Online Learning in Bandits with Predicted Context

Yongyi Guo[*]
University of Wisconsin-Madison

Ziping Xu[*]
Harvard University

Susan A. Murphy
Harvard University

Abstract

We consider the contextual bandit problem where at each time, the agent only has access to a noisy version of the context and the error variance (or an estimator of this variance). This setting is motivated by a wide range of applications where the true context for decision-making is unobserved, and only a prediction of the context by a potentially complex machine learning algorithm is available. When the context error is non-vanishing, classical bandit algorithms fail to achieve sublinear regret. We propose the first online algorithm in this setting with sublinear regret guarantees under mild conditions. The key idea is to extend the measurement error model in classical statistics to the online decision-making setting, which is nontrivial due to the policy being dependent on the noisy context observations. We further demonstrate the benefits of the proposed approach in simulation environments based on synthetic and real digital intervention datasets.

1 Introduction

Contextual bandits (Auer, 2002; Langford and Zhang, 2007) represent a classical sequential decision-making problem where an agent aims to maximize cumulative reward based on context information. At each round t, the agent observes a

context and must choose one of K available actions based on both the current context and previous observations. Once the agent selects an action, she observes the associated reward, which is then used to refine future decision-making. Contextual bandits are typical examples of reinforcement learning problems where a balance between exploring new actions and exploiting previously acquired information is necessary to achieve optimal long-term rewards. It has numerous real-world applications including personalized recommendation systems (Li et al., 2010; Bouneffouf et al., 2012), healthcare (Yom-Tov et al., 2017; Liao et al., 2020), and online education (Liu et al., 2014; Shaikh et al., 2019).

Despite the extensive existing literature on contextual bandits, in many real-world applications, the agent never observes the context *exactly*. One common reason is that the true context for decision-making can only be detected or learned approximately from observable auxiliary data. For instance, consider the Sense2Stop mobile health study, in which the context is whether the individual is currently physiologically stressed (Battalio et al., 2021). A complex predictor of current stress was constructed/validated based on multiple prior studies (Cohen et al., 1985; Sarker et al., 2016, 2017). This predictor was then tuned to each user in Sense2Stop prior to their quit-smoke attempt and then following the user's attempt to quit smoking, at each minute, the predictor inputs high-dimensional sensor data on the user and outputs a continuous likelihood of stress for use by the decision-making algorithm. In many such applications in health interventions, models using validated predictions as contexts are preferred to raw sensor data because of the high noise in these settings, and that the decision rules are interpretable so they can be critiqued by domain experts. The

---

[*]: the authors have equal contribution.

second reason why the context is not observed exactly is because of measurement error. Contextual variables, such as user preferences in online advertising, content attributes in recommendation systems, and patient conditions in clinical trials, are prone to noisy measurement. This introduces an additional level of uncertainty that must be accounted for when making decisions.

Motivated by the above, we consider the linear contextual bandit problem where at each round, the agent only has access to a noisy observation of the true context. Moreover, the agent has limited knowledge about the underlying distribution of this noisy observation, as in many practical applications (e.g. the above-mentioned ones). This is especially the case when this 'observation' is the output of a complex machine learning algorithm. We only assume that the noisy observation is unbiased, its variance is known or can be estimated, and we put no other essential restrictions on its distribution. In healthcare applications, the estimated error variance for context variables can often be derived from data in prior studies (e.g. pilot studies). This setting is intrinsically difficult for two main reasons: First, when estimating the reward model, the agent needs to take into account the misalignment between the noisy context observation and the reward which depends on the true context. Second, even if the reward model is known, the agent may suffer from making bad decisions at each round because of the inaccurate context.

Our contributions. We present the first online algorithm MEB (Measurement Error Bandit) with sublinear regret in this setting under mild conditions. MEB achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret compare to a standard benchmark and $\widetilde{\mathcal{O}}(T^{1/2})$ regret compare to a clipped benchmark with minimum exploration probability which is common in many applications (Yang et al., 2020a; Yao et al., 2021). MEB is based on a novel approach to model estimation which removes the systematic bias caused by the noisy context observation. The estimator is inspired by the measurement error literature in statistics (Carroll et al., 1995; Fuller, 2009): We extend this classical method with additional tools in the online decision-making setting due to the policy being dependent on the measurement error.

1.1 Related work

Our work complements several lines of literature in contextual bandits, as listed below.

Latent contextual bandit. In the latent contextual bandit literature (Zhou and Brunskill, 2016; Sen et al., 2017; Hong et al., 2020a,b; Xu et al., 2021; Nelson et al., 2022; Galozy and Nowaczyk, 2023), the reward is typically modeled as jointly depending on the latent state, the context, and the action. Several works (Zhou and Brunskill, 2016; Hong et al., 2020a,b; Galozy and Nowaczyk, 2023) assume no direct relation between the latent state and the context while setting a parametric reward model. For example, Hong et al. (2020a) assumes the latent state $s \in \mathcal{S}_l$ is unknown but constant over time. Hong et al. (2020b) assume that the latent state evolves through a Markov chain. Xu et al. (2021) sets a specific context as well as the latent feature for each action, and models the reward depending on them through a generalized linear model. Different from the aforementioned studies, we specify that the observed context is a noisy version of the latent context (which aligns with the applications we are addressing), and then leverage this structure to design the online algorithm.

In another line of work, Sen et al. (2017); Nelson et al. (2022) consider contextual bandit with latent confounder, where the observed context influences the reward through a latent confounder variable, which ranges within a small discrete set. Our setting is distinct from these works in that the latent context can span an infinite (or even continuous) space.

Bandit with noisy context. In bandit with noisy context (Yun et al., 2017; Kirschner and Krause, 2019; Yang et al., 2020b; Lin and Moothedath, 2022; Lin et al., 2022), the agent has access to a noisy version of the true context and/or some knowledge of the distribution of the noisy observation. Yun et al. (2017); Park and Faradonbeh (2022); Jose and Moothedath (2024) consider settings where the joint distribution of the true context and the noisy observed context is known (up to a parameter). Other works, such as Kirschner and Krause (2019); Yang et al. (2020b); Lin et al. (2022), assume that the agent knows the exact distribution of the context each time, and observes no context. By assuming a linear reward model, Kirschner and Krause (2019)

2

transforms the problem into a linear contextual bandit, and obtains $\widetilde{\mathcal{O}}(d\sqrt{T})$ regret compared to the policy maximizing the expected reward over the context distribution. Yang et al. (2020b); Lin et al. (2022); Lin and Moothedath (2022) consider variants of the problem such as multiple linear feedbacks, multiple agents, and delayed observation of the exact context. Compared to these works, we consider a practical but more challenging setting where besides an unbiased noisy observation(prediction) for each context, the agent only knows the second-moment information about the distribution. This does not transform into a standard linear contextual bandit as in Kirschner and Krause (2019).

Bandit with inaccurate/corrupted context. These works consider the setting where the context is simply inaccurate (without randomness), or is corrupted and cannot be recovered. In Yang and Ren (2021a,b), at each round, only an inaccurate context is available to the decision-maker, and the exact context is revealed after the action is taken. In Bouneffouf (2020); Galozy et al. (2020), each context $\mathbf{x}_t$ is completely corrupted with some probability and the corruption cannot be recovered. In Ding et al. (2022), the context is attacked by an adversarial agent. Because these works focus more on adversarial settings for the context observations, the application of their regret bounds to our setting generally results in a linear regret. For example, in Yang and Ren (2021a), the regret of Thompson sampling is $\widetilde{\mathcal{O}}(d\sqrt{T} + \sqrt{d}\sum_{t\in[T]}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|_2)$, where $\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|_2$ is the error of the inaccurate context. As is typical in our setting, $\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|_2$ will be non-vanishing through time, so the second term is linear in T. Given the applications we consider, we can exploit the stochastic nature of the noisy context observations in our algorithm to achieve improved performance.

1.2 Notations

Throughout this paper, we use $[n]$ to represent the set $\{1, 2, \ldots, n\}$ for $n \in \mathbb{N}^+$. For $a, b \in \mathbb{R}$, let $a \wedge b$ denote the minimum of a and b. Given $d \in \mathbb{N}^+$, $\boldsymbol{I}_d$ denotes the d-by-d identity matrix, and $\mathbf{1}_d$ denotes the d-dimensional vector with 1 in each entry. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, denote $\|\boldsymbol{v}\|_2$ as its ℓ_2 norm. For a matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, denote $\|\boldsymbol{M}\|_2$ as its operator norm. The notation $\mathcal{O}(X)$ refers to a quantity that is upper bounded by X up to constant multiplicative factors, while $\widetilde{\mathcal{O}}(X)$ refers to a quantity that is upper bounded by X up to poly-log factors.

2 Measurement error adjustment to bandit with noisy context

2.1 Problem setting

We consider a linear contextual bandit with context space $\mathcal{X} \subset \mathbb{R}^d$ and binary action space $\mathcal{A} = \{0, 1\}$[1]. Let T be the time horizon. As discussed above, we consider the setting where at each time $t \in [T]$, the agent only observes a noisy version of the context $\widetilde{\mathbf{x}}_t$ instead of the true underlying context $\mathbf{x}_t$. Thus, at time t, the observation o_t only contains $(\widetilde{\mathbf{x}}_t, a_t, r_t)$, where a_t is the action and r_t is the corresponding reward. We further assume that $\widetilde{\mathbf{x}}_t = \mathbf{x}_t + \boldsymbol{\epsilon}_t$, where the error $\boldsymbol{\epsilon}_t$ is independent of the history $\mathcal{H}_{t-1} := \{o_\tau\}_{\tau \leq t-1}$, $\mathbb{E}\boldsymbol{\epsilon}_t = 0$, $\mathrm{Var}(\boldsymbol{\epsilon}_t) = \boldsymbol{\Sigma}_{e,t}$. Here, 'e' in the subscript means 'error'. Initially we assume that $(\boldsymbol{\Sigma}_{e,t})_{t\geq1}$ is known. In Section 2.4, we consider the setting where only estimators of $(\boldsymbol{\Sigma}_{e,t})_{t\geq1}$ are available. There is no restriction that the distribution of $(\boldsymbol{\epsilon}_t)_{t\geq1}$ belongs to any known (parametric) family. The reward $r_t = \langle \boldsymbol{\theta}_{a_t}^*, \mathbf{x}_t \rangle + \eta_t$, where $\mathbb{E}[\eta_t | \mathcal{H}_{t-1}, \boldsymbol{\epsilon}_t, a_t] = 0$ and $(\boldsymbol{\theta}_a^*)_{a \in \mathcal{A}}$ are the unknown parameters. It's worth noting that besides the policy, all the randomness here comes from the reward noise η_t and the context error $\boldsymbol{\epsilon}_t$. We treat $(\mathbf{x}_t)_{t\geq1}$ as fixed throughout but unknown to the algorithm (Unlike Yun et al. (2017), we don't assume $\mathbf{x}_t$ are i.i.d.). Our goal at each time t is to design policy $\pi_t(\cdot|\mathcal{H}_{t-1}, \widetilde{\mathbf{x}}_t) \in \Delta(\mathcal{A})$ given past history $\mathcal{H}_{t-1}$ and current observed noisy context $\widetilde{\mathbf{x}}_t$, so that the agent can maximize the reward by taking action $a_t \sim \pi_t(\cdot|\mathcal{H}_{t-1}, \widetilde{\mathbf{x}}_t)$.

If $\boldsymbol{\Sigma}_{e,t}$ is non-vanishing, standard contextual bandit algorithms are generally sub-optimal. To see this, notice that $r_t = \langle \boldsymbol{\theta}_{a_t}^*, \mathbf{x}_t \rangle + \eta_t = \langle \boldsymbol{\theta}_{a_t}^*, \widetilde{\mathbf{x}}_t \rangle + (\eta_t - \langle \boldsymbol{\theta}_{a_t}^*, \boldsymbol{\epsilon}_t \rangle)$. This means the error in the reward r_t after observing the noisy context is $\eta_t - \langle \boldsymbol{\theta}_{a_t}^*, \boldsymbol{\epsilon}_t \rangle$, where $\boldsymbol{\epsilon}_t$ and $\widetilde{\mathbf{x}}_t$ are dependent. Thus, $\mathbb{E}[r_t | \widetilde{\mathbf{x}}_t, a_t] \neq \langle \boldsymbol{\theta}_{a_t}^*, \widetilde{\mathbf{x}}_t \rangle$. This is in contrast to the standard linear bandit set-

---

[1]For simplicity, we state our results under the binary-action setting, which is common in healthcare (Trella et al., 2022), economics (Athey et al., 2017; Kitagawa and Tetenov, 2018) and other applications. However, all the results presented in this paper can be extended to the setting with multiple actions. See Appendix C.

ting, where given the true context $\mathbf{x}_t$, $\mathbb{E}[r_t|\mathbf{x}_t, a_t] = \langle \boldsymbol{\theta}^*_{a_t}, \mathbf{x}_t \rangle$, which ensures the sublinear regret of classical bandit algorithms such as UCB and Thompson sampling. Therefore, it is necessary to design an online algorithm that adjusts for the errors $(\boldsymbol{\epsilon}_t)_{t \geq 1}$. We assume that the context, parameters and the reward are bounded, as below.

Assumption 2.1 (Boundedness). $\forall t \in [T]$, $\|\widetilde{\mathbf{x}}_t\|_2 \leq 1$; There exists a positive constant R_θ such that $\forall a \in \{0, 1\}$, $\|\boldsymbol{\theta}^*_a\|_2 \leq R_\theta$; There exists a positive constant R such that $\forall t \in [T]$, $|r_t| \leq R$.

For any policy $\pi = (\pi_t)_t$, we define the (standard) cumulative regret as

$$\text{Regret}(T; \pi^*) = \sum_{t \in [T]} [\mathbb{E}_{a \sim \pi^*_t} \langle \boldsymbol{\theta}^*_a, \mathbf{x}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}^*_a, \mathbf{x}_t \rangle],$$
$$(2.1)$$

where

$$\pi^*_t(a) = \begin{cases} 1, & \text{if } a = a^*_t := \text{argmax}_a \langle \boldsymbol{\theta}^*_a, \mathbf{x}_t \rangle, \\ 0, & \text{otherwise.} \end{cases} \quad (2.2)$$

We denote the standard benchmark policy $\pi^* = (\pi^*_t)_t$. This is summarized in the setting below.

Setting 1. (Standard setting) We aim to minimize $\text{Regret}(T; \pi^*)$ among the class of all policies.

In many applications including clinical trials, it's desirable to design the policy under the constraint that each action is sampled with a minimum probability p_0. One reason for maintaining exploration is that we can update and re-optimize the policy for future users to allow for potential non-stationarity (Yang et al., 2020a). Additionally, keeping the exploration is also important for after-study analysis (Yao et al., 2021), especially when the goal of the analysis is not pre-specified prior to collecting data with the online algorithm. In these situations, it is desirable to consider only the policies that always maintain an exploration probability of $p_0 > 0$ for each arm, and compare the performance to the clipped benchmark policy $(\bar{\pi}^*_t)$:

$$\bar{\pi}^*_t(a) = \begin{cases} 1 - p_0, & \text{if } a = a^*_t, \\ p_0, & \text{otherwise.} \end{cases} \quad (2.3)$$

This is summarized in the setting below.

Setting 2. (Clipped policy setting) We minimize

$$\text{Regret}(T; \bar{\pi}^*) = \sum_{t \in [T]} [\mathbb{E}_{a \sim \bar{\pi}^*_t} \langle \boldsymbol{\theta}^*_a, \mathbf{x}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}^*_a, \mathbf{x}_t \rangle]$$

among the class of policies that explore any action with probability at least p_0.

In this work, we will provide policies with sublinear regret guarantees in both settings.

2.2 Estimation using weighted measurement error adjustment

In this section, we focus on learning the reward model parameters $(\boldsymbol{\theta}^*_a)_{a \in \mathcal{A}}$ with data $\mathcal{H}_t$ after a policy $(\pi_\tau(\cdot|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1}))_{\tau \in [t]}$ has been executed up to time $t \in [T]$. Learning a consistent model is important in many bandit algorithms for achieving low regret (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013). As we shall see in Section 2.3, consistent estimation of $(\boldsymbol{\theta}^*_a)_{a \in \mathcal{A}}$ plays an essential role in controlling the regret of our proposed algorithm.

Inconsistency of the regularized least-squares (RLS) estimator. UCB and Thompson sampling, the two classical bandit algorithms, both achieve sublinear regret based on the consistency of the estimator $\widehat{\boldsymbol{\theta}}^{(t)}_{a,RLS} = (\lambda I + \sum_{\tau \in [t]} 1_{\{a_\tau = a\}} \mathbf{x}_\tau \mathbf{x}_\tau^\top)^{-1} (\sum_{\tau \in [t]} 1_{\{a_\tau = a\}} \mathbf{x}_\tau r_\tau)$ under certain norms. When $\widetilde{\mathbf{x}}_\tau = \mathbf{x}_\tau + \boldsymbol{\epsilon}_\tau$ is observed instead of $\mathbf{x}_\tau$, the RLS estimator becomes $\widehat{\boldsymbol{\theta}}^{(t)}_{a,RLSCE} = (\lambda I + \sum_{\tau \in [t]} 1_{\{a_\tau = a\}} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top)^{-1} (\sum_{\tau \in [t]} 1_{\{a_\tau = a\}} \widetilde{\mathbf{x}}_\tau r_\tau)$. Here 'RLSCE' means the RLS estimator with contextual error. However, when $\boldsymbol{\Sigma}_{e,\tau} = \text{Var}(\boldsymbol{\epsilon}_\tau)$ is non-vanishing, $\widehat{\boldsymbol{\theta}}^{(t)}_{a,RLSCE}$ is generally no longer consistent, which may lead to bad decision-making (see Appendix B for details). In the simple case where $(\mathbf{x}_\tau, \boldsymbol{\epsilon}_\tau, \eta_\tau)_{\tau \in [t]}$ are i.i.d. and there is no action (i.e. set $a_\tau \equiv 0$), the inconsistency of $\widehat{\boldsymbol{\theta}}^{(t)}_{a,RLSCE}$ is studied in the measurement error model literature in statistics (Fuller, 2009; Carroll et al., 1995), and is known as *attenuation*.

A naive measurement error adjustment. A measurement error model is a type of regression model designed to accommodate inaccuracies in the measurement of regressors (i.e., instead of observing $\mathbf{x}_t$, we observe $\mathbf{x}_t + \boldsymbol{\epsilon}_t$ where $\boldsymbol{\epsilon}_t$ is a noise term with zero mean). As conventional regression techniques yield inconsistent estimators, measurement error models rectify this issue with adjustments to the estimator that consider these errors. In the current context when we want to estimate $\boldsymbol{\theta}^*_a$ from history $\mathcal{H}_t$, $(\widetilde{\mathbf{x}}_\tau)_{\tau \in [t]}$ can be viewed as regressors 'mea-

sured with error', while $(r_\tau)_{\tau\in[t]}$ are dependent variables. If $(\boldsymbol{\epsilon}_\tau, \eta_\tau)_{\tau\in[t]}$ are i.i.d., $\boldsymbol{\Sigma}_{e,\tau} \equiv \boldsymbol{\Sigma}_e$, and there is no action (i.e. set $a_\tau \equiv 0$), $\widehat{\boldsymbol{\theta}}_{0,me}^{(t)} := \left(\frac{1}{t}\sum_{\tau\in[t]}\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top - \boldsymbol{\Sigma}_e\right)^{-1}\left(\frac{1}{t}\sum_{\tau\in[t]}\widetilde{\mathbf{x}}_\tau r_\tau\right)$ is a consistent estimator for $\boldsymbol{\theta}_0^*$. When multiple actions are present, a naive generalization of the above estimator, $\widehat{\boldsymbol{\theta}}_{a,me}^{(t)}$, is

$$\left(\sum_{\tau\in[t]}1_{\{a_\tau=a\}}(\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top - \boldsymbol{\Sigma}_e)\right)^{-1}\cdot\left(\sum_{\tau\in[t]}1_{\{a_\tau=a\}}\widetilde{\mathbf{x}}_\tau r_\tau\right). \quad (2.4)$$

Unfortunately, $\widehat{\boldsymbol{\theta}}_{a,me}^{(t)}$ is inconsistent in the multiple-action setting, even if the policy $(\pi_\tau)_{\tau\in[t]}$ is *stationary and not adaptive to history*. This difference is essentially due to the interaction between the policy and the measurement error: In the classical measurement error (no action) setting, $\mathbb{E}(\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top) = \mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Sigma}_e$, and $\frac{1}{t}\sum_{\tau\in[t]}\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top - \boldsymbol{\Sigma}_e$ concentrates around its expectation $\frac{1}{t}\sum_{\tau\in[t]}\mathbf{x}_\tau\mathbf{x}_\tau^\top$. Likewise, $\frac{1}{t}\sum_{\tau\in[t]}\widetilde{\mathbf{x}}_\tau r_\tau$ concentrates around $(\frac{1}{t}\sum_{\tau\in[t]}\mathbf{x}_\tau\mathbf{x}_\tau^\top)\boldsymbol{\theta}_0^*$. Combining the above parts thus yields a consistent estimator of $\boldsymbol{\theta}_0^*$. In our setting with multiple actions, however, the agent picks the action a_τ *based on* $\widetilde{\mathbf{x}}_\tau$, so only certain values of $\widetilde{\mathbf{x}}_\tau$ lead to $a_\tau = a$. Therefore, for those $\tau \in [t]$ when we pick $a_\tau = a$, it's more likely that $\widetilde{\mathbf{x}}_\tau$ falls in certain regions depending on the policy, and we shouldn't expect $\mathbb{E}(\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top) = \mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Sigma}_e$ anymore. In other words, the policy creates a complicated dependence between $\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top$ and $1_{\{a_\tau=0\}}$ for each τ, which changes the limit of $\frac{1}{t}\sum_{\tau\in[t]}1_{\{a_\tau=a\}}(\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top - \boldsymbol{\Sigma}_e)$ (and similarly $\frac{1}{t}\sum_{\tau\in[t]}1_{\{a_\tau=a\}}\widetilde{\mathbf{x}}_\tau r_\tau$). This leads to the inconsistency of the naive estimator (See Appendix B for a concrete example). In Section 3, we provide examples to show that (2.4) not only deviates from the true parameters, but also leads to suboptimal decision-making.

Our proposed estimator. Inspired by the above observations, for $\pi_\tau(a|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})$ positive, we construct the following estimator for $\boldsymbol{\theta}_a^*$ given $\mathcal{H}_t$, which corrects (2.4) using importance weights:

$$\widehat{\boldsymbol{\theta}}_a^{(t)} := \left(\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\boldsymbol{\Sigma}_{e,\tau}\right)^{-1}\cdot\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)}, \quad (2.5)$$

where

$$\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} = \frac{1}{t}\sum_{\tau\in[t]}\frac{\pi_\tau^{nd}(a_\tau)}{\pi_\tau(a_\tau|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})}1_{\{a_\tau=a\}}\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top,$$

$$\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)} = \frac{1}{t}\sum_{\tau\in[t]}\frac{\pi_\tau^{nd}(a_\tau)}{\pi_\tau(a_\tau|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})}1_{\{a_\tau=a\}}\widetilde{\mathbf{x}}_\tau r_\tau.$$

Here, $(\pi_\tau^{nd}(\cdot))_{\tau\in[t]}$ is a pre-specified policy (doesn't depend on $(\widetilde{\mathbf{x}}_\tau)_\tau$ or $\mathcal{H}_{\tau-1}$) that can be chosen by the algorithm. We only require the following:

Assumption 2.2. Let $\boldsymbol{w}_t = \sqrt{d}\cdot\mathbf{x}_t$, $\widetilde{\boldsymbol{w}}_t = \sqrt{d}\cdot\widetilde{\mathbf{x}}_t$ be scaled vectors of $\mathbf{x}_t$ and $\widetilde{\mathbf{x}}_t$. Then there exist positive constants ξ, ρ, λ_0 s.t. (i) $\forall\boldsymbol{u}\in\mathbb{S}^{d-1}$, $\mathbb{E}(\boldsymbol{u}^\top\widetilde{\boldsymbol{w}}_t)^4 \le \xi$, (ii) $\mathbb{E}\widetilde{\boldsymbol{w}}_t\widetilde{\boldsymbol{w}}_t^\top \preceq \nu\boldsymbol{I}_d$, and (iii) $\forall t \ge \sqrt{T}$, $a \in \{0,1\}$, $\frac{1}{t}\sum_{\tau\le t}\pi_\tau^{nd}(a)\boldsymbol{w}_\tau\boldsymbol{w}_\tau^\top \succeq \lambda_0\boldsymbol{I}_d$.

Remark 2.1. In Assumption 2.2, (i) and (ii) are standard moment assumptions. (iii) is mild. Even restricted to the choice of $\pi_\tau^{nd}(a) \equiv 1/2$, under mild conditions, the assumption can be satisfied with deterministic $(\mathbf{x}_\tau)_{\tau\ge 1}$ or stochastic $(\mathbf{x}_\tau)_{\tau\ge 1}$ such as an i.i.d. sequence, a weakly dependent stationary time series (e.g. multivariate ARMA process (Fan and Yao, 2017)), or a sequence with periodicity/seasonality with high probability (See Appendix D for details).

The theorem below gives a high-probability upper bound on $\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2$ (proof in Appendix D).

Theorem 2.1. For any $t \in [T]$, denote $q_t := \inf_{\tau\le t, a\in\{0,1\}}\pi_\tau(a|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})$. Then under Assumptions 2.1 and 2.2, there exist absolute constants C, C_1, such that as long as $q_t \ge C_1\max\{\frac{d(d+\log t)}{\lambda_0 t}, \frac{\xi(d+\log t)}{\lambda_0^2 t}\}$, with probability at least $1-\frac{8}{t^2}$, $\forall a \in \{0,1\}$, $\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2$ is upper bounded by

$$\frac{C(R+R_\theta)d}{\lambda_0}\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\nu}+\sqrt{\xi}}{\sqrt{d}}\sqrt{\frac{d+\log t}{q_t t}}\right\}. \quad (2.6)$$

Unlike the existing literature on off-policy learning in contextual bandits (e.g. Wang et al. (2017); Zhan et al. (2021); Zhang et al. (2021); Bibaut et al. (2021)), the role of the importance weights here is to correct the dependence of a policy on the observed noisy context with error. The proof idea can be generalized to a large class of off-policy method-of-moment estimators, which might be of independent interest (see Appendix D).

5

Algorithm 1 MEB (Measurement Error Bandit)

1: **Input:** $(\mathbf{\Sigma}_{e,t})_{t\in[T]}$: variance sequence of $(\boldsymbol{\epsilon}_t)_{t\in[T]}$; $(p_0^{(t)})_{t\in[T]}$: minimum selection probability at time $t \in [T]$; T_0: warm-up stage length

2: **for** time $t = 1, 2, \ldots, T$ **do**
3:    **if** $t \leq T_0$ **then**
4:       Set $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) \in [p_0^{(t)}, 1-p_0^{(t)}]$, $a \in \{0,1\}$
5:       Sample $a_t \sim \pi_t(\cdot|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$
6:    **else**
7:       Obtain $(\widehat{\boldsymbol{\theta}}_a^{(t-1)})_{a\in\{0,1\}}$ from (2.5)
8:       $\widetilde{a}_t \leftarrow \arg\max_{a\in\{0,1\}} \langle \widehat{\boldsymbol{\theta}}_a^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle$
9:       Set $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) := \begin{cases} 1 - p_0^{(t)}, & \text{if } a = \widetilde{a}_t \\ p_0^{(t)}, & \text{otherwise} \end{cases}$
10:       Sample $a_t \sim \pi_t(\cdot|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$
11:    **end if**
12: **end for**

2.3 MEB: Online bandit algorithm with measurement error adjustment

We propose MEB (Measurement Error Bandit), an online bandit algorithm with measurement error adjustment based on the estimator (2.5). The algorithm is presented in Algorithm 1 and is designed for the binary-action setting, although it can be generalized to the case with multiple actions (see Appendix C). For $t \leq T_0$, the algorithm is in a warm-up stage and can pick any policy such that there is a minimum sampling probability $p_0^{(t)}$ for each action (Here $p_0^{(t)} \in (0, \frac{1}{2}]$). For instance, the algorithm can do pure exploration with $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) \equiv \frac{1}{2}$. For $t > T_0$, given the noisy context $\widetilde{\mathbf{x}}_t$, the algorithm computes the best action $\widetilde{a}_t$ according to $(\widehat{\boldsymbol{\theta}}_a^{(t-1)})_{a\in\{0,1\}}$ calculated from (2.5). Then, it samples $\widetilde{a}_t$ with probability $1 - p_0^{(t)}$ and keeps an exploration probability of $p_0^{(t)}$ to sample the other action. In practice, we can often set $p_0^{(t)}$ to be monotonically decreasing in t, in which case $q_t = \inf_{\tau \leq t, a\in\{0,1\}} \pi_\tau(a|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1}) = p_0^{(t)}$ for all $t \in [T]$.

Before presenting the regret analysis, we should first note that our problem is harder than a standard contextual bandit: $\mathbf{x}_t$ is unknown, and only $\widetilde{\mathbf{x}}_t$ is observed. Thus, even if $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$ is known, we may still perform suboptimally if $\widetilde{\mathbf{x}}_t$ is too far from $\mathbf{x}_t$ so that it leads to a different optimal action. Example 2.1 below shows that in general, we cannot avoid a linear regret.

Example 2.1. Let $d = 1$, $(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_0^*) = (1, -1)$. $(\mathbf{x}_t)_{t\in[T]}$ are drawn i.i.d. from $\{\pm 0.2\}$ with equal probability. $\mathbb{P}(\widetilde{\mathbf{x}}_t = 1|\mathbf{x}_t = 0.2) = 0.6$, $\mathbb{P}(\widetilde{\mathbf{x}}_t = -1|\mathbf{x}_t = 0.2) = 0.4$; $\mathbb{P}(\widetilde{\mathbf{x}}_t = 1|\mathbf{x}_t = -0.2) = 0.4$, $\mathbb{P}(\widetilde{\mathbf{x}}_t = -1|\mathbf{x}_t = -0.2) = 0.6$. Intuitively, even if we know $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$, there is still a constant probability at each time t that we cannot make the right choice due to $\widetilde{\mathbf{x}}_t$ and $\mathbf{x}_t$ having different signs, and $\mathbf{x}_t$ is never known (details in Appendix E). This results in a $\Omega(T)$ regret.

Fortunately, in practice, we expect that the errors $(\boldsymbol{\epsilon}_t)_{t\in[T]}$ are relatively 'small' in the sense that the optimal action (given $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$) is not affected. Specifically, we assume the following:

Assumption 2.3. There exist a constant $\rho \in (0,1)$ such that $\forall t \in [T]$, $|\langle \boldsymbol{\delta}_\theta, \boldsymbol{\epsilon}_t \rangle| \leq \rho|\langle \boldsymbol{\delta}_\theta, \mathbf{x}_t \rangle|$ almost surely. Here $\boldsymbol{\delta}_\theta := \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*$.

Assumption 2.3 ensures that the perturbation to the suboptimality gap between the two arms caused by $\boldsymbol{\epsilon}_t$ is controlled by the true suboptimality gap. In this way, given $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$, the optimal action based on $\widetilde{\mathbf{x}}_t$ will not deviate too much from that based on $\mathbf{x}_t$. As a special case, this assumption is satisfied with if $\forall t$, $|\langle \boldsymbol{\delta}_\theta, \mathbf{x}_t \rangle| \geq B_{e,t}\|\boldsymbol{\delta}_\theta\|_2/\rho$. Here $B_{e,t}$ is an upper bound of $\|\boldsymbol{\epsilon}_t\|_2$. Assumption 2.3 can be further weakened to the inequalities holding with high probability (see Appendix E). Note that Assumption 2.3 only guarantees the optimal action is not affected by $(\boldsymbol{\epsilon}_t)_{t\in[T]}$ *given* $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$. To achieve sublinear regret, $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$ still needs to be well-estimated. Thus, even with Assumption 2.3, classical bandit algorithms such as UCB may still suffer from linear regret because of the inconsistent estimator $\widehat{\boldsymbol{\theta}}_{a,RLSCE}^{(t)}$ (see Appendix B for a concrete example).

We first prove the following theorem, which states that the regret of MEB can be directly controlled by the estimation error. In fact, this theorem holds regardless of the form or quality of the estimation procedure (i.e. in line 7 of Algorithm 1). The proof is in Appendix E.

Theorem 2.2. Let Assumption 2.1 and 2.3 hold.

(i) For the standard setting, Algorithm 1 outputs a policy with Regret$(T; \pi^*)$ no more than

$$2T_0 R_\theta + \frac{2}{1-\rho} \cdot \sum_{t=T_0+1}^{T} \left(p_0^{(t)} R_\theta + \max_{a\in\{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2\right).$$

(ii) For the clipped policy setting, Algorithm 1 with the choice of $p_0^{(t)} \equiv p_0$ outputs a policy with $\mathrm{Regret}(T; \bar{\pi}^*)$ no more than

$$2T_0 R_\theta + \frac{2(1-2p_0)}{1-\rho} \cdot \sum_{t=T_0+1}^{T} \max_{a \in \{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2.$$

The following corollary provides regret guarantees of MEB by combining Theorem 2.1 and 2.2 (proof in Appendix E).

Corollary 2.1. Let Assumption 2.1 to 2.3 hold. There exist universal constants C, C' such that:

(i) For the standard setting, $\forall T \geq C' \max\{(1 + 1/\lambda_0^{\frac{9}{4}})(d + \log T)^3, (\xi/\lambda_0)^{\frac{9}{4}}(d + \log T)^{\frac{4}{3}}\}$, with probability at least $1 - \frac{16}{\sqrt{T}}$, Algorithm 1 with the choice of $T_0 = \lceil 2dT^{\frac{2}{3}} \rceil$, $p_0^{(t)} = \min\{\frac{1}{2}, t^{-\frac{1}{3}}\}$ outputs a policy with $\mathrm{Regret}(T; \pi^*)$ no more than

$$CdT^{\frac{2}{3}} \left\{ \frac{R_\theta}{1-\rho} + \frac{(\sqrt{\nu}+\sqrt{\xi}+1)(R+R_\theta)}{(1-\rho)\lambda_0} \sqrt{1 + \frac{\log T}{d}} \right\}.$$

(ii) For the clipped policy setting, $\forall T \geq C' \max\{(d + \log T)^2/(\lambda_0 p_0)^2, \xi^2/\lambda_0^4(1+\log T/d)^2\}$, with probability at least $1 - \frac{16}{\sqrt{T}}$, Algorithm 1 with the choice of $T_0 = \lceil 2d\sqrt{T} \rceil$ and $p_0^{(t)} \equiv p_0$ outputs a policy with $\mathrm{Regret}(T; \bar{\pi}^*)$ no more than

$$CdT^{\frac{1}{2}} \left\{ R_\theta + \frac{(\sqrt{\nu}+\sqrt{\xi}+1)(1-2p_0)(R+R_\theta)}{\sqrt{p_0}(1-\rho)\lambda_0} \sqrt{1 + \frac{\log T}{d}} \right\}.$$

Ignoring other factors, the regret upper bound is of order $\widetilde{\mathcal{O}}(dT^{2/3})$ for the standard setting, and $\widetilde{\mathcal{O}}(d\sqrt{T})$ for the clipped policy setting, depending on the horizon T and dimension d.

In certain scenarios (e.g. when d is large), it is desirable to save computational resources by updating the estimates of $(\boldsymbol{\theta}_a^*)_{a \in \{0,1\}}$ less frequently in Algorithm 1. Fortunately, low regret guarantees can still be achieved: Suppose at each time t, the agent only updates the estimators according to (2.5) at selected time points $t \in \mathcal{S} \subseteq [T]$ (in line 7); Otherwise, the agent simply makes decisions based on the most recently updated estimators. In Appendix E, we show that time points to perform the updates can be very infrequent, such as $(n^k)_{k \in \mathbb{N}^+}$ $(n \geq 2, n \in \mathbb{N}^+)$, while still achieving the same rate of regret upper bound as in Corollary 2.1.

2.4 MEB given estimated error variance

In practice, the agent might not have perfect knowledge about $\boldsymbol{\Sigma}_{e,t}$, the variance of the error $\boldsymbol{\epsilon}_t$. In this section, we discuss the situation where at each time t, the agent does not know $\boldsymbol{\Sigma}_{e,t}$, and only has a (potentially adaptive) estimator $\widehat{\boldsymbol{\Sigma}}_{e,t}$ for $\boldsymbol{\Sigma}_{e,t}$. This estimator may be derived from auxiliary data or outside knowledge. In this case, in Algorithm 1, we need to replace the estimator (2.5) with the following estimator for decision-making (i.e. in line 7 of Algorithm 1):

$$\widetilde{\boldsymbol{\theta}}_a^{(t)} := \left(\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t} \sum_{\tau \in [t]} \pi_\tau^{nd}(a) \widehat{\boldsymbol{\Sigma}}_{e,\tau} \right)^{-1} \cdot \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)}. \tag{2.7}$$

In Appendix F, we show that with this modification, the additional regret of Algorithm 1 is controlled by

$$d \cdot \sum_{t=T_0+1}^{T} \max_{a \in \{0,1\}} \|\boldsymbol{\Delta}_t(a)\|_2$$

up to a constant depending on the assumptions. Here, for each t, $\boldsymbol{\Delta}_t(a) := \frac{1}{t} \sum_{\tau \in [t]} \pi_\tau^{nd}(a)(\widehat{\boldsymbol{\Sigma}}_{e,\tau} - \boldsymbol{\Sigma}_{e,\tau})$ is the weighted average of the estimation errors $(\widehat{\boldsymbol{\Sigma}}_{e,\tau} - \boldsymbol{\Sigma}_{e,\tau})_{\tau \in [t]}$. In practice, it is reasonable to assume that $\boldsymbol{\Delta}_t(a)$ is small so as not to significantly affect the overall regret: For example, suppose the agent gathers more auxiliary data over time so that $\|d(\widehat{\boldsymbol{\Sigma}}_{e,t} - \boldsymbol{\Sigma}_{e,t})\|_2 \lesssim \sqrt{d/t}$, then the additional regret term will be $\mathcal{O}(\sqrt{dT})$ up to a constant depending on the assumptions.

3 Simulation results

In this section, we complement our theoretical analyses with simulation results on a synthetic environment with artificial noise and reward models as well as a simulation environment based on the real dataset, HeartStep V1 (Klasnja et al., 2018).

Compared algorithms. In both simulation environments, we compare the following algorithms: Thompson sampling (TS) with normal priors (Russo et al., 2018), Linear Upper Confidence Bound (UCB) approach (Chu et al., 2011), MEB (Algorithm 1), and MEB-naive (MEB plugged in with the naive measurement error estimator (2.4) instead of (2.5)). See Appendix A for a detailed description of the algorithms.

3.1 Synthetic environment

We first test our algorithms on a synthetic environment. We consider a contextual bandit environment with $d = 5$, $T = 50000$. In the reward model, we set $\boldsymbol{\theta}_0^* = (5, 6, 4, 6, 4)$, $\boldsymbol{\theta}_1^* = (6, 5, 5, 5, 5)$, and η_t drawn i.i.d. from $\mathcal{N}(0, \sigma_\eta^2)$. Let $(\mathbf{x}_t)_{t \in [T]}$ be independently sampled from $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{I}_d)$, where $\boldsymbol{\mu}_x = \mathbf{1}_d$. We further set $\boldsymbol{\Sigma}_{e,t} \equiv \boldsymbol{\Sigma}_e := \boldsymbol{I}_d/4$ and consider independent $(\boldsymbol{\epsilon}_t)_{t \in [T]}$ with Normal distribution with covariance $\boldsymbol{\Sigma}_e$. We independently generate bandit data for $n_{exp} = 100$ times, and compare among the candidate algorithms in terms of estimation quality and cumulative regret with a moderate exploration probability $p_0 = 0.2$.

3.2 HeartStep V1 simulation environment

We also construct a simulation environment with HeartSteps dataset. HeartSteps is a physical activity mobile health application, whose primary goal is to help the user prevent negative health outcomes and adopt and maintain healthy behaviors, for example, higher physical activity level. HeartSteps V1 is a 42-day mobile health trial (Dempsey et al., 2015; Klasnja et al., 2015; Liao et al., 2016), where participants are provided a Fitbit tracker and a mobile phone application. One of the intervention components is a contextually-tailored physical activity suggestion that may be delivered at any of the five user-specified times during each day. The delivery times are roughly separated by 2.5 hours.

Construction of the simulated environment. We follow the simulation setups in Liao et al. (2020). The true context at the time t is denoted by $\mathbf{x}_t$ with three main components $\mathbf{x}_t = (I_t, Z_t, B_t)$. Here, I_t is an indicator variable of whether an intervention ($A_t = 1$) is feasible (e.g. I_t is 0 when the participant is driving a car, a situation where the suggestion should not be sent). Z_t contains some features at time t. B_t is the true treatment burden, which is a function of the participant's treatment history[2]. Specifically, $B_{t+1} = \lambda B_t + 1_{\{A_t=1\}}$. We assume that $(I_t)_{t \in [T]}$ and $(Z_t)_{t \in [T]}$ are sampled i.i.d with the empirical distribution from the Heartstep V1 dataset,

---

[2]Note this violates contextual bandits assumption and leads to an MDP. We believe this is a good setup to test the robustness of our proposed approach.

and $(B_t)_{t \in [T]}$ is given by the aforementioned transition model.

The reward model is $r_t(\mathbf{x}, a; \boldsymbol{\theta}) = \mathbf{x}^\top \boldsymbol{\alpha} + a f(\mathbf{x})^\top \boldsymbol{\beta} + \eta_t$, where $\mathbf{x}$ is the full context, $f(\mathbf{x})$ is a subset of $\mathbf{x}$ that is considered to have an impact on the treatment effects, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^9$. Here η_t is the Gaussian noise on the reward observation, whose variance σ_η^2 is chosen to be $0.1, 1.0$, and 5.0 respectively (Liao et al., 2016). For a detailed list of variables in the context, see Table 2 in Appendix A.

The true parameters $(\boldsymbol{\theta}_a^*)_{a \in \{0,1\}}$ is estimated from GEE (Generalized Estimating Equations) with rewards being the log-transformed step count collected 30 minutes after the decision time.

In light of the measurement error setting in this paper, we consider an observation noise on B_t for the following reasons: 1) The burden B_t can be understood as a prediction of the burden level of the participant, which is particularly crucial in mobile health studies; 2) Other variables are normally believed to have low or no observation noise. Thus, we assume that the agent only observes $\widetilde{\mathbf{x}}_t = (I_t, Z_t, \widetilde{B}_t)$, where $\widetilde{B}_t = B_t + \epsilon_t$ and ϵ_t is drawn i.i.d. from normal distribution with mean zero and variance σ_ϵ^2.

3.3 Results

Table 1 (a) and (b) shows the average regret (cumulative regret divided by T) in both the synthetic environment and the real-data environment based on HeartStep V1. We use the same set of $\sigma_\epsilon^2 \in \{0.1, 1.0, 2.0\}$, while different σ_η^2 reflect the change of absolute values in coefficients in two different environments ($\sigma_\eta^2 = 5.0$ is the level of reward noise in HeartStep V1). MEB shows significantly smaller average regret compared to other baseline methods under most combinations of σ_η^2 and σ_ϵ^2. 1. In certain instances, MEB-naive exhibits performance comparable to MEB. This is attributed to its ability to reduce the variance of model estimation while incurring some bias compared to MEB, rendering it a feasible alternative in practical contexts. Notably, in two extreme scenarios, UCB surpasses both MEB and MEB-naive. This is as expected, since when contextual noise is sufficiently negligible, traditional bandit algorithms are anticipated to outperform the proposed algorithms. An estimation error plot can

be found in Appendix A, which also demonstrates that `MEB` has a lower estimation error.

Table 1: Average regret for both synthetic environment and real-data environment under different combinations of σ_η^2 and σ_ϵ^2. The results are averages over 100 independent runs and the standard deviations are reported in the full table in Appendix A.

(a) Average regret in the synthetic environment over 50000 steps with clipping probability $p = 0.2$.

σ_η^2	σ_ϵ^2	TS	UCB	MEB	MEB-naive
0.01	0.1	0.047	0.046	**0.027**	0.038
0.1	0.1	0.047	0.047	**0.026**	0.039
1.0	0.1	0.048	0.048	**0.027**	0.038
0.01	1.0	0.757	0.647	**0.198**	0.371
0.1	1.0	0.769	0.721	**0.205**	0.392
1.0	1.0	0.714	0.697	**0.218**	0.404
0.01	2.0	1.492	1.504	**0.358**	0.616
0.1	2.0	1.195	1.333	**0.368**	0.584
1.0	2.0	1.299	1.476	**0.416**	0.625

(b) Average regret in the real-data environment over 2500 steps with clipping probability $p = 0.2$.

σ_η^2	σ_ϵ^2	TS	UCB	MEB	MEB-naive
0.05	0.1	0.027	0.027	**0.022**	0.024
0.1	0.1	0.026	0.024	**0.020**	**0.020**
5.0	0.1	1.030	**0.743**	0.831	1.173
0.05	1.0	0.412	0.408	0.117	**0.112**
0.1	1.0	0.309	0.316	**0.085**	0.087
5.0	1.0	1.321	**0.918**	1.458	1.322
0.05	2.0	0.660	0.634	**0.144**	0.148
0.1	2.0	0.740	0.704	**0.151**	0.155
5.0	2.0	1.585	2.415	1.577	**1.436**

4 Discussion and conclusions

We propose a new algorithm, `MEB`, which is the first algorithm with sublinear regret guarantees in contextual bandits with noisy context, where we have limited knowledge of the noise distribution. This setting is common in practice, especially where only predictions for unobserved context are available. `MEB` leverages the novel estimator (2.5), which extends the conventional measurement error adjustment techniques by considering the interplay between the policy and the measurement error.

Limitations and future directions. Several questions remain for future investigation. First, is $\widetilde{\mathcal{O}}(T^{2/3})$ the optimal rate of regret compared to the standard benchmark policy (2.3), as in some other bandits with semi-parametric reward model (e.g. Xu and Wang (2022))? Providing lower bounds on the regret helps us understand the limit of improvement in the online algorithm. Second, we assume that the agent has an unbiased prediction of the true context. It is important to understand how biased predictions affect the results. Last but not least, it's interesting to see our method can be extended to more complicated decision-making settings (e.g. Markov decision processes).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

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

Susan Athey, Stefan Wager, et al. Efficient policy learning. Technical report, 2017.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Marwa Banna, Florence Merlevède, and Pierre Youssef. Bernstein-type inequality for a class of dependent random matrices. *Random Matrices: Theory and Applications*, 5(02):1650006, 2016.

Samuel L Battalio, David E Conroy, Walter Dempsey, Peng Liao, Marianne Menictas, Susan Murphy, Inbal Nahum-Shani, Tianchen Qian, Santosh Kumar, and Bonnie Spring. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.

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

Djallel Bouneffouf. Online learning with corrupted context: Corrupted contextual bandits. *arXiv preprint arXiv:2006.15194*, 2020.

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

Raymond J Carroll, David Ruppert, and Leonard A Stefanski. *Measurement error in nonlinear models*, volume 105. CRC press, 1995.

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.

Sheldon Cohen, Robin Mermelstein, Tom Kamarck, and Harry M Hoberman. Measuring the functional components of social support. *Social support: Theory, research and applications*, pages 73–94, 1985.

Walter Dempsey, Peng Liao, Pedja Klasnja, Inbal Nahum-Shani, and Susan A Murphy. Randomised trials for the fitbit generation. *Significance*, 12(6): 20–23, 2015.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 7111–7123. PMLR, 2022.

Jianqing Fan and Qiwei Yao. *The elements of financial econometrics*. Cambridge University Press, 2017.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

Alexander Galozy and Sławomir Nowaczyk. Information-gathering in latent bandits. *Knowledge-Based Systems*, 260:110099, 2023.

Alexander Galozy, Slawomir Nowaczyk, and Mattias Ohlsson. A new bandit setting balancing informa-tion from state evolution and corrupted context. *arXiv preprint arXiv:2011.07989*, 2020.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. *Advances in Neural Information Processing Systems*, 33:13423–13433, 2020a.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Mohammad Ghavamzadeh, and Craig Boutilier. Non-stationary latent bandits. *arXiv preprint arXiv:2012.00386*, 2020b.

Sharu Theresa Jose and Shana Moothedath. Thompson sampling for stochastic bandits with noisy contexts: An information-theoretic regret analysis. *arXiv preprint arXiv:2401.11565*, 2024.

Johannes Kirschner and Andreas Krause. Stochastic bandits with context distributions. *Advances in Neural Information Processing Systems*, 32, 2019.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S): 1220, 2015.

Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 09 2018.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A Murphy. Sample size calculations for

micro-randomized trials in mhealth. *Statistics in medicine*, 35(12):1944–1971, 2016.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

Jiabin Lin and Shana Moothedath. Distributed stochastic bandit learning with context distributions. *arXiv preprint arXiv:2207.14391*, 2022.

Jiabin Lin, Xian Yeow Lee, Talukder Jubery, Shana Moothedath, Soumik Sarkar, and Baskar Ganapathysubramanian. Stochastic conservative contextual linear bandits. *arXiv preprint arXiv:2203.15629*, 2022.

Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pages 161–168, 2014.

Elliot Nelson, Debarun Bhattacharjya, Tian Gao, Miao Liu, Djallel Bouneffouf, and Pascal Poupart. Linearizing contextual bandits with latent state dynamics. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Hongju Park and Mohamad Kazem Shirani Faradonbeh. Worst-case performance of greedy policies in bandits with imperfect context observations. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 1374–1379. IEEE, 2022.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4489–4501, 2016.

Hillol Sarker, Karen Hovsepian, Soujanya Chatterjee, Inbal Nahum-Shani, Susan A Murphy, Bonnie Spring, Emre Ertin, Mustafa Al'Absi, Motohiro Nakajima, and Santosh Kumar. *From markers to interventions: The case of just-in-time stress intervention*. Springer, 2017.

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR, 2017.

Hammad Shaikh, Arghavan Modiri, Joseph Jay Williams, and Anna N Rafferty. Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*, 2019.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Reward design for an online reinforcement learning algorithm supporting oral self-care. *arXiv preprint arXiv:2208.07406*, 2022.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Jianyu Xu and Yu-Xiang Wang. Towards agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9643–9662. PMLR, 2022.

Xiongxiao Xu, Hong Xie, and John CS Lui. Generalized contextual bandits with latent features: Algorithms and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*, 2020a.

Jianyi Yang and Shaolei Ren. Bandit learning with predicted context: Regret analysis and selective context query. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021a.

Jianyi Yang and Shaolei Ren. Robust bandit learning with imperfect context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10594–10602, 2021b.

Luting Yang, Jianyi Yang, and Shaolei Ren. Multi-feedback bandit learning with probabilistic contexts. In *Proceedings of the Twenty-Ninth Interna-*

tional Joint Conference on Artificial Intelligence Main track, 2020b.

Jiayu Yao, Emma Brunskill, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Power constrained bandits. In *Machine Learning for Healthcare Conference*, pages 209–259. PMLR, 2021.

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19 (10):e338, 2017.

Se-Young Yun, Jun Hyun Nam, Sangwoo Mo, and Jinwoo Shin. Contextual multi-armed bandits under feature uncertainty. *arXiv preprint arXiv:1703.01347*, 2017.

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

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

Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*, 2016.

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]

    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

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

    (a) Statements of the full set of assumptions of all theoretical results. [Yes]

    (b) Complete proofs of all theoretical results. [Yes]

    (c) Clear explanations of any assumptions. [Yes]

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]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

    (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]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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. [Not Applicable]

    (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. [Yes]

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]

Appendices: Supplementary materials for Online learning in bandits with predicted context

Contents

A   Additional details for simulation studies

A.1   Compared algorithms

In both simulation environments, we compare the following algorithms: Thompson sampling (`TS`, see details in Algorithm 2) given normal priors (Russo et al., 2018), Linear Upper Confidence Bound (`UCB`, see details in Algorithm 3) approach (Chu et al., 2011), `MEB` (Algorithm 1), and `MEB-naive` (`MEB` plugged in with the naive measurement error estimator (2.4) instead of (2.5)). To make a fair comparison between algorithms, we use the same regularization parameter $l = 1$ for all algorithms. The hyper-parameter ρ, C is set to be σ_η^2 for all results for `TS` and `UCB`.

Algorithm 2 Linear Thompson Sampling (TS)

1: **Input**: T: total number of steps; ρ: variance of $(\eta_t)_{t \in [T]}$; l: prior variance; p_0: minimum selection probability; $\boldsymbol{\mu}_{0,a} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{0,a} = l\boldsymbol{I}$

2: **for** time $t = 1, 2, \ldots, T$ **do**

3:   Observe $\widetilde{\boldsymbol{x}}_t = \boldsymbol{x}_t + \boldsymbol{\epsilon}_t$

4:   Generate posterior sample $(\widetilde{\boldsymbol{\theta}}_{t,a})_{a \in \mathcal{A}}$ from $\mathcal{N}(\boldsymbol{\mu}_{t-1,a}, \boldsymbol{\Sigma}_{t-1,a})$

5:   Set $\widetilde{a}_t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \langle \widetilde{\boldsymbol{\theta}}_{t,a}, \widetilde{\mathbf{x}}_t \rangle$

6:   Sample $a_t \sim \pi_t(\cdot | \widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$, where $\pi_t(a | \widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) := \begin{cases} 1 - (K-1)p_0, & \text{if } a = \widetilde{a}_t \\ p_0, & \text{otherwise} \end{cases}$

7:   Observe reward $r_t = \langle \boldsymbol{\theta}^\star_{a_t}, \boldsymbol{x}_t \rangle + \eta_t$

8:   Set $\boldsymbol{V}_{t,a} = l\boldsymbol{I} + \sum_{t'=1}^{t} \mathbf{1}_{\{a_{t'} = a\}} \widetilde{\boldsymbol{x}}_t \widetilde{\boldsymbol{x}}_t^\top$, $\boldsymbol{b}_t = \sum_{t'=1}^{t} \mathbf{1}_{\{a_{t'} = a\}} \widetilde{\boldsymbol{x}}_t r_t$

9:   Update $\boldsymbol{\mu}_{t,a} = \boldsymbol{V}_{t,a}^{-1} \boldsymbol{b}_{t,a}$ and $\boldsymbol{\Sigma}_{t,a} = \rho \boldsymbol{V}_{t,a}^{-1}$

10: **end for**

Algorithm 3 Linear UCB (UCB)

1: **Input**: T: total number of steps; $C > 0$; l: regularization; p_0: minimum selection probability; $\boldsymbol{V}_{0,a} = l\boldsymbol{I}$ and $\boldsymbol{b}_{0,a} = \mathbf{0}$

2: **for** time $t = 1, 2, \ldots, T$ **do**

3:   Observe $\widetilde{\boldsymbol{x}}_t = \boldsymbol{x}_t + \boldsymbol{\epsilon}_t$

4:   Estimate $\widehat{\boldsymbol{\theta}}_{t,a} = \boldsymbol{V}_{t-1,a}^{-1} \boldsymbol{b}_{t-1,a}$

5:   Set $\widehat{\mu}_a = \langle \widehat{\boldsymbol{\theta}}_{t,a}, \widetilde{\boldsymbol{x}}_t \rangle + C\sqrt{\widetilde{\boldsymbol{x}}_t^\top \boldsymbol{V}_{t,a}^{-1} \widetilde{\boldsymbol{x}}_t}$

6:   Set $\widetilde{a}_t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} \widehat{\mu}_a$

7:   Sample $a_t \sim \pi_t(\cdot | \widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$, where $\pi_t(a | \widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) := \begin{cases} 1 - (K-1)p_0, & \text{if } a = \widetilde{a}_t \\ p_0, & \text{otherwise} \end{cases}$

8:   Observe reward $r_t = \langle \boldsymbol{\theta}^\star_{a_t}, \boldsymbol{x}_t \rangle + \eta_t$

9:   Set $\boldsymbol{V}_{t,a} = l\boldsymbol{I} + \sum_{t'=1}^{t} \mathbf{1}_{\{a_{t'} = a\}} \widetilde{\boldsymbol{x}}_t \widetilde{\boldsymbol{x}}_t^\top$, $\boldsymbol{b}_t = \sum_{t'=1}^{t} \mathbf{1}_{\{a_{t'} = a\}} \widetilde{\boldsymbol{x}}_t r_t$

10: **end for**

We further compare with robust linear UCB (Algorithm 1 in Ding et al. (2022)) that is shown to achieve minimax rate for adversarial linear bandit.

A.2   Additional details for HeartStep V1 study

Table 2 presents the list of variables to include in the reward model and in the feature construction for algorithms. Recall that our reward model is $r_t(\boldsymbol{x}, a, \boldsymbol{\theta}) = \boldsymbol{x}^\top \boldsymbol{\alpha} + a f(\boldsymbol{x})^\top \boldsymbol{\beta} + \eta_t$. All the variables are included in $\boldsymbol{x}$ while only those considered to have an impact on treatment effect will be included in $f(\boldsymbol{x})$.

Table 2: List of variables in HeartSteps V1 study.

Variable	Type	Treatment?
Availability (I_t)	Discrete	No
Prior 30-minute step count	Continuous	No
Yesterday's step count	Continuous	No
Prior 30-minute step count	Continuous	No
Location	Discrete	Yes
Current temperature	Continuous	No
Step variation level	Discrete	Yes
Burden variable (B_t)	Continuous	Yes

A.3 Additional results on estimation error
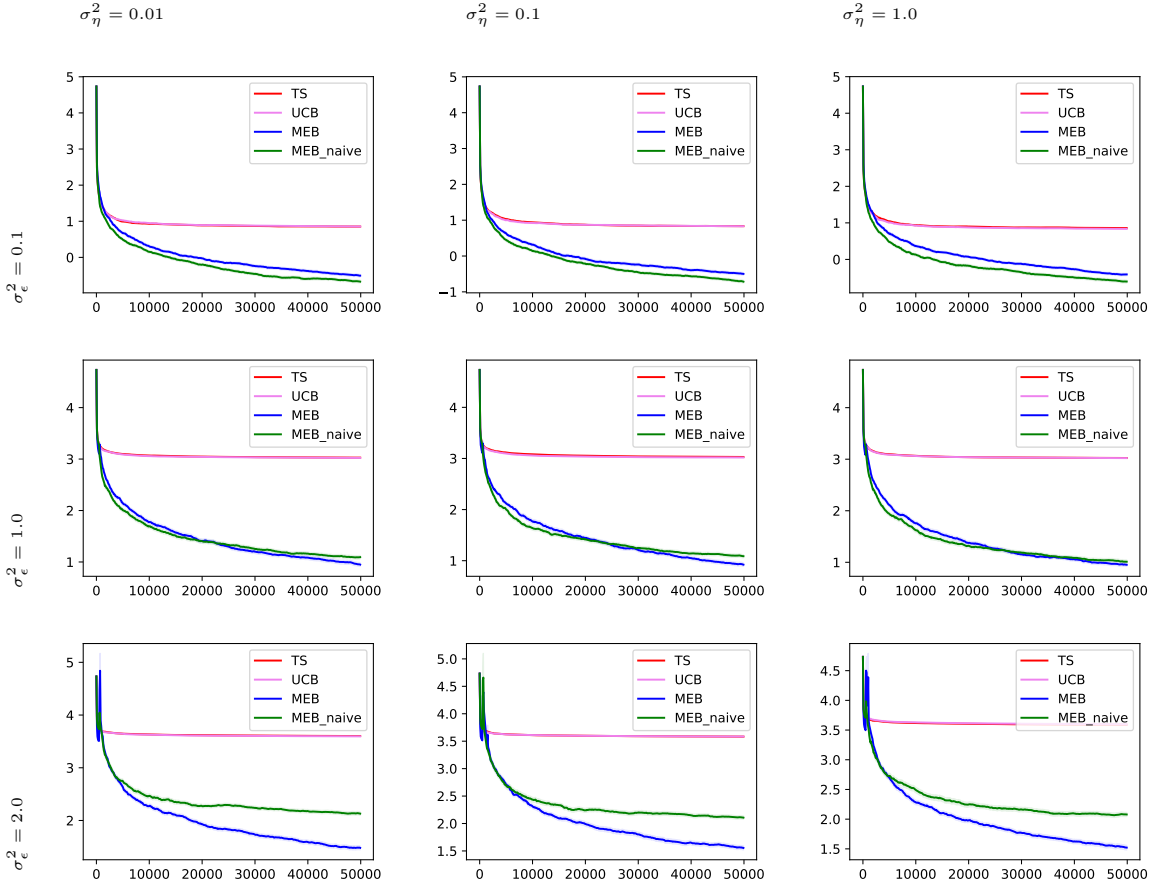
Figure 1: Log-scaled L2 norm of $\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*$ of four algorithms in the synthetic environment over 50000 steps under $\sigma_\epsilon^2 \in \{0.1, 1.0, 2.0\}$ and $\sigma_\eta^2 \in \{0.01, 0.1, 1.0\}$.

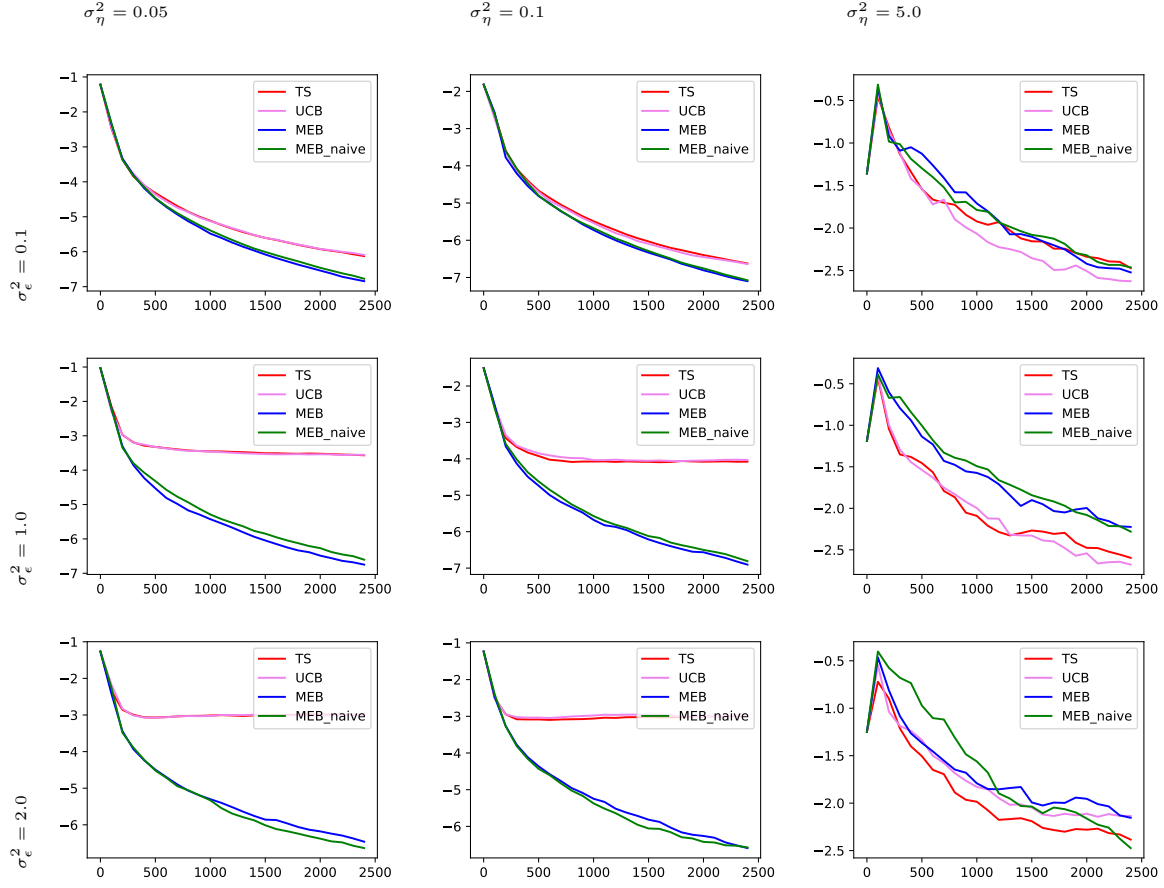Figure 2: Log-scaled L2 norm of $\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1^*$ of four algorithms in the real-data environment based on HeartStep V1 over 2500 steps under $\sigma_\epsilon^2 \in \{0.1, 1.0, 2.0\}$ and $\sigma_\eta^2 \in \{0.05, 0.1, 5.0\}$.

18

A.4 Average regret with standard deviation

Table 3: Average regret for both synthetic environment and real-data environment under different combinations of σ_η^2 and σ_ϵ^2. The numbers in parentheses are the standard deviations calculated from 100 independent runs.

(a) Average regret in the synthetic environment over 50000 steps with clipping probability $p = 0.2$.

σ_η^2	σ_ϵ^2	TS	UCB	MEB	MEB-naive	RobustUCB
0.01	0.1	0.047 (0.0015)	0.046 (0.0015)	0.027 (0.0011)	0.038 (0.0013)	0.050 (0.0051)
0.1	0.1	0.047 (0.0015)	0.047 (0.0015)	0.026 (0.0011)	0.039 (0.0013)	0.049 (0.0048)
1.0	0.1	0.048 (0.0015)	0.048 (0.0015)	0.027 (0.0011)	0.038 (0.0013)	0.044 (0.0047)
0.01	1.0	0.757 (0.0164)	0.647 (0.0145)	0.198 (0.0079)	0.371 (0.0107)	0.652 (0.0050)
0.1	1.0	0.769 (0.0160)	0.721 (0.0156)	0.205 (0.0080)	0.392 (0.0110)	0.753 (0.0056)
1.0	1.0	0.714 (0.0155)	0.697 (0.0150)	0.218 (0.0083)	0.404 (0.0112)	0.589 (0.0047)
0.01	2.0	1.492 (0.0281)	1.504 (0.0283)	0.358 (0.0129)	0.616 (0.0169)	1.608 (0.0102)
0.1	2.0	1.195 (0.0244)	1.333 (0.0260)	0.368 (0.0131)	0.584 (0.0164)	1.064 (0.0079)
1.0	2.0	1.299 (0.0257)	1.476 (0.0277)	0.416 (0.0139)	0.625 (0.0170)	1.881 (0.0114)

(b) Average regret in the real-data environment over 2500 steps with clipping probability $p = 0.2$.

σ_η^2	σ_ϵ^2	TS	UCB	MEB	MEB-naive	RobustUCB
0.05	0.1	0.027 (0.0067)	0.027 (0.0070)	0.022 (0.0057)	0.024 (0.0058)	0.025 (0.0079)
0.1	0.1	0.026 (0.0057)	0.024 (0.0053)	0.020 (0.0046)	0.020 (0.0046)	0.028 (0.0079)
5.0	0.1	1.030 (0.0287)	0.743 (0.0262)	0.831 (0.0267)	1.173 (0.0343)	1.400 (0.0447)
0.05	1.0	0.412 (0.0355)	0.408 (0.0351)	0.117 (0.0148)	0.112 (0.0143)	0.226 (0.0020)
0.1	1.0	0.309 (0.0293)	0.316 (0.0299)	0.085 (0.0112)	0.087 (0.0116)	0.206 (0.0125)
5.0	1.0	1.321 (0.0417)	0.918 (0.0304)	1.458 (0.0422)	1.322 (0.0388)	1.065 (0.0400)
0.05	2.0	0.660 (0.0343)	0.634 (0.0322)	0.144 (0.0129)	0.148 (0.0133)	0.304 (0.0241)
0.1	2.0	0.740 (0.0505)	0.704 (0.0489)	0.151 (0.0145)	0.155 (0.0149)	0.432 (0.0386)
5.0	2.0	1.585 (0.0454)	2.415 (0.0816)	1.577 (0.0508)	1.436 (0.0462)	1.345 (0.0423)

B Additional explanations on the regularized least-squares (RLS) estimator the naive estimator (2.4) under noisy context

B.1 Inconsistency of the RLS estimator

Measurement error model and attenuation. As briefly mentioned in the main text, a measurement error model is a regression model designed to accommodate inaccuracies in the measurement of regressors. Suppose that there is no action (i.e. set $a_\tau \equiv 0$), and $(\mathbf{x}_\tau, \boldsymbol{\epsilon}_\tau, \eta_\tau)_{\tau \in [t]}$ are i.i.d., then the measurement error model is a useful tool to learn $\boldsymbol{\theta}_0^*$ from $\mathcal{H}_t$ collected as follows:

$$\begin{cases} r_\tau = \langle \boldsymbol{\theta}_0^*, \mathbf{x}_\tau \rangle + \eta_\tau, \\ \widetilde{\mathbf{x}}_\tau = \mathbf{x}_\tau + \boldsymbol{\epsilon}_\tau. \end{cases} \tag{B.1}$$

Here in the measurement error model's perspective, $(\widetilde{\mathbf{x}}_\tau)_{\tau \in [t]}$ are regressors 'measured with error', and $(r_\tau)_{\tau \in [t]}$ are dependent variables.

Regression attenuation, a phenomenon intrinsic to measurement error models, refers to the observation that when the predictors are subject to measurement errors, the Ordinary Least Squares (OLS) estimators of regression coefficients become biased (see, for instance, Carroll et al. (1995)). Specifically, in simple linear

regression, the OLS estimator for the slope tends to be biased towards zero. Intuitively, this is because the measurement errors effectively 'dilute' the true relationship between variables, making it appear weaker than it actually is.

Inconsistency of the RLS estimator. Before presenting a concrete numerical example to show the inconsistency of the RLS estimator and that it leads to bad decision-making, below we first apply the theory of measurement error model to give a heuristic argument of why the RLS estimator is inconsistent even in the simplified situation where there is no action (i.e. set $a_\tau \equiv 0$), and $(\mathbf{x}_\tau, \boldsymbol{\epsilon}_\tau, \eta_\tau)_{\tau \in [t]}$ are i.i.d..

From Section 3.3.2 in Carroll et al. (1995), given data $(\widetilde{\mathbf{x}}_\tau, r_\tau)_{\tau \in [t]}$ from (B.1) with multiple covariates (in the simplified case with no action as described above), the OLS estimator for $\boldsymbol{\theta}_0^*$, denoted as $\widehat{\boldsymbol{\theta}}_0^{OLS,t}$, consistently estimates not $\boldsymbol{\theta}_0^*$ but

$$\widetilde{\boldsymbol{\theta}}_0^* = (\mathbf{\Sigma_x} + \mathbf{\Sigma}_e)^{-1}\mathbf{\Sigma_x}\boldsymbol{\theta}_0^*, \tag{B.2}$$

where $\mathbf{\Sigma_x} = \text{Var}(\mathbf{x}_\tau)$, $\mathbf{\Sigma}_e = \text{Var}(\boldsymbol{\epsilon}_\tau)$. In addition, given fixed λ, the regularized least squares (RLS) estimator

$$\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) = \left(\frac{\lambda}{t}I + \frac{1}{t}\sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top\right)^{-1}\left(\frac{1}{t}\sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau r_\tau\right) = \left(\frac{\lambda}{t}I + \frac{1}{t}\sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top\right)^{-1}\left(\frac{1}{t}\sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top\right)\widehat{\boldsymbol{\theta}}_0^{OLS,t},$$

where $\frac{1}{t}\sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top \to \text{Var}(\widetilde{\mathbf{x}}_\tau) = \text{Var}(\mathbf{x}_\tau) + \text{Var}(\boldsymbol{\epsilon}_\tau)$, and $\lambda/t \to 0$ as $t \to \infty$. This means that as $t \to \infty$, $\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda)$ and $\widehat{\boldsymbol{\theta}}_0^{OLS,t}$ converges to the same limit, which is $\widetilde{\boldsymbol{\theta}}_0^*$. Thus, for fixed λ, as t grows, $\|\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) - \boldsymbol{\theta}_0^*\|_2$ converges to $\|\widetilde{\boldsymbol{\theta}}_0^* - \boldsymbol{\theta}_0^*\|_2$ and does not converge to zero in general.

Finally, recall that in classical bandit algorithms such as UCB, the sublinear regret relies on the key property that with high probability, for all t, $\|\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) - \boldsymbol{\theta}_0^*\|_{V_t} \leq \beta$, where $V_t = \lambda I + \sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top$, $\beta = \widetilde{\mathcal{O}}(\sqrt{d})$. Here for a vector $\boldsymbol{v} \in \mathbb{R}^d$ and positive definite matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, $\|\boldsymbol{v}\|_{\mathbf{M}} := \sqrt{\boldsymbol{v}^\top \mathbf{M} \boldsymbol{v}}$. We argue that this requirement generally no longer holds in the setting with measurement error $(\boldsymbol{\epsilon}_\tau)_{\tau \geq 1}$. In fact, notice that since $\widetilde{\mathbf{x}}_\tau$ is i.i.d. in this simplified setting, and $V_t = \lambda I + \sum_{\tau \in [t]} \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top$, we expect that $\frac{1}{t}V_t$ concentrates around $\text{Var}(\widetilde{\mathbf{x}}_t)$. As long as $\lambda_{\min}(\text{Var}(\widetilde{\mathbf{x}}_t)) > 0$, with high probability, for all t, $\frac{1}{t}V_t \succ c$ for some constant c. If this holds,

$$\|\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) - \boldsymbol{\theta}_0^*\|_{V_t} \geq \sqrt{ct}\|\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) - \boldsymbol{\theta}_0^*\|_2,$$

while the last term $\|\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) - \boldsymbol{\theta}_0^*\|_2$ converges to a nonzero limit $\|\widetilde{\boldsymbol{\theta}}_0^* - \boldsymbol{\theta}_0^*\|_2$. This indicates that in general, $\|\widehat{\boldsymbol{\theta}}_0^{RLS,t}(\lambda) - \boldsymbol{\theta}_0^*\|_{V_t}$ scales with a rate of at least $\sqrt{t}$, and will not be uniformly bounded by $\widetilde{\mathcal{O}}(\sqrt{d})$ for all t.

An example. The following is an example where given the errors $(\boldsymbol{\epsilon}_\tau)_{\tau \in [t]}$, the RLS estimator in the classical bandit algorithms inconsistently estimates the true reward model, and in addition, it leads to bad decision-making (linear regret) in the classic bandit algorithms.

Example B.1. Consider the standard setting described in Section 2.1. Let $T = 10000$, $d = 2$, $\boldsymbol{\theta}_1^* = (1, 0)^\top$, $\boldsymbol{\theta}_0^* = (-1, 0)^\top$. Let $\mathbf{x}_t$ sampled i.i.d. from $\text{Unif}(\mathcal{S})$, where $\mathcal{S} = \{(1, 3)^\top, (-3, 1)^\top, (-1, -3)^\top, (3, -1)^\top\}$. Condition on $\mathbf{x}_t$, $\boldsymbol{\epsilon}_t$ is uniformly sampled from $((\rho_0 \mathbf{x}_t^{[1]}, \rho_0 \mathbf{x}_t^{[1]})^\top, (-\rho_0 \mathbf{x}_t^{[1]}, -\rho_0 \mathbf{x}_t^{[1]})^\top)$, independent from any other variable in the history. Here $\rho_0 = 0.9$, $\mathbf{x}_t^{[1]}$ denotes the first entry in $\mathbf{x}_t$. We also let η_t be i.i.d. drawn from $N(0, 0.01)$.

We conduct 100 independent experiments. For each experiment, we generate data according to the above, and test the performance of UCB (Algorithm 1 in Chu et al. (2011)) and Thompson sampling with normal priors (Russo et al., 2018) using noisy context $\widetilde{\mathbf{x}}_t$ instead of the true context. We choose the regularization parameter $\lambda = 1$ in the RLS estimator. Additionally, in UCB (Chu et al., 2011), we choose the parameter $\alpha = 1$. Figure 3 summarizes the estimation error of the RLS estimator and the cumulative regret of both algorithms with respect to time t, showing both the average and standard error across the random experiments. We see that the RLS estimator is unable to estimate the true reward model well. Moreover, it is clear that the regret of both UCB and Thompson sampling is linear in the time horizon. Intuitively,
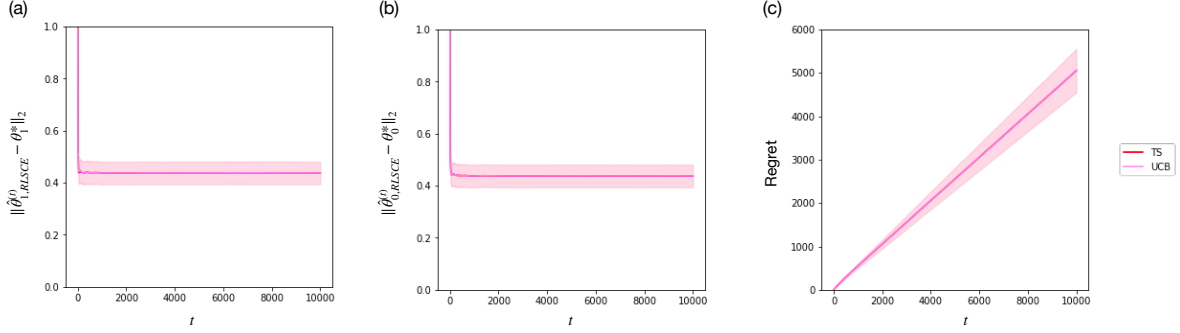
Figure 3: Estimation error of the RLS estimator and cumulative regret of UCB (Chu et al., 2011) and Thompson sampling (Russo et al., 2018) under contextual error in Example B.1. The red and pink line corresponds to Thompson sampling and UCB respectively. The solid lines indicate the mean values, while the shaded bands represent the standard error across the independent experiments.

this is because the direction of $\widetilde{\boldsymbol{\theta}}_0^*$ in (B.2) is twisted compared to $\boldsymbol{\theta}_0^*$, which not only leads to inconsistent estimators, but also the optimal action altered.

Finally, we note that in this setup, Assumption 2.3 is satisfied. This demonstrates that even if the errors $(\boldsymbol{\epsilon}_t)_{t \geq 1}$ do not affect the optimal action given $(\boldsymbol{\theta}_a^*)_{a \in \{0,1\}}$, the poor performance of the RLS estimator may still lead to linear regret in classical bandit algorithms.

B.2 Inconsistency of the naive measurement error adjusted estimator (2.4)

Example B.2. Let $d = 1$, $\mathbf{x}_\tau \equiv 1$ for all τ, and $\boldsymbol{\epsilon}_\tau \sim \text{Unif}(-2, 2)$ sampled independently. For the reward model, let $\boldsymbol{\theta}_0^* = -1$, $\boldsymbol{\theta}_1^* = 1$, $\eta_\tau \sim \text{Unif}(-0.1, 0.1)$ sampled independently. So in order to maximize expected reward, we should choose action 1 if $\mathbf{x}_\tau$ is positive and action 0 otherwise. Suppose the agent takes the following policy that is stationary and non-adaptive to history:

$$\pi_\tau(A) = \begin{cases} \frac{2}{3}1_{\{A=1\}} + \frac{1}{3}1_{\{A=0\}}, & \text{if } \widetilde{\mathbf{x}}_\tau > \rho \\ \frac{1}{3}1_{\{A=1\}} + \frac{2}{3}1_{\{A=0\}}, & \text{otherwise}. \end{cases}$$

Here, ρ is a pre-specified constant. Figure 4 (a) plots the mean and standard deviation of $\widehat{\boldsymbol{\theta}}_{0,me}^{(t)}$ (as in (2.4)) given 100 independent experiments for each $t = 1, \ldots, 10000$, where $\rho = -0.5, 0, 0.5$. Observe that as t grows, $\widehat{\boldsymbol{\theta}}_{0,me}^{(t)}$ converges to different limits for different policies. In general, the limit is not equal to $\boldsymbol{\theta}_0^* = -1$.

In contrast, Figure 4 (b) shows the mean and standard deviation of $\widehat{\boldsymbol{\theta}}_0^{(t)}$ (as in (2.5)) given 100 independent experiments under the same setting with the same three policies as in Figure 4 (a). Unlike the naive estimator (2.4), the proposed estimator (2.5) quickly converges around the true value -1 for all three candidate policies.

C Generalization to $K \geq 2$ actions

In this section, we assume that $\mathcal{A} = \{1, 2, \ldots, K\}$ instead of $\{0, 1\}$. The standard and clipped benchmark become

$$\pi_t^*(a) = \begin{cases} 1, & \text{if } a = a_t^* = \text{argmax}_a \langle \boldsymbol{\theta}_a^*, \mathbf{x}_t \rangle \\ 0, & \text{otherwise}, \end{cases} \tag{C.1}$$
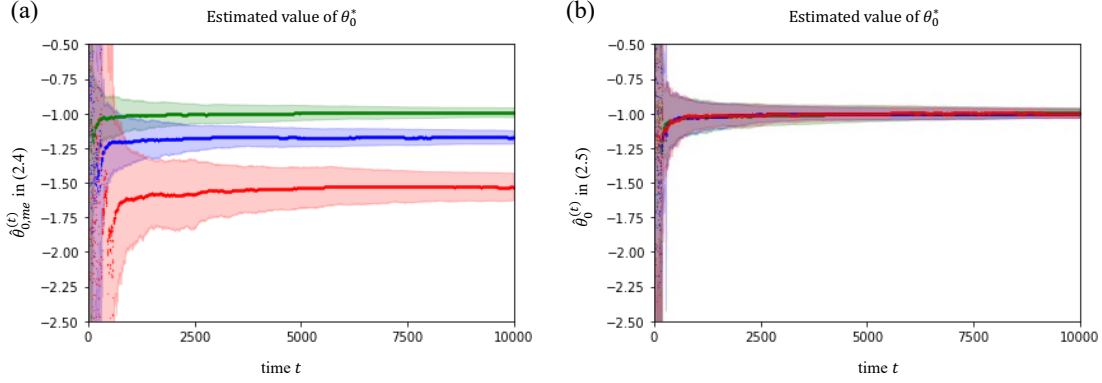
Figure 4: Estimated value of $\boldsymbol{\theta}_0^*$ given the naive estimator (2.4) in (a) and our proposed estimator (2.5) in (b) under different policies under 100 independent experiments. The green, blue, and red line corresponds to the policy with parameter $\rho = -0.5, 0$, and 0.5 respectively. The solid lines indicate the mean values, while the shaded bands represent the standard deviation across the independent experiments.

Algorithm 4 MEB with K actions

1: **Input**: $(\boldsymbol{\Sigma}_{e,t})_{t\in[T]}$: covariance sequence of $(\boldsymbol{\epsilon}_t)_{t\in[T]}$; $(p_0^{(t)})_{t\in[T]}$: minimum selection probability at each time t; T_0: length of pure exploration
2: **for** time $t = 1, 2, \ldots, T$ **do**
3:     **if** $t \leq T_0$ **then**
4:         Sample $a_t \sim \pi_t(\cdot|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$, where $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) \in [p_0^{(t)}, 1 - (K-1)p_0^{(t)}]$ for all $a \in \mathcal{A}$
5:     **else**
6:         Obtain updated estimators $(\widehat{\boldsymbol{\theta}}_a^{(t-1)})_{a\in\mathcal{A}}$ from (2.5)
7:         $\widetilde{a}_t \leftarrow \operatorname{argmax}_{a\in\mathcal{A}}\langle\widehat{\boldsymbol{\theta}}_a^{(t-1)}, \widetilde{\mathbf{x}}_t\rangle$
8:         Sample $a_t \sim \pi_t(\cdot|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$, where $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) := \begin{cases} 1 - (K-1)p_0^{(t)}, & \text{if } a = \widetilde{a}_t \\ p_0^{(t)}, & \text{otherwise} \end{cases}$
9:     **end if**
10: **end for**

---

and

$$\bar{\pi}_t^*(a) = \begin{cases} 1 - (K-1)p_0, & \text{if } a = a_t^*, \\ p_0, & \text{otherwise.} \end{cases} \tag{C.2}$$

In the K-arm setting, we can still estimate $\boldsymbol{\theta}_a^*$ using (2.5) for each $a \in \mathcal{A}$. Using the same proof ideas as Theorem 2.1, we get the following guarantees for estimation error (proof omitted):

Theorem C.1. For any $t \in [T]$, let $q_t = \inf_{\tau\in[t], a\in\mathcal{A}} \pi_\tau(a|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})$. Then under Assumption 2.1 and 2.2, there exist constants C and C_1 such that as long as $q_t \geq C_1 \max\left\{\frac{d(d+\log t)}{\lambda_0 t}, \frac{\xi(d+\log t)}{\lambda_0^2 t}\right\}$, with probability at least $1 - \frac{4K}{t^2}$,

$$\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \leq \frac{C(R + R_\theta)d}{\lambda_0} \max\left\{\frac{d + \log t}{q_t t}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}}\sqrt{\frac{d + \log t}{q_t t}}\right\}, \quad \forall a \in \mathcal{A}.$$

MEB with K actions is shown in Algorithm 4. As in Theorem 2.2, we can control the regret of Algorithm 4 by the estimation error of (2.5). Here, Assumption 2.3 needs to be generalized to the following to adapt to multiple actions:

22

Assumption C.1. There exists a constant $\rho \in (0, 1)$ such that $\forall t \in [T]$, $\forall a_1, a_2 \in \mathcal{A}$, $|\langle \boldsymbol{\theta}^*_{a_1} - \boldsymbol{\theta}^*_{a_2}, \boldsymbol{\epsilon}_t \rangle| \leq \rho |\langle \boldsymbol{\theta}^*_{a_1} - \boldsymbol{\theta}^*_{a_2}, \mathbf{x}_t \rangle|$ almost surely.

The theorem below is a generalization of Theorem 2.2 to multiple actions (The proof is only slightly different from Theorem 2.2; We briefly discuss the difference in Appendix G.1).

Theorem C.2. Let Assumption 2.1 and C.1 hold.

(i) For the standard setting, Algorithm 4 outputs a policy with $\text{Regret}(T; \pi^*)$ no more than

$$2T_0 R_\theta + \frac{2}{1-\rho} \cdot \sum_{t=T_0+1}^{T} \left[(K-1)p_0^{(t)} R_\theta + \max_{a \in \mathcal{A}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \right].$$

(ii) For the clipped policy setting, Algorithm 4 with the choice of $p_0^{(t)} \equiv p_0$ outputs a policy with $\text{Regret}(T; \bar{\pi}^*)$ no more than

$$2T_0 R_\theta + \frac{2(1-Kp_0)}{1-\rho} \cdot \sum_{t=T_0+1}^{T} \max_{a \in \mathcal{A}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2.$$

Combining Theorems C.1 and C.2, we obtain the following corollary.

Corollary C.1. Let Assumption 2.1, 2.2 and C.1 hold. There exist universal constants C, C' such that:

(i) For the standard setting, $\forall T \geq C' \max\{(1 + 1/\lambda_0^{\frac{9}{4}})(d + \log T)^3, (\xi/\lambda_0)^{\frac{9}{4}}(d + \log T)^{\frac{4}{3}}\}$, with probability at least $1 - \frac{8K}{\sqrt{T}}$, Algorithm 4 with the choice of $T_0 = \lceil 2dT^{\frac{2}{3}} \rceil$, $p_0^{(t)} = \min\{\frac{1}{2}, t^{-\frac{1}{3}}\}$ outputs a policy with $\text{Regret}(T; \pi^*)$ no more than

$$CdT^{\frac{2}{3}} \left\{ \frac{(K-1)R_\theta}{1-\rho} + \frac{(\sqrt{\nu} + \sqrt{\xi} + 1)(R + R_\theta)}{(1-\rho)\lambda_0} \sqrt{1 + \frac{\log T}{d}} \right\}.$$

(ii) For the clipped policy setting, $\forall T \geq C' \max\{(d + \log T)^2/(\lambda_0 p_0)^2, \xi^2/\lambda_0^4(1 + \log T/d)^2\}$, with probability at least $1 - \frac{8K}{\sqrt{T}}$, Algorithm 4 with the choice of $T_0 = \lceil 2d\sqrt{T} \rceil$ and $p_0^{(t)} \equiv p_0$ outputs a policy with $\text{Regret}(T; \bar{\pi}^*)$ no more than

$$CdT^{\frac{1}{2}} \left\{ R_\theta + \frac{(\sqrt{\nu} + \sqrt{\xi} + 1)(1 - Kp_0)(R + R_\theta)}{\sqrt{p_0}(1-\rho)\lambda_0} \sqrt{1 + \frac{\log T}{d}} \right\}.$$

D Analysis of the proposed estimator (2.5)

D.1 Proof of Theorem 2.1

We fix some $t \in [T]$, and control $\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2$ for $a \in \{0, 1\}$. Towards this goal, we combine analysis of the two random terms $\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}, a}^{(t)}$ and $\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}, r, a}^{(t)}$ in the lemma below.

Lemma D.1. Under the same assumptions of Theorem 2.1, there exists an absolute constant C such that with probability at least $1 - 4/t^2$, both of the followings hold:

$$\left\| \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}, a}^{(t)} - \frac{1}{t} \sum_{\tau \in [t]} \pi_\tau^{nd}(a)(\mathbf{x}_\tau \mathbf{x}_\tau^\top + \boldsymbol{\Sigma}_{e, \tau}) \right\|_2 \leq C \max \left\{ \frac{d + \log t}{q_t t}, \frac{\sqrt{\xi}}{d} \cdot \sqrt{\frac{d + \log t}{q_t t}} \right\}, \tag{D.1}$$

$$\left\| \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}}, r, a}^{(t)} - \left(\frac{1}{t} \sum_{\tau \in [t]} \pi_\tau^{nd}(a) \mathbf{x}_\tau \mathbf{x}_\tau^\top \right) \boldsymbol{\theta}_a^* \right\|_2 \leq CR \max \left\{ \frac{d + \log t}{q_t t}, \sqrt{\frac{\nu}{d}} \cdot \sqrt{\frac{d + \log t}{q_t t}} \right\}. \tag{D.2}$$

23

Proof of Lemma D.1 is in Section G.2.

Denote $\boldsymbol{\Delta}_1 = \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)(\mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Sigma}_{e,\tau})$, $\boldsymbol{\Delta}_2 = \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)} - \left(\frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\mathbf{x}_\tau\mathbf{x}_\tau^\top\right)\boldsymbol{\theta}_a^*$. Then

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_a^{(t)} &= \left(\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\boldsymbol{\Sigma}_{e,\tau}\right)^{-1}\cdot\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)}\\
&= \left(\frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Delta}_1\right)^{-1}\cdot\left[\left(\frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\mathbf{x}_\tau\mathbf{x}_\tau^\top\right)\boldsymbol{\theta}_a^* + \boldsymbol{\Delta}_2\right]\\
&= \boldsymbol{\theta}_a^* - \boldsymbol{J}_1 + \boldsymbol{J}_2,
\end{aligned}
$$

where

$$
\boldsymbol{J}_1 := \left[\frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Delta}_1\right]^{-1}\cdot\boldsymbol{\Delta}_1\boldsymbol{\theta}_a^*, \quad \boldsymbol{J}_2 := \left[\frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Delta}_1\right]^{-1}\cdot\boldsymbol{\Delta}_2.
$$

It's easy to verify that under the event where both (D.1) and (D.2) hold, whenever

$$
C\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\xi}}{d}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\} \le \frac{\lambda_0}{2d}, \tag{D.3}
$$

we have

$$
\|\boldsymbol{J}_1\|_2 \le \frac{2CR_{\boldsymbol{\theta}}d}{\lambda_0}\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\xi}}{d}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}
$$

and

$$
\|\boldsymbol{J}_2\|_2 \le \frac{2CRd}{\lambda_0}\max\left\{\frac{d+\log t}{q_t t}, \sqrt{\frac{\nu}{d}}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}.
$$

(D.3) can be ensured by $t \ge C_1\max\{\frac{d(d+\log t)}{\lambda_0 q_t}, \frac{\xi(d+\log t)}{\lambda_0^2 q_t}\}$, where $C_1 = \max\{2C, 4C^2\}$. Given these guarantees, we have with probability at least $1 - 4/t^2$,

$$
\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \le \|\boldsymbol{J}_1\|_2 + \|\boldsymbol{J}_2\|_2 \le \frac{2C(R+R_\theta)d}{\lambda_0}\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\nu}+\sqrt{\xi}}{\sqrt{d}}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}.
$$

Thus we conclude the proof.

D.2 Additional comments on Assumption 2.2

Following Remark 2.1, all the following examples of $\{\mathbf{x}_\tau\}$ allow the existence of λ_0 with $\pi_\tau^{nd}(a) \equiv 1/2$ given any reasonably big t with high probability:

- $\{\sqrt{d}\mathbf{x}_\tau\}$ is an i.i.d. sequence satisfying $d\mathbb{E}\mathbf{x}_\tau\mathbf{x}_\tau^\top \succeq \lambda_1 I_d$, $\lambda_1 > 0$;

- $\{\sqrt{d}\mathbf{x}_\tau\}$ is a weakly-dependent stationary time series (a common example is the multivariate ARMA process under regularity conditions, see e.g. (Banna et al., 2016)). The stationary distribution P satisfies $d\mathbb{E}_{\mathbf{x}\sim P}\mathbf{x}\mathbf{x}^\top \succeq \lambda_2 I_d$, $\lambda_2 > 0$;

- $\{\sqrt{d}\mathbf{x}_\tau\}$ is a periodic time series such that there exists $t_0 \in \mathbb{N}^+$ which satisfies $\frac{d}{t_0}\sum_{\tau\in(kt_0+1,(k+1)t_0]}\mathbf{x}_\tau\mathbf{x}_\tau^\top \succeq \lambda_3 I_d$ a.s. $\forall k \in \mathbb{N}$.

D.3 Generalization to off-policy method-of-moment estimation

(2.5) can be generalized to a class of method-of-moment estimators for off-policy learning. In this section, we delve into the general framework of off-policy method-of-moment estimation. This framework proves valuable in scenarios where a fully parametric model class for the reward is unavailable, yet there is a desire to estimate certain model parameters using offline batched bandit data.

For simplicity, we assume that $(X_t, Y_t(a) : a \in \mathcal{A})_{t \in [T]}$ are drawn i.i.d. from an unknown distribution $\mathcal{P}$. At each time $t \in [T]$, the action A_t is drawn from a policy $\pi_t(\cdot|X_t, \mathcal{H}_{t-1})$, and the agent observes only $o_t = (X_t, A_t, Y_t(A_t))$ together with the action selection probabilities π_t. Define the history up to time t as $\mathcal{H}_t = \{o_\tau\}_{\tau \leq t}$. For $a_0 \in \mathcal{A}$, we're interested in estimating $\boldsymbol{\theta}^*_{a_0}$, a d-dimensional parameter in $\mathcal{P}^{(a_0)}$, which is the joint distribution of $\{X_t, Y_t(a_0)\}$.

Remark D.1. When the context is i.i.d., the problem of estimating $\{\boldsymbol{\theta}^*_a\}_{a \in \mathcal{A}}$ in Section 2.2 is a special case of this setup by taking $\widetilde{\mathbf{x}}_t$ as X_t and r_t as Y_t.

The traditional method-of-moment estimator looks for functions $f_1, \ldots, f_d$ as well as a mapping $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^d$, such that

$$\boldsymbol{\theta}^*_{a_0} = \boldsymbol{\phi}\big(\mathbb{E}_{(X,Y) \sim \mathcal{P}^{(a_0)}} f_1(X,Y), \ldots, \mathbb{E}_{(X,Y) \sim \mathcal{P}^{(a_0)}} f_d(X,Y)\big).$$

Then, if given i.i.d. samples $(U_t, V_t)_{t \in [n]}$ from $\mathcal{P}^{(a_0)}$, the estimator takes the form

$$\widehat{\boldsymbol{\theta}}_{a_0} = \boldsymbol{\phi}\bigg(\frac{1}{T} \sum_{t \in [T]} f_1(U_t, V_t), \ldots, \frac{1}{T} \sum_{t \in [T]} f_d(U_t, V_t)\bigg).$$

In fact, the naive estimator (2.4) is of this form. It is clear that we cannot use this estimator for offline batched data $\mathcal{H}_T$: There are no i.i.d. samples from $\mathcal{P}^{(a_0)}$ because of the policy $\{\pi_t\}_{t \in [T]}$. Instead, we propose the following estimator:

$$\widehat{\boldsymbol{\theta}}_{a_0} = \boldsymbol{\phi}\bigg(\frac{1}{T} \sum_{t \in [T]} W_t f_1(X_t, Y_t), \ldots, \frac{1}{T} \sum_{t \in [T]} W_t f_d(X_t, Y_t)\bigg),$$

where $W_t = 1_{\{A_t = a_0\}} \frac{\pi^{nd}(A_t)}{\pi_t(A_t|X_t, \mathcal{H}_{t-1})}$ for a data-independent probability distribution π^{nd} on $\mathcal{A}$. Similar to the proof of Theorem 2.1, it's not difficult to see that $\widehat{\boldsymbol{\theta}}_{a_0}$ is consistent under mild conditions. In fact, (2.5) is a special case of this estimator when π^{nd}_τ does not depend on τ. A more detailed analysis is left for future work.

E Analysis of `MEB`

E.1 Additional comments on Example 2.1

In example 2.1, the joint distribution of $(\mathbf{x}_t, \widetilde{\mathbf{x}}_t)$ is:

$$\mathbb{P}((\mathbf{x}_t, \widetilde{\mathbf{x}}_t) = (0.2, 1)) = 0.3, \quad \mathbb{P}((\mathbf{x}_t, \widetilde{\mathbf{x}}_t) = (0.2, -1)) = 0.2,$$
$$\mathbb{P}((\mathbf{x}_t, \widetilde{\mathbf{x}}_t) = (-0.2, 1)) = 0.2, \quad \mathbb{P}((\mathbf{x}_t, \widetilde{\mathbf{x}}_t) = (-0.2, -1)) = 0.3.$$

The optimal action given $(\boldsymbol{\theta}^*_a)_{a \in \{0,1\}}$ and $\mathbf{x}_t$ is

$$a^*_t(\mathbf{x}_t) = \operatorname*{argmax}_a \langle \boldsymbol{\theta}^*_a, \mathbf{x}_t \rangle = \begin{cases} 1, & \text{if } \mathbf{x}_t = 0.2, \\ 0, & \text{if } \mathbf{x}_t = -0.2. \end{cases}$$

In the standard bandit setting, the benchmark policy is

$$\pi^*_t(a) = \begin{cases} 1, & \text{if } a = a^*_t(\mathbf{x}_t), \\ 0, & \text{otherwise.} \end{cases}$$

For any policy $\pi = (\pi_t)_{t \geq 1}$, the instantaneous regret at time t is

$$\text{Regret}_t(\pi_t, \pi_t^*) = 0.4 \mathbb{E}_{a_t \sim \pi_t(\cdot | \widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})} 1_{\{a_t \neq a_t^*(\mathbf{x}_t)\}}.$$

Even if both $(\boldsymbol{\theta}_a^*)_{a \in \{0,1\}}$ and the joint distribution of $(\mathbf{x}_t, \widetilde{\mathbf{x}}_t)$ are known, π_t can only depend on $\widetilde{\mathbf{x}}_t$ and history, and cannot be based on $\mathbf{x}_t$. Thus, there is always a (constant) positive probability that the action a_t sampled from π_t does not match $a_t^*(\mathbf{x}_t)$ (otherwise, a_t sampled from $\pi_t(\cdot | \widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$ should be equal to $a_t^*(\mathbf{x}_t)$ a.s.). Thus, the standard cumulative regret will be linear in the time horizon.

E.2 Proof of Theorem 2.2

We first prove the lemma below. Its proof is in Appendix G.3.

Lemma E.1. Under Assumption 2.3, we have $a_t^* = a_t^\dagger := \text{argmax}_a \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle$. Consequently, $\pi_t^* = \pi_t^\dagger$, $\bar{\pi}_t^* = \bar{\pi}_t^\dagger$ (given a fixed minimum action selection probability p_0), where

$$\pi_t^\dagger(a) = \begin{cases} 1, & \text{if } a = a_t^\dagger, \\ 0, & \text{otherwise}, \end{cases} \qquad \bar{\pi}_t^\dagger(a) = \begin{cases} 1 - p_0, & \text{if } a = a_t^\dagger, \\ p_0, & \text{otherwise}. \end{cases}$$

In the below, we define

$$\widehat{\text{Regret}}_t(\pi, \pi^*) := \mathbb{E}_{a \sim \pi_t^*} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle,$$

$$\widehat{\text{Regret}}_t(\pi, \bar{\pi}^*) := \mathbb{E}_{a \sim \bar{\pi}_t^*} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle.$$

Standard setting. In the standard setting, we give the lemma below (proof in Appendix G.4).

Lemma E.2. Under the assumptions of Theorem 2.2, at any time $t > T_0$,

$$\widehat{\text{Regret}}_t(\pi, \pi^*) \leq 2 p_0^{(t)} R_\theta + 2 \max_{a \in \{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2.$$

Note that for any time $t > T_0$, the instantaneous regret at time t: $\text{Regret}_t(\pi, \pi^*) = (\pi_t^*(1) - \pi_t(1)) \langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \mathbf{x}_t \rangle$, and that $\widehat{\text{Regret}}_t(\pi, \pi^*) = (\pi_t^*(1) - \pi_t(1)) \langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \widetilde{\mathbf{x}}_t \rangle$. Moreover,

$$\begin{aligned}
|\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \widetilde{\mathbf{x}}_t \rangle| =& |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \mathbf{x}_t \rangle + \langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \boldsymbol{\epsilon}_t \rangle| \\
\geq& |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \mathbf{x}_t \rangle| - |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \boldsymbol{\epsilon}_t \rangle| \\
\geq& \frac{1}{\rho} |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \boldsymbol{\epsilon}_t \rangle| - |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \boldsymbol{\epsilon}_t \rangle| \\
=& \frac{1 - \rho}{\rho} |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \boldsymbol{\epsilon}_t \rangle|.
\end{aligned}$$

Here we used Assumption 2.3. Thus we have

$$\begin{aligned}
\text{Regret}_t(\pi, \pi^*) =& \widehat{\text{Regret}}_t(\pi, \pi^*) - (\pi_t^*(1) - \pi_t(1)) \langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \boldsymbol{\epsilon}_t \rangle \\
\leq& \widehat{\text{Regret}}_t(\pi, \pi^*) + |\pi_t^*(1) - \pi_t(1)| \cdot \frac{\rho}{1 - \rho} |\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \widetilde{\mathbf{x}}_t \rangle| \\
=& \frac{1}{1 - \rho} \widehat{\text{Regret}}_t(\pi, \pi^*).
\end{aligned}$$

Combining Lemma E.2, we obtain that for any $t > T_0$,

$$\text{Regret}_t(\pi, \pi^*) \leq \frac{2}{1 - \rho} \big(p_0^{(t)} R_\theta + \max_{a \in \{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \big).$$

Finally, when $t \leq T_0$, since $\|\mathbf{x}_t\| \leq 1$, $\|\boldsymbol{\theta}_a^*\| \leq R_\theta$, the instantaneous regret $\text{Regret}_t(\pi, \pi^*) \leq 2 R_\theta$. We conclude the proof by summing up all the instantaneous regret terms.

Clipped policy setting. In the clipped policy setting, we give the lemma below (proof in Appendix G.5).

Lemma E.3. Under the assumptions of Theorem 2.2, at any time $t > T_0$,
$$\widehat{\text{Regret}}_t(\pi, \bar{\pi}^*) \leq 2(1 - 2p_0) \max_{a \in \{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2.$$

Note that the instantaneous regret at time t: $\text{Regret}_t(\pi, \bar{\pi}^*) = (\bar{\pi}_t^*(1) - \pi_t(1))\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \mathbf{x}_t \rangle$, and that $\widehat{\text{Regret}}_t(\pi, \bar{\pi}^*) = (\bar{\pi}_t^*(1) - \pi_t(1))\langle \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*, \widetilde{\mathbf{x}}_t \rangle$. Similar to the standard setting, for $t > T_0$, under Assumption 2.3, we have
$$\text{Regret}_t(\pi, \bar{\pi}^*) \leq \frac{1}{1-\rho}\widehat{\text{Regret}}_t(\pi, \bar{\pi}^*) \leq \frac{2(1 - 2p_0)}{1-\rho} \max_{a \in \{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2.$$
We conclude the proof by summing up all the instantaneous regret terms, and noticing that for $t \leq T_0$, $\text{Regret}_t(\pi, \bar{\pi}^*) \leq 2R_\theta$.

Results with a high-probability version of Assumption 2.3. As briefly mentioned in the main paper, Assumption 2.3 can be weakened to the inequalities holding with high probability. Now instead of Assumption 2.3, we assume the following:

Assumption E.1. There exist constants $\rho \in (0, 1)$, $c_e \in [0, 1]$ such that $\sum_{t=1}^T \mathbb{P}(A_t^c) \leq c_e$. Here A_t denotes the event $\{|\langle \boldsymbol{\delta}_\theta, \boldsymbol{\epsilon}_t \rangle| \leq \rho|\langle \boldsymbol{\delta}_\theta, \mathbf{x}_t \rangle|\}$, and $\boldsymbol{\delta}_\theta = \boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*$.

It's easy to see that the result of Lemma E.1 hold at time t under the event A_t. Further, following the same arguments, we obtain that under Assumption 2.1, the results for either the standard setting or the clipped policy setting hold under the event $\cap_{t=T_0+1}^T A_t$. Therefore, with Assumption 2.1 and E.1, in either setting, the results in Theorem 2.2 hold with probability at least $1 - c_e$.

E.3 Proof of Corollary 2.1

Standard setting. First, notice that $q_t = \min_{\tau \leq t, a \in \{0,1\}} \pi_\tau(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{\tau-1}) = p_0^{(t)}$, since $p_0^{(t)}$ is monotonically decreasing in t. Theorem 2.1 indicates that, as long as $\forall t > T_0$,
$$p_0^{(t)} \geq C_1 \max \left\{ \frac{d(d + \log t)}{\lambda_0 t}, \frac{\xi(d + \log t)}{\lambda_0^2 t} \right\}, \tag{E.1}$$
then with probability at least $1 - \frac{8}{t^2}$, $\forall a \in \{0, 1\}$,
$$\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \leq \frac{C(R + R_\theta)d}{\lambda_0} \max \left\{ \frac{d + \log t}{q_t t}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}}\sqrt{\frac{d + \log t}{q_t t}} \right\} \tag{E.2}$$

Plug it into Theorem 2.2, we have that with high probability,
$$\text{Regret}(T; \pi^*) \leq 2R_\theta \cdot \lceil 2dT^{2/3} \rceil + \frac{2}{1-\rho}I_1,$$
where
$$
\begin{aligned}
I_1 &= \sum_{t=T_0+1}^T \left(p_0^{(t)} R_\theta + \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \right) \\
&\leq \sum_{t=T_0+1}^T t^{-\frac{1}{3}} R_\theta + \sum_{t=T_0}^{T-1} \frac{C(R + R_\theta)d}{\lambda_0} \max \left\{ \frac{d + \log t}{t^{\frac{2}{3}}}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}}\sqrt{\frac{d + \log t}{t^{\frac{2}{3}}}} \right\} \\
&\leq 2R_\theta T^{\frac{2}{3}} + \frac{C(R + R_\theta)d}{\lambda_0} \cdot \sum_{t=T_0}^{T-1} \frac{d + \log t}{t^{\frac{2}{3}}} + \frac{C(R + R_\theta)d}{\lambda_0} \cdot \sum_{t=T_0}^{T-1} \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}}\sqrt{\frac{d + \log t}{t^{\frac{2}{3}}}} \\
&\leq 2R_\theta T^{\frac{2}{3}} + \frac{3C(R + R_\theta)}{\lambda_0}d(d + \log T)T^{\frac{1}{3}} + \frac{3C(\sqrt{\nu} + \sqrt{\xi})(R + R_\theta)}{2\lambda_0}\sqrt{d(d + \log T)}T^{\frac{2}{3}} \\
&\leq 2R_\theta T^{\frac{2}{3}} + \frac{C'}{\lambda_0}(\sqrt{\nu} + \sqrt{\xi} + 1)(R + R_\theta)\sqrt{d(d + \log T)}T^{\frac{2}{3}},
\end{aligned}
$$

27

for a universal constant C', where the last inequality holds if in addition, $T \geq [d(d + \log T)]^{\frac{3}{2}}$.

The proof is concluded by combining the above requirement for T as well as (E.1).

Clipped policy setting. Similar to the standard setting, according to Theorem 2.1, as long as $\forall t > T_0$,

$$p_0 \geq C_1 \max \left\{ \frac{d(d + \log t)}{\lambda_0 t}, \frac{\xi(d + \log t)}{\lambda_0^2 t} \right\}, \tag{E.3}$$

then with probability at least $1 - \frac{8}{t^2}$, $\forall a \in \{0, 1\}$,

$$\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \leq \frac{C(R + R_\theta)d}{\lambda_0} \max \left\{ \frac{d + \log t}{p_0 t}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{p_0 t}} \right\} \tag{E.4}$$

Plug it into Theorem 2.2, we have that with high probability,

$$\text{Regret}(T; \bar{\pi}^*) \leq 2R_\theta \cdot \lceil 2dT^{1/2} \rceil + \frac{2(1 - 2p_0)}{1 - \rho} I_2,$$

where

$$
\begin{aligned}
I_2 &= \sum_{t=T_0+1}^{T} \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \\
&\leq \sum_{t=T_0}^{T-1} \frac{C(R + R_\theta)d}{\lambda_0} \max \left\{ \frac{d + \log t}{p_0 t}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{p_0 t}} \right\} \\
&\leq \frac{C(R + R_\theta)d}{\lambda_0} \sum_{t=T_0}^{T-1} \frac{d + \log t}{p_0 t} + \frac{C(R + R_\theta)d}{\lambda_0} \sum_{t=T_0}^{T-1} \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{p_0 t}} \\
&\leq \frac{2C(R + R_\theta)}{\lambda_0} d(d + \log T) \log T + \frac{2C(R + R_\theta)}{\lambda_0} \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{p_0}} \sqrt{d(d + \log T)} \sqrt{T} \\
&\leq \frac{2C(R + R_\theta)(\sqrt{\nu} + \sqrt{\xi} + 1)}{\lambda_0 \sqrt{p_0}} \sqrt{d(d + \log T)} \sqrt{T},
\end{aligned}
$$

where the last inequality holds if in addition, $T \geq d(d + \log T) \log^2 T$.

The proof is concluded by plugging the above into the regret upper bound formula and combining the requirements for T.

Results with a high-probability version of Assumption 2.3. Recall that Assumption E.1 is a weakened version of Assumption 2.3 with a high-probability statement. Given Assumptions 2.1, 2.2, and E.1 (instead of 2.3), in the standard setting, the results of Theorem 2.1 hold as long as (E.2) or (E.4) for all $t > T_0$, $a \in \{0, 1\}$ and under the event $\cap_{t=T_0+1}^{T} A_t$. Thus, we deduce that the regret upper bound in (i) hold with probability at least $1 - 16/\sqrt{T} - c_e$. Similarly, given Assumptions 2.1, 2.2, and E.1, in the clipped benchmark setting, the regret upper bound in (ii) hold with probability at least $1 - 16/\sqrt{T} - c_e$.

E.4 MEB with infrequent model update

As mentioned at the end of Section 2.3, in certain scenarios (e.g. when d is large), we can save computational resources by updating the estimates of $(\boldsymbol{\theta}_a^*)_{a \in \{0,1\}}$ less frequently. In Algorithm 5, we propose a variant of Algorithm 1. At each time t, given the noisy context $\widetilde{\mathbf{x}}_t$, the algorithm computes the best action $\widetilde{a}_t$ according to the most recently updated estimators of $(\boldsymbol{\theta}_a^*)_{a \in \{0,1\}}$. Then, it samples $\widetilde{a}_t$ with probability $1 - p_0^{(t)}$ and keeps an exploration probability of $p_0^{(t)}$ to sample the other action. In the meantime, the agent only has

Algorithm 5 MEB with infrequent model update

Input: $(\boldsymbol{\Sigma}_{e,t})_{t\in[T]}$: covariance sequence of $(\boldsymbol{\epsilon}_t)_t$; $(p_0^{(t)})_{t\in[T]}$: minimum selection probability at each time t; $\mathcal{S} \subset [T]$: set of time points to update model estimates

$\widehat{\boldsymbol{\theta}}_a \leftarrow 0, \forall a \in \{0,1\}$ % $\widehat{\boldsymbol{\theta}}_a$ stores the most recent updated estimate of $\boldsymbol{\theta}_a^*$, only update if $t \in \mathcal{S}$

for time $t = 1, 2, \ldots, T$ **do**

    if $\min_{s\in\mathcal{S}} s \geq t$ **then**

        Sample $a_t \sim \pi_t(\cdot|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$, where $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) \in [p_0^{(t)}, 1 - p_0^{(t)}]$ for all $a \in \{0,1\}$

        % If the model has never been learned before, explore with equal probability

        continue

    end if

    $\widetilde{a}_t \leftarrow \operatorname{argmax}_{a\in\{0,1\}} \langle \widehat{\boldsymbol{\theta}}_a, \widetilde{\mathbf{x}}_t \rangle$

    Sample $a_t \sim \pi_t(\cdot|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1})$, where $\pi_t(a|\widetilde{\mathbf{x}}_t, \mathcal{H}_{t-1}) := \begin{cases} 1 - p_0^{(t)}, & \text{if } a = \widetilde{a}_t \\ p_0^{(t)}, & \text{otherwise} \end{cases}$

    if $t \in \mathcal{S}$ **then**

        $\widehat{\boldsymbol{\theta}}_a \leftarrow \widehat{\boldsymbol{\theta}}_a^{(t)}$ as in (2.5)

    end if

end for

---

to update the estimate of $(\boldsymbol{\theta}_a^*)_{a\in\{0,1\}}$ once in a while to save computation power: The algorithm specifies a subset $\mathcal{S} \subset [T]$ and updates the estimators according to (2.5) only when $t \in \mathcal{S}$.

Under mild conditions, Algorithm 5 achieves the same order of regret upper bound as Algorithm 1, as seen from Theorem E.1 and Corollary E.1 below. They are modified versions of Theorem 2.2 and Corollary 2.1.

Theorem E.1. Let $s_{\min} := \min_{s\in\mathcal{S}} s$ be the first time Algorithm 5 updates the model. Suppose Assumption 2.1 and 2.3 hold.

(i) For the standard setting, for any $T_0 \geq s_{\min}$, Algorithm 5 outputs a policy such that

$$\operatorname{Regret}(T; \pi^*) \leq 2T_0 R_\theta + \frac{2}{1-\rho} \sum_{t\in(T_0,T]} \left(p_0^{(t)} R_\theta + \max_{a\in\{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(s_t)} - \boldsymbol{\theta}_a^*\|_2 \right).$$

(ii) For the clipped policy setting, for any $T_0 \geq s_{\min}$, Algorithm 5 with the choice of $p_0^{(t)} \equiv p_0$ outputs a policy such that

$$\operatorname{Regret}(T; \bar{\pi}^*) \leq 2T_0 R_\theta + \frac{2(1 - 2p_0)}{1-\rho} \sum_{t\in(T_0,T]} \max_{a\in\{0,1\}} \|\widehat{\boldsymbol{\theta}}_a^{(s_t)} - \boldsymbol{\theta}_a^*\|_2.$$

Here for any $t \in [T]$, $s_t := \max\{s \in \mathcal{S} : s < t\}$.

The proof of Theorem E.1 is very similar to that of Theorem 2.2, and is thus omitted.

Corollary E.1. Let Assumption 2.1 to 2.3 hold. There exist constants C, C' such that:

(i) In the standard setting, as long as the set of model update times $\mathcal{S}$ satisfies: (a) $s_{\min} \leq dT^{2/3}$; (b) $\forall t \in (dT^{2/3}, T]$, $s_t = \max\{s \in \mathcal{S} : s < t\} \geq \alpha t$ for some constant $\alpha \in (e^{-d}, 1)$, then $\forall T \geq C'/\alpha \cdot \max\left\{(1 + 1/\lambda_0^{\frac{9}{4}})(d + \log T)^3, (\xi/\lambda_0)^{\frac{9}{4}}(d + \log T)^{\frac{4}{3}}\right\}$, with probability at least $1 - \frac{16}{\sqrt{T}}$, Algorithm 5 with the choice of $p_0^{(t)} = \min\{\frac{1}{2}, t^{-1/3}\}$ achieves $\operatorname{Regret}(T; \pi^*) \leq CdT^{\frac{2}{3}} \left\{ \frac{R_\theta}{\alpha} + \frac{R_\theta}{1-\rho} + \frac{(\sqrt{\nu} + \sqrt{\xi} + 1)(R + R_\theta)}{\alpha^{1/3}\lambda_0(1-\rho)} \sqrt{1 + \frac{\log T}{d}} \right\}.$

(ii) In the clipped policy setting, as long as the set of model update times $\mathcal{S}$ satisfies: (a) $s_{\min} \leq d\sqrt{T}$; (b) $\forall t \in (d\sqrt{T}, T]$, $s_t = \max\{s \in \mathcal{S} : s < t\} \geq \alpha t$ for some constant $\alpha \in (e^{-d}, 1)$, then for any T s.t. $T \geq C'/\alpha \cdot \max\{(d+\log T)^2/(\lambda_0 p_0)^2, \xi^2/\lambda_0^4(1+\log T/d)^2\}$, with probability at least $1 - \frac{16}{\sqrt{T}}$, Algorithm 5 with the choice of $p_0^{(t)} \equiv p_0$ achieves: $\mathrm{Regret}(T; \bar{\pi}^*) \leq Cd\sqrt{T}\left\{\frac{R_\theta}{\alpha} + \frac{(\sqrt{\nu}+\sqrt{\xi}+1)(1-2p_0)(R+R_\theta)}{\lambda_0(1-\rho)\sqrt{\alpha p_0}}\sqrt{1 + \frac{\log T}{d}}\right\}$.

The proof of Corollary E.1 can be directly obtained by combining Theorem 2.1 and E.1 with $T_0 = dT^{2/3}/\alpha$ in the standard setting and $T_0 = dT^{1/2}/\alpha$ in the clipped benchmark setting. Thus, the proof is omitted here.

In Corollary E.1, condition (a) and (b) essentially requires Algorithm 5 not to start learning the model too late, and to keep updating the learned model at least at time points with a 'geometric' growth rate. This covers a wide range of choices of $\mathcal{S}$ in practice. Two typical examples of $\mathcal{S}$ could be: (1) $\mathcal{S} = \{t \in [T] : t = kt_0 \text{ for some } k \in \mathbb{N}^+\}$ (the model is learned every t_0 time points routinely, where t_0 is a constant integer); (2) If $1/\alpha \in \mathbb{N}^+$, $\mathcal{S} = \{t \in [T] : t = (1/\alpha)^k \text{ for some } k \in \mathbb{N}^+\}$ (the model only needs to be learned $\mathcal{O}(\log T)$ times to save computation).

F   Analysis with estimated error variance

We consider the setting where at each time t, the agent does not have access to $\boldsymbol{\Sigma}_{e,t}$, but has a (potentially adaptive) estimator $\widehat{\boldsymbol{\Sigma}}_{e,t}$. In this setting, we estimate the model using (2.7) instead of (2.5) and plug into Algorithm 1. The following theorem controls the estimation error of $\widetilde{\boldsymbol{\theta}}_a^{(t)}$. Note that compared to Theorem 2.1, the additional error caused by inaccuracty of $\widehat{\boldsymbol{\Sigma}}_{e,t}$ can be controlled by $\boldsymbol{\Delta}_t(a) := \frac{1}{t}\sum_{\tau \in [t]} \pi_\tau^{nd}(a)(\widehat{\boldsymbol{\Sigma}}_{e,\tau} - \boldsymbol{\Sigma}_{e,\tau})$, the weighted average of the estimation errors $(\widehat{\boldsymbol{\Sigma}}_{e,\tau} - \boldsymbol{\Sigma}_{e,\tau})_{\tau \in [t]}$.

Theorem F.1. Recall that $q_t := \inf_{\tau \leq t, a \in \{0,1\}} \pi_\tau(a|\widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})$. Then under Assumptions 2.1 and 2.2, there exist constants C and C_1' such that as long as $q_t \geq C_1' \max\left\{\frac{d(d+\log t)}{\lambda_0 t}, \frac{\xi(d+\log t)}{\lambda_0^2 t}\right\}$ and $\max_{a \in \{0,1\}} \|\boldsymbol{\Delta}_t(a)\|_2 \leq \frac{\lambda_0}{4d}$, with probability at least $1 - \frac{8}{t^2}$,

$$\|\widetilde{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \leq \frac{C(R+R_\theta)d}{\lambda_0}\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\xi}+\sqrt{\nu}}{\sqrt{d}}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\} + \frac{2R_\theta d}{\lambda_0}\|\boldsymbol{\Delta}_t(a)\|_2, \quad \forall a \in \{0,1\}. \quad \text{(F.1)}$$

The proof of Theorem F.1 is in Appendix G.6.

By combining Theorem F.1 and 2.2 (with (2.7) instead of (2.5)), we obtain the following regret bounds for Algorithm 1 with (2.7) as the plug-in estimator.

Corollary F.1. Suppose Assumption 2.1 to 2.3 hold. Then there exist universal constants C, C'' such that:

(i) In the standard setting, if $\max_{t \in [dT^{2/3}, T]} \max_{a \in \{0,1\}} \|\boldsymbol{\Delta}_t(a)\|_2 \leq \frac{\lambda_0}{4d}$, $T \geq C'' \max\{(1 + 1/\lambda_0^{\frac{9}{4}})(d + \log T)^3, (\xi/\lambda_0)^{\frac{9}{4}}(d + \log T)^{\frac{4}{3}}\}$, by choosing $T_0 = \lceil 2dT^{\frac{2}{3}}\rceil$, $p_0^{(t)} = \min\{\frac{1}{2}, t^{-\frac{1}{3}}\}$, and (2.7) instead of (2.5) in Algorithm 1, with probability at least $1 - \frac{16}{\sqrt{T}}$,

$$\mathrm{Regret}(T; \pi^*) \leq CdT^{\frac{2}{3}}\left\{\frac{R_\theta}{1-\rho} + \frac{(\sqrt{\nu}+\sqrt{\xi}+1)(R+R_\theta)}{(1-\rho)\lambda_0}\sqrt{1 + \frac{\log T}{d}}\right\} + \frac{4R_\theta d}{(1-\rho)\lambda_0}\sum_{t=T_0-1}^{T}\max_{a \in \{0,1\}}\|\boldsymbol{\Delta}_t(a)\|_2.$$

(ii) In the clipped policy setting, as long as $\max_{t \in [d\sqrt{T}, T]} \max_{a \in \{0,1\}} \|\boldsymbol{\Delta}_t(a)\|_2 \leq \frac{\lambda_0}{4d}$, $T \geq C'' \max\{(d + \log T)^2/(\lambda_0 p_0)^2, \xi^2/\lambda_0^4(1 + \log T/d)^2\}$, by choosing $T_0 = \lceil 2d\sqrt{T}\rceil$, $p_0^{(t)} \equiv p_0$, and (2.7) instead of (2.5) in Algorithm 1, with probability at least $1 - \frac{16}{\sqrt{T}}$,

$$\mathrm{Regret}(T; \bar{\pi}^*) \leq CdT^{\frac{1}{2}}\left\{R_\theta + \frac{(\sqrt{\nu}+\sqrt{\xi}+1)(1-2p_0)(R+R_\theta)}{\sqrt{p_0}(1-\rho)\lambda_0}\sqrt{1 + \frac{\log T}{d}}\right\} + \frac{4(1-2p_0)R_\theta d}{(1-\rho)\lambda_0}\sum_{t=T_0-1}^{T}\max_{a \in \{0,1\}}\|\boldsymbol{\Delta}_t(a)\|_2.$$

30

The proof is in Appendix G.7.

G Additional proofs

G.1 Proof of Theorem C.2

The proof of Theorem C.2 is very similar to Theorem 2.2, and we only need to note the difference in Lemma E.2 and E.3 (for the standard setting and the clipped policy setting respectively), as stated below. Recall that

$$\widehat{\text{Regret}}_t(\pi, \pi^*) := \mathbb{E}_{a \sim \pi_t^*} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle,$$

$$\widehat{\text{Regret}}_t(\pi, \bar{\pi}^*) := \mathbb{E}_{a \sim \bar{\pi}_t^*} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle.$$

Standard setting. At any time $t > T_0$, we have

$$\begin{aligned}
\widehat{\text{Regret}}_t(\pi, \pi^*) &= \mathbb{E}_{a \sim \pi_t^*} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle \\
&= \langle \boldsymbol{\theta}_{a_t^*}^*, \widetilde{\mathbf{x}}_t \rangle - \Big[(1 - (K-1)p_0^{(t)}) \langle \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle + \sum_{a \neq \widetilde{a}_t} p_0^{(t)} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle \Big] \\
&= p_0^{(t)} \sum_{a \neq a_t^*} \langle \boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle + 1_{\{a_t^* \neq \widetilde{a}_t\}} (1 - Kp_0^{(t)}) \langle \boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle \\
&\leq 2(K-1)p_0^{(t)} R_\theta + 1_{\{a_t^* \neq \widetilde{a}_t\}} \langle \boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle.
\end{aligned} \tag{G.1}$$

Here note that Lemma E.1 still holds under Assumption C.1, so $a_t^* = \operatorname{argmax}_a \langle \boldsymbol{\theta}_a^*, \mathbf{x}_t \rangle = \operatorname{argmax}_a \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle$, $\widetilde{a}_t = \operatorname{argmax}_a \langle \widehat{\boldsymbol{\theta}}_a^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle$.

Note that $a_t^* \neq \widetilde{a}_t$ implies that

$$\begin{cases} \langle \boldsymbol{\theta}_{a_t^*}^*, \widetilde{\mathbf{x}}_t \rangle \geq \langle \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle \\ \langle \widehat{\boldsymbol{\theta}}_{a_t^*}^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle \leq \langle \widehat{\boldsymbol{\theta}}_{\widetilde{a}_t}^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle \end{cases}$$

which leads to

$$\begin{aligned}
\langle \boldsymbol{\theta}_{a_t^*}^*, \widetilde{\mathbf{x}}_t \rangle &\geq \langle \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle \geq \langle \widehat{\boldsymbol{\theta}}_{\widetilde{a}_t}^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle - \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \\
&\geq \langle \widehat{\boldsymbol{\theta}}_{a_t^*}^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle - \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \geq \langle \boldsymbol{\theta}_{a_t^*}^*, \widetilde{\mathbf{x}}_t \rangle - 2 \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2,
\end{aligned}$$

and further implies $\left| \langle \boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle \right| \leq 2 \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2$. Plugging in the above to (G.1) leads to

$$\widehat{\text{Regret}}_t(\pi, \pi^*) \leq 2(K-1)p_0^{(t)} R_\theta + 2 \max_a \|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2.$$

The rest of the proof can be done in the same way as the proof of Theorem 2.2.

Clipped policy setting. At any time $t > T_0$, we have

$$\begin{aligned}
\widehat{\text{Regret}}_t(\pi, \bar{\pi}^*) &= \mathbb{E}_{a \sim \pi_t^*} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle - \mathbb{E}_{a \sim \pi_t} \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle \\
&\leq (1 - Kp_0) 1_{\{a_t^* \neq \widetilde{a}_t\}} \left| \langle \boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle \right|.
\end{aligned} \tag{G.2}$$

Here recall that $a_t^* = \operatorname{argmax}_a \langle \boldsymbol{\theta}_a^*, \widetilde{\mathbf{x}}_t \rangle$, $\widetilde{a}_t := \operatorname{argmax}_a \langle \widehat{\boldsymbol{\theta}}_a^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle$.

Note that $a_t^* \neq \widetilde{a}_t$ implies that

$$\begin{cases} \langle \boldsymbol{\theta}_{a_t^*}^*, \widetilde{\mathbf{x}}_t \rangle \geq \langle \boldsymbol{\theta}_{\widetilde{a}_t}^*, \widetilde{\mathbf{x}}_t \rangle \\ \langle \widehat{\boldsymbol{\theta}}_{a_t^*}^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle \leq \langle \widehat{\boldsymbol{\theta}}_{\widetilde{a}_t}^{(t-1)}, \widetilde{\mathbf{x}}_t \rangle \end{cases}$$

31

which leads to

$$\langle \boldsymbol{\theta}^*_{a^*_t}, \widetilde{\mathbf{x}}_t \rangle \geq \langle \boldsymbol{\theta}^*_{\widetilde{a}_t}, \widetilde{\mathbf{x}}_t \rangle \geq \langle \widehat{\boldsymbol{\theta}}^{(t-1)}_{\widetilde{a}_t}, \widetilde{\mathbf{x}}_t \rangle - \max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2$$

$$\geq \langle \widehat{\boldsymbol{\theta}}^{(t-1)}_{a^*_t}, \widetilde{\mathbf{x}}_t \rangle - \max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2 \geq \langle \boldsymbol{\theta}^*_{a^*_t}, \widetilde{\mathbf{x}}_t \rangle - 2\max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2,$$

and further implies $\left| \langle \boldsymbol{\theta}^*_{a^*_t} - \boldsymbol{\theta}^*_{\widetilde{a}_t}, \widetilde{\mathbf{x}}_t \rangle \right| \leq 2\max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2$.

Plugging in the above to (G.2) leads to

$$\widehat{\mathrm{Regret}}_t(\pi, \bar{\pi}^*) \leq 2(1 - Kp_0)\max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2$$

for $t > T_0$. The rest of the proof can be done in the same way as the proof of Theorem 2.2.

G.2   Proof of Lemma D.1

We first analyze $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\widetilde{\mathbf{x}},a}$. Notice that $\widehat{\boldsymbol{\Sigma}}^{(t)}_{\widetilde{\mathbf{x}},a} = \frac{1}{t}\sum_{\tau \in [t]} \boldsymbol{V}_{\tau,a}$, where $\boldsymbol{V}_{\tau,a} = \frac{\pi^{nd}_\tau(A_\tau)}{\pi_\tau(A_\tau | \widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})}1_{\{A_\tau = a\}}\widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top$. For any fixed $\boldsymbol{u} \in \mathbb{S}^{d-1} := \{\boldsymbol{u}' \in \mathbb{R}^d : \|\boldsymbol{u}'\|_2 = 1\}$, $(v_{\boldsymbol{u},\tau,a} := \boldsymbol{u}^\tau[\boldsymbol{V}_{\tau,a} - \mathbb{E}[\boldsymbol{V}_{\tau,a}|\mathcal{H}_{\tau-1}]]\boldsymbol{u})_\tau$ is a martingale difference sequence. Moreover, we can verify that $|v_{\boldsymbol{u},\tau,a}| \leq \frac{2}{q_t}$ and

$$\mathrm{Var}(v_{\boldsymbol{u},\tau,a}|\mathcal{H}_{\tau-1}) \leq \mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{V}_{\tau,a}\boldsymbol{u})^2|\mathcal{H}_{\tau-1}] = \mathbb{E}\left[\left(\frac{\pi^{nd}_\tau(A_\tau)}{\pi_\tau(A_\tau | \widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})} \right)^2 1_{\{A_\tau = a\}}(\boldsymbol{u}^\top \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top \boldsymbol{u})^2 \middle| \mathcal{H}_{\tau-1} \right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_\tau, A_\tau \sim \pi^{nd}_\tau(\cdot)}\left[\left(\frac{\pi^{nd}_\tau(A_\tau)}{\pi_\tau(A_\tau | \widetilde{\mathbf{x}}_\tau, \mathcal{H}_{\tau-1})} \right) 1_{\{A_\tau = a\}}(\boldsymbol{u}^\top \widetilde{\mathbf{x}}_\tau \widetilde{\mathbf{x}}_\tau^\top \boldsymbol{u})^2 \middle| \mathcal{H}_{\tau-1} \right] \leq \frac{1}{q_t}\mathbb{E}(\boldsymbol{u}^\top \widetilde{\mathbf{x}}_\tau)^4 \leq \frac{\xi}{d^2 q_t}.$$

According to Freedman's Inequality (Freedman, 1975), for any $\gamma_1, \gamma_2 > 0$,

$$\mathbb{P}\left(\sum_{\tau \in [t]} v_{\boldsymbol{u},\tau,a} \geq \gamma_1, \sum_{\tau \in [t]} \mathrm{Var}(v_{\boldsymbol{u},\tau,a}|\mathcal{H}_{\tau-1}) \leq \gamma_2 \right) \leq e^{-\frac{\gamma_1^2}{2(\frac{2}{q_t}\gamma_1 + \gamma_2)}}.$$

Set $\gamma_2 = \frac{\xi t}{d^2 q_t}$, and we obtain $\mathbb{P}\left(\sum_{\tau \in [t]} v_{\boldsymbol{u},\tau,a} \geq \gamma_1 \right) \leq e^{-\frac{d^2 q_t \gamma_1^2}{2(2d^2\gamma_1 + \xi t)}}$. Applying the same analysis to $(-v_{\boldsymbol{u},\tau,a})_\tau$ and combining the results gives $\mathbb{P}\left(\left| \sum_{\tau \in [t]} v_{\boldsymbol{u},\tau,a} \right| \geq \gamma_1 \right) \leq 2e^{-\frac{d^2 q_t \gamma_1^2}{2(2d^2\gamma_1 + \xi t)}}$.

Denote $\boldsymbol{M}_t = \frac{1}{t}\sum_{\tau \in [t]}(\boldsymbol{V}_{\tau,a} - \mathbb{E}[\boldsymbol{V}_{\tau,a}|\mathcal{H}_{\tau-1}])$, then the above means that $\forall \boldsymbol{u} \in \mathbb{S}^{d-1}$,

$$\mathbb{P}\left(|\boldsymbol{u}^\top \boldsymbol{M}_t \boldsymbol{u}| \geq \frac{\gamma_1}{t} \right) \leq 2e^{-\frac{d^2 q_t \gamma_1^2}{2(2d^2\gamma_1 + \xi t)}}. \tag{G.3}$$

Let $\mathcal{N}$ be a $\frac{1}{4}$-net of $\mathbb{S}^{d-1}$, $|\mathcal{N}| \leq 9^d$. $\forall \boldsymbol{u} \in \mathbb{S}^{d-1}$, find $\boldsymbol{u}' \in \mathcal{N}$ s.t. $\|\boldsymbol{u} - \boldsymbol{u}'\|_2 \leq \frac{1}{4}$, and we have

$$|\boldsymbol{u}^\top \boldsymbol{M}_t \boldsymbol{u} - \boldsymbol{u}'^\top \boldsymbol{M}_t \boldsymbol{u}'| \leq |\boldsymbol{u}^\top \boldsymbol{M}_t(\boldsymbol{u} - \boldsymbol{u}')| + |\boldsymbol{u}'^\top \boldsymbol{M}_t(\boldsymbol{u} - \boldsymbol{u}')| \leq \frac{1}{2}\|\boldsymbol{M}_t\|_2.$$

This implies that

$$\|\boldsymbol{M}_t\|_2 = \sup_{\boldsymbol{u} \in \mathbb{S}^{d-1}} |\boldsymbol{u}^\top \boldsymbol{M}_t \boldsymbol{u}| \leq \sup_{\boldsymbol{u}' \in \mathcal{N}} |\boldsymbol{u}'^\top \boldsymbol{M}_t \boldsymbol{u}'| + \frac{1}{2}\|\boldsymbol{M}_t\|_2,$$

and thus $\sup_{\boldsymbol{u} \in \mathcal{N}} |\boldsymbol{u}^\top \boldsymbol{M}_t \boldsymbol{u}| \geq \frac{1}{2}\|\boldsymbol{M}_t\|_2$. Combining the above and (G.3), we obtain that for any $\gamma_1 > 0$,

$$\mathbb{P}\left(\|\boldsymbol{M}_t\|_2 \geq \frac{2\gamma_1}{t} \right) \leq \mathbb{P}\left(\sup_{\boldsymbol{u} \in \mathcal{N}} |\boldsymbol{u}^\top \boldsymbol{M}_t \boldsymbol{u}| \geq \frac{\gamma_1}{t} \right) \leq 9^d \cdot \mathbb{P}\left(|\boldsymbol{u}^\top \boldsymbol{M}_t \boldsymbol{u}| \geq \frac{\gamma_1}{t} \right) = 2 \cdot 9^d \cdot e^{-\frac{d^2 q_t \gamma_1^2}{2(2d^2\gamma_1 + \xi t)}}.$$

By choosing $\gamma_1 = 24 \max\{\frac{d+\log t}{q_t}, \frac{\sqrt{\xi}}{d} \cdot \sqrt{\frac{t(d+\log t)}{q_t}}\}$, and noticing that $\boldsymbol{M}_t = \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]} \mathbb{E}[\boldsymbol{V}_{\tau,a}|\mathcal{H}_{\tau-1}]$, we have

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]} \mathbb{E}[\boldsymbol{V}_{\tau,a}|\mathcal{H}_{\tau-1}]\right\|_2 \geq 48\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\xi}}{d}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}\right) \leq \frac{2}{t^2}. \tag{G.4}$$

At the same time, we have

$$\mathbb{E}[\boldsymbol{V}_{\tau,a}|\mathcal{H}_{\tau-1}] = \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\mathbb{E}_{A_\tau\sim\pi_\tau(\cdot|\widetilde{\mathbf{x}}_\tau,\mathcal{H}_{\tau-1})}\left[\frac{\pi_\tau^{nd}(A_\tau)}{\pi_\tau(A_\tau|\widetilde{\mathbf{x}}_\tau,\mathcal{H}_{\tau-1})}\mathbb{1}_{\{A_\tau=a\}}\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top\Big|\boldsymbol{\epsilon}_\tau,\mathcal{H}_{\tau-1}\right]\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\mathbb{E}_{A_\tau\sim\pi_\tau^{nd}(\cdot)}\left[\mathbb{1}_{\{A_\tau=a\}}\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top\Big|\boldsymbol{\epsilon}_\tau,\mathcal{H}_{\tau-1}\right]\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\pi_\tau^{nd}(a)\cdot\widetilde{\mathbf{x}}_\tau\widetilde{\mathbf{x}}_\tau^\top\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \pi_\tau^{nd}(a)(\mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Sigma}_{e,\tau}). \tag{G.5}$$

Here we've used the facts that (i) $\{\pi_\tau^{nd}\}_\tau$ is data-independent; (ii) $\mathbb{E}[\boldsymbol{\epsilon}_\tau|\mathcal{H}_{\tau-1}] = 0$, $\mathrm{Var}[\boldsymbol{\epsilon}_\tau|\mathcal{H}_{\tau-1}] = \boldsymbol{\Sigma}_{e,\tau}$. Plug (G.5) into (G.4), and we get

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]} \pi_\tau^{nd}(a)(\mathbf{x}_\tau\mathbf{x}_\tau^\top + \boldsymbol{\Sigma}_{e,\tau})\right\|_2 \geq 48\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\xi}}{d}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}\right) \leq \frac{2}{t^2}. \tag{G.6}$$

The analysis for $\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)}$ is similar. Write $\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)} = \frac{1}{t}\sum_{\tau\in[t]}\boldsymbol{Z}_{\tau,a}$, $\boldsymbol{Z}_{\tau,a} := \frac{\pi_\tau^{nd}(A_\tau)}{\pi_\tau(A_\tau|\widetilde{\mathbf{x}}_\tau,\mathcal{H}_{\tau-1})}\mathbb{1}_{\{A_\tau=a\}}\widetilde{\mathbf{x}}_\tau r_\tau$. Then for any $\boldsymbol{u}\in\mathcal{S}^{d-1}$, it's easy to verify that $|(\boldsymbol{Z}_{\tau,a} - \mathbb{E}[\boldsymbol{Z}_{\tau,a}|\mathcal{H}_{\tau-1}])^\top\boldsymbol{u}| \leq \frac{2R}{q_t}$, and $\mathrm{Var}((\boldsymbol{Z}_{\tau,a} - \mathbb{E}[\boldsymbol{Z}_{\tau,a}|\mathcal{H}_{\tau-1}])^\top\boldsymbol{u}|\mathcal{H}_{\tau-1}) \leq \mathbb{E}[(\boldsymbol{Z}_{\tau,a}^\top\boldsymbol{u})^2|\mathcal{H}_{\tau-1}] \leq \frac{\nu R^2}{q_t d}$. Applying Freedman's Inequality leads to

$$\mathbb{P}\left(|\boldsymbol{z}_t^\top\boldsymbol{u}| \geq \frac{\gamma_1}{t}\right) \leq 2e^{-\frac{dq_t\gamma_1^2}{4Rd\gamma_1+2R^2\nu t}}, \tag{G.7}$$

where $\boldsymbol{z}_t := \frac{1}{t}\sum_{\tau\in[t]}(\boldsymbol{Z}_{\tau,a} - \mathbb{E}[\boldsymbol{Z}_{\tau,a}|\mathcal{H}_{\tau-1}])$.

Recall that $\mathcal{N}$ is a $\frac{1}{4}$-net of $\mathbb{S}^{d-1}$, $|\mathcal{N}| \leq 9^d$. $\forall\boldsymbol{u}\in\mathbb{S}^{d-1}$, find $\boldsymbol{u}'\in\mathbb{S}^{d-1}$ s.t. $\|\boldsymbol{u}-\boldsymbol{u}'\| \leq 1/4$, then $|\boldsymbol{z}_t^\top\boldsymbol{u} - \boldsymbol{z}_t^\top\boldsymbol{u}'| \leq \frac{1}{4}\|\boldsymbol{z}_t\|_2$, and thus

$$\|\boldsymbol{z}_t\|_2 = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}|\boldsymbol{z}_t^\top\boldsymbol{u}| \leq \sup_{\boldsymbol{u}'\in\mathcal{N}}|\boldsymbol{z}_t^\top\boldsymbol{u}'| + \frac{1}{4}\|\boldsymbol{z}_t\|_2$$

which implies that $\sup_{\boldsymbol{u}\in\mathcal{N}}|\boldsymbol{z}_t^\top\boldsymbol{u}| \geq \frac{3}{4}\|\boldsymbol{z}_t\|_2$. Taking this and (G.7) into account, we derive that

$$\mathbb{P}\left(\|\boldsymbol{z}_t\|_2 \geq \frac{4}{3}\frac{\gamma_1}{t}\right) \leq \mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{N}}|\boldsymbol{z}_t^\top\boldsymbol{u}| \geq \frac{\gamma_1}{t}\right) \leq 9^d\cdot 2e^{-\frac{dq_t\gamma_1^2}{4Rd\gamma_1+2R^2\nu t}}$$

By choosing $\gamma_1 = 24R\max\left\{\frac{d+\log t}{q_t}, \sqrt{\nu}\cdot\sqrt{\frac{(d+\log t)t}{dq_t}}\right\}$ and noticing that $\boldsymbol{z}_t = \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]}\mathbb{E}[\boldsymbol{Z}_{\tau,a}|\mathcal{H}_{\tau-1}]$, we obtain

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)} - \frac{1}{t}\sum_{\tau\in[t]}\mathbb{E}[\boldsymbol{Z}_{\tau,a}|\mathcal{H}_{\tau-1}]\right\|_2 \geq 32R\max\left\{\frac{d+\log t}{q_t t}, \sqrt{\frac{\nu}{d}}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}\right) \leq \frac{2}{t^2}. \tag{G.8}$$

Finally, because

$$\mathbb{E}[\boldsymbol{Z}_{\tau,a}|\mathcal{H}_{\tau-1}] = \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\mathbb{E}_{A_\tau\sim\pi_\tau(\cdot|\widetilde{\mathbf{x}}_\tau,\mathcal{H}_{\tau-1}),\eta_\tau}\left[\frac{\pi_\tau^{nd}(A_\tau)}{\pi_\tau(A_\tau|\widetilde{\mathbf{x}}_\tau,\mathcal{H}_{\tau-1})}1_{\{A_\tau=a\}}\widetilde{\mathbf{x}}_\tau r_\tau\Big|\mathcal{H}_{\tau-1},\boldsymbol{\epsilon}_\tau\right]\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\mathbb{E}_{A_\tau\sim\pi_\tau^{nd}(\cdot),\eta_\tau}\left[1_{\{A_\tau=a\}}\widetilde{\mathbf{x}}_\tau r_\tau|\mathcal{H}_{\tau-1},\boldsymbol{\epsilon}_\tau\right]\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\mathbb{E}_{A_\tau\sim\pi_\tau^{nd}(\cdot),\eta_\tau}\left[1_{\{A_\tau=a\}}\widetilde{\mathbf{x}}_\tau(\mathbf{x}_\tau^\top\boldsymbol{\theta}_a^* + \eta_\tau)|\mathcal{H}_{\tau-1},\boldsymbol{\epsilon}_\tau\right]\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\pi_\tau^{nd}(a)\mathbb{E}_{\eta_\tau}\left[\widetilde{\mathbf{x}}_\tau(\mathbf{x}_\tau^\top\boldsymbol{\theta}_a^* + \eta_\tau)|\mathcal{H}_{\tau-1},\boldsymbol{\epsilon}_\tau\right]\Big|\mathcal{H}_{\tau-1}\right]$$

$$= \pi_\tau^{nd}(a)\mathbb{E}_{\boldsymbol{\epsilon}_\tau}\left[\widetilde{\mathbf{x}}_\tau\mathbf{x}_\tau^\top\boldsymbol{\theta}_a^*\Big|\mathcal{H}_{\tau-1}\right] = \pi_\tau^{nd}(a)(\mathbf{x}_\tau\mathbf{x}_\tau^\top)\boldsymbol{\theta}_a^*,$$

Plug in (G.8), and we obtain

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathbf{x}},r,a}^{(t)} - \left[\frac{1}{t}\sum_{\tau\in[t]}\pi_\tau^{nd}(a)\mathbf{x}_\tau\mathbf{x}_\tau^\top\right]\boldsymbol{\theta}_a^*\right\|_2 \geq 32R\max\left\{\frac{d+\log t}{q_t t}, \sqrt{\frac{\nu}{d}}\cdot\sqrt{\frac{d+\log t}{q_t t}}\right\}\right) \leq \frac{2}{t^2}. \qquad \text{(G.9)}$$

Combining (G.9) and (G.6), we conclude the proof.

G.3   Proof of Lemma E.1

We only need to prove

$$\text{sign}(\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\mathbf{x}_t\rangle) = \text{sign}(\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\widetilde{\mathbf{x}}_t\rangle). \qquad \text{(G.10)}$$

If $\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\mathbf{x}_t\rangle = 0$, (G.10) is a direct consequence of Assumption 2.3. If $\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\mathbf{x}_t\rangle \neq 0$, without loss of generality, suppose $\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\mathbf{x}_t\rangle > 0$. Then according to Assumption 2.3,

$$\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\widetilde{\mathbf{x}}_t\rangle = \langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\mathbf{x}_t\rangle + \langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\boldsymbol{\epsilon}_t\rangle \geq (1-\rho)\langle\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*,\mathbf{x}_t\rangle > 0.$$

Thus (G.10) is true.

G.4   Proof of Lemma E.2

We have

$$\widehat{\text{Regret}}_t(\pi,\pi^*) = \mathbb{E}_{a\sim\pi_t^*}\langle\boldsymbol{\theta}_a^*,\widetilde{\mathbf{x}}_t\rangle - \mathbb{E}_{a\sim\pi_t}\langle\boldsymbol{\theta}_a^*,\widetilde{\mathbf{x}}_t\rangle$$

$$= \langle\boldsymbol{\theta}_{a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle - [(1-p_0^{(t)})\langle\boldsymbol{\theta}_{\widetilde{a}_t}^*,\widetilde{\mathbf{x}}_t\rangle + p_0^{(t)}\langle\boldsymbol{\theta}_{1-\widetilde{a}_t}^*,\widetilde{\mathbf{x}}_t\rangle]$$

$$= 1_{\{a_t^*=\widetilde{a}_t\}}p_0^{(t)}\langle\boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{1-a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle + 1_{\{a_t^*\neq\widetilde{a}_t\}}(1-p_0^{(t)})\langle\boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{1-a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle$$

$$\leq 2p_0^{(t)}R_\theta + 1_{\{a_t^*\neq\widetilde{a}_t\}}\langle\boldsymbol{\theta}_{a_t^*}^* - \boldsymbol{\theta}_{1-a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle. \qquad \text{(G.11)}$$

Here recall that $a_t^* = \text{argmax}_a\langle\boldsymbol{\theta}_a^*,\mathbf{x}_t\rangle = \text{argmax}_a\langle\boldsymbol{\theta}_a^*,\widetilde{\mathbf{x}}_t\rangle$, $\widetilde{a}_t := \text{argmax}_{a\in\{0,1\}}\langle\widehat{\boldsymbol{\theta}}_a^{(t-1)},\widetilde{\mathbf{x}}_t\rangle$.

Note that $a_t^* \neq \widetilde{a}_t$ implies that

$$\begin{cases} \langle\boldsymbol{\theta}_{a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle \geq \langle\boldsymbol{\theta}_{1-a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle \\ \langle\widehat{\boldsymbol{\theta}}_{a_t^*}^{(t-1)},\widetilde{\mathbf{x}}_t\rangle \leq \langle\widehat{\boldsymbol{\theta}}_{1-a_t^*}^{(t-1)},\widetilde{\mathbf{x}}_t\rangle \end{cases}$$

which leads to

$$\langle\boldsymbol{\theta}_{a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle \geq \langle\boldsymbol{\theta}_{1-a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle \geq \langle\widehat{\boldsymbol{\theta}}_{1-a_t^*}^{(t-1)},\widetilde{\mathbf{x}}_t\rangle - \max_a\|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2$$

$$\geq \langle\widehat{\boldsymbol{\theta}}_{a_t^*}^{(t-1)},\widetilde{\mathbf{x}}_t\rangle - \max_a\|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \geq \langle\boldsymbol{\theta}_{a_t^*}^*,\widetilde{\mathbf{x}}_t\rangle - 2\max_a\|\widehat{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2,$$

34

and further implies $\left|\langle\boldsymbol{\theta}^*_{a^*_t} - \boldsymbol{\theta}^*_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle\right| \le 2\max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2$.

Plugging in the above to (G.11) leads to

$$\widehat{\mathrm{Regret}}_t(\pi, \pi^*) \le 2p_0^{(t)} R_\theta + 2\max_a \|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2.$$

G.5 Proof of Lemma E.3

At any time $t > T_0$, we have

$$\begin{aligned}
\widehat{\mathrm{Regret}}_t(\pi, \bar{\pi}^*) &= \mathbb{E}_{a\sim\bar{\pi}^*_t}\langle\boldsymbol{\theta}^*_a, \widetilde{\mathbf{x}}_t\rangle - \mathbb{E}_{a\sim\pi_t}\langle\boldsymbol{\theta}^*_a, \widetilde{\mathbf{x}}_t\rangle \\
&= [(1-p_0)\langle\boldsymbol{\theta}^*_{a^*_t}, \widetilde{\mathbf{x}}_t\rangle + p_0\langle\boldsymbol{\theta}^*_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle] - [(1-p_0)\langle\boldsymbol{\theta}^*_{\widetilde{a}_t}, \widetilde{\mathbf{x}}_t\rangle + p_0\langle\boldsymbol{\theta}^*_{1-\widetilde{a}_t}, \widetilde{\mathbf{x}}_t\rangle] \\
&\le (1-2p_0)1_{\{a^*_t\ne\widetilde{a}_t\}}\left|\langle\boldsymbol{\theta}^*_{a^*_t} - \boldsymbol{\theta}^*_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle\right|.
\end{aligned} \tag{G.12}$$

Here recall that $a^*_t = \mathrm{argmax}_a\langle\boldsymbol{\theta}^*_a, \mathbf{x}_t\rangle = \mathrm{argmax}_a\langle\boldsymbol{\theta}^*_a, \widetilde{\mathbf{x}}_t\rangle$, $\widetilde{a}_t := \mathrm{argmax}_{a\in\{0,1\}}\langle\widehat{\boldsymbol{\theta}}^{(t-1)}_a, \widetilde{\mathbf{x}}_t\rangle$.

Note that $a^*_t \ne \widetilde{a}_t$ implies that

$$\begin{cases}
\langle\boldsymbol{\theta}^*_{a^*_t}, \widetilde{\mathbf{x}}_t\rangle \ge \langle\boldsymbol{\theta}^*_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle \\
\langle\widehat{\boldsymbol{\theta}}^{(t-1)}_{a^*_t}, \widetilde{\mathbf{x}}_t\rangle \le \langle\widehat{\boldsymbol{\theta}}^{(t-1)}_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle
\end{cases}$$

which leads to

$$\begin{aligned}
\langle\boldsymbol{\theta}^*_{a^*_t}, \widetilde{\mathbf{x}}_t\rangle &\ge \langle\boldsymbol{\theta}^*_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle \ge \langle\widehat{\boldsymbol{\theta}}^{(t-1)}_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle - \max_a\|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2 \\
&\ge \langle\widehat{\boldsymbol{\theta}}^{(t-1)}_{a^*_t}, \widetilde{\mathbf{x}}_t\rangle - \max_a\|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2 \ge \langle\boldsymbol{\theta}^*_{a^*_t}, \widetilde{\mathbf{x}}_t\rangle - 2\max_a\|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2,
\end{aligned}$$

and further implies $\left|\langle\boldsymbol{\theta}^*_{a^*_t} - \boldsymbol{\theta}^*_{1-a^*_t}, \widetilde{\mathbf{x}}_t\rangle\right| \le 2\max_a\|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2$.

Plugging in the above to (G.12) leads to

$$\widehat{\mathrm{Regret}}_t(\pi, \bar{\pi}^*) \le 2(1-2p_0)\max_a\|\widehat{\boldsymbol{\theta}}^{(t-1)}_a - \boldsymbol{\theta}^*_a\|_2$$

G.6 Proof of Theorem F.1

Fix $t \in [T]$ such that the conditions of Theorem 2.3 hold. Fix $a \in \{0,1\}$. As in Appendix D, define $\boldsymbol{\Delta}_1 := \widehat{\boldsymbol{\Sigma}}^{(t)}_{\widetilde{\mathbf{x}},a} - \frac{1}{t}\sum_{\tau\in[t]}\pi^{nd}_\tau(a)(\mathbf{x}_\tau\mathbf{x}^\top_\tau + \boldsymbol{\Sigma}_{e,\tau})$, $\boldsymbol{\Delta}_2 := \widehat{\boldsymbol{\Sigma}}^{(t)}_{\widetilde{\mathbf{x}},r,a} - \left(\frac{1}{t}\sum_{\tau\in[t]}\pi^{nd}_\tau(a)\mathbf{x}_\tau\mathbf{x}^\top_\tau\right)\boldsymbol{\theta}^*_a$. We also let $\boldsymbol{\Delta}_3 := -\boldsymbol{\Delta}_t(a) = -\frac{1}{t}\sum_{\tau\in[t]}\pi^{nd}_\tau(a)(\widehat{\boldsymbol{\Sigma}}_{e,\tau} - \boldsymbol{\Sigma}_{e,\tau})$. Recall Lemma D.1: with probability at least $1 - 4/t^2$,

$$\|\boldsymbol{\Delta}_1\|_2 \le C\max\left\{\frac{d+\log t}{q_t t}, \frac{\sqrt{\xi}}{d}\sqrt{\frac{d+\log t}{q_t t}}\right\}, \quad \|\boldsymbol{\Delta}_2\|_2 \le CR\max\left\{\frac{d+\log t}{q_t t}, \sqrt{\frac{\nu}{d}}\sqrt{\frac{d+\log t}{q_t t}}\right\}.$$

Meanwhile,

$$\begin{aligned}
\widetilde{\boldsymbol{\theta}}^{(t)}_a &= \left(\widehat{\boldsymbol{\Sigma}}^{(t)}_{\widetilde{\mathbf{x}},a} - \frac{1}{t}\sum_{\tau\in[t]}\pi^{nd}_\tau(a)\widehat{\boldsymbol{\Sigma}}_{e,\tau}\right)^{-1} \cdot \widehat{\boldsymbol{\Sigma}}^{(t)}_{\widetilde{\mathbf{x}},r,a} \\
&= \left(\frac{1}{t}\sum_{\tau\in[t]}\pi^{nd}_\tau(a)\mathbf{x}_\tau\mathbf{x}^\top_\tau + \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_3\right)^{-1} \cdot \left[\left(\frac{1}{t}\sum_{\tau\in[t]}\pi^{nd}_\tau(a)\mathbf{x}_\tau\mathbf{x}^\top_\tau\right)\boldsymbol{\theta}^*_a + \boldsymbol{\Delta}_2\right] \\
&= \boldsymbol{\theta}^*_a - \boldsymbol{J}'_1 + \boldsymbol{J}'_2,
\end{aligned}$$

35

where

$$\boldsymbol{J}_1' := \left[\frac{1}{t} \sum_{\tau \in [t]} \pi_\tau^{nd}(a) \mathbf{x}_\tau \mathbf{x}_\tau^\top + \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_3 \right]^{-1} (\boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_3) \boldsymbol{\theta}_a^*, \quad \boldsymbol{J}_2' := \left[\frac{1}{t} \sum_{\tau \in [t]} \pi_\tau^{nd}(a) \mathbf{x}_\tau \mathbf{x}_\tau^\top + \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_3 \right]^{-1} \boldsymbol{\Delta}_2.$$

Under the events where both (D.1) and (D.2) hold, whenever

$$C \max \left\{ \frac{d + \log t}{q_t t}, \frac{\sqrt{\xi}}{d} \sqrt{\frac{d + \log t}{q_t t}} \right\} \le \frac{\lambda_0}{4d} \tag{G.13}$$

and $\|\boldsymbol{\Delta}_3\|_2 \le \frac{\lambda_0}{4d}$, we have

$$\|\boldsymbol{J}_1'\|_2 \le \frac{2dR_{\boldsymbol{\theta}}}{\lambda_0} \left(C \max \left\{ \frac{d + \log t}{q_t t}, \frac{\sqrt{\xi}}{d} \sqrt{\frac{d + \log t}{q_t t}} \right\} + \|\boldsymbol{\Delta}_3\|_2 \right)$$

and

$$\|\boldsymbol{J}_2'\|_2 \le \frac{2CdR}{\lambda_0} \max \left\{ \frac{d + \log t}{q_t t}, \sqrt{\frac{\nu}{d}} \sqrt{\frac{d + \log t}{q_t t}} \right\}.$$

(G.13) can be ensured by $t \ge C_1' \max \left\{ \frac{d(d + \log t)}{\lambda_0 q_t}, \frac{\xi(d + \log t)}{\lambda_0^2 q_t} \right\}$, where $C_1' = \max\{4C, 16C^2\}$. Given these guarantees, we have with probability at least $1 - 4/t^2$,

$$\|\widetilde{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\| \le \|\boldsymbol{J}_1'\|_2 + \|\boldsymbol{J}_2'\|_2 \le \frac{2C(R + R_{\boldsymbol{\theta}})d}{\lambda_0} \max \left\{ \frac{d + \log t}{q_t t}, \frac{\sqrt{\xi} + \sqrt{\nu}}{\sqrt{d}} \sqrt{\frac{d + \log t}{q_t t}} \right\} + \frac{2R_{\boldsymbol{\theta}} d}{\lambda_0} \|\boldsymbol{\Delta}_t(a)\|_2.$$

Thus we conclude the proof.

G.7   Proof of Corollary F.1

Standard setting. Notice that $p_0^{(t)}$ is monotonically decreasing in t. Theorem F.1 indicates that, as long as $\forall t > T_0$,

$$p_0^{(t)} \ge C' \max \left\{ \frac{d(d + \log t)}{\lambda_0 t}, \frac{\xi(d + \log t)}{\lambda_0^2 t} \right\}, \tag{G.14}$$

then with probability at least $1 - \frac{8}{t^2}$, $\forall a \in \{0, 1\}$,

$$\|\widehat{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \le \frac{C(R + R_{\boldsymbol{\theta}})d}{\lambda_0} \max \left\{ \frac{d + \log t}{t^{\frac{2}{3}}}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{t^{\frac{2}{3}}}} \right\} + \frac{2R_{\boldsymbol{\theta}} d}{\lambda_0} \|\boldsymbol{\Delta}_t(a)\|_2 \tag{G.15}$$

Plug it into Theorem 2.2, we have that with high probability,

$$\text{Regret}(T; \pi^*) \le 2R_{\boldsymbol{\theta}} \cdot \lceil 2dT^{2/3} \rceil + \frac{2}{1 - \rho} I_1',$$

where

$$\begin{aligned} I_1 &= \sum_{t=T_0+1}^{T} \left(p_0^{(t)} R_{\boldsymbol{\theta}} + \max_a \|\widetilde{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \right) \\ &\le \sum_{t=T_0+1}^{T} t^{-\frac{1}{3}} R_{\boldsymbol{\theta}} + \sum_{t=T_0}^{T-1} \left[\frac{C(R + R_{\boldsymbol{\theta}})d}{\lambda_0} \max \left\{ \frac{d + \log t}{t^{\frac{2}{3}}}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{t^{\frac{2}{3}}}} \right\} + \max_a \frac{2R_{\boldsymbol{\theta}} d}{\lambda_0} \|\boldsymbol{\Delta}_t(a)\|_2 \right] \\ &\le 2R_{\boldsymbol{\theta}} T^{\frac{2}{3}} + \frac{C'}{\lambda_0} (\sqrt{\nu} + \sqrt{\xi} + 1)(R + R_{\boldsymbol{\theta}}) \sqrt{d(d + \log T)} T^{\frac{2}{3}} + \frac{2R_{\boldsymbol{\theta}} d}{\lambda_0} \sum_{t \ge T_0} \|\boldsymbol{\Delta}_t(a)\|_2, \end{aligned}$$

for a universal constant C', where the last inequality holds if in addition, $T \geq [d(d + \log T)]^{\frac{3}{2}}$.

The proof is concluded by combining the above requirement for T as well as (G.14).

Clipped policy setting.

Similar to the standard setting, according to Theorem F.1, as long as $\forall t > T_0$,

$$p_0 \geq C_1 \max\left\{ \frac{d(d + \log t)}{\lambda_0 t}, \frac{\xi(d + \log t)}{\lambda_0^2 t} \right\}, \tag{G.16}$$

then with probability at least $1 - \frac{8}{t^2}$, $\forall a \in \{0, 1\}$,

$$\|\widetilde{\boldsymbol{\theta}}_a^{(t)} - \boldsymbol{\theta}_a^*\|_2 \leq \frac{C(R + R_\theta)d}{\lambda_0} \max\left\{ \frac{d + \log t}{p_0 t}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{p_0 t}} \right\} + \frac{2R_\theta d}{\lambda_0} \|\boldsymbol{\Delta}_t(a)\|_2 \tag{G.17}$$

Plug it into Theorem 2.2, we have that with high probability,

$$\mathrm{Regret}(T; \bar{\pi}^*) \leq 2R_\theta \cdot \lceil 2dT^{1/2} \rceil + \frac{2(1 - 2p_0)}{1 - \rho} I_2',$$

where

$$
\begin{aligned}
I_2' &= \sum_{t=T_0+1}^{T} \max_a \|\widetilde{\boldsymbol{\theta}}_a^{(t-1)} - \boldsymbol{\theta}_a^*\|_2 \\
&\leq \sum_{t=T_0}^{T-1} \left[\frac{C(R + R_\theta)d}{\lambda_0} \max\left\{ \frac{d + \log t}{p_0 t}, \frac{\sqrt{\nu} + \sqrt{\xi}}{\sqrt{d}} \sqrt{\frac{d + \log t}{p_0 t}} \right\} + \max_a \frac{2R_\theta d}{\lambda_0} \|\boldsymbol{\Delta}_t(a)\|_2 \right] \\
&\leq \frac{2C(R + R_\theta)(\sqrt{\nu} + \sqrt{\xi} + 1)}{\lambda_0 \sqrt{p_0}} \sqrt{d(d + \log T)} \sqrt{T} + \frac{2R_\theta d}{\lambda_0} \sum_{t=T_0}^{T-1} \max_a \|\boldsymbol{\Delta}_t(a)\|_2,
\end{aligned}
$$

where the last inequality holds if in addition, $T \geq d(d + \log T) \log^2 T$.

The proof is concluded by plugging the above into the regret upper bound formula and combining the requirements for T.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

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

Susan Athey, Stefan Wager, et al. Efficient policy learning. Technical report, 2017.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Marwa Banna, Florence Merlevède, and Pierre Youssef. Bernstein-type inequality for a class of dependent random matrices. *Random Matrices: Theory and Applications*, 5(02):1650006, 2016.

Samuel L Battalio, David E Conroy, Walter Dempsey, Peng Liao, Marianne Menictas, Susan Murphy, Inbal Nahum-Shani, Tianchen Qian, Santosh Kumar, and Bonnie Spring. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.

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

Djallel Bouneffouf. Online learning with corrupted context: Corrupted contextual bandits. *arXiv preprint arXiv:2006.15194*, 2020.

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

Raymond J Carroll, David Ruppert, and Leonard A Stefanski. *Measurement error in nonlinear models*, volume 105. CRC press, 1995.

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.

Sheldon Cohen, Robin Mermelstein, Tom Kamarck, and Harry M Hoberman. Measuring the functional components of social support. *Social support: Theory, research and applications*, pages 73–94, 1985.

Walter Dempsey, Peng Liao, Pedja Klasnja, Inbal Nahum-Shani, and Susan A Murphy. Randomised trials for the fitbit generation. *Significance*, 12(6):20–23, 2015.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 7111–7123. PMLR, 2022.

Jianqing Fan and Qiwei Yao. *The elements of financial econometrics*. Cambridge University Press, 2017.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

Alexander Galozy and Sławomir Nowaczyk. Information-gathering in latent bandits. *Knowledge-Based Systems*, 260:110099, 2023.

Alexander Galozy, Slawomir Nowaczyk, and Mattias Ohlsson. A new bandit setting balancing information from state evolution and corrupted context. *arXiv preprint arXiv:2011.07989*, 2020.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. *Advances in Neural Information Processing Systems*, 33:13423–13433, 2020a.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Mohammad Ghavamzadeh, and Craig Boutilier. Non-stationary latent bandits. *arXiv preprint arXiv:2012.00386*, 2020b.

Sharu Theresa Jose and Shana Moothedath. Thompson sampling for stochastic bandits with noisy contexts: An information-theoretic regret analysis. *arXiv preprint arXiv:2401.11565*, 2024.

Johannes Kirschner and Andreas Krause. Stochastic bandits with context distributions. *Advances in Neural Information Processing Systems*, 32, 2019.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220, 2015.

Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 09 2018.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A Murphy. Sample size calculations for micro-randomized trials in mhealth. *Statistics in medicine*, 35(12):1944–1971, 2016.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

Jiabin Lin and Shana Moothedath. Distributed stochastic bandit learning with context distributions. *arXiv preprint arXiv:2207.14391*, 2022.

Jiabin Lin, Xian Yeow Lee, Talukder Jubery, Shana Moothedath, Soumik Sarkar, and Baskar Ganapathysubramanian. Stochastic conservative contextual linear bandits. *arXiv preprint arXiv:2203.15629*, 2022.

Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pages 161–168, 2014.

Elliot Nelson, Debarun Bhattacharjya, Tian Gao, Miao Liu, Djallel Bouneffouf, and Pascal Poupart. Linearizing contextual bandits with latent state dynamics. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Hongju Park and Mohamad Kazem Shirani Faradonbeh. Worst-case performance of greedy policies in bandits with imperfect context observations. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 1374–1379. IEEE, 2022.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4489–4501, 2016.

Hillol Sarker, Karen Hovsepian, Soujanya Chatterjee, Inbal Nahum-Shani, Susan A Murphy, Bonnie Spring, Emre Ertin, Mustafa Al'Absi, Motohiro Nakajima, and Santosh Kumar. *From markers to interventions: The case of just-in-time stress intervention*. Springer, 2017.

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR, 2017.

Hammad Shaikh, Arghavan Modiri, Joseph Jay Williams, and Anna N Rafferty. Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*, 2019.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Reward design for an online reinforcement learning algorithm supporting oral self-care. *arXiv preprint arXiv:2208.07406*, 2022.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Jianyu Xu and Yu-Xiang Wang. Towards agnostic feature-based dynamic pricing: Linear policies vs linear valuation with unknown noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9643–9662. PMLR, 2022.

Xiongxiao Xu, Hong Xie, and John CS Lui. Generalized contextual bandits with latent features: Algorithms and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Jeremy Yang, Dean Eckles, Paramveer Dhillon, and Sinan Aral. Targeting for long-term outcomes. *arXiv preprint arXiv:2010.15835*, 2020a.

Jianyi Yang and Shaolei Ren. Bandit learning with predicted context: Regret analysis and selective context query. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021a.

Jianyi Yang and Shaolei Ren. Robust bandit learning with imperfect context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10594–10602, 2021b.

Luting Yang, Jianyi Yang, and Shaolei Ren. Multi-feedback bandit learning with probabilistic contexts. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence Main track*, 2020b.

Jiayu Yao, Emma Brunskill, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Power constrained bandits. In *Machine Learning for Healthcare Conference*, pages 209–259. PMLR, 2021.

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

Se-Young Yun, Jun Hyun Nam, Sangwoo Mo, and Jinwoo Shin. Contextual multi-armed bandits under feature uncertainty. *arXiv preprint arXiv:1703.01347*, 2017.

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

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

Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. *arXiv preprint arXiv:1604.06743*, 2016.