Contextual Bandits with Budgeted Information Reveal

Kyra Gan
Cornell Tech

Esmaeil Keyvanshokooh
Texas A&M University

Xueqing Liu
Duke-NUS Medical School

Susan Murphy
Harvard University

Abstract

Contextual bandit algorithms are commonly used in digital health to recommend personalized treatments. However, to ensure the effectiveness of the treatments, patients are often requested to take actions that have no immediate benefit to them, which we refer to as *pro-treatment* actions. In practice, clinicians have a limited budget to encourage patients to take these actions and collect additional information. We introduce a novel optimization and learning algorithm to address this problem. This algorithm effectively combines the strengths of two algorithmic approaches in a seamless manner, including 1) an online primal-dual algorithm for deciding the optimal timing to reach out to patients, and 2) a contextual bandit learning algorithm to deliver personalized treatment to the patient. We prove that this algorithm admits a sub-linear regret bound. We illustrate the usefulness of this algorithm on both synthetic and real-world data.

1 INTRODUCTION

In digital health, to ensure the effectiveness of treatments, patients are often requested to take actions that have no immediate benefit to them. We refer to these actions as *pro-treatment actions*. For instance, in personalized addiction treatment, the effectiveness of the treatment may be compromised if patients fail to complete self-reports [Carpenter et al., 2020]. Another scenario arises when utilizing commercial sensors or when there is a need to aggregate data across multiple patients. In this case, data collected from the sensor can exclusively be accessed via cloud servers, implying

that the data-collecting device, such as wearables and electronic toothbrush, may only be able to communicate with the intervention-delivery device, such as smartphones, through the cloud. To ensure the proper delivery of personalized treatments, patients may need to open a dedicated app on their smartphones, enabling the app to retrieve the latest treatment recommendations from the cloud [Trella et al., 2022]. When patients neglect pro-treatment actions, clinicians may resort to *limited*, *costly* nudges, such as clinician follow-ups to encourage compliance.

We are interested in answering the following question: *given a limited budget for expensive nudges for use when patients fail to take pro-treatment actions, when should these nudges be used?* To answer this question, we *reformulate* the problem by introducing *two agents*. The first agent functions as a **recommender**, specifically, a learning agent that uses all the *revealed* patient information up to the current time to recommend the treatment action for the subsequent time step. The second agent, a **revealer**, possesses *current* (or some surrogate of current) and past patient data, often collected by sensors, and determines whether to reveal this information to the recommender, enabling the learning of personalized treatment.

In the commercial sensor example, the recommendations (smart phone notifications) are often made by a cloud-based *reinforcement learning* (RL) algorithm. Although the RL algorithm observes all the information up to the current time, these up-to-date recommendations remain *hidden* to the patient unless they open a dedicated app on their phone (i.e., taking a pro-treatment action). The clinician serves as the revealer, prompting the patient to open the app as a means of revealing information (as depicted in Figure 1, assuming the patient never opens the app on their own). At decision points, the clinician assesses all sensor data and decides whether to contact the patient. If the clinician contacts the patient and the patient opens the mobile application, they receives the most recent recommended notification. Without such actions, the patient receives an "outdated" recommendation based on the information up to the *last reveal*. Therefore,
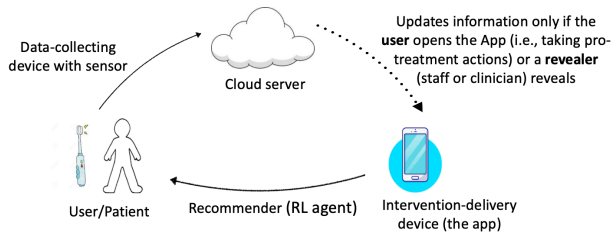
Figure 1: A depiction of the problem breakdown in the electronic toothbrush application, which we extensively investigate in our experiments. In our problem reformulation, it is equally valid to view the recommender as residing on the intervention-delivery device.

equivalently, in this problem, the recommender only has access to the history up to the last reveal. However, upon opening the app, the recommender gains access to the entire sensor data history.

Our Contributions In this work, we provide an algorithm for deciding the "optimal" timing for the revealer to take action when the number of actions that it can take is limited. We focus on the special case where the recommender acts 1) as a *linear contextual bandit* algorithm when the revealer opts to reveal information and 2) as a *multi-armed bandit* (MAB) algorithm when no additional information is revealed to the recommender (due to the missing context). We show that our problem can be decomposed into two parts: 1) an online primal-dual optimization algorithm addressing the decision of the revealer, and 2) a contextual bandit learning algorithm with delayed feedback modeling the decision of the recommender.

In the online primal-dual algorithm, we introduce a novel learning constraint and prove that the value of the objective function of our proposed algorithm, Algorithm 1, is at least $\pi_{\min}(1 - 1/c)$ times that of an offline clairvoyant benchmark, where $\pi_{\min}$ is a problem dependent constant and $1 - 1/c$ is budget dependent constant that approaches $1 - 1/e$ as the budget grows. Furthermore, by introducing the novel learning constraint in the primal-dual algorithm, we are able to separate out the effect of delayed feedback in the bandit learning loss (defined in § 2), removing the dependency of the delayed feedback effect on the context dimension. By combining these two parts, we provide the first UCB-based algorithm (Algorithm 2) that achieves a sublinear regret under suitable choice of parameters.

Related Work Our work is related to three streams of literature: (i) online optimization algorithms, (ii) contextual bandits under resource constraints, and (iii) contextual bandits with delayed feedback. Studies in

(i) typically focus on two arrival settings: stochastic and adversarial. In the stochastic setting, online algorithms either rely on the forecasted arrival pattern using historical data or assume a stochastic arrival pattern [Goel and Mehta, 2008, Karande et al., 2011, Mahdian and Yan, 2011, Zhalechian et al., 2023, Feldman et al., 2009, Jaillet and Lu, 2014, Devanur et al., 2019], while in the adversarial setting, algorithms are robust to possible changes in the arrival pattern [Mehta et al., 2007, Buchbinder et al., 2007, Aggarwal et al., 2011, Keyvanshokooh et al., 2021, Zhalechian et al., 2022, Devanur and Jain, 2012, Liu et al., 2024]. The online primal-dual mechanism is one class of algorithms that leverages the dual program to guide the decisions of the online algorithm [Buchbinder et al., 2009, Keyvanshokooh et al., 2021]. In our work, we introduce a new class of online primal-dual mechanisms with a learning component and incorporate it as an online allocation sub-routine in our proposed framework. We evaluate its performance using a *competitive ratio*, which compares its performance to that of a clairvoyant policy on the worst-case input instance.

Studies in (ii) usually assume that each action consumes a certain amount of resources. Under such resource constraints, previous works have proposed online algorithms for standard MAB [Agrawal and Devanur, 2014, Badanidiyuru et al., 2018, Ferreira et al., 2018], contextual bandit, and other RL methods [Badanidiyuru et al., 2014, Agrawal et al., 2016, Agrawal and Devanur, 2016, Wu et al., 2015, Pacchiano et al., 2021, Cao et al., 2023]. In contrast, in our algorithm, the recommender is required to take an action at each time step regardless of whether it observes the current context. A few works formulate contextual bandit with resource constraints by integrating Thompson sampling (bandit) algorithms with online optimization algorithms [Cheung et al., 2022, Zhalechian et al., 2022]. In comparison, our work provide the first analysis for UCB-based bandit algorithms, and we incorporate a learning component into the online allocation mechanisms while learning only happens in the bandit part of their algorithms. This additional level of learning helps our algorithms achieve a better performance.

Studies in (iii) include both delayed feedback in MAB [Joulani et al., 2013, Bistritz et al., 2019, Pike-Burke et al., 2018] and contextual bandit [Zhou et al., 2019, Vernade et al., 2020, Keyvanshokooh et al., 2024], where the delay patterns mostly fall into *bounded* delays or *stochastic* delays. However, in our problem, the delayed patterns are structured: unless the revealer takes an action, the recommender received no additional information. While applying algorithms from (iii) in our problem setting is feasible, our algorithm offers a significantly stronger theoretical guarantee. The order of

our regret bound is tight (optimal) up to a logarithmic factor [Chu et al., 2011]. Also, the delayed feedback only impacts our regret bound by an *additive* factor of $\sum_{t=1}^{T} \beta_t$, which is of order $\mathcal{O}(\sqrt{T})$. Therefore, unlike Zhou et al. [2019] and Vernade et al. [2020], our theoretical result removes the dependency of the delayed feedback effect on the context dimension d.

2 PROBLEM FORMULATION

We start with the worst-case setting where the recommender *never* observes any additional information *unless* the revealer takes an action at each time step. When patients occasionally take pro-treatment actions on their own, we expect the relative performance of our algorithm (with respect to the benchmark algorithms in Section 5) to stay the same. In this section, we first introduce the contextual bandit problem and then discuss the setup of each agent. Lastly, we provide an overview of our proposed framework.

Linear Contextual Bandit Let $\mathcal{S} = \{1, ..., K\}$ denote the set of contexts. Given time horizon T, at each time $t \in [T]$, a context S_t arrives. We assume that the contexts are drawn i.i.d. from a known distribution $\mathbf{p}^*$, where $\mathbf{p}_k^* := \mathbb{P}(S_t = k)$. (See Section 5, and Appendix F for the situation where $\mathbf{p}^*$ is unknown.) However, the ordering of the realized contexts can be *adversarially* chosen, that is, the adversary can choose the ordering in which the contexts appear.[1]

Let $\mathcal{A}$ denote the set of discrete actions that can be taken by the recommender. The reward X_t under context S_t and action $A_t \in \mathcal{A}$ is generated according to $X_t = \langle \theta_*, \phi(S_t, A_t) \rangle + \eta_t$, where $\theta_* \in \mathbb{R}^d$ is an *unknown* true reward parameter, $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is a *known* feature mapping, and the noise η_t is conditional mean-zero 1-sub-Gaussian.

Recommender For a given patient, when the recommender has access to the context S_t, it takes action according to a contextual bandit algorithm. When the recommender does *not* observe the current context S_t, it takes actions by treating the bandit problem as a MAB, where the expected reward of each action is now weighted by the context distribution. We elaborate on this structure in Section 3, and describe our UCB-based bandit algorithms in Section 4 (Algorithm 2). We note that this problem structure does *not* affect the reward generating process, but rather it affects whether expected reward averages out over context or not.

Revealer The revealer is given an expected budget of B information reveals to the recommender throughout the horizon T. We assume that $B > 2|\mathcal{A}|$ for technical ease. At each time, the binary decision variable for the revealer is $O_t \in \{0, 1\}$. Consider the general case where the revealer only observes part of the context. Namely, we can partition each state into two components: $S_t = [S_t^1, S_t^2]$. Let S_t^1 be the part of the state that is *always* observed by the revealer at each time t, and let S_t^2 be the part of the state that can *only* be observed when the revealer takes the action $O_t = 1$. Let $\ell(t)$ be the time of the last reveal. At each decision time t, the revealer observes the history $\mathcal{H}_t^{\text{rev}} = \{A_1, ..., A_{t-1}, X_1, ... X_{t-1}, O_1, ..., O_{t-1}, S_1, ..., S_{\ell(t)}, S_{\ell(t)+1}^1, ..., S_t^1\}$, and *decides* the revealing probability o_t; then $O_t \sim \text{Bernoulli}(o_t)$. If $O_t = 1$, the revealer additionally observes $\{S_{\ell(t)+1}^2, ..., S_t^2\}$ and the recommender observes $\mathcal{H}_t^{\text{rec}} = \{A_1, ..., A_{t-1}, X_1, ..., X_{t-1}, O_1, ..., O_{t-1}, S_1, ..., S_t\}$. Otherwise, the recommender observes the history up to time $\ell(t)$, $\mathcal{H}_{\ell(t)}^{\text{rec}}$. We highlight the key information asymmetry here lies in the fact that the revealer can always observe S_t^1, but the recommender cannot. The budget constraint requires $\sum_{t=1}^{T} o_t \leq B$. For ease of exposition, we will focus on the special case where the revealer observes the entire context at time t, i.e., $S_t^1 = S_t$, from now on. Further discussion on how our theoretical guarantees apply to the above mentioned general setting is included in Appendix B.

Framework Overview and Regret Decomposition Given that the number of actions that the revealer can take is limited, our objective is to develop a data-driven optimization and learning framework that can 1) decide the optimal timing for the revealer to reveal, and 2) learn the optimal treatment for the recommender. We achieve the former by designing an online primal-dual algorithm with a novel learning constraint (Algorithm 1) and achieve the latter by applying the UCB algorithm (Algorithm 2) which uses the online primal-dual algorithm (Algorithm 1) as a subroutine.

There are two main sources of uncertainty in this problem: (1) the unknown reward parameter, θ_*, and (2) the *sequence of future context arrivals*, $\{s_t, ..., s_T\}$. We discuss unknown context distribution $\mathbf{p}^*$ in Appendix F. We evaluate the performance of our algorithm using an *offline clairvoyant* benchmark, where both the revealer and the recommender know the reward distribution, θ_* and additionally, the revealer knows the entire context arrival sequence $\{s_1, ..., s_T\}$. Note that *no* algorithm can ever achieve this performance in practice since future contexts are inherently unknown.

We introduce a novel regret analysis to assess the theoretical performance of our algorithm in relation to

---

[1]As seen in Section 3, the performance of our proposed online algorithm depends on the context arrival *sequence*.

the clairvoyant problem. Our analysis seamlessly combines a *competitive ratio* bound for bounding the suboptimality gap of the revealer with a *regret* bound of the recommender. This integration necessitates 1) defining an *auxiliary problem* and 2) using a *bridging argument*. In the auxiliary problem, we assume knowledge of the unknown model parameter (θ_*), but the *future* context arrival sequence is *unknown*. Specifically, the auxiliary problem simulates the *online* version of the clairvoyant problem where the contexts arrive sequentially. We note that this terminology also has been used in the existing literature (see Cheung et al. [2022]).

Let V^{AUX} and V^{ALG} be the respective value functions of Algorithm 2 when θ_* is known and unknown, and let V^{CLV} be the value function of the clairvoyant problem; see Section 3 and Section 4 for formal definitions. We decompose the regret by establishing the following bridging argument:

$$\begin{aligned}\text{Regret}_T &\leq \mathbb{E}\left[V^{\text{CLV}}\right] - \mathbb{E}\left[V^{\text{ALG}}\right] \\ &= \underbrace{\mathbb{E}\left[V^{\text{CLV}} - V^{\text{AUX}}\right]}_{\text{Information Reveal Loss}} + \underbrace{\mathbb{E}\left[V^{\text{AUX}} - V^{\text{ALG}}\right]}_{\text{Bandit Learning Loss}},\end{aligned}$$

In the above decomposition, the first expression represents the loss due to the optimality gap of the information revealing mechanism for solving the auxiliary problem. The expectation is taken over the stochasticity of the algorithm, as the reward parameter is known in both problems. The second represents the loss due to contextual bandit learning, i.e., learning the unknown reward. The expectation is taken over the stochasticity in both the algorithm and the environment.

3 BOUNDING INFORMATION REVEAL LOSS

In this section, we first formally introduce the clairvoyant problem. Then, by developing an online primal-dual approach for solving the clairvoyant problem in an online fashion, we provide a feasible solution to the auxiliary problem. Finally, we provide an upper bound on the information reveal loss.

Clairvoyant Problem Recall that in the clairvoyant problem, both the revealer and recommender know θ_*, and the revealer additionally knows the entire context *realized* sequence $\{s_1, ..., s_T\}$. The optimal strategy for the recommender is to take the optimal action corresponding to context s_t when s_t is observed, and to take the action with the highest expected (where the expectation is taken over the context distribution) reward when s_t is *not* observed. The former happens when the revealer takes action $O_t = 1$ at time t, corresponding to *revealing* the history $\mathcal{H}_t^{\text{rev}}$ to the recommender, and the latter happens when the revealer decides *not to*

take action at time t. Let $u_{s_t}^* = \max_{a \in \mathcal{A}} \langle \theta_*, \phi(s_t, a) \rangle$, and $v^* = \max_{a \in \mathcal{A}} \langle \theta_*, \bar{\phi}(a) \rangle$, where $\bar{\phi}(a)$ is the weighted feature mapping, i.e., $\bar{\phi}(a) = \sum_{k=1}^{K} \phi(k, a) \mathbf{p}_k^*$.

Using this optimal strategy for the recommender, a natural objective of the revealer is to maximize the expected reward collected throughout the horizon: $\max_{o_t} \sum_{t=1}^{T} o_t \cdot u_{s_t}^* + (1 - o_t) \cdot v^*$. By removing the constant v^*, we obtain the following formulation for the clairvoyant problem (CLV):

$$\left\{ \max_{o_t} \sum_{t=1}^{T} o_t \cdot (u_{s_t}^* - v^*) : \sum_{t=1}^{T} o_t \leq B, o_t \in [0, 1], \\ \forall t \in [T]. \right\} \qquad (CLV)$$

The optimal policy of the revealer in (CLV) is first to select the contexts that yield more reward than v^*, i.e., with positive $u_{s_t}^* - v^*$, and second set $o_t = 1$ for the top B contexts that have the highest expected reward, $u_{s_t}^*$'s. Thus, V^{CLV} is the sum of the expected optimal rewards where the expectation is taken over the decision variable O_t, i.e., $\mathbb{E}[V^{\text{CLV}}] = \max_{a \in \mathcal{A}} \sum_{t=1}^{T} (o_t^{\text{CLV}} \langle \theta_*, \phi(s_t, a) \rangle + (1 - o_t^{\text{CLV}}) \langle \theta_*, \bar{\phi}(a) \rangle)$, where the sequence of $\{o_t^{\text{CLV}}\}_{t \in [T]}$ is the solution to (CLV). Note that *without* knowledge of current context S_t, each decision point becomes identical to the revealer, resulting in a trivial optimal solution for the revealer – randomly selecting B times to reveal $\mathcal{H}_t^{\text{rev}}$, highlighting the importance of information asymmetry between revealer and recommender.

Remark 1 (Objective function of CLV). *We note that if there are not enough contexts s_t's with a positive $u_{s_t}^* - v^*$, then we do not use the entire budget B. We note that our objective function is suitable for the low-budget regime, i.e., when B is less than or equal to the number of contexts with positive $u_{s_t}^* - v^*$. For larger B's, one could remove $-o_t v^*$ from the objective function, and the rest of the result still holds.*

An alternative objective function in this problem is $\max_{o_t} \sum_{t=1}^{T} o_t u_{s_t}^*$. Indeed, when the budget is low, these two objectives yield the same optimal solution to the clairvoyant problem. However, as we will see in our online primal-dual algorithm (Algorithm 1), when we do not have access to the future *context arrival sequence, v^* serves as a regularization term for spending the budget B (by ignoring the contexts that yield negative $u_{s_t}^* - v^*$), yielding better algorithmic performance.*

As a direct consequence of Remark 1, we assume that the budget is not too large when compared with the horizon length T:

Assumption 1. *We assume that $B = \mathcal{O}(\sqrt{T})$.*

In (CLV), the optimal strategy of the revealer depends on the entire *realized* context arrival sequence, including

the future arrivals $\{s_{t+1}, ..., s_T\}$. While no algorithm in practice can achieve this performance, (CLV) has two critical advantages: 1) it provides an *upper bound* for the optimal solution to the oracle problem: the oracle problem can be viewed as (CLV) with the additional constraint that the context sequence is observed up to time t, $\{s_1, ..., s_t\}$; 2) it naturally provides insight into how to incorporate online primal-dual mechanisms.

Auxiliary Problem The *auxiliary problem* is an *online* version of (CLV), where the contexts arrive sequentially, and the decision of the revealer should be made to hedge against the adversarial context arrival sequence in the future. In the auxiliary problem, both agents know θ_*, and neither has access to the future context arrival sequence (which might as well be *adversarial*). We develop an *online primal-dual* algorithm (Algorithm 1) to provide a feasible solution for the revealer in the auxiliary problem. We rigorously analyze it using the *competitive ratio analysis*.

Modified Clairvoyant Problem Intuitively, the revealing probability o_t at each time step should depend on both 1) the budget that we have spent so far and 2) the rate at which we learn the reward and context distributions. However, as it currently stands, the dual of (CLV) lacks a mechanism to connect the quality of the estimates that the *recommender* has at time t to the revealing decision o_t. Ideally, we would like o_t to increase as the time since the last reveal increases.

To solve this technical challenge, we next incorporate a novel **learning constraint**, Constraint (1). In Appendix C, we provide a road map for deriving this constraint. We first provide an algorithm that only takes the budget into account (Algorithm C.1), and then provide its theoretical guarantee (Proposition C.1).

The online primal-dual algorithm that we will develop in this section serves as a *subroutine* in our bandit learning algorithm. The bandit algorithm provides estimates of $u_{s_t}^*$ and v^* to the online primal-dual algorithm. We make the following critical observation: at each time t, the revealer has access to both $\mathcal{H}_t^{\mathrm{rev}}$ and $\mathcal{H}_t^{\mathrm{rec}}$, the revealer can calculate both the recommender's optimal action, $\hat{a}_t$, if the revealer were *not* to reveal $\mathcal{H}_t^{\mathrm{rev}}$ ($O_t = 0$), and the optimal action, $\tilde{a}_t$, when $\mathcal{H}_t^{\mathrm{rev}}$ is revealed ($O_t = 1$). We describe the calculation of the above in detail in Section 4. Next, we introduce a constraint to force the revealing probability to increase when the estimated optimal treatment differs between the two agents, i.e., $\hat{a}_t \neq \tilde{a}_t$, *and* the distance between the weighted feature mappings, $\|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2$, (recall $\bar{\phi}(a) = \sum_{k=1}^K \phi(k, a) \mathbf{p}_k^*$) is large:

$$\|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2 \, \mathbb{1}\,(\hat{a}_t \neq \tilde{a}_t)\,(1 - o_t)$$
$$\leq \beta_t(\mathcal{H}_t^{\mathrm{rev}}, \mathcal{H}_t^{\mathrm{rec}}), \; \forall t \in [T], \qquad (1)$$

where $\{\beta_t(\mathcal{H}_t^{\mathrm{rev}}, \mathcal{H}_t^{\mathrm{rec}})\}_{t=1}^T$ is a sequence of positive constants that can be *initialized adaptively* by the expert and *auto-adjusted* by our algorithm using the histories, to guarantee the feasibility of (CLV) with the above constraint. To ease notation, we abbreviate the β's using $\beta_1, ..., \beta_T$ from now on.

Appendix D.1 includes the updated primal problem. Let y, z_t's, and e_t's be the dual variables. We have the resulting dual of the updated primal problem:

$$\min_{y, x_t, z_t} \quad By + \sum_{t=1}^T z_t \qquad\qquad (\textit{Modified CLV Dual})$$
$$+ \sum_{t=1}^T \left(\beta_t - \|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2 \, \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)\right) e_t$$
$$s.t. \quad y + z_t - \|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2 \, \mathbb{1}(\hat{a}_t \neq \tilde{a}_t) e_t \geq$$
$$u_{s_t}^* - v^*, \forall t \in [T]$$
$$y, z_t, e_t \geq 0, \forall t \in [T].$$

At the margin, $y\delta$ corresponds to how the value of the optimal solution to the primal changes if we were to change the budget B by δ, z_t is the marginal value for revealing information at time step t, and e_t is the minimum value that we need to increase o_t to satisfy Constraint (1). Note that in the above dual problem, we have a separate constraint for each z_t and e_t. Within this dual structure, the online arrival of constraints corresponds to the sequential arrival of decision variables, enabling the design of online approximation algorithms.

Let $u_{\max} = \max_{s \in S} u_s^*$ and $u_{\min} = \min_{s \in S} u_s^*$. Let $\pi_{\min}$ be the smallest positive difference between $u_{s_t}^*$ and v^*, i.e., $\pi_{\min} = \min_{s \in S} \max(u_s^* - v^*, 0)$. Let $\pi_{\max} := \max_{s \in S} |u_s^* - v^*|$. We assume that an *upper bound* on $u_{\max}$ and a *lower bound* on $u_{\min}$ are known to the algorithm by domain knowledge. Without loss of generality, we assume that $0 \leq u_{s_t}^* - v^* \leq 1$ for all $s_t \in \mathcal{S}$. Otherwise, we could rescale $u_{s_t}^* - v^*$ by $u_{\max}$ and $u_{\min}$ for all $s_t \in \mathcal{S}$. We outline the online primal-dual algorithm in Algorithm 1.

At each iteration, this algorithm takes the budget (B), the optimal solution value of the recommender ($\{u_{s_t}^*\}$ when $O_t = 1$) and v^* when $O_t = 0$), and the feature mapping when s_t is not observed ($\{\bar{\phi}(a)\}_{a \in \mathcal{A}}$) as *input*. It then "simulates" the optimal solution of the recommender when $O_t = 0$ if the recommender were and were not to have access to $\mathcal{H}_t^{rev}$ ($\tilde{a}_t, \hat{a}_t$) and calculates $\|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2$. Based on this value, the algorithm decides whether to increase o_t (when compared with o_t in Algorithm C.1). Algorithm 1 provides a feasible solution to the auxiliary problem, and only depends on the history it has observed so far. With $\{o_t^{\mathrm{AUX}}\}_{t \in [T]}$

Algorithm 1 Online Primal-Dual Algorithm Revealer with Learning Component

---
1: **Input:** $B, \{u_s^*\}_{s \in S}, v^*, c = (1 + 1/B)^B, \{\bar{\phi}(a)\}_{a \in \mathcal{A}}$.
2: **Initialize:** $y \leftarrow 0, \{\beta_t\}_{t \in T}, e_j \leftarrow 0, \; \forall j \in J$.
3: **for** t = 1, ..., T **do**
4:     The new context s_t arrives, and we observe $\tilde{a}_t, \hat{a}_t$.
5:     **if** $\tilde{a}_t \neq \hat{a}_t$ **and** $\beta_t < \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2$ **and** $u_{s_t}^* > v^*$ **then**
6:        $e_t = \frac{1}{\left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2} - \frac{\beta_t}{\left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2^2}$
7:        **if** $u_{s_t}^* - v^* - y \leq 0$ **then**
8:           $\beta_t = \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2, e_t = 0$.
9:        **end if**
10:     **else**
11:        $\beta_t = \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2, e_t = 0$.
12:     **end if**
13:     **if** $y < 1$ and $u_{s_t}^* - v^* - y + \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t) e_t > 0$ **then**
14:        $\beta_t = \max \left(\beta_t, \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \left(1 - B + \sum_{i=1}^{t-1} o_i \right) \right)$
15:        $z_t = u_{s_t}^* - v^* - y + \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t) e_t$
16:        $o_t = \min(u_{s_t}^* - v^* + \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t) e_t, B - \sum_{i=1}^{t-1} o_i, 1)$
17:        $y \leftarrow y(1 + o_t/B) + o_t/((c-1) \cdot B)$.
18:     **else**
19:        $\beta_t = \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2, z_t = 0, o_t = 0$.
20:     **end if**
21: **end for**

---

chosen according to Algorithm 1, define $\mathbb{E}[V^{\text{AUX}}]$ to be

$$\max_{a \in \mathcal{A}} \sum_{t=1}^{T} (o_t^{\text{AUX}} \langle \theta_*, \phi(s_t, a) \rangle + (1 - o_t^{\text{AUX}}) \langle \theta_*, \bar{\phi}(a) \rangle).$$

Similarly, let $\{o_t^{\text{MCLV}}\}_{t \in [T]}$ be the solution to (*Modified CLV*) in Appendix D.1, then $\mathbb{E}[V^{\text{MCLV}}]$ is

$$\max_{a \in \mathcal{A}} \sum_{t=1}^{T} (o_t^{\text{MCLV}} \langle \theta_*, \phi(s_t, a) \rangle + (1 - o_t^{\text{MCLV}}) \langle \theta_*, \bar{\phi}(a) \rangle).$$

We first show the following result:

Proposition 1. *For any $u_{s_t}^*$, v^*, and context arrival sequence, $\mathbb{E}[V^{AUX}] \geq \pi_{\min}(1 - 1/c)\mathbb{E}[V^{MCLV}]$.*

The proof of Proposition 1 (Appendix D.2) consists of proving that Algorithm 1 is (1) both primal and dual feasible, and (2) in each iteration (day), the ratio between the change in the primal and dual objective functions is bounded by $\pi_{\min}(1-1/c)$. By weak duality, this implies that Algorithm 1 is $\pi_{\min}(1 - 1/c)$-competitive. Moreover, when the ratio $\tau = 1/B$ tends to zero, the competitive ratio of the algorithm tends to the best-possible competitive ratio of $(1 - 1/e)$ [Buchbinder et al., 2007]. In addition, as t increases, Algorithm 1 increasingly favors revealing information when the expected difference in rewards, $u_{s_t}^* - v^*$ is large. Thus, even though $\pi_{\min}$ appears in the competitive ratio, the performance of our algorithm is much better in practice. We illustrate this in Section 5.

Algorithm 1 contains two competing constraints for deciding o_t, and the key technical challenge is to ensure primal feasibility. Some key observations include: 1) to avoid negative competitive ratios, we increase o_t only if $u_{s_t}^* - v^*$ is positive; 2) when running out of budget, i.e., $u_{s_t}^* - v^* - y \leq 0$, we increase β_t such that the learning constraint is always satisfied; 3) when $u_{s_t}^* - v^* < y$ and e_t is not high enough to make o_t positive, we increase β_t such that the second constraint is satisfied.

To complete the regret decomposition, we show the following corollary (proof in Appendix D.3) holds:

Corollary 1. *For any $u_{s_t}^*$, v^*, and any context arrival sequence, $\mathbb{E}[V^{AUX}] \geq \pi_{\min}(1 - 1/c)\mathbb{E}[V^{CLV}]$.*

4 BOUNDING BANDIT LEARNING LOSS

Recall from Section 2 that the reward under action A_t and context S_t, $X_t(S_t, A_t) = \langle \theta_*, \phi(S_t, A_t) \rangle O_t + \eta_t$, where θ_* is the unknown reward parameter, and the noise η_t is conditional mean-zero 1-sub-Gaussian. For the purpose of proofs, we assume there exists a finite W and finite L for which $\|\theta_*\|_2 \leq W$ with Q-probability one and $\max_{a \in \mathcal{A}, s \in S} \|\phi(s, a)\|_2 \leq L$ and $\max_{a \in \mathcal{A}, s \in S} \langle \phi(s, a), \theta_* \rangle \leq 1$ with Q-probability one. Let x_t be the observed reward at time t.

In this section, we learn the unknown parameters θ_* while making *a limited number of* information-revealing

Kyra Gan, Esmaeil Keyvanshokooh, Xueqing Liu, Susan Murphy

decisions. We propose Algorithm 2, an online learning and optimization algorithm that strikes a two-way balance between (i) the exploration-exploitation dilemma for learning the unknown reward, and (ii) hedging against adversarially chosen context arrival sequence.

It consists of (i) a contextual UCB mechanism for learning the unknown θ_* for making the treatment decisions, and (ii) the online primal-dual subroutine (Algorithm 1) for making contextual-revealing decisions. At each iteration t, we maintain two uncertainty sets $\tilde{C}_t$ and $\hat{C}_t$ using histories $\mathcal{H}_t^{\text{rev}}$ and $\mathcal{H}_t^{\text{rec}}$, respectively, for the unknown θ_*, using the high-probability confidence bound that we derived in Proposition E.2. Upon observing a new context s_t by the revealer, the algorithm finds *optimistic* treatments and parameters $(\tilde{A}_t, \tilde{\theta}_t)$ and $(\hat{A}_t, \hat{\theta}_t)$ given $\tilde{C}_t$ and $\hat{C}_t$, respectively, and derives optimistic reward estimates $\tilde{u}_{s_t}^t$ and $\tilde{v}^t$. Given these values, the revealer deploys the online primal-dual subroutine (Algorithm 1) to decide the probability o_t of revealing $\mathcal{H}_t^{\text{rev}}$ to the recommender. If $O_t = 1$, the recommender updates its uncertainty set, i.e., $\hat{C}_t = \tilde{C}_t$, and take the corresponding optimistic action $\hat{A}_t^*$. Otherwise, the recommender exploits its latest uncertainty set and chooses the latest treatment. At each iteration, we update $\hat{C}_{t+1}$ after observing the new reward feedback X_t and context S_t. Let $\phi(s_0, a_0^*) = 0$. See detailed steps in Algorithm 2.

Regret To analyze the regret of Algorithm 2, we first develop a high-probability confidence bound on the regularized least-square estimator of θ_* for the revealer at time t (Proposition E.2). We note that because of Constraint (1), the bandit learning loss relies *only* on the concentration of $\tilde{C}_t$. We then bound the bandit learning loss (BLL_T) associated with learning the unknown reward (Proposition 2). Leveraging our bridging argument (Section 2), we prove our main result on bounding the regret by combining the bandit learning loss (Proposition 2) and the information reveal loss (Corollary 1) together in Theorem 1. Using the notation included in Appendix E and with $\lambda = 1/W^2$, we obtain Proposition 2 (proof in Appendix E.2). Let $BLL_T = \mathbb{E}[V^{\text{AUX}} - V^{\text{ALG}}]$, and also let $V^{\text{ALG}} = \sum_{t=1}^{T}(\langle \phi(S_t, \tilde{A}_t'), \theta_* \rangle O_t^{\text{ALG}} + \langle \bar{\phi}(\hat{A}_t), \theta_* \rangle (1 - O_t^{\text{ALG}}))$, where $\{O_t^{\text{ALG}}\}_{t \in [T]}$ is chosen according to Algorithm 2, and $\tilde{A}_t'$ and $\hat{A}_t$ are the respective recommender's decision given the history is revealed or not. The bandit learning loss BLL_T is defined as

$$BLL_T := \mathbb{E}\Big[\sum_{t=1}^{T} \langle \phi(S_t, A_t^*)O_t^{\text{AUX}} - \phi(S_t, \tilde{A}_t')O_t^{\text{ALG}}, \theta_* \rangle$$

$$+ \sum_{t=1}^{T} \langle \bar{\phi}(A^*)(1 - O_t^{\text{AUX}}) - \bar{\phi}(\hat{A}_t)(1 - O_t^{\text{ALG}}), \theta_* \rangle \Big].$$

Algorithm 2 Online Learning and Optimization Algorithm

1: **Input:** $B, \mathbf{p}^*, \{\bar{\phi}(a) = \sum_{k=1}^{K} \phi(k, a)\mathbf{p}_k^*\}_{a \in \mathcal{A}}$
2: **Initialize:** $\hat{C}_1$ and $\tilde{C}_1$ be the confidence interval for θ_* for recommender and revealer, respectively.
3: **for** $t = 1, ..., T$ **do**
4:     On each iteration t, *revealer* observes the new context s_t and calculates:
5:     $(\tilde{A}_t, \tilde{\theta}_t) = \arg\max_{(a,\theta) \in \mathcal{A} \times \tilde{C}_t} \langle \bar{\phi}(a), \theta \rangle$,
6:     $(\hat{A}_t, \hat{\theta}_t) = \arg\max_{(a,\theta) \in \mathcal{A} \times \hat{C}_t} \langle \bar{\phi}(a), \theta \rangle$,
7:     $\tilde{u}_{s_t}^t = \max_{(a,\theta) \in \mathcal{A} \times \tilde{C}_t} \langle \phi(s_t, a), \theta \rangle$,
8:     $\tilde{v}^t = \left\langle \bar{\phi}(\tilde{A}_t), \tilde{\theta}_t \right\rangle$.
9:     Given $\tilde{u}_{s_t}^t, \tilde{v}_t, \tilde{A}_t, \hat{A}_t$, and B, revealer uses Algorithm 1 to reveal the history $\mathcal{H}_t^{\text{rev}}$ to recommender with probability o_t. Let $O_t \sim \text{Bernoulli}(o_t)$.
10:     **if** $O_t = 1$ **then**
11:         *Recommender* set $\hat{C}_t = \tilde{C}_t$ and calculates:
12:         $(\hat{A}_t^*, \theta_t) = \arg\max_{(a,\theta) \in \mathcal{A} \times \tilde{C}_t} \langle \phi(s_t, a), \theta \rangle$.
13:         *Recommender* takes action $\hat{A}_t^*$.
14:     **else**
15:         *Recommender* set $\hat{C}_{t+1} = \hat{C}_t$ and takes action $\hat{A}_t^* = \hat{A}_t$.
16:     **end if**
17:     *Revealer* observes reward X_t and S_t, and update $\tilde{C}_{t+1}$ according to Proposition E.2.
18: **end for**

Finally, we need the following assumption to bound the bandit learning loss of Algorithm 2:

Assumption 2. *Assume there exists a constant $c_{\max}$ such that $\|\bar{\phi}(\tilde{A}_t)\|_{\tilde{V}_t^{-1}(\lambda)} \leq c_{\max}\|\phi(S_t, \tilde{A}_t)\|_{\tilde{V}_t^{-1}(\lambda)}$.*

Remark 2 (Existence of $c_{\max}$). *We observe that when the horizon is finite, the existence of such a constant $c_{\max}$ is implied by the fact that (i) $\tilde{V}_t^{-1}(\lambda)$ is positive semidefinite, (ii) $\tilde{V}_t^{-1}(\lambda) \preceq \tilde{V}_{t-1}^{-1}(\lambda)$ for all time periods $t \in [T]$, and (iii) $\phi(k, a)$ is bounded for any $k \in K$ and $a \in \mathcal{A}$. When the horizon is infinite, Assumption 2 holds if the bandit algorithm does not converge to selecting a single action (we expect the reveal budget to grow with respect to the horizon length).*

Proposition 2. *With probability $1 - 2\delta$, the bandit learning loss of the Algorithm 2 is bounded by:*

$$\text{BLL}_T \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d + TW^2L^2}{d}\right)}$$

$$+ W \sum_{t=1}^{T} \beta_t + 2B\pi_{\max},$$

where $\gamma = 1 + \sqrt{2\log(\frac{1}{\delta}) + d\log(1 + \frac{TW^2L^2}{d})}$, $\beta_t = \mathcal{O}(1/(\sqrt{t}\log(B)))$, $A_t^* = \arg\max_{a \in \mathcal{A}} \langle \theta_*, \phi(s_t, a) \rangle$, and
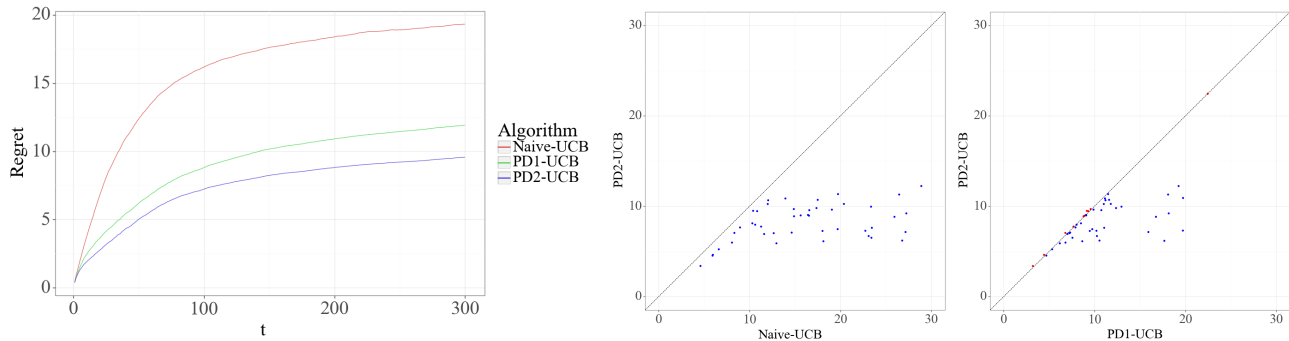
Figure 2: Regret comparison under known $\mathbf{p}^*$ with $B = 10$. Left: cumulative regret averaged over 50 instances (each with a unique θ_* value) each with 50 replications. Middle and right: regret comparison between PD2-UCB, Naive-UCB, and PD1-UCB, at $T = 300$; each dot represents one instance averaged over 50 replications.

$A^* = \arg\max_{a \in \mathcal{A}} \langle \theta_*, \sum_{k=1}^{K} \phi(k,a)\mathbf{p}_k \rangle.$

For every budget level B and horizon length T, there exists some constant α such that with $\beta_T = \alpha/(\sqrt{T}\log(B))$, Constraint (1) satisfied at time T without increasing β_T. Thus, with suitable choices of β_t's, we achieve a sub-linear regret (i.e., $BLL_T \leq \mathcal{O}(d\sqrt{T}\log(T))$, see a more detailed discussion in Remark 3). It is worth noting that the sublinearity of the regret is not the most crucial aspect of this problem. While it is possible to achieve sublinear regret by setting sufficiently large β_t's, this approach may result in a high constant in the regret. Therefore, in Section 5, we numerically evaluate the performance of our algorithm in comparison to two benchmarks, demonstrating the superior performance of our algorithm.

Lastly, the regret of Algorithm 2 (proof in Appendix E.3) is bounded by combining the results of Corollary 1 and Proposition 2. With γ and β_t defined above, and $\alpha = (1 + \frac{1}{c-1})\frac{1}{\pi_{\min}}$, we have our main result:

Theorem 1. *With probability $1 - 2\delta$, the regret of Algorithm 2 is bounded as follows:* $\text{Regret}_T \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d+TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max} + (1-\alpha)\mathbb{E}[V^{CLV}].$

We note that $\mathbb{E}[V^{CLV}] - \mathbb{E}[V^{AUX}] \leq (1-\alpha)\mathbb{E}[V^{CLV}]$ is in the order of $\mathcal{O}(B)$ and it does not depend on T. This is due to the fact that $\{O_t^{\text{CLV}}\}_{t=1}^{T}$ and $\{O_t^{\text{AUX}}\}_{t=1}^{T}$ differ in at most B entries. Thus, $\text{Regret}_T \leq \mathcal{O}(d\sqrt{T}\log(T)) + \mathcal{O}(B)$.

Remark 3 (Regret bound)**.** *In our regret upper bound, the last term $(1-\alpha)\mathbb{E}[V^{CLV}]$ that comes from the online primal-dual subroutine is constant. The first two terms come from the bandit learning algorithm. We have to highlight how a sublinear regret for our algorithm can be achieved. Intuitively, from the definition of the bandit learning loss (BLL_T), we observe that when $O_t = 1$ (i.e., classical regret of contextual bandits), we*

can readily obtain the sublinear regret. Also, when (1) $O_t = 0$ and (2) the optimal action of the recommender given the history $\mathcal{H}_T^{\text{rec}}$ is the same as the one obtained using $\mathcal{H}_T^{\text{rev}}$, then we can also achieve the sublinear regret of $\mathcal{O}(\sqrt{dT\log(T)})$. This implies that sublinearity can be achieved by 1) having a sufficiently high budget to explore all actions (i.e., we require $B > 2|\mathcal{A}|$), and 2) revealing the $\mathcal{H}_t^{\text{rev}}$ sufficiently late, i.e., when the optimal policy A^* is still being learned. We note that the latter can be achieved in Algorithm C.1 by 1) proper scaling of the revealing probability (through $u_{\max}$ so that we do not run out of budget before the horizon ends) and 2) the nature of the algorithm: when t increases, the marginal value of a context ($u_{s_t}^* - v^* - y$) required to reveal information also increases. While in Algorithm 1, the latter can be achieved with proper initialization of the parameters β_t's.*

5 EXPERIMENTS

We conduct experiments on both synthetic and real-world datasets to demonstrate the effectiveness of the proposed algorithm in minimizing regret. To show the benefit of adding the novel learning constraint, we compare two variants of the proposed algorithm: 1) PD1-UCB (UCB with primal-dual without the learning constraint, i.e., replacing the subroutine in Algorithm 2 with Algorithm C.1) and 2) PD2-UCB (Algorithm 2). We benchmark our algorithms with a naive UCB approach that reveals contexts with a fixed probability of B/T (naive-UCB). The experiments are repeated 50 times, and the cumulative regret, revealing probability, and competitive ratio are averaged and presented.

Synthetic Experiments Setup We consider a linear contextual bandit setting with 10 discrete one-dimensional contexts, i.e., $|\mathcal{S}| = 10$. For each context S_k, we sample it according to $\mathbf{p}_k$, which is drawn from a uniform distribution $U(0,1)$ and scaled by $\sum_k \mathbf{p}_k$,

For each *instance*, every coordinate of the true reward parameter $\theta_* \in \mathbb{R}^d$ is sampled from $U(0,1)$. The reward for a selected action A_t in each instance is then generated by $X_t = \langle \theta_*, \phi(S_t, A_t) \rangle + \eta_t$, where $\phi(S_t, A_t)$ includes a one-hot vector (of length $|\mathcal{A}| - 1$) denoting action A_t, a variable denoting context S_t, and $|\mathcal{A}| - 1$ interaction terms. The noise η_t is sampled from $N(0, \sigma^2)$ with $\sigma = 0.1$. We set the number of actions to be $|\mathcal{A}| = 5$, and the length of the time horizon to be $T = 300$. At each step, $\tilde{u}_{s_t}^t$'s and $\tilde{v}^t$'s are normalized using $u_{\max}$ and $u_{\min}$. Throughout this section, we choose $\beta_t = 1.2\Delta_{\min} \log(10)\sqrt{10}/(\sqrt{t}\log(B))$, where $\Delta_{\min} := \min_{k \in [K], a \in \mathcal{A}, a' \in \mathcal{A}\setminus\{a\}} \|\phi(k, a) - \phi(k, a')\|$.

Competitive Ratio We first numerically inspect the empirical competitive ratio of PD1-UCB and PD2-UCB under one instance in Table G.1, calculated by $\mathbb{E}[V^{\mathrm{AUX}}]/\mathbb{E}[V^{\mathrm{CLV}}]$. For PD1-UCB, we replace the sequence of $\{O_t\}_{t \in [T]}$ in V^{AUX} by that chosen according to Algorithm C.1. The competitive ratios are for ground truth θ_* averaged over 200 context arrival sequences. We observe that both PD1-UCB and PD2-UCB have a competitive ratio that is higher than $1 - 1/e$ as stated in Corollary 1. In addition, PD1-UCB has a slightly higher average competitive than PD2-UCB as expected, since at each step, PD2-UCB will like to increase o_t to satisfy Constraint (1).

Regret under Known $\mathbf{p}^*$ We present the cumulative regret when $B = 10$ (Figure 2), $B = 20$ (Figure G.1), and $B = 30$ (Figure G.2). We observe that PD2-UCB outperforms Naive-UCB and PD1-UCB 1) almost instance-wise (the dots above the 90 degree line in Figure 2 right is most likely due to noise), 2) by large margins on many instances (θ_* values). In addition, the benefit of our algorithm is greatest when the budget is low. We note that the regrets are in general increasing with respect to the budget since the optimal strategy for the clairvoyant is changing with respect to B. We include additional experiments where $\mathbf{p}^*$ is unknown in Figures G.3, G.4, and G.5, and observe similar results.

Real-World Experiments Our problem is motivated by the *Oralytics* mobile health application [Trella et al., 2022]. For the purpose of illustrating the concept, we use the ROBAS 2 [Shetty et al., 2020] and ROBAS 3 [Trella et al., 2022] datasets to simulate a scenario involving 10 users ($N = 10$) over 140 decision points ($T = 140$). This design follows the simulation environment presented in Trella et al. [2022] but excludes the integration of delayed effects. The primary objective of this study is to assess the effectiveness of the proposed algorithm in terms of maximizing the cumulative brushing quality, a chosen reward metric. This total brushing quality is represented by $\sum_{t=1}^{T} Q_{i,t}$,
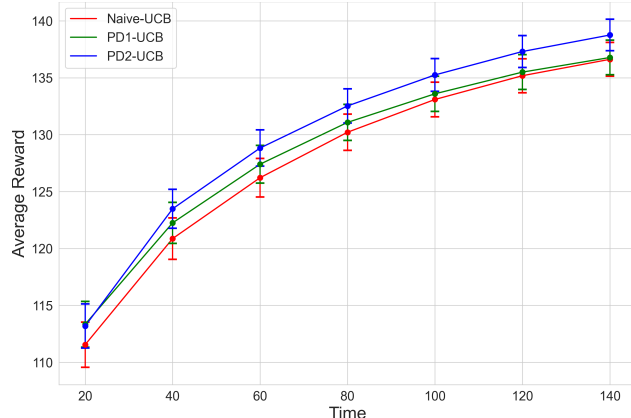

Figure 3: Average reward comparison on the ROBAS3 dataset. The y-axis is the mean and $\pm 1.96 \cdot$ standard error of the average user rewards $\left(\bar{R} = \frac{1}{10}\sum_{i=1}^{10} \frac{1}{t_0} \sum_{s=1}^{t_0} R_{i,s} \right)$ for decision times $t_0 \in [20, 40, 60, 80, 100, 120, 140]$ across 50 experiments and 10 users. Standard error is $\frac{\sum_i^{10} \hat{\sigma}_i}{10\sqrt{50}}$ where $\hat{\sigma}_i$ is the user-specific standard error.

where $Q_{i,t}$ denotes a non-negative measure of brushing quality observed following each decision point. Due to the space limitation, we include the details of the simulation environment in Appendix H.

To simulate a scenario with budgeted information disclosure, an additional budget of $B = 30$ is introduced. At each decision instance t, the revealer has the choice to disclose the historical record up to that point. If the decision is against revealing, the recommender only has access to partial historical information, such as the time of day and weekend indication. We conducted 50 simulation trials. The average reward across multiple time points $\frac{1}{N}\sum_{i=1}^{N} \frac{1}{t_0} \sum_{t=1}^{t_0} Q_{i,t}$, a metric proposed by Trella et al. [2022], is reported.

In Figure 3, we observe that PD2-UCB outperforms PD1-UCB and Naive-UCB in terms of mean performance. However, due to 1) limited trial numbers and 2) the limited number of actions, the confidence intervals of the three methods overlap. We defer larger-scale experiments to future work.

Conclusion and Future Works We develop a novel learning and optimization algorithm for the problem of jointly optimizing the timing of the pro-treatment actions and personalized treatments. This work sheds light on a new direction in online learning and optimization theory that holds promise for advancing digital health interventions. Potential future extensions include: a rigorous extension of our algorithm to other reinforcement learning algorithms, and incorporating a context predictor or noisy context.

Acknowledgments

References

Gagan Aggarwal, Gagan Goel, Chinmay Karande, and Aranyak Mehta. Online vertex-weighted bipartite matching and single-bid budgeted allocations. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 1253–1264. SIAM, 2011.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18. PMLR, 2016.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. *Advances in neural information processing systems*, 32, 2019.

Niv Buchbinder, Kamal Jain, and Joseph Naor. Online primal-dual algorithms for maximizing ad-auctions revenue. In *Algorithms–ESA 2007: 15th Annual European Symposium, Eilat, Israel, October 8-10, 2007. Proceedings 15*, pages 253–264. Springer, 2007.

Niv Buchbinder, Joseph Seffi Naor, et al. The design of competitive online algorithms via a primal–dual approach. *Foundations and Trends® in Theoretical Computer Science*, 3(2–3):93–263, 2009.

Junyu Cao, Esmaeil Keyvanshokooh, and Tian Liu. Safe reinforcement learning with contextual information: Theory and application. *Available at SSRN 4583667*, 2023.

Stephanie M Carpenter, Marianne Menictas, Inbal Nahum-Shani, David W Wetter, and Susan A Murphy. Developments in mobile health just-in-time

adaptive interventions for addiction science. *Current addiction reports*, 7:280–290, 2020.

Wang Chi Cheung, Will Ma, David Simchi-Levi, and Xinshang Wang. Inventory balancing with online learning. *Management Science*, 68(3):1776–1807, 2022.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Nikhil R Devanur and Kamal Jain. Online matching with concave returns. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 137–144, 2012.

Nikhil R Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *Journal of the ACM (JACM)*, 66(1):1–41, 2019.

Jon Feldman, Aranyak Mehta, Vahab Mirrokni, and Shan Muthukrishnan. Online stochastic matching: Beating 1-1/e. In *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 117–126. IEEE, 2009.

Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

Gagan Goel and Aranyak Mehta. Online budgeted matching in random input models with applications to adwords. In *SODA*, volume 8, pages 982–991, 2008.

Patrick Jaillet and Xin Lu. Online stochastic matching: New algorithms with better bounds. *Mathematics of Operations Research*, 39(3):624–646, 2014.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

Chinmay Karande, Aranyak Mehta, and Pushkar Tripathi. Online bipartite matching with unknown distributions. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 587–596, 2011.

Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Online advance scheduling with overtime: A primal-dual approach. *Manufacturing & Service Operations Management*, 23(1):246–266, 2021.

Esmaeil Keyvanshokooh, Mohammad Zhalechian, Cong Shi, Mark P Van Oyen, and Pooyan Kazemian. Con-

textual learning with online convex optimization: Theory and application to medical decision-making. *Forthcoming in Management Science*, 2024.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Xueqing Liu, Kyra Gan, Esmaeil Keyvanshokooh, and Susan Murphy. Online uniform risk times sampling: First approximation algorithms, learning augmentation with full confidence interval integration. *arXiv preprint arXiv:2402.01995*, 2024.

Mohammad Mahdian and Qiqi Yan. Online bipartite matching with random arrivals: an approach based on strongly factor-revealing lps. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 597–606, 2011.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Aranyak Mehta, Amin Saberi, Umesh Vazirani, and Vijay Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22–es, 2007.

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International conference on artificial intelligence and statistics*, pages 2827–2835. PMLR, 2021.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

Vivek Shetty, Douglas Morrison, Thomas Belin, Timothy Hnat, Santosh Kumar, et al. A scalable system for passively monitoring oral health behaviors using electronic toothbrushes in the home setting: development and feasibility study. *JMIR mHealth and uHealth*, 8(6):e17347, 2020.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255, 2022.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckner. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pages 9712–9721. PMLR, 2020.

Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

Mohammad Zhalechian, Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Online resource allocation with personalized learning. *Operations Research*, 70(4):2138–2161, 2022.

Mohammad Zhalechian, Esmaeil Keyvanshokooh, Cong Shi, and Mark P Van Oyen. Data-driven hospital admission control: A learning approach. *Operations Research*, 71(6):2111–2129, 2023.

Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes], please see Section 2 for the mathematical setting and assumptions, and Sections 3 and 4 for algorithms and model.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes], we conducted the regret analysis of our algorithms; see Propositions 1 and 2, and Theorem 1.

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes], specifications of all dependencies, including external libraries, have been set in our code.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes], we explained our assumptions in Sections 2, 3 and 4.

   (b) Complete proofs of all theoretical results. [Yes], all our proofs are provided in Sections C, D, E, and F of the Appendix.

   (c) Clear explanations of any assumptions. [Yes], we discussed our assumptions in Sections 2, 3 and 4.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes], our codes are included in the supplemental material.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes], these details have been included in our codes.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes], the definition of the error bar is included for Figure 3.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes], the detailed information is included in Appendix G.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes], our real-world experiments uses an existing public simulation test bed, and we cited it in the main body.

   (b) The license information of the assets, if applicable. [Not Applicable].

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable].

   (d) Information about consent from data providers/curators. [Not Applicable].

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable].

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable].

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable].

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable].

Appendix

A Additional Motivating Examples: Addiction Treatment

The objective of this section is to show an additional setting where our problem formulation can be useful. In the example of addiction treatments, the recommender could be the text message that the patient receives on their phone (produced by an RL agent sitting in the cloud). A revealer could be a staff member who observes all the information collected by the wearables. Here, we assume that completing the self-reports regularly is a mechanism that researchers use to ensure that the patient opens the App. Similar to the commercial sensor example, we consider the case where the patient does not complete the required self-reports (pro-treatment action), and a staff member could call the patient or their family to reveal all their recent health status. In this example, the context of the patient is partially observed. The sensor data is always observed by the revealer, but the survey data can only be obtained unless the staff reveals it. Once the patient completes the self-report (and thus opens the App), the recommender obtains both the sensor and survey data. We note that the staff could use the sensor data to decide whether to reach out to the patient. In summary, when the patient does not complete the self-report, the recommender observes both the sensor data and the self-reports up to the last reveal. The staff uses the sensor data to decide when to reach out to the patient to reveal. Once revealed, the recommender observes all sensor and self-report data.

B Extension to Partially Observed Context

In many digital health applications, patients need to both (1) allow, passively, sensor data through a data-collecting device to be collected, and (2) complete regular self-reports.

Our problem setting can be extended to the setting, where the revealer only observes a part of the context. Namely, we can partition each state into two components: $S_t = [S_t^1, S_t^2]$. WLOG, let S_t^1 be the part of the state that is *always* observed by the revealer at each time t, and let S_t^2 be the part of the state that can *only* be observed when the revealer takes the action $O_t = 1$. For example, S_t^1 could correspond to the data collected by sensors, and S_t^2 could correspond to the self-reports in the above example.

Let $\ell(t)$ be the time of the last reveal. At each decision time t, the revealer observes the history $\mathcal{H}_t^{\text{rev}} = \{A_1, ..., A_{t-1}, X_1, ..., X_{t-1}, O_1, ..., O_{t-1}, S_1, ..., S_{\ell(t)}, S_{\ell(t)+1}^1, ..., S_t^1\}$. Then, the reward that we observe at each time step can be decomposed as $X_t = \langle \phi(S_t^1, S_t^2, A_t), \theta_* \rangle + \eta_t$, where $\theta_* \in \mathbb{R}^d$ is an *unknown* true reward parameter, $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ is a *known* feature mapping, and the noise η_t is conditional mean-zero 1-sub-Gaussian.

At each time, the revealer decides whether to reveal $\mathcal{H}_t^{\text{rev}}$ to the recommender. If the revealer decides to take an action, i.e., $O_t = 1$, then the revealer additionally observes $\{S_{\ell(t)+1}^2, ..., S_t^2\}$, and the recommender observes $\mathcal{H}_t^{\text{rev}} = \{A_1, ..., A_{t-1}, X_1, ..., X_{t-1}, O_1, ..., O_{t-1}, S_1, ..., ..., S_t\}$. Otherwise, the recommender observes $\mathcal{H}_t^{\text{rec}} = \{A_1, ..., A_{\ell(t)-1}, X_1, ..., X_{\ell(t)-1}, O_1, ..., O_{\ell(t)-1}, S_1, ..., S_{\ell(t)}\}$.

Let M be the cardinality of the set S_t^2, i.e., $S_t^2 \in [M]$. Moreover, let Q be the cardinality of the set S_t^1, i.e., $S_t^1 \in [Q]$. Assume that the state S_t^1 is i.i.d. drawn from distribution $\mathbf{p}'$, and the state S_t^2 is drawn i.i.d. from the conditional distribution $\mathbf{p}''(S_t^1) = \mathbb{P}(S_t^2|S_t^1)$. In ($CLV$), we then have that $u_{s_t^1}^* = \arg\max_{a \in \mathcal{A}} \langle \sum_{m=1}^M \phi(s_t^1, m, a) \mathbf{p}''(s_t^1), \theta_* \rangle$, and $v^* = \arg\max_{a \in \mathcal{A}} \langle \sum_{q=1}^Q \sum_{m=1}^M \phi(q, m, a) \mathbf{p}_q' \mathbf{p}''(q)_m, \theta_* \rangle$, and the rest of problem follows. In ($Modified\ CLV$), the added constraint remains the same since the recommender in this problem setting does not get additional information than the current setting in the main body. Since observing partial information only affects the decision of the revealer, the learning part of the problem remains the same.

C Online Primal-Dual Algorithm without Learning Constraint

In this section, we lay out the road map for deriving the learning constraint and proving Proposition 1. In Algorithm C.1, we only take the budget into account and derive its competitive ratio in Proposition C.1. We first

write down the dual problem of (CLV) as:

$$\min_{y, z_t} \quad By + \sum_{t=1}^{T} z_t$$
$$s.t. \quad y + z_t \geq u^*_{s_t} - v^*, \ \forall t \in [T]$$
$$y, z_t \geq 0, \ \forall t \in [T]. \qquad\qquad (Clairvoyant\ Dual)$$

At the margin, y is the marginal value of the budget ($y\delta$ corresponds to how the value of the optimal solution to the primal change if we were to change the budget B by δ), and z_t is the marginal value for revealing the context at time step t. We note that in ($Clairvoyant\ Dual$), we have a separate constraint for each z_t. Thus, when we do not know the context arrival sequence ahead of time (as in the clairvyant problem), the constraints in ($Clairvoyant\ Dual$) are arriving in an online fashion (one-by-one). This provides a nice framework to analyze the online context arrival sequence.

Let $u_{\max} = \max_{s \in S} u^*_{s_t}$ and $\pi_{\min} = \min_{s \in S} \max(u^*_{s_t} - v^*, 0)$, where $\pi_{\min}$ is the smallest positive difference between $u^*_{s_t}$ and v^*. We assume that an *upper bound* on $u_{\max}$ is known to the algorithm by applying domain knowledge. *Without loss of generality* (WLOG), we assume that $u^*_{s_t} - v^* \leq 1$ for all $s_t \in \mathcal{S}$, since otherwise we could scale $u^*_{s_t} - v^*$ by $u_{\max}$ for all $s_t \in \mathcal{S}$. We outline the online primal-dual algorithm in Algorithm C.1. This algorithm provides a feasible solution to both the primal and dual problems, and only depends on the history it has observed so far. We show:

Proposition C.1. *For any $u^*_{s_t}$, v^*, and context arrival sequence, the value of the objective function of Algorithm C.1 is at least $\pi_{\min}(1 - 1/c)$ times that of (CLV).*

The proof of Proposition C.1 (§ C.1) consists of showing that Algorithm C.1 is (1) both primal and dual feasible, and (2) in each iteration (day), the ratio between the change in the primal and dual objective functions is bounded by $\pi_{\min}(1 - 1/c)$. By weak duality, this implies that Algorithm C.1 is $\pi_{\min}(1 - 1/c)$-competitive. Moreover, note that when the ratio $\tau = 1/B$ tends to zero, the competitive ratio of the algorithm approaches the best-possible competitive ratio of $(1 - 1/e)$ [Buchbinder et al., 2007]. In addition, we note that as t increases, Algorithm C.1 increasingly favors revealing the context information when the expected difference in rewards, $u^*_{s_t} - v^*$ is large. Thus, even though $\pi_{\min}$ appears in our competitive ratio, the performance of our algorithm is much better in practice. We illustrate this in § 5.

We can use Algorithm C.1 as a subroutine in our learning algorithm (Alg. 2) to decide the revealing probability, where we plug in the empirical estimates for $\{u^*_s\}_{s \in \mathcal{S}}$ and v^* obtained by the *revealer*[2] using $\mathcal{H}^{\text{rev}}_t$ at time t, denoting them $\{\tilde{u}^t_s\}_{s \in \mathcal{S}}$ and $\tilde{v}^t$, respectively. Since the recommender has *no* access to $\{\tilde{u}^t_s\}_{s \in \mathcal{S}}$ and $\tilde{v}^t$, as a subroutine, (CLV) lacks a mechanism to connect the quality of the estimates that the *recommender* has at time t to the revealing decision o_t. Ideally, we would like o_t to increase as the time since the last reveal increases. This leads to the Constraint (1) in § 3.

C.1 Proof of Proposition C.1

Proposition C.1. *For any $u^*_{s_t}$, v^*, and context arrival sequence, the value of the objective function of Algorithm C.1 is at least $\pi_{\min}(1 - 1/c)$ times that of (CLV).*

Proof. First, we set $z_t = u^*_{s_t} - v^* - y$ whenever $y < 1$ and $u^*_{s_t} - v^* - y > 0$, this implies that the solution is dual feasible. To show primal feasibility, we need to show $\sum_{t=1}^{T} o_t \leq B$. We prove this by showing that $y > 1$ for at most B updates. Let $\mathcal{I}$ be the set of time indices that we reveal the context. We want to show that when $\sum_{i \in \mathcal{I}} o_t \geq B$, then $y \geq 1$. We use $y^{(i)}$ to denote the value of y in the i-th iteration, where $i \in \mathcal{I}$.

We will prove the following bound:

$$y^{(i)} \geq \frac{1}{c-1} \left(c^{\frac{\sum_{t=1}^{i} o_t}{B}} - 1 \right). \qquad\qquad (2)$$

---

[2]If we plug in the estimates obtained by the *recommender* instead, i.e, $\{\hat{u}^0_s\}_{s \in \mathcal{S}}$ and $\hat{v}^0$, then our algorithm will not be robust with respect to parameter initialization. Take one extreme example, in which the confidence intervals of $\{\hat{u}^0_s\}_{s \in \mathcal{S}}$ and $\hat{v}^0$ are initialized to be the same, then the current algorithm will *never* choose to reveal the context. Thus, the recommender will never learn the optimal interventions.

Algorithm C.1 Online Primal-Dual Algorithm: From Clairvoyant to Auxiliary Revealer

---

Input: $B, \{u_s^*\}_{s \in \mathcal{S}}, v^*, c = (1 + 1/B)^B$
Initialize: $y \leftarrow 0$
On each day, a new context s_t arrives,
and the t-th constraint in the dual problem arrives
if $y < 1$ and $u_{s_t}^* - v^* - y > 0$ **then**
    $z_t = u_{s_t}^* - v^* - y$
    $o_t = \min(u_{s_t}^* - v^*, B - \sum_{i=1}^{t-1} o_i)$
    $y \leftarrow y(1 + o_t/B) + o_t/((c-1) \cdot B)$
else
    $z_t = 0$
    $o_t = 0$
end if

Thus, whenever $\sum_{i \in \mathcal{I}} o_t \geq B$, we have $y \geq 1$. We will prove by induction.

Base case $i = 0$: $y^{(0)} \geq 0$. Equation (2) is trivially true.

Induction step: Assume that Equation (2) is true for step $i - 1$, then show for step 1. We have:

$$
\begin{aligned}
y^{(i)} &= y^{(i-1)}\left(1 + \frac{u_{s_i}^* - v^*}{B}\right) + \frac{u_{s_i}^* - v^*}{(c-1)B} \\
&\geq \frac{1}{c-1}\left[c^{\frac{\sum_{t=1}^{i-1} o_t}{B}} - 1\right]\left(1 + \frac{u_{s_i}^* - v^*}{B}\right) + \frac{u_{s_i}^* - v^*}{(c-1)B} \\
&= \frac{1}{c-1}\left[c^{\frac{\sum_{t=1}^{i-1} o_t}{B}}\left(1 + \frac{u_{s_i}^* - v^*}{B}\right) - 1 - \frac{u_{s_i}^* - v^*}{B} + \frac{u_{s_i}^* - v^*}{B}\right] \\
&\geq \frac{1}{c-1}\left[c^{\frac{\sum_{t=1}^{i-1} o_t}{B}} c^{\frac{u_{s_i}^* - v^*}{B}} - 1\right] \\
&= \frac{1}{c-1}\left[c^{\frac{\sum_{t=1}^{i} o_t}{B}} - 1\right],
\end{aligned}
$$

where the second inequality follows from the induction hypothesis, and the last inequality follows from the fact that $\ln(1+m)/m \geq \ln(1+n)/n$ for all $0 \leq m \leq n \leq 1$, where $m = \frac{u_{s_i}^* - v^*}{B}$, and $n = \frac{1}{B}$. Thus, we have $c = (1 + 1/B)^B$.

Next, we will show that the ratio between the change in the dual and primal objective functions is bounded by $(1 + 1/(c-1))\pi_{\min}^{-1}$. If $y < 1$ and $u_{s_t}^* - v^* - y > 0$, then the primal objective function increases by $(u_{s_t}^* - v^*)^2$, and the increase in the dual objective function is $B\Delta y + z_t = (u_{s_t}^* - v^*)y + (u_{s_t}^* - v^*)/(c-1) + u_{s_t}^* - v^* - y = (u_{s_t}^* - v^*)/(c-1) + (u_{s_t}^* - v^* - 1)y$. Thus, the ratio between the change in the dual and primal is

$$
\frac{\Delta\text{Clairvoyant Dual}}{\Delta\text{Clairvoyant}} = \frac{(u_{s_t}^* - v^*)/(c-1) + u_{s_t}^* - v^* + (u_{s_t}^* - v^* - 1)y}{(u_{s_t}^* - v^*)^2} \leq \left(1 + \frac{1}{c-1}\right)\frac{1}{\pi_{\min}}.
$$

$\square$

D Online Primal-Dual with Learning Constraint

D.1 Clairvoyant with Learning (Primal)

$$\max_{o_t} \quad \sum_{t=1}^{T} o_t \cdot u_{s_t}^* - o_t \cdot v^*$$

$$s.t. \quad \sum_{t=1}^{T} o_t \leq B, \qquad\qquad\qquad (Modified\ CLV)$$

$$- \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t) o_t$$

$$\leq \beta_t - \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t), \ \forall t \in [T]$$

$$o_t \in [0,1], \ \forall t \in [T].$$

D.2 Proof of Proposition 1

Proposition 1. *For any $u_{s_t}^*$, v^*, and context arrival sequence, $\mathbb{E}[V^{AUX}] \geq \pi_{\min}(1 - 1/c)\mathbb{E}[V^{MCLV}]$.*

Proof. First, we observe that $z_t = u_{s_t}^* - v^* - y + \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)e_t$ whenever $y < 1$ and $u_{s_t}^* - v^* - y + \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)e_t > 0$, this implies that $z_t \geq 0$. By construction, $e_t \geq 0 \ \forall t \in [T]$. Thus, the solution is dual feasible.

To show primal feasibility, we first need to show $\sum_{t=1}^{T} o_t \leq B$. We prove this by showing that when $\sum_{i=1}^{t} o_i \geq B$, then $y \geq 1$. We use $y^{(i)}$ to denote the value of y in the i-th iteration, where $i \leq t$.

We will prove the following bound:

$$y^{(i)} \geq \frac{1}{c-1}\left(c^{\frac{\sum_{j=1}^{i} o_j}{B}} - 1 \right). \qquad\qquad (3)$$

Thus, whenever $\sum_{i=1}^{t} o_i \geq B$ for some t, we have $y \geq 1$. We will prove this by induction.

Base case $i = 0$: $y^{(0)} \geq 0$. Equation 3 is trivially true.

Induction step: assume Equation (3) is true for step $i-1$, show for step 1. We have

$$y^{(i)} = y^{(i-1)}\left(1 + \frac{o_i}{B}\right) + \frac{o_i}{(c-1)B}$$

$$\geq \frac{1}{c-1}\left[c^{\frac{\sum_{j=1}^{i-1} o_j}{B}} - 1 \right]\left(1 + \frac{o_i}{B}\right) + \frac{o_i}{(c-1)B}$$

$$= \frac{1}{c-1}\left[c^{\frac{\sum_{j=1}^{i-1} o_j}{B}}\left(1 + \frac{o_i}{B}\right) - 1 - \frac{o_i}{B} + \frac{o_i}{B} \right]$$

$$\geq \frac{1}{c-1}\left[c^{\frac{\sum_{j=1}^{i-1} o_j}{B}} c^{\frac{o_i}{B}} - 1 \right] = \frac{1}{c-1}\left[c^{\frac{\sum_{j=1}^{i} o_j}{B}} - 1 \right],$$

where the second inequality follows from the induction hypothesis, and the last inequality follows from the fact that $\ln{(1+m)}/m \geq \ln{(1+n)}/n$ for all $0 \leq m \leq n \leq 1$, where $m = \frac{o_i}{B}$, and $n = \frac{1}{B}$. Thus, we have $c = (1 + 1/B)^B$.

Next, we need to show that $\left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)(1 - o_t) \leq \beta_t \ \forall t \in [T]$. We first observe that the constraint $u_{s_t}^* > v^*$ is critical for the last step of the primal-dual algorithm where we need to bound the ratio between the change in the primal and dual objective functions. Case 1a): $\hat{a}_t = \tilde{a}_t$ or $\beta_t \geq \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2$. In these cases, this constraint is automatically satisfied and thus we set $e_t = 0$. Case 1b): $u_{s_t}^* \leq v^*$. In this case, our algorithm updates β_t to $\left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2$, and the constraint is satisfied.

Case 2: $\hat{a}_t \neq \tilde{a}_t$, $\beta_t < \left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2$, and $u_{s_t}^* > v^*$. In this case, $e_t = \frac{1}{\left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2} - \frac{\beta_t}{\left\| \bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t) \right\|_2^2}$. Below, we will show that Equation (1) holds for all 3 cases.

Case 2a): $1 \leq \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}$. This case will never happen because $\beta_t < \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2$.

Case 2b): $1 > \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}$ and $u^*_{s_t} - v^* - y + 1 - \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2} > 0$. In this case, $e_t = \frac{1}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2} - \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2^2}$ and $o_t = \min(u^*_{s_t} - v^* + 1 - \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}, B - \sum_{i=1}^{t-1} o_i, 1)$. We first show that Equation (1) holds for $o_t = u^*_{s_t} - v^* + 1 - \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}$: $1 - o_t = (v^* - u^*_{s_t}) + \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2} \leq \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}$, where the last inequality is due to the fact that $u^*_{s_t} > v^*$. Since Equation (1) holds trivially when $o_t = 1$, next we show that Equation (1) holds for $o_t = B - \sum_{i=1}^{t-1} o_i$. In other words, $1 - o_t = 1 - B + \sum_{i=1}^{t-1} o_i$. To ensure that $1 - o_t \leq \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}$, we need to impose $\beta_t \geq \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2 \left(1 - B + \sum_{i=1}^{t-1} o_i\right)$.

Case 2c): $1 > \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2}$ and $u^*_{s_t} - v^* - y + 1 - \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2} \leq 0$. In this case, $e_t = \frac{1}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2} - \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2^2}$ and $o_t = 0$. Intuitively, under this case, $u^*_{s_t} - v^*$ is less than y and e_t is not high enough to have a positive o_t. Since $u^*_{s_t} - v^* > 0$, then the setup of Case 2c implies that $1 > \frac{\beta_t}{\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2} \geq 1 - y + u^*_{s_t} - v^*$. There are two ways that we can avoid this from happening: by setting $\beta_t > \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2$ or by setting $\beta_t \leq (1-y)\left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2$. Due to the nature of budget constraint, we will set $\beta_t > \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2$.

Next, we will show that the ratio between the change in the dual and primal objective functions is bounded by $(1+1/(c-1))\pi_{\min}^{-1}$. If $y < 1$ and $u^*_{s_t} - v^* - y + \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)e_t > 0$, then the primal objective function increases by $o_t(u^*_{s_t} - v^*)$, and the increase in the dual objective function is $B\Delta y + z_t + (\beta_t - \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)e_t \leq o_t y + o_t/(c-1) + o_t - y = o_t/(c-1) + o_t + (o_t - 1)y$, where the first inequality is due to the fact that e_t is positive only when the coefficient $\beta_t - \left\|\bar{\phi}(\tilde{a}_t)-\bar{\phi}(\hat{a}_t)\right\|_2 \mathbb{1}(\hat{a}_t \neq \tilde{a}_t)$ is negative. Thus, the ratio between the change in the dual and primal is

$$\frac{\Delta \text{Clairvoyant Dual with Learning}}{\Delta \text{Clairvoyant with Learning}} = \frac{o_t/(c-1) + o_t + (o_t - 1)y}{o_t(u^*_{s_t} - v^*)} \leq \left(1 + \frac{1}{c-1}\right)\frac{1}{\pi_{\min}}.$$

Note that the last inequality is because $o_t \leq 1$, and $\frac{1}{u^*_{s_t}-v^*} \leq \frac{1}{\pi_{\min}}$ when o_t is positive. $\qquad\square$

D.3 Proof of Corollary 1

Corollary 1. *For any $u^*_{s_t}$, v^*, and any context arrival sequence, $\mathbb{E}[V^{AUX}] \geq \pi_{\min}(1 - 1/c)\mathbb{E}[V^{CLV}]$.*

Proof of Corollary 1. The proof of Proposition 1 shows that the solution provided by Algorithm 1 is feasible to (*Modified CLV*). Since the objective function in (*Modified CLV*) and (*CLV*) are the same, and the feasible set in (*Modified CLV*) is a subset of that of (*CLV*), Algorithm 1 also provides a feasible to (*CLV*). Thus, the last step in the proof of Proposition 1 holds for Corollary 1, completing the proof. $\qquad\square$

E Bandit Learning Loss under Known Context Distributions

Notation Let $x \in \mathbb{R}^d$ be a d-dimensional vector, then $\|x\|_p$ is the p-norm of vector x. Let $A \in \mathbb{R}^{d \times d}$ be a positive definite matrix, then the weighted 2-norm of x is $\|x\|_A := \sqrt{x^T A x}$.

Proposition E.2. *Let $\bar{\theta}_t$ be the L^2-regularised least-square estimator for θ_* using $\mathcal{H}_t^{\text{rec}}$, and let $\delta \in (0,1)$. Then, with probability at least $1 - \delta$, the following inequality holds for any $t \in [T]$ and $\lambda > 0$:*

$$\left\|\theta_* - \bar{\theta}_t\right\|_{\tilde{V}_t(\lambda)} \leq \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\tilde{V}_t(\lambda))}{\lambda^d}\right)} + \sqrt{\lambda}\|\theta_*\|_2,$$

where $\bar{V}_t(\lambda) = \lambda I + \sum_{i=1}^{t-1} \phi(s_i, a_i)\phi(s_i, a_t)^T$.

Note that the recommender constructs $\hat{C}_t$ by using $\mathcal{H}_t^{\text{rec}}$ instead in the above high-probability confidence bound. However, because of Constraint (1), the bandit learning loss relies *only* on the concentration of $\tilde{C}_t$ as stated in Proposition E.2.

E.1   Proof of Proposition E.2

Proof of Proposition E.2. Recall that we assumed $X_t(S_t, A_t)$ as the uncertain reward feedback of patient t with context S_t assigned to action A_t:

$$X_t(S_t, A_t) = \langle \theta_*, \phi(s_t, a_t) \rangle + \eta_t.$$

If we were to observe the realized reward up to time t (noninclusive) at each step, the L^2-regularised least-square estimator of θ_* at time t, $\tilde{\theta}_t$, can be obtained by solving the following optimization problem:

$$L_t(\theta) = \lambda \|\theta\|^2 + \sum_{i=1}^{t-1} \left(X_i - \langle \theta, \phi(s_i, a_i) \rangle \right)^2,$$

where $\lambda > 0$ is the regularization parameter. We note that since the *revealer* observed the context information at each step, so this update step is well-defined.

Minimizing the above term ($\nabla_\theta L_t(\theta) = 0$) yields the following estimator for θ_*:

$$\bar{\theta}_t = \left(\sum_{i=1}^{t-1} \phi(s_i, a_i) \phi(s_i, a_i)^T + \lambda I \right)^{-1} \left(\sum_{i=1}^{t-1} \phi(s_i, a_i) X_i \right).$$

Let

$$\tilde{V}_t(\lambda) = \lambda I + \sum_{i=1}^{t-1} \phi(s_i, a_i) \phi(s_i, a_i)^T,$$

and let $\tilde{V}_t = V_t(0)$, then we can rewrite

$$\bar{\theta}_t = \tilde{V}_t(\lambda)^{-1} \left(\tilde{V}_t \theta_* + \sum_{i=1}^{t-1} \phi(s_i, a_i) \eta_i \right).$$

Following the proof of Theorem 20.5 in Lattimore and Szepesvári [2020], we can show that the following holds for any t:

$$
\begin{aligned}
\left\| \theta_* - \bar{\theta}_t \right\|_{\tilde{V}_t(\lambda)} &= \left\| \tilde{V}_t(\lambda)^{-1} \left(\sum_{i=1}^{t-1} \phi(s_i, a_i) \eta_i \right) + \left(\tilde{V}_t(\lambda)^{-1} \tilde{V}_t - I \right) \theta_* \right\|_{\tilde{V}_t(\lambda)} \\
&\leq \left\| \tilde{V}_t(\lambda)^{-1} \left(\sum_{i=1}^{t-1} \phi(s_i, a_i) \eta_i \right) \right\|_{\tilde{V}_t(\lambda)} + \left\| \left(\tilde{V}_t(\lambda)^{-1} \tilde{V}_t - I \right) \theta_* \right\|_{\tilde{V}_t(\lambda)} \\
&= \left\| \sum_{i=1}^{t-1} \phi(s_i, a_i) \eta_i \right\|_{\tilde{V}_t(\lambda)^{-1}} + \sqrt{ \theta_*^\top (\tilde{V}_t(\lambda)^{-1} \tilde{V}_t - I) \tilde{V}_t(\lambda) (\tilde{V}_t(\lambda)^{-1} \tilde{V}_t - I) \theta_* } \\
&\leq \left\| \sum_{i=1}^{t-1} \phi(s_i, a_i) \eta_i \right\|_{\tilde{V}_t(\lambda)^{-1}} + \sqrt{\lambda} \|\theta_*\|_2
\end{aligned}
$$

where $\tilde{V}_t$ is the design matrix defined above and note that $\tilde{V}_t = \tilde{V}_t(\lambda) - \lambda I$.

Let d be the dimension of $\phi(s_t, a_t)$, and since under $\bar{\theta}_t$ we would have observed the history of the rewards up to time $t-1$ and the history of states up to time t, we consider the σ-algebra $\mathcal{F}_t := \sigma\left(X_1, ..., X_{t-1}, \phi(S_1, A_1), ..., \phi(S_t, A_t) \right)$.

Recall that the noises η_t's are conditionally 1-subgaussian: for all $\alpha \in \mathbb{R}$ and $t \geq 1$, $\mathbb{E}\left[\exp(\alpha \eta_t) | \mathcal{F}_{t-1} \right] \leq \exp\left(\frac{\alpha^2}{2} \right)$ *a.s.* Then, following Lemmas 20.2 and 20.3, and Theorem 20.4 in Lattimore and Szepesvári [2020], we have that for all $\lambda > 0$ and $\delta \in (0, 1)$:

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \left\| \sum_{i=1}^{t-1} \phi_i(a_i) \eta_i \right\|_{\tilde{V}_t(\lambda)^{-1}}^2 \geq 2 \log(1/\delta) + \log(\det(\tilde{V}_t(\lambda))/\lambda^d) \right) \leq \delta.$$

Together, we have that with probability $1 - \delta$,

$$\left\|\theta_* - \bar{\theta}_t\right\|_{\tilde{V}_t(\lambda)} \leq \sqrt{2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(\tilde{V}_t(\lambda))}{\lambda^d}\right)} + \sqrt{\lambda}\|\theta_*\|_2.$$

$\square$

E.2   Proof of Proposition 2

Proposition 2. *With probability $1 - 2\delta$, the bandit learning loss of the Algorithm 2 is bounded by:*

$$\mathrm{BLL}_T \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d + TW^2L^2}{d}\right)}$$

$$+ W \sum_{t=1}^{T} \beta_t + 2B\pi_{\max},$$

where $\gamma = 1 + \sqrt{2\log(\frac{1}{\delta}) + d\log(1 + \frac{TW^2L^2}{d})}$, $\beta_t = \mathcal{O}(1/(\sqrt{t}\log(B)))$, $A_t^* = \arg\max_{a \in \mathcal{A}}\langle\theta_*, \phi(s_t, a)\rangle$, and $A^* = \arg\max_{a \in \mathcal{A}}\langle\theta_*, \sum_{k=1}^{K}\phi(k, a)\mathbf{p}_k\rangle$.

Proof of Proposition 2. At each time t, there are 4 quantities that the information revealer and treatment recommender calculate in our algorithm:

$$\text{For } O_t^{\mathrm{ALG}} = 0 : \ (\tilde{A}_t, \tilde{\theta}_t) = \underset{(a,\theta) \in \mathcal{A} \times \tilde{C}_t}{\arg\max} \ \langle\bar{\phi}(a), \theta\rangle,$$

$$(\hat{A}_t, \hat{\theta}_t) = \underset{(a,\theta) \in \mathcal{A} \times \hat{C}_t}{\arg\max} \ \langle\bar{\phi}(a), \theta\rangle,$$

$$\text{For } O_t^{\mathrm{ALG}} = 1 : \ (\tilde{A}_t', \tilde{\theta}_t') = \underset{(a,\theta) \in \mathcal{A} \times \tilde{C}_t}{\arg\max} \ \langle\phi(S_t, a), \theta\rangle,$$

$$(\hat{A}_t', \hat{\theta}_t') = \underset{(a,\theta) \in \mathcal{A} \times \hat{C}_t}{\arg\max} \ \langle\phi(S_t, a), \theta\rangle.$$

We first note that when $O_t^{\mathrm{ALG}} = 1$, $\hat{C}_t = \tilde{C}_t$, so $\tilde{A}_t' = \hat{A}_t'$ and $\tilde{\theta}_t' = \hat{\theta}_t'$. Thus, when $O_t^{\mathrm{ALG}} = 0$, the treatment recommender takes the action $\hat{A}_t$, and when $O_t^{\mathrm{ALG}} = 1$, the treatment recommender takes the action $\tilde{A}_t'$.

Since when $O_t^{\mathrm{ALG}} = 0$, the recommender does not observe the context S_t, the best action that it can take is $A^* = \arg\max_{a \in \mathcal{A}}\langle\bar{\phi}(a), \theta_*\rangle$. Let $A_t^* = \arg\max_{a \in \mathcal{A}}\langle\phi(S_t, a), \theta_*\rangle$. Thus, we will compare the performance of our algorithm a "weaker" benchmark (who have no access to S_t when $O_t^{\mathrm{ALG}} = 0$) by decomposing the bandit learning loss as follows:

$$\mathrm{BLL}_T = \mathbb{E}\left[\sum_{t=1}^{T}\langle\phi(S_t, A_t^*)O_t^{\mathrm{AUX}} - \phi(S_t, \tilde{A}_t')O_t^{\mathrm{ALG}}, \theta_*\rangle + \sum_{t=1}^{T}\langle\bar{\phi}(A^*)(1 - O_t^{\mathrm{AUX}}) - \bar{\phi}(\hat{A}_t)(1 - O_t^{\mathrm{ALG}}), \theta_*\rangle\right]$$

$$\leq \mathbb{E}\left[\underbrace{\sum_{t=1}^{T}\langle\phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_*\rangle O_t^{\mathrm{ALG}} + \sum_{t=1}^{T}\langle\bar{\phi}(A^*) - \bar{\phi}(\hat{A}_t), \theta_*\rangle(1 - O_t^{\mathrm{ALG}})}_{\text{Term I}}\right]$$

$$+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\langle\phi(S_t, A_t^*) - \bar{\phi}(A^*), \theta_*\rangle(O_t^{\mathrm{AUX}} - O_t^{\mathrm{ALG}})\right]}_{\text{Term II}}$$

Note that in our above derivation, we replace the auxiliary revealing decision of O_t^{AUX} with $O_t^{\mathrm{AUX}} - O_t^{\mathrm{ALG}} + O_t^{\mathrm{ALG}}$.

We then prove how the first expectation (Term I) in above can be bounded as follows:

$$\text{Term I} = \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_*\right\rangle O_t^{\text{ALG}} + \sum_{t=1}^{T}\left\langle\bar{\phi}(A^*) - \bar{\phi}(\hat{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_*\right\rangle O_t^{\text{ALG}} + \sum_{t=1}^{T}\left\langle\bar{\phi}(A^*) - \bar{\phi}(\tilde{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\right]$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\bar{\phi}(\tilde{A}_t) - \bar{\phi}(\hat{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\mathbb{1}_{\tilde{A}_t \neq \hat{A}_t}\right].$$

In above, the first expectation is the standard bandit loss where we observe the entire history up to time t. The second term is obtained due to the fact that when $O_t^{\text{ALG}} = 1$, $\tilde{\theta}(t) = \hat{\theta}(t)$, implying that $\tilde{A}_t = \hat{A}_t$, and $\bar{\phi}(\tilde{A}_t) = \bar{\phi}(\hat{A}_t)$.

Consider the following historical information:

$$\mathcal{F}_t := \sigma(X_1, ..., X_{t-1}, A_1, ..., A_{t-1}, O_1, ..., O_{t-1}, S_1, ..., S_t),$$

where $A_1, ..., A_{t-1}$ are the random variables indicating the actual treatment taken from time 1 to $t-1$. Since our constraint indicates that $(1 - o_t^{\text{ALG}})\mathbb{1}_{\tilde{a}_t \neq \hat{a}_t}\|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2 \leq \beta_t$ given $\mathcal{F}_t$, then using Cauchy–Schwarz inequality, the second term above is bounded by:

$$\mathbb{E}\left[\left\langle\bar{\phi}(\tilde{A}_t) - \bar{\phi}(\hat{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\mathbb{1}_{\tilde{A}_t \neq \hat{A}_t}\right] \leq \mathbb{E}\left[\left|\left\langle\bar{\phi}(\tilde{A}_t) - \bar{\phi}(\hat{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\mathbb{1}_{\tilde{A}_t \neq \hat{A}_t}\right|\right]$$

$$= \mathbb{E}_{\mathcal{F}_t}\left[\mathbb{E}_{O_t^{\text{ALG}}}\left[\left|\left\langle\bar{\phi}(\tilde{A}_t) - \bar{\phi}(\hat{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\mathbb{1}_{\tilde{A}_t \neq \hat{A}_t}\right| |\mathcal{F}_t\right]\right]$$

$$\leq (1 - o_t^{\text{ALG}})\mathbb{1}_{\tilde{a}_t \neq \hat{a}_t}\|\bar{\phi}(\tilde{a}_t) - \bar{\phi}(\hat{a}_t)\|_2\|\theta_*\|_2 \leq W\beta_t.$$

We note that in the second line above, when conditioned on the σ-algebra $\mathcal{F}_t$, the only uncertainty coming from the inner expectation is from observation O_t^{ALG}.

By the construction of UCB and the fact that $\theta_*, \tilde{\theta}_t', \tilde{\theta}_t \in \tilde{C}_t$, we have that:

$$\langle\phi(S_t, A_t^*), \theta_*\rangle \leq \left\langle\phi(S_t, \tilde{A}_t'), \tilde{\theta}_t'\right\rangle, \text{ and}$$

$$\langle\bar{\phi}(A^*), \theta_*\rangle \leq \left\langle\bar{\phi}(\tilde{A}_t), \tilde{\theta}_t\right\rangle.$$

Thus, by Cauchy-Schwarz and the above facts, we have the following:

$$\text{Term I} \leq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_*\right\rangle O_t^{\text{ALG}} + \sum_{t=1}^{T}\left\langle\bar{\phi}(A^*) - \bar{\phi}(\tilde{A}_t), \theta_*\right\rangle(1 - O_t^{\text{ALG}})\right] + W\sum_{t=1}^{T}\beta_t$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\phi(S_t, \tilde{A}_t'), \tilde{\theta}_t' - \theta_*\right\rangle O_t^{\text{ALG}} + \sum_{t=1}^{T}\left\langle\bar{\phi}(\tilde{A}_t), \tilde{\theta}_t - \theta_*\right\rangle(1 - O_t^{\text{ALG}})\right] + W\sum_{t=1}^{T}\beta_t$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\|\phi(S_t, \tilde{A}_t')\|_{\tilde{V}_t^{-1}(\lambda)}\|\tilde{\theta}_t' - \theta_*\|_{\tilde{V}_t(\lambda)}O_t^{\text{ALG}} + \|\bar{\phi}(\tilde{A}_t)\|_{\tilde{V}_t^{-1}(\lambda)}\|\tilde{\theta}_t - \theta_*\|_{\tilde{V}_t(\lambda)}(1 - O_t^{\text{ALG}})\right] + W\sum_{t=1}^{T}\beta_t,$$

where $\tilde{V}_t(\lambda) = V_0 + \sum_{i=1}^{t-1}\phi(s_i, a_i)\phi(s_i, a_i)^T$, and $V_0 = \lambda I$.

By Assumption 2, we know there exists a constant $c_{\max}$ such that $\|\bar{\phi}(\tilde{A}_t)\|_{\tilde{V}_t^{-1}(\lambda)} \leq c_{\max}\|\phi(S_t, \tilde{A}_t)\|_{\tilde{V}_t^{-1}(\lambda)}$.

In addition, we note that since $\tilde{V}_t^{-1}(\lambda)$ does not depend on the action taken at time period $t \in [T]$, we can bound $\|\phi(S_t, \tilde{A}_t)\|_{\tilde{V}_t^{-1}(\lambda)}$ using Lemma 19.4 of Lattimore and Szepesvári [2020].

In addition, recall that we obtained a high probability confidence set for the unknown parameter θ_* in Proposition E.2 as the confidence bound $\|\theta_* - \bar{\theta}_t(\lambda)\|_{\tilde{V}_t(\lambda)} \leq \sqrt{2\log(1/\delta) + \log(\det(\tilde{V}_t(\lambda))/\lambda^d)} + \sqrt{\lambda}\|\theta_*\|_2$. Plugging in the

value $\lambda = 1/W^2$, and also utilizing the inequality that $\frac{\det V_t(\lambda)}{\lambda^d} \leq \left(\text{trace}\left(\frac{V_t(\lambda)}{\lambda d}\right)\right)^d \leq \left(1 + \frac{TL^2}{\lambda d}\right)^d$, we can get that the width of the confidence interval is bounded by $\gamma = 1 + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log(1 + \frac{TW^2L^2}{d})}$. Thus, we have that $\|\theta_* - \bar{\theta}_t(\lambda)\|_{\widetilde{V}_t(\lambda)} \leq \gamma$. Furthermore, since both parameters $\widetilde{\theta}_t$ and θ_* belong to this confidence bound, i.e., $\widetilde{\theta}_t, \theta_* \in \widetilde{\Theta}_t = \left\{\theta \in \mathbb{R}^d : \|\theta - \bar{\theta}_t(\lambda)\|_{\widetilde{V}_t(\lambda)} \leq \gamma\right\}$, we have: $\left\|\widetilde{\theta}_t - \theta_*\right\|_{\widetilde{V}_t(\lambda)} \leq \|\widetilde{\theta}_t - \bar{\theta}_t(\lambda)\|_{\widetilde{V}_t(\lambda)} + \|\theta_* - \bar{\theta}_t(\lambda)\|_{\widetilde{V}_t(\lambda)} \leq 2\gamma$.

Considering the above two explanations together, we can now continue bounding Term (I) with a probability of at least $1 - 2\delta$ as follows:

$$
\begin{aligned}
\text{Term I} \leq {}& 2\gamma\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\min\left(1, \|\phi(S_t, \tilde{A}'_t)\|_{\widetilde{V}_t^{-1}}^2\right)(O_t^{\text{ALG}})^2}\right] \\
& + 2\gamma\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\min\left(1, \|\bar{\phi}(\tilde{A}_t)\|_{\widetilde{V}_t^{-1}}^2\right)(1 - O_t^{\text{ALG}})^2}\right] + W\sum_{t=1}^{T}\beta_t \\
\leq {}& 2\gamma\max(c_{\max}, 1)\sqrt{2Td\log\left(\frac{\text{trace}(V_0) + TL^2}{d\det(V_0)^{1/d}}\right)} + W\sum_{t=1}^{T}\beta_t \\
\leq {}& \max(c_{\max}, 1)\sqrt{8Td\gamma^2\log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t,
\end{aligned}
$$

where the first inequality follows from the fact that O_t^{ALG} is binary and the last inequality follows from Lemma 19.4 of Lattimore and Szepesvári [2020].

We next prove how the second expectation (Term II) in our regret decomposition can be bounded. Let $\pi_{\max} := \max_{s \in S}|u_s^* - v^*|$, where $u_{s_t}^* = \langle\phi(s_t, A_t^*), \theta_*\rangle$, and $v^* = \langle\bar{\phi}(A^*), \theta_*\rangle$. Thus, Term II is bounded by:

$$
\text{Term II} = \mathbb{E}\left[\sum_{t=1}^{T}\langle\phi(S_t, A_t^*) - \bar{\phi}(A^*), \theta_*\rangle(O_t^{\text{AUX}} - O_t^{\text{ALG}})\right] \leq \pi_{\max}\mathbb{E}\left[\sum_{t=1}^{T}|(O_t^{\text{AUX}} - O_t^{\text{ALG}})|\right] \leq 2B\pi_{\max}. \quad (4)
$$

We note that by Assumption 1, Equation 4 is bounded by $\mathcal{O}(\sqrt{T})$.

Putting the bounds for Terms I and II, we can bound the bandit learning loss as:

$$
\text{BLL}_T \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2\log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max}.
$$

$\square$

E.3 Proof of Theorem 1

Theorem 1. *With probability $1 - 2\delta$, the regret of Algorithm 2 is bounded as follows:* $\text{Regret}_T \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2\log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max} + (1 - \alpha)\mathbb{E}[V^{CLV}].$

Proof of Theorem 2. Recall we defined three models and then establish the following bridging argument for the regret decomposition:

$$
\text{Regret}_T \leq \mathbb{E}\left[V^{\text{CLV}}\right] - \mathbb{E}\left[V^{\text{ALG}}\right] = \underbrace{\mathbb{E}\left[V^{\text{AUX}} - V^{\text{ALG}}\right]}_{\text{Bandit Learning Loss}} + \underbrace{\mathbb{E}\left[V^{\text{CLV}} - V^{\text{AUX}}\right]}_{\text{Information Reveal Loss}}.
$$

From our online primal-dual analysis, we derive the following competitive ratio:

$$
\mathbb{E}\left[V^{\text{AUX}}\right]/\mathbb{E}\left[V^{\text{CLV}}\right] \geq \left(1 + \frac{1}{c - 1}\right)\frac{1}{\pi_{\min}},
$$

which implies the following by a simple algebra and letting $\alpha = \left(1 + \frac{1}{c-1}\right)\frac{1}{\pi_{\min}}$:

$$\mathbb{E}[V^{\mathrm{CLV}}] - \mathbb{E}\left[V^{\mathrm{AUX}}\right] \leq (1 - \alpha)\,\mathbb{E}[V^{\mathrm{CLV}}].$$

Also, from our Proposition 2, we have the following:

$$\begin{aligned}
\mathrm{BLL}_T &= \mathbb{E}\left[V^{\mathrm{AUX}} - V^{\mathrm{ALG}}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle \phi(S_t, A_t^*)O_t^{\mathrm{AUX}} - \phi(S_t, \tilde{A}_t')O_t^{\mathrm{ALG}}, \theta_*\rangle + \sum_{t=1}^{T}\langle \bar{\phi}(A^*)(1 - O_t^{\mathrm{AUX}}) - \bar{\phi}(\hat{A}_t)(1 - O_t^{\mathrm{ALG}}), \theta_*\rangle\right] \\
&\leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max}.
\end{aligned}$$

So, summing the upper bounds established above for each term, we have the following:

$$\mathrm{Regret}_T \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max} + (1 - \alpha)\,\mathbb{E}\left[V^{\mathrm{CLV}}\right].$$

$\square$

F Extension to Unknown Context Distribution

We next extend our online learning setting to the case, where we also learn the unknown context distribution, $\mathbf{p}^*$, in Algorithm F.2. To derive the regret, we leverage the *empirical Berstein's inequality* to build a high-probability confidence bound $\tilde{P}_t$ for the latent context distribution (see Lemma F.2). Let $\tilde{\mathbf{p}}_k^t$ be the empirical average estimate of $\mathbf{p}_k^*$. With γ and β_t defined above, and $\zeta_t = \sqrt{\frac{2\,\tilde{\mathbf{p}}_k^t\left(1 - \tilde{\mathbf{p}}_k^t\right)\log(\frac{2KT}{\delta})}{\max\{m(k,t),1\}}} + \frac{7\log(\frac{2KT}{\delta})}{3(\max\{m(k,t)-1,1\})}$, where $m(k,t)$ is the number of times that the context k has been observed up to time t, the bandit learning loss in Algorithm F.2 is defined as $BLL_T := \mathbb{E}[\sum_{t=1}^{T}\langle \phi(S_t, A_t^*)O_t^{\mathrm{AUX}} - \phi(S_t, \tilde{A}_t')O_t^{\mathrm{ALG}}, \theta_*\rangle] + \mathbb{E}[\sum_{t=1}^{T}\sum_{k=1}^{K}\langle \phi(k, A^*)(1 - O_t^{\mathrm{AUX}}) - \phi(k, \hat{A}_t)(1 - O_t^{\mathrm{ALG}}), \theta_*\rangle\mathbf{p}_k^*]$.

Proposition F.3. *With probability $1 - 4\delta$, the bandit learning loss in Algorithm F.2 is bounded by:*

$$\begin{aligned}
\mathrm{BLL}_T &=\leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max} + (WL+1)\sum_{t=1}^{T}\sum_{k=1}^{K}\zeta_t \\
&= \mathcal{O}\left(d\sqrt{T}\log\left(T\right) + K\sqrt{T}\right),
\end{aligned}$$

where $\tilde{A}_t'$ and $\hat{A}_t$ are the respective Algorithm F.2's treatments given the history is revealed or not to the recommender.

The proof of Proposition F.3 is included in Appendix F.1. Algorithm F.2 is computationally expensive since, at each step, it involves optimizing over two convex uncertainty sets when calculating the optimal action. While this step can be solved using an existing bilinear optimization solver, the objective function in our problem is neither convex nor concave, making the problem NP-hard. Instead, we plugin the empirical mean estimate of $\mathbf{p}^*$ in Algorithm F.2, and numerically evaluate its performance in Appendix G.

F.1 Proof

Lemma F.1. *(Empirical Berstein's Inequality Maurer and Pontil [2009]) Let $X = (X_1, X_2, ..., X_n)$ be i.i.d. random vector with values in $[0,1]^n$, and let $\delta \in (0,1)$. Then, we have the following bound holds with probability at least $1 - \delta$:*

$$\mathbb{E}\left[X\right] - \frac{1}{n}\sum_{i=1}^{n}X_i \leq \sqrt{\frac{2V_n(X)\log(\frac{2}{\delta})}{n}} + \frac{7\log(\frac{2}{\delta})}{3(n-1)},$$

where $V_n(X)$ is the sample variance.

Algorithm F.2 Online θ^* and $\mathbf{p}^*$ Learning & Optimization Algorithm

---

Input: B

Initialize: $\hat{C}_1$ and $\tilde{C}_1$ be the confidence interval of θ_* of recommender and revealer, respectively; $\tilde{P}_1, \hat{P}_1$ be the confidence interval of $\mathbf{p}^*$ of recommender and revealer, respectively.

for $t = 1, ..., T$ **do**

    On each iteration t, *revealer* observes the new context s_t and calculates:

    $(\tilde{A}_t, \tilde{\theta}_t, \tilde{\mathbf{p}}^t) = \arg\max_{(a,\theta,\mathbf{p}) \in \mathcal{A} \times \tilde{C}_t \times \tilde{P}_t} \langle \sum_{k=1}^K \phi(k, a)\mathbf{p}_k, \theta \rangle,$

    $(\hat{A}_t, \hat{\theta}_t, \hat{\mathbf{p}}^t) = \arg\max_{(a,\theta,\mathbf{p}) \in \mathcal{A} \times \hat{C}_t \times \hat{P}_t} \langle \sum_{k=1}^K \phi(k, a)\mathbf{p}_k, \theta \rangle,$

    $\tilde{u}_{s_t}^t = \max_{(a,\theta) \in \mathcal{A} \times \tilde{C}_t \times} \langle \phi(s_t, a), \theta \rangle,$

    $\tilde{v}^t = \left\langle \sum_{k=1}^K \phi(k, \tilde{A}_t)\tilde{\mathbf{p}}_k^t, \tilde{\theta}_t \right\rangle.$

    Given $\tilde{u}_{s_t}^t, \tilde{v}_t, \tilde{A}_t, \hat{A}_t$, and B, revealer uses Algorithm 1 to reveal the history $\mathcal{H}_t^{\text{rev}}$ to recommender with probability: $O_t = \text{Bernoulli}(o_t)$.

    if $O_t = 1$ **then**

        Recommender set $\hat{C}_t = \tilde{C}_t$ and $\hat{P}_t = \tilde{P}_t$ and calculates:

        $(\hat{A}_t^*, \theta_t) = \arg\max_{(a,\theta) \in \mathcal{A} \times \tilde{C}_t} \langle \phi(s_t, a), \theta \rangle.$

        Recommender takes action $\hat{A}_t^*$.

    else

        Recommender set $\hat{C}_{t+1} = \hat{C}_t$, $\hat{P}_{t+1} = \hat{P}_t$, and takes action $\hat{A}_t^* = \hat{A}_t$.

    end if

    Revealer observes reward X_t and S_t, and update $\tilde{C}_{t+1}$ and $\tilde{P}_{t+1}$.

end for

Lemma F.2. *(Confidence Bound on Context Distribution) If we use the empirical average estimate $\widehat{\mathbf{p}}_k(t)$ for estimating the latent context distribution $\mathbf{p}_k^* = \mathbb{P}(S_t = k))$ for each context $k \in [K]$ at iteration $t \in [T]$, then the following bound holds with probability at least $1 - 2\delta$:*

$$|\mathbf{p}_k^* - \widehat{\mathbf{p}}_k(t)| \leq \sqrt{\frac{2\,\widehat{\mathbf{p}}_k(t)\,(1 - \widehat{\mathbf{p}}_k(t))\log(\frac{2KT}{\delta})}{\max\{m(k,t), 1\}}} + \frac{7\log(\frac{2KT}{\delta})}{3\,(\max\{m(k,t) - 1, 1\})}, \tag{5}$$

where $m(k, t)$ is the number of time that the context k has been observed up to time t.

Proof of Lemma F.2. First, we establish the following for estimating the latent context distribution (i.e., $\mathbf{p}_k = \mathbb{P}(S_t = k)$) for each context k:

$$\widehat{\mathbf{p}}_k(t) = \frac{\sum_{u=1}^{t-1} \mathbb{1}(S_u = k)}{\sum_{u=1}^{t-1} \sum_{m=1}^K \mathbb{1}(S_u = m)},$$

Using the empirical Berstein's inequality in Lemma F.1 and making a union-bound argument, the resulting bound is obtained with probability at least $1 - 2\delta$. □

Proposition F.3. *With probability $1 - 4\delta$, the bandit learning loss in Algorithm F.2 is bounded by:*

$$\text{BLL}_T = \leq \max(c_{\max}, 1)\sqrt{8Td\gamma^2 \log\left(\frac{d + TW^2L^2}{d}\right)} + W\sum_{t=1}^T \beta_t + 2B\pi_{\max} + (WL + 1)\sum_{t=1}^T \sum_{k=1}^K \zeta_t$$

$$= \mathcal{O}\left(d\sqrt{T}\log(T) + K\sqrt{T}\right),$$

where $\tilde{A}_t'$ and $\hat{A}_t$ are the respective Algorithm F.2's treatments given the history is revealed or not to the recommender.

Proof of Proposition F.3. In this proof, we take the learning of context distribution into account. Let $\tilde{P}_t$ be the uncertainty set for the context distribution that contains the ground truth context distribution $\mathbf{p}^*$ at time t.

At each time t, there are 4 quantities that the information revealer and treatment recommender calculate:

$$\text{For } O_t^{\text{ALG}} = 0: \; (\tilde{A}_t, \tilde{\theta}_t, \tilde{\mathbf{p}}^t) = \underset{(a,\theta,\mathbf{p}) \in \mathcal{A} \times \tilde{C}_t \times \tilde{P}_t}{\arg\max} \left\langle \sum_{k=1}^{K} \phi(k,a)\mathbf{p}, \theta \right\rangle, \tag{6}$$

$$(\hat{A}_t, \hat{\theta}_t, \hat{\mathbf{p}}^t) = \underset{(a,\theta,\mathbf{p}) \in \mathcal{A} \times \hat{C}_t \times \hat{P}_t}{\arg\max} \left\langle \sum_{k=1}^{K} \phi(k,a)\mathbf{p}, \theta \right\rangle, \tag{7}$$

$$\text{For } O_t^{\text{ALG}} = 1: \; (\tilde{A}_t', \tilde{\theta}_t') = \underset{(a,\theta) \in \mathcal{A} \times \tilde{C}_t}{\arg\max} \langle \phi(S_t, a), \theta \rangle, \tag{8}$$

$$(\hat{A}_t', \hat{\theta}_t') = \underset{(a,\theta) \in \mathcal{A} \times \hat{C}_t}{\arg\max} \langle \phi(S_t, a), \theta \rangle. \tag{9}$$

We first note that when $O_t^{\text{ALG}} = 1$, $\hat{C}_t = \tilde{C}_t$, so $\tilde{A}_t' = \hat{A}_t'$ and $\tilde{\theta}_t' = \hat{\theta}_t'$. Thus, when $O_t^{\text{ALG}} = 0$, the treatment recommender takes the action $\hat{A}_t$, and when $O_t^{\text{ALG}} = 1$, the treatment recommender takes the action $\tilde{A}_t'$.

Since when $O_t^{\text{ALG}} = 0$, the recommender does not observe the context S_t, the best action that it can take is $A^* = \arg\max_{a \in \mathcal{A}} \langle \bar{\phi}(a), \theta_* \rangle$. Let $A_t^* = \arg\max_{a \in \mathcal{A}} \langle \phi(S_t, a), \theta_* \rangle$. Thus, we will compare the performance of our algorithm to a "weaker" benchmark (who has no access to S_t when $O_t^{\text{ALG}} = 0$). The bandit regret at time T is the following:

$$\begin{aligned}
\text{BLL}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \phi(S_t, A_t^*)O_t^{\text{AUX}} - \phi(S_t, \tilde{A}_t')O_t^{\text{ALG}}, \theta_* \right\rangle \right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k=1}^{K} \left\langle \phi(k, A^*)(1 - O_t^{\text{AUX}}) - \phi(k, \hat{A}_t)(1 - O_t^{\text{ALG}}), \theta_* \right\rangle \mathbf{p}_k^* \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle \phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_* \rangle O_t^{\text{ALG}} + \sum_{t=1}^{T} \sum_{k=1}^{K} \langle \phi(k, A^*) - \phi(k, \hat{A}_t), \theta_* \rangle \mathbf{p}_k^*(1 - O_t^{\text{ALG}}) \right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \langle \phi(S_t, A_t^*), \theta_* \rangle (O_t^{\text{AUX}} - O_t^{\text{ALG}}) + \sum_{t=1}^{T} \sum_{k=1}^{K} \langle \phi(k, A^*), \theta_* \rangle \mathbf{p}_k^* \left(O_t^{\text{ALG}} - O_t^{\text{AUX}} \right) \right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \langle \phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_* \rangle O_t^{\text{ALG}} \right] + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k=1}^{K} \langle \phi(k, A^*) - \phi(k, \tilde{A}_t), \theta_* \rangle \mathbf{p}_k^*(1 - O_t^{\text{ALG}}) \right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k=1}^{K} \langle \phi(k, \tilde{A}_t) - \phi(k, \hat{A}_t), \theta_* \rangle \mathbf{p}_k^*(1 - O_t^{\text{ALG}}) \mathbb{1}_{\tilde{A}_t \neq \hat{A}_t} \right] \\
&+ \mathbb{E}\left[\sum_{t=1}^{T} \langle \phi(S_t, A_t^*) - \bar{\phi}(A^*), \theta_* \rangle (O_t^{\text{AUX}} - O_t^{\text{ALG}}) \right].
\end{aligned}$$

To obtain the above decomposition, we first replace the auxiliary revealing decision of O_t^{AUX} with $O_t^{\text{AUX}} - O_t^{\text{ALG}} + O_t^{\text{ALG}}$. We then add and subtract the term $\phi(k, \tilde{A}_t)$, and use the fact that $\bar{\phi}(A^*) = \sum_{k=1}^{K} \phi(k, A^*) \mathbf{p}_k^*$.

The last expectation term above can be bounded using the similar procedure for bounding term II in Proposition 2, which results in:

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \phi(S_t, A_t^*) - \bar{\phi}(A^*), \theta_* \rangle (O_t^{\text{AUX}} - O_t^{\text{ALG}}) \right] \leq 2B\pi_{\max}.$$

We then take a look at bounding the third expectation term. Consider the following information history:

$$\mathcal{F}_t := \sigma(\{X_i\}_{i \in [t-1]}, \{O_i^{\text{ALG}}\}_{i \in [t-1]}, \{\phi(k, A_i)\}_{k \in K}^{i \in [t-1]}, \{S_i\}_{i \in [t]}, \tilde{\mathbf{p}}^1, ..., \tilde{\mathbf{p}}^t, \hat{\mathbf{p}}^1, ..., \hat{\mathbf{p}}^t),$$

where the notation $\{X_i\}_{i \in [j]}$ represents the enumerations of X_i from $i = 1$ to j, and similarly for the other variables.

Our constraint indicates that $(1 - o_t^{\mathrm{ALG}}) \mathbb{1}_{\tilde{a}_t \neq \hat{a}_t} \left\| \sum_{k=1}^{K} \phi(k, \tilde{a}_t) \tilde{\mathbf{p}}_k^t - \sum_{k=1}^{K} \phi(k, \hat{a}_t) \tilde{\mathbf{p}}_k^t \right\|_2 \leq \beta_t$ given the information history $\mathcal{F}_t$. Using this inequality along with the Cauchy–Schwarz inequality, the third expectation term above can be bounded as follows:

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{k=1}^{K} \left\langle \phi(k, \tilde{A}_t) - \phi(k, \hat{A}_t), \theta_* \right\rangle \mathbf{p}_k^* (1 - O_t^{\mathrm{ALG}}) \mathbb{1}_{\tilde{A}_t \neq \hat{A}_t} \right]$$

$$\leq \sum_{t=1}^{T} \mathbb{E}\left[\left| \sum_{k=1}^{K} \left\langle \phi(k, \tilde{A}_t) - \phi(k, \hat{A}_t), \theta_* \right\rangle \left(\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t + \tilde{\mathbf{p}}_k^t \right) (1 - O_t^{\mathrm{ALG}}) \mathbb{1}_{\tilde{A}_t \neq \hat{A}_t} \right| \right]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{F}_t} \left[\mathbb{E}_{O_t^{\mathrm{ALG}}} \left[\left| \sum_{k=1}^{K} \left\langle \phi(k, \tilde{A}_t) - \phi(k, \hat{A}_t), \theta_* \right\rangle \left(\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t + \tilde{\mathbf{p}}_k^t \right) (1 - O_t^{\mathrm{ALG}}) \mathbb{1}_{\tilde{A}_t \neq \hat{A}_t} \right| \, \Big| \mathcal{F}_t \right] \right]$$

$$\leq \sum_{t=1}^{T} (1 - o_t^{\mathrm{ALG}}) \mathbb{1}_{\tilde{a}_t \neq \hat{a}_t} \left\| \sum_{k=1}^{K} \left(\phi(k, \tilde{a}_t) - \phi(k, \hat{a}_t) \right) \tilde{\mathbf{p}}_k^t \right\|_2 \|\theta_*\|_2$$

$$+ \sum_{t=1}^{T} (1 - o_t^{\mathrm{ALG}}) \mathbb{1}_{\tilde{a}_t \neq \hat{a}_t} \left\| \sum_{k=1}^{K} \left(\phi(k, \tilde{a}_t) - \phi(k, \hat{a}_t) \right) \left(\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t \right) \right\|_2 \|\theta_*\|_2$$

$$\leq W \sum_{t=1}^{T} \beta_t + WL \sum_{t=1}^{T} \left| \sum_{k=1}^{K} \left(\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t \right) \right|.$$

The last inequality is due to the fact that $\|\theta_*\|_2 \leq W$ and $\max_{a \in \mathcal{A}, s \in \mathcal{S}} \|\phi(s, a)\|_2 \leq L$. We note that in the second line above, when conditioned on the σ-algebra $\mathcal{F}_t$, the only uncertainty coming from the inner expectation is from observation O_t^{ALG}.

By the construction of UCB (optimism) and the fact that $\theta_*, \tilde{\theta}_t', \tilde{\theta}_t \in \tilde{C}_t$, we have the inequalities:

$$\langle \phi(S_t, A_t^*), \theta_* \rangle \leq \left\langle \phi(S_t, \tilde{A}_t'), \tilde{\theta}_t' \right\rangle, \text{ and}$$

$$\left\langle \sum_{k=1}^{K} \phi(k, A^*) \mathbf{p}_k^*, \theta_* \right\rangle \leq \left\langle \sum_{k=1}^{K} \phi(k, \tilde{A}_t) \tilde{\mathbf{p}}_k^t, \tilde{\theta}_t \right\rangle.$$

By Cauchy-Schwarz and the above facts, we can establish the following arguments to bound the bandit regret:

$$\mathrm{BLL}_T \leq \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \phi(S_t, A_t^*) - \phi(S_t, \tilde{A}_t'), \theta_* \right\rangle O_t^{\mathrm{ALG}} \right] + W \sum_{t=1}^{T} \beta_t + WL \sum_{t=1}^{T} \left| \sum_{k=1}^{K} \mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t \right|$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \sum_{k=1}^{K} \phi(k, A^*) \mathbf{p}_k^* - \sum_{k=1}^{K} \phi(k, \tilde{A}_t) \mathbf{p}_k^*, \theta_* \right\rangle (1 - O_t^{\mathrm{ALG}}) \right] + 2B\pi_{\max}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \phi(S_t, \tilde{A}_t'), \tilde{\theta}_t' - \theta_* \right\rangle O_t^{\mathrm{ALG}} \right] + W \sum_{t=1}^{T} \beta_t + WL \sum_{t=1}^{T} \left| \sum_{k=1}^{K} \mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t \right|$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T} \sum_{k=1}^{K} \left\langle \phi(k, \tilde{A}_t), \left(\tilde{\mathbf{p}}_k^t - \mathbf{p}_k^* + \mathbf{p}_k^* \right) \tilde{\theta}_t - \mathbf{p}_k^* \theta_* \right\rangle (1 - O_t^{\mathrm{ALG}}) \right] + 2B\pi_{\max}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \|\phi(S_t, \tilde{A}_t')\|_{\tilde{V}_t^{-1}} \|\tilde{\theta}_t' - \theta_*\|_{\tilde{V}_t} O_t^{\mathrm{ALG}} \right] + W \sum_{t=1}^{T} \beta_t + WL \sum_{t=1}^{T} \left| \sum_{k=1}^{K} \mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t \right|$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T} \left(\left\| \sum_{k=1}^{K} \phi(k, \tilde{A}_t) \mathbf{p}_k^* \right\|_{\tilde{V}_t(\lambda)^{-1}} \|\tilde{\theta}_t - \theta_*\|_{\tilde{V}_t(\lambda)} + \sum_{k=1}^{K} \left\langle \phi(k, \tilde{A}_t), \left(\tilde{\mathbf{p}}_k^t - \mathbf{p}_k^* \right) \tilde{\theta}_t \right\rangle \right) (1 - O_t^{\mathrm{ALG}}) \right] + 2B\pi_{\max}.$$

Our assumption indicates that $\max_{a \in \mathcal{A}, s \in S} \langle \phi(s, a), \theta_* \rangle \leq 1$ with Q-probability one. Then, it is also reasonable to assume that for all t, we have $\max_{a \in \mathcal{A}, s \in S, \theta \in \tilde{C}_t} \langle \phi(s, a), \theta \rangle \leq 1$ with Q-probability one. Thus, the bandit regret

above satisfies the following:

$$\text{BLL}_T \leq \mathbb{E}\left[\sum_{t=1}^{T}\|\phi(S_t,\tilde{A}'_t)\|_{\tilde{V}_t^{-1}}\|\tilde{\theta}'_t-\theta_*\|_{\tilde{V}_t}O_t^{\text{ALG}}\right] + W\sum_{t=1}^{T}\beta_t + WL\sum_{t=1}^{T}\left|\sum_{k=1}^{K}\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t\right|$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T}\left(\left\|\sum_{k=1}^{K}\phi(k,\tilde{A}_t)\mathbf{p}_k^*\right\|_{\tilde{V}_t(\lambda)^{-1}}\|\tilde{\theta}_t-\theta_*\|_{\tilde{V}_t(\lambda)} + \sum_{k=1}^{K}\left(\tilde{\mathbf{p}}_k^t - \mathbf{p}_k^*\right)\right)(1-O_t^{\text{ALG}})\right] + 2B\pi_{\max}.$$

Furthermore, from the result of Proposition E.2, if we plug $\lambda = 1/W^2$ in, and also employ the inequality that $\frac{\det V_t(\lambda)}{\lambda^d} \leq \left(\text{trace}\left(\frac{V_t(\lambda)}{\lambda d}\right)\right)^d \leq \left(1+\frac{TL^2}{\lambda d}\right)^d$, we can get that the width of the confidence interval is bounded by $\gamma = 1 + \sqrt{2\log\left(\frac{1}{\delta}\right) + d\log(1+\frac{TW^2L^2}{d})}$. We assume that the reward X_t that we get at each round is bounded by 1. Thus, we can establish the following to bound the bandit regret:

$$\text{BLL}_T \leq 2\gamma\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\min\left(1,\|\phi(S_t,\tilde{A}'_t)\|_{\tilde{V}_t^{-1}}^2\right)(O_t^{\text{ALG}})^2}\right] + (WL+1)\sum_{t=1}^{T}\left|\sum_{k=1}^{K}\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t\right|$$

$$+ 2\gamma\mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\min\left(1,\left\|\sum_{k=1}^{K}\phi(k,\tilde{A}_t)\mathbf{p}_k^*\right\|_{\tilde{V}_t^{-1}(\lambda)}^2\right)(1-O_t^{\text{ALG}})^2}\right] + W\sum_{t=1}^{T}\beta_t + 2B\pi_{\max}$$

$$\leq 2\gamma\max(c_{\max},1)\sqrt{2Td\log\left(\frac{\text{trace}(V_0)+TL^2}{d\det(V_0)^{1/d}}\right)} + W\sum_{t=1}^{T}\beta_t + (WL+1)\sum_{t=1}^{T}\left|\sum_{k=1}^{K}\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t\right| + 2B\pi_{\max}$$

$$\leq \max(c_{\max},1)\sqrt{8Td\gamma^2\log\left(\frac{d+TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + (WL+1)\sum_{t=1}^{T}\sum_{k=1}^{K}\left|\mathbf{p}_k^* - \tilde{\mathbf{p}}_k^t\right| + 2B\pi_{\max},$$

where the first inequality is due to the fact that O_t^{ALG} is binary, and the last inequality follows from Lemma 19.4 of Lattimore and Szepesvári [2020]. Finally, using the high-probability bound we established in Lemma F.2 for the latent context distribution, we complete the proof by establishing the following bound:

$$\text{BLL}_T \leq \max(c_{\max},1)\sqrt{8Td\gamma^2\log\left(\frac{d+TW^2L^2}{d}\right)} + W\sum_{t=1}^{T}\beta_t + (WL+1)\sum_{t=1}^{T}\sum_{k=1}^{K}\zeta_t + 2B\pi_{\max},$$

where $\zeta_t = \sqrt{\frac{2\,\tilde{\mathbf{p}}_k^t\left(1-\tilde{\mathbf{p}}_k^t\right)\log(\frac{2KT}{\delta})}{\max\{m(k,t),1\}}} + \frac{7\log(\frac{2KT}{\delta})}{3(\max\{m(k,t)-1,1\})}$. $\qquad\square$

G   Additional Synthetic Experimental Results

All experiments were run on a high performance computing cluster with a 12-core CPU.

B	2	4	8	16	32	64
PD1-UCB	0.805	0.832	0.849	0.861	0.870	0.879
PD2-UCB	0.759	0.807	0.836	0.854	0.867	0.878

Table G.1: Empirical competitive ratio for one value of θ_* averaged over 200 context arrival sequences.
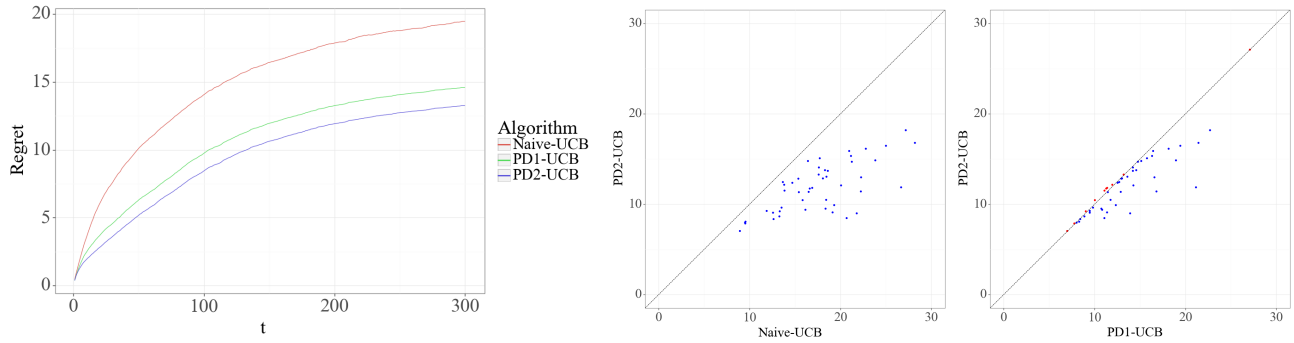
Figure G.1: Average regret (left) and scatter plot for $B = 20$ at $T = 300$ under known $\mathbf{p}^*$. Each dot corresponds to one instance averaged over 50 replications.
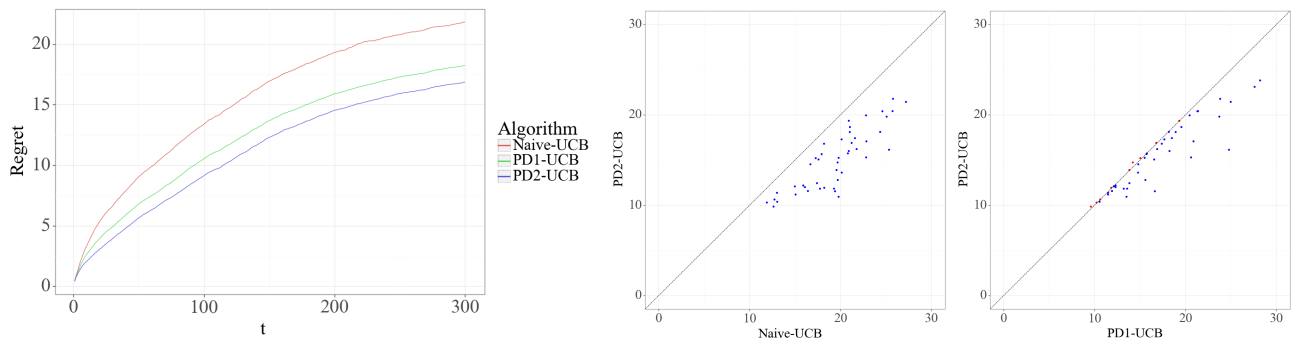

Figure G.2: Average regret (left) and scatter plots for $B = 30$ at $T = 300$ under known $\mathbf{p}^*$. Each dot corresponds to one instance averaged over 50 replications.
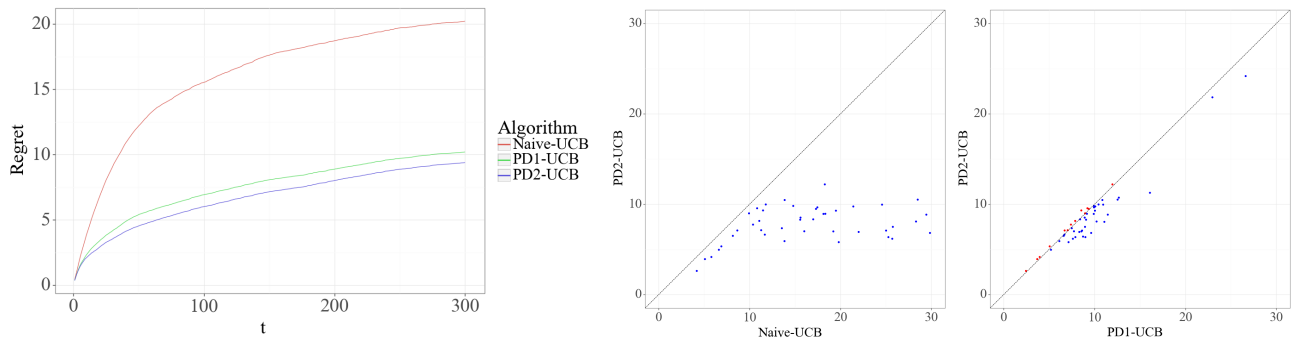

Figure G.3: Average regret (left) and scatter plots for $B = 10$ at $T = 300$ under unknown $\mathbf{p}^*$. Each dot corresponds to one instance averaged over 50 replications.
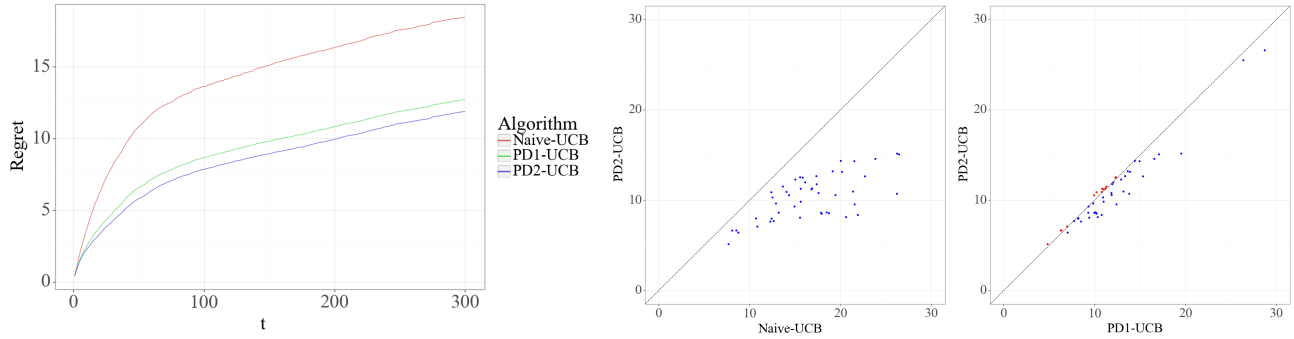
Figure G.4: Average regret (left) and scatter plots for $B = 20$ at $T = 300$ under unknown $\mathbf{p}^*$. Each dot corresponds to one instance averaged over 50 replications.
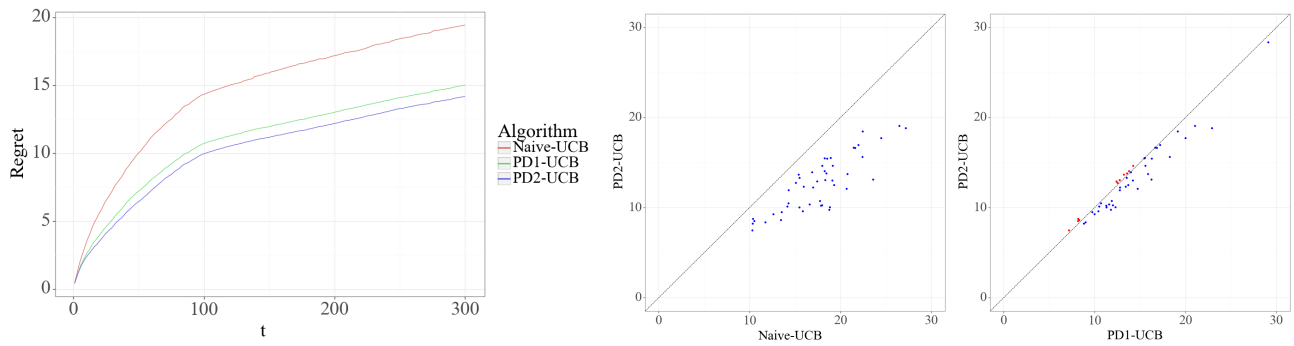

Figure G.5: Average regret (left) and scatter plots for $B = 30$ at $T = 300$ under unknown $\mathbf{p}^*$. Each dot corresponds to one instance averaged over 50 replications.

H   Real-World Experiments on ROBAS 2 and 3

The simulation environment yields brushing quality $Q_{i,t}$ in response to action $A_{i,t}$ in state $S_{i,t}$, based on the following mathematical formulation:

$$Z_{i,t} \sim \text{Bern}\left(1 - \widetilde{p}_t\right),$$
$$\widetilde{p}_t = \text{sigmoid}\left(g\left(S_{i,t}\right)^T w_b - A_{i,t} \max\left(h\left(S_{i,t}\right)^T \Delta_B, 0\right)\right),$$
$$Y_{i,t} \sim \text{Pois}\left(\lambda_{i,t}\right),$$
$$\lambda_{i,t} = \exp\left(g\left(S_{i,t}\right)^T w_p + A_{i,t} \max\left(h\left(S_{i,t}\right)^T \Delta_N, 0\right)\right),$$
$$Q_{i,t} = Z_{i,t} Y_{i,t}.$$

Here, $g\left(S_{i,t}\right)$ is the baseline feature vector, and $h\left(S_{i,t}\right)$ represents the feature vector interacting with the effect size. Similar to Trella et al. [2022], we consider binary actions (sending a message versus not sending a message). For further details, refer to Trella et al. [2022].