Learning with Incomplete Context: Linear Contextual Bandits with Pretrained Imputation

Hao Yan* **Heyan Zhang*** **Yongyi Guo**✉
Department of Statistics, University of Wisconsin–Madison

Abstract

The rise of large-scale pretrained models has made it feasible to generate predictive or synthetic features at low cost, raising the question of how to incorporate such surrogate predictions into downstream decision-making. We study this problem in the setting of online linear contextual bandits, where contexts may be complex, nonstationary, and only partially observed. In addition to bandit data, we assume access to an auxiliary dataset containing fully observed contexts—common in practice since such data are collected without adaptive interventions. We propose PULSE-UCB, an algorithm that leverages pretrained models trained on the auxiliary data to impute missing features during online decision-making. We establish regret guarantees that decompose into a standard bandit term plus an additional component reflecting pretrained model quality. In the i.i.d. context case with Hölder-smooth missing features, PULSE-UCB achieves near-optimal performance, supported by matching lower bounds. Our results quantify how uncertainty in predicted contexts affects decision quality and how much historical data is needed to improve downstream learning.

* Equal contribution, listed in alphabetical order.
✉ Corresponding author: guo98@wisc.edu.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

Contextual bandits provide a powerful framework for sequential decision-making under uncertainty, where the learner repeatedly observes a context, chooses an action, and receives a reward. The key challenge is to balance exploration and exploitation while adapting decisions to the observed context. Owing to their simplicity and flexibility, contextual bandits have been widely applied in practice, including personalized recommendations (Li et al., 2010), mobile health (Nahum-Shani et al., 2016), and online education platforms (Cai et al., 2021).

In many practical applications, the contexts required for decision-making may be missing or only partially observed during online interactions. For example, in the HeartSteps mobile health study (Liao et al., 2020), the full physiological state of a participant is unobserved, while only partial signals such as step counts, activity levels, or self-reports from wearables are available to guide intervention delivery. Similarly, in online education platforms (Lan and Baraniuk, 2016), a learners complete knowledge state across multiple concepts is latent, and the system only observes partial signals such as responses to quiz items or practice problems. At the same time, large offline datasets with substantially more complete contexts are often accessible, since they can be collected without interventions or adaptive decision-making. Such datasets have been shown to reveal richer contextual information than what is available in online interaction (Kausik et al., 2025), raising the question of how these auxiliary resources can be effectively leveraged to improve sequential decision-making when online contexts are missing.

In this work, we address the problem of linear contextual bandits when contexts are only partially observed

during online interaction, while offline auxiliary data provide full context information. The key idea is to use predictive models trained on auxiliary data to impute the missing contexts for online decisions. Even with access to auxiliary data, it is often reasonable in practice to combine pretrained imputations with simple policies such as linear bandits, since they yield stable and interpretable rules, and enable valid post-hoc statistical inference, which are crucial in applications such as healthcare and education (Rafferty et al., 2019; Zhang et al., 2024; Guo and Xu, 2025). Fundamental questions arise: how can predictive models trained on auxiliary data improve such decision rules, and how does imputation quality affect regret?

Our contributions. We propose `PULSE-UCB`, an online algorithm that uses auxiliary data to impute missing contexts and guide decision-making in linear contextual bandits. Under general context sequence distributions, we establish a regret bound of $\widetilde{\mathcal{O}}(dT^{1/2} + \delta_0 d^{3/2}T)$, where where T is the time horizon, d is the dimension of the full context, and δ_0 captures the quality of the predictive model learned from auxiliary data. In the special case of i.i.d. contexts with β-Hölder smooth missing features, we further show that $\delta_0 \lesssim N^{-\beta/(2\beta+d_S)}$, where N is the auxiliary sample size and d_S is the dimension of the observed contexts, and we complement this with a matching lower bound, establishing near-optimality in both the time horizon and the auxiliary data size.

1.1 Related works

Bandits with partially observed contexts. Given its importance, a substantial literature has studied contextual bandits with partially observed contexts. Many works impose parametric assumptions on the full context, such as i.i.d. Gaussian contexts or linear dynamical systems with additive Gaussian noise (Kim et al., 2023; Park and Faradonbeh, 2022, 2024; Zeng et al., 2025; Xu et al., 2021). In contrast, we do not impose parametric assumptions on the context distribution, which can be misspecified in complex applications. Other works allow more general distributions but with restrictions, such as fixed, time-invariant contexts (Kim et al., 2025), or contexts missing completely at random (MAR) (Jang et al., 2022). Compared to these works, we allow the context process to be dependent, nonstationary, and missing not at random (MNAR). A closely related work is Hu and Simchi-Levi (2025), which considers nonlinear bandits with i.i.d. partially observed contexts (including both MAR and MNAR scenarios) and leverages pretrained models with orthogonal statistical learning to derive regret upper bounds. Compared to Hu and Simchi-Levi (2025), we focus on linear contextual bandits with potentially dependent and nonstationary contexts and provide both upper and matching lower bounds, establishing near-optimal regret in this setting. Finally, another related line of work analyzes corrupted contexts and benchmarks against a mixture of contextual and multi-armed bandits (Bouneffouf, 2020), whereas our auxiliary data enable comparison to the stronger benchmark of the optimal full-context policy.

Connections to broader areas. Our work also relates to AI-assisted decision-making, where pretrained models support online policies (Tianhui Cai et al., 2024; Zhang et al., 2025; Chen et al., 2021; Janner et al., 2021; Lin et al., 2023; Lee et al., 2023; Ye et al., 2025; Cao et al., 2024), and to the broad literature on imputation-based methods in statistics and machine learning, from the classical EM algorithm (Dempster et al., 1977) to modern ML-based approaches (Xia and Wainwright, 2024; Angelopoulos et al., 2023). We differ by focusing specifically on the missing-context issue in online bandits, providing regret bounds that guide the principled use of pretrained imputation for sequential decisions. A more complete literature review is deferred to the Appendix A.

Notation For a positive integer n, we write $[n] = \{1, 2, \ldots, n\}$. For a vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)^\top \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$ and $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$. $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ denotes the n-by-n identity matrix. For positive functions $f(n)$ and $g(n)$, we write $f(n) \gtrsim g(n)$, $f(n) = \Omega(g(n))$ or $g(n) = \mathcal{O}(f(n))$ if for some constant $C > 0$, we have $f(n)/g(n) \geq C$ for all sufficiently large n. We write $f(n) = \widetilde{\mathcal{O}}(g(n))$ if $f(n) = \mathcal{O}(g(n)\text{polylog}(n))$, that is, there exist constants $C, k > 0$ such that $f(n) \leq Cg(n)(\log n)^k$ for all sufficiently large n.

2 PROBLEM SETUP

We consider a sequential decision-making process in contextual bandits with partially observed contexts. Given a time horizon T, for each $t = 1, 2, \ldots, T$:

(a) **Context generation.** A latent context $\boldsymbol{Y}_t \in \mathbb{R}^{d_Y}$ is generated from an unknown probability distribution $p_\star(\cdot \mid \boldsymbol{Y}_{1:t-1})$. At each time step t,

the agent observes a partial context $\boldsymbol{S}_t \in \mathbb{R}^{d_{S,t}}$. Note that we use the subscript t in the dimension $d_{S,t}$ to explicitly indicate that the dimension of the observed feature is allowed to vary over time.

(b) **Action and reward.** Based on the observed history, the agent selects an action $A_t \in \mathcal{A}$ and receives a reward $R_t = R(t, A_t)$. Here we define the potential reward as

$$R(t,a) := \langle \boldsymbol{\theta}^\star, \boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \rangle + \eta_t, \quad \forall t \in [T], a \in \mathcal{A}, \tag{2.1}$$

where $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ is an unknown parameter, $\boldsymbol{\Phi}$ is a known feature mapping, and η_t is mean-zero condition on past history. We assume that

$$R(t,a) \in [-1,1]. \tag{2.2}$$

The feature map $\boldsymbol{\Phi}$ satisfies the following assumption:

Assumption 2.1. *For any $a \in \mathcal{A}$, there exists $B > 0$*

$$\sup_{\boldsymbol{y} \in \mathbb{R}^{d_Y}} \|\boldsymbol{\Phi}(\boldsymbol{y}, a)\|_\infty \le 1, \quad \sup_{\boldsymbol{y} \in \mathbb{R}^{d_Y}} \|\boldsymbol{\Phi}(\boldsymbol{y}, a)\|_2 \le B.$$

In addition, we impose a standard assumption on the noise sequence $\{\eta_t\}_{t=1}^T$.

Assumption 2.2. *Suppose that $\{\eta_t\}_{t=1}^T$ is a σ_η^2-sub-Gaussian martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t=1}^T$. Here*

$$\mathcal{F}_t := \sigma\left(\boldsymbol{Y}_{1:t}, A_{1:t-1}, R_{1:t-1}\right), \tag{2.3}$$

where $\sigma(\cdot)$ denotes the generated σ-algebra.

Note that in this bandit setting, both the full context $\boldsymbol{Y}_t$ and its partial observation $\boldsymbol{S}_t$ are assumed to be exogenous and do not depend on the action sequence $(A_1, \ldots, A_t)$. This asymmetric context availability, where offline auxiliary data provides richer, higher-fidelity information than what can be feasibly observed during real-time online decision-making, is strongly motivated by practical constraints and naturally arises in many real-world applications. For instance, in digital health interventions, the full state of a patient $\boldsymbol{Y}_t$ may include comprehensive physiological and psychological factors such as stress level and sleep quality. However, because online systems often face strict privacy regulations, local computational limits, or frequent sensor failures, only a partial subset $\boldsymbol{S}_t$ (such as step counts or heart rate) is observed in real-time. Similarly, in online education platforms, a learner's true knowledge state across multiple concepts ($\boldsymbol{Y}_t$) is unobservable during real-time interactions, and the system only receives partial signals like answers to specific quiz items or homework questions ($\boldsymbol{S}_t$). Fortunately, archival offline data in these domains can safely store the comprehensive state variables ($\boldsymbol{Y}_t$) to aid online learning. We provide a detailed discussion of these practical motivations in Appendix B.

Our goal is to sequentially select actions $\{A_t\}_{t=1}^T$, where each A_t is chosen based only on the observed history $\{(\boldsymbol{S}_\tau, A_\tau, R_\tau)\}_{\tau=1}^{t-1}$ and the current observation $\boldsymbol{S}_t$, so as to maximize the cumulative reward. This is equivalent to minimizing the cumulative regret

$$\sum_{t=1}^T \mathbb{E}\left[R(t, A_t^\star) - R(t, A_t)\right],$$

where $A_t^\star$ is the optimal action that maximizes the expected reward, assuming the full context $\boldsymbol{Y}_t$ is observed:

$$A_t^\star := \arg\max_{a \in \mathcal{A}} \langle \boldsymbol{\theta}^\star, \boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \rangle. \tag{2.4}$$

The key challenge is that the latent contexts are only observed indirectly through the partial information $\boldsymbol{S}_{1:T}$. In general, good decision-making is impossible without adequate knowledge of the underlying contexts. In practice, however, it is often possible to obtain auxiliary historical data from related populations that include both partial observations and richer measurements of the underlying state. In the digital health example, historical studies often collect both wearable sensor streams and survey or clinical assessments. In online education, large-scale platforms frequently link fine-grained interaction logs (e.g., quiz responses, practice problems) with standardized test scores or comprehensive assessments, providing aligned data on both partial signals and richer proxies of the true knowledge state. Motivated by these settings, we assume access to an auxiliary dataset $\mathcal{D}$ consisting of i.i.d. trajectories

$$\mathcal{D} = \left\{ \left(\boldsymbol{Y}_{i,1:T_0}^{(0)}, \boldsymbol{S}_{i,1:T_0}^{(0)} \right) : i = 1, \ldots, N \right\},$$

where $T_0 \ge 1$ denotes the time horizon of the historical data, and each trajectory $(\boldsymbol{Y}_{i,1:T_0}^{(0)}, \boldsymbol{S}_{i,1:T_0}^{(0)})$ is drawn from the same joint distribution as the bandit contexts $(\boldsymbol{Y}_{1:T_0}, \boldsymbol{S}_{1:T_0})$. This dataset is assumed to reasonably capture the joint distribution of $\boldsymbol{Y}_{1:T}$ and $\boldsymbol{S}_{1:T}$. For instance, if the dependence structure between $\boldsymbol{Y}_{1:T}$ and

$\boldsymbol{S}_{1:T}$ is complex, one may require T_0 to be of the same order as the bandit horizon T in order to accurately recover this relation. In general, however, T_0 is flexible and need not be greater than T, and most of our results impose no explicit relation between them.

3 THE PULSE-UCB ALGORITHM

Under the setting introduced above, we propose *Pretrained Unobserved Latent State Estimation UCB* (`PULSE-UCB`), an algorithm that leverages auxiliary data to fill in the blanks of the missing contexts before making decisions. The main idea is as follows. We first pretrain a model $\widehat{p}$ on $\mathcal{D}$ that learns to predict the full context $\boldsymbol{Y}_t$ from the observed sequence $\boldsymbol{S}_{1:t}$[2]. Then, during online interaction, whenever we only see the partial context $\boldsymbol{S}_{1:t}$, we use $\widehat{p}$ to impute the missing parts and obtain complete feature vectors $\widehat{\boldsymbol{\phi}}_{t,a}$ for each action. With these surrogate features in hand, the problem reduces to a standard linear contextual bandit, and we apply OFUL (Abbasi-Yadkori et al., 2011): the algorithm maintains a confidence set for the unknown parameter $\boldsymbol{\theta}^\star$, chooses the action that maximizes an optimistic reward estimate, observes the payoff, and updates its estimates accordingly. In this way, the pretrained model provides the missing information, while OFUL handles the exploration-exploitation trade-off.

To formalize the imputation step, at each time t, for any action $a \in \mathcal{A}$, we define the imputed features as the conditional expectation of $\boldsymbol{\Phi}(\boldsymbol{Y}_t, a)$ under the pretrained model $\widehat{p}$:

$$\widehat{\boldsymbol{\phi}}_{t,a} := \mathbb{E}_{\widehat{p}}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}\right], \quad \text{for all } t \in [T]. \tag{3.1}$$

In practice, this conditional expectation may not admit a closed-form expression. However, a natural approximation is to draw samples $\boldsymbol{y}^{(b)} \sim \widehat{p}(\cdot \mid \boldsymbol{S}_{1:t})$ for $b \in [B]$ and compute the Monte Carlo average

$$\widehat{\boldsymbol{\phi}}_{t,a} \approx \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\Phi}\left(\boldsymbol{y}^{(b)}, a\right).$$

Such approximation can be made arbitrarily accurate, given sufficient computational resources, and we therefore assume direct access to $\widehat{\boldsymbol{\phi}}_{t,a}$ in later analysis.

[2]Here $\widehat{p}$ can be any pretrained model that provides a conditional distribution of $\boldsymbol{Y}_t$ given $\boldsymbol{S}_{1:t}$. If $\widehat{p}$ provides a deterministic prediction, one can convert it into a probabilistic model by viewing it as the mean of a suitably chosen distribution.

A full description is given in Algorithm 1.

Algorithm 1 `PULSE-UCB`

Require: Pretrained distribution $\widehat{p}$, tuning parameters $\lambda, \{\gamma_t\}_{t=1}^T$.

1: Initialize $\boldsymbol{\Sigma}_0 = \lambda \boldsymbol{I}$, $\texttt{BALL}_0 \leftarrow \left\{\boldsymbol{\theta} \mid \lambda\|\boldsymbol{\theta}\|_2^2 \le \gamma_0\right\}$.

2: **for** $t = 1$ to T **do**

3: Observe context $\boldsymbol{S}_t$, compute $\widehat{\boldsymbol{\phi}}_{t,a}$ according to Equation (3.1).

4: Choose action

$$A_t = \text{argmax}_{a\in\mathcal{A}} \max_{\boldsymbol{\theta}\in\texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,a}, \tag{3.2}$$

with ties broken arbitrarily.

5: Receive payoff R_t.

6: Update

$$\boldsymbol{\Sigma}_t \leftarrow \lambda \boldsymbol{I} + \sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}\widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top, \tag{3.3}$$

$$\widehat{\boldsymbol{\theta}}_t \leftarrow \boldsymbol{\Sigma}_t^{-1}\sum_{\tau=1}^{t} R_\tau \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}. \tag{3.4}$$

$$\texttt{BALL}_t \leftarrow \left\{\boldsymbol{\theta} \mid \left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right)^\top \boldsymbol{\Sigma}_t \left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right) \le \gamma_t\right\}. \tag{3.5}$$

7: **end for**

4 REGRET ANALYSIS

In this section, we analyze the regret of Algorithm 1. Section 4.1 characterizes the imputation error of the context from the pretrained model, which serves as a key ingredient in the analysis. Section 4.2 then establishes a general regret bound under arbitrary context distributions, and Section 4.3 specializes the result to some specific distributional settings.

4.1 Characterizing imputation error

Our first step in the regret analysis is to quantify the quality of the imputed contexts—that is, how far the predicted $\boldsymbol{Y}_t$ can deviate from the true $\boldsymbol{Y}_t$ given the current partial context $\boldsymbol{S}_{1:t}$. We capture this discrepancy through how well the pretrained model $\widehat{p}$ approximates the ground-truth distribution. Formally, let $\widehat{\mathbb{P}}$ denote the distributions of $(\boldsymbol{Y}_{1:T}, \boldsymbol{S}_{1:T})$ under $\widehat{p}$. For any $t \in [T]$, we measure the divergence between $\widehat{\mathbb{P}}$ and the ground truth $\mathbb{P}$ by

$$D_t = \tfrac{1}{2}\text{KL}\Big(\mathbb{P}(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}) \| \widehat{\mathbb{P}}(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t})\Big). \tag{4.1}$$

The next lemma establishes the theoretical basis that a small D_t ensures the imputed contexts remain close to the true contexts, enabling reliable downstream decision-making. The proof is in the Appendix E.

Lemma 4.1. For any time step $t \in [T]$ and any measurable scalar function $g : \mathbb{R}^{d_Y} \to \mathbb{R}$ with $\|g\|_\infty \leq 1$,

$$\mathbb{E}[g(\boldsymbol{Y}_t)|\, \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}] - \mathbb{E}_{\widehat{p}}[g(\boldsymbol{Y}_t)|\, \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}] \leq \sqrt{D_t}. \tag{4.2}$$

Here $\mathbb{E}[\cdot]$ denotes the expectations with respect to $\mathbb{P}$.

Lemma 4.1 can be applied to obtain an error bound for the imputed contexts used in bandit decisions. Specifically, considering Equation (4.2) and Assumption 2.1, for any action $a \in \mathcal{A}$ we have

$$\left\| \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}\right] - \widehat{\boldsymbol{\phi}}_{t,a} \right\|_\infty \leq \sqrt{D_t}. \tag{4.3}$$

Remark 4.1. *The choice of KL divergence in defining D_t in (4.1) is for convenience. In fact, the proof bounds the left-hand side of (4.2) via total variation and Pinsker's inequality. Thus, any f-divergence with a Pinsker-type bound (e.g., Hellinger distance) yields an analogous result. If the left-hand side of (4.2) can be controlled directly, no divergence measure is needed at all.*

4.2 Regret analysis under general context distributions

To establish the regret bound of Algorithm 1, we begin by decomposing the reward at round t. Specifically,

$$\begin{aligned} R_t &= \boldsymbol{\theta}^{\star\top} \boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) + \eta_t \\ &= \boldsymbol{\theta}^{\star\top} \mathbb{E}\big[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t\big] + \varepsilon_t + \eta_t, \end{aligned} \tag{4.4}$$

where in the first term, $\mathbb{E}\big[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t\big]$ can be viewed as a new effective context—the part we could recover if the underlying distribution $\mathbb{P}$ were known. The second term,

$$\varepsilon_t := \boldsymbol{\theta}^{\star\top} \left(\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) - \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \,|\, \boldsymbol{S}_{1:t}, A_t\right]\right), \tag{4.5}$$

captures the error introduced by the unobserved portion of the true context $\boldsymbol{Y}_t$. Intuitively, if ε_t has mean zero conditioned on the history, then $\varepsilon_t + \eta_t$ forms a martingale difference sequence. This structure allows us to invoke martingale self-normalized concentration techniques to analyze the regret of the resulting linear contextual bandit if $\mathbb{P}$ is known. To ensure that the error term ε_t can be properly controlled, we impose the following assumption.

Assumption 4.1. *For all $a \in \mathcal{A}$ and $t \in [T]$,*

$$\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, \boldsymbol{S}_{1:t}] = \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}]\,.$$

Assumption 4.1 is quite general. It holds whenever $\boldsymbol{Y}_t$ is conditionally independent of $(\boldsymbol{Y}_{1:t-1}, \eta_{1:t-1})$ given $\boldsymbol{S}_{1:t}$, i.e., whenever $\boldsymbol{S}_{1:t}$ provides sufficient information for predicting $\boldsymbol{Y}_t$. In particular, it is satisfied whenever $\boldsymbol{Y}_t$ is generated as an arbitrary function of $\boldsymbol{S}_{1:t}$ together with independent randomness. This already covers a broader class of models than many existing works on stochastic contextual bandits, which typically assume i.i.d. contexts (Li et al., 2021; Kim et al., 2023; Hu and Simchi-Levi, 2025). Informally, the key requirement is that the new randomness injected at time t does not depend on past randomness; whenever this holds, our algorithm and analysis apply.

Remark 4.2. *An exception arises when $\boldsymbol{Y}_t$ depends on $\boldsymbol{S}_{1:t}$ but is driven by autoregressive (AR) rather than independent noise. In such cases, latent variables exhibit temporal dependence that propagates into the rewards, and different techniques are required. Moreover, different forms of AR structure yield fundamentally different learning dynamics: action-driven AR models (Bacchiocchi et al., 2024) admit modified UCB algorithms with $\mathcal{O}(\sqrt{T})$ regret, whereas latent AR models (Trella et al., 2024) require Kalman-filter-type estimators and may incur regret exceeding $\mathcal{O}(\sqrt{T})$. Even without Assumption 4.1, however, ε_t can sometimes be controlled by alternative means: for example, if $\boldsymbol{W}_t$ is generated by a stationary process with geometrically decaying dependence (e.g., a stationary AR process), then concentration inequalities for mixing sequences may be applied, though such analysis would typically require additional structural assumptions on the feature map $\boldsymbol{\Phi}$. These distinctions underscore that extending beyond Assumption 4.1 to AR settings generally requires case-by-case, model-specific treatment; a detailed discussion is provided in Appendix C.*

Lemma 4.2. Under Assumptions 2.2 and 4.1, the random variables $\{\varepsilon_t\}_{t=1}^T$ defined in Equation (4.5) is a martingale difference sequence with respect to the filtration $\{\mathcal{G}_t\}_{t=1}^T$, which is given by

$$\mathcal{G}_{t-1} := \sigma\left(\boldsymbol{S}_{1:t}, \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, U_{1:t}\right)$$

where $U_{1:t}$ are independent auxiliary random variables in selecting $A_{1:t}$ under randomized algorithm.

As a proof sketch, the main goal is to show that

$$\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \mathcal{G}_{t-1}\right] = \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t\right].$$

Ignoring the auxiliary randomness $U_{1:t}$ and considering a simplified setting where $(A_{1:t-1}, R_{1:t-1})$, the action and reward prior to round t, can be expressed via $(\boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, \boldsymbol{S}_{1:t})$. The term $\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \mathcal{G}_t\right]$ can then be converted to a conditional expectation over $(\boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, \boldsymbol{S}_{1:t})$. Under Assumption 4.1, such conditional expectation only depends on $\boldsymbol{S}_{1:t}$, rendering Equation (4.5) a martingale difference sequence. The complete proof is given in the Appendix F.

With ε_t forming a martingale difference sequence, and considering the reward decomposition (4.4), we can then adapt self-normalized concentration techniques from linear contextual bandits (Abbasi-Yadkori et al., 2011). However, an important caveat arises: the effective context $\mathbb{E}\big[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t)|\boldsymbol{S}_{1:t}, A_t\big]$ is unknown because the true context distribution $\mathbb{P}$ is unobserved. Instead, it can only be approximated by $\mathbb{E}_{\widehat{p}}\big[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t)|\boldsymbol{S}_{1:t}, A_t\big]$. Our analysis therefore requires an additional sensitivity argument that quantifies how inaccuracies in the imputed contexts affect the cumulative regret, leading to regret bounds that explicitly depend on the approximation error between $\widehat{\mathbb{P}}$ and $\mathbb{P}$.

At each time t, define the conditional instantaneous regret between A_t and $A_t^\star$ given the observed context $\boldsymbol{S}_{1:t}$ as

$$\mathrm{reg}_t = \mathbb{E}\left[R(t, A_t^\star) - R(t, A_t) \mid \boldsymbol{S}_{1:t}\right], \tag{4.6}$$

and define the cumulative conditional regret up to horizon T as

$$\mathcal{R}_T := \sum_{t=1}^{T} \mathrm{reg}_t.$$

Remark 4.3 (Alternative Regret Formulation)**.** *Our regret is defined against an oracle observing the full context Y_t. Alternatively, one might consider a baseline that acts on the conditional expected context: $A_t^\dagger = \arg\max_a \mathbb{E}[\boldsymbol{\theta}^{\star\top}\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}]$. By Jensen's inequality, the expected reward of $A_t^\dagger$ is bounded by that of our full-context oracle:*

$$\max_a \mathbb{E}[\boldsymbol{\theta}^{\star\top}\boldsymbol{\Phi}(\boldsymbol{Y}_t, a)|\boldsymbol{S}_{1:t}] \le \mathbb{E}\Big[\max_a \boldsymbol{\theta}^{\star\top}\boldsymbol{\Phi}(\boldsymbol{Y}_t, a)|\boldsymbol{S}_{1:t}\Big].$$

Therefore, bounding the regret against the full-context oracle establishes a strictly stronger theoretical guarantee that naturally upper-bounds the regret against the conditional-expectation baseline.

We now state the main result. The next theorem provides a high-probability upper bound on the cumulative regret of Algorithm 1 under general context distributions. The proof is provided in the Appendix G.

Theorem 4.1. Suppose that $\|\boldsymbol{\theta}^\star\|_2 \le 1$, and let Assumptions 2.2 and 4.1 hold. For a given $\delta \in (0, 1)$, in Algorithm 1 choose

$$\gamma_t := \gamma_t^{(0)} + 3d^2 \sum_{\tau=1}^{t} D_t, \tag{4.7}$$

where

$$\gamma_t^{(0)} := 3\lambda + 6(\sigma_\eta + 2)^2 \log\left[\frac{4t^2}{\delta}\left(1 + \frac{tB^2}{d\lambda}\right)^d\right],$$

and D_t is defined in (4.1). Then, with probability at least $1 - \delta$, the regret of Algorithm 1 satisfies

$$\mathcal{R}_T \le \mathcal{R}_T^{(\mathtt{imp})} + \mathcal{R}_T^{(\mathtt{lin})}.$$

Here,

$$\mathcal{R}_T^{(\mathtt{lin})} := 2\sqrt{2\gamma_T^{(0)} dT \log\left(1 + \frac{TB^2}{d\lambda}\right)}$$

denotes the standard $\widetilde{\mathcal{O}}(d\sqrt{T})$ cumulative regret achieved by the vanilla `OFUL` algorithm, and

$$\mathcal{R}_T^{(\mathtt{imp})} = 4\sqrt{6d^3\left(\sum_{t=1}^{T} D_t\right) T \log\left(1 + \frac{TB^2}{d\lambda}\right)}$$

captures the additional cumulative regret due to imputing missing context with the pretrained model.

Note that both Lemma 4.1 and Theorem 4.1 remain valid regardless of how $\widehat{p}$ is trained or whether it is correctly specified. Thus, our theory is applicable to a broad range of modern machine learning models.

Remark 4.4. *In Theorem 4.1, the choice of hyperparameters γ_t depends on D_t, which may not be directly known. In many practical settings, however, D_t or its order can be reasonably estimated. For example, if the dimensions of $\boldsymbol{\Phi}(\boldsymbol{Y}_t, a)$ and $\boldsymbol{Y}_t$ are bounded and the dependence of $\boldsymbol{Y}_t$ on $\boldsymbol{S}_{1:t}$ is parametric within a fixed window (i.e., depending only on the most recent few $\boldsymbol{S}_\tau$'s), then D_t is typically of order $\widetilde{\mathcal{O}}((NT_0)^{-1/2})$, leading to $\mathcal{R}_T^{(imp)} = \widetilde{\mathcal{O}}(T(NT_0)^{-1/2})$. More generally, when the dependence is nonparametric but smooth, the order of D_t can also be derived (see Section 4.3). In both parametric and nonparametric settings—and in more general cases without structural assumptions—D_t and γ_t may also be chosen in a data-driven manner. We defer a detailed discussion to the Appendix D.*

Remark 4.5. *We also provide a variant of Algorithm 1, which combines the imputations with an alternative*

online exploration algorithm similar to the idea from He et al. (2022). We show that an improved regret guarantee is possible when the order of total imputation error is known and the error varies across time. See Appendix I for details.

Remark 4.6 (Time-Varying Context Dimensions and Structures)**.** *We explicitly emphasize that our framework does not restrict the dimension of the observed state $\boldsymbol{S}_t$ to be fixed across all time steps. The notation $\boldsymbol{S}_t \in \mathbb{R}^{d_{S,t}}$ accommodates highly dynamic real-world environments where the availability of sensors or recorded features fluctuates over time (e.g., Missing-Not-At-Random scenarios). Our theoretical guarantees (Theorem 4.1) hold even when the online learner does not share the exact same set of observed features at every round. To empirically substantiate this, we demonstrate* `PULSE-UCB`*'s robustness to dynamically changing observed feature sets via a Random Feature Masking experiment in Appendix M.3.*

4.3 Application of Theorem 4.1

We now provide several examples that yield explicit rates for D_t in Theorem 4.1 and the resulting cumulative regret of Algorithm 1. These examples are intentionally simplified for clarity but remain representative, and the ideas extend to more general settings.

Suppose that the full context $\boldsymbol{Y}_t$ can be written as $(\boldsymbol{S}_t, W_t) \in \mathbb{R}^{d_S} \times \mathbb{R}$, where $\boldsymbol{S}_t$ denotes the partially observed features in the bandit period and W_t denotes the features that are missing.

Linear Model Consider a linear model where

$$W_t = \sum_{j=0}^{m} \boldsymbol{\beta}_j^\top \boldsymbol{S}_{t-j} + \xi_t,\ \xi_t \sim \mathcal{N}(0,1),\ \forall t \in [T] \quad (4.8)$$

and $\boldsymbol{S}_{-j} = \boldsymbol{0}$ for $j \in [m]$. The historical data contains N i.i.d. observations of length T_0 from Equation (4.8):

$$\mathcal{D} = \left\{ \boldsymbol{Y}_{i,1:T_0}^{(0)} \right\}_{i=1}^{N} = \left\{ \left(\boldsymbol{S}_{i,1:T_0}^{(0)}, W_{i,1:T_0}^{(0)} \right) \right\}_{i=1}^{N} \quad (4.9)$$

Proposition 4.1. Suppose that the historical data $\mathcal{D}$ follows Equation (4.9) and the missing feature W_t follows Equation (4.8). Assume that $\boldsymbol{S}_{i,t}^{(0)} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{d_S})$ for all $i \in [N]$ and $t \in [T_0]$, and that $T_0 \geq 2m$. There exists a pretrained model $\widehat{p}$ such that

$$\mathbb{E}\sqrt{D_t} \lesssim \sqrt{\frac{m d_S}{N T_0}}.$$

Thus, the expected cumulative regret of Algorithm 1, taken over $\boldsymbol{S}_{1:T}$ and $\mathcal{D}$, satisfies

$$\mathbb{E}\left[\mathcal{R}_T\right] = \widetilde{\mathcal{O}}\left(T\sqrt{\frac{m d_S d^3}{N T_0}} + d\sqrt{T} \right).$$

Nonparametric Model As another example, consider the case where $T_0 = 1$, so that the historical dataset $\mathcal{D}$ contains N i.i.d. samples

$$\mathcal{D} = \left\{ \boldsymbol{Y}_i^{(0)} \right\}_{i=1}^{N} = \left\{ (\boldsymbol{S}_i^{(0)}, W_i^{(0)}) \right\}_{i=1}^{N} \quad (4.10)$$

where each missing feature $W_i \in \mathbb{R}$. For simplicity, assume $\boldsymbol{S}_i^{(0)} \sim \text{Unif}\left([0,1]^{d_S}\right)$. Consider a nonparametric regression model where for all $i \in [N]$,

$$W_i^{(0)} = f\left(\boldsymbol{S}_i^{(0)}\right) + \xi_i, \quad \xi_i \sim \mathcal{N}(0,1). \quad (4.11)$$

Here f is a scalar-value function that satisfies the Hölder smoothness condition with parameters (β, L).

Assumption 4.2 (Hölder Smoothness)**.** *A function $f : \mathbb{R}^{d_S} \to \mathbb{R}$ satisfies the Hölder condition with parameters (β, L) if for all $\boldsymbol{s}, \boldsymbol{s}' \in \mathbb{R}^{d_S}$,*

$$|f(\boldsymbol{s}) - f(\boldsymbol{s}')| \leq L \left\|\boldsymbol{s} - \boldsymbol{s}'\right\|_2^{\beta} \quad (4.12)$$

for some $\beta \in (0,1]$ and $L > 0$. Denote the class of such functions as $\mathcal{F}_{\beta,L}$.

Proposition 4.2. Suppose that the historical data $\mathcal{D}$ is specified by Equation (4.10) and the missing feature W follows Equation (4.11) with f satisfying Assumption 4.2. Then there exists $\widehat{p}$ such that

$$\mathbb{E}\sqrt{D_t} \lesssim N^{-\frac{\beta}{2\beta + d_S}}$$

and thus, the expected cumulative regret of Algorithm 1, taken over $\boldsymbol{S}_{1:T}$ and $\mathcal{D}$, satisfies

$$\mathbb{E}\left[\mathcal{R}_T\right] = \widetilde{\mathcal{O}}\left(T d^{\frac{3}{2}} N^{-\frac{\beta}{2\beta + d_S}} + d\sqrt{T} \right).$$

In Section 5, we show that this upper bound is near-minimax optimal in both the time horizon and the auxiliary sample size, provided the auxiliary dataset is sufficiently large.

The proof of both propositions, as well as the explicit choice of $\widehat{p}$ is included in the Appendix H. As a remark, the examples above focus on a one-dimensional missing covariate. Extending to a general d_W-dimensional missing context is straightforward and leads to a similar result, except for an additional $\sqrt{d_W}$ factor from

handling coordinates separately (e.g., via a union bound or vector-valued concentration). Furthermore, the rates in Propositions 4.1 and 4.2 are robust to mild covariate shift between the historical and online distributions; see Appendix H.3 for details.

5 LOWER BOUNDS

To demonstrate the optimality of our algorithm, we establish a minimax lower bound by analyzing a carefully constructed two-arm contextual bandit instance with partially observed contexts. The exact data-generating process, including feature construction and verification of technical conditions, is provided in the Appendix J&K. We give a high-level overview below.

Our construction highlights two fundamental sources of difficulty. First, the reward of one arm depends on an unobserved scalar W, whose conditional mean is determined by an unknown function $f \in \mathcal{F}_{\beta,L}$ defined on a $d_{\texttt{non}}$-dimensional subset of the observed context $\boldsymbol{S}$. When historical data are limited, the challenge of estimating f dominates the regret. Second, the rewards of both arms involve a linear parameter $\boldsymbol{\theta}^\star$ acting on the complementary $d_{\texttt{lin}}$-dimensional subset of $\boldsymbol{S}$. The two subsets together form a partition of $\boldsymbol{S}$, so that $d_{\texttt{lin}} + d_{\texttt{non}} = d_S$. Once f can be accurately estimated from pretraining data, the remaining difficulty reduces to online learning of $\boldsymbol{\theta}^\star$, which contributes a $\sqrt{T}$ regret term.

By alternating between these two regimes, the construction forces both sources of error to matter: historical samples provide noisy information about f, while online bandit interaction governs the estimation of $\boldsymbol{\theta}^\star$. As a result, the minimax regret necessarily includes two additive components: one tied to the nonparametric rate for learning f, and the other to the linear rate for estimating $\boldsymbol{\theta}^\star$. Full details are deferred to the Appendix J & K.

Theorem 5.1 (Informal Lower Bound)**.** Consider the two-arm contextual bandit problem with partially observed contexts $\boldsymbol{S}_t \in \mathbb{R}^{d_S}$. Under suitable regularity conditions, there exists a construction such that the minimax expected cumulative regret satisfies the rate of

$$\Omega\Big(TN^{-\frac{\beta}{2\beta+d_{\texttt{non}}}} + \sqrt{d_{\texttt{lin}}T}\Big),$$

where $d_{\texttt{non}}, d_{\texttt{lin}} > 0$ denote the nonparametric and linear dimensions of the observed context, respectively, and $d_{\texttt{non}} + d_{\texttt{lin}} = d_S$.

Remark 5.1. *When $N = \Omega\big(T^{\frac{2\beta+d_S}{2\beta}}\big)$, both the upper bound in Proposition 4.2 and the lower bound in Theorem 5.1 reduce to $\sqrt{T}$ (up to logarithmic factors). Consequently, for sufficiently large N, Algorithm 1 attains near-minimax optimality. Notably, this matches the oracle rate when the context is fully observed, indicating that with ample data there is no efficiency loss when leveraging a well-suited pretrained model.*

For small N (taking $d_{\texttt{non}} = d_S - 1$ in Theorem 5.1), we observe a slight difference in the N-dependence relative to Proposition 4.2. This stems from our proofs partition of $\boldsymbol{S}$ into complementary subsets to decouple the nonparametric and linear components. Allowing the linear part to also depend on the nonparametric coordinates would likely shift the dependence toward d_S, but entails substantially complicated analysis. Sharpening the small-N dependence is an appealing direction for future work.

6 NUMERICAL EXPERIMENTS

In this section, we validate our theory and algorithm with simulations on synthetic data and the real Taobao Ad Display/Click dataset (Alibaba, 2018). To facilitate reproducibility, the complete source code is publicly available at `https://github.com/Heyan-Zhang/PULSE-UCB`.

6.1 Synthetic Experiments

In the synthetic experiments, the full context $\boldsymbol{Y}_t = (\boldsymbol{S}_t, W_t)$, where $\boldsymbol{S}_t$ denotes the observed context and W_t the unobserved part. The observed context $\{\boldsymbol{S}_t\}_{t\geq 1}$ follows a stationary ARMA(2, 2) process: $\boldsymbol{S}_t = \phi_1 \boldsymbol{S}_{t-1} + \phi_2 \boldsymbol{S}_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$, with $(\phi_1, \phi_2, \theta_1, \theta_2) = (0.75, -0.25, 0.65, 0.35)$ and $\varepsilon_t \sim \mathcal{N}(0, 0.1^2)$. The unobserved context W_t depends on a feature vector $\boldsymbol{x}_t \in \mathbb{R}^2$ summarizing recent context history: $\boldsymbol{x}_t = (1, x_{t,2})^\top$, where $x_{t,2} = (\boldsymbol{S}_t + \boldsymbol{S}_{t-1} + \boldsymbol{S}_{t-2})/3$. We consider two cases of how W_t depends on $\boldsymbol{x}_t$: (a) **Linear**: $W_t = \boldsymbol{\beta}_*^\top \boldsymbol{x}_t + \xi_t$; (b) **Nonlinear**: $W_t = \boldsymbol{\beta}_*^\top \boldsymbol{x}_t + \sin(\rho \cdot x_{t,2}) + \xi_t$, where $\rho = 4$. In both settings, we choose $\boldsymbol{\beta}_* = (0.50, -0.14)$ and $\xi_t \sim \mathcal{N}(0, 0.1^2)$. Finally, the reward R_t follows (2.1) with $\boldsymbol{\Phi}(\boldsymbol{Y}_t, a_t) = (1, \boldsymbol{S}_t, W_t, \boldsymbol{S}_t \cdot a_t)^\top$, $\boldsymbol{\theta}^\star = (0.65, 1.52, -0.23, -0.23)$, and $\eta_t \sim \mathcal{N}(0, 0.05^2)$.

`PULSE-UCB` consists of two phases. In pretraining, a context transition model is learned from $N = 1000$ historical time series of length $T_0 = 100$ to predict the

latent context W_t. In the online evaluation, the agent runs for $T = 1000$ steps. We compare against two benchmarks: (i) `OFUL`, a naive agent that ignores W_t and uses only $\boldsymbol{S}_t$; (ii) `OFUL-Full`, an idealized agent with access to the full context $\boldsymbol{Y}_t = (\boldsymbol{S}_t, W_t)$. The cumulative regret, averaged over 30 independent trials, is shown in Figure 1. As expected, `OFUL-Full` achieves the lowest regret since it observes the full context, while `OFUL` performs worst by ignoring the missing component. In both the linear and nonlinear settings for the missing context, `PULSE-UCB` performs nearly as well as `OFUL-Full`, demonstrating the clear benefit of leveraging a predictive model for the unobserved context. To further demonstrate the performance of our approach, we also compared `PULSE-UCB` with `CLBBF` (Kim et al., 2023). However, because `CLBBF` intrinsically relies on a Missing-At-Random (MAR) assumption, it struggles to handle the more general Missing-Not-At-Random (MNAR) structures present in our setting. Consequently, `PULSE-UCB` consistently outperforms `CLBBF` and achieves near-oracle performance. Additional experimental details, including sensitivity analyses on pre-training data scales (N and T_0), are deferred to Appendix M.1.

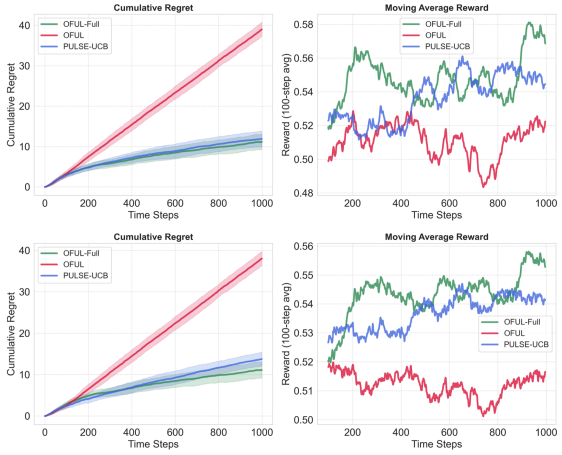

Figure 1: Comparison of algorithms in synthetic experiments. Left: cumulative regret. Right: 100-step moving average reward. Top: linear missing-feature setting (a). Bottom: nonlinear setting (b). Shaded areas denote ± one standard error over 30 trials.

6.2 Real-World Experiments

To evaluate our method in a practical setting, we use the public Taobao Ad Display/Click dataset (Alibaba, 2018), which contains 186,730 advertisement display/click records from Taobao.com. Each record includes 83 features describing user and ad attributes such as gender, age, consumption grade, brand, and category. We embed the features into a 32-dimensional space and partition them into 16 observed features ($\boldsymbol{S}_t$) and 16 unobserved features ($\boldsymbol{W}_t$). All algorithms are evaluated on 80% of the data. For `PULSE-UCB`, we additionally use the remaining 20% for pretraining the context transition model. The action corresponds to selecting an ad (adgroup ID), and the reward is the binary click feedback (1 if clicked, 0 otherwise). Further preprocessing details are deferred to the Appendix N.

We compare `PULSE-UCB` with three baselines: `OFUL`, which ignores the missing context; `OFUL-Full`, which has access to the full context; and `CLBBF` (Kim et al., 2023), designed for bandits with stochastically missing features. We compare these algorithms over $T \approx 1.5 \times 10^5$ steps, with $K = 20$ arms per step, averaging results over 5 runs. Figure 2 shows that `PULSE-UCB` greatly outperforms `OFUL`, highlighting the benefit of context reconstruction, and also surpasses `CLBBF`, whose mechanism struggles under structural missingness. Notably, `PULSE-UCB` achieves performance nearly indistinguishable from the ideal `OFUL-Full`, indicating that the pretraining step not only imputes the missing context but also produces a feature representation well suited for linear bandit learning.

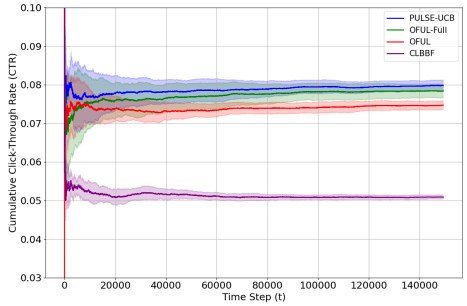

Figure 2: Algorithm comparison on the Taobao dataset. Shaded areas denote ± one standard error over 5 runs.

Discussion and future directions. We proposed a new algorithm, `PULSE-UCB` (Algorithm 1), which leverages imputation models pretrained on historical data to address linear contextual bandits with missing covariates. We established regret guarantees in Theorem 4.1 and showed near-optimality via the lower bound in Theorem 5.1. Empirical results in Section 6 demonstrate strong performance across both synthetic and real-world datasets. Future directions include extending our framework to more general decision-making problems (e.g., Markov decision processes), accommodating more complex missing data mechanisms, and developing adaptive strategies to update the pretrained model during bandit interactions.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96.

Alibaba (2018). Taobao.com dataset.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674.

Bacchiocchi, F., Genalti, G., Maran, D., Mussi, M., Restelli, M., Gatti, N., and Metelli, A. M. (2024). Autoregressive bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 937–945. PMLR.

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

Cai, C., Cai, T. T., and Li, H. (2024). Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232.

Cai, W., Grossman, J., Lin, Z. J., Sheng, H., Wei, J. T.-Z., Williams, J. J., and Goel, S. (2021). Bandit algorithms to personalize educational chatbots. *Machine Learning*, 110(9):2389–2418.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2022). Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*, 28(4):2998–3022.

Cao, J., Gao, R., and Keyvanshokooh, E. (2024). Hr-bandit: Human-ai collaborated linear recourse bandit. *arXiv preprint arXiv:2410.14640*.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564 – 1597.

Chetverikov, D. and Wilhelm, D. (2017). Nonparametric instrumental variable estimation under monotonicity. *Econometrica*, 85(4):1303–1320.

Coughlin, L. N., Campbell, M., Wheeler, T., Rodriguez, C., Florimbio, A. R., Ghosh, S., Guo, Y., Hung, P.-Y., Newman, M. W., Pan, H., et al. (2024). A mobile health intervention for emerging adults with regular cannabis use: A micro-randomized pilot trial design protocol. *Contemporary Clinical Trials*, 145:107667.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Gonzalez, R. T., Abbott, M. R., Nallamothu, B., Hummel, S., Dorsch, M., and Dempsey, W. (2025). Practical considerations when designing an online learning algorithm for an app-based mhealth intervention. *arXiv preprint arXiv:2511.08719*.

Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Number 38. Cambridge University Press.

Guo, Y. and Xu, Z. (2025). Statistical inference for misspecified contextual bandits. *arXiv preprint arXiv:2509.06287*.

He, J., Zhou, D., Zhang, T., and Gu, Q. (2022). Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625.

Hu, H. and Simchi-Levi, D. (2025). Pre-trained ai model assisted online decision-making under missing covariates: A theoretical perspective. *arXiv preprint arXiv:2507.07852*.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120.

Jang, B., Nepper, J., Chevrette, M., Handelsman, J., and Hero, A. O. (2022). High dimensional stochastic linear contextual bandit with missing covariates. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Janner, M., Li, Q., and Levine, S. (2021). Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286.

Kausik, C., Tan, K., and Tewari, A. (2025). Leveraging offline data in linear latent contextual bandits. In *Proceedings of the 2025 International Conference on Machine Learning (ICML)*. ICML 2025 Poster, Published May 1 2025, Last modified July 23 2025.

Kim, J.-H., Yun, S.-Y., Jeong, M., Nam, J., Shin, J., and Combes, R. (2023). Contextual linear bandits under noisy features: Towards bayesian oracles. In *International Conference on Artificial Intelligence and Statistics*, pages 1624–1645. PMLR.

Kim, W., Park, S., Iyengar, G., Zeevi, A., and Oh, M.-h. (2025). Linear bandits with partially observable features. *arXiv preprint arXiv:2502.06142*.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, 105(490):852–863.

Lan, A. S. and Baraniuk, R. G. (2016). A contextual bandits framework for personalized learning action selection. In *EDM*, pages 424–429.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lee, J., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., and Brunskill, E. (2023). Supervised pretraining can learn in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:43057–43083.

Li, K., Yang, Y., and Narisetty, N. N. (2021). Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. *Electronic Journal of Statistics*, 15(2):5652–5695.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Liao, P., Greenewald, K., Klasnja, P., and Murphy, S. (2020). 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.

Lin, L., Bai, Y., and Mei, S. (2023). Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779 – 3821.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2016). Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of behavioral medicine*, pages 1–17.

Park, H. and Faradonbeh, M. K. S. (2022). A regret bound for greedy partially observed stochastic contextual bandits. In *Decision Awareness in Reinforcement Learning Workshop at ICML 2022*.

Park, H. and Faradonbeh, M. K. S. (2024). Thompson sampling in partially observable contextual bandits. *arXiv preprint arXiv:2402.10289*.

Rafferty, A., Ying, H., Williams, J., et al. (2019). Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *Journal of Educational Data Mining*, 11(1):47–79.

Tianhui Cai, T., Namkoong, H., Russo, D., and Zhang, K. W. (2024). Active exploration via autoregressive generation of missing data. *arXiv e-prints*, pages arXiv–2405.

Trella, A. L., Dempsey, W., Gazi, A. H., Xu, Z., Doshi-Velez, F., and Murphy, S. A. (2024). Nonstationary latent auto-regressive bandits. *arXiv preprint arXiv:2402.03110*.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Williams, J., Li, N., Kim, J., Whitehill, J., Maldonado, S., Pechenizkiy, M., Chu, L., and Heffernan, N. (2014). The mooclet framework: Improving online education through experimentation and personalization of modules. *Available at SSRN 2523265*.

Xia, E. and Wainwright, M. J. (2024). Prediction aided by surrogate training. *arXiv preprint arXiv:2412.09364*.

Xu, X., Xie, H., and Lui, J. C. (2021). Generalized contextual bandits with latent features: Algorithms and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4763–4775.

Ye, Z., Yoganarasimhan, H., and Zheng, Y. (2025). Lola: Llm-assisted online learning algorithm for content experiments. *Marketing Science*.

Zeng, S., Bhatt, S., Koppel, A., and Ganesh, S. (2025). Partially observable contextual bandits with linear payoffs. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhang, K. W., Cai, T. T., Namkoong, H., and Russo, D. (2025). Contextual thompson sampling via generation of missing data. *arXiv preprint arXiv:2502.07064*.

Zhang, K. W., Closser, N., Trella, A. L., and Murphy, S. A. (2024). Replicable bandits for digital health interventions. *arXiv preprint arXiv:2407.15377*.

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: Our problem formulation, theoretical assumptions, and the proposed `PULSE-UCB` algorithm are detailed in Sections 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   Yes: We provide a theoretical analysis of the regret bound in Section 4 and 5.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   Yes: We provide anonymized source code for all our experiments in the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   Yes: All assumptions for our theoretical results are formally stated in Sections 2-5.
   (b) Complete proofs of all theoretical results.
   Yes: Complete proofs for all theorems and lemmas are provided in Appendix.
   (c) Clear explanations of any assumptions.
   Yes: We have provided intuitions and discussions for all our key assumptions.

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: The code and instructions to reproduce all experimental results are in the supplementary material. For the real-world experiment, our code is based on the public Taobao dataset.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   Yes: We provide all training details, including data splits and hyperparameter settings, in the paper (Section 6.1 and 6.2) as well as the Appendix.
   (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: All our results are over multiple independent runs and shown with error bars. All shaded areas denote $\pm$ one standard error over multiple runs.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   No: Our experiments are computationally inexpensive and were run on a standard desktop computer; they do not require any specialized computing infrastructure.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   Yes: We cite the source of the Taobao dataset in Section 6.2.
   (b) The license information of the assets, if applicable.
   Yes: This Taobao dataset is provided by Alibaba and distributed under the CC BY-NC-SA 4.0 license.

(c) New assets either in the supplemental material or as a URL, if applicable.
Not Applicable: We use an existing public dataset and do not release new assets.

(d) Information about consent from data providers/curators.
Not Applicable: We use a well-established public dataset that has been anonymized.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
Not Applicable: The dataset is anonymized and does not contain personally identifiable or offensive content.

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: Our research does not involve human subjects or crowdsourcing.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
Not Applicable: Our research does not involve human subjects.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
Not Applicable: Our research does not involve human subjects.

Contents

1 INTRODUCTION **1**
- 1.1 Related works . . . 2

2 PROBLEM SETUP **2**

3 THE PULSE-UCB ALGORITHM **4**

4 REGRET ANALYSIS **4**
- 4.1 Characterizing imputation error . . . 4
- 4.2 Regret analysis under general context distributions . . . 5
- 4.3 Application of Theorem 4.1 . . . 7

5 LOWER BOUNDS **8**

6 NUMERICAL EXPERIMENTS **8**
- 6.1 Synthetic Experiments . . . 8
- 6.2 Real-World Experiments . . . 9

Appendix **14**

A Additional literature review **15**

B Practical Motivations for Asymmetric Context Availability **16**

C Discussion of Assumption 4.1 and Autoregressive Dependence **16**

D A data-driven approach for choosing D_t **17**

E Proof of Lemma 4.1 **19**

F Proof of Lemma 4.2 **19**

G Proof of Theorem 4.1 **20**
- G.1 Technical Lemmas . . . 24

H Proof of Results in Section 4.3 **25**
- H.1 Proof of Proposition 4.1 . . . 25
- H.2 Proof of Proposition 4.2 . . . 26
- H.3 Robustness of Propositions 4.1 and 4.2 to Covariate Shift . . . 27

I Extension: A Confidence-Weighted Variant of PULSE-UCB **27**
- I.1 Algorithm Design: Robust PULSE-UCB . . . 28
- I.2 Theoretical Guarantees and Regret Comparison . . . 28

J Setup of the Lower Bound **30**

K Proof of Theorem J.1 **33**

K.1 Proof of Technical Lemmas . . . 35
K.1.1 Proof of Lemma K.1 . . . 35
K.1.2 Proof of Lemma K.2 . . . 36
K.1.3 Proof of Lemma K.3 . . . 37
K.1.4 Proof of Lemma K.4 . . . 38
K.1.5 Proof of Lemma K.5 . . . 39

L Derivation of Equivalent Formulations for UCB Exploration in Algorithm 1 **43**
L.1 Implication for Adaptive Exploration . . . 44

M Synthetic Data Experiments **45**
M.1 Extra Baseline Comparisons . . . 45
M.1.1 Sensitivity to Pre-training Data Scale N and T_0 . . . 45
M.1.2 Expanded Comparison with CLBBF . . . 45
M.2 Impact of Smoothness . . . 46
M.3 Robustness under Time-Varying Observation Patterns . . . 47
M.4 Robustness to Distributional and Covariate Shifts . . . 48
M.4.1 Impact of Covariate Variance Shift . . . 48
M.4.2 Rotation: Structural Shift via AR Perturbation . . . 48
M.4.3 Translation: Sensitivity to Location Shift . . . 49
M.4.4 Deformation: Sensitivity to Geometric Shape . . . 49

N Related Details about Real Dataset Experiments **50**
N.1 Dataset and Preprocessing . . . 50
N.2 Dimensionality Reduction via Autoencoder . . . 50
N.3 Partially Observed Setting and Inference Model . . . 50
N.4 Online Evaluation Protocol . . . 51

A Additional literature review

AI-assisted decision-making. Recently, there has been growing interest in applying AI, including foundation models, to enhance decision-making. For example, Tianhui Cai et al. (2024); Zhang et al. (2025) design Thompson sampling algorithms for online bandits that treat uncertainty as missing future outcomes, imputing them with pretrained generative models to optimize policies. Chen et al. (2021); Janner et al. (2021) recast offline reinforcement learning as sequence modeling over trajectories to improve decisions, while Lin et al. (2023); Lee et al. (2023) study in-context reinforcement learning, showing that supervised pretraining on past trajectories enables models to approximate algorithms such as LinUCB and Thompson sampling with regret guarantees. Applications include LLM-assisted adaptive experimentation for content delivery (Ye et al., 2025) and humanAI collaboration in linear bandits with resource constraints for healthcare (Cao et al., 2024). Our work focuses on the missing-context issue, specifically in online contextual bandits, and provides near-optimal regret guarantees that guide the use of pretrained models for imputation in this setting.

Imputation in statistics and ML. Imputation has long been a central strategy across statistics and machine learning for handling missing information. A classical example is the EM algorithm (Dempster et al., 1977), which provides a likelihood-based framework for parameter estimation with incomplete data and remains highly influential in this area. In causal inference, imputation is widely used for estimating potential outcomes under counterfactual interventions (Little and Rubin, 2019). From a statistical learning perspective, recent work has incorporated modern machine learning models to deal with missing responses: Xia and Wainwright (2024)

propose surrogate training that leverages helper covariates to impute pseudo-responses for unlabeled data, yielding prediction improvements with excess risk guarantees, while Angelopoulos et al. (2023) demonstrate that combining a small labeled set with imputed outcomes enables valid confidence intervals and hypothesis tests. Our work contributes to this line of research by extending imputation-based methods to sequential decision-making with partially observed contexts, and quantify the impact of imputation quality on online learning performance.

B Practical Motivations for Asymmetric Context Availability

A core premise of our framework is that the auxiliary offline dataset $\mathcal{D}$ contains richer or less noisy contextual information (Y_t) than what can be feasibly observed during real-time online decision-making (S_t). This asymmetric context availability is driven by how data are collected in practice. In many real-world environments, online intervention systems must operate under strict privacy constraints, tight computational limits, or high risks of missing data. Conversely, offline datasets are often collected from prior clinical trials, extensive observational cohorts, or archival systems, affording to store and process much richer state variables.

Rather than assuming the online agent can access a complete context capturing all factors that affect the reward, our framework aims for *context enrichment*: leveraging the rich contextual variables in the auxiliary data to build a more accurate reward approximation offline, which in turn improves the decisions made online using the limited context. We illustrate this setting with two concrete domains:

Digital Health. In mobile health (mHealth) and digital interventions, multiple sources of context-rich offline data are common, such as prior trials, earlier interventions, and observational cohorts equipped with detailed baseline surveys or high-burden ecological momentary assessments (Liao et al., 2020; Coughlin et al., 2024). By contrast, online intervention systems running on local devices face tight bandwidth limits, battery constraints, and strict privacy regulations. Furthermore, user disengagement and frequent sensor failures often lead to missing or coarsened context in practice (Gonzalez et al., 2025). In our setting, the offline full state Y_t represents these rich baseline variables, low-frequency assessments (e.g., stress assessed via validated multi-item scales), and high-resolution wearable signals (e.g., heart-rate variability). The online observed state S_t, however, is typically restricted to coarser, continuously available signals like step counts or simple self-reports.

Online Education. In intelligent tutoring systems and online education platforms, the online bandit algorithm typically only observes a limited set of immediate signals S_t, such as short performance summaries or recent clickstreams. However, educational platforms maintain rich archival databases containing concept-level mastery estimates, comprehensive exam scores, and detailed background information from prior assessments (Lan and Baraniuk, 2016). For instance, in large-scale MOOCs, frameworks like MOOClet utilize archival data (e.g., full homework records and fine-grained logs) to model detailed student states and design adaptive components (Williams et al., 2014). Within our framework, these archival records constitute the auxiliary data $\mathcal{D}$, where Y_t aggregates the richer, knowledge-related quantities inferred offline, while the online agent relies solely on the high-frequency, easily observable signals S_t.

Robustness to Distribution Shift. While our primary theoretical analysis focuses on the matched-distribution regime where the auxiliary dataset and the online environment share the same context distribution, the learned latent structure often generalizes robustly when this assumption is relaxed. As demonstrated empirically in Appendix M.4, our method maintains strong performance even when the online environment experiences significant covariate shifts relative to the offline auxiliary data. A theoretical justification for this robustness, showing that the rates in Propositions 4.1 and 4.2 are preserved under mild covariate shift, is given in Appendix H.3.

C Discussion of Assumption 4.1 and Autoregressive Dependence

Assumption 4.1 requires that $\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, \boldsymbol{S}_{1:t}] = \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}]$. As noted in the main text, this is satisfied whenever $\boldsymbol{Y}_t$ is generated as an arbitrary function of $\boldsymbol{S}_{1:t}$ together with independent randomness,

because the fresh noise at time t carries no information about the past beyond what is already captured by $\boldsymbol{S}_{1:t}$. This covers a broad class of models, including all i.i.d. contextual bandit settings as a special case.

The assumption fails, however, when the noise driving $\boldsymbol{Y}_t$ is itself temporally dependent, as in autoregressive (AR) models. In such cases, past noise innovations $\xi_{1:t-1}$ influence both $\boldsymbol{Y}_{1:t-1}$ and the current $\boldsymbol{Y}_t$, so that conditioning on $\boldsymbol{S}_{1:t}$ alone no longer screens off the dependence on the past. The residual ε_t defined in Equation (4.5) then fails to form a martingale difference sequence, and the self-normalized concentration machinery underlying our regret analysis breaks down. Handling ε_t in this regime requires fundamentally different tools. Moreover, existing works on AR bandits suggest that the appropriate techniques depend heavily on the specific form of AR structure, making it difficult to develop a unified treatment. We illustrate this point with two representative examples below.

First, suppose we ignore the observed context $\boldsymbol{S}_{1:t}$ and let the latent state evolve via action-dependent dynamics

$$W_{t+1} = \mu(a_t) + \rho(a_t)\, W_t + \xi_t,$$

where ξ_t is i.i.d. noise. The reward satisfies $R(t, a_t) = \boldsymbol{\theta}^{\star\top} \boldsymbol{\Phi}(\boldsymbol{Y}_t, a_t) + \widetilde{\eta}_t$, giving a model whose transition coefficients depend on the agent's action. This is the action-driven AR bandit studied by Bacchiocchi et al. (2024), for which modified UCB-type algorithms exploit the action-dependent structure to achieve the optimal $\mathcal{O}(\sqrt{T})$ regret. Crucially, however, this structure is absent in our problem: the auxiliary historical data contain no action information, and the contextual process evolves according to fixed, action-independent dynamics. Techniques designed for action-driven AR dependence therefore do not apply.

Second, consider a form of AR dependence more representative of our setting. Let

$$W_t = \alpha\, W_{t-1} + \xi_t, \qquad \xi_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad |\alpha| < 1,$$

and set $\boldsymbol{Y}_t = W_t$ with feature map $\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) = \beta_a W_t$. The reward becomes $R(t, a) = \boldsymbol{\theta}^\star\, \beta_a\, W_t + \eta_t$. Here the AR dependence resides in a latent process that influences the reward only multiplicatively through the action, producing a fundamentally different statistical structure from the first example. This setting corresponds to the latent autoregressive bandit model of Trella et al. (2024), where the analysis requires Kalman-filter-type estimators and strong linear-Gaussian assumptions, and the resulting regret may exceed $\mathcal{O}(\sqrt{T})$.

These two examples illustrate that it is currently not straightforward to propose a unified approach for handling AR dependence in bandit problems. The difficulty of the learning problem hinges critically on *how* autoregressive dependence enters the model, and seemingly minor modeling choices, such as whether the AR dependence acts on the observed reward or on a latent state, and whether it is action-driven or governed by fixed dynamics, lead to qualitatively different algorithms and regret guarantees. Furthermore, both examples above assume that the feature map $\boldsymbol{\Phi}$ is linear in the contextual variables; once nonlinear dependencies are introduced, it is unclear how to adapt existing techniques or whether the same regret rates remain achievable.

Taken together, these observations underscore that extending our framework to accommodate autoregressive noise would require a case-by-case analysis tied to the specific AR structure at hand, rather than a single unified extension of Assumption 4.1. We therefore leave this direction to future work.

D A data-driven approach for choosing D_t

For any $t \in [T]$, recall that Algorithm 1 requires an upper bound D_t to calibrate the confidence balls. In Remark 4.4 and Section 4.3 we described several cases where an explicit rate of D_t is available. Here we discuss a fully data-driven alternative based on *uniform confidence bands* (UCBs) for the conditional mean under the ground-truth conditional law.

Let p denote the ground-truth conditional distribution of $\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t}$ and let $\widehat{p}$ be an estimator of this conditional distribution built from historical data. Fix an action $a \in \mathcal{A}$. Write

$$\mu_p(\boldsymbol{s}) \;:=\; \mathbb{E}[\Phi(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}]\,, \qquad \mu_{\widehat{p}}(\boldsymbol{s}) \;:=\; \mathbb{E}_{\widehat{p}}[\Phi(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}]\,,$$

where we suppose $\Phi(\boldsymbol{Y}_t, a)$ to be one-dimensional for clarity (the multivariate case follows coordinatewise with a union adjustment). Denote our imputation error at $\boldsymbol{S}_{1:t} = \boldsymbol{s}$ as

$$\mathcal{E}_t(p, \widehat{p}; \boldsymbol{s}) \;:=\; \big|\mu_p(\boldsymbol{s}) - \mu_{\widehat{p}}(\boldsymbol{s})\big|.$$

Instead of bounding D_t, it suffices for us to control

$$\mathcal{E}_t(p, \widehat{p}; \boldsymbol{s})$$

for every $\boldsymbol{s}$ simultaneously over a compact domain $\mathcal{S}_t$ where $\boldsymbol{S}_{1:t}$ is in.

We upper bound $\mathcal{E}_t(p, \widehat{p}; \boldsymbol{s})$ by combining (i) a *uniform confidence band* for $\mu_p(\boldsymbol{s})$, centered at a reference estimator that *does* admit UCBs, and (ii) a directly computable discrepancy between $\mu_{\widehat{p}}$ and that reference estimator. Concretely, split the historical data into two folds I_0 and I_1 (sample splitting or cross-fitting):

1. On I_0, fit a reference conditional distribution $\widehat{p}_0$ using a method with established UCBs (e.g., local-polynomial with robust bias correction or penalized splines with simultaneous bands). Obtain a $(1-\alpha)$-UCB for μ_p on a grid $\mathcal{G}_t \subset \mathcal{S}_t$,
$$\mathcal{C}_{1-\alpha}(\boldsymbol{s}) \;=\; \big[\mu_{\widehat{p}_0}(\boldsymbol{s}) \pm r_{1-\alpha}(\boldsymbol{s})\big], \qquad \boldsymbol{s} \in \mathcal{G}_t,$$
where $r_{1-\alpha}(\boldsymbol{s})$ is the half-width delivered by the band construction.

2. On I_1, fit $\widehat{p}$ (any estimation strategy; no UCB requirement).

3. By the triangle inequality, for any $\boldsymbol{s} \in \mathcal{G}_t$,
$$\big|\mu_p(\boldsymbol{s}) - \mu_{\widehat{p}}(\boldsymbol{s})\big| \;\le\; \underbrace{\big|\mu_p(\boldsymbol{s}) - \mu_{\widehat{p}_0}(\boldsymbol{s})\big|}_{\text{controlled by the UCB}} \;+\; \underbrace{\big|\mu_{\widehat{p}_0}(\boldsymbol{s}) - \mu_{\widehat{p}}(\boldsymbol{s})\big|}_{\text{fully data-computable}}. \tag{D.1}$$

Taking suprema over $\mathcal{G}_t$ and, if desired, extending from the grid to $\mathcal{S}_t$ with a modulus-of-continuity bound yields a valid high-probability bound for $\sup_{\boldsymbol{s}\in\mathcal{S}_t} \mathcal{E}_t(p, \widehat{p}; \boldsymbol{s})$. Under certain regularity conditions, one can extend the bound over the grid $\mathcal{G}_t$ to the entire domain $\mathcal{S}_t$ at the cost of a discretization penalty. In practice, when the grid $\mathcal{G}_t$ is chosen sufficiently fine, this additional term becomes negligible, and one may safely restrict attention to $\mathcal{G}_t$ without loss of generality.

Suppose $\mathcal{C}_{1-\alpha}$ is a $(1-\alpha)$ uniform confidence band for μ_p over $\mathcal{G}_t$ centered at $\mu_{\widehat{p}_0}$ (constructed on I_0), i.e.,

$$\mathbb{P}\big\{\mu_p(\boldsymbol{s}) \in \mathcal{C}_{1-\alpha}(\boldsymbol{s}),\ \forall\, \boldsymbol{s} \in \mathcal{G}_t\big\} \ge 1 - \alpha.$$

Then with probability at least $1-\alpha$,

$$\mathcal{E}_t(p, \widehat{p}; \boldsymbol{s}) \;\le\; \sup_{\boldsymbol{s}\in\mathcal{G}_t} r_{1-\alpha}(\boldsymbol{s}) \;+\; \big|\mu_{\widehat{p}_0}(\boldsymbol{s}) - \mu_{\widehat{p}}(\boldsymbol{s})\big|, \tag{D.2}$$

yielding a data-driven choice for $\mathcal{E}_t(p, \widehat{p}; \boldsymbol{s})$ on $\mathcal{G}_t$.

For practical purpose, one can follow the procedure below:

1. **UCB machinery for the reference fit $\widehat{p}_0$.** Two widely used choices are: (i) local-polynomial regression with *robust bias correction* (RBC), whose studentized process admits valid simultaneous bands and is robust to MSE-optimal bandwidth choice; the quantiles are obtained via Gaussian/multiplier bootstrap of the sup-statistic; (ii) penalized splines with simultaneous bands via volume-of-tube or bootstrap calibrations.[3]

2. **Computing $\mu_{\widehat{p}_0}$ and $\mu_{\widehat{p}}$.** For general Φ (which is known), approximate $\mathbb{E}_{\widehat{p}}[\Phi(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}]$ by Monte Carlo from $\widehat{p}$ (and analogously for $\widehat{p}_0$) with negligible simulation error relative to the statistical half-widths.

[3] See, e.g., Calonico et al. (2018, 2022) for RBC-based bands and Krivobokova et al. (2010) for spline bands; multiplier bootstrap for suprema is treated in Chernozhukov et al. (2014).

3. **From grid to domain.** Choose $\mathcal{G}_t$ fine enough relative to the smoothing scale (e.g., grid spacing $\ll$ bandwidth) and, if needed, add the modulus-of-continuity correction to pass from a grid-wide error bound to a domain-wide error bound.

4. **Coordinatewise or joint control (multi-dimensional $\boldsymbol{\Phi}$).** Apply the above per coordinate and combine by Bonferroni (conservative), or calibrate a joint supremum over coordinates via multiplier bootstrap of a vector-valued process.

As a special case, if the estimator used for $\widehat{p}$ provides a valid UCB centered at $\mu_{\widehat{p}}$, one may set $\widehat{p}_0 = \widehat{p}$ and simply take $\sup_{\boldsymbol{s}\in\mathcal{G}_t} r_{1-\alpha}(\boldsymbol{s})$. RBC-based local polynomials are particularly convenient here because the same fit supplies both the point estimates and a simultaneous band with good finite-sample coverage properties.

Caveat (high-dimensional context $\boldsymbol{S}_{1:t}$). When the observed context $\boldsymbol{S}_{1:t} = \boldsymbol{s}$ lies in a high-dimensional space, it is generally impossible to obtain tight confidence bounds without additional structural assumptions. Specifically, nonparametric estimators suffer from the curse of dimensionality, causing inflated confidence bands and consequently large $\widehat{D}_t$. This reflects a fundamental limitation of nonparametric inferencewithout further assumptions, nontrivial guarantees cannot be achieved in the worst case. To address this issue, one may collect substantially more data or impose structural restrictions that effectively reduce the intrinsic dimension, such as additivity (Meier et al., 2009), single-index models (Ichimura, 1993), or shape constraints (Groeneboom and Jongbloed, 2014; Chetverikov and Wilhelm, 2017).

E Proof of Lemma 4.1

Lemma E.1. Under Assumption 4.1, for all $t \in [T]$,

$$\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid A_{1:t-1}, R_{1:t-1}, \boldsymbol{S}_{1:t}\right] = \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}\right].$$

Proof. Fixing $\boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}$, we have

$$\begin{aligned}
&\sup_{g:\|g\|_\infty\le 1} \left\{\mathbb{E}\left[g(\boldsymbol{Y}_t) \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right] - \mathbb{E}_{\widehat{p}}\left[g(\boldsymbol{Y}_t) \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right]\right\} \\
&\overset{(i)}{=} d_{\mathrm{TV}}\left(\mathbb{P}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right), \widehat{\mathbb{P}}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right)\right) \\
&\overset{(ii)}{\le} \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathbb{P}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right) \| \widehat{\mathbb{P}}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right)\right)}
\end{aligned}$$

where (i) holds by the definition of total variation, (ii) holds by Pinsker's inequality. Taking expectation with respect to $\boldsymbol{S}_{1:t}$ on both sides of the above display and applying Jensen's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{S}_{1:t}} \sup_{g:\|g\|_\infty\le 1} \left\{\mathbb{E}\left[g(\boldsymbol{Y}_t) \mid \boldsymbol{S}_{1:t}\right] - \mathbb{E}_{\widehat{p}}\left[g(\boldsymbol{Y}_t) \mid \boldsymbol{S}_{1:t}\right]\right\} \\
&\le \sqrt{\frac{1}{2}\mathbb{E}\,\mathrm{KL}\left(\mathbb{P}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right) \| \widehat{\mathbb{P}}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t} = \boldsymbol{s}_{1:t}\right)\right)} \\
&= \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathbb{P}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t}\right) \| \widehat{\mathbb{P}}\left(\boldsymbol{Y}_t \mid \boldsymbol{S}_{1:t}\right)\right)}
\end{aligned}$$

where the last equality follows from the definition of KL divergence between conditional distributions. ∎

F Proof of Lemma 4.2

Proof. Recall that for any $t \in [T]$,

$$\mathcal{G}_{t-1} = \sigma\left(\boldsymbol{S}_{1:t}, \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, U_{1:t}\right).$$

We remind the reader that U_t is a auxiliary random variable used to select A_t (e.g., U_t captures the randomness involved in algorithms such as Thompson sampling or in breaking ties when selecting actions). and U_t is

independent of $(\boldsymbol{S}_{1:t}, R_{1:t-1}, A_{1:t-1})$. To verify that $\{\varepsilon_t\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{G}_t\}_{t=1}^T$, we need to verify two conditions, namely

$$\varepsilon_t \in \mathcal{G}_t, \tag{F.1}$$

and

$$\mathbb{E}\left[\varepsilon_t \mid \mathcal{G}_{t-1}\right] = 0. \tag{F.2}$$

Noting that for all $\tau \in [t-1]$,

- A_τ is a function of the observed history and the auxiliary random variable $(\boldsymbol{S}_{1:\tau}, R_{1:\tau-1}, A_{1:\tau-1}, U_\tau)$.
- R_τ is a function of A_τ, $\boldsymbol{Y}_\tau$ and η_τ.

We conclude from the above observation that $(A_{1:t-1}, R_{1:t-1})$ is a function of $(\boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, U_{1:t-1})$ and

$$A_t \in \sigma\left(\boldsymbol{S}_{1:t}, \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, U_{1:t}\right). \tag{F.3}$$

Thus, we have

$$\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \in \sigma\left(\boldsymbol{Y}_t, A_t\right) \subset \sigma\left(\boldsymbol{Y}_{1:t}, \eta_{1:t}, U_{1:t+1}\right) \subset \mathcal{G}_t$$

and Equation (F.1) holds. For Equation (F.2) to hold, we have

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \mathcal{G}_{t-1}\right] &= \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, U_{1:t}\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, U_{1:t}, A_t\right] \\
&\overset{(ii)}{=} \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, \boldsymbol{Y}_{1:t-1}, \eta_{1:t-1}, A_t\right]
\end{aligned}$$

where equality (i) holds from Equation (F.3) and equality (ii) follows from $\boldsymbol{Y}_t$ is independent of $U_{1:t}$. Applying Assumption 4.1 with the above display, it follows that

$$\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \mathcal{G}_{t-1}\right] = \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t\right].$$

It is then straightforward to see that Equation (F.2) holds by the definition of ε_t. We then conclude that $\{\varepsilon_t\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{G}_t\}_{t=1}^T$

Additionally, we show that $\{\varepsilon_t\}_{t=1}^T$ satisfies a sub-Gaussian tail condition, we only need to verify that it is a bounded sequence. Since for any $a \in \mathcal{A}$, under Assumption 2.2 and Equation (2.2),

$$\boldsymbol{\theta}^{\star\top} \boldsymbol{\Phi}(\boldsymbol{Y}_t, a) = \mathbb{E}\left[R(t, a) \mid \mathcal{F}_t\right] \in [-1, 1],$$

it follows that

$$|\varepsilon_t| \leq \left|\boldsymbol{\theta}^{\star\top} \boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t)\right| + \left|\boldsymbol{\theta}^{\star\top} \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t\right]\right| \leq 2.$$

By Azuma-Hoeffding inequality (see Corollary 2.20 in Wainwright, 2019), we conclude that $\{\varepsilon_t\}_{t=1}^T$ is a martingale difference sequence with sub-Gaussian parameter $\sigma_\varepsilon^2 \leq 4$. ■

G Proof of Theorem 4.1

Proof. Before presenting the proof, we briefly outline the main steps. We first assume that for all $t \in [T]$, $\boldsymbol{\theta}^\star \in \texttt{BALL}_{t-1}$, where $\texttt{BALL}_{t-1}$ is the confidence ball at step $t-1$ defined in Equation (3.5). Under this assumption, we show that the regret at each step $t \in [T]$ can be decomposed into two components (see Equation (G.9)): the first reflecting the width of the confidence ellipsoid in the direction of the chosen decision $\widehat{\boldsymbol{\phi}}_{t,A_t}$, and the second capturing the imputation error. The former is bounded in Lemma G.1, while the latter is controlled by Equation (G.8). Finally, we select an appropriate sequence $\{\gamma_t\}_{t\in[T]}$ to guarantee that $\boldsymbol{\theta}^\star \in \texttt{BALL}_t$ with high probability.

Recall that $\widehat{\boldsymbol{\phi}}_{t,a}$ is the conditional expectation of the context $\boldsymbol{\Phi}(\boldsymbol{Y}_t, a)$ given the partial observation $\boldsymbol{S}_{1:t}$ under distribution $\widehat{p}$, as defined in Equation (3.1). Let $\bar{\boldsymbol{\theta}}_t \in \texttt{BALL}_{t-1}$ denote the vector which maximizes the inner product $\boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,A_t}$. Then

$$\bar{\boldsymbol{\theta}}_t^\top \widehat{\boldsymbol{\phi}}_{t,A_t} = \max_{\boldsymbol{\theta}\in\texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,A_t} = \max_{a\in\mathcal{A}} \max_{\boldsymbol{\theta}\in\texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,a} \tag{G.1}$$

where the last equality follows from the way we choose A_t as defined in Equation (3.2). Recall that $A_t^\star$ is the optimal action given by Equation (2.4). The right-hand side of Equation (G.1) is lower bounded by

$$\max_{a\in\mathcal{A}} \max_{\boldsymbol{\theta}\in\texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,a} \geq \max_{\boldsymbol{\theta}\in\texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,A_t^\star} \geq \boldsymbol{\theta}^{\star\top} \widehat{\boldsymbol{\phi}}_{t,A_t^\star}$$

Adding and subtracting $\boldsymbol{\theta}^{\star\top}\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star]$ on the right-hand side of the above display yields

$$\begin{aligned}\max_{a\in\mathcal{A}} \max_{\boldsymbol{\theta}\in\texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \widehat{\boldsymbol{\phi}}_{t,a} &\geq \boldsymbol{\theta}^{\star\top} \widehat{\boldsymbol{\phi}}_{t,A_t^\star} - \boldsymbol{\theta}^{\star\top}\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star\right] + \boldsymbol{\theta}^{\star\top}\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star] \\ &= \boldsymbol{\theta}^\star \left(\widehat{\boldsymbol{\phi}}_{t,A_t^\star} - \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star\right]\right) + \boldsymbol{\theta}^{\star\top}\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star].\end{aligned}$$

Taking the above display into Equation (G.1) gives

$$\bar{\boldsymbol{\theta}}_t^\top \widehat{\boldsymbol{\phi}}_{t,A_t} \geq \boldsymbol{\theta}^{\star\top}(\widehat{\boldsymbol{\phi}}_{t,A_t^\star} - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star]) + \boldsymbol{\theta}^{\star\top}\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star].$$

Rearranging the above display, we have

$$\boldsymbol{\theta}^{\star\top}\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star\right] \leq \bar{\boldsymbol{\theta}}_t^\top \widehat{\boldsymbol{\phi}}_{t,A_t} - \boldsymbol{\theta}^{\star\top}\left(\widehat{\boldsymbol{\phi}}_{t,A_t^\star} - \mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}, A_t^\star\right]\right). \tag{G.2}$$

Therefore, for reg_t as defined in Equation (4.6)

$$\begin{aligned}\mathrm{reg}_t &= \mathbb{E}\left[R(t, A_t^\star) - R(t, A_t) \mid \boldsymbol{S}_{1:t}\right] \\ &= \boldsymbol{\theta}^{\star\top}\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}] - \boldsymbol{\theta}^{\star\top}\mathbb{E}\left[\widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] + \boldsymbol{\theta}^{\star\top}\mathbb{E}\left[\widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] - \boldsymbol{\theta}^{\star\top}\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}] \\ &\overset{(i)}{\leq} \mathbb{E}\left[\left(\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right)^\top \widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] - \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}\left[\widehat{\boldsymbol{\phi}}_{t,A_t^\star} \mid \boldsymbol{S}_{1:t}\right] - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}]\right) \\ &\quad + \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}\left[\widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}]\right) \\ &\overset{(ii)}{=} \mathbb{E}\left[\left(\bar{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_t\right)^\top \widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] + \mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right)^\top \widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] \\ &\quad - \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}\left[\widehat{\boldsymbol{\phi}}_{t,A_t^\star} \mid \boldsymbol{S}_{1:t}\right] - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t^\star) \mid \boldsymbol{S}_{1:t}]\right) + \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}\left[\widehat{\boldsymbol{\phi}}_{t,A_t} \mid \boldsymbol{S}_{1:t}\right] - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}]\right).\end{aligned} \tag{G.3}$$

where inequality (i) follows from Equation (G.2) and equality (ii) follows from adding and subtracting $\widehat{\boldsymbol{\theta}}_t^\top \widehat{\boldsymbol{\phi}}_{t,A_t}$, where $\widehat{\boldsymbol{\theta}}_t$ is defined in Equation (3.4).

Recall $\boldsymbol{\Sigma}_t$ defined in Equation (3.3). We claim that for any $\boldsymbol{\theta} \in \texttt{BALL}_{t-1}$ and any $\boldsymbol{\phi} \in \mathbb{R}^d$,

$$\left|(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)^\top \boldsymbol{\phi}\right| \leq \sqrt{\gamma_t \boldsymbol{\phi}^\top \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\phi}}. \tag{G.4}$$

To see this, by Cauchy-Schwarz inequality, we have

$$\left|(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)^\top \boldsymbol{\phi}\right| = \left|(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t)^\top \boldsymbol{\Sigma}_t^{1/2} \boldsymbol{\Sigma}_t^{-1/2} \boldsymbol{\phi}\right| \leq \left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_t\right\|_{\boldsymbol{\Sigma}_t} \|\boldsymbol{\phi}\|_{\boldsymbol{\Sigma}_t^{-1}} \leq \sqrt{\gamma_t \boldsymbol{\phi}^\top \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\phi}},$$

where the last inequality follows from the fact that $\boldsymbol{\theta} \in \texttt{BALL}_{t-1}$ and the choice of γ_t in Equation (3.5). Applying the above display with $\boldsymbol{\theta} \in \{\boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}_t\}$ and $\boldsymbol{\phi} = \widehat{\boldsymbol{\phi}}_{t,A_t}$ yields

$$\left|(\bar{\boldsymbol{\theta}}_t - \widehat{\boldsymbol{\theta}}_t)^\top \widehat{\boldsymbol{\phi}}_{t,A_t}\right| + \left|(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star)^\top \widehat{\boldsymbol{\phi}}_{t,A_t}\right| \leq 2\sqrt{\gamma_t \widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1} \widehat{\boldsymbol{\phi}}_{t,A_t}}. \tag{G.5}$$

Let

$$\xi_{1,t} := \min\left\{\sqrt{\gamma_t \widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1} \widehat{\boldsymbol{\phi}}_{t,A_t}}, 1\right\}. \tag{G.6}$$

For any $a \in \mathcal{A}$ and $t \in [T]$, let

$$\xi_{2,t} = \max_{a\in\mathcal{A}} \left|\boldsymbol{\theta}^{\star\top}\left(\widehat{\boldsymbol{\phi}}_{t,a} - \mathbb{E}\left[\boldsymbol{\Phi}\left(\boldsymbol{Y}_t, a\right) \mid \boldsymbol{S}_{1:t}\right]\right)\right|. \tag{G.7}$$

We have

$$\begin{aligned}\xi_{2,t} \le \max_{a\in\mathcal{A}} \left|\boldsymbol{\theta}^{\star\top}\left(\widehat{\boldsymbol{\phi}}_{t,a} - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}]\right)\right| &\le \|\boldsymbol{\theta}^\star\|_2 \max_{a\in\mathcal{A}} \left\|\widehat{\boldsymbol{\phi}}_{t,a} - \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \mid \boldsymbol{S}_{1:t}]\right\|_2 \\ &\le \sqrt{dD_t}\end{aligned} \tag{G.8}$$

where the first inequality follows from Cauchy-Schwarz inequality and the last inequality follows from the assumption that $\|\boldsymbol{\theta}^\star\|_2 \le 1$ and Equation (4.3).

Taking Equations (G.5), (G.6) and (G.7) into Equation (G.3), we have

$$|\mathrm{reg}_t| = \min\left\{|\mathrm{reg}_t|, 1\right\} \le 2\mathbb{E}\left[\xi_{1,t} \mid \boldsymbol{S}_{1:t}\right] + 2\xi_{2,t}, \tag{G.9}$$

where the first equality follows from the assumption that $R(t,a) \in [-1,1]$ for any $a \in \mathcal{A}$. Summing Equation (G.9) over $t \in [T]$ gives

$$\begin{aligned}\sum_{t=1}^T \mathrm{reg}_t &\le 2\sum_{t=1}^T \mathbb{E}\left[\xi_{1,t} \mid \boldsymbol{S}_{1:t}\right] + 2\sum_{t=1}^T \xi_{2,t} \\ &\overset{(i)}{\le} 2\sqrt{T\sum_{t=1}^T \mathbb{E}\left[\xi_{1,t}^2 \mid \boldsymbol{S}_{1:t}\right]} + 2\sum_{t=1}^T \xi_{2,t} \\ &\overset{(ii)}{\le} 2\sqrt{2T\gamma_T d \log\left(1 + \frac{TB^2}{d\lambda}\right)} + 2\sum_{t=1}^T \sqrt{dD_t}\end{aligned} \tag{G.10}$$

where inequality (i) follows from Cauchy-Schwarz inequality and inequality (ii) follows from Equation (G.18) in Lemma G.1 and Equation (G.8).

It remains to choose a sequence of suitable $\{\gamma_t\}_{t=1}^T$ so that we have $\boldsymbol{\theta}^\star \in \texttt{BALL}_{t-1}$ for all $t \in [T]$ with high probability. At time $t \in [T]$, we have

$$\begin{aligned}R_t &= \boldsymbol{\theta}^{\star\top}\widehat{\boldsymbol{\phi}}_{t,A_t} + \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t] - \widehat{\boldsymbol{\phi}}_{t,A_t}\right) - \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t] - \boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t)\right) + \eta_t \\ &= \boldsymbol{\theta}^{\star\top}\widehat{\boldsymbol{\phi}}_{t,A_t} + \boldsymbol{\theta}^{\star\top}\left(\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_t, A_t) \mid \boldsymbol{S}_{1:t}, A_t] - \widehat{\boldsymbol{\phi}}_{t,A_t}\right) + \varepsilon_t + \eta_t\end{aligned} \tag{G.11}$$

where the first equality follows from the definition of R_t in Equation (2.1) and the second equality follows from Equation (4.5) By the definition of $\widehat{\boldsymbol{\theta}}_t$ given in Equation (3.4), it follows that

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star &= \boldsymbol{\Sigma}_t^{-1}\sum_{\tau=1}^t R_\tau \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} - \boldsymbol{\theta}^\star \\ &= \left[\boldsymbol{\Sigma}_t^{-1}\left(\sum_{\tau=1}^t \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}\widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top\right) - 1\right]\boldsymbol{\theta}^\star + \boldsymbol{\Sigma}_t^{-1}\sum_{\tau=1}^t \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}\left(\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau\right] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}\right)^\top \boldsymbol{\theta}^\star \\ &\quad + \boldsymbol{\Sigma}_t^{-1}\sum_{\tau=1}^t \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}(\eta_\tau + \varepsilon_\tau) \\ &= -\lambda\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\theta}^\star + \boldsymbol{\Sigma}_t^{-1}\sum_{\tau=1}^t \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}\left(\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau\right] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}\right)^\top \boldsymbol{\theta}^\star + \boldsymbol{\Sigma}_t^{-1}\sum_{\tau=1}^t \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}(\eta_\tau + \varepsilon_\tau)\end{aligned} \tag{G.12}$$

where the first equality follows from Equation (G.11) and the last equality follows from the definition of Σ_t in Equation (3.3).

Compared to standard analysis of vanilla `LinUCB`, the only different term is that we have an extra term

$$\boldsymbol{\Sigma}_t^{-1} \sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \left(\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau\right] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right)^\top \boldsymbol{\theta}^\star. \tag{G.13}$$

Following the same analysis as Equation (G.8), we arrive at

$$\left| \left(\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right)^\top \boldsymbol{\theta}^\star \right| \le \left\| \mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right\|_2 \le \sqrt{dD_t}. \tag{G.14}$$

To control Equation (G.13), we have

$$\begin{aligned}
&\left| \left(\sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top (\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau})^\top \boldsymbol{\theta}^\star \right) \boldsymbol{\Sigma}_t^{-1} \left(\sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} (\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau})^\top \boldsymbol{\theta}^\star \right) \right| \\
&= \left\| \sum_{\tau=1}^{t} \boldsymbol{\Sigma}_t^{-1/2} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} (\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau})^\top \boldsymbol{\theta}^\star \right\|_2^2 \\
&\overset{(i)}{\le} \left(\sum_{\tau=1}^{t} \left\| \boldsymbol{\Sigma}_t^{-1/2} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} (\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau})^\top \boldsymbol{\theta}^\star \right\|_2 \right)^2 \\
&\overset{(ii)}{\le} \left(\sum_{\tau=1}^{t} \left[\left(\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right)^\top \boldsymbol{\theta}^\star \right]^2 \right) \left(\sum_{\tau=1}^{t} \left\| \boldsymbol{\Sigma}_t^{-1/2} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right\|_2^2 \right) \\
&\overset{(iii)}{\le} d \left(\sum_{\tau=1}^{t} D_\tau \right) \left(\sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top \boldsymbol{\Sigma}_t^{-1} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right)
\end{aligned} \tag{G.15}$$

where inequality (i) follows from the triangle inequality, inequality (ii) follows from Cauchy-Schwarz inequality and inequality (iii) follows from Equation (G.14). Using properties of the trace operator, we continue to bound the right-hand side of Equation (G.15) using

$$\begin{aligned}
d \left(\sum_{\tau=1}^{t} D_\tau \right) \left(\sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top \boldsymbol{\Sigma}_t^{-1} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right) &= d \left(\sum_{\tau=1}^{t} D_\tau \right) \cdot \operatorname{tr} \left(\boldsymbol{\Sigma}_t^{-1} \sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top \right) \\
&\overset{(i)}{=} d \left(\sum_{\tau=1}^{t} D_\tau \right) \left(d - \lambda \operatorname{tr}\left(\boldsymbol{\Sigma}_t^{-1} \right) \right) \le d^2 \left(\sum_{\tau=1}^{t} D_\tau \right)
\end{aligned} \tag{G.16}$$

where equality (i) follows from the definition of $\boldsymbol{\Sigma}_t$ as given in Equation (3.3). Taking Equation (G.16) into Equation (G.15) yields that

$$\left\| \sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top \left(\mathbb{E}\left[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:t}, A_\tau\right] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right)^\top \boldsymbol{\theta}^\star \right\|_{\Sigma_t^{-1}} \le d \sqrt{\sum_{\tau=1}^{t} D_\tau} \tag{G.17}$$

Therefore, using standard self-normalization concentration inequalities (see Lemma G.2) with Equations (G.12) and (G.17), with probability at least $1 - \delta_t$,

$$\begin{aligned}
\left\| \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star \right\|_{\boldsymbol{\Sigma}_t} &\le \sqrt{\lambda} \|\boldsymbol{\theta}^\star\|_{\boldsymbol{\Sigma}_t^{-1}} + \left\| \sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} (\eta_\tau + \varepsilon_\tau) \right\|_{\boldsymbol{\Sigma}_t^{-1}} + \left\| \sum_{\tau=1}^{t} \widehat{\boldsymbol{\phi}}_{\tau,A_\tau}^\top \left(\mathbb{E}[\boldsymbol{\Phi}(\boldsymbol{Y}_\tau, A_\tau) \mid \boldsymbol{S}_{1:\tau}, A_\tau] - \widehat{\boldsymbol{\phi}}_{\tau,A_\tau} \right)^\top \boldsymbol{\theta}^\star \right\|_{\boldsymbol{\Sigma}_t^{-1}} \\
&\le \sqrt{\lambda} + (\sigma_\eta + \sigma_\varepsilon) \sqrt{2 \log(\det(\boldsymbol{\Sigma}_t) \det(\boldsymbol{\Sigma}_1)^{-1} / \delta_t)} + d \sqrt{\sum_{\tau=1}^{t} D_\tau} \\
&\le \sqrt{\lambda} + (\sigma_\eta + \sigma_\varepsilon) \sqrt{2 \log \left[\left(1 + \frac{tB^2}{d\lambda} \right)^d / \delta_t \right]} + d \sqrt{\sum_{\tau=1}^{t} D_\tau}
\end{aligned}$$

where the last inequality follows from Equation (G.21). It suffices to set $\delta_t := \delta(3/\pi^2)/t^2$. Hence, by taking γ_t as defined in Equation (4.7), with probability at least $1-\delta$, we have

$$\left\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^\star\right\|^2_{\boldsymbol{\Sigma}_t^{-1}} \le \gamma_t$$

holds for all $t \in [T]$. It follows from Equation (G.10) that $\sum_{t=1}^T \mathrm{reg}_t$ is bounded by

$$2\sum_{t=1}^T \sqrt{dD_t} + 2\sqrt{6T\left(\sum_{t=1}^T D_t\right) d^3 \log\left(1+\frac{TB^2}{d\lambda}\right)} + 2\sqrt{2\gamma_T^{(0)} Td \log\left(1+\frac{TB^2}{d\lambda}\right)}$$

where we use the naive bound $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Applying Cauchy-Schwarz inequality to $\sum_{t=1}^T \sqrt{dD_t}$ yields that

$$\begin{aligned}\sum_{t=1}^T \mathrm{reg}_t &\le 2\sqrt{dT\sum_{t=1}^T D_t} + 2\sqrt{6T\left(\sum_{t=1}^T D_t\right) d^3 \log\left(1+\frac{TB^2}{d\lambda}\right)} + 2\sqrt{2\gamma_T^{(0)} Td \log\left(1+\frac{TB^2}{d\lambda}\right)} \\ &\le 4\sqrt{6T\left(\sum_{t=1}^T D_t\right) d^3 \log\left(1+\frac{TB^2}{d\lambda}\right)} + 2\sqrt{2\gamma_T^{(0)} Td \log\left(1+\frac{TB^2}{d\lambda}\right)}\end{aligned}$$

as desired. ■

G.1 Technical Lemmas

Lemma G.1. For any $t \in [T]$ and $\xi_{1,t}$ defined in Equation (G.6), under the same conditions as Theorem 4.1, we have

$$\sum_{t=1}^T \xi_{1,t}^2 \le 2\gamma_T d \log\left(1+\frac{TB^2}{d\lambda}\right) \tag{G.18}$$

Proof. For $\gamma_t \ge 1$, by the definition of $\xi_{1,t}$ in the above display,

$$\sum_{t=1}^T \xi_{1,t}^2 \le \sum_{t=1}^T \gamma_t \min\left\{\widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1} \widehat{\boldsymbol{\phi}}_{t,A_t}, 1\right\} \tag{G.19}$$

To control Equation (G.19), we use the potential function bound. We include a brief proof here for completeness. By the definition of $\boldsymbol{\Sigma}_{t+1}$ in Equation (3.3), we have

$$\begin{aligned}\det \boldsymbol{\Sigma}_{t+1} &= \det\left(\boldsymbol{\Sigma}_t + \widehat{\boldsymbol{\phi}}_{t,A_t}\widehat{\boldsymbol{\phi}}_{t,A_t}^\top\right) = \det(\boldsymbol{\Sigma}_t)\det\left(\boldsymbol{I} + \boldsymbol{\Sigma}_t^{-1/2}\widehat{\boldsymbol{\phi}}_{t,A_t}\left(\boldsymbol{\Sigma}_t^{-1/2}\widehat{\boldsymbol{\phi}}_{t,A_t}\right)^\top\right) \\ &= \det(\boldsymbol{\Sigma}_t)\left(1+\widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1}\widehat{\boldsymbol{\phi}}_{t,A_t}\right),\end{aligned} \tag{G.20}$$

where the last equality follows from Sylvester's determinant theorem. By induction, it is straightforward to show that

$$\det \boldsymbol{\Sigma}_T = \det \boldsymbol{\Sigma}_0 \prod_{t=1}^T \left(1+\widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1}\widehat{\boldsymbol{\phi}}_{t,A_t}\right),$$

following Equation (G.20). Rearranging terms and taking logarithm on both sides of the above display implies that

$$\sum_{t=1}^T \log\left(1+\widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1}\widehat{\boldsymbol{\phi}}_{t,A_t}\right) = \log\left(\frac{\det \boldsymbol{\Sigma}_T}{\det \boldsymbol{\Sigma}_0}\right) \le 2\gamma_T d \log\left(1+\frac{TB^2}{d\lambda}\right) \tag{G.21}$$

where the last inequality follows from Assumption 2.1 and the potential function bound in Lemma G.3. Hence, applying the above display to Equation (G.19)

$$\begin{aligned}\sum_{t=1}^T \gamma_t \min\left\{\widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1}\widehat{\boldsymbol{\phi}}_{t,A_t}, 1\right\} &\overset{(i)}{\le} 2\gamma_T \sum_{t=1}^T \log\left(1+\widehat{\boldsymbol{\phi}}_{t,A_t}^\top \boldsymbol{\Sigma}_t^{-1}\widehat{\boldsymbol{\phi}}_{t,A_t}\right) = 2\gamma_T \log\left(\frac{\det \boldsymbol{\Sigma}_T}{\det \boldsymbol{\Sigma}_0}\right) \\ &\le 2\gamma_T d \log\left(1+\frac{TB^2}{d\lambda}\right).\end{aligned}$$

where inequality (i) follows from $\log(1+y) \geq y/2$ for all $y \in [0,1]$. Taking the above display into Equation (G.19) yields the desired bound as in Equation (G.18). ■

Lemma G.2. [Self-Normalized Bound for Vector-Valued Martingales] Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$. Let $\{\boldsymbol{X}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\boldsymbol{X}_t$ is $\mathcal{F}_{t-1}$-measurable. Assume that $\boldsymbol{V}$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define

$$\bar{\boldsymbol{V}}_t = \boldsymbol{V} + \sum_{s=1}^{t} \boldsymbol{X}_s \boldsymbol{X}_s^\top \quad S_t = \sum_{s=1}^{t} \eta_s \boldsymbol{X}_s.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \geq 0$,

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

Proof. See Theorem 1 in Abbasi-Yadkori et al. (2011). ■

Lemma G.3 (Potential Function Bound)**.** For any sequence $\boldsymbol{x}_0, \ldots \boldsymbol{x}_{T-1}$ such that, for $t < T$, $\|\boldsymbol{x}_t\|_2 \leq B$, we have

$$\log\left(\det \boldsymbol{\Sigma}_{T-1} / \det \boldsymbol{\Sigma}_0\right) = \log\det\left(\boldsymbol{I} + \frac{1}{\lambda}\sum_{t=0}^{T-1} \boldsymbol{x}_t \boldsymbol{x}_t^\top\right) \leq d \log\left(1 + \frac{TB^2}{d\lambda}\right),$$

where $\boldsymbol{\Sigma}_t = \lambda \boldsymbol{I} + \sum_{\tau=0}^{t-1} \boldsymbol{x}_\tau \boldsymbol{x}_\tau^\top$ with $\boldsymbol{\Sigma}_0 = \lambda \boldsymbol{I}$ for any $\lambda > 0$.

Proof. See Lemma 6.11 in Agarwal et al. (2019). ■

H Proof of Results in Section 4.3

H.1 Proof of Proposition 4.1

Proof. Let $\boldsymbol{b} = \left(\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_m^\top\right) \in \mathbb{R}^{(m+1)d_S}$. A standard analysis of the OLS estimator $\widehat{\boldsymbol{b}}$ yields that

$$\mathbb{E}_{\mathcal{D}}\left[\left\|\boldsymbol{b} - \widehat{\boldsymbol{b}}\right\|_2^2\right] \lesssim \frac{m d_S}{N T_0}, \tag{H.1}$$

where the expectation is taken with respect to the historical data $\mathcal{D}$. We omit the details for brevity. For a new copy $\boldsymbol{Y}_t = (W_t, \boldsymbol{S}_t)$ independent of the historical data $\mathcal{D}$, since

$$\mu_t := \mathbb{E}\left[W_t \mid \boldsymbol{S}_{1:t}\right] = \sum_{j=0}^{m} \boldsymbol{\beta}_j^\top \boldsymbol{S}_{t-j} \tag{H.2}$$

and

$$\operatorname{Var}\left(W_t \mid \boldsymbol{S}_{1:t}\right) = \operatorname{Var}\left(\xi_t \mid \boldsymbol{S}_{1:t}\right) = 1,$$

we have $W_t \mid \boldsymbol{S}_{1:t} \sim \mathcal{N}(\mu_t, 1)$. The imputed W_t is then given by

$$\widehat{W}_t = \sum_{j=0}^{m} \widehat{\boldsymbol{\beta}}_j^\top \boldsymbol{S}_{t-j} + \xi_t$$

and it follows that $\widehat{W}_t \mid \boldsymbol{S}_{1:t} \sim \mathcal{N}(\widehat{\mu}_t, 1)$, where

$$\widehat{\mu}_t := \sum_{j=0}^{m} \widehat{\boldsymbol{\beta}}_j^\top \boldsymbol{S}_{t-j} \tag{H.3}$$

It follows that

$$\begin{aligned}\sqrt{D_t} &= \sqrt{\frac{1}{2}\,\mathrm{KL}\left(\mathcal{N}(\mu_t,1)\|\mathcal{N}(\widehat{\mu}_t,1)\right)} = \frac{1}{2}\left|\mu_t-\widehat{\mu}_t\right| \\ &= \frac{1}{4}\left|\sum_{j=0}^{m}\left(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\right)^\top \boldsymbol{S}_{t-j}\right|\end{aligned} \tag{H.4}$$

where the last equality follows from the definition of μ_t and $\widehat{\mu}_t$ in Equations (H.2). Since $\boldsymbol{S}_{t-j}$ and $\widehat{\boldsymbol{\beta}}_j$ are independent, conditioned on $\widehat{\boldsymbol{\beta}}_j$, we have

$$\sum_{j=0}^{m}\left(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\right)^\top \boldsymbol{S}_{t-j} \sim \mathcal{N}\left(0, \sum_{j=0}^{m}\left\|\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\right\|_2^2\right).$$

Combining the above display with Equation (H.4) yields that

$$\begin{aligned}\mathbb{E}\left[\sqrt{D_t}\right] &= \frac{1}{4}\mathbb{E}\left|\sum_{j=0}^{m}\left(\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\right)^\top \boldsymbol{S}_{t-j}\right| \\ &= \frac{\pi}{8}\mathbb{E}_{\mathcal{D}}\sqrt{\sum_{j=0}^{m}\left\|\widehat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j\right\|_2^2} \lesssim \sqrt{\frac{md_S}{NT_0}},\end{aligned}$$

where the last equality follows from Equation (H.1).

It then follows from Theorem 4.1 that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_T\right] &\le \delta T + \mathbb{E}\mathrm{reg}_T^{(\mathtt{imp})} + \mathrm{reg}_T^{(\mathtt{lin})} \\ &\lesssim \delta T + \sqrt{\gamma_T^{(0)} Td\log\left(1+\frac{TB^2}{d\lambda}\right)} + \mathbb{E}\sqrt{T\left(\sum_{t=1}^{T}D_t\right)d^3\log\left(1+\frac{TB^2}{d\lambda}\right)} \\ &\lesssim \delta T + \sqrt{\gamma_T^{(0)} Td\log\left(1+\frac{TB^2}{d\lambda}\right)} + T\sqrt{\frac{md_Sd^3}{NT_0}\log\left(1+\frac{TB^2}{d\lambda}\right)}\end{aligned} \tag{H.5}$$

Taking $\delta = T^{-1/2}$, we have

$$\gamma_T^{(0)} = 3\lambda + 6(\sigma_\eta+2)^2\log\left[4T^{5/2}\left(1+\frac{TB^2}{d\lambda}\right)^d\right] \asymp d\log T + d\log\left(1+\frac{TB^2}{d\lambda}\right)$$

Taking the above display into Equation (H.5) yields the desired result. ■

H.2 Proof of Proposition 4.2

Proof. From the classical nonparametric statistics literature, there exists an estimator $\widehat{f}$ (such as the kernel estimator, see Chapter 1 of Tsybakov, 2008) of f that satisfies

$$\mathbb{E}_{\mathcal{D}}\left[\left\|\widehat{f}-f\right\|_{L_2}^2\right] \lesssim N^{-\frac{2\beta}{2\beta+d_S}}, \tag{H.6}$$

where the expectation is taken with respect to the historical data $\mathcal{D}$. For any pair $(\boldsymbol{S}_t, W_t)$ independent of the historical data $\mathcal{D}$, where $\boldsymbol{S}_t \sim \mathrm{Unif}([0,1]^{d_S})$, one has

$$\begin{aligned}\mathrm{KL}\left(\mathbb{P}_f\left(W_t \mid \boldsymbol{S}_t\right)\|\mathbb{P}_{\widehat{f}}\left(W_t \mid \boldsymbol{S}_t\right)\right) &= \mathbb{E}_{\boldsymbol{S}_t}\left[\mathrm{KL}\left(\mathcal{N}(f(\boldsymbol{S}_t),1)\|\mathcal{N}(\widehat{f}(\boldsymbol{S}_t),1)\right) \mid \widehat{f}\right] \\ &= \frac{1}{2}\mathbb{E}_{\boldsymbol{S}_t}\left[\left(\widehat{f}(\boldsymbol{S}_t)-f(\boldsymbol{S}_t)\right)^2\right] = \frac{1}{2}\left\|\widehat{f}-f\right\|_{L_2}^2.\end{aligned}$$

Taking expectation over the historical data and combining Equation (H.6) with the above display, we have

$$\mathbb{E}_{\mathcal{D}} \operatorname{KL}\left(\mathbb{P}_f\left(W_0 \mid \boldsymbol{S}_0\right), \mathbb{P}_{\widehat{f}}\left(W_0 \mid \boldsymbol{S}_0\right)\right) \lesssim N^{-\frac{2\beta}{2\beta+d_S}}.$$

Recall the definition of D_t in Equation (4.1), it follows that

$$\mathbb{E}\sqrt{D_t} \asymp \mathbb{E}_{\boldsymbol{S}_t}\sqrt{\operatorname{KL}\left(\mathbb{P}_f\left(\boldsymbol{Y}_t|\boldsymbol{S}_t=\boldsymbol{s}_t\right)\|\mathbb{P}_{\widehat{f}}\left(\boldsymbol{Y}_t|\boldsymbol{S}_t=\boldsymbol{s}_t\right)\right)} \lesssim \mathbb{E}_{\boldsymbol{S}_t}\left\|\widehat{f}-f\right\|_{L_2} \lesssim N^{-\frac{\beta}{2\beta+d_S}}.$$

Combining the above display with Theorem 4.1 yields that

$$\mathbb{E}\left[\mathcal{R}_T\right] \lesssim T\sqrt{d^3\log\left(1+\frac{TB^2}{d\lambda}\right)}N^{-\frac{\beta}{2\beta+d_S}} + \mathrm{reg}_T^{(\mathtt{lin})} + \delta T.$$

Taking $\delta = T^{-1/2}$ and following a similar proof of Proposition 4.1 yields the desired result. ■

H.3 Robustness of Propositions 4.1 and 4.2 to Covariate Shift

The proofs above assume that the online contexts $\boldsymbol{S}_t$ follow the same marginal distribution as the historical data. Here we show that the rates for $\mathbb{E}[\sqrt{D_t}]$ in Propositions 4.1 and 4.2 are preserved when the marginal distribution of $\boldsymbol{S}_t$ shifts, provided the conditional mechanism $W_t \mid \boldsymbol{S}_{1:t}$ remains invariant and the marginal shift is mild.

Consider the linear model of Proposition 4.1 and suppose the online contexts are independent of the historical dataset $\mathcal{D}$ and satisfy $\boldsymbol{S}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$ with a possibly time-varying covariance. The key step in the proof (see Equation (H.4)) involves bounding the distribution of the estimation error $\sum_{j=0}^{m}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top \boldsymbol{S}_{t-j}$. Under the shifted marginal, this term is distributed as

$$\sum_{j=0}^{m}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top \boldsymbol{S}_{t-j} \;\sim\; \mathcal{N}\left(0,\; \sum_{j=0}^{m}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^\top \boldsymbol{\Sigma}_{t-j}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right),$$

and the same rates for $\mathbb{E}[\sqrt{D_t}]$ continue to hold as long as $\|\boldsymbol{\Sigma}_t\|_{\mathrm{op}} = O(1)$. Thus, even if the marginal distribution of $\boldsymbol{S}_t$ shifts in the online setting, the regret guarantees of Proposition 4.1 remain valid under uniformly bounded second moments. An analogous argument applies to Proposition 4.2: the nonparametric rate depends on the L_2 estimation error $\|\widehat{f} - f\|_{L_2}^2$, which is invariant to the online marginal as long as the conditional mechanism $W_t \mid \boldsymbol{S}_t$ is unchanged.

More broadly, the key condition is that the conditional mechanism $W_t \mid \boldsymbol{S}_{1:t}$ is invariant across the historical and online phases and the marginal shift in $\boldsymbol{S}_{1:t}$ is mild (e.g., bounded moments). Under these conditions, the rates in Section 4.3 remain unchanged. Investigating more general covariate-shift models, such as transfer-exponent conditions between historical and online distributions (Cai et al., 2024), is an interesting direction that requires substantially more technical care and is left for future work.

I Extension: A Confidence-Weighted Variant of PULSE-UCB

In the main text, we employ the standard OFUL algorithm in the online stage of PULSE-UCB as a simple, well-understood baseline. As demonstrated in Theorems 4.1 and 5.1, this plug-in approach already achieves near-optimal regret for our primary settings. However, since the context recovery step inherently introduces an imputation error D_t for the feature at each step, one might ask whether we can make further improvements to mitigate this residual error.

To explore this, we draw inspiration from the literature on linear contextual bandits with corrupted rewards (He et al., 2022) to propose an alternative algorithmic variant. It is important to note that this variant is not a direct application of He et al. (2022); their work focuses on mitigating adversarial corruptions on the rewards, whereas our setting deals with misspecification in the contexts due to offline imputation. By adapting

their confidence-weighted mechanism to our missing context framework, this alternative variant can potentially yield tighter regret bounds. Crucially, this improvement is conditional: it requires a reliable estimator for the total imputation error budget $\bar{C}$ a priori, and it is primarily beneficial when the imputation error D_t varies substantially across time.

I.1 Algorithm Design: Robust PULSE-UCB

The primary modification in this robust variant lies in the construction of the confidence sets and the weighted regularized least-squares estimator. Specifically, we select actions via the modified confidence set $\widetilde{\texttt{BALL}}_t = \{\theta : \|\theta - \widetilde{\theta}_{t-1}\|^2_{\widetilde{\Sigma}_{t-1}} \le \widetilde{\gamma}_{t-1}\}$, where the covariance matrix and the parameter estimator are updated using adaptive weights:

$$\widetilde{\Sigma}_t = \lambda I + \sum_{\tau=1}^{t} w_\tau \widehat{\phi}_{\tau,a_\tau} \widehat{\phi}_{\tau,a_\tau}^\top, \quad \widetilde{\theta}_t = \widetilde{\Sigma}_t^{-1} \sum_{\tau=1}^{t} w_\tau \widehat{\phi}_{\tau,a_\tau} R_\tau.$$

The adaptive weight w_t for each step is defined as:

$$w_t = \min\left\{1, \frac{\alpha}{\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}}}\right\},$$

where $\alpha > 0$ is a tuning parameter.

I.2 Theoretical Guarantees and Regret Comparison

We establish the theoretical guarantee for this robust variant as follows:

Proposition I.1 (Regret of Robust PULSE-UCB)**.** Let $\bar{C}$ be any known upper bound such that $\bar{C} \ge \sqrt{d}\sum_{t=1}^{T}\sqrt{D_t}$. By setting the tuning parameter $\alpha = \sqrt{d}/\bar{C}$ and the confidence radius $\widetilde{\gamma}_t = \gamma_t^{(0)} + 3\alpha^2 d\big(\sum_{t=1}^{T}\sqrt{D_t}\big)^2$, the cumulative regret of the robust variant satisfies, with high probability:

$$R_T \le R_T^{(lin)} + \widetilde{R}_T^{(imp)},$$

where $R_T^{(lin)}$ matches the standard linear bandit regret term in Theorem 4.1, and the modified imputation penalty term is bounded by:

$$\widetilde{R}_T^{(imp)} = 12d\log^{3/2}\left(1 + \frac{TB^2}{d\lambda}\right)\bar{C}.$$

Proof. Let $\phi_{t,a}$ denote the true feature vector and $\widehat{\phi}_{t,a}$ denote the imputed feature vector provided by the pretrained model. We can express the observed reward at round t by explicitly isolating the stochastic noise and the context imputation error:

$$R_t = \langle\theta^\star, \phi_{t,a_t}\rangle + \eta_t = \langle\theta^\star, \widehat{\phi}_{t,a_t}\rangle + c_t + \eta_t,$$

where the adversarial corruption term induced by imputation is defined as $c_t = \langle\theta^\star, \phi_{t,a_t} - \widehat{\phi}_{t,a_t}\rangle$. Assuming the unknown environment parameter is bounded by $\|\theta^\star\|_2 \le S$ and the feature norm is bounded by L, the magnitude of this corruption is bounded by $|c_t| \le S\|\phi_{t,a_t} - \widehat{\phi}_{t,a_t}\|_2 \le S\sqrt{D_t}$. Consequently, the total cumulative corruption over the horizon T is $C = \sum_{t=1}^{T}|c_t| \le S\sum_{t=1}^{T}\sqrt{D_t}$.

We first establish the concentration of the estimation error. The weighted ridge regression estimator $\widetilde{\theta}_t$ admits the closed-form representation:

$$\widetilde{\theta}_t = \widetilde{\Sigma}_t^{-1}\sum_{\tau=1}^{t} w_\tau \widehat{\phi}_{\tau,a_\tau}(\langle\theta^\star, \widehat{\phi}_{\tau,a_\tau}\rangle + c_\tau + \eta_\tau)$$

Subtracting $\theta^\star$ from both sides and multiplying by $\widetilde{\Sigma}_t$, the Mahalanobis norm of the estimation error satisfies:

$$\|\widetilde{\theta}_t - \theta^\star\|_{\widetilde{\Sigma}_t} \le \Big\|\widetilde{\Sigma}_t^{-1}\sum_{\tau=1}^{t} w_\tau\widehat{\phi}_{\tau,a_\tau}\eta_\tau\Big\|_{\widetilde{\Sigma}_t} + \Big\|\widetilde{\Sigma}_t^{-1}\sum_{\tau=1}^{t} w_\tau\widehat{\phi}_{\tau,a_\tau}c_\tau\Big\|_{\widetilde{\Sigma}_t} + \lambda\|\widetilde{\Sigma}_t^{-1}\theta^\star\|_{\widetilde{\Sigma}_t}.$$

This decomposes the error into three absolute components: the stochastic error, the corruption error, and the regularization error. Because the adaptive weights w_τ are bounded by 1, the stochastic error can be bounded using the standard self-normalized martingale concentration inequality by $R\sqrt{d\log((1+TL^2/\lambda)/\delta)}$ with probability at least $1-\delta$. The regularization error is strictly bounded by $\sqrt{\lambda}S$.

Crucially, we bound the corruption error utilizing the definition of the adaptive weight $w_\tau = \min\{1, \alpha/\|\widehat{\phi}_{\tau,a_\tau}\|_{\widetilde{\Sigma}_{\tau-1}^{-1}}\}$. By Cauchy-Schwarz and the weight constraint, we have:

$$\left\|\widetilde{\Sigma}_t^{-1}\sum_{\tau=1}^{t} w_\tau \widehat{\phi}_{\tau,a_\tau} c_\tau\right\|_{\widetilde{\Sigma}_t} \le \sum_{\tau=1}^{t} |c_\tau| w_\tau \|\widehat{\phi}_{\tau,a_\tau}\|_{\widetilde{\Sigma}_t^{-1}} \le \sum_{\tau=1}^{t} |c_\tau|\alpha \le \alpha C \le \alpha S \sum_{t=1}^{T}\sqrt{D_t}.$$

Therefore, with probability at least $1-\delta$, the true parameter $\theta^\star$ lies in the confidence set $\widetilde{\texttt{BALL}}_t$ for all $t \in [T]$, centered at $\widetilde{\theta}_{t-1}$ with the exact confidence radius:

$$\beta_T = R\sqrt{d\log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S + \alpha C.$$

Conditioned on this high-probability event, we bound the instantaneous regret. Let $a_t^\star = \arg\max_{a\in\mathcal{A}}\langle\theta^\star, \phi_{t,a}\rangle$. By the optimism of the action selection rule, $\langle\widetilde{\theta}_{t-1}, \widehat{\phi}_{t,a_t^\star}\rangle + \beta_T\|\widehat{\phi}_{t,a_t^\star}\|_{\widetilde{\Sigma}_{t-1}^{-1}} \le \langle\widetilde{\theta}_{t-1}, \widehat{\phi}_{t,a_t}\rangle + \beta_T\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}}$. The instantaneous regret satisfies:

$$\mathrm{reg}_t = \langle\theta^\star, \phi_{t,a_t^\star}\rangle - \langle\theta^\star, \phi_{t,a_t}\rangle \le \langle\theta^\star, \widehat{\phi}_{t,a_t^\star}\rangle - \langle\theta^\star, \widehat{\phi}_{t,a_t}\rangle + 2\max_{a\in\mathcal{A}}|c_{t,a}|.$$

Applying the confidence bound and the optimism condition, this reduces to:

$$\mathrm{reg}_t \le 2\beta_T\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}} + 2\max_{a\in\mathcal{A}}|c_{t,a}|.$$

Summing over the horizon T, the cumulative regret is strictly bounded by:

$$R_T \le \sum_{t=1}^{T} 2\beta_T\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}} + 2S\sum_{t=1}^{T}\sqrt{D_t}.$$

To rigorously bound the sum of the weighted confidence widths, we partition the time indices into two disjoint subsets: $\mathcal{I}_1 = \{t \in [T] : w_t = 1\}$ and $\mathcal{I}_2 = \{t \in [T] : w_t < 1\}$. For the un-truncated subset $\mathcal{I}_1$, we directly apply the Cauchy-Schwarz inequality and the standard Elliptical Potential Lemma:

$$\sum_{t\in\mathcal{I}_1} 2\beta_T\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}} \le 2\beta_T\sqrt{|\mathcal{I}_1|\sum_{t\in\mathcal{I}_1}\|\widehat{\phi}_{t,a_t}\|^2_{\widetilde{\Sigma}_{t-1}^{-1}}} \le 2\beta_T\sqrt{2dT\log\left(1+\frac{TL^2}{d\lambda}\right)}.$$

This yields the standard linear bandit regret term $R_T^{(lin)}$.

For the truncated subset $\mathcal{I}_2$, the condition $w_t < 1$ strictly implies $w_t = \alpha/\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}}$. We algebraically rewrite the confidence width explicitly in terms of the weight:

$$\sum_{t\in\mathcal{I}_2} 2\beta_T\|\widehat{\phi}_{t,a_t}\|_{\widetilde{\Sigma}_{t-1}^{-1}} = \sum_{t\in\mathcal{I}_2} 2\beta_T\frac{w_t\|\widehat{\phi}_{t,a_t}\|^2_{\widetilde{\Sigma}_{t-1}^{-1}}}{\alpha}.$$

By applying the matrix determinant lemma and bounding the fractional terms, we obtain an absolute upper bound:

$$\frac{2\beta_T}{\alpha}\sum_{t\in\mathcal{I}_2} w_t\|\widehat{\phi}_{t,a_t}\|^2_{\widetilde{\Sigma}_{t-1}^{-1}} \le \frac{2\beta_T}{\alpha}\sum_{t=1}^{T}\min\left(1, w_t\|\widehat{\phi}_{t,a_t}\|^2_{\widetilde{\Sigma}_{t-1}^{-1}}\right) \le \frac{4\beta_T d}{\alpha}\log\left(1+\frac{TL^2}{d\lambda}\right).$$

Finally, we substitute the specified tuning parameter $\alpha = \sqrt{d}/\bar{C}$, where the known upper bound satisfies $\bar{C} \ge \sqrt{d}\sum_{t=1}^{T}\sqrt{D_t}$. The corruption penalty embedded within the confidence radius satisfies $\alpha C \le$

$(\sqrt{d}/\bar{C})(S\bar{C}/\sqrt{d}) = S$. Thus, the confidence radius simplifies to an inflation factor strictly independent of T: $\beta_T \le R\sqrt{d\log((1+TL^2/\lambda)/\delta)} + \sqrt{\lambda}S + S$.

Substituting α into the upper bound for $\mathcal{I}_2$, the imputation-induced regret becomes:

$$\widetilde{R}_T^{(imp)} = \frac{4\beta_T d\bar{C}}{\sqrt{d}} \log\left(1 + \frac{TL^2}{d\lambda}\right) = 4\sqrt{d}\beta_T \bar{C} \log\left(1 + \frac{TL^2}{d\lambda}\right) + 2S\frac{\bar{C}}{\sqrt{d}}.$$

By expanding β_T and grouping the leading polynomial and logarithmic terms, the penalty strictly respects the proposed theoretical bound $\widetilde{\mathcal{O}}(d\bar{C})$, verifying that the combination of bounded weights cleanly eliminates the $\sqrt{T}$ dependency strictly through the control of residual misspecification. This completes the proof. ■

Discussion and Comparison: To understand the advantage of this robust variant, suppose the upper bound $\bar{C}$ is tightly chosen such that $\bar{C} = \Theta\big(\sqrt{d}\sum_{t=1}^{T}\sqrt{D_t}\big)$. In this case, the imputation-induced regret becomes $\widetilde{R}_T^{(imp)} = \widetilde{\mathcal{O}}\big(d^{3/2}\sum_{t=1}^{T}\sqrt{D_t}\big)$.

Recall that for the standard PULSE-UCB (Theorem 4.1), the corresponding imputation penalty is $R_T^{(imp)} = \widetilde{\mathcal{O}}\big(d^{3/2}\sqrt{T\sum_{t=1}^{T}D_t}\big)$. By the Cauchy–Schwarz inequality, we always have:

$$\sum_{t=1}^{T}\sqrt{D_t} \le \sqrt{T\sum_{t=1}^{T}D_t}.$$

Consequently, the robust variant *never* exhibits a worse regret upper bound than the standard PULSE-UCB. More importantly, it can achieve strictly tighter bounds in environments where the imputation errors D_t vary substantially (i.e., highly heterogeneous) over time.

Practical Trade-offs: While the robust variant provides tighter theoretical guarantees for heterogeneous errors, it requires the practitioner to specify the upper bound parameter $\bar{C}$ a priori. In contrast, if the errors D_t are of the same order across time (as in the stationary settings discussed in Section 4.3), the Cauchy–Schwarz inequality is tight, and the regret rates of the two algorithms coincide. Therefore, practitioners can choose between the simpler standard PULSE-UCB and its robust variant based on whether a reliable estimate of $\bar{C}$ is available and the expected heterogeneity of the imputation errors.

J Setup of the Lower Bound

Recall that for obtaining the lower bound, we assume the action set is given by

$$\mathcal{A} = \{\pm 1\}.$$

We use a similar construction as given in Section 4.3.

Let

$$\boldsymbol{Y}_{t,a} := \boldsymbol{\Phi}(\boldsymbol{Y}_t, a) \quad \text{for all } a \in \{-1, 1\}.$$

Partitioning $\boldsymbol{S}_t$ into two parts, we have

$$\boldsymbol{Y}_t = \big(\boldsymbol{S}_t^\top, W_t\big)^\top = \big(\boldsymbol{Q}_t^\top, \boldsymbol{O}_t^\top, W_t\big)^\top \in \mathbb{R}^{d_{\texttt{lin}}} \times \mathbb{R}^{d_{\texttt{non}}} \times \mathbb{R} \tag{J.1}$$

where $W_t \in \mathbb{R}$ is a scalar representing the unobserved part of the context and $d_{\texttt{non}} + d_{\texttt{lin}} = d_S$. For action $a = 1$, we let

$$\boldsymbol{Y}_{t,1} = \boldsymbol{Y}_t.$$

For the alternative action $a = -1$, the associated feature vector is given by

$$\boldsymbol{Y}_{t,-1} = \big(-\boldsymbol{Q}_t^\top, \boldsymbol{0}^\top\big)^\top \in \mathbb{R}^{d_0}, \tag{J.2}$$

mirroring the structure of $\boldsymbol{Y}_{t,1}$, but with fixed values 0 in the coordinates corresponding to $\boldsymbol{O}_t$ and W_t.

We assume that the missing context W_t depends on $\boldsymbol{S}_t$ only through $\boldsymbol{O}_t$, and its conditional expectation is given by

$$\mathbb{E}[W_t \mid \boldsymbol{S}_t] = f(\boldsymbol{O}_t) \tag{J.3}$$

for some function $f : \mathbb{R}^{d_{\text{non}}} \to \mathbb{R}$.

The historical dataset consists of N i.i.d. samples $\left(\boldsymbol{S}_i^{(0)}, \boldsymbol{Y}_i^{(0)}\right)$. Under the above setup, it is equivalent to observing the pairs $\left(\boldsymbol{S}_i^{(0)}, W_i^{(0)}\right)$. We denote the historical dataset by

$$\begin{aligned}\mathcal{D}_N &:= \left\{\left(\boldsymbol{S}_i^{(0)}, W_i^{(0)}\right) : i \in [N]\right\} \\ &= \left\{\left(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}, W_i^{(0)}\right) : i \in [N]\right\}.\end{aligned} \tag{J.4}$$

Denote

$$\Theta := \left\{\boldsymbol{\theta} \in \mathbb{R}^{d_Y} : \|\boldsymbol{\theta}\|_2 \le 1\right\} \tag{J.5}$$

and

$$\Theta_Q := \left\{\boldsymbol{\theta}_Q \in \mathbb{R}^{d_{\text{lin}}} : \|\boldsymbol{\theta}_Q\|_2 \le \frac{\sqrt{3}}{2}\right\}. \tag{J.6}$$

To decouple the estimation of $\boldsymbol{\theta}$ and the nonparametric component f, we assume that $d_{\text{lin}} < \frac{1}{2}\sqrt{3T}$ and for all $i \in [d_{\text{lin}}]$

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_Q^\top, \mathbf{0}^\top, \frac{1}{2}\right)^\top \in \mathbb{R}^{d_Y} \quad |\boldsymbol{\theta}_{Q,i}| = \sqrt{\frac{d_{\text{lin}}}{T}}. \tag{J.7}$$

When $\boldsymbol{\theta}_Q \in \Theta_Q$, we have $\boldsymbol{\theta} \in \Theta$. Combining Equation (J.7) with the construction of $\boldsymbol{Y}_{t,1}$ in Equation (J.1) and $\boldsymbol{Y}_{t,-1}$ in Equation (J.2), we have

$$\boldsymbol{\theta}^\top \boldsymbol{Y}_{t,1} = \boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t + \frac{1}{2} W_t, \tag{J.8}$$

and

$$\boldsymbol{\theta}^\top \boldsymbol{Y}_{t,-1} = -\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t. \tag{J.9}$$

We consider the following data generating process:

Definition J.1 (Data Generating Process for Lower Bounds)**.** *Let $V_t \in \{0, 1\}$ be a latent binary variable defined as follows:*

$$(V_t, \boldsymbol{Q}_t, \boldsymbol{O}_t) = \begin{cases} (0,\ \mathbf{0},\ \boldsymbol{O}_t), & \text{with probability } \frac{1}{2} \\ (1,\ \boldsymbol{Q}_t,\ \mathbf{o}_0), & \text{with probability } \frac{1}{2}, \end{cases} \tag{J.10}$$

where $\boldsymbol{O}_t \sim \mathbb{P}_O = \text{Unif}([-1, 1]^{d_{\text{non}}})$ and $\boldsymbol{Q}_t \sim \mathbb{P}_Q$ is a distribution specified in Equation (K.4) and $\mathbf{o}_0 \in ([-1, 1]^{d_{\text{non}}})^c$ is an arbitrarily fixed vector such that $f(\mathbf{o}_0) = 0$ for f given in Equation (J.3). Under the setup in Equation (J.10):

(i) When $V_t = 0$, by Equations (J.8) and (J.9)

$$\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Y}_{t,1} \mid \boldsymbol{S}_t, V_t\right] = \frac{1}{2} f(\boldsymbol{O}_t), \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Y}_{t,-1} \mid \boldsymbol{S}_t, V_t\right] = 0.$$

Let

$$f^{(1)}(\boldsymbol{O}_t) := \frac{1}{2} f(\boldsymbol{O}_t) \quad \text{and} \quad f^{(-1)}(\boldsymbol{O}_t) \equiv 0,$$

we denote the conditional distribution of the reward R_t as

$$\mathbb{P}_{f^{(a)}(\boldsymbol{O}_t)} := \mathbb{P}\left(R_t \mid A_t = a, \boldsymbol{S}_t, V_t = 0\right) \tag{J.11}$$

for $a \in \mathcal{A}$.

(ii) When $V_t = 1$,

$$\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Y}_{t,1} \mid \boldsymbol{S}_t, V_t\right] = \boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t, \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{Y}_{t,-1} \mid \boldsymbol{S}_t, V_t\right] = -\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t.$$

Let

$$\boldsymbol{Q}_t^{(1)} = \boldsymbol{Q}_t \quad \textit{and} \quad \boldsymbol{Q}_t^{(-1)} = -\boldsymbol{Q}_t,$$

we denote the conditional distribution of R_t *as*

$$\mathbb{P}_{\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(a)}} := \mathbb{P}\left(R_t \mid A_t = a, \boldsymbol{S}_t, V_t = 1\right) \tag{J.12}$$

for $a \in \mathcal{A}$.

Recall the historical data $\mathcal{D}_N$ defined in Equation (J.4). Let $\mathbb{P}_f$ denote the distribution of a pretraining sample $(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}, W_i^{(0)})$, with density

$$\begin{aligned} p_f\left(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}, W_i^{(0)}\right) &= p_f\left(W_i^{(0)} \mid \boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}\right) p_S\left(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}\right) \\ &= p^{(0)}_{f\left(\boldsymbol{O}_i^{(0)}\right)}\left(W_i^{(0)}\right) p_S\left(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}\right) \end{aligned} \tag{J.13}$$

where $p^{(0)}_{f(\boldsymbol{O})}$ is the conditional density of W given $\boldsymbol{O}$, and p_S is the marginal density of $(\boldsymbol{Q}, \boldsymbol{O})$, defined as

$$p_S(\boldsymbol{Q}_t, \boldsymbol{O}_t) = \frac{1}{2}\delta_0(\boldsymbol{Q}_t) p_O(\boldsymbol{O}_t) + \frac{1}{2} p_Q(\boldsymbol{Q}_t)\delta_{\mathbf{o}_0}(\boldsymbol{O}_t). \tag{J.14}$$

We assume the following bounds on KL divergence.

Assumption J.1. *For any* $(\boldsymbol{\theta}_1, f_1), (\boldsymbol{\theta}_2, f_2) \in \Theta \times \mathcal{F}_{\beta,L}$, *the distributions in Equations* (J.11) *and* (J.12) *satisfy that*

$$\mathrm{KL}\left(\mathbb{P}_{f_1^{(a)}(\boldsymbol{O}_t)} \| \mathbb{P}_{f_2^{(a)}(\boldsymbol{O}_t)}\right) \le C_D\left(f_1^{(a)}(\boldsymbol{O}_t) - f_2^{(a)}(\boldsymbol{O}_t)\right)^2$$

and

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\theta}_{1,Q}^\top \boldsymbol{Q}_t} \| \mathbb{P}_{\boldsymbol{\theta}_{2,Q}^\top \boldsymbol{Q}_t}\right) \le C_D\left(\boldsymbol{\theta}_{1,Q}^\top \boldsymbol{Q}_t^{(a)} - \boldsymbol{\theta}_{2,Q}^\top \boldsymbol{Q}_t^{(a)}\right)^2$$

for some constant $C_D > 0$ *and all* $a \in \mathcal{A}$.

Remark J.1. *Assumption J.1 can be satisfied by distributions such as Gaussian or Bernoulli.*

We assume another KL divergence bound between conditional distributions over $W_i^{(0)}$

$$\mathrm{KL}\left(\mathbb{P}^{(0)}_{f\left(\boldsymbol{O}_i^{(0)}\right)}\left(W_i^{(0)}\right) \| \mathbb{P}^{(0)}_{f'\left(\boldsymbol{O}_i^{(0)}\right)}\left(W_i^{(0)}\right)\right) \le C_0\left(f\left(\boldsymbol{O}_i^{(0)}\right) - f'\left(\boldsymbol{O}_i^{(0)}\right)\right)^2. \tag{J.15}$$

Fix a policy $\pi = \{\pi_\tau\}_{\tau=1}^T$, where $\pi_\tau(A_\tau)$ is the abbreviation of

$$\pi_\tau(A_\tau) = \pi_\tau(A_\tau \mid \mathcal{H}_{\tau-1}, \boldsymbol{S}_\tau),$$

and

$$\mathcal{H}_\tau := (\mathcal{D}_N, \boldsymbol{S}_1, A_1, R_1, \cdots, \boldsymbol{S}_\tau, A_\tau, R_\tau),\ \mathcal{H}_0 := \mathcal{D}_N.$$

Let $p_{\boldsymbol{\theta},f}(\cdot \mid \boldsymbol{Q}_t, \boldsymbol{O}_t, A_t)$ denote the reward density under parameters $(\boldsymbol{\theta}, f)$. The joint density of the full observation history $\mathcal{H}_t$ up to round $t \in [T]$ is given by

$$\begin{aligned} &p^{(t)}_{\boldsymbol{\theta},f,\pi}(\mathcal{D}_N, \boldsymbol{Q}_1, \boldsymbol{O}_1, A_1, R_1, \cdots, \boldsymbol{Q}_T, \boldsymbol{O}_T, A_T, R_T) \\ =&\prod_{i=1}^N p_f\left(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}, W_i^{(0)}\right) \prod_{t=1}^t p_S(\boldsymbol{Q}_\tau, \boldsymbol{O}_\tau)\pi_\tau(A_\tau) p_{\boldsymbol{\theta},f}(R_\tau \mid \boldsymbol{Q}_\tau, \boldsymbol{O}_\tau, A_\tau) \end{aligned} \tag{J.16}$$

where the equality follows from Equations (J.13) and (J.14). Additionally, let $\mathbb{E}^{(t)}_{\boldsymbol{\theta},f,\pi}$ denote the expectation taken with respect to the joint density $p^{(t)}_{\boldsymbol{\theta},f,\pi}$.

We now state a formal definition of Theorem J.1.

Theorem J.1 (Formal Lower Bound)**.** Consider the data generating process given in Definition J.1. Suppose that $0 < d_{\texttt{lin}} < c\sqrt{T}$ for some sufficiently small constant $c > 0$ and $d_{\texttt{non}} > 0$ is a constant. Fix a policy π. For any $(\boldsymbol{\theta}, f) \in (\Theta, \mathcal{F}_{\beta,L})$, define

$$\mathcal{R}_T(\boldsymbol{\theta}, f) := \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(t-1)} \mathbb{E}\left[R_t^\star - R_t \mid \mathcal{H}_{t-1}\right]$$

where the joint density of the full observation history $\mathcal{H}_t$ up to round $t \in [T]$ is defined in Equation (J.16) and $R_t^\star = \max\{R(t,1), R(t,-1)\}$. For the class of functions $\mathcal{F}_{\beta,L}$ satisfies Assumptions 4.2, under Assumption J.1 and Equation (J.15), the expected cumulative regret is lower bounded by

$$\sup_{\boldsymbol{\theta}\in\Theta, f\in\mathcal{F}_{\beta,L}} \mathcal{R}_T(\boldsymbol{\theta}, f) = \Theta\left(TN^{-\frac{\beta}{2\beta+d_{\texttt{non}}}}\right) + \Theta\left(\sqrt{d_{\texttt{lin}}T}\right). \tag{J.17}$$

K Proof of Theorem J.1

Proof. We now introduce upper bound on the KL divergence between two distributions $\mathbb{P}_{\boldsymbol{\theta},f,\pi}^{(t)}$ and $\mathbb{P}_{\boldsymbol{\theta}',f',\pi}^{(t)}$ under a fixed policy π.

Lemma K.1. For any $t \in [T]$, let

$$\begin{aligned}\mathcal{K}_{\boldsymbol{\theta},t}^{(\texttt{non})}(f, f') := C_0 \sum_{i=1}^{N} \mathbb{E}_f\left[f\left(\boldsymbol{O}_i^{(0)}\right) - f'\left(\boldsymbol{O}_i^{(0)}\right)\right]^2 \\ + C_D \sum_{\tau=1}^{t} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(\tau-1)} \mathbb{E}\left[\left(f(\boldsymbol{O}_\tau) - f'(\boldsymbol{O}_\tau)\right)^2 \mathbb{1}\{A_\tau = 1, V_\tau = 0\} \mid \mathcal{H}_{\tau-1}\right]\end{aligned} \tag{K.1}$$

and

$$\mathcal{K}_{f,t}^{(\texttt{lin})}(\boldsymbol{\theta}, \boldsymbol{\theta}') := C_D \sum_{\tau=1}^{t} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(\tau-1)} \mathbb{E}\left[\mathbb{1}(V_\tau = 1)\left(\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_\tau^{(A_\tau)} - \boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_\tau^{(A_\tau)}\right)^2 \mid \mathcal{H}_{\tau-1}\right]. \tag{K.2}$$

Then for any fixed policy π and distribution $\mathbb{P}_{\boldsymbol{\theta},f,\pi}^{(t)}$ whose density is specified in Equation (J.16),

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\theta},f,\pi}^{(t)} \| \mathbb{P}_{\boldsymbol{\theta}',f',\pi}^{(t)}\right) \le \mathcal{K}_{\boldsymbol{\theta},t}^{(\texttt{non})}(f, f') + \mathcal{K}_{f,t}^{(\texttt{lin})}(\boldsymbol{\theta}, \boldsymbol{\theta}'). \tag{K.3}$$

The proof of all the technical lemmas are deferred to Section K.1. Lemma K.2 controls $\mathcal{K}_{f,t}^{(\texttt{lin})}$, while $\mathcal{K}_{\boldsymbol{\theta},t}^{(\texttt{non})}(f, f')$ is controlled by Equation (K.31) in the proof of Lemma K.5.

Lemma K.2. Let $\mathbf{e}_i$ be the standard basis in $\mathbb{R}^{d_{\texttt{lin}}}$, with Suppose that $\mathbb{P}_Q$ is given by

$$\mathbb{P}(\boldsymbol{Q} = \mathbf{e}_i) = \frac{1}{d_{\texttt{lin}}} \quad \text{for } i \in [d_{\texttt{lin}}]. \tag{K.4}$$

For any $t \in [T]$

$$\mathcal{K}_{f,t}^{(\texttt{lin})}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{C_D t \left\|\boldsymbol{\theta}_Q - \boldsymbol{\theta}_Q'\right\|_2^2}{2d_{\texttt{lin}}}. \tag{K.5}$$

We turn our attention to the expected cumulative regret. Let

$$R_t^\star = \max\{R(t,1), R(t,-1)\}$$

and

$$\boldsymbol{Q}_t^\star = \underset{\boldsymbol{Q}_t^{(a)} \in \left\{\boldsymbol{Q}_t^{(-1)}, \boldsymbol{Q}_t^{(1)}\right\}}{\operatorname{argmax}} \boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(a)}, \qquad f^\star = \underset{f_t^{(a)} \in \left\{f_t^{(-1)}, f_t^{(1)}\right\}}{\operatorname{argmax}} f^{(a)}(\boldsymbol{O}_t).$$

For the distribution in Equation (J.16), the expected cumulative regret is given by

$$\begin{aligned}\mathcal{R}_T(\boldsymbol{\theta}, f) &= \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(t-1)} \mathbb{E}\left[R_t^\star - R_t \mid \mathcal{H}_{t-1}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(t-1)} \mathbb{E}\left[\mathbb{1}\left(V_t = 1\right)\left(\boldsymbol{Q}_t^\star - \boldsymbol{Q}_t^{(A_t)}\right)^\top \boldsymbol{\theta}_Q \mid \mathcal{H}_{t-1}\right] \\ &\quad + \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(t-1)} \mathbb{E}\left[\mathbb{1}\left(V_t = 0\right)\left(f^\star(\boldsymbol{O}_t) - f^{(A_t)}(\boldsymbol{O}_t)\right) \mid \mathcal{H}_{t-1}\right] \\ &=: \mathcal{R}_T^{(\mathtt{lin})}(\boldsymbol{\theta}, f) + \frac{1}{2}\mathcal{R}_T^{(\mathtt{non})}(\boldsymbol{\theta}, f),\end{aligned} \tag{K.6}$$

Lemma K.3 controls $\mathcal{R}_T^{(\mathtt{lin})}(\boldsymbol{\theta}, f)$.

Lemma K.3. Suppose that $0 < d_{\mathtt{lin}} < c\sqrt{T}$ for some sufficiently small constant $c > 0$. For any $f \in \mathcal{F}_{\beta,L}$,

$$\sup_{\theta\in\Theta} \mathcal{R}_T^{(\mathtt{lin})}(\boldsymbol{\theta}, f) \geq \frac{\sqrt{d_{\mathtt{lin}} T}}{4} \exp\left\{-2C_D\right\}. \tag{K.7}$$

For $\mathcal{R}_T^{(\mathtt{non})}(\boldsymbol{\theta}, f)$, we first construct a packing set for $\mathcal{F}_{\beta,L}$. For any multi-index $\boldsymbol{k} \in [M]^{d_{\mathtt{non}}}$, define the hypercube

$$B_{\boldsymbol{k}} = \left\{\mathbf{o} \in \mathcal{O} : \frac{\boldsymbol{k}_l - 1}{M} \leq o_l \leq \frac{\boldsymbol{k}_l}{M}, l \in [d_{\mathtt{non}}]\right\} \subset \mathbb{R}^{d_{\mathtt{non}}},$$

where $M > 0$ is specified later in Equation (K.45). We index the bins by integers $k \in [M^{d_{\mathtt{non}}}]$ via the mapping

$$k = 1 + \sum_{l=1}^{d_{\mathtt{non}}} (\boldsymbol{k}_l - 1) M^{l-1}$$

and write B_k as a shorthand for $B_{\boldsymbol{k}}$. For each bin B_k, define its center $b_k \in \mathbb{R}^{d_{\mathtt{non}}}$ coordinate-wise as

$$b_{k,l} = \frac{\boldsymbol{k}_l}{M} - \frac{1}{2M}, \quad l \in [d_{\mathtt{non}}].$$

This yields a regular grid of centers $\mathcal{B} = \{b_1, \cdots, b_{M^{d_{\mathtt{non}}}}\}$ across the domain. Next, we define a smooth, compactly supported bump function $\phi_\beta : \mathbb{R}^{d_{\mathtt{non}}} \to [0, 1]$ by

$$\phi_\beta(\mathbf{o}) = \begin{cases} (1 - \|\mathbf{o}\|_\infty)^\beta & \text{if } 0 \leq \|\mathbf{o}\|_\infty \leq 1, \\ 0 & \text{if } \|\mathbf{o}\|_\infty > 1. \end{cases} \tag{K.8}$$

We will now construct localized perturbation functions supported within each bin. Let

$$m = \lceil c_m M^{d_{\mathtt{non}}} \rceil \tag{K.9}$$

for some sufficiently small constant $c_m > 0$. Define $\Omega_m = \{\pm 1\}^m$. For any $\boldsymbol{\omega} \in \Omega_m$, define the function

$$f_{\boldsymbol{\omega}}(\mathbf{o}) = \sum_{j=1}^{m} \omega_j \varphi_j(\mathbf{o}), \tag{K.10}$$

where each component function φ_j is defined as

$$\varphi_j(\mathbf{o}) = M^{-\beta} C_\phi \phi_\beta\left(2M[\mathbf{o} - \boldsymbol{b}_j]\right) \mathbb{1}(\mathbf{o} \in B_j) \tag{K.11}$$

and $C_\phi > 0$ is a constant specified in Equation (K.12). Note that for any $\mathbf{o} \in B_j$, the rescaled argument satisfies $2M(\mathbf{o} - \boldsymbol{b}_j) \in [-1, 1]^{d_{\mathtt{non}}}$, so $\|2M(\mathbf{o} - \boldsymbol{b}_j)\|_\infty \in [0, 1]$, ensuring that φ_j in Equation (K.11) is well-defined.

The function $f_{\boldsymbol{\omega}}$ is thus a linear combination of localized, smooth bump functions with disjoint supports. Lemma K.4 establishes that each $f_{\boldsymbol{\omega}}$ lies in $\mathcal{F}_{\beta,L}$ for a suitable choice of constant L.

Lemma K.4. Suppose that $\beta \in (0,1]$. For any $\boldsymbol{\omega} \in \Omega_m = \{\pm 1\}^m$, the function $f_{\boldsymbol{\omega}}$ defined in Equation (K.10) belongs to the smoothness class $\mathcal{F}_{\beta,L}$ with $L = \beta 2^\beta C_\phi > 0$.

Hence, for a given parameter $L > 0$, we set

$$C_\phi := \frac{L}{\beta 2^\beta}. \tag{K.12}$$

Based on this choice of packing set, Lemma K.5 controls $\mathcal{R}_T^{(\mathtt{non})}(\boldsymbol{\theta}, f)$.

Lemma K.5. Suppose that $d_{\mathtt{non}} > 0$ is a constant. For any fixed $\boldsymbol{\theta} \in \Theta$,

$$\sup_{f \in \mathcal{F}_{\beta,L}} \mathcal{R}_T^{(\mathtt{non})}(\boldsymbol{\theta}, f) = \Theta\left(TN^{-\frac{\beta}{2\beta+d_{\mathtt{non}}}}\right), \tag{K.13}$$

where $\mathcal{R}^{(\mathtt{non})}$ is defined in Equation (K.6).

Taking Equations (K.13) with (K.19) into Equation (K.6), we have

$$\sup_{\boldsymbol{\theta} \in \Theta, f \in \mathcal{F}_{\beta,L}} \mathcal{R}_T(\boldsymbol{\theta}, f) = \Theta\left(TN^{-\frac{\beta}{2\beta+d_{\mathtt{non}}}}\right) + \Theta\left(\sqrt{d_{\mathtt{lin}}T}\right),$$

establishing the desired result in Equation (J.17).

■

K.1 Proof of Technical Lemmas

In this section, we present the proof of the Lemmas K.1-K.5 used in the proof of Theorem J.1. We will frequently use the Bretagnolle-Huber inequality given in the following theorem.

Theorem K.1 (Bretagnolle-Huber inequality)**.** Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,

$$\mathbb{P}(A) + \mathbb{Q}(A^c) \geq \frac{1}{2}\exp\left(-\mathrm{KL}(\mathbb{P}\|\mathbb{Q})\right).$$

Proof. See Theorem 14.2 in Lattimore and Szepesvári (2020). ■

K.1.1 Proof of Lemma K.1

Proof. Recall the definition of $\mathbb{P}^{(t)}_{\boldsymbol{\theta},f,\pi}$ as stated in Equation (J.16). Eliminating the shared terms, it follows that

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^{(t)}_{\boldsymbol{\theta},f,\pi} \| \mathbb{P}^{(t)}_{\boldsymbol{\theta}',f',\pi}\right) &= \mathbb{E}^{(t)}_{\boldsymbol{\theta},f,\pi}\left[\log \frac{d\mathbb{P}^{(t)}_{\boldsymbol{\theta},f,\pi}}{d\mathbb{P}^{(t)}_{\boldsymbol{\theta}',f',\pi}}\right] \\
&= \underbrace{\sum_{i=1}^{N} \mathbb{E}_f\left[\log \frac{p_f}{p_{f'}}\left(\boldsymbol{Q}_i^{(0)}, \boldsymbol{O}_i^{(0)}, W_i^{(0)}\right)\right]}_{\mathcal{K}_1} + \underbrace{\sum_{\tau=1}^{t} \mathbb{E}^{(\tau)}_{\boldsymbol{\theta},f,\pi}\left[\log \frac{p_{\boldsymbol{\theta},f}\left(R_\tau \mid \boldsymbol{Q}_\tau, \boldsymbol{O}_\tau, A_\tau\right)}{p_{\boldsymbol{\theta}',f'}\left(R_\tau \mid \boldsymbol{Q}_\tau, \boldsymbol{O}_\tau, A_\tau\right)}\right]}_{\mathcal{K}_2}.
\end{aligned} \tag{K.14}$$

For $\mathcal{K}_1$ in Equation (K.14), by the KL divergence assumption in Equation (J.13), we have

$$\mathcal{K}_1 = \sum_{i=1}^{N} \mathbb{E}_f\left[\log \frac{p^{(0)}_{f(\boldsymbol{O}_i^{(0)})}}{p^{(0)}_{f'(\boldsymbol{O}_i^{(0)})}}\left(W_i^{(0)}\right)\right] \leq C_0 \sum_{i=1}^{N} \mathbb{E}_f\left[f\left(\boldsymbol{O}_i^{(0)}\right) - f'\left(\boldsymbol{O}_i^{(0)}\right)\right]^2. \tag{K.15}$$

To control $\mathcal{K}_2$, we note that

$$\begin{aligned}
&\mathbb{E}^{(t)}_{\boldsymbol{\theta},f,\pi}\left[\log\frac{p_{\boldsymbol{\theta},f}\left(R_t \mid \boldsymbol{Q}_t,\boldsymbol{O}_t,A_t\right)}{p_{\boldsymbol{\theta}',f'}\left(R_t \mid \boldsymbol{Q}_t,\boldsymbol{O}_t,A_t\right)}\right]\\
&=\mathbb{E}^{(t)}_{\boldsymbol{\theta},f,\pi}\left[\log\frac{p_{\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(A_t)}}\left(R_t\right)}{p_{\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_t^{(A_t)}}\left(R_t\right)}\mathbb{1}\{V_t=0\}\right]+\mathbb{E}^{(t)}_{\boldsymbol{\theta},f,\pi}\left[\log\frac{p_{f^{(A_t)}(\boldsymbol{O}_t)}\left(R_t\right)}{p_{f'^{(A_t)}(\boldsymbol{O}_t)}\left(R_t\right)}\mathbb{1}\{V_t=1\}\right]\\
&=\sum_{a\in\mathcal{A}}\mathbb{E}^{(t-1)}_{\boldsymbol{\theta},f,\pi}\mathbb{E}\left[\mathbb{1}(A_t=a,V_t=0)\,\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(a)}}\|\mathbb{P}_{\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_t^{(a)}}\right)\mid\mathcal{H}_{t-1}\right]\\
&\quad+\sum_{a\in\mathcal{A}}\mathbb{E}^{(t-1)}_{\boldsymbol{\theta},f,\pi}\mathbb{E}\left[\mathbb{1}(A_t=a,V_t=1)\,\mathrm{KL}\left(\mathbb{P}_{f^{(a)}(\boldsymbol{O}_t)}\|\mathbb{P}_{f'^{(a)}(\boldsymbol{O}_t)}\right)\mid\mathcal{H}_{t-1}\right],
\end{aligned}$$

where the last equality follows from the definition of KL divergence. Taking the above display and Equation (K.15) into Equation (K.14) yields that

$$\begin{aligned}
\mathrm{KL}\left(\mathbb{P}^{(t)}_{\boldsymbol{\theta},f,\pi},\mathbb{P}^{(t)}_{\boldsymbol{\theta}',f',\pi}\right)&=\mathcal{K}_1+\sum_{\tau=1}^{t}\sum_{a\in\mathcal{A}}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f,\pi}\mathbb{E}\left[\mathbb{1}(A_\tau=a,V_\tau=0)\,\mathrm{KL}\left(\mathbb{P}_{f^{(a)}(\boldsymbol{O}_\tau)}\|\mathbb{P}_{f'^{(a)}(\boldsymbol{O}_\tau)}\right)\mid\mathcal{H}_{\tau-1}\right]\\
&\quad+\sum_{\tau=1}^{t}\sum_{a\in\mathcal{A}}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f,\pi}\mathbb{E}\left[\mathbb{1}(A_\tau=a,V_\tau=1)\,\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_\tau^{(a)}}\|\mathbb{P}_{\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_\tau^{(a)}}\right)\mid\mathcal{H}_{\tau-1}\right]\\
&\le C_0\sum_{i=1}^{N}\mathbb{E}_f\left[f\left(\boldsymbol{O}_i^{(0)}\right)-f'\left(\boldsymbol{O}_i^{(0)}\right)\right]^2\\
&\quad+C_D\sum_{\tau=1}^{t}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f,\pi}\mathbb{E}\left[\left(f(\boldsymbol{O}_\tau)-f'(\boldsymbol{O}_\tau)\right)^2\mathbb{1}\left\{A_\tau=1,V_\tau=0\right\}\mid\mathcal{H}_{\tau-1}\right]\\
&\quad+C_D\sum_{\tau=1}^{t}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f,\pi}\mathbb{E}\left[\mathbb{1}(V_\tau=1)\left(\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_\tau^{(A_\tau)}-\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_\tau^{(A_\tau)}\right)^2\mid\mathcal{H}_{\tau-1}\right],
\end{aligned}\tag{K.16}$$

where the last inequality follows from Assumption J.1 and Equation (J.15). Taking the definition of $\mathcal{K}^{(\mathtt{non})}_{\boldsymbol{\theta},t}$ and $\mathcal{K}^{(\mathtt{lin})}_{f,t}$ in Equations (K.1) and (K.2) into Equation (K.16) yields the desired bound in Equation (K.3). ■

K.1.2 Proof of Lemma K.2

Proof. By definition of $\mathbb{P}_Q$ in Equation (K.4),

$$\left\langle \boldsymbol{Q}_t^{(1)},\boldsymbol{\theta}_Q-\boldsymbol{\theta}_Q'\right\rangle^2=\left\langle \boldsymbol{Q}_t^{(-1)},\boldsymbol{\theta}_Q-\boldsymbol{\theta}_Q'\right\rangle^2=\left\langle \boldsymbol{Q}_t^{(A_t)},\boldsymbol{\theta}_Q-\boldsymbol{\theta}_Q'\right\rangle^2.$$

It follows that for any $a\in\mathcal{A}$,

$$\mathbb{E}_Q\left[\left\langle \boldsymbol{Q}_t^{(a)},\boldsymbol{\theta}_Q-\boldsymbol{\theta}_Q'\right\rangle^2\right]=\frac{\left\|\boldsymbol{\theta}_Q-\boldsymbol{\theta}_Q'\right\|_2^2}{d_{\mathtt{lin}}},$$

and

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}(V_t=1)\left(\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(A_t)}-\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_t^{(A_t)}\right)^2\mid\mathcal{H}_{t-1}\right]&=\mathbb{E}\left[\mathbb{1}(V_t=1)\left(\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(1)}-\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_t^{(1)}\right)^2\mid\mathcal{H}_{t-1}\right]\\
&=\frac{1}{2}\mathbb{E}\left[\left(\boldsymbol{\theta}_Q^\top \boldsymbol{Q}_t^{(1)}-\boldsymbol{\theta}_Q'^\top \boldsymbol{Q}_t^{(1)}\right)^2\mid\mathcal{H}_{t-1},V_t=1\right]\\
&=\frac{\left\|\boldsymbol{\theta}_Q-\boldsymbol{\theta}_Q'\right\|_2^2}{2d_{\mathtt{lin}}}.
\end{aligned}$$

Thus, combining the above display with Equation (K.2) gives the desired result in Equation (K.5). ■

K.1.3 Proof of Lemma K.3

Proof. Noting that

$$\begin{aligned}\left(\boldsymbol{Q}_t^\star - \boldsymbol{Q}_t^{(A_t)}\right)^\top \boldsymbol{\theta}_Q &= 2\sum_{i=1}^{d_{\mathtt{lin}}} \mathbb{1}\{\boldsymbol{Q}_t = \boldsymbol{e}_i\}\mathbb{1}\{A_t \neq \operatorname{sign}(\boldsymbol{\theta}_{Q,i})\}|\boldsymbol{\theta}_{Q,i}| \\ &= 2\sqrt{\frac{d_{\mathtt{lin}}}{T}}\sum_{i=1}^{d_{\mathtt{lin}}} \mathbb{1}\{\boldsymbol{Q}_t = \boldsymbol{e}_i\}\mathbb{1}\{A_t \neq \operatorname{sign}(\boldsymbol{\theta}_{Q,i})\}\end{aligned}$$

where the last equality follows from the fact that $|\theta_{Q,i}| = \sqrt{d_{\mathtt{lin}}/T}$ as given in Equation (J.7). Recall the definition of $\mathcal{R}_t^{\mathtt{lin}}$ in Equation (K.6). Combined with the above display, it follows that

$$\begin{aligned}\mathcal{R}_t^{\mathtt{lin}}(\boldsymbol{\theta}, f) &= 2\sqrt{\frac{d_{\mathtt{lin}}}{T}}\sum_{\tau=1}^{t}\sum_{i=1}^{d_{\mathtt{lin}}} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(\tau-1)}\mathbb{E}\left[\mathbb{1}\left\{A_\tau \neq \operatorname{sign}(\boldsymbol{\theta}_{Q,i}), V_\tau = 1, \boldsymbol{Q}_\tau = \boldsymbol{e}_i\right\} \mid \mathcal{H}_{\tau-1}\right] \\ &= \sqrt{\frac{1}{d_{\mathtt{lin}}T}}\sum_{i=1}^{d_{\mathtt{lin}}}\sum_{\tau=1}^{t} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(\tau-1)}\mathbb{E}\left[\mathbb{1}\left(A_\tau \neq \operatorname{sign}(\boldsymbol{\theta}_{Q,i})\right) \mid \mathcal{H}_{\tau-1}, \boldsymbol{Q}_\tau = \boldsymbol{e}_i\right]\end{aligned} \tag{K.17}$$

where the last equality follows from

$$\mathbb{P}\left(\boldsymbol{Q}_\tau = \boldsymbol{e}_i, V_\tau = 1 \mid \mathcal{H}_{\tau-1}\right) = \frac{1}{2}\mathbb{P}\left(\boldsymbol{Q}_\tau = \boldsymbol{e}_i\right) = \frac{1}{2d_{\mathtt{lin}}}$$

as specified by the data generating process in Definition J.1 and Equation (K.4). Consider $\boldsymbol{\theta}'_Q \in \mathbb{R}^{d_{\mathtt{lin}}}$ such that $\theta'_{Q,j} = \theta_{Q,j}$ for all $j \neq i$ and $\theta'_{Q,i} = -\theta_{Q,i}$. Let $\mathbb{P}_{Q,i}^{(t-1)} := \mathbb{P}\left(\cdot \mid \mathcal{H}_{t-1}, \boldsymbol{Q}_t = \boldsymbol{e}_i\right)$. Continuing from Equation (K.17), by the Bretagnolle-Huber inequality as stated in Theorem K.1, we have for any $t \in [T]$,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left(A_t \neq \operatorname{sign}(\theta_{Q,i})\right) \mid \mathcal{H}_{t-1}, \boldsymbol{Q}_t = \boldsymbol{e}_i\right] + \mathbb{E}_{\boldsymbol{\theta}',f,\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left(A_t \neq \operatorname{sign}(\theta'_{Q,i})\right) \mid \mathcal{H}_{t-1}, \boldsymbol{Q}_t = \boldsymbol{e}_i\right] \\ &\geq \frac{1}{2}\exp\left\{-\operatorname{KL}\left(\mathbb{P}_{\boldsymbol{\theta},f,\pi}^{(t-1)} \times \mathbb{P}_{Q,i}^{(t-1)} \| \mathbb{P}_{\boldsymbol{\theta}',f,\pi}^{(t-1)} \times \mathbb{P}_{Q,i}^{(t-1)}\right)\right\} \\ &= \frac{1}{2}\exp\left\{-\operatorname{KL}\left(\mathbb{P}_{\boldsymbol{\theta},f,\pi}^{(t-1)} \| \mathbb{P}_{\boldsymbol{\theta}',f,\pi}^{(t-1)}\right)\right\} \\ &\geq \frac{1}{2}\exp\left\{-\mathcal{K}_{\boldsymbol{\theta},t}^{(\mathtt{non})}(f,f) - \mathcal{K}_{f,t}^{(\mathtt{lin})}(\boldsymbol{\theta},\boldsymbol{\theta}')\right\},\end{aligned}$$

where the last inequality follows from Equations (K.16), (K.1) and (K.2). Since $\mathcal{K}_{\boldsymbol{\theta},t}^{(\mathtt{non})}(f,f) = 0$, it follows from the above display that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left(A_t \neq \operatorname{sign}(\theta_{Q,i})\right) \mid \mathcal{H}_{t-1}, \boldsymbol{Q}_t = \boldsymbol{e}_i\right] + \mathbb{E}_{\boldsymbol{\theta}',f,\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left(A_t \neq \operatorname{sign}(\theta'_{Q,i})\right) \mid \mathcal{H}_{t-1}, \boldsymbol{Q}_t = \boldsymbol{e}_i\right] \\ &\geq \frac{1}{2}\exp\left\{-\mathcal{K}_{f,t}^{(\mathtt{lin})}(\boldsymbol{\theta},\boldsymbol{\theta}')\right\} = \frac{1}{2}\exp\left\{-\frac{C_D t\left\|\boldsymbol{\theta}_Q - \boldsymbol{\theta}'_Q\right\|_2^2}{2d_{\mathtt{lin}}}.\right\}\end{aligned} \tag{K.18}$$

where the last inequality follows from Equation (K.5). Let $\Theta_{d_{\mathtt{lin}}} \subset \mathbb{R}^{d_{\mathtt{lin}}}$ denote the set of all vectors whose coordinate are either $\beta := \sqrt{d_{\mathtt{lin}}/T}$ or $-\beta$, i.e.,

$$\Theta_{d_{\mathtt{lin}}} := \left\{\boldsymbol{\theta}_Q \in \mathbb{R}^{d_{\mathtt{lin}}} : \theta_{Q,i} \in \{\pm\beta\}, \forall i \in [d_{\mathtt{lin}}]\right\}.$$

For any vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and $j \in [d]$, denote $(\theta_1, \cdots, \theta_{j-1}, \theta_{j+1}, \cdots, \theta_d) \in \mathbb{R}^{d-1}$ as $\boldsymbol{\theta}_{[-j]}$ and $\boldsymbol{\theta}_{[-j]}^i := (\theta_1, \cdots, \theta_{j-1}, i, \theta_1, \cdots, \theta_d) \in \mathbb{R}^d$ for $i \in \mathbb{R}$. Applying an average hammer over all $\boldsymbol{\theta}_Q \in \Theta_{d_{\mathtt{lin}}}$, which satis-

fies $|\Theta_{d_{\texttt{lin}}}| = 2^{d_{\texttt{lin}}}$, it follows from Equation (K.17) that

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_t^{\texttt{lin}}(\boldsymbol{\theta}, f) &\geq \frac{1}{|\Theta_{d_{\texttt{lin}}}|} \sum_{\boldsymbol{\theta}_Q \in \Theta_{d_{\texttt{lin}}}} \sqrt{\frac{1}{d_{\texttt{lin}} T}} \sum_{i=1}^{d_{\texttt{lin}}} \sum_{\tau=1}^{t} \mathbb{E}_{\boldsymbol{\theta},f,\pi}^{(\tau-1)} \mathbb{E}\left[\mathbb{1}\left(A_\tau \neq \operatorname{sign}(\theta_{Q,i})\right) \mid \mathcal{H}_{\tau-1}, \boldsymbol{Q}_\tau = \boldsymbol{e}_i\right] \\
&\geq \frac{1}{2^{d_{\texttt{lin}}}} \sum_{i=1}^{d_{\texttt{lin}}} \sum_{\boldsymbol{\theta}^j_{Q,[-i]} \in \Theta_{d_{\texttt{lin}}}} \sum_{j\in\{\pm\beta\}} \sqrt{\frac{1}{d_{\texttt{lin}} T}} \sum_{\tau=1}^{t} \mathbb{E}_{\boldsymbol{\theta}^j_{Q,[-i]},f,\pi}^{(\tau-1)} \mathbb{E}\left[\mathbb{1}\left(A_\tau \neq \operatorname{sign}(\theta_{Q,i})\right) \mid \mathcal{H}_{\tau-1}, \boldsymbol{Q}_\tau = \boldsymbol{e}_i\right] \\
&\overset{(i)}{\geq} \frac{1}{2^{d_{\texttt{lin}}+1}} \sqrt{\frac{1}{d_{\texttt{lin}} T}} \sum_{i=1}^{d_{\texttt{lin}}} \sum_{\boldsymbol{\theta}^j_{Q,[-i]} \in \Theta_{d_{\texttt{lin}}}} \sum_{\tau=1}^{t} \exp\left\{-\frac{C_D t \left\|\boldsymbol{\theta}^{\beta}_{Q,[-i]} - \boldsymbol{\theta}^{-\beta}_{Q,[-i]}\right\|_2^2}{2 d_{\texttt{lin}}}\right\} \\
&\overset{(ii)}{=} \frac{1}{2^{d_{\texttt{lin}}+1}} \sqrt{\frac{1}{d_{\texttt{lin}} T}} \sum_{i=1}^{d_{\texttt{lin}}} \sum_{\boldsymbol{\theta}^j_{Q,[-i]} \in \Theta_{d_{\texttt{lin}}}} \sum_{\tau=1}^{t} \exp\left\{-\frac{2 C_D t}{T}\right\} \\
&= \frac{t}{4} \sqrt{\frac{d_{\texttt{lin}}}{T}} \exp\left\{-\frac{2 C_D t}{T}\right\}
\end{aligned} \tag{K.19}$$

where inequality (i) follows from Equation (K.18) and equality (ii) follows from

$$\left\|\boldsymbol{\theta}^{\beta}_{Q,[-i]} - \boldsymbol{\theta}^{-\beta}_{Q,[-i]}\right\|_2^2 = 4\beta^2 = \frac{4 d_{\texttt{lin}}}{T}.$$

Taking $t = T$ in Equation (K.19) yields the result in Equation (K.7). ■

K.1.4 Proof of Lemma K.4

Proof. To verify that $f_{\boldsymbol{\omega}} \in \mathcal{F}_{\beta,L}$ for some suitable $L > 0$, we first note that for any $0 \leq x, y \leq 1$

$$\left|x^\beta - y^\beta\right| \leq \beta \left|x - y\right|. \tag{K.20}$$

For $\mathbf{o}, \mathbf{o}' \in B_k$, by definition of $f_{\boldsymbol{\omega}}$ in Equation (K.10), we have

$$\begin{aligned}
|f_{\boldsymbol{\omega}}(\mathbf{o}) - f_{\boldsymbol{\omega}}(\mathbf{o}')| &= |\varphi_k(\mathbf{o}) - \varphi_k(\mathbf{o}')| \\
&= M^{-\beta} C_\phi |\phi_\beta(2M[\mathbf{o} - \boldsymbol{b}_k]) - \phi_\beta(2M[\mathbf{o}' - \boldsymbol{b}_k])| \\
&= M^{-\beta} C_\phi \left|(1 - \|2M[\mathbf{o} - \boldsymbol{b}_k]\|_\infty)^\beta - (1 - \|2M[\mathbf{o}' - \boldsymbol{b}_k]\|_\infty)^\beta\right|
\end{aligned}$$

where the second equality follows from the definition of φ_k in Equation (K.11) and the last equality follows from the definition of ϕ_β in Equation (K.8). Continuing from the above display,

$$\begin{aligned}
|f_{\boldsymbol{\omega}}(\mathbf{o}) - f_{\boldsymbol{\omega}}(\mathbf{o}')| &= 2^\beta C_\phi \left|\left(\frac{1}{2M} - \|\mathbf{o} - \boldsymbol{b}_k\|_\infty\right)^\beta - \left(\frac{1}{2M} - \|\mathbf{o} - \boldsymbol{b}_k\|_\infty\right)^\beta\right| \\
&\overset{(i)}{\leq} 2^\beta \beta C_\phi \left|\|\mathbf{o} - \boldsymbol{b}_k\|_\infty - \|\mathbf{o}' - \boldsymbol{b}_k\|_\infty\right| \\
&\overset{(ii)}{\leq} 2^\beta \beta C_\phi \|\mathbf{o} - \mathbf{o}'\|_\infty \leq 2^\beta \beta C_\phi \|\mathbf{o} - \mathbf{o}'\|_2
\end{aligned} \tag{K.21}$$

where equality (i) follows from Equation (K.20) and inequality (ii) follows from the triangle inequality.

If $\mathbf{o}, \mathbf{o}'$ are in different bins $B_k, B_{k'}$, then we can pick $\mathbf{p}_k \in B_k$ and $\mathbf{p}_{k'} \in B_{k'}$ each on the boundary of B_k and $B_{k'}$, such that both $f(\mathbf{p}_k) = 0$ and $f(\mathbf{p}_{k'}) = 0$, and

$$\begin{aligned}
|f(\mathbf{o}) - f(\mathbf{o}')| &\leq \max\left\{|f(\mathbf{o}) - f(\mathbf{p}_k)|, |f(\mathbf{o}') - f(\mathbf{p}_{k'})|\right\} \\
&\leq 2^\beta \beta C_\phi \max\left\{\|\mathbf{o} - \mathbf{p}_k\|_\infty, \|\mathbf{o}' - \mathbf{p}_{k'}\|_\infty\right\}
\end{aligned} \tag{K.22}$$

where the last inequality follows from Equation (K.21). We can pick p_k and $p_{k'}$ so that

$$\|\mathbf{o} - \mathbf{o}'\|_\infty \geq \max\left\{\|\mathbf{o} - \mathbf{p}_k\|_\infty, \|\mathbf{o}' - \mathbf{p}_{k'}\|_\infty\right\}$$

it then follows from Equation (K.22) that

$$|f(\mathbf{o}) - f(\mathbf{o}')| \leq 2^{\beta}\beta C_{\phi} \left\|\mathbf{o} - \mathbf{o}'\right\|_{\infty} \leq 2^{\beta}\beta C_{\phi} \left\|\mathbf{o} - \mathbf{o}'\right\|_{2}.$$

Combining the above display with Equation (K.21) finishes the proof. ■

K.1.5 Proof of Lemma K.5

Proof. Let

$$\widetilde{B}_j = B_j \cap \left\{\mathbf{o} : \phi_{\beta}(2M(\mathbf{o} - \boldsymbol{b}_j)) \geq \delta M^{\beta}\right\}.$$

For any $\mathbf{o} \in \widetilde{B}_j$, it follows from Equation (K.10) that

$$f_{\boldsymbol{\omega}}(\mathbf{o}) = \omega_j \varphi_j(\mathbf{o}) \geq C_{\phi} M^{-\beta} \delta M^{\beta} = \delta C_{\phi}. \tag{K.23}$$

For any $\boldsymbol{\omega} \in \Omega_m$ and $\delta > 0$, combining Equation (K.23) with $\mathcal{R}_T^{\text{non}}$ as defined in Equation (K.6) yields that

$$\begin{aligned}
\mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f_{\boldsymbol{\omega}}) &= \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\mathbb{1}\left\{A_t \neq \operatorname{sign}\left(f_{\boldsymbol{\omega}}(\boldsymbol{O}_t)\right), V_t = 0\right\} \left|f_{\boldsymbol{\omega}}(\boldsymbol{O}_t)\right| \mid \mathcal{H}_{t-1}\right] \\
&= \sum_{t=1}^{T} \sum_{j=1}^{m} \mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\mathbb{1}\left\{A_t \neq \operatorname{sign}\left(f_{\boldsymbol{\omega}}(\boldsymbol{O}_t)\right), V_t = 0\right\} \mathbb{1}\left\{\boldsymbol{O}_t \in B_j\right\} \left|f_{\boldsymbol{\omega}}(\boldsymbol{O}_t)\right| \mid \mathcal{H}_{t-1}\right] \\
&\geq C_{\phi}\delta \sum_{j=1}^{m} \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\mathbb{1}\left\{A_t \neq \omega_j, \boldsymbol{O}_t \in \widetilde{B}_j, V_t = 0\right\} \mid \mathcal{H}_{t-1}\right].
\end{aligned} \tag{K.24}$$

For any $\boldsymbol{\omega} \in \Omega_m$ and $\boldsymbol{O} \sim \mathbb{P}_O$, we have for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{O} \in \widetilde{B}_1\right) &= \mathbb{P}\left(\phi_{\beta}\left(2M[\boldsymbol{O} - \boldsymbol{b}_1]\right) \geq \delta M^{\beta}, \boldsymbol{O} \in B_1\right) \\
&= \int_{B_1} \mathbb{1}\left(\phi_{\beta}(2M(\mathbf{o} - \boldsymbol{b}_1)) \geq \delta M^{\beta}\right) d\mathbf{o} \\
&= \int_{B_1} \mathbb{1}\left\{(1 - 2M\|\mathbf{o} - \boldsymbol{b}_1\|_{\infty})^{\beta} \geq \delta M^{\beta}\right\} d\mathbf{o} \\
&= \int_{\left[0, \frac{1}{M}\right]^{d_{\text{non}}}} \mathbb{1}\left(\max_{l \in [d_{\text{non}}]} \left|o_l - \frac{1}{2M}\right| \leq \frac{1}{2M} - \frac{1}{2}\delta^{1/\beta}\right) d\mathbf{o} \\
&= \int_{\left[0, \frac{1}{M}\right]^{d_{\text{non}}}} \mathbb{1}\left(\mathbf{o} \in \left[\frac{1}{2}\delta^{1/\beta}, \frac{1}{M} - \frac{1}{2}\delta^{1/\beta}\right]^{d_{\text{non}}}\right) d\mathbf{o} \\
&= \left(\frac{1}{M} - \delta^{1/\beta}\right)^{d_{\text{non}}}.
\end{aligned} \tag{K.25}$$

The same probability holds for all other $\widetilde{B}_j$ where $j \in [M^d]$.
To handle $\mathcal{K}^{(\text{non})}$ as defined in Equation (K.1), take $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ so that they only differ in ω_j, we have

$$|f_{\boldsymbol{\omega}}(\mathbf{o}) - f_{\boldsymbol{\omega}'}(\mathbf{o})| = 2\varphi_j(\mathbf{o})$$

and

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\left(f_{\boldsymbol{\omega}}(\boldsymbol{O}_t) - f_{\boldsymbol{\omega}'}(\boldsymbol{O}_t)\right)^2 \mathbb{1}\left\{A_t = 1, V_t = 0\right\} \mid \mathcal{H}_{t-1}\right] \\
&= 4\mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\varphi_j(\boldsymbol{O}_t)^2 \mathbb{1}\left\{A_t = 1, \boldsymbol{O}_t \in B_j\right\} \mid \mathcal{H}_{t-1}\right] \\
&\leq \frac{4C_{\phi}^2\delta^2}{M^{d_{\text{non}}}} + 4\mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\varphi_j(\boldsymbol{O}_t)^2 \mathbb{1}\left\{A_t = 1, \boldsymbol{O}_t \in \widetilde{B}_j\right\} \mid \mathcal{H}_{t-1}\right] \\
&\leq \frac{4C_{\phi}^2\delta^2}{M^{d_{\text{non}}}} + 4C_{\phi}^2 M^{-2\beta} \mathbb{E}_{\boldsymbol{\theta}, f_{\boldsymbol{\omega}}, \pi}^{(t-1)} \mathbb{E}\left[\mathbb{1}\left\{A_t = 1, \boldsymbol{O}_t \in \widetilde{B}_j\right\} \mid \mathcal{H}_{t-1}\right]
\end{aligned} \tag{K.26}$$

where the first equality follows from the fact that $\boldsymbol{O}_t \in B_j$ already implies $V_t = 0$. Similarly, for any $i \in [N]$, applying the above argument with Equation (K.25) to the pretrained data yields

$$\mathbb{E}_{f_{\boldsymbol{\omega}}}\left[\left(f_{\boldsymbol{\omega}}\left(\boldsymbol{O}_i^{(0)}\right) - f_{\boldsymbol{\omega}'}\left(\boldsymbol{O}_i^{(0)}\right)\right)^2\right] \leq \frac{4C_\phi^2\delta^2}{M^{d_{\text{non}}}} + 4C_\phi^2 M^{-2\beta}\left(\frac{1}{M} - \delta^{1/\beta}\right)^{d_{\text{non}}}. \tag{K.27}$$

Pick δ_0 so that

$$M^{2\beta}\delta_0^2 \asymp \left(1 - M\delta_0^{1/\beta}\right)^{d_{\text{non}}}.$$

Let κ_0 be the solution to the equation

$$\kappa^{2\beta} = (1-\kappa)^{d_{\text{non}}}$$

then we set

$$\delta_0 = \kappa_0^\beta M^{-\beta}. \tag{K.28}$$

Under the assumption that d_{non} is a fixed constant in Lemma K.5, we have κ_0 is also a constant and $\delta_0 = \Theta(M^{-\beta})$. Under Equation (K.28), the bound in Equation (K.26) becomes

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}},\pi}^{(t-1)}\mathbb{E}\left[(f_{\boldsymbol{\omega}}(\boldsymbol{O}_t) - f_{\boldsymbol{\omega}'}(\boldsymbol{O}_t))^2\mathbb{1}\left\{A_t = 1, V_t = 0\right\} \mid \mathcal{H}_{t-1}\right] \\
&\lesssim M^{-2\beta-d_{\text{non}}} + M^{-2\beta}\mathbb{E}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}},\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left\{A_t = 1, \boldsymbol{O}_t \in \widetilde{B}_j\right\} \mid \mathcal{H}_{t-1}\right]
\end{aligned} \tag{K.29}$$

and Equation (K.27) becomes

$$\mathbb{E}_{f_{\boldsymbol{\omega}}}\left[\left(f_{\boldsymbol{\omega}}\left(\boldsymbol{O}_i^{(0)}\right) - f_{\boldsymbol{\omega}'}\left(\boldsymbol{O}_i^{(0)}\right)\right)^2\right] \lesssim M^{-2\beta-d_{\text{non}}}. \tag{K.30}$$

Combining Equations (K.29) and (K.30) with Equation (K.1), for any $t \in [T]$ and the choice of δ_0 given in Equation (K.28),

$$\mathcal{K}_{\boldsymbol{\theta},t}^{(\text{non})}(f_{\boldsymbol{\omega}}, f_{\boldsymbol{\omega}'}) \lesssim M^{-2\beta}\sum_{\tau=1}^{t}\mathbb{E}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}},\pi}^{(\tau-1)}\mathbb{E}\left[\mathbb{1}\left\{A_\tau = 1, \boldsymbol{O}_\tau \in \widetilde{B}_j\right\} \mid \mathcal{H}_{\tau-1}\right] + \frac{(t+N)}{M^{2\beta+d_{\text{non}}}}. \tag{K.31}$$

Using an average hammer over $\boldsymbol{\omega} \in \Omega_m$, it follows from Equation (K.24) and the choice of δ_0 in Equation (K.28) that

$$\begin{aligned}
\sup_{f\in\mathcal{F}_{\beta,L}}\mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f) &\gtrsim M^{-\beta}\sup_{\boldsymbol{\omega}\in\Omega_m}\sum_{j=1}^{m}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}},\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left\{A_t \neq \omega_j, \boldsymbol{O}_t \in \widetilde{B}_j\right\} \mid \mathcal{H}_{t-1}\right] \\
&\geq 2^{-m}M^{-\beta}\sum_{\boldsymbol{\omega}\in\Omega_m}\sum_{j=1}^{m}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}},\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left\{A_t \neq \omega_j, \boldsymbol{O}_t \in \widetilde{B}_j\right\} \mid \mathcal{H}_{t-1}\right],
\end{aligned} \tag{K.32}$$

where the last inequality follows from $|\Omega_m| = 2^m$. Let

$$G_j^t := \sum_{\boldsymbol{\omega}_{[-j]}\in\Omega_{m-1}}\sum_{i\in\{\pm 1\}}\mathbb{E}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}_{[-j]}^i},\pi}^{(t-1)}\mathbb{E}\left[\mathbb{1}\left\{A_t \neq i, \boldsymbol{O}_t \in \widetilde{B}_j\right\} \mid \mathcal{H}_{t-1}\right], \tag{K.33}$$

where we group $\boldsymbol{\omega}_{[-j]}^1$ and $\boldsymbol{\omega}_{[-j]}^{-1}$ together in the inner sum. Taking Equation (K.33) into Equation (K.32), we have

$$\sup_{f\in\mathcal{F}_{\beta,L}}\mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f) \gtrsim 2^{-m}M^{-\beta}\sum_{j=1}^{m}\sum_{t=1}^{T}G_j^t. \tag{K.34}$$

We pause to provide some intuition for introducing G_j^t. The idea is that we would like to apply Bretagnolle-Huber inequality as stated in Theorem K.1 to obtain a lower bound of the cumulative regret. To get a tighter lower bound, we would group the most similar pairs of $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega_m$ together to minimize the KL divergence between the two probability measures indexed by $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$.

By Equation (K.25) and the definition of δ_0 in Equation (K.28),

$$\mathbb{E}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^i_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_t\neq i,\boldsymbol{O}_t\in\widetilde{B}_j\right\}\mid\mathcal{H}_{t-1}\right]\asymp\frac{1}{M^{d_{\text{non}}}}\mathbb{E}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^i_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_t\neq i\right\}\mid\mathcal{H}_{t-1},\boldsymbol{O}_t\in\widetilde{B}_j\right]. \tag{K.35}$$

Denote by $\mathbb{P}^{(t-1)}_j$ the conditional probability $\mathbb{P}\left(\cdot\mid\mathcal{H}_{t-1},\boldsymbol{O}_t\in\widetilde{B}_j\right)$. We apply Bretagnolle-Huber inequality as stated in Theorem K.1 and obtain

$$\begin{aligned}
&\sum_{i\in\{\pm1\}}\mathbb{E}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^i_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_t\neq i\right\}\mid\mathcal{H}_{t-1},\boldsymbol{O}_t\in\widetilde{B}_j\right]\\
&\geq\frac{1}{2}\exp\left[-\mathrm{KL}\left(\mathbb{P}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^1_{[-j]}},\pi}\times\mathbb{P}^{(t-1)}_j\|\mathbb{P}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^{-1}_{[-j]}},\pi}\times\mathbb{P}^{(t-1)}_j\right)\right]\\
&=\frac{1}{2}\exp\left[-\mathrm{KL}\left(\mathbb{P}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^1_{[-j]}},\pi}\|\mathbb{P}^{(t-1)}_{\boldsymbol{\theta},f_{\boldsymbol{\omega}^{-1}_{[-j]}},\pi}\right)\right]\\
&\geq\frac{1}{2}\exp\left[-\mathcal{K}^{(\text{non})}_{\boldsymbol{\theta},t}\left(f_{\boldsymbol{\omega}^1_{[-j]}},f_{\boldsymbol{\omega}^{-1}_{[-j]}}\right)\right]
\end{aligned} \tag{K.36}$$

where the last inequality follows from Equations (K.16), (K.1) and (K.2). Taking Equations (K.35) and (K.36) into Equation (K.33) yields that

$$\begin{aligned}
G^t_j&\gtrsim M^{-d_{\text{non}}}\sum_{\boldsymbol{\omega}_{[-j]}\in\Omega_{m-1}}\exp\left[-\mathcal{K}^{(\text{non})}_{\boldsymbol{\theta},t}\left(f_{\boldsymbol{\omega}^1_{[-j]}},f_{\boldsymbol{\omega}^{-1}_{[-j]}}\right)\right]\\
&\overset{(i)}{\geq}\frac{1}{M^{d_{\text{non}}}}\sum_{\boldsymbol{\omega}_{[-j]}\in\Omega_{m-1}}\exp\left(-\frac{C}{M^{2\beta}}\sum_{\tau=1}^{t}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f^1_{\boldsymbol{\omega}_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_\tau=-1,\boldsymbol{O}_\tau\in\widetilde{B}_j\right\}\mid\mathcal{H}_{\tau-1}\right]-\frac{C(t+N)}{M^{2\beta+d_{\text{non}}}}\right)\\
&\geq\frac{1}{M^{d_{\text{non}}}}\sum_{\boldsymbol{\omega}_{[-j]}\in\Omega_{m-1}}\exp\left(-\frac{C}{M^{2\beta}}\sum_{\tau=1}^{T}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f^1_{\boldsymbol{\omega}_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_\tau=-1,\boldsymbol{O}_\tau\in\widetilde{B}_j\right\}\mid\mathcal{H}_{\tau-1}\right]-\frac{C(T+N)}{M^{2\beta+d_{\text{non}}}}\right)\\
&\overset{(ii)}{\geq}\frac{2^{m-1}}{M^{d_{\text{non}}}}\exp\left(-\frac{C}{M^{2\beta}2^{m-1}}\sum_{\boldsymbol{\omega}_{[-j]}\in\Omega_{m-1}}\sum_{\tau=1}^{T}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f^1_{\boldsymbol{\omega}_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_\tau=-1,\boldsymbol{O}_\tau\in\widetilde{B}_j\right\}\mid\mathcal{H}_{\tau-1}\right]-\frac{C(T+N)}{M^{2\beta+d_{\text{non}}}}\right)
\end{aligned} \tag{K.37}$$

where inequality (i) follows from Equation (K.31) and inequality (ii) follows from Jensen's inequality. Let

$$E_{j,\pi}:=\frac{1}{2^{m-1}}\sum_{\boldsymbol{\omega}_{[-j]}\in\Omega_{m-1}}\sum_{\tau=1}^{T}\mathbb{E}^{(\tau-1)}_{\boldsymbol{\theta},f^1_{\boldsymbol{\omega}_{[-j]}},\pi}\mathbb{E}\left[\mathbb{1}\left\{A_\tau=1,\boldsymbol{O}_\tau\in\widetilde{B}_j\right\}\mid\mathcal{H}_{\tau-1}\right]$$

and taking $E_{j,\pi}$ into Equation (K.37), we have

$$G^t_j\gtrsim\frac{2^{m-1}}{M^{d_{\text{non}}}}\exp\left(-CM^{-2\beta}E_{j,\pi}-C(T+N)M^{-2\beta-d_{\text{non}}}\right) \tag{K.38}$$

From the definition of G^t_j in Equation (K.33), we also have

$$\sum_{t=1}^{T}G^t_j\geq2^{m-1}E_{j,\pi}. \tag{K.39}$$

Taking Equations (K.38) and (K.39) into Equation (K.34) yields

$$\begin{aligned}\sup_{f\in\mathcal{F}_{\beta,L}} \mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f) &\gtrsim 2^{-m} M^{-\beta} \sum_{j=1}^{m}\sum_{t=1}^{T} G_j^t \\ &\geq \frac{1}{2} M^{-\beta} \sum_{j=1}^{m} \max\left\{E_{j,\pi}, \frac{1}{M^{d_{\text{non}}}} \exp\left(-CM^{-2\beta}E_{j,\pi} - \frac{C(T+N)}{M^{2\beta+d_{\text{non}}}}\right)\right\} \\ &\geq \frac{1}{4} M^{-\beta} \sum_{j=1}^{m} \left\{E_{j,\pi} + \frac{T}{M^{d_{\text{non}}}} \exp\left(-CM^{-2\beta}E_{j,\pi} - \frac{C(T+N)}{M^{2\beta+d_{\text{non}}}}\right)\right\} \\ &\gtrsim \inf_{z\geq 0} M^{-\beta} \sum_{j=1}^{m} \left\{z + \frac{T}{M^{d_{\text{non}}}} \exp\left[-CM^{-2\beta}z - \frac{C(T+N)}{M^{2\beta+d_{\text{non}}}}\right]\right\}.\end{aligned}$$

The case where $N = \Theta(T)$ can be handled similarly as the analysis below and we omit the details here. We focus on the case where $N \gg T$ and the above display can be simplified into

$$\begin{aligned}\sup_{f\in\mathcal{F}_{\beta,L}} \mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f) &\gtrsim \inf_{z\geq 0} mM^{-\beta} \left\{z + \frac{T}{M^{d_{\text{non}}}} \exp\left[-CM^{-2\beta}z - \frac{CN}{M^{2\beta+d_{\text{non}}}}\right]\right\} \\ &\gtrsim \inf_{z\geq 0} M^{-\beta+d_{\text{non}}} \left\{z + \frac{T\alpha}{M^{d_{\text{non}}}} \exp\left[-CM^{-2\beta}z\right]\right\}\end{aligned} \tag{K.40}$$

where in the last inequality, we use the definition of m as in Equation (K.9) and let

$$\alpha := \exp\left(-\frac{CN}{M^{2\beta+d_{\text{non}}}}\right).$$

The minimizer of the right-hand side of Equation (K.40) over $z \in \mathbb{R}$ is given by

$$z^\star = \frac{M^{2\beta}}{C} \log\left(\frac{CT\alpha}{M^{2\beta+d_{\text{non}}}}\right) = \frac{M^{2\beta}}{C} \log\left(\frac{CT}{M^{2\beta+d_{\text{non}}}}\right) - \frac{N}{M^{d_{\text{non}}}}. \tag{K.41}$$

For $z^\star \geq 0$ to hold, we need

$$M^{2\beta+d_{\text{non}}} \log\left(\frac{CT}{M^{2\beta+d_{\text{non}}}}\right) \geq CN. \tag{K.42}$$

Noting that when $M^{2\beta+d_{\text{non}}} > CT$, the left hand side of the above display is negative. Thus, for Equation (K.42) to hold, we must have $M^{2\beta+d_{\text{non}}} = O(T)$, implying that

$$M^{2\beta+d_{\text{non}}} \log\left(\frac{CT}{M^{2\beta+d_{\text{non}}}}\right) = O(T).$$

The maximizer of the left hand side of the above display is given by

$$M^{2\beta+d_{\text{non}}} = \frac{CT}{e}. \tag{K.43}$$

When $T \geq CN$ for some constant C sufficiently large, $z^\star \geq 0$ holds. Taking $z = z^\star$ in Equation (K.40) yields

$$\sup_{f\in\mathcal{F}_{\beta,L}} \mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f) \gtrsim M^{\beta+d_{\text{non}}} \left[\log\left(\frac{TC}{M^{2\beta+d_{\text{non}}}}\right) + 1\right] - \frac{N}{M^{d_{\text{non}}}} = \Theta\left(T^{\frac{\beta+d_{\text{non}}}{2\beta+d_{\text{non}}}}\right)$$

where the last equality holds from the choice of M in Equation (K.43).

When $T \ll N$, we have $z^\star < 0$. Noting that for any constants $a, b > 0$, the function

$$h(z) = z + a\exp(-bz)$$

attains its minimum at

$$z_0 = \frac{\log(ab)}{b},$$

and is monotonically increasing when $z > z_0$, it follows that the minimizer of $h(z)$ over $z \geq 0$ when $z_0 < 0$ is attained at $z = 0$. Comparing the form of the right-hand side of Equation (K.40) to $h(z)$ defined above yields that the minimizer is attained at $z = 0$ and

$$\sup_f \mathcal{R}_T^{(\text{non})}(\boldsymbol{\theta}, f) \geq \frac{T}{M^\beta} \exp\left(-\frac{CN}{M^{2\beta+d_{\text{non}}}}\right). \tag{K.44}$$

Let

$$g(M) := -\beta \log M + \log T - CNM^{-2\beta-d_{\text{non}}},$$

we have

$$g'(M) = -\frac{\beta}{M} + \frac{C(2\beta+d_{\text{non}})N}{M^{2\beta+d_{\text{non}}+1}}.$$

It attains its maximum at

$$M = \left[\frac{C(2\beta+d_{\text{non}})N}{\beta}\right]^{\frac{1}{2\beta+d_{\text{non}}}} = \Theta\left(N^{\frac{1}{2\beta+d_{\text{non}}}}\right). \tag{K.45}$$

Taking Equation (K.45) into the right-hand side of Equation (K.44) yields the desired result in Equation (K.13). ■

L Derivation of Equivalent Formulations for UCB Exploration in Algorithm 1

This section demonstrates that the Upper Confidence Bound (UCB) exploration strategy used in the LinUCB algorithm can be expressed in two equivalent forms. We first derive the general relationship between the exploration parameter α and the confidence set parameter γ_t in the LinUCB algorithm. We then show how this relationship leads to an adaptive exploration schedule in our specific context.

The key variables are defined as follows:

- α_t: The exploration hyperparameter at timestep t.
- γ_t: A parameter controlling the size of the confidence ellipsoid at timestep t.
- $\boldsymbol{x}_{t,a} \in \mathbb{R}^d$: The context vector for action $a \in \mathcal{A}$ at timestep t.
- $\widehat{\boldsymbol{\theta}}_{t-1} \in \mathbb{R}^d$: The ridge regression estimate of the parameter vector at the end of timestep $t-1$.
- $\boldsymbol{\Sigma}_{t-1} \in \mathbb{R}^{d\times d}$: The design matrix, defined as $\boldsymbol{\Sigma}_{t-1} = \lambda \boldsymbol{I} + \sum_{t=1}^{t-1} \boldsymbol{x}_{t,A_t} \boldsymbol{x}_{t,A_t}^\top$.
- $\texttt{BALL}_{t-1}$: The confidence ellipsoid for the true parameter vector $\boldsymbol{\theta}^*$ at timestep $t-1$. It is defined as:

$$\texttt{BALL}_{t-1} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d \mid (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{t-1})^\top \boldsymbol{\Sigma}_{t-1} (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{t-1}) \leq \gamma_{t-1}\right\}$$

The LinUCB algorithm can be formulated from two equivalent perspectives.

1. The α-based UCB formulation: The action A_t is chosen to maximize an upper confidence bound on the expected reward:

$$A_t = \arg\max_{a\in\mathcal{A}} \left(\widehat{\boldsymbol{\theta}}_{t-1}^\top \boldsymbol{x}_{t,a} + \alpha_{t-1}\sqrt{\boldsymbol{x}_{t,a}^\top \boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{x}_{t,a}}\right) \tag{L.1}$$

2. The confidence set formulation: The action A_t is chosen by finding the most optimistic parameter vector within the confidence set for each action, and then selecting the action with the highest optimistic reward:

$$A_t = \arg\max_{a\in\mathcal{A}} \max_{\boldsymbol{\theta}\in \text{BALL}_{t-1}} \boldsymbol{\theta}^\top \boldsymbol{x}_{t,a} \tag{L.2}$$

Our goal is to show the equivalence of the objective functions in (L.1) and (L.2). We focus on solving the inner maximization problem in (L.2):

$$\max_{\boldsymbol{\theta}} \quad \boldsymbol{\theta}^\top \boldsymbol{x}_{t,a} \quad \text{subject to} \quad \boldsymbol{\theta} \in \text{BALL}_{t-1}$$

Let's introduce a change of variables: $\boldsymbol{z} = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_{t-1}$, which implies $\boldsymbol{\theta} = \boldsymbol{z} + \widehat{\boldsymbol{\theta}}_{t-1}$. The optimization problem becomes:

$$\begin{aligned} \max_{\boldsymbol{z}} \quad & (\boldsymbol{z} + \widehat{\boldsymbol{\theta}}_{t-1})^\top \boldsymbol{x}_{t,a} \\ \text{subject to} \quad & \boldsymbol{z}^\top \boldsymbol{\Sigma}_{t-1} \boldsymbol{z} \leq \gamma_{t-1} \end{aligned}$$

The objective function can be split into two parts: $\boldsymbol{z}^\top \boldsymbol{x}_{t,a} + \widehat{\boldsymbol{\theta}}_{t-1}^\top \boldsymbol{x}_{t,a}$. Since $\widehat{\boldsymbol{\theta}}_{t-1}^\top \boldsymbol{x}_{t,a}$ is constant with respect to $\boldsymbol{z}$, we only need to maximize $\boldsymbol{z}^\top \boldsymbol{x}_{t,a}$.

The problem is now $\max_{\boldsymbol{z}} \boldsymbol{z}^\top \boldsymbol{x}_{t,a}$ subject to $\boldsymbol{z}^\top \boldsymbol{\Sigma}_{t-1} \boldsymbol{z} \leq \gamma_{t-1}$. By the generalized Cauchy-Schwarz inequality, which states $(u^\top v)^2 \leq (u^\top M u)(v^\top M^{-1} v)$ for a positive definite matrix M, we can set $u = \boldsymbol{z}$, $v = \boldsymbol{x}_{t,a}$, and $M = \boldsymbol{\Sigma}_{t-1}$. This gives:

$$(\boldsymbol{z}^\top \boldsymbol{x}_{t,a})^2 \leq (\boldsymbol{z}^\top \boldsymbol{\Sigma}_{t-1} \boldsymbol{z})(\boldsymbol{x}_{t,a}^\top \boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{x}_{t,a})$$

Using our constraint $\boldsymbol{z}^\top \boldsymbol{\Sigma}_{t-1} \boldsymbol{z} \leq \gamma_{t-1}$, we get:

$$(\boldsymbol{z}^\top \boldsymbol{x}_{t,a})^2 \leq \gamma_{t-1} (\boldsymbol{x}_{t,a}^\top \boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{x}_{t,a})$$

Taking the square root, the maximum value for $\boldsymbol{z}^\top \boldsymbol{x}_{t,a}$ is:

$$\max_{\boldsymbol{z}} \boldsymbol{z}^\top \boldsymbol{x}_{t,a} = \sqrt{\gamma_{t-1}} \sqrt{\boldsymbol{x}_{t,a}^\top \boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{x}_{t,a}}$$

Substituting this back into the full objective function, we have:

$$\max_{\boldsymbol{\theta} \in \texttt{BALL}_{t-1}} \boldsymbol{\theta}^\top \boldsymbol{x}_{t,a} = \widehat{\boldsymbol{\theta}}_{t-1}^\top \boldsymbol{x}_{t,a} + \sqrt{\gamma_{t-1}} \sqrt{\boldsymbol{x}_{t,a}^\top \boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{x}_{t,a}}$$

By comparing this result with the objective function in (L.1), we can directly establish the relationship:

$$\alpha_{t-1} = \sqrt{\gamma_{t-1}}$$

L.1 Implication for Adaptive Exploration

This equivalence enables us to understand how the adaptive nature of the confidence set, defined by γ_t, is directly translated into the exploration parameter α_t.

Given the definition of γ_t from Theorem 4.1:

$$\gamma_t := \gamma_t^{(0)} + 3d^2 \sum_{t=1}^{t} D_t \tag{L.3}$$

where $\gamma_t^{(0)}$ captures the baseline uncertainty from stochastic noise, the relationship is:

$$\alpha_t = \sqrt{\gamma_t^{(0)} + 3d^2 \sum_{t=1}^{t} D_t} \tag{L.4}$$

Expanding the $\gamma_t^{(0)}$ term, we get the complete expression:

$$\alpha_t = \sqrt{\left(3\lambda + 6(\sigma_\eta + \sigma_\varepsilon)^2 \log\left[\frac{4t^2}{\delta}\left(1 + \frac{tB^2}{d\lambda}\right)^d\right]\right) + 3d^2 \sum_{t=1}^{t} D_t} \tag{L.5}$$

This equation shows that the exploration parameter α_t is adaptive. It increases not only due to inherent stochasticity (the $\gamma_t^{(0)}$ term) but also in response to the accumulated uncertainty in context estimation over all past timesteps (the $\sum D_t$ term).

M Synthetic Data Experiments

M.1 Extra Baseline Comparisons

In this section, we provide extended empirical results to complement the main experiments. Specifically, we systematically evaluate the dependency of PULSE-UCB on the pre-training data scales and present an expanded benchmark against CLBBF in both linear and nonlinear settings.

M.1.1 Sensitivity to Pre-training Data Scale N and T_0

To evaluate the dependency on pre-training data, we adopted a high-dimensional setting where the unobserved context $W_t = \boldsymbol{\beta}_*^\top \boldsymbol{x}_t + \xi_t$ depends on history $\boldsymbol{x}_t = (1, \boldsymbol{S}_{t-1}, \ldots, \boldsymbol{S}_{t-19})^\top \in \mathbb{R}^{20}$, while $\{\boldsymbol{S}_t\}$ remains an ARMA$(2,2)$ process. The reward is updated to $\boldsymbol{\Phi}(\boldsymbol{Y}_t, a_t) = (1, \boldsymbol{S}_t, W_t, W_t \cdot a_t)^\top$.

Tables 1 and 2 summarize the results under varying sample sizes N and trajectory lengths T_0.

Table 1: Final cumulative regret vs. Pre-training Size N in the linear setting.

N	PULSE-UCB	OFUL-Full	OFUL	CLBBF
1	252.1 ± 87.5	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
2	150.3 ± 77.0	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
3	61.3 ± 91.9	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
5	11.7 ± 9.6	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
10	4.6 ± 1.4	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
50	4.2 ± 1.6	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
500	4.3 ± 1.6	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
1000	4.0 ± 1.4	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5

Table 2: Final cumulative regret vs. Pre-training Trajectory Length T_0 in the linear setting.

T_0	PULSE-UCB	OFUL-Full	OFUL	CLBBF
21	248.7 ± 163.8	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
25	46.9 ± 55.7	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
30	8.6 ± 2.7	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
40	5.1 ± 2.2	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
50	4.3 ± 1.4	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
100	4.4 ± 2.2	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5
200	4.6 ± 1.5	2.7 ± 2.0	333.7 ± 30.2	145.3 ± 9.5

Discussion on Data Scaling: As clearly demonstrated in Tables 1 and 2, PULSE-UCB exhibits a sharp phase transition in both cases. When pre-training data is scarce (e.g., $N = 1$ or $T_0 = 21$), the performance is limited and behaves similarly to the naive OFUL baseline. However, increasing either N (with a fixed $T_0 = 25$) or T_0 (with a fixed $N = 3$) beyond a modest threshold leads to a rapid drop in cumulative regret. In this stable and data-sufficient regime, PULSE-UCB demonstrates exceptionally high sample efficiency, achieving near-oracle performance that matches OFUL-Full (which operates with fully observed contexts) and significantly outperforming CLBBF.

M.1.2 Expanded Comparison with CLBBF

In the main text, CLBBF was primarily evaluated on the real-world dataset as that specific dataset was a suitable testbed for CLBBF's underlying assumptions. We clarify that CLBBF was originally omitted from the synthetic benchmarks because it assumes a *Missing-At-Random* (MAR) structure, whereas our synthetic setting involves structurally missing (MNAR) covariates. Following the reviewer's suggestion, we have now added CLBBF to the synthetic analysis to explicitly illustrate the limitations of the MAR assumption in our environment.

Tables 3 and 4 present the time-evolution of cumulative regret up to $T = 1000$ in both linear and nonlinear reward environments.

Table 3: Cumulative Regret Comparison of algorithms in the linear setting.

Time step	PULSE-UCB	OFUL	CLBBF	OFUL-Full
1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
10	0.1 ± 0.2	0.2 ± 0.2	0.1 ± 0.2	0.2 ± 0.2
100	2.8 ± 0.6	3.6 ± 0.8	1.2 ± 0.6	3.0 ± 0.6
300	6.1 ± 0.8	11.6 ± 1.2	4.7 ± 1.0	6.0 ± 0.9
800	10.5 ± 1.3	31.6 ± 1.7	16.0 ± 1.5	9.8 ± 1.4
1000	11.9 ± 1.4	39.6 ± 2.1	20.6 ± 1.4	11.1 ± 1.7

Table 4: Cumulative Regret Comparison of algorithms in the nonlinear setting.

Time step	PULSE-UCB	OFUL	CLBBF	OFUL-Full
1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
10	0.1 ± 0.2	0.2 ± 0.2	0.2 ± 0.2	0.2 ± 0.2
100	2.3 ± 0.6	2.8 ± 0.7	2.2 ± 0.5	2.9 ± 0.6
300	5.4 ± 0.8	10.6 ± 1.3	5.8 ± 0.8	6.0 ± 0.7
800	11.6 ± 1.6	30.6 ± 1.8	15.3 ± 1.4	9.7 ± 1.4
1000	13.7 ± 1.7	38.5 ± 2.1	19.2 ± 1.5	11.0 ± 1.8

Discussion on Baselines and MAR/MNAR Structures: Consistent with the real-world results, CLBBF outperforms the naive OFUL algorithm but consistently lags behind PULSE-UCB. The results confirm our hypothesis: online feature estimation methods like CLBBF struggle in this MNAR regime where covariates are structurally missing. Conversely, the time-evolution analysis demonstrates that by $T = 1000$, PULSE-UCB rapidly converges to the oracle (OFUL-Full) performance in both linear and nonlinear settings. This highlights that our pre-training approach successfully bridges the information gap caused by MNAR structures, maintaining near-oracle performance where traditional MAR-based methods falter.

M.2 Impact of Smoothness

To test robustness, we evaluated `PULSE-UCB` in a linear environment and three nonlinear variants controlled by a parameter ρ. The results in Figure 3 show a clear correlation between performance and the degree of nonlinearity. The agent performs well in the linear and low-nonlinearity ($\rho = 0.1$) cases, with final regrets around 9.3. As the model misspecification becomes more pronounced, the final regret increases to 9.7 for $\rho = 1.0$ and further to 14.6 for the highly nonlinear case of $\rho = 10.0$. The smoothed instant regret plot (Right) confirms this trend, showing larger and more volatile regret for higher ρ. This experiment demonstrates that while `PULSE-UCB` is robust to smooth deviations from linearity, its performance gracefully deteriorates as the environment becomes more complex.

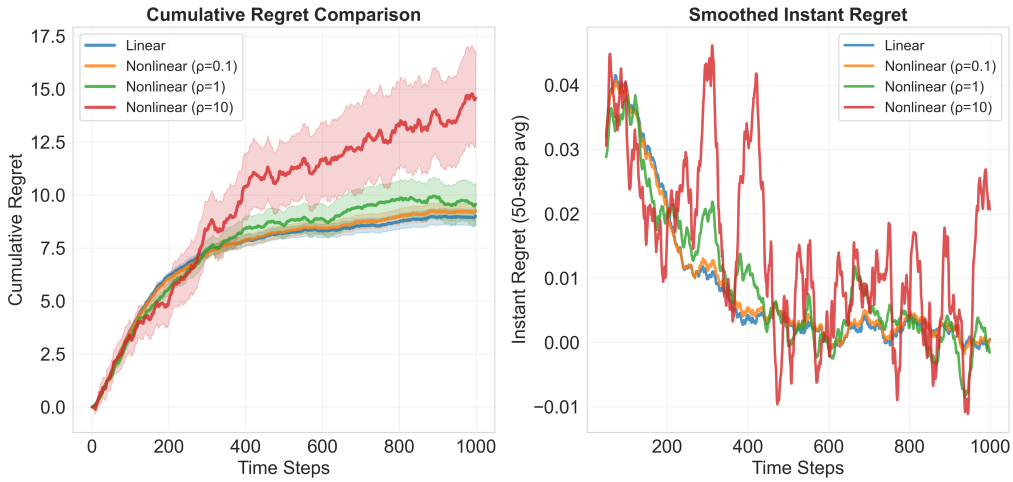

Figure 3: Comparison of PULSE-UCB agent learning results under different linearity settings

M.3 Robustness under Time-Varying Observation Patterns

To empirically substantiate our theoretical claim that PULSE-UCB handles general, time-varying observation patterns, we conducted a Random Feature Masking experiment. In this setup, we adopt the high-dimensional setting described in Appendix M.1.1, where the dimension of the underlying feature vector is fixed at $d = 20$. However, at each time step t, the agent only observes a random subset of these features. Specifically, each feature coordinate is independently masked (set to zero) with a probability $p \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. This simulates a challenging regime where the available information changes dynamically across agents or time. We trained a robust version of the β-estimator using data augmentation (dropout) and compared PULSE-UCB against all baselines.

Table 5 summarizes the final cumulative regret under varying masking probabilities.

Table 5: Robustness under Random Feature Masking with varying mask probabilities p.

Mask Prob (p)	PULSE-UCB	OFUL-Full	OFUL	CLBBF
0.0	46.52 ± 5.20	2.73 ± 1.77	346.18 ± 28.19	144.95 ± 7.19
0.2	73.18 ± 5.90	2.64 ± 1.76	323.65 ± 21.64	138.15 ± 10.06
0.4	111.29 ± 7.61	2.04 ± 2.51	332.07 ± 21.84	141.86 ± 12.60
0.6	169.47 ± 12.70	3.12 ± 1.27	335.68 ± 23.22	139.63 ± 12.02
0.8	267.07 ± 23.99	2.66 ± 1.39	322.72 ± 20.34	146.55 ± 10.72

Discussion on Algorithm Behavior and Robustness: The results summarized in Table 5 illustrate distinct behavioral patterns across the algorithms. As expected, OFUL-Full and the naive OFUL maintain their respective near-zero and high-regret baselines, as neither is influenced by history masking. Similarly, CLBBF exhibits stable but mediocre performance across all masking rates, indicating that it fails to leverage partial history to recover the latent context W_t.

In contrast, PULSE-UCB demonstrates remarkable robustness. With full observations, it outperforms CLBBF by a substantial margin. Crucially, this advantage persists even when a significant portion of features is missing ($p = 0.4$); it is only under conditions of extreme sparsity (e.g., $p \geq 0.6$) that the information loss causes PULSE-UCB to fall behind the conservative CLBBF baseline.

These results conclusively demonstrate that PULSE-UCB does not rely on a fixed, complete feature structure. As long as the observed subset of features retains predictive power regarding the latent context, our approach effectively bridges the information gap, validating its practical applicability in general regimes with time-varying observations.

M.4 Robustness to Distributional and Covariate Shifts

Identifying the boundaries of our method under distribution mismatches is essential for understanding its practical applicability. In this section, we evaluate the robustness of PULSE-UCB empirically by introducing shifts in the marginal distribution of the observed covariates between the pre-training and online phases, while keeping the latent dynamics and reward mechanism fixed. To thoroughly assess this, we first examine a basic variance shift and then extend our evaluation to more complex, multi-dimensional geometric shifts in the feature space: Rotation, Translation, and Deformation.

M.4.1 Impact of Covariate Variance Shift

We first conducted a sensitivity analysis by introducing a *variance shift*. The pre-training data are generated with a standard noise scale $\sigma = 0.1$, while the online environments use increasingly large noise levels $\sigma \in \{0.1, 0.5, 1.0, 1.5, 2.0\}$ (up to a $20\times$ increase in standard deviation), where σ denotes the standard deviation of the white noise process ε_t driving the ARMA generation of the observed context S_t. The base high-dimensional setting follows the description in Appendix M.1.1.

Table 6 summarizes the final cumulative regrets under varying noise scales.

Table 6: Impact of Covariate Variance Shift (Training $\sigma = 0.1$).

Noise Scale (σ)	PULSE-UCB	OFUL	CLBBF	OFUL-Full
0.1	5.1 ± 2.3	504.9 ± 48.0	115.3 ± 11.0	1.8 ± 1.7
0.5	4.5 ± 1.0	510.5 ± 84.5	107.2 ± 15.1	2.9 ± 2.5
1.0	5.2 ± 2.4	519.9 ± 50.1	107.4 ± 12.2	2.6 ± 1.3
1.5	3.9 ± 2.3	498.0 ± 44.2	113.9 ± 6.5	2.7 ± 1.9
2.0	6.0 ± 1.4	513.6 ± 29.8	111.7 ± 14.1	4.0 ± 1.5

Discussion on Variance Shift: Table 6 shows that PULSE-UCB remains remarkably stable under covariate variance shifts. Across all noise levels, its final cumulative regret stays in a narrow low-regret range and is consistently close to the oracle OFUL-Full agent. Importantly, despite pre-training only at the low-noise setting ($\sigma = 0.1$), the learned latent structure generalizes to much noisier online environments without noticeable degradation. This indicates that the transition model on the latent state space, rather than the raw covariates, is driving the performance of PULSE-UCB, ensuring robustness to substantial fluctuations in the scale of observed features.

In contrast, the online-only baselines struggle throughout the entire range. OFUL incurs very large final regrets due to the severe information loss caused by the unobserved context. CLBBF performs better than OFUL but plateaus at a sub-optimal level, showing no capability to adapt to the latent structure. Across all noise scales, PULSE-UCB maintains a massive performance gap—roughly two orders of magnitude relative to OFUL and more than a factor of 20 relative to CLBBF.

M.4.2 Rotation: Structural Shift via AR Perturbation

Varying the scalar noise variance σ^2 induces a one-dimensional scaling shared across all coordinates of the feature vector $\mathbf{x}_t$. To more thoroughly assess robustness to higher-dimensional covariate shifts, we introduce structural perturbations to the autoregressive (AR) coefficients that generate the scalar state process S_t. Since the feature vector $\mathbf{x}_t = (1, S_{t-1}, \ldots, S_{t-19})^\top$ collects 20 lags of this process, the AR parameters fully determine its covariance and correlation structure. Varying these parameters changes the principal axes (eigenvectors) of the 20-dimensional feature distribution, effectively rotating the feature manifold in the context space.

Setup: Historical data are generated under AR coefficients $\phi_{\text{train}} = [0.75, -0.25]$. Online data are generated with coefficients that interpolate between ϕ_{train} and a more oscillatory regime $\phi_{\text{target}} = [-0.80, -0.20]$, controlled by a shift intensity α.

Geometric Change: We quantify the structural shift using *PC Subspace Similarity* (the overlap between the top-5 principal components of the historical and online feature distributions). As shown in Table 7, this

similarity drops from 1.000 to 0.073, indicating that the feature manifold has rotated to be nearly orthogonal to the training distribution.

Table 7: Regret under Structural Shift (Rotation).

α	AR Params	PC Subspace Sim.	PULSE-UCB	OFUL	CLBBF	OFUL-Full
0.0	$[0.75, -0.25]$	1.000	5.24 ± 0.43	593.0 ± 15.8	110.5 ± 2.5	2.68 ± 0.63
0.3	$[0.28, -0.23]$	0.878	5.53 ± 0.60	418.8 ± 8.97	118.5 ± 3.2	1.90 ± 0.60
0.7	$[-0.33, -0.21]$	0.858	6.25 ± 0.63	367.3 ± 8.71	103.3 ± 1.8	1.77 ± 0.55
0.9	$[-0.65, -0.21]$	0.452	6.17 ± 0.43	249.2 ± 33.7	104.7 ± 2.0	1.81 ± 0.57
1.0	$[-0.80, -0.20]$	**0.073**	$\mathbf{5.94 \pm 0.47}$	222.0 ± 35.9	107.0 ± 2.0	1.53 ± 0.61

Notably, PULSE-UCB demonstrates superior robustness, maintaining low regret comparable to the Oracle even under severe structural shifts where standard baselines fail.

M.4.3 Translation: Sensitivity to Location Shift

To assess robustness to shifts in the domain's operating point, we introduce a global translation to the feature coordinates.

Setup: We shift the centroid of the state distribution by adding a constant $\mu \in [0, 5.0]$ to every generated state value S_t.

Geometric Change: Since the feature vector $\mathbf{x}_t$ is constructed from lagged states, adding μ to the scalar process effectively adds μ to every coordinate of the feature vector (except the fixed intercept). Geometrically, this translates the entire feature cloud along the diagonal direction $\mathbf{v} = (0, 1, \ldots, 1)^\top$ in the multi-dimensional feature space.

Table 8: Regret under Mean Shift (Translation).

α	Shift Detail	PULSE-UCB	OFUL	CLBBF	OFUL-Full
0.0	Mean +0.0	4.22 ± 0.67	598.4 ± 16.9	106.1 ± 3.3	2.83 ± 0.52
0.4	Mean +2.0	10.83 ± 1.20	10.48 ± 1.00	154.0 ± 3.3	10.86 ± 0.91
0.8	Mean +4.0	23.35 ± 0.88	22.79 ± 0.66	92.56 ± 1.3	23.19 ± 0.65
1.0	Mean +5.0	$\mathbf{28.92 \pm 0.74}$	28.54 ± 0.66	86.34 ± 0.8	28.55 ± 0.62

As shown in Table 8, PULSE-UCB tightly tracks the Oracle (OFUL-Full) performance throughout the entire range. This confirms that our method adapts effectively when the test domain is centered in a different region of the feature space compared to the training domain.

M.4.4 Deformation: Sensitivity to Geometric Shape

Finally, we evaluate robustness to changes in the shape of the distribution density, specifically moving from well-behaved Gaussian noise to heavy-tailed distributions.

Setup: We alter the noise process ε_t. The historical data retains standard Gaussian noise, while the online data switches to a heavy-tailed Student-t distribution (df = 3). Simultaneously, we introduce a scaling factor that linearly increases the dispersion of this noise up to $3\times$ the original scale.

Geometric Change: This modification fundamentally changes the density geometry of the feature manifold. The transition to heavy tails introduces frequent outliers and extreme values in the multi-dimensional space, creating a spiky distribution shape that violates the sub-Gaussian assumptions typically required by linear bandits.

This setting poses a severe challenge. As shown in Table 9, the standard OFUL algorithm collapses under this geometric deformation, with regret exploding significantly. In stark contrast, PULSE-UCB remains remarkably stable (regret ≈ 10), demonstrating superior resilience to non-Gaussian geometric distortions.

Table 9: Regret under Geometric Shift (Deformation).

α	Shift Detail	PULSE-UCB	OFUL	CLBBF	OFUL-Full
0.0	Gaussian (0.1)	5.98 ± 0.30	591.9 ± 15.2	107.7 ± 3.6	2.55 ± 0.61
0.4	T-dist $\times 1.8$	6.17 ± 0.92	1470.3 ± 48.7	69.54 ± 8.8	5.40 ± 0.58
0.8	T-dist $\times 2.6$	4.57 ± 0.97	2117.7 ± 104.0	55.16 ± 4.9	3.60 ± 1.03
1.0	T-dist $\times 3.0$	$\mathbf{10.31 \pm 1.91}$	2930.3 ± 169.7	51.44 ± 5.2	11.11 ± 1.89

Conclusion: Collectively, these experiments confirm that PULSE-UCB is highly robust across three critical dimensions of distributional shift: Rotation (Structure), Translation (Location), and Deformation (Geometry), significantly outperforming baselines in complex, multi-dimensional shift scenarios.

N Related Details about Real Dataset Experiments

This experiment evaluates the performance of the proposed PULSE-UCB algorithm against several baselines in a realistic setting using the public Taobao User Behavior dataset Alibaba (2018).

N.1 Dataset and Preprocessing

We use the Taobao dataset, which contains user interaction data from Taobao's recommender system. The raw data consists of user profiles (`user_profile.csv`), ad features (`ad_feature.csv`), and user-ad interaction logs (`raw_sample.csv`). Our preprocessing pipeline involves the following steps:

1. **Filtering**: To manage the scale and focus on active user segments and ad categories, we filter the data. We retain only the interactions from users belonging to the top 10 most frequent user segments (`cms_segid`) and ads belonging to the top 25 most frequent categories (`cate_id`) and brands (`brand`).
2. **Feature Encoding**: Categorical features for both users (e.g., age range, gender) and ads (e.g., category, brand) are converted into high-dimensional, sparse binary vectors using one-hot encoding. The numerical `price` feature for ads is logarithmically scaled and discretized.
3. **Feature Combination**: For each user-ad interaction, the corresponding user feature vector and ad feature vector are concatenated to form a single high-dimensional feature vector.
4. **Label Creation**: The `clk` column in the interaction log (1 for click, 0 for no-click) serves as the ground-truth reward signal for our online bandit simulation. The data is partitioned into two sets based on this label: X_0 for non-click events and X_1 for click events.

N.2 Dimensionality Reduction via Autoencoder

The initial one-hot encoded feature vectors are extremely high-dimensional and sparse. To create a more manageable and dense feature representation, we train an `Autoencoder` with Batch Normalization.

- **Architecture**: The model consists of an encoder that maps the raw feature dimension $d = 83$ to a dense embedding of size $d = 32$, and a decoder that reconstructs the original vector from this embedding.
- **Training**: The autoencoder is trained on the shuffled combination of all available feature vectors (X_0 and X_1) for 500 epochs with an MSE loss function, a batch size of 10,000, and an Adam optimizer.
- **Output**: After training, we use the encoder to transform all high-dimensional feature vectors into dense 32-dimensional embeddings, which are used in all subsequent steps.

N.3 Partially Observed Setting and Inference Model

To simulate a realistic scenario where only a subset of features is immediately available, we define a partially observed setting.

- **Feature Split**: Each 32-dimensional feature vector Y_t is split into two halves. The first 16 dimensions, denoted as S_t, are considered as the observed features, while the remaining 16 dimensions, S'_t, are the unobserved features.
- **Inference Model**: For PULSE-UCB, we pre-train an inference model to predict S'_t from S_t. This model is a Multi-Layer Perceptron (MLP) with two hidden layers of 128 neurons each, using ReLU activation functions.
- **Pre-training**: The MLP is trained on a dedicated pre-training set, which constitutes 20% of the total shuffled data. The model is trained for 100 epochs using an MSE loss function and an Adam optimizer to minimize the reconstruction error of S'_t. The remaining 80% of the data is reserved for the online evaluation phase.

N.4 Online Evaluation Protocol

The online simulation is performed on the held-out 80% of the dataset.

1. The simulation runs for T time steps, where T is the size of the online dataset minus K.
2. At each time step t, a set of $K = 20$ candidate arms (ads) is randomly sampled without replacement from the online dataset.
3. Each bandit agent selects one arm from the K candidates based on its internal policy.
4. The agent observes the reward (click or no-click) associated with the chosen arm.
5. The agent updates its internal parameters using the feature vector of the chosen arm and the observed reward.
6. This process is repeated over independent runs with different random seeds to ensure robust results, and the average cumulative click-through rate (CTR) is reported.