CONFOUNDED LINEAR CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

This work studies *confounded contextual bandits*, where the reward is modeled as a linear function of the action, additively confounded by a *stochastic* term that is independent of the action. When the confounder is strong, it can dominate and attenuate the treatment effect induced by the chosen action, making reward inference from bandit feedback challenging. To address this, we draw on Neyman orthogonalization and propose a Thompson sampling based algorithm, INF-TS, which achieves $\tilde{O}(d\sqrt{T \log N})$ regret over a horizon of T, where d denotes the dimension of the context-action tensor product and N is the number of actions. We also prove a lower bound of $\Omega(\sqrt{dT \log N})$ for the confounded linear contextual bandit problem, showing that INF-TS attains the optimal regret rate in T. Empirical results demonstrate that INF-TS outperforms existing methods designed for adversarial confounders in linear contextual bandits (Greenewald et al., 2017; Krishnamurthy et al., 2018) under both stochastic and adversarial confounder settings, particularly when the confounding signal dominates the treatment effect.

1 INTRODUCTION

Many online personalized services require an agent to make sequential decisions based on incoming user information, with the objective of optimizing a target metric, commonly referred to as the reward. Relying solely on observed rewards, effective algorithms must carefully balance the *exploration–exploitation* trade-off: they must explore potentially suboptimal actions to gather information while exploiting actions known to perform well. The *contextual bandit* framework provides a natural model for such problems, capturing the personalization aspect of these settings while enabling principled exploration–exploitation strategies. Among these, the *linear contextual bandit* is one of the most widely studied and applied models, owing to its simple parametric structure and efficient estimation of treatment effects via maximum likelihood methods. This framework has been successfully deployed in a variety of real-world domains, including online recommendation systems (Li et al., 2010; Schwartz et al., 2017; Wang et al., 2024) and clinical trials (Tewari & Murphy, 2017; Woodroofe, 1979).

A key component of successful learning in linear contextual bandits is the accurate estimation of the underlying linear treatment effect parameter from bandit feedback. In the classical literature (Auer, 2002; Filippi et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013), this is typically guaranteed under an *unconfounded* realizability assumption, which states that the expected reward is linear in the chosen action. However, empirical evidence suggests that, in many real-world decision-making problems, observed rewards often contain a substantial context-dependent confounding component, while the action-dependent treatment effect is comparatively small (Bareinboim et al., 2015). For instance, in recommendation systems, rewards frequently conflate user-specific contextual confounders with the effect of the recommended item itself; such contextual influences are also commonly referred to as baseline rewards (Sharma et al., 2015; Sato et al., 2020; Wang et al., 2020). In these settings, the reward includes a context-dependent but action-independent confounder.

The presence of such confounding violates the unconfounded realizability assumption and introduces significant challenges for online causal inference, as isolating the treatment effect from the confounder using only bandit feedback becomes difficult—particularly when the contextual signal is strong and the treatment effect is weak. This motivates the following question: How can one efficiently identify the treatment effect in the presence of *dominant* confounders in online contextual bandits, while retaining sublinear regret? This work answers this question in the setting of confounded contextual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bandits, where the reward is modeled as the sum of a linear function of the context–action pair and a stochastic, action-independent but context-dependent confounder that may dominate the treatment effect.

It is worth noting that causal inference under confounding has been extensively studied in the offline setting. In the double machine learning (DML) literature (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023; Robinson, 1988), causal effects are estimated via doubly robust procedures (Robinson, 1988) based on Neyman orthogonalization (Neyman, 1979), using independently and identically distributed samples. These approaches achieve parametric convergence rates despite potentially slow nuisance parameter (in this case, the confounders) estimation. Motivated by this line of research, this paper contributes the following.

Our contributions. We extend Neyman orthogonalization to the online setting and, in Section 3.2, propose INF-TS, a Thompson sampling (TS) algorithm that performs residual-on-residual regression (Robinson, 1988) to remove the confounding component for online causal inference.

Theoretically (Section 3.3), we show that under stochastic action-independent confounders, INF-TS achieves $\tilde{O}(d\sqrt{T \log N})$ regret, where d is the dimension of the context–action tensor product and N is the number of actions. This matches canonical TS methods for linear contextual bandits (Agrawal & Goyal, 2013) and their generalized-linear-model extensions (Kveton et al., 2020). We further prove a lower bound of $\Omega(\sqrt{dT \log N})$, establishing that INF-TS is optimal in the horizon T but leaving a $\tilde{O}(\sqrt{d})$ gap.

Empirically, we compare INF-TS against standard benchmarks (Agrawal & Goyal, 2013) and prior algorithms designed for adversarial confounders (Krishnamurthy et al., 2018) on synthetic and real-world datasets. Under stochastic confounding, INF-TS substantially outperforms these methods, with gains growing as the confounder strengthens; under adversarial confounding, it remains competitive with (Krishnamurthy et al., 2018). These results demonstrate that INF-TS is both effective under stochastic confounders and robust under adversarial confounders.

2 PRELIMINARIES

We consider an online linear stochastic bandit problem with action-dependent features and an action-independent confounder. In this setting, the confounding term may dominate and attenuate the treatment effect by the chosen action, thereby complicating its inference.

The agent interacts with the environment for a total of T horizon. At each time $t \in [T]$, a user arrives with context $x_t \in \mathbb{R}^{d_x}$. The agent is also shown an action set $\mathcal{A}_t = \{a_t^1, \ldots, a_t^N\}$ with N actions, where each action $a_t^j \in \mathbb{R}^{d_a}$ is represented by a continuous embedding corresponding to its features. The agent follows an online policy $\pi_t : \mathbb{R}^{d_x} \to \Delta(\mathcal{A}_t)$ at time t which decides on the action given the context x_t and conditioned on the observed history. For simplicity, we write $\pi = \{\pi_t\}_{t\in[T]}$. The agent chooses $a_t = \pi_t(x_t)$ and receives the reward

$$r_t(a_t) = x_t^\top \mathbf{\Theta}\, a_t + \phi(x_t) + \varepsilon_t, \tag{1}$$

where $\mathbf{\Theta} \in \mathbb{R}^{d_x \times d_a}$ parametrizes heterogeneous treatment effects of user context and the action, $\phi : \mathbb{R}^{d_x} \to \mathbb{R}$ is a confounding term that captures user-specific baselines that does not depend on the chosen action a_t, and ε_t is a noise term that is centered and independent of a_t.

The stochastic environment is formalized in the following assumption.

Assumption 2.1 (Stochastic environment)**.** We assume that $x_t, \mathcal{A}_t, \varepsilon_t$ are generated at the beginning of time t before a_t is chosen. The noise term ε_t's are i.i.d. σ^2-sub-Gaussian and $\mathbb{E}[\varepsilon \mid x_t, a_t] = 0$.

We present the following examples to motivate the reward structure as in (1).

Example 2.2 (Online recommendation system). Consider an online movie recommendation platform such as Netflix. At each time t, a user arrives with context x_t, which may include demographic attributes and latent preference features. The platform also observes the current catalog $\mathcal{A}_t$ of N movies, where N may be large and may grow over time as new content is added. In this setting, the confounded reward structure in (1) has been studied in the literature (Sharma et al., 2015; Sato et al., 2020; Wang et al., 2020), where ϕ captures an action-independent baseline engagement effect. For

example, some users may be generally more likely to engage positively with content regardless of which movie is recommended, and such tendencies are reflected in their contexts.

Example 2.3 (Online advertising). Consider an online advertising platform that decides which ad to show to a user visiting the page. The catalog $\mathcal{A}_t$ contains eligible ads on the platform at time t. The linear treatment effect in (1) captures the incremental match between the impression context and the ad, while the confounding term ϕ captures the user's intent to click, independent of the particular ad shown (Bottou et al., 2013). For example, a user who typed a highly commercial search query may already be likely to buy something, even before seeing any ad.

We also impose mild regularity assumptions on the problem parameters so that the reward is bounded.

Assumption 2.4 (Bounded mean reward)**.** We assume that $\|\text{vec}(\Theta)\|_2 \leq 1$ and the contexts and action embeddings are also bounded, i.e., $\|x_t\|_2 \leq 1$ and $\|a_t^j\|_2 \leq 1, \forall j \in [N]$, $t \in [T]$. The expected reward itself is always finite: $|x_t^\top \mathbf{\Theta} a_t + \phi(x_t)| \leq M$ for some positive constant $M > 0$.

The boundedness of the problem parameters ensures that the rewards are bounded in expectation—a standard assumption in the literature (Agrawal & Goyal, 2013; Li et al., 2017; Kveton et al., 2020; Krishnamurthy et al., 2018). Moreover, the boundedness of the context and action embeddings implies that $\|a_t^j \otimes x_t\|_2 \leq 1$ for all $j \in [N]$ and $t \in [T]$. We later write $a_t^j \otimes x_t \in \mathbb{R}^d$ with $d := d_a d_x$. In practice, the reward upper bound M can be enforced via normalization of the observed rewards.

Given some user context x_t and action set $\mathcal{A}_t$ at time t, in our stochastic environment described in Assumption 2.1, the optimal action a_t^* at time t maximizes the expected reward:

$$a_t^* := \arg\max_{a \in \mathcal{A}_t} r_t(a) = \arg\max_{a \in \mathcal{A}_t} x_t^\top \mathbf{\Theta} a.$$

For any online policy π that outputs a sequence of recommendation decisions $\{a_t\}_{t\in[T]}$, we evaluate its performance by its cumulative regret over T horizon:

$$\text{Regret}_T(\{a_t\}_{t\in[T]}) := \sum_{t=1}^{T} r_t(a_t^*) - r_t(a_t) = \sum_{t=1}^{T} x_t^\top \mathbf{\Theta} \, (a_t^* - a_t). \tag{2}$$

2.1 Related works

Our setting is closely related to the *semiparametric contextual bandit* literature (Greenewald et al., 2017; Krishnamurthy et al., 2018), which incorporates *adversarial* confounding terms in the reward structure. Greenewald et al. (2017) is motivated by mobile-health interventions, where user behavior has a possibly time-varying confounding term. Its methodology uses a required zero action and an action-centering trick and provides a TS algorithm using a new reward estimator that eliminates the confounding term. Krishnamurthy et al. (2018) is more closely related to our setting, which achieves near-optimal regret while remaining robust to action-independent adversarial confounding. They propose an elimination-based algorithm, bandit orthogonalized semiparametric estimation (BOSE), which centers actions so that the confounding term vanishes in expectation. Under adversarially confounded rewards, Greenewald et al. (2017) relies on a special zero-action structure to eliminate confounding, whereas Krishnamurthy et al. (2018) adopts an orthogonalization-based strategy.

In contrast, our framework focuses on *dominant confounders that overwhelm the reward signal*, rather than adversarially chosen ones. As illustrated in Examples 2.2 and 2.3, many real-world applications feature stochastic confounders that carry strong contextual signals, posing significant challenges for estimating the treatment effect $\mathbf{\Theta}$. In the absence of a zero-action structure, the TS approach of Greenewald et al. (2017) is not applicable. On the other hand, although Krishnamurthy et al. (2018) mitigates bias from adversarial confounding via action-centering, BOSE is not adaptive to the magnitude of the confounder and may therefore suffer degraded performance when the confounding signal is strong.

Other related work includes linear stochastic bandits, such as confidence-bound–based algorithms (Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011), Thompson sampling approaches (Agrawal & Goyal, 2013; Russo & Van Roy, 2014), and generalized linear bandits (Kveton et al., 2020; Filippi et al., 2010; Li et al., 2017). These methods typically assume an unconfounded reward structure; consequently, in the presence of strong confounders, they cannot reliably identify the treatment effect, leading to degraded performance.

Another line of research on *agnostic* stochastic bandits adopts weaker modeling assumptions on the reward structure and learns over a broad class of decision policies (Langford & Zhang, 2007; Agarwal et al., 2014). Such approaches are inherently more robust to model misspecification and potential confounding in the reward signal. However, this generality typically comes at the cost of weaker statistical guarantees and significant computational overhead, which can be NP-hard in the worst case. Consequently, in our parametric reward setting, these methods are often computationally prohibitive and impractical to apply.

Our approach also parallels ideas from semiparametric statistics and DML (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023; Robinson, 1988). Building on Neyman orthogonalization (Neyman, 1979) and moment-based estimation (Chernozhukov et al., 2018), these methods achieve parametric convergence rates despite potentially poor estimation of nuisance components (e.g., propensity scores and confounding terms), through double robustness (Robins & Rotnitzky, 1992). However, such estimators do not directly extend to online settings with adaptively chosen actions. To address this limitation, we generalize these techniques via residual-on-residual regression (Robinson, 1988), and show that, under adaptive online policies, the resulting estimator supports decision-making with sublinear regret.

3 Confounded linear contextual bandit learning

3.1 Treatment effect inference through orthogonalization

The main challenge of our confounded linear contextual bandit problem is the estimation of treatment effect parameter $\mathbf{\Theta}$. To this end, we borrow ideas from the semiparametric statistics literature (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023; Robinson, 1988), in particular, the Neyman orthogonalization (Neyman, 1979).

Let $\mathcal{F}_t$ denote the information available before choosing a_t, including the past observations $\{(x_s, \mathcal{A}_s, a_s, r_s)\}_{s<t}$ and the current context–action set pair $(x_t, \mathcal{A}_t)$. Since the confounding function ϕ is time-invariant and the policy randomization is known to the learner, the policy-induced conditional mean reward is identifiable. Specifically, under the learner's randomized policy, it can be written as $\mu_t(x_t) := \mathbb{E}_{a_t\sim\pi_t}[r_t(a_t) \mid \mathcal{F}_t] = x_t^\top \mathbf{\Theta}\, e_t + \phi(x_t)$, where $e_t := \mathbb{E}_{a_t\sim\pi_t}[a_t \mid \mathcal{F}_t]$ is the expected action induced by the algorithm's randomized policy at time t. To eliminate the bias from the confounder and correctly estimate $\mathbf{\Theta}$, we follow Neyman orthogonalization and regress the reward residuals $r_t(a_t) - \mu_t(x_t)$ on the centered action $a_t - e_t$. Indeed, the orthogonalized reward is no longer biased: $r_t(a_t) - \mu_t(x_t) = x_t^\top \mathbf{\Theta}(a_t - e_t) + \varepsilon_t = \langle \mathrm{vec}(\mathbf{\Theta}), z_t\rangle + \varepsilon_t = \theta^\top z_t + \varepsilon_t$, where we write $z_t := (a_t - e_t) \otimes x_t \in \mathbb{R}^d$ and $\theta := \mathrm{vec}(\mathbf{\Theta}) \in \mathbb{R}^d$.[1]

In DML literature (Chernozhukov et al., 2018; Robins et al., 1994), estimators for $\theta \in \mathbb{R}^d$ were originally developed for observational settings in which the data-collection policy π is unknown. The standard framework involves estimating the nuisance functions e and μ, followed by a residual-on-residual regression procedure (Robinson, 1988). Leveraging orthogonalization (or debiasing) techniques (Chernozhukov et al., 2018), such estimators achieve parametric convergence rates despite potentially slow convergence of the nuisance estimators.

In the online setting, however, the implemented randomized policy π_t is logged at each time step, so e_t's are directly observed. Treating the accumulated data as an offline dataset, DML therefore imposes *no* estimation rate requirements on the estimation of μ.

At each time step $t \in [T]$, we first estimate $\hat{\mu}_t$ with historical samples $\{(x_s, r_s)\}_{s<t}$ and then compute e_t. The orthogonalized estimator $\hat{\theta}$ for θ is a regularized Ridge regression version of Robinson's classical semiparametric regression estimator (Robinson, 1988):

$$\hat{\theta}_t = \arg\min_{\theta} \sum_{s<t} \big\{ \big(r_s - \hat{\mu}_s(x_s)\big) - \langle \theta, z_s\rangle \big\}^2 + \lambda \|\theta\|_2^2. \tag{3}$$

Comparison to BOSE (Krishnamurthy et al., 2018). The direct ridge estimation of θ in (3) highlights two distinctions between our approach and BOSE (Krishnamurthy et al., 2018). Beyond the fundamental difference between adversarial and stochastic confounders, the two methods address

[1] This is equivalent to $\mathrm{unvec}(\theta) = \mathbf{\Theta}$.

different objectives. In Krishnamurthy et al. (2018), the presence of adversarial confounding precludes direct estimation of a stable reward model, as considered in our setting. To eliminate bias induced by the confounder, BOSE employs an orthogonalization-type strategy by centering the actions. While this ensures robustness to adversarial confounding, which Krishnamurthy et al. (2018) aims to resolve, it leads to increased estimator variance that scales with the magnitude of the confounder. In contrast, our approach is designed to address settings with strong but stochastic confounding signals, with the primary goal of accurately identifying the treatment effect parameter $\boldsymbol{\Theta}$. To achieve this, we employ a residual-on-residual regression procedure, which effectively removes the confounding component from the observed rewards.

Algorithm 1 Inference Thompson Sampling (INF-TS)

Input: $T,\ T_0,\ \lambda,\ \nu$, regression algorithm $\mathcal{R}$ for μ
for $t = 1, \ldots, T_0$ **do** ▷ Randomized exploration period
  Observe $x_t, \mathcal{A}_t$, sample $a_t \sim \mathrm{Unif}(\mathcal{A}_t)$, observe r_t
  $e_t \leftarrow \frac{1}{N}\sum_{a_t^j \in \mathcal{A}_t} a_t^j,\ z_t \leftarrow (a_t - e_t) \otimes x_t$
end for
for $t = 1, \ldots, T_0$ **do**
  Fit $\hat{\mu}_t \sim \mathcal{R}(\{(x_s, r_s) : s < t\})$
  $\tilde{r}_t \leftarrow r_t - \hat{\mu}_t(x_t)$
end for
$\boldsymbol{V}_{T_0+1} \leftarrow \lambda \mathbf{I} + \sum_{t \le T_0} z_t z_t^\top,\ b_{T_0+1} \leftarrow \sum_{t \le T_0} z_t \tilde{r}_t$
for $t = T_0 + 1, \ldots, T$ **do** ▷ TS period
  $\hat{\theta}_t \leftarrow \boldsymbol{V}_t^{-1} b_t$, sample $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \nu^2 \boldsymbol{V}_t^{-1})$, $\tilde{\boldsymbol{\Theta}}_t \leftarrow \mathrm{unvec}(\tilde{\theta}_t)$
  Observe $x_t, \mathcal{A}_t$, choose $a_t \leftarrow \arg\max_{a_t^j \in \mathcal{A}_t} x_t^\top \tilde{\boldsymbol{\Theta}}_t a_t^j$, observe r_t
  Compute e_t as in (4), $z_t \leftarrow (a_t - e_t) \otimes x_t$
  Update $\hat{\mu}_t \sim \mathcal{R}(\{(x_s, r_s) : s < t\})$, $\tilde{r}_t \leftarrow r_t - \hat{\mu}_t(x_t)$ ▷ LOO regression
  $\boldsymbol{V}_{t+1} \leftarrow \boldsymbol{V}_t + z_t z_t^\top,\ b_{t+1} \leftarrow b_t + z_t \tilde{r}_t$
end for
Output: $\{a_t\}_{t=1}^T$

3.2 Inference Thompson sampling

An important aspect of computing (3) is the role of random π_t (Krishnamurthy et al., 2018). In particular, any deterministic action-selection strategy combined with the orthogonalized estimator – including confidence-based methods such as upper confidence bound (UCB) – can fail. Specifically, when $a_t = e_t$, the eigenvalues of $\boldsymbol{V}_t$ do not grow sufficiently, and the resulting estimator degenerates to zero.

We hence adopt Thompson sampling, a widely used approach that preserves randomized action selection at all time steps. At time period $t \in [T]$, the pseudo-posterior is updated as $\hat{\theta}_t = \boldsymbol{V}_t^{-1} b_t$, and the algorithm samples $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \nu^2 \boldsymbol{V}_t^{-1})$ and sets $\tilde{\boldsymbol{\Theta}}_t = \mathrm{unvec}(\tilde{\theta}_t)$. The algorithm then chooses the action $a_t = \arg\max_{a_t^j \in \mathcal{A}_t} x_t^\top \tilde{\boldsymbol{\Theta}}_t a_t^j$. The detailed algorithm INF-TS is given in Algorithm 1.

Propensity computation. A key component in estimating θ via (3) is the computation of e_t. Given information $\mathcal{F}_t$, define the score $S_t^j := (a_t^j \otimes x_t)^\top \tilde{\theta}_t,\ j \in [N]$, where $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_t, \nu^2 \boldsymbol{V}_t^{-1})$. According to INF-TS, the score $S_t^j \sim \mathcal{N}(s_t^j, \boldsymbol{D}_t^j)$ where

$$s_t^j = (a_t^j \otimes x_t)^\top \hat{\theta}_t, \quad \boldsymbol{D}_t^j = \nu^2 (a_t^j \otimes x_t)^\top \boldsymbol{V}_t^{-1} (a_t^j \otimes x_t), \quad j \in [N].$$

Then the propensity score

$$\mathbb{P}(a_t = a_t^j) = \mathbb{P}\big(a_t^j \in \arg\max_{i \in [N]} x_t^\top \tilde{\boldsymbol{\Theta}} a_t^i\big) = \mathbb{P}(S_j \ge S_i,\ \forall i \ne j, i \in [N]).$$

Equivalently, define matrix $\boldsymbol{B}_t^j \in \mathbb{R}^{(N-1)\times d}$ with rows $((a_t^j - a_t^i) \otimes x_t)^\top$ for $i \in [N]$ and $i \ne j$. Then we also have $\boldsymbol{B}_t^j \tilde{\theta}_t \sim \mathcal{N}(\boldsymbol{B}_t^j \hat{\theta}_t, \nu^2 \boldsymbol{B}_t^j \boldsymbol{V}_t^{-1} (\boldsymbol{B}_t^j)^\top)$, and

$$\mathbb{P}(a_t = a_t^j) = \mathbb{P}(\boldsymbol{B}_t^j \tilde{\theta}_t \ge 0) = \int_{y \ge 0} \phi_{N-1}(y; \boldsymbol{B}_t^j \hat{\theta}_t, \nu^2 \boldsymbol{B}_t^j \boldsymbol{V}_t^{-1} (\boldsymbol{B}_t^j)^\top)\, dy,$$

where $\phi_{N-1}(\cdot;\mu,\mathbf{\Sigma})$ is the $(N-1)$-dimensional Gaussian density with mean μ and covariance $\mathbf{\Sigma}$. Finally, the expected action of INF-TS at time t is

$$e_t = \sum_{a_t^j \in \mathcal{A}_t} \mathbb{P}(a_t = a_t^j) \cdot a_t^j. \tag{4}$$

In the following, we introduce some techniques adopted in INF-TS.

Randomization period. During the initial T_0 periods, INF-TS selects actions uniformly at random to facilitate a warm start for estimating $\mathbf{\Theta}$, following a standard exploration strategy in the bandit literature (Li et al., 2017; Kveton et al., 2020). To guarantee sufficient exploration, we impose the following assumption.

Assumption 3.1 (Context-action variation)**.** There exists a constant $\gamma > 0$ such that for $t \in [T_0]$,

$$\mathbf{\Gamma}_t := \mathbb{E}[((\bar{A}_t - e_t^{\text{unif}}) \otimes x_t)((\bar{A}_t - e_t^{\text{unif}}) \otimes x_t)^\top] \succeq \gamma \boldsymbol{I}, \tag{5}$$

where $\bar{A}_t \sim \text{Unif}(\mathcal{A}_t)$ and $e_t^{\text{unif}} := \mathbb{E}[\bar{A}_t \mid \mathcal{A}_t] = \frac{1}{N}\sum_{a_t^j \in \mathcal{A}_t} a_t^j$.

This assumption parallels the standard context-variation condition in the contextual bandit literature (Li et al., 2017; Zhou et al., 2020; Papini et al., 2021; Bastani et al., 2021) and the identifiability condition of Krishnamurthy et al. (2018). It requires every bilinear direction of $\mathbf{\Theta}$ to receive nonzero variation during the randomization period, for a *small* randomization period T_0, so that the initial design matrix is well conditioned.

We later show that, theoretically, without Assumption 3.1, a cold-start INF-TS with $T_0 = 0$, i.e., no exploration period, also achieves a sublinear regret rate; while empirically, the value of warm start is manifested in the smaller cumulative regrets.

Leave-One-Out (LOO) regression. In the TS phase for $t > T_0$, where $\hat{\mu}_t$ is estimated using data collected prior to time t, accurate estimation of $\hat{\mu}_t$ via leave-one-out (LOO) regression is crucial for cross-validation under limited samples. This approach effectively implements an online cross-fitting procedure and mitigates the overfitting bias and supports valid inference of $\mathbf{\Theta}$. We also note that, in response to the bounded mean reward assumption 2.4, the reward estimator $\{\hat{\mu}_t\}_{t\in[T]}$ can be clipped in practice.

3.3 Regret Upper Bound of INF-TS

In this section, we present the regret rate of INF-TS, which follows the classic TS analysis techniques presented in Agrawal & Goyal (2013).

The primary challenge lies in establishing finite-sample concentration inequalities for the orthogonalized estimator in (3). Although this estimator is structurally similar to that used in BOSE (Krishnamurthy et al., 2018), its error depends critically on the estimation accuracy of $\{\hat{\mu}_t\}_{t\in[T]}$. We address this issue by leveraging self-normalized concentration inequalities for centered vector-valued martingales, building on the techniques of (de la Pena et al., 2009; Krishnamurthy et al., 2018). Notably, our analysis does not impose explicit rate conditions on the estimation of μ, paralleling the requirements in offline DML under known propensity scores.

At a high level, we analyze the regret under the good events that $\hat{\theta}_t \to \text{vec}(\mathbf{\Theta})$, $\tilde{\theta}_t$ is close to $\hat{\theta}_t$, and that $\tilde{\theta}_t$ is optimistic for the optimal action a_t^*. We then bound the instant regret by the saturated arm argument on the good events, and finally conclude the following result.

Theorem 3.2. *Suppose Assumption 2.1, 2.4 and 3.1 hold. Set*

$$\nu = c_1 := \sqrt{\lambda} + \sigma\sqrt{d\log(\lambda + T/d) + 2\log(4/\delta)} + 2M\sqrt{2\log(4/\delta) + d/2\log(1 + 5T/(d))}$$
$$\lambda \geq 36d\log(9T) + 32\log(8T), \quad T_0 = O(d\log T).$$

Let $\{a_t\}_{t\in[T]}$ be the sequence of actions produced by INF-TS *with the parameters specified above, over a total of T horizon. Suppose* INF-TS *adopts the exact propensity score computation as in* (4)*. Then the expected regret*

$$\mathbb{E}\big[\textit{Regret}_T(\text{INF}-\text{TS})\big] \leq \frac{3(c_1 + \nu\sqrt{2\log(NT)})}{1/(4e\sqrt{\pi}) - 1/T}\sqrt{2dT\log(T/(d))} + (T_0 + 2)M = \tilde{O}(d\sqrt{T\log N}).$$

The denominator is positive for sufficiently large T. The detailed proof is deferred to Appendix C, which also inspires the following corollary of cold-start INF-TS ($T_0 = 0$) without Assumption 3.1. The proof of the corollary can be found in Appendix C.5.

Corollary 3.3. *Suppose Assumption 2.1 and 2.4 hold. Set* $\lambda \geq \max\{4, 4(9d \log 9T + 8 \log 8T)\}$ *and* ν *as in Theorem 3.2. If the cold-start* INF-TS *adopts* (4)*, then its expected regret is* $\tilde{O}(d\sqrt{T \log N})$.

3.4 A REGRET LOWER BOUND

It is well known that classic contextual linear bandits admit a lower bound of $\Omega(\sqrt{dT})$ (Chu et al., 2011; Li et al., 2023). These results do not apply directly, however, because of our bilinear reward structure (1). We nevertheless establish the following nontrivial lower bound; its proof is given in Appendix D. This bound shows that INF-TS is minimax optimal in the horizon T and the action set size N, but *not* in the context–action dimension $d = d_x d_a$.

Theorem 3.4 (Lower bound)**.** *Suppose* $T \geq \frac{\sigma^2 d^2}{8}$. *Let* $\mathcal{E}_0(d_x, d_a, N, \sigma, \phi)$ *denote the class of stochastic confounded linear contextual bandit environments that satisfy Assumption 2.1, 2.4 and 3.1. Then, for any policy* $\hat{\pi}$ *that, at each time step* $t \in [T]$*, output an action* $a_t \in \mathcal{A}_t$ *according to the history* $\mathcal{F}_t$*, there is* $\sup_{P \in \mathcal{E}_0(d_x, d_a, N, \sigma, \phi)} \mathbb{E}_P[\textit{Regret}_T(\hat{\pi})] \geq \frac{\sigma}{4\sqrt{2}} \sqrt{dT \log N}$.

4 EXPERIMENTS

We compare INF-TS against standard Thompson Sampling (TS) (Agrawal & Goyal, 2013), BOSE (Krishnamurthy et al., 2018), and a cold-start variant of INF-TS (with no randomization period, i.e., $T_0 = 0$). Our experiments evaluate the performance of these algorithms under varying types and strengths of confounding.

4.1 SYNTHETIC EXPERIMENT SETUP

We evaluate all algorithms in synthetic confounded contextual bandit environments. Unless otherwise stated, we use context dimension $d_x = 2$, action embedding dimension $d_a = 5$ (i.e., the problem dimension $d = 10$), and Gaussian noise with standard deviation $\sigma = 0.5$. Following the experiment setup in Krishnamurthy et al. (2018), user contexts and action embeddings are sampled independently and uniformly from the unit spheres in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_a}$, respectively. We consider various types of confounder terms, summarized in Table 1, in which $w_\phi \in \mathbb{R}^p$ is fixed at initialization, and $\alpha \geq 0$ controls the strength of the stochastic confounder. For each algorithm, we tune its exploration parameter by a grid search over 20 logarithmically spaced values in $[10^{-3}, 10^1]$. Results are reported with average over 10 random seeds, and shaded regions and error bars denote 95% confidence intervals.[2]

Table 1: Confounders used in the synthetic experiments.

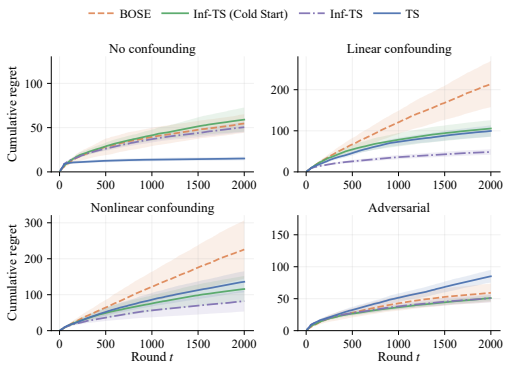
Confounder type	Unconfounded	Linear	Nonlinear	Adversarial Krishnamurthy et al. (2018)
$\phi(x)$	0	$\alpha x^\top w_\phi$	$\alpha\{\sin(x^\top w_\phi) + x_1^2/2\}$	$-\max_{a_t^j \in \mathcal{A}_t} x_t^\top \Theta a_t^j$

4.2 SYNTHETIC EXPERIMENT RESULTS AND DISCUSSIONS

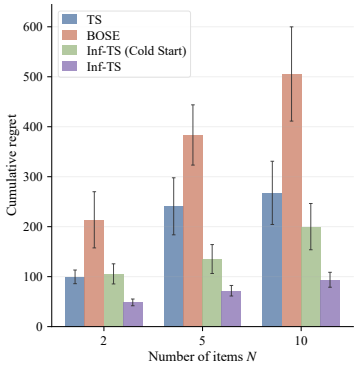
Strong confounding and variance reduction. Figure 1(a) compares the algorithms under the four representative confounders in Table 1. In the unconfounded setting, TS performs competitively, as the observed reward is well aligned with the treatment effect signal. However, once confounding is introduced, TS is unable to disentangle the action-independent confounder ϕ from the treatment effect. Consequently, it fails to reliably identify the treatment effect, resulting in degraded performance.

BOSE is designed to hedge against adversarial confounding and performs well in such settings. However, because BOSE does not residualize the observed rewards, the realized confounder signal

[2]Experiments are run on a 12th Gen Intel Core i7-12700H, 14 cores / 20 logical processors CPU with a 15.73 GB RAM.



(a) Regret curves under four representative environments with $N = 2$.



(b) Cumulative regret at time T versus the number of actions N.

Figure 1: **Strong confounding and action-set size.**

remains part of the effective regression noise. Consequently, when the confounder is strong (e.g., $\alpha = 4$ in both linear and nonlinear settings), bias reduction via action centering leads to increased variance, resulting in less efficient treatment effect estimation and unstable performance.

In contrast, INF-TS consistently outperforms all benchmarks under stochastic confounding. Notably, it also achieves the lowest regret in adversarial environments—despite BOSE being specifically designed for such settings—suggesting that residual-on-residual sampling remains effective even under adversarial confounding.

The performance gap between INF-TS and its cold-start variant highlights the importance of the initial randomization period. This reduces early-stage estimation error and provides a reliable warm start, which is critical for strong long-term performance.

Figure 1(b) evaluates scalability with respect to the action set size N. As N increases, accurate estimation of the treatment effect becomes increasingly critical for optimal decision-making. Among all methods considered, INF-TS again exhibits the slowest growth in regret, indicating strong robustness to the size of the action set $\mathcal{A}$.

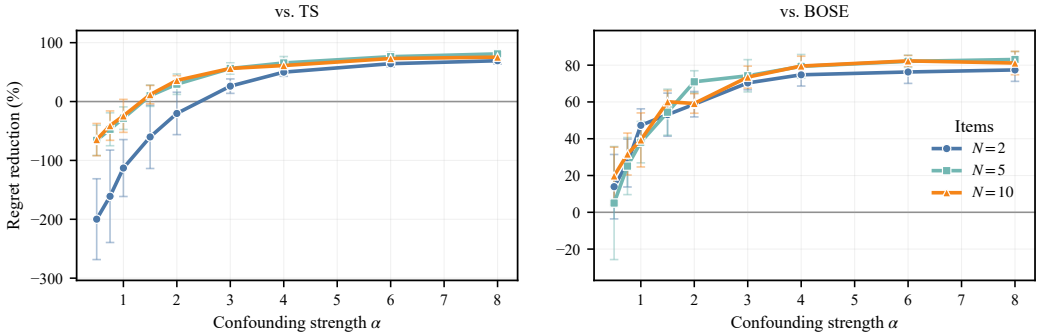


Figure 2: **Regret reduction versus confounding strength.**

Sensitivity to confounding strength. Figure 2 illustrates that the advantage of INF-TS becomes increasingly pronounced as the confounding strength α grows. When α is small, the problem closely resembles a standard linear contextual bandit, and TS remains competitive. As α increases, however, the reward signal becomes increasingly dominated by the confounder, rendering pseudo-posterior updates based on raw rewards in TS less reliable. A similar trend is observed in the comparison between TS and BOSE, where the relative advantage of TS over BOSE grows with α. In regimes with strong confounding, INF-TS consistently achieves the best performance among all algorithms.

Robustness to Monte Carlo propensity score approximation. In INF-TS, the propensity score can be computed exactly as in (4). While this is feasible for small action spaces, it becomes computationally prohibitive as the number of actions increases, in which case Monte Carlo (MC) approximation provides a practical alternative. To evaluate the impact of MC-based propensity estimation, we compare INF-TS with exact and MC propensity computation in a small-action setting $N = 2$, where exact computation remains tractable. Figure 3 presents the corresponding regret curves. The results indicate that both approaches achieve comparable performance, although the MC-based variant exhibits higher variance, reflecting a general limitation of MC methods.

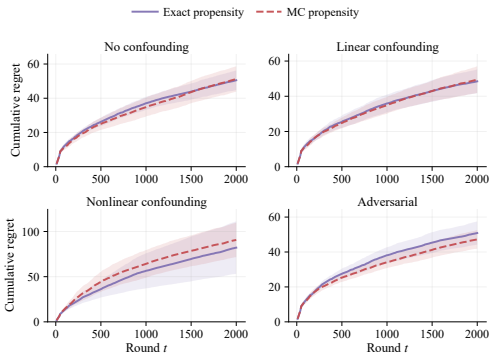
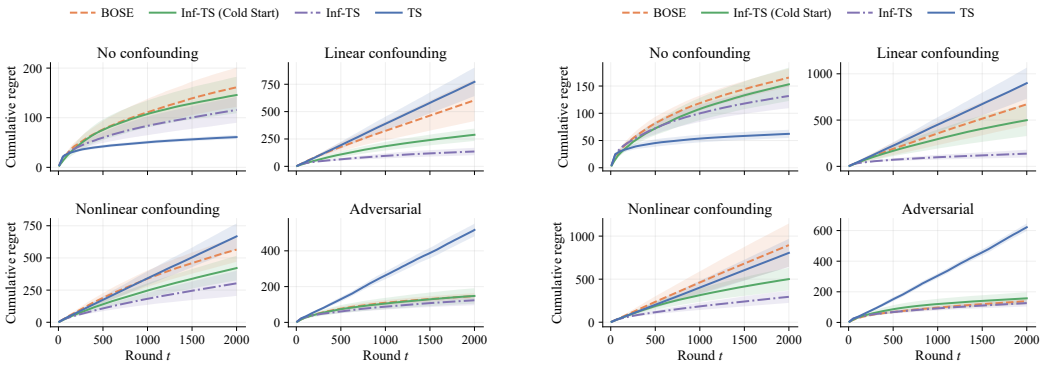


Figure 3: **Propensity-score approximation.**

4.3 Semi-synthetic case study of movie recommendation

We construct a semi-synthetic case study based on the OpenML Movies dataset, to demonstrate the advantage of INF-TS in real-wrold online recommendation systems (cf. Example 2.2).

The dataset contains 4,803 movies with rich metadata, including genres, keywords, cast, director, release date, and revenue. We derive item embeddings from these textual fields by concatenating them into one document per movie and encoding the text with term frequency–inverse document frequency (TF–IDF) weighting. To reduce the high dimensionality of the resulting representation, we apply truncated singular value decomposition (SVD) to obtain a d-dimensional dense embedding, which we normalize to unit Euclidean norm. All other components of the environment follow the synthetic setup. Because exhaustive hyperparameter tuning for BOSE is computationally prohibitive at this scale, we adopt a common exploration parameter of 0.20 across all algorithms rather than tuning each method individually. We select this value from preliminary small-scale experiments, in which it yields strong performance for BOSE.[3]



(a) $N = 100$ candidate items.

(b) $N = 500$ candidate items.

Figure 4: **Large-N semi-synthetic case study of movie recommendation.**

Figure 4 shows that the semi-synthetic results largely mirror the trends observed in the synthetic experiments. In the unconfounded setting, TS remains competitive, whereas INF-TS consistently achieves the lowest regret across all confounded settings. These findings indicate that INF-TS remains effective under realistic embedding structures and large action spaces, providing strong empirical support for its practical deployment.

[3]The fixed exploration parameter is therefore not chosen to favor INF-TS; rather, it gives BOSE its strongest preliminary setting while keeping the large-N comparison computationally feasible and consistent across methods.

AI USE STATEMENT

In this work, we only used generative AI tools for polish writing, double checking technical statements and drafting experiment code. All experiment codes are double-checked by authors and experiments are implemented by authors.

REPRODUCIBILITY STATEMENT

All of our experiment results can be reproduced by executing the code package in the supplementary materials. The OpenML movie dataset can be downloaded via this link.

REFERENCES

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

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International conference on machine learning*, pp. 1638–1646. PMLR, 2014.

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

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

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in neural information processing systems*, 28, 2015.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

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*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications*, 119(12):4210–4227, 2009.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, and Predag Klasnja. Action centered contextual bandits. *Advances in neural information processing systems*, 30, 2017.

Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. In *International Conference on Machine Learning*, pp. 2776–2785. PMLR, 2018.

Branislav Kveton, Manzil Zaheer, Csaba Szepesvari, Lihong Li, Mohammad Ghavamzadeh, and Craig Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 2066–2076. PMLR, 2020.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 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*, pp. 661–670, 2010.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. *IEEE Transactions on Information Theory*, 70(1):372–388, 2023.

Jerzy Neyman. C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 1–21, 1979.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *International Conference on Machine Learning*, pp. 8371–8380. PMLR, 2021.

James M Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology: methodological issues*, pp. 297–331. Springer, 1992.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: journal of the Econometric Society*, pp. 931–954, 1988.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 378–387, 2020.

Eric M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

Amit Sharma, Jake M Hofman, and Duncan J Watts. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 453–470, 2015.

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

Jingyuan Wang, Perry Dong, Ying Jin, Ruohan Zhan, and Zhengyuan Zhou. Adaptively learning to select-rank in online platforms. *arXiv preprint arXiv:2406.05017*, 2024.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 426–431, 2020.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International conference on machine learning*, pp. 11492–11502. PMLR, 2020.

A NOTATION

We use calligraphic letters for sets and $[n]$ to denote the discrete set $\{1, 2, \cdots, n\}$ for any $n \in \mathbb{Z}$. We use argmin and argmax to denote the minimizers and maximizers. For generic vectors $u, v \in \mathbb{R}^d$, their inner product is denoted by $\langle u, v\rangle = u^\top v$, and their tensor / Kronecker product (containing all pairwise products of components of v and u) is denoted by $u \otimes v \in \mathbb{R}^{d^2}$. When $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive semi-definite matrix, the semi-norm of a vector x is written as $\|x\|_{\mathbf{\Sigma}} = \sqrt{x^\top \mathbf{\Sigma} x}$. The vectorization of matrix $\mathbf{\Sigma}$ (with stacking columns) is denoted as $\mathrm{vec}(\mathbf{\Sigma}) \in \mathbb{R}^{d^2}$ and $\mathrm{unvec}(\mathrm{vec}(\mathbf{\Sigma})) = \mathbf{\Sigma}$. We denote the usual p-norm as $\|\cdot\|_p$, and the operator norm of $\mathbf{\Sigma}$ as $\|\mathbf{\Sigma}\|_{op}$.

B CONCLUSION

This work studies confounded contextual bandits, where the reward is modeled as a linear function of the action, additively confounded by a stochastic term that is independent of the action. In contrast to prior work, which focuses primarily on adversarial confounders (Greenewald et al., 2017; Krishnamurthy et al., 2018), we address online causal inference in the presence of strong stochastic confounders that may dominate the treatment effect. Our algorithm, INF-TS, achieves $\tilde{O}(d\sqrt{T \log N})$ regret, matching classical TS guarantees in unconfounded (generalized) linear contextual bandits while significantly outperforming existing methods in our empirical evaluation.

Our framework is currently limited to linear reward models. An important direction for future work is to extend it to nonlinear settings, which would allow for more flexible interactions among the confounder, treatment effect, and reward. A generalized linear model (GLM) extension (Filippi et al., 2010) is plausible but not immediate: under a GLM reward with a nonlinear link, the centered policy-induced conditional mean is generally not a GLM in the centered context-action feature, so the estimator and regret analysis developed here do not directly yield a guarantee in that setting.

A further direction is to close the gap between the INF-TS upper bound of $\tilde{O}(d\sqrt{T \log N})$ and our lower bound of $\Omega(\sqrt{dT \log N})$; the minimax optimality of the dimension dependence d remains open in confounded linear contextual bandit problem.

C DEFERRED PROOFS OF INF-TS REGRET UPPER BOUND

The proof of the regret upper bound of INF-TS follows the classical idea in Agrawal & Goyal (2013); Krishnamurthy et al. (2018) and as described in the proof sketch in Section 4.2. For simplicity, let

$$z_t^j := (a_t^j - e_t) \otimes x_t, \quad z_t^* := (a_t^* - e_t) \otimes x_t,$$

we define the corresponding good events:

$$\begin{aligned}
\mathcal{E}_t^{(1)} :=& \{\forall a_t^j \in \mathcal{A}_t : |(z_t^j)^\top (\hat{\theta}_t - \theta)| \le c_1 \|z_t^j\|_{\boldsymbol{V}_t^{-1}}\} \\
\mathcal{E}_t^{(2)} :=& \{\forall a_t^j \in \mathcal{A}_t : |(z_t^j)^\top (\tilde{\theta}_t - \hat{\theta}_t)| \le c_2 \|z_t^j\|_{\boldsymbol{V}_t^{-1}}\}, \\
\mathcal{E}_t^{(3)} :=& \{(z_t^*)^\top \tilde{\theta}_t > (z_t^*)^\top \theta\},
\end{aligned}$$

where constants $c_1, c_2 > 0$ are tuned later.

Throughout the proof, the minimum and maximum eigenvalue of a matrix $\mathbf{\Sigma}$ is written as $\lambda_{\min}(\mathbf{\Sigma}), \lambda_{\max}(\mathbf{\Sigma})$ respectively. We denote P to be any probability measure defined on the probability space $(\Omega, \sigma(\Omega), P)$. For any filtration $\{\mathcal{F}_t\}_t$, we write $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. For any event $\mathcal{E}$, we write $\bar{\mathcal{E}}$ as the complement of $\mathcal{E}$.

C.1 BOUNDING THE PROBABILITY OF $\mathcal{E}_t^{(1)}$

Lemma C.1. *Suppose Assumption 2.1, 2.4 and 3.1 hold. Let* $c_1 := \sqrt{\lambda} + \sigma\sqrt{d\log(\lambda + T/d) + 2\log(4/\delta)} + 2M\sqrt{2\log(4/\delta) + d\log(1 + 2T/(d))}$ *with* $\lambda \ge 36d\log(9T) + 32\log(8/\delta)$. *We choose* T_0 *such that* $T_0 = O(d\log(T))$. *Then for all time period* $t > T_0$,

$$p_1 := \mathbb{P}(\bar{\mathcal{E}}_t^{(1)},\ \|\hat{\theta}_t - \theta\|_2 \le 1) \le \delta.$$

Proof. In our setting, we define the canonical σ-algebra

$$\mathcal{F}_t := \sigma(x_1, \mathcal{A}_1, a_1, r_1, \cdots, x_{t-1}, \mathcal{A}_{t-1}, a_{t-1}, r_{t-1}, x_t, \mathcal{A}_t)$$

corresponding to the t-th time period. First fix any time $t \geq T_0$,

$$|(z_t^j)^\top(\hat{\theta}_t - \theta)| \leq \|\hat{\theta}_t - \theta\|_{\boldsymbol{V}_t} \|z_t^j\|_{\boldsymbol{V}_t^{-1}}$$

We proceed to bound the RHS with high probability.

Let $\xi_s := \mu_s(x_s) - \hat{\mu}_s(x_s)$ for $1 \leq s \leq t$. Then for any $1 \leq s \leq t$,

$$\begin{aligned}\tilde{r}_s &= r_s - \hat{\mu}_s(x_s) = \langle (a_s - e_s) \otimes x_s, \theta \rangle + \mu_s(x_s) + \varepsilon_s - \hat{\mu}_s(x_s) \\ &= z_s^\top \theta + \xi_s + \varepsilon_s.\end{aligned}$$

Hence, we can compute

$$\begin{aligned}\boldsymbol{V}_t(\hat{\theta}_t - \theta) =& \boldsymbol{V}_t\left(\boldsymbol{V}_t^{-1}\sum_{s\leq t}(z_s)(r_s^t - \hat{\mu}_t(x_s)) - \theta\right) \\ =& \sum_{s\leq t} z_s\big(z_s^\top\theta + \xi_s + \varepsilon_s\big) - \boldsymbol{V}_t\theta \\ =& \sum_{s\leq t} z_s z_s^\top\theta + \sum_{s\leq t} z_s\xi_s + \sum_{s\leq t} z_s\varepsilon_s - \boldsymbol{V}_t\theta \\ =& -\lambda\boldsymbol{I}\theta + \sum_{s\leq t} z_s\xi_s + \sum_{s\leq t} z_s\varepsilon_s,\end{aligned}$$

where the last inequality is by recalling the definition of $\boldsymbol{V}_t$. We can bound the estimator's difference in $\boldsymbol{V}_t$-induced norm:

$$\|\hat{\theta}_t - \theta\|_{\boldsymbol{V}_t} \leq \|\lambda\theta\|_{\boldsymbol{V}_t^{-1}} + \|W_t^{(1)}\|_{\boldsymbol{V}_t^{-1}} + \|W_t^{(2)}\|_{\boldsymbol{V}_t^{-1}},$$

where $W_t^{(1)} = \sum_{s\leq t} z_s\varepsilon_s$ and $W_t^{(2)} = \sum_{s\leq t} z_s\xi_s$. Since $\boldsymbol{V}_t \succeq \lambda\boldsymbol{I}$, we have that $\|\lambda\theta\|_{\boldsymbol{V}_t^{-1}} \leq \sqrt{\lambda}$. According to Lemma F.1 and F.2, there is

$$\begin{aligned}\|W_t^{(1)}\|_{\boldsymbol{V}_t^{-1}} \leq& \sigma\sqrt{2\log\left(\frac{2(\lambda + t/d)^{d/2}\lambda^{-d/2}}{\delta}\right)} \\ =& \sigma\sqrt{d\log(\lambda + t/d) + 2\log(4/\delta)},\end{aligned}$$

with probability $1 - \delta/4$.

We now turn our attention to $W_t^{(2)}$:

$$W_t^{(2)} = \sum_{s\leq t} z_s\xi_s = \sum_{s\leq t}\big((a_s - e_s) \otimes x_s\big)\xi_s$$

We note that $\{W_t^{(2)}\}_t$ is a martingale since $\mathbb{E}_s[(a_s - e_s) \otimes x_s] = 0$. scale up It remains to bound $\|W_t^{(2)}\|_{\boldsymbol{V}_t^{-1}}$. We invoke the following lemma.

Lemma C.2. *Let $\{\mathcal{F}_s\}_{s\leq t}$ be a filtration and let $\{(Z_s, \xi_s)\}_{s\leq t}$ be a stochastic process with $Z_t \in \mathbb{R}^d$ and $\xi_t \in \mathbb{R}$ such that: (i) ξ_t is $\mathcal{F}_t$ measurable, (ii) given $\mathcal{F}_t$, ξ_t is deterministic and $|\xi_s| \leq M$ a.s. for all $s \in [t]$, and (iii) $\mathbb{E}[Z_t \mid \mathcal{F}_t] = 0$. Define $\widehat{\boldsymbol{\Sigma}}_t := \sum_{s\leq t}(\xi_s)^2 Z_s Z_s^\top$, and $\overline{\boldsymbol{\Sigma}}_t := \sum_{s\leq t}\mathbb{E}[(\xi_s)^2 Z_s Z_s^\top \mid \mathcal{F}_s]$. Let $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ be a positive definite matrix, and let $\boldsymbol{A}_t := \boldsymbol{Q} + \widehat{\boldsymbol{\Sigma}}_t + \overline{\boldsymbol{\Sigma}}_t$. Then,*

$$\mathbb{P}_t\left(\left\|\sum_{s=1}^t Z_s\xi_s\right\|^2_{\boldsymbol{A}_t^{-1}} > 2\log\left(\frac{1}{\delta}\sqrt{\frac{\det(\boldsymbol{A}_t)}{\det(\boldsymbol{Q})}}\right)\right) \leq \delta.$$

Define $\widehat{\boldsymbol{\Sigma}}_t := \sum_{s\le t}(\xi_s)^2 z_s z_s^\top$ and $\overline{\boldsymbol{\Sigma}}_t := \sum_{s\le t}\mathbb{E}_s[(\xi_s z_s)(\xi_s z_s)^\top]$. Let $\boldsymbol{Q}\in\mathbb{R}^{d\times d}$ be a positive definite matrix. Noting that at any time period t, the estimation error ξ_t is deterministic given $\mathcal{F}_t$ and $|\xi_t|\le 2M$ a.s., we can apply Lemma C.2: with probability at least $1-\delta/4$,

$$\|W_t^{(2)}\|^2_{(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)^{-1}} = \left\|\sum_{s=1}^t Z_s\xi_s\right\|^2_{(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)^{-1}} \le 2\log\left(\frac{4}{\delta}\sqrt{\frac{\det(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)}{\det(\boldsymbol{Q})}}\right).$$

Define $\widehat{\boldsymbol{V}}_t := \sum_{s\le t} Z_s Z_s^\top$ and $\overline{\boldsymbol{V}}_t := \sum_{s\le t}\mathbb{E}_s[Z_s Z_s^\top]$. Since $(\xi_s)^2\le 4M^2$ a.s., there is

$$\boldsymbol{Q}+4M^2(\widehat{\boldsymbol{V}}_t+\overline{\boldsymbol{V}}_t)\succeq \boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t$$

a.s. Hence, with probability at least $1-\delta/4$,

$$\begin{aligned}\|W_t^{(2)}\|^2_{(\boldsymbol{Q}+4M^2(\widehat{\boldsymbol{V}}_t+\overline{\boldsymbol{V}}_t))^{-1}} =& \left\|\sum_{s=1}^t Z_s\xi_s\right\|^2_{(\boldsymbol{Q}+4M^2(\widehat{\boldsymbol{V}}_t+\overline{\boldsymbol{V}}_t))^{-1}}\\ \le& \left\|\sum_{s=1}^t Z_s\xi_s\right\|^2_{(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)^{-1}}\\ \le& 2\log\left(\frac{4}{\delta}\sqrt{\frac{\det(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)}{\det(\boldsymbol{Q})}}\right). \end{aligned} \tag{6}$$

Apply Lemma F.4 to $\widehat{\boldsymbol{V}}_t, \overline{\boldsymbol{V}}_t$, then with probability at least $1-\delta/4$,

$$\begin{aligned} v^\top\overline{\boldsymbol{V}}_t v \le& 2v^\top\widehat{\boldsymbol{V}}_t v + 36d\log(9t)+32\log(8/\delta)\\ v^\top(\overline{\boldsymbol{V}}_t+\widehat{\boldsymbol{V}}_t)v \le& v^\top\widehat{\boldsymbol{V}}_t v + 36d\log(9t)+32\log(8/\delta).\end{aligned}$$

The above leads to a positive semidefinite inequality

$$\lambda\boldsymbol{I}+\widehat{\boldsymbol{V}}_t \succeq (\lambda-36d\log(9t)-32\log(8/\delta))\boldsymbol{I}+(\overline{\boldsymbol{V}}_t+\widehat{\boldsymbol{V}}_t),$$

for $\lambda > 36d\log(9T)+32\log(8/\delta)$. Hence, for any vector v,

$$\|v\|^2_{\boldsymbol{V}_t^{-1}} \le \|v\|^2_{\left((\lambda-36d\log(9t)-32\log(8/\delta))\boldsymbol{I}+(\overline{\boldsymbol{V}}_t+\widehat{\boldsymbol{V}}_t)\right)^{-1}}.$$

Recall the inequality (6) with $\boldsymbol{Q} := 4M^2(\lambda-36d\log(9t)-32\log(8/\delta))\boldsymbol{I} =: \bar{\lambda}\boldsymbol{I}$, which satisfies the condition that $\mathcal{Q}\succ 0$ as $\lambda > 36d\log(9T)+32\log(8/\delta)$,

$$\begin{aligned}\|W_t^{(2)}\|^2_{\boldsymbol{V}_t^{-1}} \le& 4M^2\|W_t^{(2)}\|^2_{(\boldsymbol{Q}+4M^2(\widehat{\boldsymbol{V}}_t+\overline{\boldsymbol{V}}_t))^{-1}}\\ \le& 8M^2\log\left(\frac{4}{\delta}\sqrt{\frac{\det(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)}{\det(\boldsymbol{Q})}}\right).\end{aligned} \tag{7}$$

with probability at least $1-\delta/4$. Taking a union bound over the two results in Lemma C.2 and F.4 for any time period $t\in[T]$, the above inequality holds with probability at least $1-3\delta/4$.

It now remains to upper bound the determinant term. Following the standard argument in Abbasi-Yadkori et al. (2011),

$$\begin{aligned}\left(\frac{\det(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)}{\det(\boldsymbol{Q})}\right)^{1/(d)} \le& \frac{\mathrm{tr}(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)}{d\bar{\lambda}}\\ =& \frac{\mathrm{tr}(\bar{\lambda}\boldsymbol{I})}{d\bar{\lambda}}+\frac{1}{d\bar{\lambda}}\sum_{s\le t}\left(\mathrm{tr}((\xi_s)^2 Z_s Z_s^\top)+\mathrm{tr}(\mathbb{E}[(\xi_s)^2 Z_s Z_s\mid\mathcal{F}_s])\right)\\ \le& 1+\frac{1}{d\bar{\lambda}}\sum_{s\le t}(\|\xi_s Z_s\|_2^2+4M^2)\\ \le& 1+\frac{20tM^2}{d\bar{\lambda}}\\ =& 1+\frac{20tM^2}{d4M^2(\lambda-36d\log(9t)-32\log(8/\delta))}\\ =& 1+\frac{5t}{d(\lambda-36d\log(9t)-32\log(8/\delta))}.\end{aligned}$$

By the condition of λ, we have that

$$\left(\frac{\det(\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t)}{\det(\boldsymbol{Q})}\right)^{1/(d)} \leq 1+\frac{5t}{d}.$$

Plug the above result into the previous inequality (7),

$$\|W_t^{(2)}\|_{\boldsymbol{V}_t^{-1}} \leq 2M\sqrt{2\log(4/\delta)+d/2\log(1+5T/(d))},$$

with probability at least $1-\delta/4$.

Putting everything together, with probability at least $1-\delta$, for any $t\in[T]$,

$$\begin{aligned}\|\hat{\theta}_t-\theta\|_{\boldsymbol{V}_t} \leq \sqrt{\lambda}+\sigma\sqrt{d\log(\lambda+t/d)+2\log(4/\delta)} \\ +2M\sqrt{2\log(4/\delta)+d/2\log(1+5T/(d))},\end{aligned}$$

if we choose a $\lambda \geq 36d\log(9T)+32\log(8/\delta)$.

Now let $c_1=\sqrt{\lambda}+\sigma\sqrt{d\log(\lambda+T/d)+2\log(4/\delta)}+2M\sqrt{2\log(4/\delta)+d/2\log(1+5T/(d))}$. For any $t\geq 1$, by definition, $\boldsymbol{V}_t$ is a symmetric and positive definite matrix. Hence, by the Rayleigh Quotient property,

$$\|\hat{\theta}_t-\theta\|_2 \leq \frac{\|\hat{\theta}_t-\theta\|_{\boldsymbol{V}_t}}{\sqrt{\lambda_{\min}(\boldsymbol{V}_t)}}.$$

Since $\lambda_{\min}(\boldsymbol{V}_t)\geq\lambda_{\min}(\boldsymbol{V}_{T_0})$ for all $t\geq T_0$, we can choose T_0 such that

$$\lambda_{\min}(\boldsymbol{V}_{T_0}) \geq \max\{c_1^2,1\} \tag{8}$$

so that with high probability, $\mathcal{E}_t^{(1)}$ takes place and $\|\hat{\theta}_t-\theta\|_2\leq 1$ for all $t\geq T_0$.

It remains to find T_0 so that the condition (8) holds.

Lemma C.3. *Let $B>0$ be any positive constant. Suppose*

$$T_0 \geq 8\log(d/\delta)/\gamma^2+(2B+\frac{16}{3}\log(d/\delta)))/\gamma,$$

then under Assumption 3.1, with probability at least $1-\delta$, there is $\lambda_{\min}(\boldsymbol{V}_{T_0})\geq\max\{\lambda+B,1\}$.

We conclude by applying Lemma C.3 with $B=\sigma^2 d\log(\lambda+T/d)+2\log(4/\delta)+8M^2\log(4/\delta)+d/2\log(1+5T/(d))$. □

C.2 Bounding the probabilities of $\mathcal{E}_t^{(2)}$ and $\mathcal{E}_t^{(3)}$

Lemma C.4. *Let $c_2=\nu\sqrt{2\log(N/\delta)}$. Then for any time period $t>T_0$, on the event $\{\|\hat{\theta}_t-\theta\|_2\leq 1\}$,*

$$p_2 := \mathbb{P}_t(\bar{\mathcal{E}}_t^{(2)}) \leq \delta.$$

Proof. We fix some $t\in[T]$. Then given the filtration $\mathcal{F}_t$, by the design of INF-TS,

$$(z_t^j)^\top(\tilde{\theta}_t-\hat{\theta}_t)\sim\mathcal{N}(0,\nu^2\|z_t^j\|^2_{\boldsymbol{V}_t^{-1}}).$$

Hence, we have that

$$\mathbb{P}_t\big((z_t^j)^\top(\tilde{\theta}_t-\hat{\theta}_t)\geq c\|z_t^j\|_{\boldsymbol{V}_t^{-1}}\big)\leq\exp\left(-\frac{c^2}{2\nu^2}\right),$$

for any $c>0$. Then taking a union bound over all N arms, with probability at least $1-\delta$, the event $\mathcal{E}_t^{(2)}$ takes place with $c_2=\nu\sqrt{2\log(N/\delta)}$ for any $t\in[T]$.

□

Lemma C.5. *For any time period $t > T_0$, on the event of $\mathcal{E}_t^{(1)}$ and $\{\|\hat{\theta}_t - \theta\|_2 \leq 1\}$, setting $\nu = c_1$, we have that*

$$\mathbb{P}_t(\mathcal{E}_t^{(3)}) \geq \frac{1}{4e\sqrt{\pi}} =: p_3.$$

Proof. We fix some $t \in [T]$. Again, given the filtration $\mathcal{F}_t$, by the design of INF-TS,

$$(z_t^*)^\top(\tilde{\theta}_t - \hat{\theta}_t) \sim \mathcal{N}(0, \nu^2 \|z_t^*\|^2_{\boldsymbol{V}_t^{-1}}).$$

By the Gaussian anti-concentration inequality in Lemma F.6,

$$\begin{aligned}
&\mathbb{P}_t\big((z_t^*)^\top\tilde{\theta}_t > (z_t^*)^\top\theta\big)\\
=&\mathbb{P}_t\left(\frac{(z_t^*)^\top(\tilde{\theta}_t - \hat{\theta}_t)}{\nu\|z_t^*\|_{\boldsymbol{V}_t^{-1}}} > \frac{(z_t^*)^\top(\theta - \hat{\theta}_t)}{\nu\|z_t^*\|_{\boldsymbol{V}_t^{-1}}}\right)\\
=&\mathbb{P}_t(\mathcal{N}(0,1) \geq \zeta_t)\\
\geq&\frac{1}{4\sqrt{\pi}}e^{-\zeta_t^2},
\end{aligned}$$

where we write

$$\zeta_t := \frac{(z_t^*)^\top(\theta - \hat{\theta}_t)}{\nu\|z_t^*\|_{\boldsymbol{V}_t^{-1}}},$$

and

$$|\zeta_t| = \left|\frac{(z_t^*)^\top(\theta - \hat{\theta}_t)}{\nu\|z_t^*\|_{\boldsymbol{V}_t^{-1}}}\right| \leq \frac{c_1\|z_t^*\|_{\boldsymbol{V}_t^{-1}}}{\nu\|z_t^*\|_{\boldsymbol{V}_t^{-1}}} = 1,$$

on the event of $\mathcal{E}_t^{(1)}$ if we set $\nu = c_1$. □

C.3 The set of saturated arm

Let $\Delta_t^j := (a_t^* \otimes x_t)^\top\theta - (a_t^j \otimes x_t)^\top\theta$, and let $c_3 := c_1 + c_2$. We would like to note that by definition

$$\Delta_t^j = ((a_t^* - e_t) \otimes x_t)^\top\theta - ((a_t^j - e_t) \otimes x_t)^\top\theta$$

The set of saturated arms at time t is defined as $\mathcal{S}_t = \{a_t^j \in \mathcal{A}_t : \Delta_t^j > c_3\|(z_t^j)\|_{\boldsymbol{V}_t^{-1}}\}$.

Lemma C.6. *On the event $\mathcal{E}_t^{(1)}$ and $\{\|\hat{\theta}_t - \theta\|_2 \leq 1\}$, for any time period $t > T_0$,*

$$\mathbb{P}_t(a_t \notin \mathcal{S}_t) \geq p_3 - p_2.$$

Proof. Let that $\bar{S}_t^j := (z_t^j)^\top\tilde{\theta}_t$. At any t, INF-TS chooses the arm the highest score

$$a_t = \underset{a_t^j \in \mathcal{A}_t}{\arg\max}\,(a_t^j \otimes x_t)^\top\tilde{\theta}_t = \underset{a_t^j \in \mathcal{A}_t}{\arg\max}\,(z_t^j)^\top\tilde{\theta}_t = \underset{a_t^j \in \mathcal{A}_t}{\arg\max}\,\bar{S}_t^j.$$

If $\bar{S}_t^* := (z_t^*)^\top\tilde{\theta}_t > \bar{S}_t^j$ for all $a_t^j \in \mathcal{S}_t$, an unsaturated arms must be played. By the construction of $\mathcal{S}_t$, the unsaturated arms include the optimal action. Therefore,

$$\mathbb{P}_t(a_t \notin \mathcal{S}_t) \geq \mathbb{P}_t(\bar{S}_t^* > \bar{S}_t^j,\ \forall a_t^j \in \mathcal{S}_t).$$

On the event of $\mathcal{E}_t^{(1)}, \mathcal{E}_t^{(2)}$, for all $a_t^j \in \mathcal{A}_t$, we have that

$$\bar{S}_t^j \leq (z_t^j)^\top\theta + c_3\|z_t^j\|_{\boldsymbol{V}_t^{-1}}.$$

Hence, given any $\mathcal{F}_t$ that $\mathcal{E}_t^{(1)}$ is true, for any $a_t^j \in \mathcal{S}_t$,

$$\begin{aligned}
\bar{S}_t^j \leq&(z_t^j)^\top\theta + c_3\|z_t^j\|_{\boldsymbol{V}_t^{-1}}\\
\leq&(z_t^*)^\top\theta,
\end{aligned}$$

and we can compute

$$\begin{aligned}
\mathbb{P}_t\big(\bar{S}_t^* > \bar{S}_t^j,\ \forall a_t^j \in \mathcal{S}_t\big) \geq&\mathbb{P}_t\big(\bar{S}_t^* > (z_t^*)^\top\theta,\ \forall a_t^j \in \mathcal{S}_t\big) - \mathbb{P}_t(\bar{\mathcal{E}}_t^{(2)})\\
\geq&p_3 - p_2,
\end{aligned}$$

by Lemma C.4 and C.5. □

C.4 PUTTING EVERYTHING TOGETHER

Abusing the notation before, let $\Delta_t := (a_t^* \otimes x_t)^\top \theta - (a_t \otimes x_t)^\top \theta$. We also have $\Delta_t = (z_t^*)^\top \theta - z_t^\top \theta$.

Lemma C.7. *On the event $\mathcal{E}_t^{(1)}$ and $\{\|\hat{\theta}_t - \theta\|_2 \le 1\}$, for any time period $t > T_0$,*

$$\mathbb{E}_t[\Delta_t] \le \frac{3c_3}{p_3 - p_2} \mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big] + p_2 M.$$

Proof. Define $\bar{j}_t := \arg\min_{j \in [N], a_t^j \notin \mathcal{S}_t} \|z_t^j\|_{\boldsymbol{V}_t^{-1}}$ to be the unsaturated arm with the smallest $\|z_t^j\|_{\boldsymbol{V}_t^{-1}}$, which is $\mathcal{F}_t$-measurable. We also define j_t to be the arm chosen at time t.

Recall that on the event of $\mathcal{E}_t^{(1)}, \mathcal{E}_t^{(2)}$, for all $a_t^j \in \mathcal{A}_t$, we have that

$$\bar{S}_t^j \le (z_t^j)^\top \theta + c_3 \|z_t^j\|_{\boldsymbol{V}_t^{-1}}.$$

Hence, we can decompose

$$\begin{aligned}
\Delta_t =& (z_t^*)^\top \theta - \bar{z}_t^\top \theta + \bar{z}_t^\top \theta - z_t^\top \theta \\
=& \Delta_t^{\bar{j}_t} + \big(\bar{z}_t^\top \theta - z_t^\top \theta\big) \\
\le& \Delta_t^{\bar{j}_t} + \big(\bar{S}_t^{\bar{j}_t} - \bar{S}_t^{j_t}\big) + c_3\big(\|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + \|z_t\|_{\boldsymbol{V}_t^{-1}}\big) \\
\overset{(a)}{\le}& \Delta_t^{\bar{j}_t} + c_3\big(\|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + \|z_t\|_{\boldsymbol{V}_t^{-1}}\big) \\
\overset{(b)}{\le}& c_3 \|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + c_3\big(\|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + \|z_t\|_{\boldsymbol{V}_t^{-1}}\big),
\end{aligned}$$

where for simplicity, we write $\bar{z}_t := (a_t^{\bar{j}_t} - e_t) \otimes x_t$. Step (a) is due to the fact that $\bar{S}_t^{\bar{j}_t} < \bar{S}_t^{j_t}$ always and step (b) is because $a_t^{\bar{j}_t} \notin \mathcal{S}_t$.

By the above argument, given any $\mathcal{F}_t$, either $\Delta_t \le 2c_3 \|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + c_3 \|z_t\|_{\boldsymbol{V}_t^{-1}}$ on $\mathcal{E}_t^{(2)}$ or $\mathcal{E}_t^{(2)}$ is false. We can thus upper bound

$$\mathbb{E}_t[\Delta_t] \le \mathbb{E}\big[2c_3 \|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + c_3 \|z_t\|_{\boldsymbol{V}_t^{-1}}\big] + \mathbb{P}_t(\bar{\mathcal{E}}_t^{(2)}) M. \tag{9}$$

Now, with $\mathbb{P}_t(\bar{\mathcal{E}}_t^{(2)}) \le \delta$,

$$\begin{aligned}
\mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big] \ge& \mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}} \mid a_t \notin \mathcal{S}_t\big] \cdot \mathbb{P}_t(a_t \notin \mathcal{S}_t) \\
\ge& \|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} (p_3 - p_2).
\end{aligned}$$

Plug the above into the inequality (9),

$$\begin{aligned}
\mathbb{E}_t[\Delta_t] \le& \frac{2c_3}{p_3 - p_2} \mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big] + c_3 \mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big] + p_2 M \\
\le& \frac{3c_3}{p_3 - p_2} \mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big] + p_2 M.
\end{aligned}$$

□

We are now ready to prove the main result.

Proof of Theorem 3.2. We can decompose the regret as

$$\begin{aligned}\mathbb{E}\big[\text{Regret}_T(\{a_t\}_{t\in[T]})\big] &\le \sum_{t=T_0+1}^{T} \mathbb{E}[\Delta_t] + T_0 M \\ &\le \sum_{t=T_0+1}^{T} \mathbb{E}\big[\Delta_t \mathbb{1}\{\|\hat{\theta}_t - \theta\|_2 \le 1, \mathcal{E}_t^{(1)}\}\big] + (T_0 + p_1 T)M \\ &\le \sum_{t=T_0+1}^{T} \mathbb{E}\big[\mathbb{E}_t[\Delta_t] \mathbb{1}\{\|\hat{\theta}_t - \theta\|_2 \le 1, \mathcal{E}_t^{(1)}\}\big] + (T_0 + p_1 T)M \\ &\le \sum_{t=T_0+1}^{T} \left(\frac{3c_3}{p_3 - p_2} \mathbb{E}\big[\mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big]\big] + p_2 M\right) + (T_0 + p_1 T)M,\end{aligned}$$

where the last inequality is due to Lemma C.7. Now let $\delta = \frac{1}{T}$ and invoking Lemma F.7, we have that

$$\begin{aligned}\mathbb{E}\big[\text{Regret}_T(\{a_t\}_{t\in[T]})\big] \le& \frac{3c_3}{p_3 - p_2}\sqrt{2dT\log(T/(d))} + (T_0 + p_1 T + p_2 T)M \\ =& \frac{3c_3}{p_3 - p_2}\sqrt{2dT\log(T/(d))} + (T_0 + 2)M.\end{aligned}$$

□

C.5 Deferred Proof of Corollary 3.3

Proof of Corollary 3.3. Following the same logic as in Lemma C.1, with probability at least $1-\delta$, for any $t \in [T]$ and $\lambda \ge 36d\log(9T) + 32\log(8/\delta)$,

$$\begin{aligned}\|\hat{\theta}_t - \theta\|_{\boldsymbol{V}_t} \le& \sqrt{\lambda} + \sigma\sqrt{d\log(\lambda + t/d) + 2\log(4/\delta)} \\ &+ 2M\sqrt{2\log(4/\delta) + d/2\log(1 + 5T/(d))} = c_1.\end{aligned}$$

Consequently, we have that $\mathbb{P}(\bar{\mathcal{E}}_t^{(1)}) \le \delta$.

Again, following Lemma C.4, on the event of $\mathcal{E}_t^{(1)}$,let $c_2 = \nu\sqrt{2\log(2N/\delta)}$, $p_2 := \mathbb{P}_t(\bar{\mathcal{E}}_t^{(2)}) \le \delta$. By an analogous proof of Lemma C.5, on event $\mathcal{E}_t^{(1)}$,

$$\mathbb{P}_t(\mathcal{E}_t^{(3)}) = \mathbb{P}_t\big((z_t^*)^\top \tilde{\theta}_t > (z_t^*)^\top \theta)\big) \ge \mathbb{P}(\mathcal{N}(0,1) \ge 1) \ge \frac{1}{4e\sqrt{\pi}} =: p_3.$$

We now move on to the saturated arms argument. On event $\{\mathcal{E}_t^{(1)} \bigcap \mathcal{E}_t^{(2)}\}$, for every saturated action $j \in \mathcal{S}_t$, we similarly have

$$\bar{S}_t^j \le (z_t^j)^\top \theta + c_3 \|z_t^j\|_{\boldsymbol{V}_t^{-1}} \le (z_t^*)^\top \theta.$$

Using the arguments in Lemma C.6, $\mathbb{P}_t(a_t \notin \mathcal{S}_t) \ge p_3 - p_2$.

Recall that $\bar{j}_t = \arg\min_{j\in[N], a_t^j \notin \mathcal{S}_t} \|z_t^j\|_{\boldsymbol{V}_t^{-1}}$ to be the unsaturated arm with the smallest $\|z_t^j\|_{\boldsymbol{V}_t^{-1}}$. Since an unsaturated action is selected with probability at least $p_3 - p_2$,

$$\mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{(-1)}}\big] \ge (p_3 - p_2)\|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}},$$

where we have $\bar{z}_t = (a_t^{\bar{j}_t} - e_t) \otimes x_t$. By definition, we have that $\bar{z}_t^\top \tilde{\theta}_t \le z_t^\top \tilde{\theta}_t$, and on event $\{\mathcal{E}_t^{(1)} \bigcap \mathcal{E}_t^{(2)}\}$,

$$\Delta_t \le 2c_3\|\bar{z}_t\|_{\boldsymbol{V}_t^{-1}} + c_3\|z_t\|_{\boldsymbol{V}_t^{-1}}.$$

As in Lemma C.7, on $\mathcal{E}_t^{(1)}$,

$$\mathbb{E}_t[\Delta_t] \le \frac{3c_3}{p_3 - p_2}\mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big] + p_2 M.$$

We are now ready to upper bound the overall regret. Decomposing the regret,

$$\begin{aligned}\mathbb{E}\big[\text{Regret}_T(\{a_t\}_{t\in[T]})\big] \leq& \sum_{t=1}^{T}\mathbb{E}\big[\Delta_t \mathbb{1}\{\mathcal{E}_t^{(1)}\}\big] + p_1 TM \\ \leq& \sum_{t=1}^{T}\left(\frac{3c_3}{p_3-p_2}\mathbb{E}\big[\mathbb{E}_t\big[\|z_t\|_{\boldsymbol{V}_t^{-1}}\big]\big] + p_2 M\right) + p_1 TM.\end{aligned}$$

We proceed to bound the first sum.

Since $\boldsymbol{V}_t \succeq \lambda \boldsymbol{I}$, and $\|z_t\|_2 \leq 2$, we have that $\|z_t\|^2_{\boldsymbol{V}_t^{-1}} \leq 4/\lambda$. Choosing $\lambda \geq 4$, then we have that $\|z_t\|^2_{\boldsymbol{V}_t^{-1}} \leq 1$. By the matrix determinant lemma,

$$\det(\boldsymbol{V}_{t+1}) = \det(\boldsymbol{V}_t)(1 + \|z_t\|^2_{\boldsymbol{V}_t^{-1}}).$$

Since $\|z_t\|^2_{\boldsymbol{V}_t^{-1}} \in [0,1]$, $\log(1+\|z_t\|^2_{\boldsymbol{V}_t^{-1}}) \geq \|z_t\|^2_{\boldsymbol{V}_t^{-1}}/2$. Hence

$$\sum_{t=1}^{T}\|z_t\|^2_{\boldsymbol{V}_t^{-1}} \leq \sum_{t=1}^{T} 2\log(1+\|z_t\|^2_{\boldsymbol{V}_t^{-1}}) = \sum_{t=1}^{T} 2\log\frac{\det(\boldsymbol{V}_{t+1})}{\det(\boldsymbol{V}_t)} = 2\log\frac{\det(\boldsymbol{V}_{T+1})}{\det(\boldsymbol{V}_1)}.$$

On the other hand,

$$\text{tr}(\boldsymbol{V}_{T+1}) = \lambda d + \sum_{t=1}^{T}\|z_t\|^2_{\boldsymbol{V}_t^{-1}} \leq \lambda d + 4T.$$

Again, following the standard argument in Abbasi-Yadkori et al. (2011),

$$\det(\boldsymbol{V}_{T+1}) \leq (\lambda + 4T/d)^d.$$

Noting that $\det(\boldsymbol{V}_1) = \lambda^d$, we than have

$$\begin{aligned}\sum_{t=1}^{T}\|z_t\|_2^2 \leq& 2d\log\big(1+\frac{4T}{\lambda d}\big) \\ \sum_{t=1}^{T}\|z_t\|_2 \leq& \sqrt{2dT\log\big(1+\frac{4T}{\lambda d}\big)},\end{aligned}$$

where the second inequality is due to Cauchy-Schwartz. Now let $\delta = \frac{1}{T}$ and plug the above into the regret decomposition,

$$\begin{aligned}\mathbb{E}\big[\text{Regret}_T(\{a_t\}_{t\in[T]})\big] \leq& \frac{3c_3}{p_3-p_2}\sqrt{2dT\log\big(1+\frac{4T}{\lambda d}\big)} + (p_1 T + p_2 T)M \\ =& \frac{3c_3}{p_3-p_2}\sqrt{2dT\log\big(1+\frac{4T}{\lambda d}\big)} + 2M.\end{aligned}$$

□

D DEFERRED PROOF OF THE REGRET LOWER BOUND

Proof of Theorem 3.4. We will show the lower bound in the special case of $\phi \equiv 0$, i.e., the linear contextual bandit is unconfounded.

At any time period $t \in [T]$, let the context x_t be sampled uniformly from $\{e_1^{d_x}, \cdots, e_{d_x}^{d_x}\}$ where $e_j^d \in \mathbb{R}^d$ denotes the canonical d-dimensional basis with jth entry being 1. It is straightforward to see that $\|x_t\|_2 \leq 1$ always. Sample $B_t \sim$Unif$\{B \subseteq [d_a], |B| = \log N\}$. For simplicity, we later write $N = 2^m$ where $1 \leq m \leq d_a$ is an integer. The action set is chosen to be the set $\mathcal{A}_t = \{s/\sqrt{m} : s \in \{-1,+1\}^{B_t}\}$, i.e., $a := (a_j)_{j\in[d_A]} \in \mathcal{A}_t$ has entries

$$a_j = \begin{cases}\frac{s_j}{\sqrt{m}}, & j \in B_t \\ 0, & j \notin B_t\end{cases}.$$

It is easy to verify that $\|a_t\|_2 = 1$ and $|\mathcal{A}_t| = 2^m = N$. Let the noise $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ be independent of the context x_t, and Assumption 2.1 is satisfied. For a uniformly sampled $\bar{A}_t \in \mathcal{A}_t$, the expectation $e_t^{\text{unif}} = 0$ by construction. Hence

$$\begin{aligned}\mathbb{E}\big[\big((\bar{A}_t - e_t^{\text{unif}}) \otimes x_t\big)\big((\bar{A}_t - e_t^{\text{unif}}) \otimes x_t\big)^\top\big] =& \mathbb{E}\big[\big(\bar{A}_t \otimes x_t\big)\big(\bar{A}_t \otimes x_t\big)^\top\big] \\ =& \mathbb{E}[\bar{A}_t \bar{A}_t^\top] \otimes \mathbb{E}[x_t x_t^\top].\end{aligned}$$

Conditioned on a fixed B_t, the first term $\mathbb{E}[\bar{A}_t \bar{A}_t^\top \mid B_t] = \frac{1}{m}\sum_{j \in B_t} e_j^{d_a} (e_j^{d_a})^\top$, and

$$\mathbb{E}[\bar{A}_t \bar{A}_t^\top] = \frac{1}{d_a} \boldsymbol{I}_{d_a}.$$

Similarly the second term $\mathbb{E}[x_t x_t^\top] = \frac{1}{d_x} \boldsymbol{I}_{d_x}$. We conclude that

$$\mathbb{E}\big[\big((\bar{A}_t - e_t^{\text{unif}}) \otimes x_t\big)\big((\bar{A}_t - e_t^{\text{unif}}) \otimes x_t\big)^\top\big] = \frac{1}{d_a d_x} \boldsymbol{I}_d.$$

Hence Assumption 3.1 is satisfied with $\gamma = 1/d$.

We also let $u_{ij} \in \{-1, +1\}$ and let $\boldsymbol{\Theta} = \beta\sqrt{\frac{d}{T}}(u_{ij})_{i \in [d_x], j \in [d_a]}$, where β is a constant to be chosen later. We can verify that

$$\|\text{vec}(\boldsymbol{\Theta})\|_2 = \sqrt{d} \cdot \beta\sqrt{\frac{d}{T}} = \beta\frac{d}{\sqrt{T}}.$$

Hence, for $T \geq \beta^2 d^2$, we have $\|\text{vec}(\boldsymbol{\Theta})\|_2 \leq 1$, and Assumption 2.4 is satisfied. Finally, it is easy to verify that our constructed environment is in $\mathcal{E}_0(d_x, d_a, N, \sigma, \phi)$.

Now fix a realization of $x_t = e_i^{d_x}$ and B_t. Let Θ_i be the ith row of $\boldsymbol{\Theta}$: $\Theta_i = \beta\sqrt{\frac{d}{T}} u_i$. Then the optimal action $a_t^* \in \mathcal{A}_t$ is $a_t^* = s_t^*/\sqrt{m}$ where

$$s_t^* = \begin{cases} u_{ij}, & \text{for } j \in B_t \\ 0, & \text{for } j \notin B_t \end{cases}.$$

For any $a \in \mathcal{A}_t$, the instantaneous regret

$$\begin{aligned}x_t^\top \boldsymbol{\Theta}(a_t^* - a_t) =& \frac{2\beta}{\sqrt{m}}\sqrt{\frac{d}{T}} \cdot \Big(\sum_{j \in B_t} \mathbb{1}\{(a_t)_{i,j} \neq (a_t^*)_{i,j}\}\Big) \\ :=& \frac{2\beta}{\sqrt{m}}\sqrt{\frac{d}{T}} \cdot d_H(a_t, a_t^*).\end{aligned}$$

Here we write $d_H(a_t, a_t^*)$ as the Hamming distance between a_t and a_t^*. Using this definition, the expected regret can be decomposed as

$$\mathbb{E}[\text{Regret}_T] = \frac{1}{2^d}\sum_{\boldsymbol{U}} \Big(\frac{2\beta}{\sqrt{m}}\sqrt{\frac{d}{T}} \cdot \sum_{t=1}^T \mathbb{E}_{\boldsymbol{U}}\big[d_H(a_t, a_t^*)\big]\Big).$$

Hence, it suffices to lower bound $\mathbb{E}[d_H(a_t, a_t^*)]$ for a lower bound guarantee.

Fix some $\boldsymbol{\Theta}$ and some $(i, j) \in [d_x] \times [d_a]$. Let $\boldsymbol{\Theta}^{\oplus(i,j)}$ flips the sign of $\boldsymbol{\Theta}_{ij}$ (similarly define $\boldsymbol{U}^{\oplus(i,j)}$), and let $E_{ij,t} = \{x_t = e_i^{d_x}, j \in B_t\}$ and we can compute that $\mathbb{P}(E_{ij,t}) = \frac{1}{d_x} \cdot \frac{m}{d_a} = \frac{m}{d}$. Observe that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{U}}\big[d_H(a_t, a_t^*)\big] \\ =& \frac{1}{d_x\binom{d_a}{m}}\sum_{i=1}^{d_x} \sum_{B_t \subseteq [d_a], |B_t| = m} \mathbb{E}_{\boldsymbol{U}}\big[d_H(a_t, a_t^*) \mid x_t = e_i^{d_x}, B_t\big] \\ =& \frac{1}{d_x\binom{d_a}{m}}\sum_{i=1}^{d_x} \sum_{B_t \subseteq [d_a], |B_t| = m} \sum_{j \in B_t} \mathbb{P}_{\boldsymbol{U}}(s_{tj} \neq u_{ij} \mid E_{ij,t}).\end{aligned}$$

We can further average over all possible $\boldsymbol{U} \in \{-1,+1\}^d$:

$$\frac{1}{2^d}\sum_{\boldsymbol{U}\in\{-1,+1\}^d}\mathbb{E}_{\boldsymbol{U}}\big[d_H(a_t,a_t^*)\big]$$

$$=\frac{1}{2^d d_x\binom{d_a}{m}}\sum_{\boldsymbol{U}\in\{-1,+1\}^d}\sum_{i=1}^{d_x}\sum_{B_t\subseteq[d_a],|B_t|=m}\sum_{j\in B_t}\mathbb{P}_{\boldsymbol{U}}(s_{tj}\neq u_{ij}\mid E_{ij,t})$$

$$=\frac{1}{2^{d+1} d_x\binom{d_a}{m}}\sum_{\boldsymbol{U}\in\{-1,+1\}^d}\sum_{i=1}^{d_x}\sum_{B_t\subseteq[d_a],|B_t|=m}\sum_{j\in B_t}\Big(\mathbb{P}_{\boldsymbol{U}}(s_{tj}\neq u_{ij}\mid E_{ij,t})+\mathbb{P}_{\boldsymbol{U}^{\oplus(i,j)}}(s_{tj}=u_{ij}\mid E_{ij,t})\Big).$$

We proceed to bound the sum of the two probability in parenthesis.

Let $H_t=\{x_1,\mathcal{A}_1,a_1,r_1,\cdots,x_t,\mathcal{A}_t,a_t,r_t\}$ be the history and let $Q_{t,\boldsymbol{U}}$ be the probability distribution on H_t where the hidden parameter is $\boldsymbol{U}$. For two Normal distributions $\mathcal{N}(\mu_1,\sigma^2)$ and $\mathcal{N}(\mu_2,\sigma^2)$, it is well-known that

$$KL\big(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\big)=\frac{1}{2\sigma^2}(\mu_1-\mu_2)^2.$$

Using the fact that $2H^2(P,P')\le KL(P,P')$ for two distributions P,P', we have that

$$H^2\big(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)\big)\le\frac{1}{4\sigma^2}(\mu_1-\mu_2)^2.$$

We note that the two environment with parameters $\boldsymbol{U}$ and $\boldsymbol{U}^{\oplus(i,j)}$ only differ on the event $E_{ij,t}$. Hence, the squared Hellinger distance between the two environments at any $t\in[T]$ is at most $\frac{m}{d}\cdot\frac{4\beta^2 d}{4\sigma^2 mT}=\frac{\beta^2}{\sigma^2 T}$. By the sequential Hellinger chain inequality,

$$H^2\big(Q_{t-1,\boldsymbol{U}},Q_{t-1,\boldsymbol{U}^{\oplus(i,j)}}\big)\le\sum_{\tau=1}^{t-1}\mathbb{E}\big[H^2\big(Q_{\tau,\boldsymbol{U}},Q_{\tau,\boldsymbol{U}^{\oplus(i,j)}}\big)\big]\le\frac{T\beta^2}{\sigma^2 T}=\frac{\beta^2}{\sigma^2}.$$

Let $\sigma^2=1$, we have that

$$TV\big(Q_{t-1,\boldsymbol{U}},Q_{t-1,\boldsymbol{U}^{\oplus(i,j)}}\big)\le\sqrt{2H^2\big(Q_{t-1,\boldsymbol{U}},Q_{t-1,\boldsymbol{U}^{\oplus(i,j)}}\big)}\le\sqrt{2}\beta=\frac{1}{2},$$

where we set $\beta=\frac{\sigma}{2\sqrt{2}}=\frac{1}{2\sqrt{2}}$.

Now let $Y_t=(x_t,B_t)$ so that the filtration $\mathcal{F}_t=\sigma(H_{t-1},Y_t)$. Since Y_t is independent of the history H_{t-1} and the hidden parameter $\boldsymbol{U}$, the distribution

$$P_{\boldsymbol{U}}(H_{t-1},Y_t\mid E_{ij,t})=Q_{t-1,\boldsymbol{U}}\times\delta_{e_i^{d_x}}\times\text{Unif}\{B\subseteq[d_a]:|B|=m,j\in B\}.$$

Since the same conditional measure $\delta_{e_i^{d_x}}\times\text{Unif}\{B\subseteq[d_a]:|B|=m,j\in B\}$ appears under $\boldsymbol{U}^{\oplus(i,j)}$,

$$TV(P_{\boldsymbol{U}}(H_{t-1},Y_t\mid E_{ij,t}),P_{\boldsymbol{U}^{\oplus(i,j)}}(H_{t-1},Y_t\mid E_{ij,t}))=TV(Q_{t-1,\boldsymbol{U}},Q_{t-1,\boldsymbol{U}^{\oplus(i,j)}}).$$

According to the two-point testing inequality,

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{U}}(s_{tj}\neq u_{i,j}\mid E_{ij,t})+\mathbb{P}_{t,\boldsymbol{U}^{\oplus(i,j)}}(s_{tj}=u_{i,j}\mid E_{ij,t})\\ \ge&1-TV(P_{\boldsymbol{U}}(\mathcal{F}_t\mid E_{ij,t}),P_{\boldsymbol{U}^{\oplus(i,j)}}(\mathcal{F}_t\mid E_{ij,t}))\\ =&1-TV(Q_{t-1,\boldsymbol{U}},Q_{t-1,\boldsymbol{U}^{\oplus(i,j)}})=\frac{1}{2}.\end{aligned}$$

Hence,

$$\frac{1}{2^d}\sum_{\boldsymbol{U}\in\{-1,+1\}^d}\mathbb{E}_{\boldsymbol{U}}\big[d_H(a_t,a_t^*)\big]$$

$$=\frac{1}{2^{d+1} d_x\binom{d_a}{m}}\sum_{\boldsymbol{U}\in\{-1,+1\}^d}\sum_{i=1}^{d_x}\sum_{B_t\subseteq[d_a],|B_t|=m}\sum_{j\in B_t}\Big(\mathbb{P}_{\boldsymbol{U}}(s_{tj}\neq u_{ij}\mid E_{ij,t})+\mathbb{P}_{\boldsymbol{U}^{\oplus(i,j)}}(s_{tj}=u_{ij}\mid E_{ij,t})\Big)$$

$$\ge\frac{1}{2^{d+1} d_x\binom{d_a}{m}}\sum_{\boldsymbol{U}\in\{-1,+1\}^d}\sum_{i=1}^{d_x}\sum_{B_t\subseteq[d_a],|B_t|=m}\sum_{j\in B_t}\frac{1}{2}=\frac{m}{4}.$$

We can now lower bound

$$\begin{aligned}\sup_{P\in\mathcal{E}_0(d_x,d_a,N,\sigma,\phi)}\mathbb{E}_P[\text{Regret}_T] \geq& \sup_{\boldsymbol{U}\in\{-1,+1\}^d}\mathbb{E}_{\boldsymbol{U}}[\text{Regret}_T]\\ \geq& T\cdot\frac{m}{4}\cdot\frac{2\beta}{\sqrt{m}}\cdot\sqrt{\frac{d}{T}}=\frac{\beta}{2}\sqrt{dTm}\\ =&\frac{\sigma}{4\sqrt{2}}\sqrt{dT\log N}.\end{aligned}$$

□

E Deferred proofs of technical lemmas

E.1 Proof of Lemma C.2

Proof. Define $\widehat{\boldsymbol{\Sigma}}_t := \sum_{s\leq t}(\xi_s)^2 z_s z_s^\top$ and $\overline{\boldsymbol{\Sigma}}_t := \sum_{s\leq t}\mathbb{E}_s[(\xi_s z_s)(\xi_s z_s)^\top]$. Let $\boldsymbol{Q}\in\mathbb{R}^{d\times d}$ be any positive definite matrix. For simplicity, we write $\boldsymbol{A}_t := \boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t$.

The proof follows the recipe of de la Pena et al. (2009); Krishnamurthy et al. (2018). We start by applying the Chernoff method:

$$\begin{aligned}&\mathbb{P}\bigg(\Big\|\sum_{s=1}^t Z_s\xi_s\Big\|^2_{\boldsymbol{A}_t^{-1}}\geq 2\log\bigg(\frac{1}{\delta}\sqrt{\frac{\det(\boldsymbol{A}_t)}{\det(\boldsymbol{Q})}}\bigg)\bigg)\\ =&\mathbb{P}\bigg(\exp\bigg(\frac{1}{2}\Big\|\sum_{s=1}^t Z_s\xi_s\Big\|^2_{\boldsymbol{A}_t^{-1}}\bigg)\geq\frac{1}{\delta}\sqrt{\frac{\det(\boldsymbol{A}_t)}{\det(\boldsymbol{Q})}}\bigg)\\ \leq&\delta\mathbb{E}\bigg[\sqrt{\frac{\det(\boldsymbol{Q})}{\det(\boldsymbol{A}_t)}}\exp\bigg(\frac{1}{2}\Big\|\sum_{s=1}^t Z_s\xi_s\Big\|^2_{\boldsymbol{A}_t^{-1}}\bigg)\bigg].\end{aligned}$$

It remains to upper bound the expectation by 1.

By definition, we have $\mathbb{E}_t[Z_t]=0$. Since ξ_t is $\mathcal{F}_t$-measurable and deterministic given $\mathcal{F}_t$, we have $\mathbb{E}_t[\xi_t Z_t]=\xi_t\mathbb{E}_t[Z_t]=0$. Moreover, by Assumption 2.4 and clipping of the nuisance estimator, $|\xi_t|\leq 2M$ and $\|Z_t\|_2\leq 2$, hence $\|\xi_t Z_t\|_2\leq 4M$. Applying Lemma F.3 to $X_s=\xi_s Z_s$, we obtain that

$$U_t(\beta)=\exp\bigg(\beta^\top\sum_{s\leq t}\xi_s Z_s-\frac{1}{2}\beta^\top\sum_{s\leq t}\big\{(\xi_s)^2Z_sZ_s^\top+\mathbb{E}_s[(\xi_s)^2Z_sZ_s^\top]\big\}\beta\bigg)$$

is a supermartingale with $\mathbb{E}[U_t(\beta)]\leq 1$ for all $\beta\in\mathbb{R}^d$. By Fubini's theorem,

$$\begin{aligned}1\geq&\mathbb{E}_{\beta\sim\mathcal{N}(0,\boldsymbol{Q}^{-1})}\mathbb{E}[U_t(\beta)]\\ =&\mathbb{E}\int\frac{U_t(\beta)}{(2\pi)^{d/2}}\sqrt{\det(\boldsymbol{Q})}\exp\bigg(-\frac{1}{2}\beta^\top\boldsymbol{Q}\beta\bigg)\,d\beta\\ =&\mathbb{E}\int\sqrt{\frac{\det(\boldsymbol{Q})}{(2\pi)^d}}\exp\bigg(\beta^\top W_t^{(2)}-\frac{1}{2}\beta^\top\big(\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t+\boldsymbol{Q}\big)\beta\bigg)\,d\beta.\end{aligned}\tag{10}$$

Recall that $\boldsymbol{A}_t=\boldsymbol{Q}+\widehat{\boldsymbol{\Sigma}}_t+\overline{\boldsymbol{\Sigma}}_t$. Therefore,

$$\beta^\top W_t^{(2)}-\frac{1}{2}\beta^\top\boldsymbol{A}_t\beta=-\frac{1}{2}\big(\beta-\boldsymbol{A}_t^{-1}W_t^{(2)}\big)^\top\boldsymbol{A}_t\big(\beta-\boldsymbol{A}_t^{-1}W_t^{(2)}\big)+\frac{1}{2}(W_t^{(2)})^\top\boldsymbol{A}_t^{-1}W_t^{(2)}.$$

Hence the integral is a Gaussian integral:

$$\begin{aligned}
&\mathbb{E}\int \sqrt{\frac{\det(\boldsymbol{Q})}{(2\pi)^d}} \exp\left(\beta^\top W_t^{(2)} - \frac{1}{2}\beta^\top \boldsymbol{A}_t \beta\right) d\beta \\
=&\mathbb{E}\int \sqrt{\frac{\det(\boldsymbol{Q})}{(2\pi)^d}} \exp\Big(-\frac{1}{2}\big(\beta - \boldsymbol{A}_t^{-1}W_t^{(2)}\big)^\top \boldsymbol{A}_t \big(\beta - \boldsymbol{A}_t^{-1}W_t^{(2)}\big) + \frac{1}{2}(W_t^{(2)})^\top \boldsymbol{A}_t^{-1} W_t^{(2)}\Big) d\beta \\
=&\mathbb{E}\Big[\exp\left(\frac{1}{2}(W_t^{(2)})^\top \boldsymbol{A}_t^{-1} W_t^{(2)}\right)\sqrt{\frac{\det(\boldsymbol{Q})}{\det(\boldsymbol{A}_t)}}\Big] \\
=&\mathbb{E}\Big[\sqrt{\frac{\det(\boldsymbol{Q})}{\det(\boldsymbol{A}_t)}}\exp\left(\frac{1}{2}\Big\|\sum_{s=1}^t Z_s\xi_s\Big\|^2_{\boldsymbol{A}_t^{-1}}\right)\Big].
\end{aligned}$$

Plugging this into (10) gives

$$1 \geq \mathbb{E}\Big[\sqrt{\frac{\det(\boldsymbol{Q})}{\det(\boldsymbol{A}_t)}}\exp\left(\frac{1}{2}\Big\|\sum_{s=1}^t Z_s\xi_s\Big\|^2_{\boldsymbol{A}_t^{-1}}\right)\Big],$$

which concludes the proof. □

E.2 Proof of Lemma C.3

Proof. For any $t \leq T_0$, we have that $Z_t = (\bar{A}_t - e_t^{\text{unif}}) \otimes X_t$, $\bar{A}_t \sim \text{Unif}(\mathcal{A}_t)$ are independent and $\|Z_t\|_2 \leq 2$ and $\boldsymbol{\Gamma}_t = \mathbb{E}[Z_t Z_t^\top] \succeq \gamma \boldsymbol{I}$ due to Assumption 3.1. Now let $Y_t = \boldsymbol{\Gamma}_t - Z_t Z_t^\top$. Then we have that $\mathbb{E}[Y_t] = 0$ and $\|Y_t\|_{op} \leq 1$, and the Bernstein variance satisfies

$$\sigma_t^2 = \Big\|\sum_{s\leq t}\mathbb{E}[Y_s^2]\Big\| \leq 4t.$$

We are ready to invoke the matrix Bernstein inequality in Lemma F.5,

$$\lambda_{\max}\Big(\sum_{s\leq t} Y_s\Big) \leq 2\sqrt{2t\log(d/\delta)} + \frac{8}{3}\log(d/\delta),$$

with probability at least $1-\delta$. Since

$$\boldsymbol{V}_t = \lambda \boldsymbol{I} + \sum_{s\leq t} Z_s Z_s^\top = \lambda \boldsymbol{I} + \sum_{s\leq t}\boldsymbol{\Gamma}_s - \sum_{s\leq t} Y_s,$$

we can lower bound

$$\begin{aligned}
\lambda_{\min}(\boldsymbol{V}_t) \geq& \lambda_{\min}(\lambda \boldsymbol{I}) + \lambda_{\min}\Big(\sum_{s\leq t}\boldsymbol{\Gamma}_s\Big) - \Big(2\sqrt{2t\log(d/\delta)} + \frac{8}{3}\log(d/\delta)\Big) \\
\geq& \lambda + \gamma t - 2\sqrt{2t\log(d/\delta)} - \frac{8}{3}\log(d/\delta),
\end{aligned}$$

with probability at least $1-\delta$. Hence, if we need

$$\lambda_{\min}(\boldsymbol{V}_{T_0}) \geq \max\{\lambda + B, 1\},$$

then the exploration period needs to satisfy that

$$\begin{aligned}
\gamma T_0 - 2\sqrt{2T_0\log(d/\delta)} - \frac{8}{3}\log(d/\delta) \geq B \\
T_0 \geq 8\log(d/\delta)/\gamma^2 + (2B + \frac{16}{3}\log(d/\delta)))/\gamma.
\end{aligned}$$

□

F HELPER LEMMAS

Lemma F.1 (Concentration of Self-Normalized Processes (Abbasi-Yadkori et al., 2011, Theorem 1)). *Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtration. Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that ε_t is $\mathcal{F}_t$-measurable and ε_t's are σ^2-sub-Gaussian. Let X_t be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_t$-measurable. Assume that $\boldsymbol{Q}$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define $\boldsymbol{V}_t = \boldsymbol{Q} + \sum_{s \leq t} X_s X_s^\top$, and $S_t = \sum_{s \leq t} X_s \varepsilon_s$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,*

$$\|S_t\|^2_{\boldsymbol{V}_t^{-1}} \leq 2\sigma^2 \log\left(\frac{\det(\boldsymbol{V}_t)^{1/2} \det(\boldsymbol{Q})^{-1/2}}{\delta}\right)$$

Lemma F.2 (Determinant-Trace Inequality (Abbasi-Yadkori et al., 2011, Lemma 10)). *Suppose that $X_1, X_2, \cdots, X_t \in \mathbb{R}^d$ and $\|X_s\| \leq U$ for all $1 \leq s \leq t$. Let $\boldsymbol{V}_t = \lambda \boldsymbol{I} + \sum_{s \leq t} X_s X_s^\top$ for some $\lambda > 0$, then we have that $\det(\boldsymbol{V}_t) \leq (\lambda + tU^2/d)^d$.*

Lemma F.3 ((de la Pena et al., 2009, Lemma 7)). *Let $\{X_i\}_{i=1}^n$ be a sequence of conditionally centered vector-valued random variables adapted to the filtration $\{\mathcal{F}_i\}_{i=1}^n$ and such that $\|X_i\|_2 \leq B$ for some constant B. Then*

$$U_n(\beta) = \exp\left(\beta^\top \sum_{i=1}^n X_i - \beta^\top \left(\sum_{i=1}^n X_i X_i^\top + \mathbb{E}[X_i X_i^\top \mid \mathcal{F}_i]\right)\beta/2\right)$$

is a supermartingale with $\mathbb{E}[U_n(\beta)] \leq 1$ for all $\beta \in \mathbb{R}^d$.

Lemma F.4 ((Krishnamurthy et al., 2018, Lemma 14)). *Let $X_1, \cdots, X_t$ be conditionally centered random vectors in $\mathbb{R}^d$ adapted to a filtration $\{\mathcal{F}_s\}_{s \leq t}$ with $\|X_s\|_2 \leq U$ almost surely. Define $\widehat{\boldsymbol{V}}_t := \sum_{s \leq t} X_s X_s^\top$ and $\overline{\boldsymbol{V}}_t := \sum_{s \leq t} \mathbb{E}[X_s X_s^\top \mid \mathcal{F}_s]$. Then with probability at least $1 - \delta$, the following holds simultaneously for all unit vectors $v \in \mathbb{R}^d$:*

$$v^\top \overline{\boldsymbol{V}}_t v \leq 2 v^\top \widehat{\boldsymbol{V}}_t v + U^2(9d \log(9t) + 8 \log(2/\delta)).$$

Lemma F.5 (Matrix Bernstein inequality). *Let $X_1, \cdots, X_t$ be mean-zero, symmetric random matrices in $\mathbb{R}^{d \times d}$ such that $\|X_s\|_{op} \leq 1$ a.s. for all $s \leq t$. Then for all $c \geq 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{s \leq t} X_s\right) \geq c\right) \leq d \exp\left(-\frac{c^2}{2(\sigma^2 + c/3)}\right),$$

where $\sigma^2 = \|\sum_{s \leq t} \mathbb{E}[X_s^2]\|_{op}$ is the norm of the matrix variance of the sum.

Lemma F.6 (Anti-concentration Inequality). *For a Gaussian distributed random variable Z with mean m and variance σ^2, for any $z \geq 1$,*

$$\frac{1}{2\sqrt{\pi z}} e^{-z^2/2} \leq \mathbb{P}(|Z - m| > z\sigma) \leq \frac{1}{\sqrt{\pi z}} e^{-z^2/2}.$$

Lemma F.7 ((Li et al., 2017, Lemma 2)). *Let $\{X_t\}_t$ be a sequence in $\mathbb{R}^d$ satisfying that $\|X_t\|_2 \leq 1$. Let $\boldsymbol{V}_t = \sum_{s=0}^{t-1} X_s X_s^\top$ where we define $X_0 = 0$. Suppose there is an integer m such that $\lambda_{\min}(\boldsymbol{V}_{m+1}) \geq 1$, then for all $n > 0$,*

$$\sum_{t=m+1}^{m+n} \|X_t\|_{\boldsymbol{V}_t^{-1}} \leq \sqrt{2nd \log\left(\frac{n+M}{d}\right)}.$$