Scalable Representation Learning in Linear Contextual Bandits with Constant Regret Guarantees

Andrea Tirinzoni
META
tirinzoni@fb.com

Matteo Papini
Universitat Pompeu Fabra
matteo.papini@upf.edu

Ahmed Touati
META
atouati@fb.com

Alessandro Lazaric
META
lazaric@fb.com

Matteo Pirotta
META
pirotta@fb.com

Abstract

We study the problem of representation learning in stochastic contextual linear bandits. While the primary concern in this domain is usually to find *realizable* representations (i.e., those that allow predicting the reward function at any context-action pair exactly), it has been recently shown that representations with certain spectral properties (called *HLS*) may be more effective for the exploration-exploitation task, enabling *LinUCB* to achieve constant (i.e., horizon-independent) regret. In this paper, we propose BANDITSRL, a representation learning algorithm that combines a novel constrained optimization problem to learn a realizable representation with good spectral properties with a generalized likelihood ratio test to exploit the recovered representation and avoid excessive exploration. We prove that BANDITSRL can be paired with any no-regret algorithm and achieve constant regret whenever an *HLS* representation is available. Furthermore, BANDITSRL can be easily combined with deep neural networks and we show how regularizing towards *HLS* representations is beneficial in standard benchmarks.

1 Introduction

The contextual bandit is a general framework to formalize the exploration-exploitation dilemma arising in sequential decision-making problems such as recommendation systems, online advertising, and clinical trials [e.g., 1]. When solving real-world problems, where contexts and actions are complex and high-dimensional (e.g., users' social graph, items' visual description), it is crucial to provide the bandit algorithm with a suitable representation of the context-action space. While several representation learning algorithms have been proposed in supervised learning and obtained impressing empirical results [e.g., 2, 3], how to *efficiently* learn representations that are effective for the exploration-exploitation problem is still relatively an open question.

The primary objective in representation learning is to find features that map the context-action space into a lower-dimensional embedding that allows fitting the reward function accurately, i.e., *realizable* representations [e.g., 4–10]. Within the space of realizable representations, bandit algorithms leveraging features of smaller dimension are expected to learn faster and thus have smaller regret. Nonetheless, Papini et al. [11] have recently shown that, even among realizable features, certain representations are naturally better suited to solve the exploration-exploitation problem. In particular, they proved that LINUCB [12, 13] can achieve constant regret when provided with a "good" representation. Interestingly, this property is not related to "global" characteristics of the feature map (e.g., dimension, norms), but rather on a spectral property of the representation (the space associated to optimal actions should cover the context-action space, see HLS property in Def. 2.1). This naturally

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

raises the question whether it is possible to learn such representation at the same time as solving the contextual bandit problem. Papini et al. [11] provided a first positive answer with the LEADER algorithm, which is proved to perform as well as the best realizable representation in a given set up to a logarithmic factor in the number of representations. While this allows constant regret when a realizable HLS representation is available, the algorithm suffers from two main limitations: **1)** it is entangled with LINUCB and it can hardly be generalized to other bandit algorithms; **2)** it learns a different representation for each context-action pair, thus making it hard to extend beyond finite representations to arbitrary functional space (e.g., deep neural networks).

In this paper, we address those limitations through BANDITSRL, a novel algorithm that decouples representation learning and exploration-exploitation so as to work with any no-regret contextual bandit algorithm and to be easily extended to general representation spaces. BANDITSRL combines two components: 1) a representation learning mechanism based on a constrained optimization problem that promotes "good" representations while preserving realizability; and 2) a generalized likelihood ratio test (GLRT) to avoid over exploration and fully exploit the properties of "good" representations. The main contributions of the paper can be summarized as follows:

1. We show that adding a GLRT on the top of any no-regret algorithm enables it to exploit the properties of a HLS representation and achieve constant regret. This generalizes the constant regret result for LINUCB in [11] to any no-regret algorithm.
2. Similarly, we show that BANDITSRL can be paired with any no-regret algorithm and perform effective representation selection, including achieving constant regret whenever a HLS representation is available in a given set. This generalizes the result of LEADER beyond LINUCB. In doing this we also improve the analysis of the misspecified case and prove a tighter bound on the time to converge to realizable representations. Furthermore, numerical simulations in synthetic problems confirm that BANDITSRL is empirically competitive with LEADER.
3. Finally, in contrast to LEADER, BANDITSRL can be easily scaled to complex problems where representations are encoded through deep neural networks. In particular, we show that the Lagrangian relaxation of the constrained optimization problem for representation learning becomes a regression problem with an auxiliary representation loss promoting HLS-like representations. We test different variants of the resulting NN-BANDITSRL algorithm showing how the auxiliary representation loss improves performance in a number of dataset-based benchmarks.

2 Preliminaries

We consider a stochastic contextual bandit problem with context space $\mathcal{X}$ and finite action set $\mathcal{A}$. At each round $t \geq 1$, the learner observes a context x_t sampled i.i.d. from a distribution ρ over $\mathcal{X}$, selects an action $a_t \in \mathcal{A}$, and receives a reward $y_t = \mu(x_t, a_t) + \eta_t$ where η_t is a zero-mean noise and $\mu : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the expected reward. The objective of a learner $\mathfrak{A}$ is to minimize its pseudo-regret $R_T := \sum_{t=1}^{T} \big(\mu^\star(x_t) - \mu(x_t, a_t)\big)$ for any $T \geq 1$, where $\mu^\star(x_t) := \max_{a \in \mathcal{A}} \mu(x_t, a)$. We assume that for any $x \in \mathcal{X}$ the optimal action $a_x^\star := \operatorname{argmax}_{a \in \mathcal{A}} \mu(x, a)$ is unique and we define the gap $\Delta(x, a) := \mu^\star(x) - \mu(x, a)$. We say that $\mathfrak{A}$ is a no-regret algorithm if, for any instance of μ, it achieves sublinear regret, i.e., $R_T = o(T)$.

We consider the problem of representation learning in given a candidate function space $\Phi \subseteq \big\{\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{d_\phi}\big\}$, where the dimensionality d_ϕ may depend on the feature ϕ. Let $\theta_\phi^\star = \operatorname{argmin}_{\theta \in \mathbb{R}^{d_\phi}} \mathbb{E}_{x \sim \rho}\big[\sum_a (\phi(x, a)^\mathsf{T}\theta - \mu(x, a))^2\big]$ be the best linear fit of μ for representation ϕ. We assume that Φ contains a linearly realizable representation.

Assumption 1 (Realizability)**.** *There exists an (unknown) subset $\Phi^\star \subseteq \Phi$ such that, for each $\phi \in \Phi^\star$, $\mu(x, a) = \phi(x, a)^\mathsf{T}\theta_\phi^\star, \forall x \in \mathcal{X}, a \in \mathcal{A}$.*

Assumption 2 (Regularity)**.** *Let $\mathcal{B}_\phi := \{\theta \in \mathbb{R}^{d_\phi} : \|\theta\|_2 \leq B_\phi\}$ be a ball in $\mathbb{R}^{d_\phi}$. We assume that, for each $\phi \in \Phi$, $\sup_{x,a} \|\phi(x, a)\|_2 \leq L_\phi$, $\|\theta_\phi^\star\|_2 \leq B_\phi$, $\sup_{x,a} |\phi(x, a)^\mathsf{T}\theta| \leq 1$ for any $\theta \in \mathcal{B}_\phi$ and $|y_t| \leq 1$ almost surely for all t. We assume parameters L_ϕ and B_ϕ are known. We also assume the minimum gap $\Delta = \inf_{x \in \mathcal{X} : \rho(x) > 0, a \in \mathcal{A}, \Delta(x,a) > 0} \{\Delta(x, a)\} > 0$ and that $\lambda_{\min}\Big(\frac{1}{|\mathcal{A}|} \sum_a \mathbb{E}_{x \sim \rho}[\phi(x, a)\phi(x, a)^\mathsf{T}]\Big) > 0$ for any $\phi \in \Phi^\star$, i.e, all realizable representations are non-redundant.*

Under Asm. 1, when $|\Phi| = 1$, the problem reduces to a stochastic linear contextual bandit and can be solved using standard algorithms, such as LINUCB/OFUL [12, 13], LinTS [14], and ϵ-greedy [15], which enjoy sublinear regret and, in some cases, logarithmic problem-dependent regret. Recently, Papini et al. [11] showed that LINUCB only suffers constant regret when a *realizable* representation is HLS, i.e., when the features of optimal actions span the entire d_ϕ-dimensional space. HLS

Definition 2.1 (HLS Representation)**.** *A representation ϕ is* HLS *(the acronym refers to the last names of the authors of [16]) if*

$$\lambda^\star(\phi) := \lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\phi(x, a_x^\star)\phi(x, a_x^\star)^\mathsf{T}\right]\right) > 0$$

where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A.

Papini et al. showed that HLS, together with realizability, is a sufficient and necessary property for achieving constant regret in contextual stochastic linear bandits for non-redundant representations.

In order to deal with the general case where Φ may contain non-realizable representations, we rely on the following misspecification assumption from [11].

Assumption 3 (Misspecification)**.** *For each $\phi \notin \Phi^\star$, there exists $\epsilon_\phi > 0$ such that*

$$\min_{\theta\in\mathcal{B}_\phi}\min_{\pi:\mathcal{X}\to\mathcal{A}}\mathbb{E}_{x\sim\rho}\left[\left(\phi(x,\pi(x))^\mathsf{T}\theta - \mu(x,\pi(x))\right)^2\right] \geq \epsilon_\phi.$$

This assumption states that any non-realizable representation has a minimum level of misspecification on average over contexts and for any context-action policy. In the finite-context case, a sufficient condition for Asm. 3 is that, for each $\phi \notin \Phi^\star$, there exists a context $x \in \mathcal{X}$ with $\rho(x) > 0$ such that $\phi(x, a)^\mathsf{T}\theta \neq \mu(x, a)$ for all $a \in \mathcal{A}$ and $\theta \in \mathcal{B}_\phi$.

Related work. Several papers have focused on contextual bandits with an arbitrary function space to estimate the reward function under realizability assumptions [e.g., 4, 5, 7]. While these works consider a similar setting to ours, they do not aim to learn "good" representations, but rather focus on the exploration-exploitation problem to obtain sublinear regret guarantees. This often corresponds to recovering the maximum likelihood representation, which may not lead to the best regret. After the work in [11], the problem of representation learning with constant regret guarantees has also been studied in reinforcement learning [17, 18]. As these approaches build on the ideas in [11], they inherit the same limitations as [11].

Another related literature is the one of expert learning and model selection in bandits [e.g., 19–25], where the objective is to select the best candidate among a set of base learning algorithms or experts. While these algorithms are general and can be applied to different settings, including representation learning with a finite set of candidates, they may not be able to effectively leverage the specific structure of the problem. Furthermore, at the best of our knowledge, these algorithms suffers a polynomial dependence in the number of base algorithms ($|\Phi|$ in our setting) and are limited to worst-case regret guarantees. Whether the $\sqrt{T}$ or $\mathrm{poly}(|\Phi|)$ dependency can be improved in general is an open question (see [25] and [11, App. A]). Finally, [8, 26] studied the specific problem of model selection with nested linear representations, where the best representation is the one with the smallest dimension for which the reward is realizable.

Several works have recently focused on theoretical and practical investigation of contextual bandits with neural networks (NNs) [27–29]. While their focus was on leveraging the representation power of NNs to correctly predict the rewards, here we focus on learning representations with good spectral properties through a novel auxiliary loss. A related approach to our is [29] where the authors leverage self-supervised auxiliary losses for representation learning in image-based bandit problems.

3 A General Framework for Representation Learning

We introduce BANDITSRL (*Bandit Spectral Representation Learner*), an algorithm for stochastic contextual linear bandit that efficiently decouples representation learning from exploration-exploitation. As illustrated in Alg. 1, BANDITSRL has access to a fixed-representation contextual bandit algorithm $\mathfrak{A}$, the *base algorithm*, and it is built around two key mechanisms: ❶ a constrained optimization problem where the objective is to minimize a representation loss $\mathcal{L}$ to favor representations with HLS properties, whereas the constraint ensures realizability; ❷ a generalized likelihood ratio test (GLRT)

Algorithm 1 BANDITSRL

1: **Input:** representations Φ, no-regret algorithm $\mathfrak{A}$, confidence $\delta \in (0,1)$, update schedule $\gamma > 1$
2: Initialize $j = 0$, $\phi_j, \theta_{\phi_j,0}$ arbitrarily, $V_0(\phi_j) = \lambda I_{d_{\phi_j}}$, $t_j = 1$, let $\delta_j := \delta/(2(j+1)^2)$
3: **for** $t = 1, \dots$ **do**
4: Observe context x_t
5: **if** $\mathrm{GLR_{t-1}(x_t;\phi_j)} > \beta_{t-1,\delta/|\Phi|}(\phi_j)$ **then**
6: Play $a_t = \mathrm{argmax}_{a\in\mathcal{A}} \left\{\phi_j(x_t,a)^\mathsf{T}\theta_{\phi_j,t-1}\right\}$ and observe reward y_t
7: **else**
8: Play $a_t = \mathfrak{A}_t\big(x_t;\phi_j,\delta_j/|\Phi|\big)$, observe reward y_t, and feed it into $\mathfrak{A}$
9: **end if**
10: **if** $t = \lceil \gamma t_j \rceil$ **and** $|\Phi| > 1$ **then**
11: Set $j = j+1$ and $t_j = t$
12: Compute $\phi_j = \mathrm{argmin}_{\phi\in\Phi_t} \left\{\mathcal{L}_t(\phi)\right\}$ and reset $\mathfrak{A}$
13: **end if**
14: **end for**

to ensure that, if a HLS representation is learned, the base algorithm $\mathfrak{A}$ does not over-explore and the "good" representation is exploited to obtain constant regret.

Mechanism ❶ (line 12). The first challenge when provided with a generic set Φ is to ensure that the algorithm does not converge to selecting misspecified representations, which may lead to linear regret. This is achieved by introducing a hard constraint in the representation optimization, so that BANDITSRL only selects representations in the set (see also [11, App. F]),

$$\Phi_t := \left\{\phi \in \Phi : \min_{\theta\in\mathcal{B}_\phi} E_t(\phi,\theta) \le \min_{\phi'\in\Phi}\min_{\theta\in\mathcal{B}_{\phi'}} \left\{E_t(\phi',\theta) + \alpha_{t,\delta}(\phi')\right\}\right\} \tag{1}$$

where $E_t(\phi,\theta) := \frac{1}{t}\sum_{s=1}^t \left(\phi(x_s,a_s)^T\theta - y_s\right)^2$ is the empirical mean-square error (MSE) of model (ϕ,θ) and $\alpha_{t,\delta}(\phi) := \frac{40}{t}\log\left(\frac{8|\Phi|^2(12L_\phi B_\phi t)^{d_\phi} t^3}{\delta}\right) + \frac{2}{t}$. This condition leverages the existence of a realizable representation in Φ_t to eliminate representations whose MSE is not compatible with the one of the realizable representation, once accounted for the statistical uncertainty (i.e., $\alpha_{t,\delta}(\phi)$).

Subject to the realizability constraint, the representation loss $\mathcal{L}_t(\phi)$ favours learning a HLS representation (if possible). As illustrated in Def. 2.1, a HLS representation is such that the expected design matrix associated to the optimal actions has a positive minimum eigenvalue. Unfortunately it is not possible to directly optimize for this condition, since we have access to neither the context distribution ρ nor the optimal action in each context. Nonetheless, we can design a loss that works as a proxy for the HLS property whenever $\mathfrak{A}$ is a no-regret algorithm. Let $V_t(\phi) = \lambda I_{d_\phi} + \sum_{s=1}^t \phi(x_s,a_s)\phi(x_s,a_s)^\mathsf{T}$ be the empirical design matrix built on the context-actions pairs observed up to time t, then we define $\mathcal{L}_{\mathrm{eig},t}(\phi) := -\lambda_{\min}\big(V_t(\phi) - \lambda I_{d_\phi}\big)/L_\phi^2$, where the normalization factor ensures invariance w.r.t. the feature norm. Intuitively, the empirical distribution of contexts $(x_t)_{t\ge1}$ converges to ρ and the frequency of optimal actions selected by a no-regret algorithm increases over time, thus ensuring that $V_t(\phi)/t$ tends to behave as the design matrix under optimal arms $\mathbb{E}_{x\sim\rho}[\phi(x,a_x^\star)\phi(x,a_x^\star)^\mathsf{T}]$. As discussed in Sect. 5 alternative losses can be used to favour learning HLS representations.

Mechanism ❷ (line 5). While Papini et al. [11] proved that LINUCB is able to exploit HLS representations, other algorithms such as ϵ-greedy may keep forcing exploration and do not fully take advantage of HLS properties, thus failing to achieve constant regret. In order to prevent this, we introduce a *generalized likelihood ratio* test (GLRT). At each round t, let ϕ_{t-1} be the representation used at time t, then BANDITSRL decides whether to act according to the base algorithm $\mathfrak{A}$ with representation ϕ_{t-1} or fully exploit the learned representation and play greedily w.r.t. it. Denote by $\theta_{\phi,t-1} = V_{t-1}(\phi)^{-1}\sum_{s=1}^{t-1}\phi(x_s,a_s)y_s$ the regularized least-squares parameter at time t for representation ϕ and by $\pi_{t-1}^\star(x;\phi) = \mathrm{argmax}_{a\in\mathcal{A}}\left\{\phi(x,a)^\mathsf{T}\theta_{\phi,t-1}\right\}$ the associated greedy policy. Then, BANDITSRL selects the greedy action $\pi_{t-1}^\star(x_t;\phi_{t-1})$ when the GLR test is active, otherwise it selects the action proposed by the base algorithm $\mathfrak{A}$. Formally, for any $\phi\in\Phi$ and $x\in\mathcal{X}$, we define

the generalized likelihood ratio as

$$\mathrm{GLR}_{t-1}(x;\phi) := \min_{a \neq \pi^\star_{t-1}(x;\phi)} \frac{\big(\phi(x,\pi^\star_{t-1}(x;\phi)) - \phi(x,a)\big)^\mathsf{T} \theta_{\phi,t-1}}{\|\phi(x,\pi^\star_{t-1}(x;\phi)) - \phi(s,a)\|_{V_{t-1}(\phi)^{-1}}} \tag{2}$$

and, given $\beta_{t-1,\delta}(\phi) = \sigma\sqrt{2\log(1/\delta) + d_\phi \log(1 + (t-1)L_\phi^2/(\lambda d_\phi))} + \sqrt{\lambda} B_\phi$, the GLR test is $\mathrm{GLR}_{t-1}(x;\phi) > \beta_{t-1,\delta/|\Phi|}(\phi)$ [16, 30, 31]. If this happens at time t and ϕ_{t-1} is realizable, then we have enough confidence to conclude that the greedy action is optimal, i.e., $\pi^\star_{t-1}(x_t;\phi_{t-1}) = a^\star_{x_t}$. An important aspect of this test is that it is run on the current context x_t and it does not require evaluating global properties of the representation. While at any time t it is possible that a non-HLS non-realizable representation may pass the test, the GLRT is sound as **1)** exploration through $\mathfrak{A}$ and the representation learning mechanism work in synergy to guarantee that *eventually* a realizable representation is always provided to the GLRT; **2)** only HLS representations are guaranteed to consistently trigger the test at any context x.

In practice, BANDITSRL does not update the representation at each step but in phases. This is necessary to avoid too frequent representation changes and control the regret, but also to make the algorithm more computationally efficient and practical. Indeed, updating the representation may be computationally expensive in practice (e.g., retraining a NN) and a phased scheme with γ parameter reduces the number of representation learning steps to $J \approx \lceil \log_\gamma(T) \rceil$. The algorithm $\mathfrak{A}$ is reset at the beginning of a phase j when the representation is selected and it is run on the samples collected during the current phase when the base algorithm is selected. If $\mathfrak{A}$ is able to leverage off-policy data, at the beginning of a phase j, we can warm-start it by providing ϕ_j and all the past data $(x_s, a_s, y_s)_{s \le t_j}$. While the reset is necessary for dealing with *any* no-regret algorithm, it can be removed for algorithms such as LINUCB and ϵ-greedy without affecting the theoretical guarantees.

Comparison to LEADER. We first recall the basic structure of LEADER. Denote by $\mathrm{UCB}_t(x,a,\phi)$ the upper-confidence bound computed by LINUCB for the context-action pair (x,a) and representation ϕ after t steps. Then LEADER selects the action $a_t \in \mathrm{argmax}_{a\in\mathcal{A}} \min_{\phi\in\Phi_t} \mathrm{UCB}_t(x_t,a,\phi)$. Unlike the constrained optimization problem in BANDITSRL, this mechanism couples representation learning and exploration-exploitation and it requires optimizing a representation for the current x_t and for each action a. Indeed, LEADER does not output a single representation and possibly chooses different representations for each context-action pair. While this enables LEADER to mix representations and achieve constant regret in some cases even when Φ does not include any HLS representation, it leads to two major drawbacks: **1)** the representation selection is directly entangled with the LINUCB exploration-exploitation strategy, **2)** it is impractical in problems where Φ is an infinite functional space (e.g., a deep neural network). The mechanisms ❶ and ❷ successfully address these limitations and enable BANDITSRL to be paired with any no-regret algorithm and to be scaled to any representation class as illustrated in the next section.

3.1 Extension to Neural Networks

We now consider a representation space Φ defined by the last layer of a NN. We denote by $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ the last layer and by $f(x,a) = \phi(x,a)^\mathsf{T}\theta$ the full NN, where θ are the last-layer weights. We show how BANDITSRL can be easily adapted to work with deep neural networks (NN).

First, the GLRT requires only to have access to the current context x_t and representation ϕ_j, i.e., the features defined by the last layer of the current network, and its cost is linear in the number of actions. *Second*, the phased scheme allows lazy updates, where we retrain the network only $\log_\gamma(T)$ times. *Third*, we can run any bandit algorithm with a representation provided by the NN, including LINUCB, LinTS, and ϵ-greedy. *Fourth*, the representation learning step can be adapted to allow efficient optimization of a NN. We consider a regularized problem obtained through an approximation of the constrained problem:

$$\begin{aligned} &\operatorname*{argmin}_\phi \left\{ \mathcal{L}_t(\phi) - c_{\mathrm{reg}} \left(\min_{\phi',\theta'} \left\{ E_t(\phi',\theta') + \alpha_{t,\delta}(\phi') \right\} - \min_\theta E_t(\phi,\theta) \right) \right\} \\ &= \operatorname*{argmin}_\phi \min_\theta \left\{ \mathcal{L}_t(\phi) + c_{\mathrm{reg}}\, E_t(\phi,\theta) \right\}. \end{aligned} \tag{3}$$

where $c_{\mathrm{reg}} \ge 0$ is a tunable parameter. The fact we consider c_{reg} constant allows us to ignore terms that do not depend on either ϕ or θ. This leads to a convenient regularized loss that aims to minimize

the MSE (second term) while enforcing some spectral property on the last layer of the NN (first term). In practice, we can optimize this loss by stochastic gradient descent over a *replay buffer* containing the samples observed over time. The resulting algorithm, called NN-BANDITSRL, is a direct and elegant generalization of the theoretically-grounded algorithm.

While in theory we can optimize the regularized loss (3) with all the samples, in practice it is important to better control the sample distribution. As the algorithm progresses, we expect the replay buffer to contain an increasing number of samples obtained by optimal actions, which may lead the representation to solely fit optimal actions while increasing misspecification on suboptimal actions. This may compromise the behavior of the algorithm and ultimately lead to high regret. This is an instance of *catastrophic forgetting* induced by a biased/shifting sample distribution [e.g., 32]. To prevent this phenomenon, we store two replay buffers: *i)* an explorative buffer $\mathcal{D}_{\mathfrak{A},t}$ with samples obtained when $\mathfrak{A}$ was selected; *ii)* an exploitative buffer $\mathcal{D}_{\mathrm{glrt},t}$ with samples obtained when GLRT triggered and greedy actions were selected. The explorative buffer $\mathcal{D}_{\mathfrak{A},t}$ is used to compute the MSE $E_t(\phi, \theta)$. While this reduces the number of samples, it improves the robustness of the algorithm by promoting realizability. On the other hand, we use all the samples $\mathcal{D}_t = \mathcal{D}_{\mathfrak{A},t} \cup \mathcal{D}_{\mathrm{glrt},t}$ for the representation loss $\mathcal{L}(\phi)$. This is coherent with the intuition that mechanism ❶ works when the design matrix V_t drifts towards the design matrix of optimal actions, which is at the core of the HLS property. Refer to App. C for a more detailed description of NN-BANDITSRL.

4 Theoretical Guarantees

In this section, we provide a complete characterization of the theoretical guarantees of BANDITSRL when Φ is a finite set of representations, i.e., $|\Phi| < \infty$. We consider the update scheme with $\gamma = 2$.

4.1 Constant Regret Bound for HLS Representations

We first study the case where a realizable HLS representation is available. For the characterization of the behavior of the algorithm, we need to introduce the following times:

- τ_{elim}: an upper-bound to the time at which all non-realizable representations are eliminated, i.e., for all $t \geq \tau_{\mathrm{elim}}$, $\Phi_t = \Phi^\star$;
- τ_{HLS}: an upper-bound to the time (if it exists) after which the HLS representation is selected, i.e., $\phi_t = \phi^\star$ for all $t \geq \tau_{\mathrm{HLS}}$, where $\phi^\star \in \Phi^\star$ is the unique HLS realizable representation;
- τ_{glrt}: an upper-bound to the time (if it exists) such that the GLR test triggers for the HLS representation $\phi^\star$ for all $t \geq \tau_{\mathrm{glrt}}$.

We begin by deriving a constant problem-dependent regret bound for BANDITSRL with HLS representations. The proof and explicit values of the constants are reported in App. B.[1]

Theorem 4.1. *Let $\mathfrak{A}$ be any no-regret algorithm for stochastic contextual linear bandits, Φ satisfy Asm. 1- 3, $|\Phi| < \infty$, $\gamma = 2$, and $\mathcal{L}_t(\phi) = \mathcal{L}_{\mathrm{eig},t}(\phi) := -\lambda_{\min}(V_t(\phi) - \lambda I_{d_\phi})/L_\phi^2$. Moreover, let $\Phi^\star$ contains a unique* HLS *representation $\phi^\star$. Then, for any $\delta \in (0,1)$ and $T \in \mathbb{N}$, the regret of* BANDITSRL *is bounded, with probability at least $1 - 4\delta$, as*[2]

$$R_T \leq 2\tau_{\mathrm{elim}} + \max_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}((\tau_{\mathrm{opt}} - \tau_{\mathrm{elim}}) \wedge T, \phi, \delta_{\log_2(\tau_{\mathrm{opt}} \wedge T)}/|\Phi|) \log_2(\tau_{\mathrm{opt}} \wedge T),$$

where $\delta_j := \delta/(2(j+1)^2)$ and

$$\tau_{\mathrm{opt}} = \tau_{\mathrm{glrt}} \vee \tau_{\mathrm{HLS}} \vee \tau_{\mathrm{elim}} \lesssim \tau_{\mathrm{alg}} + \frac{L_{\phi^\star}^2 \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)} \left(\frac{L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} + \frac{d_{\phi^\star}}{\Delta^2} + \frac{d}{(\min_{\phi \notin \Phi^\star} \epsilon_\phi)\Delta} \right), \quad (4)$$

with τ_{alg} a finite *(independent from the horizon T) constant depending on algorithm $\mathfrak{A}$ (see Tab. 1) and $\overline{R}_{\mathfrak{A}}(\tau, \phi, \delta)$ an anytime bound (non-decreasing in τ and $1/\delta$) on the regret accumulated over τ steps by $\mathfrak{A}$ using representation ϕ and confidence level δ.*

[1] While Thm. 4.1 provides high-probability guarantees, we can easily derive a constant expected-regret bound by running BANDITSRL with a decreasing schedule for δ and with a slightly different proof.

[2] We denote by $a \wedge b$ (resp. $a \vee b$) the minimum (resp. the maximum) between a and b.

The key finding of the previous result is that BANDITSRL achieves constant regret whenever a realizable HLS representation is available in the set Φ, which may contain non-realizable as well as realizable non-HLS representations. The regret bound above also illustrates the "dynamics" of the algorithm and three main regimes. In the early stages, non-realizable representations may be included in Φ_t, which may lead to suffering linear regret until time τ_{elim} when the constraint in the representation learning step filters out all non-realizable representations (first term in the regret bound). At this point, BANDITSRL leverages the loss $\mathcal{L}$ to favor HLS representations and the base algorithm $\mathfrak{A}$ to perform effective exploration-exploitation. This leads to the second term in the bound, which corresponds to an upper-bound to the sum of the regrets of $\mathfrak{A}$ in each phase in between τ_{elim} and $\tau_{\text{glrt}} \vee \tau_{\text{HLS}}$, which is roughly $\sum_{j_{\tau_{\text{elim}}}<j<j_{\tau_{\text{opt}}}} \overline{R}_{\mathfrak{A}}(t_{j+1}-t_j,\phi_j) \leq \max_{\phi\in\Phi^\star} \overline{R}_{\mathfrak{A}}(\tau_{\text{opt}}-\tau_{\text{elim}},\phi)\log_2(\tau_{\text{opt}})$. In this second regime, in some phases the algorithm may still select non-HLS representations, which leads to a worst-case bound over all realizable representations in $\Phi^\star$. Finally, after $\tau_{\text{glrt}} \vee \tau_{\text{HLS}}$ the GLRT consistently triggers over time. During this last regime, BANDITSRL has reached enough accuracy and confidence so that the greedy policy of the HLS representation is indeed optimal and no additional regret is incurred.

We notice that the only dependency on the number of representations $|\Phi|$ in Thm. 4.1 is due to the rescaling of the confidence level $\delta \mapsto \delta/|\Phi|$. Since standard algorithms have a logarithmic dependence in $1/\delta$, this only leads to a logarithmic dependency in $|\Phi|$. On the other hand, due to the resets, BANDITSRL has an extra logarithmic factor in the effective regret horizon τ_{opt}.

Single HLS representation. A noteworthy consequence of Thm. 4.1 is that any no-regret algorithm equipped with GLRT achieves constant regret when provided with a realizable HLS representation.

Corollary 4.2. *Let $\Phi = \Phi^\star = \{\phi^\star\}$ and $\phi^\star$ is* HLS. *Then, $\tau_{\text{elim}} = \tau_{\text{HLS}} = 0$ and, with probability at least $1-4\delta$,* BANDITSRL *suffers constant regret: $R_T \leq \overline{R}_{\mathfrak{A}}(\tau_{\text{glrt}} \wedge T, \phi^\star, \delta)$.*

This corollary also illustrates that the performance of $\mathfrak{A}$ is not affected when $\phi^\star$ is non-HLS (i.e., $\tau_{\text{glrt}} = \infty$), as BANDITSRL achieves the same regret of the base algorithm. Note that there is no additional logarithmic factor in this case since we do not need any reset for representation learning.

4.2 Additional Results

No HLS representation. A consequence of Thm. 4.1 is that when $|\Phi| > 1$ but no realizable HLS exists ($\tau_{\text{glrt}} = \infty$), BANDITSRL still enjoys a sublinear regret.

Corollary 4.3 (Regret bound without HLS representation)**.** *Consider the same setting in Thm. 4.1 and assume that $\Phi^\star$ does not contain any* HLS *representation. Then, for any $\delta \in (0,1)$ and $T \in \mathbb{N}$, the regret of* BANDITSRL *is bounded, with probability at least $1-4\delta$, as follows:*

$$R_T \leq 2\tau_{\text{elim}} + \max_{\phi\in\Phi^\star} \overline{R}_{\mathfrak{A}}(T,\phi,\delta_{\log_2(T)}/|\Phi|)\log_2(T).$$

This shows that the regret of BANDITSRL is of the same order as the base no-regret algorithm $\mathfrak{A}$ when running with the worst realizable representation. While such worst-case dependency is undesirable, it is common to many representation learning algorithms, both in bandits and reinforcement learning [e.g. 4, 33].[3] In App. C, we show that an alternative representation loss could address this problem and lead to a bound scaling with the regret of the *best* realizable representation ($R_T \leq 2\tau_{\text{elim}} + \min_{\phi\in\Phi^\star} \overline{R}_{\mathfrak{A}}(T,\phi,\delta/|\Phi|)\log_2(T)$), while preserving the guarantees for the HLS case. Since the representation loss requires an upper-bound on the number of suboptimal actions and a carefully tuned schedule for guessing the gap Δ, it is less practical than the smallest eigenvalue, which we use as the basis for our practical version of BANDITSRL.

Algorithm-dependent instances and comparison to LEADER. Table 1 reports the regret bound of BANDITSRL for different base algorithms. These results make explicit the dependence in the number of representations $|\Phi|$ and show that the cost of representation learning is only logarithmic. In the specific case of LINUCB for HLS representations, we highlight that the upper-bound to the time τ_{opt}

[3] Notice that the worst-representation dependency is often hidden in the definition of Φ, which is assumed to contain features with fixed dimension and bounded norm, i.e., $\Phi = \{\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^d, \sup_{x,a}\|\phi(x,a)\|_2 \leq L\}$. As d and B are often the only representation-dependent terms in the regret bound $\overline{R}_{\mathfrak{A}}$, no worst-representation dependency is reported.

Algorithm	$\overline{R}_{\mathfrak{A}}(T, \phi, \delta/\lvert\Phi\rvert)$	τ_{alg}
LINUCB	$d_\phi^2 \log(\lvert\Phi\rvert T/\delta)^2/\Delta$	$\frac{L_{\phi^\star}^2 d^2 \log(\lvert\Phi\rvert/\delta)^2}{\lambda^\star(\phi^\star)\Delta^2}$
ϵ-greedy with $\epsilon_t = t^{-1/3}$	$\sqrt{d_\phi \lvert\mathcal{A}\rvert} \log(\lvert\Phi\rvert/\delta) T^{2/3}$	$\frac{L_{\phi^\star}^6 (d\lvert\mathcal{A}\rvert)^{3/2} L^3 \log(\lvert\Phi\rvert/\delta)^3}{\lambda^\star(\phi^\star)^3\Delta^3}$

Table 1: Specific regret bounds when using LINUCB or ϵ-greedy as base algorithms. We omit numerical constants and logarithmic factors.

in Thm. 4.1 improves over the result of LEADER. While LEADER has no explicit concept of τ_{alg}, a term with the same dependence of τ_{alg} in Tab. 1 appears also in the LEADER analysis. This term encodes an upper bound to the pulls of suboptimal actions and depends on the LINUCB strategy. As a result, the first three terms in Eq. 4 are equivalent to the ones of LEADER. The improvement comes from the last term (τ_{elim}), where, thanks to a refined analysis of the elimination condition, we are able to improve the dependence on the inverse minimum misspecification ($1/\min_{\phi\notin\Phi^\star}\epsilon_\phi$) from quadratic to linear (see App. B for a detailed comparison). On the other hand, BANDITSRL suffers from the worst regret among realizable representations, whereas LEADER scales with the *best* representation. As discussed above, this mismatch can be mitigated by using by a different choice of representation loss. In the case of ϵ-greedy, the $T^{2/3}$ regret upper-bound induces a worse τ_{alg} due to a larger number of suboptimal pulls. This in turns reflects into a higher regret to the constant regime. Finally, LEADER is still guaranteed to achieve constant regret by selecting different representations at different context-action pairs whenever non-HLS representations satisfy a certain mixing condition [cf. 11, Sec. 5.2]. This result is not possible with BANDITSRL, where one representation is selected in each phase. At the same time, it is the single-representation structure of BANDITSRL that allows us to accommodate different base algorithms and scale it to any representation space.

5 Experiments

We provide an empirical validation of BANDITSRL both in synthetic contextual linear bandit problems and in non-linear contextual problems [see e.g., 6, 27].

Linear Benchmarks. We first evaluate BANDITSRL on synthetic linear problems to empirically validate our theoretical findings. In particular, we test BANDITSRL with different base algorithms and representation learning losses and we compare it with LEADER.[4] We consider the "varying dimension" problem introduced in [11] which consists of six realizable representations with dimension from 2 to 6. Of the two representations of dimension $d = 6$, one is HLS. In addition seven misspecified representations are available. Details are provided in App. D. We consider LINUCB and ϵ-greedy as base algorithms and we use the theoretical parameters, but we perform warm start using all the past data when a new representation is selected. Similarly, for BANDITSRL we use the theoretical parameters ($\gamma = 2$) and $\mathcal{L}_t(\phi) := \mathcal{L}_{\mathrm{eig},t}(\phi)$. Fig. 1 shows that, as expected, BANDITSRL with both base algorithms is able to achieve constant regret when a HLS representation exists. As expected from the theoretical analysis, ϵ-greedy leads to a higher regret than LINUCB. Furthermore, empirically BANDITSRL with LINUCB obtains a performance that is comparable with the one of LEADER both with and without realizable HLS representation. Note that when no HLS exists, the regret of BANDITSRL with ϵ-greedy is $T^{2/3}$, while LINUCB-based algorithms are able to achieve $\log(T)$ regret. When Φ contains misspecified representations (Fig. 1(center-left)), we can observe that in the first regime $[1, \tau_{\mathrm{elim}}]$ the algorithm suffers linear regret, after that we have the regime of the base algorithm ($[\tau_{\mathrm{elim}}, \tau_{\mathrm{glrt}} \vee \tau_{\mathrm{HLS}}]$) up to the point where the GLRT leads to select only optimal actions.

Weak HLS. Papini et al. [11] showed that when realizable representations are redundant (i.e., $\lambda^\star(\phi^\star) = 0$), it is still possible to achieve constant regret if the representation is "weakly"-HLS, i.e., the features of the optimal actions span the features $\phi(x, a)$ associated to any context-action pair, but not necessarily $\mathbb{R}^{d_\phi}$. To test this case, we pad a 5-dimensional vector of ones to all the features of the six realizable representations in the previous experiment. To deal with the weak-HLS condition, we introduce the alternative representation loss $\mathcal{L}_{\mathrm{weak},t}(\phi) = -\min_{s\le t}\left\{\phi(x_s, a_s)^\mathsf{T}(V_t(\phi) - \lambda I_{d_\phi})\phi(x_s, a_s)/L_\phi^2\right\}$. Since, $V_t(\phi) - \lambda I_{d_\phi}$ tends to behave as $\mathbb{E}_x[\phi^\star(x)\phi^\star(x)^\mathsf{T}]$, this loss encourages representations where all the observed features are spanned by the optimal arms, thus promoting weak-HLS representations

[4] We do not report the performance of model selection algorithms. An extensive analysis can be found in [11], where the author showed that LEADER was outperforming all the baselines.

(see App. C for more details). As expected, Fig. 1(right) shows that the min-eigenvalue loss $\mathcal{L}_{\text{eig},t}$ fails in identifying the correct representation in this domain. On the other hand, BANDITSRL with the novel loss is able to achieve constant regret and converge to constant regret (we cut the figure for readability), and behaves as LEADER when using LINUCB.

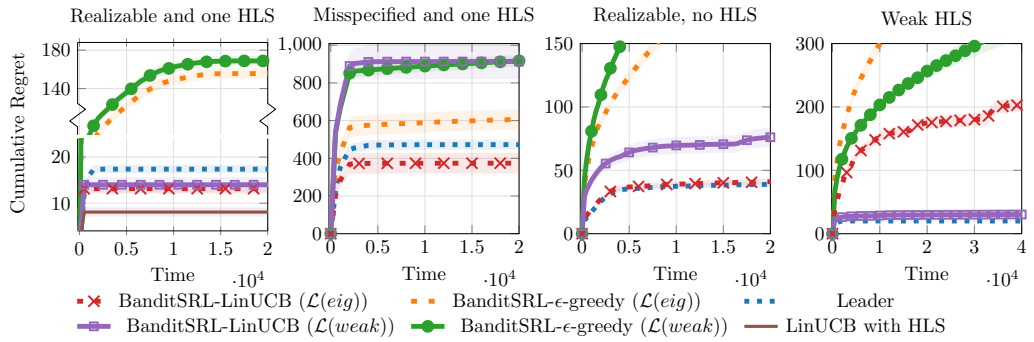

Figure 1: Varying dimension experiment with all realizable representations (left), misspecified representations (center-left), realizable non-HLS representations (center-right) and weak-HLS (right). Experiments are averaged over 40 repetitions.

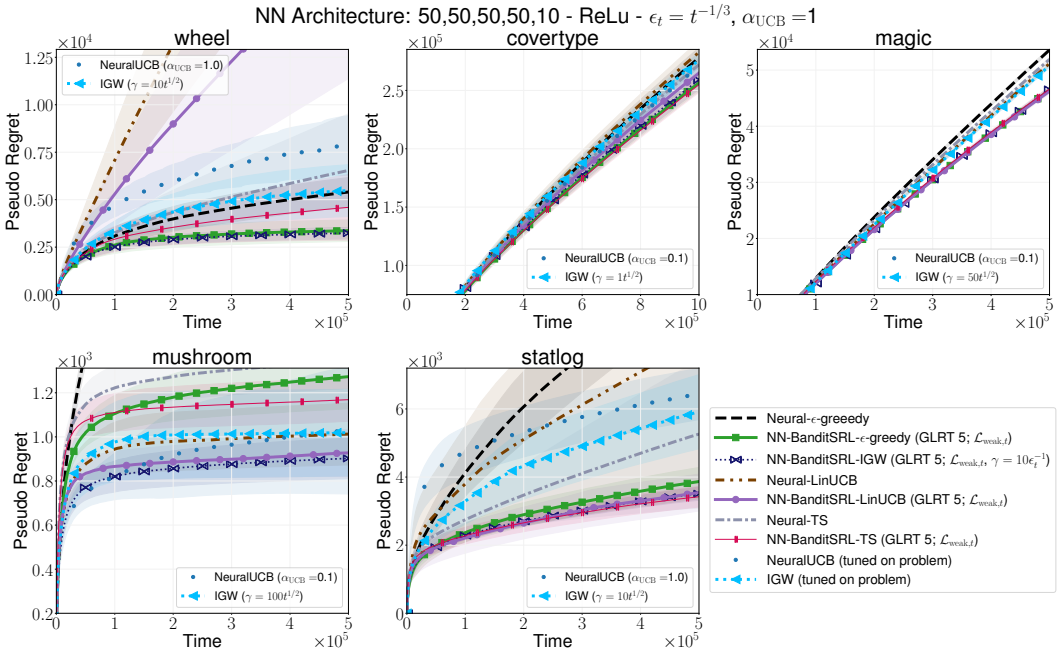

Figure 2: Average cumulative regret (over 20 runs) in non-linear domains.

Non-Linear Benchmarks. We study the performance of NN-BANDITSRL in classical benchmarks where non-linear representations are required. We only consider the weak-HLS loss $\mathcal{L}_{\text{weak},t}(\phi)$ as it is more general than full HLS. As base algorithms we consider ϵ-greedy and inverse gap weighting (IGW) with $\epsilon_t = t^{-1/3}$, and LINUCB and LINTS with theoretical parameters. These algorithms are run on the representation ϕ_j provided by the NN at each phase j. We compare NN-BANDITSRL against the base algorithms using the maximum-likelihood representation (i.e., Neural-(ϵ-greedy, LINTS) [6] and Neural-LINUCB [28]), supervised learning with the IGW strategy [e.g., 7, 10] and NeuralUCB [27][5] See App. C-D for details.

[5]For ease of comparison, all the algorithms use the same phased schema for fitting the reward and recomputing the parameters. NeuralUCB uses a diagonal approximation of the design matrix.

In all the problems[6] the reward function is highly non-linear w.r.t. contexts and actions and we use a network composed by layers of dimension $[50, 50, 50, 50, 10]$ and ReLu activation to learn the representation (i.e., $d = 10$). Fig. 2 shows that all the base algorithms (ϵ-GREEDY, IGW, LINUCB, LINTS) achieve better performance through representation learning, outperforming the base algorithms. This provides evidence that NN-BANDITSRL is effective even beyond the theoretical scenario.

For the baseline algorithms (NEURALUCB, IGW) we report the regret of the best configuration on each individual dataset, while for NN-BANDITSRL we fix the parameters across datasets (i.e., $\alpha_{\mathrm{GLRT}} = 5$). While this comparison clearly favours the baselines, it also shows that NN-BANDITSRL is a robust algorithm that behaves better or on par with the state-of-the-art algorithms. In particular, NN-BANDITSRL uses theoretical parameters while the baselines use tuned configurations. Optimizing the parameters of NN-BANDITSRL is outside the scope of these experiments.

6 Conclusion

We proposed a novel algorithm, BANDITSRL, for representation selection in stochastic contextual linear bandits. BANDITSRL combines a mechanism for representation learning that aims to recover representations with good spectral properties, with a generalized likelihood ratio test to exploit the recovered representation. We proved that, thanks to these mechanisms, BANDITSRL is not only able to achieve sublinear regret with any no-regret algorithm $\mathfrak{A}$ but, when a HLS representation exists, it is able to achieve constant regret. We demonstrated that BANDITSRL can be implemented using NNs and showed its effectiveness in standard benchmarks.

A direction for future investigation is to extend the approach to a weaker misspecification assumption than Asm. 3. Another direction is to leverage the technical and algorithmic tools introduced in this paper for representation learning in reinforcement learning, e.g., in low-rank problems [e.g. 38].

Acknowledgments and Disclosure of Funding

M. Papini was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180).

References

[1] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *CoRR*, abs/1904.10040, 2019.

[2] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[3] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Process. Mag.*, 39(3):42–62, 2022.

[4] Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *AISTATS*, volume 22 of *JMLR Proceedings*, pages 19–26. JMLR.org, 2012.

[5] Alekh Agarwal, Daniel J. Hsu, Satyen Kale, John Langford, Lihong Li, and Robert E. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1638–1646. JMLR.org, 2014.

[6] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *ICLR (Poster)*. OpenReview.net, 2018.

[6] The dataset-based problems –statlog, magic, covertype, mushroom [34–37]– are obtained from the standard multiclass-to-bandit conversion [6, 27]. See appendix D for details.

[7] Dylan J. Foster and Alexander Rakhlin. Beyond UCB: optimal and efficient contextual bandits with regression oracles. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3199–3210. PMLR, 2020.

[8] Dylan J. Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. In *NeurIPS*, pages 14714–14725, 2019.

[9] Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5662–5670. PMLR, 2020.

[10] David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *CoRR*, abs/2003.12699, 2020.

[11] Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8371–8380. PMLR, 2021.

[12] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 208–214. JMLR.org, 2011.

[13] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320, 2011.

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

[15] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[16] Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

[17] Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. In *NeurIPS*, 2021.

[18] Weitong Zhang, Jiafan He, Dongruo Zhou, Amy Zhang, and Quanquan Gu. Provably efficient representation learning in low-rank markov decision processes. *CoRR*, abs/2106.11935, 2021.

[19] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[20] Odalric-Ambrym Maillard and Rémi Munos. Adaptive bandits: Towards the best history-dependent strategy. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 570–578. JMLR.org, 2011.

[21] Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pages 12–38. PMLR, 2017.

[22] Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and RL model selection. *CoRR*, abs/2006.05491, 2020.

[23] Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvári. Model selection in contextual stochastic bandit problems. In *NeurIPS*, 2020.

[24] Jonathan N. Lee, Aldo Pacchiano, Vidya Muthukumar, Weihao Kong, and Emma Brunskill. Online model selection for reinforcement learning with function approximation. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 3340–3348. PMLR, 2021.

[25] Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and RL. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 2276–2285. PMLR, 2021.

[26] Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Problem-complexity adaptive model selection for stochastic linear bandits. In *AISTATS*, volume 130 of *Proceedings of Machine Learning Research*, pages 1396–1404. PMLR, 2021.

[27] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 11492–11502. PMLR, 2020.

[28] Pan Xu, Zheng Wen, Handong Zhao, and Quanquan Gu. Neural contextual bandits with deep representation and shallow exploration. *CoRR*, abs/2012.01780, 2020.

[29] Aniket Anand Deshmukh, Abhimanu Kumar, Levi Boyles, Denis Charles, Eren Manavoglu, and Ürün Dogan. Self-supervised contextual bandits in computer vision. *CoRR*, abs/2003.08485, 2020.

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

[31] Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

[32] Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2013.

[33] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. *CoRR*, abs/2202.00063, 2022.

[34] Jock A. Blackard. *Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types*. PhD thesis, USA, 1998. AAI9921979.

[35] R.K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, A. Vaiciulis, and W. Wittek. Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2):511–528, 2004.

[36] Jeffrey Curtis Schlimmer. *Concept acquisition through representational adjustment*. University of California, Irvine, 1987.

[37] Dheeru Dua and Casey Graff. UCI machine learning repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

[38] Alekh Agarwal, Sham M. Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: structural complexity and representation learning of low rank mdps. In *NeurIPS*, 2020.

[39] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NIPS*, pages 1177–1184. Curran Associates, Inc., 2007.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Notation **15**

B Analysis of BANDITSRL **15**
B.1 Assumptions . . . 15
B.2 Controlling the MSE . . . 15
B.3 Decomposition into phases . . . 17
B.4 Good events . . . 18
B.5 Generalized Likelihood Ratio Test . . . 18
B.6 Eliminating misspecified representations . . . 19
B.7 Regret bound without HLS representations . . . 20
B.8 Regret bound with HLS representations . . . 21
B.9 Finding explicit bounds . . . 23
B.10 Proof of the main theorems . . . 24

C Variants of BANDITSRL **25**
C.1 BANDITSRL: alternative losses . . . 25
C.2 NN-BANDITSRL: representation learning through neural networks . . . 28

D Experiments **30**
D.1 Linear Benchmarks . . . 30
D.2 Non-Linear Benchmarks . . . 31

E Examples of No-regret Algorithms **39**
E.1 LinUCB . . . 39
E.2 ϵ-greedy . . . 40

F Auxiliary Results **43**
F.1 Bounding the eigenvalues of the design matrices . . . 43
F.2 Martingale concentration . . . 43

A Notation

Symbol	Meaning
$\mathcal{X}$	Set of contexts
$\mathcal{A}$	Finite set of arms
ρ	Context distribution
$\mu : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$	Mean-reward function
Φ	Set of representations
$\Phi^\star$	Subset of realizable representations
$\pi : \mathcal{X} \to \mathcal{A}$	A policy
$\mathcal{F}_t$	σ-algebra generated by $(x_1, a_1, y_1, \dots, x_t, a_t, y_t)$
$\mathfrak{A}_t : \mathcal{X} \to \mathcal{A}$	Bandit algorithm (measurable mappings w.r.t. $\mathcal{F}_{t-1}$)
$V_t(\phi) := \sum_{k=1}^t \phi(x_k, a_k)\phi(x_k, a_k)^\mathsf{T} + \lambda I_{d_\phi}$	Design matrix for representation ϕ
$\theta_{\phi,t} = V_t(\phi)^{-1} \sum_{k=1}^t \phi(x_k, a_k) r_k$	Regularized least-square estimate for representation ϕ
$\pi_t^\star(x; \phi) := \mathrm{argmax}_{a \in \mathcal{A}} \phi(x, a)^\mathsf{T} \theta_{\phi,t}$	Empirical optimal arm for context x and representation ϕ
$\Delta(x, a) = \max_{a' \in \mathcal{A}} \mu(x, a') - \mu(x, a)$	Sub-optimality gap of arm a in context x
$a_x^\star$	Optimal arm for context x
$\pi^\star(x) = \mathrm{argmax}_{a \in \mathcal{A}} \mu(x, a)$	Optimal policy
$\lambda^\star(\phi) := \mathbb{E}_{x \sim \rho}[\phi(x, \pi^\star(x))\phi(x, \pi^\star(x))^\mathsf{T}]$	Minimum eigenvalue on optimal arms
$E_t(\phi, \theta) := \frac{1}{t} \sum_{k=1}^t \left(\phi(x_k, a_k)^\mathsf{T}\theta - y_k\right)^2$	Mean square error of model (ϕ, θ) at time t
$\mathbb{E}_t$ and $\mathbb{V}_t$	Expectation and variance conditioned on $\mathcal{F}_{t-1}$
$P_t(\phi, \theta) := \sum_{k=1}^t \mathbb{E}_k \left[\left(\phi(x_k, a_k)^\mathsf{T}\theta - \mu(x_k, a_k)\right)^2\right]$	Sum of mean prediction errors of model (ϕ, θ)
$\alpha_{t,\delta}(\phi) := \frac{40}{t} \log \frac{8\lvert\Phi\rvert^2 (12 L_\phi B_\phi t)^{d_\phi} t^3}{\delta} + \frac{2}{t}$	Threshold for MSE elimination
$D_t(\phi) := 160 d_\phi \log(12 L_\phi B_\phi t)$	Dimension factor for representation ϕ
$R_T := \sum_{t=1}^T \Delta(x_t, a_t)$	Pseudo-regret
$t_j := 2^j$	Time at which the $(j+1)$-th phase ends (with $t_0 := 0$)
$N_j(T) := \sum_{t=t_j+1}^T \mathbb{1}\{G_t\}$	Number of calls to $\mathfrak{A}$ in phase j up to time $T \le t_{j+1}$
$G_t := \{\mathrm{GLR}_{t-1}(\mathrm{x}_t; \phi_{t-1}) \le \beta_{t-1,\delta/\lvert\Phi\rvert}(\phi_{t-1})\}$	Event under which the GLRT does not trigger at time t
$S_T := \sum_{t=1}^T \mathbb{1}\{a_t \neq \pi^\star(x_t)\}$	Total number of sub-optimal pulls at time T
$\overline{R}_{\mathfrak{A}}(T, \phi, \delta)$	Regret bound of algorithm $\mathfrak{A}$ over T steps when using ϕ
$g_T(\Phi, \Delta, \delta)$	Bound on the sub-optimal pulls of $\mathfrak{A}$ (see Th. B.10)
$\delta_j := \delta/(2(j+1)^2)$	Confidence level for the base algorithm

Table 2: The notation adopted in this paper.

B Analysis of BANDITSRL

B.1 Assumptions

The analysis works under the assumptions stated in Section 2 and for any no-regret base algorithm $\mathfrak{A}$. Here we formally state the conditions required on the

Assumption 4 (No-regret algorithm)**.** *For any $\phi \in \Phi^\star$ and $\delta \in (0, 1)$, if we run algorithm $\mathfrak{A}$ with representation ϕ and confidence δ, with probability at least $1 - \delta$ we have, for any $T \in \mathbb{N}$,*

$$\sum_{t=1}^T \Delta(x_t, \mathfrak{A}_t(x_t; \phi, \delta)) \le \overline{R}_{\mathfrak{A}}(T, \phi, \delta),$$

where $\mathfrak{A}_t(x; \phi, \delta)$ denotes the policy played by $\mathfrak{A}$ at time t when instantiated with representation ϕ and confidence δ, while the function $\overline{R}_{\mathfrak{A}}(T, \phi, \delta)$ is sub-linear and non-decreasing in T and logarithmic and non-decreasing in $1/\delta$.

B.2 Controlling the MSE

The following is an extension of Lemma 4.1 in [4] and Lemma 20 in [11]. Differently from their results, which relate the empirical MSE of any model (ϕ, θ) with that of a realizable model, we also include the sum of conditional mean prediction errors $P_t(\phi, \theta) :=$

$\sum_{k=1}^{t} \mathbb{E}_k \left[\left(\phi(x_k, a_k)^\mathsf{T} \theta - \mu(x_k, a_k) \right)^2 \right]$, which roughly quantifies the misspecification of model (ϕ, θ). This shall be crucial for improving the elimination times of misspecified representations later.

Lemma B.1. *Let $\phi \in \Phi, \theta \in \mathbb{R}^{d_\phi}$. Take any realizable representation $\phi^\star \in \Phi^\star$ and let $\theta^\star := \theta^\star_{\phi^\star}$. Then, for each $t \geq 1$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(E_t(\phi^\star, \theta^\star) > E_t(\phi, \theta) + \frac{40}{t} \log \frac{4t}{\delta} - \frac{P_t(\phi, \theta)}{2t} \right) \leq \delta. \tag{5}$$

Proof. Define $Z_k := (\phi(x_k, a_k)^T \theta - y_k)^2 - (\phi^\star(x_k, a_k)^T \theta^\star - y_k)^2$. Note that, since $|\phi(x_k, a_k)^T \theta| \leq 1$, $|\phi^\star(x_k, a_k)^T \theta^\star| \leq 1$, and $|y_k| \leq 1$, we have $|Z_k| \leq 4$. Thus, $(\mathbb{E}_k[Z_k] - Z_k)_{k \geq 1}$ is a martingale difference sequence bounded by 8 in absolute value. Then, using Freedman's inequality (Lemma F.3), with probability at least $1 - \delta$, for any t,

$$\sum_{k=1}^{t} \mathbb{E}_k[Z_k] - \sum_{k=1}^{t} Z_k \leq 2 \sqrt{\sum_{k=1}^{t} \mathbb{V}_k[Z_k] \log \frac{4t}{\delta}} + 32 \log \frac{4t}{\delta}.$$

Using Lemma 4.2 in [4], we have that $\mathbb{V}_k[Z_k] \leq 4\mathbb{E}_k[Z_k]$. Solving the resulting inequality in $\sum_{k=1}^{t} \mathbb{E}_k[Z_k]$ and using $(x+y)^2 \leq 2x^2 + 2y^2$,

$$\sum_{k=1}^{t} \mathbb{E}_k[Z_k] \leq \left(2\sqrt{\log \frac{4t}{\delta}} + \sqrt{36 \log \frac{4t}{\delta} + \sum_{k=1}^{t} Z_k} \right)^2 \leq 80 \log \frac{4t}{\delta} + 2 \sum_{k=1}^{t} Z_k.$$

The proof is concluded by using $\sum_{k=1}^{t} Z_k = t(E_t(\phi, \theta) - E_t(\phi^\star, \theta^\star))$ and $\sum_{k=1}^{t} \mathbb{E}_k[Z_k] = P_t(\phi, \theta)$. □

Lemma B.2. *For each $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \geq 1, \phi \in \Phi, \phi^\star \in \Phi^\star, \theta \in \mathcal{B}_\phi : E_t(\phi^\star, \theta^\star_{\phi^\star}) > E_t(\phi, \theta) - \frac{P_t(\phi, \theta)}{4t} + \alpha_{t,\delta}(\phi) \right) \leq \delta.$$

Proof. We shall use a covering argument for each representation $\phi \in \Phi$. First note that, for any $\xi > 0$, there always exists a finite set $\mathcal{C}_\phi \subset \mathbb{R}^{d_\phi}$ of size at most $(3B_\phi/\xi)^{d_\phi}$ such that, for each $\theta \in \mathcal{B}_\phi$, there exists $\theta' \in \mathcal{C}_\phi$ with $\|\theta - \theta'\|_2 \leq \xi$ (see e.g. Lemma 20.1 in [15]). Moreover, suppose that all vectors in $\mathcal{C}_\phi$ have ℓ_2-norm bounded by B_ϕ (otherwise we can always remove vectors with large norm). Now take any two vectors $\theta, \theta' \in \mathcal{B}_\phi$ with $\|\theta - \theta'\|_2 \leq \xi$. We have

$$\begin{aligned}
E_t(\phi, \theta) &= \frac{1}{t} \sum_{k=1}^{t} \left(\phi(x_k, a_k)^\mathsf{T} \theta \pm \phi(x_k, a_k)^\mathsf{T} \theta' - y_k \right)^2 \\
&= \frac{1}{t} \sum_{k=1}^{t} \left(\phi(x_k, a_k)^\mathsf{T} (\theta - \theta') \right)^2 + \frac{1}{t} \sum_{k=1}^{t} \left(\phi(x_k, a_k)^\mathsf{T} \theta' - y_k \right)^2 \\
&\qquad + \frac{2}{t} \sum_{k=1}^{t} \left(\phi(x_k, a_k)^\mathsf{T} (\theta - \theta') \right) \left(\phi(x_k, a_k)^\mathsf{T} \theta' - y_k \right) \\
&\geq E_t(\phi, \theta') + \frac{2}{t} \sum_{k=1}^{t} \left(\phi(x_k, a_k)^\mathsf{T} (\theta - \theta') \right) \underbrace{\left(\phi(x_k, a_k)^\mathsf{T} \theta' - y_k \right)}_{|\cdot| \leq 2} \\
&\geq E_t(\phi, \theta') - \frac{4}{t} \sum_{k=1}^{t} \|\phi(x_k, a_k)\|_2 \|\theta - \theta'\|_2 \geq E_t(\phi, \theta') - 4L_\phi \xi.
\end{aligned}$$

Similarly, one can prove that

$$
\begin{aligned}
P_t(\phi,\theta) &= \sum_{k=1}^{t} \mathbb{E}_k\left[\left(\phi(x_k,a_k)^\mathsf{T}\theta - \mu(x_k,a_k)\right)^2\right] \\
&\le 2\sum_{k=1}^{t} \mathbb{E}_k\left[\left(\phi(x_k,a_k)^\mathsf{T}\theta - \phi(x_k,a_k)^\mathsf{T}\theta'\right)^2\right] + 2\sum_{k=1}^{t} \mathbb{E}_k\left[\left(\phi(x_k,a_k)^T\theta' - \mu(x_k,a_k)\right)^2\right] \\
&\le 2P_t(\phi,\theta') + 2\sum_{k=1}^{t} \mathbb{E}_k\left[\|\phi(x_k,a_k)\|_2^2\right]\|\theta-\theta'\|_2^2 \le 2P_t(\phi,\theta') + 2L_\phi^2\xi^2 t.
\end{aligned}
$$

Let us define a sequence of deterministic covers $(\mathcal{C}_{\phi,t})_{t\ge1}$ such that $\mathcal{C}_{\phi,t}$ is a ξ_t-cover with $\xi_t = \frac{1}{4L_\phi t}$. Let $\delta'_t = \frac{\delta}{2|\Phi|^2(12L_\phi S_\phi t)^{d_\phi}}$ and note that $\alpha_{t,\delta}(\phi) := \frac{40}{t}\log\frac{4t^3}{\delta'_t} + \frac{2}{t}$. Then,

$$
\begin{aligned}
&\mathbb{P}\left(\exists t\ge1, \phi\in\Phi, \phi^\star\in\Phi^\star, \theta\in\mathcal{B}_\phi : E_t(\phi^\star,\theta^\star_{\phi^\star}) > E_t(\phi,\theta) - \frac{P_t(\phi,\theta)}{4t} + \frac{40}{t}\log\frac{4t^3}{\delta'_t} + \frac{2}{t}\right) \\
&\le \sum_{t=1}^{\infty}\sum_{\phi\in\Phi}\sum_{\phi^\star\in\Phi^\star} \mathbb{P}\left(\exists\theta\in\mathcal{B}_\phi : E_t(\phi^\star,\theta^\star_{\phi^\star}) > E_t(\phi,\theta) - \frac{P_t(\phi,\theta)}{4t} + \frac{40}{t}\log\frac{4t^3}{\delta'_t} + \frac{2}{t}\right) \\
&\le \sum_{t=1}^{\infty}\sum_{\phi\in\Phi}\sum_{\phi^\star\in\Phi^\star} \mathbb{P}\left(\exists\theta'\in\mathcal{C}_\phi : E_t(\phi^\star,\theta^\star_{\phi^\star}) > E_t(\phi,\theta') - \frac{1}{t} - \frac{2P_t(\phi,\theta')+1/(8t)}{4t} + \frac{40}{t}\log\frac{4t^3}{\delta'_t} + \frac{2}{t}\right) \\
&\le \sum_{t=1}^{\infty}\sum_{\phi\in\Phi}\sum_{\phi^\star\in\Phi^\star}\sum_{\theta'\in\mathcal{C}_\phi} \mathbb{P}\left(E_t(\phi^\star,\theta^\star_{\phi^\star}) > E_t(\phi,\theta') - \frac{P_t(\phi,\theta')}{2t} + \frac{40}{t}\log\frac{4t^3}{\delta'_t}\right) \\
&\le \sum_{t=1}^{\infty}\sum_{\phi\in\Phi}\sum_{\phi^\star\in\Phi^\star}\sum_{\theta'\in\mathcal{C}_\phi} \frac{\delta'_t}{t^2} \le |\Phi|^2\sum_{t=1}^{\infty}\frac{\delta'_t}{t^2}(12L_\phi B_\phi t)^{d_\phi} \le \delta.
\end{aligned}
$$

Here the first inequality is from the union bound, the second one follows by relating θ with its closest vector in the cover as above, the third one is from another union bound, the fourth one uses Lemma B.1, the fifth one is from the maximum size of the cover, and the last one uses the definition of δ'_t. □

Corollary B.3. *For each $\delta\in(0,1)$,*

$$
\mathbb{P}\left(\exists t\ge1, \phi\in\Phi, \phi^\star\in\Phi^\star, \theta\in\mathcal{B}_\phi : E_t(\phi^\star,\theta^\star_{\phi^\star}) > E_t(\phi,\theta) + \alpha_{t,\delta}(\phi)\right) \le \delta.
$$

Proof. This is trivial from Lemma B.2 since $P_t(\phi,\theta) > 0$. □

B.3 Decomposition into phases

For $j\ge1$, let $t_j = 2^j$ be the time at which the $(j+1)$-th phase ends (i.e., when the algorithm selects a new representation for the $(j+1)$-th time). Let $t_0 = 0$. Note that, on the interval $[t_j+1, t_{j+1}]$ the algorithm uses a fixed representation ϕ_j selected at time t_j. In the remaining, we shall overload the notation used in the main paper and denote all quantities with a time subscript. Therefore, for $t\in[t_j+1,t_{j+1}]$, $\phi_{t-1} = \phi_{t_j}$ denotes the representation used a time t, i.e., ϕ_j.

Recall that G_t denotes the event under which the GLRT does not trigger at round t (i.e., the base algorithm is called). Then, for each $j\ge0$, the quantity

$$
\sum_{t=t_j+1}^{t_{j+1}} \mathbb{1}\{G_t\}\,\Delta(x_t,a_t)
$$

denotes the regret suffered by the base algorithm in phase j.

B.4 Good events

We define the following events

$$\mathcal{E}_1 = \left\{\forall t \in \mathbb{N}, \phi \in \Phi^\star : \|\theta_{\phi,t} - \theta^\star_\phi\|_{V_t(\phi)} \le \beta_{t,\delta/|\Phi|}(\phi)\right\},$$

$$\mathcal{E}_2 = \Big\{\forall t \in \mathbb{N}, \phi \in \Phi : V_t(\phi) \succeq t\,\mathbb{E}_{x\sim\rho}[\phi(x,\pi^\star(x))\phi(x,\pi^\star(x))^\mathsf{T}] + \left(\lambda - L_\phi^2 S_t - 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}\right) I_{d_\phi}\Big\},$$

$$\mathcal{E}_3 = \Big\{\forall t \in \mathbb{N}, \phi \in \Phi : V_t(\phi) \preceq t\,\mathbb{E}_{x\sim\rho}[\phi(x,\pi^\star(x))\phi(x,\pi^\star(x))^\mathsf{T}] + \left(\lambda + L_\phi^2 S_t + 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}\right) I_{d_\phi}\Big\},$$

$$\mathcal{E}_4 = \left\{\forall t \in \mathbb{N}, \phi \in \Phi, \phi^\star \in \Phi^\star, \theta \in \mathcal{B}_\phi : E_t(\phi^\star, \theta^\star_{\phi^\star}) \le E_t(\phi,\theta) - \frac{P_t(\phi,\theta)}{4t} + \alpha_{t,\delta}(\phi)\right\},$$

$$\mathcal{E}_5 = \left\{\forall j \in \mathbb{N}, T \le t_{j+1} : \sum_{t=t_j+1}^{T} \mathbb{1}\{G_t\}\,\Delta(x_t,a_t) \le \overline{R}_{\mathfrak{A}}\big(N_j(T), \phi_{t_j}, \delta_j/|\Phi|\big)\right\},$$

We define the good event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5$.

Lemma B.4 (Good event)**.** *We have* $\mathbb{P}(\mathcal{E}) \ge 1 - 4\delta$.

Proof. By using Theorem 2 in [13] together with a union bound over Φ, $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta$. Similarly, by Lemma F.1, $\mathbb{P}(E_2 \cap E_3) \ge 1-\delta$. Event $\mathcal{E}_4$ holds with probability at least $1-\delta$ by Lemma B.2.

We finally bound the probability of $\mathcal{E}_5$ failing. We have

$$\mathbb{P}(\neg\mathcal{E}_5) \le \sum_{j\in\mathbb{N}} \mathbb{P}\left\{\exists T \le t_{j+1} : \sum_{t=t_j+1}^{T} \mathbb{1}\{G_t\}\,\Delta(x_t,a_t) > \overline{R}_{\mathfrak{A}}\big(T, \phi_{t_j}, \delta_j/|\Phi|\big)\right\} \le \sum_{j\in\mathbb{N}} \delta_j \le \delta,$$

where the first inequality is from a union bound over j, the second holds from the anytime no-regret assumption (Assumption 4) together with a union bound over Φ, while the last one holds by definition of δ_j. A union bound over the 5 events proves the statement. □

Lemma B.5. *[Correctness of MSE elimination] Under event $\mathcal{E}$, for each $t \ge 1$ any realizable representation $\phi^\star \in \Phi^\star$ satisfies the constraint, i.e., $\phi^\star \in \Phi_t$.*

Proof. Under $\mathcal{E}_4$,

$$\min_{\theta\in\mathcal{B}_{\phi^\star}} E_t(\phi^\star,\theta) \le E_t(\phi^\star,\theta^\star_{\phi^\star}) \le \min_{\phi\in\Phi}\min_{\theta\in\mathcal{B}_\phi}\left(E_t(\phi,\theta) + \alpha_{t,\delta}(\phi)\right).$$

This implies the statement. □

B.5 Generalized Likelihood Ratio Test

For any $\phi \in \Phi$ and $x \in \mathcal{X}$, let us define the *generalized likelihood ratio* as

$$\mathrm{GLR}_t(x;\phi) := \min_{a\neq\pi^\star_t(x;\phi)} \frac{\big(\phi(x,\pi^\star_t(x;\phi)) - \phi(x,a)\big)^\mathsf{T}\theta_{\phi,t}}{\|\phi(x,\pi^\star_t(x;\phi)) - \phi(s,a)\|_{V_t(\phi)^{-1}}}.$$

It is known [e.g., 16, 30] that

$$\mathrm{GLR}_t(x;\phi) = \inf_{\theta\in\Lambda_t(x;\phi)} \|\theta_{\phi,t} - \theta\|_{V_t(\phi)},$$

where $\Lambda_t(x;\phi) := \{\theta \in \mathbb{R}^{d_\phi} \mid \exists a \neq \pi^\star_t(x;\phi) : \phi(x,a)^\mathsf{T}\theta > \phi(x,\pi^\star_t(x;\phi))^\mathsf{T}\theta\}$ is the set of parameters for which the optimal arm in context x is different from the one of $\theta_{\phi,t}$. In turns, the squared objective above is equivalent to

$$\frac{1}{2}\|\theta_{\phi,t} - \theta\|^2_{V_t(\phi)} = \frac{1}{2}\sum_{k=1}^{t}\left(\phi(x_k,a_k)^\mathsf{T}\theta_{\phi,t} - \phi(x_k,a_k)^\mathsf{T}\theta\right)^2,$$

which is equal to the expected (under the conditional reward distribution) log-likelihood ratio between the observations in the bandit model given by $(\phi, \theta_{\phi,t})$ and the one given by (ϕ, θ) if these were Gaussians with unit variance. This is the reason why $\mathrm{GLR}_t(x;\phi)$ is called the generalized likelihood ratio between the bandit model $(\phi, \theta_{\phi,t})$ and *any* other bandit model with a different optimal arm in context x. The generalized likelihood ratio test (GLRT) consists in checking whether

$$\mathrm{GLR_t(x;\phi)} > \beta_{t,\delta}(\phi).$$

When this happens, we have enough confidence to conclude that $\theta^\star_\phi \notin \Lambda_t(x;\phi)$, i.e., that $\pi^\star(x) = \pi^\star_t(x;\phi)$.

BANDITSRL computes, at each step, the GLRT for the currently selected representation. We can easily prove that the test is *correct* under the good event $\mathcal{E}$ if the selected representation is realizable.

Lemma B.6 (Correctness of GLRT)**.** *Under the good event $\mathcal{E}$, for any time t, if $\mathrm{GLR_{t-1}}(x_t;\phi_{t-1}) > \beta_{t-1,\delta/|\Phi|}(\phi_{t-1})$ and $\phi_{t-1} \in \Phi^\star$, then $\pi^\star(x_t) = \pi^\star_t(x_t;\phi_t)$.*

Proof. By contradiction, suppose that the statement does not hold. This means that there exists a time t, realizable feature $\phi \in \Phi^\star$, and context x such that $\pi^\star(x) \neq \pi^\star_t(x;\phi)$ while the test triggers for context x and feature ϕ. By definition, this implies that $\theta^\star_\phi \in \Lambda_t(x;\phi)$ since $\pi^\star$ is the greedy policy for the (realizable) model $(\phi, \theta^\star_\phi)$. Thus,

$$\beta_{t,\delta/|\Phi|}(\phi) < \mathrm{GLR}_t(x;\phi) = \inf_{\theta \in \Lambda_t(x;\phi)} \|\theta_{\phi,t} - \theta\|_{V_t(\phi)} \leq \|\theta_{\phi,t} - \theta^\star_\phi\|_{V_t(\phi)} \leq \beta_{t,\delta/|\Phi|}(\phi),$$

where the last inequality is from event $\mathcal{E}_1$. This is clearly a contradiction. □

B.6 Eliminating misspecified representations

Lemma B.7. *Let $\phi \in \Phi$ be any misspecified representation (i.e., $\phi \notin \Phi^\star$). Under event $\mathcal{E}$, if $\phi \in \Phi_t$ for some t, then*

$$\min_{\theta \in \mathcal{B}_\phi} P_t(\phi,\theta) \leq D_t(\phi) + \min_{\phi^\star \in \Phi^\star} D_t(\phi^\star) + 328 \log \frac{8|\Phi|^2 t^3}{\delta},$$

where $D_t(\phi) := 160 d_\phi \log(12 L_\phi B_\phi t)$.

Proof. Recall that, from Lemma B.5, under $\mathcal{E}$, any $\phi^\star \in \Phi^\star$ is always in Φ_t. Take any arbitrary $\phi^\star \in \Phi^\star$ and let $\theta^\star := \theta^\star_{\phi^\star}$. Then, by definition of Φ_t,

$$\min_{\theta \in \mathcal{B}_\phi} E_t(\phi,\theta) \leq \min_{\phi' \in \Phi} \min_{\theta \in \mathcal{B}_{\phi'}} \left\{ E_t(\phi',\theta) + \alpha_{t,\delta}(\phi') \right\} \leq E_t(\phi^\star,\theta^\star) + \alpha_{t,\delta}(\phi^\star).$$

Similarly, under $\mathcal{E}_4$ we have that

$$E_t(\phi^\star,\theta^\star) \leq \min_{\theta \in \mathcal{B}_\phi} \left(E_t(\phi,\theta) - \frac{P_t(\phi,\theta)}{4t} \right) + \alpha_{t,\delta}(\phi) \leq \min_{\theta \in \mathcal{B}_\phi} E_t(\phi,\theta) - \frac{\min_{\theta \in \mathcal{B}_\phi} P_t(\phi,\theta)}{4t} + \alpha_{t,\delta}(\phi).$$

Combining these two inequalities, we find that

$$\frac{\min_{\theta \in \mathcal{B}_\phi} P_t(\phi,\theta)}{4t} \leq \alpha_{t,\delta}(\phi) + \alpha_{t,\delta}(\phi^\star).$$

Expanding the definition of α, rearranging, and optimizing over $\phi^\star$,

$$\min_{\theta \in \mathcal{B}_\phi} P_t(\phi,\theta) \leq D_t(\phi) + \min_{\phi^\star \in \Phi^\star} D_t(\phi^\star) + 320 \log \frac{8|\Phi|^2 t^3}{\delta} + 16.$$

The proof is concluded by noting that $\log \frac{8|\Phi|^2 t^3}{\delta} \geq 2$ and, thus, $16 \leq 8 \log \frac{8|\Phi|^2 t^3}{\delta}$. □

Lemma B.8 (Elimination)**.** *Under event $\mathcal{E}$, we have $\Phi_t = \Phi^\star$ for all $t \geq \tau_{\mathrm{elim}}$, where*

$$\tau_{\mathrm{elim}} := \min_{t \in \mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > \max_{\phi \notin \Phi^\star} \frac{1}{\epsilon_\phi} \left(D_t(\phi) + \min_{\phi^\star \in \Phi^\star} D_t(\phi^\star) + 328 \log \frac{8|\Phi|^2 t^3}{\delta} \right) \right\}.$$

Let $\tau_{\mathrm{elim}} = 0$ when $\Phi = \Phi^\star$.

Proof. Let π_k be the policy played by the algorithm at round k. First note that,

$$\begin{aligned}
\min_{\theta\in\mathcal{B}_\phi} P_t(\phi,\theta) &= \min_{\theta\in\mathcal{B}_\phi} \sum_{k=1}^{t} \mathbb{E}_k\left[\left(\phi(x_k,a_k)^T\theta - \mu(x_k,a_k)\right)^2\right] \\
&= \min_{\theta\in\mathcal{B}_\phi} \sum_{k=1}^{t} \mathbb{E}_{x\sim\rho}\left[\left(\phi(x,\pi_k(x))^T\theta - \mu(x,\pi_k(x))\right)^2\right] \\
&\geq \min_{\theta\in\mathcal{B}_\phi} \sum_{k=1}^{t} \min_{\pi} \mathbb{E}_{x\sim\rho}\left[\left(\phi(x,\pi(x))^T\theta - \mu(x,\pi(x))\right)^2\right] \\
&= t \min_{\theta\in\mathcal{B}_\phi} \min_{\pi} \mathbb{E}_{x\sim\rho}\left[\left(\phi(x,\pi(x))^T\theta - \mu(x,\pi(x))\right)^2\right] \geq t\epsilon_\phi.
\end{aligned}$$

Then, under $\mathcal{E}$, from Lemma B.7, if $\phi \in \Phi_t$ and $\phi \notin \Phi^\star$,

$$t \leq \frac{1}{\epsilon_\phi}\left(D_t(\phi) + \min_{\phi^\star\in\Phi^\star} D_t(\phi^\star) + 200\log\frac{8|\Phi|^2t^3}{\delta}\right).$$

The result follows by finding the first time t at which a representation update is performed (i.e., $t = 2^j$ for some j) and the condition above is violated for all $\phi \notin \Phi^\star$. □

B.7 Regret bound without HLS representations

We first prove a general regret bound that holds for any realizable problem (in the sense of Assumption 1) without requiring the presence of HLS representations.

Theorem B.9. *Under event $\mathcal{E}$ (i.e., with probability at least $1-4\delta$), for any $T\in\mathbb{N}$, the regret of Algorithm 1 with $\gamma = 2$ and arbitrary loss $\mathcal{L}_t(\phi)$ can be bounded as*

$$R_T \leq 2\tau_{\text{elim}} + \max_{\phi\in\Phi^\star} \overline{R}_{\mathfrak{A}}(T-\tau_{\text{elim}},\phi,\delta_{\log_2(T)}/|\Phi|)\log_2(T),$$

where τ_{elim} is defined in Lemma B.8

Proof. Let $\bar{j}$ be such that $\tau_{\text{elim}} = 2^{\bar{j}}$ (which exists by definition). Using the decomposition into phases of Appendix B.3,

$$\begin{aligned}
R_T := \sum_{t=1}^{T}\Delta(x_t,a_t) &= \sum_{j=0}^{\bar{j}-1}\sum_{t=t_j+1}^{t_{j+1}\wedge T}\Delta(x_t,a_t) + \sum_{j=\bar{j}}^{\lfloor\log_2(T)\rfloor}\sum_{t=t_j+1}^{t_{j+1}\wedge T}\Delta(x_t,a_t) \\
&\leq 2\tau_{\text{elim}} + \sum_{j=\bar{j}}^{\lfloor\log_2(T)\rfloor}\sum_{t=t_j+1}^{t_{j+1}\wedge T}\Delta(x_t,a_t),
\end{aligned}$$

where the second inequality holds by definition of $\bar{j}$ and because the rewards are bounded in $[-1,1]$. It only remains to bound the regret on phases after $\bar{j}$. By Lemma B.8, we have $\phi_t \in \Phi^\star$ at all times in such phases.

Let $G_t := \{\mathrm{GLR_{t-1}(x_t;\phi_{t-1})} \leq \beta_{t-1,\delta/|\Phi|}(\phi_{t-1})\}$ be the event under which the GLRT does not trigger at time t. For any $j \geq \bar{j}$,

$$\sum_{t=t_j+1}^{t_{j+1}\wedge T}\Delta(x_t,a_t) = \sum_{t=t_j+1}^{t_{j+1}\wedge T}\mathbb{1}\{G_t\}\Delta(x_t,a_t) + \sum_{t=t_j+1}^{t_{j+1}\wedge T}\mathbb{1}\{\neg G_t\}\Delta(x_t,a_t) = \sum_{t=t_j+1}^{t_{j+1}\wedge T}\mathbb{1}\{G_t\}\Delta(x_t,a_t),$$

where the last equality holds since, under $\mathcal{E}$, if G_t does not hold, then the GLRT triggers, $a_t = \pi^\star_{t-1}(x_t;\phi_{t-1})$, and $\pi^\star_{t-1}(x_t;\phi_{t-1}) = \pi^\star(x_t)$ by Lemma B.6. Let $N_j := \sum_{t=t_j+1}^{t_{j+1}\wedge T}\mathbb{1}\{G_t\}$ be the total number of times the base algorithm $\mathfrak{A}$ is called in phase j. By event $\mathcal{E}_5$, the regret of $\mathfrak{A}$ on such steps is bounded as

$$\sum_{t=t_j+1}^{t_{j+1}\wedge T}\mathbb{1}\{G_t\}\Delta(x_t,a_t) \leq \overline{R}_{\mathfrak{A}}(N_j(t_{j+1}\wedge T),\phi_{t_j},\delta_j/|\Phi|).$$

Note that, for all $j \geq \bar{j}$, $N_j(t_{j+1} \wedge T) \leq t_{j+1} \wedge T - t_j \leq T - t_{\bar{j}} = T - \tau_{\text{elim}}$. Morevoer, the number of phases is $j \leq \log_2(T)$. Therefore, by the fact that $\overline{R}_{\mathfrak{A}}(\cdot, \phi, \cdot)$ is non-decreasing in the first and third argument,

$$\begin{aligned} R_T &\leq 2\tau_{\text{elim}} + \sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \overline{R}_{\mathfrak{A}}(T - \tau_{\text{elim}}, \phi_{t_j}, \delta_{\log_2(T)}/|\Phi|) \\ &\leq 2\tau_{\text{elim}} + \max_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(T - \tau_{\text{elim}}, \phi, \delta_{\log_2(T)}/|\Phi|) \log_2(T). \end{aligned}$$

□

Lemma B.10 (Bound on sub-optimal pulls)**.** *Under the same conditions as Theorem B.9, under event $\mathcal{E}$ (i.e., with probability at least $1 - 4\delta$), for any $T \in \mathbb{N}$,*

$$S_T = \sum_{t=1}^{T} \mathbb{1}\{a_t \neq \pi^\star(x_t)\} \leq \frac{2\tau_{\text{elim}} + \max_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(T, \phi, \delta_{\log_2(T)}/|\Phi|) \log_2(T)}{\Delta} =: g_T(\Phi, \Delta, \delta).$$

Proof. Note that, since the minimum gap is at least Δ, the event $\{a_t \neq \pi^\star(x_t)\}$ implies that $\Delta(x_t, a_t) \geq \Delta$. Then,

$$\begin{aligned} S_T \leq \sum_{t=1}^{T} \mathbb{1}\{\Delta(x_t, a_t) \geq \Delta\} &\leq \sum_{t=1}^{T} \frac{\Delta(x_t, a_t)}{\Delta} \\ &\leq \frac{2\tau_{\text{elim}} + \max_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(T, \phi, \delta_{\log_2(T)}/|\Phi|) \log_2(T)}{\Delta}, \end{aligned}$$

where the last inequality holds by Theorem B.9. □

B.8 Regret bound with HLS representations

Lemma B.11 (Selecting the HLS representation)**.** *Suppose Algorithm 1 is run with $\gamma = 2$ and $\mathcal{L}_t(\phi) = -\lambda_{\min}(V_t(\phi) - \lambda I_{d_\phi})/L_\phi^2$. Suppose that there exists a unique $\phi^\star \in \Phi^\star$ such that $\phi^\star$ is HLS. Then, under event $\mathcal{E}$ (i.e., with probability at least $1 - 4\delta$), $\phi_t = \phi^\star$ for all $t \geq \tau_{\text{hls}} \vee \tau_{\text{elim}}$, where*

$$\tau_{\text{hls}} := \min_{t \in \mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > \frac{2L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \left(g_t(\Phi, \Delta, \delta) + 8\sqrt{t \log \frac{4|\Phi| t \max_{\phi \in \Phi^\star} d_\phi}{\delta}} \right) \right\}.$$

Proof. Take any time $t \geq \tau_{\text{elim}}$. By Lemma B.8, we have $\Phi_t = \Phi^\star$ and, thus, $\phi^\star$ is the only active HLS representation. By the min-max theorem, $A \preceq B$ implies $\lambda_k(A) \leq \lambda_k(B)$ where λ_k is the k-th largest eigenvalue of the matrix. Then, from event $\mathcal{E}$, we have that, for all t,

$$\begin{aligned} \lambda_{\min}(V_t(\phi^\star) - \lambda I_{d_{\phi^\star}}) &\geq t\lambda^\star(\phi^\star) - L_{\phi^\star}^2 S_t - 8L_{\phi^\star}^2 \sqrt{t \log(4d_{\phi^\star}|\Phi|t/\delta)}, \\ \lambda_{\min}(V_t(\phi) - \lambda I_{d_\phi}) &\leq L_\phi^2 S_t + 8L_\phi^2 \sqrt{t \log(4d_\phi|\Phi|t/\delta)} \quad \forall \phi \in \Phi^\star, \phi \neq \phi^\star. \end{aligned}$$

If $t = 2^j$ for some $j \in \mathbb{N}$ (i.e., a time where representation selection is performed), $\phi^\star$ is selected if

$$\lambda_{\min}(V_t(\phi^\star) - \lambda I_{d_{\phi^\star}})/L_{\phi^\star}^2 > \max_{\phi \in \Phi^\star, \phi \neq \phi^\star} \lambda_{\min}(V_t(\phi) - \lambda I_{d_\phi})/L_\phi^2.$$

A sufficient condition based on the bounds above is

$$t\frac{\lambda^\star(\phi^\star)}{L_{\phi^\star}^2} > 2S_t + 8\sqrt{t \log(4d_{\phi^\star}|\Phi|t/\delta)} + \max_{\phi \in \Phi^\star, \phi \neq \phi^\star} \left(8\sqrt{t \log(4d_\phi|\Phi|t/\delta)} \right).$$

This, in turns, yields the simpler sufficient condition

$$t\frac{\lambda^\star(\phi^\star)}{L_{\phi^\star}^2} > 2S_t + 16\sqrt{t \log \left(\frac{4|\Phi| t \max_{\phi \in \Phi^\star} d_\phi}{\delta} \right)}.$$

Finally, using Lemma B.10 to bound S_t, it is sufficient that

$$t\frac{\lambda^\star(\phi^\star)}{L_{\phi^\star}^2} > 2g_t(\Phi,\Delta,\delta) + 16\sqrt{t\log\left(\frac{4|\Phi|t\max_{\phi\in\Phi^\star} d_\phi}{\delta}\right)}.$$

The right-hand side is a sub-linear function of t. The proof is concluded by rearringing this inequality and defining the first update time that satisfies it. □

Lemma B.12 (Triggering the GLRT)**.** *Suppose Algorithm 1 is run with $\gamma = 2$ and $\mathcal{L}_t(\phi) = -\lambda_{\min}(V_t(\phi) - \lambda I_{d_\phi})/L_\phi^2$. Suppose that there exists a unique $\phi^\star \in \Phi^\star$ such that $\phi^\star$ is HLS. Then, under the good event $\mathcal{E}$, the GLRT triggers for all for all $t \geq \tau_{\text{glrt}} \vee \tau_{\text{hls}} \vee \tau_{\text{elim}}$, where*

$$\tau_{\text{glrt}} := \min_{t\in\mathbb{N}}\left\{t \mid t \geq \frac{L_{\phi^\star}^2}{\lambda^\star(\phi^\star)}\left(\frac{16\beta_{t,\delta/|\Phi|}(\phi^\star)^2}{\Delta^2} + g_t(\Phi,\Delta,\delta) + 8\sqrt{t\log(4d_{\phi^\star}|\Phi|t/\delta)}\right) + 1\right\}.$$

Proof. From Lemma B.11, we know that $\phi_t = \phi^\star$ for all $t \geq \tau_{\text{hls}} \vee \tau_{\text{elim}}$. For simplicity, let us call $\phi := \phi^\star$. Take any time step $t \geq \tau_{\text{hls}} \vee \tau_{\text{elim}}$ (for which $\phi_t = \phi$), any $x \in \mathcal{X}$, and any $a \neq \pi_t^\star(x;\phi)$. Then, by the good event $\mathcal{E}$,

$$\begin{aligned}
\|\phi(x,\pi_t^\star(x;\phi)) - \phi(s,a)\|_{V_t(\phi)^{-1}} &\leq \frac{2L_\phi}{\sqrt{\lambda_{\min}(V_t(\phi))}} \\
&\leq \frac{2L_\phi}{\sqrt{t\lambda^\star(\phi) + \lambda - L_\phi^2 S_t - 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}}}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\big(\phi(x,\pi_t^\star(x;\phi)) - \phi(x,a)\big)^\mathsf{T}\theta_{\phi,t} &\geq \big(\phi(x,\pi^\star(x)) - \phi(x,a)\big)^\mathsf{T}\theta_{\phi,t} \\
&= \Delta(x,a) + \big(\phi(x,\pi^\star(x)) - \phi(x,a)\big)^\mathsf{T}(\theta_{\phi,t} - \theta_\phi^\star) \\
&\geq \Delta(x,a) - \|\phi(x,\pi^\star(x)) - \phi(x,a)\|_{V_t(\phi)^{-1}}\|\theta_{\phi,t} - \theta_\phi^\star\|_{V_t(\phi)} \\
&\geq \Delta(x,a) - \frac{2L_\phi\beta_{t,\delta/|\Phi|}(\phi)}{\sqrt{\lambda_{\min}(V_t(\phi))}} \\
&\geq \Delta(x,a) - \frac{2L_\phi\beta_{t,\delta/|\Phi|}(\phi)}{\sqrt{t\lambda^\star(\phi) + \lambda - L_\phi^2 S_t - 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}}} \\
&\geq \Delta - \frac{2L_\phi\beta_{t,\delta/|\Phi|}(\phi)}{\sqrt{t\lambda^\star(\phi) + \lambda - L_\phi^2 S_t - 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}}}.
\end{aligned}$$

Now suppose t is large enough so that the right-hand side is at least $\Delta/2$. Then, using the two inequalities above,

$$\begin{aligned}
\text{GLR}_t(x;\phi) &= \min_{a\neq\pi_t^\star(x;\phi)} \frac{\big(\phi(x,\pi_t^\star(x;\phi)) - \phi(x,a)\big)^\mathsf{T}\theta_{\phi,t}}{\|\phi(x,\pi_t^\star(x;\phi)) - \phi(s,a)\|_{V_t(\phi)^{-1}}} \\
&\geq \frac{\Delta}{4L_\phi}\sqrt{t\lambda^\star(\phi) + \lambda - L_\phi^2 S_t - 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}}.
\end{aligned}$$

Thus, a sufficient condition for the test trigger at time $t+1$ (recall that at time $t+1$ we perform the test with the statistics up to time t) is that the right-hand side above is larger than $\beta_{t,\delta/|\Phi|}(\phi)$. Therefore, for the test to trigger forever, we need simultaneously that

$$\frac{\Delta}{4L_\phi}\sqrt{t\lambda^\star(\phi) + \lambda - L_\phi^2 S_t - 8L_\phi^2\sqrt{t\log(4d_\phi|\Phi|t/\delta)}} \geq \beta_{t,\delta/|\Phi|}(\phi)$$

and that $t \geq \tau_{\text{hls}} \vee \tau_{\text{elim}}$. Note that this condition implies that the empirical gap is at least $\Delta/2$ as we required above. Using Lemma B.10 to bound S_t and rearranging concludes the proof. □

Theorem B.13 (Regret bound with HLS representation)**.** *Suppose Algorithm 1 is run with $\gamma = 2$ and $\mathcal{L}_t(\phi) = -\lambda_{\min}(V_t(\phi) - \lambda I_{d_\phi})/L_\phi^2$. Suppose $\phi^\star$ is the unique HLS representation in $\Phi^\star$. Under event $\mathcal{E}$ (i.e., with probability at least $1 - 4\delta$), for any $T \in \mathbb{N}$,*

$$R_T \leq 2\tau_{\text{elim}} + \max_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(\tau - \tau_{\text{elim}}, \phi, \delta_{\log_2(\tau)}/|\Phi|) \log_2(\tau),$$

where $\tau := \tau_{\text{glrt}} \vee \tau_{\text{hls}} \vee \tau_{\text{elim}}$.

Proof. Under $\mathcal{E}$, Lemma B.12 ensures that the GLRT triggers for $t \geq \tau_{\text{glrt}} \vee \tau_{\text{hls}} \vee \tau_{\text{elim}}$ with a realizable representation and, thus, the regret is zero for those times. Then, the result follows by using Theorem B.9 to bound the regret up to time $\tau_{\text{glrt}} \vee \tau_{\text{hls}} \vee \tau_{\text{elim}}$. □

B.9 Finding explicit bounds

Lemma B.14. *For $x \in \mathbb{R}$ and $c_1, c_2, c_3, c_4 \geq 0$, consider the inequality $x \leq c_1 + c_2\sqrt{x} + c_3\sqrt{x\log(x)} + c_4\log(x)$. Then, $x \lesssim c_1 + c_2^2 + c_3^2 + c_4$, where the $\lesssim$ notation hides constant and logarithmic terms.*

Proof. We can start by finding a crude bound on x by using the inequality $\log(x) \leq x^\alpha/\alpha$ for any $x, \alpha \geq 0$. Using it for $\alpha = 1/2$, we obtain

$$x \leq c_1 + c_2\sqrt{x} + \sqrt{2}c_3 x^{3/4} + 2c_4\sqrt{x}.$$

Suppose that $x \geq 1$. Then, $x \leq (c_1 + c_2 + \sqrt{2}c_3 + 2c_4)x^{3/4}$, which implies that $x \leq (c_1 + c_2 + \sqrt{2}c_3 + 2c_4)^4$. Therefore, we have $x \leq C$ for $C := \max\{(c_1 + c_2 + \sqrt{2}c_3 + 2c_4)^4, 1\}$. Plugging this into the logarithms in our initial inequality,

$$x \leq c_1 + (c_2 + c_3\sqrt{\log(C)})\sqrt{x} + c_4\log(C).$$

Solving this second-order inequality in $\sqrt{x}$ and using $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\begin{aligned} x &\leq \left(\frac{c_2 + c_3\sqrt{\log(C)}}{2} + \sqrt{\frac{(c_2 + c_3\sqrt{\log(C)})^2}{4} + c_1 + c_4\log(C)} \right)^2 \\ &\leq (c_2 + c_3\sqrt{\log(C)})^2 + 2c_1 + 2c_4\log(C) \lesssim c_1 + c_2^2 + c_3^2 + c_4. \end{aligned}$$

□

Lemma B.15. *The elimination time τ_{elim} defined in Lemma B.8 satisfies*

$$\tau_{\text{elim}} \lesssim \frac{d\log(|\Phi|/\delta)}{\min_{\phi \notin \Phi^\star} \epsilon_\phi}.$$

Proof. We know that $\tau_{\text{elim}} = 2^j$ for some specific j. Let $t = 2^{j-1}$ be the time at which the last update before τ_{elim} was performed. By definition, we have that

$$\begin{aligned} t &\leq \max_{\phi \notin \Phi^\star} \frac{1}{\epsilon_\phi} \left(D_t(\phi) + \min_{\phi^\star \in \Phi^\star} D_t(\phi^\star) + 328\log\frac{8|\Phi|^2 t^3}{\delta} \right) \\ &\leq \frac{320d\log(12BL) + 320d\log(t) + 328d\log(8|\Phi|^2/\delta) + 984\log(t)}{\min_{\phi \notin \Phi^\star} \epsilon_\phi}, \end{aligned}$$

where we used some simple crude bounds in the second inequality. Then, by Lemma B.14, $t \lesssim \frac{d\log(|\Phi|/\delta)}{\min_{\phi \notin \Phi^\star} \epsilon_\phi}$ and the same holds for τ_{elim} since $\tau_{\text{elim}} = 2t$. □

Lemma B.16. *The time τ_{hls} defined in Lemma B.11 satisfies*

$$\tau_{\text{hls}} \lesssim \tau_{\text{alg}} + \frac{L_{\phi^\star}^4 \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)^2} + \frac{\tau_{\text{elim}} L_{\phi^\star}^2}{\lambda^\star(\phi^\star)\Delta},$$

where

$$\tau_{\text{alg}} := \min_{t \in \mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > \frac{8L_{\phi^\star}^2 \log_2(t)}{\lambda^\star(\phi^\star)\Delta} \max_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(t, \phi, \delta_{\log_2(t)}/|\Phi|) \right\}.$$

Proof. By definition of τ_{hls},

$$\tau_{\text{hls}} \leq \min_{t\in\mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > 2\max\left(\frac{2L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} g_t(\Phi,\Delta,\delta), \frac{16L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \sqrt{t \log \frac{4|\Phi| t \max_{\phi\in\Phi^\star} d_\phi}{\delta}} \right) \right\}.$$

Thus, $\tau_{\text{hls}} \lesssim \tau'_{\text{hls}} + \tau''_{\text{hls}}$, where

$$\tau'_{\text{hls}} := \min_{t\in\mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > \frac{4L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} g_t(\Phi,\Delta,\delta) \right\},$$

$$\tau''_{\text{hls}} := \min_{t\in\mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > \frac{32L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \sqrt{t \log \frac{4|\Phi| t \max_{\phi\in\Phi^\star} d_\phi}{\delta}} \right\}.$$

We now bound τ''_{hls}. We know that $\tau''_{\text{hls}} = 2^j$ for some specific j. Let $t = 2^{j-1}$ be the time at which the last update before τ''_{hls} was performed. By definition, we have that

$$t \leq \frac{32L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \sqrt{t \log \frac{4|\Phi| t \max_{\phi\in\Phi^\star} d_\phi}{\delta}} \leq \frac{32L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \left(\sqrt{t \log \frac{4|\Phi| d}{\delta}} + \sqrt{t \log(t)} \right) \lesssim \frac{L_{\phi^\star}^4 \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)^2},$$

where we used Lemma B.14. The same holds for τ''_{hls} since $\tau''_{\text{hls}} = 2t$. We can now apply the same trick to τ'_{hls} by expanding the definition of $g_t(\Phi,\Delta,\delta)$. This yields

$$\tau'_{\text{hls}} \lesssim \tau_{\text{alg}} + \frac{\tau_{\text{elim}} L_{\phi^\star}^2}{\lambda^\star(\phi^\star)\Delta}.$$

□

Lemma B.17. *The time τ_{glrt} defined in Lemma B.12 satisfies*

$$\tau_{\text{glrt}} \lesssim \tau_{\text{alg}} + \frac{L_{\phi^\star}^4 \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)^2} + \frac{\tau_{\text{elim}} L_{\phi^\star}^2}{\lambda^\star(\phi^\star)\Delta} + \frac{L_{\phi^\star}^2 d_{\phi^\star} \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)\Delta^2},$$

where τ_{alg} is defined in Lemma B.16.

Proof. As we did in the proof of Lemma B.16, we can bound $\tau_{\text{glrt}} \lesssim \tau'_{\text{glrt}} + \tau''_{\text{glrt}} + \tau'''_{\text{glrt}}$, where

$$\tau'_{\text{glrt}} := \min_{t\in\mathbb{N}} \left\{ t \mid t \geq \frac{L_{\phi^\star}^2 \beta_{t,\delta/|\Phi|}(\phi^\star)^2}{\lambda^\star(\phi^\star)\Delta^2} \right\},$$

$$\tau''_{\text{glrt}} := \min_{t\in\mathbb{N}} \left\{ t \mid t \geq \frac{L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} g_t(\Phi,\Delta,\delta) \right\},$$

$$\tau'''_{\text{glrt}} := \min_{t\in\mathbb{N}} \left\{ t \mid t \geq \frac{L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \sqrt{t \log(4 d_{\phi^\star} |\Phi| t/\delta)} \right\}.$$

As before, we have

$$\tau''_{\text{glrt}} \lesssim \tau_{\text{alg}} + \frac{\tau_{\text{elim}} L_{\phi^\star}^2}{\lambda^\star(\phi^\star)\Delta} \quad \text{and} \quad \tau'''_{\text{glrt}} \lesssim \frac{L_{\phi^\star}^4 \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)^2}.$$

Regarding the first term, since $\beta_{t,\delta/|\Phi|}(\phi^\star)^2$ is of order $d_{\phi^\star} \log(t|\Phi|/\delta)$, by Lemma B.14,

$$\tau'_{\text{glrt}} \lesssim \frac{L_{\phi^\star}^2 d_{\phi^\star} \log(|\Phi|/\delta)}{\lambda^\star(\phi^\star)\Delta^2}.$$

□

B.10 Proof of the main theorems

The proof of Theorem 4.1 easily follows by using Lemma B.15, B.16, B.17 to simplify the expressions of the constant times in Theorem B.13.

Corollary 4.2 can be proved analogously to Theorem B.9 and B.13 while noting that, since $|\Phi| = 1$, the base algorithm is never reset (hence we can simply use confidence δ and remove the extra $\log_2(T)$ term) and $\tau_{\text{elim}} = \tau_{\text{hls}} = 0$.

Corollary 4.3 is simply a restatement of Theorem B.9.

C Variants of BANDITSRL

C.1 BANDITSRL: alternative losses

C.1.1 Obtaining best-in-class regret

Suppose that the upper bound $\overline{R}_{\mathfrak{A}}(T, \phi, \delta)$ to the regret of the base algorithm contains only known quantities (e.g., it could be a worst-case regret bound). Moreover, assume that the minimum gap Δ is known. This is only to simplify the notation in what follows, as we shall see at the end of this section that Δ can be estimated with a decreasing schedule without significantly altering the results. We consider the following alternative representation selection loss. For $j \in \mathbb{N}$,

$$\mathcal{L}_{\mathrm{bic},t_j}(\phi) = \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_j/|\Phi|) - \left[\frac{\lambda_{\min}(V_{t_j}(\phi) - \lambda I_{d_\phi})}{L_\phi^2} - g_{t_j}(\Phi, \Delta, \delta) - 8\sqrt{t_j \log(4d_\phi|\Phi|t_j/\delta)}\right]_+$$

where $[x]_+ := \max(x, 0)$. We show that with this selection loss we can achieve the best-in-class regret bound when no HLS realizable representation exists while preserving the constant-regret result when such a representation does exist.

Theorem C.1. *Suppose that $\Phi^\star$ does not contain any HLS representation. Under event $\mathcal{E}$ (i.e., with probability at least $1 - 4\delta$), for any $T \in \mathbb{N}$, the regret of Algorithm 1 with $\gamma = 2$ and loss $\mathcal{L}_{\mathrm{bic},t}(\phi)$ can be bounded as*

$$R_T \le 2\tau_{\mathrm{elim}} + \min_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(T, \phi, \delta_{\log_2(T)}/|\Phi|) \log_2(T),$$

where τ_{elim} is defined in Lemma B.8

Proof. Using exactly the same steps as in the proof of Theorem B.9, we have

$$R_T \le 2\tau_{\mathrm{elim}} + \sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \overline{R}_{\mathfrak{A}}(N_j(t_{j+1} \wedge T), \phi_{t_j}, \delta_j/|\Phi|),$$

where we recall that $\bar{j}$ is such that $\tau_{\mathrm{elim}} = 2^{\bar{j}}$. Note that $N_j(t_{j+1} \wedge T) \le t_{j+1} - t_j = t_j$. Moreover, under $\mathcal{E}$, for all $j \ge \bar{j}$, we have that $\Phi_{t_j} = \Phi^\star$ and, since $\Phi^\star$ does not contain any HLS representation,

$$\frac{\lambda_{\min}(V_{t_j}(\phi) - \lambda I_{d_\phi})}{L_\phi^2} - g_{t_j}(\Phi, \Delta, \delta) - 8\sqrt{t_j \log(4d_\phi|\Phi|t_j/\delta)} \le 0.$$

This implies that $\mathcal{L}_{\mathrm{bic},t_j}(\phi) = \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_j/|\Phi|)$ in such phases. Therefore,

$$\begin{aligned} R_T &\le 2\tau_{\mathrm{elim}} + \sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \overline{R}_{\mathfrak{A}}(t_j, \phi_{t_j}, \delta_j/|\Phi|) = 2\tau_{\mathrm{elim}} + \sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \mathcal{L}_{\mathrm{bic},t_j}(\phi_{t_j}) \\ &= 2\tau_{\mathrm{elim}} + \sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \min_{\phi \in \Phi^\star} \mathcal{L}_{\mathrm{bic},t_j}(\phi) = 2\tau_{\mathrm{elim}} + \sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \min_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_j/|\Phi|). \end{aligned}$$

The proof is concluded by noting that $\delta_j \ge \delta_{\log_2(T)}$ and $t_j \le T$, so that, by the properties $\overline{R}_{\mathfrak{A}}$, $\sum_{j=\bar{j}}^{\lfloor \log_2(T) \rfloor} \min_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_j/|\Phi|) \le \min_{\phi \in \Phi^\star} \overline{R}_{\mathfrak{A}}(T, \phi, \delta_{\log_2(T)}/|\Phi|) \log_2(T)$. □

Let us now derive the constant regret bound when a HLS representation exists. Note that, since we only changed the selection loss, Theorem B.9 and Lemma B.10 still hold. The only change is in the time τ_{hls} at which the HLS representation is selected. Theorem B.13 also continues to hold with the following redefinition of such time.

Lemma C.2 (Selecting the HLS representation with BIC loss)**.** *Suppose Algorithm 1 is run with $\gamma = 2$ and $\mathcal{L}_t(\phi) = \mathcal{L}_{\mathrm{bic},t}(\phi)$. Suppose that there exists a unique $\phi^\star \in \Phi^\star$ such that $\phi^\star$ is HLS. Then, under event $\mathcal{E}$ (i.e., with probability at least $1 - 4\delta$), $\phi_t = \phi^\star$ for all $t \ge \tau_{\mathrm{hls}} \vee \tau_{\mathrm{elim}}$, where*

$$\tau_{\mathrm{hls}} := \min_{t \in \mathbb{N}} \left\{ t \mid \exists j \in \mathbb{N}_{>0} : t = 2^j, t > \frac{L_{\phi^\star}^2}{\lambda^\star(\phi^\star)} \left(\overline{R}_{\mathfrak{A}}(t, \phi^\star, \delta_{\log_2(t)}/|\Phi|) + g_t(\Phi, \Delta, \delta) + 8\sqrt{t \log \frac{4|\Phi| t \max_{\phi \in \Phi^\star} d_\phi}{\delta}} \right) \right\}.$$

Proof. Take any time $t_j \geq \tau_{\text{elim}}$. By Lemma B.8, we have $\Phi_{t_j} = \Phi^\star$ and, thus, $\phi^\star$ is the only active HLS representation. Using the good event $\mathcal{E}$, we can easily see that $\mathcal{L}_{\text{bic},t_j}(\phi) \leq \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_j/|\Phi|)$ for all $\phi \in \Phi^\star, \phi \neq \phi^\star$. Moreover,

$$\frac{\lambda_{\min}(V_{t_j}(\phi) - \lambda I_{d_\phi})}{L_\phi^2} \geq t_j \frac{\lambda^\star(\phi^\star)}{L_{\phi^\star}^2} - g_{t_j}(\Phi, \Delta, \delta) - 8\sqrt{t_j \log(4d_{\phi^\star}|\Phi|t_j/\delta)}$$

and, thus,

$$\mathcal{L}_{\text{bic},t_j}(\phi^\star) \geq \overline{R}_{\mathfrak{A}}(t_j, \phi^\star, \delta_j/|\Phi|) - \left[t_j \frac{\lambda^\star(\phi^\star)}{L_{\phi^\star}^2} - 2g_{t_j}(\Phi, \Delta, \delta) - 16\sqrt{t_j \log \frac{4|\Phi|t_j \max_{\phi\in\Phi^\star} d_\phi}{\delta}}\right]_+$$

Therefore, a sufficient condition for selecting $\phi^\star$ is

$$t_j \frac{\lambda^\star(\phi^\star)}{L_{\phi^\star}^2} - 2g_{t_j}(\Phi, \Delta, \delta) - 16\sqrt{t_j \log \frac{4|\Phi|t_j \max_{\phi\in\Phi^\star} d_\phi}{\delta}} > \overline{R}_{\mathfrak{A}}(t_j, \phi^\star, \delta_j/|\Phi|).$$

The proof is concluded by rearringing this inequality. □

Dealing with unknown Δ If the minimum gap Δ is unknown, it can be easily guessed by a decreasing schedule $(1/t^\ell)_{t\geq 1}$. Then, we can replace the unknown term $g_{t_j}(\Phi, \Delta, \delta)$ in $\mathcal{L}_{\text{bic},t_j}(\phi)$ with $g_{t_j}(\Phi, 1/t_j^\ell, \delta)$. Since

$$g_{t_j}(\Phi, 1/t_j^\ell, \delta) = 2t_j^\ell \tau_{\text{elim}} + t_j^\ell \max_{\phi\in\Phi^\star} \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_{\log_2(t_j)}/|\Phi|) \log_2(t_j),$$

we only need $t_j^\ell \max_{\phi\in\Phi^\star} \overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_{\log_2(t_j)}/|\Phi|)$ to be sub-linear to derive our constant-regret result. For instance, if $\overline{R}_{\mathfrak{A}}(t_j, \phi, \delta_{\log_2(t_j)}/|\Phi|)$ is an $\widetilde{O}(\sqrt{t_j})$ regret bound, we can set $\ell = 1/4$. Then, the proofs of the two results above are the same except that we add a linear regret term $1/\Delta^{1/\ell}$ for the first time steps where $1/t^\ell > \Delta$.

C.1.2 Weak-HLS Loss

In Section 5, we introduced an alternative loss $\mathcal{L}_{\text{weak},t}(\phi) = -\min_{s\leq t}\left\{\phi(x_s, a_s)^\mathsf{T}(V_t(\phi) - \lambda I_{d_\phi})\phi(x_s, a_s)/L_\phi^2\right\}$, which is motivated by the notion of "weak-HLS" representations from [11] and appears to perform well in practice. In this section, **we will consider a slight variant**

$$\overline{\mathcal{L}}_{\text{weak},t}(\phi) = -\min_{s\leq t}\left\{\phi(x_s, a_s)^\mathsf{T}(V_t(\phi) - \lambda I_{d_\phi})\phi(x_s, a_s)/\left\|\phi(x_s, a_s)\right\|^2\right\}$$

where the features are normalized to have norm equal to one. The loss used in the experiments is $\mathcal{L}_{\text{weak},t}$ as defined in the main text.

We will show that $\overline{\mathcal{L}}_{\text{weak},t}$ does indeed select weak-HLS representations. We will assume throughout this section that both $\mathcal{X}$ and $\mathcal{A}$ are finite and $\text{supp}(\rho) = \mathcal{X}$. Let us first recall the definition of weak HLS. We abbreviate $\text{span}(\phi) = \text{span}\{\phi(x, a) \mid x \in \mathcal{X}, a \in \mathcal{A}\}$ and $\text{span}(\phi^\star) = \text{span}\{\phi(x, a_x^\star) \mid x \in \mathcal{X}\}$.

Definition C.1 (Weak-HLS Representation)**.** *A representation ϕ is weak-*HLS *if* $\text{span}(\phi^\star) = \text{span}(\phi)$.

The following characterization of the weak HLS property will be useful. We abbreviate $M_\phi^\star = \mathbb{E}_{x\sim\rho}\left[\phi(x, a_x^\star)\phi(x, a_x^\star)^\mathsf{T}\right]$.

Lemma C.3. *A representation ϕ is weak-*HLS *if and only if*

$$\min_{x\in\mathcal{X}, a\in\mathcal{A}} \left\{\frac{\phi(x, a)^\mathsf{T} M_\phi^\star \phi(x, a)}{\left\|\phi(x, a)\right\|^2}\right\} > 0. \tag{6}$$

Proof. We denote by $\text{Im}(A)$ the column space of a symmetric matrix A, and by $\ker(A)$ its kernel. Under our assumption that ρ is full-support, it is easy to see that $\text{span}(\phi^\star) = \text{Im}(M_\phi^\star)$. If ϕ is

weak-HLS, then

$$\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top M_\phi^\star\phi(x,a)}{\|\phi(x,a)\|^2}\right\} \geq \min_{v\in\mathrm{span}(\phi),\|v\|=1}\left\{v^\top M_\phi^\star v\right\} \tag{7}$$

$$= \min_{v\in\mathrm{span}(\phi^\star),\|v\|=1}\left\{v^\top M_\phi^\star v\right\} \tag{8}$$

$$= \min_{v\in\mathrm{Im}(M_\phi^\star),\|v\|=1}\left\{v^\top M_\phi^\star v\right\}, \tag{9}$$

and the latter is positive since it is the definition of the minimum *nonzero* eigenvalue of a positive semidefinite matrix.

Now assume (6) holds. We just need to show $\mathrm{span}(\phi)\subseteq\mathrm{span}(\phi^\star)$, since the other inclusion is trivial. By diagonalization, it is easy to show that the solution space of $\phi(x,a)^\top M_\phi^\star\phi(x,a)=0$ is $\ker(M_\phi^\star)$. Hence, (6) implies $\phi(x,a)\in\mathrm{Im}(M_\phi^\star)=\mathrm{span}(\phi^\star)$ for all $x\in\mathcal{X}$ and $a\in\mathcal{A}$. In turn, this implies $\mathrm{span}(\phi)\subseteq\mathrm{span}(\phi^\star)$, concluding the proof.

□

We can now show that our alternative loss does indeed select weak-HLS representations.

Lemma C.4. *Assume $\rho_{\min}>0$ is the minimum probability ρ assigns to any context, and $K=|\mathcal{A}|$. For any representation ϕ, ϵ-greedy with $\epsilon_t=t^{-1/3}$ guarantees that the following hold simultaneously with probability $1-5\delta$ for all $t\geq\left(\frac{K}{\rho_{\min}}\log\frac{1}{\delta}\right)^{3/2}$:*

$$\overline{\mathcal{L}}_{\mathrm{weak},t}(\phi)\leq -t\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top M_\phi^\star\phi(x,a)}{\|\phi(x,a)\|^2}\right\}+o(t) \qquad \textit{and} \tag{10}$$

$$\overline{\mathcal{L}}_{\mathrm{weak},t}(\phi)\geq -t\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top M_\phi^\star\phi(x,a)}{\|\phi(x,a)\|^2}\right\}-o(t) \tag{11}$$

Proof. From Lemma B.4, the good event $\mathcal{E}$ holds with probability $1-4\delta$. By $\mathcal{E}_2$, since Loewner ordering induces the same ordering on all quadratic forms:

$$\overline{\mathcal{L}}_{\mathrm{weak},t}(\phi) = -\min_{s\leq t}\left\{\frac{\phi(x_s,a_s)^\top(V_t(\phi)-\lambda I_{d_\phi})\phi(x_s,a_s)}{\|\phi(x_s,a_s)\|^2}\right\} \tag{12}$$

$$\leq -\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top(V_t(\phi)-\lambda I_{d_\phi})\phi(x,a)}{\|\phi(x,a)\|^2}\right\} \tag{13}$$

$$\leq -t\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top M_\phi^\star\phi(x,a)}{\|\phi(x,a)\|^2}\right\}+o(t), \tag{14}$$

where we have also used Lemma B.10 to bound the number of suboptimal pulls. Similarly, by $\mathcal{E}_3$:

$$\overline{\mathcal{L}}_{\mathrm{weak},t}(\phi)\geq -\min_{s\leq t}\left\{\frac{\phi(x_s,a_s)^\top M_\phi^\star\phi(x_s,a_s)}{\|\phi(x_s,a_s)\|^2}\right\}-o(t). \tag{15}$$

Let $(\overline{x},\overline{a})\in\arg\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top M_\phi^\star\phi(x,a)}{\|\phi(x,a)\|^2}\right\}$. Under our assumption, ϵ-greedy selects each context-action pair with probability at least $q=\rho_{\min}/(Kt^{1/3})$. After t rounds, the probability that it has not yet selected $(\overline{x},\overline{a})$ is at most $(1-q)^t$. A simple calculation shows that, by $t\geq\left(\frac{K}{\rho_{\min}}\log\frac{1}{\delta}\right)^{3/2}$, the algorithm has selected $(\overline{x},\overline{a})$ at least once with probability $1-\delta$, hence

$$\min_{s\leq t}\left\{\frac{\phi(x_s,a_s)^\top M_\phi^\star\phi(x_s,a_s)}{\|\phi(x_s,a_s)\|^2}\right\}=\frac{\phi(\overline{x},\overline{a})^\top M_\phi^\star\phi(\overline{x},\overline{a})}{\|\phi(\overline{x},\overline{a})\|^2}=\min_{x\in\mathcal{X},a\in\mathcal{A}}\left\{\frac{\phi(x,a)^\top M_\phi^\star\phi(x,a)}{\|\phi(x,a)\|^2}\right\}. \tag{16}$$

A union bound concludes the proof with an overall probability of $1-5\delta$. □

Now let ϕ_1 be a weak-HLS representation. Lemma C.3 and Equation 10 show that, with high probability, $\overline{\mathcal{L}}_{\text{weak},t}(\phi_1) \leq -t\widetilde{\lambda} + o(t)$ for some constant $\widetilde{\lambda} > 0$. From the proof of Lemma C.3 we can deduce that this $\widetilde{\lambda}$ is the minimum nonzero eigenvalue[7] of $M^{\star}_{\phi_1}$. On the other hand, consider a representation ϕ_2 that does not have the weak-HLS property. The other direction of Lemma C.3 and Equation 11 show that, with high probability, $\overline{\mathcal{L}}_{\text{weak},t}(\phi_1) \geq -o(t)$. Hence, the loss for the weak-HLS representations decreases (towards $-\infty$) much faster than representations that do not have this property. This justifies the use of $\overline{\mathcal{L}}$ as a loss in the BANDITSRL algorithm, when ϵ-greedy is used as a base algorithm. A more sophisticated argument allows to extend this result to any no-regret algorithm, by using the fact that they eventually sample all (finite) state-action pairs to ensure sufficient exploration.

When $\text{span}(\phi) = \mathbb{R}^d$, there is no distinction between HLS and weak-HLS. Moreover, [11] show that weak-HLS is enough for LINUCB to achieve constant regret. We could generalize the constant-regret result from this paper to weak-HLS in a similar fashion.

Empirical evaluation. We empirically compare $\overline{\mathcal{L}}_{\text{weak},t}$ and $\mathcal{L}_{\text{weak},t}$ on the same set of experiments reported in the main article. Fig. 3 shows that the loss $\mathcal{L}_{\text{weak},t}$ outperforms the theoretically grounded $\overline{\mathcal{L}}_{\text{weak},t}$ loss. We leave as open question whether the loss $\mathcal{L}_{\text{weak},t}$ is theoretically sound or not.

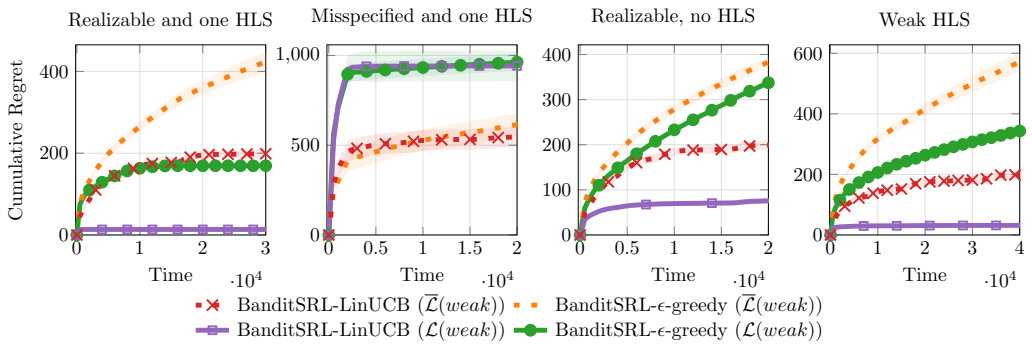

Figure 3: Varying dimension experiment with all realizable representations (left), misspecified representations (center-left), realizable non-HLS representations (center-right) and weak-HLS (right). Experiments are averaged over 40 repetitions as in the main paper.

C.2 NN-BANDITSRL: representation learning through neural networks

We recall that we consider a representation space Φ defined by the last layer of a Neural Network (NN). We denote by $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ the last layer and by $f(x, a) = \phi(x, a)^{\mathsf{T}}\omega$ the full NN, where ω are the last-layer weights. We report the pseudo code of NN-BANDITSRL in Alg. 2. The structure of NN-BANDITSRL is identical to the one of BANDITSRL, showing the generality and flexibility of the theoretical algorithm.

The GLRT is the same reported in Eq. 2. It leverages the current representation ϕ_j learnt by the NN and the regularized least squares parameters $V_t(\phi_j)$ and $\theta_{\phi_j,t}$. Note that, similarly to [28], we keep a separate estimate of the weights of the linear fitting (θ vs. ω). While the NN weights ω are learnt through the regularization loss (line 16 in Alg. 2), we compute $\theta_{\phi_j,t} = \text{argmin}_\theta \left\{ \frac{1}{t} \sum_{k=1}^{t} (\phi_{j_t}(x_t, a_t)^{\mathsf{T}}\theta - y_t)^2 + \lambda \|\theta\|_2^2 \right\}$ by RLS at each time t. This allows us to compute the best linear fit at each time t using efficient incremental updates (e.g., we can use Sherman-Morrison formula for computing directly $V_t(\phi_j)^{-1}$) and avoid to retrain the network after observing a new sample (x_t, a_t, y_t). An alternative approach is to train only the NN weights ω (i.e., keeping fix the representation ϕ) by stochastic gradient at each step, leading to an approximation of the RLS solution.

[7] Of course, an HLS representation is also weak-HLS, and $\widetilde{\lambda} = \lambda^{\star} > 0$. The converse is not true. Note also that the minimum nonzero eigenvalue $\widetilde{\lambda}$ is well-defined and positive for *all* representations, but it can only play the role of $\lambda^{\star}$ when the representation is weak-HLS.

Algorithm 2 NN-BANDITSRL

Input: Neural network f with last layer $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, no-regret algorithm $\mathfrak{A}$, confidence $\delta \in (0,1)$, update schedule $\gamma > 1$, regularizer $\lambda > 0$ and $c_{\text{reg}} > 0$
Initialize $j = 0$, f_j arbitrarily, $b_t(\phi_j) = 0$, $V_0(\phi_j) = \lambda I$
for $t = 1, \dots$ **do**
  Observe context x_t
  if $\mathrm{GLR_{t-1}(x_t; \phi_j)} > \beta_{t-1,\delta/|\Phi|}(\phi_j)$ **then**
    Play $a_t = \text{argmax}_{a \in \mathcal{A}} \left\{ \phi_j(x_t, a)^\mathsf{T} \theta_{\phi_j, t-1} \right\}$ and observe reward y_t
    $\mathcal{D}_{\text{glrt},t} = \mathcal{D}_{\text{glrt},t-1} \cup \{x_t, a_t, y_t\}$
  else
    Play $a_t = \mathfrak{A}_t(x_t; \phi_j, \delta)$, observe reward y_t, and feed it into $\mathfrak{A}$
    $\mathcal{D}_{\mathfrak{A},t} = \mathcal{D}_{\mathfrak{A},t-1} \cup \{x_t, a_t, y_t\}$
  end if
  Let $\mathcal{D}_t = \mathcal{D}_{\mathfrak{A},t} \cup \mathcal{D}_{\text{glrt},t}$
  Compute $V_t(\phi_j) = V_t(\phi_j) + \phi_j(x_t, a_t)\phi(x_t, a_t)^\mathsf{T}$, $b_t(\phi_j) = b_t(\phi_j) + \phi_j(x_t, a_t) y_t$ and $\theta_{\phi_j,t} = V_t(\phi_j)^{-1} b_t(\phi_j)$
  if $t = \lceil \gamma t_j \rceil$ **then**
    Set $j = j + 1$ and $t_j = t$
    Compute $\phi_j = \text{argmin}_\phi \min_f \left\{ \mathcal{L}_t(\phi) + c_{\text{reg}} \overline{E}_t(f) \right\}$ (see Eq. 17) and reset $\mathfrak{A}$
    Recompute least-square on the linear embedding ϕ_j using all samples

$$V_t(\phi_j) = \lambda I + \sum_{x,a,y \in \mathcal{D}_t} \phi_j(x,a)\phi_j(x,a)^\mathsf{T}, \quad b_t(\phi_j) = \sum_{x,a,y \in \mathcal{D}_t} \phi_j(x,a) y$$

    and $\theta_{\phi_j,t} = V_t(\phi_j)^{-1} b_t(\phi_j)$
  end if
end for

The phases scheme of BANDITSRL pairs very well with NN since it allows to perform the computationally costly operation of full NN training only $\log_\gamma(T)$ times. The NN is trained through a regression problem with an auxiliary representation loss promoting HLS-like representations. At the beginning of phase j, we solve the following problem

$$\begin{aligned} f, \phi_j &= \underset{\phi, f}{\text{argmin}} \left\{ \mathcal{L}_t(\phi) + c_{\text{reg}} \, \overline{E}_t(f) \right\} \\ &= \underset{\phi, \omega}{\text{argmin}} \left\{ \mathcal{L}_t(\phi) + \frac{c_{\text{reg}}}{|\mathcal{D}_{\mathfrak{A},t_j}|} \sum_{(x,a,y) \in \mathcal{D}_{\mathfrak{A},t_j}} \Big(\underbrace{\phi(x,a)^\mathsf{T} \omega}_{:= f(x,a)} - y \Big)^2 \right\} \\ &= \underset{\phi, \omega}{\text{argmin}} \left\{ c_{\text{reg},\mathcal{L}} \, \mathcal{L}_t(\phi) + \frac{1}{|\mathcal{D}_{\mathfrak{A},t_j}|} \sum_{(x,a,y) \in \mathcal{D}_{\mathfrak{A},t_j}} \Big(\underbrace{\phi(x,a)^\mathsf{T} \omega}_{:= f(x,a)} - y \Big)^2 \right\}. \end{aligned} \tag{17}$$

for some $c_{\text{reg},\mathcal{L}}, c_{\text{reg}} > 0$.[8] We recall that we compute the MSE regression loss using the explorative samples $\mathcal{D}_{\mathfrak{A},t_j}$ collected when playing the base algorithm $\mathfrak{A}$. As mentioned in the main paper, we use this separation to prevent the NN $f(x,a)$ to focus only on predicting optimal rewards when the the empirical distribution of the samples collapses towards the optimal actions (i.e., catastrophic forgetting). On the other hand, we can use all the samples $\mathcal{D}_t = \mathcal{D}_{\mathfrak{A},t} \cup \mathcal{D}_{\text{glrt},t}$ to compute the loss, where we want to leverage the bias/shift of the empirical distribution towards optimal actions to compute the empirical design matrix $V_t(\phi)$.

Concerning the loss $\mathcal{L}_t$, we leverage the same concepts used in BANDITSRL but we slightly modify them to make it more amenable for NN training. To optimize $\mathcal{L}_{\text{eig},t}(\phi)$ we leverage the fact that $\lambda_{\min}(M) = \min_z R(M,z)$, where $R(M,z) = \frac{z^\mathsf{T} M z}{z^\mathsf{T} z}$ is the Rayleigh quotient. We thus threat z as a

[8] In the experiments, we use scaling of the representation loss instead of MSE.

parameter and optimize it by gradient descent, leading to

$$\mathcal{L}_{\text{ray},t}(\phi) = \frac{-1}{|\mathcal{D}_{t_j}|} \min_{z \in \mathbb{R}^d} \frac{z^\mathsf{T}}{\|z\|_2} \left(\lambda I_d + \sum_{(x,a,y) \in \mathcal{D}_t} \frac{\phi(x,a)\phi^\mathsf{T}(x,a)}{\|\phi(x,a)\|_2^2} \right) \frac{z}{\|z\|_2} \tag{18}$$

We normalize the empirical design matrix to prevent features norms to grow unbounded. On the other hand, since the idea behind $\mathcal{L}_{\text{weak},t}(\phi)$ is to force the optimal features to span all the features we use a mixed approach to compute the loss. We leverage all the samples to compute the matrix $V_t(\phi)$, while we use the explorative samples $\mathcal{D}_{\mathfrak{A},t}$ to compute the quadratic form in V_t and avoid it collapses to evaluate only optimal actions. Then,

$$\mathcal{L}_{\text{weak},t}(\phi) = \frac{-1}{|\mathcal{D}_{t_j}|} \min_{(\overline{x},\overline{a},\overline{y}) \in \mathcal{D}_{\mathfrak{A},t}} \texttt{stop-grad}\left(\frac{\phi(\overline{x},\overline{a})^\mathsf{T}}{\|\phi(\overline{x},\overline{a})\|_2} \right) \left(\lambda I_d + \sum_{(x,a,y) \in \mathcal{D}_t} \frac{\phi(x,a)\phi^\mathsf{T}(x,a)}{\|\phi(x,a)\|_2^2} \right) \texttt{stop-grad}\left(\frac{\phi(\overline{x},\overline{a})}{\|\phi(\overline{x},\overline{a})\|_2} \right) \tag{19}$$

Where we apply the `stop-grad` operator on the outer features to only backpropagate gradient through the covariance matrix. We notice that the loss $\mathcal{L}_{\text{weak},t}$ resemble the $\mathcal{L}_{\text{eig},t}$ loss with the difference of being evaluated on the observed features rather than all the possible vectors in $\mathbb{R}^d$. We can optimize Eq. 17 by stochastic gradient descent using mini-batches but *we don't compute the gradient w.r.t. the outer features* $\phi(\overline{x},\overline{a})$.

Finally, nothing changes in term of base algorithm $\mathfrak{A}$ that now receives in input the trained NN f_j that can be used to extract the representation ϕ_j (that is fix through the entire phase). In the experiments, we use the standard LinUCB and ϵ-greedy algorithms to perform exploration given the representation ϕ_j.

D Experiments

In this section, we report additional information about the experiments. We recall that in all the experiments, we do a warm start of the base algorithm $\mathfrak{A}$ every time the representation changes using all the samples $\mathcal{D}_t$.

D.1 Linear Benchmarks

Parameters. In all the experiments, we consider all the theoretical parameters, e.g., $\gamma = 2$, $\delta = 0.01$ and $\lambda = 1$. For ϵ-greedy we use the schedule $\epsilon_t = t^{-1/3}$. For all the algorithms based on upper-confidence bound, we use the theoretical UCB value:

$$\text{UCB}_t(x,a,\phi) = \phi(x,a)^\mathsf{T}\theta_{\phi,t-1} + C_{\text{UCB},t}\|\phi(x,a)\|_{V_{t-1}^{-1}(\phi)} \tag{20}$$

where $C_{\text{UCB},t} = \alpha_{\text{UCB}}\,\sigma\sqrt{2\ln\left(\frac{\det(V_{t-1}(\phi))^{1/2}\det(\lambda I_{d_\phi})^{-1/2}}{\delta}\right)} + \sqrt{\lambda}B_\phi$, $\alpha_{\text{UCB}} = 1$ and σ is the standard deviation of the reward noise.

Varying dimension experiment. We providing additional information about the "varying dimension" problem introduced in [11]. This problem consists of six realizable representations with dimension from 2 to 6. Of the two representations of dimension $d = 6$, one is HLS. In addition seven misspecified representations are available: one considering half of the features of the HLS representation, one with a third of the same representation, and the five remaining are randomly generated representations with dimensions 3, 9, 12, 12, 18. The reward noise is drawn from a zero-mean Gaussian distribution with standard deviation $\sigma = 0.3$. All the results of the experiments can be found in the Sec. 5.

Mixing Representations. To provide a fair and comprehensive analysis, we also report the performance of the algorithms when none of the representations is HLS but a combination of them is. We consider the same problem in [11], where there are six realizable representations of the same

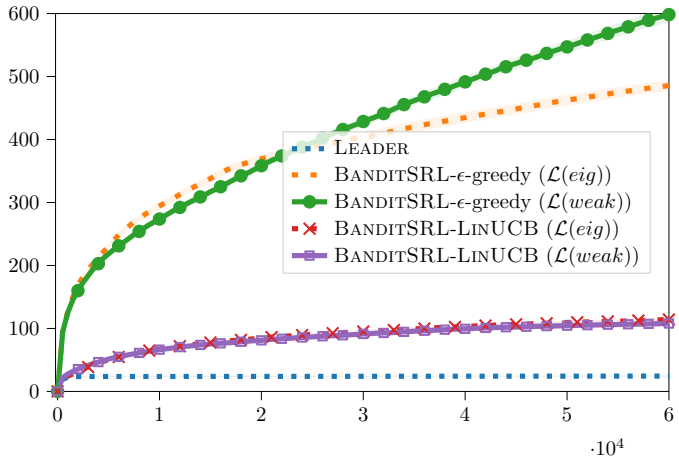

Figure 4: Cumulative regret of the algorithms in the mixing representation experiment, averaged over 40 repetitions.

dimension $d = 6$, none of which is HLS, but a mixture of them is HLS. We set $\sigma = 0.3$ for the reward noise. In this case, LEADER outperforms BANDITSRL and achieves constant regret (see Fig. 4). While LEADER is able to select a different representation for each context and mix them, BANDITSRL is only able to select a single representation for all the contexts and suffers sublinear regret. As mentioned before, this is both an advantage and drawback of LEADER since it needs to solve an optimization problem over representations for each context.

D.2 Non-Linear Benchmarks

Baselines. As baselines we consider LINUCB and ϵ-greedy with neural network and Random Fourier Features, the inverse gap weighting (IGW) strategy [e.g., 7, 10], NeuralUCB [27] and Neural-ThomposonSampling [6]. All the algorithms are implemented using the same phased schema of NN-BANDITSRL.

Neural-LinUCB fits a model to minimize the MSE and compute the UCB on the last layer of the NN.

NeuralTS performs randomized exploration on the last layer of the neural network, trained to minimize the MSE or our regularized problem. The exploration strategy is defined by the following two steps:

$$\widetilde{\theta} \sim \mathcal{N}(\theta_{\phi,t-1}, C^2_{\mathrm{UCB},t} V^{-1}_{t-1}(\phi)),$$
$$a_t = \operatorname*{argmax}_a \phi(x_t, a)^\mathsf{T} \widetilde{\theta}$$

The *IGW strategy* [e.g., 7, 10] trains the network to minimize the MSE and, at each time t, it plays an action a_t sampled from the following distribution

$$p_t(a) = \begin{cases} \frac{1}{A+\gamma_1 t^{\gamma_2}(\max_{a'} f_{j_t}(x,a') - f_{j_t}(x,a))} & \text{if } a \neq a^+_x := \operatorname{argmax}_{a'} f_{j_t}(x, a') \\ 1 - \sum_{a \neq a^+_x} p_t(a) & \text{otherwise} \end{cases}$$

Note that the network is kept fix during a phase, i.e., we do not refit the linear part at each step. We also tested the variant of IGW where we refit the last layer at each time step (see Fig. 6). We did not use the theoretical scaling factor (encoded here by γ_1 and γ_2) since it would be prohibitively large.

NeuralUCB [27] is similar to Neural-LINUCB but uses a bonus constructed with the whole gradient of the neural network. It thus selects the action that maximizes the following index

$$\mathrm{UCB}^{\mathrm{NeuralUCB}}_t(x, a) = f_{j_t}(x, a) + \alpha^{\mathrm{NeuralUCB}}_{\mathrm{UCB}} \|\nabla f_{j_t}(x, a)\|_{V^{-1}_{t-1}} \tag{21}$$

where $V^{-1}_{t-1}(f) = \sum_{k=1}^{t-1} \mathrm{diag}\Big(\nabla f_{j_k}(x_k, a_k) \nabla f_{j_k}(x_k, a_k)^\mathsf{T}\Big)$. While we use the theoretical bonus factor for Neural-LINUCB and NN-BANDITSRL, here we treat the bonus factor completely as an

hyperparameter since the true factor is prohibitively large. This is a clear advantage we provide to NeuralUCB.

We further compare our algorithm against stochastic linear bandit algorithms (i.e., ϵ-greedy and LINUCB) using random Fourier features [39]. We define $\phi(x, a) = W[x, a] + b$ with $[x, a] \in \mathbb{R}^m$ being the vector obtained from the concatenation of x and a, $W \in \mathbb{R}^{d \times m}$ is random matrix and $b \in \mathbb{R}^d$ is a random vector.

NN-BANDITSRL. We tested our algorithm with standard baseline methods: LinUCB, ϵ-greedy and IGW. LinUCB uses the theoretical parameters (see (20)) while the parameters for the other methods are reported below. As explained, we fix the representation ϕ_j for the epoch but we refit the linear parameter at each step.

Parameters. In all the experiments, we used the following parameters:

Name	Value
Phase schedule γ	1.2
Bonus parameter σ	0.2 for wheel, 0.5 for datasets
Scale factor GLRT (i.e., $\alpha_{\mathrm{GLRT}}\beta_{t-1,\delta}(\phi)$)	$\{1, 2, 5, 10, 15\}$
Scale factor UCB (i.e., α_{UCB} in Eq. 20)	$\{1, 2\}$
ϵ_t for ϵ-greedy	$\{t^{-1/3}, t^{-1/2}\}$
Loss regularization for NN-BANDITSRL ($c_{\mathrm{reg},\mathcal{L}}$)	1[9]
NN layers	$[50, 50, 50, 50, 10, 1]$
NN activation	ReLu
Batch size	128
Optimizer	SGD with learning rate 0.001 (0.0001 for Covertype)
Regularizer least-square	$\lambda = 1$
Buffer capacity	T
Scale factor for IGW (i.e., γ_1)	$\{1, 10, 50, 100\}$
Exploration rate for IGW (i.e., γ_2)	$\{1/3, 1/2\}$
Scale factor for NeuralUCB ($\alpha_{\mathrm{UCB}}^{\mathrm{NeuralUCB}}$ in Eq. 21)	$\{0.1, 1, 2, 5\}$
Random Fourier Features dimension (d)	$\{100, 300\}$

All the algorithms are implemented using Pytorch [40].

Domains. We considered the standard domains used in previous papers [e.g., 6, 27].

Wheel domain. In [6], the authors designed a synthetic non-linear contextual bandit problem where exploration is fundamental. Contexts are samples uniformly from the unit circle in $\mathbb{R}^2$ and $|\mathcal{A}| = 5$ are available. The first action a_1 has reward $\mu(x, a_1) = \mu_1$ for all x. The other actions have reward μ_i when $\|x\|_2 \leq C_r$. If $\|x\|_2 > C_r$, the sign of $x_1 x_2$ defines the optimal action. For example, a_2 is optimal when $x_1, x_2 > 0$, a_3 if $x_1 > 0$ and $x_2 < 0$ and so on. When an action $a_i \neq a_1$ is optimal the reward is μ_3, otherwise is μ_2 (a_1 has always reward μ_1). We set $\mu_1 = 1, \mu_2 = 0.8, \mu_3 = 1.2$ and $C_r = 0.5$. The reward noise is drawn from a zero-mean Gaussian distribution with standard deviation $\sigma = 0.2$. For the experiments, we consider a finite subset of contexts by sampling $X = 100$ contexts at the beginning of the experiment. All the repetitions are done with the same bandit problem (i.e., contexts are fix). We samples contexts accordingly to a uniform distribution $\rho = U(\{1, \ldots, X\})$. The features ϕ are obtained by concatenating the context with a one-hot encoding of the action ($d_\phi = 7$). Let 1_i be the vector of dimension 5 with all zeros except a one in position i, then $\phi(x, a_i) = [x, 1_{i-1}]$, for all x, $i = 1, \ldots, 5$.

Dataset-based domain. We evaluate our algorithm on standard dataset-based environments [e.g 6, 27] from the UCI repository [34–37]: MAGIC Gamma Telescope Data Set, Mushroom, Statlog (Shuttle) Data Set, Covertype Data Set. We use the classical multiclass-to-bandit conversion. We use noisy rewards with Bernoulli distribution $Bern(p)$ where $p = 0.9$ if the action is equal to the correct label for the sample x, $p = 0.1$ otherwise. The features are obtained by replicating the context $|\mathcal{A}|$-times, leading to a dimension $d = d_{\mathcal{X}}|\mathcal{A}|$ where $d_{\mathcal{X}}$ is the dimension of the context. We samples contexts accordingly to a uniform distribution $\rho = U(\mathcal{X})$. We report the characteristic of the datasets after an initial preprocessing.

[9] Note that in the code we add the regularization on the loss $\mathcal{L}_t$ and not on the MSE.

	Covertype	Magic	Mushroom	Statlog (Shuttle)
Number of contexts $\lvert\mathcal{X}\rvert$	581012	19020	8124	58000
Context dimension $d_{\mathcal{X}}$	54	10	22	9
Number of actions $\lvert\mathcal{A}\rvert$	7	2	2	7
Feature dimension d	378	20	44	63

D.2.1 Additional Experiments and Ablation

In this section we provide additional experiments and comparisons for NN-BANDITSRL. The overall message is that there always exists a configuration of NN-BANDITSRL that works well across domains and outperforms the base algorithms.

We start noticing that ϵ-greedy often outperforms LINUCB. Randomization at the level of actions is particularly efficient in these domains since the dimension of the output layer of the NN is always larger than the number of actions. This provides an advantage to ϵ-greedy since it needs to perform less exploration. Furthermore, the GLRT prevents ϵ-greedy to over explore.

In the main paper we have only reported results using the theoretical configurations of the base algorithms ($\epsilon_t = t^{-1/3}$ and $\alpha_{\mathrm{UCB}} = 1$). Fig. 5 shows that NN-BANDITSRL with $\alpha_{\mathrm{GLRT}} = 5$ is robust to variations of the base algorithm. In particular, it outperforms or performs comparably to the base algorithm and the baselines in all the experiments. The interesting thing to notice is that the different domains require a different level of exploration. The wheel domain requires a high level of exploration ($\alpha_{\mathrm{UCB}} = 2$ and $\epsilon_t = t^{-1/3}$), while the algorithms performs better with little exploration in mushroom ($\alpha_{\mathrm{UCB}} = 0.1$ and $\epsilon_t = t^{-1/2}$). We can notice that Random Fourier Features performs poorly in almost all the experiments, supporting the need of representation learning. It may be however possible to obtain better performance by using a much higher number of features. Finally, Fig. 6 shows the behavior of NN-BANDITSRL with IGW strategy for different values of γ_1 and γ_2. Interestingly, it outperforms the best version of the IGW strategy based MSE.

The second experiment aims to highlight the impact of the GLRT on the behavior of NN-BANDITSRL (Fig. 8). We can notice that the GLRT plays an important role in Neural-ϵ-greedy (see also Fig. 9), in particular when using the theoretical exploration rate $t^{-1/3}$ where it significantly improve the performance. On the other hand, the GLRT may trigger too many times when $\alpha_{\mathrm{GLRT}} = 1$, leading to under-exploration and worse regret. Note that there are potentially other confounding factors leading to this undesired behavior. For example, the fact we use only exploratory data may lead to suboptimal fitting of the reward if the GLRT triggers too early. Indeed, as soon as we increase the GLRT scale factor (i.e., $\alpha_{\mathrm{GLRT}} \geq 2$), we do not see anymore a negative impact. In general, better and more consistent results are obtained with the theoretical exploration rate $t^{-1/3}$ where over exploration is prevented by the GLRT. The GLRT plays a milder role for LINUCB-based algorithms (see also Fig. 9). Indeed, [11] showed that LINUCB is able to take advantage of the HLS property and does not requires a GLRT mechanism to achieve constant regret. The overall message is to set the GLRT scale factor to a value larger than the theoretical one (and larger than the one used for LINUCB-based algorithms). Similar results can be derived for Thompson Sampling.

To further investigate the behavior of NN-BANDITSRL, we performed an ablation study w.r.t. the losses $\mathcal{L}_{\mathrm{ray}}$ and $\mathcal{L}_{\mathrm{weak}}$ (see Eq. 18-19) and the contribution of the GLRT (i.e., $\alpha_{\mathrm{GLRT}} \in \{0, 5\}$), see Fig. 9-10. We can see for Neural-ϵ-greedy that the GLRT plays a fundamental role in avoiding over exploration. Furthermore, the regularization improves or at least does not degrade the performance of the algorithm. As mentioned before for LINUCB-based algorithms, the GLRT does not play an important role. On the other hand, these experiments show the importance of the spectral regularization. We can indeed notice a clear separation between the performance of the algorithm with and without regularization.

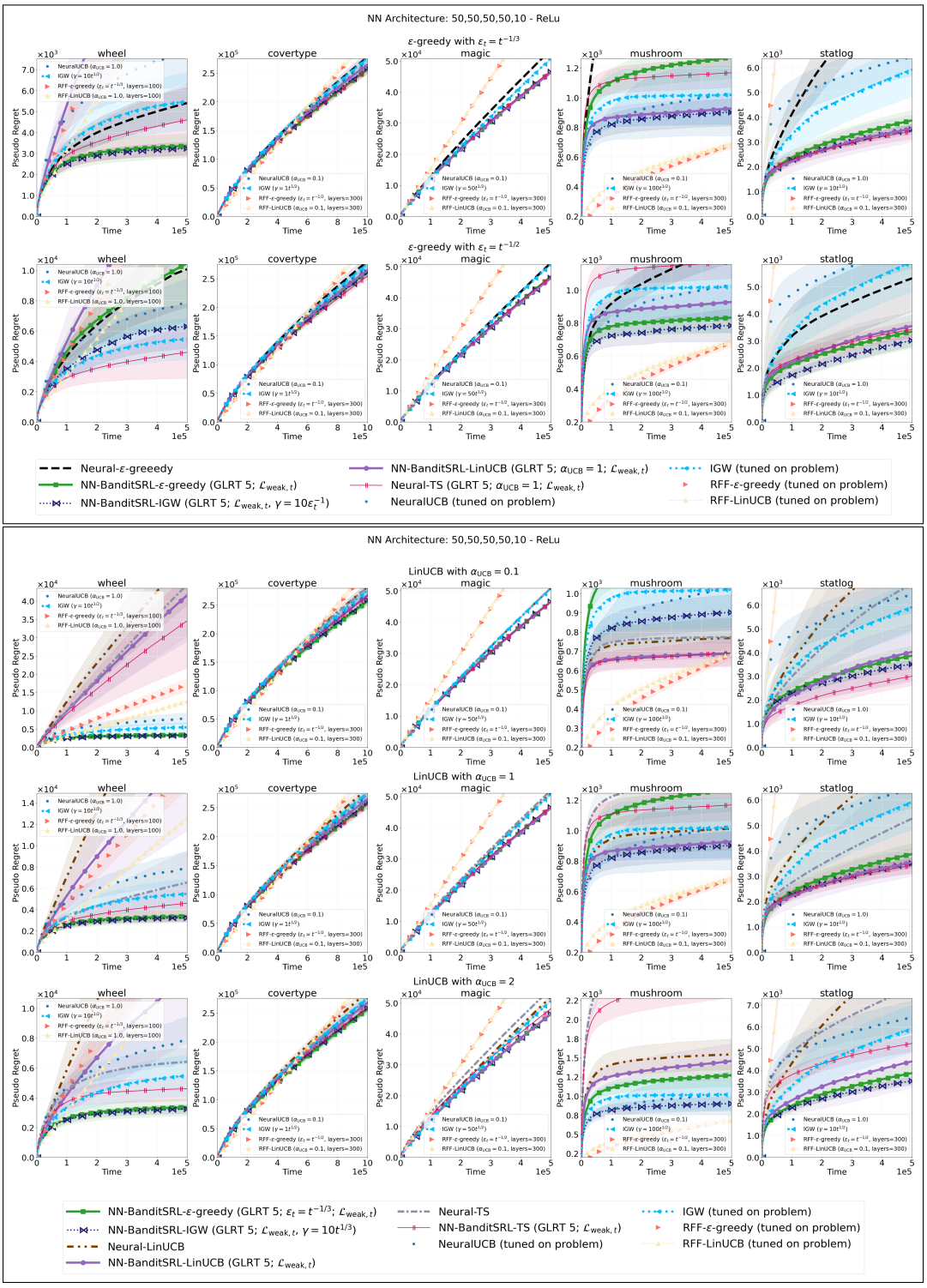

Figure 5: Ablation study of NN-BANDITSRL with $\alpha_{\text{GLRT}} = 5$ and different base algorithms (i.e., $\alpha_{\text{UCB}} \in \{0.1, 1, 2\}$, $\epsilon_t \in \{t^{-1/3}, t^{-1/2}\}$). Results are averaged over 20 runs. We report the performance of NN-BANDITSRL against the best configuration of the baselines.

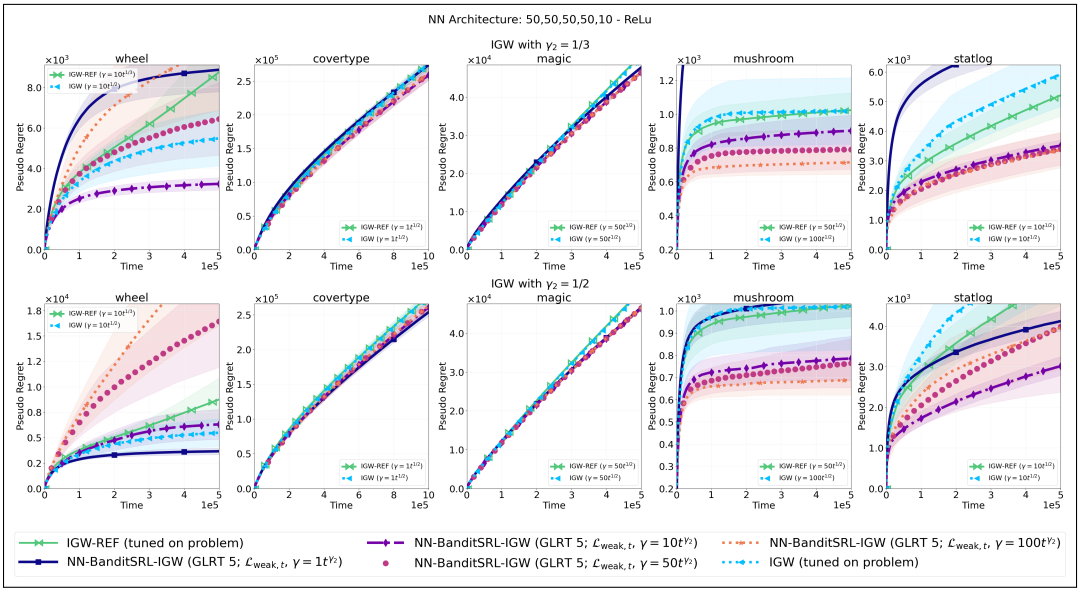

Figure 6: Ablation study of NN-BANDITSRL with $\alpha_{\text{GLRT}} = 5$ and IGW strategy for different values of γ_1 and γ_2 IGW-REF denotes the variant of IGW where we refit the last layer of the NN at each time step. Results are averaged over 20 runs.

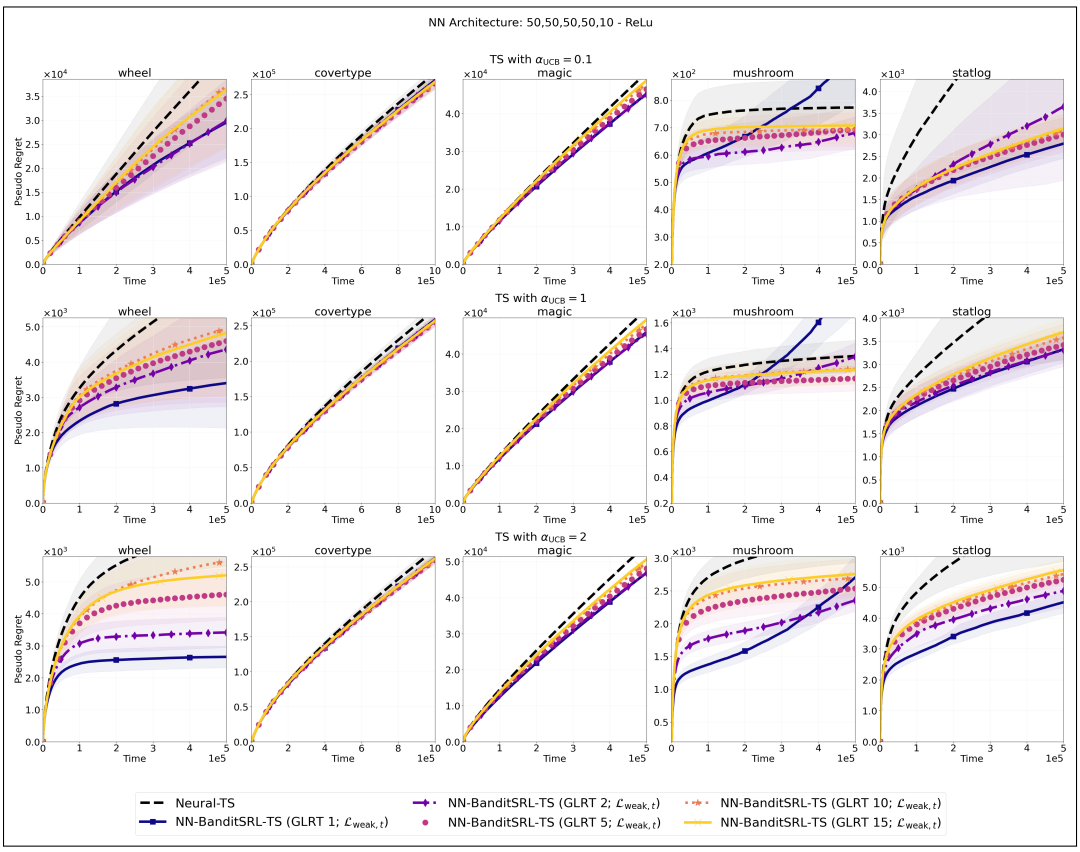

Figure 7: Ablation study of NN-BANDITSRL with different GLRT values ($\alpha_{\text{GLRT}} \in \{1, 2, 5, 10, 15\}$) for Thompson Sampling. Results are averaged over 20 runs.

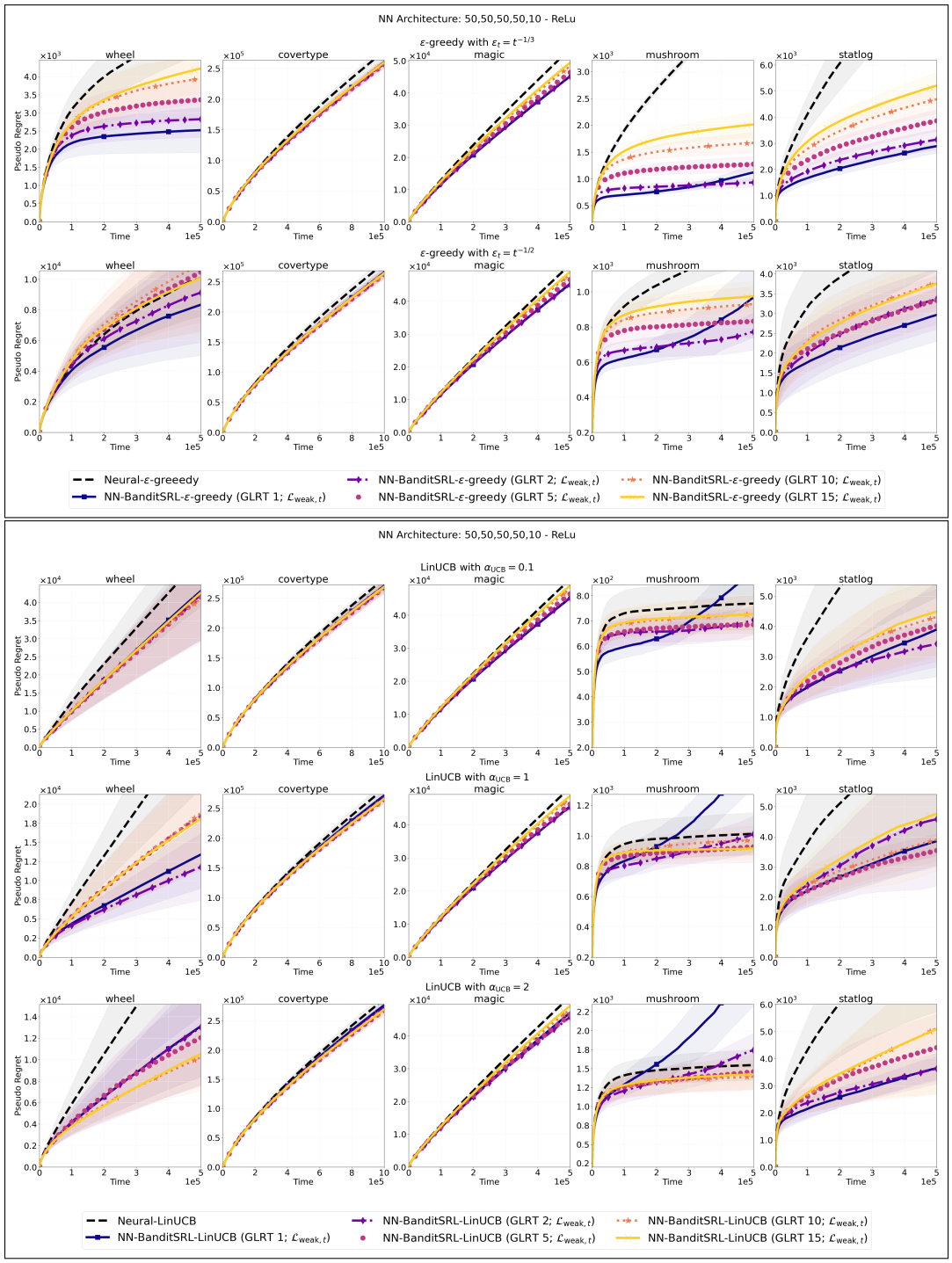

Figure 8: Ablation study of NN-BANDITSRL with different GLRT values ($\alpha_{\text{GLRT}} \in \{1, 2, 5, 10, 15\}$) and base algorithms (i.e., $\alpha_{\text{UCB}} \in \{1, 2\}$, $\epsilon_t \in \{t^{-1/3}, t^{-1/2}\}$). Results are averaged over 20 runs.

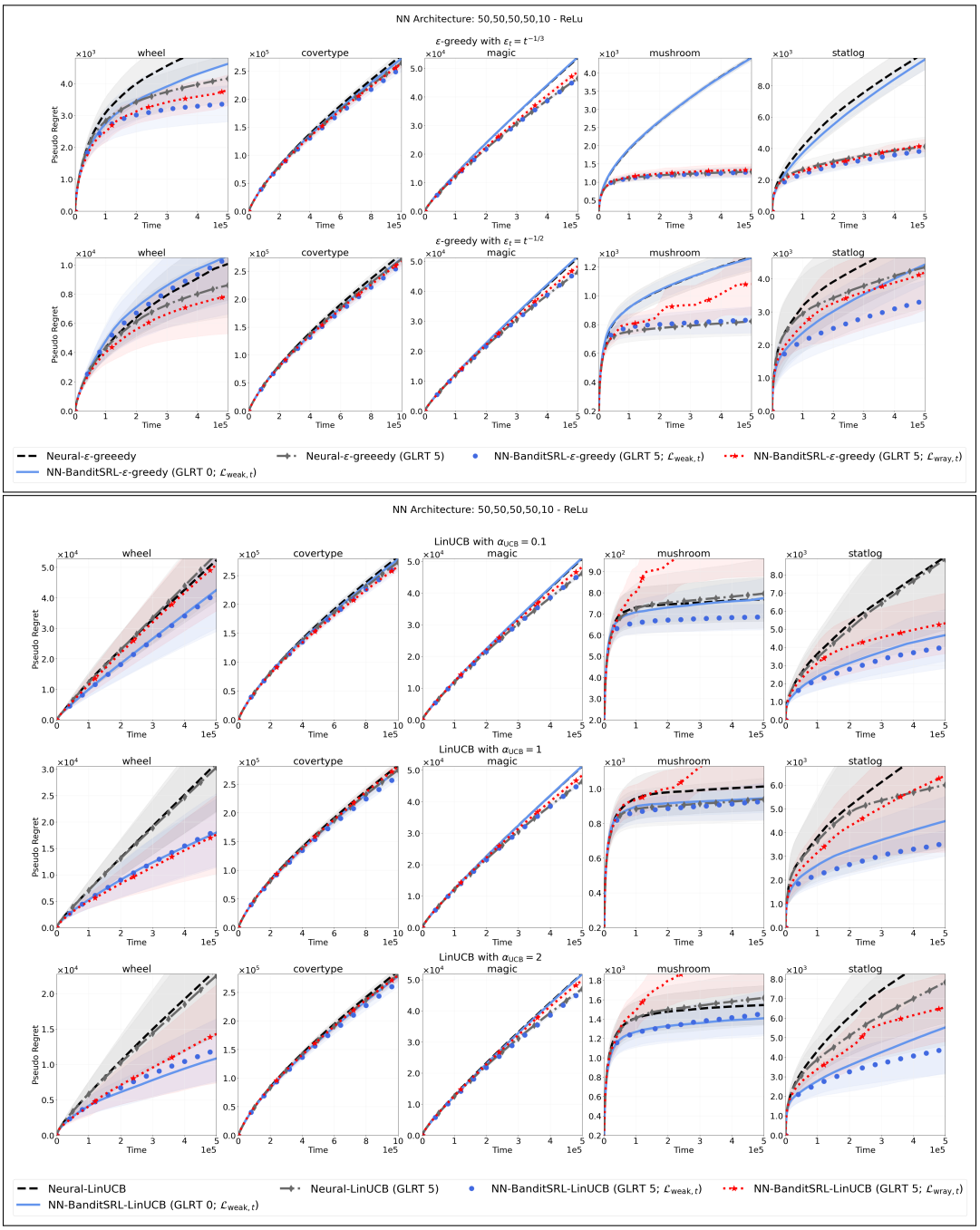

Figure 9: Ablation study of NN-BANDITSRL with different GLRT values ($\alpha_{\text{GLRT}} \in \{0, 5\}$), base algorithms (i.e., $\alpha_{\text{UCB}} \in \{1, 2\}$, $\epsilon_t \in \{t^{-1/3}, t^{-1/2}\}$) and regularization loss. Results are averaged over 20 runs.

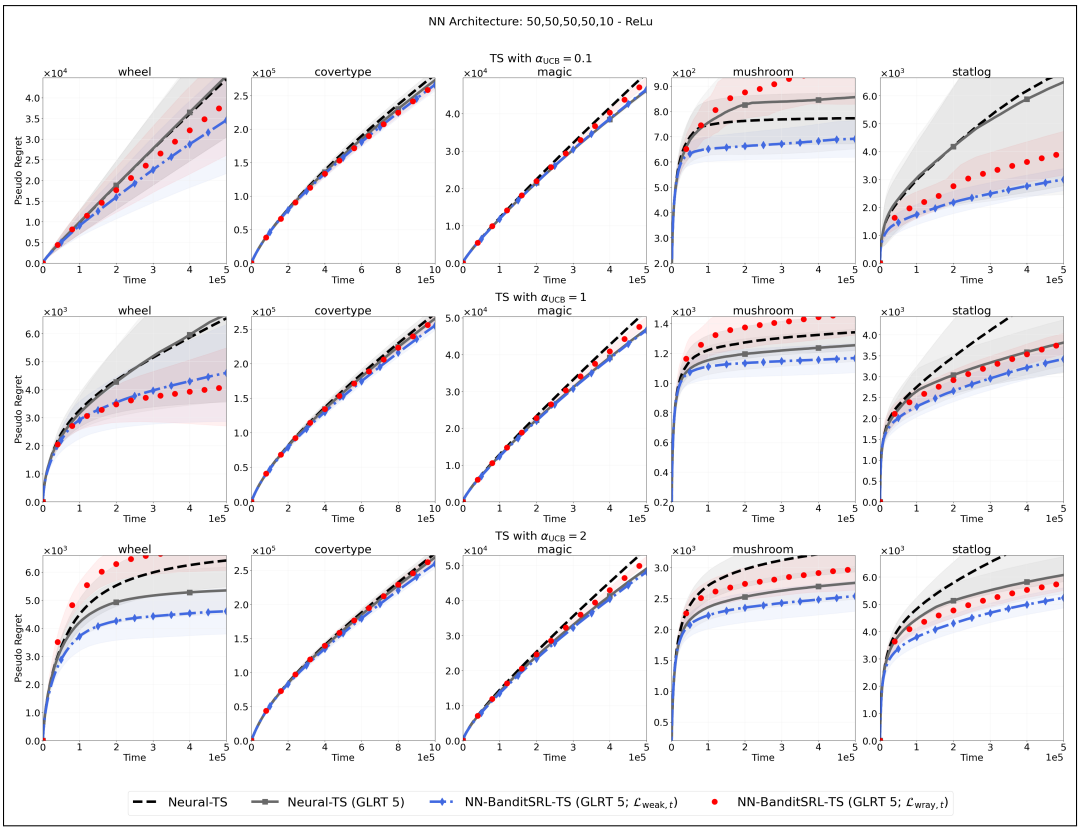

Figure 10: Ablation study of NN-BANDITSRL with different GLRT values ($\alpha_{\text{GLRT}} \in \{0, 5\}$) and regularization loss for Thompson Sampling. Results are averaged over 20 runs.

D.2.2 Network study on the Wheel Domain

To further investigate the behavior of NN-BANDITSRL, we performed an ablation study w.r.t. the network structure.

Let's start considering ϵ-greedy algorithms. Fig. 11 that the performance of these algorithms does not vary much across the experiments. However, there are interesting things to notice. When the embedding layer is large (1000,100), the regularization and GLRT do not help and NN-BANDITSRL behaves as the Neural-ϵ-greedy algorithm. Indeed it may be difficult to recover spectral properties for such a large representation (the original feature dimension is 7). Similarly the GLRT scales with the dimension d, the higher d the larger may be the time to trigger the test. When the embedding dimension is smaller, we can see an improved performance for NN-BANDITSRL compared to the base algorithm. The best regret is obtained with the deepest network and smallest embedding dimension (i.e., 10). In particular, we can see a flattening curve for NN-BANDITSRL with net $[50, 50, 50, 50, 10]$ that is not observe with embedding dimension 50.

LINUCB-based algorithms suffer when the embedding dimension is large (i.e., 1000, 100) since it needs to perform much more exploration compared to ϵ-greedy. Indeed, ϵ-greedy only needs to do exploration at the level of the 5 actions, while LINUCB needs to explore the d-dimensional space. An interesting behavior is observed with deeper networks. In particular, we observe a better performance with embedding dimension 50 rather than 10. We think that with dimension 10 the network has a larger misspecification that compromises the exploration performed by LINUCB-based algorithms. Indeed, Fig. 12 shows that both NN-BANDITSRL and Neural-LINUCB show a linear regret. This demonstrates that i) LINUCB-based algorithms are much more sensible to the misspecification than ϵ-greedy; ii) it is important to carefully select the embedding dimension d (the larger the higher the level of exploration but the smaller the misspecification). On the other hand, when $d = 50$, LINUCB-based algorithms perform comparably to ϵ-greedy. While with a shallow network (i.e., $[50, 50, 50]$) we observe a small improvement in using NN-BANDITSRL, the advantages of NN-BANDITSRL becomes extremely clear with the deep network (i.e., $[50, 50, 50, 50, 50]$) where it achieves more than half of the regret of Neural-LINUCB.

Finally, Fig. 12 shows that, similarly to ϵ-greedy, Thompson Sampling works better with smaller dimensions (in particular 10) where we can always observe a smaller regret for NN-BANDITSRL.

E Examples of No-regret Algorithms

We prove that LinUCB and ϵ-greedy satisfy Assumption 4. Then, we instantiate our general regret bounds (i.e., we bound τ_{alg} defined in Lemma B.16) for these specific algorithms.

E.1 LinUCB

Theorem E.1 (Regret bound of anytime LinUCB, Prop. 1 in [11])**.** *Let $\phi \in \Phi^\star$ be any realizable representation. With probability $1 - \delta$, for any $T \in \mathbb{N}$, the regret of anytime LinUCB run with representation ϕ, confidence δ, and threshold $\beta_{t,\delta}(\phi)$ is bounded as*

$$R_T \leq \overline{R}_{\text{LinUCB}}(T, \phi, \delta), =: \frac{128\lambda B_\phi^2 \sigma^2 \left(2\log(1/\delta) + d_\phi \log(1 + TL_\phi^2/(\lambda d_\phi))\right)^2}{\Delta}.$$

Proof. Just apply Proposition 1 in [11] while noting that the maximum per-step regret is 2 in our context. □

Lemma E.2. *When using the LinUCB algorithm, we have*

$$\tau_{\text{alg}} \lesssim \frac{L_{\phi^\star}^2 d^2 \log(|\Phi|/\delta)^2}{\lambda^\star(\phi^\star)\Delta^2}.$$

Proof. First note that, by Theorem E.1,

$$\overline{R}_{\text{LinUCB}}(t, \phi, \delta_{\log_2(t)}/|\Phi|) \lesssim \frac{d_\phi^2 \log(t|\Phi|/\delta)^2}{\Delta}.$$

Then, the result follows by applying Lemma B.14. □

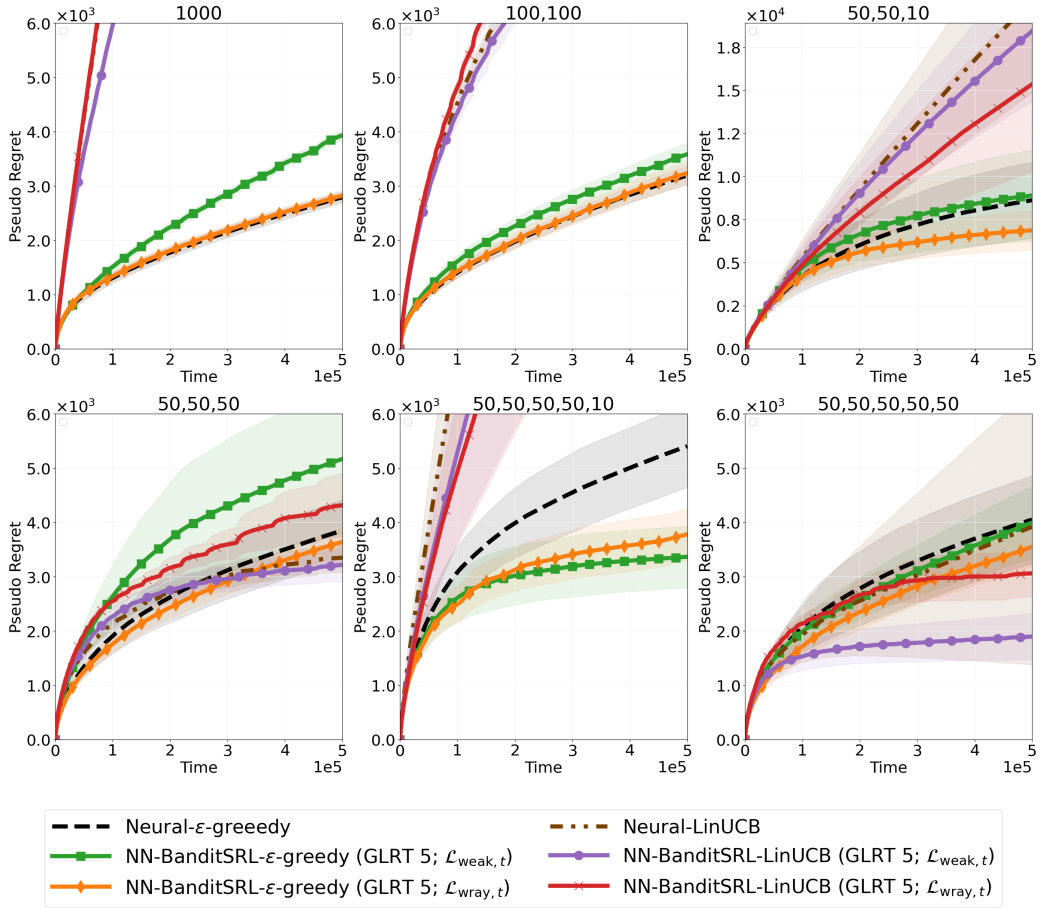

Figure 11: Ablation study of NN-BANDITSRL with ϵ-greedy and LINUCB on the Wheel domain. Results are averaged over 20 runs. The figure title corresponds to the network dimension.

E.2 ϵ-greedy

Theorem E.3 (Regret bound of ϵ-greedy)**.** *Let $\phi \in \Phi^\star$ be any realizable representation. With probability $1-\delta$, for any $T \in \mathbb{N}$, the regret of ϵ-greedy run with representation ϕ, confidence δ, and forcing schedule $(\epsilon_t)_{t\geq 1}$ with $\epsilon_t = 1/t^{1/3}$ is bounded as*

$$R_T \leq \overline{R}_{\epsilon-\text{greedy}}(T,\phi,\delta), =: 2\beta_{T,\delta/3}(\phi)\left(\frac{L_\phi}{\sqrt{\lambda}}\left(\frac{128L_\phi^2 A\sqrt{\log(12d_\phi/\delta)}}{\Gamma(\phi)}\right)^8 + \frac{2L_\phi}{\sqrt{\lambda}} + \frac{3L_\phi\sqrt{A}T^{2/3}}{\sqrt{\Gamma(\phi)}}\right) + 2\sqrt{T\log(6T/\delta)} + 3T^{2/3},$$

where $\Gamma(\phi) := \lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\sum_{a\in\mathcal{A}}\phi(x,a)\phi(x,a)^\intercal\right]\right)$ *and* $\beta_{T,\delta}(\phi) := \sigma\sqrt{2\log(1/\delta) + d_\phi\log(1+TL_\phi^2/(\lambda d_\phi))} + \sqrt{\lambda}B_\phi$.

Proof. Let F_t be the event under which the algorithm plays greedily at time t. Then,

$$R_T = \underbrace{\sum_{t=1}^{T}\mathbb{1}\{F_t\}\,\Delta(x_t,a_t)}_{(a)} + \underbrace{\sum_{t=1}^{T}\mathbb{1}\{\neg F_t\}\,\Delta(x_t,a_t)}_{(b)}.$$

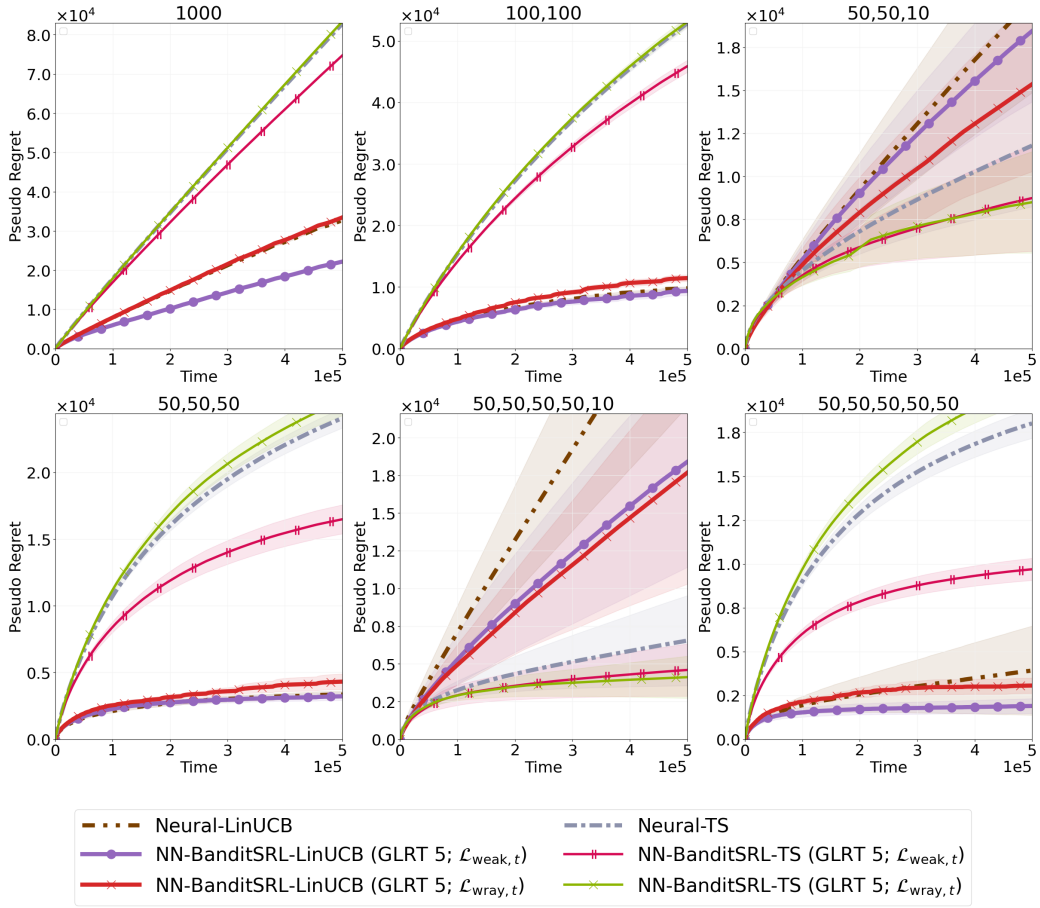

Figure 12: Ablation study of NN-BANDITSRL with LINUCB and TS on the Wheel domain. Results are averaged over 20 runs. The figure title corresponds to the network dimension.

Let us start from (a). With probability at least $1-\delta$, we have that, under F_t,

$$
\begin{aligned}
\Delta(x_t, a_t) &= \max_{a\in\mathcal{A}} \mu(x_t, a) - \mu(x_t, a_t) \\
&\leq \max_{a\in\mathcal{A}} \left(\langle \theta_{\phi,t-1}, \phi(x_t, a)\rangle + \beta_{t-1,\delta}(\phi) \|\phi(x_t, a)\|_{V_{t-1}^{-1}(\phi)} \right) - \langle \theta_{\phi,t-1}, \phi(x_t, a_t)\rangle + \beta_{t-1,\delta}(\phi) \|\phi(x_t, a_t)\|_{V_{t-1}^{-1}(\phi)} \\
&\leq \max_{a\in\mathcal{A}} \langle \theta_{\phi,t-1}, \phi(x_t, a)\rangle - \langle \theta_{\phi,t-1}, \phi(x_t, a_t)\rangle + 2 \max_{a\in\mathcal{A}} \beta_{t-1,\delta}(\phi) \|\phi(x_t, a)\|_{V_{t-1}^{-1}(\phi)} \\
&= 2 \max_{a\in\mathcal{A}} \beta_{t-1,\delta}(\phi) \|\phi(x_t, a)\|_{V_{t-1}^{-1}(\phi)},
\end{aligned}
$$

where the last equality is because a_t is greedy w.r.t. $\theta_{\phi,t-1}$ under F_t. Then,

$$
(a) \leq 2\beta_{T,\delta}(\phi) \sum_{t=1}^{T} \mathbb{1}\{F_t\} \max_{a\in\mathcal{A}} \|\phi(x_t, a)\|_{V_{t-1}^{-1}(\phi)} \leq 2\beta_{T,\delta}(\phi) \sum_{t=1}^{T} \mathbb{1}\{F_t\} \frac{L_\phi}{\sqrt{\lambda_{\min}(V_{t-1}(\phi))}}.
$$

Let $\mathbb{E}_t$ be the expectation operator conditioned on the full history up to round $t-1$ and $\pi_t(a|x) = (1-\epsilon_t)\mathbb{1}\{a = \text{argmax}_{a\in\mathcal{A}} \langle \theta_{\phi,t-1}, \phi(x_t, a)\rangle\} + \frac{\epsilon_t}{|\mathcal{A}|}$ be the stochastic policy played at time t. By

Matrix Azuma inequality (Lemma F.4) and a union bound on time, with probability at least $1-\delta$,

$$\begin{aligned}
\lambda_{\min}(V_{t-1}(\phi)) &\geq \lambda + \lambda_{\min}\left(\sum_{k=1}^{t-1}\mathbb{E}_k\left[\phi(x,a)\phi(x,a)^\mathsf{T}\right]\right) - 8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)} \\
&= \lambda + \lambda_{\min}\left(\sum_{k=1}^{t-1}\mathbb{E}_{x\sim\rho,a\sim\pi_k(\cdot|x)}\left[\phi(x,a)\phi(x,a)^\mathsf{T}\right]\right) - 8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)} \\
&\geq \lambda + \lambda_{\min}\left(\sum_{k=1}^{t-1}\epsilon_k\mathbb{E}_{x\sim\rho,a\sim\mathcal{U}(\mathcal{A})}\left[\phi(x,a)\phi(x,a)^\mathsf{T}\right]\right) - 8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)} \\
&= \lambda + \frac{\Gamma(\phi)}{A}\sum_{k=1}^{t-1}\epsilon_k - 8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)} \\
&\geq \lambda + \frac{\Gamma(\phi)}{A}(t-1)^{2/3} - 8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)},
\end{aligned}$$

where in the last step we used the definition of ϵ_k. We now seek a condition on t such that $8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)} \leq \frac{\Gamma(\phi)(t-1)^{2/3}}{2A}$, so that we have $\lambda_{\min}(V_{t-1}(\phi)) \geq \lambda + \frac{\Gamma(\phi)(t-1)^{2/3}}{2A}$. By the crude bound $\log(x) \leq x^\alpha/\alpha$, we have

$$8L_\phi^2\sqrt{(t-1)\log(4d_\phi(t-1)/\delta)} \leq 8L_\phi^2\sqrt{(t-1)\log(4d_\phi/\delta)} + 8L_\phi^2\sqrt{(t-1)^{1+\alpha}/\alpha}.$$

Thus, a sufficient condition is that

$$\begin{aligned}
8L_\phi^2\sqrt{(t-1)\log(4d_\phi/\delta)} \leq \frac{\Gamma(\phi)(t-1)^{2/3}}{4A} &\implies (t-1) \geq \left(\frac{32L_\phi^2 A\sqrt{\log(4d_\phi/\delta)}}{\Gamma(\phi)}\right)^6, \\
8L_\phi^2\sqrt{(t-1)^{1+\alpha}/\alpha} \leq \frac{\Gamma(\phi)(t-1)^{2/3}}{4A} &\implies (t-1) \geq \left(\frac{32L_\phi^2 A\sqrt{1/\alpha}}{\Gamma(\phi)}\right)^{\frac{6}{4-3(1+\alpha)}}.
\end{aligned}$$

Setting $\alpha = 1/12$, we have $\frac{6}{4-3(1+\alpha)} = 8$. Then, a sufficient condition is

$$t \geq z := \left(\frac{128L_\phi^2 A\sqrt{\log(4d_\phi/\delta)}}{\Gamma(\phi)}\right)^8 + 1.$$

Then,

$$\begin{aligned}
\sum_{t=1}^T \mathbb{1}\{F_t\}\frac{L_\phi}{\sqrt{\lambda_{\min}(V_{t-1}(\phi))}} \leq z\frac{L_\phi}{\sqrt{\lambda}} + \sum_{t=1}^T \frac{L_\phi}{\sqrt{\lambda + \frac{\Gamma(\phi)(t-1)^{2/3}}{2A}}} &\leq (z+1)\frac{L_\phi}{\sqrt{\lambda}} + \frac{\sqrt{2A}}{\sqrt{\Gamma(\phi)}}\sum_{t=1}^T\frac{L_\phi}{t^{1/3}} \\
&\leq (z+1)\frac{L_\phi}{\sqrt{\lambda}} + \frac{3L_\phi\sqrt{A}T^{2/3}}{\sqrt{\Gamma(\phi)}}.
\end{aligned}$$

Thus,

$$(a) \leq 2\beta_{T,\delta}(\phi)\left(\frac{L_\phi}{\sqrt{\lambda}}\left(\frac{128L_\phi^2 A\sqrt{\log(4d_\phi/\delta)}}{\Gamma(\phi)}\right)^8 + \frac{2L_\phi}{\sqrt{\lambda}} + \frac{3L_\phi\sqrt{A}T^{2/3}}{\sqrt{\Gamma(\phi)}}\right).$$

Let us bound (b). By Azuma's inequality (Lemma F.2), with probability at least $1-\delta$,

$$\begin{aligned}
(b) &\leq 2\sum_{t=1}^T \mathbb{1}\{\neg F_t\} = 2\sum_{t=1}^T\Big(\mathbb{1}\{\neg F_t\} - \mathbb{P}(\neg F_t)\Big) + 2\sum_{t=1}^T\mathbb{P}(\neg F_t) \\
&\leq 2\sqrt{T\log(2T/\delta)} + 2\sum_{t=1}^T\epsilon_t = 2\sqrt{T\log(2T/\delta)} + 2\sum_{t=1}^T\frac{1}{t^{1/3}} \leq 2\sqrt{T\log(2T/\delta)} + 3T^{2/3}.
\end{aligned}$$

Summing the bounds on (a) and (b) yields a regret bound that holds with probability at least $1-3\delta$ by the three concentration events used above. Then, the result follows by a union bound, i.e., by re-defining $\delta \to \delta/3$. □

Lemma E.4. *When using the ϵ-greedy algorithm (same conditions as in Theorem E.3), we have*

$$\tau_{\mathrm{alg}} \lesssim \frac{L_{\phi^\star}^6 (dA)^{3/2} L^3 \log(|\Phi|/\delta)^3}{\lambda^\star (\phi^\star)^3 \Delta^3}.$$

Proof. First note that, by Theorem E.3,

$$\overline{R}_{\epsilon-\mathrm{greedy}}(t, \phi, \delta_{\log_2(t)}/|\Phi|) \lesssim L_\phi \sqrt{d_\phi A} \log(t|\Phi|/\delta) t^{2/3},$$

where we kept only the higher-order dependences. Then, with similar steps as in the proof of Lemma B.14, one can easily show that τ_{alg} requires solving the inequality

$$t \lesssim \frac{L_{\phi^\star}^2}{\lambda^\star(\phi^\star)\Delta} \max_{\phi \in \Phi^\star} L_\phi \sqrt{d_\phi A} \log(|\Phi|/\delta) t^{2/3},$$

which proves the statement. □

F Auxiliary Results

F.1 Bounding the eigenvalues of the design matrices

The following result holds for any algorithm (i.e., any arm selection rule) any any representation ϕ (even non-realizable). It is an extension of Lemma 9 in [11].

Lemma F.1. *Under the assumption that the optimal policy is unique, with probability $1-\delta$, for all t and $\phi \in \Phi$,*

$$V_t(\phi) \succeq t\, \mathbb{E}_{x\sim\rho}[\phi(x,\pi^\star(x))\phi(x,\pi^\star(x))^\mathsf{T}] + \left(\lambda - L_\phi^2 S_t - 8L_\phi^2 \sqrt{t \log(4d_\phi|\Phi|t/\delta)}\right) I_{d_\phi}, \quad (22)$$

$$V_t(\phi) \preceq t\, \mathbb{E}_{x\sim\rho}[\phi(x,\pi^\star(x))\phi(x,\pi^\star(x))^\mathsf{T}] + \left(\lambda + L_\phi^2 S_t + 8L_\phi^2 \sqrt{t \log(4d_\phi|\Phi|t/\delta)}\right) I_{d_\phi}, \quad (23)$$

where $S_t := \sum_{k=1}^t \mathbb{1}\{a_k \neq \pi^\star(x_k)\}$.

Proof. The lower bound holds with probability $1-\delta/2$ by [11, Lemma 9]. Let us prove the upper bound. We have

$$\begin{aligned}
V_t(\phi) - \lambda I_{d_\phi} &= \sum_{k=1}^t \phi(x_k,a_k)\phi(x_k,a_k)^\mathsf{T} \\
&= \sum_{k=1}^t \mathbb{1}\{a_k \neq \pi^\star(x_k)\}\, \phi(x_k,a_k)\phi(x_k,a_k)^\mathsf{T} + \sum_{k=1}^t \mathbb{1}\{a_k = \pi^\star(x_k)\}\, \phi(x_k,a_k)\phi(x_k,a_k)^\mathsf{T} \\
&\preceq \sum_{k=1}^t \mathbb{1}\{a_k \neq \pi^\star(x_k)\}\, \phi(x_k,a_k)\phi(x_k,a_k)^\mathsf{T} + \sum_{k=1}^t \phi(x_k,\pi^\star(x_k))\phi(x_k,\pi^\star(x_k))^\mathsf{T} \\
&\preceq L_\phi^2 S_t I_{d_\phi} + \sum_{k=1}^t \phi(x_k,\pi^\star(x_k))\phi(x_k,\pi^\star(x_k))^\mathsf{T} \\
&\preceq L_\phi^2 S_t I_{d_\phi} + t\, \mathbb{E}_{x\sim\rho}[\phi(x,\pi^\star(x))\phi(x,\pi^\star(x))^\mathsf{T}] + 8L_\phi^2 \sqrt{t\log(4d_\phi t/\delta)} I_{d_\phi},
\end{aligned}$$

where the second-last inequality uses the boundedness of ϕ, while the last one holds with probability $1-\delta/2$ for all t by Lemma F.4 and a union bound. The result follows by a union bound on Φ and on the two sides of the inequality. □

F.2 Martingale concentration

We restate some well-known martingale concentration bounds.

Lemma F.2 (Azuma's inequality)**.** *Let $\{(Z_t, \mathcal{F}_t)\}_{t\in\mathbb{N}}$ be a martingale difference sequence such that $|Z_t| \leq a$ almost surely for all $t \in \mathbb{N}$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\sum_{k=1}^{t} Z_k\right| \leq a\sqrt{t\log(2t/\delta)}\right) \geq 1-\delta.$$

Lemma F.3 (Freedman's inequality)**.** *Let $\{(Z_t, \mathcal{F}_t)\}_{t\in\mathbb{N}}$ be a martingale difference sequence such that $|Z_t| \leq a$ almost surely for all $t \in \mathbb{N}$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\sum_{k=1}^{t} Z_k\right| \leq 2\sqrt{\sum_{k=1}^{t} \mathbb{V}_k[Z_k]\log(4t/\delta)} + 4a\log(4t/\delta)\right) \geq 1-\delta.$$

Lemma F.4 (Matrix Azuma's inequality)**.** *Let $\{X_k\}_{k=1}^{t}$ be a finite adapted sequence of symmetric matrices of dimension d, and $\{C_k\}_{k=1}^{t}$ a sequence of symmetric matrices such that for all k, $\mathbb{E}_k[X_k] = 0$ and $X_k^2 \preceq C_k^2$ almost surely. Then, with probability at least $1-\delta$,*

$$\lambda_{\max}\left(\sum_{k=1}^{t} X_k\right) \leq \sqrt{8\left\|\sum_{k=1}^{t} C_k^2\right\|\log(d/\delta)}. \tag{24}$$