arXiv:2305.02542v1 [stat.ME] 4 May 2023

实验中的干扰校正：抖音案例研究

VIVEK F. FARIAS，麻省理工学院，美国HAO LI，字节跳动，中国TIANYI PENG，麻省理工学院，美国
XINYUYANG REN，字节跳动，中国HUAWEI ZHANG，字节跳动，中国ANDREW ZHENG，麻省理工学院，美国

干扰是双边内容市场实验中普遍存在的问题，例如抖音（中国版 TikTok）。在许多情况下，创作者是自然的实验单位，但由于观众的时间和注意力有限，创作者之间会因争夺观众而相互干扰。目前实践中使用的“朴素方法”估计量会直接忽略干扰，但这样做会产生与处理效应同阶的偏差。我们将此类实验中的推断问题形式化为策略评估问题。离策略估计量虽然无偏，但方差高到不切实际。我们提出了一种新颖的蒙特卡洛估计量，基于“Q 值差分”（DQ）技术，其偏差是处理效应的二阶量，同时仍能通过样本高效地进行估计。在理论方面，我们的贡献是为策略评估建立一种广义泰勒展开理论，将 DQ 理论扩展到所有主要的马尔可夫决策过程（MDP）表述形式。在实践方面，我们在抖音的实验平台上实现了该估计量，并在此过程中将 DQ 发展为一种真正“即插即用”的现实场景干扰估计量：它能提供稳健、低偏差、低方差的处理效应估计；支持计算成本低且渐近精确的不确定性量化；在我们的应用中，与现有最佳替代方案相比，均方误差（MSE）降低了 99%。

其他关键词与短语：干扰、A/B 测试、实验、离策略评估、强化学习

ACM 引用格式：
Vivek F. Farias、Hao Li、Tianyi Peng、Xinyuyang Ren、Huawei Zhang 和 Andrew Zheng。2023。实验中的干扰校正：抖音案例研究。1, 1（2023年5月），18 页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

近年来，大型在线平台日益依赖实验来衡量干预的影响并提升用户体验 [8, 19, 28]。传统的随机对照试验（RCT），通常称为 A/B 测试，为处理组和对照组相互独立的情境提供了稳健的因果推断框架 [6, 29]。然而，在许多在线平台中，处理组和对照组的结果并非相互独立，而是通过共享的系统状态相互影响，这种现象称为“马尔可夫式”干扰 [7]。应对这种干扰是先进实验平台面临的一项关键挑战 [2, 12, 21, 22]。

作者地址：Vivek F. Farias，麻省理工学院，美国；Hao Li，字节跳动，中国；Tianyi Peng，麻省理工学院，美国；Xinyuyang Ren，字节跳动，中国；Huawei Zhang，字节跳动，中国；Andrew Zheng，麻省理工学院，美国。

允许出于个人或课堂用途免费制作本作品全部或部分内容的数字副本或纸质副本，但前提是这些副本不得用于营利或商业利益目的，不得为此进行复制或分发，并且副本首页必须附有本声明及完整引文。对于本作品中归其他方而非 ACM 所有的组件，必须尊重其版权。注明出处后，允许摘录。以其他方式复制、重新出版、发布到服务器或重新分发至列表，均须事先获得特定许可和/或支付费用。请通过 permissions@acm.org 申请授权。© 2023 美国计算机协会。提交至 ACM 的稿件

在这项工作中，我们提出了一种新方法，用于解决在线内容市场（如抖音）中出现的干扰问题。抖音是一家领先的社交视频平台，截至2021年初，日活跃用户达6亿。[1] 与其美国版 TikTok 一样，抖音的核心产品是短视频内容：创作者在应用上制作视频，观众则会看到一系列视频（由公司的专有推荐算法决定）。在典型的观看会话中，观众打开应用；平台向其展示一个视频；观众观看该视频，直到滑动屏幕表示想切换到下一个视频；随后平台再向其展示另一个视频。这个过程会反复进行，直到观众最终离开平台。

与许多其他双边市场一样，在抖音开展实验的一项关键挑战是，某些干预只能通过创作者侧实验进行测试，而许多关键指标却是在观众侧衡量的。举一个具体的真实案例：抖音推出了一项新功能，允许创作者以高清画质进行直播，抖音希望测试这项新功能对观众参与度的影响。一个典型指标是观众的平均“停留时间”——也就是在某个时间段内，平均每位观众在平台上花费的总时间。

由于该功能是在创作者层面推出的，平台无法控制哪些观众会接触到该功能，而这正是进行观众侧或双边随机实验所必需的。创作者之间的干扰主要是因为观众用于抖音的资源有限——这些资源包括空闲时间、注意力、电池续航以及其他因素——而处理组的创作者会与对照组的创作者争夺这些资源。说明性示例见图 1 。在这类情形下，抖音目前采用的唯一方案是朴素估计；换言之，就是忽略干扰，并像在标准随机对照试验中那样进行估计。然而，众所周知，这种朴素估计量存在相当大的偏差，其偏差甚至可能与处理效应本身一样大 [2, 11]。

贡献。 为解决干扰问题，我们首先将处理效应估计表述为马尔可夫决策过程（MDP）中的离策略评估问题，其中状态可以任意复杂，例如由观众偏好、参与度和观看历史构成。尽管现有的离策略评估方法无偏，但由于状态空间大得难以处理，它们无法适用于这种情形，而且方差过高。相比之下，朴素估计量不需要获取状态，但会产生显著偏差。为克服这些局限，我们提出了一种新颖的估计量，它无需获取状态，且相较于朴素估计量具有二阶偏差。我们提出的估计量可以视为一种差分 Q（DQ）估计量。近期的一项理论研究 [7] 表明，DQ 估计量在平均奖励马尔可夫决策过程中具有良好的偏差-方差权衡。

我们将提出的 DQ 估计量集成到抖音平台中，并结合多种推广形式，证明其性能优于现有方法。总而言之，我们的贡献有三点：

1）一种新颖且实用的估计量。我们提出了一种新颖的 DQ 估计量，用于校正抖音 A/B 测试中的干扰。与其他离策略评估方法相比，该估计量无需访问状态，因而适用范围更广。我们从理论上证明了其偏差为二阶（定理 1），并且它还具有实用优势，可适用于异质用户、部分可观测状态以及多个同时进行的实验（第 6 节）。此外，我们开发了双重稳健技术来降低方差（第 6.4 节），并实现了对方差的精确量化（第 6.4 节）。

2) 卓越的实证表现。 我们在抖音的实验平台上实现了我们的估计量，并在第 7 节报告了对该实现进行建模的大规模模拟器的结果。这些实验表明，DQ 估计量显著优于最先进的方法，将均方误差（MSE）降低了 99%

[1] https://new.qq.com/rain/a/20210329A06EP300提交至 ACM 的稿件

与现有最佳替代方法相比。我们的工作代表了 DQ 估计量在现实世界中的首次大规模实施。

3）DQ 的统一理论。最后，我们对 DQ 估计量及其基础泰勒级数理论（第5节）进行了理论扩展，使其能够更自然地适用于抖音应用中的折扣奖励和总奖励形式。我们的统一理论将 [7] 中的既有平均奖励结果作为特例推导出来，且本身也具有独立的研究价值。

2 相关工作

在公共卫生、农业和在线市场等领域，实验中的干扰 [24, 30, 32], 是指实验单位的结果受到其他单位影响的现象，这可能导致处理效应估计出现偏差。干扰已成为在线平台面临的一项关键挑战，尤其是在 eBay [2], 、Uber/Lyft [5, 36], 、Airbnb [11], 和抖音等公司中；抖音正是本研究的背景。

实验设计。现有应对干扰的方法主要侧重于复杂的实验设计。例如，通过对单位进行聚类来尽量减少干扰 [25, 37、在不同时间段交替进行随机化39], 、在双边市场中同时对供给侧和需求侧进行随机化 [3, 4, 10], ，以及采用其他设计[1, 16, 22], 。尽管这些方法有潜在优势，但成本和实施方面的顾虑等实际限制，往往妨碍了此类复杂设计的应用 [9, 15, 33]，我们在抖音的动机应用中也会看到这一点。与引入更复杂的设计不同，本文在简单的单位级随机化下提出了一种新颖的估计量，从而有效缓解干扰造成的偏差。

离策略评估（OPE）。我们的问题可以视为离策略评估（*OPE*） [26, 31], 这一强化学习中日益重要问题的一个案例。无偏离线策略评估估计量通常面临高方差[23, 34, 35], ，而在大型状态空间中还会遭遇相关的维数灾难 [13, 17, 18]。我们的方法可以被视为一种用于离策略评估的有偏方法，它以少量偏差换取方差的大幅降低（类似于 [7, 27]），并且无需访问状态信息即可提供强有力的理论保证。

3 模型

我们首先在一个简化的设定中建立理论，在该设定下，我们观察单个观众的一条轨迹（或“会话”）。在第6节中，我们将把这一核心理论扩展到一个更丰富的问题表述，以涵盖抖音应用中的复杂性。

考虑这样一种情形：平台想要测试一项面向创作者的干预——例如，推出引言中提到的高清直播功能。由于基础设施限制或对干预影响的顾虑，我们必须在创作者层面进行实验，因此实施面向创作者的 A/B 测试。

我们将观众的每个会话建模为一个马尔可夫决策过程。在每个步骤 t，平台都会向观众呈现一个视频。以概率 p，呈现的视频属于处理组，对应一个动作 $a_t = 1$；否则，视频属于对照组，对应一个动作 $a_t = 0$。当观众滑动到下一个视频时，平台会获得一个奖励 r_t——例如观众观看上一个视频所花费的时间——然后马尔可夫决策过程推进到下一步骤。奖励 r_t 是 $a_t \in \mathrm{A}$（时刻 t 的动作）和 $s_t \in \mathrm{S}$（观众在时刻 t 的状态）的一个（随机）函数。

关键在于，在实际应用中，状态可能任意复杂，且部分或完全不可观测。例如，s_t 可能由观众在平台上花费的时间（以及观众剩余时间预算的隐含信息）、其参与度，甚至观众观看过的完整视频观看历史组成。参见图1 中的示例。正如我们将看到的，我们可以设计出不受这种复杂性影响的估计量。

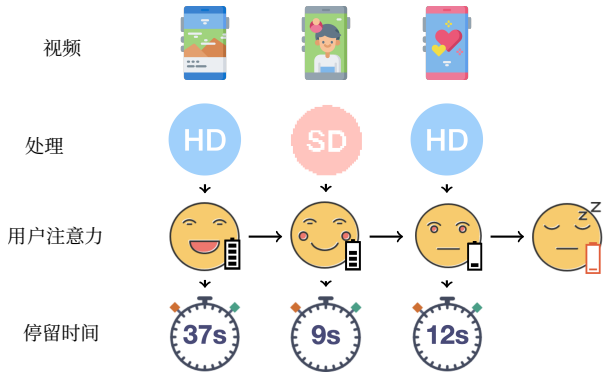

图 1. 针对特定观众的抖音马尔可夫决策过程（MDP）示意图。在每个时间步，观众都会看到一个视频；该视频要么是处理（高清，$a_t = 1$），要么是对照组（标清，$a_t = 0$）；随后，平台会根据视频和观众当前的状态（例如注意力、电池续航、数据流量预算等）获取一定的奖励（即观众的停留时间），并转到下一个视频。最终，观众会耗尽注意力并离开平台。

策略 π： $\mathcal{S} \to \mathcal{A}$ 是从状态映射到动作的（随机）映射。我们关注三种策略：

(1) 全局处理 π_1，其选择 $\pi_1(s) = 1, \forall s \in \mathcal{S}$，即展示的每个视频都接受处理。相应的转移矩阵记为 $P_1 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$。(2) 全局对照 π_0，其选择 $\pi_0(s) = 0, \forall s \in \mathcal{S}$，即每个视频都不接受处理。相应的转移矩阵记为 $P_0 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$。(3) 实验 π_p，其以 p 的概率选择 $\pi(s) = 1$；否则选择 $\pi(s) = 0$；换言之，即视频侧 A/B 测试。相应的转移矩阵记为 $P_p := (1-p)\cdot P_0 + p\cdot P_1$。为便于说明，我们设 $p = 1/2$，并在第 6 节讨论推广情形。

对于给定的策略 π，我们将该策略的总奖励定义为：$J_\pi = \sum_{t=0}^{\infty} \mathrm{E}_\pi[r_t]$，其中记号 E_π 表示对状态 $s_0 \sim \rho_{\text{init}}$、动作 $a_t|s_t \sim \pi(s_t)$、动态 $s_{t+1}|s_t \sim P_\pi(s_t, s_{t+1})$ 以及奖励取期望。

目标。给定实验策略 $\pi_{1/2}$下的观测 $\{(r_t, a_t)\}$，我们的目标是估计 π_1 和 π_0所获得的总奖励之差：

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

在上面的例子中，平均处理效应对应于将策略π_0 替换为 π_1对观众在平台上的停留时间产生的影响；在决定是否推出新策略时，这是一个关键的关注指标。

为确保 J_π 定义良好，我们对每个策略所诱导的马尔可夫链作出一项关键假设：观众最终会结束其会话；也就是说，所有马尔可夫链都是吸收式的。具体而言：若对于所有 $s \in \mathcal{S}_{\text{abs}}$，在策略 π 下，$\mathcal{S}_{\text{abs}} \subseteq \mathcal{S}_{\text{abs}}$，且若 $r_t = 0$ a.s. when $s_t \in \mathcal{S}_{\text{abs}}$，则称$\mathcal{S}_{\text{abs}} \subseteq \mathcal{S}$在策略 π 下是一个吸收类。我们还要求，从任意状态出发，几乎必然会在有限时间内到达$\mathcal{S}_{\text{abs}}$，并将到达这一吸收事件所需的期望时间定义为T^π_{abs}：$= \max_{s\in\mathcal{S}} \sum_{t=0}^{\infty} \mathrm{E}_\pi[\mathbb{I}\{s_t \notin \mathcal{S}_{\text{abs}}\} | s_0 = s]$。接下来作出如下假设：

假设 1。存在一个类别 $\mathcal{S}_{\text{abs}} \subseteq \mathcal{S}$ 以及一个时间 T_{abs}，使得 $\mathcal{S}_{\text{abs}}$ 是吸收式的，并且 $T_{\text{abs}} > T^\pi_{\text{abs}}$ 对于每个策略 π 属于 π_0， π_1， $\pi_{1/2}$。

4 干扰条件下估计的新型估计量

4.1 朴素方法估计量

我们首先介绍这一场景下的现状：朴素估计，即完全忽略干扰，并将实验视为传统的 A/B 测试。一种典型的估计量是逆倾向加权估计量。

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

该方法仅估计处理组与对照组获得的单步骤奖励之差，而忽略了每个动作对整体轨迹的影响。

然而，这个估计量的偏差可能很大。为说明这一点，请注意，期望为

$$\mathrm{E}_{\pi_{1/2}}\left[\hat{\mathrm{ATE}}_{\mathrm{Naive}}\right] = \sum_{t=0}^{\infty}\mathrm{E}_{\pi_{1/2}}\left[r(s_t,1)\right] - \mathrm{E}_{\pi_{1/2}}\left[r(s_t,0)\right] \tag{3}$$

另一方面，平均处理效应 $=\sum_{t=0}^{\infty}\mathrm{E}_{\pi_1}[r(s_t,\ 1)] - \sum_{t=0}^{\infty}\mathrm{E}_{\pi_0}[r(s_t,\ 0)]$. 平均处理效应与 $\mathrm{E}_{\pi_{1/2}}\left[\hat{}\mathrm{ATE}_{\text{朴素方法}}\right]$ 之间的差异，源于这样一个事实：我们通过从 $\pi_{1/2}$ 中抽样所计算出的期望，与我们希望在 $\pi_0,\ \pi_1$ 条件下计算的期望并不一致。

直观地说，偏差的大小取决于在 π_0 下与在 π_1 下的动态差异。我们将这一差异刻画如下：从任意状态出发，令 δ 界定下一状态分布之间的距离。具体而言，$\delta := \max_s D\mathrm{TV}(P_0(s,\cdot)\ P_1(s,\cdot))$，其中 DTV 衡量全变差距离[2]。那么，$\hat{}$ATE朴素估计量的偏差为 $\Theta(\delta)$，而平均处理效应本身也是 $\Theta(\delta)$——也就是说，朴素估计的偏差与处理效应处于同一量级。我们将通过多种方式证明这一点：第 4.3 节给出了一个简单示例，其中朴素估计量测得的效应实际上只是蚕食效应；第 7 节包含一些现实的实验，在这些实验中，朴素方法的偏差至少达到处理效应的 900%；第 5.3 节则对这一效应给出了严格的界。

4.2 Q值差分估计量

直观地说，朴素方法的偏差源于其短视性：它完全忽略了处理对未来奖励的影响。这促使我们提出一种短视性较低的估计量；下面展示其在特殊情形下的形式 $p = 1/2$[3]：

$$\hat{\mathrm{ATE}}_{\mathrm{DQ}} := \sum_{t=0}^{\infty} 2\cdot\left(1(a_t=1)\left(\sum_{t'=t}^{\infty} r_{t'}\right) - 1(a_t=0)\left(\sum_{t'=t}^{\infty} r_{t'}\right)\right), \tag{4}$$

换言之，我们只需将式（ r_t （2）中的奖励替换为长期累积奖励。我们将其称为“Q值差分”（DQ）估计量，因为它估计的是以下量：

$$\mathrm{E}_{\pi_{1/2}}\left[\hat{\mathrm{ATE}}_{\mathrm{DQ}}\right] = \sum_{t=0}^{\infty}\mathrm{E}_{\pi_{1/2}}\left[Q_{\pi_{1/2}}(s_t,1) - Q_{\pi_{1/2}}(s_t,0)\right] \tag{5}$$

其中，Q_π（$s,\ a$）：$=\sum_{t=0}^{\infty}E_\pi[r_t|s_0=s,a_0=a]$ 是马尔可夫决策过程中常见的状态-动作价值函数，即从状态 s 和动作 a出发所获得的累积长期奖励。实际上，这属于一大类DQ估计量的一个实例，其中包括先前研究 [7] 作为特例，我们将在第5节讨论。

令人惊讶的是，这一简单的改变显著改善了 平均处理效应 估计中的偏差。事实上，式（4）的偏差是 二阶的 ，关于 δ；具体而言，我们有

[2]请注意，在典型的 A/B 测试中，P_0 通常接近于 P_1，因此 δ 预期会很小。[3]这大大简化了论述；我们将在第6节讨论一般情况。

理论 命题 1（DQ 的偏差）。假设 $\forall s \in \mathcal{S}$、 $D_{\mathrm{TV}}(P_1(s,\cdot)\ P_0(s,\cdot)) \le \delta$，并令 $r_{\text{最大值}} := \text{最大值}_{s,a}|r(s,a)|$。那么，

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

将偏差从 δ 降低至 δ^2 在实践中能带来巨大收益：第 4.3 节给出了一个简单的情形，在该情形下 DQ 能完全消除偏差；正如第 7 节所示，在实证中 DQ 的偏差也可以忽略不计。

下面我们将介绍该估计量的一些显著特性。

ˆDQ 不依赖于状态。 除了偏差较低之外，$\mathrm{ATE}_{\mathrm{DQ}}$ 最大的优势在于其简单性，尤其是它不需要观测状态 s_t 即可实现。当 s_t 部分可观测或维度极高时，这一点至关重要，这在许多实际应用中都很常见——在抖音的场景中，状态包括观众的情绪、注意力、偏好、设备状态以及其他许多因素，而这些因素只能通过代理变量部分观测到，并且可能因会话而异。无需状态信息，也就无需对 $Q_{\pi_{1/2}}$的函数形式作出任何假设，也避免了使用函数逼近或状态聚合来估计$Q_{\pi_{1/2}}$所引入的偏差。

信用分配解释。 该估计量还可以被直观地理解为一种明确的“信用分配”机制。将公式（9）明确改写为：$\hat{}\mathrm{ATE}_{\mathrm{DQ}} = \sum_{t=0}^{\infty} 2 \cdot (\#_t(1) - \#_t(0)) r_t$ 其中，$\#_t(a) = \sum_{t'=0}^{t} \mathbb{I}\{a_{t'} = a\}$ 表示截至时间 t（包括该时刻）已播放的动作 a 数量。实际上，我们只是在以一种直观的方式对每个奖励 r_t 重新加权：根据每个动作在实现 r_t中的相对贡献，为其分配相应的信用。

4.3 示例

最后，我们提供一些具体示例，以帮助直观理解上述每个估计量。

示例 1。**（极端干扰）**。 考虑这样一种情形：观众将分别观看对照视频 15 分钟，以及处理视频 20 分钟；即，$r(s_t, 1) = 20$ 分钟，以及 $r(s_t, 0) = 15$ 分钟。然而，观众的注意力预算固定为 30 分钟，之后便会离开平台；如有必要，他们会在视频播放到一半时停止观看。由于总停留时间是固定的，真实的平均处理效应为 0。经检查我们还可以看到，对于任意状态，都有 $Q(s, 1) = Q(s, 0)$，因为未来总停留时间与动作无关。因此，DQ 是无偏的：

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

另一方面，可以计算得出，$\mathsf{E}_{\pi_{1/2}}\left[\hat{}\text{平均处理效应}_{\text{朴素方法}}\right] = \sum_{t=0}^{\infty} \mathsf{E}_{\pi_{1/2}}\left[r(s_t, 1) - r(s_t, 0)\right] = 2.5\text{min.}$ </style>这种差异本质上源于短视性：如果不考虑每个动作的下游效应（或没有下游效应），朴素方法测得的效应最终完全是蚕食效应的结果。

示例 2。**（无干扰）**。同样，设 $r(s_t, 1) = 20$ 分钟，且 $r(s_t, 0) = 15$ 分钟。现在假设观众每天观看三个视频，且与处理无关。在这种情况下不存在干扰：每个动作只影响单个视频的观看时长，不影响观看视频的总数；因此，处理的总效应等于各单步骤效应之和。在这种情况下，平均处理效应 $= (20 - 15) \cdot 3 = 15$分钟。显然，朴素估计量是无偏的，而在此情形下，DQ 估计量也是无偏的，因为对于任意 t,

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

即， a_t不会影响$\mathsf{E}_{\pi_{1/2}}[r_{t'}]$ 对于 $t' > t$.

5 DQ 的统一理论

<escape>虽然前一节从宏观层面阐述了 DQ 的直观原理，但在这里，我们将提供严格的分析，将 DQ 定义为平均处理效应的一阶近似。这一解释直接得出了偏差界定定理 1，以及该估计量的多种推广形式。
</escape>

更一般地，尽管现有理论涵盖了特定的奖励形式（具体而言是平均奖励 [7]），我们的主要理论贡献之一是为所有主要奖励形式中的泰勒级数展开引入一个 统一的 框架。这些形式包括具有总奖励的有限时域；具有折扣奖励的无限时域；吸收马尔可夫链中的总奖励（用于建模抖音应用）；以及遍历马尔可夫链中的平均奖励[7]。我们的分析揭示了“有效时域”在各种场景中发挥的根本作用（见表1）。此外，我们关于 DQ 的统一理论将 [7] 作为一个特例涵盖在内，采用了一种独特但更直接的证明技术，这种技术本身也具有独立的研究价值。

5.1 背景

我们首先精确定义所讨论的每种马尔可夫奖励设定：折扣奖励、平均奖励和总奖励设定。在所有设定中，我们考虑策略 π， π'，它们诱导出转移矩阵 P_π， $P_{\pi'}$ 和奖励函数r_π， $r_{\pi'}$: $\mathrm{S} \mapsto \mathbb{R}$ 其中 $r_\pi(s) = \mathrm{E}_{a\sim\pi(s)}[r(s,\ a)]$。对于每种设定和策略，我们将定义一个“价值泛函” $\tilde{\mathrm{E}}_\pi$，使得$\tilde{\mathrm{E}}_\pi r_\pi$ 是我们关注的目标（即策略 π的折扣奖励、平均奖励或总奖励），并定义相应的 Q 函数（在某个状态采取某个动作后的“后续奖励”）；我们还对π， π' 作出假设，以确保这些价值定义良好。

假设 2（折扣奖励）。令 $\gamma \in [0, 1)$。对于任意奖励函数 r，定义 $\tilde{\mathrm{E}}_\pi r = \sum_{t=0}^{\infty} \gamma^t \mathrm{E}_\pi[r(s_t)]$ 以及$Q_\pi(s, a; r) = \sum_{t=0}^{\infty} \gamma^t \mathrm{E}_\pi[r(s_t)|s_0 = s, a_0 = a]$.

假设 *3*（平均奖励）。令 $P_\pi, P_{\pi'}$ 为具有平稳分布 $\rho_\pi, \rho_{\pi'}$的遍历马尔可夫链，混合率为 $max_s\, DTV(P_\pi^k(s, \cdot)\ \rho_\pi) \le C\beta^k, \forall k \in \mathbb{N}$ 其中常数为 C和 $\beta < 1$。令 $\tilde{E}_\pi r = lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\pi[r(s_t)]$以及 $Q_\pi(s, a; r) = lim_{T\to\infty} \sum_{t=0}^{T-1} E_\pi[r(s_t) - \tilde{E}_\pi r|s_0 = s, a_0 = a]$.

假设 4（有限时域奖励）。令 $H \in \mathbb{N}$ 为时域，并令 $\tilde{\mathrm{E}}_\pi r = \sum_{t=0}^{H-1} \mathrm{E}_\pi[r(s_t)]$ 和 $Q_\pi(s_t,\ a_t;\ r) = \sum_{t'=t}^{H-1} \mathrm{E}_\pi[r(s_{t'})|s_t = s,\ a_t = a]$.

假设 *5*（吸收型马尔可夫决策过程中的总奖励）。设 $P_\pi, P_{\pi'}$ 为马尔可夫链，其吸收状态为 $S_{abs} \subseteq S$，且吸收时间有界于 T_{abs}。令 $\tilde{E}_\pi r = \sum_{t=0}^{\infty} E_\pi[r(s_t)]$ and $Q_\pi(s, a; r) = \sum_{t=0}^{\infty} E_\pi[r(s_t)|s_0 = s, a_0 = a]$。这对应于第*3*节中的设定。

5.2 统一的泰勒级数框架

现在可以陈述我们的主要定理：

定理 2。 令 π， π' 满足假设 *2*、*3*、*4*、*5*中的任意一项。

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

其中 $DQ(0)_{\pi,\pi'}(s) = r_{\pi'}(s)$，且 $DQ(k)_{\pi,\pi'}(s) = E_{a\sim\pi'(s)}[Q_\pi(s, a;\ DQ(k-1)_{\pi,\pi'})] - E_{a\sim\pi(s)}[Q_\pi(s, a;\ DQ(k-1)_{\pi,\pi'})]$。

解读：定理 2 给出了评估策略 $\pi\prime$的 K 阶近似，只需使用由 π 生成的轨迹——只需对这些轨迹上的 $DQ_{(k)\pi,\pi'}(s)$ 项求“和”[4] 即可。该定理还给出了这一近似的精确余项，我们将在下一个结果中将其界定为 $O(\delta^{K+1})$：

推论 1（K 阶余项）。设 $\pi, \pi\prime$ 满足假设 2、3、4、5中的任意一项。进一步假设$\max_s D_{\mathrm{TV}}(P_\pi(s,\cdot)\ P_{\pi'}(s,\cdot)) \le \delta$。那么，

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

其中，有效时域 H_{eff}和缩放常数 $\|\tilde{\mathrm{E}}_{\pi'}\|_1$ 定义于表 1中，而 $r_{\max} = \max_{s\in\mathcal{S}} |r(s)|$。

我们注意到，缩放常数仅仅衡量价值 $\tilde{\mathrm{E}}_\pi r$相对于$r_{\max}$的大小可以达到多少。定理 2 与推论 1 的结合为 DQ 估计量提供了一个优雅、统一的理论。在下一节中，我们将展示定理 1 如何作为该理论的一个具体实例推导出来，并将证明留到第 5.4节。

奖励	$\tilde{E}_\pi r$	H_{eff}	缩放常数 $\|\tilde{\mathrm{E}}_{\pi'}\|_1$
折扣（2）	$\sum_{t=0}^{\infty} \gamma^t \mathrm{E}_\pi[r(s_t)]$	$1/(1-\gamma)$	$1/(1-\gamma)$
平均 (3)	$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{E}_\pi[r(s_t)]$	$(2\log C + 1)/(1-\beta)$	1
有限时域 (4)	$\sum_{t=0}^{H-1} \mathrm{E}_\pi[r(s_t)]$	H	H
吸收式 (5)	$\sum_{t=0}^{\infty} \mathrm{E}_\pi[r(s_t)]$	T_{abs}	T_{abs}

T表 1. 假设2、3、4、5中的奖励设置汇总，以及推论中出现的常数 1.

5.3 从泰勒展开到 DQ

注意，推论 1 给出了一种构造平均处理效应估计量的方法，其偏差为 $O(\delta^2)$：我们只需估计 $J_{\pi_1} = \tilde{\mathrm{E}}_{\pi_1} r_{\pi_1}$的一阶展开，并减去 $J_{\pi_0} = \tilde{\mathrm{E}}_{\pi_0} r_{\pi_0}$的相同近似。偏差可直接通过界定余项得出。为便于说明，我们将在第 3 节所述的总奖励马尔可夫干扰设定中演示这一方法，该设定满足假设 5。

作为热身：考虑定理 2 的零阶展开（即令 $K = 0$ 并省略余项）。根据定义，我们立即得到

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

换句话说：朴素估计量只是对定理 2 的零阶展开进行估计。由推论 1，可以立即得到一般情况下偏差的 $O(\delta)$ 界：$\left|\text{平均处理效应} - \hat{}\text{平均处理效应}_{\text{朴素方法}}\right| \le \delta T^2_{\text{绝对值}r\text{最大值}}$。

为了减少偏差，我们自然会转而估计高阶项。继续采用这一方法，并结合 $K = 1$，通过一些代数运算可以证明，实际上

$$\tilde{\mathrm{E}}_{\pi_1}[\mathrm{DQ}^{(0)}_{\pi_{1/2},\pi_1}(s) + \mathrm{DQ}^{(1)}_{\pi_{1/2},\pi_1}(s)] - \tilde{\mathrm{E}}_{\pi_0}[\mathrm{DQ}^{(0)}_{\pi_{1/2},\pi_0}(s) + \mathrm{DQ}^{(1)}_{\pi_{1/2},\pi_0}(s)] = \mathrm{E}_{\pi_{1/2}}[\mathrm{A\hat{T}E}_{\mathrm{DQ}}]$$

换句话说：DQ 估计量只是估计了定理 2 的一阶展开，因此定理1 的 $O(\delta^2)$偏差可直接由推论 1 得出。因此，我们将 DQ 视为对朴素估计量的一阶校正；其偏差性质，以及在其他场景下构造 DQ 估计量的方法——包括其他奖励形式，[4]进行了恰当加权和归一化。提交至 ACM 的稿件

除均匀随机化以外的实验 $p = 1/2$，以及任意策略下离策略评估的一般问题，都可以直接从这一解释中得出。高阶估计量也可以通过类似的方法推导出来，不过估计这些量需要付出方差方面的代价。

<escape>5.4 定理2的证明2</escape>

定理2 的所有情形均有初等证明，且这些证明遵循统一的框架。为了直观理解，考虑假设2中定义的折扣奖励 $\tilde{E}_\pi r$。根据 P_π的定义，众所周知，

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

其中我们略微滥用记号，将 r 所诱导的向量记为 $r \in \mathbb{R}^{|\mathcal{S}|}$。其他各种设定也存在类似形式，只需将 γP_π 替换为相应的矩阵即可。因此，为了理解 $\tilde{E}_\pi r$ 如何依赖于 π，我们必须弄清逆矩阵 $(I - \gamma P_\pi)^{-1}$ 如何随 π 变化。

S基于这一动机，我们从以下基本的扰动界开始（我们注意到，该界适用于 γP π）

引理 1。 设 $A, A\prime \in \mathbb{R}^{n\times n}$ 为矩阵，使得 $(I - A)^{-1}$ 和 $(I - A\prime)^{-1}$ 存在。那么，

$$(I - A')^{-1} = (I - A)^{-1} + (I - A')^{-1}(A' - A)(I - A)^{-1} \tag{6}$$

证明。注意到 $I - A = I - A\prime + A\prime - A$。右乘 $(I - A)^{-1}$ 并左乘 $(I - A\prime)^{-1}$，即可得到结果。□

由此可立即得到矩阵 $(I - A)^{-1}$ 的级数展开：[5]

引理 2. 设 $A, A\prime \in \mathbb{R}^{n\times n}$ 为矩阵，且 $(I - A)^{-1}$ 和 $(I - A\prime)^{-1}$ 存在。则，

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

证明。只需将引理 1 反复应用于 $(I - A\prime)^{-1}$，直至 K 次，即可得证。 □

在折扣情形下，定理 2 可由引理 2 令 $A = \gamma P_\pi$ 和 $A\prime = \gamma P_{\pi'}$ 后立即推出。其他总奖励情形可通过为 γP_π 选取适当的类似项，沿用相同的证明得到；而平均奖励情形则可视为 $\gamma \to 1$ 时的极限情形。我们在附录中完整给出这些证明。

6 在抖音实施 DQ

在此，我们扩展模型，以涵盖实际应用中更复杂的情形，尤其是抖音/TikTok 这类内容市场。具体而言，现实世界带来了两个尚未讨论的主要挑战：首先，与估计单个观众的处理效应不同，我们关注的是所有观众的总体效应；其次，与针对每个观众、每个视频独立分配处理不同，处理实际上是针对每位创作者仅分配一次，而这项分配会影响所有观众。

首先，我们展示第 3 节中的估计量和理论经过少量修改后，仍适用于这一情形。其次，我们将为这些估计量补充其他技术，以满足实际需求：方差缩减、精确的方差估计，以及快速实现。最终，我们描述的估计量将是一个能够从.....中

[5]要将其理解为“泰勒”展开：设 $A' = A + \delta D$，其中方向为 D。那么，引理 2 立即给出函数 $f(\delta) = (I - A')^{-1}$ 关于 δ、在 $\delta = 0$ 附近的泰勒级数。

框架在干扰下提供了高度准确且低方差的处理效应估计，并满足实际应用的约束。

6.1 内容平台的一个更丰富模型

我们首先详细说明所考虑的精确模型。如前所述，第3 节的总奖励表述对单个观看会话的轨迹进行建模，而在平台层面，我们关注的是所有观众的总体效应。还需注意，每个观看会话都可能完全异质。换言之，每个观看会话$i \in [N]$都可能具有完全不同的奖励函数 $r(i)$、转移核 $P(i)$以及总奖励 $J(i)\pi$。为对此建模，我们假设 r, P, J是从某个关于马尔可夫决策过程的未知分布 M 中独立同分布抽取的；对于每次抽样，我们观测到一条完整轨迹 $\{s_{it}, a_{it}, r_{it}\}_{t\geq 0}$。因此，给定 N条这样的轨迹（即观看会话），我们要估计 M 下的期望平均处理效应：

$$\overline{\mathrm{ATE}} = \mathsf{E}_{\mathcal{M}}[J_{\pi_1} - J_{\pi_0}] \tag{7}$$

其次，处理分配并非在观众和时间步之间独立进行，而是针对每位创作者仅分配一次；每当任何观众遇到该创作者时，都会接受相同的处理。更正式地说，存在一个由 $j \in [M]$索引的创作者总体，其中每位创作者都有一个随机处理分配 $a(j) \sim \pi_p$。观众 i在时间步 t看到视频 j_{it}，而该时间步播放的动作则为$a_{it} = a(j_{it})$。

最后，在现实世界中，数据收集通常要复杂得多——可能存在多个处理组，而且随机化概率 p 通常远小于 1/2。将问题建模为离策略评估，可以让我们自然地将框架扩展到这类情况——数据收集策略基本上可以是任意的。相关讨论将在第 6.5 节中展开。

6.2 抖音的蒙特卡洛估计

具体来说，我们先讨论一种简单的蒙特卡洛估计量，它可应用于这个更一般的问题。回顾第4 节，式（4）中的估计量针对的是单条轨迹。这启发我们通过汇总多条轨迹上的 DQ 估计量，直接扩展该方法来估计元处理效应，即式（7）。令 $\hat{\mathrm{ATE}}^{(i)}_{\mathrm{DQ}} = \sum_{t=0}^{\infty}\left[\hat{Q}_{\mathrm{MC}}(,1) - \hat{Q}_{\mathrm{MC}}(,0)\right]$ 表示轨迹 i 的蒙特卡洛 DQ 估计量，其中 $\hat{Q}_{\mathrm{MC}}(,) \cdot \{a_{it} = a\} \sum_{t=0}^{\infty} r_{it}$是蒙特卡洛 Q函数估计。于是，我们得到如下汇总估计量：

$$\hat{\mathrm{ATE}}_{\mathrm{DQ}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathrm{ATE}}^{(i)}_{\mathrm{DQ}}. \tag{8}$$

根据期望的线性性，该估计量的偏差上界为定理 1中偏差上界的期望，其中期望是对从以下分布中抽取的马尔可夫决策过程取的：$\mathcal{M}$。

关于创作者层面的处理分配问题：只要处理不会改变某个视频展示给用户的概率（即推荐系统），那么从单个观众的角度来看，边际处理概率 $\Pr(a_{it} = a)$ 就保持不变，因此我们各种估计量的偏差特性也保持不变。然而，观众之间处理分配的相关性会给方差量化带来挑战，我们将在第 6.4 节中讨论这一问题。[6]

[6]事实上，单个观众可能会在同一会话中多次观看同一视频，从而导致不同时刻的动作概率之间存在依赖关系。在附录中，我们给出了估计量的一个通用版本，可以正确处理这种依赖关系。提交至 ACM 的稿件

6.3 双重稳健 DQ 估计量

蒙特卡洛估计量方程（8）基本上不需要任何假设，而且实现起来非常简单——它只需要奖励观测，并且只需对数据集进行一次遍历即可计算得出。然而，正如我们在实验中所展示的，这种灵活性是以方差为代价的。

受双重稳健估计量在实证中的成功启发 [13], ，我们提出一种双重稳健 DQ 估计量。假设我们有一个对 Q 函数$\hat{}$ Q_{Reg} 的近似（通常通过对实验前数据或某个留出观众集进行回归获得，正如第7节所做的那样），该近似可能存在设定错误或偏差。然后，我们将估计量定义为

$$\hat{\text{ATE}}_{\text{DQ-DR}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}\left[\hat{Q}^{\text{DR}}_{\pi_{1/2}}(s_{it},1) - \hat{Q}^{\text{DR}}_{\pi_{1/2}}(s_{it},0)\right] \tag{9}$$

其中 现在我们引入一个近似 Q函数，并改用双重稳健 Q 函数估计量：

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

其中 $\pi(s_{it},\ a)$ 是策略 π 在状态 s_{it}. 下选择动作 a 的概率。该估计量具有三个显著性质。首先，无论 $\hat{}$ $Q_{\text{Reg}}{}^{\pi_{1/2}}$ 对 $Q_{\pi_{1/2}}$的估计有多差，$\hat{}$ $Q_{\text{DR}}{}^{\pi_{1/2}}$ 仍然是 $Q_{\pi_{1/2}}$ 的无偏估计量，只要随机化概率 $\pi(s,\ a)$ 是已知的。这一点至关重要，因为在部分可观测状态的情况下，$\hat{}$ $Q_{\text{Reg}}{}^{\pi_{1/2}}$可以利用状态中可观测的部分通过启发式方法拟合，而且还可以进行函数逼近或正则化。其次，引入 $\hat{}$ $Q_{\text{Reg}}{}^{\pi_{1/2}}$ 可作为控制变量，通常能降低估计量的整体方差；在实验中，我们表明，与蒙特卡洛 DQ 估计相比，这种方法可将方差降低 95%。

最后，我们来讨论一个更微妙的问题：对随机化概率的稳健性 $\pi(s,\ a)$。具体而言，在许多实际情形中，我们只知道随机化概率的近似值 $\hat{}\pi(s,\ a)$。这种情况最常见于观察性场景，此时 $\pi(s,\ a)$ 通常事先未知，必须根据数据进行估计；但在真实世界实验场景中也很常见，因为实施过程的复杂性往往会导致程序错误，使实际随机化概率发生偏差。在 [21] 中，这种差异被称为“样本比例不匹配”（该文还提供了大量可能出现这种情况的示例）。我们将看到，蒙特卡洛估计量公式（8）只有在 $\hat{}\pi = \pi$ 时才是无偏的；如果不满足这一条件，实际中的偏差可能会非常大。

双重稳健估计量通过大幅降低估计量对 $\pi(s,\ a)$ 的依赖来解决这一问题。为便于理解，只考虑上文介绍的估计量 $\hat{}$ $Q_{\text{DR}}{}^{\pi_{1/2}}$，但将 $\pi(s,\ a)$ 替换为估计值 $\hat{}\pi$。于是，期望 $\text{E}^{\pi_{1/2}}$ $\hat{}$$Q_{\text{DR}}{}^{\pi_{1/2}}$ 为

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

我们随后可以立即看出，$\hat{}$ $Q_{\text{Reg}}{}^{\pi_{1/2}}$ 估计量 $Q_{\pi_{1/2}}$越好，$\text{E}^{\pi_{1/2}}$ $\hat{}$$Q_{\text{DR}}{}^{\pi_{1/2}}$ 对$\hat{}\pi$的敏感性就越低。在一个极端情况下，如果 $Q_{\pi_{1/2}} = \hat{}Q_{\text{Reg}}{}^{\pi_{1/2}}$，那么无论$\hat{}\pi$如何，双重稳健估计都将完全无偏；这（再加上$\hat{}$ $Q_{\text{DR}}{}^{\pi_{1/2}}$ 在$\hat{}\pi = \pi$时同样无偏这一事实）正是此类估计量众所周知的“双重”稳健性质。

在我们的实验第7节中，我们展示了，引入这种双重稳健技术后，无论 π 是已知确切值还是仅在扰动范围内已知，都能得到精确得多的估计量。

6.4 随机化检验下的方差刻画

$\hat{}$ 理解 ATE_{DQ} 方差的一个关键挑战在于：动作是在主播层面分配的；主播由多个观众共享；因此 a_{it} 可能在观众之间存在任意相关性。

虽然在这种情况下难以精确刻画方差，但我们提出使用重新随机化检验来检验尖锐零假设 H_0，即处理对奖励分布或问题动态的影响恒为零。刻画检验统计量 $\hat{}\text{ATE}_{\text{DQ}}$ 在 H_0 下分布的一种典型方法是进行模拟：随机重新抽取分配 $a(j)$（即主播 j的分配），然后在新的处理分配下重新计算检验统计量（同时保持结果不变，这正是零假设下应采取的做法）。这种“重新随机化”方法能得到检验统计量的精确分布，但在抖音的规模下通常难以实施：抖音拥有数亿观众，对一次处理分配$a(j)$进行 SQL 查询，在抖音集群上大约需要运行 12 小时。

事实证明，在零假设下，$\hat{}\text{ATE}_{\text{DQ}}$ 的方差实际上可以在线性时间内计算出来。蒙特卡洛估计量公式（9）实际上可以写成主播级项的总和，最终得出置换检验方差的精确表征 H_0。为此，请注意，估计量公式（9）等价于

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

其中，t_{ij} 是观众 i 观看主播 j 的时间段。由于每个分配 $a(j)$彼此独立，因此这些加数也各自独立。由此，我们只需刻画每个加数的方差并将它们相加，即可计算平均处理效应$_{\text{DQ}}$的方差。最后需要指出，每个加数都是伯努利随机变量的仿射函数，其方差有闭式表达式：

$$\text{Var}(\hat{\text{ATE}}_{\text{DQ}}) = \sum_{j=1}^{M} \Pr(a(j)=0)\Pr(a(j)=1) \left(\frac{1}{\Pr(a(j)=1)} - \frac{1}{\Pr(a(j)=0)} \right)^2 \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=t_{ij}}^{\infty} r_{it} \right)^2$$

利用这个方差，我们便可以进行假设检验：可以假设正态性（在实践中是合理的），也可以使用切比雪夫型界。在实验中，我们表明，这项检验在实践中具有很高的统计功效，能够检测出一天内效应量为0.2%的变化，并且实现了接近名义覆盖率的水平。

虽然这说明了为什么线性时间的置换检验应当是可行的，但对于 $p \neq \frac{1}{2}$，类似的结果要复杂得多。在这些情况下，朴素地采用相同的方法，复杂度为 $O(M^4)$；而更巧妙的精确方法需要 $O(M^2)$，只有在 $O(M)$时间内才能得到近似结果（尽管从经验上看，这种近似非常准确）。我们在附录中详细介绍了这一实现，并在第 7 节展示了该方法在经验上几乎能够实现精确覆盖。

6.5 一般数据收集策略

最后，我们举例说明如何将 DQ 估计推广到各种数据收集场景。为便于阐述，我们讨论了在特定情境下进行 DQ 估计，即通过 A/B 测试估计处理效应，处理概率为 $p = 1/2$；这种设定能得到最简洁的结果。然而，即使修改 p，推广后的结果也并不直观。一般而言，我们依据某种数据收集策略（也常称为“行为”策略）$\pi_{\text{数据}}$来收集数据，并希望评估某种新的目标策略$\pi_{\text{新的}}$（或者在某些情况下，一组目标策略）。以下是一些常见情形：

- 使用 $p \neq 1/2$ 进行 A/B 测试。令 $\pi_{\text{数据}}$ = 服从伯努利分布(p)，并将 π_0, π_1 分别视为目标策略。
- 包含多个处理组的随机对照试验 $1 \ldots K$。令 $\pi_{\text{数据}}(s)$ 为 $1 \ldots K$ 上的一个分布，且我们有目标策略 $\pi_1 \ldots \pi_K$。

• 根据日志数据评估新策略。令 $\pi_{数据}$ 为日志策略，令$\pi_{新}$ 为任意新策略（在每个状态下与$\pi_{数据}$对 A 具有相同的支持集）。

此外，我们还处理动作可能随时间相互依赖的问题。也就是说，定义历史 $\mathcal{H}_t = \{s_{t'}, a_{t'}, r_{t'}\}_{t'<t}$我们现在允许 $\pi_{数据}$ π_{new}任意依赖于此历史。我们将这种依赖记为 $\pi(s, a|\mathcal{H}_t)$。定理 2中的泰勒展开为这些推广分别提供了一个 DQ 类比。准确地说：理想化的一阶展开变为：

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

值得注意的是，这一由此得到的蒙特卡洛估计量计算起来更加复杂，涉及多个重要性抽样权重，这也为此前针对 DQ 的各项扩展带来了进一步的挑战（和机遇）。我们在抖音的实现克服了这些挑战，具体细节将在附录中说明。

7 实验

现在，我们将通过受特定应用场景（抖音）启发的模拟实验，提供实证结果，以比较 DQ 估计量及其各种改进方法。我们的实验结果旨在回答以下研究问题：

• **研究问题1**。 与现有替代方法相比，DQ 是否能提供更准确的处理效应估计?
• **研究问题 2**。 我们的假设检验方法是否具有正确的覆盖率和足够的统计功效，能够在现实时间尺度下检测出较小的处理效应?
• **研究问题3**。 在对 π进行扰动的情况下，这些结果的稳健性如何?

7.1 实验设置

抖音中一个典型的干扰来源存在于这样的实验中：由于实施限制，抖音必须在创作者（或主播）层面进行随机化，而不是在观众层面进行随机化。我们以引言中讨论的直播场景为具体示例：抖音推出了一项新功能，允许创作者以高清画质进行直播，但由于这是一项面向创作者的功能，针对该功能的实验必须在创作者层面开展。在这种场景下，高清直播几乎必然会增加单个视频的观众观看时长，但由于观众注意力蚕食，朴素估计很可能会高估其对总观看时长的影响。此外，平均处理效应有可能为负，因为该功能往往会更快地消耗观众的电量和流量额度。

虽然我们已在抖音的内部模拟器以及真实世界实验场景中测试过 DQ 估计量，但此处主要展示基于抖音用户数据校准的模拟结果。这种方法在保护隐私的同时，确保了定性结果的一致性。我们将状态表示为 $s_{it} = (w_t,\ u_i,\ v_{it})$，其中 w_t 表示观众迄今观看视频所花费的时间，$u_i \in \mathbb{R}^d$（此处我们选择 $d = 5$）是表示观众偏好的潜在向量，而 $v_{it} \in \mathbb{R}^d$ 是表示在时间 t向观众 i展示的视频的潜在向量，因此 $u_i,\ v_{it}$衡量观众 i对展示的第 t个 视频的亲和度。我们假设智能体看不到 $u,\ v$，尽管实际上智能体可能能够获取这些向量的某种含噪版本。

观众在对照组下的反事实观看时长为 $w_{it} \sim \text{Exp}(1/(ku_i^\top v_{it}))$，其中 k 是某个标量问题参数。最后，观众的实际观看时长为 $r_{it} = w_{it}(1 + \tau^* a_{it})$；也就是说，在处理 $a_{it} = 1$ 下，观看时长会增加 τ^* 倍。观众的状态要么是其累计观看时长 $s_t = \sum_{t'=1}^{t-1} r_{it'}$，要么是终止状态（记作 s_{abs}），表示观众已离开平台。在每个时期，the

观众转移到终止状态s_{abs}（即“离开平台”）的概率由 s_t as$\mathbb{P}(s_{t+1} = s_{\text{abs}}|s_t) = \exp(s_t)\alpha+\exp(s_t)$决定，其中 α是一个问题参数。我们校准子参数 k, α，使其反映抖音观众群体中观看时长和观看视频数量的实际分布；此处省略具体数值。

7.2 算法

我们将比较蒙特卡洛方法和双重稳健 DQ 两种变体的性能。对于双重稳健 DQ，我们为 $\hat{}\, Q_{\text{Reg}}{}^{\pi_{1/2}}$ 构建一个简单的回归估计量：我们留出在实验策略下观测到的前 1000 名观众，并通过最小二乘法拟合一个关于参数 β_0, β 的如下形式的回归模型：$\hat{}\, Q_{\text{Reg}}{}^{\pi_{1/2}}((w, u, v)\ a) = \beta_0 + \beta w$。

如前所述，实践中唯一采用的替代方法是朴素估计。虽然实践中并未使用，但可以通过对整条轨迹进行重要性采样来实现无偏离线策略评估，不过不难预料，这种方法的方差会大得惊人。请注意，这些方法都适用于部分可观测状态，并且与我们场景中所需的主播级随机化兼容。具体而言，典型的朴素估计量为 $\hat{}\text{ATE}_{\text{朴素方法}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{\infty}[\hat{}\, r_{\text{IS}}(s_{it},\ 1) - \hat{}\, r_{\text{IS}}(s_{it},\ 0)]$ 其中 $\hat{}\, r_{\text{IS}}$ 是重要性采样估计量 $\hat{}\, r_{\text{IS}}(s_{it},\ a) = \mathbb{I}_{\{a_{it}=a\}}{}^{\pi(s_{it},\ a)}\ r_{it}$ 而典型的逐步重要性采样离线策略评估估计量为

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

其中，$\hat{}\, r\text{IS-OPE}(s_{it}, a_{it})$ 是一个重要性采样估计量，用于估计策略 π_a 下第 t 个时期的期望奖励 $\text{E}_{\pi_a}[r(s_{it}, a_{it})]$。这里我们实现了“逐步骤”重要性采样估计量 [14],，它比朴素的轨迹级采样略有改进：$\hat{}\, r\text{IS-OPE}(s_{it}, a) = \left(\prod_{t'=0}^{t} \mathbb{I}_{\{a_{it'}=a\}}{}^{\pi(s_{it'},a)}\right) r_{it}$。

用于构造式（9）的双重稳健方法也可用于改进朴素方法和离线策略评估估计量，从而提供更精确的基线。我们实现了文献 [17] 中最先进的方法。不过值得注意的是，双重稳健变体会保留相同的偏差特性；因此，预计朴素双重稳健估计也会表现出较大的偏差，而 OPE-DR 则仍然无偏，但方差仍高得不合理。我们也使用这些估计量进行了实验，但将其推导留到附录中。

7.3 总体性能比较（研究问题1）

DQ 的平均处理效应估计比任何其他方法都准确得多。我们首先考察图 2 中各估计量的均方根误差，以及表 2 中偏差和方差的细分情况。首先要指出，干扰会导致朴素估计量产生严重偏差：朴素估计量的符号始终错误，这体现了一种直觉，即干预在短期内总是有利，但从长期来看可能产生负面影响。不仅符号错误，误差幅度也很大：在所有情况下，朴素估计值都约为实际平均处理效应的 -800%。接下来，我们来看离线策略评估的不同版本。离线策略评估确实是无偏的，但其方差极大，因此在所有考察的时间尺度上，其均方根误差都远高于其他任何估计量。双重稳健版和标准版估计量均是如此。朴素方法和离线策略评估算法的相对均方根误差都未低于 100%。

最后来看 DQ，我们发现，当效应量足够大且观测数量足够多时，蒙特卡洛 DQ 的均方根误差远低于任何朴素方法或离线策略评估变体 N。然而，即使在 10^8 位观众的情况下，其方差仍然很大，这主要是由于观测到的主播数量有限；此外，对于所测量的效应量，它也未能在提交至 ACM 的稿件中实现低于 100% 的相对均方根误差

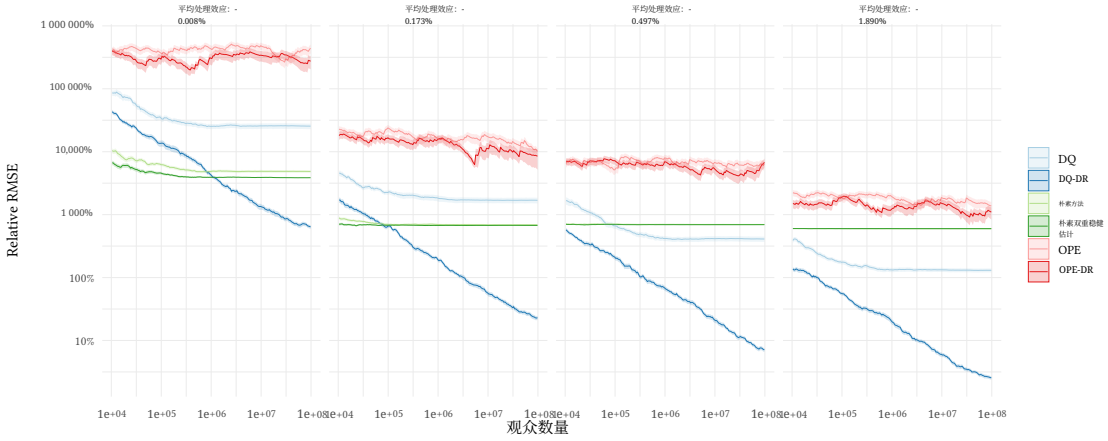

图2. 各估计量相对于 N的相对均方根误差，即观测到的观众数量，针对不同效应量。相对均方根误差的计算方式为 $\sqrt{\sum_{i=1}^{T}(\hat{}\text{ATE}_i - \text{ATE})^2}/(\sqrt{T}\text{ATE})$ 其中平均处理效应是真实处理效应，而$\hat{}\text{ATE}_i$是该估计量在第 i次实验中经过 $T = 100$ 个随机种子后得出的结果。误差条表示经过 T个随机种子后的标准误。所有估计量的偏差与方差分解见表2。

实际平均处理效应（%）	-0.01		-0.17		-0.50		-1.89	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
DQ	-0.05 (0.00)	2.01	-0.15 (0.00)	2.95	-0.49 (0.00)	2.03	-1.92 (0.00)	2.44
DQ-DR	-0.05 (0.00)	0.03	-0.15 (0.00)	0.03	-0.49 (0.00)	0.03	-1.92 (0.00)	0.04
朴素方法	0.30 (0.00)	0.17	0.99 (0.00)	0.25	2.95 (0.00)	0.17	9.43 (0.00)	0.22
朴素双重稳健估计	0.30 (0.00)	0.01	0.99 (0.00)	0.01	2.95 (0.00)	0.01	9.43 (0.00)	0.01
OPE	-0.01 (0.00)	35.42	-0.17 (0.00)	17.24	-0.50 (0.00)	32.01	-1.89 (0.00)	26.24
OPE-DR	-0.01 (0.00)	22.05	-0.17 (0.00)	14.82	-0.50 (0.00)	33.32	-1.89 (0.00)	20.87

表2。在观察到 $N = 10^8$ 观众时，各估计量的偏差和方差占基线结果 J_{π_0}的百分比，适用于具有不同效果大小（即 -0.01%、-0.17%、-0.50%和 -1.89%）的问题实例。括号中的数值为均值估计的标准误。双重稳健 DQ 几乎实现了零偏差，并将 DQ 的标准差降低了约97%。

另一方面，双重稳健 DQ 带来了显著的性能提升：在所有情况下，相较于蒙特卡洛 DQ，误差至少降低了 97%；相较于朴素估计，误差最高降低了 99%。当$N = 10^8$ （约为一天的观众活动量）时，对于所有幅度大于 0.17% 的处理效应，DQ-DR 已经能够实现低于处理效应 100% 的相对均方根误差。

7.4 假设检验的统计功效与覆盖率（研究问题 2）

使用一天的数据，重新随机化检验即可检测出效应量为 *0.2%* 的效果。现在，我们来考虑在 DQ 估计下进行假设检验的问题。图 3 展示了第 6.4 节所述重新随机化检验的效果；该方法可以直接推广到双重稳健 DQ。在分析样本量为 10_8 名观众的数据时——这与抖音约一天的流量规模相当——双重稳健 DQ 估计量展现出稳健性：即使要检测的效应量低至 0.15%，也能始终达到 80% 的统计功效。此外，当真实处理效应为零时，该检验也成功将假阳性率控制在目标值 10%。

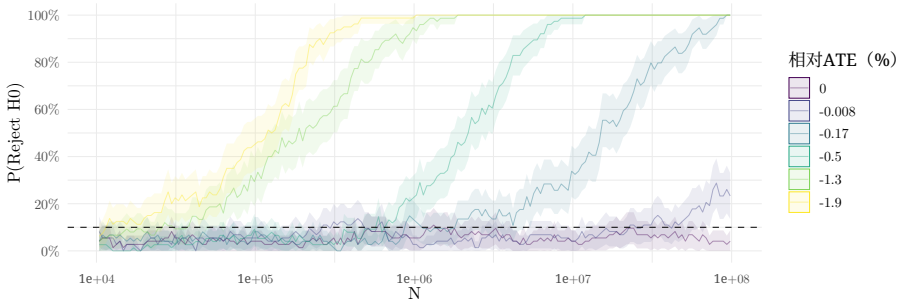

图3。对于所提出的近似置换检验和双重稳健估计量，在不同效应量和90%置信水平下，$\mathbb{P}$（拒绝 H_0）（即统计功效）与采样观众数量的关系。误差条表示基于100个随机种子计算的标准误。在 10^8 观众处，即抖音约一天的流量下，双重稳健 DQ 能够可靠地达到80%的统计功效，即使效应量低至0.15%。

7.5 对模型设定错误的稳健性（RQ3）

双重稳健。最后，我们来考察处理分配概率设定错误时各个估计量的表现；也就是说，算法并不知道环境实际执行的实验使用的概率是 $p = 1/2$，而非名义概率 $p = 1/2 + 1/1000$。对抖音而言，这种设定错误下的稳健性是首要考量。例如，干预可能导致极少数主播的表现下降，使他们不再出现在观众的信息流中。 [21] 描述了此类扰动发生的其他若干情形。

表 3 展示了该设定下所有估计量的结果。正如预期，我们看到偏差和方差问题仍然困扰着朴素方法和离线策略评估。然而，蒙特卡洛 DQ 现在也出现了严重偏差：它估计出的处理效应符号错误，且绝对值远大于实际处理效应，尽管处理概率仅发生了微小变化。相比之下，双重稳健 DQ 的偏差几乎没有变化，仍至多为平均处理效应的百分之几。尽管我们用于 $\hat{}\, Q_{\mathrm{Reg}}{}^{\pi_{1/2}}$ 的回归估计量极为粗糙，设定也极不准确，结果依然如此；这说明构造双重稳健估计量相当容易，而且这样做能产生巨大影响。

实际平均处理效应（%）	-0.01		-0.17		-0.50		-1.89	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
DQ	4.98（0.24）	2.20	4.49（0.28）	2.56	4.43（0.28）	2.52	2.70（0.26）	2.37
DQ-DR	-0.06（0.00）	0.04	-0.12 (0.00)	0.03	-0.48 (0.00)	0.04	-1.90 (0.00)	0.04
朴素方法	0.71 (0.02)	0.18	1.38 (0.02)	0.22	3.35 (0.02)	0.21	9.82 (0.02)	0.21
朴素双重稳健估计	0.30（0.00）	0.01	0.99（0.00）	0.00	2.95 (0.00)	0.01	9.43 (0.00)	0.01
OPE	1.22 (3.11)	28.80	5.94 (3.07)	28.13	-0.40 (3.29)	30.14	3.49 (2.59)	23.64
OPE-DR	-1.24 (3.32)	26.81	4.57 (3.37)	26.52	-1.69 (1.87)	14.87	-8.18 (2.31)	18.73

表3。所有估计量在不同效应大小（−0.01%、 −0.17%、 −0.50%、 −1.89%）下的估计结果，其中估计量使用 $p = 1/2$，但实际执行的策略为 $p = 1/2 + 1/1000$。朴素方法和离线策略评估估计量仍分别表现出较高的偏差和较高的方差。蒙特卡洛 DQ 现在存在显著偏差，而双重稳健 DQ 基本未受影响。

参考文献

[1] Patrick Bajari、Brian Burdick、Guido W Imbens、Lorenzo Masoero、James McQueen、Thomas Richardson 和 Ido M Rosen。多重随机化设计。arXiv 预印本 arXiv:2112.13495，2021年12月。[2] Thomas Blake 和 Dominic Coey。为什么市场实验比看起来更难：测试组-对照组干扰的作用。载于第十五届 ACM 经济学与计算会议论文集，EC '14，第 567–582 页，美国纽约州纽约市，2014年6月。美国计算机协会。[3] Iavor Bojinov 和 Neil Shephard。时间序列实验与因果估计量：精确随机化检验与交易。《美国统计学会杂志》，114(528):1665–1682，2019年。[4] Iavor Bojinov、David Simchi-Levi 和 Jinglong Zhao。切换实验的设计与分析。《管理科学》，2022年。[5] Nicholas Chamandy。网约车市场中的实验。https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e，2016年9月。[6] Angus Deaton 和 Nancy Cartwright。理解和误解随机对照试验。《社会科学与医学》，210:2–21，2018年。[7] Vivek Farias、Andrew Li、Tianyi Peng 和 Andrew Zheng。实验中的马尔可夫干扰。《神经信息处理系统进展》，35:535–549，2022年。[8] Alexandre Gilotte、Clément Calauzènes、Thomas Nedelec、Alexandre Abraham 和 Simon Dollé。推荐系统的离线 A/B 测试。载于第十一届 ACM 国际网络搜索与数据挖掘会议论文集，第 198–206 页，2018年。[9] Peter W Glynn、Ramesh Johari 和 Mohammad Rasouli。具有时间干扰的自适应实验设计：最大似然方法。《神经信息处理系统进展》，33:15054–15064，2020年。[10] Henning Hohnhold、Deirdre O'Brien 和 Diane Tang。着眼长远：这对用户和业务更有利。2015年。[11] David Holtz 和 Sinan Aral。在线市场实验中限制来自测试组-对照组干扰的偏差。SSRN 学术论文 3583596，社会科学研究网络，美国纽约州罗切斯特，2020年4月。[12] David Michael Holtz。在线市场实验中限制来自测试组-对照组干扰的偏差。博士学位论文，麻省理工学院，2018年。[13] Nan Jiang 和 Lihong Li。强化学习的双重稳健离策略价值评估。载于国际机器学习会议，第 652–661 页。PMLR，2016年。[14] Nan Jiang 和 Lihong Li。强化学习的双重稳健离策略价值评估。载于国际机器学习会议，第 652–661 页。PMLR，2016年。[15] Ramesh Johari、Pete Koomen、Leonid Pekelis 和 David Walsh。窥探 A/B 测试：为什么这很重要，以及如何应对。载于第二十三届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集，第 1517–1525 页，2017年。[16] Ramesh Johari、Hannah Li、Inessa Liskovich 和 Gabriel Y Weintraub。双边平台中的实验设计：偏差分析。《管理科学》，2022年。[17] Nathan Kallus 和 Masatoshi Uehara。用于马尔可夫决策过程高效离策略评估的双重强化学习。《机器学习研究杂志》，21(167):1–63，2020年。[18] Nathan Kallus 和 Masatoshi Uehara。利用双重强化学习高效打破离策略评估中的时域诅咒。《运筹学》，2022年。[19] Rochelle King、Elizabeth F Churchill 和 Caitlin Tan。数据驱动设计：通过 A/B 测试改善用户体验。

流行病学，19(1):187–196，1990年。[31] Richard S Sutton、Csaba Szepesvári 和 Hamid Reza Maei。采用线性函数近似进行离策略时序差分学习的一种收敛 O(n) 算法。神经信息处理系统进展，21(21):1609–1616，2008年。[32] M Talbot、AD Milner、MAE Nutkins 和 JRLaw。作物品种试验中小区间干扰对产量表现的影响。《农业科学杂志》，124(3):335–342，1995年。[33] Eric J Tchetgen Tchetgen 和 Tyler J VanderWeele。存在干扰情况下的因果推断。《医学研究统计方法》，21(1):55–75，2012年。[34] Philip Thomas 和 Emma Brunskill。用于强化学习的数据高效离策略策略评估。载于国际机器学习会议，第2139–2148页。PMLR，2016年。[35] Philip Thomas、Georgios Theocharous 和 Mohammad Ghavamzadeh。高置信度离策略评估。载于*AAAI*人工智能会议论文集，第29卷，2015年。[36] Vivek Trehan。市场实验。https://www.youtube.com/watch?v=IR000RqN7pw，2019年。访问日期：2023年3月24日。[37] Jon Vaver 和 Jim Koehler。利用地理实验衡量广告效果。2011年。[38] Dylan Walker 和 Lev Muchnik。网络中随机实验的设计。《*IEEE*会刊》，102(12):1940–1951，2014年。[39] Corwin M Zigler 和 Georgia Papadogeorgou。存在干扰情况下的二分因果推断。《统计科学：数理统计学会评论期刊》，36(1):109，2021年。