Leveraging Good Representations in Linear Contextual Bandits

Matteo Papini [*] [†] [1] **Andrea Tirinzoni** [*] [1] **Marcello Restelli** [1] **Alessandro Lazaric** [2] **Matteo Pirotta** [2]

Abstract

The linear contextual bandit literature is mostly focused on the design of efficient learning algorithms for a given representation. However, a contextual bandit problem may admit multiple linear representations, each one with different characteristics that directly impact the regret of the learning algorithm. In particular, recent works showed that there exist "good" representations for which constant problem-dependent regret can be achieved. In this paper, we first provide a systematic analysis of the different definitions of "good" representations proposed in the literature. We then propose a novel selection algorithm able to adapt to the best representation in a set of M candidates. We show that the regret is indeed never worse than the regret obtained by running LINUCB on the best representation (up to a $\ln M$ factor). As a result, our algorithm achieves constant regret whenever a "good" representation is available in the set. Furthermore, we show that the algorithm may still achieve constant regret by implicitly constructing a "good" representation, even when none of the initial representations is "good". Finally, we empirically validate our theoretical findings in a number of standard contextual bandit problems.

1. Introduction

The stochastic contextual bandit is a general framework to formalize sequential decision-making problems in which at each step the learner observes a context drawn from a fixed distribution, it plays an action, and it receives a noisy reward. The goal of the learner is to maximize the reward accumulated over n rounds, and the performance is typically measured by the regret w.r.t. playing the optimal action in each context. This paradigm has found application in a large range of domains, including recommendation systems, online advertising, and clinical trials (e.g., Bouneffouf & Rish, 2019). Linear contextual bandit (Lattimore & Szepesvári, 2020) is one of the most studied instances of contextual bandit due to its efficiency and strong theoretical guarantees. In this setting, the reward for each context x and action a is assumed to be representable as the linear combination between d-dimensional features $\phi(x, a) \in \mathbb{R}^d$ and an unknown parameter $\theta^\star \in \mathbb{R}^d$. In this case, we refer to ϕ as a realizable representation. Algorithms based on the optimism-in-the-face-of-uncertainty principle such as LINUCB (Chu et al., 2011) and OFUL (Abbasi-Yadkori et al., 2011), have been proved to achieve minimax regret bound $O\big(Sd\sqrt{n}\ln(nL)\big)$ and problem-dependent regret $O\big(\frac{S^2d^2}{\Delta}\ln^2(nL)\big)$, where Δ is the minimum gap between the reward of the best and second-best action across contexts, and L and S are upper bounds to the ℓ_2-norm of the features ϕ and $\theta^\star$, respectively.

Unfortunately, the dimension d, and the norm upper bounds L and S, are not the only characteristics of a representation to have an effect on the regret and existing bounds may fail at capturing the impact of the context-action features on the performance of the algorithm. In fact, as illustrated in Fig. 1, running LINUCB with different realizable representations with same parameters d and S may lead to significantly different performance. Notably, there are "good" representations for which LINUCB achieves *constant* regret, i.e., not scaling with the horizon n. Recent works identified different conditions on the representation that can be exploited to achieve constant regret for LINUCB (Hao et al., 2020; Wu et al., 2020). Similar conditions have also been leveraged to prove other interesting learning properties, such as sublinear regret for greedy algorithms (Bastani et al., 2020), or regret guarantees for model selection between linear and multi-arm representations (Chatterji et al., 2020; Ghosh et al., 2020). While all these conditions, often referred to as *diversity conditions*, depend on how certain context-arm features span the full $\mathbb{R}^d$ space, there is no systematic analysis of their connections and of which ones can be leveraged to achieve constant regret in linear contextual bandits.

In this paper, we further investigate the concept of "good" representations in linear bandit and we provide the following contributions: **1)** We review the diversity conditions available in the literature, clarify their relationships, and discuss how they are used. We then focus on our primary

[*]Equal contribution [†]Work done while at Facebook AI Research [1]Politecnico di Milano, Milan, Italy [2]Facebook AI Research, Paris, France. Correspondence to: Matteo Papini <matteo.papini@polimi.it>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

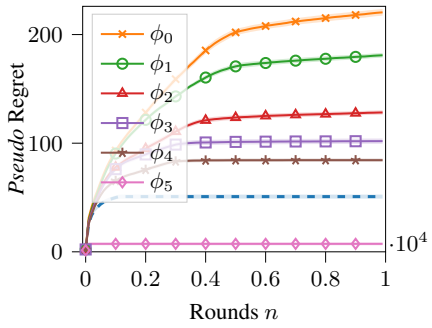

Figure 1. Regret of LINUCB with different *realizable* representations with same dimension d and parameter bound S. The dashed blue line is LEADER, our proposed representation selection algorithm. Details in App. G.1.

goal, which is to characterize the assumptions needed to achieve constant regret for LINUCB. **2)** We introduce a novel algorithm that effectively selects the best representation in a given set, thus achieving constant regret whenever at least one "good" representation is provided. **3)** Furthermore, we show that, in certain problems, the algorithm is able to combine given representations to implicitly form a "good" one, thus achieving constant problem-dependent regret even when running LINUCB on any of the representations would not. **4)** Finally, we empirically validate our theoretical findings on a number of contextual bandit problems.

Related work. The problem of selecting the best representation in a given set can be seen as a specific instance of the problem of *model selection* in bandits. In model selection, the objective is to choose the best candidate in a set of *base learning algorithms*. At each step, a *master algorithm* is responsible for selecting a base algorithm, which in turn prescribes the action to play and the reward is then provided as feedback to the base algorithms. Examples of model selection methods include adversarial masters –e.g., EXP4 (Auer et al., 2002; Maillard & Munos, 2011) and CORRAL (Agarwal et al., 2017; Pacchiano et al., 2020b)– and stochastic masters (Abbasi-Yadkori et al., 2020; Lee et al., 2020; Bibaut et al., 2020; Pacchiano et al., 2020a). For a broader discussion refer to App. A or (Pacchiano et al., 2020a, Sec. 2). Most of these algorithms achieve the regret of the best base algorithm up to a polynomial dependence on the number M of base algorithms (Agarwal et al., 2017). While existing model selection methods are general and can be applied to any type of base algorithms, [1] they may not be effective in problems with a specific structure.

An alternative approach is to design the master algorithm for a specific category of base algorithms. An instance of this case is the representation-selection problem, where the base algorithms only differ by the representation used to estimate the reward. Foster et al. (2019) and Ghosh et al. (2020) consider a set of nested representations, where the best representation is the one with the smallest dimensionality for which the reward is realizable. Finally, Chatterji et al. (2020) focus on the problem of selecting between a linear and a multi-armed bandit representation. In this paper, we consider an alternative representation-selection problem in linear contextual bandits, where the objective is to exploit constant-regret "good" representations. Differently from our work, Lattimore et al. (2020) say that a linear representation is "good" if it has a low *misspecification* (i.e., it represents the reward up to a small approximation error), while we focus on *realizable* representations for which LINUCB achieves constant-regret.

[1] Most of existing methods only require prior knowledge of the regret of the optimal base algorithm or a bound on the regret of all base algorithms. CORRAL also requires the base algorithms to satisfy certain stability conditions.

2. Preliminaries

We consider the stochastic contextual bandit problem (*contextual problem* for short) with context space $\mathcal{X}$ and finite action set $\mathcal{A} = [K] = \{1, \dots, K\}$. At each round $t \geq 1$, the learner observes a context x_t sampled i.i.d. from a distribution ρ over $\mathcal{X}$, it selects an arm $a_t \in [K]$ and it receives a reward $y_t = \mu(x_t, a_t) + \eta_t$ where η_t is a σ-subgaussian noise. The learner's objective is to minimize the pseudo-regret $R_n = \sum_{t=1}^n \mu^\star(x_t) - \mu(x_t, a_t)$ for any $n > 0$, where $\mu^\star(x_t) := \max_{a \in [K]} \mu(x_t, a)$. We define the minimum gap as $\Delta = \inf_{x \in \mathcal{X}: \rho(x) > 0, a \in [K], \Delta(x,a) > 0} \{\Delta(x, a)\}$ where $\Delta(x, a) = \mu^\star(x) - \mu(x, a)$. A *realizable* d_ϕ-dimensional linear representation is a feature map $\phi : \mathcal{X} \times [K] \to \mathbb{R}^{d_\phi}$ for which there exists an unknown parameter vector $\theta^\star_\phi \in \mathbb{R}^{d_\phi}$ such that $\mu(x, a) = \langle \phi(x, a), \theta^\star_\phi \rangle$. When a realizable linear representation is available, the problem is called (stochastic) linear contextual bandit and can be solved using, among others, optimistic algorithms like LINUCB (Chu et al., 2011) or OFUL (Abbasi-Yadkori et al., 2011).

Given a realizable representation ϕ, at each round t, LINUCB builds an estimate $\theta_{t\phi}$ of $\theta^\star_\phi$ by ridge regression using the observed data. Denote by $V_{t\phi} = \lambda I_{d_\phi} + \sum_{k=1}^{t-1} \phi(x_k, a_k) \phi(x_k, a_k)^\mathsf{T}$ the $(\lambda > 0)$-regularized design matrix at round t, then $\theta_{t\phi} = V_{t\phi}^{-1} \sum_{k=1}^{t-1} \phi(x_k, a_k) y_k$. Assuming that $\|\theta^\star_\phi\|_2 \leq S_\phi$ and $\sup_{x,a} \|\phi(x, a)\|_2 \leq L_\phi$, LINUCB builds a confidence ellipsoid $\mathcal{C}_{t\phi}(\delta) = \left\{\theta \in \mathbb{R}^{d_\phi} : \left\|\theta_{t\phi} - \theta\right\|_{V_{t\phi}} \leq \beta_{t\phi}(\delta)\right\}$. As shown in (Abbasi-Yadkori et al., 2011, Thm. 1), when

$$\beta_{t\phi}(\delta) := \sigma \sqrt{2 \ln \left(\frac{\det(V_{t\phi})^{1/2} \det(\lambda I_{d_\phi})^{-1/2}}{\delta} \right)} + \sqrt{\lambda} S_\phi,$$

then $\mathbb{P}(\forall t \geq 1, \theta^\star_\phi \in \mathcal{C}_{t\phi}(\delta)) \geq 1 - \delta$. At each step t, LINUCB plays the action with the highest upper-confidence bound $a_t = \operatorname{argmax}_{a \in [K]} \max_{\theta \in \mathcal{C}_{t\phi}(\delta)} \langle \phi(x_t, a), \theta \rangle$, and it is shown to achieve a regret bounded as reported in the

following proposition.

Proposition 1 (Abbasi-Yadkori et al., 2011, Thm. 3, 4)**.** *For any linear contextual bandit problem with d_ϕ-dimensional features, $\sup_{x,a} \|\phi(x,a)\|_2 \le L_\phi$, an unknown parameter vector $\|\theta^\star_\phi\|_2 \le S_\phi$, with probability at least $1-\delta$,* LinUCB *suffers regret $R_n = O\big(S_\phi d_\phi \sqrt{n} \ln(nL_\phi/\delta)\big)$. Furthermore, if the problem has a minimum gap $\Delta > 0$, then the regret is bounded as*[2] $R_n = O\left(\frac{S_\phi^2 d_\phi^2}{\Delta} \ln^2(nL_\phi/\delta)\right)$.

In the rest of the paper, we assume w.l.o.g. that all terms λ, $\Delta_{\max} = \max_{x,a} \Delta(x,a)$, S_ϕ, σ are larger than 1 to simplify the expression of the bounds.

3. Diversity Conditions

Several assumptions, usually referred to as *diversity conditions*, have been proposed to define linear bandit problems with specific properties that can be leveraged to derive improved learning results. While only a few of them were actually leveraged to derive constant regret guarantees for LinUCB (others have been used to prove e.g., sub-linear regret for the greedy algorithm, or regret guarantees for model selection algorithms), they all rely on very similar conditions on how certain context-action features span the full $\mathbb{R}^{d_\phi}$ space. In this section, we provide a thorough review of these assumptions, their connections, and how they are used in the literature. As diversity conditions are getting more widely used in bandit literature, we believe this review may be of independent interest. Sect. 4 will then specifically focus on the notion of *good* representation for LinUCB.

We first introduce additional notation. For a realizable representation ϕ, let $\phi^\star(x) := \phi(x, a^\star_x)$, where $a^\star_x \in \operatorname{argmax}_{a\in[K]} \mu(x,a)$ is an optimal action, be the vector of *optimal features* for context x. In the following we make the assumption that $\phi^\star(x)$ is unique. Also, let $\mathcal{X}^\star(a) = \{x \in \mathcal{X} : \mu(x,a) = \mu^\star(x)\}$ denote the set of contexts where a is optimal. Finally, for any matrix A, we denote by $\lambda_{\min}(A)$ its minimum eigenvalue. For any contextual problem with reward μ and context distribution ρ, the diversity conditions introduced in the literature are summarized in Tab. 2 together with how they were leveraged to obtain regret bounds in different settings.[3]

We first notice that all conditions refer to the smallest eigenvalue of a design matrix constructed on specific context-action features. In other words, diversity conditions require certain features to span the full $\mathbb{R}^{d_\phi}$ space. The non-redundancy condition is a common technical assumption (e.g., Foster et al., 2019) and it simply defines a problem whose dimensionality cannot be reduced without losing information. Assuming the context distribution ρ is full support, BBK and CMB are structural properties of the representation that are independent from the reward. For example, BBK requires that, for each action, there must be feature vectors lying in all orthants of $\mathbb{R}^{d_\phi}$. In the case of finite contexts, this implies there must be at least 2^{d_ϕ} contexts. WYS and HLS involve the notion of reward optimality. In particular, WYS requires that all actions are optimal for at least a context (in the continuous case, for a non-negligible set of contexts), while HLS only focuses on optimal actions.

We now review how these conditions (or variations thereof) were applied in the literature. CMB is a rather strong condition that requires the features associated with each individual action to span the whole $\mathbb{R}^{d_\phi}$ space. Chatterji et al. (2020) leverage a CMB-like assumption to prove regret bounds for OSOM, a model-selection algorithm that unifies multi-armed and linear contextual bandits. More precisely, they consider a variation of CMB, where the context distribution induces stochastic feature vectors for each action that are independent and centered. The same condition was adopted by Ghosh et al. (2020) to study representation-selection problems and derive algorithms able to adapt to the (unknown) norm of $\theta^\star_\phi$ or select the smallest realizable representation in a set of nested representations. Bastani et al. (2020, Assumption 3) introduced a condition similar to BBK for the *disjoint-parameter* case. In their setting, they prove that a non-explorative greedy algorithm achieves $O(\ln(n))$ problem-dependent regret in linear contextual bandits (with 2 actions).[4] Hao et al. (2020, Theorem 3.9) showed that HLS representations can be leveraged to prove constant problem-dependent regret for LinUCB in the shared-parameter case. Concurrently, Wu et al. (2020) showed that, under WYS, LinUCB achieves constant expected regret in the disjoint-parameter case. A WYS-like condition was also used by Bastani et al. (2020, Assumption 4) to extend the result of sublinear regret for the greedy algorithm to more than two actions. The relationship between all these conditions is derived in the following lemma.

Lemma 1. *For any contextual problem with reward μ and context distribution ρ, let ϕ be a realizable linear representation. The relationship between the diversity conditions in Tab. 2 is summarized in Fig. 3, where each inclusion is in a strict sense and each intersection is non-empty.*

This lemma reveals non-trivial connections between the diversity conditions, better understood through the examples provided in the proof (see App. B.1). BBK is indeed

[2] The logarithmic bound reported in Prop. 1 is slightly different than the one in (Abbasi-Yadkori et al., 2011) since we do not assume that the optimal feature is unique.

[3] In some cases, we adapted conditions originally defined in the disjoint-parameter setting, where features only depend on the context (i.e., $\phi(x)$) and the unknown parameter $\theta^\star_a$ is different for each action a, to the shared-parameter setting (i.e., where features are functions of both contexts and actions) introduced in Sect. 2.

[4] Whether this is enough for the optimality of the greedy algorithm in the shared-parameter setting is an interesting problem, but it is beyond the scope of this paper.

Name	Definition	Application
Non-redundant	$\lambda_{\min}\left(1/K \sum_{a\in[K]} \mathbb{E}_{x\sim\rho}\left[\phi(x,a)\phi(x,a)^\mathsf{T}\right]\right) > 0$	
CMB	$\forall a,\ \lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\phi(x,a)\phi(x,a)^\mathsf{T}\right]\right) > 0$	*Model selection*
BBK	$\forall a, \boldsymbol{u}\in\mathbb{R}^d,$ $\lambda_{\min}\left(\mathbb{E}_x\left[\phi(x,a)\phi(x,a)^\mathsf{T}\mathbb{1}\left\{\phi(x,a)^\mathsf{T}\boldsymbol{u}\geq 0\right\}\right]\right) > 0$	*Logarithmic regret for greedy*
HLS	$\lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\phi^\star(x)\phi^\star(x)^\mathsf{T}\right]\right) > 0$	*Constant regret for* LinUCB
WYS	$\forall a,\ \lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\phi(x,a)\phi(x,a)^\mathsf{T}\mathbb{1}\left\{x\in\mathcal{X}^\star(a)\right\}\right]\right) > 0$	*Constant regret for* LinUCB

Figure 2. Diversity conditions proposed in the literature adapted to the shared-parameter setting. The names refer to the authors who first introduced similar conditions.

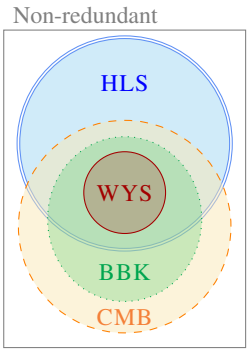

Figure 3. Categorization of diversity conditions.

stronger than CMB, and thus it is sufficient for the model selection results by Chatterji et al. (2020). By superficially examining their definitions, CMB may appear stronger than HLS, but the two properties are actually non-comparable, as there are representations that satisfy one condition but not the other. The implications of Fig. 3 on constant-regret guarantees are particularly relevant for our purposes. There are representations that satisfy BBK or CMB and are neither HLS nor WYS and thus may not enable constant regret for LinUCB. We notice that WYS is a stronger condition than HLS. Although WYS may be necessary for LinUCB to achieve constant regret in the disjoint-parameter case, HLS is sufficient for the shared-parameter case we consider in this paper. For this reason, in the following section we adopt HLS to define *good* representations for LinUCB and provide a more complete characterization.

4. *Good* Representations for Constant Regret

The HLS condition was introduced by Hao et al. (2020), who provided a first analysis of its properties. In this section, we complement those results by providing a complete proof of a constant regret bound, a proof of the fact that HLS is actually necessary for constant regret, and a novel characterization of the existence of HLS representations. In the following we define $\lambda_{\phi,\text{HLS}} := \lambda_{\min}\left(\mathbb{E}_{x\sim\rho}\left[\phi^\star(x)\phi^\star(x)^\mathsf{T}\right]\right)$, which is strictly positive for HLS representations.

4.1. Constant Regret Bound

We begin by deriving a constant problem-dependent regret bound for LinUCB under the HLS condition.

Lemma 2. *Consider a contextual bandit problem with realizable linear representation ϕ satisfying the* HLS *condition (see Tab. 2). Assume $\Delta > 0$, $\max_{x,a}\|\phi(x,a)\|_2 \leq L$ and $\|\theta^\star_\phi\|_2 \leq S$. Then, with probability at least $1-2\delta$, the regret of* OFUL *after $n \geq 1$ steps is at most*

$$R_n \leq \frac{32\lambda\Delta_{\max}^2 S_\phi^2\sigma^2}{\Delta}\left(2\ln\left(\frac{1}{\delta}\right) + d_\phi \ln\left(1+\frac{\tau_\phi L_\phi^2}{\lambda d_\phi}\right)\right)^2,$$

where $\Delta_{\max} = \max_{x,a}\Delta(x,a)$ is the maximum gap and

$$\tau_\phi \leq \max\left\{\frac{384^2 d_\phi^2 L_\phi^2 S_\phi^2\sigma^2\lambda}{\lambda_{\phi,\text{HLS}}\Delta^2}\ln^2\left(\frac{64 d_\phi^2 L_\phi^3\sigma S_\phi\sqrt{\lambda}}{\sqrt{\lambda_{\phi,\text{HLS}}}\Delta\delta}\right), \frac{768 L_\phi^4}{\lambda_{\phi,\text{HLS}}^2}\ln\left(\frac{512 d_\phi L_\phi^4}{\delta\lambda_{\phi,\text{HLS}}^2}\right)\right\}.$$

We first notice that τ_ϕ is independent from the horizon n, thus making the previous bound a constant only depending on the problem formulation (i.e., gap Δ, norms L_ϕ and S_ϕ) and the value $\lambda_{\phi,\text{HLS}}$ which measures "how much" the representation ϕ satisfies the HLS condition. Furthermore, one can always take the minimum between the constant regret in Lem. 2 and any other valid regret bound for OFUL (e.g., $O(\log(n)/\Delta)$), which may be tighter for small values of n. While Lem. 2 provides high-probability guarantees, we can easily derive a constant expected-regret bound by running LinUCB with a decreasing schedule for δ (e.g., $\delta_t \propto 1/t^3$) and with a slightly different proof (see App. C and the proof sketch below).

Proof sketch (full proof in App. C). Following Hao et al. (2020), the idea is to show that the instantaneous regret $r_{t+1} = \langle\theta^\star, \phi^\star(x_{t+1}) - \phi(x_{t+1}, a_{t+1})\rangle$ is zero for sufficiently large (but constant) time t. By using the standard regret analysis, we have

$$r_{t+1} \leq 2\beta_{t+1}(\delta)\|\phi(x_{t+1},a_{t+1})\|_{V_{t+1}^{-1}} \leq \frac{2L\beta_{t+1}(\delta)}{\sqrt{\lambda_{\min}(V_{t+1})}}.$$

Given the minimum-gap assumption, a sufficient condition for $r_{t+1} = 0$ is that the previous upper bound is smaller than Δ, which gives $\lambda_{\min}(V_{t+1}) > 4L^2\beta_{t+1}^2(\delta)/\Delta^2$. Since $\Delta > 0$, the problem-dependent regret bound in Prop. 1

holds, and the number of pulls to suboptimal arms up to time t is bounded by $g_t(\delta) = O\big((d\ln(t/\delta)/\Delta)^2\big)$. Hence, the optimal arms are pulled linearly often and, by leveraging the HLS assumption, we are able to show that the minimum eigenvalue of the design matrix grows linearly in time as

$$\lambda_{\min}(V_{t+1}) \geq \lambda + t\lambda_{\text{HLS}} - 8L^2\sqrt{t\ln\left(\frac{2dt}{\delta}\right)} - L^2 g_t(\delta).$$

By relating the last two equations, we obtain an inequality of the form $t\lambda_{\text{HLS}} - o(t) > o(t)$. If we define $\tau < \infty$ as the smallest (deterministic) time such that this inequality holds, we have that after τ the immediate regret is zero, thus concluding the proof. Note that, if we wanted to bound the expected regret, we could set $\delta_t \propto 1/t^3$ and the above inequality would still be of the same form (although the resulting τ would be slightly different).

Comparison with existing bounds. Hao et al. (2020, Theorem 3.9) prove that LINUCB with HLS representations achieves $\limsup_{n\to\infty} R_n < \infty$, without characterizing the time at which the regret vanishes. Instead, our Lem. 2 provides an explicit problem-dependent constant regret bound. Wu et al. (2020, Theorem 2) consider the disjoint-parameter setting and rely on the WYS condition. While they indeed prove a constant regret result, their bound depends on the the minimum probability of observing a context (or, in the continuous case, a properly defined meta-context). This reflects the general tendency, in previous works, to frame diversity conditions simply as a property of the context distribution ρ. On the other hand, our characterization of τ in terms of $\lambda_{\phi,\text{HLS}}$ (Lem. 2) allows relating the regret to the "goodness" of the representation ϕ for the problem at hand.

4.2. Removing the Minimum-Gap Assumption

Constant-regret bounds for LINUCB rely on a minimum-gap assumption ($\Delta > 0$). In this section we show that LINUCB can still benefit from HLS representations when $\Delta = 0$, but a margin condition holds (e.g., Rigollet & Zeevi, 2010; Reeve et al., 2018). Intuitively, we require that the probability of observing a context x decays proportionally to its minimum gap $\Delta(x) = \min_a \Delta(x,a)$.

Assumption 1 (Margin condition)**.** *There exists $C, \alpha > 2$ such that for all $\epsilon > 0$: $\rho\big(\{x \in \mathcal{X} : \Delta(x) \leq \epsilon\}\big) \leq C\epsilon^{\alpha}$.*

The following theorem provides a problem-dependent regret bound for LINUCB under this margin assumption.

Theorem 1. *Consider a linear contextual bandit problem satisfying the margin condition (Asm. 1). Assume $\max_{x,a} \|\phi(x,a)\|_2 \leq L_\phi$ and $\|\theta^\star_\phi\|_2 \leq S_\phi$. Then, given a representation ϕ, with probability at least $1-3\delta$, the regret of* OFUL *after $n \geq 1$ steps is at most*

$$R_n \leq O\left(\left(\lambda(\Delta_{\max} S_\phi \sigma d_\phi)^2 n^{1/\alpha} + \sqrt{C d_\phi}\right)\ln^2(L_\phi n/\delta)\right).$$

When ϕ is HLS *($\lambda_{\phi,\text{HLS}} > 0$), let $\tau_\phi \propto (\lambda_{\phi,\text{HLS}})^{\frac{\alpha}{2-\alpha}}$, then*

$$R_n \leq O\Big(\Delta_{\max}\tau_\phi + \sqrt{C d_\phi}\ln^2(L_\phi n/\delta)\Big).$$

We first notice that in general, LINUCB suffers $\widetilde{O}(n^{1/\alpha})$ regret, which can be significantly larger than in the minimum-gap case. On the other hand, with HLS representations, LINUCB achieves logarithmic regret, regardless of the value of α. The intuition is that, when the HLS condition holds, the algorithm collects sufficient information about $\theta^\star_\phi$ by pulling the optimal arms in rounds with large minimum gap, which occur with high probability by the margin condition. This yields at most constant regret in such rounds (first term above), while it can be shown that the regret in steps when the minimum gap is very small is at most logarithmic (second term above).

4.3. Further Analysis of the HLS Condition

While Lem. 2 shows that HLS is sufficient for achieving constant regret, the following proposition shows that it is also necessary. While this property was first mentioned by Hao et al. (2020) as a remark in a footnote, we provide a formal proof in App. C.5.

Proposition 2. *For any contextual problem with finite contexts, full-support context distribution, and given a non-redundant realizable representation ϕ,* LINUCB *achieves sub-logarithmic regret if and only if ϕ satisfies the* HLS *condition.*

As already observed in Section 4, the HLS condition can be equivalently expressed as:[5]

$$\text{span}\{\phi^\star(x) \mid x \in \mathcal{X}\} = \mathbb{R}^d,$$

i.e., optimal features must span the whole d-dimensional Euclidean space, where d is the dimension of ϕ. If we admit redundant representations, a weaker condition is sufficient to achieve constant regret:

$$\text{span}\{\phi^\star(x) \mid x \in \mathcal{X}\} = \text{span}\{\phi(x,a) \mid x \in \mathcal{X}, a \in \mathcal{A}\},$$

i.e., optimal features must span the whole feature space, which may be a subspace of $\mathbb{R}^d$ in general. We prove that this *weak HLS* condition is sufficient for LINUCB to achieve constant regret as Corollary 1 in App. E.2. This also shows that constant-regret guarantees are preserved by adding redundant features to an HLS representation.

Finally, we derive the following important existence result.

Lemma 3. *For any contextual bandit problem with optimal reward [6] $\mu^\star(x) \neq 0$ for all $x \in \mathcal{X}$, that has either* i) *a*

[5] That is assuming the context distribution is full-support. Otherwise, it is enough to replace $\mathcal{X}$ with the support of the context distribution $\text{supp}(\rho)$.

[6] This condition is technical and it can be easily relaxed.

finite context set with at least d contexts with nonzero probability, or ii) *a Borel context space and a non-degenerate context distribution*[7]*, for any dimension $d \geq 1$, there exists an infinite number of d-dimensional realizable* HLS *representations.*

This result crucially shows that the HLS condition is "robust", since in any contextual problem, it is possible to construct an infinite number of representations satisfying the HLS condition. In App. B.2, we indeed provide an oracle procedure for constructing an HLS representation. This result also supports the starting point of next section, where we assume that a learner is provided with a set of representations that may contain at least a "good" representation, i.e., an HLS representation.

5. Representation Selection

In this section, we study the problem of *representation selection* in linear bandits. We consider a linear contextual problem with reward μ and context distribution ρ. Given a set of M realizable linear representations $\{\phi_i : \mathcal{X} \times [K] \to \mathbb{R}^{d_i}\}$, the objective is to design a learning algorithm able to perform as well as the best representation, and thus achieve constant regret when a "good" representation is available. As usual, we assume $\theta_i^\star \in \mathbb{R}^{d_i}$ is unknown, but the algorithm is provided with a bound on the parameter and feature norms of the different representations.

5.1. The LEADER Algorithm

We introduce LEADER (*Linear rEpresentation bAnDit mixER*), see Alg. 1. At each round t, LEADER builds an estimate θ_{ti} of the unknown parameter $\theta_i^\star$ of each representation ϕ_i.[8] These estimates are by nature off-policy, and thus *all* the samples $(x_l, a_l, y_l)_{l<t}$ can be used to solve *all* ridge regression problems. For each ϕ_i, define $V_{ti} = \lambda I_{d_i} + \sum_{l=1}^{t-1} \phi_i(x_l, a_l)\phi_i(x_l, a_l)^\mathsf{T}$, θ_{ti} and $\mathcal{C}_{ti}(\delta/M)$ as in Sec. 2. Since all the representations are realizable, we have that $\mathbb{P}\left(\forall i \in [M], \theta_i^\star \in \mathcal{C}_{ti}(\delta/M)\right) \geq 1 - \delta$. As a consequence, for each representation ϕ_i we can build an upper-confidence bound to the reward such that, $\forall x \in \mathcal{X}, a \in \mathcal{A}$, with high probability

$$\mu(x, a) \leq \max_{\theta \in \mathcal{C}_{ti}(\delta/M)} \langle \phi_i(x, a), \theta \rangle := U_{ti}(x, a). \qquad (1)$$

Given this, LEADER uses the tightest available upper-confidence bound to evaluate each action and then it selects the one with the largest value, i.e.,

$$a_t \in \operatorname*{argmax}_{a \in [K]} \min_{i \in [M]} \{U_{ti}(x_t, a)\}. \qquad (2)$$

[7]For instance, if $\mathcal{X} = \mathbb{R}^m$ and the context distribution must have positive variance in all directions.

[8]We use the subscript $i \in [M]$ instead of ϕ_i to denote quantities related to representation ϕ_i.

Let $i_t = \operatorname{argmin}_{i \in [M]} \{U_{ti}(x_t, a_t)\}$ be the representation associated to the pulled arm a_t. Interestingly, despite a_t being optimistic, in general it may not correspond to the optimistic action of representation ϕ_{i_t}, i.e., $a_t \notin \operatorname{argmax}_a \{U_{t,i_t}(x_t, a)\}$. If a representation provides an estimate that is good along the direction associated to a context-action pair, but possibly very uncertain on other actions, LEADER is able to leverage this key feature to reduce the overall uncertainty and achieve a tighter optimism. Space and time complexity of LEADER scales linearly in the number of representations, although the updates for each representation could be carried out in parallel.

Regret bound. For ease of presentation, we assume a non-zero minimum gap ($\Delta > 0$). The analysis can be generalized to $\Delta = 0$ as done in Sec. 4.2. Thm. 2 establishes the regret guarantee of LEADER (Alg. 1).

Theorem 2. *Consider a contextual bandit problem with reward μ, context distribution ρ and $\Delta > 0$. Let (ϕ_i) be a set of M linearly realizable representations such that $\max_{x,a} \|\phi_i(x, a)\|_2 \leq L_i$ and $\|\theta_i^\star\|_i \leq S_i$. Then, for any $n \geq 1$, with probability $1 - 2\delta$,* LEADER *suffers a regret*

$$R_n \leq \min_{i \in [M]} \left\{ \frac{32\lambda\Delta_{\max}^2 S_i^2 \sigma^2}{\Delta} \times \right.$$
$$\left. \times \left(2\ln\left(\frac{M}{\delta}\right) + d_i \ln\left(1 + \frac{\min\{\tau_i, n\} L_i^2}{\lambda d_i}\right)\right)^2 \right\}$$

where $\tau_i \propto (\lambda_{i,\mathrm{HLS}}\Delta)^{-2}$ if ϕ_i is HLS *and $\tau_i = +\infty$ otherwise.*

This shows that the problem-dependent regret bound of LEADER is not worse than the one of the best representation (see Prop. 1), up to a $\ln M$ factor. This means that the cost of representation selection is almost negligible. Furthermore, Thm. 2 shows that LEADER not only achieves a constant regret bound when an HLS representation is available, but this bound scales as the one of the *best* HLS representation. In fact, notice that the "quality" of an HLS representation does not depend only on known quantities such as d_i, L_i, S_i, but crucially on HLS eigenvalue $\lambda_{i,\mathrm{HLS}}$, which is usually not known in advance, as it depends on the features of the optimal arms.

5.2. Combining Representations

In the previous section, we have shown that LEADER can perform as well as the best representation in the set. However, by inspecting the action selection rule (Eq. 2), we notice that, to evaluate the reward of an action in the current context, LEADER selects the representation with the smallest uncertainty, thus potentially using different representations for different context-action pairs. This leads to the question: *can* LEADER *do better than the best representation in the set?*

Algorithm 1 LEADER Algorithm

Input: representations $(\phi_i)_{i\in[M]}$ with values $(L_i, S_i)_{i\in[M]}$, regularization factor $\lambda \geq 1$, confidence level $\delta \in (0,1)$.
Initialize $V_{1i} = \lambda I_{d_i}$, $\theta_{1i} = 0_{d_i}$ for each $i \in [M]$
for $t = 1, \dots$ **do**
  Observe context x_t
  Pull action $a_t \in \operatorname{argmax}_{a\in[K]} \min_{i\in[M]}\{U_{ti}(x_t, a)\}$
  Observe reward r_t and, for each $i \in [M]$, set
  $V_{t+1,i} = V_{ti} + \phi_i(x_t, a_t)\phi_i(x_t, a_t)^\intercal$ and
  $\theta_{t+1,i} = V_{t+1,i}^{-1} \sum_{l=1}^{t} \phi_i(x_l, a_l) r_l$
end for

We show that, in certain cases, LEADER is able to combine representations and achieve constant regret when none of the individual representations would. The intuition is that a subset of "locally good" representations can be combined to recover a condition similar to HLS. This property is formally stated in the following definition.

Definition 1 (Mixing HLS)**.** *Consider a linear contextual problem with reward μ and context distribution ρ, and a set of M realizable linear representations $\phi_1, \dots, \phi_M$. Define $M_i = \mathbb{E}_{x\sim\rho}\left[\phi_i^\star(x)\phi_i^\star(x)^\intercal\right]$ and let $Z_i = \{(x,a) \in \mathcal{X} \times \mathcal{A} \mid \phi_i(s,a) \in \mathrm{Im}(M_i)\}$ be the set of context-action pairs whose features belong to the column space of M_i, i.e., that lie in the span of optimal features. We say that the set (ϕ_i) satisfies the mixed-*HLS *condition if $\mathcal{X} \times \mathcal{A} \subseteq \bigcup_{i=1}^{M} Z_i$.*

Let $\lambda_i^+ = \lambda_{\min}^+(M_i)$ be the minimum *nonzero* eigenvalue of M_i. Intuitively, the previous condition relies on the observation that every representation satisfies a "restricted" HLS condition on the context-action pairs (x,a) whose features $\phi_i(x,a)$ are spanned by optimal features $\phi^\star(x)$. In this case, the characterizing eigenvalue is λ_i^+, instead of the smallest eigenvalue $\lambda_{i,\mathrm{HLS}}$ (which may be zero). If every context-action pair is in the restriction Z_i of some representation, we have the mixed-HLS property. In particular, if representation i is HLS, $\lambda_i^+ = \lambda_{i,\mathrm{HLS}}$ and $Z_i = \mathcal{S} \times \mathcal{A}$. So, HLS is a special case of mixed-HLS. In App. E.2, we provide simple examples of sets of representations satisfying Def. 1. Note that, strictly speaking, there is not a single "mixed representation" solving the whole problem. Even defining one would be problematic since each representation may have a different parameter and even a different dimension. Instead, each representation "specializes" on a different portion of the context-action space. If together they cover the whole space, the benefits of HLS are recovered, as illustrated in the following theorem.

Theorem 3. *Consider a stochastic bandit problem with reward μ, context distribution ρ and $\Delta > 0$. Let (ϕ_i) be a set of M realizable linear representations satisfying the mixed-*HLS *property in Def. 1. Then, with probability at least $1 - 2\delta$, there exists a time $\tau < \infty$ independent from n such that, for any $n \geq 1$, the pseudo-regret of* LEADER *is bounded as*

$$R_n \leq \min_{i\in[M]} \left\{ \frac{32\lambda\Delta_{\max}^2 S_i^2 \sigma^2}{\Delta} \times \right.$$
$$\left. \times \left(2\ln\left(\frac{M}{\delta}\right) + d_i \ln\left(1 + \frac{\tau L_i^2}{\lambda d_i}\right)\right)^2 \right\}.$$

First, note that we are still scaling with the characteristics of the best representation in the set (i.e., d_i, L_i and S_i). However, the time τ to constant regret is a global value rather than being different for each representation. This highlights that mixed-HLS is a global property of the set of representations rather than being individual as before. In particular, whenever no representation is (globally) HLS (i.e., $\lambda_{i,\mathrm{HLS}} = 0$ for all ϕ_i), we can show that in the worst case τ scales as $(\min_i \lambda_i^+)^{-2}$. In practice, we may expect LEADER to even behave better than that since **i)** not all the representations may contribute actively to the mixed-HLS condition; and **ii)** multiple representations may cover the same region of the context-action space. In the latter case, since LEADER leverages all the representations at once, its regret would rather scale with the largest minimum non-zero eigenvalue λ_i^+ among all the representations covering such region. We refer to App. E.2 for a more complete discussion.

5.3. Discussion

Most of the model selection algorithms reviewed in the introduction could be readily applied to select the best representation for LINUCB. However, the generality of their objective comes with several shortcomings when instantiated in our specific problem (see App. A for a more detailed comparison). First, model selection methods achieve the performance of the best algorithm, up to a polynomial dependence on the number M of models. This already makes them a weaker choice compared to LEADER, which, by leveraging the specific structure of the problem, suffers only a logarithmic dependence on M. Second, model selection algorithms are often studied in a worst-case analysis, which reveals a high cost for adaptation. For instance, corralling algorithms (Agarwal et al., 2017; Pacchiano et al., 2020b) pay an extra $\sqrt{n}$ regret, which would make them unsuitable to target the constant regret of good representations. Similar costs are common to other approaches (Abbasi-Yadkori et al., 2020; Pacchiano et al., 2020a). It is unclear whether a problem-dependent analysis can be carried out and whether this could shave off such dependence. Third, these algorithms are generally designed to adapt to a specific best base algorithm. At the best of our knowledge, there is no evidence that model selection methods could combine algorithms to achieve better performance than the best candidate, a behavior that we proved for LEADER in our setting.

On the other hand, model selection algorithms effec-

tively deal with non-realizable representations in certain cases (e.g., Foster et al., 2020; Abbasi-Yadkori et al., 2020; Pacchiano et al., 2020a), while LEADER is limited to the realizable case. While a complete study of the model misspecification case is beyond the scope of this paper, in App. F, we discuss how a variation of the approach presented in (Agarwal et al., 2012b) could be paired to LEADER to discard misspecified representations and possibly recover the properties of "good" representations.

6. Experiments

In this section, we report experimental results on two synthetic and one dataset-based problems. For each problem, we evaluate the behavior of LEADER with LINUCB and model selection algorithms: EXP4.IX (Neu, 2015), CORRAL and EXP3.P in the stochastic version by Pacchiano et al. (2020b) and Regret Balancing with and without elimination (REGBALELIM and REGBAL) (Abbasi-Yadkori et al., 2020; Pacchiano et al., 2020a). See App. G for a detailed discussion and additional experiments. All results are averaged over 20 independent runs, with shaded areas corresponding to 2 standard deviations. We always set the parameters to $\lambda = 1$, $\delta = 0.01$, and $\sigma = 0.3$. All the representations we consider are normalized to have $\|\theta_i^\star\| = 1$.

Synthetic Problems. We define a randomly-generated contextual bandit problem, for which we construct sets of realizable linear representations with different properties (see App. G.1 for details). The purpose of these experiments is twofold: to show the different behavior of LINUCB with different representations, and to evaluate the ability of LEADER of selecting and mixing representations.

Varying dimension. We construct six representations of varying dimension from 2 up to 6. Of the two representations of dimension $d = 6$, one is HLS. Fig. 4(*left*) shows that in this case, LINUCB with the HLS representation outperforms any non-HLS representation, even if they have smaller dimension. This property is inherited by LEADER, which performs better than LINUCB with non-HLS representations even of much smaller dimension 2.

Mixing representations. We construct six representations of the same dimension $d = 6$, *none of which* is HLS. However, they are constructed so that together they satisfy the weaker mixed-HLS assumption (Def. 1). Fig. 4(*middle left*) shows that, as predicted by Thm. 3, LEADER leverages different representations in different context-action regions and it thus performs significantly better than any LINUCB using non-HLS representations. The superiority of LEADER w.r.t. the model-selection baselines is evident in this case (Fig. 4(*middle right*)), since only LEADER is able to mix representations, whereas model-selection algorithms target the best in a set of "bad" representations. Additional experiments in App. G confirm that LEADER consistently outperforms all model-selection algorithms.

Jester Dataset. In the last experiment, we extract multiple linear representations from the Jester dataset (Goldberg et al., 2001), which consists of joke ratings in a continuous range from -10 to 10 for a total of 100 jokes and 73421 users. For a subset of 40 jokes and 19181 users rating all these 40 jokes, we build a linear contextual problem as follows. First, we fit a 32×32 neural network to predict the ratings from features extracted via a low-rank factorization of the full matrix. Then, we take the last layer of the network as our "ground truth" linear model and fit multiple smaller networks to clone its predictions, while making sure that the resulting misspecification is small. We thus obtain 7 representations with different dimensions among which, interestingly, we find that 6 are HLS. Figure 4(*right*) reports the comparison between LEADER using all representations and LINUCB with each single representation on a log-scale. Notably, the ability of LEADER to mix representations makes it perform better than the best candidate, while transitioning to constant regret much sooner. Finally, the fact that HLS representations arise so "naturally" raises the question of whether this is a more general pattern in context-action features learned from data.

Last.fm dataset. In App. F we study a variant of LEADER that is able to handle misspecified representations, and we test it on the Last.fm music-recommendation dataset (Cantador et al., 2011). See App. F.4 for details.

7. Conclusion

We provided a complete characterization of "good" realizable representations for LINUCB, ranging from existence to a sufficient and necessary condition to achieve problem-dependent constant regret. We introduced LEADER, a novel algorithm that, given a set of realizable linear representations, is able to adapt to the best one and even leverage their combination to achieve constant regret under the milder mixed-HLS condition. While we have focused on LINUCB, other algorithms (e.g., LinTS (Abeille & Lazaric, 2017)) as well as other settings (e.g., low-rank RL (Jin et al., 2020)) may also benefit from HLS-like assumptions. We have mentioned an approach for eliminating misspecified representations, but a non-trivial trade-off may exist between the level of misspecification and the goodness of the representation. A slightly imprecise but very informative representation may be preferable to most bad realizable ones. Finally, we believe that moving from selection to representation learning –e.g., provided a class of features such as a neural network– is an important direction both from a theoretical and practical perspective.

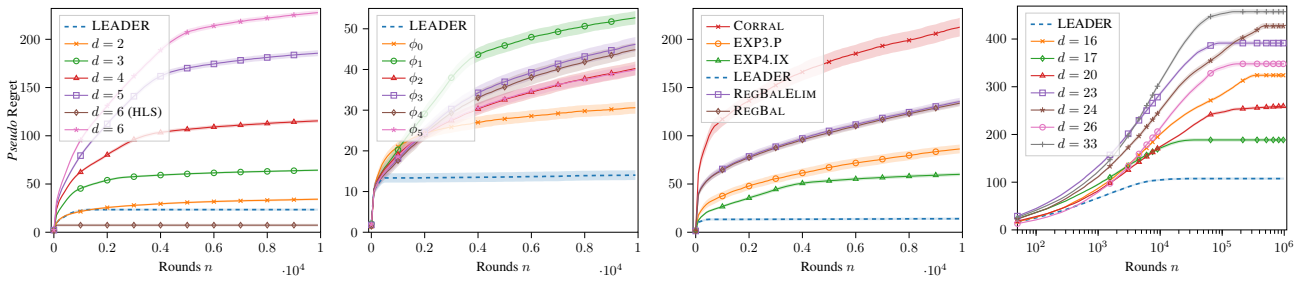

Figure 4. Regret of LEADER and model-selection baselines on different linear contextual bandit problems. (*left*) Synthetic problem with varying dimensions. (*middle left*) Representation mixing. (*middle right*) Comparison to model selection baselines. (*right*) Jester dataset.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, pp. 2312–2320, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, volume 22 of *JMLR Proceedings*, pp. 1–9. JMLR.org, 2012.

Abbasi-Yadkori, Y., Pacchiano, A., and Phan, M. Regret balancing for bandit and rl model selection, 2020.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pp. 176–184. PMLR, 2017.

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pp. 19–26. PMLR, 2012a.

Agarwal, A., Dudík, M., Kale, S., Langford, J., and Schapire, R. E. Contextual bandit learning with predictable rewards. In *AISTATS*, volume 22 of *JMLR Proceedings*, pp. 19–26. JMLR.org, 2012b.

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pp. 12–38. PMLR, 2017.

Arora, R., Marinov, T. V., and Mohri, M. Corralling stochastic bandit algorithms. *CoRR*, abs/2006.09255, 2020.

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

Bastani, H., Bayati, M., and Khosravi, K. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 2020.

Bibaut, A. F., Chambaz, A., and van der Laan, M. J. Rate-adaptive model selection over a collection of black-box contextual bandit algorithms. *CoRR*, abs/2006.03632, 2020.

Bouneffouf, D. and Rish, I. A survey on practical applications of multi-armed and contextual bandits, 2019.

Cantador, I., Brusilovsky, P., and Kuflik, T. 2nd workshop on information heterogeneity and fusion in recommender systems (hetrec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM. URL `http://ir.ii.uam.es/hetrec2011/index.html`. Last.fm website, `http://www.lastfm.com`.

Chatterji, N. S., Muthukumar, V., and Bartlett, P. L. OSOM: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1844–1854. PMLR, 2020.

Chatzigeorgiou, I. Bounds on the lambert function and their application to the outage analysis of user cooperation. *CoRR*, abs/1601.04895, 2016.

Chen, F. A note on matrix versions of kantorovich-type inequality. *JOURNAL OF MATHEMATICAL INEQUALITIES*, 7(2):283–288, 2013.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15 of *JMLR Proceedings*, pp. 208–214. JMLR.org, 2011.

Degenne, R., Ménard, P., Shang, X., and Valko, M. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pp. 2432–2442. PMLR, 2020.

Eaton, M. L. and Perlman, M. D. The non-singularity of generalized sample covariance matrices. *The Annals of Statistics*, pp. 710–717, 1973.

Foster, D. J., Krishnamurthy, A., and Luo, H. Model selection for contextual bandits. In *NeurIPS*, pp. 14714–14725, 2019.

Foster, D. J., Gentile, C., Mohri, M., and Zimmert, J. Adapting to misspecification in contextual bandits. In *NeurIPS*, 2020.

Ghosh, A., Sankararaman, A., and Ramchandran, K. Problem-complexity adaptive model selection for stochastic linear bandits. *CoRR*, abs/2006.02612, 2020.

Goldberg, K. Y., Roeder, T., Gupta, D., and Perkins, C. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.*, 4(2):133–151, 2001.

Hao, B., Lattimore, T., and Szepesvári, C. Adaptive exploration in linear contextual bandit. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3536–3545. PMLR, 2020.

Hoorfar, A. and Hassani, M. Inequalities on the lambert w function and hyperpower function. *J. Inequal. Pure and Appl. Math*, 9(2):5–9, 2008.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143. PMLR, 2020.

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

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

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pp. 5662–5670. PMLR, 2020.

Lee, J. N., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation, 2020.

Maillard, O. and Munos, R. Adaptive bandits: Towards the best history-dependent strategy. In *AISTATS*, volume 15 of *JMLR Proceedings*, pp. 570–578. JMLR.org, 2011.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NeurIPS*, pp. 3168–3176, 2015.

Pacchiano, A., Dann, C., Gentile, C., and Bartlett, P. Regret bound balancing and elimination for model selection in bandits and rl, 2020a.

Pacchiano, A., Phan, M., Abbasi-Yadkori, Y., Rao, A., Zimmert, J., Lattimore, T., and Szepesvári, C. Model selection in contextual stochastic bandit problems. In *NeurIPS*, 2020b.

Reeve, H. W. J., Mellor, J., and Brown, G. The k-nearest neighbour UCB algorithm for multi-armed bandits with covariates. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pp. 725–752. PMLR, 2018.

Rigollet, P. and Zeevi, A. Nonparametric bandits with covariates. *arXiv preprint arXiv:1003.1630*, 2010.

Tirinzoni, A., Pirotta, M., Restelli, M., and Lazaric, A. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

Wu, W., Yang, J., and Shen, C. Stochastic linear contextual bandits with diverse contexts. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2392–2401. PMLR, 2020.