Learning in Markovian bandits with non-observable states and constrained decision epochs

Anonymous authors
Paper under double-blind review

Abstract

This paper studies the problem of regret minimization in Markovian bandits with *non-observable states* and possibly *constrained* decision epochs. The focus is restricted to a "pure" regret benchmark, that compares the performance of the learning algorithm to the best *pure policy* which—akin to optimal policies of stochastic bandits—picks the optimal arm from start to finish without ever switching. We introduce a generalization of rested Markovian bandits, *self-degrading Markovian bandits*, for which pure policies are always asymptotically optimal. We show that without prior knowledge on the underlying bandit, the regret of algorithms that switch arms rarely necessarily scales super-logarithmically for every bandit, i.e., as $\omega(\log(T))$, where T is the learning horizon. Despite the unreachability of the logarithmic regime, we design `UCB-NOM`, an optimistic algorithm inspired by `UCB`, of which the regret is nearly logarithmic. Lastly, we show that given prior knowledge on the Markovian bandit in the form of a bound on the bias functions of its arm, a proper instantiation of `UCB-NOM` achieves $\mathrm{O}(\log(T))$ regret. We further show that this prior knowledge allows for a $\mathrm{O}(\sqrt{T\log(T)})$ worst-case regret bound for `UCB-NOM`. Notably, our regret bounds do not depend on the number of states of the underlying Markov chains. Our findings suggest that the non-observability of states is a mild inconvenience in self-degrading Markovian bandits.

1 Introduction

Multi-armed bandits (MABs) provide a canonical framework for sequential decision-making under uncertainty, balancing exploration and exploitation in environments with unknown reward distributions (Lattimore & Szepesvári, 2020). In standard stochastic MABs, the controller pulls arms sequentially and receives rewards generated by the reward probability distributions associated to the arms it pulls. In this work, we are interested in a variant of MABs in which (1) arms are seen as third-party services, that hold a state that we cannot observe; and (2) the controller may be locked to its decision until the reception of a specific signal. This model is motivated by communication networks, where arms are communication channels through which the controller wants to send data, similarly to Liu et al. (2010); Hira et al. (2025), although our model can also be interpreted in terms of medical trials or machine maintenance problems as well.

For our problem, the relevant Reinforcement Learning (RL) framework is the one of *Markovian bandits* (Gittins et al., 2011), that is, bandits where the reward mechanism of every arm is determined by a Markov chain with an evolving state. We supplement the Markovian bandit structure with (1) the *non-observability of states* and (2) *constrained decision epochs*, where the constraints come from the Markovian bandit itself. The assumption (1) means that the only information available to the controller are the rewards it obtained through the interaction; the other assumption (2) enforces that once an arm is chosen, the controller is to remain committed to it until a special reward is earned. Notably, we assume that a single arm is activated at a time—this meets the framework of standard stochastic bandits, although in opposition to the asymptotic theory of Markovian bandits where the controller pulls a *positive proportion* of arms (Gittins et al., 2011). In the end, the resulting setting is especially favorable to the design of learning algorithms inspired from the setting of *rested* Markovian bandits (Liu et al., 2011; 2012; Tekin & Liu, 2011; 2012).

In this paper, we present a learning algorithm for our problem, `UCB-NOM`, discuss the importance of the availability of prior information on the properties of the Markovian bandit, and provide regret guarantees for `UCB-NOM` depending on the degree of strength of that prior information. By "*regret*", we mean the regret as defined in stochastic bandits (Lattimore & Szepesvári, 2020): the difference between the reward accumulated by the controller (that is learning the best arms) and what the optimal policy can do (that has access to a complete description of the bandit before-hand).

1.1 Related work

Our work is at the intersection of several literature: (1) it is at the interplay of *rested* and *restless* Markovian bandits; (2) it incorporates considerations of *partial observability*—although the latter are pretty common already in the literature on restless Markovian bandits; and (3) it belongs to the family of RL problems with constrained decision epochs. These three literature are discussed in separated paragraphs below.

Markovian bandits (MBs). The literature to which this work is the closest to are Markovian bandits, as they stand as the foundation of our learning problem. The literature of RL on Markovian bandits can be organized along two axes: the *observation model*, ranging from full states observation to reward-only feedback; and the *regret benchmark*, ranging from the best pure policy (that plays the same arm from start to finish) to the best general policy (that may switch from one arm to another in a complex fashion). The regret benchmark is directly related to the "type" of Markovian bandit one is looking at; pure policies are optimal for *rested* bandits (for which non-activated arms stay idle), while they are sub-optimal for *restless* bandits (for which the states of non-activated arms may keep updating). In the sequel, we speak of *pure regret* when the performance of the controller is compared to those of *pure* policies, and we speak of *general regret* otherwise.

- *Fully observable MBs.* In the fully observable setting, several works have derived logarithmic upper-bounds on the pure regret by extending the principles of `UCB` (Auer, 2002), by relying on regeneration arguments (Tekin & Liu (2011; 2012)) or forced exploration mechanisms (Liu et al., 2011; 2012). The algorithms displayed by those works are arguably the closest to ours (particularly `RCA` of (Tekin & Liu, 2011; 2012)), but unlike us, they do rely on the observation of states. In the fully observable setting, other works worth mentioning include (Dai et al., 2011; 2014), that compare to the general optimal policy for a very specific family of Markovian bandits. We intend to work with general (although finite) Markovian bandits in this paper.
- *Partially observable MBs.* A second line of work assumes that only the state of the currently activated arm is observed. For this observation model, Ortner et al. (2012) compares against the best general policy and shows that regret is, in general, $\Omega(\sqrt{T})$. Then, they design `colored-UCRL2`, that matches this lower bound up to logarithmic factors. On the same observation setting, Jung & Tewari (2019) develop an algorithm based on Thompson sampling that achieves $\widetilde{\mathrm{O}}(\sqrt{T})$ Bayesian regret in the presence of periodic resets. This was later extended by Jung et al. (2019), removing the periodic resets' mechanism. These two works highlight an important aspect of this observation model: when competing against the best general policy and if only the state of the currently played arm is observed, one cannot do better than $\sqrt{T}$ general regret guarantees. The pure regret benchmark, although weaker, allows for regret of order $\log(T)$, even in more frugal observation settings.
- *"Non-observable" MBs, or, MBs with "full bandit feedback".* Assuming that only the rewards are observed, sometimes referred to as *full bandit feedback*, is the most severe observation model that one can take. It is adopted by Jiang et al. (2023) in particular, who consider restless bandits with reward-only observations and a single activated arm, similarly to this work. Their algorithm achieves a Bayesian regret upper bound of $\widetilde{\mathrm{O}}(\sqrt{T})$ with respect to the best *general* policy. Their algorithm is based on Thompson sampling with episodic explore-then-commit (`TSEETC`). However, in opposition to our work, they do not have constraints on decision epochs.

Constrained decision epochs. Decision epochs are more common in the setting of multi-armed bandits. For instance, Komiyama et al. (2013) propose a framework where decision epochs are predetermined, while Basu et al. (2019) propose one where playing an arm makes it unavailable for a fixed number of time slots

thereafter. That second work is close in spirit to *batched bandits*, where the decision maker has an upper bound on the number of times it is allowed to change arms. Batched bandits have a growing rich literature (Gao et al., 2019; Esfandiari et al., 2021; Cesa-Bianchi et al., 2013; Dekel et al., 2014; Rouyer et al., 2021). In these works however, the constraints are *exogenous* to the bandit. This is in opposition to the setting that we design in this paper, where the constraints emerge from the Markovian structure of arms directly, making the constraints *endogenous*. In particular, we will see that these constraints eventually disappear in the asymptotic regret bounds that we derive for our algorithm.

1.2 Contributions and outline of the paper

We first motivate the choice of *pure policies*, as a comparison point for our regret benchmark, that we call *pure regret* (simplified to *regret* in the text). In particular, we provide a general class of restless Markovian bandits in which such policies are asymptotically optimal: *self-degrading* Markovian bandits. This class generalizes rested Markovian bandits, but further includes non-trivial restless ones, ranging from the scheduling of non-preemptive Gilbert-Elliott channels from Hira et al. (2025) to bandits with infinitely many states inspired by the framework of Nerlove & Arrow (1962). In Section 4, we show that in general, the expected regret is $\omega(\log(T))$ where T is the number of learning steps (Theorem 4).[1] In particular, the well-known regret guarantees of $\mathrm{O}(\log(T))$ from stochastic bandits (Lattimore & Szepesvári, 2020) cannot be achieved. We then design `UCB` for `N`on-`O`bservable `M`arkovian bandits (`UCB-NOM`, Algorithm 1), that approaches the idealistic $\log(T)$-regret guarantees arbitrarily closely (Theorem 6). Finally, in Section 5, we discuss a structural assumption under which regret guarantees can be improved: If the biases of the Markov reward processes underlying every arm are uniformly bounded, `UCB-NOM` achieves regret $\mathrm{O}(\log(T))$ with the proper tuning (Theorem 8). We further provide a worst-case regret bound $\mathrm{O}(\sqrt{T\log(T)})$ under this assumption (Theorem 9). At last, Section 6 addresses the tuning of `UCB-NOM` with respect to the confidence parameters. An important take-away from our results is that constrained decision epochs do not harm the achievable regret guarantees, and that the non-observability of states does not require involved techniques from the POMDP literature as long as one sticks to the pure regret benchmark, or to self-degrading restless bandits.

2 Learning non-observable Markovian bandits

We begin this paper by describing the learning problem and introducing notations. Our learning problem is a class of Markovian bandits (Section 2.1) in which states are non-observable and decision epochs are possibly constrained. In Section 2.2, we motivate *pure policies* as a benchmark by identifying subclasses of restless bandits in which they are actually optimal. Lastly, in Section 2.3, we motivate the focus on learning algorithms that switch arms rarely.

2.1 Non-observable Markovian bandits with decision epochs

We consider a Markovian bandit with arms $\mathcal{A}$, where every arm $a \in \mathcal{A}$ is modeled by a Markov decision process $M_a \equiv (\mathcal{S}_a, \{0,1\}, \nu_a, p_a^0, p_a^1, s_a^{\mathrm{init}})$ with *activated/non-activated* actions, states $\mathcal{S}_a$, a fixed initial state s_a^{init}, a reward function agnostic to action choices $\nu_a : \mathcal{S}_a \to \mathcal{P}([0,1])$, and two transition kernels $p_a^0, p_a^1 : \mathcal{S}_a \to \mathcal{P}(\mathcal{S}_a)$ that govern a state's evolution depending on its activation or non-activation. Mean rewards are denoted by $r_a(s_a) := \int x \,\mathrm{d}\nu_a(s_a)(x)$. The resulting system is obtained by bundling arms in parallel as $M \equiv (M_a)_{a\in\mathcal{A}}$.

The interaction between the system and the controller goes as follows. At time $t \geq 1$, the controller selects a *single* arm $A(t) \in \mathcal{A}$ to activate, and the system's state evolves as

$$S_a(t+1) \sim \begin{cases} p_a^1(S_a(t)) & \text{if } A(t) = a; \\ p_a^0(S_a(t)) & \text{if } A(t) \neq a; \end{cases} .$$

Moreover, one receives an immediate reward $R(t+1) \sim \nu_a(S_a(t))$ for $a = A(t)$,[2] independently of the state's generation. By convention, $R(1) = 0$. All arms evolve independently conditionally on the current state and

[1] The notation $f = \omega(g)$ means that f grows strictly faster than g, i.e., that $g = \mathrm{o}(f)$.

[2] Because the activated arm is the only one to produce a reward, the assumption that rewards are agnostic to actions can be made without loss of generality. Equivalently, we could have assumed non-activated arms produce no reward.

activated arm. In the sequel, the *number of pulls* of arm a will be denoted $N_a(T) := \sum_{t=1}^{T-1} \mathbf{1}(A(t) = a)$. To simplify the exposition, we further assume that transition kernels under activations p_a^1 are unichain — this assumption is fairly standard in restless Markovian bandits.

Assumption 1. *For every $a \in \mathcal{A}$, p_a^1 is* unichain*, i.e., it has a unique recurrent component of states.*

2.1.1 Constrained decision epochs

In opposition to standard restless Markovian bandits, we are interested in systems in which the controller is constrained to switch arms *only at decision epochs*, induced by the stream of rewards. More specifically, epochs consist of an increasing sequence $(\tau_i)_{i\geq 1}$ of stopping times under the filtration induced by the system's rewards $(R(t))_{t\geq 1}$. In-between two decision epochs, the controller is not allowed to switch arms, that is,

$$\forall i, \forall t \in \{\tau_i, \ldots, \tau_{i+1} - 1\}, \quad A(t) = A(\tau_i). \tag{1}$$

Throughout, we make two assumptions on the structure of decision epochs. In Assumption 3 below, $H(t) = (S(1), A(1), R(2), S(2), A(2), R(3), \ldots, S(t))$ is the *history of play*, i.e., the vector of every state, every played arm and every emitted reward up to time t.

Assumption 2. *There exists a measurable set $\mathcal{U}_\tau \subseteq [0, 1]$ such that $\tau_{i+1} = \inf\{t > \tau_i : R(t) \in \mathcal{U}_\tau\}$.*

Assumption 3. *There exists $C_\tau \in \mathbf{R}_+$ such that, for all $t \geq 1$, $\mathbf{E}[\inf\{\tau_i - t : \tau_i \geq t\}|H(t)] \leq C_\tau$.*

Assumption 2 means that decision epochs happen when the immediate reward falls within a special set. For instance, if $\mathcal{U}_\tau = [0, 1]$, then the decision epochs are trivial with $\tau_i = i$ for all $i \geq 1$, so that Assumption 2 captures systems without constraints on decision epochs in particular. Assumption 3 claims that the expected time before the next decision epoch is bounded uniformly over time conditionally on the history. This assumption holds, for example, if, for every arm $a \in \mathcal{A}$, there exists a state $s_a \in \mathcal{S}_a$ that is recurrent under p_a^1 and such that $\nu(s_a)(\mathcal{U}_\tau) > 0$. Under that scenario, C_τ can typically be bounded in terms of the diameter of p_a^1 (expected time required to travel from a state to another under p_a^1) and the probability $\nu(s_a)(\mathcal{U}_\tau)$ to emit the right reward from favorable states.

2.1.2 Full bandit feedback: non-observability of states

While the number of arms is known to the controller, the latter is unaware of the current states $S_a(t)$ of arms, nor of how many states $|\mathcal{S}_a|$ they can possibly be in. In other words, states are non-observable.

Non-observability implies that the choice of arm $A(t)$ depends exclusively on the past *history of observations* $H_o(t) := (A(1), R(2), A(2), \ldots, R(t))$. In formal terms, the random arm $A(\tau_i)$ is $\sigma(H_o(t))$-measurable. Note that for each decision epoch τ_i, we have $\{\tau_i \leq t\} \in \sigma(R(2), \ldots, R(t))$. That is, decision epochs are measurable with respect to the aggregate reward vector. This assumption is absolutely necessary for our work, as the controller would be unable to properly change arm otherwise.

2.2 Our learning framework: pure policies and regret

Our objective is to design learning algorithms that maximize aggregate rewards $R(1) + \ldots + R(t)$, for $t \to \infty$, without knowledge of the underlying environment.

Let $\mathcal{M}$ be a collection of Markovian bandits as in Section 2.1, that is,

$$\mathcal{M} \subseteq \bigcup_{\substack{\mathcal{S}_a \subseteq \mathbf{N} \\ a \in \mathcal{A}}} \left\{ \left(\mathcal{S}_a, s_a^{\text{init}}, \{0, 1\}, \nu_a, p_a^0, p_a^1\right)_{a\in\mathcal{A}} : p_a^1 \text{ is unichain} \right\}.$$

In particular, all instances of $\mathcal{M}$ have the same set of arms $\mathcal{A}$, but a given arm may have different state spaces from one instance to another. The collection $\mathcal{M}$ shall contain every environment for which we want the learning algorithm to converge to optimal play.

Definition 1 (Pure policies)**.** A policy is *pure* if it plays a single action over time, i.e., if $A(t) = A(t + 1)$ for all $t \geq 1$. The pure policy that plays $a \in \mathcal{A}$ is denoted π_a.

The usual performance measure of a learning algorithm under a model $M \in \mathcal{M}$ is to compare its aggregate rewards $R(1) + \ldots + R(t)$ against the best algorithm specifically designed for M. Motivated by applications where *pure policies* are optimal, we choose those as a benchmark in order to provide much stronger guaranties for those problems than what could be attained in all generality. As such, the difference in aggregate reward between the best pure policy and the learning algorithm is called the *regret*. Formally, the (expected) regret $\mathrm{Reg}(T) \equiv \mathrm{Reg}(T; M, \Lambda)$ of a learner Λ under a model M up to $T \geq 1$ is

$$\mathrm{Reg}(T) := \max_{a \in \mathcal{A}} \left\{ V^{\pi_a}(T; M) - V^{\Lambda}(T; M) \right\} \tag{2}$$

where $V^{\Lambda}(T; M) := \mathbf{E}^{\Lambda, M}[\sum_{t=1}^{T} R(t)]$ is the expected amount of reward gathered by Λ under M within T steps. Obviously, the smaller the regret is, the closer the algorithm is to optimal play.

Rewriting policy values via the gain. The value function of the policy π_a can be expressed with respect to the *gain* and the *bias*, with the well-known formula $V^{\pi_a}(t) = (t-1)g_a(s_a^{\mathrm{init}}) + \mathbf{E}^{\pi_a}[b_a(s_a^{\mathrm{init}}) - b_a(S_a(t))]$, proved in Appendix A.1 for completeness. Both the gain and the bias depend on the initial state. Writing $\mathbf{E}_{s_a}^{\pi_a}[-]$ the expectation under the process induced by playing π_a in M with M_a initialized in $s_a \in \mathcal{S}_a$, the gain and the bias functions are given by[3]

$$g_a(s_a) := \lim_{T \to \infty} \mathbf{E}_{s_a}^{\pi_a}\left[\frac{1}{T}\sum_{t=1}^{T} R(t)\right], \tag{3}$$

$$b_a(s_a) := \lim_{T \to \infty} \mathbf{E}_{s_a}^{\pi_a}\left[\sum_{t=1}^{T} (R(t) - g_a(s_a))\right]. \tag{4}$$

Since p_a^1 is unichain, the gain $g_a(s_a)$ is independent of the initial state $s_a \in \mathcal{S}_a$, and we can simply write g_a. We denote the *optimal gain* by

$$g_* := \max_{a \in \mathcal{A}} g_a. \tag{5}$$

and *bias span* of arm a by $\mathrm{sp}(b_a) := \max_{s \in \mathcal{S}_a} b_a(s) - \min_{s \in \mathcal{S}_a} b_a(s)$.

2.3 Rarely switching algorithms and pseudo-regret

In this section, we define the concept of rarely switching policies, which provide a natural class of learning rules when considering regret with respect to pure policies. Indeed, such a learner should be able to identify the best arm by constructing a good estimator for the gain g_a for every arm. The gain of a pure policy π_a only depends on the states that are visited infinitely many times under π_a; the set of *recurrent states* of π_a. When playing an arm to estimate g_a, the process may initially lie outside this set, requiring several steps before reaching it and thus before the estimation of g_a can begin. The task is further complicated by the fact that $|\mathcal{S}_a|$, the number of states of arm a, is unknown.

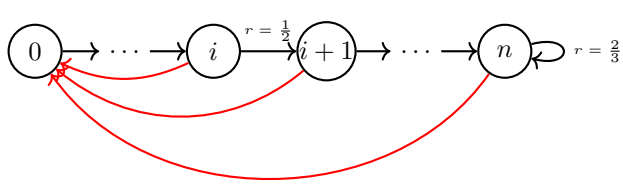


Figure 1: An arm that requires to be activated at least n times in a row to be estimated correctly. All transitions are deterministic, with p_a^1 represented in black and p_a^0 in red.

An example is provided in Figure 1, where state n is absorbing under π_a and the gain of the arm is $g_a = \frac{2}{3}$. The latter can only be estimated accurately if, infinitely often during the learning process, the controller activates the arm more than n times in a row. As $n \geq 1$ can be arbitrarily large in general, it becomes

[3]The gain and the bias, as defined in Equations (3) and (4), depend on the instance M. The dependency has been dropped for typographic convenience. When the dependency may lead to confusion, we will write $g_a(s_0; M)$ and $b_a(s_0; M)$.

tempting to reduce the number of times one switches from one arm to another as much as possible, and to measure the regret by *counting* the number of times bad arms are pulled. This leads to the introduction of the *pseudo-regret*:

$$\mathrm{PR}(T) := \sum_{a \in \mathcal{A}} (g_* - g_a) N_a(T). \tag{6}$$

It is named after its eponym version in multi-armed bandits (Lattimore & Szepesvári, 2020, §4.9).

The pseudo-regret measures the performance of the learning algorithm as a sum of instantaneous costs: the individual cost of arm a is given by the *gain-gap* $g_* - g_a$ (how far its gain is from g_*), and the pseudo regret is the sum of gain-gaps over time. The pseudo-regret is much easier to analyze than the regret, because it is directly expressed in terms of the number of pulls to sub-optimal arms. If the pseudo-regret is large, then the regret is large. The converse is false in general. We now go over a class of learning algorithms for which the regret can nevertheless be upper-bounded in terms of the pseudo-regret: *rarely-switching algorithms*.

Definition 2 (Rarely switching algorithms)**.** Given a sublinear smooth concave function $\psi : \mathbf{R}_+ \to \mathbf{R}_+$, we say that the algorithm is *ψ-rarely switching* if for all $M \in \mathcal{M}$, it holds

$$\forall a \in \mathcal{A},\ \forall T \geq 1, \quad \sum_{t=1}^{T-1} \mathbf{1}(A(t) = a, A(t+1) \neq a) \leq \psi(N_a(T)).$$

In other words, an algorithm is *rarely switching* if the number of switches to arm a grows sub-linearly with respect to $N_a(t)$, the number of pulls to arm a.

When algorithms are ψ-rarely switching, the *pseudo-regret* (6) becomes a natural proxy for the regret: If the algorithm switches arms rarely, then the two quantities are approximately the same (at first order), as stated by Proposition 1 thereafter. This result encourages to fully focus on the study of pseudo-regret.

Proposition 1 (Regret and pseudo-regret)**.** *Let $M \in \mathcal{M}$ be an arbitrary model with a unique optimal arm. Let Λ be a ψ-rarely switching algorithm. Then, there exists a positive smooth concave function $\psi_0 = \mathrm{O}(\psi)$ such that, for every horizon $T \geq 1$, we have*

$$|\mathrm{Reg}(T) - \mathbf{E}[\mathrm{PR}(T)]| \leq \psi_0(\mathbf{E}[\mathrm{PR}(T)]).$$

The proof of Proposition 1 is deferred to Appendix A.2.

Because $\psi_0 = \mathrm{O}(\psi)$ is a sublinear function, Proposition 1 states that the expected regret of a rarely switching algorithm is $\mathbf{E}[\mathrm{PR}(T)] + \mathrm{o}(\mathbf{E}[\mathrm{PR}(T)])$. Accordingly, the regret of such algorithms can be analyzed by controlling the pseudo-regret directly, which is much easier to handle.

3 Why comparing to pure policies is relevant: self-degrading restless bandits

For restless bandits in general, comparing the performance of the learning algorithm to that of pure policies is much weaker than comparing to general policies, because optimal policies are typically non-pure. The only apparent advantage is that if one asks for regret guarantees exclusively against pure policies, then one allows for simpler learning algorithms, which will make decision epochs much easier to manage. This being said, although they are sub-optimal in general, pure policies *can* be asymptotically optimal. This is what we discuss in this section: We exhibit a sub-class of restless bandits, that we name *self-degrading restless bandits* (Definition 3), for which the set of optimal policies intersects the one of pure policies. As such, defining the regret in comparison with pure policies does not weaken the learning benchmark for those classes of restless bandits.

Definition 3 (Self-degrading restless bandits)**.** Given an instance of restless bandit $M \in \mathcal{M}$ together with an activation set $\mathcal{U}_\tau$, an arm $a \in \mathcal{A}$ is said *self-degrading* if, whatever the controller Λ respecting $\mathcal{U}_\tau$,

$$A(t) \neq a \implies \mathbf{E}^{M,\Lambda}[b_a(S_a(t+1)) | H_o(t)] \leq \mathbf{E}^{M,\Lambda}[b_a(S_a(t)) | H_o(t)]\ .$$

The instance M is said *self-degrading* if all of its arms are. The set of all self-degrading instances of Markovian bandits is written $\mathcal{M}_{\mathrm{sd}}$.

Said differently, an arm is *self-degrading* if when left nonactivated between consecutive decision epochs, it tends to move towards worse and worse states conditionally on what the controller observes (the subtleties of that conditioning will appear clearly in the example on Gilbert-Elliott channels). This property is enough to make switches between arms undesirable, so that pure policies become asymptotically optimal, see Proposition 2 below. As a matter of fact, this result is a bit stronger than saying that pure policies are asymptotically optimal: it states that whatever the way the controller chooses its actions, it cannot perform better than the best pure policy up to a *constant* additive factor.

Proposition 2 (Pure policies are optimal in self-degrading instances)**.** *The regret is bounded below on self-degrading instances. That is, for every self-degrading instance of restless bandits $M \in \mathcal{M}_{\mathrm{sd}}$, there exists a constant $C_M < \infty$ such that, whatever the learning algorithm Λ, we have*

$$\mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T} R(t)\right] \leq \max_{a\in\mathcal{A}} V^{\pi_a}(T;M) + C_M \ .$$

The proof of Proposition 2 is deferred to Appendix B.1.

3.1 Regret and pseudo-regret in self-degrading instances

For self-degrading instances, the regret can be lower-bounded with respect to the pseudo-regret, as shown by Proposition 3 below, proved in Appendix B.1. This result can be seen as complementary to Proposition 1 above, as it establishes a relation between regret and pseudo-regret without relying on the *rarely switching* property described in Section 2.3.

Proposition 3 (Regret and pseudo-regret in self-degrading instances)**.** *The regret is bounded below by the pseudo-regret on self-degrading instances. That is, for every self-degrading instance of restless bandits $M \in \mathcal{M}_{\mathrm{sd}}$, there exists a constant $C_M < \infty$ such that, whatever the learning algorithm Λ,*

$$\forall T \geq 1, \qquad \mathrm{Reg}(T;M,\Lambda) \geq \mathbf{E}^{M,\Lambda}[\mathrm{PR}(T)] - C_M \ .$$

The proof of Proposition 3 invokes the same arguments as Proposition 2—in fact, it is easy to see that Proposition 2 is an immediate consequence of Proposition 3, for $\mathrm{PR}(T) \geq 0$.

3.2 Examples of self-degrading restless bandits

We now provide three examples of self-degrading restless bandits: rested bandits with Example 1, Gilbert-Elliott channels with Example 2, and a toy version of the Nerlove-Arrow model with Example 3.

Example 1 (Rested bandits)**.** A Markovian bandit is said *rested* if inactive arms remain frozen in their current state. That is, for all arm $a \in \mathcal{A}$,

$$p_a^0 = I$$

where I is the identity matrix. For nonactivated arms, $S_a(t+1) = S_a(t)$, so such instances are self-degrading for every activation set $\mathcal{U}_\tau$.

Example 2 (Gilbert-Elliott channels)**.** A Gilbert-Elliott channel (Gilbert, 1960) is a Markov chain with states `ON`/`OFF` representing the availability of the channel, and a non-null reward function from the state `ON`. Those channels evolve with the same transition kernel whether active or inactive. Figure 2 below represents such a channel. Gilbert–Elliott channels have been widely studied in the literature (Elliott, 1963; Mushkin & Bar-David, 1989; Haßlinger & Hohlfeld, 2011; Polyanskiy et al., 2011) etc... We present one instantiation from Hira et al. (2025) where the authors study non-observable Gilbert–Elliott channels with a non-preemptive scheduling constraint, that is, where decision epochs correspond to times at which the received reward is not zero. Those assumptions on the observation framework and the decision epochs define a *self-degrading* restless bandit.

Formally, given $x_a, y_a, \mu_a \in (0,1)$, the associated Gilbert-Elliott channel has transition kernels and reward distributions given by

$$p_a^1 = p_a^0 = \begin{pmatrix} 1-y_a & y_a \\ x_a & 1-x_a \end{pmatrix}, \qquad \nu_a(\texttt{ON}) = \mathrm{Bernoulli}(\mu_a) \quad \text{and} \quad \nu_a(\texttt{OFF}) = \mathrm{Dirac}(0),$$

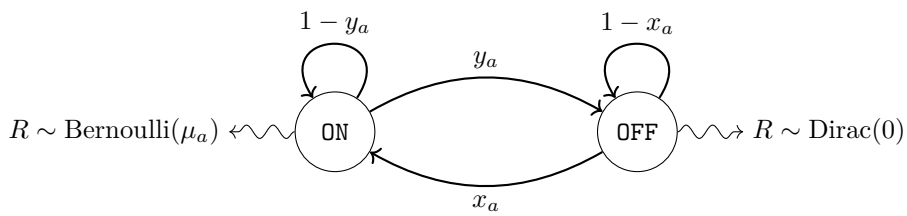


Figure 2: A Gilbert-Elliott channel with parameters x_a, y_a, μ_a.

with an initialization $s_a^{\text{init}} = \texttt{ON}$.[4] Although the state of the channel cannot be observed, receiving a positive reward reveals that the channel was ON. On the other hand, the absence of a reward does not unequivocally identify the state of the channel. The controller can choose a different channel only after receiving a reward, that models a successful transmission through the channel. Accordingly, the activation set is $\mathcal{U}_\tau = \{1\}$ (see Assumption 2). We prove in Appendix B.2.1 that if the chain p_a is *positively auto-correlated*, the resulting restless bandit is self-degrading. *Positive auto-correlation* is a more general concept that translates, in this case, to the algebraic condition $1 - x_a - y_a > 0$ (every state is absorbing enough). This condition represents some inertia to change: a channel has a tendency to stick to the state it is on rather than chattering between states. This inertia is only relevant when states are non-observable. Indeed, if b_a is the bias function of the channel, then the inequality

$$b_a(S_a(t+1)) \leq b_a(S_a(t))$$

is *not* true in general (when $S_a(t) = \texttt{OFF}$). This implies that pure policies are *not* optimal for Gilbert-Elliott channels in the setting where states are observed. However, for $s_a^{\text{init}} = \texttt{ON}$, we show in the Appendix that

$$\mathbf{E}^{M,\Lambda}[b_a(S_a(t+1))|H_o(t)] \leq \mathbf{E}^{M,\Lambda}[b_a(S_a(t))|H_o(t)]$$

holds indeed. As controllers that do not observe states take their decisions from $H_o(t)$ alone, the above property is enough to guarantee that such controllers can never outperform pure policies.

Example 3 (Nerlove-Arrow toy model)**.** For the last example, we introduce a Markovian bandit inspired from the framework of Nerlove & Arrow (1962), that models the growth of skills or interest of human subjects.

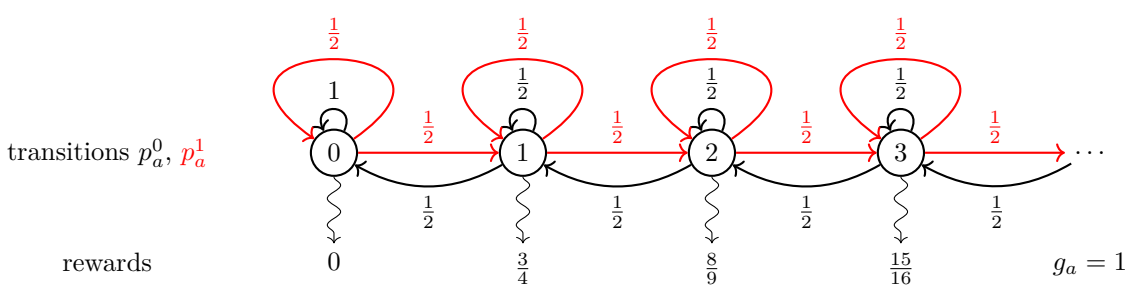


Figure 3: An instance of Nerlove-Arrow toy model.

In human resources or advertisement, skill or goodwill is earned by investing effort into it. When a high quantity is accumulated, more reward is earned. But when abandoned, those fade away. This idea can be modeled as a restless bandit as such: Arms are subjects, states $\mathcal{S}_a \equiv \mathbf{N}$ model the amount of skills, and rewards model how well the target task is performed. For each arm $a \in \mathcal{A}$, if $A(t) = a$ then the state of this arm tends to increase in expectation, otherwise it tends to decrease. We assume $\mathcal{U}_\tau = [0, 1]$, so that every step $t \in \mathbf{N}$ is a decision epoch. Formally, the evolution of skills satisfies

$$\mathbf{E}[S_a(t+1)|S_a(t), A(t) = a] \geq S_a(t) \qquad \text{and} \qquad \mathbf{E}[S_a(t+1)|S_a(t), A(t) \neq a] \leq S_a(t)\ .$$

[4]This choice of initial state is made for simplicity. The channel is not self-degrading if initialized in state OFF, but it can be shown that it satisfies the self-degrading property as soon as *one* non-null reward is observed. In other words, Gilbert-Elliott channels are *eventually self-degrading* regardless of their initialization.

The asymptotic skill threshold is $\mu_a > 0$ and represents how well the subject a performs when fully trained. From state $s \in \mathcal{S}_a$, the subject does not perform at its peak, so that mean reward is of the form

$$r_a(s) = \mu_a - f_a(s)$$

where $f_a : \mathcal{S}_a \to \mathbf{R}_+$ is non-negative. So that subject $a \in \mathcal{A}$ approaches its peak in finite time, we assume that from any initial point $s \in \mathcal{S}_a$,

$$\mathbf{E}_s^{\pi_a}\left[\sum_{t=1}^{\infty}(\mu_a - R(t+1))\right] = \left(\sum_{t=1}^{\infty}(p_a^1)^t f_a\right)(s) < \infty \ ,$$

i.e., that the chain has finite bias. The induced restless bandit is *self-degrading*, as shown in Appendix B.2.2. An example of our Nerlove-Arrow toy model is displayed in Figure 3.

4 General regret bounds

As a first sketch of the landscape of the theory, we attack the problem from an *instance-dependent* viewpoint. That is, given an instance $M \in \mathcal{M}$ and a learning algorithm Λ, we aim at bounding $\mathrm{Reg}(T; M, \Lambda)$ as accurately as possible when $T \to \infty$. This is in opposition to *worst-case*, or *minimax* approaches, for which we would bound $\sup_{M\in\mathcal{M}} \mathrm{Reg}(T; M, \Lambda)$. The minimax approach is proved to be ill-behaved when $\mathcal{M}$ is too large, while non-trivial instance-dependent bounds can always be obtained.

Starting with the instance-dependent setting, we show in Section 4.1 that the expected regret (2) of every rarely-switching algorithm grows super-logarithmically (Theorem 4). We further conjecture that the rarely-switching assumption can even be dropped on the space of self-degrading instances. These lower bounds provide a baseline for learning algorithms. In particular, regret guarantees of order $\mathrm{O}(\log(T))$ are not achievable. In Section 4.2, we present our algorithm `UCB-NOM`, a variant of `UCB2` (Auer (2002)), that takes as input a diverging *span-proxy* function f as well as a confidence function $\delta : \mathbf{N} \to [0, 1]$. In Section 4.3, we show that the instance-dependent regret of `UCB-NOM` is $\mathrm{O}(f(T)^2 \log(T))$, which implies that it can approximate the idealistic logarithmic regret regime arbitrarily close. We conclude in Section 4.4 by showing that minimax regret bounds are unattainable for general $\mathcal{M}$.

4.1 A super-logarithmic regret lower bound

To establish *instance-dependent* lower bounds, we must assume that the learning algorithm of interest is *consistent*, that is, that the algorithm learns the optimal policy in every possible instance in $\mathcal{M}$. Formally, a learning algorithm Λ is *consistent* on $\mathcal{M}$ if, for all $M \in \mathcal{M}$ and $\varepsilon > 0$, it achieves regret $\mathrm{Reg}(T; M, \Lambda) = \mathrm{o}(T^\varepsilon)$, when $T \to \infty$. This definition is fairly standard in the multi-armed bandit literature (Lattimore & Szepesvári, 2020, §16) and is necessary to derive the famous lower bound of Lai & Robbins (1985). Below, we show that consistency is enough to derive a super-logarithmic lower bound, provided that the algorithm is ψ-rarely switching in addition.

Theorem 4 (Lower bound, rarely switching algorithms)**.** *Let Λ be a learning algorithm that (1) is consistent on $\mathcal{M}$, and (2) is ψ-rarely switching. For every instance $M \in \mathcal{M}$ with $g_* < 1$ and unique optimal pure policy, it holds that*

$$\forall s_0 \in \mathcal{S}, \quad \mathrm{Reg}(T; M, \Lambda) = \omega(\log(T)). \tag{7}$$

In particular, Equation (7) implies that no algorithm achieves $\mathrm{O}(\log(T))$ regret, the gold standard for a large variety of RL problems, ranging from multi-armed bandits (Lai & Robbins (1985); Auer (2002)) to Markov decision processes (Auer & Ortner (2006)).

We further believe that the rarely switching property of Λ can be changed to $M \in \mathcal{M}_{\mathrm{sd}}$, i.e., that the regret scales super-logarithmically on the class of self-degrading instances, without requiring the rarely switching property. We note that super-logarithmic lower bounds have also been found in other settings, including multi-armed bandits problems for which the eligible set of reward distributions is too large, see (Lattimore & Szepesvári, 2020, Exercise 16.4).

Proof sketch of Theorem 4 (see complete proof in Appendix C.2.) The proof involves the construction of an alternative model M_a^ε that (*1*) is nearly indistinguishable from the original M_a from observations only, and (*2*) has a different optimal arm. The alternative model M_a^ε is constructed by adding an ε-transition to an absorbing state s_∞ that yields maximal reward $R = 1$, see Figure 4. We construct M^ε from M by picking a suboptimal arm a and changing M_a to M_a^ε, hence making a the best arm for M^ε.

Due to consistency, the algorithm learns the optimal policy in both M and in M^ε—yet the two instances do not have the same optimal policy. Hence, to play optimal in M, the algorithm must reject the plausibility that the true underlying environment is M^ε. As M and M^ε only differ at arm a, the algorithm can only reject M^ε by activating arm a. As a result, this gives a lower bound on the number of pulls of a.

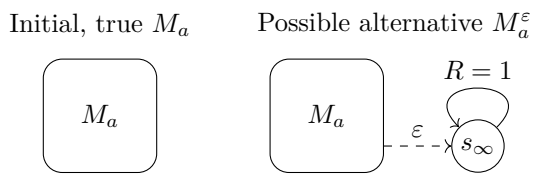


Figure 4: An illustration of the transformation M_a^ε.

More precisely, we show that $\mathbf{E}^{M,\Lambda}[N_a(T)] \geq \frac{1}{\varepsilon}\log(T)$. Since $\varepsilon > 0$ is arbitrary, we deduce that $\mathbf{E}^{M,\Lambda}[N_a(T)] = \omega(\log(T))$, which directly lower bounds the expected pseudo-regret. Equation (7) now follows by linking the pseudo-regret to the regret using the ψ-rarely switching property (Proposition 1). □

4.2 An indexed optimistic algorithm: `UCB-NOM`

In this Section 4.2, we present `UCB-NOM` (Algorithm 1). `UCB-NOM` follows the *optimism-in-the-face-of-uncertainty* (OFU) principle, and begins by constructing an *optimistic index* $\tilde{g}_a(t)$ for every arm a. To construct that index, the rationale goes as follows. First, we provide a natural estimator $\hat{g}_a(t)$ for the value g_a, then bound by how much that estimator tends to deviate from g_a, and based on that bound, we finally construct a high-probability upper-bound $\tilde{g}_a(t)$ of g_a. Following the OFU principle, our algorithm simply activates the arm maximizing $\tilde{g}_a(t)$, and keeps it active until the number of visits $N_a(t)$ doubles.

We now provide details on how the index of `UCB-NOM` is constructed.

Algorithm 1 `UCB` for `Non-Observable` `Markovian` bandits (`UCB-NOM`).

Require: A span proxy function f.
A confidence function $\delta : \mathbf{N} \to [0, 1]$.

$t \leftarrow 1$;
for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k \leftarrow t$;
  Compute indexes $\tilde{g}_a(t_k)$ as

$$\tilde{g}_a(t_k) := \hat{g}_a(t_k) + \frac{f(t_k) \cdot \xi_a(t_k, \delta)}{N_a(t_k)} \tag{8}$$

  Pick an arm $A(t_k) \in \arg\max \tilde{g}_a(t_k)$ maximizing the index;
  while $N_{A(t_k)}(t) < 2N_{A(t_k)}(t_k)$ or t isn't an activation time **do**
    Activate the arm $A(t) := A(t_k)$;
    Observe the new reward $R(t + 1)$;
    Increment $t \leftarrow t + 1$;
  end while
end for

The empirical value of an arm. To estimate the value of an arm (its gain g_a), we simply aggregate and take the average of all the rewards collected by pulling the said arm. We call it the *empirical value* of arm a,

and denote it $\hat{g}_a(t)$. Formally,

$$\hat{g}_a(t) := \frac{1}{N_a(t)} \sum_{i=1}^{t-1} \mathbf{1}(A(i) = a)\, R(i+1), \tag{9}$$

with the convention $\hat{g}_a(t) = 1$ when $N_a(t) = 0$. As dictated by the OFU principle, we are interested in constructing a *high probability upper bound* of g_a from $\hat{g}_a(t)$. We do so by bounding $|\hat{g}_a(t) - g_a|$ via a concentration result (Proposition 5), that bounds that error as a function of the number of pulls to arm a, the number of switches to arm a, and the bias span.

Proposition 5 is at the heart of the construction of the index $\tilde{g}_a(t)$ of our algorithm as such, we provide its proof in the main body.

Proposition 5 (Error of the empirical estimator)**.** *For all arm $a \in \mathcal{A}$ and confidence threshold $\delta > 0$,*

$$\mathbf{P}\big(\exists t \geq 1, N_a(t)|\hat{g}_a(t) - g_a| > D \cdot \xi_a(t,\delta)\big) \leq \delta,$$

where $D \equiv 1 + \max_{a\in\mathcal{A}} \mathrm{sp}(b_a)$ and the error function is given by

$$\xi_a(t,\delta) := \sqrt{N_a(t) \log\Big(\frac{1+N_a(t)}{\delta}\Big)} + \sum_{i=1}^{t-1} \mathbf{1}(A(i-1) \neq a, A(i) = a). \tag{10}$$

Proof. Consider the sequence of stopping times given by $\tau_1^a := \inf\{t \geq 1 : A(t) = a\}$ and $\tau_{k+1}^a := \inf\{t > \tau_k^a : A(t) = a\}$, that enumerates the time-instants when the played arm is $a \in \mathcal{A}$. First, introduce the noise terms

$$W_a^r(t) := \sum_{k=1}^{N_a(t)} \Big(R(\tau_k^a + 1) - r_a(S(\tau_k^a))\Big) \qquad \text{and} \qquad W_a^p(t) := \sum_{k=1}^{N_a(t)} \Big(b_a(S_a(\tau_k^a + 1)) - p_a^1(S_a(\tau_k^a))b_a\Big),$$

that respectively measure how much the immediate rewards and transitions—both obtained by activating arm a—deviate from their means. Further, introduce their sum $W_a(t) := W_a^r(t) + W_a^p(t)$.

For $t \geq 1$, we have

$$\begin{aligned}
N_a(t)\hat{g}_a(t) &= \sum_{i=1}^{t-1} \mathbf{1}(A(i) = a) R(i+1) \\
&= \sum_{k=1}^{\infty} \mathbf{1}(\tau_k^a < t) R(\tau_k^a + 1) \\
&\overset{\circ 1}{=} \sum_{k=1}^{N_a(t)} R(\tau_k^a + 1) \\
&\overset{\circ 2}{=} \sum_{k=1}^{N_a(t)} \Big(g_a + \Big(e_{S_a(\tau_k^a)} - p_a^1(S_a(\tau_k^a))\Big) b_a\Big) + W_a^r(t) \\
&\overset{\circ 3}{=} N_a(t) g_a + \sum_{k=1}^{N_a(t)} \Big(e_{S_a(\tau_k^a)} - e_{S_a(\tau_k^a+1)}\Big) b_a + W_a(t)
\end{aligned} \tag{11}$$

where ($\circ$1) uses that $N_a(t) := \sup\{k : \tau_k^a < t\}$; and ($\circ$2) introduces the canonical basis $(e_s)_{s\in\mathcal{S}_a}$ of $\mathbf{R}^{\mathcal{S}_a}$ and invokes the Poisson equation under π^a, given by $g_a + b_a(s_a) = r(s_a) + p_a^1(s_a) b_a$[5]. In ($\circ$3), we recognize two error terms. The first is linked to the number of switches, while the second is a martingale noise.

First, the martingale noise $W_a(t)$ is bounded using a time-uniform Azuma-Hoeffding's inequality (Bourel et al. (2020)). We get that, uniformly for $t \geq 1$ and with probability at least $1 - \delta$,

$$|W_a^r(t) + W_a^p(t)| \leq (1 + \mathrm{sp}(b_a)) \sqrt{N_a(t) \log\Big(\tfrac{1+N_a(t)}{\delta}\Big)}.$$

[5]Here a compact vector notation is used.

For the first error term of Equation (11), we have:

$$\begin{aligned}\left|\sum_{k=1}^{N_a(t)}\left(e_{S(\tau_k^a)}-e_{S(\tau_k^a+1)}\right)b_a\right| &\leq \mathrm{sp}(b_a)\left(1+\sum_{k=1}^{N_a(t)}\mathbf{1}\big(S(\tau_k^a+1)\neq S(\tau_{k+1}^a)\big)\right)\\ &\leq \mathrm{sp}(b_a)\left(1+\sum_{k=1}^{N_a(t)}\mathbf{1}\big(\tau_k^a+1\neq \tau_{k+1}^a\big)\right)\\ &\leq \mathrm{sp}(b_a)\sum_{i=1}^{t-1}\mathbf{1}(A(i-1)\neq a, A(i)=a).\end{aligned}$$

We conclude by plugging everything back in Equation (11). □

Building an optimistic estimator. The bound on the estimation error established in Proposition 5 allows for the construction of an *optimistic estimator* $\tilde{g}_a(t)$ of g_a. By optimistic, we mean that $\tilde{g}_a(t)$ serves as a high probability upper bound on the arm's value g_a, that is, $\mathbf{P}(\exists t\geq 1, \tilde{g}_a(t)<g_a)\leq\delta$. A direct consequence of Proposition 5 is that

$$\mathbf{P}\left(\exists t\geq 1 : \hat{g}_a(t)+\frac{D\cdot\xi_a(t,\delta)}{N_a(t)}<g_a\right)\leq\delta. \tag{12}$$

Equation (12) suggests that $\hat{g}_a(t)+\frac{D\cdot\xi_a(t,\delta)}{N_a(t)}$ is a natural choice for $\tilde{g}_a(t)$. Unfortunately, the quantity D is unknown and varies depending on the underlying instance of Markovian bandit. To circumvent this difficulty, we consider a *span proxy function* $f:\mathbf{N}\to\mathbf{R}_+$ diverging to $+\infty$. Eventually, $f(t)$ exceeds D, so that for every instance $M\in\mathcal{M}$, there is a time threshold $\beta_f\in\mathbf{R}_+$, s.t.,

$$\mathbf{P}\left(\exists t\geq \beta_f : \hat{g}_a(t)+\frac{f(t)\cdot\xi_a(t,\delta)}{N_a(t)}<g_a\right)\leq\delta. \tag{13}$$

That time threshold is $\beta_f := 1+\sup\{t\geq 1 : f(t)\leq D\}$. The above justifies our choice of the index $\tilde{g}_a(t)$ in Equation (8) of Algorithm 1.

Taming the number of switches. It is preferable for the optimistic estimator to converge to the empirical estimator when $N_a(t)$ clearly overshoots, or said differently, that the confidence region for g_a shrinks to a singleton when a is the most visited arm for instance. To make sure of it, we need the associated error $\xi_a(t,\delta)/N_a(t)$ to vanish asymptotically. In Equation (10), we see that the error function $\xi_a(t,\delta)$ increases with the number of switches $\sum_{i=1}^{t-1}\mathbf{1}(A(i-1)\neq a, A(i)=a)$. So, $\xi_a(t,\delta)/N_a(t)$ can only be vanishing if the number of switches grows sublinearly in $N_a(t)$; In other words, when the algorithm satisfies a rarely switching property (Definition 2).

A simple method to obtain a rarely switching algorithm is to rely on the famous *doubling trick* (see Auer et al. (1995) and (Auer, 2002, `UCB2`)): Upon activating arm a, activate it until the visit counts of $N_a(t)$ doubles. Doing so, the number of switches to arm a becomes of order $\log(1+N_a(t))$, which makes the algorithm O(log)-rarely switching.

Combining the above, we recover `UCB-NOM` (Algorithm 1).

Comparison with `RCA`. Our algorithm `UCB-NOM` is closely related to `RCA` of Tekin & Liu (2012), as both are generalizations of `UCB` (Auer, 2002) to Markovian bandits. Two major differences with `RCA` are worth mentioning.

- *Observability of states and regeneration.* `RCA` stands for *Regenerative Cycle Algorithm*, and the algorithm of Tekin & Liu (2012) observes the stream of states to calibrate its arm estimates better. In particular, `RCA` always make sure to switch off from arm a while arm a is in a reference state $s_a\in\mathcal{S}_a$. Moreover, when it switches onto arm a, it ignores the first pool of rewards until s_a is reached. This makes the empirical estimator of g_a — the $\hat{g}_a(t)$ — of Tekin & Liu (2012) different from our

own. Their estimator is unbiased—hence better—because it does not require the correction term $\sum_{i=1}^{t-1} \mathbf{1}(A(i-1) \neq a, A(i) = a)$ due to switching arms in the optimistic bonus. However, because states are non-observable in our setting, the regeneration technique is unachievable.

- *Optimistic bonus.* In the end, the optimistic bonus is a correction term. It is about how much the aggregate rewards of a Markov reward process tends to deviate from its mean, i.e., is about isolating the trajectory of rewards $(R_i^a)_{i\geq 1}$ generated by arm $a \in \mathcal{A}$ and comparing $\sum_{i=1}^{t} R_i^a$ to $t \cdot g_a$. For both, the optimistic bonus is of the form $\sqrt{L/N_a(t)}$, where

$$L \asymp \begin{cases} \frac{|\mathcal{S}_a|^2 \hat{\pi}_{\max}^2}{\varepsilon_{\min}} & \text{for Tekin \& Liu (2012);} \\ 1 + \mathrm{sp}(b_a)^2 & \text{for us,} \end{cases}$$

where $\hat{\pi}_{\max} \leq 1$ is related to the stationary distribution of the chain M_a and $\varepsilon_{\min} > 0$ is the spectral gap, see (Tekin & Liu, 2012). The presence of the spectral gap in the bound of Tekin & Liu (2012) naturally emerges from their analysis: to compare $\sum_{i=1}^{t} R_i^a$ and $t \cdot g_a$, the authors focus on the speed of convergence of the empirical measure of presence of $(S_i^a)_{i\geq 1}$, relying on standard Markov chain techniques. In general, the spectral gap is related to the mixing time as $t_{\mathrm{mix}} = \widetilde{O}(|\mathcal{S}|/\varepsilon_{\min})$,[6] see (Jerison, 2013), while the bias span is related to the mixing time with $\mathrm{sp}(b) = \mathrm{O}(t_{\mathrm{mix}})$, see (Wang et al., 2022). A dependency in $\mathrm{sp}(b)$ is better than one in $(|\mathcal{S}|, \varepsilon_{\min})$: we have no dependency in the number of states and $\mathrm{sp}(b)$ ties the transition structure together with the reward mechanism. However, our $\mathrm{sp}(b)^2$ is a bit loose — it could be improved using variance reduction techniques (Boone, 2024, Chapter 5), but proof details would then be more delicate.

4.3 An asymptotic regret bound for `UCB-NOM` matching the general regret lower bound

We have shown that the regret of every algorithm must be $\omega(\log(T))$ and we have described an algorithm (Algorithm 1), parameterized by a bias proxy f. The next theorem demonstrates that the idealistic $\log(T)$ regime is approached arbitrarily closely by `UCB-NOM`.

Theorem 6 (Upper bound)**.** *Let $f : \mathbf{R}_+ \to \mathbf{R}_+$ an increasing function diverging to infinity and a confidence function $\delta(t) : \mathbf{N} \to \mathbf{R}_+^*$ such that $\sum_{t\geq 1} \delta(t) < \infty$ and $\log(\frac{1}{\delta(t)}) = \Theta(\log(t))$.[7] For all instance $M \in \mathcal{M}$,*

$$\mathrm{Reg}(T; M, \texttt{UCB-NOM}(f)) = \mathrm{O}\left(\sum_{a \notin \mathcal{A}^*} \frac{f(T)^2 \log(T)}{g_* - g_a} + \sum_{a \notin \mathcal{A}^*} (g_* - g_a)(\beta_f)^2 \right), \tag{14}$$

where $\mathcal{A}^* := \arg\max_{a\in\mathcal{A}} g_a$ is the set of optimal arms and $\beta_f := 1 + \sup\{t \geq 1 : f(t) \leq D\}$ is a burn-in cost (see Equation (13)), with $D := \max\{\max_a \mathrm{sp}(b_a) + 1, C_\tau\}$ a threshold that depends on M.

The proof, together with error terms of higher orders, can be found in Appendix D.1. Note that the dominant term is of order $f(T)^2 \log(T)$ and that notably, the delay constant C_τ is absent from it.

In particular Theorem 6 shows that the lower bound of Theorem 4 is *tight*, and that `UCB-NOM` achieves it. Indeed, the definition of $\omega(\log(T))$ provided by the lower bound reads as follows: for a given instance M and an algorithm Λ, there exists a function $f_0 \to \infty$ such that $\liminf_{T\to\infty} \mathrm{Reg}(T; M, \Lambda)/(f_0(T)\log(T)) > 0$. For `UCB-NOM` with span proxy $f_1 \equiv (f_0)^{\frac{1}{3}}$, it holds that,

$$\limsup_{T\to\infty} \frac{\mathrm{Reg}(T; M, \texttt{UCB-NOM}(f_1))}{\mathrm{Reg}(T; M, \Lambda)} = 0.$$

Hence, our `UCB-NOM`(f_1) has better regret than Λ on M. It can even be further improved by running it with span proxy $f_2 \equiv (f_1)^{\frac{1}{3}}$, or $f_3 \equiv (f_2)^{\frac{1}{3}}$, and so and so forth. This means that whatever the algorithm is, it can be forever improved, getting closer and closer to the idealistic (but non-reachable) $\log(T)$ regret regime.

[6] $\widetilde{O}(-)$ hides logarithmic factors.

[7] For instance, any $\delta(t)$ of the form $\delta(t) = (\frac{1}{t})^{1+\varepsilon}$ with $\varepsilon > 0$ works.

4.4 Worst case regret: what is the best way to choose the span proxy f?

From an asymptotic viewpoint, the regret is simply $\mathrm{O}(f(T)^2 \log(T))$. The best choice for f is thus a function diverging $+\infty$ as slowly as possible. Unfortunately, there is no such function, because $\sqrt{f}$ always diverges slower than f. However, this observation only tells half of the story.

As it turns out, due to the reliance of `UCB-NOM` on an optimistic estimator $\tilde{g}_a(t)$ to estimate the value g_a of arm a, the asymptotic regret guarantee of Theorem 6 hides a burn-in cost $\beta_f := 1 + \sup\{t \geq 1 : f(t) \leq D\}$ (see Equation (13)) that is eventually negligible, but dominant nonetheless in the early stages of the learning process. The regret upper-bound of Equation (14) should therefore be read as such: The first term (in $f(T)^2 \log(T)/(g_* - g_a)$) holds the asymptotic behavior regret, while the second term (in β_f) represents the regret that may be accumulated before the optimistic indices become valid—a *burn-in cost*.

From a non-asymptotic viewpoint, no proper tradeoff between the asymptotic $f(T)^2 \log(T)/(g_* - g_a)$ and the burn-in cost β_f can be found. Theorem 7 states that independently of the choice of span proxy f, for every horizon $T \geq 1$, there will be an instance where the regret scales linearly.

Theorem 7 (Worst case lower bound)**.** *Let Λ be a learning algorithm. Let $\mathcal{U}_\tau \subseteq [0,1]$ be arbitrary. For $T \geq 1$, we have*

$$\sup_{M \in \mathcal{M}} \mathrm{Reg}(T; M, \Lambda) \geq \frac{11}{4800} \cdot T. \tag{15}$$

The proof can be found in Appendix C.3.

Theorem 7 shows that our class of models is so broad that *every* algorithm has linear regret in the worst case (in particular `UCB-NOM`, whatever the choice of f). Hence, no single span proxy f could be uniformly optimal. Different choices of f will perform better on some instances and worse on others.

5 Regret bounds with uniformly bounded bias span

In Section 4, we have shown that for the space of all Markovian bandit problems, the regret of consistent algorithms must scale as $\omega(\log(T))$ and that this bound is achieved by `UCB-NOM` (Algorithm 1). To achieve this result, we have argued how important it is to make the span proxy f diverge to $+\infty$, so that f eventually upper-bounds the unknown worst bias span $\max_{a \in \mathcal{A}} \mathrm{sp}(b_a)$.

What if a bound on the worst bias span is known?

In this section, we show that given prior knowledge on $\max_{a \in \mathcal{A}} \mathrm{sp}(b_a)$, the span proxy function does not need to diverge to $+\infty$ anymore and the logarithmic regret regime can even be achieved. In particular, by running `UCB-NOM` with span proxy $f \equiv \max_{a \in \mathcal{A}} \mathrm{sp}(b_a)$, we obtain instance dependent regret $\mathrm{O}(\log(T))$ (Section 5.1) and worst-case regret $\mathrm{O}(\sqrt{T \log(T)})$ (Section 5.2).

For $D \in \mathbf{R}_+$, we introduce

$$\mathcal{M}_D := \left\{ M \in \mathcal{M} : \max\left\{ \max_{a \in \mathcal{A}} \mathrm{sp}(b_a) + 1, C_\tau \right\} \leq D \right\},$$

the collection of instances with bias span and decision epoch delays C_τ (Assumption 3) both bounded by D. The second assumption "$C_\tau \leq D$" is not strictly necessary, but will significantly simplify the theory.

5.1 Learning with a bound on the span proxy

Under the assumption that $M \in \mathcal{M}_D$, the idealistic logarithmic regret regime is achieved by `UCB-NOM`, when equipped with the span proxy function $f \equiv D$, see Theorem 8 below.

Theorem 8 (Instance-dependent regret)**.** *Let $D \in \mathbf{R}_+$ and $M \in \mathcal{M}_D$ with a* unique *optimal arm a^*. Let $\delta(t) : \mathbf{N} \to \mathbf{R}_+^*$ be a confidence function such that $\sum_{t \geq 1} \delta(t) < \infty$ and $\log(\frac{1}{\delta(t)}) = \Theta(\log(t))$. When run with*

span proxy $f \equiv D = 2\max_{a\in\mathcal{A}} \max\{1+\mathrm{sp}(b_a), 3\,\mathrm{sp}(b_a)\}$*, the regret of* `UCB-NOM` *satisfies*

$$\mathrm{Reg}(T; M, \texttt{UCB-NOM}(D)) = \mathrm{O}\left(\sum_{a\neq a^*} \frac{D^2 \log(T)}{g_* - g_a}\right).$$

Proof. The proof essentially reduces to the analysis performed in the proof of Theorem 6. Details can be found in the Appendix D.2. □

The regret remains logarithmic even if the optimal arm is not unique, but the regret upper bound is slightly degraded. See Appendix D.2 for more details. A discussion on the relation between our results and similar ones from the stochastic bandits literature can be found in Section 6.3.

5.2 Worst-case regret guarantees

Furthermore, with the assumption $M \in \mathcal{M}_D$, we can move beyond the instance-dependent logarithmic regret previously presented in Theorem 8 and establish robust guarantees that hold for every instance uniformly.

Theorem 9 (Worst-case regret)**.** *Let* $D \in \mathbf{R}_+$ *and* $\delta(t) : \mathbf{N} \to \mathbf{R}_+^*$ *be a confidence function such that* $\sum_{t\geq 1} t\cdot\delta(t) < \infty$ *and* $\log(\frac{1}{\delta(t)}) = \Theta(\log(t))$.[8] *When running with span proxy* $f \equiv D$*,* `UCB-NOM`(D) *has the following worst-case regret:*

$$\sup_{M\in\mathcal{M}_D} \mathrm{Reg}(T; M, \texttt{UCB-NOM}(D)) = \mathrm{O}\Big(D\sqrt{|\mathcal{A}| T \log(T)}\Big).$$

Proof. The proof is deferred to Appendix D.3. □

The tightness of this result is discussed in more detail in Appendix D.3.

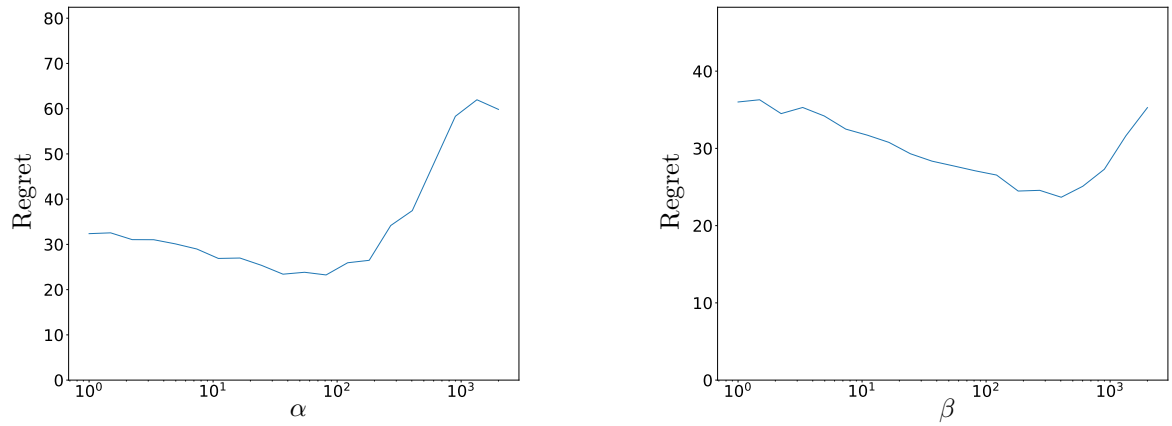
6 Tuning $\delta(t)$: Being under-optimistic is preferable in Markovian bandits

Section 5 provides the natural choice $f \equiv D$ for the span proxy—when bounds on maximal span of the bias function and the delays of the decision epochs are available—and removes the need for a diverging span proxy. In this last section, we are interested in what the best choice for the confidence threshold $\delta(t)$ is, given that the span-proxy is fixed to the value recommended by Section 5. Indeed, while from an asymptotic viewpoint the regret is simply $\mathrm{O}((\frac{D}{\Delta})^2 \log(T))$ for all summable confidence function, in finite horizon, a trade-off appears.

- Be $\delta(t)$ too large, and the confidence intervals are too narrow to contain the true gains, leading to a failure of the optimistic property. As the optimistic bonus $\log(\frac{1}{\delta(t)})$ vanishes, the algorithm's behavior converges toward a greedy algorithm, that pulls the arm with the highest empirical estimator. Then, `UCB-NOM` is very likely to pull bad up to $\approx T$ times, resulting in disastrous performance.
- Be $\delta(t)$ too small, and the algorithm is too conservative: The optimistic bonuses dictate everything that the algorithm does. As the optimistic bonus $\log(\frac{1}{\delta(t)})$ explodes to $+\infty$, the algorithm's behavior converges towards a round-robin process.

In Figure 5, we display the regret of `UCB-NOM`(D) for a reasonably short horizon, as a function of the confidence parameter and for different parameterizations of $\delta(t)$. To avoid running into deceptive conclusions, the environment is specifically designed so that the greedy algorithm (`UCB-NOM` with $\delta = 1$) has no chance to work. But then, even though the environment is unfavorable to greedy algorithms, these experiments suggest that very high confidence parameters are preferred, such as $\delta(t) = \frac{\alpha(t)}{t}$ or $\frac{\beta(t)}{t\log(t)}$ with very high $\alpha(t)$ or $\beta(t)$. In the remaining of this section, we explain why.

[8] For instance, any $\delta(t) = (\frac{1}{t})^{2+\varepsilon}$ with $\varepsilon > 0$ is fine.



(a) Regret of UCB-NOM(D) with horizon $T = 2000$ with confidence parameter $\delta(t) := \frac{\alpha}{t}$. Each point is obtained by averaging over 1000 trajectories.

(b) Regret of UCB-NOM(D) with horizon $T = 2000$ with confidence parameter $\delta(t) := \frac{\beta}{t \log t}$. Each point is obtained by averaging over 1000 trajectories.

Figure 5: Regret of UCB-NOM(D) on a Markovian bandits with two arms. The first arm $a = 1$ is a stochastic bandit with reward $\nu_1 = \text{Bernoulli}(0.3)$. The second arm $a = 2$ is a two states Markovian arm whose transition are, when active, uniform over $\{0, 1\}$, and when passive, deterministically directed to 0. Associated rewards are $\nu_2(0) = 0$ and $\nu_2(1) = \text{Bernoulli}(0.8)$.

In the bonus, the term due to switching cost cannot be improved in general. The environment in Figure 5 is chosen so that this term is chosen in a tight fashion: every time the arm is dropped, the arm goes back to the worst state instantly. For this environment, the switching bonus perfectly corrects the empirical estimator and we can fully focus on the study of the optimistic bonus. So, switching terms are casted aside by looking at bounds on the pseudo-regret directly.

From the proof of Theorem 8, keeping dominant terms, we get the upper-bound of the pseudo-regret is of the form:

$$\mathbf{E}[\mathrm{PR}(T)] \lesssim C_1 \|\Delta\|_1 \int_1^\infty \delta(t)\ \mathrm{d}t + C_2 D^2 \left(\sum_{a \notin \mathcal{A}^*} \frac{1}{\Delta_a} \right) \log\left(\frac{1}{\delta(T)} \right) . \tag{16}$$

where $C_1, C_2 > 0$ are universal constants.[9]

6.1 Confidence functions of the form $\delta(t) = \frac{\alpha}{t}$

The bound (16) encourages to choose $\delta(t) \asymp \frac{1}{t}$, so that the two terms are of the same order—we recover confidence parameters that are polynomial in $\frac{1}{t}$, as suggested by standard stochastic bandits theory (Lattimore & Szepesvári, 2020). Therefore, as a first choice for the confidence function, we take $\delta(t) := \frac{\alpha}{t}$ for some $\alpha \in \mathbf{R}$.

Rewriting the previous upper-bound and trashing a few negligible terms, we obtain

$$\mathbf{E}[\mathrm{PR}(T)] \lesssim \varphi(\alpha) := C_1 \|\Delta\|_1 \alpha \log(T) + C_2 D^2 \left(\sum_{a \notin \mathcal{A}^*} \frac{1}{\Delta_a} \right) \log\left(\frac{T}{\alpha} \right) .$$

Now, computing derivative of φ with respect to α gives the optimal value:

$$\alpha^* = \frac{C_2 D^2 \sum_{a \notin \mathcal{A}^*} \frac{1}{\Delta_a}}{C_1 \|\Delta\|_1 \log(T)} . \tag{17}$$

[9]It is not easy to find a real meaning to the precise values of these constants C_1, C_2, because they are the results of computations that are pretty conservative. Therefore, the discussion revolves around the dependency in Δ and D exclusively.

We should highlight the presence of the $\log(T)$ term which emphasis that our class of function for δ is not fully satisfactory. However, this remains a suitable choice as long as the horizon is not too large, i.e., for horizon that are sub-exponential, typically $T \ll \exp\{D^2 \sum_a \frac{1}{\Delta_a} / \|\Delta\|_1\}$. For such horizons and in the *two-armed case*, we get an optimal optimism of order $(D/\Delta)^2$ up to numerical values: this is a quartic in the problem parameters, and therefore is large very easily. For the environments on which the experiments of Figure 5, we have $(D/\Delta)^2 = 400$ already.

6.2 Confidence functions of the form $\delta(t) = \frac{\beta}{t\log(t)}$

The logarithmic term in the denominator of α^* in Equation (17) is a bit unsatisfactory. To correct it, a simple approach is to take a confidence function of the form $\delta(t) := \frac{\beta}{t\log(t)}$ for $\beta \in \mathbf{R}$. Repeating the same calculation, we find a pseudo-regret upper-bound of the form

$$\mathbf{E}[\mathrm{PR}(T)] \lesssim \varphi(\beta) := C_1 \|\Delta\|_1 \beta \underbrace{\int_{\mathrm{e}}^{T} \frac{1}{t\log t}\,\mathrm{d}t}_{\log\log(T)} + C_2 D^2 \left(\sum_{a \notin \mathcal{A}^*} \frac{1}{\Delta_a}\right) \log\left(\frac{T\log(T)}{\beta}\right).$$

Optimizing φ with respect to β, we obtain

$$\beta^* = \frac{C_2 D^2 \sum_{a \notin \mathcal{A}^*} \frac{1}{\Delta_a}}{C_1 \|\Delta\|_1 \log\log(T)}\,.$$

Again, β^* depends on T, but the dependency in T is mild, as $\log\log(T)$ is basically a numerical constant for any reasonable time horizon (say $T \leq 10^{10}$).[10] We recover the quartic critical value of $(D/\Delta)^2$, as in the parameterization $\delta(t) = \frac{\alpha}{\log(t)}$.

6.3 Comparison with stochastic bandits

The critical value $(D/\Delta)^2$ is reminiscent of what is known in stochastic bandits. Indeed, for such learning problems, it is known that choosing a confidence parameter scaling as $\delta(t) = \frac{\sigma^2}{\Delta^2}\frac{1}{t}$ leads to better regret bounds, see for instance (Lattimore & Szepesvári, 2020, Chapter 9.2).[11] In our setting, the quantity D plays a similar role to the variance proxy of subgaussian random variables: according to Proposition 5, the noise on the empirical estimator of the gain scales as $D\sqrt{n\log(1/\delta)}$, while it is $\sigma^2\sqrt{n\log(1/\delta)}$ for σ^2-subgaussian bandits. However—and this is the crucial difference with stochastic bandits—the noise can have high magnitude even though rewards remain $[0,1]$. This is something that is exclusive to Markovian bandits, because by Hoeffding's Lemma, every random variable supported in $[0,1]$ is $\frac{1}{2}$-subgaussian. Accordingly, Markovian bandits can "simulate" reward distributions with higher noise *yet* values in $[0,1]$, and the optimistic bonuses must be inflated accordingly.

7 Conclusion

In this paper, we considered a class of Markovian bandit learning problems that allow for non-observable states and constrained decision epochs. We argued that pure policies provide an appropriate benchmark for Markovian bandits by introducing a subclass, self-degrading bandits, for which they are optimal. We first provided a lower bound for the regret of rarely-switching algorithms, revealing that strict logarithmic regret remains unreachable—in opposition to the stochastic bandit case. We then introduced `UCB-NOM` and showed that, despite the non-observability of the environment's inner state, non-trivial regret guarantees can be achieved. We have further discussed a structural assumption—a prior knowledge of the bias span—under which the classical $\mathrm{O}(\log(T))$ instance-dependent and $\mathrm{O}(\sqrt{T\log(T)})$ worst-case regrets can be recovered, as soon as the delay between decision epochs remain reasonable.

[10] The reason why we have remaining $\log(-)$ terms is that we are solving equations of the form $ax + b\log(x) = c$, of which the solution involves the Lambert W function, that cannot be expressed analytically.

[11] Lattimore & Szepesvári (2020) is focusing on the case $\sigma = 1$. We refer the reader to their discussion in Chapter 9.2, where they optimize a regret bound of order $n\Delta\delta + \frac{1}{\Delta}\log(\frac{1}{\delta})$. For σ^2-subgaussian bandits with $\sigma \neq 1$, the analogue of their bound is $n\Delta\delta + \frac{\sigma^2}{\Delta}\log(\frac{1}{\delta})$, that is optimized for $\delta \approx \sigma^2/(n\Delta^2)$.

We conclude by mentioning several possible directions for future work. A central assumption in our theory is that the state space $\mathcal{S}_a$ of every arm can be arbitrarily large. It would be interesting to study what happens when the state space $\mathcal{S}_a$ is known in advance. Our preliminary intuition is that in such a setting, the theory would become considerably more intricate. Another avenue is to explore alternative information settings, for instance when states are organized into clusters, and rewards depend on the cluster of the current state. Finally, in line with the well-studied restless bandit framework, it would be interesting to consider scenarios where both activated and non-activated arms yield rewards, possibly dependent on the action.

References

Rajeev Agrawal, MV Hedge, and Demosthenis Teneketzis. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Transactions on Automatic Control*, 33(10): 899–906, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331, 1995.

Peter Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *J. Mach. Learn. Res.*, 3: 397–422, 2002.

Peter Auer and Ronald Ortner. Logarithmic Online Regret Bounds for Undiscounted Reinforcement Learning. *Proceedings of the 19th International Conference on Neural Information Processing Systems*, December 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal Regret Bounds for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.

Soumya Basu, Rajat Sen, Sujay Sanghavi, and Sanjay Shakkottai. Blocking bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/88fee0421317424e4469f33a48f50cb0-Paper.pdf`.

Victor Boone. Optimal regrets in markov decision processes, November 2024. URL `https://victor-boone.github.io/assets/pdf/VBOONE-manuscript-v1.1.pdf`. PhD thesis, see victor-boone.github.io.

Victor Boone and Odalric-Ambrym Maillard. The regret lower bound for communicating markov decision processes, 2025. URL `https://arxiv.org/abs/2501.13013`.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening Exploration in Upper Confidence Reinforcement Learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1056–1066. PMLR, July 2020.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(2):119–137, 1979.

Nicolo Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. *Advances in Neural Information Processing Systems*, 26, 2013.

Wenhan Dai, Yi Gai, Bhaskar Krishnamachari, and Qing Zhao. The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2940–2943. IEEE, 2011.

Wenhan Dai, Yi Gai, and Bhaskar Krishnamachari. Online learning for multi-channel opportunistic access over unknown markovian channels. In *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 64–71. IEEE, 2014.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 459–467, 2014.

E. O. Elliott. Estimates of error rates for codes on burst-noise channels. *IEEE Transactions on Information Theory*, 1963. URL `https://api.semanticscholar.org/CorpusID:60858435`.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7340–7348, 2021.

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

Gilbert. Capacity of burst-noise channels. *Bell System Technical Journal*, 39:1253–1265, 1960.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

Gerhard Haßlinger and Oliver Hohlfeld. The gilbert-elliott model for packet loss in real time services on the internet. In *Messung, Modellierung und Bewertung von Rechen- und Kommunikationssystemen*, 2011. URL `https://api.semanticscholar.org/CorpusID:9995770`.

Thomas Hira, Urtzi Ayesta, Rhonda Righter, and Ina Maria Verloop. Optimal non-preemptive scheduling with time-varying non-observable channels. *ACM SIGMETRICS Performance Evaluation Review*, 53(2): 134–136, 2025.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Daniel Jerison. General mixing time bounds for finite markov chains via the absolute spectral gap, 2013.

Bowen Jiang, Bo Jiang, Jian Li, Tao Lin, Xinbing Wang, and Chenghu Zhou. Online restless bandits with unobserved states. In *International Conference on Machine Learning*, pp. 15041–15066. PMLR, 2023.

Tackseung Jun. A survey on the bandit problem with switching costs. *de Economist*, 152(4):513–541, 2004.

Young Hun Jung and Ambuj Tewari. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Young Hun Jung, Marc Abeille, and Ambuj Tewari. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*, 2019.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the Complexity of Best-Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Junpei Komiyama, Issei Sato, and Hiroshi Nakagawa. Multi-armed bandit problem with lock-up periods. In *Asian Conference on Machine Learning*, pp. 100–115. PMLR, 2013.

Vikram Krishnamurthy. *Partially observed Markov decision processes*. Cambridge university press, 2016.

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

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

Haoyang Liu, Keqin Liu, and Qing Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1968–1971. IEEE, 2011.

Haoyang Liu, Keqin Liu, and Qing Zhao. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3):1902–1916, 2012.

K. Liu, Q. Zhao, and B. Krishnamachari. Dynamic multichannel access with imperfect channel state detection. *IEEE Transactions on Signal Processing*, 58(5):2795–2808, 2010.

Mordechai Mushkin and Israel Bar-David. Capacity and coding for the gilbert-elliot channels. *IEEE Trans. Inf. Theory*, 35:1277–1290, 1989. URL `https://api.semanticscholar.org/CorpusID:7043451`.

Marc Nerlove and Kenneth J Arrow. Optimal advertising policy under dynamic conditions. *Economica*, pp. 129–142, 1962.

Ronald Ortner. Online regret bounds for Markov decision processes with deterministic transitions. *Theoretical Computer Science*, 411(29):2684–2695, 2010. ISSN 0304-3975.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pp. 214–228. Springer, 2012.

Yury Polyanskiy, H Vincent Poor, and Sergio Verdú. Dispersion of the gilbert-elliott channel. *IEEE Transactions on Information Theory*, 57(4):1829–1848, 2011.

Chloé Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9127–9135. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/rouyer21a.html`.

Cem Tekin and Mingyan Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *2011 Proceedings IEEE INFOCOM*, pp. 2462–2470. IEEE, 2011.

Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

Jinghan Wang, Mengdi Wang, and Lin F Yang. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

A General results

In this Appendix, we provide general results on our class of Markovian bandit problems. In Appendix A.1, we prove a simple expression for the score of pure policies, that we use to provide a simple expression for the regret. In Appendix A.2, we deep-dive into the theory of rarely switching algorithms (Definition 2), in which the regret and the pseudo-regret can be related to one another (Proposition 1).

A.1 The score of pure policies and a simple formula for the regret

In the main text, we have stated without proof that the score of pure policies could be linked to their gain and bias functions, with $V^{\pi_a}(t) = (t-1)g_a + \mathbf{E}^{\pi_a}[b_a(s_a^{\text{init}}) - b_a(S_a(t))]$. This formula is a well-known result in Markov Reward Processes. Below we restate it as a lemma and include its proof for completeness.

Lemma A.1 (Score of pure policies)**.** *Let $M \in \mathcal{M}$ be an instance. Let π_a be the pure policy pulling arm $a \in \mathcal{A}$ exclusively. For every horizon $T \geq 1$, we have*

$$\mathbf{E}^{M,\pi_a}\left[\sum_{t=1}^{T-1} R(t+1)\right] = (T-1)g_a + \mathbf{E}^{M,\pi_a}[b_a(s_0) - b_a(S_a(T))]. \tag{A.1}$$

Note that in particular, this implies that $|V^{\pi_a}(t) - (t-1)g_a| \leq \mathrm{sp}(b_a)$ for all $t \geq 1$ and $a \in \mathcal{A}$.

Proof. The result follows using the Poisson equation of $\pi \equiv \pi_a$ that states that for all $s_a \in \mathcal{S}_a$, we have $g_a + b_a(s_a) = r_a(s_a) + p_a^1(s_a)b_a$. More precisely, we have

$$\begin{aligned}
\mathbf{E}^{M,\pi}\left[\sum_{t=1}^{T-1} R(t+1)\right] - (T-1)g_a &= \mathbf{E}^{M,\pi}\left[\sum_{t=1}^{T-1}(R(t+1) - g_a)\right] \\
&\overset{\circ 1}{=} \begin{cases} \mathbf{E}^{M,\pi}\left[\sum_{t=1}^{T-1}(R(t+1) - r_a(S_a(t)))\right] \\ +\mathbf{E}^{M,\pi}\left[\sum_{t=1}^{T-1}(b_a(S_a(t)) - b_a(S_a(t+1)))\right] \\ +\mathbf{E}^{M,\pi}\left[\sum_{t=1}^{T-1}\left(b_a(S_a(t+1)) - p_a^1(S_a(t))b_a\right)\right] \end{cases} \\
&\overset{\circ 2}{=} \mathbf{E}^{M,\pi}[b_a(S_a(1)) - b_a(S_a(T))],
\end{aligned}$$

where ($\circ$1) follows from the Poisson equation of $\pi \equiv \pi_a$; and ($\circ$2) uses that the first and third terms in ($\circ$1) are martingale difference sequences, and that the second term is a telescopic sum. □

As a direct consequence of Lemma A.1, the regret can be approximated by $\sum_{t=1}^{T-1}(g_* - R(t))$, where $g_* = \max_{a\in\mathcal{A}} g_a$. By Lemma A.1 indeed, by letting π be the optimal pure policy, we have $V^{\pi}(T) = (T-1)g^* \pm \max_{a\in\mathcal{A}} \mathrm{sp}(b_a)$. This approximation can be traced back at least to (Auer et al., 2009), and will be the starting point for our upper and lower bounds on the regret.

Corollary A.2 (Simple form for the regret)**.** *Let $M \in \mathcal{M}$ be an instance. For every learning algorithm Λ and every horizon $T \geq 1$, the regret can be approximated by $\sum_{t=1}^{T-1}(g_* - R(t))$, with an error term of at most $\max_{a\in\mathcal{A}} \mathrm{sp}(b_a)$, that is,*

$$\left|\mathrm{Reg}(T; M, \Lambda) - \mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T-1}(g_* - R(t+1))\right]\right| \leq \max_{a\in\mathcal{A}} \mathrm{sp}(b_a). \tag{A.2}$$

The error $\max_{a\in\mathcal{A}} \mathrm{sp}(b_a)$ in Equation (A.2) does not grow with $T \geq 1$. Since the regret is eventually growing as $\log(T)$ in the instance dependent setting and as $\sqrt{T}$ in the worst-case setting, the latter constant can safely be ignored.

A.2 Proof of Proposition 1: Rarely switching algorithms

The link between the regret and the pseudo-regret for rarely switching algorithms plays a crucial role in our analysis, as it connects the regret to the number of times suboptimal arms are pulled. For an arbitrary learner, such a link does not generally exist, since pulling suboptimal arms does *not* necessarily imply that the regret increases accordingly: a smart controller could deliberately pull suboptimal arms when the optimal arm is in a bad state. However, ψ-rarely switching algorithms preclude such behavior, which allows us to obtain a direct connection between the regret and the pseudo-regret, via the switching function ψ. This result was stated in Proposition 1 and we prove a more precise version of it below in Proposition A.3.

Proposition A.3 (Regret and pseudo-regret)**.** *Let $M \in \mathcal{M}$ be an arbitrary model with a unique optimal arm $a^* \in \mathcal{A}$. Let Λ be a learning algorithm that satisfies a ψ-rarely switching property. Denote $\Delta_0 := \min\{g_* - g_a : g_a < g_*\}$ the minimum gain-gap. For every horizon $T \geq 1$, we have*

$$|\mathrm{Reg}(T) - \mathbf{E}[\mathrm{PR}(T)]| \leq \max_{a\in\mathcal{A}}\big(\mathrm{sp}(b_a)\big) \cdot \left(\frac{2|\mathcal{A}|}{\Delta_0} \cdot \psi(\mathbf{E}[\mathrm{PR}(T)]) + \left(\frac{2|\mathcal{A}|}{\Delta_0} - 1\right)C_\psi + 1\right), \tag{A.3}$$

where PR *is the pseudo-regret* (6) *and $C_\psi \in \mathbf{R}_+$ is the constant given by Lemma A.6 below.*

To prove Proposition A.3, we first state Lemma A.4, which shows that, for any learner, the expected regret and pseudo-regret can be linked through the number of switches. The result then follows by bounding that number of switches, for which we rely on the technical Lemmas A.5 and A.6.

Lemma A.4 (Regret and number of switches)**.** *Let $M \in \mathcal{M}$ be an arbitrary model. Let Λ be a learning algorithm. For every horizon $T \geq 1$, we have*

$$\Big|\mathrm{Reg}(T; M, \Lambda) - \mathbf{E}^{M,\Lambda}[\mathrm{PR}(T)]\Big| \leq \max_{a\in\mathcal{A}} \mathrm{sp}(b_a) \cdot \left(1 + \mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T-1} \mathbf{1}(A(t-1) \neq A(t))\right]\right). \tag{A.4}$$

Proof. Let (τ_k) be the sequence of decision epochs. Since Λ is non-preemptive, we know that $A(t) = A(\tau_k)$ for all $t \in \{\tau_k, \ldots, \tau_{k+1} - 1\}$. We clip decision epochs to $\{1, \ldots, T-1\}$ by introducing $\tau'_k := \min\{\tau_k, T-1\}$. We have

$$\begin{aligned}
\mathrm{Reg}(T; M, \Lambda) &\overset{\circ 1}{\leq} \mathbf{E}^{M,\Lambda}\left[\sum_{k=1}^{\infty}\sum_{t=\tau'_k}^{\tau'_{k+1}-1}(g_* - R(t+1))\right] + \max_{a\in\mathcal{A}} \mathrm{sp}(b_a) \\
&\overset{\circ 2}{=} \mathbf{E}^{M,\Lambda}\left[\sum_{k=1}^{\infty}\sum_{t=\tau'_k}^{\tau'_{k+1}-1}\Big(g_* - g_{A(\tau'_k)} - b_{A(\tau'_k)}(S_{A(\tau'_k)}(t)) + b_{A(\tau'_k)}(S_{A(\tau'_k)}(t+1))\Big)\right] + \max_{a\in\mathcal{A}} \mathrm{sp}(b_a) \\
&= \mathbf{E}^{M,\Lambda}\left[\sum_{k=1}^{\infty}(\tau'_{k+1} - \tau'_k)\Big(g_* - g_{A(\tau'_k)}\Big)\right] + \mathbf{E}^{M,\Lambda}\left[\sum_{k=1}^{\infty}\Big(e_{S(\tau'_{k+1})} - e_{S(\tau'_k)}\Big)b_{A(\tau'_k)}\right] + \max_{a\in\mathcal{A}} \mathrm{sp}(b_a) \\
&\overset{\circ 3}{\leq} \mathbf{E}^{M,\Lambda}\left[\sum_{k=1}^{\infty}(\tau'_{k+1} - \tau'_k)\Big(g_* - g_{A(\tau'_k)}\Big)\right] + \max_{a\in\mathcal{A}} \mathrm{sp}(b_a) \cdot \left(1 + \mathbf{E}^{M,\Lambda}\left[\sum_{k=1}^{\infty}\mathbf{1}\big(A(\tau'_{k-1}) \neq A(\tau'_k)\big)\right]\right) \\
&\overset{\circ 4}{\leq} \mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}}(g_* - g_a)N_a(T)\right] + \max_{a\in\mathcal{A}} \mathrm{sp}(b_a) \cdot \left(1 + \mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T-1}\mathbf{1}(A(t-1) \neq A(t))\right]\right),
\end{aligned} \tag{A.5}$$

where (°1) invokes Corollary A.2; (°2) uses that at time t with $\tau'_k \leq t < \tau'_{k+1}$ the arm $A(\tau'_k)$ is activated and hence, $R(t+1)$ can be replaced using the Poisson equation of the Markov reward process under the pure policy $\pi_{A(\tau'_k)}$; (°3) uses that the right-most term is nearly telescopic (see Lemma D.9); and (°4) follows by definition of visit counts. In Equation (A.5), we recognize the expected pseudo-regret in the first term, providing the desired upper bound of the regret with respect to the pseudo-regret. The lower bound is obtained similarly. □

Lemma A.5 (Bound on the number of switches)**.** *Let $M \in \mathcal{M}$ be an arbitrary model. Consider a learning algorithm that satisfies a ψ-rarely switching property. Then, for all horizon $T \geq 1$ and arm $a \in \mathcal{A}$, we have*

$$\sum_{t=1}^{T-1} \mathbf{1}(A(t-1) \neq a, A(t) = a) \leq \sum_{a' \neq a} \psi(N_{a'}(T)). \tag{A.6}$$

Proof. Because the algorithm is rarely switching, there is a non-decreasing sublinear function $\psi : \mathbf{R}_+ \to \mathbf{R}$ such that the number of switches off is bounded as $\sum_{t=1}^{T-1} \mathbf{1}(A(t-1) = a, A(t) \neq a) \leq \psi(N_a(t))$ for all arm $a \in \mathcal{A}$ and $T \geq 1$. Then, we have

$$\begin{aligned}\sum_{t=1}^{T-1} \mathbf{1}(A(t-1) \neq a, A(t) = a) &\overset{\circ 1}{\leq} \sum_{a' \neq a} \sum_{t=1}^{T-1} \mathbf{1}(A(t-1) = a', A(t) \neq a') \\ &\overset{\circ 2}{\leq} \sum_{a' \neq a} \psi(N_{a'}(T)),\end{aligned}$$

where (°1) holds because switching onto arm a means that we are switching off some arm $a' \neq a$; (°2) follows from the rarely switching property. This concludes the proof. □

Lemma A.6 (Homogeneity of concave functions)**.** *Let $f : \mathbf{R}_+ \to \mathbf{R}_+$ be a non-negative smooth concave function and let $C_f := \sup_{x \neq 1}(f(x) - xf(1))/(x-1)$. Then $C_f < \infty$ and*

$$\forall \lambda \geq 0, \forall x \geq 0, \quad f((1+\lambda)x) \leq (1+\lambda)f(x) + \lambda \cdot C_f. \tag{A.7}$$

Proof. First, we check that $C_f < \infty$. Let $\psi(x) := (f(x) - xf(1))/(x-1)$ for $x \neq 1$, which is a smooth function on its domain. As f is concave, it is bounded by a linear function $\alpha x + \beta$ and we have $\limsup_{x \to \infty} \psi(x) \leq \alpha - f(1)$. We are left to study the convergence properties of $\psi(x)$ as $x \to 1$. Because f is smooth, we have the Taylor approximation

$$\psi(1+u) = \frac{f(1+u) - (1+u)f(1)}{u} = \frac{u(f'(1) - f(1)) + \mathrm{O}(u^2)}{u} = f'(1) - f(1) + \mathrm{O}(u)$$

and ψ is therefore continuous at 1. So ψ is bounded and $C_f < \infty$.

Now, we establish Equation (A.7). Since f is smooth and concave, we have

$$f((1+\lambda)x) \leq f(x) + \lambda x f'(x). \tag{A.8}$$

Now, we bound $xf'(x)$ with respect to $f(x)$ and C_f. By definition of C_f, we have $(x-1)C_f \geq f(x) - xf(1)$ for all $x \geq 0$. Reordering terms, we get $f(x) \leq x(f(1) + C_f \cdot (1 - \frac{1}{x}))$. From a straight-forward computation, we check that the function $g(x) := f(1) + C_f \cdot (1 - \frac{1}{x})$ is smooth and satisfies $x^2 g'(x) = C_f$, while $f'(x) = xg'(x) + g(x)$. Combined, we get

$$xf'(x) = x^2 g'(x) + xg(x) = C_f + f(x). \tag{A.9}$$

Combining Equations (A.8) and (A.9), this completes the proof. □

We can now prove Proposition A.3.

Proof of Proposition A.3. By Lemma A.4, controlling the difference between the expected regret and the pseudo-regret reduces to bounding the number of switches in Equation (A.4). Let $a^* = \arg\max_{a \in \mathcal{A}} g_a$ be the unique optimal arm. Let $\Delta_0 := \min\{g_* - g_a : a \neq a^*\}$ be the gain-gap between the optimal and the best suboptimal arms. We have

$$\mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T-1} \mathbf{1}(A(t-1) \neq A(t))\right]$$

$$
\begin{aligned}
&= \mathbf{E}^{M,\Lambda}\left[\sum_{a\neq a^*}\sum_{t=1}^{T-1}\mathbf{1}(A(t-1)\neq a, A(t)=a)\right] + \mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T-1}\mathbf{1}(A(t-1)\neq a^*, A(t)=a^*)\right] \\
&\overset{\circ 1}{\leq} \mathbf{E}^{M,\Lambda}\left[\sum_{a\neq a^*}\psi(N_a(T))\right] + \mathbf{E}^{M,\Lambda}\left[\sum_{a\neq a^*}\psi(N_a(T))\right] \\
&\overset{\circ 2}{\leq} 2|\mathcal{A}|\cdot \mathbf{E}^{M,\Lambda}\left[\psi\left(\frac{1}{\Delta_0}\sum_{a\in\mathcal{A}}(g_*-g_a)N_a(T)\right)\right] \\
&\overset{\circ 3}{\leq} 2|\mathcal{A}|\cdot \psi\left(\mathbf{E}^{M,\Lambda}\left[\frac{1}{\Delta_0}\sum_{a\in\mathcal{A}}(g_*-g_a)N_a(T)\right]\right) \\
&\overset{\circ 4}{\leq} \frac{2|\mathcal{A}|}{\Delta_0}\cdot \psi\left(\mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}}(g_*-g_a)N_a(T)\right]\right) + \left(\frac{2|\mathcal{A}|}{\Delta_0}-1\right)C_\psi,
\end{aligned}
$$

where ($^\circ$1) is obtained by bounding the first term using the ψ-rarely switching property and the second using Lemma A.5; ($^\circ$2) follows by monotonicity of ψ; ($^\circ$3) follows by Jensen's inequality; and ($^\circ$4) holds following a homogeneity property of smooth concave functions, where $C_\psi < \infty$ is a constant, see Lemma A.6. □

B Self-degrading restless bandits

In this appendix, we aim to provide proofs for the claims of Section 3. This includes the proof of asymptotic optimality of pure policies within self-degrading bandits. In particular, we prove Propositions 2 and 3, and show that the models described in Example 2 and Example 3 are instances of *self-degrading* restless bandits.

Recall the definition of a *self-degrading* restless bandit from the main body:

Definition B.1 (Self-degrading restless bandits)**.** Given an instance of restless bandit $M \in \mathcal{M}$ together with an activation set $\mathcal{U}_\tau$, an arm $a \in \mathcal{A}$ is said *self-degrading* if, whatever the controller Λ respecting $\mathcal{U}_\tau$,

$$A(t) \neq a \implies \mathbf{E}^{M,\Lambda}[b_a(S_a(t+1))|H_o(t)] \leq \mathbf{E}^{M,\Lambda}[b_a(S_a(t))|H_o(t)] \ .$$

The instance M is said *self-degrading* if all of its arms are. The set of all self-degrading instances of Markovian bandits is written $\mathcal{M}_{\mathrm{sd}}$.

B.1 Proof of Propositions 2 and 3: Regret and pseudo-regret in self-degrading instances

Here, we show that the regret is lower-bounded by the expected pseudo-regret in self-degrading instances, by proving Proposition 3 from the main text. Coupled with Lemma A.1 that expresses the score of the pure policies with respect to the gain and the bias, the optimality of pure policies in self-degrading instances (Proposition 2) is immediate.

Proposition 3 (Regret and pseudo-regret in self-degrading instances)**.** *The regret is bounded below by the pseudo-regret on self-degrading instances. That is, for every self-degrading instance of restless bandits $M \in \mathcal{M}_{\mathrm{sd}}$, there exists a constant $C_M < \infty$ such that, whatever the learning algorithm Λ,*

$$\forall T \geq 1, \qquad \mathrm{Reg}(T; M, \Lambda) \geq \mathbf{E}^{M,\Lambda}[\mathrm{PR}(T)] - C_M \ .$$

Proof of Proposition 3. For all arm $a \in \mathcal{A}$ consider the sequences of stopping times

$$\tau_k^a := \inf\{t > \tilde{\tau}_k^a : A(t) = a\} \wedge T, \qquad \tilde{\tau}_{k+1}^a := \inf\{t > \tau_k^a : A(t) \neq a\} \wedge T$$

which are respectively the first For any learning algorithm Λ and $s_0 \in \mathcal{S}$,

$$V^\Lambda(T; M) = \mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T} R(t)\right]$$

$$
\begin{aligned}
&= \mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}}\sum_{k=1}^{\infty}\sum_{t=\tau_k^a}^{\tilde{\tau}_{k+1}^a-1} R(t)\right] \\
&\overset{\circ 1}{=} \mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}}\sum_{k=1}^{\infty}\sum_{t=\tau_k^a}^{\tilde{\tau}_{k+1}^a-1} (g_a + b(S_a(t)) - b(S_a(t+1)))\right] \\
&\leq \mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}}\left(g_a N_a(T) + \sum_{k=1}^{\infty}\big(b_a(S_a(\tau_k^a)) - b_a(S_a(\tilde{\tau}_{k+1}^a))\big)\right)\right] \\
&= \mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}} g_a N_a(T)\right] + \mathbf{E}^{M,\Lambda}\left[\sum_{a\in\mathcal{A}}\left(b_a(S_a(1)) + \sum_{k=2}^{\infty}\big(b_a(S_a(\tau_{k+1}^a)) - b_a(S_a(\tilde{\tau}_{k+1}^a))\big)\right)\right] \\
&\overset{\circ 2}{\leq} (T-1)g_* - \mathbf{E}^{M,\Lambda}[\mathrm{PR}(T)] + \mathrm{O}(1)
\end{aligned}
$$

where (°1) follows from the Poisson equation; (°2) stems from both the definition of the pseudo-regret and the *self-degrading* property. It is now sufficient to note that

$$
V^{\pi_*}(T;M) = \mathbf{E}^{\pi_*,M}\left[\sum_{t=1}^{T} R(t)\right] \overset{\circ 3}{=} Tg_* + \mathrm{O}(1),
$$

where (°3) follows directly from the Poisson equation with π_* the policy playing only the optimal arm, to conclude:

$$
\mathrm{Reg}(T) = V^{\pi_*}(T;M) - V^{\Lambda}(T;M) \geq \mathbf{E}^{M,\Lambda}[\mathrm{PR}(T)] - \mathrm{O}(1). \qquad \square
$$

Proposition 2 (Pure policies are optimal in self-degrading instances)**.** *The regret is bounded below on self-degrading instances. That is, for every self-degrading instance of restless bandits $M \in \mathcal{M}_{\mathrm{sd}}$, there exists a constant $C_M < \infty$ such that, whatever the learning algorithm Λ, we have*

$$
\mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T} R(t)\right] \leq \max_{a\in\mathcal{A}} V^{\pi_a}(T;M) + C_M \ .
$$

Proof of Proposition 2. Let π be an optimal pure policy, i.e., maximizing g_a. Let Λ be an arbitrary learning algorithm. We have

$$
\begin{aligned}
\mathbf{E}^{M,\Lambda}\left[\sum_{t=2}^{T} R(t)\right] &\overset{\circ 1}{=} V^{\pi}(T) - \mathrm{Reg}(T;M,\Lambda) \\
&\overset{\circ 2}{\leq} V^{\pi}(T) - \mathbf{E}^{M,\Lambda}[\mathrm{PR}(T)] + C_M \\
&\overset{\circ 3}{\leq} V^{\pi}(T) + \mathrm{O}(1)
\end{aligned}
$$

where (°1) simply unfolds the definition of the regret; (°2) bounds the right-term via Proposition 3; and (°3) follows by non-negativity of the pseudo-regret. So π is asymptotically optimal. $\square$

B.2 Examples of self-degrading restless bandits

In the following, we come back to the examples from Section 3.2 and prove they belong to the self-degrading restless bandit family. As in the main text, we begin with the case of Gilbert-Elliott communication channels and follow up by a toy model inspired from the Nerlove-Arrow concept in marketing.

B.2.1 Example: Non-preemptive scheduling with Gilbert-Elliott channels

Proposition B.1. *Consider a restless bandit model in which, every arm $a \in \mathcal{A}$ corresponds to a two-states MDP with $\mathcal{S}_a = \{\texttt{OFF}, \texttt{ON}\}$, transition kernels*

$$p_a^1 = p_a^0 = \begin{pmatrix} 1 - y_a & y_a \\ x_a & 1 - x_a \end{pmatrix}, \qquad x_a, y_a \in (0, 1)$$

and reward distribution satisfying

$$\nu_a(\texttt{OFF}) = \text{Dirac}(0), \qquad \nu_a(\texttt{ON}) = \text{Bernoulli}(\mu_a), \quad \mu_a \in (0, 1).$$

Moreover, if each arm is positively autocorrelated, i.e. $1 - x_a - y_a > 0$, the activation set is $\mathcal{U}_\tau = \{1\}$ and the initial state is $s_a^{init} = \texttt{ON}$. Then, the model $M_a \equiv (\mathcal{S}_a, s_a^{init}, \{0, 1\}, \nu_a, p_a^0, p_a^1)$ is a self-degrading restless bandit.

Before proving Proposition B.1, we first introduce a few lemmas. An important part of the proof is the reformulation of the problem: we replace the full history of observations by a *belief state*, that is, a probability distribution over the hidden states representing the information gathered from past observations. Then, considering the evolution of said belief state, we obtain a *belief MDP*—an MDP on the simplex. The belief state is known to be a *sufficient statistic* of the original problem: an optimal policy for the belief MDP can always be translated back into an optimal policy for the original problem. This is a standard approach from the POMDP literature (see (Krishnamurthy, 2016) for a comprehensive introduction to the subject). In the context of Gilbert-Elliott channel, there are only two states. Therefore, we can characterize the whole belief MDP by considering the evolution of the probability of the current state being ON given past observations. It should be noted that to prove the *self-degrading* property, we only need to consider the dynamic of the belief state under the passive action.

Under the passive action, nothing is observed about the arm. Therefore, the belief state just tends to drift towards the stationary distribution of the Markov chain. The following lemma quantifies this drift.

Lemma B.2 (Dynamics of belief state under passive action)**.** *For all arm $a \in \mathcal{A}$ and all $t \in \mathbf{N}$,*

$$A(t) \neq a \implies \beta_a(t+1) = \beta_a(t)(1 - x_a - y_a) + y_a$$

where $\beta_a(t) := \mathbf{P}[S_a(t) = \texttt{ON} \mid H_o(t)]$ is the belief state, i.e., the probability to be in state ON under the current observations.

Proof. We write $\mathbf{P} \equiv \mathbf{P}^{M,\Lambda}$ and $\mathbf{E} \equiv \mathbf{E}^{M,\Lambda}$ all throughout. From the definition of β_a,

$$\begin{aligned}
\beta_a(t+1) &:= \mathbf{P}[S_a(t+1) = \texttt{ON} \mid H_o(t+1)] \\
&\overset{\circ 1}{=} \mathbf{E}[\mathbf{P}[S_a(t+1) = \texttt{ON} \mid H_o(t)] \mid H_o(t+1)] \\
&\overset{\circ 2}{=} \mathbf{P}[S_a(t) = \texttt{ON} \mid H_o(t)] p_a(\texttt{ON} \mid \texttt{ON}) + \mathbf{P}[S_a(t) = \texttt{OFF} \mid H_o(t)] p_a(\texttt{ON} \mid \texttt{OFF}) \\
&= \beta_a(t)(1 - x_a) + (1 - \beta_a(t)) y_a \\
&= (1 - x_a - y_a)\beta_a(t) + y_a
\end{aligned}$$

where ($\circ$1) stems from the tower property and ($\circ$2) follows from the fact that, under $A(t) \neq a$, nothing is observed for arm a. Hence, update is just the one-step transition of the Markov chain. □

Lemma B.3 (Lower bound on the belief state)**.** *Fix $a \in \mathcal{A}$, and let $x_a, y_a \in (0, 1)$ such that $1 - x_a - y_a > 0$. Then,*

$$A(\tau_k) \neq a \implies \forall t \in \{\tau_k, \dots, \tau_{k+1}\}, \quad \beta_a(t) > \frac{y_a}{x_a + y_a}.$$

In particular, due to the non-preemptive constraint,

$$A(t) \neq a \implies \beta_a(t) > \frac{y_a}{x_a + y_a}.$$

Proof. We write $\mathbf{P} \equiv \mathbf{P}^{M,\Lambda}$ and $\mathbf{E} \equiv \mathbf{E}^{M,\Lambda}$ all throughout. We first note that, due to the constraint induced by the activation set $\mathcal{U}_\tau = 1$, if arm a was activated at τ_k, on the next decision epoch τ_{k+1}, the belief state of that arm is

$$\beta_a(\tau_{k+1}) = \mathbf{P}[S_a(t) = \texttt{ON}|H_o(\tau_{k+1})] \overset{\circ 1}{=} \mathbf{P}[S_a(\tau_{k+1}) = \texttt{ON}|S_a(\tau_{k+1}-1) = \texttt{ON}] \overset{\circ 2}{=} 1 - x_a \overset{3}{>} \frac{y_a}{x_a + y_a}$$

where (°1) follows from the Markov property; (°2) is due to the impossibility a reward if $S_a(\tau_{k+1}-1) = \texttt{OFF}$; and (°3) stems from the positive autocorrelation assumption.

Therefore, we only need to cover what happens when the belief state follows a passive dynamic. We proceed by induction. Due to the fixed initial state $s_a^{\text{init}} = \texttt{ON}$, we have that $\beta(0) = 1 > \frac{y_a}{x_a+y_a}$. Suppose now that for $t > 0$ we have $\beta_a(t) > \frac{y_a}{x_a+y_a}$, then

$$\beta_a(t+1) \overset{\circ 1}{=} (1 - x_a - y_a)\beta_a(t) + y_a \overset{\circ 2}{>} (1 - x_a - y_a)\frac{y_a}{x_a + y_a} + y_a = \frac{y_a}{x_a + y_a}$$

where (°1) follows from Lemma B.2; and (°2) stems from the induction hypothesis $\beta_a(t) > \frac{y_a}{x_a+y_a}$ coupled with the positive autocorrelation assumption $1 - x_a - y_a > 0$. □

Lemma B.4 (Monotony of belief state under passive action)**.** *Let $a \in \mathcal{A}$, x_a and y_a are such that $1-x_a-y_a > 0$. Then for all $t > 0$,*

$$A(t) \neq a \implies \beta_a(t) \geq \beta_a(t+1) \qquad \mathbf{P}\text{-}a.s.$$

where $\beta_a(t) := \mathbf{P}[S_a(t) = \texttt{ON} \mid H_o(t)]$.

Proof. Fix $a \in \mathcal{A}$. Under $A(t) \neq a$ we have,

$$\beta_a(t+1) - \beta_a(t) \overset{\circ 1}{=} -\beta_a(t)(x_a + y_a) + y_a \overset{\circ 2}{\leq} 0$$

where (°1) follows from Lemma B.2; and (°2) stems from Lemma B.3. □

The last lemma allows us to compare the bias of the two hidden states ON and OFF.

Lemma B.5 (Bias of the Gilbert-Elliott channel)**.** *For all arm $a \in \mathcal{A}$ we have*

$$b_a(\texttt{ON}) \geq b_a(\texttt{OFF}).$$

Proof. Using Poisson equation, we obtain for all arm $a \in \mathcal{A}$

$$\frac{\mu_a y_a}{x_a + y_a} + b_a(\texttt{ON}) = \mu_a + (1 - x_a)b_a(\texttt{ON}) + x_a b_a(\texttt{OFF}).$$

Which gives after a few algebraic manipulations

$$\begin{aligned} b_a(\texttt{ON}) &= \frac{\mu_a x_a}{x_a + y_a} + (1 - x_a)b_a(\texttt{ON}) + x_a b_a(\texttt{OFF}) \\ &= \frac{\mu_a}{x_a + y_a} + b_a(\texttt{OFF}) \geq b_a(\texttt{OFF}). \end{aligned}$$ □

We may now prove Proposition B.1.

Proof of Proposition B.1. We write $\mathbf{P} \equiv \mathbf{P}^{M,\Lambda}$ and $\mathbf{E} \equiv \mathbf{E}^{M,\Lambda}$ all throughout. Fix arm $a \in \mathcal{A}$. For all $t > 0$ such that $A(t) \neq a$, we have

$$\begin{aligned} \mathbf{E}[b_a(S(t+1)) \mid H_o(t)] &\overset{\circ 1}{=} \sum_{s \in \{\texttt{ON},\texttt{OFF}\}} b(s)\, \mathbf{E}[\mathbf{1}(S_a(t+1) = s) \mid H_o(t)] \\ &\overset{\circ 2}{=} \mathbf{E}[\beta_a(t+1) \mid H_o(t)]b_a(\texttt{ON}) + (1 - \mathbf{E}[\beta_a(t+1) \mid H_o(t)])b_a(\texttt{OFF}) \\ &\overset{\circ 3}{\leq} \beta_a(t)b_a(\texttt{ON}) + (1 - \beta_a(t))b_a(\texttt{OFF}) \\ &= \mathbf{E}[b_a(S(t)) \mid H_o(t)] \end{aligned}$$

where (°1) follows from Lemma B.4 and Lemma B.5. □

B.2.2 Example: Nerlove-Arrow toy model

In this paragraph, we prove that the Nerlove-Arrow toy model from Example 3 is self-degrading.

Proposition B.6. *Consider a restless bandit model in which, for all arm $a \in \mathcal{A}$,*

$$\forall s \leq s', \;\; p_a^1(s) \leq_{\text{st}} p_a^1(s') \quad \textit{and} \quad p_a^0(s) \leq_{\text{st}} e_s,$$

where $\leq_{\text{st}}$ is the inequality with respect to the first-order stochastic ordering; the mean reward satisfies

$$r_a(s) = \mu_a - f_a(s)$$

where $\mu_a > 0$ and $f_a : \mathcal{S}_a \to \mathbf{R}_+$ is an decreasing function such that

$$\sum_{t=0}^{\infty} (p_a^1)^t(s) f_a < \infty$$

for any initial point $s \in \mathcal{S}_a$; and the activation set is $\mathcal{U}_\tau = [0, 1]$ i.e. every step is a decision epoch.

Then, the induced restless bandit is self-degrading.

Lemma B.7 (Bias representation)**.** *The gain and the bias of the Nerlove-Arrow toy model are respectively*

$$g_a := \mu_a, \quad \textit{and} \quad b_a : s \mapsto -\sum_{t=0}^{\infty} (p_a^1)^t(s) f_a.$$

Proof. For all arm $a \in \mathcal{A}$, let us consider the function

$$\ell_a : s \mapsto -\sum_{t=0}^{\infty} (p_a^1)^t(s) f_a. \tag{B.1}$$

Then, we can show the couple (μ_a, ℓ_a) satisfy the Poisson equation,

$$\mu_a + \ell_a(s) = \mu_a - \sum_{t=0}^{\infty} (p_a^1)^t(s) f_a = \mu_a - \sum_{t=1}^{\infty} (p_a^1)^t(s) f_a - f_a(s) = r_a(s) + p_a^1 \ell(s)$$

hence, μ_a and ℓ_a are respectively the gain and the bias of the arm. □

Lemma B.8 (Monotonicity of the bias function)**.** *For the Nerlove-Arrow toy model, the bias function $s \mapsto b_a(s)$ is non-increasing.*

Proof. Fix an arm a. Let $s \leq s' \in \mathbf{N}$. Then,

$$b_a(s) \overset{\circ 1}{=} -\sum_{t=0}^{\infty} (p_a^1)^t(s) f_a \overset{\circ 2}{\leq} -\sum_{t=0}^{\infty} (p_a^1)^t(s') f_a \overset{\circ 1}{=} b_a(s')$$

where (°1) stems from Lemma B.7; and (°2) follows from the stochastic dominance $p_a^1(s) \leq_{\text{st}} p_a^1(s')$ and the fact that f_a is decreasing. □

Proof of Proposition B.6. We write $\mathbf{P} \equiv \mathbf{P}^{M,\Lambda}$ and $\mathbf{E} \equiv \mathbf{E}^{M,\Lambda}$ all throughout. Fix arm $a \in \mathcal{A}$. For all $t > 0$ such that $A(t) \neq a$ we have

$$\begin{aligned}
\mathbf{E}[b_a(S_a(t+1)) \mid H_o(t)] &\overset{\circ 1}{=} \mathbf{E}[\mathbf{E}[b_a(S_a(t+1)) \mid S_a(t)] \mid H_o(t)] \\
&\overset{\circ 2}{=} \mathbf{E}\big[p_a^0(S_a(t)) b_a \mid H_o(t)\big] \\
&\overset{\circ 3}{\leq} \mathbf{E}[b_a(S_a(t)) \mid H_o(t)]
\end{aligned}$$

where (°1) follows from the tower property; (°2) stems from $A(t) \neq a$; and (°3) is a consequence of both Lemma B.8 and the stochastic dominance property $p_a^0(s) \leq_{\text{st}} e_s$. □

C Regret lower bounds (Theorems 4 and 7)

In this Appendix, we provide the proofs of the two regret lower bounds of the main text (Theorems 4 and 7). Our proof technique follows a well-established line that we have presented in the proof sketch of Theorem 4. The core of the argument consists in finding two instances M_1 and M_2 that (1) are difficult to distinguish from a statistical view points and (2) do not have the same optimal arm. Using a change of measure (Appendix C.1, see Theorems C.1 and C.2), we can bind the behavior of the learning algorithm in M_1 to its behavior on M_2. Then, the consistency of the algorithm forces it to play mostly the optimal arm of M_2 when running on M_2; However, because the behaviors on M_1 and M_2 are bounded, this gives information on what the algorithm is doing on M_1. For a well-chosen M_2 (see Lemma C.4), we derive a lower bound on the number of pulls of suboptimal arms. In turn, this provides a lower bound on the pseudo-regret that we convert to a lower bound on the regret using the ψ-rarely switching property (Proposition A.3).

Articulating this argument around two changes of measure of different style, we obtain the worst-case regret lower bound (Appendix C.3) and the instance dependent regret lower bound (Appendix C.2).

C.1 Changes of measure and an idealistic state transform

We start by providing a refresher on the change of measure machinery.

Given two probability measures $\mathbf{P}$ and $\mathbf{P}'$ over the same probability space satisfying $\mathbf{P} \ll \mathbf{P}'$, recall that the *relative entropy* from $\mathbf{P}'$ to $\mathbf{P}$ is given by $\mathrm{KL}(\mathbf{P}||\mathbf{P}') := \int \log(\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{P}')\,\mathrm{d}\mathbf{P}$, where $\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{P}'$ is the Radon-Nikodym derivative of $\mathbf{P}$ with respect to $\mathbf{P}'$.

In this work, we will use a different style of change of measure, depending on the kind of regret lower bound of interest. The first is known as Bretagnolle-Huber inequality (Theorem C.1) and will be used for the worst-case regret lower bound (Theorem 7).

Theorem C.1 (Bretagnolle-Huber inequality, Bretagnolle & Huber (1979))**.** *Let $\mathbf{P} \ll \mathbf{P}'$ be two probability measures over the same probability space, and let $\mathcal{E}$ be an event. Then*

$$\mathbf{P}(\mathcal{E}) + \mathbf{P}'(\mathcal{E}^c) \geq \frac{1}{2}\exp\big\{-\mathrm{KL}(\mathbf{P}||\mathbf{P}')\big\}. \tag{C.1}$$

The second has no name unfortunately (Theorem C.2), and is closer to the lower bound of (Kaufmann et al., 2016, Lemma 1). It will be used for the instance-dependent regret lower bound (Theorem 4).

Theorem C.2 (Boone & Maillard (2025))**.** *Let $\mathbf{P} \ll \mathbf{P}'$ be two probability measures over the same probability space with respective expectation operators $\mathbf{E}$ and $\mathbf{E}'$. For every $X > 0$ positive random variable, we have*

$$\log \mathbf{E}'[X] + \mathrm{KL}(\mathbf{P}||\mathbf{P}') \geq \mathbf{E}[\log(X)]. \tag{C.2}$$

To make use of Theorems C.1 and C.2, we provide an expression of the relative entropy $\mathrm{KL}(\mathbf{P}\|\mathbf{P}')$ between the two probability measures $\mathbf{P}, \mathbf{P}'$ of the stochastic processes induced by the models M, M', see Lemma C.3.

Given two instances $M \equiv (\mathcal{S}_a, \nu_a, p_a)_{a\in\mathcal{A}}$ and $M' \equiv (\mathcal{S}'_a, \nu'_a, p'_a)_{a\in\mathcal{A}}$, we write $M \ll M'$ if for every arm $a \in \mathcal{A}$, (1) $\mathcal{S}_a \subseteq \mathcal{S}'_a$; (2) $\nu_a(s_a) \ll \nu'_a(s_a)$ for all $s_a \in \mathcal{S}_a$; and (3) $\mathrm{supp}(p_a(s_a)) \subseteq \mathrm{supp}(p'_a(s_a))$ for all $s_a \in \mathcal{S}_a$.

Lemma C.3 (Relative entropy)**.** *Let $M \equiv (\mathcal{S}_a, \nu_a, p_a)_{a\in\mathcal{A}} \ll M' \equiv (\mathcal{S}'_a, \nu'_a, p'_a)_{a\in\mathcal{A}}$ be two instances. Let Λ be a learning algorithm, $T \geq 1$ a horizon and $s_0 \in \mathcal{S}$ an initial state. Upon running Λ initialized at s_0 for T steps on M (respectively M'), we obtain a probability measure $\mathbf{P}$ (respectively $\mathbf{P}'$) on the possible histories up to time T. We have $\mathbf{P} \ll \mathbf{P}'$ and by denoting $\mathrm{d}\mathbf{P}/\mathrm{d}\mathbf{P}'$ the Radon-Nikodym derivative of $\mathbf{P}$ with respect to $\mathbf{P}'$,*

the relative entropy from $\mathbf{P}'$ *to* $\mathbf{P}$ *satisfies*

$$\mathrm{KL}(\mathbf{P}||\mathbf{P}') = \begin{cases} \mathbf{E}_{s_0}^{M,\Lambda}\left[\sum_{t=1}^{T-1} \mathrm{KL}\Big(\nu_{A(t)}(S_{A(t)}(t))\Big\|\nu'_{A(t)}(S_{A(t)}(t))\Big)\right] \\ +\mathbf{E}_{s_0}^{M,\Lambda}\left[\sum_{t=1}^{T-1} \mathrm{KL}\Big(p^1_{A(t)}(S_{A(t)}(t))\Big\|p'^1_{A(t)}(S_{A(t)}(t))\Big)\right] \\ +\mathbf{E}_{s_0}^{M,\Lambda}\left[\sum_{t=1}^{T-1}\sum_{a\neq A(t)} \mathrm{KL}\big(p^0_a(S_a(t))\big\|p'^0_a(S_a(t))\big)\right] \end{cases}. \tag{C.3}$$

Proof. The fact that $\mathbf{P} \ll \mathbf{P}'$ comes from $M \ll M'$.
Denote $h(T) \equiv (s(1), a(1), r(2), s(2), \ldots, r(T), s(T))$ an instance of history at time T. By Markov's property, the Radon-Nikodym derivative can be written as

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{P}'}(h(T)) &= \frac{\mathbf{P}(H(T)=h(T))}{\mathbf{P}'(H(T)=h(T))} \\
&= \prod_{t=1}^{T-1}\left(\frac{\nu_{a(t)}(r(t+1)|s_{a(t)}(t))}{\nu'_{a(t)}(r(t+1)|s_{a(t)}(t))}\frac{p^1_{a(t)}(s_{a(t)}(t+1)|s_{a(t)}(t))}{p'^1_{a(t)}(s_{a(t)}(t+1)|s_{a(t)}(t))}\cdot\prod_{a\neq a(t)}\frac{p^0_a(s_a(t+1)|s_a(t))}{p'^0_a(s_a(t+1)|s_a(t))}\right) \\
&= \exp\left\{\sum_{t=1}^{T-1}\log\left(\frac{\nu_{a(t)}(r(t+1)|s_{a(t)}(t))}{\nu'_{a(t)}(r(t+1)|s_{a(t)}(t))}\frac{p^1_{a(t)}(s_{a(t)}(t+1)|s_{a(t)}(t))}{p'^1_{a(t)}(s_{a(t)}(t+1)|s_{a(t)}(t))}\cdot\prod_{a\neq a(t)}\frac{p^0_a(s_a(t+1)|s_a(t))}{p'^0_a(s_a(t+1)|s_a(t))}\right)\right\}
\end{aligned}$$

with the convention that $\log(0) = -\infty$. Taking the expectation of the $\log(-)$, we obtain

$$\mathrm{KL}(\mathbf{P}\|\mathbf{P}') = \begin{cases} \mathbf{E}\left[\sum_{t=1}^{T-1}\log\left(\frac{\nu_{A(t)}(R(t+1)|S_{A(t)}(t))}{\nu'_{A(t)}(R(t+1)|S_{A(t)}(t))}\right)\right] \\ +\mathbf{E}\left[\sum_{t=1}^{T-1}\log\left(\frac{p^1_{A(t)}(S_{A(t)}(t+1)|S_{A(t)}(t))}{p'^1_{A(t)}(S_{A(t)}(t+1)|S_{A(t)}(t))}\right)\right] \\ +\mathbf{E}\left[\sum_{t=1}^{T-1}\sum_{a\neq A(t)}\log\left(\frac{p^0_a(S_a(t+1)|S_a(t))}{p'^0_a(S_a(t+1)|S_a(t))}\right)\right] \end{cases} \tag{C.4}$$

To explain how to go from Equation (C.4) to sums of Kullback-Leibler divergences, we focus on the first term of the above, as all the others are handled similarly. We have

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T-1}\log\left(\frac{\nu_{A(t)}(R(t+1)|S_{A(t)}(t))}{\nu'_{A(t)}(R(t+1)|S_{A(t)}(t))}\right)\right] &\overset{\circ 1}{=} \mathbf{E}\left[\sum_{t=1}^{T-1}\mathbf{E}\left[\log\left(\frac{\nu_{A(t)}(R(t+1)|S_{A(t)}(t))}{\nu'_{A(t)}(R(t+1)|S_{A(t)}(t))}\right)\middle| S_{A(t)}(t), A(t)\right]\right] \\
&\overset{\circ 2}{=} \mathbf{E}\left[\sum_{t=1}^{T-1}\mathrm{KL}\Big(\nu_{A(t)}(S_{A(t)}(t))\Big\|\nu'_{A(t)}(S_{A(t)}(t))\Big)\right]
\end{aligned}$$

where (°1) applies the Tower Property; and (°2) follows by definition of $\mathrm{KL}(\nu_a(s)\|\nu'_a(s))$. Rewriting the other two terms similarly and injecting it back to Equation (C.4), we obtain Equation (C.3). □

Lastly, for lower bounds (both instance-dependent and worst-case), the proof strategy revolves around an *idealistic state transform*, that adds an idealistic absorbing state of full reward which is difficult to access (Lemma C.4). The idea of the construction was sketched in the main body with Figure 4, that we recall in Figure 6.

The precise statement is given below.

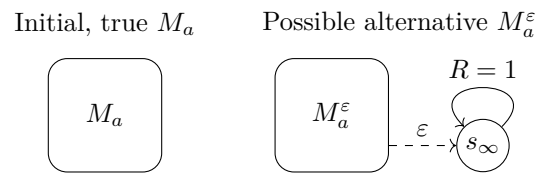


Figure 6: An illustration of the transformation M_a^ε.

Lemma C.4 (Idealistic state transform)**.** *Let $M \equiv (\mathcal{S}, \nu, p)$ a Markov reward process*[12] *with $[0,1]$-rewards. For all $\varepsilon > 0$, there exists a Markov reward process $M^\varepsilon \equiv (\mathcal{S} + \{s_\infty\}, \nu^\varepsilon, p^\varepsilon)$ such that*

$$\max_{s \in \mathcal{S}} \{\mathrm{KL}(\nu(s) || \nu^\varepsilon(s)) + \mathrm{KL}(p(s) || p^\varepsilon(s))\} \leq \varepsilon$$

with $[0,1]$-rewards, $g_*(M^\varepsilon) = 1$ and $\mathrm{sp}(b^*(M^\varepsilon)) \leq 2/\varepsilon$.

Proof. Let $\varepsilon > 0$. Add an extra state s_∞ to M. In $M^\varepsilon \equiv (\mathcal{S} + \{s_\infty\}, \nu^\varepsilon, p^\varepsilon)$, the new state s_∞ is absorbing with maximal reward $r^\varepsilon(s_\infty) = 1$; Then, for $s \in \mathcal{S}$, we set $\nu^\varepsilon(s) = \nu(s)$ and

$$p^\varepsilon(s'|s) := (1-\varepsilon)p(s'|s) + \varepsilon \mathbf{1}(s' = s_\infty).$$

We have $g_*(M^\varepsilon) = 1$ and $\mathrm{sp}(b^*(M^\varepsilon)) \leq 1/\varepsilon$. Continuing, we have

$$\begin{aligned}
\max_{s \in \mathcal{S}} \{\mathrm{KL}(\nu(s) \| \nu^\varepsilon(s)) + \mathrm{KL}(p(s) \| p^\varepsilon(s))\} &= \max_{s \in \mathcal{S}} \mathrm{KL}(p(s) \| (1-\varepsilon)p(s) + \varepsilon \cdot \mathrm{Dirac}(s_\infty)) \\
&\leq \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} p(s'|s) \log\left(\frac{p(s'|s)}{(1-\varepsilon)p(s'|s)}\right) \\
&= \log\left(\frac{1}{1-\varepsilon}\right) \overset{\circ 1}{\leq} 2\varepsilon \ ,
\end{aligned}$$

where ($\circ$1) holds when $\varepsilon \leq 1/2$. Accordingly, the family $(M_{\varepsilon/2})_{\varepsilon > 0}$ satisfies the desired properties. □

C.2 Proof of Theorem 4: Instance dependent lower bound

In this section, we provide the complete proof of Theorem 4, restated below for convenience.

Theorem 4. *Let Λ be a learning algorithm that (1) is consistent on $\mathcal{M}$, and (2) is ψ-rarely switching. For every instance $M \in \mathcal{M}$ with $g_* < 1$ and unique optimal pure policy, it holds that*

$$\forall s_0 \in \mathcal{S}, \quad \liminf_{T \to \infty} \frac{\mathrm{Reg}(T; M, \Lambda)}{\log(T)} = +\infty. \tag{C.5}$$

Proof. Let Λ be a consistent learning algorithm on the universal ambient space $\mathcal{M}$ and assume that it is ψ-rarely switching. Let $M \in \mathcal{M}$ be a model such that $g_* < 1$ and fix $a \in \mathcal{A}$ a sub-optimal arm in M. For $\varepsilon > 0$, consider M^ε the model in which M_a^1 (the Markov reward process of arm a induced by M_a upon playing the activating action 1 exclusively) is changed according to Lemma C.4, i.e., upon activating arm a, the associated Markov chain eventually falls into an absorbing state of maximal reward. By definition of M^ε (Lemma C.4), we see that

$$\begin{aligned}
\mathbf{E}^{M,\Lambda}\left[\sum_{t=1}^{T} \mathrm{KL}\Big(\big(\nu_{A(t)} \otimes p_{A(t)}\big)\big(S_{A(t)}(t)\big) \Big\| \big(\nu^\varepsilon_{A(t)} \otimes p^\varepsilon_{A(t)}\big)\big(S_{A(t)}(t)\big)\Big)\right] \\
\leq \varepsilon \mathbf{E}^{M,\Lambda}[N_a(T)],
\end{aligned} \tag{C.6}$$

[12] i.e., a single-action Markov decision process.

where $N_a(T) := \sum_{t=1}^{T-1} \mathbf{1}(A(t) = a)$ is the number of visits of arm a up to time T. Consider the random variable

$$X_T := T - N_a(T)$$

tracking the number times a isn't played. Combining Equation (C.6) with standard information theory (Boone & Maillard, 2025, Corollary C.3), we have

$$\log \mathbf{E}^{M^\varepsilon,\Lambda}[X_T] + \varepsilon \mathbf{E}^{M,\Lambda}[N_a(T)] \geq \mathbf{E}^{M,\Lambda}[\log(X_T)]. \tag{C.7}$$

Our objective is to rework Equation (C.7) to obtain a lower bound on $N_a(T)$. Using that X_t tracks the number of times a isn't played, and that a is optimal (respectively sub-optimal) in M^ε (respectively M), we rely on the consistency of the algorithm Λ to show that $\mathbf{E}^{M,\Lambda}[\log(X_T)] \sim \log(T)$, while $\log \mathbf{E}^{M^\varepsilon,\Lambda}[X_T] = \mathrm{o}(\log(T))$.

On the one hand, a is the unique optimal arm in M^ε by construction. As $\mathrm{Reg}(T; M^\varepsilon, \Lambda) = \mathrm{o}(T^\eta)$ when $T \to \infty$, we deduce $\log \mathbf{E}^{M^\varepsilon,\Lambda}[X_T] \leq \eta \log(T) + \mathrm{o}(\log(T))$, where $\eta > 0$ is arbitrary. This is proved as follows. Binding the regret with the number of suboptimal pulls, we know from Proposition A.3 that there is a positive smooth concave function φ such that

$$\mathbf{E}^{M^\varepsilon,\Lambda}\left[\sum_{a' \in \mathcal{A}} \Delta_{a'}(M^\varepsilon) N_{a'}(T)\right] \leq \mathrm{Reg}(T; M^\varepsilon, \Lambda) + \varphi\left(\mathbf{E}^{M^\varepsilon,\Lambda}\left[\sum_{a' \in \mathcal{A}} \Delta_{a'}(M^\varepsilon) N_{a'}(T)\right]\right).$$

Now, letting $\Delta_0(M^\varepsilon) := \min\{\Delta_{a'}(M^\varepsilon) : a' \neq a\}$ the smallest sub-optimal gap of M^ε, we have

$$\Delta_0(M^\varepsilon) X_T \leq \sum_{a' \in \mathcal{A}} \Delta_{a'}(M^\varepsilon) N_{a'}(T) \leq X_T.$$

Combining the two previous equation and using that φ is non-increasing, we find that for all $\eta > 0$, we have

$$\mathbf{E}^{M^\varepsilon,\Lambda}[X_T] - \frac{1}{\Delta_0(M^\varepsilon)} \varphi\Big(\mathbf{E}^{M^\varepsilon,\Lambda}[X_T]\Big) \leq \mathrm{Reg}(T; M^\varepsilon, \Lambda) = \mathrm{o}(T^\eta).$$

As φ is a sublinear function, it follows that $\mathbf{E}^{M^\varepsilon,\Lambda}[X_T] = \mathrm{o}(T^\eta)$. So $\log \mathbf{E}^{M^\varepsilon,\Lambda}[X_T] \leq \eta \log(T) + \mathrm{o}(\log(T))$. This holds for all $\eta > 0$ and X_t is non-negative, so that $\log \mathbf{E}^{M^\varepsilon,\Lambda}[X_T] = \mathrm{o}(\log(T))$.

Continuing by proving that $\mathbf{E}^{M,\Lambda}[\log(X_T)] = \log(T) + \mathrm{o}(\log(T))$. Using the same technique as above, we show that because the arm a is sub-optimal in M, we have $\mathbf{E}^{M,\Lambda}[T - X_T] = \mathbf{E}^{M,\Lambda}[N_a(T)] = \mathrm{o}(T^\eta)$. So, we have

$$\begin{aligned}
\mathbf{E}^{M,\Lambda}[\log(X_T)] &\overset{\circ 1}{\geq} \sup_{\lambda > 0}\Big\{\lambda \mathbf{P}^{M,\Lambda}(\log(X_T) \geq \lambda)\Big\} \\
&\geq \sup_{\delta > 0}\Big\{\log(\delta T)\Big(1 - \mathbf{P}^{M,\Lambda}(X_T \leq \delta T)\Big)\Big\} \\
&\overset{\circ 2}{\geq} \sup_{\delta > 0}\left\{\log(\delta T)\left(1 - \frac{\mathbf{E}^{M,\Lambda}[T - X_T]}{(1-\delta)T}\right)\right\} \\
&\overset{\circ 3}{\geq} \log\left(\frac{1}{2}T\right)\Big(1 - \mathrm{o}\Big(T^{-\frac{1}{2}}\Big)\Big) = \log(T) + \mathrm{o}(\log(T)),
\end{aligned}$$

where ($\circ$1) and ($\circ$2) follow from Markov's inequality; and ($\circ$3) sets $\delta \equiv \frac{1}{2}$ and uses $\mathbf{E}^{M,\Lambda}[T - X_T] = \mathrm{o}(T^{1/2})$.

Going back to Equation (C.7) and injecting everything, we finally obtain

$$\mathbf{E}^{M,\Lambda}[N_a(T)] \geq \frac{\log(T) + \mathrm{o}(\log(T))}{\varepsilon}. \tag{C.8}$$

As $\varepsilon > 0$ is arbitrary, we conclude accordingly that $\mathbf{E}^{M,\Lambda}[N_a(T)] = \omega(\log(T))$. Binding the regret to the number of visits of suboptimal arms again with Proposition A.3, we conclude that $\mathrm{Reg}(T; M, \Lambda) = \omega(\log(T))$. $\square$

C.3 Proof of Theorem 7: Worst-case lower bound

In this section, we provide a complete proof of Theorem 7, restated below for convenience.

Theorem 7. *Let Λ be a learning algorithm. Let $\mathcal{U}_\tau \subseteq [0,1]$ be arbitrary. Then*

$$\forall T \geq 1, \quad \sup_{M \in \mathcal{M}} \mathrm{Reg}(T; M, \Lambda) \geq \frac{11}{4800} \cdot T. \tag{C.9}$$

Proof. We introduce the following constants:

$$\alpha := 12, \quad \lambda := \frac{1}{10}, \quad \delta := \frac{1}{12} \log\left(\frac{12}{11}\right), \quad c := \frac{11}{4800}. \tag{C.10}$$

Consider the arm space $\mathcal{A} := \{1, 2\}$ and let $M \equiv (\mathcal{S}, \mathcal{A}, \nu, p^0, p^1)$ be the Bernoulli multi-armed bandits with arms $\mathcal{A}$ and mean reward $r_1 = \frac{1}{2}$ and $r_2 = 0$. That is, $\mathcal{S}_a = \{1\}$ for all arm, with $\nu_1(1) = \mathrm{Ber}(\frac{1}{2})$ and $\nu_2(1) = \mathrm{Ber}(0)$. Write $\mathbf{P}, \mathbf{E}$ the probability (resp. the expectation) under M. We see that $a = 1$ is the optimal arm of M and that, in M, both the expected regret and pseudo-regret are equal to $\frac{1}{2}\mathbf{E}[N_2(T)]$.

Let $\varepsilon \equiv \alpha/T$ with $\alpha := 12$ as given by Equation (C.10). Consider the alternative model M^ε where the model of arm 2 undergoes the idealistic transform of Lemma C.4, that is, an extra state is added and $p_2^{\varepsilon,1}$ goes to that state with probability $\varepsilon/2$. Write $\mathbf{P}^\varepsilon, \mathbf{E}^\varepsilon$ the probability (resp. the expectation) under M^ε. We see that $g_2(M^\varepsilon) = 1$. Moreover, the reaching time to the idealistic state is $1/(2\varepsilon)$, while one scores 0 instead of 1, so $\mathrm{sp}(b_2^\varepsilon) \leq 2/\varepsilon$. Also, we have

$$\mathrm{KL}(\mathbf{P} || \mathbf{P}^\varepsilon) \leq \varepsilon \mathbf{E}[N_2(T)].$$

The proof is split on whether $\mathbf{E}[N_2(T)] \leq \delta T$ or not, and proves that in both cases, we have

$$\max\{\mathrm{Reg}(T; M), \mathrm{Reg}(T; M^\varepsilon)\} \geq cT. \tag{C.11}$$

First, if $\mathbf{E}[N_2(T)] > \delta T$ then $\mathrm{Reg}(T; M) \geq T/(2\delta) \geq cT$, hence Equation (C.11).

We now focus on the case $\mathbf{E}[N_2(T)] \leq \delta T$. By Bretagnolle-Huber inequality (Theorem C.1), we have

$$\mathbf{P}(N_2(T) \geq \lambda T) + \mathbf{P}^\varepsilon(N_1(T) \geq (1-\lambda)T) \geq \frac{1}{2}\exp(-\varepsilon\delta T) = \frac{1}{2\mathrm{e}^{\delta\alpha}}. \tag{C.12}$$

We have one out of two cases.

- Case 1: $\mathbf{P}(N_2(T) \geq \lambda T) \geq \lambda/(2\mathrm{e}^{\delta\alpha})$, in which case we trivially obtain

$$\mathrm{Reg}(T; M) = \frac{1}{2}\mathbf{E}[N_2(T)] \geq \frac{1}{2} \cdot \lambda T \cdot \frac{\lambda}{2\mathrm{e}^{\delta\alpha}} = cT. \tag{C.13}$$

- Case 2: $\mathbf{P}(N_1(T) \geq (1-\lambda)T) \geq (1-\lambda)/(2\mathrm{e}^{\delta\alpha})$. In M^ε, the gain gap between arm 1 and arm 2 is $1/2$. So $\mathbf{E}^\varepsilon[\mathrm{PR}^\varepsilon(T)] \geq (1-\lambda)/(4\mathrm{e}^{\delta\alpha}) \cdot T$ for the same reason as in Case 1. To link this to the regret, we use $\mathrm{Reg}(T; M^\varepsilon) \geq \mathbf{E}^\varepsilon[\mathrm{PR}^\varepsilon(T)] - 2/\varepsilon$, that holds by Corollary A.2. We have

$$\mathrm{Reg}(T; M^\varepsilon) \geq \left(\frac{(1-\lambda)^2}{4\mathrm{e}^{\delta\alpha}} - \frac{2}{\alpha}\right)T \geq \left(\frac{1-2\lambda}{4\mathrm{e}^{\delta\alpha}} - \frac{2}{\alpha}\right)T = \frac{1}{60} \cdot T \geq cT. \tag{C.14}$$

This proves the result. □

D Regret upper bounds (Theorems 6, 8 and 9)

In this Appendix, we provide the proofs of the regret upper bounds of the main text. In Appendix D.1, we prove Theorem 6, which gives the instance dependent bound. We improve on this bound in Appendix D.2 when restricting to bounded span proxy, that is, Theorem 8. Lastly, in Appendix D.3, we establish the worst-case regret bound of Theorem 9.

Additional notations and terminology. Throughout the section, we introduce a supplementary set of notations. Given an arm a, we write (τ_k^a) the enumeration of pulls of arm a, given by $\tau_1^a := \inf\{t \geq 1 : A(t) = a\}$ and $\tau_{k+1}^a := \inf\{t > \tau_k^a : A(t) = a\}$. We further write $\hat{g}_{a,n}$ the empirical estimate $\hat{g}_a(t)$ conditionally on $N_a(t) = n$. Formally,

$$\hat{g}_{a,n} := \frac{1}{n}\sum_{k=1}^{n} R(\tau_k^a + 1). \tag{D.1}$$

Also, we will call *episode* k the time segment $\{t_k, \dots, t_{k+1} - 1\}$ over which `UCB-NOM` plays the fixed arm $A(t_k)$. The *number of episodes* up to time T is $K(T) := \sum_{k=1}^{\infty} \mathbf{1}(t_k \leq T)$.

D.1 Proof of Theorem 6: Instance-dependent regret upper bound

In this section, we provide a general upper bound for the instance dependent regret of `UCB-NOM`, when no information on the bias span of the underlying instance is provided. In that scenario, `UCB-NOM` is fed a span proxy function f that diverges to $+\infty$. The regret bound is sketched in Theorem 6 of the main text, which essentially claims that the performance scale as $\mathrm{O}(f(T)^2 \log(T))$. We give a more detailed version of Theorem 6 in Theorem D.1, incorporating second-order terms in the bound.

Theorem D.1 (Instance-dependent regret of `UCB-NOM`)**.** *Let $f : \mathbf{R}_+ \to \mathbf{R}_+$ an increasing function diverging to infinity such that $f(T) = \mathrm{o}(\log(T)^{\frac{1}{4}})$. Let $\delta : \mathbf{N} \to \mathbf{R}_+$ be a confidence function such that $\sum_{t=1}^{\infty} \delta(t) < \infty$ and $\log(\frac{1}{\delta(t)}) = \Theta(\log(t))$. For all instance $M \in \mathcal{M}$, $\mathrm{Reg}(T) \equiv \mathrm{Reg}(T; M, \texttt{UCB-NOM}(f))$ is bounded as*

$$\begin{aligned}
\mathrm{Reg}(T) = \mathrm{O}\Bigg(\sum_{a \notin \mathcal{A}^*} \left(\frac{f(T)^2 \log(T)}{\Delta_a} + \frac{f(T)^4}{\Delta_a^3}\right) + |\mathcal{A}| \max_{a \in \mathcal{A}} \mathrm{sp}(b_a) \log(T)\Bigg) & \quad \begin{pmatrix}\textit{DIVERGING}\\ \textit{TERMS}\end{pmatrix} \\
+ \mathrm{O}\Bigg(\sum_{a \notin \mathcal{A}^*} \Delta_a \left((\beta_f)^2 + C_\tau\left(\beta_f + \frac{1}{\Delta_a^4} + \left(\frac{1 + \mathrm{sp}(b_a)}{\Delta_a}\right)^2\right) + \left(\frac{1 + \mathrm{sp}(b_a)}{\Delta_a}\right)^4\right)\Bigg), & \quad \begin{pmatrix}\textit{CONSTANT}\\ \textit{TERMS}\end{pmatrix}
\end{aligned}$$

where $\mathcal{A}^* := \arg\max_{a \in \mathcal{A}} g_a$ is the set of optimal arms, $\Delta_a := g_* - g_a$ is the gain-gap and β_f is the same as in Theorem 6. Moreover, if there is a single optimal arm, $|\mathcal{A}^*| = 1$, the bound of $\mathrm{Reg}(T)$ is improved to

$$\begin{aligned}
\mathrm{Reg}(T) = \mathrm{O}\Bigg(\sum_{a \notin \mathcal{A}^*} \left(\frac{f(T)^2 \log(T)}{\Delta_a} + \frac{f(T)^4}{\Delta_a^3}\right) + |\mathcal{A}| \max_{a \in \mathcal{A}} \mathrm{sp}(b_a) \log\big(f(T)^2 \log(T)\big)\Bigg) & \quad \begin{pmatrix}\textit{DIVERGING}\\ \textit{TERMS}\end{pmatrix} \\
+ \mathrm{O}\Bigg(\sum_{a \notin \mathcal{A}^*} \Delta_a \left((\beta_f)^2 + C_\tau\left(\beta_f + \frac{1}{\Delta_a^4} + \left(\frac{1 + \mathrm{sp}(b_a)}{\Delta_a}\right)^2\right) + \left(\frac{1 + \mathrm{sp}(b_a)}{\Delta_a}\right)^4\right)\Bigg). & \quad \begin{pmatrix}\textit{CONSTANT}\\ \textit{TERMS}\end{pmatrix}
\end{aligned}$$

About Theorem 6. Theorem 6 follows immediately by extracting the dominant diverging term and selecting the constant term $\sum_{a \notin \mathcal{A}^*} \Delta_a (\beta_f)^2$. This term is independent of Δ_a, and grows larger the slower is f diverging to infinity. In particular, $(\beta_f)^2$ can be significantly large if f is poorly initialized with respect to M.

Discussion. Remarkably, we can provide better regret guarantees when the optimal arm is unique. In cases where there are multiple optimal arms, the third order term of the regret upper bound scales with the number of arm switches (which is bounded by $\mathrm{O}(|\mathcal{A}| \log(T))$, see Lemma D.7). However, the dependency in the number of arm switches is pushed to the second order when the optimal arm is unique — this is because when $|\mathcal{A}^*| = 1$, the total number of switches is provably $\mathrm{O}(\log(f(T) \log(T)))$.

About notations. In the remaining of the section, we fix an instance $M \in \mathcal{M}$, a span-proxy function f and the algorithm $\Lambda \equiv \texttt{UCB-NOM}(f)$. Now that they are all fixed, we drop the dependency in all these quantities in notations. In particular, we write $\mathrm{Reg}(T), \mathbf{E}[-]$ and $\mathbf{P}(-)$ instead of $\mathrm{Reg}(T; M, \Lambda), \mathbf{E}^{M,\Lambda}[-]$ and $\mathbf{P}^{M,\Lambda}(-)$.

Idea of the proof. To bound the regret, we bound the pseudo-regret (Equation (6)), then convert back the pseudo-regret bound to a regret bound using Proposition A.3. To bound the pseudo-regret, we upper-bound

the number of pulls to every suboptimal arms via Lemma D.2 down-below, that incarnates most of the difficulties behind Theorem D.1.

Let $\varepsilon < \frac{1}{2}\min_{a\notin\mathcal{A}^*}(g_* - g_a)$ be an arbitrary (but small enough) constant.

Lemma D.2 (Number of visits, f diverging)**.** *Consider a suboptimal arm $a \notin \mathcal{A}^*$. Then $\mathbf{E}[N_a(T)]$ is of order:*

$$\frac{f(T)^2\log(T)}{(g_* - g_a - 2\varepsilon)^2} + \left(\frac{f(T)}{g_* - g_a - 2\varepsilon}\right)^4 + (\beta_f)^2 + C_\tau\left(\beta_f + \frac{1}{\varepsilon^4} + \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^2\right) + \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^4. \tag{D.2}$$

Proof. Denote $C_b^0 := 1 + \max_{a\in\mathcal{A}}\mathrm{sp}(b_a)$. Fix $\varepsilon > 0$ and $T \geq 1$, and introduce the function

$$\chi_\varepsilon(T) := 1 + \sup\left\{n \geq 1 : \sqrt{\frac{\log((1+n)/\delta(T))}{n}} + \frac{\log_2(1+n)}{n} \geq \frac{g_* - g_a - 2\varepsilon}{f(T)}\right\}. \tag{D.3}$$

We will upper bound this function in time.

We define the sequence of episodes *relative to arm a* $(t_\ell^a)_{\ell\geq 1}$ as the episodes such that $A(t_k) = a$, further clipped to $\{1,\dots,T-1\}$, i.e., $t_1^a := \min\{\inf\{t_k : A(t_k) = a\}, T-1\}$ and $t_{\ell+1}^a := \min\{\inf\{t_k : A(t_k) = a, t_k > t_\ell^a\}, T-1\}$. The number of visits to arm a is decomposed as such. To determine why the algorithm pulls $A(t) = a$, we look at what happens at the start of the current episode, i.e., at the unique time t_ℓ^a such that $t_\ell^a \leq t \leq t_\ell^a - 1$. We write,

$$\begin{aligned}
N_a(T) &= \sum_{t=1}^{T-1}\mathbf{1}(A(t) = a)\\
&= \sum_{\ell=1}^{\infty}\sum_{t=1}^{\infty}\mathbf{1}\begin{pmatrix} A(t) = a \\ t_{\ell+1}^a \leq t < t_{\ell+1}^a\end{pmatrix}\\
&= \sum_{\ell=1}^{\infty}\sum_{t=1}^{\infty}\Bigg(\underbrace{\mathbf{1}\begin{pmatrix} A(t) = a \\ t_\ell^a \leq t < t_{\ell+1}^a \\ \hat{g}_a(t_\ell^a) > g_a + \varepsilon\end{pmatrix}}_{\substack{\textsc{Empirical}\\ \textsc{Errors}\\ \text{(I)}}} + \underbrace{\mathbf{1}\begin{pmatrix} A(t) = a \\ t_\ell^a \leq t < t_{\ell+1}^a \\ g_{a^*} > \tilde{g}_{a^*}(t_\ell^a) + \varepsilon\end{pmatrix}}_{\substack{\textsc{Optimism}\\ \textsc{Failures}\\ \text{(II)}}} + \underbrace{\mathbf{1}\begin{pmatrix} A(t) = a \\ t_\ell^a \leq t < t_{\ell+1}^a \\ N_a(t_\ell^a) \leq \chi_\varepsilon(T)\end{pmatrix}}_{\substack{\textsc{Optimistic}\\ \textsc{Pulls}\\ \text{(III)}}} + \underbrace{\mathbf{1}\begin{pmatrix} A(t) = a \\ t_\ell^a \leq t < t_{\ell+1}^a \\ N_a(t_\ell^a) > \chi_\varepsilon(T) \\ \hat{g}_a(t_\ell^a) \leq g_a + \varepsilon \\ g_{a^*} \leq \tilde{g}_{a^*}(t_\ell^a) + \varepsilon\end{pmatrix}}_{\substack{\textsc{Late Bloomers}\\ \text{(IV)}}}\Bigg)
\end{aligned} \tag{D.4}$$

In the above, we state that arm $A(t) = a$ may be pulled for four reasons. Either (I) the empirical estimate of arm a is far away from its true value g_a; Or (II) the suboptimal arm is pulled because the index of the optimal arm is not an upper bound of g^*, meaning that optimism failed; Or (III) the number of pulls is still small, so that the optimism bonus $\xi_a(t, \delta(t))$ (see Proposition 5) is still large; Or finally (IV) optimism didn't fail, the empirical estimate of arm a is well concentrated and the optimism bonus is small.

In what follows, we show that (I) accounts for O(1) pulls in expectation, (II) for O(1), (III) is of order $\mathrm{O}(f(T)\log(T))$ and dominates all the other terms, and (IV) is 0 by design of `UCB-NOM` and of $\chi_\varepsilon(T)$, see Equation (D.3).

Empirical estimator failure (I). We show that this term is bounded.

$$\begin{aligned}
\mathbf{E}\left[\sum_{\ell=1}^{\infty}\sum_{t=1}^{\infty}(\mathrm{I})\right] &= \mathbf{E}\left[\sum_{\ell=1}^{\infty}\sum_{t=1}^{\infty}\mathbf{1}\begin{pmatrix} A(t) = a \\ t_\ell^a \leq t \leq t_{\ell+1}^a \\ \hat{g}_a(t_\ell^a) > g_a + \varepsilon\end{pmatrix}\right]\\
&\overset{\circ 1}{=} \mathbf{E}\left[\sum_{\ell=1}^{\infty}\big(N_a(t_{\ell+1}^a) - N_a(t_\ell^a)\big)\mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right]
\end{aligned}$$

$$\overset{\circ 2}{=} \mathbf{E}\left[\sum_{\ell=1}^{\infty} N_a(t_\ell^a)\mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right] + \mathbf{E}\left[\sum_{\ell=1}^{\infty}\big(N_a(t_{\ell+1}^a) - 2N_a(t_\ell^a)\big)\mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right], \quad \text{(D.5)}$$

where (°1) uses that only arm a is pulled during the episode $\{t_\ell^a, \dots, t_{\ell+1}^a - 1\}$; and (°2) splits the sum of (°1) according to the doubling trick. Both terms in Equation (D.5) are bounded separately.

For the left term of Equation (D.5), we proceed as follows:

$$\begin{aligned}
\mathbf{E}\left[\sum_{\ell=1}^{\infty} N_a(t_\ell^a)\mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right] &\overset{\circ 1}{\leq} \mathbf{E}\left[\sum_{n=1}^{\infty} n \cdot \mathbf{1}(\hat{g}_{a,n} > g_a + \varepsilon)\right] \\
&= \sum_{n=1}^{\infty} n \cdot \mathbf{P}(\hat{g}_{a,n} > g_a + \varepsilon) \\
&\overset{\circ 2}{\leq} \sum_{n=1}^{\infty} n \cdot \exp\left(-\frac{n\varepsilon^2 - 2\varepsilon\,\mathrm{sp}(b_a)(1 + \log_2(1+n))}{(1+\mathrm{sp}(b_a))^2}\right) = \mathrm{O}\left(\frac{1}{\varepsilon^4}\right), \quad \text{(D.6)}
\end{aligned}$$

where (°1) uses that the LHS is a subset of the RHS; and (°2) invokes Lemma D.6.

For the right term of Equation (D.5), we have

$$\begin{aligned}
\mathbf{E}\left[\sum_{\ell=1}^{\infty}\big(N_a(t_{\ell+1}^a) - 2N_a(t_\ell^a)\big)\mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right] &\overset{\circ 1}{\leq} \mathbf{E}\left[\sum_{\ell=1}^{\infty} \inf\Big\{\tau_i - (t_\ell^a + N_a(t_\ell^a)) \underset{:\ \tau_i \geq t_\ell^a + N_a(t_\ell^a)}{}\Big\}\mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right] \\
&\overset{\circ 2}{\leq} C_\tau \mathbf{E}\left[\sum_{\ell=1}^{\infty} \mathbf{1}(\hat{g}_a(t_\ell^a) > g_a + \varepsilon)\right] \\
&\leq C_\tau \mathbf{E}\left[\sum_{n=1}^{\infty} \mathbf{1}(\hat{g}_{a,n} > g_a + \varepsilon)\right] \\
&= C_\tau \sum_{n=1}^{\infty} \mathbf{P}(\hat{g}_{a,n} > g_a + \varepsilon) \\
&\overset{\circ 3}{\leq} C_\tau \sum_{n=1}^{\infty} \exp\left(-\frac{n\varepsilon^2 - 2\varepsilon\,\mathrm{sp}(b_a)(1 + \log_2(1+n))}{(1+\mathrm{sp}(b_a))^2}\right) \\
&= \mathrm{O}\left(\frac{C_\tau}{\varepsilon^2}\right),
\end{aligned}$$

where (°1) uses that $t_{\ell+1}^a$ corresponds to the first decision epoch after which the Doubling Trick does not force the continuation of the episode anymore, which is at time $t_\ell^a + N_a(t_\ell^a)$; (°2) conditions on that time $t_\ell^a + N_a(t_\ell^a)$, which is a stopping-time, and invokes the Tower Property; and finally, (°3) follows from Lemma D.6.

We conclude that

$$\text{(I)} = \mathrm{O}\left(\frac{C_\tau}{\varepsilon^2} + \frac{1}{\varepsilon^4}\right). \quad \text{(D.7)}$$

Optimism failure (II). We now aim to bound the term (II) which correspond to the case where the optimistic estimator of the optimal arm $\tilde{g}_{a^*}$ turns out to be smaller than the true gain g_{a^*}.

$$\begin{aligned}
&\mathbf{E}\left[\sum_{\ell=1}^{\infty}\sum_{t=1}^{\infty}\text{(II)}\right] \\
&\quad = \mathbf{E}\left[\sum_{\ell \geq 1}\big(N_a(t_{\ell+1}^a) - N_a(t_\ell^a)\big)\mathbf{1}(g_{a^*} > \tilde{g}_{a^*}(t_\ell^a) + \varepsilon)\right] \quad \text{(D.8)}
\end{aligned}$$

$$= \underbrace{\mathbf{E}\left[\sum_{\ell\geq 1} N_a(t^a_\ell)\mathbf{1}(g_{a^*} > \tilde{g}_{a^*}(t^a_\ell)+\varepsilon)\right]}_{\text{(II.1)}} + \underbrace{\mathbf{E}\left[\sum_{\ell\geq 1}\big(N_a(t^a_{\ell+1}) - 2N_a(t^a_\ell)\big)\mathbf{1}(g_{a^*} > \tilde{g}_{a^*}(t^a_\ell)+\varepsilon)\right]}_{\text{(II.2)}}. \tag{D.9}$$

We bound those two terms (II.1) and (II.2) separately. We will introduce the following magical constant:

$$\beta_f := 1 + \sup\left\{t \geq 1 : f(t) \leq 2\max_{a\in\mathcal{A}}\{1+\mathrm{sp}(b_a), 3\,\mathrm{sp}(b_a)\}\right\}. \tag{D.10}$$

Note that $\beta_f < \infty$, since $f \to \infty$.

For the first term (II.1) in Equation (D.9), we write

$$\begin{aligned}
\text{(II.1)} &= \mathbf{E}\left[\sum_{\ell=1}^{\infty}\sum_{n=1}^{+\infty} n\cdot\mathbf{1}(g_{a^*} > \tilde{g}_{a^*}(t^a_\ell)+\varepsilon, N_a(t^a_\ell)=n)\right] \\
&\overset{\circ 1}{\leq} \mathbf{E}\left[\sum_{t=1}^{\infty}\sum_{n=1}^{+\infty} n\cdot\mathbf{1}\left(g_{a^*} > \hat{g}_{a^*} + \varepsilon + f(t)\sqrt{\frac{1}{n}\log\left(\frac{1+n}{\delta(t)}\right)}, N_a(t)=n\right)\right] \\
&\overset{\circ 2}{\leq} \mathbf{E}\left[\sum_{t=1}^{\infty}\sum_{n=1}^{t} n\cdot\mathbf{1}\left(g_{a^*} > \hat{g}_{a^*} + \varepsilon + f(t)\sqrt{\frac{1}{n}\log\left(\frac{1+n}{\delta(t)}\right)}\right)\right]
\end{aligned}$$

where (°1) follows by union bound over all the values that t^a_ℓ can take; and (°2) uses that $N_a(t) \leq n$.

We split the term in two depending on the value of $f(t)$. The regret contribution before time β_f is trivially bounded by β_f and is thus O(1). Consequently, we can restrict ourselves to evaluating the regret past β_f.

$$\begin{aligned}
\text{(II.1)} &\overset{\circ 3}{\leq} \sum_{t=1}^{\beta_f} n + \sum_{t=\beta_f}^{\infty}\sum_{n=1}^{t} n\cdot\exp\left\{-\frac{n\cdot\left(\left(\varepsilon + f(t)\sqrt{\frac{\log((1+n)/\delta(t))}{n}} - \frac{3\,\mathrm{sp}(b_a)}{\sqrt{n}}\right)^2 - 9\,\mathrm{sp}(b_a)^2\right)}{(1+\mathrm{sp}(b_a))^2}\right\} \\
&\overset{\circ 4}{\leq} (\beta_f)^2 + \sum_{t=\beta_f}^{\infty}\sum_{n=1}^{t} n\cdot e^9\exp\left\{-\frac{n}{(1+\mathrm{sp}(b_a))^2}\left(\varepsilon + \frac{1}{2}f(t)\sqrt{\frac{\log((1+n)/\delta(t))}{n}}\right)^2\right\} \\
&\leq (\beta_f)^2 + \sum_{t=\beta_f}^{\infty}\cdot\sum_{n=1}^{t} n\cdot e^9\exp\left\{-\frac{n\varepsilon^2 + \frac{1}{4}f(t)^2\log\left(\frac{1}{\delta(t)}\right)}{(1+\mathrm{sp}(b_a))^2}\right\} \\
&\overset{\circ 5}{\leq} (\beta_f)^2 + \sum_{t=\beta_f}^{\infty}\sum_{n=1}^{t} n\cdot e^9\delta(t)\exp\left\{-\frac{n\varepsilon^2}{(1+\mathrm{sp}(b_a))^2}\right\} \\
&\leq (\beta_f)^2 + \mathrm{O}\left(\left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^4\sum_{t=1}^{\infty}\delta(t)\right),
\end{aligned} \tag{D.11}$$

where (°3) follows from Lemma D.6 inequalities; and (°4) and (°5) stem from the definition of β_f in Equation (D.10), as for $t \geq \beta_f$, we have $\frac{1}{2}f(t)\sqrt{\frac{1}{n}\log((1+n)/\delta(t))} \geq \frac{3\,\mathrm{sp}(b_a)}{\sqrt{n}}$ and $\frac{1}{4}f(t)^2 \geq (1+\mathrm{sp}(b_a))^2$.

We continue with the second term (II.2) from Equation (D.9). We have

$$\begin{aligned}
\text{(II.2)} &\overset{\circ 1}{\leq} C_\tau\mathbf{E}\left[\sum_{\ell=1}^{\infty}\mathbf{1}(g_{a^*} > \tilde{g}_{a^*}(t^a_\ell)+\varepsilon)\right] \\
&\overset{\circ 2}{\leq} C_\tau\mathbf{E}\left[\sum_{t=1}^{\infty}\sum_{n=1}^{t}\mathbf{1}\left(g_{a^*} > \hat{g}_{a^*,n} + \varepsilon + f(t)\sqrt{\frac{1}{n}\log\left(\frac{1+n}{\delta(t)}\right)}\right)\right].
\end{aligned}$$

Again, we decompose depending on the value of $f(t)$. This derivation is similar to the one of the previous paragraph.

$$
\begin{aligned}
\text{(II.2)} &\overset{\circ 3}{\leq} C_\tau \beta_f + C_\tau \sum_{t=\beta_f}^{\infty} \sum_{n=1}^{t} \exp\left\{ -\frac{n \cdot \left(\left(\varepsilon + f(t) \sqrt{\frac{\log((1+n)/\delta(t))}{n}} - \frac{3\,\mathrm{sp}(b_a)}{\sqrt{n}} \right)^2 - 9\,\mathrm{sp}(b_a)^2 \right)}{(1+\mathrm{sp}(b_a))^2} \right\} \\
&\overset{\circ 4}{\leq} C_\tau \beta_f + C_\tau \sum_{t=\beta_f}^{\infty} \sum_{n=1}^{t} \mathrm{e}^9 \exp\left\{ -\frac{n}{(1+\mathrm{sp}(b_a))^2} \left(\varepsilon + \frac{1}{2} f(t) \sqrt{\frac{\log((1+n)/\delta(t))}{n}} \right)^2 \right\} \\
&\overset{\circ 5}{\leq} C_\tau \beta_f + C_\tau \sum_{t=\beta_f}^{\infty} \sum_{n=1}^{t} \mathrm{e}^9 \exp\left\{ -\frac{n\varepsilon^2 + \frac{1}{4} f(t)^2 \log\left(\frac{1}{\delta(t)}\right)}{(1+\mathrm{sp}(b_a))^2} \right\} \\
&\overset{\circ 6}{\leq} C_\tau \beta_f + C_\tau \sum_{t=\beta_f}^{\infty} \sum_{n=1}^{t} \mathrm{e}^9 \delta(t) \exp\left\{ -\frac{n\varepsilon^2}{(1+\mathrm{sp}(b_a))^2} \right\} \\
&\leq C_\tau \beta_f + \mathrm{O}\left(C_\tau \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon} \right)^2 \sum_{t=1}^{\infty} \delta(t) \right), \qquad \text{(D.12)}
\end{aligned}
$$

where, likewise, (°3) follows from Lemma D.6 and (°4), (°5) stem from the definition of β_f. We put the bound of (II) together by combining Equations (D.11) and (D.12) into Equation (D.9). We conclude that

$$
\text{(II)} = \mathrm{O}\left((C_\tau + \beta_f)\beta_f + \left(C_\tau \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon} \right)^2 + \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon} \right)^4 \right) \sum_{t=1} \delta(t) \right). \qquad \text{(D.13)}
$$

Optimistic pulls (III). We show that this term is dominant, and accounts for a total of $2\chi_\varepsilon(T) + C_\tau$ pulls in expectation. Given $t \geq 1$ such that $A(t) = a$, we write $\ell(t)$ the current episodes t falls in, i.e., the unique $\ell \geq 1$ such that $t^a_\ell \leq t < t^a_{\ell+1}$. Introduce the stopping time

$$
\tau_\chi := \inf \left\{ t \geq 1 : \begin{array}{c} A(t) = a, \\ N_a(t^a_{\ell(t)}) \leq \chi_\varepsilon(T), \quad N_a(t) > \chi_\varepsilon(T), \\ N_a(t) \geq 2N_a(t^a_{\ell(t)}) \end{array} \right\}. \qquad \text{(D.14)}
$$

That is, τ_χ is the first time at which $A(t) = a$ is pulled, and such that (1) the current started below the threshold with $N_a(t^a_{\ell(t)}) \leq \chi_\varepsilon(T)$, (2) the current visited count has exceeded the threshold with $N_a(t) \geq \chi_\varepsilon(T)$, and (3) the continuation of the current episode $\ell(t)$ is not triggered by the Doubling Trick since $N_a(t) \geq 2N_a(t^a_{\ell(t)})$.

We then bound (III) as follows.

$$
\begin{aligned}
\mathbf{E}\left[\sum_{\ell=1}^{\infty} \sum_{t=1}^{\infty} \mathbf{1}\begin{pmatrix} A(t) = a \\ t^a_\ell \leq t < t^a_{\ell+1} \\ N_a(t^a_\ell) \leq \chi_\varepsilon(T) \end{pmatrix} \right] &\overset{\circ 1}{=} \mathbf{E}\left[\sum_{\ell=1}^{\infty} \sum_{t=1}^{\infty} \mathbf{1}\begin{pmatrix} A(t) = a \\ t^a_\ell \leq t < t^a_{\ell+1} \\ N_a(t^a_\ell) \leq \chi_\varepsilon(T) \\ t < \tau_\chi \end{pmatrix} \right] + \mathbf{E}\left[\sum_{\ell=1}^{\infty} \sum_{t=1}^{\infty} \mathbf{1}\begin{pmatrix} A(t) = a \\ t^a_\ell \leq t < t^a_{\ell+1} \\ N_a(t^a_\ell) \leq \chi_\varepsilon(T) \\ t \geq \tau_\chi \end{pmatrix} \right] \\
&\overset{\circ 2}{\leq} \mathbf{E}[\tau_\chi] + \mathbf{E}\left[\sum_{t=1}^{\infty} \mathbf{1}\begin{pmatrix} A(t) = a \\ t^a_{\ell(\tau_\chi)} \leq t < t^a_{\ell(\tau_\chi)+1} \\ t \geq \tau_\chi \end{pmatrix} \right] \\
&\overset{\circ 3}{\leq} 2\chi_\varepsilon(T) + \mathbf{E}[\inf\{\tau_i - \tau_\chi : \tau_i \geq \tau_\chi\}] \overset{\circ 4}{\leq} 2\chi_\varepsilon(T) + C_\tau,
\end{aligned}
$$

where (°1) splits the sum according to the stopping time τ_χ introduces by Equation (D.14); (°2) bounds the first term by τ_χ by using that every pull of a makes its visit counts N_a increase by 1; (°2) bounds the second term using the observation that, by definition of τ_χ, there is at most one episode $\ell \geq 1$ such that $N_a(t^a_\ell) \leq \chi_\varepsilon(T)$ that contains some $t \geq \tau_\chi$; (°3) bounds the first term using that $N_a(\tau_\chi) \leq 2x$ a.s., because if

$\tau_\chi \geq 2x+1$, one checks that $\tau_\chi - 1$ also satisfies Equation (D.14); (°3) bounds the second term by observing that, by definition of τ_χ, the current episode $\ell(\tau_\chi)$ has passed the doubling trick rule at time τ_χ, so that the episode will stop at the next decision epoch τ_i; and finally, (°4) bounds the second term by invoking Assumption 3.

Late bloomers (IV). This term is almost surely 0.
Indeed, when term (IV) $= 1$ we have

$$\begin{aligned}\tilde{g}_{a^*}(t_\ell^a) - \tilde{g}_a(t_l^a) &\overset{\circ 1}{\geq} g_{a^*} - \varepsilon - (\hat{g}_a(t_\ell^a) + \xi(t_\ell^a, \delta(t_\ell^a)) \\ &\overset{\circ 2}{\geq} g_{a^*} - g_a - 2\varepsilon - \xi(t_\ell^a, \delta(t_\ell^a)) \overset{\circ 3}{>} 0,\end{aligned}$$

where (°1) follow from $g_{a^*} \leq \tilde{g}_{a^*}(t_\ell^a) + \varepsilon$; (°2) from $\hat{g}_a(t_\ell^a) \leq g_a + \varepsilon$; and (°3) from $N_a(t_\ell^a) > \chi_\varepsilon(T)$. Thus, by definition, the algorithm will not select arm a at time t_ℓ^a.

Bounding the visit threshold χ_ε. Last but not least, we must upper-bound $\chi_\varepsilon(T)$ from Equation (D.3), that appears in the bound of (III). Denote $n \equiv \chi_\varepsilon(T) - 1$. Then n is such that

$$\sqrt{\frac{\log((1+n)/\delta(T))}{n}} + \frac{\log_2(1+n)}{n} \geq \frac{g_* - g_a - 2\varepsilon}{f(T)}.$$

From straight-forward algebra, we find that n must satisfy one of the three conditions below:

$$\frac{\log(1+n)}{n} \geq \frac{1}{8}\left(\frac{g_* - g_a - 2\varepsilon}{f(T)}\right)^2, \text{ or } \frac{\log\left(\frac{1}{\delta(T)}\right)}{n} \geq \frac{1}{8}\left(\frac{g_* - g_a - 2\varepsilon}{f(T)}\right)^2, \text{ or } \frac{\log_2(1+n)}{n} \geq \frac{1}{2}\frac{g_* - g_a - 2\varepsilon}{f(T)}.$$

Using $\log(1+x) \leq \sqrt{x}$, we find that

$$\begin{aligned}\chi_\varepsilon(T) &\leq 1 + \max\left\{\left(\frac{2\sqrt{2}f(T)}{g_* - g_a - 2\varepsilon}\right)^4, \frac{8f(T)^2\log\left(\frac{1}{\delta(T)}\right)}{(g_* - g_a - 2\varepsilon)^2}, \left(\frac{2f(T)}{g_* - g_a - 2\varepsilon}\right)^2\right\} \\ &= \mathrm{O}\left(\frac{f(T)^2\log(T)}{(g_* - g_a - 2\varepsilon)^2} + \left(\frac{f(T)}{g_* - g_a - 2\varepsilon}\right)^4\right). \end{aligned} \tag{D.15}$$

Note that because $\log(\frac{1}{\delta(t)}) = \Theta(\log(t))$ and $f(T)^4 = \mathrm{o}(\log(T))$, the term in $f(T)\log(\frac{1}{\delta(T)})$ is dominant.

Conclusion. Combining everything, we find that

$$\begin{aligned}\mathbf{E}[N_a(T)] &\leq 2\chi_\varepsilon(T) + \mathrm{O}\left((\beta_f)^2 + C_\tau\left(\beta_f + \frac{1}{\varepsilon^4}\right) + \left(C_\tau\left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^2 + \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^4\right)\sum_{t=1}^{\infty}\delta(t)\right) \\ &\overset{\circ 1}{=} \mathrm{O}\left(\frac{f(T)^2\log(T)}{(g_* - g_a - 2\varepsilon)^2} + \left(\frac{f(T)}{g_* - g_a - 2\varepsilon}\right)^4 + (\beta_f)^2 + C_\tau\left(\beta_f + \frac{1}{\varepsilon^4} + \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^2\right) + \left(\frac{1+\mathrm{sp}(b_a)}{\varepsilon}\right)^4\right)\end{aligned}$$

Hence the result. □

Proof of Theorem D.1. For $a \in \mathcal{A}$ such that $g_a < g_*$, denote $\Delta_a := g_* - g_a$ the gain-gap.

We apply Lemma D.2 for every sub-optimal arm independently, with $\varepsilon_a := \frac{1}{4}\Delta_a$ to bound $\mathbf{E}[N_a(T)]$. By invoking that $\mathbf{E}[\mathrm{PR}(T)] = \sum_{a \notin \mathcal{A}^*}\mathbf{E}[N_a(T)]\Delta_a$, we obtain that the expected pseudo-regret is bounded as:

$$\sum_{a \notin \mathcal{A}^*}\left(\frac{f(T)^2\log(T)}{\Delta_a} + \frac{f(T)^4}{\Delta_a^3} + (\beta_f)^2\Delta_a + C_\tau\Delta_a\left(\beta_f + \frac{1}{\Delta_a^4} + \left(\frac{1+\mathrm{sp}(b_a)}{\Delta_a}\right)^2\right) + \left(\frac{1+\mathrm{sp}(b_a)}{\Delta_a}\right)^4\right). \tag{D.16}$$

We are left to relate the regret to the pseudo-regret. Because UCB-NOM relies on the Doubling Trick to manage episodes, we see that it is $\log_2$-rarely switching. Therefore, we can relate the regret to the pseudo-regret up to the number of switches via Lemma A.4. Then, we bound the number of switches via the number of episodes with Lemma D.7, and obtain that, in general,

$$\begin{aligned}\mathrm{Reg}(T) &= \mathrm{O}\left(\sum_{a\notin\mathcal{A}^*}\left(\frac{f(T)^2\log(T)}{\Delta_a}+\frac{f(T)^4}{\Delta_a^3}\right)+|\mathcal{A}|\max_{a\in\mathcal{A}}\mathrm{sp}(b_a)\log(T)\right) && \begin{pmatrix}\text{DIVERGING}\\ \text{TERMS}\end{pmatrix}\\ &+\mathrm{O}\left(\sum_{a\notin\mathcal{A}^*}\Delta_a\left((\beta_f)^2+C_\tau\left(\beta_f+\frac{1}{\Delta_a^4}+\left(\frac{1+\mathrm{sp}(b_a)}{\Delta_a}\right)^2\right)+\left(\frac{1+\mathrm{sp}(b_a)}{\Delta_a}\right)^4\right)\right). && \begin{pmatrix}\text{CONSTANT}\\ \text{TERMS}\end{pmatrix}\end{aligned}$$

Moreover, in case where there is a single optimal arm, i.e., $|\mathcal{A}^*|=1$, the bound can be improved by using a better bound on the number of switches. Using Proposition A.3, we obtain

$$\begin{aligned}\mathrm{Reg}(T) &= \mathrm{O}\left(\sum_{a\notin\mathcal{A}^*}\left(\frac{f(T)^2\log(T)}{\Delta_a}+\frac{f(T)^4}{\Delta_a^3}\right)+|\mathcal{A}|\max_{a\in\mathcal{A}}\mathrm{sp}(b_a)\log\big(f(T)^2\log(T)\big)\right) && \begin{pmatrix}\text{DIVERGING}\\ \text{TERMS}\end{pmatrix}\\ &+\mathrm{O}\left(\sum_{a\notin\mathcal{A}^*}\Delta_a\left((\beta_f)^2+C_\tau\left(\beta_f+\frac{1}{\Delta_a^4}+\left(\frac{1+\mathrm{sp}(b_a)}{\Delta_a}\right)^2\right)+\left(\frac{1+\mathrm{sp}(b_a)}{\Delta_a}\right)^4\right)\right). && \begin{pmatrix}\text{CONSTANT}\\ \text{TERMS}\end{pmatrix}\end{aligned}$$

This concludes the proof of Theorem D.1. □

D.2 Proof of Theorem 8: Regret bound under bounded span proxy

In this section, we show that when a uniform bound on the bias span is known, the diverging function f introduced in Appendix D.1 is no longer necessary. We can now run algorithm UCB-NOM with a *constant* function instead. This additional knowledge allow us to improve on the regret obtained in Theorem D.1 and to reach a logarithmic regret. The theorem below summarize the improvement on the regret allowed by this additional assumption.

Theorem D.3 (Instance-dependent regret of UCB-NOM when an upper bound on the bias is available.)**.** *Let* $C_b := 1+2\max(1+C_b^0, 3C_b^0)$. *Let* $\delta:\mathbf{N}\to\mathbf{R}_+$ *be a confidence function such that* $\sum_{t=1}^{\infty}\delta(t)<\infty$ *and* $\log(\frac{1}{\delta(t)})=\mathrm{O}(\log(t))$. *For all instance* $M\in\mathcal{M}$, $\mathrm{Reg}(T;M,\texttt{UCB-NOM}(C_b))$ *is of order*

$$\mathrm{Reg}(T;M,\texttt{UCB-NOM}(C_b))=\mathrm{O}\left(\sum_{a\notin\mathcal{A}^*}\frac{(C_b^0)^2\log(T)}{g_*-g_a}+|\mathcal{A}|\max_{a\in\mathcal{A}}\mathrm{sp}(b_a)\log(T)\right). \tag{D.17}$$

Moreover, in case where there is a single optimal arm $|\mathcal{A}^*|=1$, *the bound is improved to*

$$\mathrm{Reg}(T;M,\texttt{UCB-NOM}(C_b))=\mathrm{O}\left(\sum_{a\notin\mathcal{A}^*}\frac{(C_b^0)^2\log(T)}{g_*-g_a}\right). \tag{D.18}$$

Note that the number of episodes may be dominant this time around: If there are multiple optimal arms, UCB-NOM switches $\log(T)$ many times between them and may endure an extra cost of $|\mathcal{A}|\max_{a\in\mathcal{A}}\mathrm{sp}(b_a)\log(T)$. This term is comparable to the main term in $(C_b^0)^2\log(T)$. However, switching costs are negligible if there is a unique optimal arm, because one can show that there are $\mathrm{O}(\log\log(T))$ of them. Making the switching costs $\mathrm{o}(\log(T))$ when $|\mathcal{A}^*|>1$ is far from easy, and is the reason why in multi-armed bandits with switching costs Ortner (2010); Agrawal et al. (2002); Jun (2004), the assumption $|\mathcal{A}|=1$ is often made to achieve asympotically optimal regret Agrawal et al. (2002).

Regarding the proof of Theorem D.3, it is almost completely identical to the one of Theorem D.1.

Lemma D.4 (Number of vistis, f bounded)**.** *Consider a suboptimal arm* $a\notin\mathcal{A}^*$. *We have*

$$\mathbf{E}[N_a(T)]=\mathrm{O}\left(\frac{(C_b^0)^2\log\left(\frac{1}{\delta(T)}\right)}{(g_*-g_a-2\varepsilon)^2}+\frac{1}{\varepsilon^4}+\frac{(C_\tau)^2+(1+C_b^0)^2}{\varepsilon^2}\right). \tag{D.19}$$

Proof. The proof of Lemma D.4 follows closely the line of Lemma D.2. Except for the changes noted below, the reasoning carries over verbatim after replacing f with C_b. We therefore isolate the steps that must be altered, specify the required modifications, and derive the corresponding implications for our algorithm's regret.

The analysis of $\mathbf{E}[N_a(T)]$ splits the number of visits into for terms: empirical errors (I), optimism failures (II), optimistic pulls (III) and late bloomers (IV), see Equation (D.4).

The computations for the terms (I), (III) and (IV) carry on from Lemma D.2 to Lemma D.4, albeit f must be changed to C_b. The only real change can be found in term (II): The original proof relied on the diverging property of f, by introducing the ad-hoc constant β_f in Equation (D.10). The analysis of (II) then split the sums between what happens before and after the times at which $f(t)$ is large enough to upper bound C_b. If a bound on the bias is already known, this is no longer necessary, because $f(t) \geq C_b$. This simplifies the calculation by half.

Putting everything together, we obtain an equation of the form of Equation (D.19). □

Proof of Theorem D.3. Once Lemma D.4 is established, the proof is concluded just as the one of Theorem D.1, by picking $\varepsilon = ((\min_{a \notin \mathcal{A}^*}(g_* - g_a))^2/((C_b)^2 \log(T)))^{\frac{1}{5}}$ and simplifying the bound. □

D.3 Proof of Theorem 9: regret upper bound under bounded span proxy

The goal of this section is to derive worst-case regret of `UCB-NOM` when a bound on the bias span and the decision epoch delays is known. Just like for the instance dependent upper bound, we provide a finer result below (Theorem D.5) and explain how the result from the main body (Theorem 9) is an instantiation of it. First, we refine the set $\mathcal{M}_D$ from the main body by differentiating between the constraints $1 + \max_{a \in \mathcal{A}} \mathrm{sp}(b_a) \leq D$ and $C_\tau \leq D$. For $C_b^0, C_\tau^0 \in \mathbf{R}_+$, let

$$\mathcal{M}_{C_b^0, C_\tau^0} := \left\{ M \in \mathcal{M} : 1 + \max_{a \in \mathcal{A}} \mathrm{sp}(b_a) \leq C_b^0 \text{ and } C_\tau \leq C_\tau^0 \right\}. \tag{D.20}$$

On $\mathcal{M}_{C_b^0, C_\tau^0}$, we derive the following worst-case regret guarantees.

Theorem D.5 (Worst-case regret of `UCB-NOM`)**.** *Let $C_b^0 \in \mathbf{R}_+$. Let $\delta : \mathbf{N} \to \mathbf{R}_+$ be a confidence function such that $\sum_{t=1}^{\infty} t \cdot \delta(t) < \infty$ and $\log(\frac{1}{\delta(t)}) = \Theta(\log(t))$. For all instances $M \in \mathcal{M}_{C_b^0, C_\tau^0}$, `UCB-NOM`$(C_b^0)$ has the following worst-case regret:*

$$\mathrm{Reg}(T; M, \texttt{UCB-NOM}(C_b^0)) = \mathrm{O}\Big(C_b^0 \sqrt{|\mathcal{A}| T \log(T)} + C_b^0 C_\tau^0 |\mathcal{A}| \log(T)^2 \Big). \tag{D.21}$$

Observe that, in opposition to the instance-dependent bound, the cost due to delays (Assumption 3) scales with the number of episodes as a second order term. The dominant term $C_b^0 \sqrt{|\mathcal{A}| T \log(T)}$ scales exclusively with the span bound. Because the span bound C_b^0 is to be analogous to a diameter-like quantity D, this term is to be compared with the famous $D|\mathcal{S}|\sqrt{|\mathcal{A}| T \log(T)}$ regret bound of `UCRL2` (Auer et al. (2009)), for which the dependency in D rather than $\sqrt{D}$ is well-known to be suboptimal.

Proof of Theorem D.5. Let $M \in \mathcal{M}$ such that $1 + \max_{a \in \mathcal{A}} \mathrm{sp}(b_a) \leq C_b^0$, and such that $C_\tau \leq C_\tau^0$. Fix a horizon $T \geq 1$ and an initial state $s_0 \in \mathcal{S}$. Because the dependencies in M, $\Lambda \equiv$ `UCB-NOM`, T and s_0 are fixed from now on, these objects are dropped in notations.

We write $(t'_k)_{k \geq 1}$ the sequence of episodes clipped to $\{1, \ldots, T-1\}$. That is, $t'_k := \min\{t_k, T-1\}$.

We start by rewriting the regret relatively to the optimal gain and episodes. We have

$$\mathrm{Reg}(T) \overset{\circ 1}{\leq} \mathbf{E}\left[\sum_{t=1}^{T-1} (g_* - R(t+1)) \right] + C_b^0 \overset{\circ 2}{=} \mathbf{E}\left[\sum_{k=1}^{\infty} \sum_{t=t'_k}^{t'_{k+1}-1} (g_* - R(t+1)) \right] + C_b^0 \tag{D.22}$$

where (°1) follows by Corollary A.2; and (°2) rewrites the sum according to episodes.

Then, the summand $g_* - R(t+1)$ is expanded into sub-terms, each one corresponding to an error of a different nature. Fix any episode k and time $t \in [t_k, t_{k+1} - 1]$. We can write

$$\begin{aligned} & g_* - R(t+1) \\ &= \underbrace{g_* - \tilde{g}_*(t'_k)}_{\substack{\text{Optimism}\\\text{Failure}}\\(\text{I})} + \underbrace{\tilde{g}_*(t'_k) - \tilde{g}_{A(t_k)}(t'_k)}_{\substack{\text{Arm Selection}\\\text{Slackness}}\\(\text{II})} + \underbrace{\tilde{g}_{A(t'_k)}(t'_k) - \hat{g}_{A(t'_k)}(t'_k)}_{\substack{\text{Optimism}\\\text{Overshot}}\\(\text{III})} + \underbrace{\hat{g}_{A(t'_k)}(t'_k) - g_{A(t'_k)}}_{\substack{\text{Empirical}\\\text{Error}}\\(\text{IV})} + \underbrace{g_{A(t'_k)} - R(t+1)}_{\substack{\text{Navigation}\\\text{Error}}\\(\text{V})} \end{aligned}$$

We proceed to bound each term's contribution separately in expectation over all episodes and times.

Optimism failure (I) This term is shown to be O(1). We first define the "good event" at episode k by:

$$\mathcal{E}(t'_k) := \{\tilde{g}_*(t'_k) > g_*\}. \tag{D.23}$$

The main idea of this first part of the proof is that, in the "good event" $\mathcal{E}(t_k)$ where the gain of the optimal arm lie below its optimistic estimator — and therefore below its upper-confidence bound — (Term I) is negative. We then use the concentration inequality (Proposition 5) to show that the probability of the "bad event" $\mathcal{E}(t'_k)^c$ decays sufficiently quickly, so that its total error contribution is small.

$$\begin{aligned} & \mathbf{E}\left[\sum_{k=1}^{\infty}\sum_{t=t'_k}^{t'_{k+1}-1}(\text{I})\right] \\ & \overset{\circ 1}{\leq} \mathbf{E}\left[\sum_{k=1}^{\infty}\sum_{t=t'_k}^{t'_{k+1}-1}\mathbf{1}(\tilde{g}_*(t'_k) < g_*)\right] \\ & \overset{\circ 2}{\leq} \mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T-1)\sum_{t=t_k}^{t_{k+1}-1}\mathbf{1}(\tilde{g}_*(t_k) < g_*)\right] \\ & \overset{\circ 3}{\leq} \underbrace{\mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T-1)\sum_{t=t_k}^{t_k^{\mathrm{DT}}-1}\mathbf{1}(\tilde{g}_*(t_k) < g_*)\right]}_{(\text{I.1})} + \underbrace{\mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T-1)\sum_{t=t_k^{\mathrm{DT}}}^{t_{k+1}-1}\mathbf{1}(\tilde{g}_*(t_k) < g_*)\right]}_{(\text{I.2})} \end{aligned}$$

where (°1) holds since $g_* - \tilde{g}_*(t_k) \leq 1$ a.s., and on the good event $\mathcal{E}(t_k)$ from Equation (D.23), we have $g_* - \tilde{g}_*(t_k) \leq 0$; (°2) just rewrites the sum of (°1); and (°3) splits every episode according to whether the doubling trick rule is passed or not, by setting $t_k^{\mathrm{DT}} := t_k + N_{A(t_k)}(t_k)$. Note that t_k^{DT} is a stopping time.

We then bound the terms (I.1) and (I.2) separately. For the first term (I.1), we have

$$(\text{I.1}) \overset{\circ 1}{=} \mathbf{E}\left[\sum_{k=1}^{\infty} t_k \cdot \mathbf{1}\begin{pmatrix} t_k < T \\ \tilde{g}_*(t_k) < g_* \end{pmatrix}\right] \overset{\circ 2}{=} \mathbf{E}\left[\sum_{t=1}^{T-1} t \cdot \mathbf{1}(\tilde{g}_*(t) < g_*)\right] \overset{\circ 3}{\leq} \sum_{t=1}^{\infty} t \cdot \delta(t) = \mathrm{O}(1)$$

where in (°1) we state that $t_k^{\mathrm{DT}} \leq 2t_k$, hence that $t_k^{\mathrm{DT}} - t_k \leq t_k$; (°2) we use that the left sum is a subset of the right one; and (°2) we invoke Proposition 5.

For the second term (I.2), we have

$$\begin{aligned} (\text{I.2}) &= \mathbf{E}\left[\sum_{k=1}^{\infty}(t_{k+1} - t_k^{\mathrm{DT}})\mathbf{1}\begin{pmatrix} t_k < T \\ \tilde{g}_*(t_k) < g_* \end{pmatrix}\right] \\ &\overset{\circ 1}{=} \mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}\begin{pmatrix} t_k < T \\ \tilde{g}_*(t_k) < g_* \end{pmatrix}\mathbf{E}\left[t_{k+1} - t_k^{\mathrm{DT}} \middle| H(t_k^{\mathrm{DT}})\right]\right] \end{aligned}$$

$$\overset{\circ 2}{=} \mathbf{E}\left[\sum_{k=1}^{\infty} \mathbf{1}\begin{pmatrix} t_k < T \\ \tilde{g}_*(t_k) < g_* \end{pmatrix} \mathbf{E}\left[\inf\{t - t_k^{\mathrm{DT}} : t \geq t_k^{\mathrm{DT}}, t \in \{\tau_i : i \geq 1\}\} \middle| H(t_k^{\mathrm{DT}})\right]\right]$$

$$\overset{\circ 3}{\leq} C_\tau \mathbf{E}\left[\sum_{k=1}^{\infty} \mathbf{1}\begin{pmatrix} t'_k < T \\ \tilde{g}_*(t'_k) < g_* \end{pmatrix}\right] \leq C_\tau \mathbf{E}\left[\sum_{t=1}^{T-1} \mathbf{1}(\tilde{g}_*(t) < g_*)\right] \overset{\circ 4}{\leq} C_\tau \sum_{t=1}^{\infty} \delta(t) = \mathrm{O}(1)$$

where (°1) stems from the tower property, and uses that $\mathbf{1}(t_k < T, \tilde{g}_*(t_k) < g_*)$ is $\sigma(H(t_k^{\mathrm{DT}}))$-measurable; (°2) uses that once t_k^{DT} is passed, episode k ends as soon as a decision epoch τ_i triggers; (°3) bounds the delay before the decision epoch via Assumption 3; and (°4) follows from Proposition 5.

Combining (I.1) and (I.2) we conclude that the total contribution to the regret from term I is O(1).

Slackness of arm selection (II). By construction of our algorithm `UCB-NOM`, at the beginning of every episodes k, `UCB-NOM` picks the arm $A(t_k)$ achieving $\tilde{g}_a(t_k) := \max_{a \in \mathcal{A}} \tilde{g}_a(t_k)$. So

$$\mathbf{E}\left[\sum_{k=1}^{\infty} \sum_{t=t'_k}^{t'_{k+1}-1} (\mathrm{II})\right] = 0.$$

Optimism overshot (III). This term is the dominant one. We start by writing it as follows:

$$\mathbf{E}\left[\sum_{k=1}^{\infty} \sum_{t=t'_k}^{t'_{k+1}-1} (\mathrm{III})\right]$$

$$\overset{\circ 1}{=} \mathbf{E}\left[\sum_{k=1}^{\infty} (t'_{k+1} - t'_k) \cdot C_b^0 \left(\sqrt{\frac{1}{N_{A(t'_k)}(t'_k)} \log\left(\frac{1 + N_{A(t'_k)}(t'_k)}{\delta(t'_k)}\right)} + \frac{\log\left(1 + N_{A(t'_k)}(t'_k)\right)}{N_{A(t'_k)}(t'_k)}\right)\right] \tag{D.24}$$

$$\overset{\circ 2}{\leq} \mathbf{E}\left[\sum_{k=1}^{\infty} \mathbf{1}(t_k \leq T)(t_{k+1} - t_k) \cdot C_b^0 \left(\sqrt{\frac{1}{N_{A(t'_k)}(t'_k)} \log\left(\frac{1+T}{\delta(T)}\right)} + \frac{\log(1+T)}{N_{A(t'_k)}(t'_k)}\right)\right]$$

$$\overset{\circ 3}{\leq} C_b^0 \cdot \mathbf{E}\left[\sum_{k=1}^{\infty} \sum_{a \in \mathcal{A}} \mathbf{1}\begin{pmatrix} t_k \leq T \\ A(t_k) = a \end{pmatrix} \sum_{t=t_k}^{t_k^{\mathrm{DT}}-1} \left(\sqrt{\frac{\log\left(\frac{1+T}{\delta(T)}\right)}{N_a(t_k)}} + \frac{\log(1+T)}{N_a(t_k)}\right)\right] \Bigg\} \text{(III.1)}$$

$$+ C_b^0 \cdot \mathbf{E}\left[\sum_{k=1}^{\infty} \sum_{a \in \mathcal{A}} \mathbf{1}\begin{pmatrix} t_k \leq T \\ A(t_k) = a \end{pmatrix} \sum_{t=t_k^{\mathrm{DT}}}^{t_{k+1}-1} \left(\sqrt{\frac{\log\left(\frac{1+T}{\delta(T)}\right)}{N_a(t_k)}} + \frac{\log(1+T)}{N_a(t_k)}\right)\right] \Bigg\} \text{(III.2)} \tag{D.25}$$

where (°1) unfolds the definition of the index $\tilde{g}_a(t)$; (°2) simplifies the expression using that $\delta(-)$ is non-increasing and the non-negativity of the summand; and (°3) splits every episode according to whether the doubling trick rule is passed or not, by setting $t_k^{\mathrm{DT}} := t_k + N_{A(t_k)}(t_k)$. We continue by bounding (III.1) and (III.2) in Equation (D.25).

We begin with (III.1), which is shown to dominate (III.2). We have

$$(\mathrm{III.1}) \overset{\circ 1}{\leq} 2C_b^0 \cdot \mathbf{E}\left[\sum_{k=1}^{\infty} \sum_{a \in \mathcal{A}} \mathbf{1}\begin{pmatrix} t_k \leq T \\ A(t_k) = a \end{pmatrix} \sum_{t=t_k}^{t_k^{\mathrm{DT}}-1} \left(\sqrt{\frac{\log\left(\frac{1+T}{\delta(T)}\right)}{N_a(t)}} + \frac{\log(1+T)}{N_a(t)}\right)\right]$$

$$\overset{\circ 2}{\leq} 2C_b^0 \sqrt{\log\left(\frac{1+T}{\delta(T)}\right)} \cdot \mathbf{E}\left[\sum_{a \in \mathcal{A}} \sum_{n=1}^{N_a(t)} \sqrt{\frac{1}{n}}\right] + 2C_b^0 \log(1+T) \cdot \mathbf{E}\left[\sum_{a \in \mathcal{A}} \sum_{n=1}^{N_a(t)} \frac{1}{n}\right]$$

$$\overset{\circ 3}{\leq} 4C_b^0 \sqrt{\log\left(\frac{1+T}{\delta(T)}\right)} \cdot \mathbf{E}\left[\sum_{a \in \mathcal{A}} \sqrt{N_a(T)}\right] + 4C_b^0 \log(1+T) \cdot \mathbf{E}\left[\sum_{a \in \mathcal{A}} \log(1 + N_a(T))\right]$$

$$\overset{\circ 4}{\leq} 4C_b^0\sqrt{|\mathcal{A}|T\log\left(\frac{1+T}{\delta(T)}\right)} + 4C_b^0|\mathcal{A}|\log(1+T)^2 \tag{D.26}$$

where (°1) changes $N_a(t_k)$ to $N_a(t)$ in (III.1) by making the observation that, for $t_k \leq t \leq t_k^{\mathrm{DT}} - 1$ and writing $a = A(t_k)$, we have $N_a(t) \leq 2N_a(t_k)$ by definition of the doubling trick; (°2) rearranges both sums; (°3) follows from standard analysis; and (°4) uses that $\sum_{a\in\mathcal{A}} N_a(T) = T$ and invokes Cauchy-Schwartz inequality.

Continuing with (III.2), we have

$$\begin{aligned}
\text{(III.2)} &\overset{\circ 1}{\leq} C_b^0 \cdot \mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T)\sum_{t=t_k^{\mathrm{DT}}}^{t_{k+1}-1}\left(\sqrt{\log\left(\frac{1+T}{\delta(T)}\right)} + \log(1+T)\right)\right] \\
&\overset{\circ 2}{=} C_b^0 \cdot \left(\sqrt{\log\left(\frac{1+T}{\delta(T)}\right)} + \log(1+T)\right) \cdot \mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T)\inf\{\tau_i - t_k^{\mathrm{DT}} : \tau_i \geq t_k^{\mathrm{DT}}\}\right] \\
&\overset{\circ 3}{\leq} C_b^0 C_\tau \cdot \left(\sqrt{\log\left(\frac{1+T}{\delta(T)}\right)} + \log(1+T)\right) \cdot \mathbf{E}[K(T)] \\
&\overset{\circ 4}{=} C_b^0 C_\tau |\mathcal{A}| \cdot \left(\sqrt{\log\left(\frac{1+T}{\delta(T)}\right)} + \log(1+T)\right) \cdot \log(2T)
\end{aligned} \tag{D.27}$$

where (°1) merely simplifies the expression by removing the visit counts; (°2) uses that once t_k^{DT} is passed, episode k ends as soon as a decision epoch τ_i triggers; (°3) bounds the expected delay before the next decision epoch using Assumption 3 and the Tower Property — t_k is $\sigma(H(t_k^{\mathrm{DT}}))$-measurable indeed; and (°4) bounds the number of episodes using Lemma D.7.

Combining Equations (D.26) and (D.27) into Equation (D.25), we obtain a bound on (III).

Empirical error (IV). For this term, introduce the good event

$$\mathcal{E}(T) := \left(\forall t \leq T, \forall a \in \mathcal{A}: \quad |g_a(t) - g_a| \leq C_b^0 \cdot \left(\sqrt{\frac{2\log(1+T)}{N_a(t)}} + \frac{\log(1+T)}{N_a(t)}\right)\right). \tag{D.28}$$

By Proposition 5, we have $\mathbf{P}(\mathcal{E}(T)^c) \leq T^{-1}$. Now, we bound (IV) as follows.

$$\text{(IV)} \leq \mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T)\sum_{t=t_k}^{t_{k+1}-1}\left|\hat{g}_{A(t_k)}(t_k) - g_{A(t_k)}\right|\right] \tag{D.29}$$

$$\overset{\circ 1}{\leq} \mathbf{E}[T \cdot \mathbf{1}(\mathcal{E}(T)^c)] + \mathbf{E}\left[\sum_{k=1}^{\infty}\mathbf{1}(t_k \leq T)\sum_{t=t_k}^{t_{k+1}-1} C_b^0 \cdot \left(\sqrt{\frac{2\log(1+T)}{N_a(t)}} + \frac{\log(1+T)}{N_a(t)}\right)\right] \tag{D.30}$$

$$\overset{\circ 2}{\leq} 1 + 4C_b^0 \cdot \left(\sqrt{2|\mathcal{A}|T\log(1+T)} + |\mathcal{A}|\log(1+T)\right) \tag{D.31}$$

where (°1) cuts the summand through the event $\mathcal{E}(T)$ of Equation (D.28) and uses that $\left|\hat{g}_{A(t_k)}(t_k) - g_{A(t_k)}\right| \leq 1$ a.s.; and (°2) is obtained via a similar line as the one (III.2), see Equation (D.26).

Navigation error (V). This term is due to switches from one arm to another, and is bounded by the number of episodes. We have

$$\begin{aligned}
\mathbf{E}\left[\sum_{k=1}^{\infty}\sum_{t=t'_k}^{t'_{k+1}-1}(\mathrm{V})\right] &\overset{\circ 1}{=} \mathbf{E}\left[\sum_{k=1}^{\infty}\sum_{t=t'_k}^{t'_{k+1}-1}\left(b_{A(t'_k)}(S_{A(t'_k)}(t+1)) - b_{A(t'_k)}(S_{A(t'_k)}(t))\right)\right] \\
&\overset{\circ 2}{=} \mathbf{E}\left[\sum_{k=1}^{\infty}\left(b_{A(t'_k)}(S_{A(t'_k)}(t'_{k+1})) - b_{A(t'_k)}(S_{A(t'_k)}(t'_k))\right)\right]
\end{aligned}$$

$$\leq \max_{a\in\mathcal{A}} \mathrm{sp}(b_a)\cdot \mathbf{E}[K(T)] \overset{\circ 3}{\leq} 2|\mathcal{A}|C_b^0 \log(T)$$

where (°1) invokes the Poisson equation of the pure policy $\pi_{A(t'_k)}$; (°2) recognizes a telescopic sum; and (°3) bounds the expected number of episodes via Lemma D.7.

Conclusion. We conclude by merging the bounds of (I), (II), (III), (IV) and (V) together. □

D.4 Auxiliary structural results on `UCB-NOM`

In this section we provide a large range of results about `UCB-NOM` that are used throughout its regret analysis.

D.4.1 Concentration results of the empirical gain estimator

In this paragraph, we provide a deviation inequality on the quality of the empirical gain estimator $\hat{g}_{a,n}$ with Lemma D.6. This result is a variant of Proposition 5, which is expressed by putting the emphasis on the error probability of error rather than on the expected deviation from the mean.

Lemma D.6. *The empirical estimator of arm a satisfies, for all $n \geq 1$ and $x \geq 0$:*

$$\begin{aligned}\mathbf{P}(\hat{g}_{a,n} - g_a \geq x) &\leq \exp\left(-\frac{nx^2 - 2x\,\mathrm{sp}(b_a)(1+\log_2(1+n))}{(1+\mathrm{sp}(b_a))^2}\right)\\ &\leq \exp\left(-\frac{n\left(x - \frac{3\,\mathrm{sp}(b_a)}{\sqrt{n}}\right)^2 - 9\,\mathrm{sp}(b)^2}{(1+\mathrm{sp}(b_a))^2}\right).\end{aligned}$$

Proof. Let (τ_k^a) be the sequence of stopping times enumerating the pulls of arm a, that is, $\tau_1^a := \inf\{t \geq 1 : A(t) = a\}$ and $\tau_{i+1}^a := \inf\{t > \tau_i^a : A(t) = a\}$. To begin with, introduce the following martingale noise terms:

$$\begin{aligned}W_{a,n}^r &:= \sum_{k=1}^n \Big(R(\tau_k^a + 1) - r_a(S_a(\tau_k^a))\Big),\\ W_{a,n}^p &:= \sum_{k=1}^n \Big(b_a(S_a(\tau_k^a + 1)) - p_a^1(S_a(\tau_k^a))b_a\Big).\end{aligned}$$

The quantities $W_{a,n}^r$ and $W_{a,n}^p$ respectively measure how much the immediate rewards (resp. state transitions) obtained by activating arm a deviate from their mean after n pulls of arm a. The aggregate noise is denoted $W_{a,n} := W_{a,n}^r + W_{a,n}^p$. Now, we bound the deviations of $g_{a,n}$ as follows. For $x \geq 0$, we have

$$\begin{aligned}\mathbf{P}(\hat{g}_{a,n} - g_a \geq x) &\overset{\circ 1}{=} \mathbf{P}\left(\sum_{k=1}^n R(\tau_k^a + 1) - g_a \geq nx\right)\\ &\overset{\circ 2}{=} \mathbf{P}\left(\sum_{k=1}^n \Big(W_{a,n} + \Big(e_{S_a(\tau_k^a)} - e_{S_a(\tau_k^a+1)}\Big)b_a\Big) \geq nx\right)\\ &\overset{\circ 3}{\leq} \mathbf{P}\left(\sum_{k=1}^n W_{a,n} + \mathrm{sp}(b_a)(1+\log_2(1+n)) \geq nx\right)\\ &\overset{\circ 4}{\leq} \exp\left(-\frac{n\left(x - \frac{\mathrm{sp}(b_a)(1+\log_2(1+n))}{n}\right)^2}{(1+\mathrm{sp}(b_a))^2}\right)\\ &\overset{\circ 5}{\leq} \exp\left(-\frac{nx^2 - 2x\,\mathrm{sp}(b_a)(1+\log_2(1+n))}{(1+\mathrm{sp}(b_a))^2}\right)\end{aligned}$$

$$\overset{\circ 6}{\leq} \exp\left(-\frac{n\left(x - \frac{3\,\mathrm{sp}(b_a)}{\sqrt{n}}\right)^2 - 9\,\mathrm{sp}(b)^2}{(1+\mathrm{sp}(b_a))^2}\right)$$

where (°1) unfolds the definition of the empirical estimate $g_{a,n}$ (see Equation (9)); (°2) follows from the Poisson equation of π_a; (°3) bounds the switching term using Lemma D.9; (°4) stems from Hoeffding inequality (Lemma D.10); and (°5) uses the identity $(u-v)^2 \geq u^2 - 2uv$. As for (°6), it follows from simple algebraic manipulations, that we detail below: Using that $1 + \log_2(1+n) \leq 3\log(1+n) \leq 3\sqrt{n}$, we have

$$\begin{aligned} nx^2 - 2x\,\mathrm{sp}(b_a)(1+\log(1+n)) &\geq nx^2 - 6x\,\mathrm{sp}(b_a)\log(1+n) \\ &\geq nx^2 - 6x\,\mathrm{sp}(b_a)\sqrt{n} \\ &= n\left(x - \frac{3\,\mathrm{sp}(b_a)}{\sqrt{n}}\right)^2 - 9\,\mathrm{sp}(b_a)^2. \end{aligned}$$

Hence (°6) above. This concludes the proof. □

D.4.2 A bound on the number of episodes

At multiple point in the regret analysis, we use the following upper bound of the number of episodes. The number of episodes can be bounded by $\mathrm{O}(\log(T))$ thanks to the doubling trick that `UCB-NOM` uses.

Lemma D.7 (Number of episodes)**.** *The number of episodes of* `UCB-NOM` *up to horizon $T \geq 1$ is bounded by*

$$K(T) := |\{k \geq 1 : t_k \leq T-1\}| \leq |\mathcal{A}|\log(2T) \tag{D.32}$$

Proof. Given an arm $a \in \mathcal{A}$, let $\mathcal{K}_a(T) := \{k \geq 1 : t_k \leq T-1, A(t_k) = a\}$ be the set of episodes prior to time T where a is pulled. Denote $(t^a_\ell)_{\ell\geq 1}$ the enumeration of episodes initial times where arm a is pulled, i.e., $t^a_{\ell+1} := \inf\{t_k > t^a_\ell : A(t_k) = a\}$. Thanks to the Doubling Trick rule, we have $N_a(t^a_\ell) \geq 2^{\ell-1} - 1$. So,

$$|\mathcal{K}_a(T)| \leq 1 + \log_2(1 + N_a(T)).$$

Summing over arms, we obtain:

$$K(T) \leq |\mathcal{A}| + \sum_{a\in\mathcal{A}} \log_2(1+N_a(T)) =: |\mathcal{A}| + \psi(N(T)). \tag{D.33}$$

Note that the above function $\psi : \mathbf{R}_+^{\mathcal{A}} \to \mathbf{R}_+$ is the parallel sum of the concave increasing function $x \mapsto \log_2(1+x)$. The maximum is reached when $N_a(T) = N_{a'}(T)$ for all $a, a' \in \mathcal{A}$. We conclude using that $\sum_{a\in\mathcal{A}} N_a(T) = T$ and that $\log_2(1+x) \leq \log(2x)$ for $x \geq 1$. □

D.4.3 About the number of switches

In the previous paragraph, we have shown that the number of episodes is $\mathrm{O}(\log(T))$. Because the number of switches is bounded by the number of episodes, the number of switches is $\mathrm{O}(\log(T))$. However, this bound proves to be insufficient to improve the instance dependent regret bounds under $|\mathcal{A}^*| = 1$ in Theorem D.3. To that end, we will use the more precise result below.

Lemma D.8 (Number of switches under Doubling Trick)**.** *After n pulls of arm a, the number of switches of that arm is bounded by $\log_2(1+n)$, i.e., for all $T \geq 1$,*

$$\sum_{t=1}^{T-1} \mathbf{1}(A(t) = a, A(t+1) \neq a) = \sum_{i=1}^{N_a(T)} \mathbf{1}\big(\tau^a_i + 1 \neq \tau^a_{i+1}\big) \leq \log_2(1+N_a(T)). \tag{D.34}$$

Proof. By design of the algorithm, arms are played in contiguous segments of doubling durations. So, $\tau^a_{2^i+j} = \tau^a_{2^i} + j$ for all $i \geq 0$ and $j \in \{0, \ldots, 2^j - 1\}$. In particular, there are at most i switches of arm a prior to time $\tau^a_{2^i+j}$. We conclude using $N_a(\tau^a_k) = k-1$. □

Lemma D.9 (Almost-telescopic bias sum with non-consecutive pulls). *Let $(e_s)_{s\in\mathcal{S}}$ be the canonical basis of $\mathbf{R}^{\mathcal{S}}$. Recall that (τ_k^a) denotes the enumeration of pulls of arm a, that is, $\tau_1^a := \inf\{t \geq 1 : A(t) = a\}$ and $\tau_{k+1}^a := \inf\{t > \tau_k^a : A(t) = a\}$. For every arm $a \in \mathcal{A}$ and all $n \geq 1$ we have*

$$\left|\sum_{k=1}^{n}\Big(e_{S(\tau_k^a)} - e_{S(\tau_k^a+1)}\Big)b_a\right| \leq \mathrm{sp}(b_a)\left(1 + \sum_{k=1}^{n}\mathbf{1}\big(\tau_k^a + 1 \neq \tau_{k+1}^a\big)\right)$$

Proof. The sum of the LHS is *nearly* a telescopic sum. More precisely, this sum *is* telescopic inside any block of consecutive pulls. Only when there is a *gap* — when the algorithm leaves arm a and come back later — can the telescopic sum "break", and this break cost at most $\mathrm{sp}(b_a)$. Formally,

$$\begin{aligned}
&\left|\sum_{k=1}^{n}\Big(e_{S(\tau_k^a)} - e_{S(\tau_k^a+1)}\Big)b_a\right| \\
&= \left|b_a(S(\tau_1^a)) - b_a\big(S(\tau_{n+1}^a)\big) + \sum_{k=1}^{n}\Big(e_{S(\tau_{k+1}^a)} - e_{S(\tau_k^a+1)}\Big)b_a\right| \\
&= \left|b_a(S(\tau_1^a)) - b_a\big(S(\tau_{n+1}^a)\big) + \sum_{k=1}^{n}\mathbf{1}\big(S(\tau_k^a+1) \neq S(\tau_{k+1}^a)\big)\Big(e_{S(\tau_{k+1}^a)} - e_{S(\tau_k^a+1)}\Big)b_a\right| \\
&\leq \mathrm{sp}(b_a)\left(1 + \sum_{k=1}^{n}\mathbf{1}\big(S(\tau_k^a+1) \neq S(\tau_{k+1}^a)\big)\right) \\
&\leq \mathrm{sp}(b_a)\left(1 + \sum_{k=1}^{n}\mathbf{1}\big(\tau_k^a + 1 \neq \tau_{k+1}^a\big)\right).
\end{aligned}$$

□

D.5 Auxiliary probability results

In paragraph, we recall the version of Hoeffding inequality that we use in the paper.

Lemma D.10 (Hoeffding Lemma, Hoeffding (1963)). *Let (X_k) be a sequence of independent random variables of common mean μ such that $a \leq X_k \leq b$ a.s. Denoting $\hat{\mu}_n := \frac{1}{n}\sum_{k=1}^{n} X_k$ for empirical mean, for all $n \geq 1$, $x \geq 0$,*

$$\mathbf{P}(\hat{\mu}_n - \mu \geq x) \leq \exp\left(-\frac{2nx^2}{(b-a)^2}\right).$$