FRONT: Foresighted Online Policy Optimization with Interference

Liner Xiang, Jiayi Wang, Hengrui Cai

Keywords: Contextual bandits, Interference, Online policy optimization, Causal Inference, Statistical inference, Regret bound

Summary

Contextual bandits, which leverage baseline features of sequentially arriving individuals to optimize cumulative rewards while balancing exploration and exploitation, are critical for online decision-making. Existing approaches typically assume no interference, where each individual's action affects only their own reward. Yet, such an assumption can be violated in many practical scenarios, and the oversight of interference can lead to short-sighted policies that focus solely on maximizing the immediate outcomes for individuals, which further results in suboptimal decisions and potentially increased regret over time. To address this significant gap, we introduce the foresighted online policy with interference (FRONT) that innovatively considers the long-term impact of the current decision on subsequent decisions and rewards.

Contribution(s)

1. We exhibit the **online additive outcome model with heterogeneous treatment effects and homogeneous interference effects** as mean outcome model.
   Context: No existing work that models mean outcome using interference over time.
2. We propose an optimal *foresight* policy for online decision-making, which we name **foresighted online policy with interference (FRONT)**. FRONT addresses interference over time, where the actions of prior individuals influence subsequent individuals.
   Context: Prior work considered batch bandits and helped multiple individuals in the batch make coordinated decisions at the same time step (Bargiacchi et al., 2018; Verstraeten et al., 2020; Dubey et al., 2020; Jia et al., 2024; Agarwal et al., 2024; Xu et al., 2024).
3. We develop the online estimator with valid inference under two distinct dependence structures: **the sequential dependence** arising from policy updates and adaptive data, and **the spatial dependence** induced by the growing network interference. We propose a two-level exploration mechanism to break the two-layer independencies using $\epsilon-$Greedy and force pulls triggers.
   Context: Prior work only consider the sequential dependence caused from policy updates and adaptive data when making statistical inference (Chen et al., 2021; Shen et al., 2024; Xu et al., 2024).
4. We propose two regret definitions for FRONT, accounting for **future-regret effects introduced by interference**, and prove sublinear regret in both cases.
   Context: Traditional regret only compares to an optimal policy's cumulative rewards (Chen et al., 2021; Shen et al., 2024).

FRONT: Foresighted Online Policy Optimization with Interference

Liner Xiang[1], **Jiayi Wang**[2], **Hengrui Cai**[1]
linerx1@uci.edu, jiayi.wang2@utdallas.edu, hengrc1@uci.edu

[1]**Department of Statistics, University of California Irvine**
[2]**Department of Mathematical Sciences, University of Texas at Dallas**

Abstract

Contextual bandits, which leverage baseline features of sequentially arriving individuals to optimize cumulative rewards while balancing exploration and exploitation, are critical for online decision-making. Existing approaches typically assume no interference, where each individual's action affects only their own reward. Yet, such an assumption can be violated in many practical scenarios, and the oversight of interference can lead to short-sighted policies that focus solely on maximizing the immediate outcomes for individuals, which further results in suboptimal decisions and potentially increased regret over time. To address this significant gap, we introduce the foresighted online policy with interference (FRONT) that innovatively considers the long-term impact of the current decision on subsequent decisions and rewards. The proposed FRONT method employs a sequence of exploratory and exploitative strategies to manage the intricacies of interference, ensuring robust parameter inference and regret minimization. Theoretically, we establish the tail bound of the online estimation and derive the asymptotic distribution of parameters of interest. We further show how FRONT manages to maintain sublinear regret under two different definitions concerning interference, accounting for both immediate and consequential impacts of decisions. The effectiveness of FRONT is well demonstrated through extensive simulations and a real-world application to urban hotel profits.

1 Introduction

In the online decision-making process, contextual bandit algorithms (Langford & Zhang, 2007) aim to maximize cumulative rewards for sequentially arriving individuals by taking the optimal action based on their baseline features, while also carefully balancing the trade-off between exploitation and exploration. Contextual bandits have been widely applied in various fields, such as recommendation systems (Li et al., 2011; Bouneffouf et al., 2012), precision medicine (Tewari & Murphy, 2017; Durand et al., 2018; Lu et al., 2021), and dynamic pricing (Misra et al., 2019; Tajik et al., 2024). Most existing works (Chen et al., 2021; Bibaut et al., 2021; Zhan et al., 2021; Dimakopoulou et al., 2021; Zhang et al., 2021; Khamaru et al., 2021; Ramprasad et al., 2023; Shen et al., 2024) typically assume that the mean outcome of interest is determined solely by the individual's current action and characteristics, leading to approaches by directly modeling the mean outcome function, such as Upper Confidence Bound (Li et al., 2011) and Thompson Sampling (Agrawal & Goyal, 2013). However, in practice, actions may also affect other individuals' outcomes—a phenomenon called *interference* (Cox, 1958), which poses significant challenges for online decision-making.

Interference among sequential individuals arises in real-world applications. In HPV vaccination campaigns, social media messages create network effects, where one individual's treatment (e.g., an ad) influences others' decisions (Hopfer et al., 2022; Athey et al., 2023). In hotel pricing (Cho

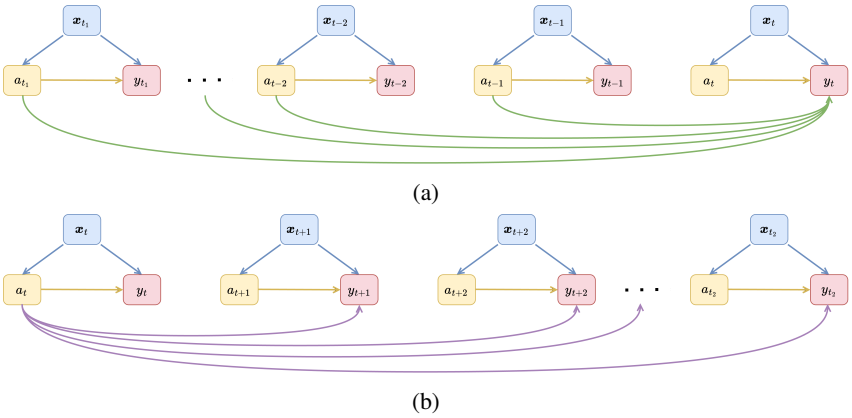

Figure 1: Individuals arrive sequentially, with each triple $\{\boldsymbol{x}_t, a_t, y_t\}$ denoting the context, action, and outcome for the t-th individual. The causal graphs illustrate interference among individuals. **Upper plot (a)**: Green arrows show how earlier actions (from t_1) affect the t-th individual's outcome. **Lower plot (b)**: Purple arrows depict the t-th individual's influence on subsequent ones (up to t_2, possibly infinite). Effects from other actions are omitted for clarity.

et al., 2018), current pricing trends affect future profitability across booking dates. Here, outcomes depend on past actions (Figure 1, upper panel), and current actions affect future outcomes (Figure 1, lower panel). This *long-term interference* challenges traditional contextual bandits, which focus on immediate outcomes and may yield *short-sighted*, suboptimal policies for the population.

Despite extensive literature on causal inference and contextual bandits, research directly addressing our focus remains limited. Existing work primarily studies offline interference with pre-collected data (Bajari et al., 2023), relaxing SUTVA (Rosenbaum, 2007; Forastiere et al., 2021). These approaches assume observable networks, exchangeable ordering, and homogeneous interference effects (Forastiere et al., 2021; Qu et al., 2021; Bargagli-Stoffi et al., 2025). Standard methods employ partial interference and exposure mapping to estimate causal effects both randomized experiments (Sobel, 2006; Hudgens & Halloran, 2008; Aronow, 2012; Liu & Hudgens, 2014; Aronow & Samii, 2017) and observational studies (Manski, 2013; Forastiere et al., 2021; Qu et al., 2021; Lee et al., 2024; Bargagli-Stoffi et al., 2025). However, these techniques require complete network knowledge - a critical limitation for online settings where subsequent individuals are unknown and arrival order becomes crucial. Consequently, existing offline methods cannot handle adaptively collected data in sequential decision-making scenarios.

On the side of online decision making, most contextual bandit works treat each decision independently with no interference assumption, and focus on either minimizing regret (see e.g., Li et al., 2011; Agarwal et al., 2024) or deriving parameter inference (see e.g., Chen et al., 2021; Shen et al., 2024). Recently, several studies considered more general reward models for *batched bandits* where the action of one individual can affect the rewards of others in the batch by extending the multi-agent cooperative game. Pioneering works (Bargiacchi et al., 2018; Verstraeten et al., 2020; Dubey et al., 2020) developed algorithms that help multiple individuals in the batch make coordinated decisions at the same time step. Bargiacchi et al. (2018); Verstraeten et al. (2020); Dubey et al. (2020) proposed coordinated decision-making algorithms, while Jia et al. (2024) studied grid-structured interference and Agarwal et al. (2024) addressed sparse networks. Xu et al. (2024) further considered within-cluster heterogeneous actions. Unlike existing approaches that focus on batched bandits with group-level interference, we explicitly model interference over sequential decision points and derive individualized policies optimized for long-term outcomes.

Contribution 1: To the best of our knowledge, we are the first to propose an optimal *foresight* policy for online decision-making, which we name **foresighted online policy with interference (FRONT)**. As illustrated in Figure 1, our work addresses interference over time, where the actions

of prior individuals influence subsequent individuals. Considering real-world applications (Zhao et al., 2015; Hopfer et al., 2022), where individual interactions and influences expand progressively, we focus on a growing interference scale. Our proposed method, FRONT, explicitly incorporates the long-term effects of current actions into each decision-making step, utilizing the critical role of foresight in achieving optimal performance.

Contribution 2: Secondly, we develop the online estimator with valid inference under the online additive outcome model with heterogeneous treatment effects and homogeneous interference effects. To avoid a growing dimension of features from expanding neighbor interactions, we extend exposure mapping to online settings. Our framework **simultaneously accounts for two distinct dependence structures**: *the sequential dependence* arising from policy updates and adaptive data collected online, and *the spatial dependence* induced by the growing network interference from neighboring individuals' actions. We propose a two-level exploration mechanism using ϵ-Greedy with proper rate and force-pull triggers. Remarkably, even with limited samples and growing interference under online setting, we can still prove *consistency* and *asymptotic normality* of our estimates.

Contribution 3: Thirdly, we propose two regret definitions for FRONT and prove *sublinear regret* in both cases. While traditional regret compares to an optimal policy's cumulative rewards (Chen et al., 2021), interference introduces future-regret effects. We analyze: (1) immediate observed outcomes, and (2) total long-term impact including future manifestations. FRONT achieves sublinear regret under both definitions, demonstrating its efficacy for interference-aware sequential decision-making.

In this paper, we propose a novel policy FRONT, designed to optimize decision-making and maximize long-term benefits. To ensure practical applicability, we provide a comprehensive algorithm accompanied by exploration strategies. Furthermore, we establish the consistency and asymptotic normality of the online estimator under the assumption of ISO convergence. Finally, we analyze the regret under two different definitions and derive sublinear regret bounds for both scenarios.

2 Problem Formulation

Framework. Suppose a sequence of individuals arrives in an order. Let T denote the total number of individuals that can go to infinity. At each time t, where $t \in [T] = \{1, 2, \ldots, T\}$, a new individual arrives, and we observe their contextual covariates $\boldsymbol{x}_t \in \mathbb{R}^{d_1}$. Here, $\boldsymbol{x}_t$ includes 1 for the intercept, and $\boldsymbol{x}_t \overset{\text{i.i.d}}{\sim} \mathcal{P}_X$. We assume that the action set individuals can take is $\{0, 1\}$. After selecting an action a_t, we observe the reward y_t. Denote all the observed information up to time t as $\mathcal{H}_t = \{\boldsymbol{x}_1, a_1, y_1, \boldsymbol{x}_2, a_2, y_2, \cdots, \boldsymbol{x}_t, a_t, y_t\}$. We use $\mathcal{N}_t$ to denote the neighborhood of the t-th individual and $|\mathcal{N}_t|$ denotes the cardinality of $\mathcal{N}_t$. In the online setting, we assume that $\mathcal{N}_t$ consists of previous individuals from $t - |\mathcal{N}_t|$ to $t - 1$, intuitively because they are the observed individuals closest to the t-th individual in time. For simplicity, we denote $g(t) = |\mathcal{N}_t|$.

Including all neighbors' actions as features leads to high and even infinite dimensionality in the online setting. To address this issue, we assume that the effects are mediated through an exposure mapping function (Van der Laan, 2014; Aronow & Samii, 2017; Forastiere et al., 2021). In this work, we consider a simple case where all neighbors' actions contribute equally to the t-th individual. We then define interference action $\bar{a}_t$ as the average action of the previous $g(t)$ individuals, i.e., $\bar{a}_t = \frac{1}{g(t)} \sum_{s=t-g(t)}^{t-1} a_s$. where $g(t)$ should satisfy the following conditions: 1. $0 < g(t) \leq t - 1$, ensuring that each individual is influenced by at most all previous individuals; 2. $g(t)$ is non-decreasing and $g(t) \to \infty$ as $t \to \infty$, guaranteeing that subsequent individuals are influenced by a growing number of previous individuals. Following Forastiere et al. (2021); Xu et al. (2024), we treat $g(t)$ as known, and set $\bar{a}_t = 0$ when $g(t) = 0$, indicating no interference.

Suppose that the outcome given $\boldsymbol{x}_t$, $\bar{a}_t$ and a_t follows by $y_t \equiv \mu(\boldsymbol{x}_t, \bar{a}_t, a_t) + e_t$, where $\mu(\boldsymbol{x}_t, \bar{a}_t, a_t) \equiv \mathbb{E}(y_t | \boldsymbol{x}_t, \bar{a}_t, a_t)$ is the conditional mean outcome, also known as the Q-function (Murphy, 2003; Sutton, 2018). The term e_t represents sub-Gaussian noise, and conditioned on a_t is independent of all previous information $\mathcal{H}_{t-1}$, contextual covariates $\boldsymbol{x}_t$, and the interference action

$\bar{a}_t$. We further assume that the conditional variance is $\mathbb{E}(e_t|a_t = i) = \sigma_i^2$ for $i = 0, 1$. It is worth noting, to our knowledge, that there is *no* existing work that models y_t using $\bar{a}_t$ over time t.

We aim to identify the optimal sequence of actions that maximizes the cumulative outcome, i.e., $\arg\max_{a_1,a_2,\ldots,a_T} \sum_{t=1}^{T} \mu(\boldsymbol{x}_t, \bar{a}_t, a_t)$, which can be solved by backward inductive reasoning (see e.g., Chakraborty & Murphy, 2014).

Working Model. We begin by examining two intuitive policies, which ignore potential impact on subsequent individuals and restrict attention to the suboptimal policy class, introduced in Appendix A in details. In settings with interference, the action a_t affects both immediate outcome but also the future outcomes. However, deriving a general closed-form solution for the *long-term* optimal policy is infeasible, as it depends critically on the specific functional form of the interference mechanism. Following the current online inference literature (Deshpande et al., 2018; Zhang et al., 2020; Chen et al., 2021; Shen et al., 2024; Xu et al., 2024), we start with the online additive outcome model for the conditional mean outcome, which also accounts for the interference effect. Specifically, we propose **the online additive outcome model with heterogeneous treatment effects and homogeneous interference effects** as follows:

$$\mu(\boldsymbol{x}_t, \bar{a}_t, a_t) = \mathbb{E}(y_t|\boldsymbol{x}_t, \bar{a}_t, a_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \bar{a}_t \gamma, \tag{1}$$

where model parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1 \in \mathbb{R}^{d_1}$ are treatment effect parameters associated with choosing action 0 or 1, and $\gamma \in \mathbb{R}$ measures the homogeneous interference effect arising from neighbors' actions. Extensions of the working model (1) can be found in Appendix C.2.

3 The Proposed Method: FRONT

In this section, we derive the optimal policy and formally present the FRONT method along with its implementation via the ϵ-Greedy strategy. Under the online additive outcome model (1), the optimal policy is given by,

$$a_t = \begin{cases} \mathbb{I}\left\{\boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t} \frac{1}{g(s)}\right]\gamma \geq 0\right\}, & \text{if } 1 \leq t \leq T - g(T) - 1 \\ \mathbb{I}\left\{\boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t, s\leq T} \frac{1}{g(s)}\right]\gamma \geq 0\right\}, & \text{if } T - g(T) \leq t \leq T \end{cases}, \tag{2}$$

where $\mathcal{A}_t = \{s : s - g(s) \leq t \leq s - 1\}$ represents all subsequent individuals whose neighbors include the t-th individual. We observe that in the second piecewise solution, the summation term within the indicator function is truncated by the termination time T. However, in online decision-making scenarios, we assume that individuals will arrive indefinitely and T is unknown. Consequently, the optimal policy $\pi^*(\cdot)$ at time t is determined by the dominant first piece of (2).

Proposition 3.1 (*Optimal policy with interference*)**.** *Under the conditional mean outcome model in* (1)*, the optimal policy* $\pi^*(\cdot)$ *at time* t *is given by*

$$a_t^* = \pi^*(\boldsymbol{x}_t) = \mathbb{I}\left\{\boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t} \frac{1}{g(s)}\right]\gamma \geq 0\right\}. \tag{3}$$

We define the term $\left[\sum_{s\in\mathcal{A}_t} \frac{1}{g(s)}\right]\gamma$ as the **interference effect on subsequent outcome (ISO)** for the t-th individual. This represents the contribution of the t-th individual's action to the interference effect, excluding its direct impact on the long-term reward. As shown in Proposition 3.1, the optimal action at time t depends on ISO. Unlike the myopic policy, (3) incorporates ISO as an additional term within the indicator function. (3) can be interpreted as the difference in the long-term cumulative reward between choosing action 0 and action 1 at time t.

In online learning, we fit the proposed online additive outcome model and adopt the derived optimal policy to make decisions. Denote $\boldsymbol{w}_t = (\boldsymbol{x}_t^\top, \bar{a}_t)^\top \in \mathbb{R}^d$ and $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i^\top, \gamma_i)^\top \in \mathbb{R}^d$, $i = 0, 1$, where $d = d_1 + 1$, $\gamma_0 = \gamma_1 = \gamma$. Then (1) can be simplified to $\mu(\boldsymbol{w}_t, a_t) = a_t \boldsymbol{w}_t^\top \boldsymbol{\theta}_1 + (1 - a_t)\boldsymbol{w}_t^\top \boldsymbol{\theta}_0$. We

assume an initial warm-up phase with T_0 samples, at each time $t = T_0 + 1, T_0 + 2, \ldots$, we estimate the parameters $\boldsymbol{\theta}_i$ based on $\mathcal{H}_t$ by

$$\widehat{\boldsymbol{\theta}}_{i,t} = \left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\} \boldsymbol{w}_s \boldsymbol{w}_s^\top \right)^{-1} \left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\} \boldsymbol{w}_s y_s \right), \quad i = 0, 1, \tag{4}$$

if the first term $\frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\} \boldsymbol{w}_s \boldsymbol{w}_s^\top$ is invertible. In this paper, we extend the ϵ-Greedy policy (Chambaz et al., 2017; Chen et al., 2021) with interference to avoid getting trapped in a single action and being led in the wrong direction. To be specific, at each step after the warm-up samples, we first update the online estimators and then apply the ϵ-Greedy to make random choices with probability ϵ_t. The action a_t is then generated by Bernoulli$(\widehat{\pi}_t)$, where

$$\widehat{\pi}_t = (1 - \epsilon_t) \mathbb{I}\left\{ \boldsymbol{x}_t^\top (\widehat{\boldsymbol{\beta}}_{1,t-1} - \widehat{\boldsymbol{\beta}}_{0,t-1}) + \left[\sum_{s \in \mathcal{A}_t} \frac{1}{g(s)} \right] \widehat{\gamma}_{t-1} \geq 0 \right\} + \frac{\epsilon_t}{2}. \tag{5}$$

We name our method as **foresighted online policy with interference (FRONT)**, with the detailed pseudocode provided in Algorithm 1. To enhance the exploration efficiency, particularly for the interference-related parameter γ, we further include the force pulls step in Step (7). Furthermore, during the warm-up period in Step (2), we increase the variability of $\bar{a}_t$ to ensure a robust initial estimator. Details of practical strategies can be found in Appendix C.1.

Algorithm 1 FRONT under ϵ-Greedy

Input: total number of individuals T, clipping parameter C, number of force pulls at one trigger K, warm-up period T_0, warm-up parameter L;
for $t = 1, 2, \ldots, T_0$ **do**
  (1) Sample $d-1$-dimensional context $\boldsymbol{x}_t \in \mathcal{P}_X$;
  (2) Set $a_t = 1$ when $t \in \bigcup_{i=1}^{L} (\frac{i-1}{L} T_0, \frac{2i-1}{2L} T_0]$, and $a_t = 0$ when $t \in \bigcup_{i=1}^{L} (\frac{2i-1}{2L} T_0, \frac{i}{L} T_0]$;
  (3) Update $\bar{a}_t$ by $\bar{a}_t = \frac{1}{g(t)} \sum_{s=t-g(t)}^{t-1} a_s$;
end for
for $t = T_0 + 1, T_0 + 2, \ldots, T$ **do**
  (4) Sample $d-1$-dimensional context $\boldsymbol{x}_t \in \mathcal{P}_X$;
  (5) Update $\widehat{\boldsymbol{\theta}}_{i,t-1}$ by (4) and $\widehat{\mu}_{t-1}(\cdot)$;
  (6) Update $\widehat{\pi}_t(\cdot)$ by (5) and a_t by following Bernoulli $(\widehat{\pi}_t)$;
  if $\lambda_{\min} \left\{ \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\} \boldsymbol{w}_s^\top \boldsymbol{w}_s \right\} \leq C \epsilon_t$ **then**
    if $\bar{a}_{t-1} \leq 0.5$ **then**
      (7a) Perform force pulls by setting $a_t = a_{t+1} = \ldots a_{t+K} = 1$;
    else
      (7b) Perform force pulls by setting $a_t = a_{t+1} = \ldots a_{t+K} = 0$;
    end if
  end if
end for

4 Theoretical Results

In this section, we present our theoretical results. We first derive the tail bound of the online estimator, followed by its asymptotic normality. Next, we analyze the regret rate of our proposed FRONT. All the proofs are provided in supplementary material E.

Parameter Inference. Our theoretical analysis starts from the following assumptions, which enable us to establish the tail bound of the online estimator.

Assumption 4.1 (*Bound*). *For all $\boldsymbol{x} \sim \mathcal{P}_X$, there exist positive constants L_x, λ such that $\|\boldsymbol{x}\|_\infty \leq L_x$ and $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top) > \lambda$. Let $L_w = \max\{1, L_x\}$, then $\|\boldsymbol{w}\|_\infty \leq L_w$ always holds for all $\boldsymbol{w} = (\boldsymbol{x}^\top, \bar{a})^\top$, where $\bar{a} \in [0, 1]$.*

Assumption 4.2 (*Clipping*). *For any time step $t \geq 1$ and any action $i \in \{0,1\}$, there exists a constant C such that $\lambda_{\min}\left\{\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}\{a_s = i\}\boldsymbol{w}_s^\top \boldsymbol{w}_s\right\} > C\epsilon_t$.*

Assumption 4.1, commonly used in contextual bandits (Zhang et al., 2020; Chen et al., 2021; Shen et al., 2024), restricts the contextual covariates, ensuring that the mean of the martingale differences converges to zero. Assumption 4.2 is a technical condition necessary for the consistency and asymptotic normality of the least squares estimators (Deshpande et al., 2018; Zhang et al., 2020; Hadad et al., 2021; Shen et al., 2024). We first derive the consistency of the online estimator.

Theorem 4.1 (*Tail bound for the online estimator*). *In the online decision-making using FRONT, suppose Assumptions 4.1 and 4.2 are satisfied, for $\forall h > 0$, we have*

$$\mathbb{P}\left\{\left\|\widehat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_i\right\|_1 \leq h\right\} \geq 1 - \exp\left\{-\frac{t\epsilon_t^2 C^2 h^2}{2d^2\sigma^2 L_w^2}\right\}, \; i = 0, 1. \tag{6}$$

Theorem 4.1 demonstrates that the consistency of the online estimator can be established if $t\epsilon_t^2 \to \infty$. Corollary 4.1 can be derived straightforwardly as follows.

Corollary 4.1 (*Consistency of the online estimator*). *If conditions in Theorem 4.1 are satisfied and $t\epsilon_t^2 \to \infty$ as $t \to \infty$, then the online estimator is consistent, i.e., $\widehat{\boldsymbol{\theta}}_{i,t} \to \boldsymbol{\theta}_i$ as $t \to \infty$, $i = 0, 1$.*

After deriving the consistent online estimator, we obtain point estimates. We further explore the trend of the interference effect and the statistical properties of the estimator.

Assumption 4.3 (*Force pulls*). *When executing Algorithm 1, the total count of all forced pulls is $\mathcal{O}(\sqrt{T})$.*

Assumption 4.4 (*Convergence of ISO*). *For any known well-specified non-decreasing function $g(t)$ satisfying $0 < g(t) \leq t - 1$, $g(t) \to \infty$ as $t \to \infty$, and $|\mathcal{A}_t| < \infty$, $\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}$ converges to a finite constant. We denote $\kappa_g = \lim_{t\to\infty}\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}$.*

Assumption 4.3 ensures that force pulls are not performed too many times, and it is necessary to derive the convergence of $\bar{a}_t$ and also the regret bound later. Assumption 4.4 is used to establish the convergence of $\bar{a}_t$ and the valid parameter inference. For the general function $g(t)$, it is challenging to derive a closed-form expression for κ_g except in certain special cases. Instead, we perform numerical experiments to verify that Assumption 4.4 holds across various scenarios, as presented in supplementary material D.3. Here, we present two specific cases where closed-form expressions for κ_g can be derived (with details provided in supplementary material E.3), which also correspond to the scenarios used in our simulations: (1) $g(t) = \lfloor \rho t \rfloor$, $\kappa_g = \frac{1}{\rho}\ln\frac{1}{1-\rho}$; (2) $g(t) = \lfloor \rho\sqrt{t} \rfloor$, $\kappa_g = 1$. Note that $0 < \rho < 1$ and $\lfloor t \rfloor$ denotes the greatest integer less than or equal to t.

Corollary 4.2 (*Convergence of interference action $\bar{a}_t$*). *With conditions in Corollary 4.1, Assumptions 4.3 and 4.4 hold, we have $\bar{a}_t \xrightarrow{p} \bar{a}_\infty = (1-\epsilon_\infty)\mathbb{P}(\boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma \geq 0) + \frac{\epsilon_\infty}{2}$, as $t \to \infty$. If $\epsilon_t \to 0$, $\bar{a}_t$ will converge to $\bar{a}_\infty^* = \mathbb{P}(\boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma \geq 0)$, which is the limit of interference action under the optimal policy.*

As shown in Corollary 4.2, when the interference scale diverges and the limit of ISO exists, the interference action $\bar{a}_t$ approaches to a constant $\bar{a}_\infty$. Then the asymptotic normality of estimated parameters can be derived by Martingale Central Limit Theorem.

Theorem 4.2 (*Inference for the online estimator*). *Suppose the conditions in Corollary 4.2 are satisfied, when $\bar{a}_\infty \neq 0$, we have*

$$\sqrt{t}(\widehat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_i) = \sqrt{t}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{i,t} - \boldsymbol{\beta}_i \\ \widehat{\gamma}_{i,t} - \gamma_i\end{pmatrix} \xrightarrow{D} \mathcal{N}_d(\mathbf{0}, S_i), \quad i = 0, 1,$$

where

$$S_i = \sigma_i^2\begin{pmatrix}\frac{\epsilon_\infty}{2}\int \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty)\int_{\mathcal{X}_i}\boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X & \bar{a}_\infty\left[\frac{\epsilon_\infty}{2}\int \boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty)\int_{\mathcal{X}_i}\boldsymbol{x}^\top d\mathcal{P}_X\right] \\ \bar{a}_\infty\left[\frac{\epsilon_\infty}{2}\int \boldsymbol{x}\, d\mathcal{P}_X + (1-\epsilon_\infty)\int_{\mathcal{X}_i}\boldsymbol{x}\, d\mathcal{P}_X\right] & \bar{a}_\infty^2\left[\frac{\epsilon_\infty}{2} + (1-\frac{\epsilon_\infty}{2})\mathbb{P}(\boldsymbol{x}\in\mathcal{X}_i)\right]\end{pmatrix}^{-1}, \tag{7}$$

with $\mathcal{X}_1 = \left\{\boldsymbol{x} : \boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma \geq 0\right\}$ *and* $\mathcal{X}_0 = \left\{\boldsymbol{x} : \boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma < 0\right\}$. *A consistent estimator for* S_i *is given by*

$$\frac{\sum_{s=1}^{t} \mathbb{I}\{a_s = i\}\, \widehat{e}_s^2}{\sum_{s=1}^{t} \mathbb{I}\{a_s = i\}} \left(\frac{1}{t}\sum_{s=1}^{t} \mathbb{I}\{a_s = i\}\, \boldsymbol{w}_s \boldsymbol{w}_s^\top\right)^{-1}, \tag{8}$$

where $\widehat{e}_s = y_s - a_s \boldsymbol{w}_s^\top \widehat{\boldsymbol{\theta}}_{1,t} - (1 - a_s)\boldsymbol{w}_s^\top \widehat{\boldsymbol{\theta}}_{0,t}$.

Regret Analysis. We discuss two definitions of cumulative regret, both of which are reasonable and depend on different perspectives of understanding the problem.

Assumption 4.5 (*Margin*)**.** *With Assumption 4.4 holds, for any* $\boldsymbol{x}_t \sim \mathcal{P}_X$ *and any* t*, there exists a positive constant* M*, such that* $\mathbb{P}\left(0 < \left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t} \frac{1}{g(s)}\right]\gamma\right| < l\right) \leq Ml, \forall l > 0.$

Assumption 4.5 is a margin condition commonly proposed in the contextual bandit literature (Goldenshluger & Zeevi, 2013; Chambaz et al., 2017; Bastani & Bayati, 2020; Chen et al., 2021; Shen et al., 2024; Xu et al., 2024) to restrict the probability of encountering covariates close to these boundaries, thereby reducing the variance caused by incorrect decisions.

Our two definitions of regret depend on the perspective—immediate or consequential—determined by whether the regret manifests by time T or is caused before time T but may not have appeared yet. We first define the cumulative regret $R_1(T)$ based on the first idea as $R_1(T) = \sum_{t=1}^{T} \mathbb{E}\{\mu(\boldsymbol{x}_t, \bar{a}_t^*, a_t^*) - \mu(\boldsymbol{x}_t, \bar{a}_t, a_t)\}$.

We then consider the cumulative regret $R_2(T)$ based on another idea—caused before time T. We define $R_2(T) = \sum_{t=1}^{T} \mathbb{E}\{\nu(\boldsymbol{x}_t, a_t^*) - \nu(\boldsymbol{x}_t, a_t)\}$, where $\nu(\boldsymbol{x}_t, a_t)$ represents the total contribution to the long-term cumulative outcome from the t-th individual, i.e., $\nu(\boldsymbol{x}_t, a_t) = \boldsymbol{x}_t^\top\boldsymbol{\beta}_0 + a_t\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + a_t\left[\sum_{s\in\mathcal{A}_t} \frac{1}{g(s)}\right]\gamma$.

Theorem 4.3 (Regret bound)**.** *Suppose the conditions in Theorem 4.2 and Assumption 4.5 are satisfied. When applying FRONT, the regrets can be bounded respectively as follows:*

$$R_1(T) = \mathcal{O}\left(\sum_{t=1}^{T} \epsilon_t + g(T)T^{-\frac{1}{4}}\right), \quad R_2(T) = \mathcal{O}\left(\sum_{t=1}^{T} \epsilon_t\right).$$

When $t\epsilon_t^2 \to \infty$ as $t \to \infty$, we have $\mathcal{O}\left(\sum_{t=1}^{T} \epsilon_t\right) \geq c\sqrt{t}$, where c is some positive constant. Then the regret exhibits sublinear behavior under both definitions.

5 Numerical Studies

In this section, we evaluate our proposed FRONT policy in comparison to the naïve policy and the myopic policy introduced in Appendix A, and validate the statistical inference results for FRONT. For the interference scale function g, we consider three different scenarios: (1) $g(t) = \lfloor 0.2t \rfloor$; (2) $g(t) = \lfloor 5\sqrt{t} \rfloor$; (3) $g(t) = \lfloor 20t^{0.2} \rfloor$. Details about data generation are provided in Appendix B.1.

Policy Performance. As shown in Figure 2, which presents the cumulative average reward trajectories, FRONT consistently achieves the lowest regret and progressively approaches to the optimal reward over time. The reward under FRONT converges to that of the optimal policy across all three scenarios of $g(t)$, demonstrating the universality of FRONT.

Evaluation of Statistical Inference. The reported results for FRONT include the following metrics: the ratio of the average standard error (SE) to the Monte Carlo standard deviation (MCSD) across 500 replications, the average parameters estimation bias computed 500 experiments, and the coverage probability of the 95% two-sided Wald confidence interval. The confidence intervals are constructed using the parameter estimates and their corresponding estimated standard errors from

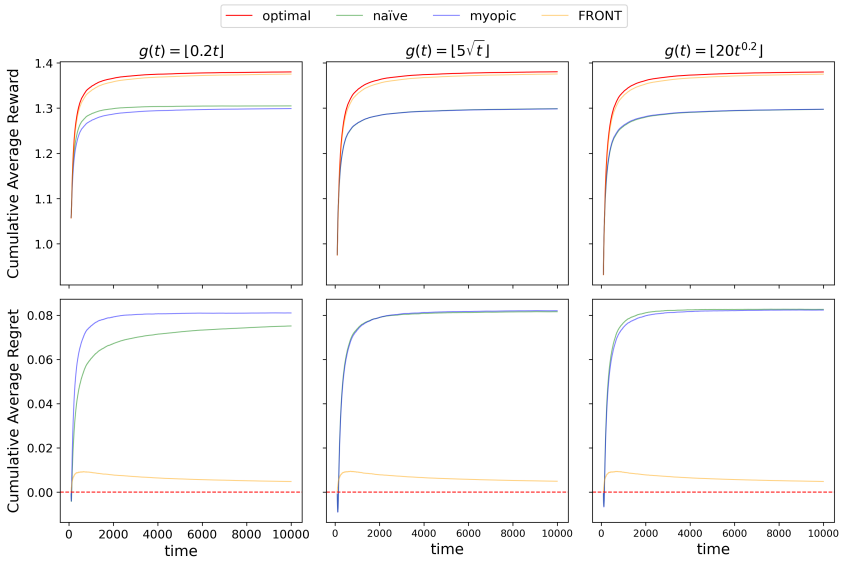

Figure 2: **Upper panel**: the cumulative average reward over time. **Lower panel**: the cumulative average regret, quantifying the gap between the optimal policy and the other three policies (shown in the upper panel), with regret defined by $R_1(T)$.

each experiment, based on (8) in Theorem 4.2. The results are summarized in Figures 4, 5, and 6 in Appendix B.1, which show that FRONT performs well in all three cases: the ratio between SE and MCSD converges to 1, the average bias approaches 0, and the coverage probabilities for $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ are close to the nominal level of 95%. Performances of $\bar{a}_t$ and sensitivity analyzes are provided in supplementary material D.1 and D.2.

6 Real Data Application

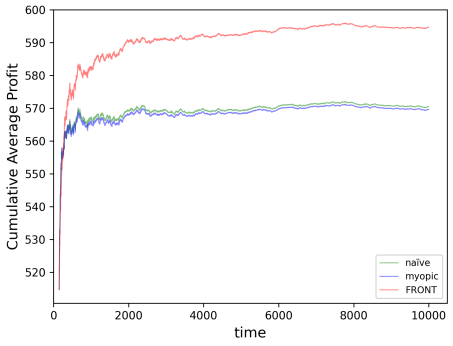

Figure 3: The cumulative average profit achieved by three policies on the hotel dataset.

In this section, we evaluate the performance of our proposed method, FRONT, using publicly available data from a large urban hotel chain ("Hotel 1") (Bodea et al., 2009). It includes records of room purchases and associated revenues with check-in dates between March 12, 2007, to April 15, 2007.

We consider binary actions where $a_t = 1$ ($a_t = 0$) denotes increasing (decreasing) the nightly rate relative to the room-type average. The profit outcome y_t equals (rate - cost) $\times$ rooms $\times$ stay length. Five context features ($d = 7$ with intercept and interference) include: room type (ordinal), booking lead-time (log-transformed), membership status (binary), party size, and rate type (ordinal). After cleaning, we analyze 1,961 purchased-room entries. Assuming a linear model with $g(t) = \lfloor 0.5\sqrt{t} \rfloor$, we simulate online decisions: at each t, we observe $\boldsymbol{x}_t$, compute $\bar{a}_t$, take action a_t, and receive profit $\sim \mathcal{N}(\mu(\boldsymbol{x}_t, \bar{a}_t, a_t), 10^2)$. Comparing naïve, myopic and FRONT policies over $T = 10,000$ steps, FRONT achieves superior cumulative profit (Fig. 3), demonstrating its foresight advantage. Details can be found in Appendix B.2.

A Intuitive Policies

We begin by examining two intuitive policies that could be applied to the online decision-making problem. In classical contextual bandit frameworks, interference effects are typically neglected (Chen et al., 2021; Shen et al., 2024), resulting in a restrictive model $\phi(\boldsymbol{x}_t, a_t)$. In this model, we consider a policy from a suboptimal policy class that evaluates the expected outcomes of available actions without accounting for their potential impact on subsequent individuals. At each time step t, we estimate the parameters using all available historical data $\mathcal{H}_{t-1}$ (Deshpande et al., 2018; Zhang et al., 2020; Chen et al., 2021). We denote this approach as the *naïve* policy, formally expressed as $\mathbb{I}\left\{\tilde{\phi}_{t-1}(\boldsymbol{x}_t, 1) - \tilde{\phi}_{t-1}(\boldsymbol{x}_t, 0) \geq 0\right\}$, where $\tilde{\phi}_{t-1}(\boldsymbol{x}_t, a_t)$ is the estimated Q-function with plugged-in parameter estimates.

Next, we consider an approach that accounts for interference effects in the mean outcome model, while still restricting attention to the suboptimal policy class. Such a policy focuses on maximizing only the immediate outcome at time t, yielding what we name the *myopic* policy. The policy is expressed as $\mathbb{I}\left\{\widehat{\mu}_{t-1}(\boldsymbol{x}_t, \bar{a}_t, 1) - \widehat{\mu}_{t-1}(\boldsymbol{x}_t, \bar{a}_t, 0) \geq 0\right\}$, where $\widehat{\mu}_{t-1}(\boldsymbol{x}_t, \bar{a}_t, a_t)$ denotes the conditional mean outcome estimated correctly using historical information $\mathcal{H}_{t-1}$.

Both the naïve policy and myopic policy ignore ISO, potentially leading to suboptimal decisions at each time step t and resulting in linear regret, as shown in the lower panel of Figure 2 (Section 5). The cumulative average regret $(R_1(t)/t)$ shown in this figure demonstrates that both policies exhibit regret converging to a non-zero constant, showing their $\mathcal{O}(T)$ regret rate.

B Details on Experiments

B.1 Details on Numerical Studies

The context $\boldsymbol{x} = (1, x_1, x_2)^\top$ is defined as follows: x_1 follows a truncated normal distribution with mean zero, variance one, and support $[-10, 10]$, and $x_2 \overset{\text{i.i.d}}{\sim} \text{Uniform}(0, 2)$, ensuring Assumption 4.1 is satisfied. Then, $\boldsymbol{w}$ is four-dimensional, including the interference action $\bar{a}$, i.e., $d = 4$. The true parameters are $\boldsymbol{\theta}_0 = (0.3, -0.1, 0.7, 0.6)^\top$ and $\boldsymbol{\theta}_1 = (0.2, 0.7, 0.3, 0.6)^\top$, which implies $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0 = (-0.1, 0.8, -0.4)^\top$ and $\gamma = 0.6$. Additionally, the noise term is modeled as $e_t | a_t = i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 0.1^2)$ for $i = 0, 1$. For the interference scale function g, we consider three different scenarios: (1) $g(t) = \lfloor 0.2t \rfloor$; (2) $g(t) = \lfloor 5\sqrt{t} \rfloor$; (3) $g(t) = \lfloor 20t^{0.2} \rfloor$.

To ensure a fair comparison, all policies are implemented with identical configurations. Specifically, the exploration rate ϵ_t is set as $\log(t)/10\sqrt{t}$ with the clipping parameter of $C = 0.01$. The termination time is $T = 10{,}000$, with a warm-up period of $T_0 = 100$ and $L = 6$. The force-pull parameter is set to $K = 50$, and we run 500 replications following Algorithm 1. The comparative analysis uses cumulative average rewards as the performance metric, computed as the total reward accumulated up to a given time step divided by the number of steps.

The inference results are summarized in Figures 4, 5, and 6, for three scenarios: (1) $g(t) = \lfloor 0.2t \rfloor$; (2) $g(t) = \lfloor 5\sqrt{t} \rfloor$; (3) $g(t) = \lfloor 20t^{0.2} \rfloor$, respectively. The average bias for β_{01}, γ_0, β_{11}, and γ_1 until the termination time T does not perform as well as for other parameters. The reason is that $\bar{a}_t$ stabilizes to $\bar{a}_\infty$ over time, leading to weak collinearity in the design matrix. Although we use force pulls to introduce variability in $\bar{a}_t$, it still exhibits less variation compared to other covariates, which impacts the accuracy of the estimation. Moreover, a smaller interference scale g_t yields better performance in parameter inference because greater variation in $\bar{a}_t$ increases the variability of samples.

B.2 Details on Real Data Application

We define the action space as binary, where $a_t = 1$ represents an increase in the nightly rate decided by the hotel, and $a_t = 0$ represents a decrease, compared to the average nightly rate for the specific

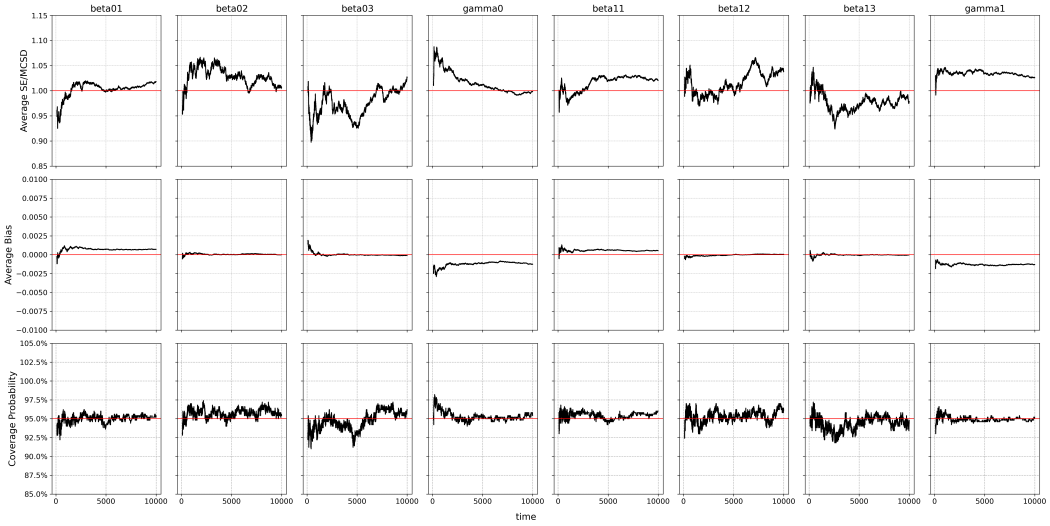

Figure 4: The performance of the online estimator when $g(t) = \lfloor 0.2t \rfloor$ is shown in the figure. **Upper panel**: the ratio between the standard error and the Monte Carlo standard deviation, with the red line indicating the nominal level of 1. **Middle panel**: the bias between the estimated value and the true value. **Lower panel**: the coverage probabilities of the 95% two-sided Wald-type confidence interval, with the red line indicating the nominal level of 95%.

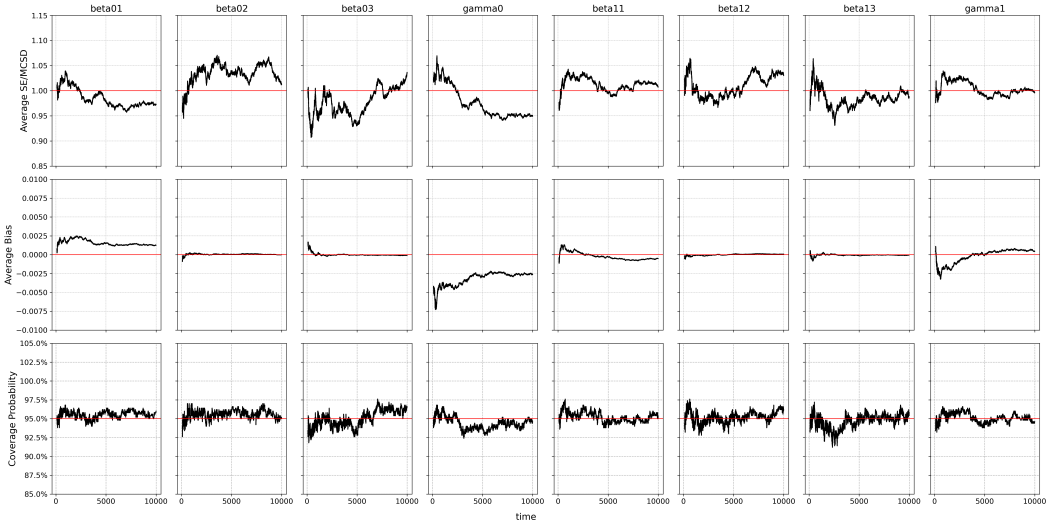

Figure 5: The performance of the online estimator when $g(t) = \lfloor 5\sqrt{t} \rfloor$.

room type. The lowest nightly rate is assumed as the cost for a given room type. The outcome y_t is defined as the profit from each purchase, calculated by multiplying the difference between the nightly rate and the cost by the number of rooms and the length of stay. Additionally, we consider five context features (with $d = 7$ including the intercept and the interference action): room type, advanced purchase dates, membership status, party size, and rate type (e.g., including other services, activities, or rewards). For this offline dataset, we retain only entries where the product was purchased by the customer and exclude any with invalid data, resulting in a total of 1,961 entries. To reduce the number of dummy variables and address imbalances they cause, we treat room type and rate type as ordinal variables, ranking them based on their corresponding average nightly rates. We apply a logarithmic transformation to advanced purchase dates to rescale this

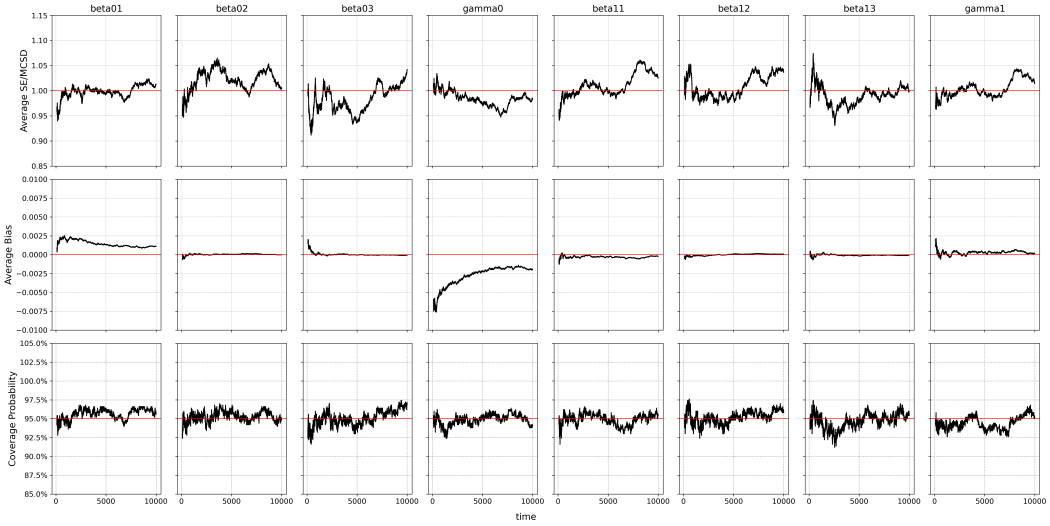

Figure 6: The performance of the online estimator when $g(t) = \lfloor 20t^{0.2} \rfloor$.

feature. Membership status is set to 1 for customers with either standard or VIP memberships, regardless of their membership tier.

Since the true model is unknown, we assume a linear conditional mean outcome model with an interference scale given by $g(t) = \lfloor 0.5\sqrt{t} \rfloor$. Using this model, we derive the fitted coefficients and simulate an online decision-making process. At each time t, a customer arrives, and we draw context features $\boldsymbol{x}_t$, calculate $\bar{a}_t$, and select an action a_t based on the given context and updated policy. The profit is then observed from the distribution $\mathcal{N}(\mu(\boldsymbol{x}_t, \bar{a}_t, a_t), 10^2)$, where $\mu(\boldsymbol{x}_t, \bar{a}_t, a_t)$ represents the outcome model we fit earlier. The warm-up period is set to $T_0 = 150$ with $L = 4$, the force pulls parameter is $K = 50$, the clipping parameter is $C = 0.01$, and the exploration rate is $\epsilon_t = \log(t)/10\sqrt{t}$. Suppose the termination time T is 10,000. Our goal is to maximize cumulative profit, and we apply three policies in this online setting: naïve, myopic, and FRONT to compare their performances. Figure 3 shows the cumulative average profit under the three policies. FRONT achieves higher cumulative profit, demonstrating the strength of FRONT and the importance of foresighted decision-making.

C Discussion

C.1 Practical Strategies

We can select appropriate values for parameters T_0, L and L to ensure sufficient exploration initially and satisfy Assumption 4.3.

We divide the warm-up duration T_0 into $2L$ intervals, alternating actions: even-numbered intervals take action 0, while odd-numbered intervals take action 1. This periodic assignment expands the variability of observed values for $\bar{a}_t$, reducing the need for future exploration. By choosing the suitable parameters T_0, K, and L, we reduce consecutive force pulls and prevent frequent triggers. Previous methods apply force pulls at fixed time points (see e.g., Bastani & Bayati, 2020; Hu & Kallus, 2020), where a single force pull often does not substantially perturb $\bar{a}_t$. Instead, we conduct consecutive pulls whenever the condition in Step (7) in Algorithm 1 is triggered to meet the clipping assumption (i.e., Assumption 4.2, which serves as one theoretical ground for online statistical inference). Specifically, in Step (7), we perform K consecutive force pulls from a_t to a_{t+K}. If $\bar{a}_{t-1} \leq 0.5$, we set the force pulls action to 1; otherwise, we set them to be 0, in order to increase the variability of $\bar{a}_t$.

We divide the warm-up duration T_0 into $2L$ intervals, alternating actions: even-numbered intervals take action 0, while odd-numbered intervals take action 1. This periodic assignment expands the variability of observed values for $\bar{a}_t$, reducing the need for future exploration. By choosing the suitable parameters T_0, K, and L, we reduce consecutive force pulls and prevent frequent triggers.

Specifically, we recommend more exploration during the warm-up phase to establish a reliable estimator early, followed by a reduction in force pulls in later stages to sustain ongoing exploration.

C.2 Other Working Models

In Section 3, we establish the method based on the online additive outcome model with heterogeneous treatment effects and homogeneous interference effects (described as (1)). We now extend our discussion to more general model formulations.

First, consider a model incorporating interaction between $\bar{a}_t$ and a_t: $\mu(\boldsymbol{x}_t, \bar{a}_t, a_t) = \mathbb{E}(y_t|\boldsymbol{x}_t, \bar{a}_t, a_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \bar{a}_t \gamma_0 + a_t \bar{a}_t (\gamma_1 - \gamma_0)$. The optimal action for the t-th user is a nested structure of indicator functions and they should be derived reversely. This formulation is detailed in (12), (13), and (14) in supplementary material E.7. Consequently, determining the optimal policy becomes intractable, due to the unknown termination time T. Second, we consider a model with interaction between $\bar{a}_t$ and $\boldsymbol{x}_t$, i.e., $\mu(\boldsymbol{x}_t, \bar{a}_t, a_t) = \mathbb{E}(y_t|\boldsymbol{x}_t, \bar{a}_t, a_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \bar{a}_t \gamma_0 + \bar{a}_t \boldsymbol{x}_t^\top \boldsymbol{\gamma}$. In this case, the optimal policy depends on the contextual features of future individuals (see (15) in the supplementary material). Due to the uncertainty in these features and their varying influence, deriving the optimal action remains challenging. However, we can develop practical strategies for the special case where $g(t)$ is a constant, denoted by N. Following the similar argument in Proposition 3.1, when $t \leq T - N$, the optimal policy is

$$a_t^* = \mathbb{I}\left\{ \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \gamma_0 + \left(\frac{1}{N} \sum_{s=t+1}^{t+N} \boldsymbol{x}_s^\top \right) \boldsymbol{\gamma} \geq 0 \right\},$$

where $\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2}, \ldots, \boldsymbol{x}_{t+N}$ are unknown at time t. The term $\frac{1}{K} \sum_{s=t+1}^{t+N} \boldsymbol{x}_s^\top$ represents the average of the features from subsequent N individuals. This quantity can be estimated by the sample average, expressed as $\frac{1}{t} \sum_{s=1}^{t} \boldsymbol{x}_s^\top$ or through resampling from the observed feature vectors $\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_t^\top$ to generate predictions for $\boldsymbol{x}_{t+1}^\top, \boldsymbol{x}_{t+2}^\top, \ldots, \boldsymbol{x}_{t+N}^\top$, followed by averaging the results.

C.3 Future Work

There are several directions in which we can extend our work in the future. First, in contextual bandits, policy evaluation is crucial to determine when to stop updating the policy (see e.g., Chen et al., 2021; Shen et al., 2024; Xu et al., 2024). We plan to define policy value that aligns with the two regret definitions and develop value estimation and inference when interference persists over time. Second, model misspecification presents a challenge in the interference setting, and we expect to address this issue in future research. Third, our current framework assumes a binary action set, and we aim to extend it to continuous action spaces. Finally, with respect to considering the interference effect on future outcomes, we have already incorporated the idea from reinforcement learning and will extend our proposal within the RL framework.

Acknowledgments

The authors thank the Action Editor and anonymous reviewers for their constructive and insightful feedback. This work was supported by the National Science Foundation under grant DMS-CDS&E-MSS No. 2401271.

References

Abhineet Agarwal, Anish Agarwal, Lorenzo Masoero, and Justin Whitehouse. Mutli-armed bandits with network interference. *Advances in Neural Information Processing Systems*, 37:36414–

36437, 2024.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Peter M Aronow. A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research*, 41(1):3–16, 2012.

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

Susan Athey, Kristen Grabarz, Michael Luca, and Nils Wernerfelt. Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to covid vaccines. *Proceedings of the National Academy of Sciences*, 120(5):e2208110120, 2023.

Patrick Bajari, Brian Burdick, Guido W Imbens, Lorenzo Masoero, James McQueen, Thomas S Richardson, and Ido M Rosen. Experimental design in marketplaces. *Statistical Science*, 38(3): 458–476, 2023.

Falco J Bargagli-Stoffi, Costanza Tortù, and Laura Forastiere. Heterogeneous treatment and spillover effects under clustered network interference. *The Annals of Applied Statistics*, 19(1):28–55, 2025.

Eugenio Bargiacchi, Timothy Verstraeten, Diederik Roijers, Ann Nowé, and Hado Hasselt. Learning to coordinate with coordination graphs in repeated single-stage multi-agent decision problems. In *International conference on machine learning*, pp. 482–490. PMLR, 2018.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Aurélien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, and Mark van Der Laan. Post-contextual-bandit inference. *Advances in neural information processing systems*, 34:28548–28559, 2021.

Tudor Bodea, Mark Ferguson, and Laurie Garrow. Data set—choice-based revenue management: Data from a major hotel chain. *Manufacturing & Service Operations Management*, 11(2):356–361, 2009.

Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part III 19*, pp. 324–331. Springer, 2012.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1(1):447–464, 2014.

Antoine Chambaz, Wenjing Zheng, and Mark J van der Laan. Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Annals of statistics*, 45(6): 2537, 2017.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021.

Sungjin Cho, Gong Lee, John Rust, and Mengkai Yu. Optimal dynamic hotel pricing. In *2018 Meeting Papers*, volume 179, 2018.

David Roxbee Cox. Planning of experiments. 1958.

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pp. 1194–1203. PMLR, 2018.

Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951, 2021.

Abhimanyu Dubey et al. Kernel methods for cooperative multi-agent contextual bandits. In *International Conference on Machine Learning*, pp. 2740–2750. PMLR, 2020.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*, pp. 67–82. PMLR, 2018.

Laura Forastiere, Edoardo M Airoldi, and Fabrizia Mealli. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, 2021.

Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

P. Hall and C.C. Heyde. *Martingale Limit Theory and Its Application*. Communication and Behavior. Academic Press, 1980. ISBN 9780123193506. URL https://books.google.com/books?id=xxbvAAAAMAAJ.

Suellen Hopfer, Kalani Kieu-Diem Phillips, Maxwell Weinzierl, Hannah E Vasquez, Sarah Alkhatib, and Sanda M Harabagiu. Adaptation and dissemination of a national cancer institute hpv vaccine evidence-based cancer control program to the social media messaging environment. *Frontiers in Digital Health*, 4:819228, 2022.

Yichun Hu and Nathan Kallus. Dtr bandit: Learning to make response-adaptive decisions with low regret. *arXiv preprint arXiv:2005.02791*, 2020.

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

Su Jia, Peter Frazier, and Nathan Kallus. Multi-armed bandits with interference. *arXiv preprint arXiv:2402.01845*, 2024.

Koulik Khamaru, Yash Deshpande, Tor Lattimore, Lester Mackey, and Martin J Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Chanhwa Lee, Donglin Zeng, and Michael G Hudgens. Efficient nonparametric estimation of stochastic policy effects with clustered interference. *Journal of the American Statistical Association*, pp. 1–13, 2024.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Lan Liu and Michael G Hudgens. Large sample randomization inference of causal effects in the presence of interference. *Journal of the american statistical association*, 109(505):288–301, 2014.

Yangyi Lu, Ziping Xu, and Ambuj Tewari. Bandit algorithms for precision medicine. In *Handbook of Statistical Methods for Precision Medicine*, pp. 260–294. Chapman and Hall/CRC, 2021.

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

Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Zhaonan Qu, Ruoxuan Xiong, Jizhou Liu, and Guido Imbens. Efficient treatment effect estimation in observational studies under heterogeneous partial interference. *arXiv preprint arXiv:2107.12420*, 2021.

Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544):2901–2914, 2023.

Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200, 2007.

Ye Shen, Hengrui Cai, and Rui Song. Doubly robust interval estimation for optimal policy evaluation in online learning. *Journal of the American Statistical Association*, pp. 1–20, 2024.

Michael E Sobel. What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407, 2006.

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

Mahmoud Tajik, Babak Mohamadpour Tosarkani, Ahmad Makui, and Rouzbeh Ghousi. A novel two-stage dynamic pricing model for logistics planning using an exploration–exploitation framework: A multi-armed bandit problem. *Expert Systems with Applications*, 246:123060, 2024.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pp. 495–517, 2017.

Mark J Van der Laan. Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2(1):13–74, 2014.

Timothy Verstraeten, Eugenio Bargiacchi, Pieter JK Libin, Jan Helsen, Diederik M Roijers, and Ann Nowé. Multi-agent thompson sampling for bandit applications with sparse neighbourhood structures. *Scientific reports*, 10(1):6728, 2020.

Yang Xu, Wenbin Lu, and Rui Song. Linear contextual bandits with interference. *arXiv preprint arXiv:2409.15682*, 2024.

Ruohan Zhan, Vitor Hadad, David A Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2125–2135, 2021.

Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829, 2020.

Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with m-estimators on adaptively collected data. *Advances in neural information processing systems*, 34:7460–7471, 2021.

Xinyuan Zhao, Liang Wang, Xiao Guo, and Rob Law. The influence of online reviews to online hotel booking intentions. *International Journal of Contemporary Hospitality Management*, 27 (6):1343–1364, 2015.

Supplementary Materials

The following content was not necessarily subject to peer review.

D Extended Simulation Results

D.1 Plot of $\bar{a}_t$

As Figure 7 shows, $\bar{a}_t$ approaches the optimal interference action $\bar{a}^*_\infty$ and will converge over time, which also verifies convergence of ISO. Note that the convergence in the left panel of Figure 7 is much more obvious due to its larger interference scale. When $g(t)$ is smaller, convergence takes longer, even though $g(t) \to \infty$ as $t \to \infty$.

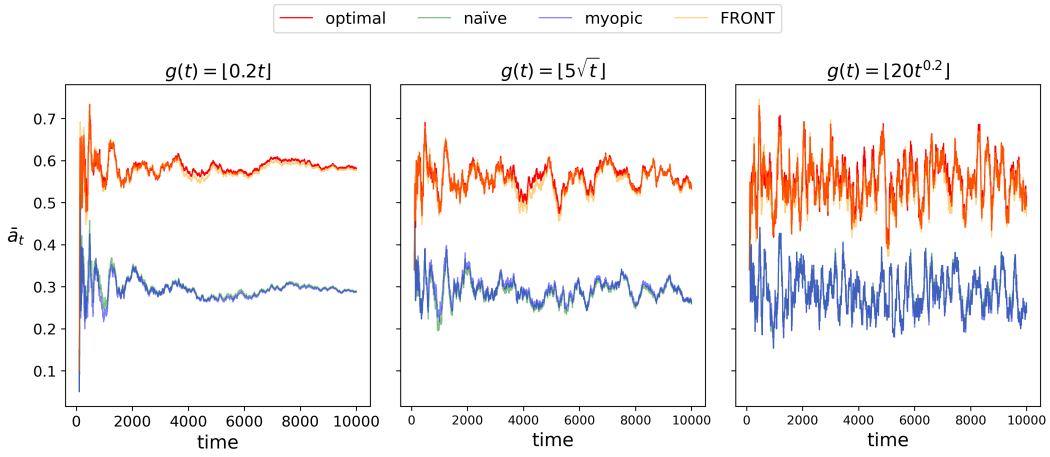

Figure 7: The trend of $\bar{a}_t$ from one replication.

D.2 Sensitivity Analyses

We perform a sensitivity analysis for different values of T_0, L and K, repeating the simulations 500 times for each setting. As illustrated in Figure 8, the coverage probabilities confirm that Algorithm 1 is not sensitive to to variations in these parameters.

D.3 Numerical Verification of the Existence of κ_g

We assume the existence of κ_g and and use it to establish the convergence of ISO. Aside from the special cases proven in E.3, we provide numerical results to support the existence of κ_g. Defining $\zeta_t = \sum_{s \in \mathcal{A}_t} \frac{1}{g(s)}$.

Figure 9 illustrates its trends for when $100 \leq t \leq 10000$ under four different scenarios: (1) $g(t) = \lfloor 20t^{0.2} \rfloor$(a case used in simulations), (2) $g(t) = \lfloor 10t^{0.4} \rfloor$, (3) $g(t) = \lfloor 5t^{0.6} \rfloor$, (4) $g(t) = \lfloor 2t^{0.8} \rfloor$. The stabilization of ζ_t over time supports our assumption, indicating the existence of the limit of ζ_t, i.e., κ_g.

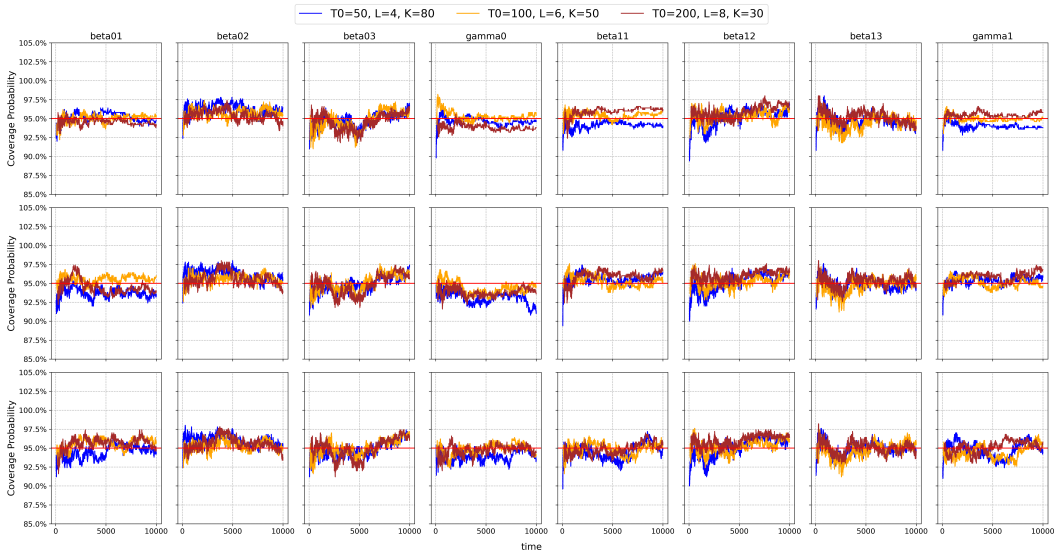

Figure 8: The coverage probabilities of the 95% two-sided CI under different values of T_0, L and K. The red lines indicate the nominal level of 95%.

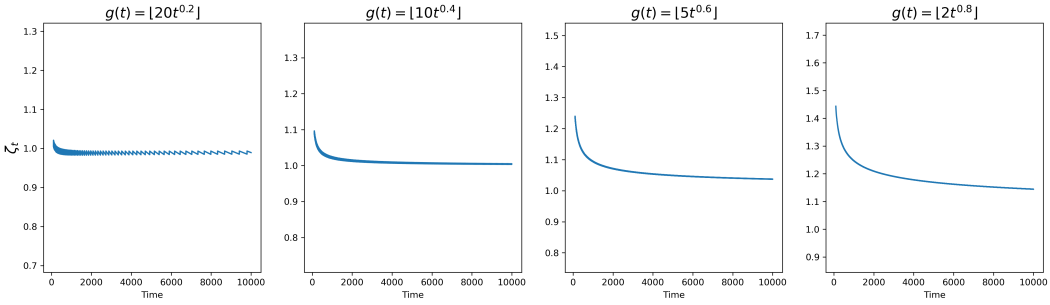

Figure 9: Convergence of ζ_t under different $g(t)$.

E Proof of Main Results

E.1 Optimal Policy with Interference

Proof. The expected cumulative outcome is given by $\sum_{t=1}^{T} \mu(\boldsymbol{x}_t, \bar{a}_t, a_t)$. Under our working model, it can be expanded as

$$\sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \bar{a}_t \gamma.$$

Substituting $\bar{a}_t$ with $\frac{1}{g(t)}\sum_{s=t-g(t)}^{t-1} a_s$, we have

$$\begin{aligned}\sum_{t=1}^{T}\mu(\boldsymbol{x}_t,\bar{a}_t,a_t) &= \sum_{t=1}^{T}\boldsymbol{x}_t^\top\boldsymbol{\beta}_0 + a_t\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \bar{a}_t\gamma \\ &= \sum_{t=1}^{T}\boldsymbol{x}_t^\top\boldsymbol{\beta}_0 + a_t\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \frac{1}{g(t)}\left[\sum_{s=t-g(t)}^{t-1} a_s\right]\gamma \\ &= \sum_{t=1}^{T-g(T)-1}\boldsymbol{x}_t^\top\boldsymbol{\beta}_0 + a_t\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right\} \\ &+ \sum_{t=T-g(T)}^{T}\boldsymbol{x}_t^\top\boldsymbol{\beta}_0 + a_t\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t,s\le T}\frac{1}{g(s)}\right]\gamma\right\},\end{aligned}$$

where $\mathcal{A}_t = \{s : s-g(s)\le t\le s-1\}$. Because we want to maximize the cumulative outcome, then the optimal action should be

$$a_t^* = \begin{cases}\mathbb{I}\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma \ge 0\right\}, & \text{if } 1\le t\le T-g(T)-1 \\ \mathbb{I}\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0) + \left[\sum_{s\in\mathcal{A}_t,s\le T}\frac{1}{g(s)}\right]\gamma \ge 0\right\}, & \text{if } T-g(T)\le t\le T\end{cases}.$$

The expansion of the cumulative reward is also used in the cumulative regret later, where we simply need to ignore the first term $\boldsymbol{x}_t^\top\boldsymbol{\beta}_0$ in both summations. □

E.2 Tail Bound for the Online Estimator

Proof. Denote $\widehat{\Sigma}_i(t) = \frac{1}{t}\sum_{s=1}^{t}\mathbb{I}\{a_s=i\}\boldsymbol{w}_s\boldsymbol{w}_s^\top$. To derive the tail bound of

$$\widehat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_i = \widehat{\Sigma}_i^{-1}(t)\left(\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}\{a_s=i\}\boldsymbol{w}_s e_s\right)\ i=0,1,$$

we need to apply Assumption 4.2 to bound the minimum eigenvalue. By Assumption 4.2,

$$\lambda_{\min}\left\{\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}\{a_s=i\}\boldsymbol{w}_s^\top\boldsymbol{w}_s\right\} = \lambda_{\min}(\widehat{\Sigma}_i(t)) > C\epsilon_t,$$

By Lemma 2 in Chen et al. (2021), we have

$$\mathbb{P}\left\{\left\|\widehat{\boldsymbol{\theta}}_{i,t}-\boldsymbol{\theta}_i\right\|_1 \le h\right\} \ge 1-\exp\left\{-\frac{t\epsilon_t^2C^2h^2}{2d^2\sigma^2L_w^2}\right\}.$$

□

E.3 Special Cases for the Expression of κ_g

Proof. We derive the closed-form expression of κ_g for two special cases: linear scale and square root scale.

1. $g(t) = \lfloor\rho t\rfloor$.

The set of individuals receiving the influence from the t-th individual is

$$\mathcal{A}_t = \{s : s-g(s)\le t\le s-1\} = \{s : s-\lfloor\rho s\rfloor\le t\le s-1\} = \left\{s : t+1\le s\le\left\lfloor\frac{t}{1-\rho}\right\rfloor\right\},$$

and then $\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}$ is given by $\sum_{s=t+1}^{\lfloor\frac{t}{1-\rho}\rfloor}\frac{1}{g(s)}$. For this summation, we can bound it by

$$\int_{t+1}^{\lfloor\frac{t}{1-\rho}\rfloor+1}\frac{1}{\lfloor\rho s\rfloor}\,ds\le\sum_{s=t+1}^{\lfloor\frac{t}{1-\rho}\rfloor}\frac{1}{\lfloor\rho s\rfloor}\le\int_{t}^{\lfloor\frac{t}{1-\rho}\rfloor}\frac{1}{\lfloor\rho s\rfloor}\,ds.$$

The lower bound can be formulated as

$$\int_{t+1}^{\lfloor\frac{t}{1-\rho}\rfloor+1}\frac{1}{\lfloor\rho s\rfloor}\,ds\ge\int_{t+1}^{\frac{t}{1-\rho}}\frac{1}{\rho s}\,ds=\frac{1}{\rho}\ln\rho s\Big|_{t+1}^{\frac{t}{1-\rho}}=\frac{1}{\rho}\ln\frac{\frac{\rho t}{1-\rho}}{\rho(t+1)},$$

then let $t\to\infty$ and by L'Hopital's rule, we have

$$\frac{1}{\rho}\lim_{t\to\infty}\ln\frac{\frac{\rho t}{1-\rho}}{\rho(t+1)}=\frac{1}{\rho}\ln\frac{1}{1-\rho}.$$

Similarly, the upper bound can be formulated as

$$\int_{t}^{\lfloor\frac{t}{1-\rho}\rfloor}\frac{1}{\lfloor\rho s\rfloor}\,ds\le\int_{t}^{\frac{t}{1-\rho}}\frac{1}{\rho s-1}\,ds=\frac{1}{\rho}\ln\left(\rho s-1\right)\Big|_{t}^{\frac{t}{1-\rho}}=\frac{1}{\rho}\ln\frac{\frac{\rho t}{1-\rho}-1}{\rho t-1},$$

then let $t\to\infty$ and by L'Hopital's rule, we have

$$\frac{1}{\rho}\lim_{t\to\infty}\ln\frac{\frac{\rho t}{1-\rho}-1}{\rho t-1}=\frac{1}{\rho}\ln\frac{1}{1-\rho}.$$

Finally, by Squeeze Theorem, when $g(t)=\lfloor\rho t\rfloor$, we have

$$\kappa_g=\lim_{t\to\infty}\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}=\frac{1}{\rho}\ln\frac{1}{1-\rho}.$$

2. $g(t)=\lfloor\rho\sqrt{t}\rfloor$.

By the quadratic formula, the set of individuals that receive the influence of the t-th individual is

$$\begin{aligned}\mathcal{A}_t=\{s:s-g(s)\le t\le s-1\}&=\left\{s:s-\lfloor\rho\sqrt{s}\rfloor\le t\le s-1\right\}\\&=\left\{s:t+1\le s\le\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor\right\},\end{aligned}$$

then

$$\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}=\sum_{s=t+1}^{\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor}\frac{1}{g(s)}.$$

Similarly, we obtain the lower and upper bounds as follows:

$$\int_{t+1}^{\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor+1}\frac{1}{\lfloor\rho\sqrt{s}\rfloor}\,ds\le\sum_{s=t+1}^{\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor}\frac{1}{\lfloor\rho\sqrt{s}\rfloor}\le\int_{t}^{\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor}\frac{1}{\lfloor\rho\sqrt{s}\rfloor}\,ds.$$

The lower bound can be formulated as

$$\int_{t+1}^{\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor+1}\frac{1}{\lfloor\rho\sqrt{s}\rfloor}\,ds\ge\int_{t+1}^{\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}}\frac{1}{\rho\sqrt{s}}\,ds=\frac{2}{\rho}\sqrt{s}\Big|_{t+1}^{\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}}=\frac{2}{\rho}\left(\frac{\rho+\sqrt{\rho^2+4t}}{2}-\sqrt{t+1}\right),$$

then let $t \to \infty$, we have

$$\lim_{t\to\infty}\left(\frac{\sqrt{\rho^2+4t}}{2}-\sqrt{t+1}\right)=\lim_{t\to\infty}\left(\frac{\rho^2/4+t-(t+1)}{\sqrt{\rho^2/4+t}+\sqrt{t+1}}\right)=\lim_{t\to\infty}\left(\frac{\rho^2/4-1}{\sqrt{\rho^2/4+t}+\sqrt{t+1}}\right)=0.$$

Finally, the limit is given by

$$\frac{2}{\rho}\lim_{t\to\infty}\left(\frac{\rho+\sqrt{\rho^2+4t}}{2}-\sqrt{t+1}\right)=\frac{2}{\rho}\lim_{t\to\infty}\left(\frac{\rho}{2}+\frac{\sqrt{\rho^2+4t}}{2}-\sqrt{t+1}\right)=1.$$

We next consider the upper bound,

$$\int_t^{\left\lfloor\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}\right\rfloor}\frac{1}{\lfloor\rho\sqrt{s}\rfloor}\,ds\le\int_t^{\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}}\frac{1}{\rho\sqrt{s}-1}\,ds=\left(\frac{2}{\rho}\sqrt{s}+\frac{2}{\rho}\ln\left(\rho\sqrt{s}-1\right)\right)\Big|_t^{\frac{(\rho+\sqrt{\rho^2+4t})^2}{4}}.$$

Following a similar approach as in deriving the lower bound, when $t \to \infty$, the first term becomes

$$\frac{2}{\rho}\lim_{t\to\infty}\left(\frac{\rho+\sqrt{\rho^2+4t}}{2}-\sqrt{t}\right)=1.$$

For the second term, by L'Hopital's rule, we have,

$$\frac{2}{\rho}\lim_{t\to\infty}\ln\left(\frac{\rho(\rho+\sqrt{\rho^2+4t})/2-1}{\rho\sqrt{t}-1}\right)=0,$$

thus the limit of upper bound is also 1. By Squeeze Theorem, we have $\kappa_g = 1$. The proof is hence completed.

□

E.4 Convergence of Interference Action $\bar{a}_t$

Proof. **Step 1**: We first prove the convergence of the optimal interference action $\bar{a}_t^*$. The optimal policy is defined as

$$a_t^*=\pi^*(\boldsymbol{x}_t)=\mathbb{I}\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\ge 0\right\}.$$

Then optimal interference action is

$$\bar{a}_t^*=\frac{1}{g(t)}\sum_{s=t-g(t)}^{t-1}a_s^*=\frac{1}{g(t)}\sum_{s=t-g(t)}^{t-1}\mathbb{I}\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\ge 0\right\}.$$

Because $\boldsymbol{x}_t$ are i.i.d samples drawn from $\mathcal{P}_X$, the optimal actions a_t^* are independent. Besides, with Assumption 4.4,

$$\begin{aligned}\mathbb{E}(a_t^*)&=\mathbb{E}\left(\mathbb{I}\left\{\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\ge 0\right\}\right)\\&=\mathbb{P}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\ge 0\right)\\&\to\mathbb{P}\left(\boldsymbol{x}^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\kappa_g\gamma\ge 0\right),\end{aligned}$$

as $t \to \infty$. Then we have $\lim_{t\to\infty} \sum_{s=1}^{t} \mathbb{E}a_s^* = \mathbb{P}\left(\boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma \geq 0\right)$. By Kolmogorov Strong Law of Large Numbers, we have

$$\mathbb{P}\left\{\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t}(a_s^* - \mathbb{E}a_s^*)\right\} = 1,$$

which can be simplified as

$$\frac{1}{t}\sum_{s=1}^{t} a_s^* \xrightarrow{p} \mathbb{P}\left(\boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma \geq 0\right),$$

i.e., $\frac{1}{t}\sum_{s=1}^{t} a_s^* \xrightarrow{p} \bar{a}_\infty^*$. Now that we have the convergence of the optimal action average, we can derive the convergence of the optimal interference action easily. If $t - g(t) \to \infty$ as $t \to \infty$, we have

$$\frac{1}{t-g(t)-1}\sum_{s=1}^{t-g(t)-1} a_s^* \xrightarrow{p} \bar{a}_\infty^*,$$

then

$$\frac{1}{g(t)}\sum_{s=t-g(t)}^{t-1}(a_s^* - \bar{a}_\infty^*) = \frac{1}{g(t)}\left[\sum_{s=1}^{t-1} a_s^* - (t-1)\bar{a}_\infty^* - \left(\sum_{s=1}^{t-g(t)-1} a_s^* - (t-g(t)-1)\bar{a}_\infty^*\right)\right] \xrightarrow{p} 0.$$

If $t - g(t)$ is bounded by some constant N_0, i.e., $g(t) \geq t - N_0$, then we have $|\mathcal{A}_t| = \infty$. This case is excluded under Assumption 4.4, since it causes an unbounded ISO and an infinite interference effect.

Step 2: Then we will prove the convergence of $\bar{a}_t$. Our proposed policy is defined as $a_t \sim$ Bernoulli$(\widehat{\pi}_t)$, where

$$\widehat{\pi}_t = (1-\epsilon_t)\mathbb{I}\left\{\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1} - \widehat{\boldsymbol{\beta}}_{0,t-1}) + \left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1} \geq 0\right\} + \frac{\epsilon_t}{2}, \tag{9}$$

where ϵ_t is the exploration probability under ϵ-greedy. First, we assume that all actions taken follow (9) rather than force pulls, then

$$\begin{aligned}
\mathbb{E}(a_t|\mathcal{H}_{t-1}) &= \mathbb{E}[\mathbb{E}(a_t|\mathcal{H}_{t-1}, \boldsymbol{x}_t)|\mathcal{H}_{t-1}] = \mathbb{E}[\widehat{\pi}_t|\mathcal{H}_{t-1}] \\
&= \mathbb{E}\left((1-\epsilon_t)\mathbb{I}\left\{\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1} - \widehat{\boldsymbol{\beta}}_{0,t-1}) + \left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1} \geq 0\right\} + \frac{\epsilon_t}{2}\Big|\mathcal{H}_{t-1}\right) \\
&= (1-\epsilon_t)\mathbb{P}\left\{\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1} - \widehat{\boldsymbol{\beta}}_{0,t-1}) + \left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1} \geq 0\Big|\widehat{\boldsymbol{\beta}}_{0,t-1}, \widehat{\boldsymbol{\beta}}_{1,t-1}, \widehat{\gamma}_{t-1}\right\} + \frac{\epsilon_t}{2} \\
&= (1-\epsilon_t)\mathbb{P}_{\boldsymbol{X}}\left\{\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1} - \widehat{\boldsymbol{\beta}}_{0,t-1}) + \left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1} \geq 0\right\} + \frac{\epsilon_t}{2}.
\end{aligned}$$

By Assumption 4.4, Corollary 4.2 and Continuous Mapping Theorem, we have

$$\mathbb{E}(a_t|\mathcal{H}_{t-1}) \to (1-\epsilon_\infty)\mathbb{P}\left\{\boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g\gamma \geq 0\right\} + \frac{\epsilon_\infty}{2}, \tag{10}$$

i.e., $\mathbb{E}(a_t|\mathcal{H}_{t-1}) \to (1-\epsilon_\infty)\bar{a}_\infty^* + \frac{\epsilon_\infty}{2}$ as $t \to \infty$.
Let A_0 be an r.v. such that $\mathbb{P}(A_0 = 1) = 1$, then we have $\mathbb{E}(A_0) = 1 < \infty$. For any t and $h > 0$, we have $\mathbb{P}(|a_t| > h) = \mathbb{P}(a_t > h) \leq \mathbb{P}(A_0 > h)$.
Now all conditions of Theorem 2.19 from Hall & Heyde (1980) are satisfied, then we have

$$\frac{1}{t}\sum_{s=1}^{t}[a_s - \mathbb{E}(a_s|\mathcal{H}_{s-1})] \xrightarrow{p} 0.$$

By (10) and Lemma 4 in Chen et al. (2021), we have $\frac{1}{t}\sum_{s=1}^{t}\mathbb{E}(a_s|\mathcal{H}_{s-1}) \to (1-\epsilon_\infty)\bar{a}_\infty^* + \frac{\epsilon_\infty}{2}$. Based on Assumption 4.3, the count of force pulls are $\mathcal{O}(\sqrt{t})$, so the difference between $\sum_{s=1}^{t} a_t$ in the practical setting and actions drawn following (9) is at most $\mathcal{O}(\sqrt{t})$. After being divided by t, this difference becomes $\mathcal{O}(1/\sqrt{t})$, i.e. $o(1)$. Therefore, we have $\frac{1}{t}\sum_{s=1}^{t} a_s \xrightarrow{p} (1-\epsilon_\infty)\bar{a}_\infty^* + \frac{\epsilon_\infty}{2}$. Similarly, as shown in the proof in Step 1, because $g(t) \to \infty$, we still have

$$\bar{a}_t \xrightarrow{p} (1-\epsilon_\infty)\bar{a}_\infty^* + \frac{\epsilon_\infty}{2},$$

i.e. $\bar{a}_t \xrightarrow{p} \bar{a}_\infty$. □

E.5 Asymptotic Normality of the Online Estimator

Proof. Based on our online estimator, we have

$$\sqrt{t}(\widehat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_i) = \left(\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}\{a_s = i\}\boldsymbol{w}_s\boldsymbol{w}_s^\top\right)^{-1}\left(\frac{1}{\sqrt{t}}\sum_{s=1}^{t}\mathbb{I}\{a_s = i\}\boldsymbol{w}_s e_s\right), \quad i = 0, 1.$$

1. First we will show that

$$\frac{1}{\sqrt{t}}\sum_{s=1}^{t}\mathbb{I}\{a_s = i\}\boldsymbol{w}_s e_s \xrightarrow{D} \mathcal{N}_d(0, G_i).$$

Using Cramer-Wold device, it suffices to show for any $\boldsymbol{v} \in \mathbb{R}^d$,

$$\frac{1}{\sqrt{t}}\sum_{s=1}^{t}\mathbb{I}\{a_s = i\}\boldsymbol{v}^\top\boldsymbol{w}_s e_s \xrightarrow{D} \mathcal{N}_d(0, \boldsymbol{v}^\top G_i \boldsymbol{v}).$$

For $1 \le j \le t$ and $t \ge 1$, define $\mathcal{H}_{tj} = \mathcal{H}_j$, and

$$M_{tj} = \frac{1}{\sqrt{t}}\sum_{s=1}^{t}\mathbb{I}\{a_s = i\}\boldsymbol{v}^\top\boldsymbol{w}_s e_s.$$

Note that $\mathbb{E}(\mathbb{I}\{a_s = i\}\boldsymbol{v}^\top\boldsymbol{w}_s e_s|\mathcal{H}_{t,s-1}) = \mathbb{E}(\mathbb{I}\{a_s = i\}\boldsymbol{v}^\top\boldsymbol{w}_s|\mathcal{H}_{t,s-1})\mathbb{E}(e_s|\mathcal{H}_{t,s-1}, a_s = i) = 0$, indicating that $\{M_{tj}, \mathcal{H}_{tj}, 1 \le j \le t, t \ge 1\}$ is a martingale array.
We will now prove the convergence of M_{tt} using Martingale Central Limit Theorem (see Theorem 3.2 in Hall & Heyde (1980)). The proof proceeds in two steps: firstly we will verify the conditional Lindeberg condition is satisfied and secondly we will find the limit of conditional variance.
(a) Check the conditional Lindeberg condition. For $\forall\delta > 0$,

$$\begin{aligned}
&\sum_{s=1}^{t}\mathbb{E}\left[\frac{1}{t}\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}^\top\boldsymbol{w}_s e_s\right)^2\mathbb{I}\left\{\mathbb{I}\{a_s = i\}\boldsymbol{v}^\top\boldsymbol{w}_s e_s > \delta\sqrt{t}\right\}\middle|\mathcal{H}_{t,s-1}\right] \\
\le&\frac{\|\boldsymbol{v}\|_2^2 L_w^2 d}{t}\sum_{s=1}^{t}\mathbb{E}\left[\mathbb{I}\{a_s = i\}e_s^2\mathbb{I}\left\{\mathbb{I}\{a_s = i\}e_s^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|^2 L_w^2 d}\right\}\middle|\mathcal{H}_{s-1}\right] \\
=&\frac{\|\boldsymbol{v}\|_2^2 L_w^2 d}{t}\sum_{s=1}^{t}\mathbb{E}\left(\mathbb{I}\{a_s = i\}|\mathcal{H}_{s-1}\right)\mathbb{E}\left[e_s^2\mathbb{I}\left\{\mathbb{I}\{a_s = i\}e_s^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|^2 L_w^2 d}\right\}\middle|\mathcal{H}_{s-1}\right] \\
\le&\frac{\|\boldsymbol{v}\|_2^2 L_w^2 d}{t}\sum_{s=1}^{t}\mathbb{E}\left[e_{s(i)}^2\mathbb{I}\left\{e_{s(i)}^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|^2 L_w^2 d}\right\}\middle|\mathcal{H}_{s-1}\right],
\end{aligned}$$

where $e_{s(i)} = e_s$ when $a_s = i$ and 0 otherwise. e_s conditioned on a_s are i.i.d distributed, so the right hand side comes to be

$$\|\boldsymbol{v}\|_2^2 L_w^2 d\mathbb{E}\left[e^2\mathbb{I}\left\{e^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|^2 L_w^2 d}\right\}\right],$$

where e is a random variable defined by $e_s|\mathcal{H}_{s-1}$. Since $e^2\mathbb{I}\left\{e^2 > \frac{\delta^2 t}{\|\boldsymbol{v}\|^2 L_w^2 d}\right\}$ is bounded by e^2 with $\mathbb{E}(e^2) < \infty$ and convergences to 0 almost surely as $t \to \infty$. Therefore, by Dominated Convergence Theorem, we get

$$\sum_{s=1}^{t} \mathbb{E}\left[\frac{1}{t}\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}^\top \boldsymbol{w}_s e_s\right)^2 \mathbb{I}\left\{\mathbb{I}\{a_s = i\}\boldsymbol{v}^\top \boldsymbol{w}_s e_s > \delta\sqrt{t}\right\}\Big|\mathcal{H}_{t,s-1}\right] \to 0, \text{ as } t \to \infty.$$

(b) Derive the limit of the conditional variance. The conditional variance is given by

$$\begin{aligned}\widehat{\eta}_t^2 &= \sum_{s=1}^{t} \mathbb{E}\left\{\frac{1}{t}\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}^\top \boldsymbol{w}_s\right)^2 \Big|\mathcal{H}_{t,s-1}\right\} \\ &= \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}\left\{\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}^\top \boldsymbol{w}_s\right)^2 \mathbb{E}\left[e_s^2 \mid a_s = i, \boldsymbol{w}_s\right] \mid \mathcal{H}_{s-1}\right\}.\end{aligned}$$

Given $\mathcal{H}_{s-1}$, $\bar{a}_s$ is known, so the expectation is take with respect to $\boldsymbol{x}_s$. As for the random noise, given a_s, e_s is independent of $\mathcal{H}_{s-1}$ and $\boldsymbol{w}_s$, and $\mathbb{E}(e_s^2|a_s = i, \boldsymbol{w}_s) = \sigma_i^2$. Thus we have

$$\begin{aligned}\widehat{\eta}_t^2 &= \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}\left\{\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}^\top \boldsymbol{w}_s\right)^2 \sigma_i^2 \mid \mathcal{H}_{s-1}\right\} \\ &= \frac{1}{t}\sum_{s=1}^{t} \sigma_i^2 \mathbb{E}\left\{\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v}\right) \mid \mathcal{H}_{s-1}\right\}.\end{aligned}$$

Let $\boldsymbol{v} = (\boldsymbol{v}_1^\top, v_2)^\top$ to align with $\boldsymbol{w}_t = (\boldsymbol{x}_t^\top, \bar{a}_t)^\top$, where $\boldsymbol{v}_1 \in \mathbb{R}^{d-1}$ and $v_2 \in \mathbb{R}$. Then we have $\left(\boldsymbol{v}^\top \boldsymbol{w}_s\right)^2 = \left(\boldsymbol{v}_1^\top \boldsymbol{x}_s + v_2 \bar{a}_s\right)^2$.

$$\widehat{\eta}_t^2 = \frac{\sigma_i^2}{t}\sum_{s=1}^{t} \mathbb{E}\left\{\mathbb{I}\{a_s = i\}\left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right)\Big|\mathcal{H}_{s-1}\right\}.$$

Denote the difference between estimated parameter, $\widehat{\boldsymbol{\beta}}_{1,s-1} - \widehat{\boldsymbol{\beta}}_{0,s-1}$, as $\widehat{\boldsymbol{\beta}}_{s-1}$. We have defined $\zeta_t = \sum_{s\in\mathcal{A}_t} \frac{1}{g(s)}$ in Section D.3. Then the expectation term in the above equation can be expressed as a continuous function p_i of $\widehat{\boldsymbol{\beta}}_{s-1}$, $\widehat{\gamma}$, $\bar{a}_s$, ζ_s and ϵ_s,

$$\begin{aligned}p_i(\widehat{\boldsymbol{\beta}}_{s-1}, \widehat{\gamma}, \bar{a}_s, \zeta_s, \epsilon_s) = \frac{\epsilon_s}{2}\int &\left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right) d\mathcal{P}_X \\ &+ (1-\epsilon_s)\int_{\mathcal{X}_{is}} \left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right) d\mathcal{P}_X,\end{aligned}$$

where $\mathcal{X}_{1s} = \left\{\boldsymbol{x} : \boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{1,s-1} - \widehat{\boldsymbol{\beta}}_{0,s-1}) + \zeta_s\widehat{\gamma}_{s-1} \geq 0\right\}$ and $\mathcal{X}_{0s} = \mathbb{R}^{d-1}\backslash\mathcal{X}_{1s}$. We can rewrite $p_i(\widehat{\boldsymbol{\beta}}_{s-1}, \widehat{\gamma}, \bar{a}_s, \zeta_s, \epsilon_s)$ as

$$\begin{aligned}p_1(\widehat{\boldsymbol{\beta}}_{s-1}, \widehat{\gamma}, \bar{a}_s, \zeta_s, \epsilon_s) = \frac{\epsilon_s}{2}\int &\left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right) d\mathcal{P}_X \\ &+ (1-\epsilon_s)\int \mathbb{I}\left\{\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{1,s-1} - \widehat{\boldsymbol{\beta}}_{0,s-1}) + \zeta_s\widehat{\gamma}_{s-1} \geq 0\right\} \\ &\left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right) d\mathcal{P}_X,\end{aligned}$$

and similarly,

$$\begin{aligned}p_0(\widehat{\boldsymbol{\beta}}_{s-1}, \widehat{\gamma}, \bar{a}_s, \zeta_s, \epsilon_s) = \frac{\epsilon_s}{2}\int &\left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right) d\mathcal{P}_X \\ &+ (1-\epsilon_s)\int \mathbb{I}\left\{\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{1,s-1} - \widehat{\boldsymbol{\beta}}_{0,s-1}) + \zeta_s\widehat{\gamma}_{s-1} < 0\right\} \\ &\left(\boldsymbol{v}_1^\top \boldsymbol{x}_s \boldsymbol{x}_s^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_s^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x}_s v_2 \bar{a}_s\right) d\mathcal{P}_X.\end{aligned}$$

Then by Corollary 4.1, 4.2, Assumption 4.4 and Continuous Mapping Theorem,

$$p_i(\widehat{\boldsymbol{\beta}}_{s-1}, \widehat{\gamma}, \bar{a}_s, \zeta_s, \epsilon_s) \to p_i(\boldsymbol{\beta}, \gamma, \bar{a}_\infty, \kappa_g, \epsilon_\infty),$$

as $s \to \infty$, where $\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$. Moreover, $p_i(\widehat{\boldsymbol{\beta}}_{s-1}, \widehat{\gamma}, \bar{a}_s, \zeta_s, \epsilon_s)$ is bounded by $\boldsymbol{v}^\top \mathbb{E}_X(\boldsymbol{w}_s \boldsymbol{w}_s^\top)\boldsymbol{v} \leq d\|\boldsymbol{v}\|_2^2 L_w^2$, then by Lemma 4 in Chen et al. (2021), we have

$$\begin{aligned}\widehat{\eta}_t^2 \xrightarrow{p} \sigma_i^2 p_i(\boldsymbol{\beta}, \gamma, \bar{a}_\infty, \kappa_g, \epsilon_\infty) = \sigma_i^2 \Bigg\{ \frac{\epsilon_\infty}{2} \int (\boldsymbol{v}_1^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_\infty^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x} v_2 \bar{a}_\infty)\, d\mathcal{P}_X \\ + (1-\epsilon_\infty) \int_{\mathcal{X}_1} (\boldsymbol{v}_1^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{v}_1 + v_2^2 \bar{a}_\infty^2 + 2\boldsymbol{v}_1^\top \boldsymbol{x} v_2 \bar{a}_\infty)\, d\mathcal{P}_X \Bigg\},\end{aligned}$$

where $\mathcal{X}_1 = \{\boldsymbol{x} : \boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g \gamma \geq 0\}$ and $\mathcal{X}_0 = \{\boldsymbol{x} : \boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \kappa_g \gamma < 0\}$. To move $\boldsymbol{v}$ out of the integral, we can rewrite the above equation as:

$$\widehat{\eta}_t^2 \xrightarrow{p} \sigma_i^2 \boldsymbol{v}^\top \begin{pmatrix} \frac{\epsilon_\infty}{2} \int \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X & \bar{a}_\infty \left[\frac{\epsilon_\infty}{2} \int \boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}^\top d\mathcal{P}_X \right] \\ \bar{a}_\infty \left[\frac{\epsilon_\infty}{2} \int \boldsymbol{x}\, d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}\, d\mathcal{P}_X \right] & \bar{a}_\infty^2 \left[\frac{\epsilon_\infty}{2} + (1 - \frac{\epsilon_\infty}{2}) \mathbb{P}(\boldsymbol{x} \in \mathcal{X}_i) \right] \end{pmatrix} \boldsymbol{v}.$$

Finally, by Central Limit Theorem, we have

$$\frac{1}{\sqrt{t}} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\}\, \boldsymbol{w}_s e_s \xrightarrow{D} \mathcal{N}_d(\boldsymbol{0}, G_i),$$

with

$$G_i = \sigma_i^2 \begin{pmatrix} \frac{\epsilon_\infty}{2} \int \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X & \bar{a}_\infty \left[\frac{\epsilon_\infty}{2} \int \boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}^\top d\mathcal{P}_X \right] \\ \bar{a}_\infty \left[\frac{\epsilon_\infty}{2} \int \boldsymbol{x}\, d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}\, d\mathcal{P}_X \right] & \bar{a}_\infty^2 \left[\frac{\epsilon_\infty}{2} + (1 - \frac{\epsilon_\infty}{2}) \mathbb{P}(\boldsymbol{x} \in \mathcal{X}_i) \right] \end{pmatrix}.$$

2. Secondly, we will show the limit of the second moment term. By Lemma 6 in Chen et al. (2021), for $\forall \boldsymbol{v} = (\boldsymbol{v}_1^\top, v_2)^\top \in \mathbb{R}^d$, we need to find the limit of

$$\widehat{\xi}_t = \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\}\, \boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v}.$$

Let $\boldsymbol{x} \sim \mathcal{P}_X$ and define $\tilde{\boldsymbol{w}}$ as

$$\tilde{\boldsymbol{w}} = \begin{cases} (\boldsymbol{x}^\top, 1)^\top, & \text{if } v_2 > 0 \\ (\boldsymbol{x}^\top, 0)^\top, & \text{if } v_2 \leq 0 \end{cases}.$$

Then for any $h > 0$ and each s, we have

$$\mathbb{P}(\mathbb{I}(a_s = i)\boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v} > h) \leq \mathbb{P}(\boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v} > h) = \mathbb{P}(\|\boldsymbol{v}^\top \boldsymbol{w}_s\|_2^2 > h) \leq \mathbb{P}(\|\boldsymbol{v}^\top \tilde{\boldsymbol{w}}\|_2^2 > h),$$

and $\mathbb{E}(\boldsymbol{v}^\top \tilde{\boldsymbol{w}} \tilde{\boldsymbol{w}}^\top \boldsymbol{v}) \leq \boldsymbol{v}^\top \boldsymbol{1}\boldsymbol{1}^\top \boldsymbol{v} L_w^2 < \infty$. Then by Theorem 2.19 from Hall & Heyde (1980), we have

$$\frac{1}{t} \sum_{s=1}^{t} \left\{ \mathbb{I}\{a_s = i\}\, \boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v} - \mathbb{E}(\mathbb{I}(a_s = i)\boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v} \middle| \mathcal{H}_{s-1}) \right\} = \widehat{\xi}_t - \frac{1}{\sigma_i^2} \widehat{\eta}_t^2 \xrightarrow{p} 0.$$

According to the results from the first part, we have $\widehat{\xi}_t \xrightarrow{p} p_i(\boldsymbol{\beta}, \gamma, \bar{a}_\infty, \kappa_g, \epsilon_\infty)$. Using Lemma 6 in Chen et al. (2021) and Continuous Mapping Theorem, we derive

$$\left(\frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\{a_s = i\}\, \boldsymbol{v}^\top \boldsymbol{w}_s \boldsymbol{w}_s^\top \boldsymbol{v} \right)^{-1} \xrightarrow{p} \sigma_i^2 G_i^{-1}. \tag{11}$$

3. Combining results from the first two parts and using Slutsky's Theorem, we have

$$\sqrt{t}(\widehat{\boldsymbol{\theta}}_{i,t} - \boldsymbol{\theta}_i) \xrightarrow{D} \sigma_i^2 G_i^{-1} \mathcal{N}_d(\mathbf{0}, G_i) G_i^{-1} \sigma_i^2 = \mathcal{N}_d(\mathbf{0}, \sigma_i^4 G_i^{-1}) = \mathcal{N}_d(\mathbf{0}, S_i),$$

with

$$S_i = \sigma_i^2 \begin{pmatrix} \frac{\epsilon_\infty}{2} \int \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}\boldsymbol{x}^\top d\mathcal{P}_X & \bar{a}_\infty \left[\frac{\epsilon_\infty}{2} \int \boldsymbol{x}^\top d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}^\top d\mathcal{P}_X \right] \\ \bar{a}_\infty \left[\frac{\epsilon_\infty}{2} \int \boldsymbol{x}\, d\mathcal{P}_X + (1-\epsilon_\infty) \int_{\mathcal{X}_i} \boldsymbol{x}\, d\mathcal{P}_X \right] & \bar{a}_\infty^2 \left[\frac{\epsilon_\infty}{2} + (1 - \frac{\epsilon_\infty}{2}) \mathbb{P}(\boldsymbol{x} \in \mathcal{X}_i) \right] \end{pmatrix}^{-1}.$$

4. Finally, we will show the consistency of variance estimator

$$\frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \widehat{e}_s^2}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} \left(\frac{1}{t} \sum_{s=1}^t \mathbb{I}\{a_s = i\} \boldsymbol{w}_s \boldsymbol{w}_s^\top \right)^{-1}.$$

Based on (11), it is sufficient to show

$$\widehat{\sigma}_i^2 := \frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \widehat{e}_s^2}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} = \frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \left\{ \boldsymbol{w}_s^\top (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t}) + e_s \right\}^2}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} \xrightarrow{p} \sigma_i^2.$$

We will expand the quadratic term and analyze each of the three resulting components,

$$\widehat{\sigma}_i^2 := \frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \left[\left\{ \boldsymbol{w}_s^\top (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t}) \right\}^2 + 2\boldsymbol{w}_s^\top (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t}) e_s + e_s^2 \right]}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}}.$$

By Corollary 4.1, the first term

$$(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t})^\top \frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \boldsymbol{w}_s \boldsymbol{w}_s^\top}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} (\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t}) \leq L_w^2 \|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t}\|_2^2 \xrightarrow{p} 0.$$

Since $\mathbb{I}\{a_s = i\} \boldsymbol{w}_s e_s = \mathbb{I}\{a_s = i\} \boldsymbol{w}_s e_s^{(i)}$, where $e_s^{(i)}$ represents i.i.d. random noise from $\mathcal{P}_{ei}$, and by Corollary 4.1, $\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t} \xrightarrow{p} 0$, the estimation error converges to zero in probability. Furthermore, by Lemma 1 in Chen et al. (2021), we have

$$\frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \boldsymbol{w}_s e_s^{(i)}}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} \xrightarrow{p} \mathbf{0}.$$

Thus, the second term

$$2(\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{i,t})^\top \frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} \boldsymbol{w}_s e_s^{(i)}}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} \xrightarrow{p} 0.$$

with $\widehat{\boldsymbol{\theta}}_{i,t} \xrightarrow{p} \boldsymbol{\theta}_i$ by Corollary 4.1. Finally, by Weak Law of Large Numbers, we have the last term

$$\frac{\sum_{s=1}^t \mathbb{I}\{a_s = i\} (e_s^{(i)})^2}{\sum_{s=1}^t \mathbb{I}\{a_s = i\}} \xrightarrow{p} \mathbb{E}(e_s^{(i)})^2 = \sigma_i^2.$$

The proof is hence completed by combining the limit of the three terms. □

E.6 Regret Bound

Proof. Because $R_2(T)$ can be trivially derived by using the same argument in Chen et al. (2021), we focus on proving $R_1(T)$.

The regret $R_1(T)$ can be bounded by

$$R_1(T) \leq \sum_{t=1}^{T-g(T)-1} \mathbb{E}\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\}$$
$$+\sum_{t=T-g(T)}^{T} \mathbb{E}\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t,s\leq T}\frac{1}{g(s)}\right]\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\}.$$

For each term in the second component, we have

$$\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t,s\leq T}\frac{1}{g(s)}\right]\gamma\right| \leq \left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right| + \left|\sum_{s\in\mathcal{A}_t,s>T}\frac{1}{g(s)}\gamma\right|,$$

then we can reorganize and bound $R_1(T)$ as

$$\begin{aligned}
R_1(T) \leq & \sum_{t=1}^{T-g(T)-1} \mathbb{E}\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\} \\
& + \sum_{t=T-g(T)}^{T} \mathbb{E}\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t,}\frac{1}{g(s)}\right]\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\} \\
& + \sum_{t=T-g(T)}^{T} \mathbb{E}\left|\sum_{s\in\mathcal{A}_t,s>T}\frac{1}{g(s)}\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\} \\
= & \sum_{t=1}^{T-g(T)-1} \mathbb{E}\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\} \\
& + \sum_{t=1}^{T} \mathbb{E}\left|\sum_{s\in\mathcal{A}_t,s>T}\frac{1}{g(s)}\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\}.
\end{aligned}$$

With Assumption 4.4, i.e. $\kappa_g = \lim_{t\to\infty}\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}$, then $\left|\sum_{s\in\mathcal{A}_t,s>T}\frac{1}{g(s)}\gamma\right|$ is bounded by some constant, which we denote as ξ_g. Then, combining the first two components, we have

$$\begin{aligned}
R_1(T) \leq & \sum_{t=1}^{T} \mathbb{E}\left|\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right| \mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\} \\
& + \xi_g \sum_{t=T-g(T)}^{T} \mathbb{E}\left\{\mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right\}\right\},
\end{aligned}$$

where the first component is $\mathcal{O}(\sum_{t=1}^{T}\epsilon_t)$ by the same argument in Chen et al. (2021). Then, we need to find the rate of the second part, which can be separated into regret from exploration

$$\eta_1 = \xi_g \sum_{t=T-g(T)}^{T} \mathbb{E}\left\{\mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1}-\widehat{\boldsymbol{\beta}}_{0,t-1})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1}\right)\right\}\right\},$$

and regret from estimation

$$\eta_2 = \xi_g \sum_{t=T-g(T)}^{T} \mathbb{E}\left|\mathbb{I}\left(\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1}-\widehat{\boldsymbol{\beta}}_{0,t-1})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1}\right) \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right|.$$

Since $\mathbb{E}\left\{\mathbb{I}\left\{a_t \neq \mathbb{I}\left(\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1}-\widehat{\boldsymbol{\beta}}_{0,t-1})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1}\right)\right\}\right\} = \epsilon_t/2$, η_1 is bounded by $\xi_g\sum_{t=T-g(T)}^{T}\epsilon_t$, which is dominated by $\sum_{t=1}^{T}\epsilon_t$. Use the fact that

$$\begin{aligned}
&\left|\mathbb{I}\left(\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{1,t-1}-\widehat{\boldsymbol{\beta}}_{0,t-1})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\widehat{\gamma}_{t-1}\right)\neq\mathbb{I}\left(\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right)\right| \\
\leq&\mathbb{I}\left\{\left|\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right](\widehat{\gamma}_{t-1}-\gamma)\right|>\left|\boldsymbol{x}_t^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|\right\},
\end{aligned}$$

where $\boldsymbol{\beta}\equiv\boldsymbol{\beta}_1-\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_{t-1}\equiv\widehat{\boldsymbol{\beta}}_{1,t-1}-\widehat{\boldsymbol{\beta}}_{0,t-1}$. Then η_2 is bounded by

$$\eta_2\leq\xi_g\sum_{T-g(T)}^{T}\mathbb{E}\left\{\mathbb{I}\left\{\left|\boldsymbol{x}_t^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right](\widehat{\gamma}_{t-1}-\gamma)\right|>\left|\boldsymbol{x}_t^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|\right\}\right\}.$$

We then split η_2 into two parts:

$$\begin{aligned}
J_1=&\sum_{t=T-g(T)}^{T}\int\mathbb{I}\left\{0<\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|<T^{-\frac{1}{4}}\right\} \\
&\times\mathbb{I}\left\{\left|\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right](\widehat{\gamma}_{t-1}-\gamma)\right|>\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|\right\}d\mathcal{P}_X, \\
J_2=&\sum_{t=T-g(T)}^{T}\int\mathbb{I}\left\{\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|>T^{-\frac{1}{4}}\right\} \\
&\times\mathbb{I}\left\{\left|\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right](\widehat{\gamma}_{t-1}-\gamma)\right|>\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|\right\}d\mathcal{P}_X.
\end{aligned}$$

Following Assumption 4.5,

$$\begin{aligned}
J_1\leq&\sum_{t=T-g(T)}^{T}\int\mathbb{I}\left\{0<\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|<T^{-\frac{1}{4}}\right\}d\mathcal{P}_X \\
\leq&\sum_{t=T-g(T)}^{T}MT^{-\frac{1}{4}}=\mathcal{O}(g(T)T^{-\frac{1}{4}}).
\end{aligned}$$

For J_2, we have

$$\begin{aligned}
J_2\leq&\sum_{t=T-g(T)}^{T}\int\mathbb{I}\left\{\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|>T^{-\frac{1}{4}}\right\}\frac{\left|\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right](\widehat{\gamma}_{t-1}-\gamma)\right|}{\left|\boldsymbol{x}^\top\boldsymbol{\beta}+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right]\gamma\right|}d\mathcal{P}_X \\
\leq&T^{\frac{1}{4}}\sum_{t=T-g(T)}^{T}\int\left|\boldsymbol{x}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta})+\left[\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}\right](\widehat{\gamma}_{t-1}-\gamma)\right|d\mathcal{P}_X \\
\leq&T^{\frac{1}{4}}\sum_{t=T-g(T)}^{T}C_1\left\|\left((\widehat{\boldsymbol{\beta}}_t-\boldsymbol{\beta})^\top,\widehat{\gamma}_{1,t}-\gamma_1\right)^\top\right\|_1,
\end{aligned}$$

where C_1 is a positive constant related to L_x and the upper bound of $\sum_{s\in\mathcal{A}_t}\frac{1}{g(s)}$. By Theorem 4.2, Because there is no overlap between the data used for deriving $(\widehat{\boldsymbol{\beta}}_{1,t}^\top,\widehat{\gamma}_1)^\top$ and $\widehat{\boldsymbol{\beta}}_{0,t}$, the joint distribution of $(\widehat{\boldsymbol{\beta}}_{1,t}^\top,\widehat{\boldsymbol{\beta}}_{0,t}^\top,\widehat{\gamma}_1)^\top$ is asymptotically normal, which is

$$\sqrt{t}\begin{pmatrix}\widehat{\boldsymbol{\beta}}_{1,t}-\boldsymbol{\beta}_1\\ \widehat{\boldsymbol{\beta}}_{0,t}-\boldsymbol{\beta}_0\\ \widehat{\gamma}_{1,t}-\gamma_1\end{pmatrix}\xrightarrow{D}\mathcal{N}_d(\mathbf{0},Q),$$

Denote $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ as $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{\beta}}_{1,t} - \widehat{\boldsymbol{\beta}}_{0,t}$ as $\widehat{\boldsymbol{\beta}}_t$. Then, by the Slutsky's Theorem, we are able to derive the asymptotic distribution of $(\widehat{\boldsymbol{\beta}}_t^\top, \widehat{\gamma}_1)^\top$, which is given by

$$\sqrt{t}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta} \\ \widehat{\gamma}_{1,t} - \gamma_1 \end{pmatrix} \xrightarrow{D} \mathcal{N}_d(\mathbf{0}, H),$$

where Q and H are two symmetric matrices and can be derived trivially according to S_0 and S_1. Then we have $\left((\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta})^\top, \widehat{\gamma}_{1,t} - \gamma_1\right)^\top = \mathcal{O}_p(t^{-\frac{1}{2}})$, so $\left\|\left((\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta})^\top, \widehat{\gamma}_{1,t} - \gamma_1\right)^\top\right\|_1 = \mathcal{O}_p(t^{-\frac{1}{2}})$, thus the upper bound becomes

$$J_2 \leq C_1 T^{\frac{1}{4}} \left(\mathcal{O}_p(T^{\frac{1}{2}} - (T - g(T))^{\frac{1}{2}})\right).$$

By applying

$$T^{\frac{1}{2}} - (T - g(T))^{\frac{1}{2}} = \frac{T - (T - g(T))}{T^{\frac{1}{2}} + (T - g(T))^{\frac{1}{2}}} = \frac{g(T)}{T^{\frac{1}{2}} + (T - g(T))^{\frac{1}{2}}},$$

we have $\mathcal{O}_p(T^{\frac{1}{2}} - (T - g(T))^{\frac{1}{2}}) = \mathcal{O}_p(g(T)T^{-\frac{1}{2}})$. Thus $J_2 = \mathcal{O}_p(g(T)T^{-\frac{1}{4}})$. Then $\eta_2 = J_1 + J_2 = \mathcal{O}_p(g(T)T^{-\frac{1}{4}})$. Therefore, $R_1(T) \leq \eta_1 + \eta_2 = \mathcal{O}_p(\sum_{t=1}^T \epsilon_t + g(T)T^{-\frac{1}{4}})$. Besides, under Assumption 4.3, considering the force pulls regret $\mathcal{O}(\sqrt{T})$ in practice, which is dominated by $\mathcal{O}\left(\sum_{t=1}^T \epsilon_t\right)$, we obtain the regret bound $\mathcal{O}(\sum_{t=1}^T \epsilon_t + g(T)T^{-\frac{1}{4}})$. □

E.7 Optimal Policy for Other Working Models

Here, we demonstrate how to derive the optimal policy for alternative working models, as discussed in the Discussion section, using backward inductive reasoning (see, e.g., Chakraborty & Murphy, 2014).

1. $\mu(\boldsymbol{x}_t, \bar{a}_t, a_t) = \mathbb{E}(y_t|\boldsymbol{x}_t, \bar{a}_t, a_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \bar{a}_t \gamma_0 + a_t \bar{a}_t (\gamma_1 - \gamma_0)$.

Solution.

$$\sum_{t=1}^T \mu(\boldsymbol{x}_t, \bar{a}_t, a_t) = \sum_{t=1}^T \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + \bar{a}_t \gamma_0 + a_t \bar{a}_t (\gamma_1 - \gamma_0).$$

Using backward induction, we assume that the optimal decisions $a_1, a_2, \ldots, a_{T-1}$ have aleady been determined. The optimal action a_T^* is then selected to maximize the cumulative reward:

$$\sum_{t=1}^{T-1} \mu(\boldsymbol{x}_t, \bar{a}_t, a_t) + \mu(\boldsymbol{x}_T, \bar{a}_T, a_T).$$

Because $a_1, a_2, \ldots, a_{T-1}$ are determined, the corresponding interference actions $\bar{a}_1, \bar{a}_2, \ldots, \bar{a}_{T-1}$, $\bar{a}_T$ are also fixed. We denote $\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ and $\gamma = \gamma_1 - \gamma_0$ for simplicity. So the optimal action at time T is given by,

$$a_T^* = \arg\max_{a_T} a_T \boldsymbol{x}_T^\top \boldsymbol{\beta} + a_T \bar{a}_T \gamma = \mathbb{I}\left\{\boldsymbol{x}_T^\top \boldsymbol{\beta} + \bar{a}_T \gamma \geq 0\right\}.$$

Then we will select a_{T-1}^* to maximize the expected outcome that would result from choosing the option at time T optimally given the history available at that point. Suppose we have determined the decisions $a_1, a_2, \ldots, a_{T-2}$. We plug in $\bar{a}_T = \frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2} a_s\right) + a_{T-1}\right]$ and $a_T^* = \mathbb{I}\left\{\boldsymbol{x}_T^\top \boldsymbol{\beta} + \bar{a}_T \gamma \geq 0\right\} = \mathbb{I}\left\{\boldsymbol{x}_T^\top \boldsymbol{\beta} + \frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2} a_s\right) + a_{T-1}\right]\gamma \geq 0\right\}$. Then we need to solve:

$$\arg\max_{a_{T-1}} \sum_{t=1}^{T-2} \mu(\boldsymbol{x}_t, \bar{a}_t, a_t) + \mu(\boldsymbol{x}_{T-1}, \bar{a}_{T-1}, a_{T-1})$$
$$+ \mu\left(\boldsymbol{x}_T, \frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2} a_s\right) + a_{T-1}\right], \mathbb{I}\left\{\boldsymbol{x}_T^\top \boldsymbol{\beta} + \frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2} a_s\right) + a_{T-1}\right]\gamma \geq 0\right\}\right).$$

Removing all terms that are unrelated to a_{T-1}, we obtain:

$$\begin{aligned}
&\arg\max_{a_{T-1}} a_{T-1}\left(\boldsymbol{x}_{T-1}^\top\boldsymbol{\beta}+\bar{a}_{T-1}\gamma\right)+\mathbb{I}\left\{\boldsymbol{x}_T^\top\boldsymbol{\beta}+\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+a_{T-1}\right]\gamma\geq 0\right\}\boldsymbol{x}_T^\top\boldsymbol{\beta}\\
&+\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+a_{T-1}\right]\gamma_0\\
&+\mathbb{I}\left\{\boldsymbol{x}_T^\top\boldsymbol{\beta}+\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+a_{T-1}\right]\gamma\geq 0\right\}\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+a_{T-1}\right]\gamma.
\end{aligned}$$

Plugging in $a_{T-1}=1$ and $a_{T-1}=0$ and comparing the results, we obtain the solution:

$$a_{T-1}^*=\mathbb{I}\left\{\tau_1\geq\tau_0\right\}, \tag{12}$$

where τ_1 with $a_{T-1}=1$ and τ_0 with $a_{T-1}=0$. τ_1 is given by,

$$\begin{aligned}
\tau_1=&\boldsymbol{x}_{T-1}^\top\boldsymbol{\beta}+\bar{a}_{T-1}\gamma+\mathbb{I}\left\{\boldsymbol{x}_T^\top\boldsymbol{\beta}+\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+1\right]\gamma\geq 0\right\}\boldsymbol{x}_T^\top\boldsymbol{\beta}\\
&+\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+1\right]\gamma_0\\
&+\mathbb{I}\left\{\boldsymbol{x}_T^\top\boldsymbol{\beta}+\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+1\right]\gamma\geq 0\right\}\frac{1}{g(T)}\left[\left(\sum_{s=T-g(T)}^{T-2}a_s\right)+1\right]\gamma.
\end{aligned} \tag{13}$$

τ_0 is given by,

$$\begin{aligned}
\tau_0=&\mathbb{I}\left\{\boldsymbol{x}_T^\top\boldsymbol{\beta}+\frac{1}{g(T)}\left(\sum_{s=T-g(T)}^{T-2}a_s\right)\gamma\geq 0\right\}\boldsymbol{x}_T^\top\boldsymbol{\beta}\\
&+\frac{1}{g(T)}\left(\sum_{s=T-g(T)}^{T-2}a_s\right)\gamma_0\\
&+\mathbb{I}\left\{\boldsymbol{x}_T^\top\boldsymbol{\beta}+\frac{1}{g(T)}\left(\sum_{s=T-g(T)}^{T-2}a_s\right)\gamma\geq 0\right\}\frac{1}{g(T)}\left(\sum_{s=T-g(T)}^{T-2}a_s\right)\gamma.
\end{aligned} \tag{14}$$

We observe that the optimal action at time $T-1$ is a nested indicator function. Similarly, the optimal actions for times $T-2, \ldots, 1$ can be determined through the same recursive process. However, when the termination time T is unknown, deriving a general form for the optimal policy becomes infeasible. □

2. $\mu(\boldsymbol{x}_t,\bar{a}_t,a_t)=\mathbb{E}(y_t|\boldsymbol{x}_t,\bar{a}_t,a_t)=\boldsymbol{x}_t^\top\boldsymbol{\beta}_0+a_t\boldsymbol{x}_t^\top(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)+\bar{a}_t\gamma_0+\bar{a}_t\boldsymbol{x}_t^\top\boldsymbol{\gamma}$

Solution. *We denote* $\boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$.

$$\begin{aligned}
\sum_{t=1}^{T} \mu(\boldsymbol{x}_t, \bar{a}_t, a_t) &= \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top \boldsymbol{\beta} + \bar{a}_t \gamma_0 + \bar{a}_t \boldsymbol{x}_t^\top \boldsymbol{\gamma} \\
&= \sum_{t=1}^{T} \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \boldsymbol{x}_t^\top \boldsymbol{\beta} + \frac{1}{g(t)} \left[\left(\sum_{s=t-g(t)}^{t-1} a_s \right) \gamma_0 + \left(\sum_{s=t-g(t)}^{t-1} a_s \boldsymbol{x}_s^\top \right) \boldsymbol{\gamma} \right] \\
&= \sum_{t=1}^{T-g(T)-1} \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \left\{ \boldsymbol{x}_t^\top \boldsymbol{\beta} + \left[\sum_{s \in \mathcal{A}_t} \frac{1}{g(s)} \right] \gamma_0 + \left[\sum_{s \in \mathcal{A}_t} \frac{1}{g(s)} \boldsymbol{x}_s^\top \right] \boldsymbol{\gamma} \right\} \\
&\quad + \sum_{t=T-g(T)}^{T} \boldsymbol{x}_t^\top \boldsymbol{\beta}_0 + a_t \left\{ \boldsymbol{x}_t^\top \boldsymbol{\beta} + \left[\sum_{s \in \mathcal{A}_t, s \le T} \frac{1}{g(s)} \right] \gamma_0 + \left[\sum_{s \in \mathcal{A}_t, s \le T} \frac{1}{g(s)} \boldsymbol{x}_s^\top \right] \boldsymbol{\gamma} \right\}.
\end{aligned}$$

The optimal action should be

$$a_t^* = \begin{cases} \mathbb{I}\left\{ \boldsymbol{x}_t^\top \boldsymbol{\beta} + \left[\sum_{s \in \mathcal{A}_t} \frac{1}{g(s)} \right] \gamma_0 + \left[\sum_{s \in \mathcal{A}_t} \frac{1}{g(s)} \boldsymbol{x}_s^\top \right] \boldsymbol{\gamma} \ge 0 \right\}, & \text{if } 1 \le t \le T - g(T) - 1 \\ \mathbb{I}\left\{ \boldsymbol{x}_t^\top \boldsymbol{\beta} + \left[\sum_{s \in \mathcal{A}_t, s \le T} \frac{1}{g(s)} \right] \gamma_0 + \left[\sum_{s \in \mathcal{A}_t, s \le T} \frac{1}{g(s)} \boldsymbol{x}_s^\top \right] \boldsymbol{\gamma} \ge 0 \right\}, & \text{if } T - g(T) \le t \le T \end{cases}. \tag{15}$$

We observe that the optimal action at time t *depends on feature contextual features that are not available* t*. For example, for* $1 \le t \le T - g(T) - 1$*,* a_t^* *is influenced by* x_s *indices satisfying* $s - g(s) \le t \le s - 1$*. If interference scale is fixed as* N*. Then the optimal actions turn to be:*

$$a_t^* = \begin{cases} \mathbb{I}\left\{ \boldsymbol{x}_t^\top \boldsymbol{\beta} + \gamma_0 + \frac{1}{N} \left[\sum_{s=t+1}^{t+N} \boldsymbol{x}_s^\top \right] \boldsymbol{\gamma} \ge 0 \right\}, & \text{if } 1 \le t \le T - N - 1 \\ \mathbb{I}\left\{ \boldsymbol{x}_t^\top \boldsymbol{\beta} + \frac{T-t}{N} \gamma_0 + \frac{1}{N} \left[\sum_{s=t+1}^{T} \boldsymbol{x}_s^\top \right] \boldsymbol{\gamma} \ge 0 \right\}, & \text{if } T - N \le t \le T \end{cases}. \tag{16}$$

□