Bandits for Efficient Experimentation: Adapting to Control Group, Preferences, and Context Drifts

Udvas Das, Waris Radji, Debabrota Basu, Odalric-Ambrym Maillard

Keywords: Heteroskedastic Bandits, Context-drift, Customised preference, Control group, Minimum empirical divergence, Instance-dependent regret.

Summary

We consider a variant of the linear contextual stochastic multi-armed bandits, where the learner must provide recommendations to a group of users, each having its personalised preference vector, and in the presence of context distributions that are drifting over time. Under practitioner-friendly assumptions, we reduce this setting to linear bandit with stationary mean but heteroscedastic and non-stationary noise. We further study the case when the learner must satisfy a *baseline performance threshold*. We introduce `Dri-MED`, an algorithm inspired from the linear version of the MED strategy, and carefully adapted to handle the non-stationary heteroscedastic noise. We prove upper bounds on instance-dependent regret and also expected number of constraint violations. Our numerical results suggest that `Dri-MED` significantly outperforms conservative baselines that ignores the drift and preference structure.

Contribution(s)

1. In this work, we propose `Dri-MED` , the first MED-like algorithm that handles non-stationarity of contexts in every episode. We also provide instance-dependent upper bound on regret.
   Context: No prior algorithm in the existing literature handles non-stationarity in contexts.
2. We impose a safety constraint with respect to a base-policy (control-group). As we do not know the true mean parameter θ beforehand, this constraint is also stochastic in nature. We show `Dri-MED` can adapt to this constraint and the instance dependent regret upper bound involves a constraint-aware gap $\tilde{\Delta}$. We also prove with high probability, `Dri-MED` enjoys $\tilde{\mathcal{O}}(d)$ constraint violation
   Context: Existing literature on non-stationary bandits and safe bandits are presently unwed. We affirmatively bridge these two literature.
3. We propose a novel constraint-aware Lagrangian dual penalised version of the G-optimal design to efficiently compute exact probabilities for each arm at every step.
   Context: To best of our knowledge, no prior work handles customised preferences per user and propose a preference divergent allocation strategy.
4. We conduct numerical experiments across synthetic datasets with varying types of context drifts. `Dri-MED` outperforms the standard baseline algorithm significantly.
   Context: Our numerical results suggest that `Dri-MED` significantly outperforms conservative baselines that ignores the drift and preference structure.

Bandits for Efficient Experimentation: Adapting to Control Group, Preferences, and Context Drifts

Udvas Das, Waris Radji, Debabrota Basu, Odalric-Ambrym Maillard
udvas.das@inria.fr

Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRIStAL F-59000 Lille, France

Abstract

We consider a variant of the linear contextual stochastic multi-armed bandits, where the learner must provide recommendations to a group of users, each having its personalised preference vector, and in the presence of context distributions that are drifting over time. Under practitioner-friendly assumptions, we reduce this setting to linear bandit with stationary mean but heteroscedastic and non-stationary noise. We further study the case when the learner must ensure the mean reward of each decision must exceed that of a baseline strategy $\boldsymbol{\pi}_0$ at each decision step. We introduce `Dri-MED`, an algorithm inspired from the linear version of the MED strategy, and carefully adapted to handle the non-stationary heteroscedastic noise. We show that the instance-dependent regret scales as $\tilde{\mathcal{O}}\left(\frac{\kappa}{\tilde{\Delta}} d^2(\log(T)\right)$, where $\tilde{\Delta}$ is the constraint-aware sub-optimality gap subject to policy π_0, with variance-aware multiplicative term κ that we carefully handle using heteroscedastic regression. We further show `Dri-MED` enjoys $\tilde{\mathcal{O}}(d)$ expected constraint violations. Our numerical results suggest that `Dri-MED` significantly outperforms conservative baselines that ignores the drift and preference structure.

1 Introduction

Design of Experiments (DoE) has been an integral component of agricultural research, clinical trials, and natural sciences (Neyman, 1923; Fisher, 1926; Diggle & Chetwynd, 2011; Hoshmand, 2018). Suppose a farmer testing whether a new fertilizer, seed variety, or irrigation protocol outperforms the current standard cannot simply apply the new treatment to one plot and draw conclusions. Numerous factors such as soil fertility gradients, microclimatic variation, pest pressure, might confound any naive comparison (Gomez & Gomez, 1984; Tanner et al., 2023). This fundamental challenge DoE aims to address lies in separating the signal of a treatment effect from the noise of environmental heterogeneity. Pioneered by (Fisher, 1926) at the Rothamsted Experimental Station, DoE provides a principled statistical framework for planning agricultural field trials so that *valid, reliable, and efficient* conclusions can be drawn from inherently noisy observations (Rangaswamy, 1995).

The first and most fundamental objective is *Validity*, i.e., ensuring that an observed difference between a treated plot and a control plot can be causally attributed to the treatment itself, and not to pre-existing differences between plots. Thus, randomized experiment is at the core of sound statistical validation (Banerjee et al., 2020; Raccuglia et al., 2016; Kohavi & Longbotham, 2015) of a experimental design. Treatments must be assigned to experimental units by a chance mechanism, also referred to as an allocation distribution.

Reliability: Validity alone is insufficient in practice. Even a perfectly randomized experiment may fail to be *reliable* if, in the process of exploring new treatments, some field sites systematically receive practices that perform far worse than the current default. In agriculture, this is not merely a statistical concern but an operational and ethical one. A farmer whose plot is assigned a poorly performing treatment suffers an economically damaging yield loss, potentially threatening food security

or livelihood. Thus, it becomes imperative to enforce a *performance/safety constraint* (Pacchiano et al., 2021; Das & Basu, 2024; Amani et al., 2019), to ensure reliability. Safety constraints are typically imposed by performing a baseline action/policy over a control group in each site.

Efficiency: Finally, as real-world experimentation is costly, we want to minimize the suboptimal decisions over time and avoid unsafe decisions underperforming that of default baseline.

This common, albeit still idealized situation naturally fits the sequential decision-making (Sutton et al., 1998; Lattimore & Szepesvári, 2020) framework, where one observes the impact of an action or intervention by deploying it in an agricultural site and then adapts the future decisions accordingly. Now, suppose we observe a fixed set of farmers across different season. The change in environmental factors can be modelled as drifts in the decision making contexts over multiple interaction episodes. In addition, a critical but frequently overlooked dimension of agricultural experimentation is that different farmers evaluate the outcomes of a treatment differently (Fisher, 1926; Rangaswamy, 1995). A smallholder farmer may weigh yield stability above raw yield maximization, or a farmer in a water-scarce region may prioritize water-use efficiency over all other metrics. Hence, it calls for a design diverse enough that can adapt to these customised preferences.

In this work, we study this problem of designing valid, reliable, and efficient sequential experiments for a finite set of farmers experimenting across multiple agricultural sites with control groups and preferences, and across multiple episodes with plausible context drifts to find out optimal interventions over time. Though we elucidate our problem motivation with the agricultural example, the same problem reoccurs across domains, such as clinical trials for drugs and treatments (Réda et al., 2020; Munro et al., 2021), aircraft design optimisation (Mavrotas, 2009), recommender systems (Feijer et al., 2025; Saha et al., 2026) to name a few.

1.1 Problem Setting

We formalise this problem as a variant of the multi-armed bandit problem with a finite set of arms $\mathcal{A}$, from which a learner must recommend actions to be played to a group of H users (or experimental sites), in L successive rounds called *episodes*.

Contexts. The rewards depend on a context that might evolve over episodes. Precisely, within an episode $\ell \in [L]$, the contexts $C_{h,\ell}$ are generated from a context distribution $\mathcal{C}_\ell$ for all users $h \in [H]$. But the context distribution is non-stationary, i.e. it varies over the episodes ℓ. Then, the learner must recommend one action for each user $h \in [H]$. Upon recommending the actions $A_{h,\ell}$, the learner receives M-dimensional feedbacks $Y_{h,\ell}$, where $Y_{h,\ell}$ is a random vector of dimension M generated by an unknown distribution, with a specific linear structure (Assumption 2).

Preferences. Each user h further has a preference $\mathbf{p}_h \in \mathbb{R}^M$ over the vectorial feedback. This further yields a reward $r_{h,\ell}(a) \triangleq Y_{h,\ell}^\top \mathbf{p}_h$ for user h upon receiving an action $A_{h,\ell} = a$. The goal of the learner is to suggest a sequence of actions $\{A_{h,\ell}\}$ to the users over multiple episodes that maximises the expected cumulative reward, i.e., $\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} r_{h,\ell}(A_{h,\ell})\right]$.

Baseline policy. To simulate the control group and action, a mixed baseline policy $\pi_0 \in \mathcal{P}(\mathcal{A})$ is given to the learner. Given π_0 and a tolerance level $(1-\epsilon) \in [0,1]$, the learner must ensure to accumulate at least $(1-\epsilon)\mathbb{E}_{\pi_0}[r_{h,\ell}]$ reward for each user in each round. This additional constraint forces the learner to recommend policies that are at least as good or better than the baseline policy applied on the control group.

Performance metrics: In this setting, performance of an algorithm is measured with two metrics: cumulative regret and expected cumulative number of constraint violations. Lower values of both the metrics indicate better performance for an algorithm.

Regret. The cumulative regret is defined as

$$\mathrm{Reg} = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \Delta_{A_{h,\ell},h,\ell}\right] = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}\left(\mathbf{m}^\star_{h,\ell} - \mathbf{m}_{h,\ell}(A_{h,\ell})\right)\right]. \tag{1}$$

Here, the sub-optimality gaps $\Delta_{a,h,\ell} \triangleq \mathbf{m}^\star_{h,\ell} - \mathbf{m}_{h,\ell}(a)$, where $\mathbf{m}_{h,\ell}(a)$ denotes the mean of reward $r_{h,\ell}(a)$ and $\mathbf{m}^\star_{h,\ell} \triangleq \max_a \mathbf{m}_{h,\ell}(a)$. Also, Δ_0 denotes the minimum sub-optimality gap from π_0.

Constraint Violation. On the other hand, the safety is enforced by measuring the cumulated constraint violation with respect to a stationary baseline policy $\pi_0 \in \mathcal{P}(\mathcal{A})$,

$$\text{Violation}(\pi_0) = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\Big\{\mathbf{m}_{h,\ell}(A_{h,\ell}) < (1-\epsilon)\mathbb{E}_{a\sim\pi_0}[\mathbf{m}_{h,\ell}(a)]\Big\}\right]. \tag{2}$$

We observe that while the works on safe linear bandits need to assume explicitly existence of a safe arm satisfying the constraint (Pacchiano et al., 2021), we do not need such assumption. This is due to existence of a mixed baseline policy and the constraint defined by it. Thus, $\forall h \in [H], \ell \in [L]$, there exists at leasts one action $a \in \mathcal{A}$ satisfying $\mathbf{m}_{a,h,\ell} \geq (1-\epsilon)\mathbb{E}_{a\sim\pi_0}[\mathbf{m}_{h,\ell}(a)]$ for any $\epsilon < 1$.

Now, we elaborate the structural assumptions that we adhere to for the rest of the paper.

Assumption 1 (Gaussian feedback signal with context-independent mean)**.** *The feedback signal is Gaussian with mean assumed context-independent. Formally, for any $h \in [H]$ and $\ell \in [L]$, $Y_{h,\ell} \mid (A_{h,\ell} = a, C_{h,\ell} = c) \sim \mathcal{N}(m_a, \Sigma_{a,c})$, where $m_a \in \mathbb{R}^M$ is **constant** in c and $\Sigma_{a,c} \in \mathbb{R}^{M\times M}$ is **a positive semi-definite matrix** that varies with both arm a and context c. In particular*[1]

$$\mathbb{E}[Y_{h,\ell} \mid A_{h,\ell} = a, C_{h,\ell} = c] = m_a\,, \quad \text{Cov}[Y_{h,\ell} \mid A_{h,\ell} = a,\, C_{h,\ell} = c] = \Sigma_{a,c}. \tag{3}$$

In particular, under Assumption 1, the mean rewards are ℓ-independent, i.e., $\forall \ell, \mathbf{m}_{h,\ell}(a) = m_a \mathbf{p}_h$. Hence, the optimal mean $\mathbf{m}^\star_\ell$ and action $a^\star_h$ are also episode independent and user dependent. Assumption 1 is prevalent in many real-life settings. For example, in clinical trials mean response for a given dose remains stationary but variance in patients' responses is heteroskedastic (Legedza & Ibrahim, 2001). In case of recommender systems having arms as content categories (sports, politics), average engagement (click, watch-time) or reward for a category is stable over weeks but per-user variance can vary (Saha et al., 2026; Feijer et al., 2025).

Assumption 2 (Linear structure)**.** *The mean signal satisfies $m_a = \Phi(a)\theta$, for some unknown parameter $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 \leq S_*$, and a **known** feature matrix $\Phi(a) \in \mathbb{R}^{M\times d}$. In particular, the normalized reward $\tilde{r}_{h,\ell}(a) = r_{h,\ell}(a)/\sigma_{a,h,\ell}$ satisfies $\tilde{r}_{h,\ell}(a) \mid (C_{h,\ell} = c) \sim \mathcal{N}(\theta^\top \phi_{a,h,\ell}, 1)$, where*

$$\phi_{a,h,\ell} = \frac{1}{\sigma_{a,h,\ell}}\Phi(a)^\top \mathbf{p}_h \in \mathbb{R}^d, \quad \sigma^2_{a,h,\ell} = \mathbf{p}_h^\top \Sigma_{a,C_{h,\ell}}\,\mathbf{p}_h > 0. \tag{4}$$

The variance $\sigma^2_{a,h,\ell}$ in Assumption 2, is heteroskedastic in three senses simultaneously: it varies across actions a, users h (through $\mathbf{p}_h$), and episodes ℓ (through $C_{h,\ell} \sim \mathcal{C}_\ell$). We further introduce three quantities: (a) maximum variance per step $\bar{\sigma}^2_{h,\ell} = \max_a \sigma^2_{a,h,\ell}$, (b) condition number of covariances over the episodes $\kappa \triangleq \frac{\max_{a,h,\ell}\sigma^2_{a,h,\ell}}{\min_{a,h,\ell}\sigma^2_{a,h,\ell}} \leq \infty$, and (c) discounted cumulative variance $\Sigma_{LH} \triangleq \sum_{\ell=1}^{L}(\gamma_{\text{decay}})^{L-\ell}\sum_{h=1}^{H}\bar{\sigma}^2_{h,\ell}$ for some discount factor $\gamma_{decay} \in (0,1]$.

Discussion. Context-independence in Assumption 1 is naturally satisfied in applications where one base action, called the control, is available at each site location h, and the practitioner considers as feedback signal the difference of observed effects of action a and of the control. While we assume Gaussian noise, our theory and algorithm design extend to the general sub-Gaussian noise.

1.2 Outline and Contributions

We consider the problem of linear contextual bandits (LinCB) under non-stationary or drifting context distribution for H users encountered in each of L episodes, where each user comes with a customised preference feedback. We investigate the following questions:

[1] This implies independence between sites, since the distribution of $Y_{h,\ell}$ does not depend on $(A_{h',\ell'}, C'_{h',\ell})_{(h',\ell')\neq(h,\ell)}$

Algorithms	Instance Dependent Regret	Minimax Regret	Context Drift	Safety Constraint
OFUL (Abbasi-Yadkori et al., 2011)	$\tilde{\mathcal{O}}\left(\frac{d^2}{\Delta_{\min}}\log^3(T)\right)$	$\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$	✗	✗
SpannerIGW (Zhu et al., 2022)	$\Omega\left(\frac{1}{\Delta_{\min}}\sqrt{T}\right)$	$\mathcal{O}\left(\sqrt{dT\log K}\right)$	✗	✗
OPLB (Pacchiano et al., 2021)	✗	$\tilde{\mathcal{O}}\left(\frac{d\sqrt{T}}{\Delta_{\min}}\right)$	✗	✓
SOLID (Tirinzoni et al., 2020)	✗	$\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$	✗	✗
LinIMED (Bian & Tan, 2024)	✗	$\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$	✗	✗
LinMED (Balagopalan & Jun, 2024)	$\tilde{\mathcal{O}}\left(\frac{d^2}{\Delta_{\min}}\log^2(T)\right)$	$\tilde{\mathcal{O}}\left(d\sqrt{T}\right)$	✗	✗
Dri−MED (This work)	$\tilde{\mathcal{O}}\left(\frac{\kappa d^2}{\tilde{\Delta}}\log^2(T)\right)$	$\tilde{\mathcal{O}}\left(d\sqrt{\kappa T}\right)$	✓	✓

Table 1: Comparison of Regret Upper Bounds of Dri−MED against SOTA methods. Here, $T = LH$ and $\tilde{\Delta} = \min\{\Delta_{\min}, \Delta_0\}$, and $\kappa \triangleq \frac{\max_{a,h,\ell}\sigma^2_{a,h,\ell}}{\min_{a,h,\ell}\sigma^2_{a,h,\ell}}$.

1. Can we design an algorithm to solve this problem of safe LinCB under context drifts and customized preferential feedback, and derive an instance-dependent upper bound on its regret?
2. If we have a performance constraint subject to a control-group, or equivalently a safety constraint with respect to a base-policy π_0, can we adapt to it, i.e., can we derive an upper bound on expected number of constraint violations?
3. Can we optimally design sampling probabilities at each step that will be simultaneously constraint-aware and preference divergent (sufficient exploration in all customised preference directions)?

We aim to design a randomized strategy that efficiently computes the closed form probabilities of selecting each arm, since they prove to be beneficial in downstream tasks such as offline evaluation via inverse propensity score. In this work, we affirmatively answer the above questions and state them as main contributions below:

1. Context Drift and Customised Preference. In this work, we propose Dri−MED , the first MED-like algorithm that handles non-stationarity of contexts in every episode. We also prove an upper bound on the instance dependent regret of order $\tilde{\mathcal{O}}\left(\frac{\kappa d^2}{\tilde{\Delta}}\log^2(LH)\right)$. This non-trivial condition number captures sensitivity due to strength of drift in contexts across episodes.

2. Safety Constraint. In our setting, we impose a safety constraint with respect to a base-policy (control-group). As we do not know the true mean parameter θ beforehand, this constraint is also stochastic in nature. We show Dri−MED can adapt to this constraint and the instance dependent regret upper bound involves a constraint-aware gap $\tilde{\Delta}$. We also prove with high probability, Dri−MED enjoys $\tilde{\mathcal{O}}(d)$ constraint violation.

3. Constraint-aware and Preference Divergent Design. We propose a novel constraint-aware Lagrangian dual penalised version of the G-optimal design (Balagopalan & Jun, 2024) to efficiently compute exact probabilities for each arm at every step. To best of our knowledge, no prior work handles customised preferences per user and propose a preference divergent allocation strategy. We also verify this claim in our experimental analysis, as Dri−MED assigns more allocation on the best arm consistently over baselines.

4. Empirical Performance Gain. We conduct numerical experiments across synthetic datasets with varying types of context drifts, namely abrupt, periodical, gradual or no drift. We observe Dri−MED outperforms the standard baseline algorithm significantly. Additionally, we propose an IMED (Baudry et al., 2023; Honda & Takemura, 2015) version of Dri−MED for empirical comparison, namely Dri−IMED (see Section E). The performance of Dri−IMED proves to be competitive consistently with Dri−MED across all types of drift.

2 Related Work

This problem lies at the crossroad of different settings of bandit literature. While the linear parametrisation of rewards allows us to leverage the rich literature of linear contextual bandits, the

Algorithm 1 LinMED (Balagopalan & Jun, 2024)

Initialise $\hat{\theta}_0 = 0$, $V_0 = \lambda I$
for $t = 1, 2, \ldots$ **do**
  Observe arm set $\mathcal{A}_t$; compute $\hat{a}_t = \arg\max_{a' \in \mathcal{A}_t} \langle \hat{\theta}_{t-1}, a' \rangle$
  Compute gaps $\hat{\Delta}_{a,t} \leftarrow \langle \hat{\theta}_{t-1},\ \hat{a}_t - a \rangle$ for all $a \in \mathcal{A}_t$
  Compute weights $f_t(a)$ via for all $a \in \mathcal{A}_t$
  $p'_t(a) \leftarrow g(q_t^{\mathrm{opt}}(a)) f_t(a) \,/ \sum_{b \in \mathcal{A}_t} g(q_t^{\mathrm{opt}}(b)) f_t(b)$
  $\mathcal{B}_t \leftarrow \{a \in \mathcal{A}_t : \|a\|^2_{V_{t-1}^{-1}} > 1\}$
  if $|\mathcal{B}_t| > 0$ **then**
    $p_t(a) \leftarrow \frac{1}{2} p'_t(a) + \frac{1}{2} \mathbb{1}\{a = B_t\},\ B_t \in \mathcal{B}_t$ arbitrary
  else
    $p_t(a) \leftarrow p'_t(a)$
  end if
  Sample $A_t \sim p_t$; observe Y_t; update $V_t \leftarrow V_{t-1} + A_t A_t^\top,\ \hat{\theta}_t \leftarrow V_t^{-1} \sum_{s=1}^{t} A_s Y_s$
end for

context drifts over episodes are related to non-stationary and heteroskedastic linear bandits. Finally, the baseline policy requires us to bring techniques from bandits with safety constraints and extend them further. Here, we summarise the relevant literature.

Since the seminal OFUL (homoskedastic) strategy is proposed by Abbasi-Yadkori et al. (2011), the literature in linear multi-armed bandits has considerably expanded over the last decade (Krause & Ong, 2011; Valko et al., 2014; Lattimore & Szepesvari, 2017; Chowdhury & Gopalan, 2017; Abeille, 2017; Durand et al., 2018), culminating in provably instance-dependent optimal and efficient strategies LinIMED (Bian & Tan, 2024) and LinMED (Balagopalan & Jun, 2024), respectively inspired from the IMED (Honda & Takemura, 2015; Baudry et al., 2023) and MED (aka Maillard sampling) (Honda & Takemura, 2011; Qin et al., 2023; 2025) for unstructured bandits.

But the classical linear contextual bandits do not consider non-stationarity of contexts and also assume the variance across contexts and rounds to be static. To mitigate non-stationarity, Russac et al. (2019) and Faury et al. (2021) propose time-discounted regression estimates in (generalized) linear bandits under bounded variation budget.

Heteroskedasticity in linear bandits has been studied in (He & Gu, 2025), establishing variance-aware lower performance bounds. Kirschner & Krause (2018); Kim et al. (2026) assume known local variance to provide worst-case regret bounds in $\mathcal{O}((d\log(T) + \log(1/\delta))\sqrt{T})$, while Brégère et al. (2019) extends to the unknown heteroskedastic variance.

Batch bandits have received increasing attention over the past decade (Perchet et al., 2016; Gao et al., 2019; Jin et al., 2021), where the goal is to perform similarly to the pure sequential bandit problem with a minimal number of adaptively chosen batch sizes. Closer to this setting, Zhang et al. (2020) study when constant batch sizes are given as a constraint. This problem is also reminiscent of episodic reinforcement learning and combinatorial bandits, although with simpler structure.

Regarding the performance constraint, safety with respect to constraint violation has been studied in Amani et al. (2019), Pacchiano et al. (2021). Finally, we mention the somewhat related notion of satisficing objective Feng et al. (2025), as well as the study of bandits under linear constraints from Das & Basu (2024), for the different pure exploration objective.

A Primer on MED and LinMED. MED (Minimum Empirical Divergence) is a family of randomised bandit algorithms based on the probability matching philosophy (Honda & Takemura, 2011). MED-type algorithms pull each arm according to a probability dictating how likely it might be the optimal arm. Though originally proposed for bounded rewards and unstructured bandits, it has been extended to sub-Gaussian rewards, and is often referred to as Maillard sampling (Bian & Tan, 2024). Balagopalan & Jun (2024) extended the MED strategy to the linear bandits and proposed Lin-

MED. At each round t, it identifies the empirical best arm $\hat{a}_t = \arg\max_{a \in \mathcal{A}_t} \langle \hat{\theta}_{t-1}, a \rangle$ and computes exponential weights $f_t(a) = \exp\left(-\frac{\hat{\Delta}_{a,t}^2}{\beta_{t-1}(\delta_{t-1}) \, \|\hat{a}_t - a\|^2_{V_{t-1}^{-1}}} \right)$, where $\hat{\Delta}_{a,t} := \langle \hat{\theta}_{t-1}, \, \hat{a}_t - a \rangle$ is the estimated sub-optimality gap and $\beta_t(\delta_t)$ is a confidence scaling factor. These weights are used to compute an approximate G-optimal design q_t^{opt} over $\mathcal{A}_t$, which is then blended with a mass on the empirical best arm and a uniform component to form the final sampling distribution p_t. A saturation check $\mathcal{B}_t = \{a \in \mathcal{A}_t : \|a\|^2_{V_{t-1}^{-1}} > 1\}$ forces uniform exploration of under-sampled arms. The pseudocode is given in Algorithm 1. For further details, we refer to (Balagopalan & Jun, 2024).

3 Dri-MED: Drift Adaptive Minimum Empirical Divergence Algorithm

Now we are ready to propose our algorithm Dri-MED: drift adaptive minimum empirical divergence for safe linear contextual bandits under customised preference feedback. Dri-MED extends the MED-type algorithms in this setting. We choose MED-type algorithms for two reasons: (a) it yields tight problem-dependent regret bound for both independent arms (Baudry et al., 2023) and linear contextual settings (Balagopalan & Jun, 2024), (b) it allows an experimental design over the actions (Zhu et al., 2022; Balagopalan & Jun, 2024), i.e., it controls a non-zero probability of pulling each action for every user that leads to enough data collection for each of the actions and facilitates downstream evaluation of goodness of different actions against the baseline policy. The main challenges that Dri-MED resolve in addition to the LinMED algorithm (Balagopalan & Jun, 2024) are adapting to heteroskedastic noise across users with preferences, context drifts across episodes, and constraint violations against the baseline policies. We resolve them in three phases.

Phase 1: At the beginning of ℓ-th episode, we observe the contexts $\{C_{h,\ell}\}_{h=1}^H \sim \mathcal{C}_\ell$ and preferences $\{\mathbf{p}_h\}_{h=1}^H$ for all the H users. We compute per-user feature vectors $\Phi(a)^\top \mathbf{p}_h$ and then normalise with their corresponding $\sigma_{a,h,\ell}$ to tackle heteroskedasticity.

Phase 2: Since the true reward mean of the baseline policy $\boldsymbol{\pi}_0$, i.e., $\mu_0 \triangleq \mathbb{E}_{a \sim \pi_0}\left[\frac{1}{H}\sum_{h=1}^H \tilde{r}_h(a)\right]$ is unknown in the beginning, we construct an optimistic estimate of it at the beginning of each episode ℓ using the observations available till that episode. The optimistic estimate is built in two steps.

First, we use the available renormalized rewards and features $\{\tilde{r}_{h,s}, \phi_{A_{h,s},h,s}\}_{s=1,h=1}^{\ell-1,H}$ to create a ridge estimate of the underlying linear parameter θ. Specifically,

$$\hat{\theta}_{\ell-1} = \arg\min_{\theta} \sum_{s=1}^{\ell-1} \gamma_{\text{decay}}^{\ell-1-s} \sum_{h=1}^{H} \left(\theta^\top \phi_{A_{h,s},h,s} - \tilde{r}_{h,s}\right)^2 + \frac{\lambda}{2}\|\theta\|_2^2 = (V_{\ell-1}^{\gamma_{\text{decay}}})^{-1} \sum_{s,h=1}^{\ell-1,H} \phi_{A_{h,s},h,s} \tilde{r}_{h,s} \,, \tag{5}$$

where $V_{\ell-1}^{\gamma_{\text{decay}}} \triangleq \sum_{s=1}^{\ell-1} \gamma_{\text{decay}}^{\ell-1-s} \sum_{h=1}^{H} \phi_{A_{h,s},h,s} \phi_{A_{h,s},h,s}^\top + \lambda \gamma_{\text{decay}}^{\ell-1} I$ is an episodically discounted design matrix (aka Gram matrix) to tackle episodic context drifts. We use $\hat{\theta}_{\ell-1}$ to construct an empirical estimate $\hat{\mu}_{0,\ell} \triangleq \langle \hat{\theta}_{\ell-1}, \sum_a \pi_0(a) \sum_{h=1}^H \phi_{a,h,\ell} \rangle$ of the reward of the baseline policy π_0 at the beginning of episode ℓ. The proposed estimate extends the homoskedastic weighted linear regression (Russac et al., 2019; Kim et al., 2026) for non-stationary linear bandits to this episodic setting with heteroskedastic noise.

Second, we compute the confidence width around the empirical estimate $\hat{\mu}_{0,\ell}$ using $V_{\ell-1}^{\gamma_{\text{decay}}}$. Then, we inflate the empirical mean estimate $\hat{\mu}_0$ using this confidence width and saturation information of arms to get an *optimistic* estimate

$$\hat{\mu}_{0,\ell}^+ = \hat{\mu}_{0,\ell} + \sqrt{\beta_\ell(\delta_\ell)} \left\| \sum\nolimits_a \pi_0(a) \bar{\phi}_{a,\ell} \right\|_{(V_{\ell-1}^{\gamma_{\text{decay}}})^{-1}} , \tag{6}$$

where the confidence width $\beta_\ell(\delta_\ell) \triangleq \left(\sqrt{\log \frac{\det V_{\ell-1}^{\gamma_{\text{decay}}}}{\det V_0} + 2 \log \frac{1}{\delta_\ell}} + \sqrt{\lambda} S_* \right)^2$. Being optimistic in terms of base policy mean is being *pessimistic about constraint satisficing*. This design resonates with the existing safe bandit literature (Pacchiano et al., 2021; Das & Basu, 2024).

Algorithm 2 Dri-MED: **Dri**ft adaptive **M**inimum **E**mpirical **D**ivergence

Require: Baseline parameters π_0, ϵ, regularization parameter λ, S_*
Require: Discount over drifts $\gamma_{\text{decay}} \in [0,1]$ and confidence levels $\{\delta_\ell\}$
Require: Initial Lagrangian multiplier $\nu_0 > 0$, Lagrangian step size schedule $\{\eta_{h,\ell} = (\ell h)^{-\frac{1}{2}}\}$

1: Initialize $\hat{\theta}_{0,0} = 0$, $V_0 = \lambda I_d$, $\bar{C}_0 = 0$, and $N_0(a) = 0\ \forall a$
2: **for** episode $\ell = 1$ to L **do**
3: Initialize an episode with $V_{0,\ell}^{\gamma_{\text{decay}}} = \gamma_{\text{decay}}^{L-\ell} V_{\ell-1}^{\gamma_{\text{decay}}}$, $\hat{\theta}_{0,\ell} = \frac{1}{\gamma_{\text{decay}}}\hat{\theta}_{\ell-1}$, and $\nu_{h,\ell} = \nu_{\ell-1}$
Phase 1: Context observation & drift detection
4: Observe contexts $\{C_{h,\ell}\}_{h=1}^H \sim \mathcal{C}_\ell$ and preferences $\{\mathbf{p}_h\}_{h=1}^H$
5: For all $a \in \mathcal{A}$, compute per-user features $\phi_{a,h,\ell} = \frac{1}{\sigma_{a,h,\ell}}\Phi(a)^\top \mathbf{p}_h$ and mean feature $\bar{\phi}_{a,\ell} = \frac{1}{H}\sum_{h=1}^H \phi_{a,h,\ell}$
Phase 2: Episode-specific baseline & constraint
6: Confidence width on baseline $\sqrt{\beta_\ell(\delta_\ell)} = \sqrt{\log \frac{\det V_{\ell-1}^{\gamma_{\text{decay}}}}{\det V_0} + 2\log\frac{1}{\delta_\ell}} + \sqrt{\lambda} S_*$
7: Compute the optimistic baseline $\hat{\mu}_{0,\ell}^+$ with Equation (6)
8: Set the constraint threshold $\tau_\ell = (1-\epsilon)\hat{\mu}_{0,\ell}^+$
Phase 3: Within-episode decisions
9: **for** user $h = 1$ to H **do**
10: Play an action $A_{h,\ell} \leftarrow$ Apply Algorithm 3
11: Observe $Y_{h,\ell}$, $\mathbf{p}_h$, and $\tilde{r}_{h,\ell} = \frac{1}{\sigma_{A_{h,\ell},h,\ell}}\mathbf{p}_h^\top Y_{h,\ell}$
Update Lagrangian multiplier, Gram matrix, and parameter estimate
12: Constraint violation indicator: $\xi_{h,\ell} = \mathbb{1}\left\{\langle \hat{\theta}_{h,\ell}, \phi_{A_{h,\ell},h,\ell}\rangle < \tau_\ell\right\}$
13: Update the Lagrangian multiplier $\nu_{h,\ell} \leftarrow \max\left\{0, \nu_{h,\ell} + \eta_{h,\ell}\left(\xi_{h,\ell} - \frac{\epsilon}{1-\epsilon}\right)\right\}$
14: Update $N_{h+1,\ell}(A_{h,\ell}) = N_{h,\ell}(A_{h,\ell}) + 1$, $V_{h,\ell}^{\gamma_{\text{decay}}} = V_{h-1,\ell}^{\gamma_{\text{decay}}} + \phi_{A_{h,\ell},h,\ell}\phi_{A_{h,\ell},h,\ell}^\top$
15: Compute $\hat{\theta}_{h,\ell} = (V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}\sum_{s=1}^h \phi_{A_{s,\ell},s,\ell} r_{A_{s,\ell}}$
16: **end for**
17: Store episodic statistics $V_\ell^{\gamma_{\text{decay}}} \leftarrow V_{H,\ell}^{\gamma_{\text{decay}}}, \hat{\theta}_\ell \leftarrow \hat{\theta}_{H,\ell}, \bar{C}_\ell, \mu_{0,\ell}, \nu_{h,\ell}$
18: **end for**

Finally, this yields the constraint threshold $\tau_\ell = (1-\epsilon)\hat{\mu}_{0,\ell}^+$ for the present episode.

Phase 3: The baseline adaptive MED step. In beginning of this phase, we play an arm according to Algorithm 3. The sampling strategy uses Minimum Empirical Divergence at its core. The novelty lies in the adaption for our episodic setting with non-stationary contexts and baseline. We first compute the empirical best arm and estimate the gaps $\hat{\Delta}_{A_{h,\ell},h,\ell}$. We use these estimated gaps to calculate exponential weights over all arms as

$$f_h(a) = \exp\Big(-\hat{\Delta}_{a,h,\ell}^2 / \left(\beta_{h,\ell}(\alpha_{h,\ell})\|\phi_{\hat{a}_{h,\ell},h,\ell} - \phi_{a,h,\ell}\|^2_{(V_{h,\ell}^\gamma)^{-1}}\right)\Big) \tag{7}$$

$$\tilde{f}_h(a) = f_h(a)\cdot \exp\Big(-\nu_{h,\ell}\cdot \max\{0,\ \tau_\ell - \text{LCB}(a,h,\ell)\}\Big) \tag{8}$$

To make the weights constraint-aware, we introduce a Lagrangian dual based penalty with Lagrangian multiplier $\nu_{h,\ell}$. Specifically, we show in Lemma 3 this dual parameter grows at a rate $\mathcal{O}(\sqrt{\ell h})$. Next, we rescale the arms using normalized features. We introduce a novel baseline-aware ApproxDesign() algorithm (refer Section G) or C_{opt}-optimal design that ensures saturation of the true best arm efficiently, i.e., $\left\|a_{h,\ell}^*\right\|^2_{(V(p_h)^{\gamma_{\text{decay}}})^{-1}} \le C_{\text{opt}}\mathcal{O}\left(d\log(d)\right)$.

Though the baseline-awareness is adapted in the optimal design via $q_h(a)$. While we compute the final design probability in Line 9, we show in Lemma 10, the denominator is lower bounded by a dual-penalised term. Specifically $\sum_b q_h(b)\tilde{f}_h(b) > \alpha_{\text{opt}} e^{-\nu_{h,\ell}}$. This prevents unsafe arms from saturating and efficiently discards them. Finally, upon checking the saturation status of each arm and construct the augmented arm set as $B_{h,\ell} = \left\{a : \|\phi_{a,h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} > 1\right\}$. Any arm in this set

Algorithm 3 Baseline Adaptive Minimum Empirical Divergence (MED) Step

Require: Episode $\ell \in [L]$, user id $h \in [H]$, user's context $\{\phi_{a,h,\ell}\}_{a=1}^{|\mathcal{A}|}$, $\hat{\theta}_{h,\ell}, V_{h,\ell}^{\gamma_{\text{decay}}}, \lambda$.
Require: Constraint threshold τ_ℓ, Lagrangian multiplier $\nu_{h,\ell}, \alpha_{\text{emp}}, \alpha_{\text{opt}}$.

Compute confidence radius: $\beta_{h,\ell}(\alpha_{h,\ell})^{1/2} = \sqrt{\log \frac{\det V_{h,\ell}^{\gamma_{\text{decay}}}}{\det V_0} + 2\log \frac{1}{\alpha_{h,\ell}}} + \sqrt{\lambda} S_*$

For each arm $a \in \mathcal{A}$: $\text{LCB}(a,h,\ell) = \langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell} \rangle - \beta_{h,\ell}(\alpha_{h,\ell})^{1/2} \|\phi_{a,h,\ell}\|_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}}$

Empirical best action: $\hat{a}_{h,\ell} = \arg\max_{a \in \mathcal{A}} \langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell} \rangle$

Empirical gaps: $\hat{\Delta}_{a,h,\ell} = \langle \hat{\theta}_{h,\ell},\ \phi_{\hat{a}_{h,\ell},h,\ell} - \phi_{a,h,\ell} \rangle$

Compute Constraint-aware pulling probabilities: For every $a \in \mathcal{A}$, compute $\tilde{f}_h(a)$ with Equation (8).

Rescale arms: $\mathcal{A}^{(h)} = \left\{ \sqrt{\tilde{f}_h(a)} \cdot \phi_{a,h,\ell} \;\middle|\; a \in \mathcal{A} \right\}$

$q_h^{\text{opt}} = \text{ApproxDesign}\left(\mathcal{A}^{(h)}\right) \leftarrow$ Algorithm 6

$q_h(a) = \alpha_{\text{opt}}\, q_h^{\text{opt}}(a) + \alpha_{\text{emp}}\, \mathbf{1}[a = \hat{a}_{h,\ell}] + (1 - \alpha_{\text{opt}} - \alpha_{\text{emp}})\, \frac{1}{|\mathcal{A}|}$

Compute final design probabilities: $p'_h(a) = q_h(a)\, \tilde{f}_h(a) / \sum_b q_h(b)\, \tilde{f}_h(b)$

Check saturation event for each arms: $B_{h,\ell} = \left\{ a : \|\phi_{a,h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} > 1 \right\}$

$$p_h(a) = \begin{cases} \frac{1}{2}\, p'_h(a) + \frac{1}{2}\, \mathbf{1}[a \in B_{h,\ell}] & \text{if } B_{h,\ell} \neq \emptyset \\ p'_h(a) & \text{otherwise} \end{cases}$$

Sample $A_{h,\ell} \sim p_h$

has not been explored enough to be sure of it's mean parameter, thus we pull this arm with a toss of a fair coin. Otherwise we play by the design probabilities. At the end, we observe the feedback signal, normalize the reward we compute. Then we check whether the constraint has been violated by the playing arm and update the Lagrangian multiplier $\nu_{h,\ell}$ accordingly.

4 Regret Analysis of Dri-MED

In this section, we provide upper bounds on both instance dependent regret, and expected constraint violation for the strategy Dri-MED, along with very high-level proof intuitions.

Assumption 3 (Boundedness). *1.* $\max_{a,h,\ell} \|\phi_{a,h,\ell}\|_2 \leq 1$, *2.* $\max_h \|\mathbf{p}_h\|_2 \leq 1$, *and 3.* $\max_{a,h,\ell} \Delta_{a,h,\ell} \leq B$.

Standard Regret. We first decompose the original regret in Equation 1 into two parts based on the event $\mathcal{V}_{h,\ell}(A_{h,\ell}) = \{\mathbf{m}_{h,\ell}(A_{h,\ell}) < (1-\epsilon)\mathbb{E}_{a\sim\pi_0}[\mathbf{m}_{h,\ell}(a)]\}$ as:

$$\text{Reg} = \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \Delta_{A_{h,\ell},h,\ell} \mathbb{1}\left\{\overline{\mathcal{V}_{h,\ell}(A_{h,\ell})}\right\}\right]}_{\triangleq \text{Reg}^{\text{safe}}} + \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \Delta_{A_{h,\ell},h,\ell} \mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\right]}_{\triangleq \text{Reg}^{\text{viol}}} \tag{9}$$

The novel decomposition in Equation (9) allows to analyze the regret of Dri-MED in two stages: 1. bound on sum of suboptimality over time under the event that we do not pull any truly unsafe arm, 2. under the event we pull arms that are unsafe.

Reg^{safe}. We formalize the regret upper bound guaranty under safety constraint below in Theorem 1.

Theorem 1 (Instance Dependent Regret Upper Bound under Constraint Satisfaction). *Under Assumption 1,2, and 3, Algorithm 2 and 3 jointly satisfies –*

$$\text{Reg}^{\text{safe}} = \tilde{\mathcal{O}}\left(\kappa \frac{d^2\left(\log^2(LH) + \log\left(\Sigma_{LH}\right)\right)}{\Delta_{\min}}\right)$$

Proof Concept. The main challenge while analysing regret in our setting lies in elegantly incorporate the constraint violation event, and also to capture the effect of the episodic drift in context. To prove upper bound on the regret under constraint satisficing, we adapt the proof structure of Balagopalan & Jun (2024) to our episodic interaction setting with known, but shifting reward noise variance. The adaptation comes in two folds: 1. while analysing $\mathrm{Reg}^{\mathrm{safe}}$ the first challenge is handled by controlling the true gap $\Delta_{A_{h,\ell},h,\ell}$ using peeling technique around $(\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_{\boldsymbol{\pi}_0})$, instead of the maximum gap B, 2. we realise the effect drifting context comes implicitly through the non-stationary noise variance involved in the feedback signal $Y_{A_{h,\ell},h,\ell}$. Thus, the log-dependency on sum of variances in Theorem 1 comes naturally while we try to bound the saturation event of feature of $A_{h,\ell}$. i.e., $\mathbb{E}\left[\|\phi_{A_{h,\ell}}\|_{(V_{h,\ell}^{\gamma_{\mathrm{decay}}})^{-1}} > 1\right]$. We prove this by using (Abbasi-Yadkori et al., 2011, Lemma 4) . For more details, refer to Section B in the supplementary materials.

$\mathrm{Reg}^{\mathrm{viol}}$. This part of the whole regret is about characterising the suboptimality cost of the arms that are unsafe in true sense. We formally state the upper bound for this part below:

Theorem 2 (Instance Dependent Regret Upper Bound under Constraint Violation)**.** *Under Assumption 1,2, and 3, Algorithm 2 and 3 jointly satisfies –*

$$\mathrm{Reg}^{\mathrm{viol}} = \tilde{\mathcal{O}}\left(\sigma_{\max}^2 \frac{B - \Delta_{\boldsymbol{\pi}_0}}{(\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)^2} d^2 \left(\log^2(LH) + \log\left(\Sigma_{LH}\right)\right)\right)$$

Proof Concept. Existing literature (Bian & Tan, 2024; Balagopalan & Jun, 2024) use peeling technique to leverage a lower bound on true gap that in turn, helps upper bounding the pulling probability of arm $A_{h,\ell}$. From the definition of the event $\mathcal{V}_{h,\ell}(\alpha_{h,\ell})$, we immediately show $\Delta_{A_{h,\ell},h,\ell} > \Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_{\boldsymbol{\pi}_0}$. Thus, 1. we do not need additional peeling over the value of $\Delta_{A_{h,\ell},h,\ell}$, 2. we again leverage this natural lower bound while handling the conditioning event on concentration of $\hat{\theta}_{h,\ell}$ around true parameter θ. For the detailed proof, refer to Section B.

Final Regret. We combine $\mathrm{Reg}^{\mathrm{safe}}$ and $\mathrm{Reg}^{\mathrm{viol}}$ to state the final regret upper bound of `Dri-MED`:

Theorem 3 (Instance Dependent Regret Upper Bound of `Dri-MED`)**.** *Let us denote* $\frac{1}{\tilde{\Delta}} \triangleq \max\left\{\frac{1}{\Delta_{\min}}, \frac{B-\Delta_{\boldsymbol{\pi}_0}}{(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}\right\}$. *Under Assumption 1,2, and 3, `Dri-MED` exhibits*

$$\mathrm{Reg} = \tilde{\mathcal{O}}\left(\kappa \frac{1}{\tilde{\Delta}} d^2 \left(\log^2(LH) + \log\left(\Sigma_{LH}\right)\right)\right).$$

Discussion. 1. While (Balagopalan & Jun, 2024) achieves a regret of $\tilde{\mathcal{O}}\left(\kappa \frac{1}{\Delta_{\min}} d^2 \log^2(LH)\right)$, our bound also depends on the constraint gap Δ_0 due to the baseline policy and the condition number of the variances over the episodes due to context drifts. 2. While the safe contextual bandit literature provides a minimax bound on the regret under the assumption of a safe action Amani et al. (2019); Pacchiano et al. (2021), we remove the assumption due to access to a baseline policy while achieving a problem-dependent regret upper bound as well as a minimax regret bound of same order. 3. On the other hand, if we consider the implication of our results for heteroskedastic linear bandits, we observe that we derive a problem-dependent regret bound under heteroskedasticity while the present literature focuses on minimax regret (Kirschner & Krause, 2018; Kim et al., 2026).

Constraint Violation. It is imperative for a safe bandit algorithm (Amani et al., 2019; Pacchiano et al., 2021) to have an upper bound on expected number of true constraint violations (defined in Equation (2) for the whole interaction. For `Dri-MED`, we state this guaranty formally below:

Theorem 4 (Expected Constraint Violation of `Dri-MED`)**.** *Under Assumption 1,2, and 3, `Dri-MED` suffers expected number of constraint violations* $\mathrm{Violation}(\boldsymbol{\pi}_0) = \tilde{\mathcal{O}}(d)$.

Proof Concept. In Lemma 5, we prove under constraint violation, the true gap is lower bounded by the gap subject to base-policy $\boldsymbol{\pi}_0$. That means $\Delta_{A_{h,\ell},h,\ell} \geq \Delta_{\boldsymbol{\pi}_0}$. While analysing standard regret, we decompose $\mathrm{Reg}^{\mathrm{viol}}$ into two parts,

$$\mathrm{Reg}^{\mathrm{viol}} = \underbrace{\Delta_{\boldsymbol{\pi}_0} \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\right]}_{\mathrm{Violation}(\boldsymbol{\pi}_0)} + \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} (\Delta_{A_{h,\ell},h,\ell} - \Delta_{\boldsymbol{\pi}_0}) \mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\right]$$

one of which is the expected constraint violation. We prove upper bound on this by simply using Corollary 1 adapted (Jun & Kim, 2024, Lemma C.2). We discuss it at length in Section B.

Implication. (1) The variance dependency is captured by κ and $\Sigma_{LH} \triangleq \sum_{\ell=1}^{L} {\gamma_{\mathrm{decay}}}^{L-\ell} \sum_{h=1}^{H} \bar{\sigma}^2_{h,\ell}$ appearing in the logarithmic term. This logarithmic part is reminiscent of the term $\sum_t \omega_t^2$ appearing in (Russac et al., 2019, Thm.1), further adapted here to the heteroskedastic setup.

(2) Our bound is instance-dependent, scaling as polylog of T, contrasting with minimax regret bounds. From He & Gu (2025); Kim et al. (2026), the minimax lower bound is of order $d\sqrt{\sum_t \sigma_t^2}$ where σ_t^2 is the variance of reward r_t. Compare to the simplified setting of Balagopalan & Jun (2024), we replace the term $1/\Delta_{\min}$ with the larger $\kappa/\tilde{\Delta}$, mainly due to the safety constraint, and the handling of heteroskedastic variance. We conjecture the dependency on κ might be further reduced.

(3) The intricate threshold within $\tilde{\Delta}$ appears as the constraint induced by π_0 does not vanish even as $\epsilon \to 1$. Indeed, even in this case, the mean rewards must stay positive. We pay the largest regret when $\Delta_{\pi_0} = 0$, forcing a maximal factor $B/(\epsilon^2 \mu_0^2)$, while as Δ_{π_0} approaches the largest gap B, we recover the leading factor $1/\Delta_{\min}$ from the unconstrained setup (Balagopalan & Jun, 2024).

5 Numerical Experiments

We evaluate `Dri-MED` and `Dri-IMED`, a deterministic variant of our algorithm that we do not theoretically analyse (Appendix E), on a synthetic episodic contextual linear bandit with preference feedback and context drift, focusing on gradual and periodic drift regimes. Full environment details, additional drift regimes, ablation studies, and hyperparameter settings are provided in Appendix F.

Baselines. We compare against four stationary linear bandit algorithms that ignore both drift and preference structure: `OFUL` (Abbasi-Yadkori et al., 2011), `LinMED` (Balagopalan & Jun, 2024), `LinIMED` (Bian & Tan, 2024), and `LinTS` (Agrawal & Goyal, 2013). To our knowledge, no prior algorithm addresses this combined setting. Thus, the baselines serve as the natural reference points[2].

Environment. We consider a synthetic environment with $H = 10$ users, $K = 5$ arms, and $L = 1000$ episodes; full parameter details are given in Appendix F (Algorithm 5). We study two drift regimes, both parametrised by magnitude $\kappa = 100$: *gradual drift*, where the context scale grows linearly as $f(\ell) = \kappa \cdot \ell/L$, producing a smooth and monotone increase in noise; and *periodic drift*, where $f(\ell) = \kappa \cdot \frac{1}{2}(1+\sin(2\pi\ell/(L/50)))$, producing recurring fluctuations that repeatedly challenge the constraint mechanism. All results are averaged over 128 independent seeds.

Regret Evolution and Constraint Violation. Figure 1 reports cumulative regret for both drift regimes. `Dri-MED` and `Dri-IMED` achieve substantially lower regret than all baselines, with the gap widening as episodes progress. Among the baselines, `OFUL` performs best yet still incurs regret an order of magnitude larger than that of our methods. `LinMED` and `LinTS` perform the worst, confirming that ignoring heteroskedasticity and preference structure is costly. Performance is stable across both drift types, validating that the drift-adaptive discounted regression successfully absorbs the non-stationarity, whether it is smooth or oscillatory. Both `Dri-MED` and `Dri-IMED` rapidly drive the true constraint violation rate to zero after an initial exploration phase, while all stationary baselines exhibit persistent violations throughout.

[2]Source code can be found on https://github.com/riiswa/context_drift_lin_bandits/

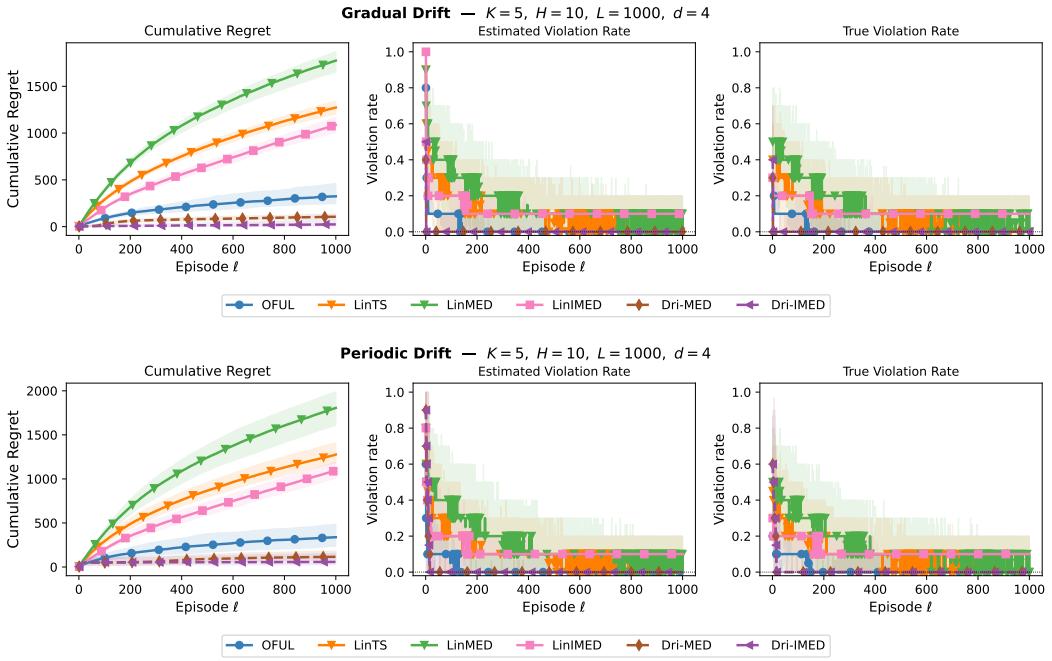

Figure 1: Cumulative regret (left), estimated violation rate (center), and true violation rate (right) for gradual and periodic drift. Shaded bands show the 5–95% quantile range over 128 seeds. Results for no-drift and abrupt drift are provided in Appendix F.

Design Probabilities. The arm allocation analysis (Appendix F, Figure 4) reveals a clear behavioural distinction between algorithms. `Dri-MED` concentrates almost all pull mass on the oracle-optimal arm $a_h^\star$ for every user. In contrast, `LinMED` spreads significant mass across suboptimal arms, unable to distinguish arm quality under the heteroskedastic preference structure it ignores.

For more experimental results and ablation studies, refer to Section F in supplementary materials.

6 Discussions, Limitations, and Future Works

We study a class of sequential experimental problems with a group of users having heterogeneous preferences and temporally evolving context distributions as a linear contextual bandit with heteroskedastic and non-stationary noise but stationary mean. In this context, we propose `Dri-MED` that achieves (a) validity, i.e. a distribution of scores over its actions, (b) reliability, i.e. improved performance than a baseline action/policy applied over a control group, and (c) efficiency, i.e. low regret. We prove that `Dri-MED` achieves logarithmic regret while exhibiting better numerical performance than the existing baselines.

The main limitation of this work is the stationary reward mean assumption. Though this assumption with heteroskedastic noise holds true for multiple applications, it would interesting to generalise our analysis for non-stationary means. In addition, the present MED strategies work only for sub-Gaussian rewards, whereas in real-life experiments, the noise distribution can exhibit heavy tails. Designing heavy-tail robust MED algorithms in this context, and in general remain open questions.

Broader Impact Statement

This work mainly aimed for theoretical advancements in the safe, heteroscedastic contextual bandit literature. The authors does not see any potential negative impact that should be mentioned here.

Acknowledgments

The authors would like to acknowledge PEPR project FOUNDRY (ANR23-PEIA-0003) for their support. DB and UD would like to acknowledge ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01). We would also like to acknowledge the anonymous reviewers for their constructive feedback. Authors are members of the Inria team-project Scool.

References

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

Marc Abeille. *Exploration-exploitation with Thompson sampling in linear systems*. PhD thesis, Université de Lille 1, 2017.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

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

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Kapilan Balagopalan and Kwang-Sung Jun. Minimum empirical divergence for sub-gaussian linear bandits. *arXiv preprint arXiv:2411.00229*, 2024.

Abhijit V Banerjee, Sylvain Chassang, Sergio Montero, and Erik Snowberg. A theory of experimenters: Robustness, randomization, and balance. *American Economic Review*, 110(4):1206–1230, 2020.

Dorian Baudry, Fabien Pesquerel, Rémy Degenne, and Odalric-Ambrym Maillard. Fast asymptotically optimal algorithms for non-parametric stochastic bandits. *Advances in Neural Information Processing Systems*, 36:11469–11514, 2023.

Jie Bian and ntnt YF Tan. Indexed minimum empirical divergence-based algorithms for linear bandits. *arXiv preprint arXiv:2405.15200*, 2024.

Margaux Brégère, Pierre Gaillard, Yannig Goude, and Gilles Stoltz. Target tracking for contextual bandits: Application to demand side management. In *International Conference on Machine Learning*, pp. 754–763. PMLR, 2019.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Udvas Das and Debabrota Basu. Learning to explore with lagrangians for bandits under unknown constraints. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

Peter J. Diggle and Amanda G. Chetwynd. 5 experimental design: agricultural field experiments and clinical trials. In *Statistics and Scientific Method: An Introduction for Students and Researchers*. Oxford University Press, 08 2011. ISBN 9780199543182. DOI: 10.1093/acprof:oso/9780199543182.003.0005.

Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *Journal of Machine Learning Research*, 19 (17):1–34, 2018.

Louis Faury, Yoan Russac, Marc Abeille, and Clément Calauzènes. Regret bounds for generalized linear bandits under parameter drift. *arXiv preprint arXiv:2103.05750*, 2021.

Diego Feijer, Himan Abdollahpouri, Sanket Gupta, Alexander Clare, Yuxiao Wen, Todd Wasson, Maria Dimakopoulou, Zahra Nazari, Kyle Kretschman, and Mounia Lalmas. Calibrated recommendations with contextual bandits. *arXiv preprint arXiv:2509.05460*, 2025.

Qing Feng, Tianyi Ma, and Ruihao Zhu. Satisficing regret minimization in bandits. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ronald Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33: 503–515, 1926.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.

Kwanchai A Gomez and Arturo A Gomez. *Statistical procedures for agricultural research*. John wiley & sons, 1984.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Jiafan He and Quanquan Gu. Variance-dependent regret lower bounds for contextual bandits. *arXiv preprint arXiv:2503.12020*, 2025.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *The Journal of Machine Learning Research*, 16(1):3721–3756, 2015.

Reza Hoshmand. *Design of experiments for agriculture and the natural sciences*. Chapman and Hall/CRC, 2018.

Tianyuan Jin, Jing Tang, Pan Xu, Keke Huang, Xiaokui Xiao, and Quanquan Gu. Almost optimal anytime algorithm for batched multi-armed bandits. In *International Conference on Machine Learning*, pp. 5065–5073. PMLR, 2021.

Kwang-Sung Jun and Jungtaek Kim. Noise-adaptive confidence sets for linear bandits and application to bayesian optimization. *arXiv preprint arXiv:2402.07341*, 2024.

Sanghwa Kim, Junghyun Lee, and Se-Young Yun. A jointly efficient and optimal algorithm for heteroskedastic generalized linear bandits with adversarial corruptions. *arXiv preprint arXiv:2602.10971*, 2026.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pp. 358–384. PMLR, 2018.

Ron Kohavi and Roger Longbotham. Online controlled experiments and a/b tests. *Encyclopedia of machine learning and data mining*, pp. 1–11, 2015.

Andreas Krause and Cheng Ong. Contextual gaussian process bandit optimization. *Advances in neural information processing systems*, 24, 2011.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pp. 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. DOI: 10.1017/9781108571401.

Anna TR Legedza and Joseph G Ibrahim. Heterogeneity in phase i clinical trials: prior elicitation and computation using the continual reassessment method. *Statistics in Medicine*, 20(6):867–882, 2001.

George Mavrotas. Effective implementation of the ε-constraint method in multi-objective mathematical programming problems. *Applied mathematics and computation*, 213(2):455–465, 2009.

Alasdair PS Munro, Leila Janani, Victoria Cornelius, Parvinder K Aley, Gavin Babbage, David Baxter, Marcin Bula, Katrina Cathie, Krishna Chatterjee, Kate Dodd, et al. Safety and immunogenicity of seven covid-19 vaccines as a third dose (booster) following two doses of chadox1 ncov-19 or bnt162b2 in the uk (cov-boost): a blinded, multicentre, randomised, controlled, phase 2 trial. *The Lancet*, 398(10318):2258–2276, 2021.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on. *Statistical Science*, 5(4,465-480), 1923. reprinted in 1990.

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

Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, pp. 660–681, 2016.

Hao Qin, Kwang-Sung Jun, and Chicheng Zhang. Kullback-leibler maillard sampling for multi-armed bandits with bounded rewards. *Advances in Neural Information Processing Systems*, 36: 60514–60526, 2023.

Hao Qin, Kwang-Sung Jun, and Chicheng Zhang. Achieving adaptivity and optimality for multi-armed bandits using exponential-kullback leibler maillard sampling. *arXiv preprint arXiv:2502.14379*, 2025.

Paul Raccuglia, Katherine C Elbert, Philip DF Adler, Casey Falk, Malia B Wenny, Aurelio Mollo, Matthias Zeller, Sorelle A Friedler, Joshua Schrier, and Alexander J Norquist. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, 2016.

R Rangaswamy. *A text book of agricultural statistics*. New age international, 1995.

Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Machine learning applications in drug development. *Computational and structural biotechnology journal*, 18:241–252, 2020.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

Aadirupa Saha, Amith Bhat, and Haipeng Luo. One good source is all you need: Near-optimal regret for bandits under heterogeneous noise. *arXiv preprint arXiv:2602.14474*, 2026.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Kari Christine Tanner, Gordon B Jones, and Elizabeth Claire Verhoeven. *Experimenting on the Farm: Introduction to Experimental Design*. Oregon State University Extension Service, 2023.

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33:1417–1427, 2020.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International conference on machine learning*, pp. 46–54. PMLR, 2014.

Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829, 2020.

Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pp. 27428–27453. PMLR, 2022.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Notations

Notation	Description
ℓ	Episode index $\in [1, L]$
h	User index $\in [1, H]$
θ	True reward parameter with $\|\theta\|_2 \leq S_*$
$\hat{\theta}_{h,\ell}$	Estimate of the true parameter θ for h-th user in ℓ-th episode
K	Number of actions
$\mathbf{c}_\ell$	Context distribution at episode ℓ
$\Phi(a)$	$M \times d$ feature matrix for arm $a \in \mathcal{A}$
$Y_{h,\ell}$	M-dimensional feedback signal $\mathbb{E}[Y_{h,\ell} \mid A_{h,\ell} = a, C_{h,\ell} = c] = \Phi(a)\theta$
$\mathbf{p}_h$	d-dimensional preference vector for h-th user
$r_{h,\ell}$	Scalar reward for h-th user in ℓ-th episode $= Y_{h,\ell}^\top \mathbf{p}_h$
$\phi_{a,h,\ell}$	d-dimensional context for arm $a = \frac{1}{\sigma_{a,h,\ell}} \Phi(a)^\top \mathbf{p}_h$
$\Sigma_{a,h,\ell}$	$= \mathrm{Cov}[Y_{h,\ell} \mid A_{h,\ell} = a, C_{h,\ell} = c]$
$\sigma_{a,h,\ell}$	$= \mathrm{Var}(\eta_{h,\ell} \mid A_{h,\ell} = a)$ (Variance of noise in the feedback, if arm a is pulled)
$\sigma^2_{h,\ell}$	$= \min_{a \in \mathcal{A}} \sigma^2_{a,h,\ell}$
$\sigma^2_{\max}$	$= \max_{a,h,\ell} \sigma^2_{a,h,\ell}$
$\sigma^2_{\min}$	$= \min_{a,h,\ell} \sigma^2_{a,h,\ell}$
$\Delta_{a,h,l}$	$= \theta^\top(\phi_{a^*_\ell,h,\ell} - \phi_{a,h,\ell}) \geq 0$
$\hat{\Delta}_{a,h,l}$	$= \hat{\theta}_t^\top(\phi_{a^*_\ell,h,\ell} - \phi_{a,h,\ell})$
π_0	Baseline policy
$\Delta_{\boldsymbol{\pi}_0}$	$= \min_{\ell \in [L]} \theta^\top(\phi_{a^*_\ell,h,\ell} - \phi_{\pi_0})$
$\mu_{0,\ell}$	Estimated reward of the baseline policy $\frac{1}{H}\sum_{h=1}^H \sum_a \pi_0(a)\langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell}\rangle$
$\mu^+_{0,\ell}$	Optimistic reward of baseline $\mu_{0,\ell} + \sqrt{\beta_\ell(\delta_\ell)}\left\|\frac{1}{H}\sum_h \sum_a \pi_0(a)\phi_{a,h,\ell}\right\|_{(V^\gamma_\ell)^{-1}}$
τ_ℓ	Constraint threshold $= (1-\epsilon)\mu^+_{0,\ell}$, for $\epsilon \geq 0$
τ	True constraint threshold $= (1-\epsilon)\mu_0$, where μ_0 is the true mean reward for π_0
$\xi_{h,\ell}$	Constraint violation indicator $= \mathbb{1}\left\{\langle \hat{\theta}_{h,\ell}, \phi_{A_{h,\ell},h}\rangle < \tau_\ell\right\}$
$\mathcal{V}_{h,\ell}(A_{h,\ell})$	True constraint violating event $\{(h,\ell) \in [h,\ell] : \langle \theta, \phi_{A_{h,\ell},h}\rangle < \tau\}$
μ_0	True mean of base-policy $= \theta^\top \phi_{\boldsymbol{\pi}_0}$

For the supplementary materials, whenever we write $\underline{\Delta}_{a,h,\ell}$, it denotes the feature normalized true suboptimal gap.

B Proof of Regret Upper Bound of Dri-MED (Algorithm 2)

B.1 Good Event: Concentration and Confidence Width

We start by defining the following good events:

$$\mathcal{G}_1 \triangleq \left\{ \forall h \geq 1, \ell \geq 1 : \|\theta - \hat{\theta}_{h,\ell}\|^2_{V^{\gamma_{\text{decay}}}_{h,\ell}} \leq \beta_{h,\ell}(\alpha_{h,\ell}) \right\} \tag{10}$$

Lemma 1 (Heteroscedastic weighted (h,ℓ)-th confidence set)**.** *Following the information acquisition rule with action $A_{h,\ell} = a$ and it's rescaled feature vector $\phi_{a,h,\ell}$, i.e., $V^{\gamma_{\text{decay}}}_{h,\ell} = V^{\gamma_{\text{decay}}}_{h-1,\ell} + \phi_{a,h,\ell}\phi^\top_{a,h,\ell}$, we define*

$$\beta_{h,\ell}(\alpha_{h,\ell})^{1/2} \triangleq \sqrt{2\log\frac{\det V^{\gamma_{\text{decay}}}_{h,\ell}}{\det V_0} + 2\log\frac{1}{\alpha_{h,\ell}}} + \sqrt{\lambda}\, S_* \,,$$

where $\alpha_{h,\ell} \in (0,1)$ is to characterised later on. Then $\mathbb{P}(\mathcal{G}_1) \geq 1 - \sum_{\ell=1}^{L}\sum_{h=1}^{H}\alpha_{h,\ell}$.

Proof. We apply the self-normalized martingale inequality of (Abbasi-Yadkori et al., 2011, Theorem 2) to the Gram matrix $V^{\gamma_{\text{decay}}}_{h,\ell}$. At round (h,ℓ), the noise involved in the reward signal $\eta_{h,\ell}$ is $\sigma^2_{a,h,\ell}$-sub-Gaussian and $\mathcal{F}_{h,\ell}$-measurable. The weighted process $\sum_{\ell=1}^{L}\gamma^\ell\sum_{h=1}^{H}\eta_{h,\ell}\phi_{A_{h,\ell},h,\ell}$ is a martingale, and the theorem yields the stated bound round-by-round. Thus, a union bound over all (h,ℓ) gives $\mathbb{P}(\mathcal{G}_1) \geq 1 - \sum_{\ell=1}^{L}\sum_{h=1}^{H}\alpha_{h,\ell}$. □

B.2 Regret Decomposition

Definition 1 (Constraint-violating rounds)**.** *Let $\mathcal{V}_{h,\ell}(A_{h,\ell}) \triangleq \left\{\langle\theta, \phi_{A_{h,\ell},h,\ell}\rangle < \tau\right\}$ denotes whether (h,ℓ) is a true constraint-violating index.*

We should note, as we are normalising each feature by the noise variance, the gaps are also inherently scaled by the noise variance. Thus, we define $\Delta_{A_{h,\ell},h,\ell} = \sigma_{A_{h,\ell},h,\ell}\underline{\Delta}_{A_{h,\ell},h,\ell}$, $\hat{\Delta}_{A_{h,\ell},h,\ell} = \sigma_{A_{h,\ell},h,\ell}\hat{\underline{\Delta}}_{A_{h,\ell},h,\ell}$, and finally $\underline{\Delta}_{\boldsymbol{\pi}_0}\sigma_{\min} \leq \Delta_{\boldsymbol{\pi}_0} \leq \underline{\Delta}_{\boldsymbol{\pi}_0}\sigma_{\max}$. We decompose the total expected regret as–

$$\begin{aligned}
\text{Reg} &= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}\Delta_{A_{h,\ell},h,\ell}\right] \\
&= \sigma_{\max}\,\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}\underline{\Delta}_{A_{h,\ell},h,\ell}\mathbb{1}\left\{\overline{\mathcal{V}_{h,\ell}(A_{h,\ell})}\right\}\right] + \sigma_{\max}\,\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}\underline{\Delta}_{A_{h,\ell},h,\ell}\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\right] \\
&\triangleq \text{Reg}^{\text{safe}} + \text{Reg}^{\text{viol}}
\end{aligned}$$

B.3 Part I: Regret Upper Bound under Constraint Violation

Theorem 2. *Algorithm 2 and 3 jointly satisfies –*

$$\text{Reg}^{\text{viol}} = \tilde{\mathcal{O}}\left(\sigma^2_{\max}\frac{B - \Delta_{\boldsymbol{\pi}_0}}{(\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)^2}d^2\left(\log^2(LH) + \log\left(\sum_{\ell=1}^{L}(\gamma_{\text{decay}})^{L-\ell}\sum_{h=1}^{H}\sigma^2_{h,\ell}\right)\right)\right)$$

In this section, we provide a complete proof of the regret upper bound under constraint violation, i.e, upper bound on Reg^{viol}. The proof structure involves two steps:

1. Then for the upper bound, we first define the conditioning event that will be used to decompose the regret expression further.
2. We bound each part of the decomposition to get a final accumulated upper bound.

B.3.1 Defining the Conditioning Events

Definition 2. *For any arm $a \in \mathcal{A}$, we define the following conditioning events:*

$$\begin{aligned}
&\textit{Concentration of gaps: } \mathcal{U}_{h,\ell}(a) \triangleq \left\{ \hat{\underline{\Delta}}_{A_{h,\ell},h,\ell} \geq \frac{\underline{\Delta}_{A_{h,\ell},h,\ell}}{1+c} \right\} \qquad \textit{for } c \geq 0 \\
&\textit{Saturation of arm features: } \mathcal{E}_{h,\ell} \triangleq \{|B_{h,\ell}| > 0\}, \textit{where } B_{h,\ell} = \left\{ a : \|\phi_{a,h,\ell}\|^2_{(V_{h,\ell}^{\gamma\text{decay}})^{-1}} > 1 \right\} \\
&\textit{Concentration of parameter estimate: } \mathcal{J}_{h,\ell}(a) \triangleq \left\{ \theta^\top \phi_{a^*_{h,\ell},h,\ell} - \hat{\theta}_{h,\ell}^\top \phi_{a^*_{h,\ell},h,\ell} \leq \varepsilon_1 \right\} \quad \textit{for } \varepsilon_1 \geq 0 \\
&\textit{Partial Saturation of arm features: } \mathcal{F}_{h,\ell}(a) \triangleq \left\{ \|\phi_{a,h,\ell}\|_{(V_{h,\ell}^{\gamma\text{decay}})^{-1}} > \varepsilon_2 \right\} \qquad \textit{for } \varepsilon \geq 0
\end{aligned} \tag{11}$$

The event $\mathcal{U}_{h,\ell}(a)$ signifies how accurately we are being able to estimate the sub-optimality gaps for arm a, $\mathcal{E}_{h,\ell}$ measures the saturation level for every arm at index (h,ℓ), $\mathcal{J}_{h,\ell}(a)$ assigns a precision in estimating the true parameter θ, and finally $F_{h,\ell}$ sets a lower bound on the saturation (number of pulls) level of arm a.

B.3.2 Decomposition of $\mathrm{Reg}^{\mathrm{viol}}$

We rewrite $\underline{\Delta}_{A_{h,\ell},h,\ell} = \underline{\Delta}_{\boldsymbol{\pi}_0} + (\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})$ for $(h,\ell) \in \mathcal{V}$. Thus, we have:

$$\begin{aligned}
\mathrm{Reg}^{\mathrm{constr}} =& \sigma_{\max} \mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \underline{\Delta}_{A_{h,\ell},h,\ell} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \right] \\
=& \sigma_{\max} \underline{\Delta}_{\boldsymbol{\pi}_0} \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \right]}_{\triangleq A_1} + \sigma_{\max} \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} (\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0}) \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \right]}_{\triangleq A_2}
\end{aligned} \tag{12}$$

Both the terms in this decomposition are non-negative as Lemma 5 yields a natural lower bound on the regret due to constraint-violation $\mathrm{Reg}^{\mathrm{constr}} \geq \Delta_{\boldsymbol{\pi}_0} \mathbb{E}\Big[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \Big]$.

Step 1: Upper bound on A_1. We further decompose A_1 using the conditioning event $\mathcal{U}_{h,\ell}(A_{h,\ell})$ as below:

$$\begin{aligned}
A_1 =& \mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \right] \\
\leq& \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \mathbb{1}\{ \mathcal{U}_{h,\ell}(A_{h,\ell}) \} \right]}_{\triangleq B_1} + \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \mathbb{1}\{ \overline{\mathcal{U}_{h,\ell}(A_{h,\ell})} \} \right]}_{\triangleq B_2}
\end{aligned} \tag{13}$$

Upper bound on B_1. We first divide the term B_1 in two parts: one under the good event $\mathcal{G}_1$ and when $\mathcal{G}_1$ is false. We denote them by $B_1 \mid_{\mathcal{G}_1}$ and $B_1 \mid_{\overline{\mathcal{G}_1}}$, respectively.

First, we bound $B_1 \mid_{\mathcal{G}_1}$ as follows:

$$B_1 \mid_{\mathcal{G}_1} = \mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{h=1}^{H} \mathbb{1}\left\{ \mathcal{V}_{h,\ell}(A_{h,\ell}) \right\} \mathbb{1}\{ \mathcal{U}_{h,\ell}(A_{h,\ell}) \} \right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\underline{\Delta}_{A_{h,\ell},h,\ell} > \underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0\right\} \mathbb{1}\left\{\hat{\Delta}_{A_{h,\ell},h,\ell} \geq \frac{\underline{\Delta}_{A_{h,\ell},h,\ell}}{1+c}\right\}\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\hat{\Delta}_{A_{h,\ell},h,\ell} > \frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c}\right\}\right] \\
&\overset{(a)}{\leq} \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\hat{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{A_{h,\ell},h,\ell} > \frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c} - B\right\}\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{(\theta - \hat{\theta}_{h,\ell})^\top(\phi_{A_{h,\ell},h,\ell} - \phi_{a^*_{h,\ell},h,\ell}) > \frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c} - B\right\}\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\beta_{h,\ell}(\alpha_{h,\ell})^{1/2}\|\phi_{A_{h,\ell},h,\ell} - \phi_{a^*_{h,\ell},h,\ell}\|_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} > \frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c} - B\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} + \|\phi_{a^*_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}}\right.\right. \\
&\qquad\qquad \left.\left. > \quad \frac{1}{\beta_{h,\ell}(\alpha_{h,\ell})}\left(\frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c} - B\right)^2\right\}\right] \\
&\overset{(c)}{\leq} 2\,\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} > \frac{1}{2\beta_{1,1}(\alpha_{1,1})}\left(\frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c} - B\right)^2\right\}\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} > \frac{1}{2\lambda S_*^2}\left(\frac{\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0}{1+c} - B\right)^2\right\}\right],
\end{aligned}
$$

where (a) holds due to the Assumption 3, and the fact that

$$
\begin{aligned}
\left\{\theta^\top\phi_{A_{h,\ell},h,\ell} < \tau\right\} &= \left\{\theta^\top\phi_{A_{h,\ell},h,\ell} < (1-\epsilon)\theta^\top\phi_{\boldsymbol{\pi}_0}\right\} \\
&= \left\{\theta^\top\phi_{A_{h,\ell},h,\ell} - \theta^\top\phi_{a^*_{h,\ell},h,\ell} + \theta^\top\phi_{a^*_{h,\ell},h,\ell} - \theta^\top\phi_{\boldsymbol{\pi}_0} < -\epsilon\mu_0\right\} \\
&= \left\{-\underline{\Delta}_{A_{h,\ell},h,\ell} + \theta^\top\phi_{a^*_{h,\ell},h,\ell} - \theta^\top\phi_{\boldsymbol{\pi}_0} < -\epsilon\mu_0\right\} \\
&\subseteq \left\{-\underline{\Delta}_{A_{h,\ell},h,\ell} + \min_{\ell\in[L]}\theta^\top\phi_{a^*_{h,\ell},h,\ell} - \theta^\top\phi_{\boldsymbol{\pi}_0} < -\epsilon\mu_0\right\} \\
&= \left\{-\underline{\Delta}_{A_{h,\ell},h,\ell} + \underline{\Delta}_{\boldsymbol{\pi}_0} < -\epsilon\mu_0\right\} \\
&= \left\{\underline{\Delta}_{A_{h,\ell},h,\ell} > \underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0\right\},
\end{aligned}
$$

(b) holds due to the definition of good event $\mathcal{G}_1$ in Equation (10), and finally (c) holds because we apply Boole's inequality, $\forall h, \ell \geq 1, \beta_{h,\ell}(\alpha_{h,\ell}) \geq \beta_{1,1}(\alpha_{1,1})$.

It is easy to see if we put $\lambda = \frac{2S_*^2}{\min\left\{\left(\frac{\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0}{1+c} - B\right)^2, \left(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0 - \frac{B}{1+c}\right)^2\right\}}$, then

$$
B_1 \mid_{\mathcal{G}_1} = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} > 1\right\}\right]
$$

Leveraging the elliptical potential count lemma (Lemma 7) and Lemma 6, we get the final bound on B_1 as:

$$
B_1 \leq B_1 \mid_{\mathcal{G}_1} + B_1 \mid_{\overline{\mathcal{G}_1}} \leq 6d\log\left(1 + \frac{2}{\lambda}\right) + \sum_{\ell=1}^{L}\sum_{h=1}^{H}\alpha_{h,\ell}
$$

Upper Bound on B_2. We again divide the second term B_2 in two parts: one under the good event $\mathcal{G}_1$ and otherwise. We similarly denote them by $B_2 \mid_{\mathcal{G}_1}$ and $B_2 \mid_{\overline{\mathcal{G}_1}}$, respectively.

First, we bound $B_2 \mid_{\mathcal{G}_1}$ as follows:

$$
\begin{aligned}
B_2 \mid_{\mathcal{G}_1} &= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\} \mathbb{1}\{\overline{\mathcal{U}_{h,\ell}(A_{h,\ell})}\}\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\underline{\Delta}_{A_{h,\ell},h,\ell} > \underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0\right\} \mathbb{1}\left\{\hat{\Delta}_{A_{h,\ell},h,\ell} < \frac{\underline{\Delta}_{A_{h,\ell},h,\ell}}{1+c}\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\hat{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{A_{h,\ell},h,\ell} \leq \frac{\underline{\Delta}_{A_{h,\ell},h,\ell}}{1+c} - \underline{\Delta}_{\boldsymbol{\pi}_0} - \epsilon\mu_0\right\}\right] \\
&\leq \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\hat{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{A_{h,\ell},h,\ell} \leq \frac{B}{1+c} - \underline{\Delta}_{\boldsymbol{\pi}_0} - \epsilon\mu_0\right\}\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\underline{\Delta}_{A_{h,\ell},h,\ell} - \hat{\Delta}_{A_{h,\ell},h,\ell} \geq \underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0 - \frac{B}{1+c}\right\}\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} \geq \frac{1}{2\beta_{h,\ell}(\alpha_{h,\ell})}\left(\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0 - \frac{B}{1+c}\right)^2\right\}\right] \\
&\leq 2\,\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} \geq \frac{1}{\lambda S_*^2}\left(\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0 - \frac{B}{1+c}\right)^2\right\}\right]
\end{aligned}
$$

Similar to the analysis B_1, we use Lemma 7 with $\lambda = \frac{2S_*^2}{\min\left\{\left(\frac{\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0}{1+c} - B\right)^2, \left(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0 - \frac{B}{1+c}\right)^2\right\}}$ to get the final bound on B_2 as:

$$
B_2 \leq 6d\log\left(1+\frac{2}{\lambda}\right) + \sum_{\ell=1}^{L}\sum_{h=1}^{H}\alpha_{h,\ell}\,.
$$

Step 2: Upper Bound on A_2.

To find upper bound on A_2, we closely follow the proof structure of the standard regret in (Balagopalan & Jun, 2024, Lemma 1 and Theorem 1) but remove the need of the peeling event. Balagopalan & Jun (2024) leverage peeling to lower bound the gap $\Delta_{A_{h,\ell},h,\ell}$, since it appears in the denominator of the pulling probability of arm $A_{h,\ell}$. We get the lower bound due to the presence of $\boldsymbol{\pi}_0$.

We use the event $\mathcal{E}_{h,\ell}$ to decompose A_2 as:

$$
\begin{aligned}
A_2 &= \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\right] \\
&\leq \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\mathcal{E}_{h,\ell}\right\}\right]}_{\triangleq C_1} \\
&\quad + \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\right]}_{\triangleq C_2}
\end{aligned}
$$

Upper bound on C_1. Now, we focus on the term C_1:

$$\begin{aligned} C_1 =& \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\mathcal{E}_{h,\ell}\right\}\right] \\ =& \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\underline{\Delta}_{A_{h,\ell},h,\ell} > \underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0\right\}\mathbb{1}\left\{\mathcal{E}_{h,\ell}\right\}\right] \end{aligned}$$

As we assign the probability of pulling $A_{h,\ell}$ from $B_{h,\ell}$ as exactly $\frac{1}{2}$, upper bound analysis of term C_1 closely matches with the analysis for term A_2 in (Balagopalan & Jun, 2024). Thus, we have–

$$C_1 \leq 6(B - \underline{\Delta}_{\boldsymbol{\pi}_0})d\log\left(1 + \frac{2}{\lambda}\right) \leq \frac{6(B - \Delta_{\boldsymbol{\pi}_0})d}{\sigma_{\max}}\log\left(1 + \frac{2}{\lambda}\right)$$

Upper bound on C_2. We further decompose C_2 based on the event based on the event $\mathcal{F}_{h,\ell}(A_{h,\ell})$–

$$\begin{aligned} C_2 =& \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0}\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\right] \\ \leq& \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\mathcal{F}_{h,\ell}(A_{h,\ell})\right\}\right]}_{\triangleq D_1} \\ &+ \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\right]}_{\triangleq D_2} \end{aligned}$$

Upper bound on D_1. Now, we focus on the term D_1–

$$D_1 = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\mathcal{F}_{h,\ell}(A_{h,\ell})\right\}\right]$$

To upper bounding this term, we follow the upper bound formalisation of Term D_2 in (Balagopalan & Jun, 2024) without the peeling argument to get–

$$\begin{aligned} D_1 \leq& 3d\frac{(B - \underline{\Delta}_{\boldsymbol{\pi}_0})}{\varepsilon_2}\log\left(1 + \frac{2}{\lambda\varepsilon_2}\right) + 3(B - \underline{\Delta}_{\boldsymbol{\pi}_0})d\log\left(1 + \frac{2}{\lambda}\right) \\ \leq& 3d\frac{(B - \Delta_{\boldsymbol{\pi}_0})}{\sigma_{\max}\varepsilon_2}\log\left(1 + \frac{2}{\lambda\varepsilon_2}\right) + \frac{3(B - \Delta_{\boldsymbol{\pi}_0})d}{\sigma_{\max}}\log\left(1 + \frac{2}{\lambda}\right) \end{aligned}$$

Upper bound on D_2. To upper bound this term, we again decompose it by leveraging the event $\mathcal{U}_{h,\ell}(A_{h,\ell})$–

$$D_2 = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell} - \underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\right]$$

$$\leq \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\mathcal{U}_{h,\ell}(A_{h,\ell})\right\}\right]}_{\triangleq E_1}$$

$$\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\overline{\mathcal{U}_{h,\ell}(A_{h,\ell})}\right\}\right]}_{\triangleq E_2}$$

Upper bound on E_1. This upper bound can be obtained following the steps in (Balagopalan & Jun, 2024) for their term F_1 ignoring the peeling steps. Thus, we have–

$$\begin{aligned}
E_1 \leq & \frac{1}{\alpha_{\text{emp}}}(B-\underline{\Delta}_{\boldsymbol{\pi}_0})HL\exp\left(-\frac{(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}{4\varepsilon_2\beta_{h,\ell}(\alpha_{h,\ell})(1+c)^2}\right) \\
& + \frac{3d(B-\underline{\Delta}_{\boldsymbol{\pi}_0})}{\alpha_{\text{emp}}^2}\left(\frac{1}{\varepsilon_2}\log\left(1+\frac{2}{\lambda\varepsilon_2}\right)+\log\left(1+\frac{2}{\lambda}\right)\right) \\
\leq & \frac{1}{\sigma_{\max}\alpha_{\text{emp}}}(B-\Delta_{\boldsymbol{\pi}_0})HL\exp\left(-\frac{(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}{4\sigma_{\max}^2\varepsilon_2\beta_{h,\ell}(\alpha_{h,\ell})(1+c)^2}\right) \\
& + \frac{3d(B-\Delta_{\boldsymbol{\pi}_0})}{\sigma_{\max}\alpha_{\text{emp}}^2}\left(\frac{1}{\varepsilon_2}\log\left(1+\frac{2}{\lambda\varepsilon_2}\right)+\log\left(1+\frac{2}{\lambda}\right)\right)
\end{aligned}$$

Upper bound on E_2. We finally use the event $\mathcal{J}_{h,\ell}(h,\ell)$ to decompose the event E_2 as–

$$E_2 = \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\overline{\mathcal{U}_{h,\ell}(A_{h,\ell})}\right\}\right]$$

$$\leq \underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\overline{\mathcal{U}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\mathcal{J}_{h,\ell}(A_{h,\ell})\right\}\right]}_{\triangleq F_1}$$

$$+\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\mathbb{1}\left\{\overline{\mathcal{E}_{h,\ell}}\right\}\mathbb{1}\left\{\overline{\mathcal{F}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\overline{\mathcal{U}_{h,\ell}(A_{h,\ell})}\right\}\mathbb{1}\left\{\overline{\mathcal{J}_{h,\ell}(A_{h,\ell})}\right\}\right]}_{\triangleq F_2}$$

Upper bound on F_1. It is easy to see that in F_1, if the good event in Equation (10) does not happen, then under $\overline{\mathcal{G}_1}$,

$$F_1\mid_{\overline{\mathcal{G}_1}} \leq (B-\underline{\Delta}_{\boldsymbol{\pi}_0})\sum_{h=1}^{H}\sum_{\ell=1}^{L}\alpha_{h,\ell} \leq\leq \frac{(B-\Delta_{\boldsymbol{\pi}_0})}{\sigma_{\max}}\sum_{h=1}^{H}\sum_{\ell=1}^{L}\alpha_{h,\ell}.$$

On the other hand, under the event $\mathcal{G}_1$, following (Balagopalan & Jun, 2024), we write

$$\begin{aligned}
F_1\mid_{\mathcal{G}_1} \leq & \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{(\hat{\theta}_{h,\ell}-\theta)^\top\phi_{A_{h,\ell},h,\ell} > \frac{c(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0)}{1+c}-\varepsilon_1\right\}\mathbb{1}\{\mathcal{G}_1\}\right] \\
\leq & \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} \geq \frac{1}{\beta_{h,\ell}(\alpha_{h,\ell})}\left(\frac{c(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0)}{2(1+c)}\right)^2\right\}\mathbb{1}\{\mathcal{G}_1\}\right] \\
\overset{(a)}{\leq} & \mathbb{E}\left[\sum_{\ell=1}^{L}\sum_{h=1}^{H}(\underline{\Delta}_{A_{h,\ell},h,\ell}-\underline{\Delta}_{\boldsymbol{\pi}_0})\mathbb{1}\left\{\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} \geq \frac{1}{\lambda S_*^2}\left(\frac{c(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0)}{2(1+c)}\right)^2\right\}\mathbb{1}\{\mathcal{G}_1\}\right]
\end{aligned}$$

$$\leq 3(B-\underline{\Delta}_{\boldsymbol{\pi}_0})d\log\left(1+\frac{2}{\lambda}\right) \leq \frac{3(B-\Delta_{\boldsymbol{\pi}_0})d}{\sigma_{\max}}\log\left(1+\frac{2}{\lambda}\right),$$

choosing $\varepsilon_1 = \frac{c(\underline{\Delta}_{\boldsymbol{\pi}_0}+\epsilon\mu_0)}{2(1+c)}$.

Thus, final upper on F_1 is given by,

$$F_1 \leq \frac{3(B-\Delta_{\boldsymbol{\pi}_0})d}{\sigma_{\max}}\log\left(1+\frac{2}{\lambda}\right) + \frac{(B-\Delta_{\boldsymbol{\pi}_0})}{\sigma_{\max}}\sum_{h=1}^{H}\sum_{\ell=1}^{L}\alpha_{h,\ell}$$

Bound on F_2. We follow the analysis of term F_3 of (Balagopalan & Jun, 2024) to state the final upper bound on F_2 with $H_{\max} = e$ in our case–

$$F_2 \leq \frac{128e(B-\Delta_{\boldsymbol{\pi}_0})C_{\text{opt}}d\log d}{\sigma_{\max}\alpha_{\text{opt}}\varepsilon_1^2}\left(\frac{\lambda S_*^2}{2} + d\log\left(1+\frac{\sum_{h=1}^{H}\sum_{\ell=1}^{L}\sigma_{h,\ell}^2}{d\lambda}\right)\right)$$

Hence, putting all the components in Equation (12), we get the final bound on $\text{Reg}^{\text{constr}}$ as:

$$\begin{aligned}
\text{Reg}^{\text{constr}} \leq & \frac{128e(B-\Delta_{\boldsymbol{\pi}_0})C_{\text{opt}}d\log d}{\alpha_{\text{opt}}\varepsilon_1^2}\left(\frac{\lambda S_*^2}{2} + d\log\left(1+\frac{\sum_{h=1}^{H}\sum_{\ell=1}^{L}\sigma_{h,\ell}^2}{d\lambda}\right)\right) \\
& + 12(B-\Delta_{\boldsymbol{\pi}_0})d\log\left(1+\frac{2}{\lambda}\right) + (B-\Delta_{\boldsymbol{\pi}_0})\sum_{h=1}^{H}\sum_{\ell=1}^{L}\alpha_{h,\ell} \\
& + \frac{1}{\alpha_{\text{emp}}}(B-\Delta_{\boldsymbol{\pi}_0})HL\exp\left(-\frac{(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}{4\sigma_{\min}^2\varepsilon_2\beta_{h,\ell}(\alpha_{h,\ell})(1+c)^2}\right) \\
& + \frac{3d(B-\Delta_{\boldsymbol{\pi}_0})}{\alpha_{\text{emp}}^2}\left(\frac{1}{\varepsilon_2}\log\left(1+\frac{2}{\lambda\varepsilon_2}\right) + \log\left(1+\frac{2}{\lambda}\right)\right) \\
& + 3d\frac{(B-\Delta_{\boldsymbol{\pi}_0})}{\varepsilon_2}\log\left(1+\frac{2}{\lambda\varepsilon_2}\right) \\
& + 12\Delta_{\boldsymbol{\pi}_0}d\log\left(1+\frac{2}{\lambda}\right) + 2\Delta_{\boldsymbol{\pi}_0}\sum_{\ell=1}^{L}\sum_{h=1}^{H}\alpha_{h,\ell} \\
& = \tilde{\mathcal{O}}\left(\frac{B-\Delta_{\boldsymbol{\pi}_0}}{(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}d^2\left(\log^2(LH) + \log\left(\sum_{h=1}^{H}\sum_{\ell=1}^{L}\sigma_{h,\ell}^2\right)\right)\right)
\end{aligned}$$

where the last inequality holds due to putting $\varepsilon_2 = \frac{(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}{4\sigma_{\max}^2\beta_{h,\ell}(\alpha_{h,\ell})\log(LH)}$, $c=1$, and $\alpha_{h,\ell} = \frac{1}{LH}$.
Recall, $\lambda = \frac{2S_*^2}{\sigma_{\max}\min\left\{\left(\frac{\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0}{2}-B\right)^2, \left(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0-\frac{B}{2}\right)^2\right\}}$.

Thus, Algorithm 2 and 3 jointly satisfies –

$$\text{Reg}^{\text{constr}} = \tilde{\mathcal{O}}\left(\frac{\sigma_{\max}^2(B-\Delta_{\boldsymbol{\pi}_0})}{(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}d^2\left(\log^2(LH) + \log\left(\sum_{h=1}^{H}\sum_{\ell=1}^{L}\sigma_{h,\ell}^2\right)\right)\right).$$

B.4 Part II: Constraint Satisfying Regret Upper Bound

Theorem 1. *Thus, Algorithm 2 and 3 jointly satisfies –*

$$\mathrm{Reg}^{\mathrm{safe}} = \tilde{\mathcal{O}}\left(\frac{d^2\left(\log^2(LH) + \log\left(\sum_{\ell=1}^{L}\left(\gamma_{\mathrm{decay}}\right)^{L-\ell}\sum_{h=1}^{H}\sigma_{h,\ell}^2\right)\right)}{\Delta_{\min}}\right)$$

Proof. In this section, we provide a complete proof of the regret upper bound under constraint satisfaction, i.e, upper bound on $\mathrm{Reg}^{\mathrm{opt}}$. The proof structure involves two steps:

1. Then for the upper bound, we first define the conditioning event that will be used to decompose the regret expression further down the line.
2. We bound each part of the decomposition step by step to get a final accumulated upper bound.

B.4.1 Step 1: Defining the Conditioning Events

To prove upper bound on $\mathrm{Reg}^{\mathrm{safe}}$, we recall the events defined in Part I (Definition 2) with additional peeling events.

$$\begin{aligned}
&\mathcal{U}_{h,\ell}(a) \triangleq \left\{\hat{\underline{\Delta}}_{A_{h,\ell},h,\ell} \geq \frac{\underline{\Delta}_{A_{h,\ell},h,\ell}}{1+c}\right\} \qquad \text{for some } c \geq 0\\
&\mathcal{E}_{h,\ell} \triangleq \{|B_{h,\ell}| > 0\}, \quad \text{where } B_{h,\ell} = \left\{a : \|\phi_{a,h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\mathrm{decay}}})^{-1}} > 1\right\}\\
&\mathcal{C}_{h,\ell,k}(A_{h,\ell}) \triangleq \left\{\|\phi_{a,h,\ell}\|_{(V_{h,\ell}^{\gamma_{\mathrm{decay}}})^{-1}} > \varepsilon_k\right\}\\
&\mathcal{D}_{h,\ell,k}(A_{h,\ell}) \triangleq \left\{(\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0)2^{-k} < \bar{\Delta}_{A_{h,\ell},h,\ell} < (\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0)2^{-k+1}\right\}\\
&\overline{\mathcal{D}_{h,\ell,K}(A_{h,\ell})} \triangleq \left\{\bar{\Delta}_{A_{h,\ell}} \leq (\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0)2^{-K}\right\}\\
&\mathcal{J}_{h,\ell}(a) \triangleq \left\{\theta^\top\phi_{a^*_{h,\ell},h,\ell} - \hat{\theta}_{h,\ell}^\top\phi_{a^*_{h,\ell},h,\ell} \leq \varepsilon_{2,k}\right\},
\end{aligned}$$

where we define $\bar{\Delta}_{A_{h,\ell},h,\ell} \triangleq \max\{\underline{\Delta}_{A_{h,\ell},h,\ell}, \underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0\}$.

B.4.2 Decomposition of $\mathrm{Reg}^{\mathrm{safe}}$

We follow the proof structure from (Balagopalan & Jun, 2024) for this part. We use similar peeling technique and decomposition of regret using conditioning events. As we have constraint satisficing in this case, we use $(\underline{\Delta}_{\boldsymbol{\pi}_0} + \epsilon\mu_0)$ as the upper bound for $\underline{\Delta}_{A_{h,\ell},h,\ell}$.

Final Regret Bound under Constraint Satisficing.

$$\begin{aligned}
\frac{1}{\sigma_{\max}}\mathrm{Reg}^{\mathrm{safe}} \leq\; & 18(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)d\log\left(1+\frac{2}{\lambda}\right)\\
&+ LH(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)2^{-P}\mathbb{1}\left\{(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)2^{-P} > \Delta_{\min}\right\}\\
&+ \frac{12d(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)}{2^{-P}\varepsilon}\log\left(1+\frac{2}{\lambda 2^{-2P}\varepsilon}\right)\\
&+ \frac{1}{\alpha_{\mathrm{emp}}}4LH(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)\exp\left(-\frac{(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)^2}{16\varepsilon\beta_{h,\ell}(\alpha_{h,\ell})}\right)\\
&+ \frac{1}{\alpha_{\mathrm{emp}}^2}\left(\frac{12d(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)}{2^{-P}\varepsilon}\log\left(1+\frac{2}{\lambda 2^{-2P}\varepsilon}\right)\right.\\
&\left.+ 6(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)d\log\left(1+\frac{2}{\lambda}\right)\right)\\
&+ 2(\sigma_{\max}\Delta_{\boldsymbol{\pi}_0} + \epsilon\mu_0)\log(LH)
\end{aligned}$$

$$+\frac{192\beta_{h,\ell}(\alpha_{h,\ell})d}{(\sigma_{\min}\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)2^{-P}}\log\left(1+\frac{32\beta_{h,\ell}(\alpha_{h,\ell}}{\lambda(\sigma_{\min}\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2 2^{-2P}}\right)$$
$$+\frac{512d\log(d)eC_{\text{opt}}}{\alpha_{\text{opt}}(\sigma_{\min}\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)2^{-P}}\left(\frac{\lambda S_*}{2}+d\log\left(1+\frac{\sum_{\ell=1}^{L}\sum_{h=1}^{H}\sigma^2_{a,h,\ell}}{d\lambda}\right)\right)$$

□

Putting $\epsilon = \frac{(\sigma_{\min}\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0)^2}{16\beta_{h,\ell}(\alpha_{h,\ell})}$

Algorithm 2 and 3 jointly satisfy–

$$\text{Reg}^{\text{safe}} = \tilde{\mathcal{O}}\left(\kappa\frac{d^2\left(\log^2(LH)+\log\left(\sum\sum_{h=1}^{H}\sigma^2_{a,h,\ell}\right)\right)}{\Delta_{\min}}\right)$$

where $\kappa \triangleq \frac{\max_{a,h,\ell}\sigma^2_{a,h,\ell}}{\min_{a,h,\ell}\sigma^2_{a,h,\ell}}$.

C Upper Bound on Expected Constraint Violation

In this section, we prove the upper bound guaranty on the expected number of constraint violation of Dri-MED.

Theorem 4. *Dri-MED enjoys*

$$\textit{Violation}(\boldsymbol{\pi}_0) = \tilde{\mathcal{O}}(d)\ ,$$

Proof. We start from the definition of expected constraint violation–

$$\text{Violation}(\boldsymbol{\pi}_0) = \mathbb{E}\left[\mathbb{1}\left\{\mathcal{V}_{h,\ell}(A_{h,\ell})\right\}\right]$$

From the standard regret analysis in Section B, we know $\text{Violation}(\boldsymbol{\pi}_0) = A_1$.

Thus, setting $\lambda = \frac{2S_*^2}{\min\left\{\left(\frac{\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0}{2}-B\right)^2,\left(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0-\frac{B}{2}\right)^2\right\}}$ to get the final bound on B_2 as:

$$\text{Violation}(\boldsymbol{\pi}_0) \leq 12d\log\left(1+\frac{\min\left\{\left(\frac{\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0}{2}-B\right)^2,\left(\Delta_{\boldsymbol{\pi}_0}+\epsilon\mu_0-\frac{B}{2}\right)^2\right\}}{S_*}\right)+2 = \tilde{\mathcal{O}}(d)\,.$$

□

D Discussion on the Lagrangian Dual Variable

Lyapunov Drift for the Dual Variable. Recall the dual update with $\rho \triangleq \frac{\epsilon}{(1-\epsilon)}$ in Algorithm 2:

$$\nu_{h+1,\ell} = \max\{0,\ \nu_{h,\ell} + \eta_h(\xi_{h,\ell} - \rho)\}, \qquad \xi_{h,\ell} := \mathbb{1}\left[\langle \hat{\theta}_{h,\ell},\, \phi_{A_{h,\ell},h}\rangle < \tau_\ell\right], \tag{14}$$

with $\eta_{h,\ell} = \eta/\sqrt{h}$, $\eta > 0$ to be chosen. Note, $\xi_{h,\ell} \in \{0,1\}$ regardless of the reward distribution. Thus, the dual update depends only on whether the *estimated* reward crosses the threshold, not on the magnitude of the reward itself.

Lemma 2 (Lyapunov drift inequality)**.** *For all $t \geq 1$:*

$$\mathbb{E}[\nu_{h+1,\ell}^2 - \nu_{h,\ell}^2 \mid \mathcal{F}_t] \ \leq\ 2\eta_{h,\ell}\big(\mathbb{E}[\xi_{h,\ell} \mid \mathcal{F}_{h,\ell}] - \rho\big)\nu_{h,\ell} + \eta_{h,\ell}^2.$$

Proof. Using $\max\{0,x\}^2 \leq x^2$:

$$\nu_{h+1,\ell}^2 \leq \big(\nu_{h,\ell} + \eta_{h,\ell}(\xi_{h,\ell} - \rho)\big)^2 = \nu_{h,\ell}^2 + 2\eta_{h,\ell}(\xi_{h,\ell} - \rho)\nu_{h,\ell} + \eta_{h,\ell}^2(\xi_{h,\ell} - \rho)^2.$$

Since $\xi_{h,\ell} \in \{0,1\}$, we have $(\xi_{h,\ell} - \rho)^2 \leq 1$. Taking conditional expectation gives the result. □

Lemma 3 (Dual variable growth)**.** *With $\nu_0 = 0$ and $\eta_{h,\ell} = \eta/\sqrt{\ell h}$, we have $\nu_{h,\ell} \leq 2\eta\sqrt{\ell h}$.*

Proof. Since $\xi_{h,\ell} \leq 1$ we have $\nu_{h+1,\ell} \leq \nu_{h,\ell} + \eta_{h,\ell}$, so $\nu_{h,\ell} \leq \sum_{\ell=1}^{L}\sum_{h=1}^{H} \eta/\sqrt{s} \leq 2\eta\sqrt{LH}$. □

The dual update in Equation (14) is exactly projected online gradient descent (OGD) on the sequence of losses $\ell_{h,\ell}(\nu) \triangleq -\nu(\xi_{h,\ell} - \rho)$ over $\nu \in [0,\infty)$. The gradient at $\nu_{h,\ell}$ is $-(\xi_{h,\ell} - \rho)$, which lies in $(-1,1]$ since $\xi_{h,\ell} \in \{0,1\}$ and $\rho \in [0,1)$.

Lemma 4 (OGD regret bound)**.** *For any $\nu^* \geq 0$:*

$$\sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathcal{L}_t(\nu_{h,\ell}) - \sum_{\ell=1}^{L}\sum_{h=1}^{H} \mathcal{L}_t(\nu^*) \ \leq\ \frac{\nu^{*2}}{2\eta_{LH}} + \eta\sqrt{LH}.$$

Proof. The loss $\mathcal{L}_{h,\ell}(\nu) = -\nu(\xi_{h,\ell} - \rho)$ is linear in ν, so it is convex, and its gradient is $\nabla_t = \nabla_\nu \mathcal{L}_{h,\ell}(\nu_{h,\ell}) = -(\xi_{h,\ell} - \rho)$. Since $\xi_{h,\ell} \in \{0,1\}$ and $\rho \in [0,1)$, we have $\nabla_{h,\ell} \in [-(1-\rho), \rho] \subset (-1,1]$, so $\nabla_{h,\ell}^2 \leq 1$. We use (Hazan, 2016, Theorem 3.1) with $\nu_{1,1} = 0$ to directly get the desired result. □

E Dri-IMED: Drift Adaptive Indexed Minimum Empirical Divergence

For empirical evaluation, in this section we provide the implemented pseudocode of the `Dri-IMED` algorithm, that is an Indexed Minimum Empirical Divergence version of MED strategy used in this paper. The main difference from Algorithm 3 is in Line 4 of `Dri-IMED`. It computes the Lagrangian dual-augmented IMED index being optimistic about the action, but pessimistic with respect to the constraint. Whenever a pulled arm violates the constraint the IMED index is penalised. At the end it plays the arm that has the minimum index.

Algorithm 4 `Dri-IMED`: **Dri**ft adaptive **I**ndexed **M**inimum **E**mpirical **D**ivergence (IMED) Step

Require: Episode $\ell \in [L]$, user id $h \in [H]$, user's rescaled features $\{\phi_{a,h,\ell}\}_{a=1}^{|\mathcal{A}|}$, $\hat{\theta}_{h,\ell}$, $V_{h,\ell}^{\gamma_{\text{decay}}}$, λ.
Require: Constraint threshold τ_ℓ, Lagrangian multiplier ν_ℓ.

1: **Compute confidence radius and upper and lower confidence bound:** For each arm $a \in \mathcal{A}$,

$$\beta_{h,\ell}(\alpha_{h,\ell}) = \left(\sqrt{\log \frac{\det V_{h,\ell}^{\gamma_{\text{decay}}}}{\det V_0}} + 2 \log \frac{1}{\alpha_{h,\ell}} + \sqrt{\lambda} S \right)^2$$

$$\text{UCB}(a,h,\ell) = \langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell} \rangle + \sqrt{\beta_{h,\ell}(\alpha_{h,\ell})} \|\phi_{a,h,\ell}\|_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}}$$

$$\text{LCB}(a,h,\ell) = \langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell} \rangle - \sqrt{\beta_{h,\ell}(\alpha_{h,\ell})} \|\phi_{a,h,\ell}\|_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}}$$

2: Empirical best action: $\hat{a}_{h,\ell} = \arg\max_{a \in \mathcal{A}} \langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell} \rangle$

3: Maximum empirical reward: $\hat{\mu}_{\max}(h,\ell) = \max_{a \in \mathcal{A}} \langle \hat{\theta}_{h,\ell}, \phi_{a,h,\ell} \rangle$

4: **Lagrangian-augmented IMED index:** For each arm $a \in \mathcal{A}$:

$$I_{h,\ell}^{\text{aug}}(a) = \begin{cases} \nu_\ell \cdot \max\{0, \tau_\ell - \text{LCB}(a,h,\ell)\} & \text{if } a = \hat{a}_{h,\ell} \\ N_{h,\ell}(a) \cdot \frac{(\text{UCB}(a,h,\ell) - \hat{\mu}_{\max}(h,\ell))^2}{2\sigma_{h,\ell}^2} & \\ \quad + \nu_\ell \cdot \max\{0, \tau_\ell - \text{LCB}(a,h,\ell)\} & \text{if } \text{UCB}(a,h,\ell) \geq \mu_{\max}(h,\ell) \end{cases}$$

5: Pull $A_{h,\ell} = \arg\min_{a \in \mathcal{A}} I_{h,\ell}^{\text{aug}}(a)$

F Experimental Analysis

To assess the performance of `Dri-MED` and `Dri-IMED`, we conduct numerical experiments on a synthetic episodic contextual linear bandit with preference feedback and context drift. The environment consists of H users, A arms, and L episodes, with a d-dimensional parameter θ^* and M-dimensional feedback signals. The mean feedback signal $\mathbb{E}[Y_{h,\ell} \mid a] = \Phi(a)\theta^*$ is context-independent, while the covariance scales with the observed context norm, inducing heteroscedastic noise. User preference vectors $\{p_h\}$ are drawn from a Dirichlet distribution and held fixed across episodes.

Setup. The full environment generation is detailed in Algorithm 5. The baseline policy π_0 is set as the arm at the median performance quantile for each user, i.e. $\pi_0(h) = \arg\min_a |\phi_{a,h}^\top \theta^\star - q_{0.5}^h|$ where $q_{0.5}^h$ is the median of $\{\phi_{a,h}^\top \theta^\star\}_{a\in[A]}$. We use the following environment parameters throughout all experiments: feature dimension $d = 4$, feedback dimension $M = 4$, number of arms $A = 5$, number of users $H = 10$, number of episodes $L = 1000$, parameter norm $B = 1$, preference scale $s = 2$, context dimension $D = 2$, context noise $\sigma_C = 1$, and reward noise $\sigma_r = 0.1$. Results that are presented in Figure 2 are averaged over 128 independent seeds.

Algorithm 5 Environment Generation

Require: $d, M, A, H, L, f, \sigma_C, \sigma_r, s, B$

Draw $\theta^\star \sim \mathcal{N}(0, I_d)$, normalise to $\|\theta^\star\| = B$
Draw row-normalised $\Phi(a) \in \mathbb{R}^{M\times d}$ and $\Sigma_a = S_a S_a^\top / M$ for each $a \in [A]$
Draw $\omega_h \sim \mathrm{Dir}(s \cdot p_h)$ with $p_h \sim \mathrm{Dir}(\mathbf{1}_M)$ for each $h \in [H]$
Set $\kappa_\ell \leftarrow 1 + f(\ell, L)$ for $\ell \in [L]$
for $\ell = 1$ **to** L **do**
    Sample $C_{h,\ell} \sim \mathcal{N}(0,\, \sigma_C^2 \kappa_\ell^2\, I_D)$ for each $h \in [H]$
    Set $\phi_{a,h} \leftarrow \Phi(a)^\top \omega_h$ for all $a \in [A]$, $h \in [H]$
    Draw $Y_{h,\ell} \sim \mathcal{N}\big(\Phi(A_{h,\ell})\theta^\star,\ \sigma_r(1 + \|C_{h,\ell}\|)\, \Sigma_{A_{h,\ell}}\big)$, set $r_{h,\ell} \leftarrow \omega_h^\top Y_{h,\ell}$
end for

Drifting. We consider four drift regimes for the context distribution, parameterized by a magnitude $\kappa > 0$ and a shared drift direction $v \in \mathbb{R}^C$, $\|v\| = 1$, so that $\mu_\ell = f(\ell) \cdot v$:

- *No drift*: $f(\ell) = 0$,
- *Gradual drift*: $f(\ell) = \kappa \cdot \ell / L$,
- *Periodic drift*: $f(\ell) = \kappa \cdot \frac{1}{2}\left(1 + \sin\left(\frac{2\pi\ell}{L/50}\right)\right)$,
- *Abrupt drift*: $f(\ell) = \kappa \cdot \sum_i i \cdot \mathbf{1}\{\ell \geq \tau_i\}$, where $\tau_1 = \lfloor L/3 \rfloor$ and $\tau_2 = \lfloor 2L/3 \rfloor$ are fixed change-points.

Baselines. To the best of our knowledge, no prior algorithm addresses this combined setting of preference structure, heteroscedastic noise, and non-stationary context distributions; Since no existing algorithm is designed for this setting, we evaluate `Dri-MED` and `Dri-IMED` against four stationary linear bandit baselines that ignore both the drift and the preference structure: `OFUL` (Abbasi-Yadkori et al., 2011), `LinMED` (Balagopalan & Jun, 2024), `LinIMED` (Bian & Tan, 2024), and `LinTS` (Agrawal & Goyal, 2013). For all drift types, we set magnitude $\kappa = 100$, so the maximum context scale is $s_{\max} = 1 + \kappa = 101$. The per-arm noise standard deviation is $\sigma_{a,h,\ell} = \sqrt{\omega_h^\top \Sigma_a \cdot \sigma_r \cdot (1 + \|C_{h,\ell}\|) \cdot \omega_h}$, which is bounded by

$$\bar{\sigma} = \sqrt{\sigma_r \cdot (1 + \sigma_c \cdot s_{\max} \cdot \sqrt{D}) \cdot \max_{a,i} [\Sigma_a]_{ii}}$$

where $\sigma_r = 0.1$ is the reward noise, $\sigma_c = 1.0$ the context noise scale, $D = 2$ the context dimension, and $\max_{a,i}[\Sigma_a]_{ii}$ the largest diagonal entry across all arm covariance matrices (evaluated once at initialisation). In our experiments this yields $\bar{\sigma} \approx 1.92$, which we use as the sub-Gaussianity parameter for all algorithms. We report cumulative regret and cumulative satisficing constraint violations, averaged over independent random seeds. All specific hyperparameters used by the algorithm are summarized in Table 3

Table 3: Hyperparameters used per algorithm. Parameters shared by all algorithms: regularisation $\lambda = \bar{\sigma}^2/S^2$, confidence $\delta = 0.01$, $S = 1$.

Algorithm	γ_{decay}	ε	ν_0	α_{opt}	α_{emp}	C
OFUL	–	–	–	–	–	–
LinTS	–	–	–	–	–	–
LinMED	–	–	–	0.99	0.005	–
LinIMED	–	–	–	–	–	30
Dri-MED	0.99	0.1	1.0	0.99	0.005	–
Dri-IMED	0.99	0.1	1.0	–	–	–

Results. Figure 2 report cumulatieve regret and constraint violation rates across all drift regimes. Dri-MED and and Dri-IMED consistently achieve very low cumulative regret across all settings, outperforming all stationary baselines by a large margin. Among baselines, OFUL performs best but still incurs regret an order of magnitude larger than our methods; LinMED and LinTS perform worst, highlighting the cost of ignoring the preference and heteroscedastic structure. Importantly, both Dri-MED and Dri-IMED maintain near-zero true constraint violation throughout all episodes, whereas baselines show persistent violations. The results are stable across drift regimes: even under abrupt drift with magnitude $\kappa = 100$, the performance of our methods is largely unaffected, confirming that the drift-adaptive design successfully absorbs the non-stationarity through the discounted regression and context-scaled normalisation. Figure 4 further illustrates that Dri-MED concentrates pulls on the oracle-optimal arm for every user, while LinMED spreads mass across suboptimal arms, and π_0 is always fixed at the median arm.

Ablations. Figure 3 examines the sensitivity of Dri-MED and Dri-IMED to the baseline quantile q and satisficing tolerance ε under abrupt drift. As q increases, the baseline policy π_0 becomes stronger, tightening the constraint and reducing the feasible arm set; this forces the algorithm to focus on higher-quality arms earlier, yielding lower final regret. Conversely, increasing ε loosens the constraint, allowing more exploration of suboptimal arms and increasing regret. Dri-IMED consistently achieves lower regret and tighter interquartile ranges than Dri-MED across all ablation settings, suggesting it is more sample-efficient under tighter constraints.

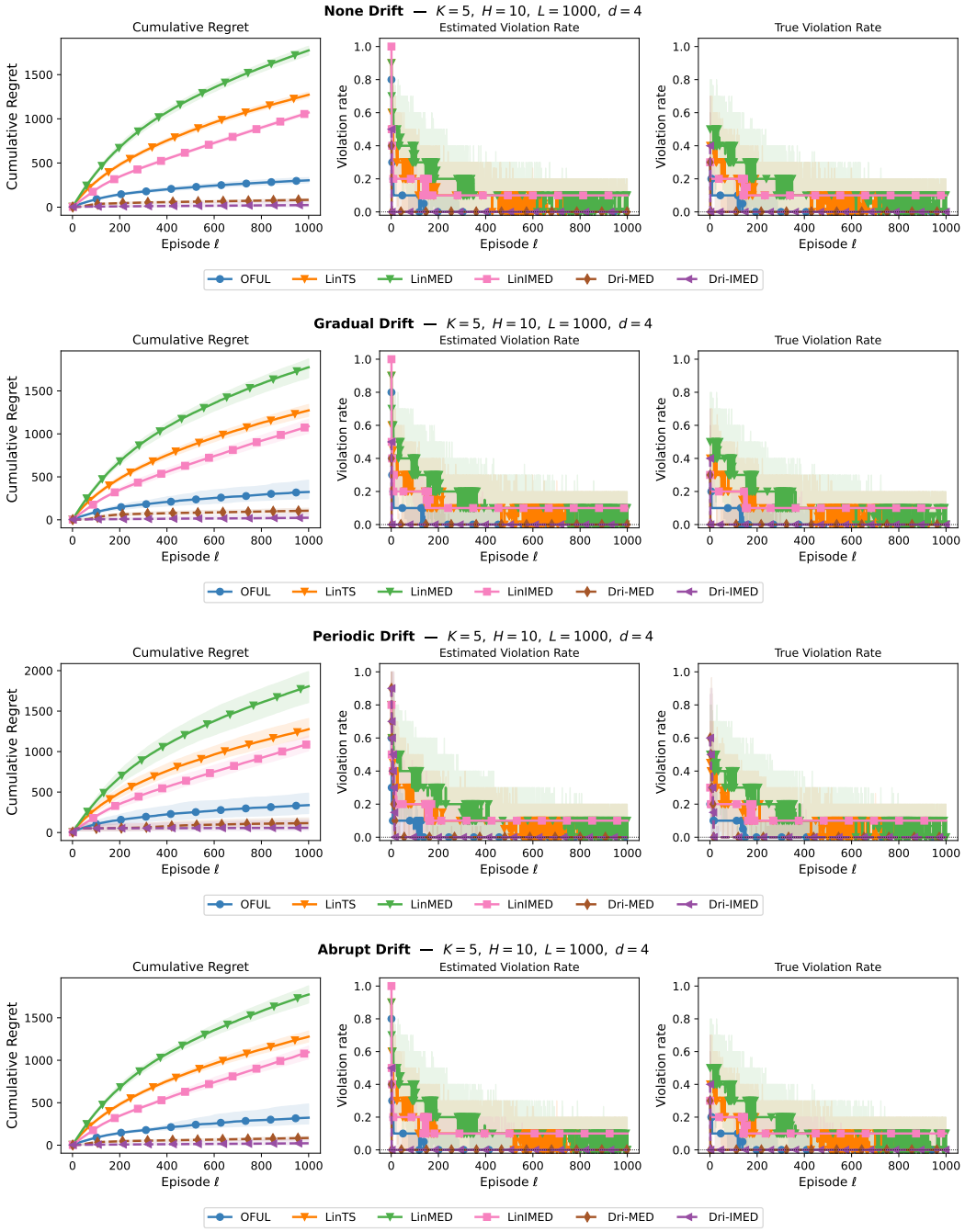

Figure 2: Cumulative regret (left), estimated violation rate (center), and true violation rate (right) for each drift regime. Shaded bands show the 5–95% quantile range over 128 seeds.

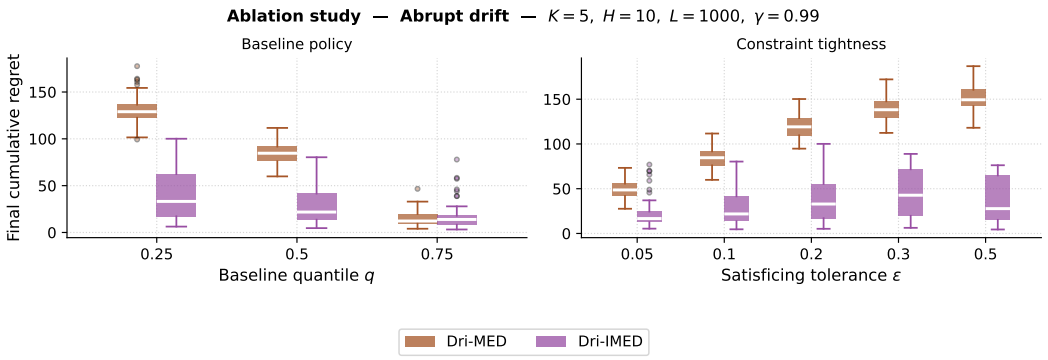

Figure 3: Ablation study on abrupt drift: effect of the baseline policy quantile q (left) and the satisficing tolerance ε (right) on the final cumulative regret of `Dri-MED` and `Dri-IMED`.

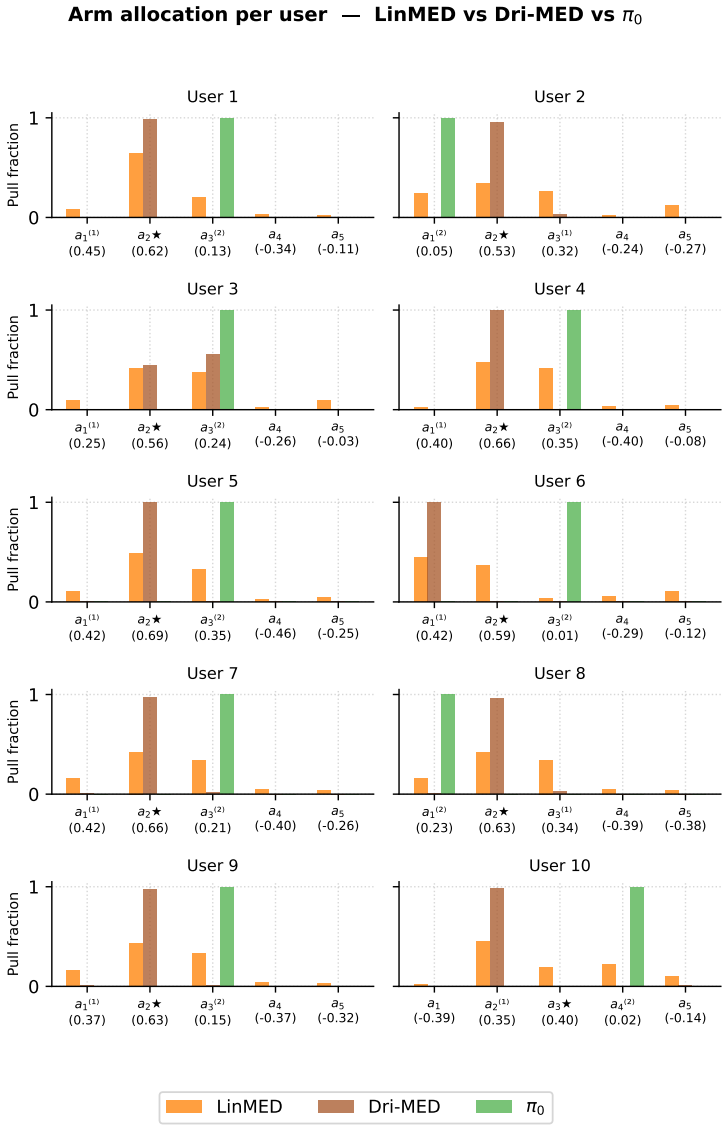

Figure 4: Arm pull fraction per user over 1000 episodes on abrupt drift. $\star$ denotes the oracle best arm; $^{(1)}$ and $^{(2)}$ the first and second suboptimal arms. Mean rewards are shown in parentheses.

G ApproxDesign

Algorithm 6 ApproxDesign

Require: Scaled arm matrix $\mathcal{A}^{(h)} = \{\tilde{\phi}_a\}_{a=1}^{|\mathcal{A}|}$

Project onto active subspace: compute PCA of $\mathcal{A}^{(h)}$, retain eigenvectors with eigenvalue $> \epsilon_0$, obtain low-dimensional matrix $\tilde{X}$

Initialise via volume approximation: select $2d'$ arms $\mathcal{S}_0 \subseteq \mathcal{A}$ spanning the subspace by greedily picking $\arg\max_a \langle \tilde{\phi}_a, b_i \rangle$ and $\arg\min_a \langle \tilde{\phi}_a, b_i \rangle$ along each basis direction b_i

Set $V \leftarrow \sum_{a \in \mathcal{S}_0} \tilde{\phi}_a \tilde{\phi}_a^\top$

while $\max_{a \in \mathcal{A}} \|\tilde{\phi}_a\|^2_{V^{-1}} > 1$ **do**

  $a^\star \leftarrow \arg\max_{a \in \mathcal{A}} \|\tilde{\phi}_a\|^2_{V^{-1}}$

  $V \leftarrow V + \tilde{\phi}_{a^\star} \tilde{\phi}_{a^\star}^\top, \quad \mathcal{S}_0 \leftarrow \mathcal{S}_0 \cup \{a^\star\}$

end while

$q_h^{\text{opt}}(a) \leftarrow |\{i : \mathcal{S}_0[i] = a\}| \;/\; |\mathcal{S}_0|$

return q_h^{opt}

By the virtue of design of `Dri-MED`, we retain the exact guaranty of eliminating highly suboptimal arms as same as (Balagopalan & Jun, 2024). By the virtue of design of `Dri-MED`, we retain every guaranty on arm saturation by scaling both α_{emp} and α_{opt} by $\exp(-\boldsymbol{\nu}_{h,\ell})$ per step. Intuitively, more the value of the Lagrangian dual, more unsafe is the arm. As the Lagrangian dual grows at a rate $\mathcal{O}(\sqrt{\ell h})$ (Lemma 3), it ensures a truly unsafe arm is eliminated while augmentation of the arm set. This is a novel adaptation to make the design constraint-aware.

We omit these proofs as they directly follow from (Balagopalan & Jun, 2024) and we do not want to remain repetitive.

H Useful Technical Results

Lemma 5 (Lower bound on gap for violating rounds)**.** *For any constraint violating index* (h, ℓ)*,* $\Delta_{A_{h,\ell},h,\ell} \geq \Delta_{\boldsymbol{\pi}_0}$.

Proof. The proof is straightforward. We decompose the gap as:

$$\begin{aligned}\Delta_{A_{h,\ell},h,\ell} &= \theta^\top(\phi_{a^*_\ell,h,\ell} - \phi_{A_{h,\ell},h,\ell}) \\ &= \theta^\top(\phi_{a^*_\ell,h,\ell} - \phi_{\pi_0}) + \underbrace{\theta^\top(\phi_{\pi_0} - \phi_{A_{h,\ell},h,\ell})}_{\geq 0} \qquad \text{, iff } (h,\ell) \in \mathcal{V} \\ &\geq \min_{\ell \in [L]} \theta^\top(\phi_{a^*_\ell,h,\ell} - \phi_{\pi_0}) = \Delta_{\boldsymbol{\pi}_0}\end{aligned}$$

□

Definition 3 (Good event at the end of ℓ-th episode)**.**

$$\mathcal{G}_2 \triangleq \left\{h = H, \forall \ell \geq 1 : \|\theta - \hat{\theta}_\ell\|^2_{V_\ell^{\gamma_{\text{decay}}}} \leq \beta_\ell(\delta_\ell)\right\} \tag{15}$$

Note, the index (H, ℓ) *denotes end of the episode* ℓ*. Thus for brevity, we remove the index* h *for the event* $\mathcal{G}_2$*.*

Lemma 6 (Heteroscedastic weighted (H, ℓ)-th confidence set)**.** *Following the information acquisition rule at the end of episode* $\ell \in [L]$*, i.e.,* $V_\ell^{\gamma_{\text{decay}}} = \gamma_{\text{decay}} V_{\ell-1}^{\gamma_{\text{decay}}}$*, we define*

$$\beta_\ell(\delta_\ell)^{1/2} \triangleq \sqrt{2\log\frac{\det V_{\ell-1}^{\gamma_{\text{decay}}}}{\det V_0} + 2\log\frac{1}{\delta_\ell}} + \sqrt{\lambda}\, S_*\,,$$

where $\delta_\ell \in (0,1)$ *is to characterised later on. Then* $\mathbb{P}(\mathcal{G}_2) \geq 1 - \sum_{\ell=1}^L \delta_\ell$.

Proof. For this proof, we again apply the self-normalized martingale inequality of (Abbasi-Yadkori et al., 2011, Theorem 2) to the discounted Gram matrix $V_\ell^{\gamma_{\text{decay}}}$. At round (H, ℓ), the accumulated noise over H reward signal $\eta_{H,\ell} = \sum_{h=1}^H \eta_{h,\ell}$ is $\sum_{h=1}^H \sigma^2_{A_{h,\ell},h,\ell}$-sub-Gaussian and $\mathcal{F}_\ell$-measurable. The weighted process $\sum_{\ell=1}^L \gamma^\ell \eta_{H,\ell}\bar{\phi}_\ell$ is a martingale, and the theorem yields the stated bound round-by-round. Thus, a union bound over all ℓ gives $\mathbb{P}(\mathcal{G}_2) \geq 1 - \sum_{\ell=1}^L \delta_\ell$. □

Lemma 7 (Elliptical Potential Count Lemma: Lemma C.2 of (Jun & Kim, 2024))**.** *Let* $x_1, x_2, \ldots, x_t \in \mathbb{R}^d$ *be a sequence of vectors with* $\|x_s\|_2 \leq 1, \forall s \in [t]$*. Let* $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$ *for some* $\lambda > 0$*. Let* $J = \left\{s \in [t] : \|x_s\|^2_{V_{s-1}^{-1}} \geq L^2\right\}$ *for some* $L^2 \leq 1$*. Then,*

$$|J| \leq 3\frac{d}{L^2}\ln\left(1 + \frac{2}{L^2\lambda}\right)$$

Lemma 8 (Elliptical Potential Lemma: Proposition 2 of (Abeille & Lazaric, 2017))**.** *Let* $x_1, x_2, \ldots, x_t \in \mathbb{R}^d$ *be a sequence of vectors with* $\|x_s\|_2 \leq 1, \forall s \in [t]$*. Let* $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$ *for some* $\lambda > 0$*. Then,*

$$\sum_{s=1}^t \|x_s\|^2_{V_s^{-1}} \leq 2d\log\left(1 + \frac{t}{d\lambda}\right).$$

Corollary 1 (Elliptical Potential Lemma for rescaled features)**.** *For* $\|\phi_{A_{h,\ell},h,\ell}\|2^2 \leq 1, \forall h, \ell ge 1$,

$$\|\phi_{A_{h,\ell},h,\ell}\|^2_{(V_{h,\ell}^{\gamma_{\text{decay}}})^{-1}} \leq 2d\log\left(1 + \frac{t}{d\lambda}\right)$$

Lemma 9 (OFUL confidence bound lemma adapted from Theorem 2 of (Abbasi-Yadkori et al., 2011)). *Assume* $\forall s \in [t], \quad \|a_s\| \leq 1$, *and* $\|\theta^*\|_2 \leq S$, *for some fixed* $\mathrm{S} > 0$. *We also assume* $\Delta_a := \max_{a' \in \mathcal{A}_t} \langle a, \theta^* \rangle - \langle a, \theta^* \rangle \leq 1$, $\forall a \in \mathcal{A}$

$$\forall t \geq 1, \quad \mathbb{P}\left(\left\| \hat{\theta}_{t-1} - \theta^* \right\|_{V_{t-1}} \leq \sqrt{\beta_{t-1}\left(\delta_{t-1}\right)} \right) \geq 1 - \delta.$$

Lemma 10 (Adapted from Lemma 3 in Balagopalan & Jun (2024)). *Let* $\tilde{f}_h(a), q_h$ *are defined as per Algorithm 3 where* $A_{h,\ell} \neq a$. *Then* $\forall h > 1$,

$$1 \geq \sum_{b \in B_h} q_h(b) f_t(b) \geq \alpha_{\text{emp}} e^{-\nu_{h,\ell}}$$

Lemma 11 (Adapted from Lemma 5 in (Balagopalan & Jun, 2024)). *For* `Dri-MED`, *we have–*

$$\left\| a^*_{h,\ell} \right\|^2_{(V(p_h)^{\gamma_{\text{decay}}})^{-1}} \leq \frac{2}{\alpha_{\text{opt}}} \exp\left(\frac{\left\| \hat{\theta}_{h,\ell} - \theta \right\|^2_{V_{h,\ell}^{\gamma_{\text{decay}}}}}{\beta_{h,\ell}\left(\alpha_{h,\ell}\right)} \right) \cdot C_{\text{opt}} \cdot d \log(d)$$

where $a^*_{h,\ell}$ *is true best arm at user* h *in episode* ℓ.

Lemma 12 ((Balagopalan & Jun, 2024)).

$$\left\| a^*_{h,\ell} \right\|^2_{(V(p_h)^{\gamma_{\text{decay}}})^{-1}} \leq \frac{2e}{\alpha_{\text{opt}}} \cdot C_{\text{opt}} \cdot d \log(d)$$