Probabilistic Guarantees for Adversarial Linear Contextual Bandits

Guyon Phoelich
Delft University of Technology
g.p.phoelich@tudelft.nl

Julia Olkhovskaya
Delft University of Technology
julia.olkhovskaya@gmail.com

Tim van Erven
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
tim@timvanerven.nl

Abstract

We study the adversarial linear contextual bandit problem, in which an oblivious adversary fixes a sequence of loss vectors while the per-round context (action set) is generated stochastically. In the standard setting where contexts are drawn i.i.d. from a fixed distribution, near-optimal regret bounds are known, but only in expectation. We prove the first such bound that holds with high probability, resolving an open problem stated by Neu and Olkhovskaya [2020], Olkhovskaya et al. [2023] and Dai et al. [2024]. Specifically, the regret is at most $O\big(d^3\sqrt{\log(K)T}\operatorname{polylog}(T/\delta)\big)$ with probability at least $1-\delta$ over T rounds, where K is the number of actions and d is the dimension of the actions. We then ask whether high-probability guarantees can be extended beyond i.i.d. contexts. Although fully adversarial contexts make sublinear regret impossible in this setting, we show that two natural relaxations remain tractable: *(i)* non-stationary but independent contexts, whose per-round distributions do not deviate too much from a fixed reference distribution, for which we obtain a high-probability regret bound of $O\big(d^3\sqrt{\log(K)T}\operatorname{polylog}(T/\delta) + d^{5/2}V\operatorname{polylog}(T/\delta)\big)$, where V is the cumulative total variation distance of the per-round distributions from the reference; and *(ii)* stationary but dependent exponentially mixing contexts, like, for instance, uniformly ergodic Markov chains started from their stationary distribution, for which we obtain a high-probability bound of $O\big(C_{\text{mix}}^2 d^3\sqrt{\log(K)T}\operatorname{polylog}(T/\delta)\big)$, where C_{mix} is a constant that depends on the mixing process. Technically, our results build on the recent reduction from adversarial linear contextual bandits to misspecified linear bandits by Van Erven et al. [2025]. All three results are obtained from a single base algorithm under different hyperparameter choices, some of which are functions of data-dependent quantities that are unknown in advance. We resolve this by automatically tuning the hyperparameters using the CORRAL algorithm of Agarwal et al. [2017]. For this purpose, we use a high-probability analysis of CORRAL, since the original version only has guarantees in expectation.

1 Introduction

The contextual bandit problem is a sequential decision-making framework in which a learner repeatedly observes a *context* summarizing the current decision environment, selects an action, and incurs a loss that depends on both the chosen action and the observed context. By conditioning its decisions on the context, a contextual bandit algorithm can substantially outperform any context-agnostic strategy, which makes the framework a natural fit for a wide range of real-world applications,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

including adaptive treatment allocation in clinical trials [Tewari and Murphy, 2017], personalized recommender systems [Li et al., 2010, Beygelzimer et al., 2011], and online advertising [Abe et al., 2003]. Compared to full reinforcement learning, contextual bandits are considerably more tractable, which has driven an active line of work establishing strong theoretical guarantees under a variety of assumptions on how the contexts and losses are generated.

The most extensively studied variant is the *stochastic* linear contextual bandit, where the loss of each action is a *fixed* but unknown linear function of the context, observed up to zero-mean noise, while the contexts themselves may follow an arbitrary, possibly adversarial, sequence [Abe and Long, 1999, Auer, 2002]. By now this setting is well-understood in many respects: computationally efficient algorithms are available that are guaranteed to achieve the minimax regret up to logarithmic factors with high probability, both via the optimism-in-the-face-of-uncertainty principle [Chu et al., 2011, Abbasi-Yadkori et al., 2011] and via Thompson sampling [Agrawal and Goyal, 2013].

In this paper we focus on the alternative *adversarial* linear contextual bandit variant, first introduced by Neu and Olkhovskaya [2020]. In this setting the contexts are stochastic (drawn independently from a common distribution) but nothing is assumed about the loss vectors $\theta_1, \dots, \theta_T$, which may be chosen adversarially. We restrict attention to the oblivious adversary model, in which the losses are fixed before play begins. This adversarial setting is attractive whenever the loss-generating mechanism is non-stationary, strategic, or too complex to fit a tractable stochastic model, since adversarial guarantees do not rely on the correctness of any assumed loss model. We adopt the general formulation in which the learner is provided with a stochastically generated action set $\mathcal{A}_t \subset \mathbb{R}^d$ at the start of each round, which represents the context. They then choose $a_t \in \mathcal{A}_t$, and observe the corresponding loss $\langle a_t, \theta_t \rangle$. This formulation subsumes both the simpler adversarial K-armed bandit setting without contexts, for which $\mathcal{A}_t = \{e_1, \dots, e_K\}$ is fixed to be the standard basis, and the setting originally considered by Neu and Olkhovskaya [2020], in which the actions in each round are obtained from a random context vector via fixed feature maps. When the learner has access to a *simulator* that returns free draws from the context distribution, the near-optimal rate $\tilde{O}(\min\{d\sqrt{T}, \sqrt{dT \log K}\})$ was shown by Dai et al. [2023] (when converting their approach to the contextual linear bandit setting), where $K \geq \max_t |\mathcal{A}_t|$ upper-bounds the number of actions per round. Without simulator access, the best known results are by Liu et al. [2023]: they provide an algorithm with near-optimal regret $\tilde{O}\big(d\sqrt{T}\big)$, but run-time $T^{\Omega(d)}$, and another algorithm with worse regret $\tilde{O}\big(d^2\sqrt{T}\big)$ but polynomial run-time in d, T and K. Using a general reduction from adversarial contextual linear bandits to misspecified adversarial linear bandits without contexts, Van Erven et al. [2025] obtain $\tilde{O}\big(\min\{d^2\sqrt{T}, \sqrt{d^3 T \log K}\}\big)$ regret with an algorithm that runs in polynomial time even for many combinatorial problems. Neu et al. [2024] generalize the original setting of Neu and Olkhovskaya [2020] to loss functions that belong to a reproducing kernel Hilbert space, obtaining rates that depend on the eigenvalue decay of the corresponding kernel. All of these guarantees, however, hold only *in expectation*, and how to extend any of them to a bound that holds with high probability over the randomness of the algorithm and the contexts has remained open, as pointed out by Neu and Olkhovskaya [2020], Olkhovskaya et al. [2023] and Dai et al. [2024]. Closing this gap is of practical importance: in-expectation bounds can mask large fluctuations on individual runs, so that a learner with small average regret may nonetheless suffer large regret on a non-negligible fraction of realizations. This phenomenon is already well known in the simpler multi-armed bandit setting, where the vanilla EXP3 algorithm without explicit exploration is known to frequently incur much larger regret than its optimal in-expectation guarantee suggests [Auer et al., 2002, Stoltz, 2005].

Main Contributions In this paper we resolve this open problem. Our main contribution is a high-probability regret bound of order $O\big(d^3\sqrt{\log(K)T}\mathrm{polylog}(T/\delta)\big)$ that holds with probability at least $1-\delta$, for adversarial linear contextual bandits with independent, identically distributed (i.i.d.) contexts, without requiring a simulator. Compared to the known rates in expectation, this matches the optimal dependence on T, but suffers a worse dependence on the dimension d. We further ask whether the i.i.d. assumption itself can be weakened. Some restriction on the contexts is then unavoidable: Sublinear regret is impossible when both the contexts and the losses are fully adversarial, as is shown for adversarial online classification by Syrgkanis et al. [2016] (Theorem 16); in Appendix I we adapt this result to the contextual bandit setting with oblivious adversary for completeness. We therefore consider two natural relaxations of the i.i.d. assumption, each weakening one of its two components. In the first, the contexts are still independent across rounds but no longer identically distributed: the per-round distributions are only required to have cumulative total variation distance bounded

by some quantity V from some (unknown) fixed distribution. For this setting we establish a high-probability regret bound of order $O\big(d^3\sqrt{\log(K)T}\mathrm{polylog}(T/\delta) + d^{5/2}V\mathrm{polylog}(T/\delta)\big)$, which recovers our i.i.d. guarantee when $V = 0$. In our second relaxation, the contexts have a fixed marginal distribution but are no longer independent across rounds: they are drawn from a stationary process that is exponentially mixing, a class that includes, for instance, any uniformly ergodic Markov chain [Meyn and Tweedie, 1993, Section 16.2] started from its stationary distribution. Here we establish a high-probability regret bound of order $O\big(C_{\text{mix}}^2 \ln^{2/r}\left(12T^3/\delta\right) d^3\mathrm{polylog}(T/\delta)\sqrt{T\log(K)}\big)$, where C_{mix} and r are constants that depend on the mixing process.

Approach Technically, all three results follow from a single base algorithm. The starting point is the reduction from adversarial linear contextual bandits to adversarial linear bandits with misspecified stochastic feedback introduced by Van Erven et al. [2025]. We instantiate this reduction with a different linear bandit subroutine: the corruption-robust algorithm of Liu et al. [2024], which requires an upper bound on the *total* misspecification across rounds rather than a per-round misspecification bound. This is essential for the non-stationary independent setting, where misspecification may concentrate on a few rounds rather than spreading evenly across them. This does come at the cost of worse d dependence since we only consider their computationally efficient algorithm, leading to a factor d^3 rather than $d^{3/2}$. To handle the stationary exponentially mixing setting, we further dilute the dependence between contexts by interleaving κ independent copies of the linear bandit subroutine, so that each copy sees contexts that are κ time-steps apart and therefore approximately independent. The optimal choice of κ and another hyperparameter depends on properties of the context-generating process that the learner does not know in advance, so no single set of hyperparameters can be optimal across all three settings simultaneously. We resolve this by wrapping the base algorithm in the CORRAL meta-algorithm of Agarwal et al. [2017], which adapts to the best hyperparameter from a finite grid. Unlike the original analysis of CORRAL, which only provides in-expectation guarantees, in Appendix D we adapt the analysis of Lee et al. [2020] to show our variant has high-probability guarantees, which may be of independent interest. We further show that, although our base algorithm does not satisfy the assumption originally required by Agarwal et al. [2017], it does satisfy a weaker variant that is still sufficient for the analysis to go through.

Related Work A first high-probability regret guarantee for our problem can be obtained by applying EXP4.P [Beygelzimer et al., 2011] to a finite expert class that discretizes the parameter space of linear policies and is guaranteed to contain a near-optimal policy. However, the cardinality of this expert class is exponential in d.

The reduction to non-contextual linear bandits for the adversarial setting by Van Erven et al. [2025] is conceptually closely related to an earlier similar reduction for the stochastic setting by Hanna et al. [2023]. Using this reduction, Buyukkalayci et al. [2026] have very recently obtained high-probability guarantees for context distributions that follow a Markov chain in the stochastic setting. When the chain is started from its stationary distribution, these are an important special case of our exponentially mixing setting, and indeed it seems likely that their approach could be extended to handle the general exponentially mixing case as well. Although weakening distributional assumptions for the contexts is perhaps less urgent in the stochastic setting, because Chu et al. [2011], Abbasi-Yadkori et al. [2011], Agrawal and Goyal [2013] already allow arbitrary adversarial contexts, Buyukkalayci et al.'s approach via the reduction to non-contextual linear bandits contains strong parallels with ours. They further make the interesting observation that, for the stochastic setting, any linear bandit that can handle feedback delayed by κ time steps can be used in their analysis. Our approach of interleaving κ copies of a regular linear bandit algorithm may be viewed as one well-known way of handling such delayed feedback [Weinberger and Ordentlich, 2002], which does not require developing new algorithms.

The idea of relaxing the i.i.d. assumption for contexts has a natural extension to reinforcement learning, where Jin et al. [2023] obtain regret bounds depending on the variation of the transition probabilities from a fixed reference distribution in adversarial tabular Markov decision processes (MDPs). The relation is that contexts can be mapped to states in an MDP, but since Jin et al. [2023] assume that the number of possible states is finite, whereas we allow the contexts to be continuous, their results do not carry over.

2 Main Results

We first formalize the adversarial linear contextual bandit setting and then present our three main high-probability regret guarantees: for the standard i.i.d. regime, for non-stationary but independent contexts, and for stationary, exponentially mixing contexts. Proofs for the impossibility of sublinear regret in the fully adversarial setting are provided in Appendix I for both nonlinear (Theorem 20) and linear (Theorem 21) comparator policies.

2.1 General Setting

An oblivious adversary fixes a sequence of loss vectors $\theta_1, \dots, \theta_T \in \mathbb{R}^d$ s.t. $\|\theta_t\|_2 \le \sqrt{d}$ without loss of generality (see Lemma 8 of Wei et al. [2021]) before play starts. In each round $t = 1, \dots, T$, nature draws a context $\mathcal{A}_t \subset \mathbb{R}^d$, which is a finite set of at most $|\mathcal{A}_t| \le K$ action vectors of dimension d. The learner observes $\mathcal{A}_t$, then selects a (randomized) action $a_t \in \mathcal{A}_t$ s.t. $\|a_t\|_2 \le 1$ and observes the linear loss $\ell_t(a_t) = \langle a_t, \theta_t \rangle \in [0, 1]$. A *policy* π is any function that maps contexts $\mathcal{A}$ to actions $\pi(\mathcal{A}) \in \mathrm{conv}(\mathcal{A})$, where we interpret elements of the convex hull $\mathrm{conv}(\mathcal{A})$ of $\mathcal{A}$ as randomized actions from $\mathcal{A}$. We further let $\mathbb{E}_t$ denote expectation conditional on the history of the rounds up to and including $t-1$. The regret of the learner against a comparator policy π is

$$\mathrm{Reg}_T(\pi) = \sum_{t=1}^{T} \big(\ell_t(a_t) - \ell_t(\pi(\mathcal{A}_t))\big) = \sum_{t=1}^{T} \langle a_t - \pi(\mathcal{A}_t), \theta_t \rangle.$$

A central role is played by the class of *linear* policies, parametrized by a vector $\phi \in \mathbb{R}^d$:

$$\Pi_{\mathrm{lin}} = \{\pi_\phi : \phi \in \mathbb{R}^d\}, \qquad \pi_\phi(\mathcal{A}) = \arg\min_{a \in \mathcal{A}} \langle a, \phi \rangle.$$

We will compete against the optimal policy $\pi^* = \arg\min_{\pi \in \Pi} \mathbb{E}[\sum_{t=1}^{T} \ell_t(\pi(\mathcal{A}_t))]$ that minimizes the expected cumulative loss over all possible policies Π. For the settings with i.i.d. and with stationary contexts, π^* is a linear policy; for non-stationary but independent contexts we will show that π^* can be well approximated by a linear policy in terms of cumulative losses.

2.2 Independent, Identically Distributed Contexts

We first consider the standard i.i.d. regime, where the contexts $\mathcal{A}_1, \dots, \mathcal{A}_T$ are i.i.d., with $\mathcal{A}_t \sim Q$ for some fixed distribution Q. In this case the optimal policy is known to be linear: $\pi^* = \pi_{\phi^*}$ for $\phi^* = \sum_{t=1}^{T} \theta_t$ [Neu and Olkhovskaya, 2020]. For this setting, we obtain the following result:

Theorem 1. *In the i.i.d. contextual setting, the algorithm of Section 3.1 (Algorithm 4) guarantees that, with probability at least* $1 - \delta$*,*

$$\mathrm{Reg}_T(\pi^*) \le O\big(d^3 \mathrm{polylog}(T/\delta) \sqrt{T \log(K)}\big).$$

This is the first computationally efficient algorithm with a guaranteed regret bound in the adversarial linear contextual bandit setting that holds with high probability rather than only in expectation, resolving the open problem of Neu and Olkhovskaya [2020], Olkhovskaya et al. [2023] and Dai et al. [2024].

2.3 Non-Stationary, Independent Contexts

We next relax the i.i.d. assumption on contexts while keeping independence across rounds: the contexts $\mathcal{A}_1, \dots, \mathcal{A}_T$ are independent, with $\mathcal{A}_t \sim P_t$, where the per-round distributions P_t may drift. This is the contextual-bandit analogue of the non-stationary tabular MDP setting of Jin et al. [2023]. We measure the amount of non-stationarity by the cumulative total variation distance from a fixed reference distribution Q. Recall that the total variation distance between two distributions P and Q on a common space is

$$\mathrm{TV}(P, Q) = \int |p - q| \, d\mu,$$

where p and q are densities of P and Q with respect to any common dominating σ-finite measure μ (e.g. $\mu = P + Q$); the integral does not depend on the choice of μ. We then set

$$V_t = \mathrm{TV}(P_t, Q), \qquad V = \sum_{t=1}^{T} V_t.$$

For the analysis we choose Q to minimize V; the learner does not need to know Q.

Choosing the comparator is more delicate here than in the i.i.d. case, because, with time-varying P_t, the optimal policy is no longer linear:

Proposition 2. *In the non-stationary independent setting, the optimal policy is*

$$\pi^*(\mathcal{A}) = \arg\min_{a \in \mathcal{A}} \langle a, \sum_{t=1}^{T} p_t(\mathcal{A})\, \theta_t \rangle,$$

where p_t is the density of P_t w.r.t. $\mu = \sum_{t=1}^{T} P_t$ (or any other σ-finite measure that dominates all P_t).

Proof. Note that, for any policy π,

$$\begin{aligned}\mathbb{E}\,\big[\sum_{t=1}^{T} \ell_t(\pi(\mathcal{A}_t))\big] &= \sum_{t=1}^{T} \mathbb{E}[\langle \pi(\mathcal{A}_t), \theta_t \rangle] = \sum_{t=1}^{T} \int p_t(\mathcal{A}) \langle \pi(\mathcal{A}), \theta_t \rangle \,\mathrm{d}\mu(\mathcal{A}) \\ &= \int \sum_{t=1}^{T} p_t(\mathcal{A}) \langle \pi(\mathcal{A}), \theta_t \rangle \,\mathrm{d}\mu(\mathcal{A}) = \int \langle \pi(\mathcal{A}), \sum_{t=1}^{T} p_t(\mathcal{A}) \theta_t \rangle \,\mathrm{d}\mu(\mathcal{A}),\end{aligned}$$

which is minimized by π^* as claimed. □

This is a non-linear policy in general, because the weight vector on the right-hand side depends on $\mathcal{A}$, but in terms of cumulative losses it can be well-approximated by the linear policy π_{ϕ^*} with $\phi^* = \sum_{t=1}^{T} \theta_t$:

Lemma 3. *In the non-stationary independent setting, with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} \ell_t(\pi_{\phi^*}(\mathcal{A}_t)) \le \sum_{t=1}^{T} \ell_t(\pi^*(\mathcal{A}_t)) + 2V + O\big(\sqrt{T \log(1/\delta)}\big).$$

The proof is in Appendix A.1. Combining this with a regret bound against π_{ϕ^*} yields the following high-probability bound, proved in Section 4.2.

Theorem 4. *In the non-stationary independent setting, the algorithm of Section 3.1 (Algorithm 4) guarantees that, with probability at least $1 - \delta$,*

$$\mathrm{Reg}_T(\pi^*) \le O\big(d^3 \mathrm{polylog}(T/\delta) \sqrt{T \log(K)} + d^{5/2} \mathrm{polylog}(T/\delta)\, V\big).$$

The bound recovers the i.i.d. rate of Theorem 1 when $V = 0$, and is sublinear whenever $V = o(T)$.

2.4 Stationary, Exponentially Mixing Contexts

In our third setting we keep stationarity but drop independence. A stationary stochastic process is normally defined to be an infinite sequence $\ldots, \mathcal{A}_{-2}, \mathcal{A}_{-1}, \mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, \ldots$ such that, for any t and non-negative m and k, the random vectors $(\mathcal{A}_t, \ldots, \mathcal{A}_{t+m})$ and $(\mathcal{A}_{t+k}, \ldots, \mathcal{A}_{t+m+k})$ have the same distribution. In particular, the marginal distribution of $\mathcal{A}_t$ does not depend on t; we denote this common marginal distribution by Q. We assume that the contexts $\mathcal{A}_1, \ldots, \mathcal{A}_T$ are a subsequence of such an infinite stationary sequence, and write P_t^κ for the distribution of $\mathcal{A}_t$ conditional on $\mathcal{A}_1, \ldots, \mathcal{A}_{t-\kappa}$, with $P_t := P_t^1$ for short and P_t^κ taken to be the unconditional distribution Q when $t - \kappa \le 0$.

Using the definition of β-mixing in Mohri and Rostamizadeh [2008], we measure the dependence between contexts by the β-mixing coefficient for any positive integer k:

$$\beta(k) = \sup_n \mathop{\mathbb{E}}_{B \in \sigma_{-\infty}^n} \Big[\sup_{A \in \sigma_{n+k}^\infty} |\Pr(A \mid B) - \Pr(A)| \Big], \tag{1}$$

where for any $i, j \in \mathbb{Z} \cup \{-\infty, \infty\}$ we let σ_i^j denote the σ-algebra generated by the random variables A_k, $i \le k \le j$. The process is *exponentially mixing* if there exist real numbers $\beta_0, \beta_1, r > 0$ such that

$$\beta(k) \le \beta_0 \exp(-\beta_1 k^r) \qquad \text{for all } k. \tag{2}$$

This class includes, for example, all uniformly ergodic Markov chains [Meyn and Tweedie, 1993, Section 16.2] initialized from their stationary distribution.

Under stationarity the optimal policy is again linear, just as in the i.i.d. case:

Proposition 5. *In the stationary, exponentially mixing contexts setting, the optimal policy is linear:* $\pi^* = \pi_{\phi^*}$ *for* $\phi^* = \sum_{t=1}^{T} \theta_t$.

Proof. For any policy π,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(\pi(\mathcal{A}_t))\right] = \sum_{t=1}^{T} \mathbb{E}[\langle \pi(\mathcal{A}_t), \theta_t \rangle] = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathcal{A}\sim Q}[\langle \pi(\mathcal{A}), \theta_t \rangle] = \mathop{\mathbb{E}}_{\mathcal{A}\sim Q}[\langle \pi(\mathcal{A}), \sum_{t=1}^{T} \theta_t \rangle],$$

which is minimized by π_{ϕ^*} as claimed. □

Compared to the i.i.d. setting, the main challenge is the dependence between contexts: applying the reduction with a single instance of the linear bandit subroutine yields an additive regret term that grows linearly in T. We circumvent this by *diluting* the dependence: we interleave κ independent copies of the linear bandit subroutine, each becoming active every κ-th round, so that within a single copy the contexts are κ time steps apart and therefore approximately independent for sufficiently large κ. This is a standard technique in the analysis of mixing processes [Weinberger and Ordentlich, 2002, Mohri and Rostamizadeh, 2008]. Tuning κ to the mixing coefficient gives the following bound, proved in Section 4.2:

Theorem 6. *Suppose the contexts form a stationary, exponentially mixing process satisfying* (2)*. Then the algorithm of Section 3.1 (Algorithm 4) guarantees that, with probability at least* $1 - \delta$*,*

$$\mathrm{Reg}_T(\pi^*) \le O\big(C_{\mathrm{mix}}^2 \ln^{2/r}\left(12T^3/\delta\right) d^3 \mathrm{polylog}(T/\delta)\sqrt{T\log(K)}\big),$$

where $C_{\mathrm{mix}} = \max\{1, ((1 + \max\{0, \ln(\beta_0)\})/\beta_1)^{1/r}\}$.

The C_{mix}^2 and $\ln^{2/r}(12T^3/\delta)$ factors reflect the strength of the dependence: weaker mixing demands a larger κ, hence more interleaved copies and a larger constant.

3 Algorithm

Our algorithm builds on the reduction from adversarial linear contextual bandits to non-contextual misspecified linear bandits introduced by Van Erven et al. [2025]. The reduction rests on the following observation: The conditionally expected loss of a policy π in round t is $\mathbb{E}_t[\langle \pi(\mathcal{A}_t), \theta_t \rangle] = \langle \Psi_t(\pi), \theta_t \rangle$, where

$$\Psi_t(\pi) = \mathbb{E}_t[\pi(\mathcal{A}_t)] \in \mathbb{R}^d$$

is the (conditional) mean action of π. Hence, applying Ψ_t to every policy turns the contextual problem into a non-contextual linear bandit over the action set $\Omega_t = \{\Psi_t(\pi) : \pi \in \Pi\} \subset \mathbb{R}^d$. Since P_t is unknown and Ω_t varies over time, we replace Ψ_t by its empirical counterpart

$$\widehat{\Psi}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \pi(\mathcal{A}_i),$$

based on a sample $\mathcal{S}_N = (\mathcal{A}_1, \ldots, \mathcal{A}_N)$ of N previously observed contexts, and we work over the empirical action set $\widehat{\Omega} = \{\widehat{\Psi}(\pi) : \pi \in \Pi\}$, which does not vary over time. The gap between Ψ_t and $\widehat{\Psi}$ is the source of misspecification in the resulting linear bandit problem. We can control it in all three settings by the following approximations:

$$\Psi_t(\pi) = \mathbb{E}_t[\pi(\mathcal{A}_t)] \approx \mathop{\mathbb{E}}_{\mathcal{A}\sim Q}[\pi(\mathcal{A})] \approx \frac{1}{N}\sum_{i=1}^{N} \pi(\mathcal{A}_i) = \widehat{\Psi}(\pi),$$

which hold uniformly over all *linear* policies π with high probability. Concretely, the linear bandit algorithm produces an action $y_t \in \widehat{\Omega}$, which is decomposed as a convex combination of vertices of $\widehat{\Omega}$; sampling a vertex ψ_t according to the corresponding distribution α_t with mean y_t, and a direction $\phi_t \in -\mathcal{N}(\widehat{\Omega}, \psi_t)$ in its (negated) normal cone yields a linear policy π_{ϕ_t} for which $\widehat{\Psi}(\pi_{\phi_t}) = \psi_t$, so that $\mathbb{E}_{\alpha_t}[\Psi_t(\pi_{\phi_t})] \approx \mathbb{E}_{\alpha_t}[\widehat{\Psi}(\pi_{\phi_t})] = \mathbb{E}_{\alpha_t}[\psi_t] = y_t$. The learner then plays $a_t = \pi_{\phi_t}(\mathcal{A}_t)$ on the actual context.

3.1 Algorithm description

The formal pseudocode of our full algorithm is introduced in Appendix G (Algorithm 4) and computational efficiency is discussed in Appendix G.1. We first describe the main contextual algorithm, which interleaves κ copies of the reduction above to handle dependence between consecutive contexts. We then specify the misspecification-robust linear bandit subroutine that drives each copy. These misspecification-robust linear bandits require a hyperparameter $\bar{C}$ that upper bounds the misspecification that arises from the fact that $\widehat{\Psi}(\pi)$ is not equal to $\Psi_t(\pi)$. Finally, we explain how to handle the unknown hyperparameters κ and $\bar{C}$ via a CORRAL+ wrapper. The full algorithm uses a doubling trick to obtain a sample of previous contexts $\mathcal{S}_N$. After every epoch the algorithm restarts and combines the previous sample with the contexts from the epoch that just finished.

3.2 Contextual Bandit Base Algorithm

Algorithm 1 describes the contextual bandit base learner used by CORRAL+. For a fixed pair of hyperparameters $(\kappa, \bar{C})$, it implements the reduction of Van Erven et al. [2025] using κ interleaved instances of the misspecification-robust linear bandit subroutine of Liu et al. [2024]. It takes as input a sample $\mathcal{S}_N = (\mathcal{A}_1, \ldots, \mathcal{A}_N)$ of N previously observed contexts, a dilution parameter $\kappa \geq 1$, and an upper bound $\bar{C} \geq 0$ on the total misspecification incurred by the reduction. It maintains κ independent copies $\mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(\kappa)}$ of the misspecification-robust linear bandit subroutine of Section 3.3, each operating on the empirical action set $\widehat{\Omega} = \{\widehat{\Psi}(\pi) : \pi \in \Pi\}$ derived from $\mathcal{S}_N$. In round t, copy $\mathcal{L}^{(m)}$ with $m = ((t-1) \bmod \kappa) + 1$ is queried for an action $y_t \in \widehat{\Omega}$, which is inverted through the reduction of Van Erven et al. [2025] to a linear policy π_{ϕ_t} played on the actual context $\mathcal{A}_t$; the resulting loss is fed back to $\mathcal{L}^{(m)}$. In Algorithm 1, t indexes the rounds of the current epoch. The m-th linear bandit instance $\mathcal{L}^{(m)}$ is activated in rounds $(m, m+\kappa, m+2\kappa, \ldots)$. Thus, the r-th internal round of the algorithm $\mathcal{L}^{(m)}$ corresponds to round $t = m + (r-1)\kappa$ of Algorithm 1.

Algorithm 1: Contextual Bandit Base Algorithm

Input: Sample $\mathcal{S}_N = (\mathcal{A}_1, \ldots, \mathcal{A}_N)$ of previously observed contexts; hyperparameters $\bar{\rho}$, $\bar{\delta}$, T and a pair of hyperparameter values $(\kappa, \bar{C})$.

Construct $\widehat{\Omega} = \{\widehat{\Psi}(\pi) : \pi \in \Pi\}$ from $\mathcal{S}_N$, where $\widehat{\Psi}(\pi) = \frac{1}{N}\sum_{n=1}^{N} \pi(\mathcal{A}_n)$.

For each $m = 1, \ldots, \kappa$, instantiate a copy $\mathcal{L}^{(m)}$ of Algorithm 3 over $\widehat{\Omega}$ with inputs $\bar{T} = \lceil T/\kappa \rceil$, $\bar{C}$, $\bar{\rho}$, κ, $\bar{\delta}$.

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

- Set $m \leftarrow ((t-1) \bmod \kappa) + 1$.
- Obtain $y_t \in \widehat{\Omega}$ from $\mathcal{L}^{(m)}$.
- Find $\alpha_t \in \Delta(\mathrm{vert}(\widehat{\Omega}))$ such that $\mathbb{E}_{\psi_t \sim \alpha_t}[\psi_t] = y_t$.
- Sample $\psi_t \sim \alpha_t$ and let ϕ_t be an arbitrary element in the interior of $-\mathcal{N}(\widehat{\Omega}, \psi_t)$.
- Receive context $\mathcal{A}_t$, play $a_t = \arg\min_{a \in \mathcal{A}_t} \langle a, \phi_t \rangle$, and observe loss $\ell_t \in [0,1]$.
- Send ℓ_t to $\mathcal{L}^{(m)}$.

The algorithm requires the sample $\mathcal{S}_N$ as input, but the learner has no prior contexts at the start of play. We therefore run the algorithm in epochs of doubling length: epoch p uses $\widehat{\Omega}$ built from all contexts collected during the first $p-1$ epochs, so that $|\mathcal{S}_N|$ doubles between consecutive epochs. At each epoch boundary, $\widehat{\Omega}$ is reconstructed from the enlarged sample.

3.3 Misspecification-robust Linear Bandit Subroutine

Each copy $\mathcal{L}^{(m)}$ is an instance of the misspecification-robust adversarial linear bandit algorithm of Liu et al. [2024] (their Algorithm 6), restated for convenience as Algorithm 3 in Appendix E. The subroutine operates on a fixed action set $\widehat{\Omega} \subset \mathbb{R}^d$. The time index t in this subsection refers to the internal rounds of a single linear bandit instance, rather than the rounds of the full algorithm. It plays actions $y_t \in \widehat{\Omega}$ over $\bar{T}$ rounds and observes losses $\ell_t \in [0,1]$ whose conditional expectation satisfies $|\,\mathbb{E}_t[\ell_t \mid y_t] - \langle y_t, \theta_t \rangle| \leq \epsilon_t$ for some adversarial misspecification sequence $(\epsilon_t)_{t=1}^{\bar{T}}$ with

$\sum_{t=1}^{\bar{T}} \epsilon_t \le \bar{C}$. Regret is measured against the best fixed action $u \in \widehat{\Omega}$ in terms of the uncorrupted loss:

$$\max_{u \in \widehat{\Omega}} \sum_{t=1}^{\bar{T}} \langle y_t - u, \theta_t \rangle.$$

Liu et al. [2024] prove a high-probability regret bound for this algorithm with explicit dependence on d, $\bar{T}$ and $\bar{C}$. The version of this guarantee that we use, suitably adapted to the importance-weighted feedback received from the CORRAL+ wrapper of Section 3.4, is Lemma 17 in Appendix E.1.

3.4 Tuning Hyperparameters κ and $\bar{C}$

The optimal dilution parameter κ is unknown in the stationary mixing setting (and equals 1 in the i.i.d. and non-stationary independent settings), and a tight upper bound $\bar{C}$ on the misspecification of $\widehat{\Psi}$ is unavailable in the non-stationary independent setting. We address both at once by instantiating one copy of Algorithm 1 for every pair $(\kappa, \bar{C})$ in a finite grid $\mathcal{K} \times \mathcal{C}$, and combining the resulting base learners with CORRAL+: a high-probability variant of the CORRAL meta-algorithm of Agarwal et al. [2017] that we provide in Appendix D.2. The grid specification and pseudocode are deferred to Appendix F and Appendix G; the grid has size $|\mathcal{K}| \cdot |\mathcal{C}| = O(\log^2 T)$, so this wrapping adds only the $\mathrm{polylog}(T/\delta)$ factors already present in our theorems. The wrapper is restarted at each of the $O(\log T)$ epoch boundaries so that it can adapt to the freshly reconstructed $\widehat{\Omega}$, which again only costs $\mathrm{polylog}(T)$.

4 Analysis

In this section we outline the proofs of Theorems 1, 4 and 6. The three results share a common high-probability regret bound, Lemma 7 below, which we then specialize to each setting.

4.1 General regret bound

We state the bound for a single doubling epoch. Recall that with the use of the doubling trick the algorithm partitions the rounds into $P = \lceil \log_2 T \rceil$ epochs. The first epoch contains the first two rounds. For $p \ge 2$, epoch p consists of rounds $\mathcal{T}_p = \{2^{p-1} + 1, 2^{p-1} + 2, \ldots, \min(2^p, T)\}$ and has length $T_p = |\mathcal{T}_p| \le 2^{p-1}$ and the estimates $\widehat{\Psi}_p$ and $\widehat{\Omega}_p$ used during epoch p are formed from data collected before $\mathcal{T}_p$. Throughout this section, we fix one such epoch p and write $\widehat{\Psi} = \widehat{\Psi}_p$, $\mathcal{T} = \mathcal{T}_p$, T_p for the epoch length, and $\mathrm{Reg}_{T_p}(\pi) = \sum_{t \in \mathcal{T}_p} (\ell_t(a_t) - \ell_t(\pi(\mathcal{A}_t)))$ for the regret accumulated within the epoch. We omit the epoch index p from $\widehat{\Omega}_p$ for brevity, since the analysis is restricted to a fixed epoch. Within the analysis of epoch p we also use a local t such that $t = 1$ is the first round of epoch p, not the first overall.

Adapting the linear bandit reduction of Van Erven et al. [2025] to time-varying context distributions, let $\Psi_t^\kappa(\pi) = \mathbb{E}_{t-\kappa+1}[\pi(\mathcal{A}_t)]$ denote the conditional mean action of π in round t given the history up to and including round $t - \kappa$, clamped at 0 for $t < \kappa$. Note that conditioning on 0 still includes all rounds from previous epochs. The total misspecification provided to the m'th instance of Algorithm 3 within epoch p is then

$$C_p^m(\kappa) = \sum_{t=1}^{T_p} \sup_{\phi} \left| \langle \Psi_t^\kappa(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle \right| \mathbb{I}\{(t-1) \bmod (\kappa) = m - 1\}, \tag{3}$$

which corresponds to the weak corruption measure C_∞ of Liu et al. [2024].

Lemma 7. *Fix any doubling epoch p of length T_p. Algorithm 4 guarantees that, with probability at least $1 - \frac{2}{3P}\delta$,*

$$\mathrm{Reg}_{T_p}(\pi_{\phi^*}) \le O\left(\kappa^2 d^{5/2} \sqrt{T_p}\, \mathrm{polylog}(T_p/\delta) + \kappa d^{5/2}\, \mathrm{polylog}(T_p/\delta) \sum_{m=1}^{\kappa} C_p^m(\kappa) \right)$$

for all $\kappa \in \mathcal{K}$ simultaneously, where $C_p^m(\kappa)$ is as defined in (3).

The total regret follows by summing over the $P = O(\log T)$ doubling epochs and a union bound, which only inflates the $\mathrm{polylog}(T/\delta)$ factors. The remaining work in each setting is to choose $\kappa \in \mathcal{K}$ and bound $\sum_{m=1}^{\kappa} C_p^m(\kappa)$ within an epoch with high probability.

4.2 Instantiating the three settings

We now specialize the general regret bound of Lemma 7 to each of the three settings. For each setting, we choose an appropriate value of κ and derive a high-probability bound on the total misspecification incurred within a single epoch p. Substituting these bounds into Lemma 7 and summing over the doubling epochs yields the corresponding regret guarantees. Note that for the final epoch we could have $T_p < 2^{p-1}$ and then the bounds would hold for 2^{p-1} instead of T_p, which costs us a constant factor at most as $2^{p-1} < T$.

IID contexts (Theorem 1). We pick $\kappa = 1$. The following per-epoch bound holds (proof in Appendix C.1):

Lemma 8. *Fix any doubling epoch p of length T_p. With probability at least $1 - \delta$,*

$$C_p^1(1) \le 2\sqrt{2dT_p \ln(T_p K^2)} + \sqrt{2T_p \ln(4T_p/\delta)}.$$

Plugging this with $\delta/(3P)$ into Lemma 7, summing the per-epoch bound over the $P = O(\log T)$ doubling epochs (which is absorbed into the $\mathrm{polylog}(T/\delta)$ factor), and a union bound yield Theorem 1.

Non-stationary independent contexts (Theorem 4). We again pick $\kappa = 1$. Setting the size of the prior context sample to $N = T_p$, the following per-epoch bound holds (proof in Appendix C.2):

Lemma 9. *Fix any doubling epoch p of length T_p. With probability at least $1 - \delta$,*

$$C_p^1(1) \le V_p + \sum_{l=1}^{p-1} V_l + 2\sqrt{2dT_p \ln(T_p K^2)} + \sqrt{2T_p \ln(4T_p/\delta)},$$

where V_l is the total time-variation within epoch l.

Since the optimal policy π^* need not be linear, we additionally invoke Lemma 3, applied to the rounds of epoch p only, to translate the bound from the best linear comparator π_{ϕ^*} to π^*, picking up an additive $2V_p + \tilde{O}(\sqrt{T_p \log(1/\delta)})$. Combining Lemmas 7, 9 and 3 by a union bound (with $\delta/(6P)$ for the latter two) and summing over the $P = O(\log T)$ doubling epochs (absorbed into the $\mathrm{polylog}(T/\delta)$ factor) gives Theorem 4.

Stationary, exponentially mixing contexts (Theorem 6). We then choose the value of $\kappa = \max\{1, \lceil (\ln(6P\beta_0 T_p/\delta)/\beta_1)^{1/r} \rceil\}$ to make the dependence within any single instance of Algorithm 3 negligible. The following bound then holds (proof in Appendix C.4):

Lemma 10. *Fix any doubling epoch p of length T_p, set $\kappa = \max\{1, \left\lceil (\ln(6\beta_0 P T_p/\delta)/\beta_1)^{1/r} \right\rceil\}$. With probability at least $1 - \delta/(3P)$,*

$$\sum_{m=1}^{\kappa} C_p^m(\kappa) \le 96\sqrt{dT_p \ln(48PT_p K^2/\delta)} C_{\mathrm{mix}} \ln^{1/r}\left(\frac{12PT_p^2}{\delta}\right),$$

with $C_{\mathrm{mix}} = \max\{1, ((1 + \max\{0, \ln(\beta_0)\})/\beta_1)^{1/r}\}$, for exponentially mixing contexts with β_0, β_1, r as in (2)*.*

Plugging into Lemma 7, filling in $\kappa = \max\{1, \lceil (\ln(6P\beta_0 T_p/\delta)/\beta_1)^{1/r} \rceil\} \le 2C_{\mathrm{mix}} \ln^{1/r}\left(\frac{12PT_p^2}{\delta}\right)$ and summing over the $P = O(\log T)$ doubling epochs (absorbed into the $\mathrm{polylog}(T/\delta)$ factor) yields Theorem 6. The value of κ might not be in $\mathcal{K}$ but then we take the next value which is at most a constant factor 2 larger. If there is no such larger value the bound is vacuous due to the κ^2 factor.

4.3 Proof sketch of Lemma 7

Recall that CORRAL+ combines M contextual base learners, each corresponding to a different hyperparameter configuration $(\kappa, \bar{C})$. We split the regret of the meta algorithm into the regret of CORRAL+ against an arbitrary base i for some $i \in [M]$ and the regret of that same base under importance-weighted feedback against the best fixed linear comparator:

$$\sum_{t=1}^{T}\big(\ell_t(a_t) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))\big) \;=\; \underbrace{\sum_{t=1}^{T}\big(\ell_t(a_t) - \ell_t(\pi_t^i(\mathcal{A}_t))\big)}_{\mathrm{CORRAL}} \;+\; \underbrace{\sum_{t=1}^{T}\big(\ell_t(\pi_t^i(\mathcal{A}_t)) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))\big)}_{\mathrm{BASE}_i}. \tag{4}$$

Unlike in Agarwal et al. [2017], BASE_i measures the regret on the *real* losses under importance-weighted feedback, not the regret on the importance-weighted losses themselves.

Bounding BASE_i**.** Each contextual base learner (Algorithm 1) maintains κ instances of the linear bandit subroutine (the algorithm of Liu et al. [2024]), operating on the empirical action set $\widehat{\Omega}$. Adapting the analysis of Liu et al. [2024] to the importance-weighted feedback supplied by CORRAL+, with misspecification bounded by $C_p^m(\kappa)$ for each instance m, and controlling five additional terms via Azuma's inequality and the definition (3), yields (as shown in Lemma 17 in Appendix E.1), with probability at least $1 - \frac{1}{2P}\delta$ jointly over $\kappa \in \mathcal{K}$,

$$\begin{aligned}\mathrm{BASE}_i \;&\le O\Big(d^{3/2}\,\rho_{T,i}\sqrt{T}\,\mathrm{polylog}(T/\delta)\Big) + O\Big(\kappa d^{5/2}\,\mathrm{polylog}(T/\delta)\textstyle\sum_{m=1}^{\kappa} C_p^m(\kappa)\Big)\\ &\quad + O\Big(\kappa d^{5/2}\sqrt{\rho_{T,i}T}\,\mathrm{polylog}(T/\delta)\Big),\end{aligned} \tag{5}$$

where $\rho_{T,i}$ is as determined by the CORRAL+ algorithm and bounds the maximum importance-weight given to base i throughout the T rounds. This bound assumes that the value of $\bar{C}$ for this base is large enough, which will generally not be the case for all base models, but for any value of $\kappa \in \mathcal{K}$ there will be a base model that attains this bound.

Bounding CORRAL **and combining.** Recall from Section 3.4 that CORRAL+ is our high-probability variant of CORRAL [Agarwal et al., 2017]. The full pseudocode and high-probability analysis are in Appendix D; here we state the two guarantees we use in the regret decomposition above. The first is (5); the second is Lemma 15, which bounds CORRAL by $\tilde{O}(M/\eta^* + M\sqrt{T} + \rho_{T,i} + \eta^* T) - \rho_{T,i}/(20\eta^* \ln T) + O(\sqrt{\rho_{T,i} T \log(PTM/\delta)})$, uniformly over $i \in [M]$, with probability at least $1 - \delta/(6P)$, for $\eta^* > 0$, which is the starting value of the learning rates of CORRAL+. Adding BASE_i and CORRAL in (5) and choosing learning rate η^* in CORRAL+ so as to cancel the $\rho_{T,i}$-dependent terms eliminates the dependence on the unknown $\rho_{T,i}$ and yields the bound stated in Lemma 7. We choose η^* specifically to cancel out the $\rho_{T,i}$ term. Finding such a value of η^* that does not depend on the hyperparameters is still possible because the $\rho_{T,i}$ term only depends on d and T. The full proof of Lemma 7 is given in Appendix F.

5 Conclusion

We obtained the first computationally efficient high-probability regret bounds for adversarial linear contextual bandits, both in the i.i.d. setting (resolving the open problem of Neu and Olkhovskaya [2020], Olkhovskaya et al. [2023], Dai et al. [2024]) and under non-stationary independent or stationary mixing contexts. A natural direction for future work is to improve the dependence on the dimension d, since the current d^3 dependence is much worse than the $d^{3/2}$ dependence of Van Erven et al. [2025] for expected regret in the i.i.d. setting. The obvious approach would be to try different linear bandit algorithms for the linear bandit reduction. This should be more straightforward for the i.i.d. and the stationary mixing setting as the per-round misspecification can be bounded for those settings. For example, Algorithm 1 from Zimmert and Lattimore [2022] could be a good candidate as it is computationally efficient, has good dependence on d and already has high-probability bounds.

Acknowledgments and Disclosure of Funding

Olkhovskaya and Van Erven were supported by the Netherlands Organization for Scientific Research (NWO) under grants https://doi.org/10.61686/QVRIJ08667 and https://doi.org/10.61686/OWREN85146, respectively.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320, 2011.

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 3–11, 1999.

N. Abe, A. W. Biermann, and P. M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135. PMLR, 2013. URL https://proceedings.mlr.press/v28/agrawal13.html.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 19–26. PMLR, 2011.

K. Buyukkalayci, O. Hanna, and C. Fragouli. A reduction algorithm for Markovian contextual linear bandits. *arXiv preprint arXiv:2603.12530*, 2026.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 208–214. PMLR, 2011.

Y. Dai, H. Luo, C.-Y. Wei, and J. Zimmert. Refined regret for adversarial MDPs with linear function approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6726–6759, 2023. URL https://proceedings.mlr.press/v202/dai23b.html.

Y. Dai, Q. Cui, and S. S. Du. Refined sample complexity for Markov games with independent linear function approximation. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 1260–1261. PMLR, 2024.

O. A. Hanna, L. F. Yang, and C. Fragouli. Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. In G. Neu and L. Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1791–1821. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/hanna23a.html.

S. Ito, S. Hirahara, T. Soma, and Y. Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33:2028–2038, 2020.

T. Jin, J. Liu, C. Rouyer, W. Chang, C.-Y. Wei, and H. Luo. No-regret online reinforcement learning with adversarial losses and transitions. *Advances in Neural Information Processing Systems*, 36: 38520–38585, 2023.

C.-W. Lee, H. Luo, C.-Y. Wei, and M. Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs. *Advances in neural information processing systems*, 33: 15522–15533, 2020.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. Association for Computing Machinery, 2010. doi: 10.1145/1772690.1772758.

H. Liu, C.-Y. Wei, and J. Zimmert. Bypassing the simulator: Near-optimal adversarial linear contextual bandits. *Advances in Neural Information Processing Systems*, 36:52086–52131, 2023.

H. Liu, A. Tajdini, A. Wagenmaker, and C.-Y. Wei. Corruption-robust linear bandits: Minimax optimality and gap-dependent misspecification. *Advances in Neural Information Processing Systems*, 37:24277–24325, 2024.

S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer Verlag, 2nd edition, 1993.

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-IID processes. *Advances in neural information processing systems*, 21, 2008.

G. Neu and J. Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory*, pages 3049–3068. PMLR, 2020.

G. Neu, J. Olkhovskaya, and S. Vakili. Adversarial contextual bandits go kernelized. In *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 907–929. PMLR, 2024. URL `https://proceedings.mlr.press/v237/neu24a.html`.

J. Olkhovskaya, J. Mayo, T. Van Erven, G. Neu, and C.-Y. Wei. First-and second-order bounds for adversarial linear contextual bandits. *Advances in Neural Information Processing Systems*, 36: 61625–61644, 2023.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

G. Stoltz. *Information incomplète et regret interne en prédiction de suites individuelles*. PhD thesis, Université Paris–XI, 2005.

V. Syrgkanis, A. Krishnamurthy, and R. Schapire. Efficient algorithms for adversarial contextual learning. In *International Conference on Machine Learning*, pages 2159–2168. PMLR, 2016.

A. Tewari and S. A. Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health: Sensors, Analytic Methods, and Applications*, pages 495–517. Springer, 2017.

T. Van Erven, J. Mayo, J. Olkhovskaya, and C.-Y. Wei. An improved algorithm for adversarial linear contextual bandits via reduction. *arXiv preprint arXiv:2508.11931*, 2025.

C.-Y. Wei, M. J. Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

M. J. Weinberger and E. Ordentlich. On delayed prediction of individual sequences. *IEEE Transactions on Information Theory*, 48(7):1959–1976, 2002. doi: 10.1109/TIT.2002.1013136.

J. Zimmert and T. Lattimore. Return of the bias: Almost minimax optimal high probability bounds for adversarial linear bandits. In *Conference on Learning Theory*, pages 3285–3312. PMLR, 2022.

A Proofs for Section 2

A.1 Proof of Lemma 3

Lemma 3. *In the non-stationary independent setting, with probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T}\ell_t(\pi_{\phi^*}(\mathcal{A}_t)) \le \sum_{t=1}^{T}\ell_t(\pi^*(\mathcal{A}_t)) + 2V + O\big(\sqrt{T\log(1/\delta)}\big).$$

Proof. We start with the difference between the cumulative losses under the best fixed linear policy π_{ϕ^*} and the best fixed policy π^*. By adding and subtracting the per-round expectations $\mathbb{E}_t[\cdot]$ under P_t and the expectations under the reference distribution Q for the two policies, we decompose the difference of cumulative losses into five terms that we then bound individually with high probability:

$$\begin{aligned}
\sum_{t=1}^{T}\ell_t(\pi_{\phi^*}(\mathcal{A}_t)) - \sum_{t=1}^{T}\ell_t(\pi^*(\mathcal{A}_t)) = &\underbrace{\sum_{t=1}^{T}\big(\ell_t(\pi_{\phi^*}(\mathcal{A}_t)) - \mathbb{E}_t\left[\ell_t(\pi_{\phi^*}(\mathcal{A}_t))\right]\big)}_{\text{(I)}}\\
&+\underbrace{\sum_{t=1}^{T}\big(\mathbb{E}_t\left[\ell_t(\pi_{\phi^*}(\mathcal{A}_t))\right] - \mathbb{E}_{\mathcal{A}\sim Q}\left[\ell_t(\pi_{\phi^*}(\mathcal{A}))\right]\big)}_{\text{(II)}}\\
&+\underbrace{\sum_{t=1}^{T}\big(\mathbb{E}_{\mathcal{A}\sim Q}\left[\ell_t(\pi_{\phi^*}(\mathcal{A}))\right] - \mathbb{E}_{\mathcal{A}\sim Q}\left[\ell_t(\pi^*(\mathcal{A}))\right]\big)}_{\text{(III)}}\\
&+\underbrace{\sum_{t=1}^{T}\big(\mathbb{E}_{\mathcal{A}\sim Q}\left[\ell_t(\pi^*(\mathcal{A}))\right] - \mathbb{E}_t\left[\ell_t(\pi^*(\mathcal{A}_t))\right]\big)}_{\text{(IV)}}\\
&+\underbrace{\sum_{t=1}^{T}\big(\mathbb{E}_t\left[\ell_t(\pi^*(\mathcal{A}_t))\right] - \ell_t(\pi^*(\mathcal{A}_t))\big)}_{\text{(V)}}.
\end{aligned}$$

Terms (I) and (V) are sums of bounded martingale differences in $[-1,1]$; by Azuma's inequality and a union bound, each is at most $O(\sqrt{T\log(1/\delta)})$ with probability at least $1-\delta/2-\delta/2 = 1-\delta$.

For terms (II) and (IV), fix any policy $\pi\in\Pi$ and round t. Letting $\mu = P_t + Q$ and writing p_t, q for the corresponding densities, since $|\ell_t(\pi(\mathcal{A}))|\le 1$,

$$\begin{aligned}
\big|\,\mathbb{E}_t\left[\ell_t(\pi(\mathcal{A}_t))\right] - \mathbb{E}_{\mathcal{A}\sim Q}\left[\ell_t(\pi(\mathcal{A}))\right]\big| &= \left|\int\big(p_t(\mathcal{A}) - q(\mathcal{A})\big)\,\ell_t(\pi(\mathcal{A}))\,\mathrm{d}\mu(\mathcal{A})\right|\\
&\le \int |p_t(\mathcal{A}) - q(\mathcal{A})|\,\mathrm{d}\mu(\mathcal{A}) = V_t.
\end{aligned}$$

Instantiating with $\pi=\pi_{\phi^*}$ for (II) and $\pi=\pi^*$ for (IV), and summing over t, bounds each of them by $V=\sum_{t=1}^{T}V_t$.

For term (III), we have that

$$\sum_{t=1}^{T}\mathbb{E}_{\mathcal{A}\sim Q}\left[\ell_t(\pi(\mathcal{A}))\right] = \mathbb{E}_{\mathcal{A}\sim Q}\left[\left\langle\pi(\mathcal{A}), \textstyle\sum_{t=1}^{T}\theta_t\right\rangle\right] = \mathbb{E}_{\mathcal{A}\sim Q}\left[\langle\pi(\mathcal{A}),\phi^*\rangle\right],$$

is minimized over all policies by the linear policy π_{ϕ^*}. Hence term (III) is non-positive.

Combining all bounds completes the proof. □

B Uniform Convergence Lemmas

This section collects three uniform convergence bounds used in the proofs of the total misspecification high-probability bounds in Appendix C: one for the i.i.d. setting, one for the non-stationary independent setting, and one for the stationary dependent setting.

B.1 Independent, Identically Distributed Contexts

Lemma 11. *With probability at least* $1-\delta$*,*

$$\sup_{\phi} |\langle \Psi(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \leq 2\sqrt{\frac{2d\ln(NK^2)}{N}} + \sqrt{\frac{2\ln(4/\delta)}{N}},$$

where $\widehat{\Psi}$ *is based on a sample of* N *previous contexts.*

See derivation of Lemma 1 in Van Erven et al. [2025]

B.2 Non-Stationary, Independent Contexts

Lemma 12. *With probability at least* $1-\delta$*,*

$$\sup_{\phi} |\langle \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \leq \frac{1}{N}\sum_{i=1}^{N} \tilde{V}_i + 2\sqrt{\frac{2d\ln(NK^2)}{N}} + \sqrt{\frac{2\ln(4/\delta)}{N}},$$

where $\Psi_{\min}(\pi)$ *is the expected action played by policy* π *under the reference/minimizer distribution* Q *(see Section 2.3) and* $\tilde{V}_i$ *is the total variation distance from the context distribution corresponding to the* i*'th context from the set of previous contexts* S_N *from distribution* Q*.*

Proof. We start by decomposing the expression as follows

$$\begin{aligned}\sup_{\phi} |\langle \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \leq &\sup_{\phi} |\langle \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \Psi_{1:N}(\pi_\phi), \theta_t \rangle| \\ &+ \sup_{\phi} |\langle \Psi_{1:N}(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle|,\end{aligned}$$

where we have defined $\Psi_{1:N}(\pi) = \frac{1}{N}\sum_{i=1}^{N} \Psi_i(\pi)$ and $\Psi_i(\pi)$ is the expected action played by policy π under the context distribution corresponding to the i'th context from the sample of previous contexts S_N. We then bound the first term as follows

$$\begin{aligned}|\langle \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \Psi_{1:N}(\pi_\phi), \theta_t \rangle| &= |\langle \frac{1}{N}\sum_{i=1}^{N} \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \frac{1}{N}\sum_{i=1}^{N} \Psi_i(\pi_\phi), \theta_t \rangle| \\ &= |\frac{1}{N}\sum_{i=1}^{N} (\langle \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \Psi_i(\pi_\phi), \theta_t \rangle)| \\ &\leq \frac{1}{N}\sum_{i=1}^{N} |\langle \Psi_{\min}(\pi_\phi), \theta_t \rangle - \langle \Psi_i(\pi_\phi), \theta_t \rangle| \\ &\leq \frac{1}{N}\sum_{i=1}^{N} \tilde{V}_i,\end{aligned}$$

where $\tilde{V}_i$ is the time-variation for the i'th round in the sample of previous contexts S_N. The second term can be bounded with probability at least $1-\delta$ by $2\sqrt{\frac{2d\ln(NK^2)}{N}} + \sqrt{\frac{2\ln(4/\delta)}{N}}$, since the derivation of Lemma 11 from the i.i.d. setting still applies to non-identically distributed contexts: both McDiarmid's inequality and the symmetrization argument go through unchanged. See Appendix C in Van Erven et al. [2025] (and Chapter 26 in Shalev-Shwartz and Ben-David [2014]) for the i.i.d. setting. □

B.3 Stationary, Exponentially Mixing Contexts

Lemma 13. *With probability at least* $1-\delta$*,*

$$\sup_{\phi} |\langle \Psi_0(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \leq 16\sqrt{\frac{d \ln(NK^2) C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}} + 4\sqrt{\frac{\ln(8/\delta) C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}} + \frac{16 C_{\mathrm{mix}}(\frac{\delta}{2N})}{N},$$

where $C_{\mathrm{mix}}(\delta) = \max\{1, \left\lceil \left(\frac{\ln(\beta_0/\delta)}{\beta_1}\right)^{1/r} \right\rceil\}$ *and where* $\Psi_0(\pi)$ *is the expected action of policy* π *under the stationary context distribution and* $\widehat{\Psi}$ *is based on a sample of* N *contexts from an exponentially mixing process with* β_0, β_1, r *as in* (2)*.*

Proof. By Theorem 1 of Mohri and Rostamizadeh [2008], for any sample from a stationary, exponentially mixing process of size N, with probability at least $1-\delta$,

$$\sup_{h \in \mathcal{H}} \left(R_N(h) - \hat{R}_N(h) \right) \leq \mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(2/\delta')}{2\mu}},$$

where $\delta' = \delta - 2(\mu - 1)\beta(a)$ needs to be positive and where the sample size N has been divided into blocks such that $N = 2\mu a$. $\mathcal{R}_N(\mathcal{H})$ is the expectation of the empirical Rademacher complexity of the hypothesis class $\mathcal{H}$ for a sample of size N. Unlike the Rademacher complexity in Van Erven et al. [2025] theirs already contains a factor 2 so it doesn't appear in the expression above explicitly, but it is present. $R_N(h)$ and $\hat{R}_N(h)$ are the true risk and the empirical risk of hypothesis h on a sample of size N, respectively.

We let $a(\delta/N) = \min_{p \in \mathbb{N}}\{2^p : 2^p \geq \max\{2, \left\lceil \left(\frac{\ln(\beta_0 N/\delta)}{\beta_1}\right)^{1/r} \right\rceil\}\}$ so that $\delta' \geq \delta - 2\mu\beta(a) \geq \delta - 2\mu\frac{\delta}{N} = \delta/2$ so we can fill that into the formula for δ'. Note that this is strictly for the analysis, the algorithm does not need to know a or the mixing parameters. We choose a to be a power of 2 so that we guarantee that $\mu = \frac{N}{2a}$ is always an integer for $N \geq 2a$. This works because of how we collect the samples through the doubling trick (we first start with 2 samples). We then fill in $\mu = N/(2a)$ into the expression obtained from Mohri and Rostamizadeh [2008] using that $a(\delta/N) \leq 8C_{\mathrm{mix}}(\delta/N)$. We also add $\frac{2a(\delta/N)}{N} \leq 16C_{\mathrm{mix}}(\delta/N)/N$ to compensate for the case where N is too small for a positive integer value of μ.

Then letting h range over the linear classifiers ϕ and $\mathcal{H}$ be the set of all such classifiers, we have

$$\begin{aligned}\sup_{\phi} |\langle \Psi_0(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| &= \sup_{h \in \mathcal{H}} \left| R_N(h) - \hat{R}_N(h) \right| \\ &\leq 16\sqrt{\frac{d \ln(NK^2) C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}} + 4\sqrt{\frac{\ln(8/\delta) C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}} \\ &\quad + \frac{16 C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}\end{aligned}$$

similarly to Van Erven et al. [2025], who define a multiclass classification problem (including the cost functions) in such a way that they obtain the first equality. They then bound the expected Rademacher complexity $\mathcal{R}_N(\mathcal{H})$ by bounding the empirical Rademacher complexity first, which means we can do the same as it doesn't take the distribution of the sample into account. Finally, we replace δ with $\delta/2$, which comes from a union bound to handle the absolute value, applying the result from Mohri and Rostamizadeh [2008] twice at confidence $1 - \delta/2$. See Appendix C in Van Erven et al. [2025] for details. □

C Total Misspecification High-Probability Bounds

This appendix proves Lemmas 8, 9, and 10. Each proof works within a single doubling epoch p of length T_p, indexing its rounds as $t = 1, \ldots, T_p$ for notational simplicity, and uses an estimator $\widehat{\Psi}$

built from the $N = T_p$ contexts of the previous epochs. The relevant uniform convergence lemma from Appendix B is invoked with δ rescaled by $1/T_p$, so that a union bound over the T_p rounds of the epoch keeps the total failure probability at δ.

C.1 Independent, Identically Distributed Contexts

Lemma 8. *Fix any doubling epoch p of length T_p. With probability at least $1-\delta$,*

$$C_p^1(1) \le 2\sqrt{2dT_p \ln(T_p K^2)} + \sqrt{2T_p \ln(4T_p/\delta)}.$$

Proof. Apply Lemma 11 with δ/T_p in place of δ, take a union bound over the T_p rounds, and sum the per-round bound to obtain the lemma. □

C.2 Non-Stationary, Independent Contexts

Lemma 9. *Fix any doubling epoch p of length T_p. With probability at least $1-\delta$,*

$$C_p^1(1) \le V_p + \sum_{l=1}^{p-1} V_l + 2\sqrt{2dT_p \ln(T_p K^2)} + \sqrt{2T_p \ln(4T_p/\delta)},$$

where V_l is the total time-variation within epoch l.

Proof. We start with the uniform convergence lemma for this setting:

Lemma 12. *With probability at least $1-\delta$,*

$$\sup_{\phi} |\langle \Psi_{\min}(\pi_\phi), \theta_t\rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t\rangle| \le \frac{1}{N}\sum_{i=1}^{N} \tilde{V}_i + 2\sqrt{\frac{2d\ln(NK^2)}{N}} + \sqrt{\frac{2\ln(4/\delta)}{N}},$$

where $\Psi_{\min}(\pi)$ is the expected action played by policy π under the reference/minimizer distribution Q (see Section 2.3) and $\tilde{V}_i$ is the total variation distance from the context distribution corresponding to the i'th context from the set of previous contexts S_N from distribution Q.

For the proof see Appendix B.2. Applying it with δ/T_p in place of δ and taking a union bound over the T_p rounds of the epoch, we obtain for the misspecification in the non-stationary setting (with $\kappa = 1$):

$$\begin{aligned}
C_p^1(1) &= \sum_{t=1}^{T_p} \sup_{\phi} |\langle \Psi_t(\pi_\phi), \theta_t\rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t\rangle| \\
&\le \sum_{t=1}^{T_p} \sup_{\phi} |\langle \Psi_t(\pi_\phi), \theta_t\rangle - \langle \Psi_{\min}(\pi_\phi), \theta_t\rangle| + \sum_{t=1}^{T_p} \sup_{\phi} |\langle \Psi_{\min}(\pi_\phi), \theta_t\rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t\rangle| \\
&\le \sum_{t=1}^{T_p} V_t + \sum_{t=1}^{T_p} \frac{1}{N}\sum_{i=1}^{N} \tilde{V}_i + 2\sqrt{\frac{2d\ln(NK^2)}{N}} + \sqrt{\frac{2\ln(4T_p/\delta)}{N}} \\
&= V_p + \sum_{l=1}^{p-1} V_l + 2\sqrt{2dT_p \ln(T_p K^2)} + \sqrt{2T_p \ln(4T_p/\delta)},
\end{aligned}$$

where V_l is the total time-variation within epoch l. We have used the fact that due to the doubling the number of previous context samples N is equal to T_p. □

C.3 High-Probability Bound on Mixing Time-Variation

For exponentially mixing contexts, the cumulative κ-step-ahead conditional total-variation distance from the stationary distribution is small with high probability, provided the lookahead κ is large enough. The following lemma is used in the proof of Lemma 10.

Lemma 14. *Let the context process be exponentially mixing in the sense of* (2)*, with stationary distribution Q and parameters $\beta_0, \beta_1, r > 0$. Set*

$$\kappa = \max\{1, \left\lceil \left(\frac{\ln(\beta_0 T/\delta)}{\beta_1} \right)^{1/r} \right\rceil \}.$$

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

$$\sum_{t=1}^{T-\kappa} \mathrm{TV}(P_{t+\kappa}^{\kappa}, Q) \;\le\; 2\sqrt{T},$$

where P_t^{κ} is the distribution of the context $\mathcal{A}_t$ of round t conditional on $\mathcal{A}_1, \ldots, \mathcal{A}_{t-\kappa}$.

Proof. Fix $t \in \{1, \ldots, T-\kappa\}$ and let $\mathcal{F}_t$ denote the σ-algebra generated by $\mathcal{A}_1, \ldots, \mathcal{A}_t$, which is contained in the full past $\sigma_{-\infty}^{t}$. For any measurable set B of contexts, stationarity gives $\Pr(\mathcal{A}_{t+\kappa} \in B) = Q(B)$, so by the tower property of conditional expectation

$$\begin{aligned} P_{t+\kappa}^{\kappa}(B) - Q(B) &= \Pr(\mathcal{A}_{t+\kappa} \in B \mid \mathcal{F}_t) - \Pr(\mathcal{A}_{t+\kappa} \in B) \\ &= \mathbb{E}\left[\Pr(\mathcal{A}_{t+\kappa} \in B \mid \sigma_{-\infty}^{t}) - \Pr(\mathcal{A}_{t+\kappa} \in B) \;\middle|\; \mathcal{F}_t \right]. \end{aligned}$$

Since our total variation distance equals twice the supremum over measurable sets, Jensen's inequality for the conditional expectation yields

$$\begin{aligned} \mathrm{TV}(P_{t+\kappa}^{\kappa}, Q) &= 2 \sup_{B} \left| P_{t+\kappa}^{\kappa}(B) - Q(B) \right| \\ &\le 2\, \mathbb{E}\left[\sup_{B} \left| \Pr(\mathcal{A}_{t+\kappa} \in B \mid \sigma_{-\infty}^{t}) - \Pr(\mathcal{A}_{t+\kappa} \in B) \right| \;\middle|\; \mathcal{F}_t \right] \\ &\le 2\, \mathbb{E}\left[\sup_{A \in \sigma_{t+\kappa}^{\infty}} \left| \Pr(A \mid \sigma_{-\infty}^{t}) - \Pr(A) \right| \;\middle|\; \mathcal{F}_t \right], \end{aligned}$$

where the last step enlarges the supremum from events of the form $\{\mathcal{A}_{t+\kappa} \in B\}$ to all events in $\sigma_{t+\kappa}^{\infty}$. Taking expectations on both sides and recognizing the term with $n = t$ in the definition of $\beta(\kappa)$ in (1), we obtain

$$\mathbb{E}\left[\mathrm{TV}(P_{t+\kappa}^{\kappa}, Q) \right] \le 2\, \mathbb{E}\left[\sup_{A \in \sigma_{t+\kappa}^{\infty}} \left| \Pr(A \mid \sigma_{-\infty}^{t}) - \Pr(A) \right| \right] \le 2\beta(\kappa).$$

Summing over t and using linearity of expectation then gives

$$\mathbb{E}\left[\sum_{t=1}^{T-\kappa} \mathrm{TV}(P_{t+\kappa}^{\kappa}, Q) \right] \;\le\; 2T\, \beta(\kappa).$$

For the chosen κ, exponential mixing gives $\beta(\kappa) \le \beta_0 \exp(-\beta_1 \kappa^r) \le \delta/T$. Markov's inequality then yields

$$\Pr\left(\sum_{t=1}^{T-\kappa} \mathrm{TV}(P_{t+\kappa}^{\kappa}, Q) \ge 2\sqrt{T} \right) \;\le\; \frac{T\, \beta(\kappa)}{\sqrt{T}} \;\le\; \frac{\delta}{\sqrt{T}} \;\le\; \delta. \qquad \square$$

C.4 Stationary, Exponentially Mixing Contexts

Lemma 10. *Fix any doubling epoch p of length T_p, set $\kappa = \max\{1, \left\lceil (\ln(6\beta_0 P T_p/\delta)/\beta_1)^{1/r} \right\rceil\}$. With probability at least $1 - \delta/(3P)$,*

$$\sum_{m=1}^{\kappa} C_p^m(\kappa) \le 96\sqrt{d T_p \ln(48 P T_p K^2/\delta)}\, C_{\mathrm{mix}} \ln^{1/r}\left(\frac{12 P T_p^2}{\delta} \right),$$

with $C_{\mathrm{mix}} = \max\{1, ((1 + \max\{0, \ln(\beta_0)\})/\beta_1)^{1/r}\}$, for exponentially mixing contexts with β_0, β_1, r as in (2)*.*

Proof. We start with the uniform convergence lemma for this setting:

Lemma 13. *With probability at least* $1-\delta$*,*

$$\sup_{\phi} |\langle \Psi_0(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \leq 16\sqrt{\frac{d \ln(NK^2) C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}} + 4\sqrt{\frac{\ln(8/\delta) C_{\mathrm{mix}}(\frac{\delta}{2N})}{N}} + \frac{16 C_{\mathrm{mix}}(\frac{\delta}{2N})}{N},$$

where $C_{\mathrm{mix}}(\delta) = \max\{1, \left\lceil \left(\frac{\ln(\beta_0/\delta)}{\beta_1}\right)^{1/r} \right\rceil\}$ *and where* $\Psi_0(\pi)$ *is the expected action of policy* π *under the stationary context distribution and* $\widehat{\Psi}$ *is based on a sample of* N *contexts from an exponentially mixing process with* β_0, β_1, r *as in* (2)*.*

See Appendix B.3 for the derivation based on results from Mohri and Rostamizadeh [2008]. We then have with probability at least $1 - \delta/(3P)$

$$\begin{aligned}
\sum_{m=1}^{\kappa} C_p^m(\kappa) &= \sum_{t=1}^{T_p} \sup_{\phi} |\langle \Psi_t^\kappa(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \\
&\leq \sum_{t=1}^{T_p} \sup_{\phi} |\langle \Psi_t^\kappa(\pi_\phi), \theta_t \rangle - \langle \Psi_0(\pi_\phi), \theta_t \rangle| + \sum_{t=1}^{T_p} \sup_{\phi} |\langle \Psi_0(\pi_\phi), \theta_t \rangle - \langle \widehat{\Psi}(\pi_\phi), \theta_t \rangle| \\
&\leq \kappa + \sum_{t=1}^{T_p - \kappa} \mathrm{TV}(P_{t+\kappa}^\kappa, Q) + 16 T_p \sqrt{\frac{d \ln(NK^2) C_{\mathrm{mix}}(\frac{\delta}{12PNT_p})}{N}} \\
&\quad + 4T_p \sqrt{\frac{\ln(48PT_p/\delta) C_{\mathrm{mix}}(\frac{\delta}{12PNT_p})}{N}} + 16 T_p \frac{C_{\mathrm{mix}}(\frac{\delta}{12PNT_p})}{N} \\
&\leq \kappa + 2\sqrt{T_p} + 16\sqrt{dT_p \ln(T_p K^2) C_{\mathrm{mix}}\left(\frac{\delta}{12PT_p^2}\right)} \\
&\quad + 4\sqrt{T_p \ln(48PT_p/\delta) C_{\mathrm{mix}}\left(\frac{\delta}{12PT_p^2}\right)} + 16 C_{\mathrm{mix}}\left(\frac{\delta}{12PT_p^2}\right) \\
&\leq 48\sqrt{dT_p \ln(48PT_p K^2/\delta)} C_{\mathrm{mix}}\left(\frac{\delta}{12PT_p^2}\right) \\
&\leq 96\sqrt{dT_p \ln(48PT_p K^2/\delta)} C_{\mathrm{mix}} \ln^{1/r}\left(\frac{12PT_p^2}{\delta}\right),
\end{aligned}$$

where the first term is bounded by Lemma 14 applied with $\delta/(6P)$ (so $\kappa = \max\{1, \lceil (\ln(6\beta_0 PT_p/\delta)/\beta_1)^{1/r} \rceil\}$), and the second uses Lemma 13 with $\delta/(6PT_p)$ together with a union bound over the T_p rounds. A final union bound combines the two failure events into a total probability of at most $\delta/(3P)$. We have also used the fact that due to the doubling the number of previous context samples N is equal to T_p. The added κ arises because the first κ rounds of an epoch were excluded from Lemma 14. That is because the first $\kappa - 1$ rounds do not have the required dilution of κ because they use the data from the previous epochs through $\widehat{\Omega}$. And while round κ does have the dilution of κ it was still added because Lemma 14 started the sum from $t = 1$. The misspecification for a round can be at most 1 and we had κ of these rounds. The final inequality uses that $C_{\mathrm{mix}}\left(\frac{\delta}{12PT_p^2}\right) \leq 2 C_{\mathrm{mix}} \ln^{1/r}\left(\frac{12PT_p^2}{\delta}\right)$, where we have split it into a constant that only depends on the mixing parameters and a factor that also depends on T_p and δ. □

D CORRAL+

In this section we present CORRAL+, a high-probability variant of CORRAL [Agarwal et al., 2017] which serves as a generic wrapper for combining base algorithms. We first describe the idea behind

CORRAL+ (Section D.1), then provide the algorithm pseudocode (Section D.2), and finally state and prove a high-probability regret bound for the multi-armed bandit part of the regret decomposition (Section D.3). The proof of Lemma 7, which uses CORRAL+ to tune the hyperparameters of our linear bandit base algorithm, is given in Appendix F.

D.1 High-Level Idea CORRAL+

The idea behind CORRAL from Agarwal et al. [2017] is that we want to be able to compete with the best model/base algorithm in a set of models/base algorithms. We allow the regret to depend on the number of models as well as the regret of the base algorithms. The idea behind CORRAL is that we can treat the base algorithms as arms in an adversarial multi-armed bandit algorithm and then use regret bounds for that algorithm to obtain results against the best algorithm. This decomposition of the regret of the learner against the best fixed policy is as follows

$$\begin{aligned}\sum_{t=1}^{T}(\ell_t(a_t) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))) &= \sum_{t=1}^{T}(\ell_t(a_t) - \ell_t(\pi_t^i(\mathcal{A}_t))) + \sum_{t=1}^{T}(\ell_t(\pi_t^i(\mathcal{A}_t)) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))) \\ &\leq \mathrm{CORRAL} + \mathrm{BASE}_i,\end{aligned}$$

where BASE_i is the regret of the base algorithm corresponding to arm i in the multi-armed bandit and CORRAL is the regret of the multi-armed bandit against arm i. But one subtle problem is that the best arm that the multi-armed bandit algorithm is competing with is not the base algorithm if it had been played for all rounds (and importantly had received feedback in all rounds), but the base algorithm under importance-weighted feedback. So in the above decomposition BASE_i cannot be replaced by standard regret bounds for this base algorithm. For many base algorithms the regret on such importance-weighted feedback can be shown to depend on the maximum importance-weight given to the base algorithm in some manner. The log-barrier algorithm they use produces a negative term that depends on that maximum weight as well. They show that in some cases, this will cancel out and still allow for useful regret bounds.

But the log-barrier is also used in Lee et al. [2020] to obtain high-probability bounds for the multi-armed bandit setting. They also use the negative term to cancel out other terms. The idea behind our version CORRAL+ is to instead use their analysis to obtain high-probability bounds for the CORRAL part in the decomposition, including a negative term with the maximum importance-weight. And we also obtain a high-probability bound for BASE_i. We do this by adjusting the analysis of Liu et al. [2024]. We obtain high-probability regret bounds for the base algorithms that depend on the maximum importance-weight and as in Agarwal et al. [2017] we show that this will cancel out with the negative term that is part of CORRAL. This gives us high-probability regret bounds that do not depend on the maximum importance-weights anymore and we can do this for base algorithms with different values of the hyperparameter jointly and thus we can tune the hyperparameter as in Agarwal et al. [2017], except now with high probability as well.

D.2 Algorithm CORRAL+

The CORRAL+ algorithm is very similar to the one in Lee et al. [2020]. The main difference is that our base algorithms need the losses to be within $[0, 1]$, so we cannot use the original importance weighting. Instead we rescale the importance-weighted losses with $\frac{1}{\rho_{t,i}}$. Every time that the value for $\rho_{t,i}$ is updated we restart the base algorithm for this new rescaling. This restarting has a slight delay of $\kappa - 1$ rounds because we want to make sure that the policy is fixed κ steps ahead to disentangle policy and misspecification level. If we restart the base algorithm immediately the policy is never truly fixed long enough. This delay does mean that some rounds will be played without the right rescaling which hurts the FTRL guarantees for those rounds. This requires careful consideration during the regret decomposition and the bounding of the terms. Furthermore, due to the manner of updating (at least a doubling) and the truncation ensuring that $p_{t+1,i} \geq \frac{2}{\sqrt{T}}$ for all i and all t, we can guarantee at most $2\log_2(T)$ restarts for the base algorithms.

D.3 High-Probability Regret Multi-Armed Bandit

In this subsection we prove a high-probability bound for the CORRAL part of the decomposition (4). We prove the following lemma:

Algorithm 2: CORRAL+

Input: δ, T, a sample of previous contexts $\mathcal{S}$, M base algorithms $\mathcal{B}_1, \ldots, \mathcal{B}_M$ and a set G of M pairs of hyperparameter values

Define $\Xi = \{\boldsymbol{p} \in \Delta_M : p_i \geq \frac{2}{\sqrt{T}} \forall i \in [M]\}$.

Initialize: $\beta = \exp\left(\frac{1}{\ln(T)}\right), \rho_{1,i} = 2M$ for all $i \in [M], \boldsymbol{p_1} = \frac{\iota}{M}$,
$\eta^* = \frac{1}{d^{5/2}\sqrt{T}+5120\ln(10dT)\sqrt{2}d^{3/2}\ln^3(T)\sqrt{T}}$, $\eta_{1,i} = \eta^*$ for all $i \in [M]$.

Initialize base algorithm i with $(\kappa, \bar{C})_i \in G$, T, $\rho_{1,i}$, $\bar{\delta} = \frac{\delta}{4\kappa_i M}$, $\mathcal{S}$ for all $i \in [M]$.

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

  Receive $\mathcal{A}_t$ and send it to $\mathcal{B}_i$ for all $i \in [M]$.

  Obtain a_t^i from $\mathcal{B}_i$ for all $i \in [M]$.

  Sample $i_t \sim \boldsymbol{p_t}$, play/return $a_t^{i_t}$.

  Receive loss ℓ_t and return $\ell_t^i = \ell_t$ and $s_t^i = \frac{\mathbb{I}\{i_t = i\}}{p_{t,i}}$ to $\mathcal{B}_i$ for all $i \in [M]$.

  Update $\boldsymbol{p_{t+1}} = \arg\min_{\boldsymbol{p} \in \Xi} \langle \boldsymbol{p}, \frac{\ell_t}{p_{t,i_t}} e_{i_t} \rangle + D_{\xi_t}(\boldsymbol{p}, \boldsymbol{p_t})$, where $\xi_t(\boldsymbol{p}) = \sum_{i=1}^M \frac{1}{\eta_{t,i}} \ln\left(\frac{1}{p_i}\right)$.

  for $i = 1, ..., M$ **do**

    if $\frac{1}{p_{t+1,i}} > \rho_{t,i}$ **then**

      $\rho_{t+1,i} = \frac{2}{p_{t+1,i}}, \eta_{t+1,i} = \beta\eta_{t,i}$ and restart $\mathcal{B}_i$ with $\rho_{t+1,i}$ after round $t + \kappa_i - 1$.

    else

      $\rho_{t+1,i} = \rho_{t,i}, \eta_{t+1,i} = \eta_{t,i}$.

Lemma 15. *Choosing arm i_t in round t according to Algorithm 2 guarantees that with a probability of at least $1 - \delta$*

$$\sum_{t=1}^{T}(\ell_{t,i_t} - \ell_{t,i}) \leq O\left(\frac{M\ln(T)}{\eta} + \eta T + M\sqrt{T}\ln\left(\frac{4TM}{\delta}\right)\right) - \frac{\rho_{T,i}}{20\eta\ln(T)} + 3\sqrt{\rho_{T,i}T\ln\left(\frac{4TM}{\delta}\right)} + 2\rho_{T,i}\ln\left(\frac{4TM}{\delta}\right)$$

for all $i \in [M]$ and where $\rho_{T,i}$ is as determined by the algorithm and bounds the maximum importance-weight given to arm i throughout the T rounds.

Proof. We can first decompose the regret against arm i as follows. Fix i^*. Let $u = (1 - \frac{2M}{\sqrt{T}})e_{i^*} + \frac{2}{\sqrt{T}}\iota$, with ι a vector of ones.

$$\sum_{t=1}^{T}(\ell_{t,i_t} - \ell_{t,i^*}) = \sum_{t=1}^{T}(\ell_{t,i_t} - \langle u, \hat{\ell}_t\rangle) + \sum_{t=1}^{T}\langle u, \hat{\ell}_t - \ell_t\rangle + \sum_{t=1}^{T}\langle u - e_{i^*}, \ell_t\rangle \tag{6}$$

$$\leq \sum_{t=1}^{T}(\ell_{t,i_t} - \langle u, \hat{\ell}_t\rangle) + \sum_{t=1}^{T}\langle u, \hat{\ell}_t - \ell_t\rangle + 2M\sqrt{T} \tag{7}$$

Using Lemma 16 we can bound the first term by

$$\sum_{t=1}^{T}(\ell_{t,i_t} - \langle u, \hat{\ell}_t\rangle) \leq O\left(\frac{M\ln(T)}{\eta} + \eta T\right) - \frac{\rho_{T,i^*}}{20\eta\ln(T)},$$

where we have used that $\langle \rho_T, u\rangle \geq \frac{1}{2}\rho_{T,i^*}$ assuming $M \leq \sqrt{T}/4$. In case this doesn't hold we would have that the term on the left could cancel out the negative term anyway as $\rho_{T,i^*} \leq \sqrt{T}$, so the expression as a whole still holds. For the second term we use the Strengthened Freedman's inequality from Zimmert and Lattimore [2022] (see Lemma 18). We set $X_t = \langle u, \hat{\ell}_t - \ell_t\rangle$. We have

$$\mathbb{E}_t\left[X_t^2\right] \leq \sum_{i=1}^{M} u_i^2 \ell_{t,i}^2 \rho_{T,i} \leq \rho_{T,i^*} + 4M/\sqrt{T}$$

using that $\rho_{T,i} \leq \sqrt{T}$ and splitting between arm i^* and all the other ones. Note that for the other ones $u_i = 2/\sqrt{T}$. So we have $\sum_{t=1}^{T} \mathbb{E}_t\left[X_t^2\right] \leq \rho_{T,i^*} T + 4M\sqrt{T}$. We also have that the maximum value of X_t can be upper bounded by ρ_{T,i^*}, using that $u_i \rho_{T,i} \leq 2$ if $i \neq i^*$. We can then fill these into the Strengthened Freedman's inequality (Lemma 18) to get

$$\sum_{t=1}^{T} \langle u, \hat{\ell}_t - \ell_t \rangle \leq 3\sqrt{(\rho_{T,i^*} T + 4M\sqrt{T}) \ln(4T/\delta)} + 2\rho_{T,i^*} \ln(4T/\delta)$$

which holds with a probability of at least $1 - \delta$. We can complete the proof by filling in this result and Lemma 16 into the final result of Equation 7. Replace δ by δ/M to get the result for all arms simultaneously. □

Lemma 16. CORRAL+ *(Algorithm 2) ensures*

$$\sum_{t=1}^{T} (\ell_{t,i_t} - \langle u, \hat{\ell}_t \rangle) \leq O\left(\frac{M \ln(T)}{\eta} + \eta T\right) - \frac{\langle \rho_T, u \rangle}{10\eta \ln(T)}$$

for any $u \in \Xi$.

See Lemma 2.1 from Lee et al. [2020].

E Misspecification-robust Linear Bandit Algorithm

For convenience, we restate the misspecification-robust adversarial linear bandit algorithm of Liu et al. [2024] (their Algorithm 6) here as Algorithm 3, and then derive in Section E.1 a high-probability regret bound for it under the importance-weighted feedback supplied by CORRAL+ (Appendix D). This bound is invoked in the proof of Lemma 7 in Appendix F.

Algorithm 3: Adversarial linear bandit robust to misspecification

Input: $\bar{C}, \kappa, \bar{T}, \bar{\rho}, \widehat{\Omega}, \bar{\delta}$.

Initialize: $\gamma = \frac{\ln(\bar{T}/\bar{\delta})}{\bar{T}}$, $\zeta = 4\ln(10d\bar{T})$, $\eta = \frac{\sqrt{\kappa}}{16\sqrt{2d}\zeta\sqrt{\bar{T}} + 32\sqrt{d}\alpha\sqrt{\kappa}}$,
$\alpha = 8\sqrt{d}\left(\frac{1}{\bar{\rho}}\left(\sqrt{\gamma}\bar{T} + 2\zeta\bar{C} + 26\zeta\sqrt{\kappa\bar{T}}\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right) + 6\zeta\sqrt{\frac{\bar{T}}{\bar{\rho}}}\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right)$.

for $t = 1, \ldots, \bar{T}$ **do**

Let $q_t(y) = \dfrac{\exp\left(-\eta\sum_{\tau<t}\langle y, s_\tau\hat{\theta}_\tau/\bar{\rho} - b_\tau\rangle\right)}{\int_{\widehat{\Omega}}\exp\left(-\eta\sum_{\tau<t}\langle z, s_\tau\hat{\theta}_\tau/\bar{\rho} - b_\tau\rangle\right)\mathrm{d}z}$, $x_t = \mathbb{E}_{y\sim q_t}[y]$ and $\Sigma_t = \mathbb{E}_{y\sim q_t}[yy^\intercal]$.

Let $\hat{q}_t(y) = \dfrac{q_t(y)\,\mathbb{I}\left[\|y\|^2_{\Sigma_t^{-1}} \leq d\zeta^2\right]}{\int_{\widehat{\Omega}} q_t(z)\,\mathbb{I}\left[\|z\|^2_{\Sigma_t^{-1}} \leq d\zeta^2\right]\mathrm{d}z}$ and $\hat{x}_t = \mathbb{E}_{y\sim\hat{q}_t}[y]$ and $\hat{\Sigma}_t = \gamma I + \mathbb{E}_{y\sim\hat{q}_t}[yy^\intercal]$.

Play $y_t \sim \hat{q}_t$, receive loss $\ell_t \in [0,1]$ and weight s_t.

Let $\hat{\theta}_t = \hat{\Sigma}_t^{-1} y_t \ell_t$ and $B_t = I + \begin{bmatrix} \hat{\Sigma}_t^{-1} & -\hat{\Sigma}_t^{-1}x_t \\ -x_t^\intercal\hat{\Sigma}_t^{-1} & x_t^\intercal\hat{\Sigma}_t^{-1}x_t \end{bmatrix}$.

if $\lambda_{\max}(B_t - \sum_{\tau\in\mathcal{I}} B_\tau) > 0$ **then**

$\mathcal{I} \leftarrow \mathcal{I} \cup \{t\}$.

$b_t = -\alpha\nabla F(x_t)$, where $\nabla F(x_t) = -\eta\sum_{\tau<t}(s_\tau\hat{\theta}_\tau/\bar{\rho} - b_\tau)$.

else

$b_t = 0$.

E.1 High-Probability Bound Linear Bandit under CORRAL+ Feedback

In this subsection we prove a high-probability regret bound for the BASE_i part of the decomposition (4). We prove the following lemma:

Lemma 17. *For the linear bandit setting with κ_i instances of Algorithm 3, with total misspecification of $C_p^m(\kappa_i)$ provided to the m'th instance ($1 \le m \le \kappa_i$), with $\bar{\delta}_i$ set to $\delta/(4\kappa_i M)$ and under the feedback provided by* CORRAL+*, we have with probability at least $1-\delta$, against any comparator $u \in \widehat{\Omega}$,*

$$\sum_{t=1}^{T} \langle y_t - u, \theta_t \rangle \le 64\sqrt{2} d^{3/2} \zeta \ln^2(T) \rho_{T,i} \sqrt{T} + O\left(\kappa d^{5/2} \mathrm{polylog}(T/\delta) \sum_{m=1}^{\kappa} C_p^m(\kappa)\right) + O\left(\kappa d^{5/2} \sqrt{\rho_{T,i} T} \mathrm{polylog}(T/\delta)\right),$$

for all $\kappa \in \mathcal{K}$ simultaneously, where for each κ the bound holds for the base algorithm i for which $\kappa_i = \kappa$ and for which the misspecification hyperparameter $\bar{C}$ is the smallest value still large enough. Furthermore, $\rho_{T,i}$ is the maximum rescaling used for this base algorithm i and M is the number of base algorithms used.

The proof of Lemma 17 occupies the remainder of this subsection. We first bound the regret of a single instance of Algorithm 3: Section E.1.1 decomposes it into a bias term, a deviation term, a bonus term and an FTRL term, which are bounded in turn in Sections E.1.2–E.1.5. Section E.1.6 then combines these bounds, sums over the κ instances and takes a union bound over the base algorithms to complete the proof.

E.1.1 Regret Decomposition

We start with a regret decomposition before bounding the individual components. The analysis is similar to Liu et al. [2024], except we are working with losses instead of rewards and we are rescaling our losses to make sure the importance-weighted feedback is still between 0 and 1. However, as mentioned before, due to the delayed restarts, we cannot guarantee this for all rounds. For every start there will be at most one round (per copy) that we cannot guarantee the right scaling for. So we need to carefully take that into account in this subsection for both the decomposition and the later bounding of the terms.

We first consider the regret of a single linear bandit algorithm (of which base algorithm i will have κ_i copies). Note that $\mathbb{E}_t[\cdot]$ is conditioned one step back from this particular copy's perspective (so a different t as well) but from the meta algorithm's perspective it would be κ steps back and a different t and thus equivalent to $\mathbb{E}_{(t-2)\kappa+j+1}[\cdot]$ for the j'th copy. And for the current round we use t from the perspective of the j'th copy but from the meta algorithm's perspective this current round would be round $(t-1)\kappa + j$. Also recall from Section 4.1 that the indexing is clamped at 0 for rounds $t < \kappa$ in terms of the epoch's local t. This means that for the first $\kappa - 1$ rounds of the epoch, this conditioning is not actually κ steps back. We deal with this problem in Section C.4 when bounding the total misspecification in the stationary mixing setting with Lemma 10.

We will start with regret bounds for a single copy and work from its perspective and only switch to the learner's perspective when useful. We will split the rounds into the (re-)starts and then take out the final round for each one which will be bounded by 1. We also introduce $\bar{\rho}_{t,i}$ which is the rescaling used for round t for base algorithm i. Note that this equals $\rho_{t,i}$ for every round that is not the last one in the (re-)start. We need to introduce it because that is what the algorithm uses but we will later convert them to $\rho_{t,i}$.

For regret comparator $u^* \in \widehat{\Omega}$, define $x^* = \arg\min_{x \in \widehat{\Omega}} F(x)$ and $u = (1 - 1/\bar{T})u^* + \frac{1}{\bar{T}}x^*$. We have with probability at least $1 - \bar{\delta}$

$$\begin{aligned}
\mathrm{Reg}_{\bar{T}} &= \sum_{t=1}^{\bar{T}} \langle y_t - u^*, \theta_t \rangle = \sum_{t=1}^{\bar{T}} \langle y_t - u, \theta_t \rangle + \frac{1}{\bar{T}} \sum_{t=1}^{\bar{T}} \langle x^* - u^*, \theta_t \rangle \\
&\leq \sum_{t=1}^{\bar{T}} \langle \hat{x}_t - u, \theta_t \rangle + O\left(\sqrt{\bar{T} \ln(\bar{T}/\bar{\delta})}\right) + 1 \\
&= \sum_{t=1}^{\bar{T}} \langle x_t - u, \theta_t \rangle + \sum_{t=1}^{\bar{T}} \langle \hat{x}_t - x_t, \theta_t \rangle + O\left(\sqrt{\bar{T} \ln(\bar{T}/\bar{\delta})}\right) + 1 \\
&\leq \sum_{t=1}^{\bar{T}} \langle x_t - u, \theta_t \rangle + O\left(\sqrt{\bar{T} \ln(\bar{T}/\bar{\delta})}\right) + 2,
\end{aligned}$$

using Azuma's inequality and Lemma 19 for the first and second inequality, respectively. Now we focus on decomposing the first term

$$\begin{aligned}
\sum_{t=1}^{\bar{T}} \langle x_t - u, \theta_t \rangle &= \sum_{h=1}^{H} \sum_{t \in \bar{\mathcal{T}}_h} \langle x_t - u, \theta_t \rangle \leq H + \sum_{h=1}^{H} \sum_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \langle x_t - u, \theta_t \rangle \\
&= H + \sum_{h=1}^{H} \sum_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} (\langle x_t - u, \theta_t - \mathbb{E}_t\left[s_t \hat{\theta}_t\right] \rangle + \langle x_t - u, \mathbb{E}_t\left[s_t \hat{\theta}_t\right] - s_t \hat{\theta}_t \rangle \\
&\qquad + \langle x_t - u, s_t \hat{\theta}_t - \bar{\rho}_{t,i} b_t \rangle + \langle x_t - u, \bar{\rho}_{t,i} b_t \rangle),
\end{aligned}$$

where we have set $s_t = \frac{\mathbb{I}\{i_t = i\}}{p_{t,i}}$, $\bar{\mathcal{T}}_h$ is the set of rounds included in the h'th (re-)start and $\bar{T}_h$ is the final round of the h'th (re-)start and H is the number of (re-)starts, which is at most of order $\log(T)$ due to the doubling. The terms that are not yet bounded are referred to as the bias, deviation, FTRL and bonus terms, respectively. We now bound them individually before combining the results.

E.1.2 Bounding Bias Term

Due to the different perspectives on the rounds we mentioned before, we have to be a bit more careful when dealing with the importance-weighted feedback. The importance-weighting and $\hat{\theta}_t$ are dependent because they both depend on which base algorithm the meta algorithm plays. This is why we switch $\hat{\theta}_t$, which is the estimator that the linear bandit algorithm uses, with the hypothetical $\tilde{\theta}_t$, which is the same estimator but with the hypothetical loss of this base algorithm. However, the importance-weighting and $\tilde{\theta}_t$ are still dependent (κ steps back) through the intermediate contexts. The importance-weighting is determined one step back from the meta algorithm's perspective, not from the current linear bandit algorithm copy we are analyzing now and so it will depend on intermediate contexts. This round's context will depend on the intermediate contexts as well and in turn impacts $\tilde{\theta}_t$. But the importance-weighting and $\tilde{\theta}_t$ are conditionally independent when we condition on $p_{t,i}$, which is what we now use to get rid of the importance-weight. We first define $\tilde{\theta}_t = \hat{\Sigma}_t^{-1} y_t \tilde{\ell}_t$, where $\tilde{\ell}_t$ is the loss associated with this algorithm's action, which is unobserved whenever this base algorithm is not selected by the meta algorithm and thus we generally cannot create $\tilde{\theta}_t$ either. But we do have that $\frac{\mathbb{I}\{i_t = i\}}{p_{t,i}} \hat{\theta}_t = \frac{\mathbb{I}\{i_t = i\}}{p_{t,i}} \tilde{\theta}_t$, which we will now use for the analysis.

$$\begin{aligned}
\mathbb{E}_t\left[s_t \hat{\theta}_t\right] &= \mathbb{E}_t\left[\frac{\mathbb{I}\{i_t = i\}}{p_{t,i}} \hat{\theta}_t\right] = \mathbb{E}_t\left[\frac{\mathbb{I}\{i_t = i\}}{p_{t,i}} \tilde{\theta}_t\right] = \mathbb{E}_t\left[\mathbb{E}_t\left[\frac{\mathbb{I}\{i_t = i\}}{p_{t,i}} \tilde{\theta}_t \mid p_{t,i}\right]\right] \\
&= \mathbb{E}_t\left[\frac{1}{p_{t,i}} \mathbb{E}_t\left[\mathbb{I}\{i_t = i\} \tilde{\theta}_t \mid p_{t,i}\right]\right] = \mathbb{E}_t\left[\frac{1}{p_{t,i}} \mathbb{E}_t\left[\mathbb{I}\{i_t = i\} \mid p_{t,i}\right] \mathbb{E}_t\left[\tilde{\theta}_t \mid p_{t,i}\right]\right] \\
&= \mathbb{E}_t\left[\mathbb{E}_t\left[\tilde{\theta}_t \mid p_{t,i}\right]\right] = \mathbb{E}_t\left[\tilde{\theta}_t\right],
\end{aligned}$$

where we could use the conditional independence of $\tilde{\theta}_t$ and $\mathbb{I}\{i_t = i\}$ because we replaced the actual observed loss with the hypothetical loss.

We have $\mathbb{E}_t\left[\tilde{\ell}_t|y_t\right] = y_t^\intercal\theta_t + \epsilon_t(y_t)$, where $\epsilon_t(y_t)$ is the misspecification when the linear bandit algorithm plays action y_t. Therefore, using the law of total expectation and conditioning on y_t, we have for the expectation of the estimator:

$$\begin{aligned}\mathbb{E}_t\left[\tilde{\theta}_t\right] &= \hat{\Sigma}_t^{-1}\,\mathbb{E}_t\left[y_t y_t^\intercal\right]\theta_t + \hat{\Sigma}_t^{-1}\,\mathbb{E}_t\left[y_t\epsilon_t(y_t)\right]\\ &= \hat{\Sigma}_t^{-1}(\hat{\Sigma}_t - \gamma I)\theta_t + \hat{\Sigma}_t^{-1}\,\mathbb{E}_t\left[y_t\epsilon_t(y_t)\right]\\ &= \theta_t + \hat{\Sigma}_t^{-1}(\mathbb{E}_t\left[y_t\epsilon_t(y_t)\right] - \gamma\theta_t),\end{aligned}$$

where we used that $\hat{\Sigma}_t = \gamma I + \mathbb{E}_t\left[y_t y_t^\intercal\right]$. So for the bias term we have:

$$\begin{aligned}&\sum_{h=1}^{H}\sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\langle x_t - u, \theta_t - \mathbb{E}_t\left[s_t\hat{\theta}_t\right]\rangle\\ &= \sum_{h=1}^{H}\left(\sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\langle x_t - u, \gamma\hat{\Sigma}_t^{-1}\theta_t\rangle + \sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\langle x_t - u, -\hat{\Sigma}_t^{-1}\,\mathbb{E}_t\left[y_t\epsilon_t(y_t)\right]\rangle\right)\\ &\le \sum_{h=1}^{H}\gamma\sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}}\|\theta_t\|_{\hat{\Sigma}_t^{-1}} + \sum_{h=1}^{H}\sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}}\,\mathbb{E}_t\left[\|y_t\|_{\hat{\Sigma}_t^{-1}}|\epsilon_t(y_t)|\right]\\ &\le \sum_{h=1}^{H}\max_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}}\left(\sqrt{\gamma d}\bar{T} + 2\sqrt{d}\zeta\sum_{t=1}^{\bar{T}}\mathbb{E}_t\left[|\epsilon_t(y_t)|\right]\right)\\ &\le \sum_{h=1}^{H}\max_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}}\left(\sqrt{\gamma d}\bar{T} + 2\sqrt{d}\zeta\sum_{t=1}^{\bar{T}}\max_{y_t}|\epsilon_t(y_t)|\right)\\ &= \sum_{h=1}^{H}\max_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}}\left(\sqrt{\gamma d}\bar{T} + 2\sqrt{d}\zeta C_\infty\right),\end{aligned}$$

where we used $\gamma I \preceq \hat{\Sigma}_t$, $\|\theta_t\|_2 \le \sqrt{d}$, and $\|y_t\|_{\hat{\Sigma}_t^{-1}} \le 2\sqrt{d}\zeta$. The last bound combines the clipping of Algorithm 3 (which gives $\|y_t\|_{\Sigma_t^{-1}} \le \sqrt{d}\zeta$) together with the PSD comparison $\hat{\Sigma}_t^{-1} \preceq \frac{4}{3}\Sigma_t^{-1}$ from Lemma 19 (giving an extra factor $2/\sqrt{3}$).

E.1.3 Bounding Deviation Term

Recall that $s_t = \frac{\mathbb{I}\{i_t = i\}}{p_{t,i}}$. Then

$$|\langle x_t - u, s_t\hat{\theta}_t\rangle| \le s_t|(x_t - u)^\intercal\hat{\Sigma}_t^{-1}y_t| \le s_t\|u - x_t\|_{\hat{\Sigma}_t^{-1}}\|y_t\|_{\hat{\Sigma}_t^{-1}} \le 2\sqrt{d}\zeta s_t\|u - x_t\|_{\hat{\Sigma}_t^{-1}}.$$

We thus have

$$\begin{aligned}\mathbb{E}_t\left[\langle x_t - u, s_t\hat{\theta}_t\rangle^2\right] &\le 4d\zeta^2\|u - x_t\|_{\hat{\Sigma}_t^{-1}}^2\,\mathbb{E}_t\left[s_t^2\right]\\ &= 4d\zeta^2\|u - x_t\|_{\hat{\Sigma}_t^{-1}}^2\,\mathbb{E}_t\left[\frac{1}{p_{t,i}}\right] \le 4d\zeta^2\|u - x_t\|_{\hat{\Sigma}_t^{-1}}^2\,\mathbb{E}_t\left[\rho_{t,i}\right],\end{aligned}$$

where we conditioned on $p_{t,i}$ and then used $\rho_{t,i} \ge \frac{1}{p_{t,i}}$. We again had to be careful as the importance-weights and thus the scaling are no longer known one step ahead from the perspective of the copy of the linear bandit algorithm. This is why we are left with an expectation of the scaling rather than the actual value.

We can then use the Strengthened Freedman's inequality (see Lemma 18) for a given set of consecutive rounds $\tilde{\mathcal{T}} \subseteq \bar{\mathcal{T}}$ so that we have with a probability of at least $1 - \bar{\delta}/\bar{T}^2$

$$\begin{aligned}
\sum_{t\in\tilde{\mathcal{T}}} \langle x_t - u, \mathbb{E}_t\left[s_t\hat{\theta}_t\right] - s_t\hat{\theta}_t\rangle &\leq 3\sqrt{\sum_{t\in\tilde{\mathcal{T}}} \mathbb{E}_t\left[\langle x_t - u, s_t\hat{\theta}_t\rangle^2\right] \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)} \\
&\quad + \left(8\sqrt{d}\zeta\rho_{\bar{T},i}\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}} + 2\right)\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \\
&\leq 3\sqrt{4d\zeta^2\sum_{t\in\tilde{\mathcal{T}}} \mathbb{E}_t\left[\rho_{t,i}\right]\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)} \\
&\quad + \left(8\sqrt{d}\zeta\sqrt{\kappa\bar{T}}\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|_{\hat{\Sigma}_t^{-1}} + 2\right)\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right),
\end{aligned} \tag{8}$$

where we have used that $\rho_{t,i} \leq \sqrt{T} = \sqrt{\kappa\bar{T}}$. We also used that $\|u - x_t\|_{\hat{\Sigma}_t^{-1}} \leq 2\sqrt{\bar{T}}$ due to how $\hat{\Sigma}_t^{-1}$ is constructed and filling in γ. This is only used for the part inside of the $\ln(\cdot)$. The problem is that we now have $\mathbb{E}_t\left[\rho_{t,i}\right]$ terms in the expression, while we need $\rho_{t,i}$ or $\rho_{T,i}$ instead. So that is why we need to use the Strengthened Freedman's inequality again to bound $\sum_{t\in\tilde{\mathcal{T}}} \mathbb{E}_t\left[\rho_{t,i}\right]\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}}$ in terms of $\sum_{t\in\tilde{\mathcal{T}}}\rho_{t,i}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}}$. We have with a probability of at least $1 - \bar{\delta}/\bar{T}^2$

$$\begin{aligned}
&\sum_{t\in\tilde{\mathcal{T}}}\left(\mathbb{E}_t\left[\rho_{t,i}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}}\right] - \rho_{t,i}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}}\right) \\
&\leq 3\sqrt{\sum_{t\in\tilde{\mathcal{T}}}\mathbb{E}_t\left[\rho_{t,i}^2\|u - x_t\|^4_{\hat{\Sigma}_t^{-1}}\right]\ln(8\kappa\bar{T}^4/\bar{\delta})} + 2\left(\max_{t\in\tilde{\mathcal{T}}}\mathbb{E}_t\left[\rho_{t,i}\right]\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} + 1\right)\ln(8\kappa\bar{T}^4/\bar{\delta}) \\
&\leq 3\sqrt{\sum_{t\in\tilde{\mathcal{T}}}\kappa\bar{T}\|u - x_t\|^4_{\hat{\Sigma}_t^{-1}}\ln(8\kappa\bar{T}^4/\bar{\delta})} + 2\left(\sqrt{\kappa\bar{T}}\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} + 1\right)\ln(8\kappa\bar{T}^4/\bar{\delta}) \\
&\leq 3\sqrt{\kappa\bar{T}^2\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|^4_{\hat{\Sigma}_t^{-1}}\ln(8\kappa\bar{T}^4/\bar{\delta})} + 2\left(\sqrt{\kappa\bar{T}}\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} + 1\right)\ln(8\kappa\bar{T}^4/\bar{\delta}) \\
&\leq \left(\left(3\sqrt{\kappa\bar{T}} + 2\sqrt{\kappa\bar{T}} + 2\right)\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} + 2\right)\ln(8\kappa\bar{T}^4/\bar{\delta}) \\
&\leq \left(9\kappa\bar{T}\max_{t\in\tilde{\mathcal{T}}}\|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} + 2\right)\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right),
\end{aligned}$$

where we have used that $\rho_{t,i} \leq \sqrt{T} = \sqrt{\kappa \bar{T}}$, which is the reason we slightly modified the truncation in Algorithm 2. Filling this into the final result of Equation 8 we get

$$
\begin{aligned}
&\sum_{t \in \tilde{\mathcal{T}}} \langle x_t - u, \mathbb{E}_t \left[s_t \hat{\theta}_t \right] - s_t \hat{\theta}_t \rangle \\
&\leq 6\sqrt{d}\zeta \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \sqrt{\sum_{t \in \tilde{\mathcal{T}}} \rho_{t,i} \|u - x_t\|^2_{\hat{\Sigma}_t^{-1}} + 9\kappa\bar{T} \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|^2_{\hat{\Sigma}_t^{-1}}} \\
&\quad + \left(8\sqrt{d}\zeta\sqrt{\kappa\bar{T}} \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} + 4\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right) \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \\
&\leq 6\sqrt{d}\zeta \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \left(\sqrt{\rho_{\tilde{T},i}\bar{T} \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|^2_{\hat{\Sigma}_t^{-1}}} + 3\sqrt{\kappa\bar{T}} \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}}\right) \\
&\quad + \left(8\sqrt{d}\zeta\sqrt{\kappa\bar{T}} \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} + 4\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right) \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \\
&\leq 26\sqrt{d}\zeta\sqrt{\kappa\bar{T}} \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} \\
&\quad + 6\sqrt{d}\zeta \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \sqrt{\bar{T}}\sqrt{\rho_{\tilde{T},i}} \max_{t \in \tilde{\mathcal{T}}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} \\
&\quad + 4\ln^2\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right),
\end{aligned}
$$

which holds with a probability of at least $1 - 2\bar{\delta}/\bar{T}^2$ because we need both high-probability events to hold. Now we will take a union bound over any possible start and end of this consecutive set of rounds $\tilde{T}$ of which there are at most $\bar{T}^2$ so that we get a probability of at least $1 - 2\bar{\delta}$ that this holds for all of those sets. This allows us to use the above result for all the (re-)starts to obtain the following with a probability of at least $1 - 2\bar{\delta}$

$$
\begin{aligned}
&\sum_{h=1}^{H} \sum_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \langle x_t - u, \mathbb{E}_t \left[s_t \hat{\theta}_t \right] - s_t \hat{\theta}_t \rangle \\
&\leq \sum_{h=1}^{H} 26\sqrt{d}\zeta\sqrt{\kappa\bar{T}} \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \max_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} \\
&\quad + \sum_{h=1}^{H} 6\sqrt{d}\zeta \ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right) \sqrt{\bar{T}}\sqrt{\bar{\rho}_{\bar{T}_h,i}} \max_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} + O\left(\text{polylog}(\bar{T}/\bar{\delta})\right),
\end{aligned}
$$

where we have used that $\rho_{\tilde{T},i} = \bar{\rho}_{\bar{T}_h,i}$ for $\tilde{\mathcal{T}} = \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}$.

E.1.4 Bounding Bonus Term

$$
\begin{aligned}
\sum_{h=1}^{H} \sum_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \langle x_t - u, \bar{\rho}_{t,i} b_t \rangle &= \sum_{h}^{H} \bar{\rho}_{\bar{T}_h,i} \sum_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \langle x_t - u, b_t \rangle \\
&\leq \sum_{h}^{H} \bar{\rho}_{\bar{T}_h,i} \left(-\frac{\alpha_h}{8} \max_{t \in \bar{\mathcal{T}}_h \setminus \{\bar{T}_h\}} \|u - x_t\|_{\hat{\Sigma}_t^{-1}} + 6\alpha_h d^2 \log_2(4\bar{T}/\gamma)\right),
\end{aligned}
$$

where α_h is the value of α for the h'th rescaling. The first equality uses that $\bar{\rho}_{t,i} = \rho_{t,i}$ for the rounds we sum over. The second inequality uses that it remains constant throughout the rounds we are summing over. For the derivation see Lemma J.9 in Liu et al. [2024].

E.1.5 Bounding FTRL Term

The FTRL term for a given (re-)start can be bounded as follows:

$$
\begin{aligned}
&\sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}} \langle x_t - u, s_t\hat{\theta}_t - \bar{\rho}_{t,i} b_t\rangle \\
&= \sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}} \bar{\rho}_{t,i}\langle x_t - u, \frac{s_t}{\bar{\rho}_{t,i}}\hat{\theta}_t - b_t\rangle \\
&= \bar{\rho}_{\bar{T}_h,i} \sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}} \langle x_t - u, \frac{s_t}{\bar{\rho}_{\bar{T}_h,i}}\hat{\theta}_t - b_t\rangle \\
&\leq \bar{\rho}_{\bar{T}_h,i}\left(\frac{d\ln(\bar{T})}{\eta_h} + 2\eta_h \sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}} \|\frac{s_t\hat{\theta}_t}{\bar{\rho}_{\bar{T}_h,i}} - b_t\|^2_{\nabla^{-2}F(x_t)}\right) \\
&\leq \bar{\rho}_{\bar{T}_h,i}\left(\frac{d\ln(\bar{T})}{\eta_h} + 4\eta_h \sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}} \left(\|\frac{s_t\hat{\theta}_t}{\rho_{\bar{T}_h,i}}\|^2_{\nabla^{-2}F(x_t)} + \|b_t\|^2_{\nabla^{-2}F(x_t)}\right)\right) \\
&\leq \frac{d\ln(\bar{T})\bar{\rho}_{\bar{T}_h,i}}{\eta_h} + 4\sum_{t\in\bar{\mathcal{T}}_h}\eta_t\rho_{t,i}\|\frac{s_t\hat{\theta}_t}{\rho_{t,i}}\|^2_{\nabla^{-2}F(x_t)} + 4\sum_{t\in\bar{\mathcal{T}}_h}\eta_t\bar{\rho}_{t,i}\|b_t\|^2_{\nabla^{-2}F(x_t)},
\end{aligned}
$$

where we have used Lemma 17 from Zimmert and Lattimore [2022] to bound the FTRL term. We know that for the rounds we consider $\bar{\rho}_{t,i} = \rho_{t,i} \geq s_t$ so as we will check later, the stability condition holds for these rounds even if it potentially doesn't hold for the final round of a (re-)start. This equality is then also used to replace $\bar{\rho}$ with ρ for the final inequality. We define η_h to be the value of the learning rate η used for the h'th rescaling and η_t is the value used for round t. We will now bound the term on the right side. We start by bounding the first part so that with a probability of at least $1-\bar{\delta}$ for all (re-)starts jointly

$$
\begin{aligned}
&\sum_{t=1}^{\bar{T}} \eta_t\rho_{t,i}\left\|\frac{s_t\hat{\theta}_t}{\rho_{t,i}}\right\|^2_{\nabla^{-2}F(x_t)} \\
&\leq \max_h \eta_h \sum_{t=1}^{\bar{T}} \rho_{t,i}\rho_{t,i}^{-1/2}\left\|\frac{s_t\hat{\theta}_t}{\rho_{t,i}^{3/4}}\right\|^2_{\nabla^{-2}F(x_t)} \\
&= \max_h \eta_h \sum_{t=1}^{\bar{T}} \sqrt{\rho_{t,i}}\left\|\frac{s_t\hat{\theta}_t}{\rho_{t,i}^{3/4}}\right\|^2_{\nabla^{-2}F(x_t)} \\
&\leq \max_h \eta_h \sqrt{\rho_{\bar{T},i}}\sum_{t=1}^{\bar{T}}\left\|\frac{s_t\hat{\theta}_t}{\rho_{t,i}^{3/4}}\right\|^2_{\nabla^{-2}F(x_t)} \\
&\leq \max_h \eta_h \sqrt{\rho_{\bar{T},i}}\Bigg(\sum_{t=1}^{\bar{T}}\mathbb{E}_t\left[\left\|\frac{s_t\hat{\theta}_t}{\rho_{t,i}^{3/4}}\right\|^2_{\nabla^{-2}F(x_t)}\right] \\
&\qquad + \left(6d\zeta^2\sqrt{\bar{T}} + 4d\zeta^2\sqrt{\rho_{\bar{T},i}}\right)\ln\left(8d\zeta^2\kappa\bar{T}/\bar{\delta}\right)\Bigg) \\
&\leq \max_h \eta_h \Bigg(2d\zeta^2\sqrt{\rho_{\bar{T},i}}\bar{T} \\
&\qquad + \left(6d\zeta^2\sqrt{\rho_{\bar{T},i}}\sqrt{\bar{T}} + 4d\zeta^2\rho_{\bar{T},i}\right)\ln\left(8d\zeta^2\kappa\bar{T}/\bar{\delta}\right)\Bigg) \\
&\leq \tilde{O}\left(\sqrt{\kappa}\sqrt{d}\zeta\sqrt{\rho_{\bar{T},i}}\sqrt{\bar{T}}\right).
\end{aligned}
$$

We use the Strengthened Freedman's inequality (Lemma 18) to bound $\sum_{t=1}^{\bar{T}} \left\| \frac{s_t \hat{\theta}_t}{\rho_{t,i}^{3/4}} \right\|^2_{\nabla^{-2} F(x_t)}$. We let $Z_t = \left\| \frac{s_t \hat{\theta}_t}{\rho_{t,i}^{3/4}} \right\|^2_{\nabla^{-2} F(x_t)}$. Then $\mathbb{E}_t[Z_t] \leq 2d\zeta^2$, $\mathbb{E}_t\left[Z_t^2\right] \leq 4d^2\zeta^4$ and $\max_t Z_t \leq 2d\zeta^2\sqrt{\rho_{\bar{T},i}}$ (see Lemma J.12 of Liu et al., 2024). Then we have $\|b_t\|^2_{\nabla^{-2}F(x_t)} \leq \alpha_h^2 d$ if $t \in \bar{T}_h$ and $\sum_{t \in \bar{T}_h} \|b_t\|^2_{\nabla^{-2}F(x_t)} \leq \alpha_h^2 d^2 \log_2(4\bar{T}/\gamma)$ (see Lemmas J.8 and J.12 of Liu et al., 2024). Where α_h is the value of α used for the rescaling h. We can then bound the second part of the stability term for all (re-)starts combined

$$
\begin{aligned}
\sum_{h=1}^{H} \sum_{t \in \bar{\mathcal{T}}_h} \eta_t \bar{\rho}_{t,i} \|b_t\|^2_{\nabla^{-2}F(x_t)} &\leq \sum_h \eta_h \bar{\rho}_{\bar{T}_h,i} \sum_{t \in \bar{\mathcal{T}}_h} \|b_t\|^2_{\nabla^{-2}F(x_t)} \\
&\leq \sum_h \eta_h \bar{\rho}_{\bar{T}_h,i} \alpha_h^2 d^2 \log_2(4\bar{T}/\gamma) \\
&\leq \sum_h d^2 \eta_h \alpha_h d^{1/2} \zeta \tilde{O}\left(\sqrt{\bar{\rho}_{\bar{T}_h,i}}\sqrt{\bar{T}} + \bar{C} + \sqrt{\kappa \bar{T}}\right) \\
&\leq \sum_h d^2 \zeta \tilde{O}\left(\sqrt{\bar{\rho}_{\bar{T}_h,i}}\sqrt{\bar{T}} + \bar{C} + \sqrt{\kappa \bar{T}}\right) \\
&\leq \tilde{O}\left(d^2 \zeta \sqrt{\rho_{\bar{T},i}}\sqrt{\bar{T}} + d^2\zeta\bar{C} + d^2\zeta\sqrt{\kappa\bar{T}}\right),
\end{aligned}
$$

where we filled in the value of α_h and also used that $\eta_h \alpha_h \leq \frac{1}{32\sqrt{d}}$. For the last inequality we used that $\rho_{t,i} \geq \bar{\rho}_{t,i}$ for all t, i. We still need to check the stability condition $\eta_t \|s_t \hat{\theta}_t \bar{\rho}_{t,i}^{-1} - b_t\|_{\nabla^{-2}F(x_t)} \leq 1/16$ (Lemma 17 of Zimmert and Lattimore, 2022). We let α_t be the value of α in round t. Then we have:

$$
\begin{aligned}
\eta_t \|s_t \hat{\theta}_t \bar{\rho}_{t,i}^{-1} - b_t\|_{\nabla^{-2}F(x_t)} &\leq \eta_t \left(\|s_t \hat{\theta}_t \bar{\rho}_{t,i}^{-1}\|_{\nabla^{-2}F(x_t)} + \|b_t\|_{\nabla^{-2}F(x_t)}\right) \\
&\leq \eta_t \left(\|\hat{\theta}_t\|_{\nabla^{-2}F(x_t)} + \|b_t\|_{\nabla^{-2}F(x_t)}\right) \\
&\leq \frac{\sqrt{\kappa}}{16\sqrt{2d}\zeta\sqrt{\bar{T}} + 32\sqrt{d}\alpha_t\sqrt{\kappa}} \left(\sqrt{2d}\zeta + \alpha_t\sqrt{d}\right) \\
&\leq \frac{\sqrt{2d}\zeta + \alpha_t\sqrt{d}}{16\left(\sqrt{2d}\zeta\sqrt{\bar{T}/\kappa} + 2\sqrt{d}\alpha_t\right)} \\
&\leq \frac{1}{16},
\end{aligned}
$$

for every round t that is not the final round of a (re-)start. The second inequality uses that $s_t/\bar{\rho}_{t,i} \leq 1$ for these rounds. Further we have assumed that $\bar{T} \geq \kappa$. Since the base algorithm runs κ instances over a total horizon T, we have $\bar{T} = T/\kappa$, so this assumption becomes $\kappa \leq \sqrt{T}$. This assumption is reasonable because otherwise we would have more instances than rounds per instance. Now we will bound the first part of the FTRL term.

$$
\begin{aligned}
\sum_h \frac{\bar{\rho}_{\bar{T}_h,i}}{\eta_h} &\leq \sum_h \kappa^{-1/2}\left(16\sqrt{2d}\zeta\sqrt{\bar{T}}\bar{\rho}_{\bar{T}_h,i} + 32\sqrt{d}\kappa^{1/2}\alpha_h\bar{\rho}_{\bar{T}_h,i}\right) \\
&= \sum_h \left(16\sqrt{2d}\kappa^{-1/2}\zeta\sqrt{\bar{T}}\bar{\rho}_{\bar{T}_h,i} + 32\sqrt{d}\alpha_h\bar{\rho}_{\bar{T}_h,i}\right) \\
&\leq 64\ln\left(\bar{T}\right)\sqrt{2d}\kappa^{-1/2}\zeta\sqrt{\bar{T}}\bar{\rho}_{\bar{T},i} + \tilde{O}\left(d\zeta\sqrt{\kappa\bar{T}} + d\zeta\bar{C} + d\zeta\sqrt{\bar{\rho}_{\bar{T}_h,i}}\sqrt{\bar{T}}\right) \\
&\leq 64\ln\left(\bar{T}\right)\sqrt{2d}\kappa^{-1/2}\zeta\sqrt{\bar{T}}\rho_{\bar{T},i} + \tilde{O}\left(d\zeta\sqrt{\kappa\bar{T}} + d\zeta\bar{C} + d\zeta\sqrt{\rho_{\bar{T}_h,i}}\sqrt{\bar{T}}\right),
\end{aligned}
$$

where we first filled in η_h, cancelled out the κ on the right side and finally filled in α_h and used that the number of rescalings is at most $4\ln(\bar{T})$ for $\bar{T} \geq 2$ and $\rho_{\bar{T},i} \geq \bar{\rho}_{\bar{T},i} \geq \bar{\rho}_{\bar{T}_h,i}$. We thus have for the

full FTRL term with a probability of at least $1-\bar{\delta}$

$$\sum_{h=1}^{H}\sum_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\langle x_t-u, s_t\hat{\theta}_t-\rho_{t,i}b_t\rangle \le 64\ln^2\left(\bar{T}\right)\sqrt{2}d^{3/2}\kappa^{-1/2}\zeta\sqrt{\bar{T}}\rho_{\bar{T},i} + O\left(d^2\zeta\left(\sqrt{\kappa\bar{T}\rho_{\bar{T},i}}+\bar{C}\right)\mathrm{polylog}(\bar{T}/\bar{\delta})\right)$$

E.1.6 Combining Terms

When we fill in the bias, deviation, bonus and FTRL terms into the regret decomposition we obtain the following with a probability of at least $1-4\bar{\delta}$:

$$\begin{aligned}
&\sum_{t=1}^{\bar{T}}\langle y_t-u^*,\theta_t\rangle\\
&\le \sum_h \max_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u-x_t\|_{\hat{\Sigma}_t^{-1}}\sqrt{d}\left(\sqrt{\gamma}\bar{T}+2\zeta C_\infty+26\zeta\sqrt{\kappa\bar{T}}\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right)\\
&\quad+\sum_h \max_{t\in\bar{\mathcal{T}}_h\setminus\{\bar{T}_h\}}\|u-x_t\|_{\hat{\Sigma}_t^{-1}}\left(6\sqrt{d}\zeta\sqrt{\bar{T}\bar{\rho}_{\bar{T}_h,i}}\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)-\frac{\bar{\rho}_{\bar{T}_h,i}\alpha_h}{8}\right)\\
&\quad+64\ln^2\left(\bar{T}\right)\sqrt{2}d^{3/2}\kappa^{-1/2}\zeta\sqrt{\bar{T}}\rho_{\bar{T},i}+O\left(d^2\zeta\left(\sqrt{\kappa\bar{T}\rho_{\bar{T},i}}+\bar{C}\right)\mathrm{polylog}(\bar{T}/\bar{\delta})\right)\\
&\quad+O\left(\sqrt{\bar{T}}\mathrm{polylog}(\bar{T}/\bar{\delta})\right)+6\sum_h\bar{\rho}_{\bar{T}_h,i}\alpha_h d^2\log_2(4\bar{T}/\gamma).
\end{aligned}$$

We then set

$$\alpha_h = 8\sqrt{d}\left(\frac{1}{\bar{\rho}_{\bar{T}_h,i}}\left(\sqrt{\gamma}\bar{T}+2\zeta\bar{C}+26\zeta\sqrt{\kappa\bar{T}}\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right)+6\zeta\sqrt{\frac{\bar{T}}{\bar{\rho}_{\bar{T}_h,i}}}\ln\left(16\sqrt{d}\zeta\kappa\bar{T}^4/\bar{\delta}\right)\right)$$

so that the term with $\max_t\|u-x_t\|_{\hat{\Sigma}_t^{-1}}$ gets cancelled out due to the bonuses if $\bar{C}\ge C_\infty$. Filling in α_h we get

$$\sum_{t=1}^{\bar{T}}\langle y_t-u^*,\theta_t\rangle \le \frac{64\ln^2\left(\bar{T}\right)\sqrt{2}d^{3/2}\kappa^{-1/2}\zeta\sqrt{\bar{T}}\rho_{\bar{T},i}}{\sqrt{\kappa}} + O\left(\left(d^{5/2}\zeta\sqrt{\kappa\rho_{\bar{T},i}\bar{T}}+d^{5/2}\bar{C}\right)\mathrm{polylog}\left(\bar{T}/\bar{\delta}\right)\right),$$

where we have used $\rho_{\bar{T},i}\ge\bar{\rho}_{\bar{T}_h,i}$ for any h. This is the regret of one instance of a linear bandit but we have κ instances. Substituting $\bar{T}=T/\kappa$ and summing over the κ instances gives the total regret. Note that $\bar{C}_m=\bar{C}$ for all instances $1\le m\le\kappa$. A union bound over the κ instances yields, with probability at least $1-4\kappa\bar{\delta}$,

$$\sum_{t=1}^{T}\langle y_t-u^*,\theta_t\rangle \le 64\ln^2\left(T\right)\sqrt{2}d^{3/2}\zeta\sqrt{T}\rho_{T,i} + O\left(\left(d^{5/2}\kappa\sqrt{\rho_{T,i}T}+d^{5/2}\kappa\bar{C}\right)\mathrm{polylog}\left(T/\bar{\delta}\right)\right)$$

By construction of the grid, for every κ there is a $\bar{C}\in\mathcal{C}$ that upper-bounds the misspecification of all κ instances and exceeds the largest among them by at most a factor of two (because of the doubling discretization) unless that value is below 1/2. Hence $\bar{C}\le 1+2\max_{m\in[\kappa]}C_p^m(\kappa)\le 1+2\sum_{m=1}^{\kappa}C_p^m(\kappa)$, and we substitute this bound for $\bar{C}$ in the expression above. We require the result to hold for every base algorithm in CORRAL+ (with an accurate total misspecification bound), of which there are at most M, so a union bound yields probability at least $1-\sum_{i=1}^{M}4\kappa_i\bar{\delta}_i$. Since CORRAL+ initializes the base algorithm i with $\bar{\delta}_i=\frac{\delta}{4\kappa_i M}$, the lemma holds with probability at least $1-\delta$. This completes the proof of Lemma 17.

F Proof of Lemma 7

The proof has two main ingredients. First, we set up a discretized grid of hyperparameter values $(\kappa, \bar{C})$ on which CORRAL+ runs one base algorithm per grid point (below). Second, in Section F.1 we use the regret decomposition $\mathrm{CORRAL} + \mathrm{BASE}_i$ of (4), controlling CORRAL by the generic Lemma 15 from Appendix D and BASE_i by Lemma 17 from Appendix E (combined with an adaptation of the linear bandit reduction of Van Erven et al. [2025] to time-varying context distributions). The two bounds are then combined, and the CORRAL+ learning rate is tuned to cancel the dependence on the unknown importance-weight scaling.

Hyperparameter grid. For a given epoch p, let $\mathcal{K}$ be a set of values for κ, the parameter passed to each base algorithm that controls the number of internal instances of Algorithm 3. Since the misspecification is at most T_p, we discretize the set of values for $\bar{C}$ as $\mathcal{C} = \{1, 2, 4, \ldots, T_p\}$, which has size $O(\log T_p)$. The grid consists of all pairs $(\kappa_i, \bar{C}_i) \in \mathcal{K} \times \mathcal{C}$, with $1 \le i \le M_p$, and we instantiate one base algorithm per pair, so $M_p = O(\log(T_p)\,|\mathcal{K}|)$. If $\mathcal{K}$ is also discretized by doubling, then $M_p = O(\log^2(T_p))$, since values $\kappa \ge T_p$ would mean more base instances than rounds. The stability analysis in Section E.1.5 additionally requires $\kappa \le \sqrt{T_p}$, so the largest value of κ that we include in $\mathcal{K}$ is the largest power of two that is at most $\sqrt{T_p}$.

F.1 Combining CORRAL+ and Base Algorithm Bounds

In this subsection we combine Lemma 15 (controlling CORRAL) with Lemma 17 (controlling BASE_i) and tune the CORRAL+ learning rate to cancel the dependence on the unknown importance-weight scaling, completing the proof of Lemma 7:

Lemma 7. *Fix any doubling epoch p of length T_p. Algorithm 4 guarantees that, with probability at least $1 - \frac{2}{3P}\delta$,*

$$\mathrm{Reg}_{T_p}(\pi_{\phi^*}) \le O\left(\kappa^2 d^{5/2}\sqrt{T_p}\,\mathrm{polylog}(T_p/\delta) + \kappa d^{5/2}\,\mathrm{polylog}(T_p/\delta)\sum_{m=1}^{\kappa} C_p^m(\kappa)\right)$$

for all $\kappa \in \mathcal{K}$ simultaneously, where $C_p^m(\kappa)$ is as defined in (3).

Proof. The proof analyzes the algorithm within a single doubling epoch p, during which $\widehat{\Psi}$ and $\widehat{\Omega}$ are held fixed and CORRAL+ is run on a fresh instance. We index the rounds of the epoch as $t = 1, \ldots, T_p$ for notational simplicity; the global $O(\log T)$ union bound across epochs is applied at the invocation sites in Section 4.2. We use the decomposition (4), applied per-epoch with T replaced by T_p, to split the regret into $\mathrm{CORRAL} + \mathrm{BASE}_i$ for an arbitrary $i \in [M]$, and bound each part separately. Recall that BASE_i measures the regret of base algorithm i on the *real* losses under importance-weighted feedback, not the regret on the importance-weighted losses themselves as in Agarwal et al. [2017].

Recall from Section 4.1 that $\Psi_t^\kappa(\pi) = \mathbb{E}_{t-\kappa+1}\left[\pi(\mathcal{A}_t)\right]$ denotes the conditional mean action played by policy π conditioned on rounds up to and including $t - \kappa$ and the index is clamped at 0 for $t \le \kappa - 1$ so that for the first $\kappa - 1$ rounds of the epoch the conditioning is not κ steps back. The total misspecification bound takes this into account. We then decompose the regret of any base algorithm $\mathcal{B}_i$, which plays the contextual linear bandit policy π_t^i in round t, against the best fixed linear policy

π_{ϕ^*} as follows: with probability at least $1 - \frac{1}{2P}\delta$,

$$\begin{aligned}
\sum_{t=1}^{T_p}(\ell_t^i - \ell_t(\pi_{\phi^*})) &= \sum_{t=1}^{T_p}(\ell_t^i - \langle \Psi_t^{\kappa_i}(\pi_t^i), \theta_t\rangle) + \sum_{t=1}^{T_p}(\langle \Psi_t^{\kappa_i}(\pi_t^i), \theta_t\rangle - \langle \widehat{\Psi}(\pi_t^i), \theta_t\rangle) \\
&+ \sum_{t=1}^{T_p}(\langle \widehat{\Psi}(\pi_t^i), \theta_t\rangle - \langle y_t, \theta_t\rangle) + \sum_{t=1}^{T_p}(\langle y_t, \theta_t\rangle - \langle \widehat{\Psi}(\pi_{\phi^*}), \theta_t\rangle) \\
&+ \sum_{t=1}^{T_p}(\langle \widehat{\Psi}(\pi_{\phi^*}), \theta_t\rangle - \langle \Psi_t^{\kappa_i}(\pi_{\phi^*}), \theta_t\rangle) + \sum_{t=1}^{T_p}(\langle \Psi_t^{\kappa_i}(\pi_{\phi^*}), \theta_t\rangle - \ell_t(\pi_{\phi^*})) \\
&\leq O\left(\sqrt{\kappa_i T_p \ln(\kappa_i M_p/\delta)}\right) + 2\sum_{m=1}^{\kappa_i} C_p^m(\kappa_i) + 64\ln^2\left(\bar{T}\right)\sqrt{2}d^{3/2}\zeta\sqrt{T}\rho_{T,i} \\
&+ O\left(\kappa_i d^{5/2}\mathrm{polylog}(PT_p/\delta)\sum_{m=1}^{\kappa_i} C_p^m(\kappa_i)\right) \\
&\quad + O\left(\kappa_i d^{5/2}\sqrt{\rho_{T_p,i}T_p}\mathrm{polylog}(PT_p/\delta)\right),
\end{aligned}$$

where we used the definition of $C_p^m(\kappa)$ in (3) for the second and fifth terms; the first, third and sixth are bounded using Azuma's inequality (with a union bound over κ interleaved martingales due to the κ steps back conditioning for the first and sixth), and the fourth is bounded by Lemma 17. We want these high-probability inequalities to hold for all base algorithms (with correct misspecification hyperparameter), of which there are at most M_p. These Azuma inequalities are each invoked with $\delta/(8P\kappa_i M_p)$ in place of δ and Lemma 17 with $\delta/(8P)$ so the combined statement holds with probability at least $1 - \frac{1}{2P}\delta$ for all base algorithms jointly, where P is the number of epochs. Algorithm 4 indeed divides δ by an upper bound on the number of epochs $1 + \log_2(T)$ multiplied by 8.

Combining the bounds on CORRAL and BASE_i via (4), we get with probability at least $1 - \frac{2}{3P}\delta$

$$\begin{aligned}
\sum_{t=1}^{T_p}(\ell_t(a_t) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))) &\leq O\left(\frac{M\ln(T_p)}{\eta^*} + \eta^* T_p + M\sqrt{T_p}\ln\left(\frac{24PT_pM}{\delta}\right)\right) - \frac{\rho_{T_p,i}}{20\eta^*\ln(T_p)} \\
&+ 3\sqrt{\rho_{T_p,i}T_p\ln\left(\frac{24PT_pM}{\delta}\right)} + 2\rho_{T_p,i}\ln\left(\frac{24PT_pM}{\delta}\right) \\
&+ 64\sqrt{2}d^{3/2}\zeta_i\ln^2(T_p)\rho_{T_p,i}\sqrt{T_p} + O\left(\kappa_i d^{5/2}\mathrm{polylog}(PT_p/\delta)\sum_{m=1}^{\kappa_i} C_p^m(\kappa_i)\right) \\
&+ O\left(\kappa_i d^{5/2}\sqrt{\rho_{T_p,i}T_p}\mathrm{polylog}(PT_p/\delta)\right),
\end{aligned}$$

where we have filled in $\delta/(6P)$ in Lemma 15 (which already holds for all arms) and used the union bound to get $1 - \frac{1}{2P}\delta - \delta/(6P) = 1 - \frac{2}{3P}\delta$. P is absorbed into the $\mathrm{polylog}(T/\delta)$ factor). We first define $\tilde{\eta} = \frac{1}{d^{5/2}\sqrt{T_p}}$ and then we pick

$$\eta^* = \frac{1}{d^{5/2}\sqrt{T_p} + 5120\ln(10dT_p)\sqrt{2}d^{3/2}\ln^3(T_p)\sqrt{T_p}}$$

as the learning rate for CORRAL+, so that we get with a probability of at least $1 - \frac{2}{3P}\delta$

$$\begin{aligned}
\sum_{t=1}^{T_p}(\ell_t(a_t) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))) &\leq O\left(d^{5/2}\text{polylog}(T_p/\delta)\sqrt{T_p}\right) - \frac{\rho_{T_p,i}}{20\eta^*\ln(T_p)} \\
&\quad + O\left(\kappa_i d^{5/2}\text{polylog}(T_p/\delta)\sqrt{T_p}\right)\sqrt{\rho_{T_p,i}} + 64\sqrt{2}d^{3/2}\zeta_i\ln^3(T_p)\rho_{T_p,i}\sqrt{T_p} \\
&\quad + O\left(\kappa_i d^{5/2}\text{polylog}(T_p/\delta)\sum_{m=1}^{\kappa_i} C_p^m(\kappa_i)\right) \\
&\leq O\left(\kappa_i d^{5/2}\text{polylog}(T_p/\delta)\sqrt{T_p}\right)\sqrt{\rho_{T_p,i}} - 64\sqrt{2}d^{3/2}\zeta_i\ln^3(T_p)\rho_{T_p,i}\sqrt{T_p} \\
&\quad - \frac{\rho_{T_p,i}}{20\tilde{\eta}\ln(T_p)} \\
&\quad + 64\sqrt{2}d^{3/2}\zeta_i\ln^3(T_p)\rho_{T_p,i}\sqrt{T_p} + O\left(\kappa_i d^{5/2}\text{polylog}(T_p/\delta)\sum_{m=1}^{\kappa_i} C_p^m(\kappa_i)\right) \\
&\leq O\left(\kappa_i d^{5/2}\text{polylog}(T_p/\delta)\sqrt{T_p}\right)\sqrt{\rho_{T_p,i}} - \frac{\rho_{T_p,i}}{20\tilde{\eta}\ln(T_p)} \\
&\quad + O\left(\kappa_i d^{5/2}\text{polylog}(T_p/\delta)\sum_{m=1}^{\kappa_i} C_p^m(\kappa_i)\right),
\end{aligned}$$

where we have suppressed the dependence on M_p, the number of base algorithms used in CORRAL+ for epoch p, because this is of polylog(T) order at most. We also used that $\zeta_i \leq 4\ln(10dT_p)$ for all $i \in [M]$. More importantly, we have split the negative term in two parts by filling in η^*. One part to cancel out the $\rho_{T_p,i}$ term and one part that looks like the original but with $\tilde{\eta}$ instead. We then use a result from Agarwal et al. [2017] (see the proof of their Theorem 7) that says $\rho_{T_p,i}^{1/2}R - \frac{\rho_{T_p,i}}{40\ln(T_p)\tilde{\eta}} \leq \tilde{O}(\tilde{\eta}R^2)$. Which means we need to square the terms with $\sqrt{\rho_{T_p,i}}$ in them and then remove the $\sqrt{\rho_{T_p,i}}$ and multiply them with $\tilde{\eta}$. This gives us (still with a probability of at least $1 - \frac{2}{3P}\delta$)

$$\sum_{t=1}^{T_p}(\ell_t(a_t) - \ell_t(\pi_{\phi^*}(\mathcal{A}_t))) \leq O\left(\kappa^2 d^{5/2}\sqrt{T_p}\text{polylog}(T_p/\delta) + \kappa d^{5/2}\text{polylog}(T_p/\delta)\sum_{m=1}^{\kappa} C_p^m(\kappa)\right),$$

where we have suppressed the i indexing again. Note that this regret bound holds for all base algorithms with an accurate total misspecification parameter (for which every value of κ has one due to how we have constructed the set of base algorithms) and thus it holds for all $\kappa \in \mathcal{K}$. □

G Main Algorithm

In this section, we give the pseudocode of the main algorithm of this work, Algorithm 4. The algorithm uses a doubling trick to obtain a sample of previous contexts. After every epoch the contexts from that epoch are added to the current sample of contexts to create the sample for the next epoch. In every epoch the meta algorithm CORRAL+ (Algorithm 2) is restarted and supplied with a new grid of hyperparameters κ and $\bar{C}$. This grid depends on the size of the current epoch and that is why we create a new one each epoch. For every combination of hyperparameters in the grid, we create an instance of Algorithm 1 as the base algorithms and provide these to this epoch's version of CORRAL+. We then play according to this meta algorithm for every round of this current epoch. At the start, we play arbitrary actions as we do not have enough contexts yet. We play for two rounds to make sure that the sample size is a power of 2. We also play arbitrary actions if the epoch size is below $16M_p^2$. This ensures that the $M_p \leq \sqrt{T^*}/4$ condition holds. The regret bound already absorbs this. The pseudocode for this full algorithm and a discussion of its computational complexity is written below:

G.1 Computational efficiency

The contextual to linear bandit reduction is exactly the reduction of Van Erven et al. [2025], including the construction and use of the empirical action set $\widehat{\Omega}$. Section 4 of Van Erven et al. [2025] gives

Algorithm 4: Full Algorithm

Input: δ, T

for $t = 1, 2$ **do**
  Observe $\mathcal{A}_t$.
  Play arbitrary action $a_t \in \mathcal{A}_t$.
Create $\mathcal{S} = (\mathcal{A}_1, \mathcal{A}_2)$.
for $p = 2, \ldots, \lceil \log_2 T \rceil$ **do**
  Let $T^* = 2^{p-1}, \delta^* = \frac{\delta}{8(1+\log_2(T))}$.
  Create $\mathcal{K} = \{1, 2, 4, \ldots, 2^{\lfloor \log_2(\sqrt{T^*}) \rfloor}\}$.
  Create $\mathcal{C} = \{1, 2, 4, \ldots, T^*\}$.
  Create $G = \{(\kappa, \bar{C}) : \kappa \in \mathcal{K}, \bar{C} \in \mathcal{C}\}$.
  Let $M = |G|$.
  Create $\mathcal{B}_i$ as an instance of Algorithm 1 for all $i \in [M]$.
  Start Algorithm 2 with $T^*, \delta^*, \mathcal{S}, \mathcal{B}_1, \ldots, \mathcal{B}_M$ and G .
  for $t = 2^{p-1} + 1, 2^{p-1} + 2, \ldots, \min(2^p, T)$ **do**
    Observe $\mathcal{A}_t$ and send it to Algorithm 2.
    if $T^* \leq 16M^2$ **then**
      Play arbitrary action $a_t \in \mathcal{A}_t$.
    else
      Play a_t according to Algorithm 2.
  $\mathcal{S} = (\mathcal{A}_1, \ldots, \mathcal{A}_{\min(2^p, T)})$.

polynomial-time implementations of the operations appearing in the reduction under the assumption that $\mathrm{conv}(A)$ can be represented using at most $C \leq K + 1$ linear constraints. In particular, $\widehat{\Omega}$ does not have to be explicitly computed. Van Erven et al. [2025] constructs a separation oracle for it in $\mathrm{poly}(d, C, N)$ time. Further, the decomposition of a point $y_t \in \widehat{\Omega}$ can be implemented in polynomial time. Finally, an interior point of the relevant normal cone can be obtained by solving a linear program with a polynomial number of variables and constraints. Therefore, building and using the empirical action set, computing the convex decomposition, and recovering the contextual policy have the same polynomial-time implementation and assumptions as in Van Erven et al. [2025]. Our high-probability analysis does not alter these computational steps.

The computation of Corral+ (Algorithm 2) is likewise the same type of M-dimensional log-barrier update as in the original Corral algorithm of Agarwal et al. [2017]. The truncation of the simplex instead of the mixing is useful for the high-probability analysis, but it does not introduce a qualitatively harder optimization problem. Step 9 of Algorithm 2, computing p_{t+1} can be done as follows. As $D_{\xi_t}(\boldsymbol{p}, \boldsymbol{p_t})$ is the Bregman divergence, where $\xi_t(\boldsymbol{p}) = \sum_{i=1}^M \frac{1}{\eta_{t,i}} \ln\left(\frac{1}{p_i}\right)$, $D_{\xi_t}(\boldsymbol{p}, \boldsymbol{p_t}) = \sum_{i=1}^M \frac{1}{\eta_{t,i}}\left(\frac{p_i}{p_{t,i}} - 1 - \ln\left(\frac{p_i}{p_{t,i}}\right)\right)$. Then we can write a Lagrangian $L(q, \lambda, \mu) = \langle q, \ell \rangle + D_{\xi_t}(q, p) + \lambda(\sum_{i=1}^M q_i - 1) + \sum_{i=1}^M \mu_i(\frac{2}{\sqrt{T}} - q_i)$. Stationarity gives $\ell_i + \frac{1}{\eta_{t,i}}\left(\frac{1}{p_i} - \frac{1}{q_i}\right) + \lambda - \mu_i = 0$. By complementary slackness, $q_i^+(\lambda) = \max\left\{\frac{2}{\sqrt{T}}, \frac{1}{1/p_i + \eta_{t,i}(\ell_i + \lambda)}\right\}$. The function $G(\lambda) = \sum_{i=1}^M \max\left\{\frac{2}{\sqrt{T}}, \frac{1}{1/p_i + \eta_{t,i}(\ell_i + \lambda)}\right\}$ is monotonically decreasing in λ, so λ can be found by line search.

For the adversarial linear bandit algorithm (Algorithm 3), the subroutine used in our paper is specifically Algorithm 6 of Liu et al. [2024], which is introduced there as the polynomial time computed alternative to their computationally inefficient fixed-point algorithm. For a general convex action set these quantities cannot in general be computed exactly. Following Ito et al. [2020], sampling y_t can be approximated using polynomial-time log-concave sampling.

H Miscellaneous Lemmas

Strengthened Freedman's inequality from Zimmert and Lattimore [2022]:

Lemma 18. *Let $X_1, X_2, ...$ be a martingale difference sequence with respect to a filtration $F_1 \subseteq F_2 \subseteq ...$ such that $\mathbb{E}[X_t|F_t] = 0$ and assume $\mathbb{E}[|X_t||F_t] < \infty$ a.s. Then with probability at least $1-\delta$*

$$\sum_{t=1}^{T} X_t \leq 3\sqrt{V_T \ln\left(\frac{2\max\{\sqrt{V_T}, U_T\}}{\delta}\right)} + 2U_T \ln\left(\frac{2\max\{\sqrt{V_T}, U_T\}}{\delta}\right),$$

where $V_T = \sum_{t=1}^{T} \mathbb{E}_t\left[X_t^2\right]$ and $U_T = \max\{1, \max_{1\leq t\leq T} X_t\}$.

Lemma 19. *With the choice of $\zeta = 4\ln(10dT)$, we have*

$$|\,\mathbb{E}_{y\sim q_t}[f(y)] - \mathbb{E}_{y\sim \hat{q}_t}[f(y)]\,| \leq 10d\exp(-\zeta) \leq \frac{1}{2T^2}$$

for any $f : \widehat{\Omega} \to [-1, 1]$ and

$$\frac{3}{4}\,\mathbb{E}_{y\sim q_t}[yy^\intercal] \preceq \mathbb{E}_{y\sim \hat{q}_t}[yy^\intercal] \preceq \frac{4}{3}\,\mathbb{E}_{y\sim q_t}[yy^\intercal],$$

with q_t and $\hat{q}_t$ as in Algorithm 3.

See Lemma J.2 in Liu et al. [2024].

I Impossibility Results

I.1 Nonlinear Impossibility Result

Theorem 20. *If the adversary can pick both the contexts and the loss vectors, $o(T)$ expected regret is impossible in the contextual linear bandit setting.*

Proof. Let $K = 2$ and $d = 2$. In each round t let $\mathcal{A}_t = \{a_{t,1}, a_{t,2}\}$ such that $a_{t,1} = [x_t, 0]^\intercal$ and $a_{t,2} = [0, x_t]^\intercal$, where x_t is between 0.5 and 1 and unique in every round. For every round t, the adversary generates θ_t as either $[0, 1]^\intercal$ or $[1, 0]^\intercal$, both with equal probability and independent of everything else. This can be done beforehand just as the contexts can be generated beforehand. The fixed policy we use as comparator is the following

$$\pi^*(\mathcal{A}) = \begin{cases} a_1 & \text{if } \sum_{t=1}^{T} \mathbb{I}\{|a_1| = x_t\}\mathbb{I}\{\theta_{t,1} = 0\} = 1 \\ a_2 & \text{else} \end{cases}$$

for a context $\mathcal{A} = \{a_1, a_2\}$. This policy will always play $a_{t,1} = [x_t, 0]^\intercal$ whenever $\theta_t = [0, 1]^\intercal$ and $a_{t,2} = [0, x_t]^\intercal$ whenever $\theta_t = [1, 0]^\intercal$. This means the losses will equal 0 in each round t. This policy is able to perfectly pick the best action in each round despite being fixed. This is because of the uniqueness of x_t and the fact that it knows both x_t and θ_t before play starts. The adversary picking the contexts before hand removes the randomness and that was essential to making things harder for the comparator. The cumulative loss of the comparator is 0, regardless of the outcome of the randomization of the θ_t's done at the start as the best fixed policy is determined after that process.

Now for the learner, regardless of its policy in each round, the expected loss will be at least $0.5 \times 0.5 = 0.25$ per round. So the expected regret will be at least $T/4$ (expectation over the randomness of the θ_t's at the start, randomness of the learner's policy itself does not matter) over T rounds. If the learner's algorithm has linear regret in expectation (over random θ_t's) then there must exist a sequence of θ_t's for which the algorithm has linear regret as well. □

I.2 Linear Impossibility Result

Theorem 21. *If the adversary can pick both the contexts and the loss vectors, $o(T)$ expected regret against the best fixed linear policy is impossible.*

The idea behind the proof is similar to the one behind Theorem 16 in Syrgkanis et al. [2016], except we convert the setting to the contextual linear bandit setting and we consider an oblivious rather than an adaptive adversary.

Algorithm 5: Adversary Algorithm

Set $x_{\min} = 1$.
Set $x_{\max} = 0.5$.
for $t = 1, 2, ..., T$ **do**
  Set $x_t = \frac{x_{\min} + x_{\max}}{2}$.
  Set $\mathcal{A}_t = \{ \begin{bmatrix} 1/\sqrt{2} \\ x_t/\sqrt{2} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1/\sqrt{2} \\ x_t/\sqrt{2} \end{bmatrix} \}$.
  Generate $\theta_t = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ or $\theta_t = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ with equal probability and independently.
  if $\theta_t = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ **then**
    $x_{\max} = x_t$.
  else
    $x_{\min} = x_t$.

Proof. The (oblivious) adversary generates all of the contexts and loss vectors beforehand according to Algorithm 5.

Let $x^*_{\min}$ and $x^*_{\max}$ be the final values of $x_{\min}$ and $x_{\max}$ respectively, and then let $x^* = \frac{x^*_{\min} + x^*_{\max}}{2}$. We have that $x^*_{\min} < x^*_{\max}$ and for any round t either $x_t \leq x^*_{\min}$ or $x_t \geq x^*_{\max}$. We compare against the linear policy $\pi_{\tilde{\phi}}$ with $\tilde{\phi} = \begin{bmatrix} -1 & \frac{1}{x^*} & 0 & 0 \end{bmatrix}^\mathsf{T}$.

For any round t for which $x_t \leq x^*_{\min} < x^*$, we would have that $\begin{bmatrix} 1/\sqrt{2} & x_t/\sqrt{2} & 0 & 0 \end{bmatrix} \tilde{\phi} < 0$ and $\begin{bmatrix} 0 & 0 & 1/\sqrt{2} & x_t/\sqrt{2} \end{bmatrix} \tilde{\phi} = 0$. Because of how linear policies work this means that we have $\pi_{\tilde{\phi}}(\mathcal{A}_t) = \begin{bmatrix} 1/\sqrt{2} & x_t/\sqrt{2} & 0 & 0 \end{bmatrix}^\mathsf{T}$. But $x_t \leq x^*_{\min}$ means that we must have had $\theta_t = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^\mathsf{T}$. Otherwise $x_{\max}$ would have been set to x_t but $x_{\max}$ can only decrease or remain the same so we would have had $x^*_{\max} \leq x_t$ but this would contradict $x_t \leq x^*_{\min}$ since $x^*_{\min} < x^*_{\max}$. So we are guaranteed to have $\theta_t = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^\mathsf{T}$ which means that $\langle \pi_{\tilde{\phi}}(\mathcal{A}_t), \theta_t \rangle = 0$. Similarly, we can show that for rounds with $x_t \geq x^*_{\max}$ this linear policy would also get a loss of 0, so that it will have a cumulative loss of 0.

In every round there is one action with a loss of 0 and one action with a loss of $x_t/\sqrt{2}$ which is at least $\frac{1}{2\sqrt{2}}$. But due to the completely random nature of the θ_t's, the expected loss (over the θ_t generation) is at least $\frac{1}{2} \times \frac{1}{2\sqrt{2}} = \frac{1}{4\sqrt{2}}$ per round for both actions. Therefore, regardless of the policy of the learner, the expected loss per round would be at least $\frac{1}{4\sqrt{2}}$. The comparator manages to avoid this $\frac{1}{4\sqrt{2}}$ expected loss because it is determined with the knowledge of θ_t in each round, whereas the learner has no way of knowing/learning the value because it is independent of anything that is known to the learner at that point in time. This means an expected regret of at least $\frac{T}{4\sqrt{2}}$ over the T rounds. And since this holds in expectation, regardless of what the learner's algorithm does, for any algorithm there must exist a sequence of θ_t's, such that the regret (expectation only over learner's randomization) is at least $\frac{T}{4\sqrt{2}}$. Thus sublinear regret is impossible in the fully adversarial setting even when restricted against linear comparators. □

NeurIPS Paper Checklist

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction state the three main contributions (regret bounds for non-stationary independent contexts, stationary mixing contexts, and high-probability bounds in the IID setting), and these match the formal theorems proved in the paper.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses the assumptions of each setting (linear realizability, boundedness, mixing assumptions, knowledge of the reference distribution or variation budget) and the resulting scaling of the regret bounds.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are stated explicitly in Section 2 and in the statements of the theorems and lemmas. Complete proofs are provided in Section 4 and in the appendix, with cross-references to the supporting lemmas.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This is a purely theoretical paper and does not contain any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper is theoretical and contains no experiments, code, or data.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper is theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper is theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper is theoretical and contains no experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is purely theoretical and conforms with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a foundational theoretical work on regret bounds for contextual bandits; we do not foresee any direct positive or negative societal impacts beyond those of bandit theory in general.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use any existing code, data, or model assets. All prior theoretical results that are built upon are properly credited via citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the methodology or any of the theoretical contributions, so a declaration is not required.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.