Correcting for Interference in Experiments: A Case Study at Douyin

Vivek F. Farias
Massachusetts Institute of Technology
USA

Hao Li
ByteDance
China

Tianyi Peng*
tianyipeng95@gmail.com
Massachusetts Institute of Technology
USA

Xinyuyang Ren
ByteDance
China

Huawei Zhang
ByteDance
China

Andrew Zheng*
azheng92@gmail.com
Massachusetts Institute of Technology
USA

ABSTRACT

Interference is a ubiquitous problem in experiments conducted on two-sided content marketplaces, such as Douyin (China's analog of TikTok). In many cases, creators are the natural unit of experimentation, but creators interfere with each other through competition for viewers' limited time and attention. "Naive" estimators currently used in practice simply ignore the interference, but in doing so incur bias on the order of the treatment effect. We formalize the problem of inference in such experiments as one of policy evaluation. Off-policy estimators, while unbiased, are impractically high variance. We introduce a novel Monte-Carlo estimator, based on "Differences-in-Qs" (DQ) techniques, which achieves bias that is second-order in the treatment effect, while remaining sample-efficient to estimate. On the theoretical side, our contribution is to develop a generalized theory of Taylor expansions for policy evaluation, which extends DQ theory to all major MDP formulations. On the practical side, we implement our estimator on Douyin's experimentation platform, and in the process develop DQ into a truly "plug-and-play" estimator for interference in real-world settings: one which provides robust, low-bias, low-variance treatment effect estimates; admits computationally cheap, asymptotically exact uncertainty quantification; and reduces MSE by 99% compared to the best existing alternatives in our applications.

CCS CONCEPTS

• **Applied computing** → **Marketing**; • **Computing methodologies** → *Artificial intelligence*; **Reinforcement learning**; **Dynamic programming for Markov decision processes**; • **Mathematics of computing** → *Probability and statistics.*

*Corresponding authors: tianyipeng95@gmail.com, azheng92@gmail.com

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09…$15.00
https://doi.org/10.1145/3604915.3608808

KEYWORDS

Interference, A/B testing, Experimentation, Off-policy Evaluation, Reinforcement Learning

ACM Reference Format:
Vivek F. Farias, Hao Li, Tianyi Peng, Xinyuyang Ren, Huawei Zhang, and Andrew Zheng. 2023. Correcting for Interference in Experiments: A Case Study at Douyin. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3604915.3608808

1 INTRODUCTION

In recent years, large-scale online platforms have increasingly relied on experimentation to measure the impact of interventions and enhance user experiences [8, 18, 27]. Traditional randomized controlled trials (RCTs), commonly known as A/B tests, offer a robust framework for causal inference in settings where treatment and control groups are independent [6, 28]. However, in many online platforms, treatment and control outcomes are not independent but interact through a shared system state, a phenomenon known as "Markovian" interference [7]. Addressing this interference is a key challenge for state-of-the-art experimentation platforms [2, 12, 20, 21].

In this work, we propose a novel method to address problems of interference that arise in online content marketplaces such as Douyin, a leading social video platform with 600 million daily active users as of early 2021.[1] Like its US analog TikTok, Douyin's core product is short-form video content: creators create videos on the app, and viewers are presented with a sequence of these videos (determined by the company's proprietary recommendation algorithm). In a typical viewer session, the viewer opens the app; the platform presents them with a video; the viewer watches this video until they swipe to indicate they would like to advance to the next video; and the platform then presents them with another video. The process repeats until the viewer eventually leaves the platform.

As with many other two-sided marketplaces, a key challenge in experimentation at Douyin is that certain interventions can *only* be tested via creator-side experiments, while many key metrics are measured on the viewer side. To give a specific, real example: Douyin has a new feature which will allow creators to deliver livestreams in high definition, and Douyin wants to test the impact

[1] https://new.qq.com/rain/a/20210329A06EP300

of this new feature on viewer engagement. A typical metric will be average "dwell time" for a viewer – that is, the total time an average viewer spends on the platform in some interval.

Because the feature is introduced at the creator level, the platform cannot control which viewers will be exposed, as would be necessary for viewer-side or two-side randomized experiments. Interference between creators occurs primarily because viewers have a limited budget of resources to spend on Douyin – a budget consisting of free time, attention, battery life, and other factors – and creators in treatment compete for this budget with those in control. See Fig. 1 for an illustrative example. In such settings, the only existing option deployed at Douyin is naive estimation; in other words, to ignore the interference and proceed with estimation as in a standard RCT. However, this naive estimator is known to be considerably biased, with the bias potentially being as large as the treatment effect itself [2, 11].

Contributions. To address the interference problem, we first formulate treatment effect estimation as a problem of off-policy evaluation in a Markovian Decision Process (MDP), where states can be arbitrarily complex – consisting for example of a viewer's preferences, engagement level, and viewing history. Existing off-policy evaluation methods, despite being unbiased, fail in this setting due to the prohibitively large state space, and are excessively high variance. In contrast, the naive estimator does not require access to the state, but it suffers from a significant bias. To overcome these limitations, we propose a novel estimator that does not require state access and has a second-order bias compared to the naive estimator. Our proposed estimator can be viewed as an instance of Differences-in-Qs (DQ) estimators, proposed in a very recent theoretical work [7] which showed that DQ estimators enjoy a favorable bias-variance tradeoff in average-reward MDPs.

We integrate our proposed DQ estimator into the Douyin platform, incorporating various generalizations, and demonstrate its superior performance compared to existing methods. To summarize, our contributions are threefold:

1) A Novel and Practical Estimator. We propose a novel DQ estimator to correct for interference in A/B tests at Douyin. This estimator does not require access to the state, making it more broadly applicable compared to other off-policy evaluation methods. It has provably second-order bias (Theorem 1) and offers practical benefits, as it accommodates heterogeneous users, partially observable states, and multiple simultaneous experiments (Section 6). Additionally, we develop doubly robust techniques to reduce variance (Section 6.3) and enable a precise variance quantification (Section 6.4).

2) Superior Empirical Performance. We implement our estimator on Douyin's experimentation platform, and report results from a large-scale simulator modeling this implementation in Section 7. These experiments show that the DQ estimator significantly outperforms state-of-the-art approaches, reducing mean squared error (MSE) by 99% compared to the best existing alternatives. Our work represents the first large-scale implementation of the DQ estimator in the real world.

3) A Unified Theory for DQ. Finally, we provide theoretical extensions to the DQ estimator and its underlying Taylor series theory (Section 5), allowing it to accommodate the discounted and total reward formulations more naturally suited to Douyin's applications. Our unifying theory yields existing average reward results in [7] as a special case, and are of independent interest.

2 RELATED WORK

Interference in experiments, across fields such as public medicine, agriculture, and online markets [23, 29, 31], occurs when the outcome of an experimental unit is affected by others, potentially leading to biases in treatment effect estimation. Interference has emerged as a key challenge in online platforms in particular, across companies including eBay [2], Uber/Lyft [5, 35], Airbnb [11], and Douyin, the setting which motivates this study.

Experimental Design. Existing approaches to interference have primarily focused on sophisticated experimental designs. Examples include minimizing interference through clustering units [24, 36–38], alternating randomization across time periods [3, 4, 10], conducting randomization on both supply and demand sides in two-sided markets [1, 15, 21], and other designs [9, 14, 32]. Despite their potential, practical limitations such as cost and implementation concerns often restrict the use of such sophisticated designs [19, 20], as we will also see in motivating applications at Douyin. In contrast, as opposed to introducing more complex designs, this paper introduces a novel estimator under *simple unit-level randomization*, effectively mitigating bias induced by interference.

Off-Policy Evaluation (OPE). Our problem can be viewed as a case of *off-policy evaluation (OPE)* [25, 30], an increasingly important problem in reinforcement learning. Unbiased OPE estimators typically suffer from high variance [22, 33, 34], and the related curse of dimensionality in large state spaces [13, 16, 17]. Our approach can be construed as a *biased* method for OPE, which incurs a small amount of bias for a massive reduction in variance (similarly to [7, 26]) and offers strong theoretical guarantees without the need to access state information.

3 MODEL

We begin by developing our theory in a simplified setting, where we observe a single trajectory (or "session") from *a single viewer*. In Section 6, we will extend this core theory to a much richer problem formulation, capturing the complexities of applications at Douyin.

Consider a scenario in which the platform wants to test a creator-side intervention – for example, rolling out the HD streaming feature discussed in the introduction. Due to infrastructure constraints or concerns about the impact of the intervention, we must experiment at a creator level, and therefore implement a creator-side A/B test.

We model each session for the viewer as an MDP. At each step t, a video is presented to the viewer. With probability p, the presented video is in the treatment group, corresponding to an action $a_t = 1$; otherwise the video is in the control group, corresponding to an action $a_t = 0$. When the viewer swipes to the next video, the platform realizes a reward r_t – e.g., the time the viewer spent watching the previous video – and the MDP advances to the next step. The reward r_t is a (random) function of $a_t \in \mathcal{A}$, the action at time t, and $s_t \in \mathcal{S}$, the viewer's state at time t.

Critically, in real applications, the state may be arbitrarily complex, and partially or totally unobservable. For example, s_t may consist of the time the viewer has spent on the platform (and implicitly, the remaining time budget of the viewer), their engagement

level, and even the whole history of videos the viewer has watched. See Fig. 1 for an example. As we will see, we can design estimators for which this complexity poses no challenge.

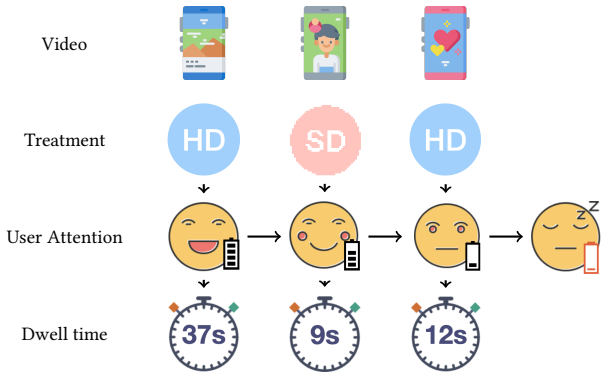

Figure 1: A graphical depiction of the Douyin MDP for a given viewer. At each timestep, the viewer is shown a video; the video is either treatment (high definition, $a_t = 1$) or control (standard definition, $a_t = 0$); and depending on the video and the viewer's current state (consisting of e.g., attention, battery life, data budget, etc.), the platform then collects some reward (i.e., the viewer's dwell time), and moves to the next video. Eventually the viewer runs out of attention and leaves the platform.

A policy $\pi : \mathcal{S} \to \mathcal{A}$ is a (random) mapping from states to actions. We are interested in three policies:

(1) **Global treatment** π_1, which chooses $\pi_1(s) = 1, \forall s \in \mathcal{S}$, i.e., every video presented is treated. The associated transition matrix is denoted as $P_1 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$.
(2) **Global control** π_0, which chooses $\pi_0(s) = 0, \forall s \in \mathcal{S}$, i.e., every video is untreated. The associated transition matrix is denoted as $P_0 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$.
(3) **The experiment** π_p, which chooses $\pi(s) = 1$ with probability p; otherwise $\pi(s) = 0$; in other words, video-side A/B testing. The associated transition matrix is denoted as $P_p := (1-p)\cdot P_0 + p \cdot P_1$. For ease of exposition, we set $p = 1/2$ and discuss generalizations in Section 6.

For a given policy π, we define the total reward of the policy as: $J_\pi = \sum_{t=0}^{\infty} \mathsf{E}_\pi[r_t]$, where the notation E_π denotes an expectation over states $s_0 \sim \rho_{\text{init}}$, actions $a_t|s_t \sim \pi(s_t)$, dynamics $s_{t+1}|s_t \sim P_\pi(s_t, s_{t+1})$, and rewards.

Goal. Given observations $\{(r_t, a_t)\}$ under the experiment policy $\pi_{1/2}$, our goal is to estimate the difference between the total rewards obtained by π_1 and π_0:

$$\text{ATE} := J_{\pi_1} - J_{\pi_0}. \tag{1}$$

In our example above, ATE corresponds to the impact to the viewer's dwell time in the platform by replacing the policy π_0 by π_1, a key quantity of interest when deciding whether rolling out a new policy.

To ensure that J_π is well-defined, we make one key assumption on the Markov chain induced by each policy, which is that the viewer eventually terminates their session; i.e., all Markov chains are *absorbing*. To be precise: we say that $\mathcal{S}_{\text{abs}} \subseteq \mathcal{S}$ is an absorbing class under policy π if $P_\pi(s, s') = \mathbb{I}\{s' \in \mathcal{S}_{\text{abs}}\}$ for all $s \in \mathcal{S}_{\text{abs}}$, and if $r_t = 0$ a.s. when $s_t \in \mathcal{S}_{\text{abs}}$. We also require that $\mathcal{S}_{\text{abs}}$ is reached in finite time almost surely starting from any state and define the expected time to this absorption event as $T^\pi_{\text{abs}} := \max_{s\in\mathcal{S}} \sum_{t=0}^{\infty} \mathsf{E}_\pi[\mathbb{I}\{s_t \notin \mathcal{S}_{\text{abs}}\} | s_0 = s]$. We then make the following assumption:

Assumption 1. *There exists a class $\mathcal{S}_{\text{abs}} \subseteq \mathcal{S}$ and a time T_{abs}, such that $\mathcal{S}_{\text{abs}}$ is absorbing and $T_{\text{abs}} > T^\pi_{\text{abs}}$ for each policy π in $\pi_0, \pi_1, \pi_{1/2}$.*

4 A NOVEL ESTIMATOR FOR ESTIMATION UNDER INTERFERENCE

4.1 A Naive Estimator

We begin by describing the status quo for this setting: Naive estimation, wherein interference is simply ignored and the experiment is treated as a traditional A/B test. A typical estimator is the inverse propensity weighted estimator

$$\widehat{\text{ATE}}_{\text{Naive}} := \sum_{t=0}^{\infty} \left(\frac{1(a_t = 1)}{\Pr(a_t = 1)} r_t - \frac{1(a_t = 0)}{\Pr(a_t = 0)} r_t \right) \tag{2}$$

$$= \sum_{t=0}^{\infty} 2 \cdot (1(a_t = 1) r_t - 1(a_t = 0) r_t) \tag{3}$$

which simply estimates the difference between the single-step rewards obtained under treatment and control – ignoring the impact of each action on the overall trajectory.

However, the bias of this estimator can be significant. To see this, note that in expectation

$$\mathsf{E}_{\pi_{1/2}}\left[\widehat{\text{ATE}}_{\text{Naive}}\right] = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_{1/2}}[r(s_t, 1)] - \mathsf{E}_{\pi_{1/2}}[r(s_t, 0)] \tag{4}$$

On the other hand, $\text{ATE} = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_1}[r(s_t, 1)] - \sum_{t=0}^{\infty} \mathsf{E}_{\pi_0}[r(s_t, 0)]$. The difference between ATE and $\mathsf{E}_{\pi_{1/2}}\left[\widehat{\text{ATE}}_{\text{Naive}}\right]$ then results from the fact that the expectation we compute by sampling from $\pi_{1/2}$ does not match the expectations we would like to compute, under π_0, π_1.

Intuitively, then, the magnitude of the bias depends on the difference between the dynamics under π_0 and under π_1. We characterize this difference as follows: starting from any state, let δ bound the distance between distributions over next states. Precisely, $\delta := \max_s D_{\text{TV}}(P_0(s, \cdot), P_1(s, \cdot))$ where D_{TV} measures the total variation distance[2]. Then, the bias of $\widehat{\text{ATE}}_{\text{Naive}}$ is $\Theta(\delta)$, whereas the ATE itself is also $\Theta(\delta)$ – i.e., the bias of Naive estimation is on the same order as the treatment effect. We will demonstrate this in a number of ways: Section 4.3 gives a simple example where the Naive estimator measures an effect which is in fact only cannibalization; Section 7 contains realistic experiments in which the bias of Naive is at least 900% of the treatment effect; and Section 5.3 provides a rigorous bound to this effect.

[2]Note that in a typical A/B testing, P_0 is often close to P_1 so that δ is expected to be small.

4.2 The Differences-in-Qs Estimator

Intuitively, Naive's bias results from its myopia: Naive simply ignores the impact of the treatment on future rewards. This inspires us to propose a less myopic estimator, which we show here for the special case $p = 1/2$[3]:

$$\hat{\text{ATE}}_{\text{DQ}} := \sum_{t=0}^{\infty} 2 \cdot \left(1(a_t = 1) \left(\sum_{t'=t}^{\infty} r_{t'} \right) - 1(a_t = 0) \left(\sum_{t'=t}^{\infty} r_{t'} \right) \right), \quad (5)$$

In other words, we simply replace the rewards r_t in Eq. (2) by *long-term* accumulated rewards. We refer to this as a "Differences-in-Qs" (DQ) estimator, so called because it estimates the quantity

$$\mathsf{E}_{\pi_{1/2}} \left[\hat{\text{ATE}}_{\text{DQ}} \right] = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_{1/2}} \left[Q_{\pi_{1/2}}(s_t, 1) - Q_{\pi_{1/2}}(s_t, 0) \right] \quad (6)$$

where $Q_\pi(s, a) := \sum_{t=0}^{\infty} E_\pi[r_t | s_0 = s, a_0 = a]$ is the usual state-action value function in MDPs; i.e., the accumulated long-term rewards starting from the state s and the action a. This is actually an instantiation of a broad class of DQ estimators, including prior work [7] as a special case, which we discuss in Section 5.

Surprisingly, this simple change yields a dramatic improvement in the bias in estimating ATE. In fact, the bias of Eq. (5) is *second-order* in δ; precisely, we have

Theorem 1 (Bias of DQ). *Assume that $D_{\text{TV}}(P_1(s, \cdot), P_0(s, \cdot)) \leq \delta$ for all $s \in \mathcal{S}$, and let $r_{\max} := \max_{s,a} |r(s, a)|$. Then,*

$$\left| \text{ATE} - \mathsf{E}_{\pi_{1/2}} \left[\hat{\text{ATE}}_{\text{DQ}} \right] \right| \leq T_{\text{abs}}^3 r_{\max} \cdot \delta^2.$$

Reducing bias from δ to δ^2 produces huge gains in practice: Section 4.3 gives a simple setting in which DQ removes bias entirely, and empirically the bias of DQ is negligible, as we show in Section 7.

Below we remark on some salient properties of the estimator.

DQ is agnostic to state. Beyond its low bias, the greatest advantage of $\hat{\text{ATE}}_{\text{DQ}}$ is its simplicity, in particular the fact that it does not require observations of state s_t to implement. This is critical when s_t is partially observed or extremely high dimensional, as in many real applications – in the Douyin setting state consists of a viewer's sentiment, attention, preferences, device status, and any number of other factors, which are only partially observable by proxy and may be totally distinct from session to session. This lack of state also avoids any assumptions on the functional form of $Q_{\pi_{1/2}}$ and avoids any bias introduced by using function approximation or state aggregation to estimate $Q_{\pi_{1/2}}$.

Credit Assignment Interpretation. This estimator also has a surprising and intuitive interpretation as an explicit "credit assignment" mechanism. Rewriting Eq. (5) explicitly, we have $\hat{\text{ATE}}_{\text{DQ}} = \sum_{t=0}^{\infty} 2 \cdot (\#_t(1) - \#_t(0))\, r_t$ where $\#_t(a) = \sum_{t'=0}^{t} \mathbb{I}\{a_{t'} = a\}$ is the count of actions a played up to and including time t. In effect: we are simply reweighting each reward r_t in an intuitive way, by assigning credit to each action according to the relative contribution of that action in realizing r_t.

4.3 Examples

Finally, we provide some specific examples to build intuition on each of these estimators.

[3]This greatly simplifies exposition; we discuss the general case in Section 6.

Example 1. (Extreme interference). Consider a scenario where the viewer will watch control videos for 15 minutes each, and treatment videos for 20 minutes each; however, the viewer has a fixed attention budget of 30 minutes, after which they leave the platform, stopping in the middle of a video if necessary. In other words, letting s_t be the viewer's remaining time budget upon reaching the t^{th} video, we have $r(s_t, 1) = \min\{20, s_t\}$ min and $r(s_t, 0) = \min\{15, s_t\}$ min. Since the total dwell time is fixed, the true ATE is 0. Upon inspection, we also see that $Q(s, 1) = Q(s, 0)$ for any state, since the total future dwell time does not depend on actions. As a result, DQ is unbiased:

$$\mathsf{E}_{\pi_{1/2}} \left[\hat{\text{ATE}}_{\text{DQ}} \right] = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_{1/2}} \left[Q(s_t, 1) - Q(s_t, 0) \right] = 0.$$

On the other hand, one can compute that[4]

$$\mathsf{E}_{\pi_{1/2}} \left[\hat{\text{ATE}}_{\text{Naive}} \right] = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_{1/2}} \left[r(s_t, 1) - r(s_t, 0) \right] = 5\text{min}$$

This discrepancy results essentially from myopia: without accounting for the downstream effects (or lack thereof) of each action, the effect measured by Naive turns out to be fully a result of cannibalization.

Example 2. (No interference). Again, let $r(s_t, 1) = 20$ min and $r(s_t, 0) = 15$ min. Now suppose that a viewer will watch three videos a day, independent of the treatments. There is no interference in this case: each action impacts watch time for a single video, and has no impact on total videos watched; and thus the total effect of treatment is the sum of the single-step effects. In this case ATE $= (20-15) \cdot 3 = 15$min. The naive estimator is clearly unbiased, while the DQ estimator is unbiased as well in this scenario since for any t,

$$\mathsf{E}_{\pi_{1/2}} \left[1(a_t = 1) \left(\sum_{t'=t}^{\infty} r_{t'} \right) - 1(a_t = 0) \left(\sum_{t'=t}^{\infty} r_{t'} \right) \right] = \mathsf{E}_{\pi_{1/2}} \left[1(a_t = 1) r_t - 1(a_t = 0) r_t \right],$$

i.e., a_t does not impact the values of $\mathsf{E}_{\pi_{1/2}}[r_{t'}]$ for $t' > t$.

5 A UNIFIED THEORY FOR DQ

While the previous section develops a high-level intuition for DQ, here we provide a rigorous analysis which frames DQ as a first-order approximation to the ATE. This interpretation immediately yields the bias bound Theorem 1, as well as a variety of generalizations of the estimator.

More generally, whereas existing theory covers specific reward formulations (specifically average reward [7]), one of our main theoretical contributions is to introduce a *unified* framework for Taylor series expansions in all major reward formulations. These include finite horizon with total reward; infinite horizon with discounted reward; total reward in absorbing Markov chains (which models the Douyin application); and average reward in ergodic Markov chains [7]. Our analysis reveals the fundamental role played by the 'effective horizon' in various scenarios (see Table 1). In addition, our unifying theory of DQ covers [7] as a special case by adopting

[4]To see this: note that $r(s_0, 1) - r(s_0, 0) = 5$, and the user has at most 15 minutes to watch the second video, and therefore $r(s_1, 1) - r(s_1, 0) = 0$; after which the trajectory ends.

a distinct, yet more straightforward proof technique that can be of independent interest.

5.1 Background

We begin by defining precisely each of the Markovian reward settings we address: discounted, average, and total reward settings. In all settings, we consider policies π, π' inducing transition matrices $P_\pi, P_{\pi'}$, and reward functions $r_\pi, r_{\pi'} : \mathcal{S} \mapsto \mathbb{R}$ where $r_\pi(s) = \mathsf{E}_{a\sim\pi(s)}[r(s,a)]$. For each setting and policy, we will define a "value functional" $\tilde{\mathsf{E}}_\pi$, such that $\tilde{\mathsf{E}}_\pi r_\pi$ is the objective of interest (either the discounted reward, average reward, or total reward for policy π) and a corresponding Q function (the 'reward-to-go' after taking some action at a certain state); and we make assumptions on π, π' to ensure that these values are well-defined.

Assumption 2 (Discounted reward). *Let $\gamma \in [0,1)$. For any reward function r, define $\tilde{\mathsf{E}}_\pi r = \sum_{t=0}^\infty \gamma^t \mathsf{E}_\pi[r(s_t)]$ and*

$$Q_\pi(s,a;r) = \sum_{t=0}^\infty \gamma^t \mathsf{E}_\pi[r(s_t)|s_0 = s, a_0 = a].$$

Assumption 3 (Average reward). *Let $P_\pi, P_{\pi'}$ be ergodic Markov chains with stationary distributions $\rho_\pi, \rho_{\pi'}$, and mixing rate $\max_s D_{\mathrm{TV}}(P_\pi^k(s,\cdot), \rho_\pi) \le C\beta^k, \forall k \in \mathbb{N}$ for constants C and $\beta < 1$.*
Let $\tilde{\mathsf{E}}_\pi r = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{E}_\pi[r(s_t)]$ and

$$Q_\pi(s,a;r) = \lim_{T\to\infty} \sum_{t=0}^{T-1} \mathsf{E}_\pi[r(s_t) - \tilde{E}_\pi r|s_0 = s, a_0 = a].$$

Assumption 4 (Finite horizon reward). *Let $H \in \mathbb{N}$ be the horizon, and let $\tilde{\mathsf{E}}_\pi r = \sum_{t=0}^{H-1} \mathsf{E}_\pi[r(s_t)]$ and*

$$Q_\pi(s_t,a_t;r) = \sum_{t'=t}^{H-1} \mathsf{E}_\pi[r(s_t')|s_t = s, a_t = a].$$

Assumption 5 (Total reward in absorbing MDPs). *Let $P_\pi, P_{\pi'}$ be Markov chains with absorbing states $\mathcal{S}_{\mathrm{abs}} \subseteq \mathcal{S}$ and time to absorption bounded by T_{abs}. Let $\tilde{\mathsf{E}}_\pi r = \sum_{t=0}^\infty \mathsf{E}_\pi[r(s_t)]$ and $Q_\pi(s,a;r) = \sum_{t=0}^\infty \mathsf{E}_\pi[r(s_t)|s_0 = s, a_0 = a]$. This corresponds to the setting in Section 3.*

5.2 A Unified Taylor Series Framework

We can now state our main theorem:

Theorem 2. *Let π, π' satisfy any one of Assumptions 2, 3, 4, 5. Then,*

$$\tilde{E}_{\pi'} r = \left(\sum_{k=0}^K \tilde{E}_\pi[\mathrm{DQ}^{(k)}_{\pi,\pi'}(s)]\right) + \tilde{E}_{\pi'}[\mathrm{DQ}^{(K+1)}_{\pi,\pi'}(s)]$$

where $\mathrm{DQ}^{(0)}_{\pi,\pi'}(s) = r_{\pi'}(s)$ and

$$\mathrm{DQ}^{(k)}_{\pi,\pi'}(s) = \mathsf{E}_{a\sim\pi'(s)}[Q_\pi(s,a;\mathrm{DQ}^{(k-1)}_{\pi,\pi'})] - \mathsf{E}_{a\sim\pi(s)}[Q_\pi(s,a;\mathrm{DQ}^{(k-1)}_{\pi,\pi'})]$$

To interpret: Theorem 2 provides a K-th order approximation to evaluating the policy π', using only trajectories generated by π – simply by taking a "sum"[5] of $DQ^{(k)}_{\pi,\pi'}(s)$ terms along such trajectories. The theorem also gives an exact remainder to this approximation, which we bound as $O(\delta^{K+1})$ in the next result:

[5]Appropriately weighted and normalized.

Corollary 1 (K^{th} order remainder). *Let π, π' satisfy any one of Assumptions 2, 3, 4, 5. Further assume that $\max_s D_{\mathrm{TV}}(P_\pi(s,\cdot), P_{\pi'}(s,\cdot)) \le \delta$. Then,*

$$\left|\tilde{\mathsf{E}}_{\pi'} r - \sum_{k=0}^K \tilde{\mathsf{E}}_\pi[\mathrm{DQ}^{(k)}_{\pi,\pi'}(s)]\right| \le (\delta H_{\mathrm{eff}})^{K+1} \|\tilde{\mathsf{E}}_{\pi'}\|_1 r_{\max}$$

where the effective horizon H_{eff} and scaling constant $\|\tilde{\mathsf{E}}_{\pi'}\|_1$ are defined in Table 1, and $r_{\max} = \max_{s\in\mathcal{S}} |r(s)|$.

We note that the scaling constant simply measures how large the value $\tilde{\mathsf{E}}_\pi r$ can be, relative to $r_{\max}$. The combination of Theorem 2 and Corollary 1 offers an elegant, unified theory for DQ estimators. In the following section, we demonstrate how Theorem 1 can be derived as a specific instance of this theory, and we reserve the proof for Section 5.4.

Reward	$\tilde{E}_\pi r$	H_{eff}	$\|\tilde{\mathsf{E}}_{\pi'}\|_1$
Discounted (2)	$\sum_{t=0}^\infty \gamma^t \mathsf{E}_\pi[r(s_t)]$	$1/(1-\gamma)$	$\frac{1}{1-\gamma}$
Average (3)	$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathsf{E}_\pi[r(s_t)]$	$\frac{2\log C+1}{1-\beta}$	1
Finite Horizon (4)	$\sum_{t=0}^{H-1} \mathsf{E}_\pi[r(s_t)]$	H	H
Absorbing (5)	$\sum_{t=0}^\infty \mathsf{E}_\pi[r(s_t)]$	T_{abs}	T_{abs}

Table 1: Summary of the reward settings in Assumptions 2, 3, 4, 5, along with the constants appearing in Corollary 1.

5.3 From the Taylor expansion to DQ

Note that Corollary 1 suggests a recipe for obtaining estimators of the ATE with bias $O(\delta^2)$: we simply estimate the first-order expansion of $J_{\pi_1} = \tilde{\mathsf{E}}_{\pi_1} r_{\pi_1}$, and subtract it from the same approximation of $J_{\pi_0} = \tilde{\mathsf{E}}_{\pi_0} r_{\pi_0}$. The bias results immediately arises from bounding the remainder. For illustrative purposes, we will now execute this recipe in the context of total reward Markovian interference setting described in Section 3, which implements Assumption 5.

As a warm-up: consider the zeroth-order expansion of Theorem 2 (i.e., we set $K = 0$ and omit the remainder). Then, from definitions, we immediately have

$$\begin{aligned}\tilde{\mathsf{E}}_{\pi_1}[\mathrm{DQ}^{(0)}_{\pi_{1/2},\pi_1}(s)] - \tilde{\mathsf{E}}_{\pi_0}[\mathrm{DQ}^{(0)}_{\pi_{1/2},\pi_0}(s)] &= \sum_{t=0}^\infty \mathsf{E}_{\pi_{1/2}}[r_{\pi_1}(s_t) - r_{\pi_0}(s_t)] \\ &= \mathsf{E}_{\pi_{1/2}}\hat{\mathrm{ATE}}_{\mathrm{Naive}}.\end{aligned}$$

In other words: the Naive estimator simply estimates the zeroth-order expansion of Theorem 2. Corollary 1 then immediately yields the $O(\delta)$ bound on bias in general: $\left|\mathrm{ATE} - \hat{\mathrm{ATE}}_{\mathrm{Naive}}\right| \le \delta T_{\mathrm{abs}}^2 r_{\max}$.

To reduce bias, we naturally turn to estimating higher order terms. Continuing with this recipe with $K = 1$, we can show via some algebra that in fact

$$\begin{aligned}\tilde{\mathsf{E}}_{\pi_1}[\mathrm{DQ}^{(0)}_{\pi_{1/2},\pi_1}(s) + \mathrm{DQ}^{(1)}_{\pi_{1/2},\pi_1}(s)] - \tilde{\mathsf{E}}_{\pi_0}[\mathrm{DQ}^{(0)}_{\pi_{1/2},\pi_0}(s) + \mathrm{DQ}^{(1)}_{\pi_{1/2},\pi_0}(s)] \\ = \mathsf{E}_{\pi_{1/2}}[\hat{\mathrm{ATE}}_{\mathrm{DQ}}]\end{aligned}$$

In other words: the DQ estimator simply estimates the first-order expansion of Theorem 2, and as a result the $O(\delta^2)$ bias of Theorem 1 follows immediately from Corollary 1. Thus, we view DQ as the first-order correction to the Naive estimator; its bias properties, and

a recipe for DQ estimators in other settings – including other reward formulations, experiments other than uniform randomization $p = 1/2$, and in the general problem of off-policy evaluation for arbitrary policies – follow immediately from this interpretation. Higher-order estimators can also be derived through similar means, although one pays a cost in variance to estimate them.

5.4 Proof of Theorem 2

All cases of Theorem 2 have elementary proofs, which follow a unified framework. For intuition, consider the discounted reward as defined in Assumption 2, $\tilde{E}_\pi r$. From the definition of P_π, it is well known that

$$\tilde{E}_\pi r = \sum_{t=0}^{\infty} \gamma^t P_\pi^t r = \rho_{\text{init}}^\top (I - \gamma P_\pi)^{-1} r$$

where we overload notation slightly to let $r \in \mathbb{R}^{|\mathcal{S}|}$ be the vector induced by the function r. Similar forms exist for each of the other settings, replacing γP_π with an analogous matrix. To understand how $\tilde{E}_\pi r$ depends on π, then, we must understand how the inverse $(I - \gamma P_\pi)^{-1}$ varies with π.

So motivated, we start with the following elementary perturbation bound (which we note, applies to γP_π)

Lemma 1. *Let $A, A' \in \mathbb{R}^{n\times n}$ be matrices such that $(I - A)^{-1}$ and $(I - A')^{-1}$ exist. Then,*

$$(I - A')^{-1} = (I - A)^{-1} + (I - A')^{-1}(A' - A)(I - A)^{-1} \quad (7)$$

Proof. Observe that $I - A = I - A' + A' - A$. Right-multiply by $(I - A)^{-1}$ and left-multiply by $(I - A')^{-1}$. □

This immediately leads to a series expansion for the matrix $(I - A)^{-1}$:[6]

Lemma 2. *Let $A, A' \in \mathbb{R}^{n\times n}$ be matrices such that $(I - A)^{-1}$ and $(I - A')^{-1}$ exist. Then,*

$$(I - A')^{-1} = (I - A)^{-1} \sum_{k=0}^{K} \left[(A' - A)(I - A)^{-1}\right]^k + (I - A')^{-1} \left[(A' - A)(I - A)^{-1}\right]^{K+1}$$

Proof. The proof follows simply from applying Lemma 1 to $(I - A')^{-1}$ repeatedly up to K times. □

In the discounted case, Theorem 2 then follows immediately from Lemma 2 by letting $A = \gamma P_\pi$ and $A' = \gamma P_{\pi'}$. The other total reward cases follow from the same proof by choosing appropriate analogs of γP_π; whereas the average reward case can be shown as a limiting regime where $\gamma \to 1$. We provide these proofs in full in the online supplement[7].

[6] To see this as a "Taylor" expansion: suppose we let $A' = A + \delta D$ for some direction D. Then, Lemma 2 immediately yields a Taylor series of the function $f(\delta) = (I - A')^{-1}$ in δ, around $\delta = 0$.

[7] https://arxiv.org/abs/2305.02542

6 IMPLEMENTING DQ AT DOUYIN

Here, we expand the model to allow for much more of the complexity present in real applications, in particular in content marketplaces like Douyin/TikTok. In particular, the real world introduces two main challenges not yet discussed: first, as opposed to estimating the treatment effect on a single viewer, we are interested in the aggregate effect across all viewers; and second, as opposed to assigning treatments independently for each viewer for each video, treatments are instead assigned once for each *creator*, and this assignment propagates to all viewers.

First, we show that the estimators and theory from Section 3 continue to work in this setting, with slight modification. Second, we will augment these estimators with additional techniques to fulfill practical requirements: variance reduction, tight variance estimation, and fast implementations. Ultimately, the estimator we describe will be one that, out of the box, provides highly accurate and low variance treatment effect estimates under interference, that satisfy the constraints of real applications.

6.1 A Richer Model of Content Platforms

We begin by detailing the precise model we consider. As mentioned, the total reward formulation of Section 3 models the trajectory of a single viewer session, whereas at the platform level, we are interested in an aggregate effect across all viewers. Note also that each viewer session may be completely heterogeneous. In other words, each viewer session $i \in [N]$ may correspond to a completely distinct MDP, with a distinct reward function $r^{(i)}$, transition kernel $P^{(i)}$, and total reward $J_\pi^{(i)}$. For each session, we observe an entire trajectory $\{s_{it}, a_{it}, r_{it}\}_{t\geq 0}$, where these trajectories *need not be independent across sessions*. Our problem is then, given N such trajectories (i.e., viewer sessions), to estimate the ATE averaged over sessions $i \in [N]$:

$$\overline{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left[J_{\pi_1}^{(i)} - J_{\pi_0}^{(i)}\right] \quad (8)$$

Second, the treatment assignment occurs not independently across viewers and timesteps, but rather once for each creator; and any viewer who encounters that creator, each time they encounter that creator, will receive the same treatment. More formally, there exists a population of creators indexed by $j \in [M]$, each with a random treatment assignment $a(j) \sim \pi_p$. Viewer i at timestep t is shown video j_{it}, and the action played at that timestep is then $a_{it} = a(j_{it})$.

Finally, data collection is typically much more complicated in the real world – there may be multiple treatment groups, and the randomization probability p is typically much smaller than $1/2$. Casting the problem as one of off-policy evaluation allows us to extend the framework naturally to such cases – the data collection policy can essentially be arbitrary. We defer the discussion to Section 6.5.

6.2 Monte-Carlo Estimation At Douyin

To be precise, we begin by discussing a simple Monte-Carlo estimator which can be applied to this more general problem. Recall from Section 4 that this the estimator Eq. (5) is for a single trajectory. This then suggests a straightforward extension to estimate the meta-treatment effect Eq. (8), by simply aggregating DQ estimators across

trajectories. Let $\hat{\text{ATE}}^{(i)}_{\text{DQ}} = \sum_{t=0}^{\infty}\left[\hat{Q}^{\text{MC}}_{\pi_{1/2}}(s_{it},1) - \hat{Q}^{\text{MC}}_{\pi_{1/2}}(s_{it},0)\right]$ be the Monte-Carlo DQ estimator for trajectory i, where $\hat{Q}^{\text{MC}}_{\pi_{1/2}}(s_{it},a) = 2\cdot\mathbb{I}\{a_{it}=a\}\sum_{t=0}^{\infty} r_{it}$ is the Monte-Carlo Q-function estimate. We then have the aggregate estimator:

$$\hat{\text{ATE}}_{\text{DQ}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\text{ATE}}^{(i)}_{\text{DQ}}. \tag{9}$$

By linearity of expectation, the bias of this estimator will be upper bounded by the average of the bias bound in Theorem 1, averaged over MDPs $i \in [N]$.

Regarding the issue of creator-level treatment assignments: As long as the treatment does not alter a video's probability of being shown to the user (i.e., the recommendation system), then from a single viewer's perspective, the marginal treatment probabilities $\Pr(a_{it}=a)$ remain unchanged, and as a result the bias properties of our various estimators remain unchanged. However, this correlation of treatment assignments across viewers introduces challenges for variance quantification, which we address in Section 6.4.[8]

6.3 A Doubly Robust DQ Estimator

The Monte-Carlo estimator Eq. (9) is essentially free of assumptions, and is trivial to implement – it requires only reward observations, and can be computed in a single pass through the dataset. However, this flexibility comes at a cost in terms of variance, as we show in our experiments.

Motivated by the empirical success of doubly robust estimators [13], we introduce a doubly robust DQ estimator. Suppose we have an approximation to the Q function, $\hat{Q}^{\text{Reg}}$ (typically obtained via regression on pre-experiment data, or on some held out set of viewers, as we do in Section 7), which may be misspecified or biased. Then, we define the estimator

$$\hat{\text{ATE}}_{\text{DQ-DR}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\left[\hat{Q}^{\text{DR}}_{\pi_{1/2}}(s_{it},1) - \hat{Q}^{\text{DR}}_{\pi_{1/2}}(s_{it},0)\right] \tag{10}$$

where we now introduce an approximate Q-function and use instead the doubly robust Q-function estimator:

$$\hat{Q}^{\text{DR}}_{\pi_{1/2}}(s_{it},a) = \hat{Q}^{\text{Reg}}_{\pi_{1/2}}(s_{it},a) + \frac{\mathbb{I}\{a_{it}=a\}}{\pi_{1/2}(s_{it},a)}\left[\sum_{t'=t}^{\infty} r_{it'} - \hat{Q}^{\text{Reg}}_{\pi_{1/2}}(s_{it},a)\right]$$

where $\pi(s_{it},a)$ is the probability of selecting action a under the policy π and state s_{it}. This estimator has three salient properties. First, $\hat{Q}^{\text{DR}}_{\pi_{1/2}}$ remains an unbiased estimator of $Q_{\pi_{1/2}}$ regardless of how poorly $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}$ estimates $Q_{\pi_{1/2}}$, as long as the randomization probabilities $\pi(s,a)$ are known. This fact is key, as in the case of partially observable state, $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}$ may be fit heuristically using the observable portion of state, and may in addition perform function approximation or regularization. Second, the introduction of $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}$ serves as a control variate, typically decreasing the variance of the estimator overall; in the experiments we show that this reduces variance by 95% relative to Monte-Carlo DQ estimation.

[8] In fact, a single viewer may view the same video multiple times in the same session, creating dependence in action probabilities across time. In the online supplement, we give a general version of our estimator which accounts for this dependence correctly.

Finally, we consider a subtler issue: robustness to the randomization probabilities $\pi(s,a)$. In particular, in many real situations, only an approximation of the randomization probabilities $\hat{\pi}(s,a)$ is known. This arises most often in observational settings, where $\pi(s,a)$ is simply unknown a priori and must be estimated from data; but is also common in real-world experimental settings, where the complexity of implementation can often result in bugs that skew the realized randomization probability, a discrepancy referred to as "sample ratio mismatch" in [20] (which also provides a wide range of examples under which this can occur). As we will see, the Monte-Carlo estimator Eq. (9) is only unbiased when $\hat{\pi}=\pi$, and when this does not hold the bias can be extremely large in practice.

Doubly robust estimators address this issue by sharply reducing the dependence of the estimator on $\pi(s,a)$. For intuition, consider only the estimator $\hat{Q}^{\text{DR}}_{\pi_{1/2}}$ described above, but replacing $\pi(s,a)$ with an estimate $\hat{\pi}$. The expectation $\mathsf{E}_{\pi_{1/2}}\hat{Q}^{\text{DR}}_{\pi_{1/2}}$ is then

$$\mathsf{E}_{\pi_{1/2}}\hat{Q}^{\text{DR}}_{\pi_{1/2}}(s,a) = \frac{\pi(s,a)}{\hat{\pi}(s,a)}\left[Q_{\pi_{1/2}}(s,a) - \hat{Q}^{\text{Reg}}_{\pi_{1/2}}(s,a)\right] + \hat{Q}^{\text{Reg}}_{\pi_{1/2}}(s,a)$$

We can then immediately see that the better $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}$ estimates $Q_{\pi_{1/2}}$, the less sensitive $\mathsf{E}_{\pi_{1/2}}\hat{Q}^{\text{DR}}_{\pi_{1/2}}$ will be to $\hat{\pi}$. At one extreme, if $Q_{\pi_{1/2}} = \hat{Q}^{\text{Reg}}_{\pi_{1/2}}$, then the doubly robust estimate will be completely unbiased regardless of $\hat{\pi}$; this (together with the fact that $\hat{Q}^{\text{DR}}_{\pi_{1/2}}$ is also unbiased if $\hat{\pi}=\pi$) is the well-known "doubly" robust property of such estimators.

In our experiments Section 7, we show that that the introduction of this doubly robust technique yields dramatically more precise estimators, both when π is known exactly and up to perturbation.

6.4 Variance Characterization under Randomization Testing

A key challenge in understanding the variance of $\hat{\text{ATE}}_{\text{DQ}}$ is that actions are assigned at the streamer level; streamers are shared across viewers; and as result a_{it} can be arbitrarily correlated across viewers.

Whereas exact variance characterizations are difficult in this situation, we propose the use of rerandomization tests to test the sharp null hypothesis H_0 that the treatment has identically zero effect on the reward distribution or problem dynamics. One typical approach to characterizing the distribution of the test statistic $\hat{\text{ATE}}_{\text{DQ}}$ under H_0 is via simulation, where we randomly redraw the assignments $a(j)$ (the assignment of the streamer j), and recompute the test statistic under the new treatment assignments (holding the outcomes constant, as would be prescribed under the null). This "re-randomization" approach yields an exact distribution of the test statistic, but is typically infeasible at Douyin's scale: with on the order of hundreds of millions of viewers, a SQL query for a single realization of the treatment assignments $a(j)$ took around 12 hours to run on Douyin's cluster.

As it turns out, the variance of $\hat{\text{ATE}}_{\text{DQ}}$ under the null hypothesis can actually be computed in *linear* time. The Monte-Carlo estimator Eq. (10) can actually be written as a sum over streamer-level terms, which eventually leads to an *exact* characterization of the permutation testing variance under H_0. To see this, note that the

estimator Eq. (10) is equivalent to

$$\hat{\mathrm{ATE}}_{\mathrm{DQ}} = \sum_{j=1}^{M} \left(\frac{\mathbb{I}\{a(j)=1\}}{p_{j1}} - \frac{\mathbb{I}\{a(j)=0\}}{p_{j0}} \right) \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=t_{ij}}^{\infty} r_{it} \right)$$

where $p_{j0} = \Pr(a(j)=0)$, $p_{j1} = \Pr(a(j)=1)$, and t_{ij} is the time period in which viewer i viewed streamer j. Each of these summands is independent since each assignment $a(j)$ is independent from each other, and as a result we need only to characterize the variance of each summand and sum them in order to compute the variance of $\mathrm{ATE}_{\mathrm{DQ}}$. To conclude this line of thought, note that each summand is an affine function of a Bernoulli random variable, which has closed form expression for variance:

$$\mathrm{Var}(\hat{\mathrm{ATE}}_{\mathrm{DQ}}) = \sum_{j=1}^{M} p_{j0} p_{j1} \left(\frac{1}{p_{j1}} - \frac{1}{p_{j0}} \right)^2 \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=t_{ij}}^{\infty} r_{it} \right)^2$$

Using this variance, we can then perform a hypothesis test, either assuming normality (reasonable in practice) or a Chebyshev-type bound. In the experiments, we show that this test is very high-powered in practice, detecting effect sizes of 0.2% in a day, and achieves nearly nominal coverage.

While this gives a flavor of why linear-time permutation testing should be possible, analogous results for $p \neq \frac{1}{2}$ are much more involved. In those cases, naively pursuing the same approach has complexity $O(M^4)$; while a smarter, exact approach is $O(M^2)$ and only an approximation (albeit empirically a very good one) is possible in $O(M)$ time. We provide an implementation in the supplementary source code, and show in Section 7 that this approach empirically provides nearly exact coverage.

6.5 General Data Collection Policies

Finally, we provide some examples of how to generalize DQ estimation to variety of data collection scenarios. For exposition we have discussed DQ estimation in the specific setting of treatment effect estimation from a A/B test with treatment probability $p = 1/2$, which yields the cleanest results. Even modifying p, however, results in a non-intuitive generalization. In full generality, we collect data under some data-collecting (or, commonly, "behavioral") policy π_{data}, and we would like to evaluate some new target policy π_{new} (or, in some cases, a set of target policies). A range of common scenarios fall under this umbrella:

- A/B testing with $p \neq 1/2$. Let $\pi_{\mathrm{data}} = \mathrm{Bern}(p)$, and we treat π_0, π_1 individually as target policies.
- RCTs with multiple treatment groups $1 \dots K$. Let $\pi_{\mathrm{data}}(s)$ be a distribution over $1 \dots K$, and we have target policies $\pi_1 \dots \pi_K$.
- Evaluating a new policy from logged data. Let π_{data} be the logging policy, and let π_{new} be an arbitrary new policy (with the same support over $\mathcal{A}$ in each state as π_{data}).

On top of this, we also deal with the issue of actions possibly being dependent across time. That is, defining the history $\mathcal{H}_t = \{s_{t'}, a_{t'}, r_{t'}\}_{t'<t}$ we now allow $\pi_{\mathrm{data}}, \pi_{\mathrm{new}}$ to depend arbitrarily on this history. We denote this dependence as $\pi(s, a|\mathcal{H}_t)$.

The Taylor expansion in Theorem 2 provides a DQ analog for each of these generalizations. To be precise: the idealized first-order expansion becomes:

$$J^{\mathrm{DQ}}_{\pi_{\mathrm{new}}} = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_{\mathrm{data}}} \left[\mathsf{E}_{a\sim\pi_{\mathrm{new}}} [r(s_t, a_t) + Q_{\pi_{\mathrm{data}}}(s_t, a_t; r_{\pi_{\mathrm{new}}})] - Q_{\pi_{\mathrm{data}}}(s_t, a_t; r_{\pi_{\mathrm{new}}}) \right]$$

Notably, the resulting Monte-Carlo estimator is more complicated to compute, with multiple importance sampling weights, and this introduces further challenges (and opportunities) for each of the previous extensions to DQ. We defer details to the online supplement. A key limitation of our approach, however, is that the probabilities $\pi_{\mathrm{data}}(s, a|\mathcal{H}_t)$ and $\pi_{\mathrm{new}}(s, a|\mathcal{H}_t)$ must be known, or at least well-estimated; this is trivial in the A/B testing case were the system has perfect control over these probabilities, but potentially difficult in scenarios where prior treatments affect future treatment probabilities. While our doubly-robust estimator exhibits some robustness to misspecified treatment probabilities, the problem merits more thorough investigation in future work.

7 EXPERIMENTS

We now provide empirical results comparing the DQ estimator (and various improvements) on simulations motivated by our particular application at Douyin. Our experimental results aim to address the following research questions:

- **RQ1.** Does DQ provide more accurate treatment effect estimates than existing alternatives?
- **RQ2.** Do our hypothesis testing approaches have correct coverage and sufficient power to detect realistic treatment effects at realistic timescales?
- **RQ3.** How robust are these results under perturbations to π?

7.1 Experimental setting

A typical source of interference at Douyin is in experiments where, due to implementation constraints, Douyin must randomize at the creator (or streamer) level, rather than the viewer level. We take the streaming setting discussed in the introduction as a concrete example: Douyin has a new feature which will allow creators to deliver livestreams in high definition (HD), but as a creator-facing feature, experiments with this feature must be conducted at the creator level. A priori, HD streaming is expected to increase viewer watch time on a per-video basis, but the effect on total watch time is likely to be overestimated by Naive estimation due to cannibalization of viewer attention. Furthermore, the ATE can plausibly be negative because the feature will tend to consume viewers' battery and data budgets much more quickly.

While we have tested the DQ estimator on internal simulators at Douyin, as well as in real-world experimental settings, here we primarily present results from a simulation calibrated to Douyin user data. This approach ensures privacy protection while maintaining the consistency of the qualitative outcomes. We represent state as $s_{it} = (w_t, u_i, v_{it})$, where w_t represents the amount of time the viewer has spent watching videos so far, $u_i \in \mathbb{R}^d$ (here we choose $d = 5$) is a latent vector representing the viewer's preferences, and $v_{it} \in \mathbb{R}^d$ is a latent vector representing the video shown at time t to viewer i, such that $\langle u_i, v_{it} \rangle$ measures the affinity of viewer i

for the t^{th} video shown. We assume that u, v are invisible to the agent, although in reality the agent may have access to some noisy version of these vectors.

The viewer's counterfactual watch time under control is then generated as $w_{it} \sim \text{Exp}(1/(ku_i^\top v_{it}))$, for some scalar problem parameter k. Finally, the viewer's actual watch time is $r_{it} = w_{it}(1 + \tau^* a_{it})$; i.e. it increases by a multiplicative factor of τ^* under treatment $a_{it} = 1$. The viewer's state is either their cumulative watch time $s_t = \sum_{t'=1}^{t-1} r_{it'}$, or the terminal state (denoted by s_{abs}) which indicates that the viewer has left the platform. At each epoch the viewer has a probability of transitioning to the terminal state s_{abs} (i.e., "leaving the platform") determined by s_t as $\mathbb{P}(s_{t+1} = s_{\text{abs}}|s_t) = \frac{\exp(s_t)}{\alpha+\exp(s_t)}$ where α is a problem parameter. We calibrated the parameters k, α to reflect the actual distribution of watch durations and number of videos watched in Douyin's viewer population; we omit the actual values here. In the experimental results below, we generate problem instances with varying treatment effect sizes by varying the parameter τ^*.

7.2 Algorithms

We will compare the performance of both Monte-Carlo and Doubly Robust DQ variants. For Doubly Robust DQ, we construct a simple regression estimator for $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}$: we hold out the first 1000 viewers observed under the experiment policy, and fit via least squares a regression model of the form $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}((w, u, v), a) = \beta_0 + \beta w$ with respect to the parameters β_0, β.

As mentioned, the only alternative being applied in practice is Naive estimation. Unbiased OPE is possible via importance sampling at the level of the entire trajectory; however this is not used in practice, and as one might expect the variance of such an approach is astronomical. Note that each of these approaches work for partially observable state, and are compatible with the streamer-level randomization required in our scenario. To be precise, the typical Naive estimator is $\hat{\text{ATE}}_{\text{Naive}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}[\hat{r}_{\text{IS}}(s_{it}, 1) - \hat{r}_{\text{IS}}(s_{it}, 0)]$ where $\hat{r}_{\text{IS}}$ is the importance sampling estimator $\hat{r}_{\text{IS}}(s_{it}, a) = \frac{\mathbb{I}\{a_{it}=a\}}{\pi(s_{it},a)} r_{it}$ and the typical, stepwise importance sampling OPE estimator is

$$\hat{\text{ATE}}_{\text{OPE}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}[\hat{r}_{\text{IS-OPE}}(s_{it}, 1) - \hat{r}_{\text{IS-OPE}}(s_{it}, 0)]$$

where $\hat{r}_{\text{IS-OPE}}(s_{it}, a_{it})$ is an importance sampling estimator for the expected reward in the t^{th} period under policy π_a, $\mathsf{E}_{\pi_a}[r(s_{it}, a_{it})]$. Here we implement the "stepwise" importance sampling estimator [13], which improves slightly over naive trajectory-level sampling: $\hat{r}_{\text{IS-OPE}}(s_{it}, a) = \left(\prod_{t'=0}^{t} \frac{\mathbb{I}\{a_{it'}=a\}}{\pi(s_{it'},a)}\right) r_{it}$.

The doubly robust approaches used to construct Eq. (10) can also be applied to Naive and OPE estimators, to provide a more refined baseline. We implemented state-of-the-art methods from [16]. It is worth noting, however, that the DR variants will retain the same bias properties; and therefore one will expect Naive-DR to also exhibit large bias, and for OPE-DR to remain unbiased but to still have unreasonably high variance. We also experiment with these estimators, but defer their derivation to the online supplement.

Actual ATE (%)	-0.01		-0.17	
	Mean	SD	Mean	SD
DQ	-0.05 (0.00)	2.01	-0.15 (0.00)	2.95
DQ-DR	-0.05 (0.00)	0.03	-0.15 (0.00)	0.03
Naive	0.30 (0.00)	0.17	0.99 (0.00)	0.25
Naive-DR	0.30 (0.00)	0.01	0.99 (0.00)	0.01
OPE	-0.01 (0.00)	35.42	-0.17 (0.00)	17.24
OPE-DR	-0.01 (0.00)	22.05	-0.17 (0.00)	14.82
Actual ATE (%)	-0.50		-1.89	
	Mean	SD	Mean	SD
DQ	-0.49 (0.00)	2.03	-1.92 (0.00)	2.44
DQ-DR	-0.49 (0.00)	0.03	-1.92 (0.00)	0.04
Naive	2.95 (0.00)	0.17	9.43 (0.00)	0.22
Naive-DR	2.95 (0.00)	0.01	9.43 (0.00)	0.01
OPE	-0.50 (0.00)	32.01	-1.89 (0.00)	26.24
OPE-DR	-0.50 (0.00)	33.32	-1.89 (0.00)	20.87

Table 2: Bias and variance of each estimator, at $N = 10^8$ viewers observed, as a % of the baseline outcome J_{π_0}, for problem instances with various effect sizes (i.e., -0.01%, -0.17%, -0.50%, and -1.89%). Parenthetical quantities are standard errors of the mean estimation. Doubly Robust DQ achieves nearly zero bias and reduces standard deviation of DQ by about 97%.

7.3 Overall Performance Comparison (RQ1)

DQ provides much more accurate estimates of ATE than any alternative. We first consider the root mean-squared error of each estimator in Fig. 2, as well as a breakdown into bias and variance in Table 2. First, we note that interference leads to heavy bias in the Naive estimator: the sign of the Naive estimator is always wrong, a manifestation of the intuition that the intervention is always myopically positive, but can be negative in the long run. Not only is the sign wrong, but the magnitude of the error is large: the Naive estimate is around -800% of the actual ATE in all cases. Next, we turn to the OPE variants. OPE is indeed unbiased, but the variance is so large that its RMSE is by far the largest out of any of the estimators considered, at all timescales considered. This holds for both doubly robust and vanilla versions of the estimator. Neither Naive nor OPE algorithms achieve RMSE below 100%.

Finally turning to DQ, we see that Monte-Carlo DQ achieves much lower RMSE than any Naive or OPE variant, for sufficiently large effect sizes and sufficient observations N. However, its variance remains large even at 10^8 viewers, primarily due to the limited number of streamers observed, and it also fails to achieve relative RMSE below 100% on the effect sizes measured. Doubly Robust DQ, on the other hand, provides striking performance gains: error is reduced by at least 97% compared to Monte-Carlo DQ in all instances, and by up to 99% compared to Naive estimation. By $N = 10^8$ (on the order of one day's worth of viewer activity), DQ-DR already achieves relative RMSE below 100% of the treatment effect on all treatment effects of magnitude greater than 0.17%.

Figure 2: Relative RMSE of each estimator vs. N, i.e., the number of viewers observed, for various effect sizes. Relative RMSE is measured as $\sqrt{\sum_{i=1}^{T}(\hat{\text{ATE}}_i - \text{ATE})^2}/(\sqrt{T}\text{ATE})$ where ATE is the true treatment effect and $\hat{\text{ATE}}_i$ is the estimator's output for the i-th experiment over $T = 100$ seeds. Error bars indicate standard errors over T seeds. See Table 2 for a breakdown of bias vs. variance for all estimators.

7.4 Power and Coverage of Hypothesis Tests (RQ2)

Rerandomization testing detects effect sizes of 0.2% using a day of data. We now consider the problem of hypothesis testing under DQ estimation. Fig. 3 illustrates of the effectiveness of our rerandomization tests as outlined in Section 6.4, which generalizes straightforwardly to Doubly Robust DQ. When analyzing a sample size of 10^8 viewers, which is on the same order as about one day of traffic on Douyin, the Doubly Robust DQ estimator demonstrates its robustness by consistently achieving 80% power for detecting effect sizes as small as 0.15%. In addition, the test successfully adheres to the target false positive rate of 10% when the true treatment effect is zero.

7.5 Robustness to Misspecification (RQ3)

Doubly Robust. Finally, we consider what happens to each estimator when the treatment assignment probabilities are misspecified; i.e., unbeknownst to the algorithm, rather than the nominal probability $p = 1/2$, the environment actually executes an experiment with $p = 1/2 + 1/1000$. Robustness to such misspecification is a first-order concern at Douyin: suppose, for example, that the intervention causes a regression for a tiny proportion of streamers, who then leave the platform and no longer appear in viewers' feeds. The realized treatment probability (i.e., percentage of videos from treated streamers) will then be smaller than nominal. [20] describes a number of other scenarios in which such perturbations occur.

Table 3 shows results for all estimators in this setting. Here, we see that the bias and variance issues continue to plague Naive and OPE, as expected. However, Monte-Carlo DQ now becomes heavily biased as well, estimating a treatment effect of the wrong sign and of substantially larger magnitude than the actual treatment effect, despite the small change in the treatment probabilities. In contrast, the bias of Doubly Robust DQ is almost unchanged, remaining at most a few percent of the ATE. This is despite the fact that the regression estimator we use for $\hat{Q}^{\text{Reg}}_{\pi_{1/2}}$ is extremely crude, and extremely poorly specified, speaking to the ease with which doubly robust estimators can be constucted, and to the outsize impact of doing so.

8 CONCLUSION

In conclusion, our study addresses the critical issue of interference in experiments conducted on two-sided content marketplaces, such as Douyin. We have demonstrated the limitations of naive estimators and the impracticality of off-policy estimators due to their high variance. To overcome these challenges, we introduced a novel Monte-Carlo estimator based on Differences-in-Qs (DQ) techniques, which provides second-order bias in the treatment effect while remaining sample-efficient. Our theoretical contribution lies in the development of a generalized theory of Taylor expansions for policy evaluation, which extends DQ theory to all major MDP formulations. This advancement significantly broadens the applicability of the DQ approach in various experimental settings. On the practical side, we successfully implemented our estimator on Douyin's experimentation platform, transforming DQ into a truly "plug-and-play" estimator for interference in real-world settings.

REFERENCES

[1] Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas Richardson, and Ido M Rosen. 2021. Multiple Randomization Designs. *arXiv preprint arXiv:2112.13495* (Dec. 2021). arXiv:2112.13495 [cs, econ, math, stat]

[2] Thomas Blake and Dominic Coey. 2014. 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 (EC '14)*. Association for Computing Machinery, New York, NY, USA, 567–582. https://doi.org/10.1145/2600057.2602837

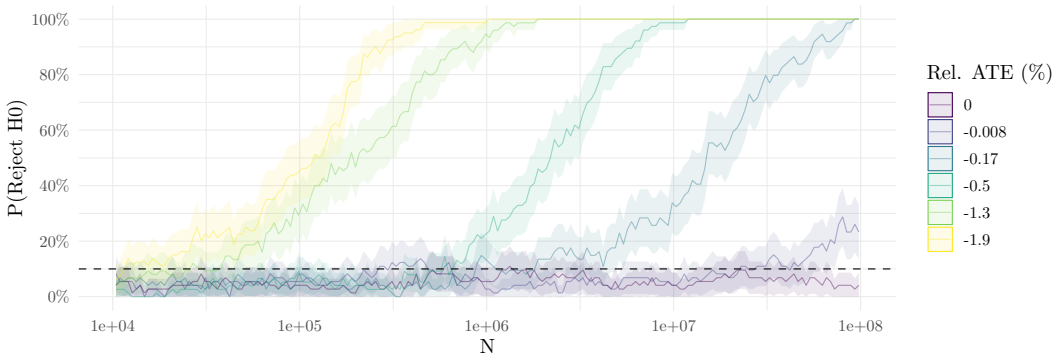

Figure 3: $\mathbb{P}$(Reject H_0) (i.e., power) vs. number of viewers sampled for the proposed approximate permutation test, for the doubly robust estimator, for various effect sizes, at a 90% confidence level. Error bars show standard errors over 100 random seeds. At 10^8 viewers, roughly one day of traffic at Douyin, Doubly Robust DQ reliably achieves 80% power for effect sizes as small as 0.15%.

Actual ATE (%)	-0.01		-0.17	
	Mean	SD	Mean	SD
DQ	4.98 (0.24)	2.20	4.49 (0.28)	2.56
DQ-DR	-0.06 (0.00)	0.04	-0.12 (0.00)	0.03
Naive	0.71 (0.02)	0.18	1.38 (0.02)	0.22
Naive-DR	0.30 (0.00)	0.01	0.99 (0.00)	0.00
OPE	1.22 (3.11)	28.80	5.94 (3.07)	28.13
OPE-DR	-1.24 (3.32)	26.81	4.57 (3.37)	26.52
Actual ATE (%)	-0.50		-1.89	
	Mean	SD	Mean	SD
DQ	4.43 (0.28)	2.52	2.70 (0.26)	2.37
DQ-DR	-0.48 (0.00)	0.04	-1.90 (0.00)	0.04
Naive	3.35 (0.02)	0.21	9.82 (0.02)	0.21
Naive-DR	2.95 (0.00)	0.01	9.43 (0.00)	0.01
OPE	-0.40 (3.29)	30.14	3.49 (2.59)	23.64
OPE-DR	-1.69 (1.87)	14.87	-8.18 (2.31)	18.73

Table 3: Estimation results for all estimators and various effects sizes (−0.01%, −0.17%, −0.50%, −1.89%), where the estimator uses $p = 1/2$ but in fact the policy executes $p = 1/2+1/1000$. Naive and OPE estimators continue to display high bias / high variance respectively. Monte-Carlo DQ is now significantly biased, while Doubly Robust DQ remains essentially unperturbed.

[3] Iavor Bojinov and Neil Shephard. 2019. Time series experiments and causal estimands: exact randomization tests and trading. *J. Amer. Statist. Assoc.* 114, 528 (2019), 1665–1682.

[4] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. 2022. Design and analysis of switchback experiments. *Management Science* (2022).

[5] Nicholas Chamandy. 2016. Experimentation in a Ridesharing Marketplace. https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e.

[6] Angus Deaton and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social science & medicine* 210 (2018), 2–21.

[7] Vivek Farias, Andrew Li, Tianyi Peng, and Andrew Zheng. 2022. Markovian interference in experiments. *Advances in Neural Information Processing Systems* 35 (2022), 535–549.

[8] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.

[9] Peter W Glynn, Ramesh Johari, and Mohammad Rasouli. 2020. Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems* 33 (2020), 15054–15064.

[10] Henning Hohnhold, Deirdre O'Brien, and Diane Tang. 2015. Focus on the long-term: It's better for users and business. (2015).

[11] David Holtz and Sinan Aral. 2020. *Limiting Bias from Test-Control Interference in Online Marketplace Experiments.* SSRN Scholarly Paper 3583596. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3583596

[12] David Michael Holtz. 2018. *Limiting bias from test-control interference in online marketplace experiments.* Ph. D. Dissertation. Massachusetts Institute of Technology.

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

[14] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. 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*. 1517–1525.

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

[16] Nathan Kallus and Masatoshi Uehara. 2020. Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes. *J. Mach. Learn. Res.* 21, 167 (2020), 1–63.

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

[18] Rochelle King, Elizabeth F Churchill, and Caitlin Tan. 2017. *Designing with data: Improving the user experience with A/B testing.* " O'Reilly Media, Inc.".

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

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

[21] Hannah Li, Geng Zhao, Ramesh Johari, and Gabriel Y Weintraub. 2022. Interference, bias, and variance in two-sided marketplace experimentation: Guidance for platforms. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 182–192. https://doi.org/10.1145/3485447.3512063

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

[23] David Lucking-Reiley. 1999. Using field experiments to test equivalence between auction formats: Magic on the Internet. *American Economic Review* 89, 5 (Dec. 1999), 1063–1080. https://doi.org/10.1257/aer.89.5.1063

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

[25] Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. 2001. Off-policy temporal-difference learning with function approximation. In *ICML*. 417–424.
[26] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *Intl. conf. on machine learning*. PMLR, 1889–1897.
[27] Dan Siroker and Pete Koomen. 2015. *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.
[28] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. 2004. Randomized controlled trials. *AJR Am J Roentgenol* 183, 6 (2004), 1539–44.
[29] Claudio J Struchiner, Mary Elizabeth Halloran, James M Robins, and Andrew Spielman. 1990. The behaviour of common measures of association used to assess a vaccination programme under complex disease transmission patterns—a computer simulation study of malaria vaccines. *Intl. journal of epidemiology* 19, 1 (1990), 187–196.
[30] Richard S Sutton, Csaba Szepesvári, and Hamid Reza Maei. 2008. A convergent O (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Adv. in neural information processing systems* 21, 21 (2008), 1609–1616.
[31] M Talbot, AD Milner, MAE Nutkins, and JR Law. 1995. Effect of interference between plots on yield performance in crop variety trials. *The J. of Agricultural Science* 124, 3 (1995), 335–342.
[32] Eric J Tchetgen Tchetgen and Tyler J VanderWeele. 2012. On causal inference in the presence of interference. *Statistical methods in medical research* 21, 1 (2012), 55–75.
[33] Philip Thomas and Emma Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *Intl. Conf. on Machine Learning*. PMLR, 2139–2148.
[34] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High-confidence off-policy evaluation. In *Proc. of the AAAI Conf. on Artificial Intelligence*, Vol. 29.
[35] Vivek Trehan. 2019. Marketplace Experimentation. https://www.youtube.com/watch?v=IR000RqN7pw. Accessed: 2023-03-24.
[36] Jon Vaver and Jim Koehler. 2011. Measuring ad effectiveness using geo experiments. (2011).
[37] Dylan Walker and Lev Muchnik. 2014. Design of randomized experiments in networks. *Proc. of the IEEE* 102, 12 (2014), 1940–1951.
[38] Corwin M Zigler and Georgia Papadogeorgou. 2021. Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics* 36, 1 (2021), 109.