Estimation of Treatment Effects under Nonstationarity via the Truncated Policy Gradient Estimator

Ramesh Johari
Stanford University

Tianyi Peng
Columbia University

Wenqian Xing
Stanford University

Abstract

Randomized experiments (A/B tests) are widely used to evaluate interventions in dynamic systems such as recommendation platforms, marketplaces, and digital health. In these settings, interventions affect both current and future system states, so estimating the global average treatment effect (GATE) requires accounting for temporal dynamics. Existing estimators—including difference-in-means (DM), off-policy evaluation methods, and difference-in-Q's (DQ)—perform poorly in nonstationary environments due to high bias and variance. We address this challenge with the truncated policy gradient (TPG) estimator, which replaces instantaneous outcomes with truncated outcome trajectories. Theoretically, it corresponds to a truncated policy gradient that approximates the GATE to first order, yielding provable bias and variance improvements in nonstationary Markovian settings. We validate our theory through a ride-sharing simulation calibrated to New York City taxi data. The results show that a well-calibrated TPG estimator achieves low bias and variance in practical nonstationary settings.

1 Introduction

Randomized controlled experiments ("A/B tests") are essential tools for evaluating the impact of interventions across dynamic, technology-driven environments such as recommendation systems, online marketplaces, and digital health platforms. Typically, these interventions do not merely influence immediate outcomes but also propagate effects over subsequent system states. This leads to a breakdown of the standard assumption of unit-level independence, a phenomenon known as *temporal interference* [8]. For instance, in a ride-sharing platform, user assignments to treatment and control affect the availability of drivers: a booking decision by one user influences the system state—e.g., driver distribution or availability—which in turn affects the experience and behavior of subsequent users. In such settings, applying a naive estimator (e.g., the difference-in-means or Horvitz-Thompson estimators [10]) can result in substantial bias, as it fails to account for temporal dependencies and state dynamics [4]. One way to model this dynamic effect is through a Markovian system that evolves over time. At each step, the state represents the current environment (e.g., the number and location of available drivers), the action indicates whether an arriving user is assigned to control or treatment, and the outcome captures the user's response (e.g., booking a trip). The system dynamics depend on both the state and the action, and crucially, *the underlying transition*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

rules may change over time. This nonstationarity—driven by shifting user preferences, evolving supply dynamics, or external shocks like seasonal trends or viral events—forms the central challenge we address in this work.

Our main contribution is the construction of a simple low-bias, low-variance estimator for the global treatment effect in such settings. We build on the recently introduced *difference-in-Q's (DQ) estimator*, which addresses interference in Markovian settings by leveraging Q-values [6, 7]. However, in nonstationary environments, estimating Q-functions is generally infeasible, and the bias of DQ grows with the horizon. We show that replacing long-horizon Q-values with truncated k-step accumulated rewards yields a practical, low-bias, low-variance alternative, which we call the *truncated policy gradient (TPG) estimator*. This modification transforms the theoretical appeal of DQ into a broadly usable tool. Our contributions are threefold. (1) *Theoretical:* we prove that the TPG estimator achieves provably low bias and variance under mild mixing assumptions, via a novel truncated policy gradient analysis. (2) *Practical:* the estimator is simple, requiring only observed rewards from a single trajectory, and permits flexible post-experiment choices of the truncation level. (3) *Empirical:* in large-scale ride-sharing simulations, it consistently outperforms standard OPE and switchback methods, yielding lower bias and variance in nonstationary settings.

Related work. Experimental design has been widely studied in *networked systems* [5, 16, 2, 24] and *online marketplaces* [13, 26, 14, 3, 25]. To capture temporal dependence, recent work models experiments in *Markovian environments* [8, 11, 6, 7, 15], where the difference-in-Q's (DQ) estimator offers a favorable bias–variance tradeoff. Extensions to *nonstationary settings* include time series models [18, 27] and Markovian environments with mixing properties [11], which are closest to our setup. Meanwhile, off-policy evaluation (OPE) has been studied in fully and partially observed MDPs [12, 23, 22, 17, 21, 19], but most approaches assume multiple independent trajectories and struggle with a single nonstationary path due to the curse of horizon [23].

2 Preliminaries

We consider a finite-horizon setting with $t = 1, \dots, T$, a finite state space $\mathcal{X}$, and binary actions $\mathcal{Z} = \{0, 1\}$ (i.e., treatment and control). For $x, x' \in \mathcal{X}$ and $z \in \mathcal{Z}$, let $P_z^t(x, x') = \mathbb{P}(X_{t+1} = x' \mid X_t = x, Z_t = z)$ denote the transition kernel at time t. The initial state is distributed as $X_1 \sim \rho$, and thereafter the dynamics satisfy the Markov property. At each t, a bounded reward $Y_t \in [-M, M]$ is generated with mean $r(x, z)$ when action z is taken in state x, independent of the past given (X_t, Z_t). Finally, we impose a *mixing-time* assumption on the transition kernels, following [6] and [11] in the context of switchback experiments.

Assumption 1 (Mixing time)**.** *There exists $\gamma \in (0, 1)$ such that for any time $t \in [T]$, action $z \in \mathcal{Z}$, and distributions f, f' over the state space $\mathcal{X}$, we have $\|f' P_z^t - f P_z^t\|_{\mathrm{TV}} \leqslant \gamma \|f' - f\|_{\mathrm{TV}}$, where $(f P_z^t)(x') := \sum_{x \in \mathcal{X}} f(x) P_z^t(x, x')$ and $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance.*

We consider a family of Markov policies $\{\pi_\theta : \theta \in [0, 1]\}$. Under π_θ, action $z = 1$ is chosen with probability θ (and $z = 0$ with probability $1 - \theta$), independently of state and past randomness. The extremes π_1 and π_0 correspond to the *treatment* and *control* policies, respectively, while policies with $0 < \theta < 1$ assign actions i.i.d. as Bernoulli(θ), which we refer to as *Bernoulli randomization*.

For such a policy, the transition kernel at time t is $P_\theta^t = \theta P_1^t + (1 - \theta) P_0^t$, and the expected reward in state x is $r_\theta(x) = \theta r(x, 1) + (1 - \theta) r(x, 0)$. We denote by $\mathcal{L}_\theta^t$ the law of Y_t under π_θ, i.e.,

$$\mathcal{L}_\theta^t = \mathrm{Law}(Y_t \mid X_1 \sim \rho,\ Z_u \sim \text{i.i.d. Bernoulli}(\theta),\ 1 \leqslant u \leqslant t)\,. \tag{1}$$

Let $X_1 \sim \rho$ be the initial state. The treatment effect at time t is $\tau_t := \mathbb{E}_{\mathcal{L}_1^t}[Y_t] - \mathbb{E}_{\mathcal{L}_0^t}[Y_t]$, and the global average treatment effect (GATE) over horizon $[T]$ is $\tau := T^{-1}\sum_{t=1}^{T}\tau_t$.

3 Estimation via truncated policy gradient

The TPG estimator is based on a simple idea: instead of using only the immediate outcome, we aggregate outcomes over a window of length k. Formally, for each $0 \leqslant k < T$, we define:

$$\hat{\tau}_k := \frac{1}{T}\sum_{u=1}^{T}\left(\frac{1\{Z_u = 1\}}{1/2} - \frac{1\{Z_u = 0\}}{1/2}\right)\sum_{t=u}^{\min(u+k,T)} Y_t. \tag{2}$$

When $k = 0$, $\hat{\tau}_k$ reduces to the naive DM estimator, ignoring Markovian dynamics. For $k \geqslant 1$, it incorporates short-term future outcomes, capturing these dynamics. Theorem 1 establishes bias and variance bounds for the TPG estimator $\hat{\tau}_k$ in nonstationary Markovian environments.

Theorem 1. *Under Assumption 1, the bias of the TPG estimator $\hat{\tau}_k$ with respect to the global average treatment effect τ, for any truncation size $0 \leqslant k < T$, is bounded by*

$$|\mathbb{E}[\hat{\tau}_k] - \tau| = O\left(k^2\delta^2 M + \frac{\gamma^k}{1-\gamma}\delta M\right), \tag{3}$$

where $\delta := \sup_{1\leqslant t\leqslant T}\sup_{x\in\mathcal{X}}\|P_1^t(x,\cdot) - P_0^t(x,\cdot)\|_{\mathrm{TV}}$. The variance of $\hat{\tau}_k$ is bounded by

$$\mathrm{Var}(\hat{\tau}_k) = O\left(\frac{(k+1)^3 M^2}{T} + \frac{\gamma(k+1)^2 M^2}{T(1-\gamma)}\right). \tag{4}$$

The idea of the TPG estimator is not ad hoc. It can be shown that the difference between Q-values is closely connected to the policy gradient in a stationary MDP [20]. The estimator can thus be interpreted as the policy gradient of a truncated policy in a nonstationary MDP, a connection we establish below.

Let $\mathcal{L}_\theta^{t,k}$ denote the distribution of the outcome Y_t when policy $\pi_{1/2}$ is followed for the first $(t-k)$ states (if $t > k$), and policy π_θ is followed for the remaining k states (i.e., a truncated policy), conditioned on the initial state distribution $X_1 \sim \rho$, we have:

$$\mathcal{L}_\theta^{t,k} = \mathrm{Law}\left(Y_t \mid X_1 \sim \rho,\ \{Z_u\}_{u=1}^{t-k} \sim \text{i.i.d. Bern}(1/2),\ \{Z_u\}_{u=t-k+1}^{t} \sim \text{i.i.d. Bern}(\theta)\right). \tag{5}$$

The *truncated policy value function* $J_k(\theta)$ is then defined as

$$J_k(\theta) := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{L}_\theta^{t,k}}[Y_t] = \frac{1}{T}\left(\sum_{t=1}^{k}\mathbb{E}_{\mathcal{L}_\theta^{t}}[Y_t] + \sum_{t=k+1}^{T}\mathbb{E}_{\mathcal{L}_\theta^{t,k}}[Y_t]\right). \tag{6}$$

We refer to the gradient of $J_k(\theta)$ as the *truncated policy gradient* and use $J(\theta) := J_{T-1}(\theta)$ to denote the full-horizon policy gradient. For truncation size $0 \leqslant k < T$, define the *truncated Q-value* as

$$Q_{1/2}^{t,k}(z) := \sum_{u=t}^{\min(t+k,T)}\mathbb{E}_{\mathcal{L}_{1/2}^t}[Y_u \mid Z_t = z]. \tag{7}$$

Proposition 1. *The truncated policy gradient $\nabla J_k(\theta)$ evaluated at the uniform randomized policy $\pi_{1/2}$ exists and is given by the average difference in truncated Q-values in (7) over the horizon, i.e.,*

$$\nabla J_k(1/2) = \frac{1}{T}\sum_{t=1}^{T}\left(Q_{1/2}^{t,k}(1) - Q_{1/2}^{t,k}(0)\right).$$

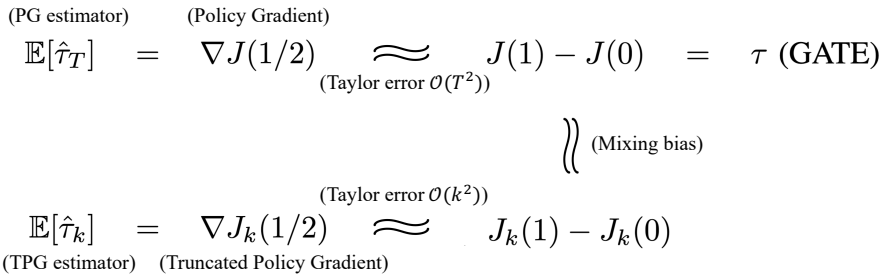

Figure 1: Connections among the estimand GATE, policy gradients $\nabla J_k(1/2)$, and the TPG estimators $\hat{\tau}_k$. The estimator $\hat{\tau}_k$ trades additional mixing bias for reduced Taylor error $O(k^2\delta^2)$, as treatment probability affects only the last k states rather than the full trajectory.

Moreover, observe that with the TPG estimator $\hat{\tau}_k$ defined in (2), we have $\mathbb{E}[\hat{\tau}_k] = \nabla J_k(1/2)$. This indicates that the TPG estimator $\hat{\tau}_k$ can be viewed as an estimation of τ via the truncated policy gradient $\nabla J_k(1/2)$. Accordingly, the bias of $\hat{\tau}_k$ relative to τ decomposes as

$$\mathbb{E}\left[\hat{\tau}_k\right] - \tau = \underbrace{\nabla J_k(1/2) - (J_k(1) - J_k(0))}_{\text{Taylor error w.r.t. } J_k(1/2)} + \underbrace{(J_k(1) - J(1)) + (J(0) - J_k(0))}_{\text{mixing bias}},$$

which includes a Taylor error between $\nabla J_k(1/2)$ and $J_k(1) - J_k(0)$, and an additional mixing bias between the original value function $J(\cdot)$ and its truncated counterpart $J_k(\cdot)$.

Asymptotic normality. Adapting [1], we establish a central limit theorem; the exact statement is omitted for space. This further allows asymptotic variance estimation via a nonstationary Newey–West estimator [9] under the strong mixing condition (i.e., Assumption 1) and a mild Cesàro-L^2 mean stability condition, enabling the construction of confidence intervals and statistical inference.

Case study: NYC ride-sharing simulation (Figure 2). We adapt a large-scale NYC ride-sharing simulator[1] using real data from [16]. We evaluate a pricing policy where the fare equals a fixed per-minute rate times the trip duration; the treatment policy uses a higher rate than the control.

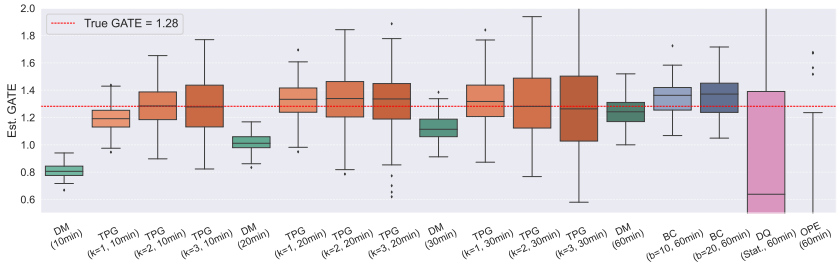

Figure 2: Estimation under NYC ride-sharing simulation with real data. The box represents the interquartile range (IQR), and the error bars extend to the farthest values within 1.5×IQR.

We use a Bernoulli switchback design, assigning each interval (e.g., 10–20 minutes) to treatment or control with probability $1/2$, with system state given by the average number of available drivers. The DM estimator is highly sensitive to interval length: short intervals incur large bias, while longer ones reduce bias but require prior knowledge to choose the interval length. The TPG estimators consistently mitigate bias across intervals. By contrast, switchback estimators show limited benefit: the BC estimator matches DM under 1-hour intervals and overestimates with short burn-in, while stationary DQ and OPE fail to learn reliable Q-values, yielding both high bias and variance.

[1]Code available at https://github.com/wenqian-xing/TPG-Estimator.

References

[1] Alessandro Arlotto and J Michael Steele. A central limit theorem for temporally nonhomogenous markov chains with applications to dynamic programming. *Mathematics of Operations Research*, 41(4):1448–1468, 2016.

[2] Peter M Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. 2017.

[3] Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023.

[4] Nicholas Chamandy. Experimentation in a ridesharing marketplace, September 2016. Lyft Engineering Blog.

[5] Dean Eckles, Brian Karrer, and Johan Ugander. Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1):20150021, 2017.

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

[7] Vivek Farias, Hao Li, Tianyi Peng, Xinyuyang Ren, Huawei Zhang, and Andrew Zheng. Correcting for interference in experiments: A case study at douyin. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 455–466, 2023.

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

[9] Bruce E Hansen. Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica: Journal of the Econometric Society*, pages 967–972, 1992.

[10] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[11] Yuchen Hu and Stefan Wager. Switchback experiments under geometric mixing. *arXiv preprint arXiv:2209.00197*, 2022.

[12] Yuchen Hu and Stefan Wager. Off-policy evaluation in partially observed markov decision processes under sequential ignorability. *The Annals of Statistics*, 51(4):1561–1585, 2023.

[13] Ramesh Johari, Hannah Li, Inessa Liskovich, and Gabriel Y Weintraub. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 68(10):7069–7089, 2022.

[14] Ramesh Johari, Hannah Li, Anushka Murthy, and Gabriel Y Weintraub. When does interference matter? decision-making in platform experiments. *arXiv preprint arXiv:2410.06580*, 2024.

[15] Shuangning Li, Ramesh Johari, Xu Kuang, and Stefan Wager. Experimenting under stochastic congestion. *arXiv preprint arXiv:2302.12093*, 2023.

[16] Tianyi Peng, Naimeng Ye, and Andrew Zheng. Differences-in-neighbors for network interference in experiments. *arXiv preprint arXiv:2503.02271*, 2025.

[17] Chengchun Shi, Xiaoyu Wang, Shikai Luo, Hongtu Zhu, Jieping Ye, and Rui Song. Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association*, 118(543):2059–2071, 2023.

[18] David Simchi-Levi, Chonghuan Wang, and Zeyu Zheng. Non-stationary experimental design under linear trends. *Advances in Neural Information Processing Systems*, 36:32102–32116, 2023.

[19] Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pages 9196–9205. PMLR, 2020.

[20] Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

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

[22] Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. *Advances in neural information processing systems*, 36:15991–16008, 2023.

[23] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

[24] Davide Viviano, Lihua Lei, Guido Imbens, Brian Karrer, Okke Schrijvers, and Liang Shi. Causal clustering: design of cluster experiments under network interference. *arXiv preprint arXiv:2310.14983*, 2023.

[25] Stefan Wager and Kuang Xu. Experimenting in equilibrium. *Management Science*, 67(11):6694–6715, 2021.

[26] Yifan Wu, Ramesh Johari, Vasilis Syrgkanis, and Gabriel Y Weintraub. Switchback price experiments with forward-looking demand. *arXiv preprint arXiv:2410.14904*, 2024.

[27] Yuhang Wu, Zeyu Zheng, Guangyu Zhang, Zuohua Zhang, and Chu Wang. Nonstationary a/b tests: Optimal variance reduction, bias correction, and valid inference. *Management Science*, 2024.