An Asymptotically Optimal Primal-Dual Incremental Algorithm for Contextual Linear Bandits

Andrea Tirinzoni[*]
Politecnico di Milano
andrea.tirinzoni@polimi.it

Matteo Pirotta
Facebook AI Research
pirotta@fb.com

Marcello Restelli
Politecnico di Milano
marcello.restelli@polimi.it

Alessandro Lazaric
Facebook AI Research
lazaric@fb.com

Abstract

In the contextual linear bandit setting, algorithms built on the optimism principle fail to exploit the structure of the problem and have been shown to be asymptotically suboptimal. In this paper, we follow recent approaches of deriving asymptotically optimal algorithms from problem-dependent regret lower bounds and we introduce a novel algorithm improving over the state-of-the-art along multiple dimensions. We build on a reformulation of the lower bound, where context distribution and exploration policy are decoupled, and we obtain an algorithm robust to unbalanced context distributions. Then, using an incremental primal-dual approach to solve the Lagrangian relaxation of the lower bound, we obtain a scalable and computationally efficient algorithm. Finally, we remove forced exploration and build on confidence intervals of the optimization problem to encourage a minimum level of exploration that is better adapted to the problem structure. We demonstrate the *asymptotic optimality* of our algorithm, while providing both problem-dependent and worst-case finite-time regret guarantees. Our bounds scale with the logarithm of the number of arms, thus avoiding the linear dependence common in all related prior works. Notably, we establish *minimax optimality* for any learning horizon in the special case of non-contextual linear bandits. Finally, we verify that our algorithm obtains better empirical performance than state-of-the-art baselines.

1 Introduction

We study the contextual linear bandit (CLB) setting [e.g., 1], where at each time step t the learner observes a context X_t drawn from a context distribution ρ, pulls an arm A_t, and receives a reward Y_t drawn from a distribution whose expected value is a linear combination between d-dimensional features $\phi(X_t, A_t)$ describing context and arm, and an unknown parameter $\theta^\star$. The objective of the learner is to maximize the reward over time, that is to minimize the cumulative regret w.r.t. an optimal strategy that selects the best arm in each context. This setting formalizes a wide range of problems such as online recommendation systems, clinical trials, dialogue systems, and many others [2]. Popular algorithmic principles, such as optimism-in-face-of-uncertainty and Thompson sampling [3], have been applied to this setting leading to algorithms such as OFUL [4] and LINTS [5, 6] with strong finite-time worst-case regret guarantees. Nonetheless, Lattimore & Szepesvari [7] recently showed that these algorithms are not asymptotically optimal (in a problem-dependent sense) as they fail to adapt to the structure of the problem at hand. In fact, in the CLB setting, the values of

[*]Work done while at Facebook.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

different arms are tightly connected through the linear assumption and a possibly suboptimal arm may provide a large amount of information about $\theta^\star$ and thus the optimal arm. Optimistic algorithms naturally discard suboptimal arms and thus may miss the chance to acquire information about $\theta^\star$ and significantly reduce the regret.

Early attempts to exploit general structures in MAB either adapted UCB-based strategies [8, 9] or focused on different criteria, such as regret to information ratio [10]. While these approaches succeed in improving the finite-time performance of optimism-based algorithms, they still do not achieve asymptotic optimality. An alternative approach to exploit the problem structure was introduced in [7] for (non-contextual) linear bandits. Inspired by approaches for regret minimization [11, 12, 13] and best-arm identification [14] in MAB, Lattimore & Szepesvari [7] proposed to compute an exploration strategy by solving the (estimated) optimization problem characterizing the asymptotic regret lower bound for linear bandits. While the resulting algorithm matches the asymptotic logarithmic lower bound with tight leading constant, it performs rather poorly in practice. Combes et al. [15] followed a similar approach and proposed OSSB, an asymptotically optimal algorithm for bandit problems with general structure (including, e.g., linear, Lipschitz, unimodal). Unfortunately, once instantiated for the linear bandit case, OSSB suffers from poor empirical performance due to the large dependency on the number of arms. Recently, Hao et al. [16] introduced OAM, an asymptotically optimal algorithm for the CLB setting. While OAM effectively exploits the linear structure and outperforms other bandit algorithms, it suffers from major limitations. From an algorithmic perspective, at each exploration step, OAM requires solving the optimization problem of the regret lower bound, which can hardly scale beyond problems with a handful of contexts and arms. Furthermore, OAM implements a forcing exploration strategy that often leads to long periods of linear regret and introduces a linear dependence on the number of arms $|\mathcal{A}|$. Finally, the regret analysis reveals a critical dependence on the inverse of the smallest probability of a context (i.e., $\min_x \rho(x)$), thus suggesting that OAM may suffer from poor finite-time performance in problems with unbalanced context distributions.[2] Degenne et al. [17] recently introduced SPL, which significantly improves over previous algorithms for MAB problems with general structures. Inspired by algorithms for best-arm identification [18], Degenne et al. reformulate the optimization problem in the lower bound as a saddle-point problem and show how to leverage online learning methods to avoid recomputing the exploration strategy from scratch at each step. Furthermore, SPL removes any form of forced exploration by introducing optimism into the estimated optimization problem. As a result, SPL is computationally efficient and achieves better empirical performance in problems with general structures.

Contributions. In this paper, we follow similar steps as in [17] and introduce SOLID, a novel algorithm for the CLB setting. Our main contributions can be summarized as follows.

- We first reformulate the optimization problem associated with the lower bound for contextual linear bandits [15, 19, 16] by introducing an additional constraint to guarantee bounded solutions and by explicitly decoupling the context distribution and the exploration policy. While we bound the bias introduced by the constraint, we also illustrate how the resulting exploration policy is better adapted to unbalanced context distributions.
- Leveraging the Lagrangian dual formulation associated with the constrained lower-bound optimization problem, we derive SOLID, an efficient primal-dual learning algorithm that incrementally updates the exploration strategy at each time step. Furthermore, we replace forced exploration with an optimistic version of the optimization problem by specifically leveraging the linear structure of the problem. Finally, SOLID does not require any explicit tracking step and it samples directly from the current exploration strategy.
- We establish the *asymptotic optimality* of SOLID, while deriving a finite-time problem-dependent regret bound that scales only with $\log |\mathcal{A}|$ and without any dependence on $\min_x \rho(x)$. To this purpose, we introduce a new concentration bound for regularized least-squares that scales as $\mathcal{O}(\log t + d \log \log t)$, hence removing the $d \log t$ dependence of the bound in [4]. Moreover, we establish a $\widetilde{\mathcal{O}}(|\mathcal{X}|\sqrt{dn})$ worst-case regret bound for any CLB problem with $|\mathcal{X}|$ contexts, d features, and horizon n. Notably, this is implies that SOLID is the first algorithm to be simultaneously *asymptotically optimal* and *minimax optimal* in non-contextual linear bandits.
- We empirically compare to a number of state-of-the-art methods for contextual linear bandits and show how SOLID is more computationally efficient and often has the smallest regret.

[2] Interestingly, Hao et al. [16] explicitly mention in their conclusions the importance of properly managing the context distribution to achieve satisfactory finite-time performance.

A thorough comparison between SOLID and related work is reported in App. B.

2 Preliminaries

We consider the contextual linear bandit setting. Let $\mathcal{X}$ be the set of contexts and $\mathcal{A}$ be the set of arms with cardinality $|\mathcal{X}| < \infty$ and $|\mathcal{A}| < \infty$, respectively. Each context-arm pair is embedded into $\mathbb{R}^d$ through a feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. For any reward model $\theta \in \mathbb{R}^d$, we denote by $\mu_\theta(x, a) = \phi(x, a)^\mathsf{T}\theta$ the expected reward for each context-arm pair. Let $a^\star_\theta(x) := \operatorname{argmax}_{a \in \mathcal{A}} \mu_\theta(x, a)$ and $\mu^\star_\theta(x) := \max_{a \in \mathcal{A}} \mu_\theta(x, a)$ denote the optimal arm and its value for context x and parameter θ. We define the sub-optimality gap of arm a for context x in model θ as $\Delta_\theta(x, a) := \mu^\star_\theta(x) - \mu_\theta(x, a)$. We assume that every time arm a is selected in context x, a random observation $Y = \phi(x, a)^\mathsf{T}\theta + \xi$ is generated, where $\xi \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.[3] Given two parameters $\theta, \theta' \in \mathbb{R}^d$, we define $d_{x,a}(\theta, \theta') := \frac{1}{2\sigma^2}(\mu_\theta(x, a) - \mu_{\theta'}(x, a))^2$, which corresponds to the Kullback-Leibler divergence between the Gaussian reward distributions of the two models in context x and arm a.

At each time step $t \in \mathbb{N}$, the learner observes a context $X_t \in \mathcal{X}$ drawn i.i.d. from a context distribution ρ, it pulls an arm $A_t \in \mathcal{A}$, and it receives a reward $Y_t = \phi(X_t, A_t)^\mathsf{T}\theta^\star + \xi_t$, where $\theta^\star \in \mathbb{R}^d$ is unknown to the learner. A bandit strategy $\pi := \{\pi_t\}_{t \geq 1}$ chooses the arm A_t to pull at time t as a measurable function $\pi_t(H_{t-1}, X_t)$ of the current context X_t and of the past history $H_{t-1} := (X_1, Y_1, \dots, X_{t-1}, Y_{t-1})$. The objective is to define a strategy that minimizes the expected cumulative regret over n steps, $\mathbb{E}^\pi_{\xi,\rho}\big[R_n(\theta)\big] := \mathbb{E}^\pi_{\xi,\rho}\left[\sum_{t=1}^n \left(\mu^\star_\theta(X_t) - \mu_\theta(X_t, A_t)\right)\right]$, where $\mathbb{E}^\pi_{\xi,\rho}$ denotes the expectation w.r.t. the randomness of contexts, the noise of the rewards, and any randomization in the algorithm. We denote by $\theta^\star$ the reward model of the bandit problem at hand, and without loss of generality we rely on the following regularity assumptions.

Assumption 1. *The realizable parameters belong to a compact subset Θ of $\mathbb{R}^d$ such that $\|\theta\|_2 \leq B$ for all $\theta \in \Theta$. The features are bounded, i.e., $\|\phi(x, a)\|_2 \leq L$ for all $x \in \mathcal{X}, a \in \mathcal{A}$. The context distribution is supported over the whole context set, i.e., $\rho(x) \geq \rho_{\min} > 0$ for all $x \in \mathcal{X}$. Finally, w.l.o.g. we assume $\theta^\star$ has a unique optimal arm in each context [see e.g., 15, 16].*

Regularized least-squares estimator. We introduce the regularized least-square estimate of $\theta^\star$ using t samples as $\widehat{\theta}_t := \overline{V}_t^{-1} U_t$, where $\overline{V}_t := \sum_{s=1}^t \phi(X_s, A_s)\phi(X_s, A_s)^\mathsf{T} + \nu I$, with $\nu \geq \max\{L^2, 1\}$ and I the $d \times d$ identity matrix, and $U_t := \sum_{s=1}^t \phi(X_s, A_s) Y_s$. The estimator $\widehat{\theta}_t$ satisfies the following concentration inequality (see App. J for the proof and exact formulation).

Theorem 1. *Let $\delta \in (0, 1)$, $n \geq 3$, and $\widehat{\theta}_t$ be a regularized least-square estimator obtained using $t \in [n]$ samples collected using an arbitrary bandit strategy $\pi := \{\pi_t\}_{t \geq 1}$. Then,*

$$\mathbb{P}\left\{\exists t \in [n] : \|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t} \geq \sqrt{c_{n,\delta}}\right\} \leq \delta,$$

where $c_{n,\delta}$ is of order $\mathcal{O}(\log(1/\delta) + d \log\log n)$.

For the usual choice $\delta = 1/n$, $c_{n,1/n}$ is of order $\mathcal{O}(\log n + d \log\log n)$, which illustrates how the dependency on d is on a lower-order term w.r.t. n (as opposed to the well-known concentration bound derived in [4]). This result is the counterpart of [7, Thm. 8] for the concentration on the reward parameter estimation error instead of the prediction error and we believe it is of independent interest.

3 Lower Bound

We recall the asymptotic lower bound for multi-armed bandit problems with structure from [20, 15, 19]. We say that a bandit strategy π is *uniformly good* if $\mathbb{E}^\pi_{\xi,\rho}\big[R_n\big] = o(n^\alpha)$ for any $\alpha > 0$ and any contextual linear bandit problem satisfying Asm. 1.

Proposition 1. *Let $\pi := \{\pi_t\}_{t \geq 1}$ by a uniformly good bandit strategy then,*

$$\liminf_{n \to \infty} \frac{\mathbb{E}^\pi_{\xi,\rho}\big[R_n(\theta^\star)\big]}{\log(n)} \geq v^\star(\theta^\star), \tag{1}$$

[3] This assumption can be relaxed by considering sub-Gaussian rewards.

where $v^\star(\theta^\star)$ is the value of the optimization problem

$$\inf_{\eta(x,a)\geq 0} \quad \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a)\Delta_{\theta^\star}(x,a) \quad \text{s.t.} \quad \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a) d_{x,a}(\theta^\star,\theta') \geq 1, \tag{P}$$

where $\Theta_{\text{alt}} := \{\theta' \in \Theta \mid \exists x \in \mathcal{X},\ a^\star_{\theta^\star}(x) \neq a^\star_{\theta'}(x)\}$ is the set of alternative reward parameters such that the optimal arm changes for at least a context x.[4]

The variables $\eta(x,a)$ can be interpreted as the number of pulls allocated to each context-arm pair so that enough information is obtained to correctly identify the optimal arm in each context while minimizing the regret. Formulating the lower bound in terms of the solution of (P) is not desirable for two main reasons. First, (P) is not a well-posed optimization problem since the inferior may not be attainable, i.e., the optimal solution may allocate an infinite number of pulls to some optimal arms. Second, (P) removes any dependency on the context distribution ρ. In fact, the optimal solution $\eta^\star$ of (P) may prescribe to select a context-arm (x,a) pair a large number of times, despite x having low probability of being sampled from ρ. While this has no impact on the asymptotic performance of $\eta^\star$ (as soon as $\rho_{\min} > 0$), building on $\eta^\star$ to design a learning algorithm may lead to poor finite-time performance. In order to mitigate these issues, we propose a variant of the previous lower bound obtained by adding a constraint on the cumulative number of pulls in each context and explicitly decoupling the context distribution ρ and the *exploration policy* $\omega(x,a)$ defining the probability of selecting arm a in context x. Given $z \in \mathbb{R}_{>0}$, we define the optimization problem

$$\min_{\omega\in\Omega} \quad z\mathbb{E}_\rho\left[\sum_{a\in\mathcal{A}} \omega(x,a)\Delta_{\theta^\star}(x,a)\right] \quad \text{s.t.} \quad \inf_{\theta'\in\Theta_{\text{alt}}} \mathbb{E}_\rho\left[\sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta')\right] \geq 1/z \tag{P_z}$$

where $\Omega = \{\omega(x,a) \geq 0 \mid \forall x \in \mathcal{X} : \sum_{a\in\mathcal{A}} \omega(x,a) = 1\}$ is the probability simplex. We denote by $\omega^\star_{z,\theta^\star}$ the optimal solution of (P_z) and $u^\star(z,\theta^\star)$ its associated value (if the problem is unfeasible we set $u^\star(z,\theta^\star) = +\infty$). Inspecting ($\text{P}_z$), we notice that z serves as a global constraint on the number of samples. In fact, for any $\omega \in \Omega$, the associated number of samples $\eta(x,a)$ allocated to a context-arm pair (x,a) is now $z\rho(x)\omega(x,a)$. Since ρ is a distribution over $\mathcal{X}$ and $\sum_a \omega(x,a) = 1$ in each context, the total number of samples sums to z. As a result, (P_z) admits a minimum and it is more amenable to designing a learning algorithm based on its Lagrangian relaxation. Furthermore, we notice that z can be interpreted as defining a more "finite-time" formulation of the lower bound. Finally, we remark that the total number of samples that can be assigned to a context x is indeed constrained to $z\rho(x)$. This constraint crucially makes (P_z) more context aware and forces the solution ω to be more adaptive to the context distribution. In Sect. 4, we leverage these features to design an incremental algorithm whose finite-time regret does not depend on $\rho_{\min}$, thus improving over previous algorithms [7, 16], as supported by the empirical results in Sect. 6. The following lemma provides a characterization of (P_z) and its relationship with (P) (see App. C for the proof and further discussion).

Lemma 1. *Let* $\underline{z}(\theta^\star) := \min\{z > 0 : (\text{P}_z) \text{ is feasible}\}$, $\overline{z}(\theta^\star) = \max_{x\in\mathcal{X}} \sum_{a\neq a^\star_{\theta^\star}(x)} \frac{\eta^\star(x,a)}{\rho(x)}$ *and* $z^\star(\theta^\star) := \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)$. *Then* $\frac{1}{\underline{z}(\theta^\star)} = \max_{\omega\in\Omega} \inf_{\theta'\in\Theta_{\text{alt}}} \mathbb{E}_\rho\left[\sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta')\right]$ *and there exists a constant $c_\Theta > 0$ such that, for any $z \in (\underline{z}(\theta^\star), +\infty)$,*

$$u^\star(z,\theta^\star) \leq v^\star(\theta^\star) + \frac{2zBL\underline{z}(\theta^\star)}{z - \underline{z}(\theta^\star)} \cdot \begin{cases} 1 & \text{if } z < \overline{z}(\theta^\star) \\ \min\left\{\max\left\{\frac{c_\Theta\sqrt{2z^\star(\theta^\star)}}{\sigma\sqrt{z}}, \frac{z^\star(\theta^\star)}{z}\right\}, 1\right\} & \text{otherwise} \end{cases}$$

The first result characterizes the range of z for which (P_z) is feasible. Interestingly, $\underline{z}(\theta^\star) < +\infty$ is the inverse of the sample complexity of the best-arm identification problem [21] and the associated solution is the one that maximizes the amount of information gathered about the reward model $\theta^\star$. As z increases, $\omega^\star_{z,\theta^\star}$ becomes less aggressive in favoring informative context-arm pairs and more sensitive to the regret minimization objective. The second result quantifies the bias w.r.t. the optimal solution of (P_z). For $z \geq \overline{z}(\theta^\star)$, the error decreases approximately at a rate $1/\sqrt{z}$ showing that the solution of (P_z) can be made arbitrarily close to $v^\star(\theta^\star)$.

[4] The infimum over this set can be computed in closed-form when the alternative parameters are allowed to lie in the whole $\mathbb{R}^d$ (see App. K.1). When these parameters are forced to have bounded ℓ_2-norm, the infimum has no closed-form expression, though its computation reduces to a simple convex optimization problem (see [21]).

In designing our learning algorithm, we build on the Lagrangian relaxation of (P_z). For any $\omega \in \Omega$, let $f(\omega;\theta^\star)$ denote the objective function and $g(\omega,z;\theta^\star)$ denote the KL constraint

$$f(\omega;\theta^\star) = \mathbb{E}_\rho\Big[\sum_{a\in\mathcal{A}} \omega(x,a)\mu_{\theta^\star}(x,a)\Big], \quad g(\omega;z,\theta^\star) = \inf_{\theta'\in\Theta_{\mathrm{alt}}} \mathbb{E}_\rho\Big[\sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta')\Big] - \frac{1}{z}.$$

We introduce the Lagrangian relaxation problem

$$\min_{\lambda\geq 0}\max_{\omega\in\Omega}\Big\{h(\omega,\lambda;z,\theta^\star) := f(\omega;\theta^\star) + \lambda g(\omega;z,\theta^\star)\Big\}, \tag{P_λ}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ is a multiplier. Notice that $f(\omega;\theta^\star)$ is not equal to the objective function of (P_z), since we replaced the gap $\Delta_{\theta^\star}$ by the expected value $\mu_{\theta^\star}$ and we removed the constant multiplicative factor z in the objective function. The associated problem is thus a concave maximization problem. While these changes do not affect the optimality of the solution, they do simplify the algorithmic design. Refer to App. D for details about the Lagrangian formulation.

4 Asymptotically Optimal Linear Primal Dual Algorithm

We introduce SOLID (aSymptotic Optimal Linear prImal Dual), which combines a primal-dual approach to incrementally compute the solution of an optimistic estimate of the Lagrangian relaxation (P_λ) within a scheme that, depending on the accuracy of the estimate $\widehat{\theta}_t$, separates *exploration* steps, where arms are pulled according to the exploration policy ω_t, and *exploitation* steps, where the greedy arm is selected. The values of the input parameters for which SOLID enjoys regret guarantees are reported in Sect. 5. In the following, we detail the main ingredients composing the algorithm (see Alg. 1).

Algorithm 1: SOLID

Input: Multiplier λ_1, confidence values $\{\beta_t\}_t$ and $\{\gamma_t\}_t$, maximum multiplier $\lambda_{\max}$, normalization factors $\{z_k\}_{k\geq 0}$, phase lengths $\{p_k\}_{k\geq 0}$, step sizes $\alpha^\lambda_k, \alpha^\omega_k$

Set $\omega_1 \leftarrow \frac{\mathbf{1}_{\mathcal{X},\mathcal{A}}}{|\mathcal{A}|}$, $\overline{V}_0 \leftarrow \nu \boldsymbol{I}$, $U_0 \leftarrow \mathbf{0}$, $\widetilde{\theta}_0 \leftarrow \mathbf{0}$, $S_0 \leftarrow 0$
Phase index: $K_1 \leftarrow 0$
for $t = 1,\dots,n$ **do**
  Receive context $X_t \sim \rho$
  Set $K_{t+1} \leftarrow K_t$
  if $\inf_{\theta'\in\overline{\Theta}_{t-1}} \|\widetilde{\theta}_{t-1} - \theta'\|^2_{\overline{V}_{t-1}} > \beta_{t-1}$ **then**
    // EXPLOITATION STEP
    $A_t \leftarrow \mathrm{argmax}_{a\in\mathcal{A}} \mu_{\widetilde{\theta}_{t-1}}(X_t,a)$
    $\lambda_{t+1} \leftarrow \lambda_t$, $\omega_{t+1} \leftarrow \omega_t$
  else
    // EXPLORATION STEP
    Sample arm: $A_t \sim \omega_t(X_t,\cdot)$
    Set $S_t \leftarrow S_{t-1} + 1$
    // UPDATE SOLUTION
    Compute $q_t \in \partial h_t(\omega_t,\lambda_t,z_{K_t})$ (see Eq. 4)
    Update policy
    $\omega_{t+1}(x,a) \leftarrow \frac{\omega_t(x,a)e^{\alpha^\omega_{K_t} q_t(x,a)}}{\sum_{a'\in\mathcal{A}}\omega_t(x,a')e^{\alpha^\omega_{K_t} q_t(x,a')}}$
    Update multiplier
    $\lambda_{t+1} \leftarrow \min\{[\lambda_t - \alpha^\lambda_{K_t} g_t(\omega_t,z_{K_t})]_+, \lambda_{\max}\}$
    // PHASE STOPPING TEST
    if $S_t - S_{T_{K_t}-1} = p_k$ **then**
      Change phase: $K_{t+1} \leftarrow K_t + 1$
      Reset solution: $\omega_{t+1} \leftarrow \omega_1$, $\lambda_{t+1} \leftarrow \lambda_1$
  Pull A_t and observe outcome Y_t
  Update $\overline{V}_t, U_t, \widehat{\theta}_t, \widehat{\rho}_t$ using X_t, A_t, Y_t
  Set $\widetilde{\theta}_t := \mathrm{argmin}_{\theta\in\Theta\cap\mathcal{C}_t} \|\theta - \widehat{\theta}_t\|_{\overline{V}_t}$

Estimation. SOLID stores and updates the regularized least-square estimate $\widehat{\theta}_t$ using all samples observed over time. To account for the fact that $\widehat{\theta}_t$ may have large norm (i.e., $\|\widehat{\theta}_t\|_2 > B$ and $\widehat{\theta}_t \notin \Theta$), SOLID explicitly projects $\widehat{\theta}_t$ onto Θ. Formally, let $\mathcal{C}_t := \{\theta \in \mathbb{R}^d : \|\theta - \widehat{\theta}_t\|^2_{\overline{V}_t} \leq \beta_t\}$ be the confidence ellipsoid at time t. Then, SOLID computes $\widetilde{\theta}_t := \mathrm{argmin}_{\theta\in\Theta\cap\mathcal{C}_t} \|\theta - \widehat{\theta}_t\|^2_{\overline{V}_t}$. This is a simple convex optimization problem, though it has no closed-form expression.[5] Note that, on those steps where $\theta^\star \notin \mathcal{C}_t$, $\Theta \cap \mathcal{C}_t$ might be empty, in which case we can set $\widetilde{\theta}_t = \widetilde{\theta}_{t-1}$. Then, SOLID uses $\widetilde{\theta}_t$ instead of $\widehat{\theta}_t$ in all steps of the algorithm. SOLID also computes an empirical estimate of the context distribution as $\widehat{\rho}_t(x) = \frac{1}{t}\sum_{s=1}^t \mathbb{1}\{X_s = x\}$.

Accuracy test and tracking. Similar to previous algorithms leveraging asymptotic lower bounds, we build on the generalized likelihood ratio test [e.g., 18] to verify the accuracy of the estimate $\widehat{\theta}_t$. At the beginning of each step t, SOLID first computes $\inf_{\theta'\in\overline{\Theta}_{t-1}} \|\widetilde{\theta}_{t-1} - \theta'\|^2_{\overline{V}_{t-1}}$, where $\overline{\Theta}_{t-1} = \{\theta' \in \Theta \mid \exists x \in \mathcal{X},\ a^\star_{\widetilde{\theta}_{t-1}}(x) \neq a^\star_{\theta'}(x)\}$ is the set of alternative models. This quantity

[5] The projection is required to carry out the analysis, while we ignore it in our implementation (see App. K.1).

measures the accuracy of the algorithm, where the infimum over alternative models defines the problem θ' that is closest to $\widetilde{\theta}_{t-1}$ and yet different in the optimal arm of at least one context.[6] This serves as a worst-case scenario for the true $\theta^\star$, since if $\theta^* = \theta'$ then selecting arms according to $\widetilde{\theta}_{t-1}$ would lead to linear regret. If the accuracy exceeds a threshold β_{t-1}, then SOLID performs an exploitation step, where the estimated optimal arm $a^\star_{\widetilde{\theta}_{t-1}}(X_t)$ is selected in the current context. On the other hand, if the test fails, the algorithm moves to an exploration step, where an arm A_t is sampled according to the estimated exploration policy $\omega_t(X_t, \cdot)$. While this approach is considerably simpler than standard tracking strategies (e.g., selecting the arm with the largest gap between the policy ω_t and the number of pulls), in Sect. 5 we show that sampling from ω_t achieves the same level of tracking efficiency.

Optimistic primal-dual subgradient descent. At each step t, we define an estimated optimistic version of the Lagrangian relaxation (P_λ) as

$$f_t(\omega) := \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) \left(\mu_{\widetilde{\theta}_{t-1}}(x,a) + \sqrt{\gamma_t} \|\phi(x,a)\|_{\overline{V}^{-1}_{t-1}} \right), \tag{2}$$

$$g_t(\omega, z) := \inf_{\theta'\in\overline{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) \left(d_{x,a}(\widetilde{\theta}_{t-1}, \theta') + \frac{2BL}{\sigma^2} \sqrt{\gamma_t} \|\phi(x,a)\|_{\overline{V}^{-1}_{t-1}} \right) - \frac{1}{z}, \tag{3}$$

$$h_t(\omega, \lambda, z) := f_t(\omega) + \lambda g_t(\omega, z), \tag{4}$$

where γ_t is a suitable parameter defining the size of the confidence interval.

Notice that we do not use optimism on the context distribution, which is simply replaced by its empirical estimate. Therefore, h_t is not necessarily optimistic with respect to the original Lagrangian function h. Nonetheless, we prove in Sect. 5 that this level of optimism is sufficient to induce enough exploration to have accurate estimates of $\theta^\star$. This is in contrast with the popular forced exploration strategy [e.g. 7, 15, 19, 16], which prescribes a minimum fraction of pulls ϵ such that at any step t, any of the arms with less than ϵS_t pulls is selected, where S_t is the number of exploration rounds so far. While this strategy is sufficient to guarantee a minimum level of accuracy for $\widehat{\theta}_t$ and to obtain asymptotic regret optimality, in practice it is highly inefficient as it requires selecting all arms in each context regardless of their value or amount of information.

At each step t, SOLID updates the estimates of the optimal exploration policy ω_t and the Lagrangian multiplier λ_t. In particular, given the sub-gradient q_t of $h_t(\omega_t, \lambda_t, z_{K_t})$, SOLID updates ω_t and λ_t by performing one step of projected sub-gradient descent with suitable learning rates $\alpha^\omega_{K_t}$ and $\alpha^\lambda_{K_t}$. In the update of ω_t, we perform the projection onto the simplex Ω using an entropic metric, while the multiplier is clipped in $[0, \lambda_{\max}]$. While this is a rather standard primal-dual approach to solve the Lagrangian relaxation (P_λ), the interplay between estimates $\widehat{\theta}_t$, ρ_t, the optimism used in h_t, and the overall regret performance of the algorithm is at the core of the analysis in Sect. 5.

This approach significantly reduces the computational complexity compared to [15, 16], which require solving problem P at each exploratory step. In Sect. 6, we show that the incremental nature of SOLID allows it to scale to problems with much larger context-arm spaces. Furthermore, we leverage the convergence rate guarantees of the primal-dual gradient descent to show that the incremental nature of SOLID does not compromise the asymptotic optimality of the algorithm (see Sect. 5).

The z parameter. While the primal-dual algorithm is guaranteed to converge to the solution of (P_z) for any fix z, it may be difficult to properly tune z to control the error w.r.t. (P). SOLID leverages the fact that the error scales as $1/\sqrt{z}$ (Lem. 1 for z sufficiently large) and it increases z over time. Given as input two non-decreasing sequences $\{p_k\}_k$ and $\{z_k\}_k$, at each phase k, SOLID uses z_k in the computation of the subgradient of h_t and in the definition of f_t and g_t. After p_k explorative steps, it resets the policy ω_t and the multiplier λ_t and transitions to phase $k+1$. Since $p_k = S_{T_{k+1}-1} - S_{T_k-1}$ is the number of *explorative* steps of phase k starting at time T_k, the actual number of steps during k may vary. Notice that at the end of each phase only the optimization variables are reset, while the learning variables (i.e., $\widehat{\theta}_t$, $\overline{V}_t$, and $\widehat{\rho}_t$) use all the samples collected through phases.

[6] In practice, it is more efficient to take the infimum only over problems with different optimal arm in the last observed context X_t. This is indeed what we do in our experiments and all our theoretical results follow using this alternative definition with only minor changes.

5 Regret Analysis

Before reporting the main theoretical result of the paper, we introduce the following assumption.

Assumption 2. *The maximum multiplier used by* SOLID *is such that* $\lambda_{\max} \geq 2BL\underline{z}(\theta^\star)$.

While an assumption on the maximum multiplier is rather standard for the analysis of primal-dual projected subgradient [e.g., 22, 23], we conjecture that it may be actually relaxed in our case by replacing the fixed $\lambda_{\max}$ by an increasing sequence as done for $\{z_k\}_k$.

Theorem 2. *Consider a contextual linear bandit problem with contexts $\mathcal{X}$, arms $\mathcal{A}$, reward parameter $\theta^\star$, features bounded by L, zero-mean Gaussian noise with variance σ^2 and context distribution ρ satisfying Asm. 1. If* SOLID *is run with confidence values $\beta_{t-1} = c_{n,1/n}$ and $\gamma_t = c_{n,1/S_t^2}$, where $c_{n,\delta}$ is defined as in Thm. 1, learning rates $\alpha_k^\lambda = \alpha_k^\omega = 1/\sqrt{p_k}$ and increasing sequences $z_k = z_0 e^k$ and $p_k = z_k e^{2k}$, for some $z_0 \geq 1$, then it is* ***asymptotically optimal*** *with the same constant as in the lower bound of Prop. 1. Furthermore, for any finite n the regret of* SOLID *is bounded as*

$$\mathbb{E}^{\pi}_{\xi,\rho}\big[R_n(\theta^\star)\big] \leq v^\star(\theta^\star)\frac{c_{n,1/n}}{2\sigma^2} + C_{\log}(\log\log n)^{\frac{1}{2}}(\log n)^{\frac{3}{4}} + C_{\text{const}}, \tag{5}$$

where $C_{\log} = lin_{\geq 0}(v^\star(\theta^\star), |\mathcal{X}|, L^2, B^2, \sqrt{d}, 1/\sigma^2)$ *and* $C_{\text{const}} = v^\star(\theta^\star)\frac{B^2L^2}{\sigma^2} + lin_{\geq 0}(L, B, z_0(\underline{z}(\theta^\star)/z_0)^3, (\overline{z}(\theta^\star)/z_0)^2)$.[7]

The first result shows that SOLID run with an exponential schedule for z is asymptotic optimal, while the second one provides a bound on the finite-time regret. We can identify three main components in the finite-time regret. **1)** The first term scales with the logarithmic term $c_{n,1/n} = O(\log n + d\log\log n)$ and a leading constant $v^\star(\theta^\star)$, which is optimal as shown in Prop. 1. In most cases, this is the dominant term of the regret. **2)** Lower-order terms in $o(\log n)$. Notably, a regret of order $\sqrt{\log n}$ is due to the incremental nature of SOLID and it is directly inherited from the convergence rate of the primal-dual algorithm we use to optimize (P_z). The larger term $(\log n)^{3/4}$ that we obtain in the final regret is actually due to the schedule of $\{z_k\}$ and $\{p_k\}$. While it is possible to design a different phase schedule to reduce the exponent towards $1/2$, this would negatively impact the constant regret term. **3)** The constant regret C_{const} is due to the exploitation steps, burn-in phase and the initial value z_0. The regret due to z_0 takes into account the regime when (P_z) is unfeasible ($z_k < \underline{z}(\theta^\star)$) or when z_k is too small to assess the rate at which $u^\star(z_k, \theta^\star)$ approaches $v^\star(\theta^\star)$ ($z < \overline{z}(\theta^\star)$), see Lem. 1. Notably, the regret due to the initial value z_0 vanishes when $z_0 > \overline{z}(\theta^\star)$. A more aggressive schedule for z_k reaching $\overline{z}(\theta^\star)$ in few phases would reduce the initial regret at the cost of a larger exponent in the sub-logarithmic terms.

The sub-logarithmic terms in the regret have only logarithmic dependency on the number of arms. This is better than existing algorithms based on exploration strategies built from lower bounds. OSSB [15] indeed depends on $|\mathcal{A}|$ directly in the main $\mathcal{O}(\log n)$ regret terms. While the regret analysis of OAM is asymptotic, it is possible to identify several lower-order terms depending linearly on $|\mathcal{A}|$. In fact, OAM as well as OSSB require forced exploration on each context-arm pair, which inevitably translates into regret. In this sense, the dependency on $|\mathcal{A}|$ is hard-coded into the algorithm and cannot be improved by a better analysis. SPL depends linearly on $|\mathcal{A}|$ in the explore/exploit threshold (the equivalent of our β_t) and in other lower-order terms due to the analysis of the tracking rule. On the other hand, SOLID never requires all arms to be repeatedly pulled and we were able to remove the linear dependence on $|\mathcal{A}|$ through a refined analysis of the sampling procedure (see App. E). This is inline with the experimental results where we did not notice any explicit linear dependence on $|\mathcal{A}|$.

The constant regret term depends on the context distribution through $\overline{z}(\theta^\star)$ (Lem. 1). Nonetheless, this dependency disappears whenever z_0 is a fraction $\overline{z}(\theta^\star)$. This is in striking contrast with OAM, whose analysis includes several terms depending on the inverse of the context probability $\rho_{\min}$. This confirms that SOLID is able to better adapt to the distribution generating the contexts. While the phase schedule of Thm. 2 leads to an asymptotically-optimal algorithm and sublinear-regret in finite time, it may be possible to find a different schedule having the same asymptotic performance and better finite-time guarantees, although this may depend on the horizon n. Refer to App. G.3 for a regret bound highlighting the explicit dependence on the sequences $\{z_k\}$ and $\{p_k\}$.

[7] *lin*$(\cdot)$ denotes any function with linear or sublinear dependence on the inputs (ignoring logarithmic terms). For example, $lin_{\geq 0}(x, y^2) \in \{a_0 + a_1 x + a_2 y + a_3 y^2 + a_4 x y^2 : a_i \geq 0\}$.

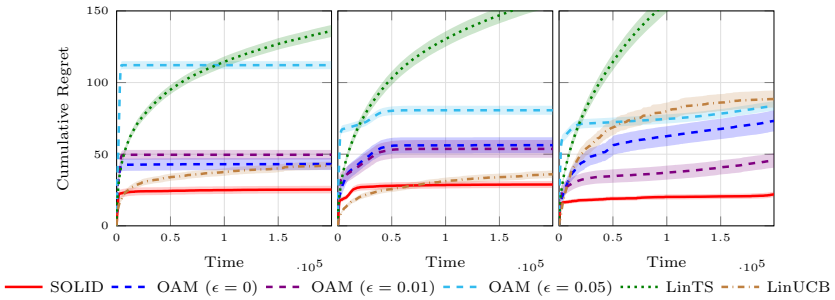

Figure 1: Toy problem with 2 contexts and (left) $\rho(x_1) = .5$, (center) $\rho(x_1) = .9$, (right) $\rho(x_1) = .99$.

As shown in [16], when the features of the optimal arms span $\mathbb{R}^d$, the asymptotic lower bound vanishes (i.e., $v^\star(\theta^\star) = 0$). In this case, selecting optimal arms is already informative enough to correctly estimate $\theta^\star$ and no explicit exploration is needed and SOLID, like OAM, has sub-logarithmic regret.

Worst-case analysis. The constant terms in Thm. 2 are due to a naive bound which assumes linear regret in those phases where z_k is small (e.g., when the optimization problem is infeasible). While this simplifies the analysis for asymptotic optimality, we verify that SOLID always suffers sub-linear regret, regardless of the values of z_k. For the following result, we do not require Asm. 2 to hold.

Theorem 3 (Worst-case regret bound)**.** *Let z_k be arbitrary, $p_k = e^{rk}$ for some constant $r \geq 1$, and the other parameters be the same as in Thm. 2. Then, for any n the regret of* SOLID *is bounded as*

$$\mathbb{E}^{\pi}_{\xi,\rho}\big[R_n(\theta^\star)\big] \leq 3BL\pi^2 \left(4 + \frac{\lambda_{\max} BL}{\sigma^2}\right) + \frac{2e^r \lambda_{\max}^2}{r}\sqrt{n} + C_{\mathrm{sqrt}} \left(1 + \frac{\lambda_{\max} BL}{\sigma^2}\right) \log(n)\sqrt{n},$$

where $C_{\mathrm{sqrt}} = lin_{\geq 0}(|\mathcal{X}|, \sqrt{d}, B, L)$.

Notably, this bound removes the dependencies on $\underline{z}(\theta^\star)$ and $\overline{z}(\theta^\star)$, while its derivation is agnostic to the values of z_k. Interestingly, we could set $\lambda_{\max} = 0$ and the algorithm would completely ignore the KL constraint, thus focusing only on the objective function. This is reflected in the worst-case bound since all terms with a dependence on σ^2 or a quadratic dependence on BL disappear. The key result is that the objective function alone, thanks to optimism, is sufficient for proving sub-linear regret but not for proving asymptotic optimality. More precisely, the resulting bound is $\widetilde{\mathcal{O}}(|\mathcal{X}|\sqrt{nd})$, which matches the minimax optimal rate apart from the dependence on $|\mathcal{X}|$. The latter could be reduced to $\sqrt{|\mathcal{X}|}$ by a better analysis. It remains an open question how to design an asymptotically optimal algorithm for the contextual case whose regret does not scale with $|\mathcal{X}|$.

6 Numerical Simulations

We compare SOLID to LinUCB, LinTS, and OAM. For SOLID, we set $\beta_t = \sigma^2(\log(t) + d\log\log(n))$ and $\gamma_t = \sigma^2(\log(S_t) + d\log\log(n))$ (i.e., we remove all numerical constants) and we use the exponential schedule for phases defined in Thm. 2. For OAM, we use the same β_t for the explore/exploit test and we try different values for the forced-exploration parameter ϵ. LinUCB uses the confidence intervals from Thm. 2 in [4] with the log-determinant of the design matrix, and LinTS is as defined in [5] but without the extra-sampling factor $\sqrt{d}$ used to prove its frequentist regret. All plots are the results of 100 runs with 95% Student's t confidence intervals. See App. K for additional details and results on a real dataset.

Toy contextual linear bandit with structure. We start with a CLB problem with $|\mathcal{X}| = 2$ and $|\mathcal{A}|, d = 3$. Let x_i (a_i) be the i-th context (arm). We have $\phi(x_1, a_1) = [1, 0, 0]$, $\phi(x_1, a_2) = [0, 1, 0]$, $\phi(x_1, a_3) = [1 - \xi, 2\xi, 0]$, $\phi(x_2, a_1) = [0, 0.6, 0.8]$, $\phi(x_2, a_2) = [0, 0, 1]$, $\phi(x_2, a_3) = [0, \xi/10, 1 - \xi]$ and $\theta^\star = [1, 0, 1]$. We consider a balanced context distribution $\rho(x_1) = \rho(x_2) = 0.5$. This is a two-context counterpart of the example presented by [7] to show the asymptotic sub-optimality of optimism-based strategies. The intuition is that, for ξ small, an optimistic strategy pulls a_2 in x_1 and a_1 in x_2 only a few times since their gap is quite large, and suffers high regret (inversely proportional to ξ) to figure out which of the remaining arms is optimal. On the other hand, an asymptotically optimal strategy allocates more pulls to "bad" arms as they bring information to identify $\theta^\star$, which in

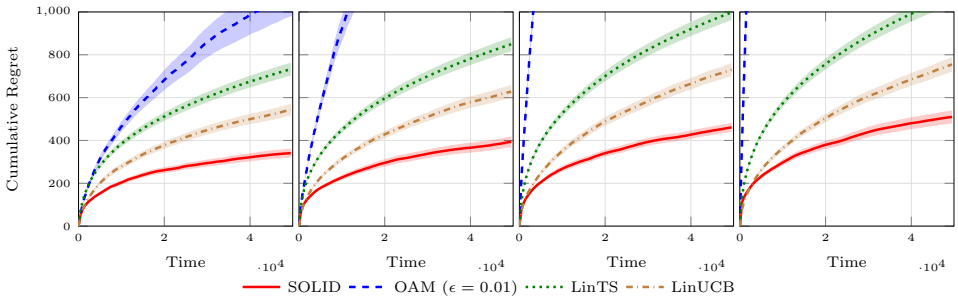

Figure 2: Randomly generated bandit problems with $d = 8$, $|\mathcal{X}| = 4$, and $|\mathcal{A}| = 4, 8, 16, 32$.

turns avoids a regret scaling with ξ. This indeed translates into the empirical performance reported in Fig. 1-(left), where SOLID effectively exploits the structure of the problem and significantly reduces the regret compared to LinTS and LinUCB. Actually, not only the regret is smaller but the "trend" is better. In fact, the regret curves of LinUCB and LinTS have a larger slope than SOLID's, suggesting that the gap may increase further with n, thus confirming the theoretical finding that the asymptotic performance of SOLID is better. OAM has a similar behavior, but the actual performance is worse than SOLID and it seems to be very sensitive to the forced exploration parameter, where the best performance is obtained for $\epsilon = 0.0$, which is not theoretically justified.

We also study the influence of the context distribution. We first notice that solving (P) leads to an optimal exploration strategy $\eta^\star$ where the only sub-optimal arm with non-zero pulls is a_1 in x_2 since it yields lower regret and similar information than a_2 in x_1. This means that the lower bound prescribes a greedy policy in x_1, deferring exploration to x_2 alone. In practice, tracking this optimal allocation might lead to poor finite-time performance when the context distribution is unbalanced towards x_1, in which case the algorithm would take time proportional to $1/\rho(x_2)$ before performing any meaningful exploration. We verify these intuitions empirically by considering the case of $\rho(x_1) = 0.9$ and $\rho(x_1) = 0.99$ (middle and right plots in Fig. 1 respectively). SOLID is consistently better than all other algorithms, showing that its performance is not negatively affected by $\rho_{\min}$. On the other hand, OAM is more severely affected by the context distribution. In particular, its performance with $\epsilon = 0$ significantly decreases when increasing $\rho(x_1)$ and the algorithm reduces to an almost greedy strategy, thus suffering linear regret in some problems. In this specific case, forcing exploration leads to slightly better finite-time performance since the algorithm pulls the informative arm a_2 in x_1, which is however not prescribed by the lower bound.

Random problems. We evaluate the impact of the number of actions $|\mathcal{A}|$ in randomly generated structured problems with $d = 8$ and $|\mathcal{X}| = 4$. We run each algorithm for $n = 50000$ steps. For OAM, we set forced-exploration $\epsilon = 0.01$ and solve (P) every 100 rounds to speed-up execution as computation becomes prohibitive. The plots in Fig. 2 show the regret over time for $|\mathcal{A}| = 4, 8, 16, 32$. This test confirms the advantage of SOLID over the other methods. Interestingly, the regret of SOLID does not seem to significantly increase as a function of $|\mathcal{A}|$, thus supporting its theoretical analysis. On the other hand, the regret of OAM scales poorly with $|\mathcal{A}|$ since forced exploration pulls all arms in a round robin fashion.

7 Conclusion

We introduced SOLID, a novel asymptotically-optimal algorithm for contextual linear bandits with finite-time regret and computational complexity improving over similar methods and better empirical performance w.r.t. state-of-the-art algorithms in our experiments. The main open question is whether SOLID is minimax optimal for contextual problems with $|\mathcal{X}| > 1$. In future work, our method could be extended to continuous contexts, which would probably require a reformulation of the lower bound and the adoption of parametrized policies. Furthermore, it would be interesting to study finite-time lower bounds, especially for problems in which bounded regret is achievable [9, 24, 25]. Finally, we could use algorithmic ideas similar to SOLID to go beyond the realizable linear bandit setting.

Broader Impact

This work is mainly a theoretical contribution. We believe it does not present any foreseeable societal consequence.

Funding Transparency Statement

Marcello Restelli was partially funded by the Italian MIUR PRIN 2017 Project ALGADIMAR "Algorithms, Games, and Digital Market".

Acknowledgements

The authors would like to thank Rémy Degenne, Han Shao, and Wouter Koolen for kindly sharing the draft of their paper before publication. We also would like to thank Pierre Ménard for carefully reading the paper and for providing insightful feedback.

References

[1] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[2] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

[3] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[4] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[5] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[6] Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 2017.

[7] Tor Lattimore and Csaba Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 728–737. PMLR, 2017.

[8] Mohammad Azar, Alessandro Lazaric, and Emma Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2220–2228. 2013.

[9] Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558, 2014.

[10] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.

[11] Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled markov chains: Finite parameter space. In *Proceedings of the 27th IEEE Conference on Decision and Control*, pages 1198–1203. IEEE, 1988.

[12] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[13] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[14] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

[15] Richard Combes, Stefan Magureanu, and Alexandre Proutière. Minimal exploration in structured stochastic bandits. In *NIPS*, pages 1763–1771, 2017.

[16] Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. volume 108 of *Proceedings of Machine Learning Research*, pages 3536–3545, Online, 26–28 Aug 2020. PMLR.

[17] Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, Vienna, Austria, 2020. Virtual conference.

[18] Rémy Degenne, Wouter M. Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *NeurIPS*, pages 14465–14474, 2019.

[19] Jungseul Ok, Alexandre Proutière, and Damianos Tranos. Exploration in structured reinforcement learning. In *NeurIPS*, pages 8874–8882, 2018.

[20] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[21] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, Vienna, Austria, 2020. Virtual conference.

[22] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

[23] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[24] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *arXiv preprint arXiv:1704.09011*, 2017.

[25] Andrea Tirinzoni, Alessandro Lazaric, and Marcello Restelli. A novel confidence-based algorithm for structured bandits. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 3175–3185. PMLR, 2020.

[26] Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward mdps. In *NeurIPS*, pages 4891–4900. 2019.

[27] Richard Combes and Alexandre Proutière. Unimodal bandits: Regret lower bounds and optimal algorithms. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 521–529. JMLR.org, 2014.

[28] Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

[29] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

[31] Xiequan Fan, Ion Grama, Quansheng Liu, et al. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20, 2015.

[32] Rasul A Khan. L p-version of the dubins–savage inequality and some exponential inequalities. *Journal of Theoretical Probability*, 22(2):348, 2009.

[33] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *information retrieval*, 4(2):133–151, 2001.

Appendix

Table of Contents

A Notation and Definitions

We provide this table for easy reference. Notation will also be defined as it is introduced.

Table 1: Symbols

$\theta^\star$	The true reward parameter
$\mathcal{X}$	Finite set of contexts
$\mathcal{A}$	Finite set of arms
σ^2	Variance of the Gaussian reward noise
B	Maximum l_2-norm of realizable reward parameters
L	Maximum l_2-norm of the features
ρ	Context distribution
$\widehat{\rho}_t(x) := \frac{1}{t}\sum_{s=1}^t \mathbb{1}\{X_s = x\}$	Estimated context distribution
$\mu_\theta(x,a)$	Mean reward of context x and arm a
$\Delta_\theta(x,a) := \max_{a' \in \mathcal{A}} \mu_\theta(x,a') - \mu_\theta(x,a)$	Gap of context x and arm a
$a^\star_\theta(x) := \operatorname{argmax}_{a\in\mathcal{A}} \mu_\theta(x,a)$	Optimal arm of context x
$\mu^\star_\theta(x) := \max_{a\in\mathcal{A}} \mu_\theta(x,a)$	Optimal reward value of context x
$d_{x,a}(\theta,\theta') := \frac{1}{2\sigma^2}(\mu_\theta(x,a) - \mu_{\theta'}(x,a))^2$	KL divergence between θ and θ' at x, a
$\Theta_{\text{alt}} := \{\theta' \in \Theta \mid \exists x \in \mathcal{X},\ a^\star_{\theta^\star}(x) \neq a^\star_{\theta'}(x)\}$	Set of alternative reward models
$\overline{\Theta}_{t-1} = \{\theta' \in \Theta \mid \exists x \in \mathcal{X},\ a^\star_{\widetilde{\theta}_{t-1}}(x) \neq a^\star_{\theta'}(x)\}$	Estimated set of alternative reward models
$v^\star(\theta^\star)$	Optimal value of the optimization problem (P)
$\eta^\star$	Optimal solution of the optimization problem (P)
$u^\star(z,\theta^\star)$	Optimal value of the optimization problem (P_z)
$\omega^\star_{z,\theta^\star}$	Optimal solution of the optimization problem (P_z)
$\underline{z}(\theta^\star) := \min\{z > 0 : (\mathrm{P}_z) \text{ is feasible}\}$	Feasibility threshold of (P_z)
$h(\omega,\lambda; z,\theta^\star) := f(\omega;\theta^\star) + \lambda g(\omega; z,\theta^\star)$	Lagrangian relaxation of (P_z)
$f(\omega;\theta^\star)$	Objective function
$f_t(\omega)$	Estimated (optimistic) objective function (see Eq. 2)
$g(\omega; z,\theta^\star)$	Constraint function
$g_t(\omega, z)$	Estimated (optimistic) constraint (see Eq. 3)
$E_t := \mathbb{1}\left\{\inf_{\theta'\in\overline{\Theta}_{t-1}} \lVert\widetilde{\theta}_{t-1} - \theta'\rVert^2_{\overline{V}_{t-1}} \le \beta_{t-1}\right\}$	Exploration round
$N_t(x,a) := \sum_{s=1}^t \mathbb{1}\{X_t = x, A_t = a\}$	Total number of visits to (x,a)
$N^E_t(x,a) := \sum_{s=1}^t \mathbb{1}\{X_t = x, A_t = a, E_t\}$	Number of visits to (x,a) in exploration rounds
$S_t := \sum_{s=1}^t \mathbb{1}\{E_t\}$	Total number of exploration rounds
$\beta_{t-1} := c_{n,1/n}$	Theoretical threshold for the exploitation test in SOLID
$\gamma_t := c_{n,1/S^2_t}$	Theoretical value for the confidence intervals in SOLID
$K_t \in \{0,1,\dots\}$	Phase index at time t
T_k	Time at which phase k starts
$\mathcal{T}_k := \{t \in [n] : K_t = k\}$	Time steps in phase k
$\mathcal{T}^E_k := \{t \in \mathcal{T}_k : E_t\}$	Exploration rounds in phase k
$\{p_k\}_{k\ge 0}$	Total number of exploration rounds in each phase
$\alpha^\lambda_k, \alpha^\omega_k$	Step sizes
$V_t := \sum_{s=1}^t \phi(X_s,A_s)\phi(X_s,A_s)^\mathsf{T}$	Design matrix
$\overline{V}_t := V_t + \nu I$	Regularized design matrix ($\nu \ge 1$)
$U_t := \sum_{s=1}^t \phi(X_s,A_s)Y_s$	Sum of reward-weighted features
$\widehat{\theta}_t := \overline{V}_t^{-1} U_t$	Regularized least-squares estimate
$\widetilde{\theta}_t := \operatorname{argmin}_{\theta\in\Theta\cap\mathcal{C}_t} \lVert\theta - \widehat{\theta}_t\rVert^2_{\overline{V}_t}$	Projected least-squares estimates
$\mathcal{C}_t := \{\theta \in \mathbb{R}^d : \lVert\theta - \widehat{\theta}_t\rVert^2_{\overline{V}_t} \le \beta_t\}$	Confidence ellipsoid at time t
G_t	Good event (see App. F)
$M_n = \sum_{t=1}^n \mathbb{1}\{E_t, \neg G_t\}$	Number of exploration rounds without good event
$M_{n,k} = \sum_{t\in\mathcal{T}^E_k} \mathbb{1}\{\neg G_t\}$	Number of exploration rounds in phase k without good event

Feature/Algorithm	OSSB	OAM	SPL	SOLID
Setting	general MAB	linear contextual	general MAB	linear contextual
Objective fun.	constrained	constrained	saddle (ratio)	saddle (Lagrangian)
Opt. variables	counts	counts	rates	policies
Asympt. optimality	order-opt	opt	opt	opt
Finite-time bound	✓	✗	✓	✓
Explore/exploit	tracking test	glrt	glrt	glrt
Tracking	direct	direct	cumulative	sampling
Optimization	exact	exact	incr. and best-response	incr.
Exp. level	forcing	forcing	unstruct. optimism	optimism
Parameters	forcing, test	forcing, test	gaps clip, test, conf. values	$\lambda_{\max}$, test, conf. values, phases

Table 2: Comparison of structured bandit algorithms. OSSB [15], OAM [16], SPL [17] and SOLID (this paper).

B Comparison to Related Work

In Table 2 we compare several bandit algorithms along several dimensions:

- *Setting* refers to whether the algorithm is designed for general multi-armed bandit (non-contextual) structured problems or it is for the linear contextual case.
- *Objective function* refers to the optimization problem solved by the algorithm. It can be either the original constrained optimization in (P) or a saddle point problem (either obtained by taking the ratio of objective and constraints or the Lagrangian relaxation in (P_z)).
- *Optimization variables* refers to the variables that are optimized by the algorithm: *counts* is the η variables in (P), *rates* is the ratio fraction of regret, *policies* is the ω variables in (P_z).
- *Asymptotic optimality* is either *order optimal* when only a logarithmic rate is proved with non-optimal constants, or *optimal*, in which case the leading constant is $v^\star(\theta)$ as in Prop. 1.
- *Finite-time bound* is whether finite-time guarantees are reported.
- *Explore/exploit* refers to the separation between exploration and exploitation steps and whether it is based on a *tracking performance test* or on the generalized likelihood ratio test (GLRT).[8]
- *Tracking* refers to how arms are selected during the exploration phase.
- *Optimization* refers to whether the optimization problem is solved exactly at each step or using an incremental method. SPL combines an incremental method using an exact computation of a best response solution.
- *Exploration level* refers to the technique used during exploration steps to guarantee a minimum level of exploration. The first option is *forcing* all arms to satisfy a hard threshold of minimal pulls. The second option is to include a form of *optimism* in the optimization problem.
- *Parameters* list the major parameters in the definition of the algorithm. This is often difficult since some algorithms directly pick theoretical values for some input parameters, while others may provide specific values only during the analysis. OSSB requires tuning the forcing parameter and the parameter used in the exploration/exploitation test. OAM has a forcing parameter and needs to properly tune the GLRT. SPL requires clipping the gap estimates from below, tuning the GLRT, and designing suitable confidence intervals for optimism. SOLID requires an upper bound for the multiplier, tuning of the GLRT, confidence intervals, and phases to tune the normalization factor z.

The major insights from this comparison can be summarized as follows:

[8] Notice that none of the algorithms implement the exact form of the GLRT, but slight variations that provide equivalent guarantees.

- *Comparison* SOLID*/OAM:* This is the more direct comparison, since both algorithms are designed for contextual linear (see Sect. 6 for the empirical comparison). SOLID improves over OAM in almost all dimensions. On the theoretical side, we provide explicit finite-time regret bounds showing that SOLID successfully adapts to the context distribution, while the performance of OAM is significantly affected by $\rho_{\min}$. Furthermore, in many lower-order regret terms in the analysis of OAM the cardinality of the arm space appears linearly, while the regret of SOLID only depends on $\log(|\mathcal{A}|)$. On the algorithmic side, SOLID leverages a primal-dual gradient descent that greatly improves the computational complexity compared to the exact solution of the constrained optimization problem done in OAM at each exploration step. Furthermore, replacing the forcing strategy with an optimistic version of the optimization problem allows SOLID to better adapt to the problem and avoid pulling highly suboptimal/non-informative arms.

- *Comparison* SOLID*/SPL:* The comparison is more on the algorithmic and theoretical properties rather than the actual algorithms, since they are designed for different settings.[9] While both algorithms replace the constrained problem in the lower bound by a saddle point problem, SPL takes the ratio between constraints and regret, while in SOLID we take a more straightforward Lagrangian relaxation. As a result, in SOLID we rely on a rather standard primal-dual gradient approach to optimize (P_z), while SPL relies on online learning algorithms for the solution of the saddle-point problem. Finally, both algorithms replace forcing by an optimistic version of the optimization problem. Nonetheless, SPL uses separate confidence intervals for each arm that ignore the structure of the problem, while SOLID relies on confidence intervals build specifically for the linear case. Finally, the regret bound of SPL, similarly to the one of OAM, depends linearly on $|\mathcal{A}|$ in several lower-order terms, even when instantiated for linear structures. SOLID, on the other hand, has only $\log(|\mathcal{A}|)$ dependence.

C Lower Bound

C.1 Proof of Lem. 1

Feasibility of (P_z)**.** We start from the first result in Lem. 1, which states the minimal value of z for which (P_z) is feasible. Clearly, the maximal value that the left-hand side of the KL constraint can assume is

$$\max_{\omega \in \Omega} \inf_{\theta' \in \Theta_{\mathrm{alt}}} \mathbb{E}_{\rho} \left[\sum_{a \in \mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star, \theta') \right],$$

which can also be interpreted as the solution to the associated pure-exploration (or best-arm identification) problem [e.g., 18]. Therefore,

$$\begin{aligned}
\underline{z}(\theta^\star) &:= \min \{ z > 0 : (\mathrm{P}_z) \text{ is feasible} \} \\
&= \min \left\{ z > 0 : \max_{\omega \in \Omega} \inf_{\theta' \in \Theta_{\mathrm{alt}}} \mathbb{E}_{\rho} \left[\sum_{a \in \mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star, \theta') \right] \geq \frac{1}{z} \right\} \\
&= \frac{1}{\max_{\omega \in \Omega} \inf_{\theta' \in \Theta_{\mathrm{alt}}} \mathbb{E}_{\rho} \left[\sum_{a \in \mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star, \theta') \right]}.
\end{aligned}$$

This proves the first statement in Lem. 1.

[9] While the general structured bandit problem does contain the linear case, it is unclear how it can manage the contextual linear case.

Connection between (P) **and** (P_z)**.** In order to prove the second result, let us rewrite (P_z) in the following more convenient form:

$$\begin{aligned} \underset{\eta(x,a)\geq 0}{\text{minimize}} \quad & \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \eta(x,a)\Delta_{\theta^\star}(x,a) \\ \text{subject to} \quad & \inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \eta(x,a) d_{x,a}(\theta^\star,\theta') \geq 1, \\ & \sum_{a\in\mathcal{A}} \eta(x,a) = z \quad \forall x\in\mathcal{X}. \end{aligned} \tag{P'_z}$$

Note that (P'_z) is obtained from (P_z) in the main paper by performing the change of variables $\eta(x,a) = z\omega(x,a)$, hence the two problems are equivalent. Recall that $v^\star(\theta^\star)$ is the optimal value of (P) and $u^\star(z,\theta^\star)$ is the optimal value of (P'_z) and (P_z) (if there exists one). We are interested in bounding the deviation between $u^\star(z,\theta^\star)$ and $v^\star(\theta^\star)$ as a function of z.

Let us first define the following set of *confusing* models:

$$\widetilde{\Theta}_{alt} := \{\theta' \in \Theta_{alt} : \forall x\in\mathcal{X}, \mu^\star_{\theta^\star}(x) = \mu_{\theta'}(x, a^\star_x)\},$$

where, for the sake of readability, we abbreviate $a^\star_x = a^\star_{\theta^\star}(x)$. These models are indistinguishable from $\theta^\star$ by pulling only optimal arms. The following proposition, which was proved in [17], connects models in the alternative set Θ_{alt} with the confusing ones in $\widetilde{\Theta}_{\mathrm{alt}}$.

Proposition 2 ([17])**.** *There exists a constant $c_\Theta > 0$ such that, for all $\theta' \in \Theta_{\mathrm{alt}}$, there exists $\theta'' \in \widetilde{\Theta}_{\mathrm{alt}}$ such that,*

$$\forall x\in\mathcal{X}, a\in\mathcal{A} \qquad |\mu_{\theta'}(x,a) - \mu_{\theta''}(x,a)| \leq c_\Theta |\mu^\star_{\theta^\star}(x) - \mu_{\theta'}(x, a^\star_{\theta^\star}(x))|.$$

We now prove the bound on $u^\star(z,\theta)$ reported in Lem. 1.

Proof of Lem. 1. We start from the Lagrangian version of (P'_z).

$$u^\star(z,\theta) = \min_{\eta\geq 0} \left\{ \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \eta(x,a)\Delta_{\theta^\star}(x,a) + \lambda^\star(z,\theta^\star) \left(1 - \inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \eta(x,a) d_{x,a}(\theta^\star,\theta')\right)\right\},$$

subject to $\sum_{a\in\mathcal{A}} \eta(x,a) = z$ for each context $x\in\mathcal{X}$. Here $\lambda^\star(z,\theta^\star)$ is the optimal value of the Lagrange multiplier for the same problem. We distinguish two cases.

Case 1: $z < \max_{x\in\mathcal{X}} \frac{1}{\rho(x)} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)$**.** Let

$$\overline{\eta}(x,a) = z \cdot \begin{cases} \frac{\eta^\star(x,a)/\rho(x)}{\max_{x\in\mathcal{X}} \frac{1}{\rho(x)} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)} & \text{if } a \neq a^\star_{\theta^\star}(x) \\ 1 - \frac{\sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)/\rho(x)}{\max_{x\in\mathcal{X}} \frac{1}{\rho(x)} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)} & \text{otherwise} \end{cases}$$

where $\eta^\star$ is the optimal solution of (P). Since $\sum_a \overline{\eta}(x,a) = z$, we have that $u^\star(z,\theta^\star)$ is less or equal to the value of the Lagrangian for $\eta = \overline{\eta}$, i.e.,

$$u^\star(z,\theta^\star) \leq v^\star(\theta^\star) + \lambda^\star(z,\theta^\star) \left(1 - \inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta')\right),$$

where we used the fact that

$$\sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a)\Delta_{\theta^\star}(x,a) = \underbrace{\frac{z}{\max_{x\in\mathcal{X}} \frac{1}{\rho(x)} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)}}_{<1} \underbrace{\sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)\Delta_{\theta^\star}(x,a)}_{=v^\star(\theta^\star)}$$

since $\Delta_{\theta^\star}(x, a^\star_{\theta^\star}(x)) = 0$. Since the KL divergence $d_{x,a}(\theta^\star,\theta')$ is lower-bounded by zero, in case 1 we have

$$u^\star(z,\theta^\star) \leq v^\star(\theta^\star) + \lambda^\star(z,\theta^\star).$$

Case 2: $z \geq \max_{x\in\mathcal{X}} \frac{1}{\rho(x)} \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)$**.** Let

$$\overline{\eta}(x,a) = \begin{cases} \eta^\star(x,a)/\rho(x) & \text{if } a \neq a^\star_{\theta^\star}(x) \\ z - \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)/\rho(x) & \text{otherwise} \end{cases}$$

where, as before, $\eta^\star$ is the optimal solution of (P). Since $z \geq \sum_{a\neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)/\rho(x)$ for any $x \in \mathcal{X}$, $\overline{\eta}$ is well defined. Since $\overline{\eta}$ also sums to z for each context, we have that $u^\star(z,\theta)$ is less or equal to the value of the Lagrangian for $\eta = \overline{\eta}$, i.e.,

$$u^\star(z,\theta^\star) \leq v^\star(\theta^\star) + \lambda^\star(z,\theta^\star)\left(1 - \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta')\right).$$

We first lower bound the infimum on the right hand side. We have

$$\inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta') = \min\Bigg\{ \underbrace{\inf_{\theta'\in\widetilde{\Theta}_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta')}_{I_{\widetilde{\Theta}_{\text{alt}}}}, \underbrace{\inf_{\theta'\in\Theta_{\text{alt}}\setminus\widetilde{\Theta}_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta')}_{I_{\Theta_{\text{alt}}\setminus\widetilde{\Theta}_{\text{alt}}}} \Bigg\}. \tag{6}$$

By definition of $\overline{\eta}$ and $\eta^\star$, the infimum over the set of confusing models can be written as

$$I_{\widetilde{\Theta}_{\text{alt}}} = \inf_{\theta'\in\widetilde{\Theta}_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta') = \inf_{\theta'\in\widetilde{\Theta}_{\text{alt}}} \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a) d_{x,a}(\theta^\star,\theta') \geq 1, \tag{7}$$

where the equality holds since the KLs are zero in the optimal arms, which are the only arms where the values of $\overline{\eta}$ differ from those of $\eta^\star$, and the inequality holds since $\eta^\star$ is feasible. Regarding the infimum over the non-confusing models,

$$I_{\Theta_{\text{alt}}\setminus\widetilde{\Theta}_{\text{alt}}} = \inf_{\theta'\in\Theta_{\text{alt}}\setminus\widetilde{\Theta}_{\text{alt}}} \left(\underbrace{\sum_{x\in\mathcal{X}} \rho(x)\overline{\eta}(x,a^\star_x) d_{x,a^\star_x}(\theta^\star,\theta')}_{(i)} + \underbrace{\sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a) d_{x,a}(\theta^\star,\theta')}_{(ii)} \right). \tag{8}$$

We partition the set of non-confusing models in two subsets:

$$\widetilde{\Theta}^{(1)}_{\text{alt}} := \left\{\theta' \in \Theta_{\text{alt}} \setminus \widetilde{\Theta}_{\text{alt}} : \forall x \in \mathcal{X}, |\mu^\star_{\theta^\star}(x) - \mu_{\theta'}(x, a^\star_{\theta^\star}(x))| < \epsilon_z\right\}, \tag{9}$$

$$\widetilde{\Theta}^{(2)}_{\text{alt}} := \left\{\theta' \in \Theta_{\text{alt}} \setminus \widetilde{\Theta}_{\text{alt}} : \exists x \in \mathcal{X}, |\mu^\star_{\theta^\star}(x) - \mu_{\theta'}(x, a^\star_{\theta^\star}(x))| \geq \epsilon_z\right\}. \tag{10}$$

The value of ϵ_z will be specified later. We have, for $\theta'' \in \widetilde{\Theta}_{\text{alt}}$,

$$\inf_{\theta'\in\widetilde{\Theta}^{(1)}_{\text{alt}}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star,\theta') \overset{(a)}{\geq} \inf_{\theta'\in\widetilde{\Theta}^{(1)}_{\text{alt}}} \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a) d_{x,a}(\theta^\star,\theta') \tag{11}$$

$$\overset{(b)}{\geq} \inf_{\theta'\in\widetilde{\Theta}^{(1)}_{\text{alt}}} \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a) \left(d_{x,a}(\theta^\star,\theta'') - \frac{1}{\sigma^2}|\mu_{\theta'}(x,a) - \mu_{\theta''}(x,a)| \right) \tag{12}$$

$$\overset{(c)}{\geq} 1 - \frac{1}{\sigma^2} \sup_{\theta'\in\widetilde{\Theta}^{(1)}_{\text{alt}}} \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a) |\mu_{\theta'}(x,a) - \mu_{\theta''}(x,a)| \tag{13}$$

$$\overset{(d)}{\geq} 1 - \frac{c_\Theta}{\sigma^2} \sup_{\theta'\in\widetilde{\Theta}^{(1)}_{\text{alt}}} \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a) \underbrace{|\mu^\star_{\theta^\star}(x) - \mu_{\theta'}(x,a^\star_x)|}_{<\epsilon_z} \tag{14}$$

$$\overset{(e)}{\geq} 1 - \frac{c_\Theta \epsilon_z}{\sigma^2} \sum_{x\in\mathcal{X}} \sum_{a\neq a^\star_x} \eta^\star(x,a), \tag{15}$$

where (a) uses the fact that $(i) \geq 0$ and the definition of $\overline{\eta}$, (b) uses the Lipschitz property of the KL divergence between Gaussians, (c) uses the fact that $\overline{\eta}$ is feasible for confusing models (see Eq. 7), (d) uses Prop. 2 and (e) uses the definition of $\widetilde{\Theta}^{(1)}_{\text{alt}}$. Regarding the second set of alternative models,

$$\inf_{\theta' \in \widetilde{\Theta}^{(2)}_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star, \theta') \tag{16}$$

$$\overset{(f)}{\geq} \inf_{\theta' \in \widetilde{\Theta}^{(2)}_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \left(z - \sum_{a \neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)/\rho(x) \right) d_{x,a^\star_x}(\theta^\star, \theta') \tag{17}$$

$$\overset{(g)}{=} \inf_{\theta' \in \widetilde{\Theta}^{(2)}_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \left(z - \sum_{a \neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)/\rho(x) \right) \frac{1}{2\sigma^2} \underbrace{(\mu_{\theta^\star}(x, a^\star_x) - \mu_{\theta'}(x, a^\star_x))^2}_{\geq \epsilon_z^2} \tag{18}$$

$$\overset{(k)}{=} \frac{\epsilon_z^2}{2\sigma^2} \left(z - \sum_{x \in \mathcal{X}} \sum_{a \neq a^\star_{\theta^\star}(x)} \eta^\star(x,a) \right). \tag{19}$$

where (f) uses the fact that $(ii) \geq 0$ and the definition of $\overline{\eta}$, (g) uses the definition of KL for Gaussian distributions and (k) uses the definition of $\widetilde{\Theta}^{(2)}_{\text{alt}}$. Let $z^\star(\theta^\star) := \sum_{x \in \mathcal{X}} \sum_{a \neq a^\star_{\theta^\star}(x)} \eta^\star(x,a)$. Putting together the results so far, we have

$$\inf_{\theta' \in \Theta_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star, \theta') \geq \min \left\{ 1, 1 - \frac{c_\Theta \epsilon_z z^\star(\theta^\star)}{\sigma^2}, \frac{\epsilon_z^2}{2\sigma^2} \Big(z - z^\star(\theta^\star) \Big) \right\}. \tag{20}$$

Setting $\epsilon_z = \sqrt{\frac{2\sigma^2}{z}}$,

$$\inf_{\theta' \in \Theta_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \overline{\eta}(x,a) d_{x,a}(\theta^\star, \theta') \geq \max \left\{ \min \left\{ 1 - \frac{c_\Theta \sqrt{2} z^\star(\theta^\star)}{\sigma \sqrt{z}}, 1 - \frac{z^\star(\theta^\star)}{z} \right\}, 0 \right\}, \tag{21}$$

Therefore, in case 2 we have

$$u^\star(z, \theta^\star) \leq v^\star(\theta^\star) + \lambda^\star(z, \theta^\star) \min \left\{ \max \left\{ \frac{c_\Theta \sqrt{2} z^\star(\theta^\star)}{\sigma \sqrt{z}}, \frac{z^\star(\theta^\star)}{z} \right\}, 1 \right\}.$$

Bounding $\lambda^\star(z, \theta^\star)$. Finally, we show that the optimal multiplier $\lambda^\star(z, \theta^\star)$ is bounded (regardless of which case z falls into). Let $\underline{\eta} = z\underline{\omega}$, where $\underline{\omega} = \omega^\star_{\underline{z}, \theta^\star}$ is the pure-exploration solution obtained solving problem (P_z) with $\underline{z}(\theta^\star)$. Recall from the first statement of Lem. 1 that

$$\inf_{\theta' \in \Theta_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \underline{\omega}(x,a) d_{x,a}(\theta^\star, \theta') = \frac{1}{\underline{z}(\theta^\star)}.$$

Thus, $\underline{\eta}$ is strictly feasible for problem ($\tilde{\text{P}}_z$) and has constraint value

$$\inf_{\theta' \in \Theta_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \underline{\eta}(x,a) d_{x,a}(\theta^\star, \theta') = \frac{z}{\underline{z}(\theta^\star)} > 1 \tag{22}$$

since $z > \underline{z}(\theta^\star)$ by assumption. Using the Slater's condition (see e.g., Lem. 3 in [22]),

$$0 \leq \lambda^\star(z, \theta^\star) \leq \frac{\sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \Delta_{\theta^\star}(x,a) (\underline{\eta}(x,a) - \eta^\star_z(x,a))}{\inf_{\theta' \in \Theta_{\text{alt}}} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \underline{\eta}(x,a) d_{x,a}(\theta^\star, \theta') - 1} \tag{23}$$

$$\leq \frac{z}{\frac{z}{\underline{z}(\theta^\star)} - 1} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \underbrace{\Delta_{\theta^\star}(x,a)}_{\geq 0} \Big(\underline{\omega}(x,a) - \underbrace{\eta^\star_z(x,a)/z}_{\geq 0} \Big) \tag{24}$$

$$\leq \frac{z}{\frac{z}{\underline{z}(\theta^\star)} - 1} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \underbrace{\Delta_{\theta^\star}(x,a)}_{\in [0, 2BL]} \underline{\omega}(x,a) \leq 2BL \frac{z \underline{z}(\theta^\star)}{z - \underline{z}(\theta^\star)}. \tag{25}$$

□

C.2 Discussion About Problem (P_z)

In this section we provide more intuition about the effect of explicitly adding the context distribution in the formulation of the lower bound. As mentioned in Sect. 3 the infimum in the original problem (P) may not be attainable, thus making it difficult to solve it and build a learning algorithm around it. A simple way to address this issue is to introduce a global constraint so that the sum of η is constrained to a parameter z. This leads to the optimization

$$\begin{aligned} \inf_{\eta(x,a)\geq 0} \quad & \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a)\Delta_{\theta^\star}(x,a) \\ \text{s.t.} \quad & \inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a) d_{x,a}(\theta^\star,\theta') \geq 1 \\ & \sum_{x,a}\eta(x,a) = z \end{aligned} \tag{$\widetilde{P}_z$}$$

Let $\widetilde{\eta}_z^\star$ be the optimal solution of ($\widetilde{P}_z$) and $\widetilde{u}_z^\star$ be its associated optimal value. On the other hand, the problem (P_z) we propose can be easily rewritten as

$$\begin{aligned} \inf_{\eta(x,a)\geq 0} \quad & \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a)\Delta_{\theta^\star}(x,a) \\ \text{s.t.} \quad & \inf_{\theta'\in\Theta_{\mathrm{alt}}} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \eta(x,a) d_{x,a}(\theta^\star,\theta') \geq 1 \\ & \sum_{a}\eta(x,a) = z\rho(x) \quad \forall x\in\mathcal{X} \end{aligned} \tag{P_z}$$

where the constraint is now on each context and it depends on the context distribution ($\omega(x,a) = \frac{\eta(x,a)}{z\rho(x)}$).[10] The crucial difference w.r.t. ($\widetilde{P}_z$) is that now the number of samples prescribed by η needs to be "compatible" with the amount of samples that can be collected within z steps from each context x depending on its probability $\rho(x)$. Let $\eta_z^\star$ be the optimal solution of (P_z) and $u_z^\star$ be its associated objective value. In order to understand how this difference may translate into a different behavior when integrated in an actual algorithm, let compare the two solutions $\widetilde{\eta}_z^\star$ and $\eta_z^\star$ if executed for z steps.[11] Since neither of them can be "played" (i.e., only one arm can be selected at each step), we need to define a specific *execution strategy* to "realize" an allocation η. For the ease of exposition, let consider a simple strategy where in each context x, an arm a is pulled at random proportionally to $\eta(x,a)$. Let $\widetilde{\zeta}_z(x,a)$ and $\zeta_z(x,a)$ the expected number of samples generated in each context-arm pair (x,a) when sampling from $\widetilde{\eta}_z^\star$ and $\eta_z^\star$ respectively. Then we have

$$\widetilde{\zeta}_z(x,a) = \widetilde{\eta}_z^\star(x,a)\, \overbrace{\frac{z\rho(x)}{\sum_{a'}\widetilde{\eta}_z^\star(x,a')}}^{\text{mismatch } \alpha_z(x,\,a)} \tag{26}$$

$$\zeta_z(x,a) = z\rho(x)\frac{\eta_z^\star(x,a)}{\sum_{a'}\eta_z^\star(x,a')} = \eta_z^\star(x,a) \tag{27}$$

which reveals how $\widetilde{\eta}_z^\star(x,a)$, which was explicitly optimized under the constraint that the total number of samples was z, may not really be "realizable" in practice, since it ignores the context distribution and the number of samples that can be actually generated at each context x. On the other hand, on average the desired allocation $\eta_z^\star$ can always be realized within z steps. Interestingly, the mismatch between $\widetilde{\eta}_z^\star(x,a)$ and $\widetilde{\zeta}_z(x,a)$ would no longer guarantee neither the performance $\widetilde{u}_z^\star$ "promised" by $\widetilde{\eta}_z^\star$ nor the feasibility for ($\widetilde{P}_z$) (i.e., $\widetilde{\zeta}_z(x,a)$ may not satisfy the KL-information constraint). This would make considerably more difficult to build a learning algorithm on $\widetilde{\eta}_z^\star$ than on $\eta_z^\star$.

As it can be noticed in Eq. 26, the level mismatch is due to the execution strategy used to realize the allocation $\widetilde{\eta}_z^\star$ (in this case, a simple sampling approach) and better solutions may exist. We could even consider to directly optimize the execution strategy so as to achieve a mismatch $\alpha_z(x,a)$ that induce

[10] Notice that the constraint directly implies $\sum_{x,a}\eta(x,a) = z$.

[11] We recall that, as discussed in Sect. 3, z introduces a more finite-time flavor into the lower bound, where pulls should now be allocated so as to satisfy the KL-information constraint within z steps.

an allocation $\widetilde{\zeta}_z(x,a)$ that performs best in terms of regret minimization under the KL-information constraint. Given the $\widetilde{\eta}_z^\star$ obtained from $(\widetilde{P}_z)$, we define the optimization problem

$$\begin{aligned} \inf_{\alpha(x,a)\geq 0} \quad & \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \widetilde{\eta}_z^\star(x,a)\alpha(x,a)\Delta_\theta(x,a) \\ \text{s.t.} \quad & \inf_{\theta'\in\Theta_{\text{alt}}} \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} \widetilde{\eta}_z^\star(x,a)\alpha(x,a)d_{x,a}(\theta,\theta') \geq 1 \\ & \sum_a \widetilde{\eta}_z^\star(x,a)\alpha(x,a) = z\rho(x) \end{aligned} \tag{$\widetilde{P}_\alpha$}$$

Interestingly, a simple change of variables reveals that $(\widetilde{P}_\alpha)$ does coincide with (P_z) that we originally introduced (i.e., $\alpha^\star(x,a) = \frac{\eta_z^\star(x,a)}{\widetilde{\eta}_z^\star(x,a)}$ minimizes the problem). This illustrates that solving (P_z) indeed leads to the optimal allocation compatible with the context distribution and the constraint of z realizations.

D Lagrangian Formulation

We discuss in more details the Lagrangian formulation presented in Section 3. Consider the following variant of (P_z):

$$\max_{\omega\in\Omega} \quad \mathbb{E}_\rho\Big[\sum_{a\in\mathcal{A}} \omega(x,a)\mu_{\theta^\star}(x,a)\Big] \quad \text{s.t.} \quad \inf_{\theta'\in\Theta_{\text{alt}}} \mathbb{E}_\rho\Big[\sum_{a\in\mathcal{A}} \omega(x,a)d_{x,a}(\theta^\star,\theta')\Big] \geq 1/z \tag{$\overline{P}_z$}$$

This problem differs from (P_z) since we replaced the action gaps with the means in the objective function and avoided scaling the latter by z. Let $\overline{\omega}_{z,\theta^\star}^\star$ the optimal solution of $(\overline{P}_z)$ and $\overline{u}^\star(z,\theta^\star)$ be its associated value (if the problem is unfeasible we set $\overline{u}^\star(z,\theta^\star) = +\infty$). Since the feasibility set is equivalent in (P_z) and $(\overline{P}_z)$ as we only changed the objective function, the following proposition is immediate.

Proposition 3. *The following properties hold:*

1. *Both* (P_z) *and* $(\overline{P}_z)$ *are feasible for* $z \geq \underline{z}(\theta^\star)$*;*
2. $\omega_{z,\theta^\star}^\star = \overline{\omega}_{z,\theta^\star}^\star$.
3. $u^\star(z,\theta^\star) = z\,(\mu^\star - \overline{u}^\star(z,\theta^\star))$ *where* $\mu^\star = \mathbb{E}_\rho[\mu_{\theta^\star}^\star(x)]$*;*

Due to the equivalence demonstrated in Prop. 3, in the remaining we shall occasionally write $\omega_z^\star$ to denote both $\omega_{z,\theta^\star}^\star$ and $\overline{\omega}_{z,\theta^\star}^\star$.

We recall the Lagrangian relaxation problem of Sec. 3. For any $\omega \in \Omega$, let $f(\omega;\theta^\star)$ denote the objective function and $g(\omega,z;\theta^\star)$ denote the KL constraint

$$f(\omega;\theta^\star) = \mathbb{E}_\rho\Big[\sum_{a\in\mathcal{A}} \omega(x,a)\mu_{\theta^\star}(x,a)\Big], \;\; g(\omega;z,\theta^\star) = \inf_{\theta'\in\Theta_{\text{alt}}} \mathbb{E}_\rho\Big[\sum_{a\in\mathcal{A}} \omega(x,a)d_{x,a}(\theta^\star,\theta')\Big] - \frac{1}{z}.$$

The Lagrangian relaxation problem of $(\overline{P}_z)$ is[12]

$$\min_{\lambda\geq 0}\max_{\omega\in\Omega} \Big\{h(\omega,\lambda;z,\theta^\star) := f(\omega;\theta^\star) + \lambda g(\omega;z,\theta^\star)\Big\}, \tag{P_λ}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ is a multiplier. We denote by $\lambda^\star(z,\theta^\star)$ the optimal multiplier for problem (P_λ). We note that f is linear in ω, while g is concave since it is an infimum of affine functions. Hence, the maximization in (P_λ) is a non-smooth concave optimization problem.

[12] In the main text we actually state that (P_λ) is the Lagrangian relaxation of (P_z) instead of $(\overline{P}_z)$. This is motivated by the fact that (P_λ) and (P_z) have the same optimal solution (see Prop. 3), though different optimal objective values.

Strong duality. We now verify that strong duality holds for the Lagrangian formulation (P_λ) (with respect to ($\overline{P}_z$)) when $z > \underline{z}(\theta^\star)$. This is immediate from the existence of a Slater point, as shown in the following proposition.

Proposition 4 (Slater Condition)**.** *For any $z > \underline{z}(\theta^\star)$, there exists a* strictly *feasible solution $\underline{\omega}$, i.e., $g(\underline{\omega}; z, \theta^\star) > 0$.*

Proof. This is a direct consequence of the fact that

$$\max_{\omega \in \Omega} \inf_{\theta' \in \Theta_{\mathrm{alt}}} \mathbb{E}_\rho \left[\sum_{a \in \mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star, \theta') \right] = \frac{1}{\underline{z}(\theta^\star)} > \frac{1}{z}. \tag{28}$$

See Lem. 1 and App. C. □

Thus, the optimal solution of (P_λ) is $(\lambda^\star(z, \theta^\star), \omega_z^\star)$.

Boundedness of the optimal multipliers. We recall the following basic result.

Lemma 2 (Lemma 3 of [22])**.** *For any $z > \underline{z}(\theta^\star)$, if $\overline{\omega}_z$ is a Slater point for ($\overline{P}_z$),*

$$\lambda^\star(z, \theta^\star) \leq \frac{f(\omega_z^\star; \theta^\star) - f(\overline{\omega}_z; \theta^\star)}{g(\overline{\omega}_z; z, \theta^\star)}$$

Using Lemma 2, we can prove the following result which will be very useful for the regret analysis.

Lemma 3. *For any $z \geq 2\underline{z}(\theta^\star)$,*

$$\lambda^\star(z, \theta^\star) \leq 2BL\underline{z}(\theta^\star). \tag{29}$$

Proof. From Prop. 4, $\underline{\omega}$ (the solution of the associated pure-exploration problem) is a Slater point for problem (P_z). Then, by Lemma 2,

$$\lambda^\star(z, \theta^\star) \leq \frac{f(\omega_z^\star; \theta^\star) - f(\underline{\omega}; \theta^\star)}{g(\underline{\omega}; z, \theta^\star)}.$$

Let $\mathrm{kl}(\omega)$ denote the expected KL of ω, so that $g(\omega; z, \theta^\star) = \mathrm{kl}(\omega) - 1/z$. Then,

$$\frac{f(\omega_z^\star; \theta^\star) - f(\underline{\omega}; \theta^\star)}{\mathrm{kl}(\underline{\omega}) - 1/z} \leq \frac{f(\omega_z^\star; \theta^\star)}{\mathrm{kl}(\underline{\omega}) - 1/z} \leq \frac{BL}{\mathrm{kl}(\underline{\omega}) - 1/z}. \tag{30}$$

Furthermore, since $\mathrm{kl}(\underline{\omega}) = 1/\underline{z}(\theta^\star)$,

$$\lambda^\star(z, \theta^\star) \leq \frac{BLz\underline{z}(\theta^\star)}{z - \underline{z}(\theta^\star)} \leq 2BL\underline{z}(\theta^\star),$$

where the last inequality holds for $z \geq 2\underline{z}(\theta^\star)$. This concludes the proof. □

E Action Sampling

SOLID does not use standard tracking approaches for action selection (e.g., cumulative tracking [14, 18] or direct tracking [15, 16]) but a sampling strategy. Despite being simpler and more practical than tracking, we show that sampling from ω_t enjoys nice theoretical guarantees.

In the following lemmas we define the filtration $\mathcal{F}_t$ as the σ-algebra generated by the t-step history, $H_t = (X_1, A_1, Y_1, \ldots, X_t, A_t, Y_t)$.

Lemma 4. *Let $\{\omega_t\}_{t \geq 1}$ be such that $\omega_t \in \Omega$ and ω_t is $\mathcal{F}_{t-1}$-measurable. Let $\{X_t\}_{t \geq 1}$ be a sequence of i.i.d. contexts distributed according to ρ and $\{A_t\}_{t \geq 1}$ be such that $A_t \sim \omega_t(X_t, \cdot)$. Then,*

$$\sum_{t \geq 1} \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \mathbb{P}\left\{ E_t, \left| N_t^E(x,a) - \rho(x) \sum_{s \leq t : E_s} \omega_s(x,a) \right| > \sqrt{\frac{S_t}{2} \log\left(S_t^2 |\mathcal{X}||\mathcal{A}|\right)} \right\} \leq \frac{\pi^2}{3}.$$

Proof. Let $Z_t := \mathbb{1}\{X_t = x, A_t = a\}$ and τ_s be a random variable such that the s-th exploration round occurs at time $\tau_s + 1$. Notice that $\{\tau_s\}_{s\geq 1}$ is a strictly-increasing sequence (i.e., $\tau_{s+1} > \tau_s$) of stopping times w.r.t. $\{\mathcal{F}_t\}_{t\geq 1}$. Furthermore, define

$$W_s := Z_{\tau_s+1} - \rho(x)\omega_{\tau_s+1}(x,a)$$

and let $\mathcal{G}_s := \mathcal{F}_{\tau_{s+1}}$. Using Lem. 10 in [26], we have that $\{W_s, \mathcal{G}_s\}_{s\geq 1}$ is a martingale difference sequence. Therefore, by Azuma's inequality

$$\mathbb{P}\left\{\left|\sum_{i=1}^{s} W_i\right| > \sqrt{\frac{s}{2}\log\frac{2}{\delta}}\right\} \leq \delta.$$

Let $a_t := \sqrt{\frac{S_t}{2}\log\left(S_t^2|\mathcal{X}||\mathcal{A}|\right)}$ and rewrite $N_t^E(x,a) = \sum_{s\leq t: E_s} Z_s$. Fix any $\bar{t} \geq 1$. Then,

$$\begin{aligned}
&\sum_{t=1}^{\bar{t}} \mathbb{1}\left\{E_t, \left|\sum_{s\leq t:E_s} (Z_s - \rho(x)\omega_s(x,a))\right| > a_t\right\} \\
&\quad\leq \sum_{s\geq 1} \mathbb{1}\left\{\left|\sum_{i=1}^{s} (Z_{\tau_i+1} - \rho(x)\omega_{\tau_i+1}(x,a))\right| > a_{\tau_s+1}, \tau_s + 1 \leq \bar{t}\right\} \\
&\quad\leq \sum_{s\geq 1} \mathbb{1}\left\{\left|\sum_{i=1}^{s} W_i\right| > \sqrt{\frac{s}{2}\log\left(s^2|\mathcal{X}||\mathcal{A}|\right)}\right\}.
\end{aligned}$$

In the last inequality, we used the fact that $a_{\tau_s+1} = \sqrt{s\log s}$. Taking expectations and applying Azuma's inequality with $\delta = \frac{2}{s^2|\mathcal{X}||\mathcal{A}|}$,

$$\sum_{t=1}^{\bar{t}} \mathbb{P}\left\{E_t, \left|\sum_{s\leq t:E_s} (Z_s - \rho(x)\omega_s(x,a))\right| > a_t\right\} \leq \frac{1}{|\mathcal{X}||\mathcal{A}|}\sum_{s\geq 1}\frac{2}{s^2} = \frac{\pi^2}{3|\mathcal{X}||\mathcal{A}|}.$$

The results holds for all $\bar{t}$, and the proof is concluded by summing over contexts and arms. □

Lemma 5. *Let $\{\omega_t\}_{t\geq 1}$ be such that $\omega_t \in \Omega$ and ω_t is $\mathcal{F}_{t-1}$-measurable. Let $\{X_t\}_{t\geq 1}$ be a sequence of i.i.d. contexts distributed according to ρ and $\{A_t\}_{t\geq 1}$ be such that $A_t \sim \omega_t(X_t, \cdot)$. Let $\{\varphi_t^i\}_{t\geq 1, i\in[m]}$ be a sequence of functions $\varphi_t^i : \mathcal{X}\times\mathcal{A} \to [-b, b]$ such that $\varphi_t^i(x,a)$ is $\mathcal{F}_{t-1}$-measurable for all $i \in [m]$. Then,*

$$\sum_{t\geq 1}\sum_{i=1}^{m} \mathbb{P}\left\{E_t, \left|\sum_{s\leq t:E_s}\left(\varphi_s^i(X_s,A_s) - \sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\varphi_s^i(x,a)\right)\right| > b\sqrt{\frac{S_t}{2}\log(mS_t^2)}\right\} \leq \frac{\pi^2}{3}.$$

Proof. The proof follows the same steps as the one of Lemma 4. Fix $i \in [m]$. Let $Z_t := \varphi_t^i(X_t, A_t)$ and τ_s be a random variable such that the s-th exploration round occurs at time $\tau_s + 1$. Notice that $\{\tau_s\}_{s\geq 1}$ is a strictly-increasing sequence (i.e., $\tau_{s+1} > \tau_s$) of stopping times w.r.t. $\{\mathcal{F}_t\}_{t\geq 1}$. Furthermore, define

$$W_s := Z_{\tau_s+1} - \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\rho(x)\omega_{\tau_s+1}(x,a)\varphi_{\tau_s+1}^i(x,a)$$

and let $\mathcal{G}_s := \mathcal{F}_{\tau_{s+1}}$. Using Lem. 10 in [26], we have that $\{W_s, \mathcal{G}_s\}_{s\geq 1}$ is a martingale difference sequence (with differences bounded by b). Therefore, by Azuma's inequality

$$\mathbb{P}\left\{\left|\sum_{i=1}^{s} W_i\right| > b\sqrt{\frac{s}{2}\log\frac{2}{\delta}}\right\} \leq \delta.$$

Let $a_t := b\sqrt{\frac{S_t}{2}\log\left(mS_t^2\right)}$ and fix some $\bar{t} \geq 1$. Then,

$$\begin{aligned}
&\sum_{t=1}^{\bar{t}} \mathbb{1}\left\{E_t, \left|\sum_{s\leq t: E_s}\left(Z_s - \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\rho(x)\omega_s(x,a)\varphi_s^i(x,a)\right)\right| > a_t\right\} \\
&\quad\leq \sum_{s\geq 1}\mathbb{1}\left\{\left|\sum_{j=1}^{s}\left(Z_{\tau_j+1} - \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\rho(x)\omega_{\tau_j+1}(x,a)\varphi_{\tau_j+1}^i(x,a)\right)\right| > a_{\tau_s+1}, \tau_s+1\leq\bar{t}\right\} \\
&\quad\leq \sum_{s\geq 1}\mathbb{1}\left\{\left|\sum_{j=1}^{s} W_j\right| > b\sqrt{\frac{s}{2}\log(ms^2)}\right\}.
\end{aligned}$$

In the last inequality, we used the fact that $a_{\tau_s+1} = b\sqrt{\frac{s}{2}\log(ms^2)}$. Taking expectations and applying Azuma's inequality with $\delta = \frac{2}{ms^2}$,

$$\sum_{t=1}^{\bar{t}}\mathbb{P}\left\{E_t, \left|\sum_{s\leq t: E_s}\left(Z_s - \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}}\rho(x)\omega_s(x,a)\varphi_s^i(x,a)\right)\right| > a_t\right\} \leq \sum_{s\geq 1}\frac{2}{ms^2} = \frac{\pi^2}{3m}.$$

The results holds for all $\bar{t}$ and the proof follows by summing over all $i \in [m]$. □

Discussion. Lemma 4 provides an analogous result to those obtained by tracking strategies, where the empirical pull counts are shown close to the sequence of conditional probabilities computed by the optimizer. Despite being simpler, our sampling rule achieves similar efficiency as existing tracking rules. In particular, our bound scales with $\log|\mathcal{A}|$, a factor that appears in the tightest known analysis of cumulative tracking [17]. The factor $\sqrt{S_t\log S_t}$ is not typically found in tracking strategies for MABs. However, we note that such dependency would naturally appear when generalizing these strategies to the contextual case.

Lemma 5 extends Lemma 4 to bound the deviation between expectations of measurable functions under the sequence of conditional probabilities and the same functions evaluated at the observed contexts/arms. This result will be very useful in the regret analysis to avoid undesirable linear dependencies on the number of arms.

F High-Probability Events

In this section, we report the high-probability events used through the paper. Refer to App. I.1 for concentration inequalities.

Let $\Phi_{x,a} := \phi(x,a)\phi(x,a)^T$. We define the following events:

true regret close to objective values

$$G_t^{\Delta} := \left\{ \left| \sum_{s \le t : E_s} \left(\Delta_{\theta^\star}(X_s, A_s) - \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_s(x,a) \Delta_{\theta^\star}(x,a) \right) \right| \le 2LB\sqrt{S_t \log S_t} \right\}, \tag{31}$$

true confidence intervals close to expected confidence intervals

$$G_t^{\phi} := \left\{ \left| \sum_{s \le t : E_s} \left(\|\phi(X_s, A_s)\|_{\bar{V}_{s-1}^{-1}} - \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_s(x,a) \|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}} \right) \right| \le \frac{L}{\nu} \sqrt{S_t \log S_t} \right\}, \tag{32}$$

true design matrix close to expected design matrix

$$G_t^{d} := \left\{ \left\| \sum_{s \le t : E_s} \left(\Phi_{X_s, A_s} - \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_s(x,a) \Phi_{x,a} \right) \right\|_{\infty} \le L^2 \sqrt{S_t \log(dS_t)} \right\}, \tag{33}$$

well-estimated context distribution

$$G_t^{\rho} := \left\{ \forall x \in \mathcal{X} : |\widehat{\rho}_{t-1}(x) - \rho(x)| \le 2 \max\left(\sqrt{\frac{\log(|\mathcal{X}| S_t^2)}{2S_t}}, \frac{2}{t} \right) \right\}, \tag{34}$$

well-estimated parameters

$$G_t^{\theta} := \left\{ \|\widehat{\theta}_{t-1} - \theta^\star\|_{\bar{V}_{t-1}} \le \sqrt{\gamma_t} \right\}. \tag{35}$$

Furthermore, we define $G_t := \{G_t^{\Delta}, G_t^{\phi}, G_t^{d}, , G_t^{\rho}, G_t^{\theta}\}$ as the *"good" event* and let $M_t = \sum_{s=1}^{t} \mathbb{1}\{E_s, \neg G_s\}$ be the number of exploration rounds in which the good event does not hold. This can be bounded in expectation as follows.

Lemma 6. *Let $M_t = \sum_{s=1}^{t} \mathbb{1}\{E_s, \neg G_s\}$ be the number of exploration rounds in which the good event does not hold, then*

$$\mathbb{E}[M_t] \le \frac{3\pi^2}{2}.$$

Proof. Using the definition of G_s together with the union bound,

$$\mathbb{E}[M_t] = \sum_{s=1}^{t} \mathbb{P}\{E_s, \neg G_s\} \le \sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\Delta}\right\} + \sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\phi}\right\} + \sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{d}\right\} + \sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\rho}\right\} + \sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\theta}\right\}.$$

The first and second term can be bounded by Lemma 5 by noticing that $\Delta_{\theta^\star}(x,a) \le 2LB$ and that $\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}$ is $\mathcal{F}_{s-1}$-measurable and upper-bounded by $\frac{L}{\nu}$ at all time steps. Thus,

$$\sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\Delta}\right\} + \sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\phi}\right\} \le \frac{2\pi^2}{3}.$$

Similarly, the third term can be bounded by Lemma 5 by taking a union bound over all elements of $\Phi_{x,a}$ (for a total of d^2 elements) and noting that each term is bounded by L^2. Thus,

$$\sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{d}\right\} \le \frac{\pi^2}{3}.$$

The fourth term is

$$\begin{aligned}
\sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\rho}\right\} &\leq \sum_{x\in\mathcal{X}} \sum_{s\geq 1} \mathbb{P}\left\{E_s, |\widehat{\rho}_{s-1}(x) - \rho(x)| > 2\max\left(\sqrt{\frac{\log(|\mathcal{X}|S_s^2)}{2S_s}}, \frac{2}{s}\right)\right\} \\
&\leq \sum_{x\in\mathcal{X}} \sum_{s\geq 1} \mathbb{P}\left\{E_s, |\widehat{\rho}_{s-1}(x) - \widehat{\rho}_s(x)| + |\widehat{\rho}_s(x) - \rho(x)| > 2\max\left(\sqrt{\frac{\log(|\mathcal{X}|S_s^2)}{2S_s}}, \frac{2}{s}\right)\right\} \\
&\leq \sum_{x\in\mathcal{X}} \sum_{s\geq 1} \mathbb{P}\left\{E_s, |\widehat{\rho}_{s-1}(x) - \widehat{\rho}_s(x)| > \frac{2}{s}\right\} + \sum_{x\in\mathcal{X}} \sum_{s\geq 1} \mathbb{P}\left\{E_s, |\widehat{\rho}_s(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|S_s^2)}{2S_s}}\right\} \\
&\leq \frac{\pi^2}{3}.
\end{aligned}$$

Here we used the fact that the absolute difference between two consecutive empirical means with samples bounded by 1 cannot be larger than $\frac{2}{s}$. We also used Lemma 7 to bound the second term. Finally, the fifth term can be directly bounded by Lemma 8:

$$\sum_{s=1}^{t} \mathbb{P}\left\{E_s, \neg G_s^{\theta}\right\} \leq \frac{\pi^2}{6}.$$

Combining the five bounds concludes the proof. □

G Regret Proof

We start decomposing the regret based on whether E_t holds or not:

$$R_n = \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{\neg E_t\right\} + \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{E_t\right\} = R_n^{\text{exploit}} + R_n^{\text{explore}}.$$

Throughout the proof, as stated in the main theorem, we use $\beta_{t-1} := c_{n,1/n}$ and $\gamma_t := c_{n,1/S_t^2}$.

G.1 Outline

An outline of our proof is as follows.

Step 1. (App. G.2) Using the confidence set derived in App. J, we show that the regret suffered when the algorithm enters the exploitation step is finite;

Step 2. (App. G.3.1) Using the properties of our action sampling strategy, we reduce the regret incurred during exploration rounds to the sum of objective values of the policies computed incrementally by primal-dual gradient ascent;

Step 3. (App. G.3.2) By combining standard tools from convex optimization with the properties of our confidence intervals, we relate the sum of objective values at each phase to the corresponding optimal value and constraint violations;

Step 4. (App. G.3.3) We relate the sum of constraints to the exploitation test used by SOLID. In particular, using the fact that the algorithm is not in the exploitation step, we show that the sum of constraints cannot be larger than $\mathcal{O}(\log n)$;

Step 5. (App. G.3.4) We combine the results obtained in the previous steps to show a first bound on the expected regret suffered during the exploration rounds. Our bound has the optimal dependency on $v^\star(\theta^\star)\log n$ but scales with the expected number $\mathbb{E}\left[K_n\right]$ of phases executed by the algorithm;

Step 6. (App. G.3.5) By relating the upper bound on the sum of constraints computed at Step 3 and a lower bound on the same quantity, we obtain an upper bound on K_n as a function of the chosen sequences p_k, z_k;

Step 7. (App. G.3.6) We derive the final result by combining the bound on K_n of Step 5 using the exponential schedule for p_k, z_k with the partial regret bound of Step 4.

G.2 Regret during Exploitation

We show that the regret suffered when exploitation occurs is finite. Let $\beta_{t-1} := c_{n,1/n}$, where $c_{n,\delta}$ was defined in Thm. 1. Then $F_t := \mathbb{1}\left\{\|\widehat{\theta}_{t-1} - \theta^\star\|^2_{\overline{V}_{t-1}} \leq c_{n,1/n}\right\}$ is the event under which the true model belongs to the confidence set, which holds with probability at least $1 - 1/n$ by the same theorem. We leverage this to decompose the regret during exploitation as:

$$R_n^{\text{exploit}} = \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{\neg E_t, F_t\right\} + \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{\neg E_t, \neg F_t\right\}.$$

The expectation of the second term is bounded by

$$\mathbb{E}\left[\sum_{t=1}^{n} \underbrace{\Delta_{\theta^\star}(X_t, A_t)}_{\leq 2LB}\mathbb{1}\left\{\neg E_t, \neg F_t\right\}\right] \leq 2LB \cdot \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{1}\left\{\neg F_t\right\}\right] \leq 2LB\sum_{t=1}^{n}\underbrace{\mathbb{P}\left\{\neg F_t\right\}}_{\leq 1/n} \leq 2LB,$$

where we bounded $\mathbb{P}\left\{\neg F_t\right\} \leq \frac{1}{n}$ by using Thm. 1 with $\delta = 1/n$. Regarding the first term, we have two possible cases. If $a^\star_{\widetilde{\theta}_{t-1}}(X_t) = a^\star_{\theta^\star}(X_t)$, then the algorithm suffers no regret since by definition it pulls the empirically optimal arm (which is the optimal arm in this case). If $a^\star_{\widetilde{\theta}_{t-1}}(X_t) \neq a^\star_{\theta^\star}(X_t)$, then it must be that $\theta^\star \in \overline{\Theta}_{t-1}$, that is, the true model is in the set of alternative models for the current context. Under $\neg E_t$, this implies that

$$\|\widehat{\theta}_{t-1} - \theta^\star\|^2_{\overline{V}_{t-1}} \geq \|\widetilde{\theta}_{t-1} - \theta^\star\|^2_{\overline{V}_{t-1}} \geq \inf_{\theta' \in \overline{\Theta}_{t-1}} \|\widehat{\theta}_{t-1} - \theta'\|^2_{\overline{V}_{t-1}} > \beta_{t-1} = c_{n,1/n},$$

where the first inequality is due to the fact that the good event F_t holds and Cor. 1. This is a contradiction with respect to F_t. Therefore, $\neg E_t$ and F_t cannot hold at the same time and the algorithm suffers no regret. Combining these results, we conclude

$$\mathbb{E}\left[R_n^{\text{exploit}}\right] \leq 2LB.$$

G.3 Regret under Exploration

The key challenge is to bound the regret during the exploration rounds. We proceed by following the steps outlined in App. G.1.

G.3.1 From Regret to Objective Values

We decompose the regret incurred during exploration as

$$R_n^{\text{explore}} := \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{E_t\right\} \leq \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{E_t, G_t\right\} + 2LB\underbrace{\sum_{t=1}^{n}\mathbb{1}\left\{E_t, \neg G_t\right\}}_{:=M_n}.$$

Refer to App. F for the definition of G_t. The second term is M_n, the number of exploration rounds in which the good event does not hold, and can be bounded in expectation by using Lem. 6. The first one can be bounded by using the good event. Suppose, without loss of generality, that E_n and G_n hold (if they do not, the following reasoning can be repeated for the last time step at which these events hold). Then, using G_t^Δ (see App. F),

$$\begin{aligned}\sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t)\mathbb{1}\left\{E_t, G_t\right\} &= \sum_{t \leq n : E_t} \Delta_{\theta^\star}(X_t, A_t) \\ &\leq \sum_{t \leq n : E_t}\sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a)\Delta_{\theta^\star}(x, a) + 2LB\sqrt{S_n \log S_n}.\end{aligned}$$

Using the definition of phase, we can rewrite the first summation as

$$\sum_{t \leq n : E_t}\sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a)\Delta_{\theta^\star}(x, a) = \sum_{k=0}^{K_n}\sum_{t \in \mathcal{T}_k^E}\sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a)\Delta_{\theta^\star}(x, a).$$

Recall that K_t is the (random) phase index at time t, while $\mathcal{T}_k^E$ is the set of exploration rounds in phase k. See App. A for a summary of notation. Let $\underline{k} := \min\{k \in \mathbb{N} | z_k \geq 2\underline{z}(\theta^\star)\}$. We split the sum into phases before and after $\underline{k}$. For those before, we have

$$\sum_{k<\underline{k}} \sum_{t\in\mathcal{T}_k^E} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Delta_{\theta^\star}(x,a) \leq 2LB \sum_{k<\underline{k}} |\mathcal{T}_k^E| \leq 2LB \sum_{k<\underline{k}} p_k,$$

which yields at most finite regret since $\{p_k\}$ is increasing. Let us now fix a phase $k \geq \underline{k}$ and bound the regret during its exploration rounds ($\mathcal{T}_k^E$). Note that the optimization problem in each phase $k \geq \underline{k}$ is feasible (see App. D). We have

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_k^E} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Delta_{\theta^\star}(x,a) \\
&\quad= \sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Delta_{\theta^\star}(x,a) + \sum_{t\in\mathcal{T}_k^E:\neg G_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)(\mu^\star_{\theta^\star}(x) - \mu_{\theta^\star}(x,a)) \\
&\quad\leq \sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Delta_{\theta^\star}(x,a) + M_{n,k}\mu^\star - \sum_{t\in\mathcal{T}_k^E:\neg G_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\mu_{\theta^\star}(x,a).
\end{aligned}$$

Here we defined $\mu^\star := \sum_{x\in\mathcal{X}} \rho(x)\mu^\star_{\theta^\star}(x)$ and $M_{n,k}$ as the number of exploration rounds during phase k where the good event does not hold. The last term can be bounded by $M_{n,k}BL$. Regarding the remaining two,

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Delta_{\theta^\star}(x,a) + M_{n,k}\mu^\star \\
&= (p_k - M_{n,k})\mu^\star + M_{n,k}\mu^\star - \sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\mu_{\theta^\star}(x,a) \\
&= p_k\mu^\star + \underbrace{\sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} (\hat{\rho}_{t-1}(x) - \rho(x)) \sum_{a\in\mathcal{A}} \omega_t(x,a)\mu_{\theta^\star}(x,a)}_{(a)} - \underbrace{\sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\mu_{\theta^\star}(x,a)}_{(b)}.
\end{aligned}$$

Term (a) can be bounded by

$$(a) \leq LB \underbrace{\sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} |\hat{\rho}_{t-1}(x) - \rho(x)|}_{\zeta_{n,k}}.$$

The second term $\zeta_{n,k}$ will be bounded shortly over all phases by means of Lemma 12. We now provide a lower bound to term (b). The first step is to relate this to the objective function optimized by the algorithm. Using the definition of G_t and Lem. 10,

$$\begin{aligned}
(b) &\geq \sum_{t\in\mathcal{T}_k^E:G_t} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) \left(\mu_{\widetilde{\theta}_{t-1}}(x,a) - \sqrt{\gamma_t}\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}\right) \\
&\quad \pm \sum_{t\in\mathcal{T}_k^E:\neg G_t} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) \underbrace{\mu_{\widetilde{\theta}_{t-1}}(x,a)}_{|\cdot|\leq LB} \pm \sum_{t\in\mathcal{T}_k^E} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\sqrt{\gamma_t}\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}} \\
&\geq \sum_{t\in\mathcal{T}_k^E} f_t(\omega_t) - M_{n,k}BL - 2\sum_{t\in\mathcal{T}_k^E} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\sqrt{\gamma_t}\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}} \\
&\geq \sum_{t\in\mathcal{T}_k^E} f_t(\omega_t) - M_{n,k}BL - 2\sqrt{\gamma_n}\Psi_{n,k}. \qquad (36)
\end{aligned}$$

In the last step, we used $\sqrt{\gamma_t} \leq \sqrt{\gamma_n}$ (which is by definition $\mathcal{O}(\log S_n)$) and defined $\Psi_{n,k} := \sum_{t\in\mathcal{T}_k^E} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}$.

To wrap-up the regret bound we have obtained so far, summing over all phases,

$$R_n^{\text{explore}} \leq 2LB \sum_{k<\underline{k}} p_k + \sum_{k\geq\underline{k}}^{K_n} p_k \mu^\star + LB \underbrace{\sum_{k\geq\underline{k}}^{K_n} \zeta_{n,k}}_{\leq \zeta_n} - \sum_{k\geq\underline{k}}^{K_n} \sum_{t\in\mathcal{T}_k^E} f_t(\omega_t)$$
$$+ 2LB \underbrace{\sum_{k\geq\underline{k}}^{K_n} M_{n,k}}_{\leq M_n} + 2LBM_n + 2\sqrt{\gamma_n} \underbrace{\sum_{k\geq\underline{k}}^{K_n} \Psi_{n,k}}_{\leq \Psi_n} + 2LB\sqrt{S_n \log S_n}.$$

Here we defined

$$\zeta_n := \sum_{t\leq n: E_t, G_t} \sum_{x\in\mathcal{X}} |\hat{\rho}_{t-1}(x) - \rho(x)|$$

and

$$\Psi_n := \sum_{t\leq n: E_t} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) \|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}.$$

ζ_n can be bounded by Lemma 12 and Ψ_n by Lemma 13. Both terms are of order $\mathcal{O}(\sqrt{S_n \log S_n})$. In order to simplify notation, we keep the specific bounds implicit in the remaining. Therefore, our partial regret bound is

$$R_n^{\text{explore}} \leq 2LB \sum_{k<\underline{k}} p_k + \sum_{k\geq\underline{k}}^{K_n} p_k \mu^\star - \sum_{k\geq\underline{k}}^{K_n} \sum_{t\in\mathcal{T}_k^E} f_t(\omega_t)$$
$$+ 4LBM_n + 2\sqrt{\gamma_n}\Psi_n + LB\zeta_n + 2LB\sqrt{S_n \log S_n}. \tag{37}$$

G.3.2 Bounding the Sum of Objective Values

Our goal here is to lower bound the sum of objective values. As before, fix some phase index $k \geq \underline{k}$ and let $\lambda \geq 0$ be arbitrary. By recalling that the optimization process is reset at the beginning of each phase and using Corollary 2 with $\alpha_k^\lambda = \alpha_k^\omega = 1/\sqrt{p_k}$ and $\omega = \omega_{z_k}^\star$ (the optimal solution of problem (P_{z_k})),

$$\sum_{t\in\mathcal{T}_k^E} f_t(\omega_t) \geq \sum_{t\in\mathcal{T}_k^E} h_t(\omega_{z_k}^\star, \lambda_t, z_k) - \lambda \sum_{t\in\mathcal{T}_k^E} g_t(\omega_t, z_k) - \left(\log|\mathcal{A}| + \frac{b_\omega^2 + b_\lambda^2}{2} + \frac{(\lambda-\lambda_1)^2}{2}\right)\sqrt{p_k}. \tag{38}$$

We recall that b_λ and b_ω are the maximum sub-gradients in λ and ω, respectively. We now lower-bound the first term on the right-hand side. Since $h_t(\omega_{z_k}^\star, \lambda_t, z_k) = f_t(\omega_{z_k}^\star) + \lambda_t g_t(\omega_{z_k}^\star, z_k)$, $f_t(\omega_{z_k}^\star) \geq -LB$, $g_t(\omega_{z_k}^\star, z_k) \geq -\frac{1}{z_k}$, and $\lambda_t \leq \lambda_{\max}$, this term, evaluated on those steps where G_t does not hold, can be lower-bounded by $\sum_{t\in\mathcal{T}_k^E:\neg G_t} h_t(\omega_{z_k}^\star, \lambda_t, z_k) \geq -(LB + \lambda_{\max}/z_k)M_{n,k}$. For any step $t \in \mathcal{T}_k^E$ in which G_t holds, the optimism property (Lemma 11) yields

$$f_t(\omega_{z_k}^\star) \geq \sum_{x\in\mathcal{X}} (\hat{\rho}_{t-1}(x) - \rho(x)) \underbrace{\sum_{a\in\mathcal{A}} \omega_{z_k}^\star(x,a)\mu_{\theta^\star}(x,a)}_{|\cdot|\leq LB} + f(\omega_{z_k}^\star)$$
$$\geq f(\omega_{z_k}^\star) - LB \sum_{x\in\mathcal{X}} |\hat{\rho}_{t-1}(x) - \rho(x)|,$$

and

$$g_t(\omega_{z_k}^\star, z_k) \geq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_{z_k}^\star(x,a) d_{x,a}(\theta^\star, \theta') - \frac{1}{z_k} \pm g(\omega_{z_k}^\star)$$
$$\geq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} (\hat{\rho}_{t-1}(x) - \rho(x)) \sum_{a\in\mathcal{A}} \omega_{z_k}^\star(x,a) d_{x,a}(\theta^\star, \theta') + g(\omega_{z_k}^\star)$$
$$\geq g(\omega_{z_k}^\star) - \frac{2L^2B^2}{\sigma^2} \sum_{x\in\mathcal{X}} |\hat{\rho}_{t-1}(x) - \rho(x)|.$$

Combining these two and using $\lambda_t \le \lambda_{\max}$,

$$\sum_{t\in\mathcal{T}_k^E:G_t} h_t(\omega^\star_{z_k},\lambda_t,z_k) \ge \sum_{t\in\mathcal{T}_k^E:G_t} \big(f(\omega^\star_{z_k})+\lambda_t g(\omega^\star_{z_k})\big) - LB\left(1+\frac{2LB\lambda_{\max}}{\sigma^2}\right)\zeta_{n,k}.$$

Note that $g(\omega^\star_{z_k}) \ge 0$ since by assumption $\omega^\star_{z_k}$ is feasible for the optimization problem (P_{z_k}). Furthermore, $\sum_{t\in\mathcal{T}_k^E:G_t} f(\omega^\star_{z_k}) = \sum_{t\in\mathcal{T}_k^E} f(\omega^\star_{z_k}) - \sum_{t\in\mathcal{T}_k^E:\neg G_t} \underbrace{f(\omega^\star_{z_k})}_{|\cdot|\le LB} \ge p_k f(\omega^\star_{z_k}) - LBM_{n,k}$.
Therefore, we obtain the following lower-bound on the sum of optimal objective values:

$$\sum_{t\in\mathcal{T}_k^E} h_t(\omega^\star_{z_k},\lambda_t,z_k) \ge p_k f(\omega^\star_{z_k}) - LB\left(1+\frac{2LB\lambda_{\max}}{\sigma^2}\right)\zeta_{n,k} - (2LB+\lambda_{\max}/z_k)M_{n,k}.$$

Plugging this back into (38),

$$\begin{aligned}\sum_{t\in\mathcal{T}_k^E} f_t(\omega_t) \ge p_k f(\omega^\star_{z_k}) - \lambda\sum_{t\in\mathcal{T}_k^E} g_t(\omega_t,z_k) - a_\lambda\sqrt{p_k}\\ - LB\left(1+\frac{2LB\lambda_{\max}}{\sigma^2}\right)\zeta_{n,k} - (2LB+\lambda_{\max}/z_k)M_{n,k},\end{aligned} \tag{39}$$

where, for simplicity, we defined $a_\lambda := \left(\log|\mathcal{A}| + \frac{b_\omega^2+b_\lambda^2}{2} + \frac{(\lambda-\lambda_1)^2}{2}\right)$. Summing over all phases,

$$\begin{aligned}\sum_{k\ge\underline{k}}^{K_n}\sum_{t\in\mathcal{T}_k^E} f_t(\omega_t) \ge \sum_{k\ge\underline{k}}^{K_n} p_k f(\omega^\star_{z_k}) - \lambda\sum_{k\ge\underline{k}}^{K_n}\sum_{t\in\mathcal{T}_k^E} g_t(\omega_t,z_k) - a_\lambda\sum_{k\ge\underline{k}}^{K_n}\sqrt{p_k}\\ - LB\left(1+\frac{2LB\lambda_{\max}}{\sigma^2}\right)\zeta_n - (2LB+\lambda_{\max})M_n,\end{aligned} \tag{40}$$

where we used $\sum_{k\ge\underline{k}}^{K_n} M_{n,k} \le M_n$, $\sum_{k\ge\underline{k}}^{K_n}\zeta_{n,k} \le \zeta_n$, and $z_k \ge 1$.

G.3.3 Bounding the sum of constraints

Our next step is to upper bound $\sum_{k\ge\underline{k}}^{K_n}\sum_{t\in\mathcal{T}_k^E} g_t(\omega_t,z_k)$, the sum of constraints of the policies played by the algorithm during feasible phases (those with $z_k \ge 2\underline{z}(\theta^\star)$). The intuition is that this term cannot be large (i.e., it cannot be above $\mathcal{O}(\log n)$), otherwise the exploitation test would trigger and we would not be exploring at step n. Using the definition of $g_t(\omega,z_k)$ (Eq. 3) and splitting the sum based on the good event

$$\begin{aligned}&\sum_{k\ge\underline{k}}^{K_n}\sum_{t\in\mathcal{T}_k^E} g_t(\omega_t,z_{K_t})\\ &\le \sum_{t\le n:E_t}\inf_{\theta'\in\bar\Theta_{t-1}}\sum_{x\in\mathcal{X}}\hat\rho_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)d_{x,a}(\widetilde\theta_{t-1},\theta') + \frac{2LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n - \sum_{k\ge\underline{k}}^{K_n}\sum_{t\in\mathcal{T}_k^E}\frac{1}{z_k}\\ &\le \underbrace{\sum_{t\le n:E_t,G_t}\inf_{\theta'\in\bar\Theta_{t-1}}\sum_{x\in\mathcal{X}}\hat\rho_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)d_{x,a}(\widetilde\theta_{t-1},\theta')}_{①} + \frac{2L^2B^2}{\sigma^2}M_n + \frac{2LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n - \sum_{k\ge\underline{k}}^{K_n}\frac{p_k}{z_k}.\end{aligned}$$

Note that in the first step above we implicitly upper bounded the sum of KLs on the feasible phases with the sum of KLs over all exploration rounds. We can use the definition of G_t and the optimism

(Lemma 11) to upper bound the first sum by

$$
\begin{aligned}
① &\leq \sum_{t\leq n:E_t,G_t} \inf_{\theta'\in\bar{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) d_{x,a}(\theta^\star,\theta') + \frac{2LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n \\
&\leq \underbrace{\sum_{t\leq n:E_t,G_t} \inf_{\theta'\in\bar{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) d_{x,a}(\theta^\star,\theta')}_{②} + \frac{2L^2B^2}{\sigma^2} \underbrace{\sum_{t\leq n:E_t,G_t} \sum_{x\in\mathcal{X}} |\rho(x)-\hat{\rho}_{t-1}(x)|}_{=\zeta_n} \\
&\quad + \frac{2LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n.
\end{aligned}
$$

Furthermore, the first term can be upper bounded by replacing each set $\bar{\Theta}_{t-1}$ over which the infimum is taken by Θ_{alt} (if the two sets were different, such term would be zero). Therefore,

$$
\begin{aligned}
② &\leq \sum_{t\leq n:E_t,G_t} \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) d_{x,a}(\theta^\star,\theta') \\
&\leq \underbrace{\inf_{\theta'\in\Theta_{alt}} \sum_{t\leq n:E_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) d_{x,a}(\theta^\star,\theta')}_{③}, \qquad (41)
\end{aligned}
$$

where we moved the infimum outside the outer sum and added the remaining steps where G_t does not hold. Let $\Phi_{x,a} := \phi(x,a)\phi(x,a)^T$ and $V_{n,e} := \sum_{t\leq n:E_t} \Phi_{X_t,A_t}$ be the design matrix of the exploration rounds. Using the definition of $d_{x,a}$,

$$
\begin{aligned}
③ &= \frac{1}{2\sigma^2} \inf_{\theta'\in\Theta_{alt}} (\theta^\star-\theta')^T \left(\sum_{t\leq n:E_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Phi_{x,a} \pm V_{n,e} \right) (\theta^\star-\theta') \\
&\leq \inf_{\theta'\in\Theta_{alt}} \left\{ \frac{1}{2\sigma^2}(\theta^\star-\theta')^T V_{n,e}(\theta^\star-\theta') + \frac{1}{2\sigma^2}\|\theta^\star-\theta'\|_2^2 \left\| \sum_{t\leq n:E_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Phi_{x,a} - V_{n,e} \right\|_2 \right\} \\
&\leq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} N_n^E(x,a) d_{x,a}(\theta^\star,\theta') + \frac{2B^2}{\sigma^2} \left\| \sum_{t\leq n:E_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Phi_{x,a} - V_{n,e} \right\|_2 \\
&\leq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} N_n^E(x,a) d_{x,a}(\theta^\star,\theta') + \frac{2B^2\sqrt{d}}{\sigma^2} \left\| \sum_{t\leq n:E_t} \sum_{x\in\mathcal{X}} \rho(x) \sum_{a\in\mathcal{A}} \omega_t(x,a)\Phi_{x,a} - V_{n,e} \right\|_\infty .
\end{aligned}
$$

Recall that G_n holds. Then, by using the definition of G^d to bound the norm,

$$
③ \leq \underbrace{\inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} N_{n-1}^E(x,a) d_{x,a}(\theta^\star,\theta')}_{④} + \frac{2B^2L^2}{\sigma^2} + \frac{2B^2L^2}{\sigma^2}\sqrt{dS_n \log(dS_n)}.
$$

Here we used $N_n(x,a) = N_{n-1}(x,a) + \mathbb{1}\{X_n = x, A_n = a\}$ and upper bounded the KL at round n by its maximum value. Moreover, similarly to Lem. 11 we can show that

$$
④ \leq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} N_{n-1}^E(x,a) d_{x,a}(\widetilde{\theta}_{n-1},\theta') + \frac{2LB\sqrt{\gamma_n}}{\sigma^2} \underbrace{\sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} N_{n-1}^E(x,a)\|\phi(x,a)\|_{\bar{V}_{n-1}^{-1}}}_{\leq\Psi_n}.
$$

The upper bound on the second term can be extracted from the proof of Lemma 13. The first term can be finally related to the exploitation test:

$$\begin{aligned}\inf_{\theta' \in \Theta_{alt}} \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} N_{n-1}^E(x,a) d_{x,a}(\widetilde{\theta}_{n-1}, \theta') &\leq \inf_{\theta' \in \bar{\Theta}_{n-1}} \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} N_{n-1}^E(x,a) d_{x,a}(\widetilde{\theta}_{n-1}, \theta') \\ &= \frac{1}{2\sigma^2} \inf_{\theta' \in \bar{\Theta}_{n-1}} \|\widetilde{\theta}_{n-1} - \theta'\|_{V_{n-1}}^2 \\ &\leq \frac{1}{2\sigma^2} \inf_{\theta' \in \bar{\Theta}_{n-1}} \|\widetilde{\theta}_{n-1} - \theta'\|_{\bar{V}_{n-1}}^2 \leq \frac{\beta_{n-1}}{2\sigma^2},\end{aligned}$$

where the second-last inequality holds since $\bar{V}_{n-1} \succeq V_{n-1}$, and the last inequality holds since the algorithm is exploring at step n. By gathering all the results together, we get

$$\begin{aligned}\sum_{k \geq \underline{k}}^{K_n} \sum_{t \in \mathcal{T}_K^E : E_t} g_t(\omega_t, z_{K_t}) \leq \frac{\beta_{n-1}}{2\sigma^2} - \sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} + \frac{2L^2B^2}{\sigma^2} M_n + \frac{6LB}{\sigma^2} \sqrt{\gamma_n} \Psi_n + \frac{2B^2L^2}{\sigma^2} \zeta_n \\ + \frac{2B^2L^2}{\sigma^2} \left(\sqrt{dS_n \log(dS_n)} + 1 \right). \end{aligned} \tag{42}$$

G.3.4 Back to the regret during exploration

So far we have (1) reduced the total regret during exploration to the sum of objective values (Eq. 37), (2) related this quantity to the optimal values of each phase (Eq. 40), and (3) derived an upper bound to the total sum of constraints (Eq. 42). We now combine all these results. If we first plug (40) into (37),

$$\begin{aligned}R_n^{\text{explore}} \leq 2LB \sum_{k < \underline{k}} p_k + \sum_{k \geq \underline{k}}^{K_n} p_k \mu^\star - \sum_{k \geq \underline{k}}^{K_n} p_k f(\omega_{z_k}^\star) + \lambda \sum_{k \geq \underline{k}}^{K_n} \sum_{t \in \mathcal{T}_k^E} g_t(\omega_t, z_k) + a_\lambda \sum_{k \geq \underline{k}}^{K_n} \sqrt{p_k} \\ + (6LB + \lambda_{\max}) M_n + 2\sqrt{\gamma_n} \Psi_n + LB \left(2 + \frac{2LB\lambda_{\max}}{\sigma^2} \right) \zeta_n + 2LB \sqrt{S_n \log S_n}. \end{aligned} \tag{43}$$

Then, plugging (42) into this inequality,

$$\begin{aligned}R_n^{\text{explore}} \leq 2LB \sum_{k < \underline{k}} p_k + \sum_{k \geq \underline{k}}^{K_n} p_k \mu^\star - \sum_{k \geq \underline{k}}^{K_n} p_k f(\omega_{z_k}^\star) + \lambda \frac{\beta_{n-1}}{2\sigma^2} - \lambda \sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} + a_\lambda \sum_{k \geq \underline{k}}^{K_n} \sqrt{p_k} \\ + \left(\lambda \frac{2L^2B^2}{\sigma^2} + 6LB + \lambda_{\max} \right) M_n + \left(2 + \frac{6LB\lambda}{\sigma^2} \right) \sqrt{\gamma_n} \Psi_n + 2LB \sqrt{S_n \log S_n} \\ + LB \left(2 + \frac{2LB(\lambda_{\max} + \lambda)}{\sigma^2} \right) \zeta_n + \frac{2\lambda B^2 L^2}{\sigma^2} \left(\sqrt{dS_n \log(dS_n)} + 1 \right). \end{aligned} \tag{44}$$

Let us simplify this expression so that it becomes more readable. First, we note that

$$\sum_{k \geq \underline{k}}^{K_n} p_k \mu^\star - \sum_{k \geq \underline{k}}^{K_n} p_k f(\omega_{z_k}^\star) = \sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} \underbrace{z_k(\mu^\star - f(\omega_{z_k}^\star))}_{= u^\star(z_k, \theta^\star)} = \sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} u^\star(z_k, \theta^\star).$$

Taking the expectation of both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[R_n^{\text{explore}}\right] \leq 2LB \sum_{k < \underline{k}} p_k + \mathbb{E}\left[\sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} u^\star(z_k, \theta^\star) \right] + \lambda \frac{\beta_{n-1}}{2\sigma^2} - \lambda \mathbb{E}\left[\sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} \right] \\ + a_\lambda \mathbb{E}\left[\sum_{k \geq \underline{k}}^{K_n} \sqrt{p_k} \right] + \mathbb{E}\left[\mathcal{O}(\sqrt{S_n \log S_n}) \right]. \end{aligned}$$

The remaining expectations on the right-hand side are due to the fact that K_n (hence S_n) is still random. Setting $\lambda = v^\star(\theta^\star)$ and combining the second and fourth terms, we get

$$\sum_{k\geq\underline{k}}^{K_n} \frac{p_k}{z_k} u^\star(z_k,\theta^\star) - \lambda \sum_{k\geq\underline{k}}^{K_n} \frac{p_k}{z_k} = \sum_{k\geq\underline{k}}^{K_n} \frac{p_k}{z_k} \left(u^\star(z_k,\theta^\star) - v^\star(\theta^\star)\right)$$

$$= \sum_{k\geq\underline{k}:z_k<\bar{z}(\theta^\star)} \frac{p_k}{z_k} \left(u^\star(z_k,\theta^\star) - v^\star(\theta^\star)\right) + \sum_{k:z_k\geq\bar{z}(\theta^\star)}^{K_n} \frac{p_k}{z_k} \left(u^\star(z_k,\theta^\star) - v^\star(\theta^\star)\right),$$

where $\bar{z}(\theta^\star) := \max_{x\in\mathcal{X}} \sum_{a\neq a^\star_{\theta^\star}(x)} \frac{\eta^\star(x,a)}{\rho(x)}$ was defined in Lem. 1. For $k \geq \underline{k}$, we can use the perturbation bound (Lem. 1) on both terms. We obtain,

$$\sum_{k\geq\underline{k}:z_k<\bar{z}(\theta^\star)} \frac{p_k}{z_k} \left(u^\star(z_k,\theta^\star) - v^\star(\theta^\star)\right) \leq BL\underline{z}(\theta^\star) \sum_{k\geq\underline{k}:z_k<\bar{z}(\theta^\star)} \frac{p_k}{z_k - \underline{z}(\theta^\star)}$$

and

$$\sum_{k\geq\underline{k}:z_k\geq\bar{z}(\theta^\star)}^{K_n} \frac{p_k}{z_k} \left(u^\star(z_k,\theta^\star) - v^\star(\theta^\star)\right) \leq BL\underline{z}(\theta^\star)z^\star(\theta^\star) \sum_{k\geq\underline{k}:z_k\geq\bar{z}(\theta^\star)}^{K_n} \frac{p_k}{z_k - \underline{z}(\theta^\star)} \max\left\{\frac{c_\Theta\sqrt{2}}{\sigma\sqrt{z_k}}, \frac{1}{z_k}\right\}$$

Partial regret bound Plugging these bounds into the expected regret,

$$\mathbb{E}\left[R_n^{\text{explore}}\right] \leq \underbrace{2BL\sum_{k<\underline{k}} p_k}_{\text{I}} + \underbrace{BL\underline{z}(\theta^\star) \sum_{k\geq\underline{k}:z_k<\bar{z}(\theta^\star)} \frac{p_k}{z_k - \underline{z}(\theta^\star)}}_{\text{II}} + \underbrace{v^\star(\theta^\star)\frac{\beta_{n-1}}{2\sigma^2}}_{\text{III}} + \underbrace{a_\lambda \mathbb{E}\left[\sum_{k\geq\underline{k}}^{K_n} \sqrt{p_k}\right]}_{\text{IV}}$$

$$\underbrace{+ BL\underline{z}(\theta)z^\star(\theta^\star)\mathbb{E}\left[\sum_{k:z_k\geq\bar{z}(\theta^\star)}^{K_n} \frac{p_k}{z_k - \underline{z}(\theta^\star)} \max\left\{\frac{c_\Theta\sqrt{2}}{\sigma\sqrt{z_k}}, \frac{1}{z_k}\right\}\right]}_{\text{V}} + \underbrace{\mathbb{E}\left[\mathcal{O}(\sqrt{S_n \log S_n})\right]}_{\text{VI}}. \tag{45}$$

The six terms constituting the bound are (from left to right):

I. finite regret suffered in the phases where the optimization problem is infeasible;

II. finite regret suffered in the phases in which we do not know much about the convergence rate of $u^\star(z,\theta^\star)$ to $v^\star(\theta^\star)$. This term is likely an artefact of the analysis;

III. asymptotically-optimal regret rate;

IV. regret suffered due to the incremental gradient updates and inversely proportional to the step sizes;

V. regret suffered due to the fact that we solve (P_z) instead of (P);

VI. other low-order terms mostly due to the concentration bounds.

Note that, since $\beta_{n-1} = c_{n,1/n}$ and $c_{n,1/n} \to 2\sigma^2 \log n$ as $n \to \infty$,

$$\limsup_{n\to\infty} \frac{v^\star(\theta^\star)\beta_{n-1}}{2\sigma^2 \log n} = v^\star(\theta^\star),$$

which is the asymptotically-optimal regret rate as prescribed by (P).

G.3.5 Bounding the total number of phases

So far we proved an upper bound on the regret incurred during exploration which depends on the (random) number of phases. We now upper bound this random variable as a function of z_k and p_k. In particular, we achieve this by focusing on the constraints only. The intuition is that, if the primal-dual algorithm works, then the sequence of policies played cannot violate the constraints *at each phase* too much. At the same time, these policies cannot satisfy the constraints too much, otherwise the

exploitation test would trigger and the algorithm would not be exploring at step n. Relating these two we obtain a bound on K_n.

Recall that, as we assumed before, n is an exploration step in which the good event G_n holds. Using (41) and the equations thereafter, we have

$$\inf_{\theta' \in \Theta_{alt}} \sum_{t \le n : E_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x,a) d_{x,a}(\theta^\star, \theta') \\ \le \frac{\beta_{n-1}}{2\sigma^2} + \frac{2LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n + \frac{2B^2L^2}{\sigma^2}\left(\sqrt{dS_n \log(dS_n)} + 1\right). \tag{46}$$

where the last two terms are $\mathcal{O}(\sqrt{S_n \log S_n})$.

We now provide a lower-bound on the same quantity. Fix a phase index $k \ge \underline{k}$. From (39), we have

$$\sum_{t \in \mathcal{T}_k^E} (f_t(\omega_t) + \lambda g_t(\omega_t, z_k)) \ge p_k f(\omega^\star_{z_k}) - a_\lambda \sqrt{p_k} - (2LB + \lambda_{\max}) M_{n,k} \\ - LB\left(1 + \frac{2LB\lambda_{\max}}{\sigma^2}\right)\zeta_{n,k}, \tag{47}$$

The left-hand side can be upper-bounded by using the optimism property to obtain the true objective and constraint. Regarding the objective function, we have

$$\sum_{t \in \mathcal{T}_k^E} f_t(\omega_t) = \sum_{t \in \mathcal{T}_k^E : G_t} f_t(\omega_t) + \sum_{t \in \mathcal{T}_k^E : \neg G_t} f_t(\omega_t) \\ \le \sum_{t \in \mathcal{T}_k^E : G_t} \sum_{x \in \mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a \in \mathcal{A}} \omega_t(x,a) \mu_{\widetilde{\theta}_{t-1}}(x,a) + \sum_{t \in \mathcal{T}_k^E : \neg G_t} \sum_{x \in \mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a \in \mathcal{A}} \omega_t(x,a) \mu_{\widetilde{\theta}_{t-1}}(x,a) + \sqrt{\gamma_n}\Psi_{n,k} \\ \le \underbrace{\sum_{t \in \mathcal{T}_k^E : G_t} \sum_{x \in \mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a \in \mathcal{A}} \omega_t(x,a) \mu_{\widetilde{\theta}_{t-1}}(x,a)}_{(a)} + BLM_{n,k} + \sqrt{\gamma_n}\Psi_{n,k}.$$

Regarding the sum over the good events, using Lem. 11,

$$(a) \le \sum_{t \in \mathcal{T}_k^E : G_t} \sum_{x \in \mathcal{X}} \hat{\rho}_{t-1}(x) \sum_{a \in \mathcal{A}} \omega_t(x,a) \mu_{\theta^\star}(x,a) + \sqrt{\gamma_n}\Psi_{n,k} \tag{48}$$

$$\le \sum_{t \in \mathcal{T}_k^E} \underbrace{\sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x,a) \mu_{\theta^\star}(x,a)}_{= f(\omega_t)} + BL \underbrace{\sum_{t \in \mathcal{T}_k^E : G_t} \sum_{x \in \mathcal{X}} |\hat{\rho}_t(x) - \rho(x)|}_{\zeta_{n,k}} + \sqrt{\gamma_n}\Psi_{n,k}. \tag{49}$$

Therefore,

$$\sum_{t \in \mathcal{T}_k^E} f_t(\omega_t) \le \sum_{t \in \mathcal{T}_k^E} f(\omega_t) + BL\zeta_{n,k} + 2\sqrt{\gamma_n}\Psi_{n,k} + BLM_{n,k}.$$

We can follow the same reasoning to upper bound the sum of constraints. Since the KLs are upper-bounded by $2B^2L^2/\sigma^2$,

$$\sum_{t \in \mathcal{T}_k^E} g_t(\omega_t, z_k) \le \sum_{t \in \mathcal{T}_k^E} g(\omega_t, z_k) + \frac{2B^2L^2}{\sigma^2}\zeta_{n,k} + \frac{4BL}{\sigma^2}\sqrt{\gamma_n}\Psi_{n,k} + \frac{2B^2L^2}{\sigma^2}M_{n,k}.$$

Combining the bounds on f and g,

$$\sum_{t \in \mathcal{T}_k^E} (f(\omega_t) + \lambda g(\omega_t, z_k)) \ge p_k f(\omega^\star_{z_k}) - \left(3BL + \lambda_{\max} + \lambda\frac{2B^2L^2}{\sigma^2}\right) M_{n,k} - a_\lambda\sqrt{p_k} \\ - 2BL\left(1 + \frac{(\lambda_{\max} + \lambda)BL}{\sigma^2}\right)\zeta_{n,k} - \left(2 + \frac{4BL\lambda}{\sigma^2}\right)\sqrt{\gamma_n}\Psi_{n,k}.$$

Let $\bar{\omega}_{t,k} := \frac{1}{p_k}\sum_{t\in\mathcal{T}_k^E}\omega_t$ be the average policy played in phase k. Since f is linear and g is concave, $\sum_{t\in\mathcal{T}_k^E}\left(f(\omega_t)+\lambda g(\omega_t,z_k)\right) \le p_k f(\bar{\omega}_{t,k}) + \lambda p_k g(\bar{\omega}_{t,k},z_k)$. We now set

$$\lambda = \begin{cases} 2\lambda_{\max} & \text{if } [g(\bar{\omega}_{t,k},z_k)]_- \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $[x]_- = \min\{x,0\}$. Therefore,

$$p_k\left(f(\bar{\omega}_{t,k}) - f(\omega^\star_{z_k}) + 2\lambda_{\max}[g(\bar{\omega}_{t,k},z_k)]_-\right) \ge -\left(3BL+\lambda_{\max}+\lambda_{\max}\frac{4B^2L^2}{\sigma^2}\right)M_{n,k} - a_{\lambda_{\max}}\sqrt{p_k} \\ - 2BL\left(1+\frac{3\lambda_{\max}BL}{\sigma^2}\right)\zeta_{n,k} - \left(2+\frac{8BL\lambda_{\max}}{\sigma^2}\right)\sqrt{\gamma_n}\Psi_{n,k}.$$

Lemma 3 together with Asm. 2 ensures that, for $k \ge \underline{k}$, $\lambda^\star(z_k,\theta^\star) \le \lambda_{\max}$. Thus, we can apply Theorem 42 of [23] and obtain

$$p_k g(\bar{\omega}_{t,k},z_k) \ge p_k[g(\bar{\omega}_{t,k},z_k)]_- \ge -\left(3BL+\lambda_{\max}+\lambda_{\max}\frac{4B^2L^2}{\sigma^2}\right)\frac{M_{n,k}}{2\lambda_{\max}} - \frac{a_{\lambda_{\max}}\sqrt{p_k}}{2\lambda_{\max}} \\ - 2BL\left(1+\frac{3\lambda_{\max}BL}{\sigma^2}\right)\frac{\zeta_{n,k}}{2\lambda_{\max}} - \left(2+\frac{8BL\lambda_{\max}}{\sigma^2}\right)\frac{\sqrt{\gamma_n}\Psi_{n,k}}{2\lambda_{\max}}.$$

Summing both sides over all phases,

$$\begin{aligned}\sum_{k\ge\underline{k}}^{K_n} p_k g(\bar{\omega}_{t,k},z_k) &= \sum_{k\ge\underline{k}}^{K_n} p_k\left(\inf_{\theta'\in\Theta_{alt}}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\bar{\omega}_{t,k}(x,a)d_{x,a}(\theta^\star,\theta') - \frac{1}{z_k}\right) \\ &= \sum_{k\ge\underline{k}}^{K_n}\left(\inf_{\theta'\in\Theta_{alt}}\sum_{t\in\mathcal{T}_k^E}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)d_{x,a}(\theta^\star,\theta') - \frac{p_k}{z_k}\right) \\ &\le \inf_{\theta'\in\Theta_{alt}}\sum_{t\le n:E_t}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)d_{x,a}(\theta^\star,\theta') - \sum_{k\ge\underline{k}}^{K_n}\frac{p_k}{z_k}.\end{aligned}$$

Therefore,

$$\inf_{\theta'\in\Theta_{alt}}\sum_{t\le n:E_t}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)d_{x,a}(\theta^\star,\theta') \ge \sum_{k\ge\underline{k}}^{K_n}\frac{p_k}{z_k} - \left(3BL+\lambda_{\max}+\lambda_{\max}\frac{4B^2L^2}{\sigma^2}\right)\frac{M_n}{2\lambda_{\max}} \\ - \frac{a_{\lambda_{\max}}\sqrt{p_k}}{2\lambda_{\max}} - 2BL\left(1+\frac{3\lambda_{\max}BL}{\sigma^2}\right)\frac{\zeta_n}{2\lambda_{\max}} - \left(2+\frac{8BL\lambda_{\max}}{\sigma^2}\right)\frac{\sqrt{\gamma_n}\Psi_n}{2\lambda_{\max}}.$$

Combining this with (46), we obtain the following inequality:

$$\begin{aligned}\sum_{k\ge\underline{k}}^{K_n}\frac{p_k}{z_k} \le \frac{\beta_{n-1}}{2\sigma^2} + \frac{a_{\lambda_{\max}}\sqrt{p_k}}{2\lambda_{\max}} + \frac{2LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n + \frac{2B^2L^2}{\sigma^2}\sqrt{dS_n\log(dS_n)} \\ + \left(3BL+\lambda_{\max}+\lambda_{\max}\frac{4B^2L^2}{\sigma^2}\right)\frac{M_n}{2\lambda_{\max}} \\ + 2BL\left(1+\frac{3\lambda_{\max}BL}{\sigma^2}\right)\frac{\zeta_n}{2\lambda_{\max}} + \left(2+\frac{8BL\lambda_{\max}}{\sigma^2}\right)\frac{\sqrt{\gamma_n}\Psi_n}{2\lambda_{\max}}.\end{aligned}$$

Recall that, by definition, $S_n = \sum_{k=0}^{K_n} p_k$. Furthermore, by Cauchy-Schwartz inequality, $\sum_{k=0}^{K_n}\sqrt{p_k} \le \sqrt{K_n\sum_{k=0}^{K_n}p_k}$. Simplifying this a little,

$$\sum_{k\ge\underline{k}}^{K_n}\frac{p_k}{z_k} \le \frac{\beta_{n-1}}{2\sigma^2} + \mathcal{O}\left(\sqrt{K_n\sum_{k=0}^{K_n}p_k}\right) + \mathcal{O}\left(\sqrt{\left(\sum_{k=0}^{K_n}p_k\right)\log\left(\sum_{k=0}^{K_n}p_k\right)}\right). \tag{50}$$

G.3.6 Choosing z_k and p_k

We choose the exponential schedule $z_k = z_0 e^k$ and $p_k = z_k e^{rk}$, where r will be specified later. The left-hand side of (50) is

$$\sum_{k \geq \underline{k}}^{K_n} \frac{p_k}{z_k} = \sum_{k \geq \underline{k}}^{K_n} e^{rk} \geq e^{rK_n},$$

while the right-hand side is

$$\begin{aligned}
&\frac{\beta_{n-1}}{2\sigma^2} + \mathcal{O}\left(\sqrt{K_n \sum_{k=0}^{K_n} e^{(r+1)k}}\right) + \mathcal{O}\left(\sqrt{\left(\sum_{k=0}^{K_n} e^{(r+1)k}\right) \log\left(\sum_{k=0}^{K_n} e^{(r+1)k}\right)}\right) \\
&\qquad \leq \frac{\beta_{n-1}}{2\sigma^2} + \mathcal{O}\left(\sqrt{K_n^2 e^{(r+1)K_n}}\right).
\end{aligned}$$

For $r > 1$, the resulting inequality yields $K_n \leq \mathcal{O}(\frac{1}{r}\log \beta_{n-1})$, i.e., $K_n \leq \mathcal{O}(\frac{1}{r}\log\log n)$ by definition of β_{n-1}. Let us recall (45):

$$\begin{aligned}
\mathbb{E}\left[R_n^{\text{explore}}\right] \leq{}& \underbrace{2BL \sum_{k<\underline{k}} p_k}_{\text{I}} + \underbrace{BL\underline{z}(\theta^\star) \sum_{k \geq \underline{k}: z_k < \bar{z}} \frac{p_k}{z_k - \underline{z}(\theta^\star)}}_{\text{II}} + \underbrace{v^\star(\theta^\star)\frac{\beta_{n-1}}{2\sigma^2}}_{\text{III}} + \underbrace{a_\lambda \mathbb{E}\left[\sum_{k \geq \underline{k}}^{K_n} \sqrt{p_k}\right]}_{\text{IV}} \\
&+ \underbrace{BL\underline{z}(\theta) z^\star(\theta^\star) \mathbb{E}\left[\sum_{k: z_k \geq \bar{z}(\theta^\star)}^{K_n} \frac{p_k}{z_k - \underline{z}(\theta^\star)} \max\left\{\frac{c_\Theta\sqrt{2}}{\sigma\sqrt{z_k}}, \frac{1}{z_k}\right\}\right]}_{\text{V}} + \underbrace{\mathbb{E}\left[\mathcal{O}(\sqrt{S_n \log S_n})\right]}_{\text{VI}}. \quad (51)
\end{aligned}$$

We bound the remaining terms separately.

Term I

$$\sum_{k<\underline{k}} p_k = z_0 \sum_{k<\underline{k}} e^{(r+1)k} \leq z_0 e^{(r+1)\log(\frac{2\underline{z}(\theta^\star)}{z_0})} \log(2\underline{z}(\theta^\star)/z_0) = z_0(2\underline{z}(\theta^\star)/z_0)^{r+1}\log(2\underline{z}(\theta^\star)/z_0),$$

where we used that, from the definition of $\underline{k}$ and z_k, it must be that $k < \log(2\underline{z}(\theta^\star)/z_0)$. Thus,

$$\text{I} \leq 2BLz_0(2\underline{z}(\theta^\star)/z_0)^{r+1}\log(2\underline{z}(\theta^\star)/z_0).$$

Term II

$$\begin{aligned}
\sum_{k \geq \underline{k}: z_k < \bar{z}(\theta^\star)} \frac{p_k}{z_k - \underline{z}(\theta^\star)} &= \sum_{\log\left(\frac{2\underline{z}(\theta^\star)}{z_0}\right) \leq k < \log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)} \frac{z_0 e^{(r+1)k}}{z_0 e^k - \underline{z}(\theta^\star)} \\
&= \sum_{\log\left(\frac{2\underline{z}(\theta^\star)}{z_0}\right) \leq k < \log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)} \underbrace{\frac{z_0 e^k}{z_0 e^k - \underline{z}(\theta^\star)}}_{\leq 2} e^{rk} \leq 2(\bar{z}(\theta^\star)/z_0)^r \log(\bar{z}(\theta^\star)/z_0).
\end{aligned}$$

Thus,

$$\text{II} \leq 2BL\underline{z}(\theta^\star)(\bar{z}(\theta^\star)/z_0)^r \log(\bar{z}(\theta^\star)/z_0).$$

Term IV The total number of exploration rounds is

$$S_n = \sum_{k=0}^{K_n} p_k = z_0 \sum_{k=0}^{K_n} e^{(r+1)k} \leq z_0 e^{(r+1)(K_n+1)} \leq \mathcal{O}((\log n)^{\frac{r+1}{r}}).$$

Therefore,

$$\text{IV} \leq \sqrt{K_n \sum_{k=0}^{K_n} p_k} \leq \mathcal{O}((\log\log n)^{1/2}(\log n)^{\frac{r+1}{2r}}).$$

Term V We consider two cases, based on which of the inner terms is the maximum. In the first case, we need to bound

$$\begin{aligned}
\sum_{k:z_k\geq\bar{z}(\theta^\star)}^{K_n}\frac{p_k}{(z_k-\underline{z}(\theta^\star))\sqrt{z_k}} &= \sum_{k\geq\log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)}^{K_n}\frac{z_0e^{(r+1)k}}{(z_0e^k-\underline{z}(\theta^\star))\sqrt{z_0e^k}} \\
&= \frac{1}{\sqrt{z_0}}\sum_{k\geq\log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)}^{K_n}\underbrace{\frac{z_0e^k}{(z_0e^k-\underline{z}(\theta^\star))}}_{\leq 2}e^{(r-1/2)k} \leq \frac{2}{\sqrt{z_0}}\sum_{k\geq\log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)}^{K_n}e^{(r-1/2)k} \\
&\leq \frac{2}{\sqrt{z_0}}\int_{\log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)}^{K_n+1}e^{(r-1/2)k}\mathrm{d}k = \frac{2}{\sqrt{z_0}}\left[\frac{e^{(r-1/2)k}}{r-1/2}\right]_{\log\left(\frac{\bar{z}(\theta^\star)}{z_0}\right)}^{K_n+1} \\
&= \frac{2}{(r-1/2)\sqrt{z_0}}\left(e^{(r-1/2)(K_n+1)}-(\bar{z}(\theta^\star)/z_0)^{r-1/2}\right).
\end{aligned}$$

Since $K_n \leq \mathcal{O}(\frac{1}{r}\log\log n)$, this term is $\mathcal{O}((\log n)^{\frac{r-1/2}{r}})$. If the other term is the maximum, then the same procedure yields a $\mathcal{O}((\log n)^{\frac{r-1}{r}})$ dependency. Thus,

$$\mathrm{V} \leq \mathcal{O}((\log n)^{\frac{r-1/2}{r}}).$$

Term VI We have $\mathrm{VI} \leq \mathcal{O}((\log n)^{\frac{r+1}{2r}})$ as in Term IV.

Final Bound Using $r = 2$, we obtain the following bound on the expected regret during exploration:

$$\begin{aligned}
\mathbb{E}\left[R_n^{\mathrm{explore}}\right] &\leq 2BLz_0(2\underline{z}(\theta^\star)/z_0)^3\log(2\underline{z}(\theta^\star)/z_0) \\
&\quad + 2BL\underline{z}(\theta^\star)(\bar{z}(\theta^\star)/z_0)^2\log(\bar{z}(\theta^\star)/z_0) + v^\star(\theta^\star)\frac{\beta_{n-1}}{2\sigma^2} + \mathcal{O}((\log\log n)^{\frac{1}{2}}(\log n)^{\frac{3}{4}}),
\end{aligned}$$

which is asymptotically optimal.

H Worst-case Analysis (Proof of Thm. 3)

H.1 Outline

The proof follows a similar argument as the one of Thm. 2 but it is considerably simpler and shorter. In particular, the main simplifications come from two worst-case arguments. (1) While bounding the regret during exploration rounds, we use the naive bound $S_n \leq n$. This is equivalent to assuming that SOLID never enters the exploitation step and it allows us to entirely avoid the bound on the number of phases of App. G.3.5. (2) We completely ignore the sequence z_k and proceed as if the optimization problem (P_z) was infeasible in all phases. This makes the multiplier saturate to $\lambda_{\max}$ and facilitate the analysis of the resulting Lagrangian[13]. An outline of the proof, together with the main differences w.r.t. the one of Thm. 2, is as follows.

1. We decompose the regret suffered during exploitation and exploration rounds. Using the same steps as in App. G, we bound the former by a constant and reduce the latter to the sum of objective values.
2. Instead of relating to the objective values of the optimal policies $\omega^\star_{z_k}$ at each phase k (as was done in App. G.3.2, we reduce our bound to the optimal solution of our bandit problem, i.e., the policy that only pulls optimal arms. This makes the sum of objective values cancel since the optimal policy achieves zero regret.
3. Using the results of App. G.3.3, we show that the sum of constraints is $\mathcal{O}(\log n)$.
4. We use the naive bound $S_n \leq n$ to conclude the proof.

[13]Recall that the regret of SOLID is not defined in terms of the optimization problem (P_z) or its Lagrangian, but only in terms of the rewards of the chosen arms compared to those of the optimal arms. This makes it possible to obtain good regret guarantees even when solving an infeasible optimization problem.

H.2 Proof

We start from the same regret decomposition as in App. G,

$$R_n = \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t) \mathbb{1}\left\{\neg E_t\right\} + \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t) \mathbb{1}\left\{E_t\right\} = R_n^{\text{exploit}} + R_n^{\text{explore}}.$$

The regret suffered during the exploitation rounds was bounded in App. G.2 as $\mathbb{E}\left[R_n^{\text{exploit}}\right] \leq 2LB$. Regarding the regret suffered during the exploration rounds, we have

$$R_n^{\text{explore}} := \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t) \mathbb{1}\left\{E_t\right\} \leq \sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t) \mathbb{1}\left\{E_t, G_t\right\} + 2LB \underbrace{\sum_{t=1}^{n} \mathbb{1}\left\{E_t, \neg G_t\right\}}_{:= M_n}. \tag{52}$$

Refer to App. F for the definition of G_t. The second term is M_n, the number of exploration rounds in which the good event does not hold, and can be bounded in expectation by using Lem. 6. The first one can be bounded by using the good event. Suppose, without loss of generality, that E_n and G_n hold (if they do not, the following reasoning can be repeated for the last time step at which these events hold). Then, using G_t^Δ (see App. F),

$$\begin{aligned}\sum_{t=1}^{n} \Delta_{\theta^\star}(X_t, A_t) \mathbb{1}\left\{E_t, G_t\right\} &\leq \sum_{t \leq n : E_t} \Delta_{\theta^\star}(X_t, A_t) \\ &\leq \sum_{t \leq n : E_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \Delta_{\theta^\star}(x, a) + 2LB\sqrt{S_n \log S_n}. \end{aligned} \tag{53}$$

We now proceed using similar steps as in App. G.3.1, except that we ignore the phases. We decompose the first term as

$$\begin{aligned}&\sum_{t \leq n : E_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \Delta_{\theta^\star}(x, a) \\ &\quad = \sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \Delta_{\theta^\star}(x, a) + \sum_{t \leq n : E_t, \neg G_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) (\mu^\star_{\theta^\star}(x) - \mu_{\theta^\star}(x, a)) \\ &\quad \leq \sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \Delta_{\theta^\star}(x, a) + M_n \mu^\star - \sum_{t \leq n : E_t, \neg G_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \mu_{\theta^\star}(x, a).\end{aligned}$$

Here we defined

$$\mu^\star := \sum_{x \in \mathcal{X}} \rho(x) \mu^\star_{\theta^\star}(x). \tag{54}$$

The last term can be bounded by $M_n BL$. Regarding the remaining two,

$$\begin{aligned}&\sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \Delta_{\theta^\star}(x, a) + M_n \mu^\star \\ &= (S_n - M_n)\mu^\star + M_n \mu^\star - \sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} \rho(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \mu_{\theta^\star}(x, a) \\ &= S_n \mu^\star + \underbrace{\sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} (\widehat{\rho}_{t-1}(x) - \rho(x)) \sum_{a \in \mathcal{A}} \omega_t(x, a) \mu_{\theta^\star}(x, a)}_{(a)} - \underbrace{\sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a \in \mathcal{A}} \omega_t(x, a) \mu_{\theta^\star}(x, a)}_{(b)}.\end{aligned}$$

Term (a) can be bounded as

$$(a) \leq LB \underbrace{\sum_{t \leq n : E_t, G_t} \sum_{x \in \mathcal{X}} \left|\widehat{\rho}_{t-1}(x) - \rho(x)\right|}_{\zeta_n}.$$

For the sake of readability, we keep the dependence on ζ_n explicit. We will bound this term by Lem. 12 at the end of the proof. Regarding term (b), using the definition of G_t and Lem. 10,

$$
\begin{aligned}
(b) &\geq \sum_{t\leq n:E_t,G_t}\sum_{x\in\mathcal{X}}\widehat{\rho}_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)\left(\mu_{\widetilde{\theta}_{t-1}}(x,a)-\sqrt{\gamma_t}\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}\right)\\
&\pm\sum_{t\leq n:E_t,\neg G_t}\sum_{x\in\mathcal{X}}\widehat{\rho}_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)\underbrace{\mu_{\widetilde{\theta}_{t-1}}(x,a)}_{|\cdot|\leq LB}\pm\sum_{t\leq n:E_t}\sum_{x\in\mathcal{X}}\widehat{\rho}_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)\sqrt{\gamma_t}\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}\\
&\geq\sum_{t\leq n:E_t}f_t(\omega_t)-M_nBL-2\sum_{t\leq n:E_t}\sum_{x\in\mathcal{X}}\widehat{\rho}_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)\sqrt{\gamma_t}\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}\\
&\geq\sum_{t\leq n:E_t}f_t(\omega_t)-M_nBL-2\sqrt{\gamma_n}\Psi_n.
\end{aligned}
$$

We recall that $\sqrt{\gamma_t}\leq\sqrt{\gamma_n}$ and $\Psi_n:=\sum_{t\leq n:E_t}\sum_{x\in\mathcal{X}}\widehat{\rho}_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_t(x,a)\|\phi(x,a)\|_{\bar{V}_{t-1}^{-1}}$. As for ζ_n, we keep the dependence on Ψ_n explicit and defer bounding this term to the end of the proof. Using the bounds on (a) and (b) and plugging everything back into (53) and then into (52), we obtain

$$
R_n^{\text{explore}}\leq S_n\mu^\star-\sum_{t\leq n:E_t}f_t(\omega_t)+4M_nBL+\zeta_nBL+2\sqrt{\gamma_n}\Psi_n+2BL\sqrt{S_n\log S_n}. \tag{55}
$$

We now lower bound the sum of objective values. Here we proceed in a slightly different way with respect to the proof of the asymptotically optimal regret bound. Instead of relating to the objective values of the optimal policies $\omega_{z_k}^\star$ at each phase k, we reduce our bound to the optimal solution of our bandit problem, i.e., the policy that only pulls optimal arms. Let

$$
\omega_{\theta^\star}^\star(x,a):=\begin{cases}1 & \text{if } a=a_{\theta^\star}^\star(x)\\ 0 & \text{otherwise}\end{cases} \tag{56}
$$

Recall that $\sum_{t\leq n:E_t}f_t(\omega_t)=\sum_{k=0}^{K_n}\sum_{t\in\mathcal{T}_k^E}f_t(\omega_t)$. Fix some phase index $k\geq 0$ and let $\lambda\geq 0$ be arbitrary. Using Corollary 2 with $\alpha_k^\lambda=\alpha_k^\omega=1/\sqrt{p_k}$ and $\omega=\omega_{\theta^\star}^\star$,

$$
\sum_{t\in\mathcal{T}_k^E}f_t(\omega_t)\geq\sum_{t\in\mathcal{T}_k^E}h_t(\omega_{\theta^\star}^\star,\lambda_t,z_k)-\lambda\sum_{t\in\mathcal{T}_k^E}g_t(\omega_t,z_k)-a_\lambda\sqrt{p_k}, \tag{57}
$$

where $a_\lambda:=\left(\log|\mathcal{A}|+\frac{b_\omega^2+b_\lambda^2}{2}+\frac{(\lambda-\lambda_1)^2}{2}\right)$ and b_λ and b_ω are the maximum sub-gradients in λ and ω, respectively. Note that, since we apply Corollary 2 to bound the sum of objective values over the whole phase, we have $S_{n,k}=p_k$. We now lower-bound the first term on the right-hand side. We have

$$
\begin{aligned}
\sum_{t\in\mathcal{T}_k^E}h_t(\omega_{\theta^\star}^\star,\lambda_t,z_k)&\overset{(c)}{=}\sum_{t\in\mathcal{T}_k^E}f_t(\omega_{\theta^\star}^\star)+\sum_{t\in\mathcal{T}_k^E}\lambda_tg_t(\omega_{\theta^\star}^\star,z_k)\overset{(d)}{\geq}\sum_{t\in\mathcal{T}_k^E}f_t(\omega_{\theta^\star}^\star)-\sum_{t\in\mathcal{T}_k^E}\frac{\lambda_t}{z_k}\\
&\overset{(e)}{\geq}\sum_{t\in\mathcal{T}_k^E}f_t(\omega_{\theta^\star}^\star)-\frac{\lambda_{\max}S_{n,k}}{z_k},
\end{aligned} \tag{58}
$$

where (c) uses the definition of h_t and g_t (see Eq. 3 and Eq. 4), (d) uses the positivity of KL divergences and confidence intervals, and (e) uses $\lambda_t\leq\lambda_{\max}$ and $S_{n,k}:=|\mathcal{T}_k^E|$. Let us focus on the sum of objective values. Since $f_t(\omega_{\theta^\star}^\star)\geq -LB$, we have $\sum_{t\in\mathcal{T}_k^{E:\neg G_t}}f_t(\omega_{\theta^\star}^\star)\geq -M_{n,k}BL$. For any step $t\in\mathcal{T}_k^E$ in which G_t holds, the optimism property (see App. F and Lem. 11) yields

$$
\begin{aligned}
\sum_{t\in\mathcal{T}_k^{E:G_t}}f_t(\omega_{\theta^\star}^\star)&\geq\sum_{t\in\mathcal{T}_k^{E:G_t}}\sum_{x\in\mathcal{X}}\widehat{\rho}_{t-1}(x)\sum_{a\in\mathcal{A}}\omega_{\theta^\star}^\star(x,a)\mu_{\theta^\star}(x,a)\\
&=\sum_{t\in\mathcal{T}_k^{E:G_t}}\sum_{x\in\mathcal{X}}(\widehat{\rho}_{t-1}(x)-\rho(x))\underbrace{\sum_{a\in\mathcal{A}}\omega_{\theta^\star}^\star(x,a)\mu_{\theta^\star}(x,a)}_{|\cdot|\leq BL}+\sum_{t\in\mathcal{T}_k^{E:G_t}}\underbrace{f(\omega_{\theta^\star}^\star)}_{=\mu^\star}\\
&\geq(S_{n,k}-M_{n,k})\mu^\star-BL\underbrace{\sum_{t\in\mathcal{T}_k^{E:G_t}}\sum_{x\in\mathcal{X}}|\widehat{\rho}_{t-1}(x)-\rho(x)|}_{:=\zeta_{n,k}},
\end{aligned}
$$

where we used the fact that $f(\omega^\star_{\theta^\star}) = \mu^\star$ by definition (56) and (54) and $\sum_{t\in\mathcal{T}_k^E} \mathbb{1}\{G_t\} = \sum_{t\in\mathcal{T}_k} \mathbb{1}\{E_t\} - \sum_{t\in\mathcal{T}_k} \mathbb{1}\{E_t, \neg G_t\} = S_{n,k} - M_{n,k}$. Plugging this back into (58) and then into (57),

$$\sum_{t\in\mathcal{T}_k^E} f_t(\omega_t) \geq (S_{n,k} - M_{n,k})\mu^\star - BL\zeta_{n,k} - \frac{\lambda_{\max} S_{n,k}}{z_k} - \lambda \sum_{t\in\mathcal{T}_k^E} g_t(\omega_t, z_k) - a_\lambda\sqrt{p_k} - M_{n,k}BL.$$

Summing over all phases and recalling that $\sum_{k=0}^{K_n} S_{n,k} = S_n$, $\sum_{k=0}^{K_n} M_{n,k} = M_n$, and $\sum_{k=0}^{K_n} \zeta_{n,k} = \zeta_n$, we obtain

$$\sum_{t\leq n:E_t} f_t(\omega_t) \geq (S_n - M_n)\mu^\star - BL\zeta_n - \sum_{k=0}^{K_n} \frac{\lambda_{\max} S_{n,k}}{z_k} - \lambda \sum_{t\leq n:E_t} g_t(\omega_t, z_{K_t}) - a_\lambda \sum_{k=0}^{K_n} \sqrt{p_k} - M_n BL. \tag{59}$$

Using the definition of g_t (see Eq. 3),

$$\sum_{t\leq n:E_t} g_t(\omega_t, z_{K_t}) := \sum_{t\leq n:E_t} \inf_{\theta'\in\overline{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega_t(x,a) \left(d_{x,a}\left(\widetilde{\theta}_{t-1}, \theta'\right) + \frac{2LB}{\sigma^2}\sqrt{\gamma_t}\|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}} \right) - \sum_{t\leq n:E_t} \frac{1}{z_{K_t}}.$$

By the definition of phase, the second term is $\sum_{t\leq n:E_t} \frac{1}{z_{K_t}} = \sum_{k=0}^{K_n} \frac{S_{n,k}}{z_k}$. The first term can be bounded using exactly the same steps as in App. G.3.3.[14] We obtain

$$\sum_{t\leq n:E_t} g_t(\omega_t, z_{K_t}) \leq \frac{\beta_{n-1}}{2\sigma^2} - \sum_{k=0}^{K_n} \frac{S_{n,k}}{z_k} + \frac{2L^2B^2}{\sigma^2} M_n + \frac{6LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n + \frac{2L^2B^2}{\sigma^2}\zeta_n + \frac{2B^2L^2}{\sigma^2}\left(\sqrt{dS_n \log(dS_n)} + 1\right). \tag{60}$$

If we now set $\lambda = \lambda_{\max}$ and plug (60) into (59),

$$\sum_{t\leq n:E_t} f_t(\omega_t) \geq (S_n - M_n)\mu^\star - BL\left(1 + \frac{2\lambda_{\max}BL}{\sigma^2}\right)(\zeta_n + M_n) + \underbrace{\sum_{k=0}^{K_n} \frac{\lambda_{\max} S_{n,k}}{z_k} - \sum_{k=0}^{K_n} \frac{\lambda_{\max} S_{n,k}}{z_k}}_{=0}$$
$$- \frac{\lambda_{\max}\beta_{n-1}}{2\sigma^2} - a_{\lambda_{\max}} \sum_{k=0}^{K_n} \sqrt{p_k} - \frac{6\lambda_{\max}LB}{\sigma^2}\sqrt{\gamma_n}\Psi_n - \frac{2\lambda_{\max}B^2L^2}{\sigma^2}\left(\sqrt{dS_n\log(dS_n)} + 1\right).$$

We can finally plug this into (55), thus obtaining

$$R_n^{\text{explore}} \leq M_n \underbrace{\mu^\star}_{|\cdot|\leq BL} + BL\left(5 + \frac{2\lambda_{\max}BL}{\sigma^2}\right)(\zeta_n + M_n) + \frac{\lambda_{\max}\beta_{n-1}}{2\sigma^2} + a_{\lambda_{\max}} \sum_{k=0}^{K_n} \sqrt{p_k}$$
$$+ \left(2 + \frac{6\lambda_{\max}LB}{\sigma^2}\right)\sqrt{\gamma_n}\Psi_n + \frac{2\lambda_{\max}B^2L^2}{\sigma^2}\left(\sqrt{dS_n\log(dS_n)} + 1\right) + 2BL\sqrt{S_n \log S_n}.$$

Let $\bar{k}_n := \min\{k : p_k \geq n\}$, then $K_n \leq \bar{k}_n$. Using the exponential schedule $p_k = e^{rk}$, $\bar{k}_n = \lceil \frac{1}{r}\log n \rceil$ and

$$\sum_{k=0}^{K_n} \sqrt{p_k} \leq \sum_{k=0}^{\bar{k}_n} e^{\frac{r}{2}k} \leq \int_0^{\bar{k}_n+1} e^{\frac{r}{2}x}\mathrm{d}x = \left[\frac{2}{r}e^{\frac{r}{2}x}\right]_0^{\bar{k}_n+1} = \frac{2}{r}e^{\frac{r}{2}(\lceil\frac{1}{r}\log n\rceil+1)} - \frac{2}{r} \leq \frac{2e^r}{r}\sqrt{n}.$$

[14] Note that the bound on the sum of constraints of App. G.3.3 uses only the properties of the confidence intervals and of the exploitation test. Thus, it is applicable regardless of the feasibility of the optimization problems at each phase.

Taking expectations of both sides of the regret bound above and using $S_n \leq n$ and $\mathbb{E}[M_n] \leq \frac{3\pi^2}{2}$ by Lem. 6,

$$\mathbb{E}\left[R_n^{\text{explore}}\right] \leq \frac{3BL\pi^2}{2}\left(6 + \frac{2\lambda_{\max}BL}{\sigma^2}\right) + \frac{\lambda_{\max}\beta_{n-1}}{2\sigma^2} + \frac{2e^r a_{\lambda_{\max}}}{r}\sqrt{n} + 2BL\sqrt{n\log n}$$
$$+ \left(2 + \frac{6\lambda_{\max}LB}{\sigma^2}\right)\mathbb{E}\left[\sqrt{\gamma_n}\Psi_n\right] + \frac{2\lambda_{\max}B^2L^2}{\sigma^2}\left(\sqrt{nd\log(nd)} + 1\right) + BL\left(5 + \frac{2\lambda_{\max}BL}{\sigma^2}\right)\mathbb{E}\left[\zeta_n\right].$$

After bounding $S_n \leq n$, by Lem. 13, $\Psi_n \leq \mathcal{O}\left(L|\mathcal{X}|\sqrt{n\log n} + \sqrt{nd\log n}\right)$ while, by Lem. 12, $\zeta_n \leq \mathcal{O}\left(|\mathcal{X}|\sqrt{n\log n}\right)$. Therefore, recalling that the regret during exploitation rounds was bounded by $2BL$ and noting that $2 < \frac{3\pi^2}{2}$,

$$\mathbb{E}[R_n] \leq 3BL\pi^2\left(4 + \frac{\lambda_{\max}BL}{\sigma^2}\right) + \frac{2e^r\lambda_{\max}^2}{r}\sqrt{n} + C_{\text{sqrt}}\left(1 + \frac{\lambda_{\max}BL}{\sigma^2}\right)\log(n)\sqrt{n},$$

where $C_{\text{sqrt}} = lin_{\geq 0}(|\mathcal{X}|, \sqrt{d}, B, L)$. Here we included $\frac{\lambda_{\max}\beta_{n-1}}{2\sigma^2}$ and the components of $a_{\lambda_{\max}}$ (except $\lambda_{\max}^2$ which is kept explicit) into the last term above. This concludes the proof.

I Auxiliary Results

I.1 Concentration Inequalities

Lemma 7 (Concentration of ρ during exploration)**.** *For any context $x \in \mathcal{X}$,*

$$\sum_{t\geq 1}\sum_{x\in\mathcal{X}}\mathbb{P}\left\{E_t, |\widehat{\rho}_t(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|S_t^2)}{2S_t}}\right\} \leq \frac{\pi^2}{3}. \tag{61}$$

Proof. The proof follows Lem. B.1 in [27]. Fix some $\bar{t} \geq 1$ and $x \in \mathcal{X}$. Then,

$$\sum_{t=1}^{\bar{t}}\mathbb{1}\left\{E_t, |\widehat{\rho}_t(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|S_t^2)}{2S_t}}\right\} \leq \sum_{s\geq 1}\mathbb{1}\left\{|\widehat{\rho}_{\tau_s}(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|s^2)}{2s}}, \tau_s \leq \bar{t}\right\}.$$

where τ_s is the random time the s-th exploration round occurs. Thus, by taking the expectation of both sides,

$$\sum_{t=1}^{\bar{t}}\mathbb{P}\left\{E_t, |\widehat{\rho}_t(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|S_t^2)}{2S_t}}\right\} \leq \sum_{s\geq 1}\mathbb{P}\left\{|\widehat{\rho}_{\tau_s}(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|s^2)}{2s}}, \tau_s \leq \bar{t}\right\}.$$

Since τ_s is a stopping-time upper bounded by $\bar{t}$ and the number of samples used to compute $\widehat{\rho}_{\tau_s}(x)$ is at least s, we can apply Lemma 4.3 of [27]:

$$\sum_{t=1}^{\bar{t}}\mathbb{P}\left\{E_t, |\widehat{\rho}_t(x) - \rho(x)| > \sqrt{\frac{\log(|\mathcal{X}|S_t^2)}{2S_t}}\right\} \leq \sum_{s\geq 1} 2e^{-2s\frac{\log(|\mathcal{X}|s^2)}{2s}} = \frac{2}{|\mathcal{X}|}\sum_{s\geq 1}\frac{1}{s^2} = \frac{\pi^2}{3|\mathcal{X}|}.$$

The reasoning above holds for any $\bar{t}$ and $x \in \mathcal{X}$. Summing over $\mathcal{X}$ concludes the proof. □

Lemma 8 (Confidence set for exploration)**.** *With some abuse of notation, let $\gamma_t := c_{n,1/S_t^2}$. Then, under the same conditions as in Theorem 1,*

$$\sum_{t=1}^{n}\mathbb{P}\left\{E_t, \|\widehat{\theta}_{t-1} - \theta^\star\|_{\overline{V}_{t-1}} > \sqrt{\gamma_t}\right\} \leq \frac{\pi^2}{6}.$$

Proof. Let $\{\tau_s\}_{s\geq 1}$ be a sequence of stopping times with respect to $\mathcal{F}$ such that if $\tau_s = t$, then the s-th exploration round occurs at time $t+1$. Then,

$$\sum_{t=1}^{n}\mathbb{1}\left\{E_t, \|\widehat{\theta}_{t-1} - \theta^\star\|_{\overline{V}_{t-1}} > \sqrt{\gamma_t}\right\} \leq \sum_{s\geq 1}\mathbb{1}\left\{\|\widehat{\theta}_{\tau_s} - \theta^\star\|_{\overline{V}_{\tau_s}} > \sqrt{\gamma_{\tau_s+1}}, \tau_s \leq n\right\}. \tag{62}$$

Since $S_{\tau_s+1} = s$, we have $\gamma_{\tau_s+1} = c_{n,1/s^2}$. Taking expectations and applying Theorem 1,

$$\sum_{s\geq 1} \mathbb{P}\left\{\|\widehat{\theta}_{\tau_s} - \theta^\star\|_{\overline{V}_{\tau_s}} > \sqrt{\gamma_{\tau_s+1}}, \tau_s \leq n\right\} \leq \sum_{s\geq 1} \frac{1}{s^2} = \frac{\pi^2}{6}.$$

□

I.2 Supporting Lemmas

The following result shows that any projection onto a non-empty convex set using a norm weighted by a positive definite matrix is a *non-expansion*. That is, the distance (in the chosen weighted norm) between the projected vector and any point in the set cannot increase w.r.t. the unprojected vector. We are not sure about a suitable citation for this result, so we include its proof.

Lemma 9 (Non-expansion of weighted projection)**.** *Let $\widehat{\theta} \in \mathbb{R}^d$ be any vector, $V \in \mathbb{R}^{d\times d}$ be a positive definite matrix, and $\mathcal{B} \subset \mathbb{R}^d$ be a non-empty convex set. Let $\widetilde{\theta}$ be the weighted projection of $\widehat{\theta}$ onto $\mathcal{B}$,*

$$\widetilde{\theta} := \operatorname*{argmin}_{\theta\in\mathcal{B}} \|\theta - \widehat{\theta}\|_V. \tag{63}$$

Then, for all $\theta \in \mathcal{B}$,

$$\|\widetilde{\theta} - \theta\|_V \leq \|\widehat{\theta} - \theta\|_V. \tag{64}$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as $f(x) := \|x - \widehat{\theta}\|_V^2$, so that $\widetilde{\theta} = \operatorname{argmin}_{x\in\mathcal{B}} f(x)$. Note that f is a convex function that is differentiable on $\mathbb{R}^d$. Therefore, using the first-order optimality conditions for convex functions (see, e.g., Theorem 2.8 in [28]), we have $\widetilde{\theta} = \operatorname{argmin}_{x\in\mathcal{B}} f(x)$ if and only if

$$\forall\theta \in \mathcal{B} : \langle \nabla f(\widetilde{\theta}), \theta - \widetilde{\theta}\rangle \geq 0. \tag{65}$$

Since $\nabla f(x) = 2V(x - \widehat{\theta})$,

$$\forall\theta \in \mathcal{B} : \langle V(\widetilde{\theta} - \widehat{\theta}), \theta - \widetilde{\theta})\rangle \geq 0. \tag{66}$$

Fix any $\theta \in \mathcal{B}$. We have

$$\|\widehat{\theta} - \theta\|_V^2 = \|\widehat{\theta} \pm \widetilde{\theta} - \theta\|_V^2 = \|\widehat{\theta} - \widetilde{\theta}\|_V^2 + \|\widetilde{\theta} - \theta\|_V^2 + 2(\widehat{\theta} - \widetilde{\theta})^T V(\widetilde{\theta} - \theta) \geq \|\widetilde{\theta} - \theta\|_V^2.$$

This concludes the proof. □

Corollary 1. *Let $t \in [n]$ be any time step in which the good event G_t holds. Then,*

$$\|\widetilde{\theta}_{t-1} - \theta^\star\|_{\overline{V}_{t-1}} \leq \|\widehat{\theta}_{t-1} - \theta^\star\|_{\overline{V}_{t-1}}. \tag{67}$$

Proof. If G_t holds, then $\theta^\star \in \mathcal{C}_{t-1}$. Since $\|\theta^\star\|_2 \leq B$ by definition, the set $\mathcal{C}_{t-1} \cap \Theta$ is non-empty (it contains $\theta^\star$ itself). Then, the result follows from Lem. 9. □

The following result is immediate from the definition of good event and the non-expansion property of the projection used to compute $\widetilde{\theta}_t$.

Lemma 10. *Let $t \in [n]$ be any time step in which the good event G_t holds. Then,*

$$\forall x \in \mathcal{X}, a \in \mathcal{A} : |\mu_{\widetilde{\theta}_{t-1}}(x, a) - \mu_{\theta^\star}(x, a)| \leq \sqrt{\gamma_t}\|\phi(x, a)\|_{\overline{V}_{t-1}^{-1}}.$$

Proof. Fix any $x \in \mathcal{X}$ and $a \in \mathcal{A}$. Then,

$$\begin{aligned}
|\mu_{\widetilde{\theta}_{t-1}}(x, a) - \mu_{\theta^\star}(x, a)| &= |\phi(x, a)^T(\widetilde{\theta}_{t-1} - \theta^\star)| = |\phi(x, a)^T \bar{V}_{t-1}^{-1/2} \bar{V}_{t-1}^{1/2}(\widetilde{\theta}_{t-1} - \theta^\star)| \\
&\overset{(a)}{\leq} \|\phi(x, a)\|_{\bar{V}_{t-1}^{-1}} \|\widetilde{\theta}_{t-1} - \theta^\star\|_{\bar{V}_{t-1}} \\
&\overset{(b)}{\leq} \|\phi(x, a)\|_{\bar{V}_{t-1}^{-1}} \|\widehat{\theta}_{t-1} - \theta^\star\|_{\bar{V}_{t-1}} \overset{(c)}{\leq} \sqrt{\gamma_t}\|\phi(x, a)\|_{\bar{V}_{t-1}^{-1}},
\end{aligned}$$

where (a) is from Cauchy-Schwartz inequality, (b) from Cor. 1, and (c) from the definition of G_t. □

Lemma 11. *Let $\gamma_t := c_{n,1/S_t^2}$ and $n \geq 3$. Then, for any time step t in which the good event G_t (see App. F) holds,*

$$f_t(\omega) := \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a)\left(\mu_{\widetilde{\theta}_{t-1}}(x,a) + \sqrt{\gamma_t}\|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}}\right) \geq \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a)\mu_{\theta^\star}(x,a), \tag{68}$$

and

$$g_t(\omega) := \inf_{\theta'\in\overline{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a)\left(d_{x,a}\left(\widetilde{\theta}_{t-1},\theta'\right) + \frac{2LB}{\sigma^2}\sqrt{\gamma_t}\|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}}\right) \geq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta'). \tag{69}$$

Proof. Since $\widehat{\rho}$ and ω are non-negative, the first inequality is trivial by upper bounding the true mean $\mu_{\theta^\star}(x,a)$ for each x,a by using the definition of G_t^θ and Lemma 10. Let us prove the second one. Fix any model $\theta' \in \Theta$. By using the definition of KL divergence of Gaussians with fixed variance, we have that:

$$\begin{aligned} d_{x,a}(\theta^\star,\theta') = \frac{(\mu_{\theta'}(x,a) - \mu_{\theta^\star}(x,a))^2}{2\sigma^2} &\leq d_{x,a}(\widetilde{\theta}_{t-1},\theta') + \frac{2LB}{\sigma^2}|\mu_{\widetilde{\theta}_{t-1}}(x,a) - \mu_{\theta^\star}(x,a)| \\ &\leq d_{x,a}(\widetilde{\theta}_{t-1},\theta') + \frac{2LB}{\sigma^2}\sqrt{\gamma_t}\|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}}, \end{aligned}$$

where the first inequality is from $|(a-c)^2 - (b-c)^2| = |(a+b-2c)(a-b)| \leq 4LB|a-b|$ and the second one is once again from the definition of G_t and Lemma 10. Therefore,

$$\begin{aligned} &\inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\theta^\star,\theta') \\ &\leq \inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a)\left(d_{x,a}(\widetilde{\theta}_{t-1},\theta') + \frac{2LB}{\sigma^2}\sqrt{\gamma_t}\|\phi(x,a)\|_{\overline{V}_{t-1}^{-1}}\right). \end{aligned} \tag{70}$$

We now upper bound the infimum over models in the alternative set. Note that such set can be fully specified once we assign an optimal arm to each context. Let $\{a_x\}_{x\in\mathcal{X}}$ and define

$$\Theta(\{a_x\}_{x\in\mathcal{X}}) = \{\theta' \in \Theta | \exists x \in \mathcal{X} : a^\star_{\theta'}(x) \neq a_x\}.$$

Note that $\Theta_{alt} = \Theta(\{a^\star_{\theta^\star}(x)\}_{x\in\mathcal{X}})$. Then,

$$\begin{aligned} &\inf_{\theta'\in\Theta_{alt}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\widetilde{\theta}_{t-1},\theta') && (71) \\ &\leq \max_{\{a_x\}_{x\in\mathcal{X}}} \inf_{\theta'\in\Theta(\{a_x\}_{x\in\mathcal{X}})} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\widetilde{\theta}_{t-1},\theta') && (72) \\ &\leq \inf_{\theta'\in\overline{\Theta}_{t-1}} \sum_{x\in\mathcal{X}} \widehat{\rho}_{t-1}(x) \sum_{a\in\mathcal{A}} \omega(x,a) d_{x,a}(\widetilde{\theta}_{t-1},\theta'). && (73) \end{aligned}$$

To see the last inequality, note that for all $\{a_x\}_{x\in\mathcal{X}}$ which do not contain only the optimal arms of $\widetilde{\theta}_{t-1}$ (i.e., $\{a_x\}_{x\in\mathcal{X}} \neq \{a^\star_{\widetilde{\theta}_{t-1}}(x)\}_{x\in\mathcal{X}}$), we have $\widetilde{\theta}_{t-1} \in \Theta(\{a_x\}_{x\in\mathcal{X}})$[15], and therefore the infimum is zero. Thus, the maximum must be attained by $\{a^\star_{\widetilde{\theta}_{t-1}}(x)\}_{x\in\mathcal{X}}$, which yields $\Theta(\{a^\star_{\widetilde{\theta}_{t-1}}(x)\}_{x\in\mathcal{X}}) = \overline{\Theta}_{t-1}$. This concludes the proof. □

Lemma 12. *For all time steps t,*

$$\sum_{s\leq t: E_s, G_s} \sum_{x\in\mathcal{X}} |\hat{\rho}_{s-1}(x) - \rho(x)| \leq 4|\mathcal{X}|\left(\sqrt{S_t \log(|\mathcal{X}|S_t^2)} + \log S_t + 1\right). \tag{74}$$

[15]Recall that, by definition, $\widetilde{\theta}_{t-1} \in \Theta$.

Proof. Using the definition of G_s,

$$\begin{aligned}
\sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}|\hat{\rho}_{s-1}(x)-\rho(x)| &\le |\mathcal{X}|\sum_{s\le t:E_s,G_s} 2\max\left(\sqrt{\frac{\log(|\mathcal{X}|S_s^2)}{2S_s}},\frac{2}{s}\right)\\
&\le 2|\mathcal{X}|\sum_{s=1}^{S_t}\max\left(\sqrt{\frac{\log(|\mathcal{X}|s^2)}{2s}},\frac{2}{s}\right)\\
&\le 2|\mathcal{X}|\sqrt{\frac{\log(|\mathcal{X}|S_t^2)}{2}}\sum_{s=1}^{S_t}\frac{1}{\sqrt{s}}+4|\mathcal{X}|\sum_{s=1}^{S_t}\frac{1}{s}\\
&\le 4|\mathcal{X}|\left(\sqrt{S_t\log(|\mathcal{X}|S_t^2)}+\log S_t+1\right),
\end{aligned}$$

where the last inequality holds since

$$\sum_{t=1}^{m}\sqrt{\frac{1}{t}}\le 1+\int_1^m x^{-1/2}dx = 1+[2x^{1/2}]_1^m = 2\sqrt{m}-1<2\sqrt{m}$$

and $\sum_{t=1}^m \frac{1}{t}\le \log m+1$. □

Lemma 13. *Let t be such that both E_t and G_t occur and suppose $\nu\ge 1$. Define*

$$\Psi_t := \sum_{s\le t:E_s}\sum_{x\in\mathcal{X}}\hat{\rho}_{s-1}(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}.$$

Then,

$$\Psi_t \le \frac{4L|\mathcal{X}|}{\sqrt{\nu}}\left(\sqrt{S_t\log(|\mathcal{X}|S_t^2)}+\log S_t+1\right)+\frac{M_tL}{\sqrt{\nu}}+\frac{L}{\nu}\sqrt{S_t\log S_t}.+\sqrt{2dS_t\log\frac{\nu+S_tL^2/d}{\nu}}.$$

Proof. We start by noticing that, for all x, a and $s\ge 0$,

$$\|\phi(x,a)\|^2_{\bar{V}_{s-1}^{-1}} = \phi(x,a)^T\bar{V}_{s-1}^{-1}\phi(x,a)\le \sigma_{\max}(\bar{V}_{s-1}^{-1})\underbrace{\|\phi(x,a)\|_2^2}_{\le L}\le \frac{L^2}{\sigma_{\min}(\bar{V}_{s-1})}\le\frac{L^2}{\nu},$$

and thus $\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}\le L/\sqrt{\nu}$. Here $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ denote the maximum and minimum eigenvalue of a matrix, respectively. Splitting the steps where the good event does and does not hold,

$$\begin{aligned}
\Psi_t &= \sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}\hat{\rho}_{s-1}(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}+\sum_{s\le t:E_s,\neg G_s}\sum_{x\in\mathcal{X}}\hat{\rho}_{s-1}(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}\\
&\le \sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}(\hat{\rho}_{s-1}(x)-\rho(x))\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}+\frac{M_tL}{\sqrt{\nu}}\\
&\quad+\sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}\\
&\le \frac{L}{\sqrt{\nu}}\sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}|\hat{\rho}_{s-1}(x)-\rho(x)|+\frac{M_tL}{\sqrt{\nu}}+\sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}\\
&\le \frac{4L|\mathcal{X}|}{\sqrt{\nu}}\left(\sqrt{S_t\log(|\mathcal{X}|S_t^2)}+\log S_t+1\right)+\frac{M_tL}{\sqrt{\nu}}+\sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}},
\end{aligned}$$

where in the first and second inequality we bounded the expected feature-norms by their maximum value and added/subtracted the first term with the true context distribution. In the last step we applied Lemma 12. We now focus exclusively on the third term. Using the fact that the good event holds at time t,

$$\begin{aligned}
\sum_{s\le t:E_s,G_s}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}} &\le \sum_{s\le t:E_s}\sum_{x\in\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\omega_s(x,a)\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}}\\
&\le \sum_{s\le t:E_s}\|\phi(X_s,A_s)\|_{\bar{V}_{s-1}^{-1}}+\frac{L}{\nu}\sqrt{S_t\log S_t}.
\end{aligned}$$

Finally, let $\bar{V}_{e,t} := \sum_{s \le t: E_s} \phi(X_s, A_s)\phi(X_s, A_s)^T + \nu I$ denote the regularized design matrix computed using only the exploration rounds. Then, we have $\bar{V}_t \succeq \bar{V}_{e,t}$ (since sum of rank-one matrices), which implies $\bar{V}_t^{-1} \preceq \bar{V}_{e,t}^{-1}$ and thus $\|\phi(x,a)\|_{\bar{V}_{s-1}^{-1}} \le \|\phi(x,a)\|_{\bar{V}_{e,s-1}^{-1}}$. Here $\succeq$ denotes the Loewner ordering, i.e., for two symmetric matrices A, B we have $A \succeq B$ ($A \succ B$) if $A - B$ is positive semi-definite (positive definite). Therefore,

$$\sum_{s \le t: E_s} \|\phi(X_s, A_s)\|_{\bar{V}_{s-1}^{-1}} \le \sum_{s \le t: E_s} \|\phi(X_s, A_s)\|_{\bar{V}_{e,s-1}^{-1}} \overset{(a)}{\le} \sqrt{S_t \sum_{s \le t: E_s} \|\phi(X_s, A_s)\|^2_{\bar{V}_{e,s-1}^{-1}}}$$
$$\overset{(b)}{\le} \sqrt{2 S_t \log \frac{\det(\bar{V}_{e,t})}{\nu^d}} \overset{(d)}{\le} \sqrt{2 d S_t \log \frac{\nu + S_t L^2/d}{\nu}},$$

where in (a) we equivalently rewritten the first term as a sum over exploration rounds, (b) is from Cauchy-Schwartz inequality, in (c) we used Lemma 11 of [4], and in (d) we used the determinant-trace inequality (Lemma 10 of [4]) to bound the determinant of $\bar{V}_{e,t}$ by $(\nu + S_t L^2/d)^d$. The final statement follows by combining the previous bounds. □

I.3 Online Convex Optimization

Here we recall some basic results from online convex optimization. See [e.g., 29] for detailed proofs and discussion of these results.

Lemma 14 (Recursion bound for subgradient descent)**.** *Let $\sup_{t \ge 1: E_t} |g_t(\omega_t, z_k)|^2 \le b_\lambda$. For any phase $k \ge 0$, $t \in \mathcal{T}_k^E$, and $\lambda \in \mathbb{R}_+$, the incremental updates to the Lagrange multiplier $\{\lambda_t\}_{t \in \mathcal{T}_k^E}$ of Algorithm 1 satisfy*

$$\sum_{s \le t: s \in \mathcal{T}_k^E} g_s(\omega_s, z_k)(\lambda_s - \lambda) \le \frac{1}{2\alpha_k^\lambda}(\lambda - \lambda_1)^2 + \frac{\alpha_k^\lambda b_\lambda^2}{2} S_{t,k}.$$

Proof. Recall that the optimization process is reset at the beginning of each phase. Let $\tau_{s,k}$ be a random variable indicating the time at which the s-th exploration round of phase k occurs. Note that $\lambda_{\tau_{1,k}} = \lambda_1$. In order to simplify the exposition, and with some abuse of notation, let $\lambda_s = \lambda_{\tau_{s,k}}$ and $g_s = g_{\tau_{s,k}}(\omega_{\tau_{s,k}}, z_k)$. By definition of the update rule, for each $s \ge 1$,

$$(\lambda_{s+1} - \lambda)^2 = (\min\{[\lambda_s - \alpha_k^\lambda g_s]_+, \lambda_{\max}\} - \lambda)^2 = \min\{[\lambda_s - \alpha_k^\lambda g_s]_+ - \lambda, \lambda_{\max} - \lambda\}^2$$
$$\le (\lambda_s - \alpha_k^\lambda g_s - \lambda)^2 = (\lambda_s - \lambda)^2 + (\alpha_k^\lambda g_s)^2 + 2\alpha_k^\lambda(\lambda - \lambda_s)g_s.$$

Dividing by $2\alpha_k^\lambda$ and rearranging,

$$(\lambda_s - \lambda)g_s \le \frac{(\lambda_s - \lambda)^2 - (\lambda_{s+1} - \lambda)^2}{2\alpha_k^\lambda} + \frac{\alpha_k^\lambda}{2} g_s^2.$$

Summing over all s up to S_t and noting that the first sum on the right-hand side is telescopic,

$$\sum_{s=1}^{S_t} (\lambda_s - \lambda)g_s \le \frac{1}{2\alpha_k^\lambda}(\lambda_1 - \lambda)^2 - \frac{1}{2\alpha_k^\lambda}(\lambda_{S_t+1} - \lambda)^2 + \frac{\alpha_k^\lambda}{2} \sum_{s=1}^{S_t} g_s^2.$$

The proof is concluded by upper-bounding the second term by zero and mapping the exploration counter s back to time steps. □

Lemma 15. *[Recursion bound for Online Mirror Descent (OMD)] Let ω_1 be the uniform distribution over actions for each context and $\sup_{t \ge 1: E_t} \|q_t\|_\infty \le b_\omega$. For any phase $k \ge 0$, $t \in \mathcal{T}_k^E$, and $\omega \in \Omega$, the OMD updates of Algorithm 1 satisfy*

$$\sum_{s \le t: s \in \mathcal{T}_k^E} h_s(\omega_s, \lambda_s, z_k) - \sum_{s \le t: s \in \mathcal{T}_k^E} h_s(\omega, \lambda_s, z_k) \ge -\frac{\log|\mathcal{A}|}{\alpha_k^\omega} - \frac{\alpha_k^\omega b_\omega^2}{2} S_{t,k}.$$

Proof. We can follow the same steps as before, mapping time steps to exploration counters and then applying the standard recursion bound for OMD [e.g., 29]. □

Corollary 2. *[Recursion bound for primal-dual algorithm] For any phase $k \geq 0$, $t \in \mathcal{T}_k^E$, $\omega \in \Omega$, and $\lambda \in \mathbb{R}_+$, under the same conditions as in Lemma 15 and 14,*

$$\sum_{s \leq t : s \in \mathcal{T}_k^E} f_s(\omega_s) \geq \sum_{s \leq t : s \in \mathcal{T}_k^E} h_s(\omega, \lambda_s, z_k) - \lambda \sum_{s \leq t : s \in \mathcal{T}_k^E} g_s(\omega_s, z_k) - \frac{\log |\mathcal{A}|}{\alpha_k^\omega} - \frac{\alpha_k^\omega b_\omega^2}{2} S_{t,k} - \frac{1}{2\alpha_k^\lambda}(\lambda - \lambda_1)^2 - \frac{\alpha_k^\lambda b_\lambda^2}{2} S_{t,k}.$$

Proof. The proof is straightforward by expanding $\sum_{s \leq t : s \in \mathcal{T}_k^E} h_s(\omega_s, \lambda_s, z_k) = \sum_{s \leq t : s \in \mathcal{T}_k^E} (f_s(\omega_s) + \lambda_s g_s(\omega_s, z_k))$ and combining Lemma 15 with Lemma 14. □

J Confidence Set for Regularized Least-Squares (Proof of Thm. 1)

The following theorem is the extended version of Thm. 1. It provides a refined confidence set for the parameters estimated by regularized least-squares.

Theorem 4 (Confidence set over parameters)**.** *Let $\delta \in (0, 1)$ and $n \geq 3$. Then,*

$$\mathbb{P}\left\{\exists t \in [n] : \|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t} \geq \sqrt{c_{n,\delta}}\right\} \leq \delta,$$

where $\sqrt{c_{n,\delta}} := \frac{\gamma_n}{1 - \frac{1}{\log n}} \sqrt{\kappa_{n,\delta}}$, $\gamma_n := 1 + \frac{1}{\log n}$, and

$$\sqrt{\kappa_{n,\delta}} = B\sqrt{\nu} + \sqrt{\frac{2\sigma^2 \log\left(\frac{2 + \frac{2nL^2}{d\nu}}{\delta}\right)}{(\log n)^2}} + \sqrt{2\sigma^2 \gamma_n^3 \log\left(\frac{2(1 + \log(n/\chi_n)\log(n))}{\delta}\right) + 2\gamma_n^3 \Upsilon_n}.$$

Finally, we set $\Upsilon_n := d \log\left(\frac{5}{2} + 2\log n \sqrt{d}\right) + d \log\left(2 + 4d \log\left(4\gamma_n d (\log n)^2 \sqrt{\frac{\nu + L^2 n}{d\nu}}\right) \log n\right)$ and $\chi_n := \frac{\nu^2 v_{\min}^2}{16 d L^2 (\nu + L^2 n)(\log n)^4 \gamma_n^4}$.

Asymptotic dependence It is important to note that $\lim_{n \to \infty} \frac{c_{n,1/n}}{2\sigma^2 \log n} = 1$.

J.1 Proof of Thm. 4

The proof can be summarized in three main steps:

1. We reduce the problem of bounding $\|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t}$ to one in which we need to bound $(\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} v$ for any $v \in \mathcal{C}_1$, where $\mathcal{C}_1 \subset \mathbb{R}^d$ is a (finite) ϵ_1-cover of the d-dimensional Euclidean unit ball. We build this cover in such a way that all its elements have norm bounded from below by a strictly positive constant and from above.
2. We extend Theorem 8 of [7] to bound $(\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} v$ uniformly over all $v \in \mathcal{C}_1$, instead of the prediction errors $(\widehat{\theta}_t - \theta^\star)^T \phi(x, a)$ uniformly over all contexts/arms. This requires a second ϵ_2-cover (we shall call it $\mathcal{C}_2$) of the set $\{\overline{V}_t^{-1/2} v : t \in [n], v \in \mathcal{C}_1\}$. The result is reported in Lemma 16.
3. The resulting bound is of order $\mathcal{O}(\log(1/\delta) + d \log(1/\epsilon_1))$, which requires tuning $\epsilon_1 = \frac{1}{\log n}$ to cancel the bias of the first cover asymptotically without compromising the size of the cover itself.

Step 1. We start from the fact that

$$\|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t} = \frac{(\widehat{\theta}_t - \theta^\star)^T \overline{V}_t (\widehat{\theta}_t - \theta^\star)}{\|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t}} = (\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} z_t, \tag{75}$$

where $z_t = \frac{\overline{V}_t^{1/2}(\widehat{\theta}_t - \theta^\star)}{\|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t}}$ is such that $\|z_t\|_2 = 1$. To handle the fact that z_t is random, we build a *linear* ($\epsilon_1 > 0$)-cover of the space $\mathcal{Z} = \{z \in \mathbb{R}^d : \|z\|_2 \le 1\}$, which includes z_t for all $t = 1, \dots, n$ Let $\epsilon_1' > 0$, $\{e_1, e_2, \dots, e_d\}$ be the canonical basis of $\mathbb{R}^d$, and define

$$\widetilde{\mathcal{C}}_1 := \left\{ \sum_{i=1}^d a_i e_i : a_i \in \left\{ \pm\epsilon_1' \Big(\frac{1}{2} + j \Big) : j = 0, 1, \dots, \overline{j} \right\} \forall i \in [d] \right\},$$

where $\overline{j} := \left\lceil \frac{1}{\epsilon_1'} - \frac{1}{2} \right\rceil$. For any vector $z \in \mathcal{Z}$, we can find a vector in $\widetilde{\mathcal{C}}_1$ with at most ϵ_1' error on each component of z, which leads to $\min_{v \in \widetilde{\mathcal{C}}_1} \|v - z\|_2 \le \epsilon_1' \sqrt{d}$ [see e.g., 30, Chap. 27]. Setting $\epsilon_1' = \epsilon_1 / \sqrt{d}$ gives an ϵ_1-cover of the unit ball in ℓ_2-norm. The only problem with this cover is that it contains vectors with norm bigger than 1 and scaling with d,[16] which may lead to an undesirable dependency later on. However, we can safely remove the vectors with large norm without affecting the desired accuracy of the cover. Without loss of generality, select $z \in \mathcal{Z}$ in the positive orthant (i.e., $z_i \ge 0$, for any $i \in [d]$) such that we make an error of ϵ_1' on each component (i.e., the worst-case) and let $w = z + \epsilon_1'$. Then

$$\|w\|_2^2 = \sum_{i=1}^d (z_i + \epsilon_1')^2 = \underbrace{\|z\|_2^2}_{\le 1} + d(\epsilon_1')^2 + 2\epsilon_1' \underbrace{\sum_{i=1}^d z_i}_{\le \|z\|_1 \le \sqrt{d}} \le 1 + \epsilon_1^2 + 2\epsilon_1 = (1 + \epsilon_1)^2.$$

Hence vectors with norm at most $(1 + \epsilon_1)$ actually suffice and thus we can set $\mathcal{C}_1 = \widetilde{\mathcal{C}}_1 \setminus \{v \in \widetilde{\mathcal{C}}_1 : \|v\|_2 > (1 + \epsilon_1)\}$. Then we upper bound the size of this cover as

$$|\mathcal{C}_1| \le |\widetilde{\mathcal{C}}_1| = 2^d (1 + \overline{j})^d \le \left(\frac{5}{2} + \frac{2}{\epsilon_1'} \right)^d = \left(\frac{5}{2} + \frac{2\sqrt{d}}{\epsilon_1} \right)^d.$$

To recap, our cover $\mathcal{C}_1$ has the following properties:

1. $\forall z \in \mathcal{Z} = \{z \in \mathbb{R}^d : \|z\|_2 \le 1\}$, $\exists v \in \mathcal{C}_1 : \|z - v\|_2 \le \epsilon_1$
2. $|\mathcal{C}_1| \le \left(\frac{5}{2} + \frac{2\sqrt{d}}{\epsilon_1} \right)^d$
3. $\forall v \in \mathcal{C}_1 : \|v\|_2 \le v_{\max} := 1 + \epsilon_1$
4. $\forall v \in \mathcal{C}_1, i \in [d] : |v_i| \ge v_{\min} := \frac{\epsilon_1}{2\sqrt{d}}$ (this follows from the discretization used in $\widetilde{\mathcal{C}}_1$ and it implies that $\|v\|_2 \ge v_{\min} \sqrt{d} = \frac{\epsilon_1}{2}$)

Step 2. We use an extension of Thm. 8 of [7] to bound the prediction error at vectors in the cover $\mathcal{C}_1$ after applying the linear transformation $\overline{V}_t^{1/2}$.

Lemma 16. *Let $\mathcal{C} \subset \mathbb{R}^d$ be a finite set such that, for any $v \in \mathcal{C}$, $\|v\|_2 \le v_{\max} < \infty$ and $|v_i| \ge v_{\min} > 0$, $\forall i \in [d]$. Suppose that $n \ge 2$. Then, for any $\delta \in (0, 1)$,*

$$\mathbb{P}\left\{ \exists t \le n, v \in \mathcal{C} : (\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} v \ge \sqrt{\kappa_{n,\delta}} \|v\|_2 \right\} \le \delta,$$

where

$$\sqrt{\kappa_{n,\delta}} = B\sqrt{\nu} + \sqrt{\frac{2\sigma^2 \log\left(\frac{2 + \frac{2nL^2}{d\nu}}{\delta} \right)}{(\log n)^2}} + \sqrt{2\sigma^2 \gamma_n^3 \log\left(\frac{2(1 + \log(n/\chi_n)\log(n))}{\delta} \right) + 2\gamma_n^3 \Upsilon_n}$$

and $\Upsilon_n = \log(|\mathcal{C}|) + d \log\left(2 + 4d \log n \frac{v_{\max}}{v_{\min}} \sqrt{\frac{\nu + L^2 n}{d\nu}} \right) \log n$ *and* $\chi_n = \frac{\nu^2 v_{\min}^2}{4L^2(\nu + L^2 n)(\log n)^2 v_{\max}^2 \gamma_n^2}$.

The specific shape of the bound is obtained by exploiting the properties of the cover $\mathcal{C}_1$ derived in the first step, where $v_{\max} = 1 + \epsilon_1$ and $v_{\min} = \frac{\epsilon_1}{2\sqrt{d}}$.

[16] Consider the vector with all components equal to 1, whose norm is $\sqrt{d}$.

Step 3. We finally tune ϵ_1 to obtain the final bound. With probability at least $1-\delta$, we have that

$$\begin{aligned}
\|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t} &\overset{(a)}{=} (\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} z_t \overset{(b)}{\leq} \max_{z\in\mathcal{Z}} (\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} z \\
&= \max_{z\in\mathcal{Z}} \min_{v\in\mathcal{C}_1} \left\{ (\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} (z-v) + (\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} v \right\} \\
&\overset{(c)}{\leq} \max_{z\in\mathcal{Z}} \min_{v\in\mathcal{C}_1} \left\{ \|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t} \|z-v\|_2 + \sqrt{\kappa_{n,\delta}} \|v\|_2 \right\} \\
&\overset{(d)}{\leq} \epsilon_1 \|\widehat{\theta}_t - \theta^\star\|_{\overline{V}_t} + (1+\epsilon_1)\sqrt{\kappa_{n,\delta}},
\end{aligned}$$

where (a) follows from Eq. 75, (b) from the fact that $z_t \in \mathcal{Z}$, (c) holds with probability at least $1-\delta$ by Lem. 16 and (d) by properties 1 and 3 of the cover $\mathcal{C}_1$. The statement of the theorem follows by setting $\epsilon_1 = \frac{1}{\log n}$ and rearranging.

J.2 Proof of Lem. 16

The proof follows similar steps as in [7, Thm. 8].

Proof. Take any $v \in \mathcal{C}_1$ and $t \in [n]$. Then,

$$\begin{aligned}
(\widehat{\theta}_t - \theta^\star)^T \overline{V}_t^{1/2} v &\overset{(a)}{=} \left(\overline{V}_t^{-1} \sum_{s=1}^t \phi(X_s, A_s) Y_s - \theta^\star \right)^T \overline{V}_t^{1/2} v \\
&\overset{(b)}{=} \left(\overline{V}_t^{-1} \sum_{s=1}^t \phi(X_s, A_s)(\phi(X_s, A_s)^T \theta^\star + \xi_s) - \theta^\star \right)^T \overline{V}_t^{1/2} v \\
&\overset{(c)}{=} \left(\overline{V}_t^{-1} V_t \theta^\star + \overline{V}_t^{-1} \sum_{s=1}^t \phi(X_s, A_s)\xi_s - \theta^\star \right)^T \overline{V}_t^{1/2} v \\
&\overset{(d)}{=} \underbrace{\left(\overline{V}_t^{-1} V_t \theta^\star - \theta^\star \right)^T \overline{V}_t^{1/2} v}_{(i)} + \underbrace{\sum_{s=1}^t v^T \overline{V}_t^{-1/2} \phi(X_s, A_s)\xi_s}_{(ii)},
\end{aligned} \tag{76}$$

where (a) is from the definition of $\widehat{\theta}_t$, (b) since $Y_s = \phi(X_s, A_s)^T \theta^\star + \xi_s$ with $\xi_s \sim \mathcal{N}(0, \sigma^2)$, (c) from the definition of V_t, and (d) after rearranging. Let us bound (i). Since $\theta^\star = \overline{V}_t^{-1} \overline{V}_t \theta^\star$, we have

$$(i) = v^T \overline{V}_t^{-1/2} (V_t - \overline{V}_t)\theta^\star = -\nu v^T \overline{V}_t^{-1/2} \theta^\star,$$

where we used $\overline{V}_t = \nu I + V_t$. Therefore,

$$|(i)| \leq \nu |v^T \overline{V}_t^{-1/2} \theta^\star| \leq \nu \|v\|_2 \|\overline{V}_t^{-1/2} \theta^\star\|_2 = \nu \|v\|_2 \|\theta^\star\|_{\overline{V}_t^{-1}},$$

where the second inequality is by Cauchy-Schwartz inequality. Since $\overline{V}_t \succeq \nu I$, $\|\theta^\star\|_{\overline{V}_t^{-1}} \leq \frac{1}{\sqrt{\nu}} \|\theta^\star\|_2 \leq \frac{B}{\sqrt{\nu}}$. This yields

$$|(i)| \leq B\sqrt{\nu} \|v\|_2.$$

Let us consider the second term. Since $\overline{V}_t^{-1/2}$ is random, we proceed using the same covering argument as in the proof in [7, Thm. 8]. Let $\epsilon_2 > 0$ (whose value will be specified later). Recall that our input is a finite set of d-dimensional vectors $\mathcal{C}_1$ such that $\|v\|_2 \leq v_{\max} < \infty$ and $|v_i| \geq v_{\min} > 0$ hold for all $v \in \mathcal{C}_1$ and $i \in [d]$. Note that the latter condition implies $\|v\|_2 \geq v_{\min}\sqrt{d}$. Our goal is to build an ϵ_2-covering set of $\{\overline{V}_t^{-1/2} v : t \in [n], v \in \mathcal{C}_1\}$. Since this set is random, we build a deterministic one that contains the former almost surely and cover it instead. Note that, for

any $t \in [n]$, $\overline{V}_t^{-1/2}$ is such that (1) $\overline{V}_t^{-1/2} \succ 0$, (2) $\|\overline{V}_t^{-1/2}\|_2 = \sigma_{\max}(\overline{V}_t^{-1/2}) \leq \frac{1}{\sqrt{\nu}}$, and (3) $\sigma_{\min}(\overline{V}_t^{-1/2}) \geq \frac{1}{\sqrt{\nu+L^2n}}$. Let $\mathcal{D}$ denote the set of $d \times d$ matrices with these properties, that is,

$$\mathcal{D} := \left\{ D \in \mathbb{R}^{d\times d} : D \succ 0, \|D\|_2 \leq \frac{1}{\sqrt{\nu}}, \sigma_{\min}(D) \geq \frac{1}{\sqrt{\nu + L^2 n}} \right\}.$$

Then, $\overline{V}_t^{-1/2} \in \mathcal{D}$ for all $t \in [n]$ and our initial set to be covered is almost surely contained into $\mathcal{B} := \{Dv : D \in \mathcal{D}, v \in \mathcal{C}_1\}$. Furthermore, $v_{\min}\sqrt{\frac{d}{\nu+L^2n}} \leq \|b\|_2 \leq \frac{v_{\max}}{\sqrt{\nu}}$ for all $b \in \mathcal{B}$. We shall now cover $\mathcal{B}$. Let $\{e_1, \ldots, e_d\}$ be the canonical basis of $\mathbb{R}^d$ and, for all $v \in \mathcal{C}_1$ we introduce a cover with *geometric scale* as

$$\tilde{\mathcal{C}}_{2,v} := \left\{ \sum_{i=1}^{d} a_i e_i \Big| \, \forall i \in [d] : a_i \in \left\{ \pm \frac{\epsilon_2 \|v\|_2 (1+\epsilon_2)^j}{\sqrt{\nu + L^2 n}} : j = 0, 1, \ldots, \bar{j} \right\} \right\},$$

where $\bar{j} := \left\lceil \frac{\log\left(\frac{v_{\max}}{\epsilon_2 v_{\min}} \sqrt{\frac{\nu+L^2n}{d\nu}}\right)}{\log(1+\epsilon_2)} \right\rceil$ is such that $\frac{\epsilon_2\|v\|_2(1+\epsilon_2)^{\bar{j}}}{\sqrt{\nu+L^2n}} \geq \frac{v_{\max}}{\sqrt{\nu}}$ (i.e., the maximum absolute value of each element in $\mathcal{B}$). Then, our cover is $\tilde{\mathcal{C}}_2 = \bigcup_{v\in\mathcal{C}_1} \tilde{\mathcal{C}}_{2,v}$. Let us analyze some its properties. First its size is

$$|\tilde{\mathcal{C}}_2| \leq |\mathcal{C}_1| \left(2 + \frac{\log\left(\frac{v_{\max}}{\epsilon_2 v_{\min}} \sqrt{\frac{\nu+L^2n}{d\nu}}\right)}{\log(1+\epsilon_2)} \right)^d. \tag{77}$$

Then, we can show the following covering property in l_∞-norm.

Proposition 5. *For all $v \in \mathcal{C}_1$, $t \in [n]$, there exists $\overline{w}_{v,t} \in \tilde{\mathcal{C}}_2$ such that*

$$\forall i \in [d] : \left| \left[\overline{V}_t^{-1/2} v - \overline{w}_{v,t} \right]_i \right| \leq \epsilon_2 \max\left\{ \left| \left[\overline{V}_t^{-1/2} v \right]_i \right|, \frac{\|v\|_2}{\sqrt{\nu + L^2 n}} \right\}.$$

Proof. For simplicity, denote $b := \overline{V}_t^{-1/2} v$. By definition, we have $b \in \mathcal{B}$ (i.e., the deterministic set that we actually covered). We shall build a vector $w \in \mathcal{C}_2$ which has the desired property. Take any component b_i, with $i \in [d]$, then
(1) If $|b_i| < \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}$, then we can set $w_i = \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}\text{sign}(b_i)$ and we have

$$|w_i - b_i| \leq |w_i| = \frac{\epsilon_2 \|v\|_2}{\sqrt{\nu + L^2 n}}.$$

(2) If $|b_i| \geq \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}$, by the geometrical cover, we can find a point w_i such that $1 \leq \frac{|w_i|}{|b_i|} \leq 1 + \epsilon_2$. Too see this, suppose, without loss of generality, that b_i is positive. Note that, since b_i lies in the range $[\frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}, \frac{v_{\max}}{\sqrt{\nu}}]$ which is covered geometrically, there exists a real value $0 \leq k \leq \bar{j}$ such that $b_i = \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}(1+\epsilon_2)^k$. Then, if we set $w_i = \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}(1+\epsilon_2)^{\lceil k \rceil}$, we can easily verify the desired property. This implies

$$|w_i - b_i| \leq |w_i| - |b_i| \leq \epsilon_2 |b_i|,$$

where the left-hand side is from the reverse triangle inequality. The statement follows by combining the two cases. □

An immediate consequence of Proposition 5 is that, for all $v \in \mathcal{C}_1$, $t \in [n]$, there exists $\overline{w}_{v,t} \in \tilde{\mathcal{C}}_2$ which can be written as $\overline{w}_{v,t} = \overline{V}_t^{-1/2} v + \zeta$, where $\zeta \in \mathbb{R}^d$ is a vector of errors such that $|\zeta_i| \leq \epsilon_2 \max\left\{ \left| \left[\overline{V}_t^{-1/2} v \right]_i \right|, \frac{\|v\|_2}{\sqrt{\nu+L^2n}} \right\}$ for all $i \in [d]$.

Note that, by definition, $\tilde{\mathcal{C}}_2$ contains vectors with norm that scales in $\sqrt{d}$ (e.g., the vector with all components larger or equal to $v_{\max}/\sqrt{\nu}$, which has norm $v_{\max}\sqrt{d/\nu}$ belongs to $\tilde{\mathcal{C}}_2$). These vectors will create an undesirable dependency on d later on, and so we need to perform some pruning before proceeding. Take any $b \in \mathcal{B}$ and suppose that $b = Dv$ for $v \in \mathcal{C}_1$ and $D \in \mathcal{D}$. Let $\mathcal{I} := \{i \in [d] : |b_i| < \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}\}$ be the set of components i such that $|b_i|$ is below the starting point of our geometrical grid $\tilde{C}_{2,v}$ and $\mathcal{I}^c = [d] \setminus \mathcal{I}$. From the proof of Proposition 5, we know that the vector $w \in \tilde{\mathcal{C}}_2$ that is the closest to b is such that $|w_i| \le \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}$ for $i \in \mathcal{I}$ and $|w_i|/|b_i| \le 1+\epsilon_2$ for $i \in \mathcal{I}^c$. Therefore,

$$\|w\|_2^2 = \sum_{i\in\mathcal{I}} |w_i|^2 + \sum_{i\in\mathcal{I}^c} |w_i|^2 \le |\mathcal{I}|\frac{\epsilon_2^2\|v\|_2^2}{\nu+L^2n} + (1+\epsilon_2)^2 \sum_{i\in\mathcal{I}^c} |b_i|^2 \le \frac{d\epsilon_2^2\|v\|_2^2}{\nu+L^2n} + (1+\epsilon_2)^2\|b\|_2^2.$$

This implies that $\|w\|_2 \le \frac{\sqrt{d}\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}} + (1+\epsilon_2)\|b\|_2$. Recall that $\|b\|_2 \le \frac{v_{\max}}{\sqrt{\nu}}$ and $\|v\|_2 \le v_{\max}$. Thus, $\|w\|_2 \le \frac{\sqrt{d}\epsilon_2 v_{\max}}{\sqrt{\nu+L^2n}} + (1+\epsilon_2)\frac{v_{\max}}{\sqrt{\nu}} \le \frac{v_{\max}}{\sqrt{\nu}}\left(1+\epsilon_2(1+\sqrt{d})\right)$. This condition holds for all "useful" vectors in our cover (i.e., those that are the closest to some of the vectors we need to cover). Therefore, we can safely set $\mathcal{C}_2 = \left\{w \in \tilde{\mathcal{C}}_2 : \|w\|_2 \le \frac{v_{\max}}{\sqrt{\nu}}\left(1+\epsilon_2(1+\sqrt{d})\right)\right\}$ as our final cover. Note that Proposition 5 still holds for $\mathcal{C}_2$ since we removed only vectors that cannot be the closest to any of the points to be covered. In the following, we set $w_{\max} := \frac{v_{\max}}{\sqrt{\nu}}\left(1+\epsilon_2(1+\sqrt{d})\right)$ as the maximum norm of any vector in $\mathcal{C}_2$.

Let us now go back to bounding term (ii) in Eq. 76. Let $\overline{w}_{v,t} := \operatorname{argmin}_{w\in\mathcal{C}_2}\left\|\overline{V}_t^{-1/2}v - w\right\|_1$ be the vector in our cover $\mathcal{C}_2$ which is the closest to $\overline{V}_t^{-1/2}v$ uniformly over all components. Then,

$$\begin{aligned}(ii) &:= \sum_{s=1}^t v^T\overline{V}_t^{-1/2}\phi(X_s,A_s)\xi_s = \left(\overline{V}_t^{-1/2}v\right)^T\sum_{s=1}^t \phi(X_s,A_s)\xi_s \\ &= \left(\overline{V}_t^{-1/2}v - \overline{w}_{v,t}\right)^T W_t + \overline{w}_{v,t}^T W_t \le \underbrace{\left\|\overline{V}_t^{-1/2}v - \overline{w}_{v,t}\right\|_{\overline{V}_t}}_{(a)}\underbrace{\|W_t\|_{\overline{V}_t^{-1}}}_{(b)} + \underbrace{\overline{w}_{v,t}^T W_t}_{(c)},\end{aligned}$$

where we defined $W_t := \sum_{s=1}^t \phi(X_s,A_s)\xi_s$. We start from (a). Using the error-decomposition property from Proposition 5, we can write $\left\|\overline{V}_t^{-1/2}v - \overline{w}_{v,t}\right\|_{\overline{V}_t} = \|\zeta\|_{\overline{V}_t}$ for some vector $\zeta \in \mathbb{R}^d$ with $|\zeta_i| \le \epsilon_2 \max\left\{|[\overline{V}_t^{-1/2}v]_i|, \frac{\|v\|_2}{\sqrt{\nu+L^2n}}\right\}$ for all $i \in [d]$. Since this implies $|\zeta_i| \le \epsilon_2\left(|[\overline{V}_t^{-1/2}v]_i| + \frac{\|v\|_2}{\sqrt{\nu+L^2n}}\right)$, we have

$$\|\zeta\|_{\overline{V}_t} \overset{(d)}{\le} \epsilon_2\left\|\overline{V}_t^{-1/2}v\right\|_{\overline{V}_t} + \frac{\epsilon_2\|v\|_2}{\sqrt{\nu+L^2n}}\|\mathbf{1}_d\|_{\overline{V}_t} \overset{(e)}{\le} \epsilon_2\|v\|_2 + \frac{\epsilon_2\|v\|_2\sqrt{d}}{\sqrt{\nu+L^2n}}\left\|\overline{V}_t^{1/2}\right\|_2 \overset{(f)}{\le} \epsilon_2\|v\|_2 + \epsilon_2\|v\|_2\sqrt{d},$$

where in (d) we used the triangle inequality ($\mathbf{1}_d$ denotes the d-dimensional vector of ones), in (e) we used $\|\mathbf{1}_d\|_{\overline{V}_t} \le \left\|\overline{V}_t^{1/2}\right\|_2\|\mathbf{1}_d\|_2$, and in (f) we upper bounded the maximum eigenvalue of $\left\|\overline{V}_t^{1/2}\right\|_2$ by $\sqrt{\nu+L^2n}$. Therefore, we conclude,

$$(a) := \left\|\overline{V}_t^{-1/2}v - \overline{w}_{v,t}\right\|_{\overline{V}_t} \le \epsilon_2(1+\sqrt{d})\|v\|_2.$$

Term (b) can be bounded by Lemma 17. For any $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$(b) := \|W_t\|_{\overline{V}_t^{-1}} \le \sqrt{2\sigma^2 d\log\left(\frac{1+\frac{tL^2}{d\nu}}{\delta'}\right)}.$$

Term (c) can be bounded by Lemma 20 (whose bound holds uniformly over all elements in $\mathcal{C}_2$). Recall that $\|w\|_2 \le w_{\max}$ for all $w \in \mathcal{C}_2$. For any $\chi > 0$ and $\delta' \in (0,1)$, with probability at least

$1-\delta'$,

$$(c) := \overline{w}_{v,t}^T W_t \leq \sqrt{2\sigma^2\gamma_n \max\left\{\chi, \|\overline{w}_{v,t}\|_{\overline{V}_t}^2\right\} \log\left(\frac{\Gamma_{w_{\max}^2 L^2,\chi}|\mathcal{C}_2|}{\delta'}\right)}.$$

Note that, by definition of $\mathcal{C}_2$, $\|\overline{w}_{v,t}\|_{\overline{V}_t}^2 \geq \sigma_{\min}(\overline{V}_t)\|\overline{w}_{v,t}\|_2^2 \geq \frac{\nu d\epsilon_2^2\|v\|^2}{\nu+L^2 n} \geq \frac{\nu d^2\epsilon_2^2 v_{\min}^2}{\nu+L^2 n}$. Hence, setting $\chi \leftarrow \chi_n' := \frac{\nu d^2\epsilon_2^2 v_{\min}^2}{\nu+L^2 n}$,

$$\Gamma_{w_{\max}^2 L^2,\chi_n'} = 1 + \frac{\log(w_{\max}^2 L^2 n/\chi_n')}{\log\gamma_n} \leq 1 + \log(w_{\max}^2 L^2 n/\chi_n')\log(n)$$

where the last inequality is from $\log(1+\frac{1}{\log n}) \geq \frac{1}{2\log n}$ for $n \geq 2$. This yields

$$(c) \leq \sqrt{2\sigma^2\gamma_n\|\overline{w}_{v,t}\|_{\overline{V}_t}^2 \log\left(\frac{(1+\log(w_{\max}^2 L^2 n/\chi_n')\log(n))|\mathcal{C}_2|}{\delta'}\right)}.$$

Let us now bound $\|\overline{w}_{v,t}\|_{\overline{V}_t}^2$. We have

$$\begin{aligned}
\|\overline{w}_{v,t}\|_{\overline{V}_t}^2 &= \left\|\overline{w}_{v,t} \pm \overline{V}_t^{-1/2}v\right\|_{\overline{V}_t}^2 = \left\|\overline{w}_{v,t} - \overline{V}_t^{-1/2}v\right\|_{\overline{V}_t}^2 + \left\|\overline{V}_t^{-1/2}v\right\|_{\overline{V}_t}^2 + 2\left(\overline{w}_{v,t} - \overline{V}_t^{-1/2}v\right)^T \overline{V}_t\left(\overline{V}_t^{-1/2}v\right) \\
&\leq \left\|\overline{w}_{v,t} - \overline{V}_t^{-1/2}v\right\|_{\overline{V}_t}^2 + \|v\|_v^2 + 2\left\|\overline{w}_{v,t} - \overline{V}_t^{-1/2}v\right\|_{\overline{V}_t}\|v\|_2 \\
&\leq (\epsilon_2)^2(1+\sqrt{d})^2\|v\|_2^2 + \|v\|_2^2 + 2\epsilon_2(1+\sqrt{d})\|v\|_2^2 = \left(1+\epsilon_2(1+\sqrt{d})\right)^2\|v\|_2^2,
\end{aligned}$$

where in the last inequality we used the previous bound on $(a) = \left\|\overline{w}_{v,t} - \overline{V}_t^{-1/2}v\right\|_{\overline{V}_t}$.

Putting (a), (b), and (c) together we obtain the following bound on (ii):

$$\begin{aligned}
(ii) = v^T\overline{V}_t^{-1/2}W_t &\leq \|v\|_2\epsilon_2(1+\sqrt{d})\sqrt{2\sigma^2 d\log\left(\frac{1+\frac{nL^2}{d\nu}}{\delta'}\right)} \\
&+ \|v\|_2\left(1+\epsilon_2(1+\sqrt{d})\right)\sqrt{2\sigma^2\gamma_n\log\left(\frac{(1+\log(w_{\max}^2 L^2 n/\chi_n')\log(n))|\mathcal{C}_2|}{\delta'}\right)}.
\end{aligned}$$

If we now set $\epsilon_2 \leftarrow \frac{1}{2d\log n}$, we have $\chi_n' = \frac{\nu v_{\min}^2}{4(\nu+L^2 n)(\log n)^2}$. Setting $\chi_n'' = \chi_n'/(w_{\max}^2 L^2)$ and using $w_{\max} = \frac{v_{\max}}{\sqrt{\nu}}\left(1+\epsilon_2(1+\sqrt{d})\right) \leq \frac{v_{\max}}{\sqrt{\nu}}\gamma_n$, $\chi_n'' \geq \frac{\nu^2 v_{\min}^2}{4L^2(\nu+L^2 n)(\log n)^2 v_{\max}^2\gamma_n^2} = \chi_n$. Thus,

$$(ii) \leq \|v\|_2\left(\sqrt{\frac{2\sigma^2\log\left(\frac{1+\frac{nL^2}{d\nu}}{\delta'}\right)}{(\log n)^2}} + \sqrt{2\sigma^2\gamma_n^3\log\left(\frac{(1+\log(n/\chi_n)\log(n))}{\delta'}\right) + 2\gamma_n^3\log|\mathcal{C}_2|}\right).$$

Furthermore, using (77), the log-size of the cover $\mathcal{C}_2$ is

$$\begin{aligned}
\Upsilon_n = \log|\mathcal{C}_2| &\leq \log(|\mathcal{C}_1|) + d\log\left(2 + \frac{\log\left(2d\log n\frac{v_{\max}}{v_{\min}}\sqrt{\frac{\nu+L^2 n}{d\nu}}\right)}{\log(1+\frac{1}{2d\log n})}\right) \\
&\leq \log(|\mathcal{C}_1|) + d\log\left(2 + 4d\log\left(2d\log n\frac{v_{\max}}{v_{\min}}\sqrt{\frac{\nu+L^2 n}{d\nu}}\right)\log n\right)
\end{aligned}$$

To conclude the proof, we notice that the derivation above holds uniformly for all $v \in \mathcal{C}_1$ and $t \in [n]$ with probability at least $1-2\delta'$ since we applied both Lemma 17 (for term (b) in (ii)) and Lemma 20 (for term (c) in (ii)). Thus, the statement follows by setting $\delta = 2\delta'$. □

J.3 Auxiliary Results

Lemma 17. *[Lemma 9 of [4]] Let τ be a stopping time with respect to filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$ and $W_t := \sum_{s=1}^{t} \phi(X_s, A_s)\xi_s$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta'$,*

$$\|W_\tau\|_{\overline{V}_\tau^{-1}} \leq \sqrt{2\sigma^2 \log\left(\frac{\det(\overline{V}_\tau)^{1/2}\nu^{-d/2}}{\delta}\right)} \leq \sqrt{2\sigma^2 d \log\left(\frac{1+\frac{\tau L^2}{d\nu}}{\delta}\right)}.$$

The following result is a specialization of Lemma 2.6 of [31] or Lemma 4.2 of [32].

Lemma 18. *Let $n \in \mathbb{N}$ and $\{Y_t\}_{t=1}^{n}$ be a sequence of sub-Gaussian random variables adapted to filtration $\mathcal{F}$ such that $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$ and*

$$\forall \zeta \in \mathbb{R} : \mathbb{E}\left[e^{\zeta Y_t}|\mathcal{F}_{t-1}\right] \leq e^{\frac{\zeta^2\sigma_t^2}{2}},$$

where $\sigma_t^2 := \mathbb{V}\mathrm{ar}[Y_t|\mathcal{F}_{t-1}]$. Then, for all $\epsilon \geq 0, v > 0$,

$$\mathbb{P}\left\{\exists t \leq n : \sum_{s=1}^{t} Y_s \geq \epsilon, \sum_{s=1}^{t} \sigma_s^2 \leq v\right\} \leq e^{-\frac{\epsilon^2}{2v}}.$$

Proof. The result follows straightforwardly from Lemma 2.6 of [31] or Lemma 4.2 of [32] after optimizing for ζ. □

Lemma 19 (Lemma 14 of [7])**.** *Let $n \in \mathbb{N}$ and $\epsilon > 0$. Let $\{Y_t\}_{t=1}^{n}$ be a sequence of Gaussian random variables adapted to filtration $\mathcal{F}$ such that $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$ and $\mathbb{V}\mathrm{ar}[Y_t|\mathcal{F}_{t-1}] \leq b$ for some $b > 0$. Then*

$$\mathbb{P}\left\{\exists t \leq n : \sum_{s=1}^{t} Y_s \geq \sqrt{2\gamma_n P_t \log\frac{\Gamma_{b,\epsilon}}{\delta}}\right\} \leq \delta,$$

where $P_t = \max\{\epsilon, \sum_{s=1}^{t} \mathbb{V}\mathrm{ar}[Y_t|\mathcal{F}_{t-1}]\}$, $\gamma_n = 1 + \frac{1}{\log n}$, and $\Gamma_{b,\epsilon} = 1 + \frac{\log(nb/\epsilon)}{\log\gamma_n}$.

Proof. The proof uses the same peeling argument as in [7] but follows different steps.

Let $\tau \leq n$ be a stopping time with respect to $\mathcal{F}$ whose value will be specified later. Define $\Upsilon_t := \sum_{s=1}^{t} \mathbb{V}\mathrm{ar}[Y_t|\mathcal{F}_{t-1}]$ as the sum of predictable variances and $f(v) := \sqrt{2\gamma_n \max\{v,\epsilon\}\log\frac{1}{\delta'}}$. Let us define a sequence of scalars $v_{-1}, v_0, \ldots v_{k_n}$, which will be used to discretize the predictable variances, with $v_{-1} = 0$, v_0 to be specified later, $v_j = \gamma_n v_{j-1}$ for $j \geq 1$, and k_n such that $v_{k_n} \geq nb$ (which implies $v_{k_n} \geq \Upsilon_n$). Note that the theorem holds trivially when $\Upsilon_\tau = 0$, so we consider the case where this variable is positive. We have

$$\begin{aligned}
\mathbb{P}\left\{\sum_{s=1}^{\tau} Y_s \geq f(\Upsilon_\tau)\right\} &\overset{(a)}{\leq} \sum_{j=0}^{k_n} \mathbb{P}\left\{\sum_{s=1}^{\tau} Y_s \geq f(\Upsilon_\tau), \Upsilon_\tau \in (v_{j-1}, v_j]\right\} \\
&\overset{(b)}{\leq} \sum_{j=0}^{k_n} \mathbb{P}\left\{\sum_{s=1}^{\tau} Y_s \geq f(v_{j-1}), \Upsilon_\tau \leq v_j\right\} \overset{(c)}{\leq} \sum_{j=0}^{k_n} e^{-\frac{f(v_{j-1})^2}{2v_j}},
\end{aligned}$$

where (a) uses a union bound, (b) holds since f is non-decreasing, and (c) is from Lemma 18. Using the definition of $\{v_j\}_{j\geq -1}$,

$$\sum_{j=0}^{k_n} e^{-\frac{f(v_{j-1})^2}{2v_j}} = e^{-\frac{\gamma_n\epsilon\log\frac{1}{\delta'}}{v_0}} + \sum_{j=1}^{k_n} e^{-\frac{2\gamma_n\max\{v_j/\gamma_n,\epsilon\}\log\frac{1}{\delta'}}{2v_j}} \leq (\delta')^{\frac{\gamma_n\epsilon}{v_0}} + k_n\delta'.$$

Since $v_{k_n} = \gamma_n^{k_n} v_0$, we have that $k_n = \left\lceil \frac{\log(nb/v_0)}{\log(\gamma_n)} \right\rceil$ suffices to have $v_{k_n} \geq nb$. Setting $v_0 \leftarrow \gamma_n\epsilon$,

$$\mathbb{P}\left\{\sum_{s=1}^{\tau} Y_s > f(\Upsilon_\tau)\right\} \leq \delta'\left(1 + \left\lceil \frac{\log(nb/\epsilon) - \log(\gamma_n)}{\log(\gamma_n)} \right\rceil\right) \leq \delta'\left(1 + \frac{\log(nb/\epsilon)}{\log(\gamma_n)}\right) = \delta'\Gamma_{b,\epsilon}.$$

The result follows by setting $\delta \leftarrow \delta'\Gamma_{b,\epsilon}$ and $\tau \leftarrow \min\left\{t \leq n : \sum_{s=1}^{t} Y_s > \sqrt{2\gamma_n P_t \log\frac{\Gamma_{b,\epsilon}}{\delta}}\right\}$. □

The following result can be derived using a similar argument as in the proof of Lemma 15 of [7].

Lemma 20. *Let $\mathcal{C} \subset \{w \in \mathbb{R}^d : \|w\| \le b\}$ be a finite set of vectors in $\mathbb{R}^d$ with norm bounded by $b > 0$ and W_t as defined in Lemma 17. Then, for all $\epsilon > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left\{\exists t \le n, w \in \mathcal{C} : w^T W_t \ge \sqrt{2\sigma^2 \gamma_n \max\{\epsilon, \|w\|^2_{\overline{V}_t}\} \log\left(\frac{\Gamma_{b^2L^2,\epsilon}|\mathcal{C}|}{\delta}\right)}\right\} \le \delta,$$

where γ_n and $\Gamma_{b^2L^2,\epsilon}$ are those defined in Lemma 19.

Proof. Fix $w \in \mathcal{C}$. Note that

$$\frac{w^T W_t}{\sigma} = \sum_{s=1}^{t} \frac{w^T \phi(X_s, A_s)\xi_s}{\sigma}$$

is a sum of Gaussian random variables adapted to $\mathcal{F}$ such that

$$\mathbb{V}\mathrm{ar}\left[\frac{w^T \phi(X_s, A_s)\xi_s}{\sigma} | \mathcal{F}_{s-1}\right] = \frac{(w^T \phi(X_s, A_s))^2}{\sigma^2} \underbrace{\mathbb{V}\mathrm{ar}[\xi_s | \mathcal{F}_{s-1}]}_{=\sigma^2} \le \|w\|^2 \|\phi(Y_s, A_s)\|^2 \le b^2 L^2.$$

Furthermore,

$$\sum_{s=1}^{t} (w^T \phi(Y_s, A_s))^2 = \sum_{s=1}^{t} w^T \phi(Y_s, A_s)\phi(Y_s, A_s)^T w = \|w\|^2_{V_t} \le \|w\|^2_{\overline{V}_t},$$

where the last inequality is from $V_t \preceq \overline{V}_t$. Therefore, using Lemma 19, with probability at least $1 - \delta'$,

$$w^T W_t \le \sqrt{2\sigma^2 \gamma_n \max\{\epsilon, \|w\|^2_{\overline{V}_t}\} \log \frac{\Gamma_{b^2L^2,\epsilon}}{\delta'}}.$$

The result follows after taking a union bound over all elements in $\mathcal{C}$. □

K Additional Experiments

K.1 Implementation Details

In our implementation of SOLID, we ignore the projection of the parameters computed by regularized least squares onto Θ. Moreover, we remove the restriction that the alternative parameters should lie in Θ. That is, we use

$$\Theta_{\mathrm{alt}} := \{\theta' \in \mathbb{R}^d \mid \exists x \in \mathcal{X},\ a^\star_{\theta^\star}(x) \ne a^\star_{\theta'}(x)\}, \tag{78}$$

and similarly for $\overline{\Theta}_t$. In this case, for linear bandits with Gaussian noise, the infimum over alternative models in the constraint of (P) can be computed in closed form as

$$2\sigma^2 \inf_{\theta' \in \Theta_{\mathrm{alt}}} \sum_{x,a} \eta(x,a) d_{x,a}(\theta^\star, \theta') = \inf_{\theta' \in \Theta_{\mathrm{alt}}} \|\theta^\star - \theta'\|^2_{V_\eta} = \min_{\substack{x \in \mathcal{X}, \\ a \ne a^\star_{\theta^\star}(x)}} \frac{\Delta_{\theta^\star}(x,a)^2}{\|\phi(x,a) - \phi^\star_{\theta^\star}(x))\|^2_{V_\eta^{-1}}}, \tag{79}$$

where $V_\eta = \sum_{x,a} \eta(x,a)\phi(x,a)\phi(x,a)^\mathsf{T}$ and $\phi^\star_{\theta^\star}(x) = \phi(x, a^\star_{\theta^\star}(x))$. The same closed-form can be used for the infimum in the constraint (3). Regarding the exploitation test, we restrict the set of alternative reward parameters to those with "incompatible" optimal arm in the last observed context. That is, we use the test

$$\inf_{\theta' \in \widetilde{\Theta}_{t-1}} \|\widehat{\theta}_{t-1} - \theta'\|^2_{\overline{V}_{t-1}} > \beta_{t-1}, \tag{80}$$

where $\widetilde{\Theta}_{t-1} = \{\theta' \in \mathbb{R}^d \mid a^\star_{\widehat{\theta}_{t-1}}(X_t) \ne a^\star_{\theta'}(X_t)\}$. Once again, the infimum can be computed in closed form as before (without the minimum over contexts).

K.2 Experiment Configurations

We provide the detailed configurations of the experiments reported in the main paper. We use the same confidence intervals in all experiments. For SOLID, we set $\beta_t = \sigma^2(\log(t) + d\log\log(n))$ and $\gamma_t = \sigma^2(\log(S_t) + d\log\log(n))$ as prescribed by Thm. 1 (without numerical constants). For OAM, we use the same β_t for the exploitation test. For LinUCB, we use the confidence set of [4] without numerical constants. Similarly, we implement LinTS as defined in [5] but without the extra-sampling factor $\sqrt{d}$ used to prove its frequentist regret. All plots are the results of 100 runs with 95% Student's t confidence intervals.

In both experiments, for SOLID we set $\alpha_\omega = 1$, $\alpha_\lambda = 0.5$, and we normalize the gradients by context in l_2-norm. We do not reset the optimizer at the beginning of each phase. We use the theoretical exponential schedule for z_k and p_k as defined in Thm. 2. We set $z_0 = 1$, $\lambda_1 = 0$ for the first experiment and $z_0 = |\mathcal{A}|$, $\lambda_1 = 50$ for the second one. The reward noise is $\sigma = 0.5$ in the first experiment and $\sigma = 1$ in the second one.

Generation of Random Problems We adopt the following procedure in order to generate the random bandit models for the second experiment. We first randomly sample a sparse $|\mathcal{X}||\mathcal{A}| \times d$ feature matrix and a sparse vector $\theta^\star$ with entries uniformly distributed in $[0, 1]$. We then compute the resulting optimal arms for each context and check whether they span $\mathbb{R}^d$. If they do, we discard the generated features/parameter and repeat the previous procedure. Otherwise we keep the bandit problem. Discarding problems where the features of the optimal arms span $\mathbb{R}^d$ is done in order to avoid easy bandit problems in which exploration is not necessary (see [16])[17].

K.3 Parameter Analysis

We provide an empirical study of how different choices for the relevant parameters of SOLID affect the algorithm's performance in the toy problem of Sec. 6. We note that the purpose of this section is to build some intuition on how SOLID behaves with different parameters rather than assessing which configurations are globally better.

We use the two-context toy problem of Sec. 6 with $\xi = 0.1$ and $\sigma^2 = 1$. We study the effect of the following parameters, with corresponding default values.

- z_0 (default 30): the initial normalization factor;
- λ_1 (default 0): the initial multiplier;
- α^ω (default 0.1): learning rate for ω. We keep it fixed instead of decreasing with the phase length as suggested by the theory;
- α^λ (default 0.5): learning rate for λ. We keep it fixed as for α^ω;
- z_k, p_k (default $z_k = z_0 e^k$, $p_k = z_k e^{2k}$): the schedule for the phase length. We use the one for which we derive regret guarantees by default but we also experiment with other schedules. By default we do not reset the optimizer at the beginning of each phase.

We vary each parameter in a suitable range while keeping all the others fixed to their default values. The results are described in the following paragraphs.

Changing z_0 As mentioned in the main paper, the initial value of the parameter z controls both the feasibility of the optimization problem and the trade-off between minimizing regret and gathering information about the optimal arms when t is small. While a small value of z_0 might lead SOLID to collect a large amount of information, this might bring high finite regret as derived in the regret bound. Fig. 3(left) confirms this claim, where the value $z_0 = 1$ suffers high initial regret but the resulting curve has a better slope.

Changing λ_1 Though the initial multiplier has no particular impact on the regret bound, in practice it induces a behavior similar to z_0, where larger values lead SOLID to collect more information about $\theta^\star$ in the very first learning steps (see Fig. 3(right)).

[17] Problems that can be solved by a greedy strategy would not reveal any interesting empirical difference between SOLID and the other baselines.

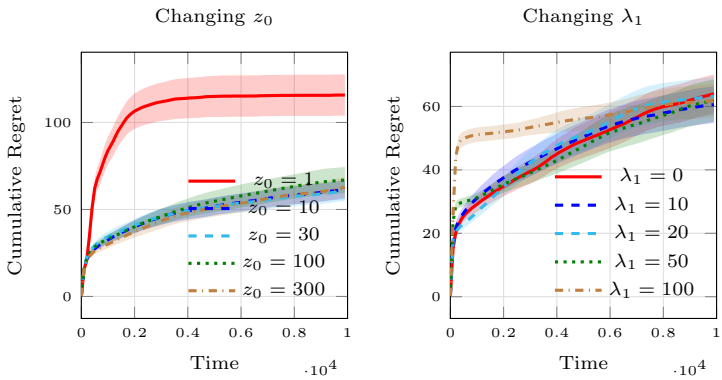

Figure 3: The effect of changing z_0 (left) and λ_1 (right).

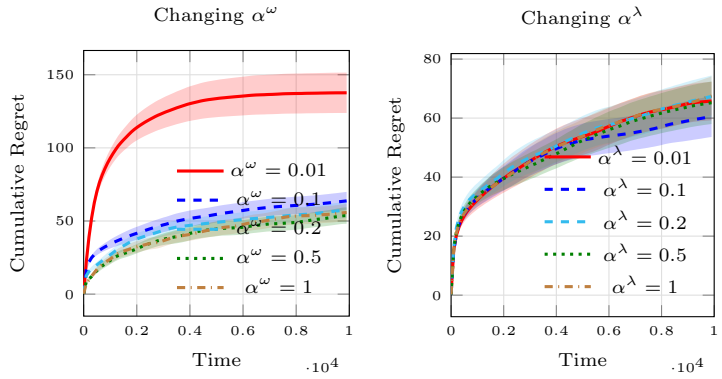

Figure 4: The effect of changing α^ω (left) and α^λ (right).

Changing the step sizes Fig. 4 shows the effect of varying α^ω and α^λ. In this particular case, α^λ seems to have no remarkable effect on SOLID's performance. On the other hand, the algorithm is quite sensible to the choice of α^ω, with very small values performing poorly since the policy is updated rarely and remains close to uniform for a long time. More aggressive step sizes seem to yield the best performance.

Phase schedule We test different schedules for z_k and p_k with respect to the one prescribed by the theory. We have $z_k = z_0 e^k, p_k = z_k e^{2k}$ (exp-exp), $z_k = z_0(1+k), p_k = z_k e^k$ (lin-exp), $z_k = z_0(1+k), p_k = z_k(1+k)^2$ (lin-pol), and $z_k = z_0(1+k), p_k = z_k(1+k)$ (lin-lin). Fig. 5(left) shows the result (here we set $z_0 = 1$ to better highlight the contribution of the different schedules). The exponential schedules are as expected more conservative since the algorithm spends more time optimizing with small values of z (i.e., seeks more information). The linear and polynomial schedules behave, on the other hand, more greedily and suffer less regret, though the resulting curve has larger slope.

We also test the effect of resetting the optimizer (middle and right plots in Fig. 5). We see that resetting the optimizer does not significantly affect the algorithm's performance both in case $z = 1$ and $z = 30$. This is likely due to the fact that phases are long (thanks to the exponential schedule) and that the algorithm spends many steps in the exploit phase, where no optimization is performed.

Tracking We compare the sampling strategy adopted by SOLID with the popular direct and cumulative tracking rules. Interestingly, Fig. 6(left) shows that sampling from ω constitutes a nice trade-off between cumulative tracking and the more aggressive direct tracking. Note that, while our theoretical results can be easily derived for cumulative tracking, we do not know whether the same can be done for direct tracking.

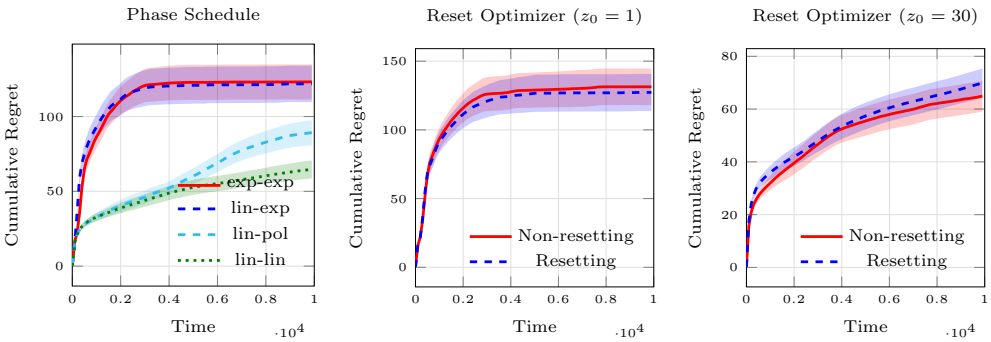

Figure 5: Different phase schedules (left) and effect of resetting the optimizer (middle and right plots).

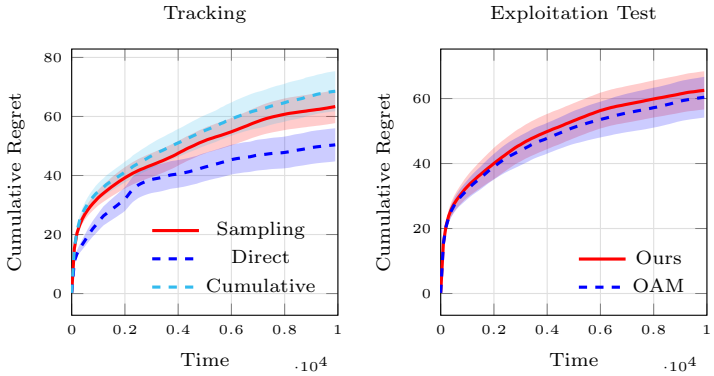

Figure 6: Different tracking strategies (left) and comparison with the exploitation test used in OAM.

Exploitation test We note that the test performed by SOLID in order to decide whether to explore or exploit is slightly different from the one adopted in OAM. In fact, the closed-form of the infimum over the alternative set (Eq. 79) leads to terms of the form $\Delta_{\widehat{\theta}_t}(x,a)^2/\|\phi(x,a)-\phi(x)^\star\|^2_{V_t^{-1}}$ while OAM uses $\Delta_{\widehat{\theta}_t}(x,a)^2/\|\phi(x,a)\|^2_{V_t^{-1}}$. We verify empirically (Fig. 6(right)) that the two tests lead to very similar performance.

K.4 Real Dataset

We report additional results on real data. We use the Jester Dataset [33] which consists of joke ratings in a continuous range from -10 to 10 for a total of 100 jokes and 73421 users. We select a subset of 40 jokes and 19181 users rating all these 40 jokes.

We build a linear contextual problem as follows. We first extract separate 36-dimensional user (context) and joke (arm) features via a low-rank matrix factorization. Then, we concatenate these user and joke features (thus obtaining vectors with 72 entries) and fit a 64×64 neural-network with ReLU non-linearities to predict the ratings of a random subset of 75% of the users, using these feature vectors as inputs. We obtain $R^2 \simeq 0.95$ on the remaining 25% users. Finally, we take the features extracted in the last layer of the network as the features for our bandit problem and the parameters of the same layer as $\theta^\star$. Rewards in our bandit problem are generated from this linear model by perturbing the prediction with $\mathcal{N}(0, 0.5^2)$ noise. We thus obtain a problem with $d = 65$ (the 64 hidden neurons plus the bias term), 40 arms (the jokes), and a total of 19181 users.

We run the algorithms for $2 \cdot 10^6$ steps, with each run randomizing a subset of 1% of the total users (hence $|\mathcal{X}| = 191$) and using all 40 arms. For SOLID, we use the same parameters as in the experiment with random models. Due to the computational bottleneck demonstrated in the previous experiments, we could not run OAM on this problem. The results are shown in Figure 7 and confirm that SOLID achieves superior performance than the other baselines.

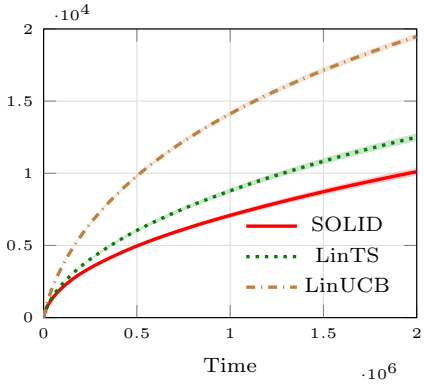

Figure 7: Experiment on a real dataset (Jester).