this is the price to pay for the un-biased off-policy evaluation.

On the other hand, one can verify that the DQ estimator is simply

$$\text{ATE}_D = \rho_{1/2}^\top (P_0 - P_1)(I - P_{1/2})^{\#}\hat{r}.$$

This leads to the limiting variance of ATE_{DQ}:

$$\sqrt{T}\left(\text{ATE}_D - \mathsf{E}_{\rho_{1/2}}[\text{ATE}_D]\right) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2)$$

where

$$\begin{aligned}\sigma_1^2 &:= \rho_{1/2}^\top (P_0 - P_1)(I - P_{1/2})^{\#}D^{-1}(\rho_{1/2}^\top (P_0 - P_1)(I - P_{1/2})^{\#})^\top\\ &= \|\rho_{1/2}^\top (P_0 - P_1)(I - P_{1/2})^{\#}D^{-1/2}\|^2.\end{aligned}$$

By the definition of W mapping, we then have

$$\begin{aligned}\sigma_1^2 &= \sum_s \left(W(\rho_{1/2})(s) \frac{1}{\rho_{1/2}(s)^{1/2}} \right)^2 \\ &\overset{(i)}{\leq} \sum_s \frac{1}{c^2} \rho_{1/2}(s) \\ &= \frac{1}{c^2}.\end{aligned}$$

where (i) is due to Lemma 3. Then σ_1 is in the order of $\log(1/\rho_{\min})$. In fact, without Lemma 3, a loose analysis will provide $\sigma_1^2 = \Omega(1/\rho_{\min})$, which is in the same order of σ_0^2, that shows no advantage of using DQ estimator. Essentially Lemma 3 characterizes the explicit superiority of evaluating on-policy quantities over off-policy quantities. We believe this novel lemma is of independent interest for the field of OPE. The proof is postponed to the end of the section.

C.2 Outline of the Proof

In this section, we present the outline of the proof for Theorem 2. We aim to use Markov chain CLT ([16]) 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 $\hat{\text{ATE}}_{\text{DQ}}$ can not be directly written as an empirical average of some function u. To address this issue, we use "delta method" (traced back to [8], see Lemma 5). In particular, we write $\hat{\text{ATE}}_{\text{DQ}} = f(u_T)$ 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, "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 Entry-wise Non-expansive Lemma. To bound σ_f, we will invoke Lemma 4, which states that

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

where $\tilde{f}_{\max} := \max_s |\tilde{f}(s)|$. Then the problem boils down to bound $\tilde{f}_{\max}$, which will be controlled by Lemma 3.

Next, we present the proof in full details.

C.3 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

$$F^{(a)}(X_t) := 2E_{s_t}E_{s_{t+1}}^{\top}\cdot 1(a_t = a) \tag{8}$$

$$h^{(a)}(X_t) := 2r(s_t,a_t)\cdot E_{s_t}\cdot 1(a_t = a) \tag{9}$$

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 $\hat{\text{ATE}}_{\text{DQ}} := f(F_T^{(0)}, F_T^{(1)}, h_T^{(0)}, h_T^{(1)})$ 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. (3). Further, one can verify that with a plugging-in estimator for Q, the DQ estimator is given by

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

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

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

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[7]

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

which is equivalent to

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

since $f(F_0^*, F_1^*, h_0^*, h_1^*) = \mathsf{E}_{\rho_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}]$. 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 $c := \mathsf{E}_{\rho_{1/2}}[2(1(a = 1) - 1(a = 0))(V(s') + r(s, a) - \lambda)]$. By Lemma 6, we have

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

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

Here σ_f^2 is explicitly given by (by Markov Chain CLT)

$$\sigma_f^2 := \sum_{s,a,s'} \tilde{f}(s,a,s')^2 \rho_{1/2}(s) P_a(s,s') \frac{1}{2}$$
$$+ 2 \sum_{s,a,s'} \rho_{1/2}(s) P_a(s,s') \frac{1}{2} \tilde{f}(s,a,s') \sum_{s_1} (I - P_{1/2})^{\#}_{s',s_1} t(s_1)$$

where $t(s) = \sum_{a,s'} \tilde{f}(s,a,s') P_a(s,s') \rho_{1/2}(s) \frac{1}{2}$.

C.4 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 $z_{\max}$, note that

$$\begin{aligned} z_s &= \mathbf{1}^\top D(P_1 - P_0)(I - P_{1/2})^{\#} D^{-1}) E_s \\ &= \rho_{1/2}^\top (P_1 - P_0)(I - P_{1/2})^{\#} D^{-1} E_s \\ &= \rho_{1/2}^\top (P_1 - P_0)(I - P_{1/2})^{\#} E_s \frac{1}{\rho_{1/2}(s)}. \end{aligned}$$

Then, we can invoke Lemma 3 to obtain that

$$z_s = W(\rho_{1/2})(s) \frac{1}{\rho_{1/2}(s)} \leq 4 \frac{\ln(C) + \ln(1/\rho_{\min}) + 1}{1-\lambda}.$$

Combining all together, we have

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

for some constant C' that depends (polynomially) on $\log(C)$, which completes the proof of Theorem 2.

C.5 Proof of Lemma 3

The only thing remaining is the proof of Lemma 3.

Note that we have $W(\rho^{1/2}) = (I - P_{1/2})^{\#\top}(P_1 - P_0)^\top \rho_{1/2}$. Let $v := (P_1 - P_0)^\top \rho_{1/2}$. We will show first (i) $\frac{1}{2}v$ is entry-wise non-expansive; and then (ii) $(I - P_{1/2})^\top v$ is entry-wise bounded.

To begin, 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',

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

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

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

Next, consider

$$\begin{aligned}
v^\top (I - P_{1/2})^\# E_s &= \sum_{k=0}^\infty v^\top (P_{1/2}^k - \mathbf{1}\rho_{1/2}^\top) E_s \\
&=: \sum_{k=0}^\infty a_k.
\end{aligned}$$

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

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

Using the same trick in proving Lemma 2, we have

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

Then $\left|W(\rho_{1/2})(s)\right| = |\sum_k a_k| \le c\rho_{1/2}(s)$ with $c := 4\frac{\ln(C)+\ln(1/\rho_{\min})+1}{1-\lambda}$. This completes the proof.

D Proof of Theorem 3

The proof is based on multi-variate Cramér-Rao bound. To begin, we assume $P_0(s,s') > 0, P_1(s,s') > 0$ for all (s,s').[8]

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

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

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

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

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 [36], 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.

D.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. (3),

$$(\hat{V}_1, \hat{\lambda}^{\pi_1}) = \arg\min_{\hat{V},\hat{\lambda}} \sum_{s\in\mathcal{S}} \left(\sum_{t,s_t=s,a_t=1} r(s_t,a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t) \right)^2 \tag{10}$$

$$(\hat{V}_0, \hat{\lambda}^{\pi_0}) = \arg\min_{\hat{V},\hat{\lambda}} \sum_{s\in\mathcal{S}} \left(\sum_{t,s_t=s,a_t=0} r(s_t,a_t) - \hat{\lambda} + \hat{V}(s_{t+1}) - \hat{V}(s_t) \right)^2. \tag{11}$$

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 **??**).

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

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

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

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

which completes the proof.

E 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 $P(s,s) = 1/2 + \delta$ (exception at 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 [5], we have the spectral gap of the chain is in the order of $\gamma = O(1/n)$. Furthermore, the mixing time of the chain is bounded by

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

Following the same proof in Theorem 2, we have 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 sufficient small constant c'. Let P_1 be the same as P except $\forall s \geq k$

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

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

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

for some constant c. Therefore,

$$\frac{\sigma_{\text{on}}}{\sigma_{\text{off}}} = O\left(\frac{n}{e^{c'n}}\right)$$

for some constant c'. Next, consider the bias of 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 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

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

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

F 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,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left(f(X_t) - f^*\right) \xrightarrow{d} N(0, \sigma^2) \tag{12}$$

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

$$\sigma \le \sqrt{2}(b-a)\sqrt{\frac{2\ln(C)+1}{1-\lambda}}. \tag{13}$$

Proof. Note that Eq. (10) is simply due to the Markov chain CLT ([16]). Let D be an diagonal matrix with entries $D_{ii} = \rho_i$. [16] further states that

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

where (i) is due to Lemma 2. Therefore,

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

■

Lemma 5 (Theorem 6.2 [24]). *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 Markov Chain CLT (Corollary 5 [16]), 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,

$$\Sigma := E_\rho[(u(X_1)-u^*)(u(X_1)-u^*)^\top] + 2\sum_{k=2}^{\infty} E_\rho[(u(X_1)-u^*)(u(X_k)-u^*)^\top] \tag{14}$$

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

Then, since f is twice differentiable in a neighbor 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 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. (12). This implies that F (the linearization of f at the point u^*) will converge with the same limiting variance as f. ■

G Experiment details

G.1 Synthetic example

G.1.1 Environment

We replicate exactly the environment of [14]. 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 $\frac{s_t v(a_t)}{N+s_t v(a_t)}$ 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 $\frac{N}{N+s_t v(a_t)}$.
3. No state change occurs; i.e. $s_{t+1} = s_t$ and $r_t = 0$.

[14] 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 $s_t = (s_t^{\mathrm{co}}, s_t^{\mathrm{tr}})$, 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 $v_{\mathrm{co}}(0) = v_{\mathrm{co}}(1) = v(0)$, while the average utility of a treatment listing is $v_{\mathrm{tr}}(0) = v(0)$ and $v_{\mathrm{tr}}(1) = v(1)$. We defer to [14] for further details of this scheme.

G.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 [14], 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 [14]). We set p, p_L, β assuming λ, μ are known, exactly as prescribed in [14]. Specifically, we compute the values reported in Table **??** 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 (3). 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:

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

   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 literature. In Section C.1 we describe this algorithm, provide convergence guarantees, and show that this algorithm is efficient. This can be construed as a direct analog of [35]'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 [44]. 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 [47] 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)	$\alpha \in \{0.01, \mathbf{0.1}, 1, 10, 100\}$
Off-Policy (TD)	$\beta \in \{0.2, \mathbf{0.5}\}, \gamma \in \{0.001, \mathbf{0.01}, 0.1, 1.\}$

Table 2: Hyperparameters for the synthetic example of [14]. Parameter settings reported in the main text are in bold.

G.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. [14] 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 [14]), 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). .

G.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.

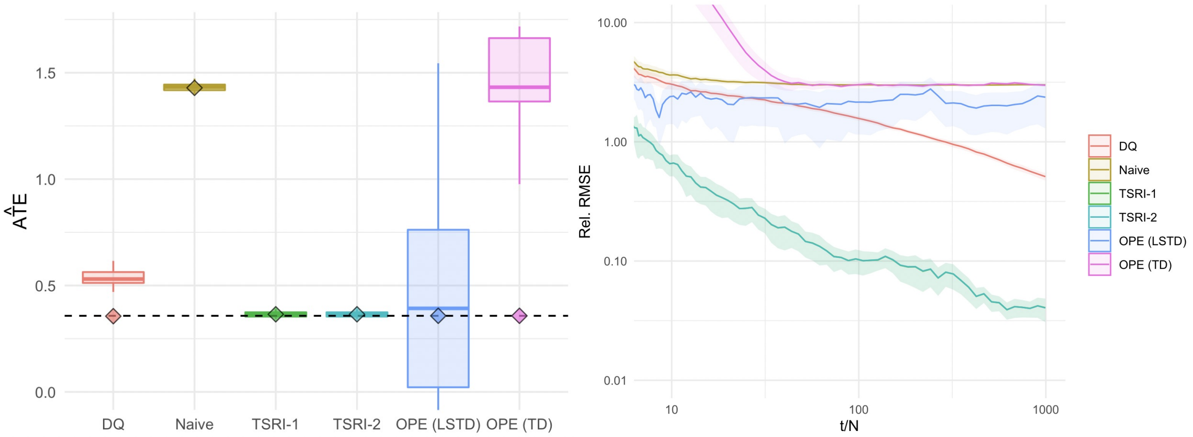

Figure 5: Simple example from [14], with $\lambda = 10$. *Left*: Estimated ATE at time $t/N = 10^3$ across 100 trajectories. Dashed line indicates actual ATE. Diamonds indicate the asymptotic mean for each estimator. Over this horizon, TSRI-1 and TSRI-2 exhibit small bias and variance, although asymptotically DQ still has lower bias.

G.2 Ridesharing Simulator

G.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] (specifically, from January 11, 2015), with a random willingness-to-pay per second distributed as $\text{LogNormal}(\log(0.01), 1.)$. 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 one hour, 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.

G.2.2 Algorithms

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.

The algorithms 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, 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 α. Here we solve for θ by approximating and solving for a fixed point to the regularized least-squares problem:

$$\theta = \arg\min_{\theta'} \|\Phi\theta' - r - P\Phi\theta + \lambda\|_2^2 + \alpha\|\theta'\|_2^2$$

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 [44] to linear function approximation, as provided in [47]. 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 [47]. 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)	$\alpha \in \{0.01, 0.1, \mathbf{1}, 10, 100\}$
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 3: Hyperparameters for the ridesharing setting. Parameter settings reported in the main text are in bold.

G.2.3 Computing environment

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