# Learning Algorithms for Markovian Bandits: Is Posterior Sampling more Scalable than Optimism?

**Anonymous authors**
**Paper under double-blind review**

## Abstract

In this paper, we study the scalability of model-based algorithms learning the optimal policy of a discounted Markovian bandit problem with $n$ arms. There are two categories of model-based reinforcement learning algorithms: Bayesian algorithms (like PSRL), and optimistic algorithms (like UCRL2 or UCBVI). While a naive application of these algorithms is not scalable because the state-space is exponential in $n$, we construct variants specially tailored to Markovian bandits (MB) that we call MB-PSRL, MB-UCRL2, and MB-UCBVI. They all have a low regret in $\tilde{O}(S\sqrt{nK})$ – where $K$ is the number of episodes, $n$ is the number of arms and $S$ is the number of states of each arm. Up to a factor $\sqrt{S}$, these regrets match the lower bound of $\Omega(\sqrt{SnK})$ that we also derive.

Even if their theoretical regrets are comparable, the *time complexity* of these algorithms varies greatly: We show that MB-UCRL2, as well as all algorithms that use bonuses on transition matrices have a time complexity that grows exponentially in $n$. In contrast, MB-UCBVI does not use bonuses on transition matrices and we show that it can be implemented efficiently, with a time complexity linear in $n$. However, our numerical experiments show that its empirical regret is large. Finally, our Bayesian algorithm, MB-PSRL, enjoys the best of both worlds: its running time is linear in the number of arms and its empirical regret is the smallest of all algorithms. This is a new confirmation of the power of Bayesian algorithms, that can often be easily tailored to the structure of the problems to learn.

## 1 Introduction

Markov decision processes (MDPs) are a powerful model to solve stochastic optimization problems. They suffer, however, from what is called the *curse of dimensionality*: the state size of a Markov process is exponential in its number of dimensions, so that the complexity of computing an optimal policy is exponential in the number of dimensions of the problem. The same holds for general purpose reinforcement learning algorithm: they all have a regret and a runtime exponential in the number of dimensions, so they also suffer from the same curse. Very few MDPs are known to escape from this curse of dimensionality. One of the most famous examples is the markovian bandit problem for which an optimal policy and its value can be computed in $O(n)$, where $n$ is the number of arms: The optimal policy can be computed by using the Gittins indices (computed locally) and its value can be computed by using retirement values (see for example Whittle (1996)).

In this paper, we study a specialization of PSRL (Osband et al., 2013) to markovian bandits, that we call markovian bandit posterior sampling (MB-PSRL) that consists in using PSRL with a prior tailored to markovian bandits. We show that the regret of MB-PSRL is sub-linear in the number of episodes and of arms. We also provide a regret guarantee for two optimistic algorithms that we call MB-UCRL2 and MB-UCBVI, and that are based respectively on UCRL2 (Auer et al., 2008) and UCBVI (Azar et al., 2017). They both use modified confidence bounds adapted to markovian bandit problems. The upper bound for their regret is similar to the bound for MB-PSRL. This shows that in terms of regret, the posterior sampling approach (MB-PSRL) and the optimistic approach (MB-UCRL2 and MB-UCBVI) scale well with the number of arms. We also provide a lower bound on the regret of any learning algorithm in markovian bandit problems, which shows that the regret bounds that we obtain for all algorithms are close to optimal.

The situation is radically different when considering the processing time: the runtime of MB-PSRL is linear in the number of arms, while the runtime of MB-UCRL2 is exponential in $n$. We show that this is not an artifact of our implementation of MB-UCRL2 by exhibiting a Markovian bandit problem for which being optimistic in each arm is not optimistic in the global MDP. This implies that UCRL2 and its variants (Bourel et al., 2020; Fruit et al., 2018; Talebi & Maillard, 2018; Filippi et al., 2010) cannot be adapted to have linear runtime in Markovian bandit problem unless an oracle gives the optimal policy. We argue that this non-scalability of UCRL2 and variants is not a limitation of all optimistic approach but comes from the fact that UCRL2 relies on extended value iteration (Auer et al., 2008) needed to deal with upper confidence bounds on the transition matrices. We show that MB-UCBVI, an optimistic algorithm that does not add bonus on transition probabilities and hence does not rely on extended value iteration, does not suffer from the same problem. Its regret is sub-linear in the number of episodes, and arms (although larger than the regret of both MB-PSRL and MB-UCRL2), and its runtime is linear in the number of arms.

We also conduct a series of numerical experiments to compare the performance of MB-PSRL, MB-UCRL2 and MB-UCBVI. They confirm the good behavior of MB-PSRL, both in terms of regret and computational complexity. These numerical experiments also show that the empirical regret of MB-UCBVI is larger than the regret of MB-PSRL and MB-UCRL2, confirming the comparisons between the upper bounds derived in Theorem 1. All this makes MB-PSRL the better choice between the three learning algorithms.

**Related work** In this paper, we focus on markovian bandit problem with discount factor $\beta < 1$ and all reward functions and transition matrices $(\boldsymbol{r}^a, Q^a)_{a\in\{1,\dots,n\}}$ are unknown. A possible approach to learn under these conditions is to ignore the problem structure and view the markovian bandit problem as a generic MDP. There are two main families of generic reinforcement learning algorithms with regret guarantees. The first one uses the *optimism in face of uncertainty* (OFU) principle. OFU methods build a confidence set for the unknown MDP and compute an optimal policy of the "best" MDP in the confidence set, *e.g.*, Bourel et al. (2020); Fruit et al. (2017); Azar et al. (2017); Auer et al. (2008); Bartlett & Tewari (2012). UCRL2 (Auer et al., 2008) is a well known OFU algorithm. The second family uses a bayesian approach, the posterior sampling method introduced by Thompson (1933). Such algorithms keep a posterior distribution over possible MDPs and execute the optimal policy of a sampled MDP, see *e.g.*, Ouyang et al. (2017); Agrawal & Jia (2017); Gopalan & Mannor (2015); Osband et al. (2013). PSRL (Osband et al., 2013) is a classical example of bayesian learning algorithm. All these algorithms, based on OFU or on bayesian principles, have sub-linear bounds on the regret, which means that they provably learn the optimal policy. Yet, applied as-is to markovian bandit problems, these bounds grow exponentially with the number of arms.

Our work is not the first attempt to exploit the structure of a MDP to improve learning. Factored MDPs (the state space can be factored into $n$ components) are investigated in Guestrin et al. (2003), where asymptotic convergence to the optimal policy is proved to scale polynomially in the number of components. The regret of learning algorithms in factored MDP with a factored action space is considered by Tian et al. (2020); Rosenberg & Mansour (2020); Xu & Tewari (2020); Osband & Van Roy (2014). Our work differs substantially from these. First, the markovian bandit problem is not a factored MDP because the action space is global and cannot be factored. Second, our reward is discounted over an infinite horizon while factored MDPs have been analyzed with no discount. Finally, and most importantly, the factored MDP framework assumes that the successive optimal policies are computed by an unspecified solver. There is no guarantee that the time complexity of this solver scales linearly with the number of components, especially for OFU-based algorithms. For markovian bandits, we get an additional leverage: when all parameters are known, the Gittins index policy is known to be an optimal policy and its computational complexity is linear in the number of arms. This reveals an interesting difference between bayesian and extended value based algorithms (the former being scalable and not the latter), which is not present in the literature about factored MDPs because such papers do not consider the time complexity.

Our markovian bandit setting is known in the literature as *rested* or *restful* bandit or a *family of alternative bandit processes*. Tekin & Liu (2012) consider a non-discounted setting, $\beta = 1$, and provide algorithms with logarithmic regret guarantee for *rested* as well as *restless* settings (a generalization of rested). However, they consider a notion of regret known as *weak regret* that measures how fast the learning algorithm identifies the best arm in stationary regime. So, it ignores the learning behavior at the beginning learning process.

In contrast, we consider the discounted rested bandit setting in which the regret of Tekin & Liu (2012) makes no more senses due to the discount factor and we propose a regret definition that is frequently used in reinforcement learning literature and captures the performance of a learning algorithm during the whole learning process. In addition, Ortner et al. (2012); Jung & Tewari (2019); Wang et al. (2020) consider a non-discounted restless bandit setting in which only the state of chosen arms are observed by the learner. Ortner et al. (2012); Wang et al. (2020) propose optimistic algorithms for infinite-horizon setting and provide regret bounds that are sub-linear in time. Again the discounted case is not considered in these papers while it is particularly interesting because learning algorithms can leverage the optimal Gittins index policy. Jung & Tewari (2019) propose a bayesian algorithm in the episodic finite-horizon setting and also provide a regret bound that is sub-linear in the number of episodes. However, the computational complexity is not studied in their work (the algorithm of Ortner et al. (2012) is intractable while the ones of Jung & Tewari (2019); Wang et al. (2020) rely on the unspecified problem solver called *oracle*). Contrarily, we provide both performance guarantee and computational complexity analysis of each algorithm that we consider in this paper. Finally, Killian et al. (2021) consider a more general setting of restless bandits in which each arm is itself a MDP and the learner has to decide which arms to choose and which action to execute on each chosen arm under a global action constraint. The authors propose a Lagrangian suboptimal policy to solve the restless bandit problem with known parameters and a sampling algorithm to learn their Lagrangian policy when the parameters are unknown. Unfortunately, no performance guarantee is provided in their work.

Since index policies scale with the number of arms, using Q-learning approaches to learn such a policy is also popular, see *e.g.*, Avrachenkov & Borkar (2022); Fu et al. (2019); Duff (1995). Duff (1995) addresses the same markovian bandit problem as we do: their algorithm learns the optimal value in the restart-in-state MDP (Katehakis & Veinott Jr, 1987) for each arm and uses Softmax exploration to solve the exploration-exploitation dilemma. As mentioned on page 250 of Auer et al. (2002), however, there exists no finite-time regret bounds for this algorithm. Furthermore, tuning its hyperparameters (learning rate and temperature) is rather delicate and unstable in practice.

## 2 Markovian bandit problem

In this section, we introduce the markovian bandit problem and recall the notion of Gittins index when the parameters $(\boldsymbol{r}^a, Q^a)$ of all arms are known.

### 2.1 Definitions and main notations

We consider a markovian bandit problem with $n$ arms. Each arm $\langle \mathcal{S}^a, \boldsymbol{r}^a, Q^a \rangle$ for $a \in \{1, \dots, n\} =: [n]$ is a Markov reward process with a finite state space $\mathcal{S}^a$ of size $S$. Each arm has a mean reward vector, $\boldsymbol{r}^a \in [0,1]^S$, and a transition matrix $Q^a$. When Arm $a$ is activated in state $x_a \in \mathcal{S}^a$, it moves to state $y_a \in \mathcal{S}^a$ with probability $Q^a(x_a, y_a)$. This provides a reward whose expected value is $r^a(x_a)$. Without loss of generality, we assume that the state spaces of the arms are pairwise distinct: $\mathcal{S}^a \cap \mathcal{S}^b = \emptyset$ for $a \neq b$. In the following, the state of an arm $a$ will always be denoted with an index $a$: we will denote such a state by $x_a$ or $y_a$. As state spaces are disjoint, this allows us to simplify the notation by dropping the index $a$ from the reward and transition matrix: when convenient, we will denote them by $r(x_a)$ instead of $r^a(x_a)$ and by $Q(x_a, y_a)$ instead of $Q^a(x_a, y_a)$ since no confusion is possible.

At time 1, the global state $\boldsymbol{X}_1$ is distributed according to some initial distribution $\rho$ over the global state space $\mathcal{X} = \mathcal{S}^1 \times \ldots \times \mathcal{S}^n$. At time $t$, the decision maker observes the states[1] of all arms, $\boldsymbol{X}_t = (X_{t,1} \dots X_{t,n})$, and chooses which arm $A_t$ to activate. This problem can be cast as a MDP – that we denote by $M$ – with state space $\mathcal{E}$ and action space $[n]$. Let $a \in [n]$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{E}$. If the state at time $t$ is $\boldsymbol{X}_t = \boldsymbol{x}$, the chosen arm is $A_t = a$, then the agent receives a random reward $R_t$ drawn from some distribution on $[0,1]$ with mean $r(x_a)$ and the MDP $M$ transitions to state $\boldsymbol{X}_{t+1} = \boldsymbol{y}$ with probability $P^a(\boldsymbol{x}, \boldsymbol{y})$ that satisfies:

$$P^a(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} Q(x_a, y_a) & \text{if } x_b = y_b \text{ for all } b \neq a; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

[1]Throughout the paper, we use capital letters (like $X_t$) to denote random variables and small letter (like $\boldsymbol{x}$) to denote their realizations. Bold letters ($\boldsymbol{X}_t$ or $\boldsymbol{x}$) design vectors. Normal letters ($X_{t,a}$ or $x_a$) are for scalar values.

That is, the active arm makes a transition while the other arms remain in the same state.

Let $\Pi$ be the set of deterministic policies, *i.e.,* the set of functions $\pi : \mathcal{X} \mapsto [n]$. For the MDP $M$, we denote by $V_M^\pi(\boldsymbol{x})$ the expected cumulative discounted reward of $M$ under policy $\pi$ starting from an initial state $\boldsymbol{x}$:

$$V_M^\pi(\boldsymbol{x}){=}\mathbb{E}\left[\sum_{t=0}^{\infty}\beta^t R_t \mid \boldsymbol{X}_0{=}\boldsymbol{x}, A_t{=}\pi(\boldsymbol{X}_t)\right].$$

An alternative definition of $V$ is to consider a finite-horizon problem with a geometrically distributed length. Indeed, let $H$ be a time-horizon geometrically distributed with parameter $1-\beta > 0$. We have

$$V_M^\pi(\boldsymbol{x}){=}\mathbb{E}\left[\sum_{t=1}^{H} R_t \mid \boldsymbol{X}_1{=}\boldsymbol{x}, A_t{=}\pi(\boldsymbol{X}_t)\right]. \tag{2}$$

**Problem 1.** *Given a markovian bandit $M$ with $n$ arms, each is a Markov reward process $\langle \mathcal{S}^a, \boldsymbol{r}^a, Q^a\rangle$ with a finite state space of size $S$, find a policy $\pi\colon \mathcal{S}^1{\times}\ldots{\times}\mathcal{S}^n \mapsto [n]$ that maximizes $V_M^\pi(\boldsymbol{x})$ for any state $\boldsymbol{x}$ distributed according to initial global state distribution $\rho$.*

A policy $\pi_*$ is optimal for Problem 1 if $V_M^{\pi_*}(\boldsymbol{x}) \geq V_M^\pi(\boldsymbol{x})$ for all $\pi \in \Pi$ and $\boldsymbol{x} \in \mathcal{E}$. By Puterman (2014), such a policy exists and does not depend on $\boldsymbol{x}$ (or $\rho$). It is given by Gittins index policy, defined below.

### 2.2 Gittins index policy

It is possible to compute an optimal policy $\pi_*$ for Problem 1 in a reasonable amount of time using the so called Gittins indices: Gittins (1979) defines the *Gittins index* for any arm $a$ in state $x_a \in \mathcal{S}_a$ as

$$\mathrm{GIndex}(x_a) = \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t r^a(Z_t) \mid Z_0 = x_a\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1}\beta^t \mid Z_0 = x_a\right]}, \tag{3}$$

where $Z$ is a Markov chain whose transitions are given by $Q^a$ and $\tau$ can be any stopping time adapted to the natural filtration of $(Z_t)_{t\geq 0}$. So, Gittins index can be considered as the maximal reward density over time of an arm at the given state.

Gittins (1979) shows that activating the arm having the largest current index is an optimal policy. Such a policy can be computed very efficiently: The computation of the indices of an arm with $S$ states can be done in $O(S^3)$ arithmetic operations, which means that the computation of the Gittins index policy is linear in the number of arms as it takes $O(nS^3)$ arithmetic operations. For more details about Gittins indices and optimality, we refer to Gittins et al. (2011); Weber (1992). For a survey on how to compute Gittins indices, we refer to Chakravorty & Mahajan (2014), and to Gast et al. (2022) for a recent paper that shows how to compute Gittins index in subcubic time (*i.e.*, $o(S^3)$) for each of the $n$ arms).

## 3 Online learning and episodic regret

We now consider an extension of Problem 1 in which the decision maker does not know the transition matrices nor the rewards. Our goal is to design a reinforcement learning algorithm that learns the optimal policy from past observations. Similarly to what is done for finite-horizon reinforcement learning with deterministic horizon – see *e.g.*, Jin et al. (2018); Azar et al. (2017); Osband et al. (2013) – we consider a decision maker that faces a sequence of independent replicas of the same markovian bandit problem, where the transitions and the rewards are drawn independently for each episode. What is new here is that the time horizon $H$ is random and has a geometric distribution. It is drawn independently for each episode. This implies that Gittins index policy is optimal for a decision maker that would know the transition matrices and rewards.

In this paper, we consider *episodic learning algorithms.* Let $H_1, \ldots, H_k$ be the sequence of random episode lengths and let $t_k := 1{+}\sum_{i=1}^{k-1} H_i$ be the starting time of the $k$th episode. Let $\mathcal{O}_{k-1} := (\boldsymbol{X}_1, A_1, R_1, \ldots, \boldsymbol{X}_{t_k-1}, A_{t_k-1}, R_{t_k-1})$ denote the observations made prior and up to episode $k$. An *Episodic*

*Learning Algorithm* $\mathcal{L}$ is a function that maps observations $\mathcal{O}_{k-1}$ to $\mathcal{L}(\mathcal{O}_{k-1})$, a probability distribution over all policies. At the beginning of episode $k$, the algorithm samples $\pi_k \sim \mathcal{L}(\mathcal{O}_{k-1})$ and uses this policy during the whole $k$th episode. Note that one could also design algorithms where learning takes place inside each episode. We will see later that episodic learning as described here is enough to design algorithms that are essentially optimal, in the sense given by Theorem 1 and Theorem 2.

For an instance $M$ of a markovian bandit problem and a total number of episodes $K$, we denote by $\mathrm{Reg}(K, \mathcal{L}, M)$ the regret of a learning algorithm $\mathcal{L}$, defined as

$$\mathrm{Reg}(K, \mathcal{L}, M) := \sum_{k=1}^{K} V_M^{\pi_*}(\boldsymbol{X}_{t_k}) - V_M^{\pi_k}(\boldsymbol{X}_{t_k}). \tag{4}$$

It is the sum over all episodes of the value of the optimal policy $\pi_*$ minus the value obtained by applying the policy $\pi_k$ chosen by the algorithm for episode $k$. In what follows, we will provide bounds on the expected regret.

A no-regret algorithm is an algorithm $\mathcal{L}$ such that its expected regret $\mathbb{E}\left[\mathrm{Reg}(K, \mathcal{L}, M)\right]$ grows sub-linearly in the number of episodes $K$. This implies that the expected regret over episode $k$ converges to 0 as $k$ goes to infinity. Such an algorithm learns an optimal policy of Problem 1.

Note that, for discounted MDPs, an alternative regret definition (used for instance by He et al. (2021)) is to use the non-episodic version $\sum_{t=1}^{T}(V_M^{\pi_*}(\boldsymbol{X}_t) - V_M^{\pi_t}(\boldsymbol{X}_t))$. In our definition at Equation 4, we use an episodic approach where the process is restarted according to $\rho$ after each episode of geometrically distributed length $H_k$.

## 4 Learning algorithms for markovian bandits

In what follows, we present three algorithms having a regret that grows like $\tilde{O}(S\sqrt{nK})$, that we call MB-PSRL, MB-UCRL2 and MB-UCBVI. As their names suggest, these algorithms are adaptation of PSRL, UCRL2 and UCBVI to markovian bandit problems that intend to overcome the exponentiality in $n$ of their regret. The structure of the three MB-* algorithm is similar and is represented in Algorithm 1. All algorithms are episodic learning algorithms. At the beginning of each episode, a MB-* learning algorithm computes a new policy $\pi_k$ that will be used during an episode of geometrically distributed length. The difference between the three algorithms lies in the way this new policy $\pi_k$ is computed. MB-PSRL uses posterior sampling while MB-UCRL2 and MB-UCBVI use optimism. We detail the three algorithms below.

**Algorithm 1** Pseudo-code of the three MB-* algorithms.

**input** Discount factor $\beta$, initial distribution $\rho$ (and a prior distribution $(\phi^a)_{a\in[n]}$ for MB-PSRL)

```
for episodes k = 1, 2, ... do
    Compute a new policy π_k (using posterior sampling or optimism).
    Set t_k ← 1 + Σ_{i=1}^{k-1} H_i, sample X_{t_k} ~ ρ and H_k ~ Geom(1 − β).
    for t ← t_k to t_k + H_k − 1 do
        Activate arm A_t = π_k(X_t).
        Observe R_t and X_{t+1}.
    end for
end for
```

### 4.1 MB-PSRL

MB-PSRL starts with a prior distribution $\phi^a$ over the parameters $(\boldsymbol{r}^a, Q^a)$. At the start of each episode $k$, MB-PSRL computes a posterior distribution of parameters $\phi^a(\cdot \mid \mathcal{O}_{k-1})$ for each arm $a \in [n]$ and samples parameters $(\boldsymbol{r}_k^a, Q_k^a)$ from $\phi^a(\cdot \mid \mathcal{O}_{k-1})$ for each arm. Then, MB-PSRL uses $(\boldsymbol{r}_k^a, Q_k^a)_{a\in[n]}$ to compute the Gittins index policy $\pi_k$ that is optimal for the sampled problem. The policy $\pi_k$ is then used for the whole episode $k$. Note that as $\pi_k$ is a Gittins index policy, it can be computed efficiently.

The difference between PSRL and MB-PSRL is mostly that MB-PSRL uses a prior distribution tailored to markovian bandit. The only hyperparameter of MB-PSRL is the prior distribution $\phi$. As we see in Appendix E, MB-PSRL seems robust to the choice of the prior distribution, even if a coherent prior gives a better performance than a misspecified prior, similarly to what happens for Thompson's sampling (Russo et al., 2018).

### 4.2 MB-UCRL2

At the beginning of each episode $k$, MB-UCRL2 computes the following quantities for each state $x_a \in \mathcal{S}^a$: $N_{k-1}(x_a)$ the number of times that Arm $a$ is activated before episode $k$ while being in state $x_a$, and $\hat{r}_{k-1}(x_a)$, and $\hat{Q}_{k-1}(x_a, \cdot)$ are the empirical means of $r(x_a)$ and $Q(x_a, \cdot)$. We define the confidence bonuses $b^r_{k-1}(x_a) = \sqrt{\frac{\log(2SnKt_k)}{2\max\{1, N_{k-1}(x_a)\}}}$ and $b^Q_{k-1}(x_a) = \sqrt{\frac{2\log(SnK2^S t_k)}{\max\{1, N_{k-1}(x_a)\}}}$. This defines a confidence set $\mathbb{M}_k$ as follows: a markovian bandit problem $M'$ is in $\mathbb{M}_k$ if for all $a \in [n]$ and $x_a \in \mathcal{S}^a$:

$$|r'(x_a) - \hat{r}_{k-1}(x_a)| \leq b^r_{k-1}(x_a) \text{ and } \|Q'(x_a, \cdot) - \hat{Q}_{k-1}(x_a, \cdot)\|_1 \leq b^Q_{k-1}(x_a). \tag{5}$$

MB-UCRL2 then chooses a policy $\pi_k$ that is optimal for the most optimistic problem $M_k \in \mathbb{M}_k$:

$$\pi_k \in \arg\max_{\pi} \max_{M' \in \mathbb{M}_k} V^{\pi}_{M'}(\rho). \tag{6}$$

Note that as we explain later in Section 6.1, we believe that there is no efficient algorithm to compute the best optimistic policy $\pi_k$ of Equation 6.

Compared to a vanilla implementation of UCRL2, MB-UCRL2 uses the structure of the markovian bandit problem: The constraints Equation 5 are on $Q$ whereas vanilla UCRL2 uses constraints on the full matrix $P$ (defined in Equation 1). This leads MB-UCRL2 to use the bonus term that scales as $\sqrt{S/N_{k-1}(x_a)}$ whereas vanilla UCRL2 would use the term in $\sqrt{S^n/N_{k-1}(\boldsymbol{x}, a)}$.

### 4.3 MB-UCBVI

At the beginning of episode $k$, MB-UCBVI uses the same quantities $N_{k-1}(x_a)$, $\hat{r}_{k-1}(x_a)$, and $\hat{Q}_{k-1}(x_a, \cdot)$ as MB-UCRL2. The difference lies in the definition of the bonus terms. While MB-UCRL2 uses a bonus on the reward and on the transition matrices, MB-UCBVI defines a bonus $b_{k-1}(x_a) := \frac{1}{1-\beta}\sqrt{\frac{\log(2SnKt_k)}{2\max\{1, N_{k-1}(x_a)\}}}$ that is used on the reward only. MB-UCBVI computes the Gittins index policy $\pi_k$ that is optimal for the bandit problem $(\hat{\boldsymbol{r}}^a_{k-1} + b^a_{k-1}, \hat{Q}^a_{k-1})_{a \in [n]}$.

Similarly to the case of UCRL2, a vanilla implementation of UCBVI would use a bonus that scales exponentially with the number of arms. MB-UCBVI makes an even better use of the structure of the learned problem because the optimistic MDP $(\hat{\boldsymbol{r}}^a_{k-1} + b^a_{k-1}, \hat{Q}^a_{k-1})_{a \in [n]}$ is still a markovian bandit problem. This implies that the optimistic policy $\pi_k$ is a Gittins index policy, and that can therefore be computed efficiently.

## 5 Regret analysis

In this section, we first present upper bounds on the expected regret of the three learning algorithms. These bounds are sub-linear in the number of episodes (hence the three algorithms are no-regret algorithms) and sub-linear in the number of arms. We then derive a minimax lower bound on the regret of any learning algorithm in the markovian bandit problem.

### 5.1 Upper bounds on regret

The theorem below provides upper bounds on the expected regret of the three algorithms presented in Section 4. Note that since MB-PSRL is a bayesian algorithm, we consider its *bayesian regret*, that is the expectation over all possible model. More precisely, if the unknown MDP $M$ is drawn from a prior distribution $\phi$, the *bayesian regret* of a learning algorithm $\mathcal{L}$ is $\text{BayReg}(K, \mathcal{L}, \phi) = \mathbb{E}[\text{Reg}(K, \mathcal{L}, M)]$, where the expectation is taken over all possible values of $M \sim \phi$ and all possible runs of the algorithm. The expected regret $\mathbb{E}[\text{Reg}(K, \mathcal{L}, M)]$ is defined by taking the expectation over all possible runs of the algorithm.

**Theorem 1.** *Let* $f(S,n,K,\beta) = Sn\left(\log K/(1-\beta)\right)^2 + \sqrt{SnK}\left(\log K/(1-\beta)\right)^{3/2}$. *There exists universal constants* $C, C'$ *and* $C''$ *independent of the model (i.e., that do not depend on* $S$*,* $n$*,* $K$ *and* $\beta$*) such that:*

- *For any prior distribution* $\phi$*:*

$$\text{BayReg}(K, \textit{MB-PSRL}, \phi) \leq C\left(\sqrt{S} + \log\frac{SnK\log K}{1-\beta}\right) f(S,n,K,\beta),$$

- *For any markovian bandit model* $M$*:*

$$\mathbb{E}\left[\text{Reg}(K, \textit{MB-UCRL2}, M)\right] \leq C'\left(\sqrt{S} + \log\frac{SnK\log K}{1-\beta}\right) f(S,n,K,\beta),$$
$$\mathbb{E}\left[\text{Reg}(K, \textit{MB-UCBVI}, M)\right] \leq C''\left(\frac{\sqrt{S}}{1-\beta}\right)\left(\log\frac{SnK\log K}{1-\beta}\right) f(S,n,K,\beta),$$

We provide a sketch of proof below. The detailed proof is provided in Appendix A in the supplementary material.

This theorem calls for several comments. First, it shows that when $K \geq Sn/(1-\beta)$, the regret of MB-PSRL and MB-UCRL2 is smaller than

$$\tilde{O}\left(\frac{S\sqrt{nK}}{(1-\beta)^{3/2}}\right), \tag{7}$$

where the notation $\tilde{O}$ means that all logarithmic terms are removed. The regret of MB-UCBVI has an extra $1/(1-\beta)$ factor.

Hence, the regret of the three algorithms is sub-linear in the number of episodes $K$ which means that they all are no-regret algorithms. This regret bound is sub-linear in the number of arms which is very significant in practice when facing a large number of arms. Note that directly applying PSRL, UCRL2 or UCBVI would lead to a regret in $\tilde{O}\left(S^n\sqrt{nK}\right)$ or $\tilde{O}\left(\sqrt{nS^nK}\right)$, which is exponential in $n$.

Second, the upper bound on the expected regret of MB-UCRL2 (and of MB-UCBVI) is a guarantee for a specific problem $M$ while the bound on bayesian regret of MB-PSRL is a guarantee in average overall the problems drawn from the prior $\phi$. Hence, the bounds of MB-UCRL2 and MB-UCBVI are stronger guarantee compared to the one of MB-PSRL. Yet, as we will see later in the numerical experiments reported in Section 7, MB-PSRL seems to have a smaller regret in practice, even when the problem does not follow the correct prior.

Finally, our bound Equation 7 is linear in $S$, the state size of each arm. Having a regret bound linear in the state space size is currently state-of-the-art for bayesian algorithms, see *e.g.*, Agrawal & Jia (2017); Ouyang et al. (2017). For optimistic algorithms, the best regret bounds are linear in the square root of the state size because they use Bernstein's concentration bounds instead of Weissman's inequality (Azar et al., 2017), yet this approach does not work in the discounted case because of the random length of episodes. UCBVI has also been studied in the discounted case by He et al. (2021). However they use with a different definition of regret, making their bound on the regret hard to compare with ours.

### Sketch of proof

A crucial ingredient of our proof is to work with the value function over a random finite time horizon ($W$ defined below), instead of working directly with the discounted value function $V$. For a given model $M$, and a stationary policy $\pi$, a horizon $H$ and a time step $h \leq H$, we define by $W^{\pi}_{M,h:H}(\boldsymbol{x})$ the value function of a policy $\pi$ over the finite time horizon $H-h+1$ when starting in $\boldsymbol{x}$ at time $h$. It is defined as

$$W^{\pi}_{M,h:H}(\boldsymbol{x}) = r^{\pi}(\boldsymbol{x}) + \sum_{\boldsymbol{y}\in\mathcal{E}} P^{\pi}(\boldsymbol{x},\boldsymbol{y}) W^{\pi}_{M,h+1:H}(\boldsymbol{y}), \tag{8}$$

with $W^{\pi}_{M,H:H}(\boldsymbol{x}) = r^{\pi}(\boldsymbol{x})$ and where $r^{\pi}$ and $P^{\pi}$ are reward vector and state transition matrix when following policy $\pi$.

By definitions of $W$ in Equation 8 and $V$ in Equation 2, for a fixed model $M$, a policy $\pi$ and a state $\boldsymbol{x}$, and a time horizon $H$ that is geometrically distributed, one has $V^{\pi}_{M}(\boldsymbol{x}) = \mathbb{E}\left[W^{\pi}_{M,1:H}(\boldsymbol{x})\right]$.

This characterization is important in our proof. Since the episode length $H_k$ is independent of the observations available before episode $k$, $\mathcal{O}_{k-1}$, for any policy $\pi_k$ that is independent of $H_k$, one has

$$\mathbb{E}\left[V^{\pi_k}_{M}(\boldsymbol{X}_{t_k}) \mid \mathcal{O}_{k-1}, \pi_k\right] = \mathbb{E}\left[W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k}) \mid \mathcal{O}_{k-1}, \pi_k\right]. \tag{9}$$

In the above Equation 9, the expectation is taken over all initial state $\boldsymbol{X}_{t_k}$ and all possible horizon $H_k$.

Equation 9 will be very useful in our analysis as it allows us to work with either $V$ or $W$ interchangeably. While the proof of MB-PSRL could be done by only studying the function $W$, the proof of MB-UCRL2 and MB-UCBVI will use the expression of the regret as a function of $V$ to deal with the non-determinism. Indeed, at episode $k$, all algorithms compare the optimal policy $\pi_*$ (that is optimal for the true MDP $M$) and a policy $\pi_k$ chosen by the algorithm (that is optimal for a MDP $M_k$ that is either sampled by MB-PSRL or chosen by an optimistic principle). The quantity $\Delta_k := W^{\pi_*}_{M,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k})$ equals:

$$\underbrace{W^{\pi_*}_{M,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{X}_{t_k})}_{(A)} + \underbrace{W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k})}_{(B)}. \tag{10}$$

The analysis of the term (B) is similar for the three algorithms: it is bounded by the distance between the sampled MDP $M_k$ and the true MDP $M$ that can in turn be bounded by using a concentration argument (Lemma 1) based on Hoeffding's and Weissman's inequalities. Compared with the literature (Azar et al., 2017; Ouyang et al., 2017), our proof leverages on taking conditional expectations, making all terms whose conditional expectation is zero disappear. One of the main technical hurdle is to deal with the $K$ random episodes $H_1, \ldots, H_k$. This is also new in our approach compared to the classical analysis of finite horizons regrets.

The analysis of (A) depends heavily on the algorithm used. The easiest case is PSRL: As our setting is bayesian, the expectation of the first term (A) with respect to the model is zero (see Lemma 5). The case of MB-UCRL2 and MB-UCBVI are harder. In fact, our bonus terms are specially designed so that $V^{\pi_k}_{M_k}(\boldsymbol{x})$ is an optimistic upper bound of the true value function with high probability, that is:

$$V^{\pi_k}_{M_k}(\boldsymbol{x}) = \max_{\pi} \max_{M' \in \mathbb{M}_k} V^{\pi}_{M'}(\boldsymbol{x}) \geq V^{\pi_*}_{M}(\boldsymbol{x}). \tag{11}$$

This requires the use of $V$ and not $W$ and it is used to show that the expectation of the term (A) of Equation 10 cannot be positive.

### 5.2 Minimax lower bound

After obtaining upper bounds on the regret, a natural question is: can we do better? Or in other terms, does there exist a learning algorithm with a smaller regret? To answer this question, the metric used in the literature is the notion of minimax lower bound: for a given set of parameters $(S, n, K, \beta)$, a minimax lower bound is a lower bound on the quantity $\inf_{\mathcal{L}} \sup_{M} \mathrm{Reg}(K, \mathcal{L}, M)$, where the supremum is taken among all possible models that have parameters $(S, n, K, \beta)$ and the infimum is taken over all possible learning algorithms. The next theorem provides a lower bound on the bayesian regret. It is therefore stronger than a minimax bound for two reasons: First, the bayesian regret is an average over models, which means that there exists at least one model that has a larger regret than the bayesian lower bound; And second, in Theorem 2, we allow the algorithm to depend on the prior distribution $\phi$ and to use this information.

**Theorem 2** (Lower bound)**.** *For any state size $S$, number of arms $n$, discount factor $\beta$ and number of episodes $K \geq 16S$, there exists a prior distribution $\phi$ on markovian bandit problems with parameters $(S, n, K, \beta)$ such that, for any learning algorithm $\mathcal{L}$:*

$$\mathrm{BayReg}(K, \mathcal{L}, \phi) \geq \frac{1}{60}\sqrt{\frac{SnK}{(1-\beta)}}. \tag{12}$$

The proof is given in Appendix B and uses a counterexample inspired by the one of Auer et al. (2008). Note that for general MDPs, the minimax lower bound obtained by Osband & Van Roy (2016); Auer et al. (2008) says that a learning algorithm cannot have a regret smaller than $\Omega(\sqrt{\tilde{S}\tilde{A}\tilde{T}})$, where $\tilde{S}$ is the number of states of the MDP, $\tilde{A}$ is the number of actions and $\tilde{T}$ is the number of time steps. Yet, the lower bound of Osband & Van Roy (2016); Auer et al. (2008) is not directly applicable to our case with $\tilde{S} = S^n$ because markovian bandit problems are very specific instances of MDPs and this can be exploited by the learning algorithm. Also note that this lower bound on the bayesian regret is also a lower bound on the expected regret of any non-bayesian algorithm for any MDP model $M$.

Apart from the logarithmic terms, the lower bound provided by Theorem 2 differs from the bound of Theorem 1 by a factor $\sqrt{S}/(1-\beta)$. This factor is similar to the one observed for PSRL and UCRL2 (Osband et al., 2013; Auer et al., 2008). There are various factors that could explain this. We believe that the extra factor $1/(1-\beta)$ might be half due to the episodic nature of MB-PSRL and MB-UCRL2 (when $1/(1-\beta)$ is large, algorithms with internal episodic updates might have smaller regret) and half due to the fact that the lower bound of Theorem 2 is not optimal and could include a term $1/\sqrt{1-\beta}$ (similar to the term $O(\sqrt{D})$ of the lower bound of Osband & Van Roy (2016); Auer et al. (2008)). The factor $\sqrt{S}$ between our two bounds comes from our use of Weissman's inequality. It might be possible that our regret bounds are not optimal with respect to this term although such an improvement cannot be obtained using the same approach of Azar et al. (2017).

## 6 Scalability of learning algorithms for markovian bandits

Historically, Problem 1 was considered unresolved until Gittins (1979) proposed Gittins indices. This is because previous solutions were based on Dynamic Programming in the global MDP which are computationally expensive. Hence, after establishing regret guarantees, we are now interested in the computational complexity of our learning algorithms, which is often disregarded in the learning literature.

### 6.1 MB-PSRL and MB-UCBVI are scalable

If one excludes the simulation of the MDP, the computational cost of MB-PSRL and MB-UCBVI of each episode is low. For MB-PSRL, its cost is essentially due to three components: Updating the observations, sampling from the posterior distribution and computing the optimal policy. The first two are relatively fast when the conjugate posterior has a closed form: updating the observation takes $O(1)$ at each time, and sampling from the posterior can be done in $O(nS^2)$ – more details on posterior distributions are given in Appendix D. When the conjugate posterior is implicit (*i.e.*, under the integral form), the computation can be higher but remains linear in the number of arms. For MB-UCBVI, the cost is due to two components: computing the bonus terms and computing the Gittins policy for the optimistic MDP. Computing the bonus is linear in the number of bandits and the length of the episode. As explained in Section 2.2, the computation of the Gittins index policy for a given problem can be done in $O(nS^3)$. Hence, MB-PSRL and MB-UCBVI successfully escape from the curse of dimensionality.

### 6.2 MB-UCRL2 is not scalable because it cannot use an Index Policy

While MB-UCRL2 has a regret equivalent to the one of MB-PSRL, its computational complexity, and in particular the complexity of computing an *optimistic* policy that maximizes Equation 6 does not scale with $n$. Such a policy can be computed by using *extended value iteration* (Auer et al., 2008). This computation is polynomial in the number of states of the global MDP and is therefore exponential in the number of arms, precisely $O(nS^{2n})$. For MB-PSRL (or MB-UCBVI), the computation is easier because the sampled (optimistic) MDP is a markovian bandit problem. Hence, using Gittins Theorem, computing the optimal policy can be done by computing local indices. In the following, we show that it is not possible to solve Equation 6 by using local indices. This suggests that MB-UCRL2 (nor any of the modifications of UCRL2's variants that would use extended value iteration) cannot be implemented efficiently.

More precisely, to find an optimistic policy (that satisfies Equation 11), UCRL2 and its variants, *e.g.*, KL-UCRL (Filippi et al., 2010), compute a policy $\pi_k$ that is optimal for the most optimistic MDP in $\mathbb{M}_k$. This

can be done by using extended value iteration. We now show that this cannot be replaced by the computation of local indices.

Let us consider that the estimates and confidence bounds for a given arm $a$ are $\hat{\mathcal{B}}^a = (\hat{\boldsymbol{r}}^a, \hat{Q}^a, b_a^r, b_a^Q)$. We say that an algorithm computes indices locally for Arm $a$ if for each $x_a \in \mathcal{S}^a$, it computes an index $I^{\hat{\mathcal{B}}^a}(x_a)$ by using only $\hat{\mathcal{B}}^a$ but not $\hat{\mathcal{B}}^b$ for any $b \neq a$. We denote by $\pi^{I(\hat{\mathcal{B}})}$ the index policy that uses index $I^{\hat{\mathcal{B}}^a}$ for arm $a$ and by $\mathbb{M}(\hat{\mathcal{B}})$ the set of markovian bandit problems $M'$ that satisfy Equation 5.

**Theorem 3.** *For any algorithm that computes indices locally, there exists a markovian bandit problem $M$, an initial state $\boldsymbol{x}$ and estimates $\hat{\mathcal{B}}^a = (\hat{\boldsymbol{r}}^a, \hat{Q}^a, b_a^r, b_a^Q)$ such that $M \in \mathbb{M}(\hat{\mathcal{B}})$ and*

$$\sup_{M' \in \mathbb{M}(\hat{\mathcal{B}})} V_{M'}^{\pi^{I(\hat{\mathcal{B}})}}(\boldsymbol{x}) < \sup_{\pi} V_M^{\pi}(\boldsymbol{x}).$$

*Proof.* The proof presented in Appendix C is obtained by constructing a set $\mathbb{M}$ and two MDPs $M_1$ and $M_2$ in $\mathbb{M}$ such that Equation 11 cannot hold simultaneously for both $M_1$ and $M_2$. □

This theorem implies that one cannot define local indices such that Equation 11 holds for all bandit problems $M \in \mathbb{M}_k$. Yet, the use of this inequality is central in the regret analysis of UCRL2 (see the proof of UCRL2 (Auer et al., 2008)). This implies that the current methodology to obtain regret bounds for UCRL2 and its variants, *e.g.*, Bourel et al. (2020); Fruit et al. (2018); Talebi & Maillard (2018); Filippi et al. (2010), that use Extended Value Iteration is not applicable to bound the regret of their modified version that computes indices locally.

Note that for any set $\mathbb{M}$ such that $M \in \mathbb{M}$, there still exists an index policy $\pi^{\text{ind}}$ that is optimistic because all MDPs in $\mathbb{M}$ are markovian bandit problems. This optimistic index policy satisfies

$$\sup_{M' \in \mathbb{M}} V_{M'}^{\pi^{\text{ind}}} \geq \sup_{\pi} V_M^{\pi}.$$

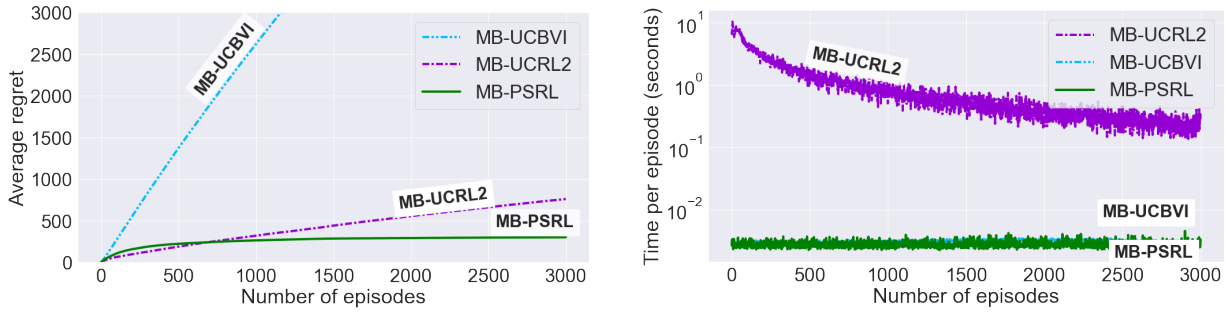
This means that restricting to index policies is not a restriction for optimism. What Theorem 3 shows is that an optimistic index policy can be defined only after the most optimistic MDP $M \in \mathbb{M}$ is computed and computing optimistic policy and $M$ simultaneously depends on the confidence sets of all arms.

Therefore, we believe that UCRL2 and its variants cannot compute optimistic policy locally: they should all require the joint knowledge of all $(\hat{\mathcal{B}}^a)_{a \in [n]}$.

## 7 Numerical experiments

In complement to our theoretical analysis, we report, in this section, the performance of our three algorithms in a model taken from the literature. The model is an environment with 3 arms, all following a Markov chain that is obtained by applying the optimal policy on the river swim MDP. A detailed description is given in Appendix D, along with all hyperparameters that we used. Our numerical experiments suggest that MB-PSRL outperforms other algorithms in term of average regret and is computationally less expensive than other algorithms. To ensure reproducibility, the code and data of our experiments are available (link to GitHub repository hidden for double blind review).

**Performance result** We investigate the average regret and policy computation time of each algorithm. To do so, we run each algorithm for 80 simulations and for $K = 3000$ episodes per simulation. We arbitrarily choose the discount factor $\beta = 0.99$. In Figure 1(a), we show the average cumulative regret of the 3 algorithms. We observe that the average regret of MB-UCBVI is larger than those of MB-PSRL and MB-UCRL2. Moreover, we observe that MB-PSRL obtains the best performance and that its regret seems to grow slower than $O(\sqrt{K})$. This is in accordance to what was observed for PSRL (Osband et al., 2013). Note that the expected number of time steps after $K$ episodes is $K/(1-\beta)$ which means that in our setting with $K = 3000$ episodes there are 300 000 time steps in average. In Figure 1(b), we compare the computation time of the various algorithms. We observe that the computation time (the $y$-axis is in log-scale) of MB-PSRL




(a) Average cumulative regret in function of the number of episodes.

(b) Average runtime per episode. The vertical axis is in log-scale.

Figure 1: Experimental result for the three 4-state random walk arms given in Table 1. The $x$-axis is the number of episodes. Each algorithm is identified by a unique color for all figures.

and MB-UCBVI, the index-based algorithms, are the fastest by far. Moreover, the computation time of these algorithms seem to be independent of the number of episodes. These two figures show that MB-PSRL has the smallest regret and computation time among all compared algorithms.

**Robustness (larger models and different priors)** To test the robustness of MB-PSRL, we conduct two more sets of experiments that are reported in Appendix E. They confirm the superiority of MB-PSRL. The first experiment is an example from Duff (1995) with 9 arms each having 11 states. This model illustrates the effect of the curse of dimensionality: the global MDP has $11^9$ states which implies that the runtime of MB-UCRL2 makes it impossible to use, while MB-PSRL and MB-UCBVI take a few minutes to complete 3000 episodes. Also in this example, MB-PSRL seems to converge faster to the optimal policy than MB-UCBVI. The second experiment tests the robustness of MB-PSRL to the choice of prior distribution. We provide numerical evidences that show that, even when MB-PSRL is run with a prior $\phi$ that is not the one from which $M$ is drawn, the regret of MB-PSRL remains acceptable (around twice the regret obtained with a correct prior).

## 8 Conclusion

In this paper, we present MB-PSRL, a modification of PSRL for markovian bandit problems. We show that its regret is close to the lower bound that we derive for this problem while its runtime scales linearly with the number of arms. Furthermore, and unlike what is usually the case, MB-PSRL does not have an optimistic counterpart that scales well: we prove that MB-UCRL2 also has a sub-linear regret but has a computational complexity exponential in the number of arms. This result generalizes to all the variants of UCRL2 that rely on extended value iteration. We nevertheless show that OFU approach may still be pertinent for markovian bandit problem: MB-UCBVI, a version of UCBVI can use Gittins indices and does not suffer from the dimensionality curse: it has a sub-linear regret in terms of the number of episodes and number of arms as well as a linear time complexity. However its regret remains larger than with MB-PSRL.

## References

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Konstantin E Avrachenkov and Vivek S Borkar. Whittle index based q-learning for restless bandits with average reward. *Automatica*, 139:110186, 2022.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, pp. 1056–1066. PMLR, 2020.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Jhelum Chakravorty and Aditya Mahajan. Multi-armed bandits, gittins index, and its calculation. *Methods and applications of statistics in clinical trials: Planning, analysis, and inferential methods*, 2(416-435): 455, 2014.

Michael O Duff. Q-learning for bandit problems. In *Machine Learning Proceedings 1995*, pp. 209–217. Elsevier, 1995.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 115–122. IEEE, 2010.

Daniel Fink. A compendium of conjugate priors. *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf*, 46, 1997.

Ronan Fruit. *Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge*. PhD thesis, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, 2019.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in mdps with options without prior knowledge. *Advances in Neural Information Processing Systems*, 30, 2017.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586. PMLR, 2018.

Jing Fu, Yoni Nazarathy, Sarat Moka, and Peter G Taylor. Towards q-learning the whittle index for restless bandits. In *2019 Australian & New Zealand Control Conference (ANZCC)*, pp. 249–254. IEEE, 2019.

Nicolas Gast, Bruno Gaujal, and Kimang Khun. Computing whittle (and gittins) index in subcubic time. *arXiv preprint arXiv:2203.05207*, 2022.

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

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pp. 861–898. PMLR, 2015.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34, 2021.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

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

Michael N Katehakis and Arthur F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

Jackson A Killian, Andrew Perrault, and Milind Tambe. Beyond" to act or not to act": Fast lagrangian approaches to general multi-action restless bandits. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 710–718, 2021.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1($2\sigma$2):16, 2007.

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.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. *Advances in Neural Information Processing Systems*, 27, 2014.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Concentration inequalities for multinoulli random variables. *arXiv preprint arXiv:2001.11595*, 2020.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient reinforcement learning in factored mdps with unknown structure. *arXiv preprint arXiv:2009.05986*, 2020.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pp. 770–805. PMLR, 2018.

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

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

Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33:19896–19907, 2020.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Siwei Wang, Longbo Huang, and John Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33:11878–11889, 2020.

Richard Weber. On the gittins index for multiarmed bandits. *The Annals of Applied Probability*, pp. 1024–1033, 1992.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

Peter Whittle. *Optimal control: basics and beyond*. John Wiley & Sons, Inc., 1996.

Ziping Xu and Ambuj Tewari. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. *Advances in Neural Information Processing Systems*, 33:18226–18236, 2020.

**All appendix are given in the supplementary material.**

The appendix are organized as follows:

- In Appendix A, we prove Theorem 1.
- In Appendix B, we obtain a lower bound of the regret of any reinforcement learning algorithm for markovian bandits (Theorem 2).
- In Appendix C, we show that Equation 6 cannot be solved by local indices (Theorem 3).
- In Appendix D, we provide a detailed description of the algorithms that we use in our numerical comparisons.
- In Appendix E, we provide additional numerical experiments that show the good behavior of MB-PSRL.
- In Appendix F, we provide details about the experimental environment and the computation time needed.

# A Proof of Theorem 1

The proof of the regret bounds for our three algorithms share a common structure but with different technical details. In this section, we do a detailed proof of the three algorithms by factorizing as much as possible what can be factorized in the different proofs. This proof is organized as follows:

- In Section A.1, we give an overview of the proof that is common to all algorithms.
- In Section A.2, we provide technical lemmas that are used in the detailed proofs of each algorithms.
- In Section A.3, A.4 and A.5, we provide detailed analysis of MB-PSRL, MB-UCRL2, and MB-UCBVI.

## A.1 Overview of the Proof

Let $\pi_*$ be the optimal policy of the true MDP $M$ and $\pi_k$ the optimal policy for $M_k$, the sampled MDP at episode $k$. Recall that the expected regret is $\sum_{k=1}^{K} \mathbb{E}[\Delta_k]$, where $\Delta_k = W^{\pi_*}_{M,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k})$. For each of the three algorithms, we will define an event $\mathcal{E}^{\text{Algo}}_{k-1}$ that is $\mathcal{O}_{k-1}$-measurable. $\mathcal{E}^{\text{Algo}}_{k-1}$ is true with high probability and guarantees that $M$ and $M_k$ are close. We have:

$$\begin{aligned}\mathbb{E}[\Delta_k] &= \mathbb{E}\left[\Delta_k \mathbb{I}_{\{\neg\mathcal{E}^{\text{Algo}}_{k-1}\}} + \Delta_k \mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}\right] \\ &\leq \mathbb{E}[H_k]\,\mathbb{P}\left(\neg\mathcal{E}^{\text{Algo}}_{k-1}\right) + \mathbb{E}\left[\Delta_k \mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}\right] \end{aligned} \tag{13}$$

because $\Delta_k \leq H_k$ and the random variables $H_k$ and $\mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}$ are independent. For each of the three algorithms, the policy $\pi_k$ used at episode $k$ is optimal for a model $M_k$, that is either sampled from the posterior distribution for MB-PSRL, or computed by extended value iteration for MB-UCRL2, or equal to the model with the bonus for MB-UCBVI. We have

$$\Delta_k = \underbrace{W^{\pi_*}_{M,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{X}_{t_k})}_{:=\Delta_k^{model}} + \underbrace{W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k})}_{:=\Delta_k^{conc}}.$$

As we deal with the expected regret and $H_k$ is independent of the model $M_k$ and of the policy $\pi_k$, we have:

$$\mathbb{E}\left[\Delta_k^{model}\right] = V^{\pi_*}_M(\boldsymbol{X}_{t_k}) - V^{\pi_k}_{M_k}(\boldsymbol{X}_{t_k}) \tag{14}$$

As we see later, the above equation can be used to show that $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}\right]$ is either 0 (for MB-PSRL) or non-positive (for MB-UCRL2 or MB-UCBVI).

We are then left with $\mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}\right]$. To do so, we use Lemma 2 to show that there exists a constant $B_k$ (equal to $H_k$ for MB-PSRL and MB-UCRL2 and $H_k L_{k-1}/(2(1-\beta))$ for MB-UCBVI) such that

$$\begin{aligned}&\mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}\right] = \mathbb{E}\left[\mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}}\left(W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k})\right)\right] \\ &\leq \mathbb{E}\left[\mathbb{I}_{\{\mathcal{E}^{\text{Algo}}_{k-1}\}} \sum_{t=t_k}^{t_{k+1}-1} |r_k(X_{t,A_t}) - r(X_{t,A_t})| + B_k \left\|Q_k(X_{t,A_t},\cdot) - Q(X_{t,A_t},\cdot)\right\|_1\right] \end{aligned} \tag{15}$$

where $\|Q_k(x_a,\cdot) - Q(x_a,\cdot)\|_1 = \sum_{y_a} |Q_k(x_a,y_a) - Q(x_a,y_a)|$. For an arm $a$ and a state $x_a \in \mathcal{S}^a$, we denote[2] by $N_{k-1}(x_a) = \sum_{t=1}^{t_k-1} \mathbb{I}_{\{X_{t,A_t}=x_a\}}$ the number of times that Arm $a$ is activated before episode $k$ while being in state $x_a$. Equation 15 relates the performance gap to the distance between the reward functions and

[2]In the paper, we use the notation $\mathbb{I}_{\{E\}}$ to denote a random variable that equals 1 if $E$ is true and 0 otherwise. For instance, $\mathbb{I}_{\{Y_i=y\}} = 1$ if $Y_i = y$ and 0 otherwise.

transition matrices of the MDPs $M$ and $M_k$. With $L_K{=}\sqrt{2\log\frac{4SnK^2\log K}{1-\beta}}$, the event $\mathcal{E}_{k-1}^{\text{Algo}}$ guarantees that for all $a, x_a$ and $k \geq 1$,

$$|r_k(x_a){-}r(x_a)| \leq \frac{L_K}{\sqrt{\max\{1, N_{k-1}(x_a)\}}} \text{ and } \|Q_k(x_a,\cdot){-}Q(x_a,.)\|_1 \leq \frac{2L_K{+}3\sqrt{S}}{\sqrt{\max\{1, N_{k-1}(x_a)\}}} \tag{16}$$

We use this with Equation 15 to show that:

$$\sum_{k=1}^{K}\mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{Algo}}\}}\right] \leq \mathbb{E}\left[C_K^{\text{Algo}}\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\frac{1}{\sqrt{\max\{1, N_{k-1}(x_a)\}}}\right], \tag{17}$$

where $C_K^{\text{Algo}}$ is a random variable that depends on the algorithm studied.

The final analysis takes care of the right term of Equation 17 and is more technical. It uses the fact that there cannot be too many large terms in this sum because if an arm is activated many times, then $1/\sqrt{N_{k-1}(X_{t,A_t})}$ is small. The main technical hurdle here is to deal with the $K$ random episodes $H_1, \ldots, H_K$. This is specific to our approach compared to the analysis of finite horizons. To bound this, one needs to bound terms of the form $\mathbb{E}\left[\max_{1\leq k\leq K}(H_k)^\alpha\right]$ with $\alpha \in \{1.5, 2\}$ (see Equation 32). To bound this, we use the geometric distribution of $H_k$ to show that $\mathbb{E}\left[\max_{1\leq k\leq K}(H_k)^\alpha\right] = O((\frac{\log K}{1-\beta})^\alpha)$ (see Lemma 4).

### A.2 Technical lemmas common to the three algorithms

In this section, we establish a series of lemmas that are true for any learning algorithm used. They show that:

- The estimates $\hat{r}$ and $\hat{Q}$ concentrates on their true values (Lemma 1);
- One can transform $\Delta_k^{conc}$ into Equation 15 (Lemma 2);
- The sum Equation 17 can be analyzed (Lemma 3).

#### A.2.1 High Probability Events

Recall that $\mathcal{O}_{k-1}$ are the observations collected by the decision maker before episode $k$. Based on $\mathcal{O}_{k-1}$, we compute the empirical estimators of reward vector and transition matrix as the following: For all $a \in [n]$ and any $x_a \in \mathcal{S}^a$, let $N_{k-1}(x_a) = \sum_{t=1}^{t_k-1}\mathbb{I}_{\{X_{t,A_t}=x_a\}}$ be the number of times so far that an arm $a$ was activated in state $x_a$ (at episode 1, we have $N_0(x_a) = 0$). Recall that $t_k{:=}1{+}\sum_{i=1}^{k-1}H_i$, and that $\hat{r}_{k-1}$ and $\hat{Q}_{k-1}$ are the empirical mean reward vector and transition matrix. More precisely, $\hat{r}_{k-1}(x_a)$ is the empirical mean reward earned when arm $a$ is chosen while being in state $x_a$:

$$\hat{r}_{k-1}(x_a) = \frac{1}{N_{k-1}(x_a)}\sum_{t=1}^{t_k-1}R_t\mathbb{I}_{\{A_t=a\wedge X_{t,A_t}=x_a\}},$$

and $\hat{Q}_{k-1}(x_a, y_a)$ is the fraction of times that arm $a$ moved from $x_a$ to $y_a$:

$$\hat{Q}_{k-1}(x_a, y_a) = \frac{1}{N_{k-1}(x_a)}\sum_{t=1}^{t_k-1}\mathbb{I}_{\{A_t=a\wedge X_{t,A_t}=x_a\wedge X_{t+1,A_t}=y_a\}}.$$

We design confidence sets similar to Auer et al. (2008); Bartlett & Tewari (2012).

**Lemma 1.** *For any $k \leq K$, let $L_{k-1} = \sqrt{2\log(\frac{2SnK(k-1)\log(K(k-1))}{1-\beta})}$. Let*

$$\mathcal{E}^H_{k-1} := \left\{\forall k' \leq k-1 : H_{k'} \leq \frac{\log(K(k-1))}{1-\beta}\right\} \tag{18}$$

$$\mathcal{E}^r_{k-1} := \left\{\forall a \in [n], x_a \in \mathcal{S}^a, k' \leq k-1 : |\hat{r}_{k'}(x_a) - r(x_a)| \leq \frac{L_{k-1}}{2\sqrt{\max\{1, N_{k'}(x_a)\}}}\right\} \tag{19}$$

$$\mathcal{E}^Q_{k-1} := \left\{\forall a \in [n], x_a \in \mathcal{S}^a, k' \leq k-1 : \left\|\hat{Q}_{k'}(x_a, \cdot) - Q(x_a, \cdot)\right\|_1 \leq \frac{L_{k-1} + 1.5\sqrt{S}}{\sqrt{\max\{1, N_{k'}(x_a)\}}}\right\}. \tag{20}$$

$$\mathcal{E}^V_{k-1} := \Bigg\{\forall a \in [n], \boldsymbol{x} \in \mathcal{X}, k' \leq k-1 : |\hat{r}_{k'}(x_a) - r(x_a) + \beta \sum_{\boldsymbol{y}} (\hat{P}^a_{k'}(\boldsymbol{x}, \boldsymbol{y}) - P^a(\boldsymbol{x}, \boldsymbol{y})) V^{\pi_*}_M(\boldsymbol{y})| \leq \frac{L_{k-1}}{2(1-\beta)\sqrt{\max\{1, N_{k'}(x_a)\}}}\Bigg\} \tag{21}$$

*Then, the above events are all $\mathcal{O}_{k-1}$-measurable. Moreover:*

$$\mathbb{P}\left(\neg\mathcal{E}^H_{k-1}\right) \leq 1/K$$
$$\mathbb{P}\left(\neg\mathcal{E}^r_{k-1}\right) \leq 2/K$$
$$\mathbb{P}\left(\neg\mathcal{E}^Q_{k-1}\right) \leq 2/K$$
$$\mathbb{P}\left(\neg\mathcal{E}^V_{k-1}\right) \leq 2/K.$$

*Proof.* For event $\mathcal{E}^H_{k-1}$, since $\{H_{k'}\}_{k' \leq k-1}$ are i.i.d. and geometrically distributed with parameter $(1-\beta)$, we have that

$$\mathbb{P}\left(\exists k' \leq k-1 : H_{k'} > \epsilon\right) \leq \sum_{k'=1}^{k-1} \mathbb{P}\left(H_{k'} > \epsilon\right) = (k-1)\beta^{\lfloor \epsilon \rfloor}.$$

Then, with $\epsilon = \frac{\log(1/(K(k-1)))}{\log(\beta)}$, we get $\mathbb{P}\left(\exists k' \leq k-1 : H_{k'} > \epsilon\right) \leq 1/K$. Moreover,

$$\epsilon = \frac{\log(1/(K(k-1)))}{\log(\beta)} = \frac{\log(K(k-1))}{\log(1/\beta)} < \frac{\log(K(k-1))}{1-\beta}.$$

Then, $\mathbb{P}\left(\exists k' \leq k-1 : H_{k'} > \frac{\log(K(k-1))}{1-\beta}\right) \leq 1/K$.

Let $\tau_k = (k-1)\frac{\log(K(k-1))}{1-\beta}$. Under event $\mathcal{E}^H_{k-1}$, the random variable $t_k$ is upper bounded by the deterministic quantity $\tau_k$. In what follows, we assume that event $\mathcal{E}^H_{k-1}$ holds.

For event $\mathcal{E}^r_{k-1}$, let $\tilde{r}_\ell(x_a)$ be a random variable that is the empirical mean of $\ell$ i.i.d. realization of the reward when the arm in state $x_a$ is chosen. In particular, $\hat{r}_{k-1}(x_a) = \tilde{r}_{N_{k-1}(x_a)}(x_a)$. By Hoeffding's inequality, for any $\epsilon > 0$, one has:

$$\mathbb{P}\left(|\tilde{r}_\ell(x_a) - r(x_a)| \geq \epsilon\right) \leq 2e^{-2\ell\epsilon^2}.$$

In particular, this holds for $\epsilon = \sqrt{\frac{\log(2SnK\tau_k)}{2\ell}}$. As $N_{k-1}(x_a) < \tau_k$, by using the union-bound, this implies that:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}^H_{k-1} \wedge \exists a, x_a, k' \leq k-1 : |\hat{r}_{k'}(x_a) - r(x_a)| \geq \sqrt{\frac{\log(2SnK\tau_k)}{2N_{k'}(x_a)}}\right) \qquad (22)\\
&\leq \sum_a \sum_{x_a} \mathbb{P}\left(\exists \ell \in \{1, \dots, \tau_k - 1\} : |\tilde{r}_\ell(x_a) - r(x_a)| \geq \sqrt{\frac{\log(2SnK\tau_k)}{2\ell}}\right)\\
&\leq \sum_{\ell=1}^{\tau_k} \sum_a \sum_{x_a} \mathbb{P}\left(|\tilde{r}_\ell(x_a) - r(x_a)| \geq \sqrt{\frac{\log(2SnK\tau_k)}{2\ell}}\right)\\
&\leq nS \sum_{\ell=1}^{\tau_k} 2e^{-2\ell \frac{\log(2SnK\tau_k)}{2\ell}} = 1/K,
\end{aligned}$$

where the second and third line is the union on all possible events $N_{k'}(x_a)=\ell$ for all $\ell \in \{1, \dots, \tau_k - 1\}$. In total this says $\mathbb{P}\left(\mathcal{E}^H_{k-1} \wedge \neg\mathcal{E}^r_{k-1}\right) \leq 1/K$. Now, $\neg\mathcal{E}^r_{k-1} = (\mathcal{E}^H_{k-1} \wedge \neg\mathcal{E}^r_{k-1}) \vee (\neg\mathcal{E}^H_{k-1} \wedge \neg\mathcal{E}^r_{k-1})$. Then, using union bound,

$$\begin{aligned}
\mathbb{P}\left(\neg\mathcal{E}^r_{k-1}\right) &\leq \mathbb{P}\left(\neg\mathcal{E}^r_{k-1} \wedge \mathcal{E}^H_{k-1}\right) + \mathbb{P}\left(\neg\mathcal{E}^r_{k-1} \wedge \neg\mathcal{E}^H_{k-1}\right)\\
&\leq \mathbb{P}\left(\neg\mathcal{E}^r_{k-1} \wedge \mathcal{E}^H_{k-1}\right) + \mathbb{P}\left(\neg\mathcal{E}^H_{k-1}\right) \leq 2/K
\end{aligned}$$

The event $\mathcal{E}^Q_{k-1}$ is similar but by using Weissman's inequality (Weissman et al., 2003) instead of Hoeffding's bound. Indeed, by using Equation (8) in Theorem 2.1 of Weissman et al. (2003), if $N_{k-1}(x_a)$ was not a random variable, one would have

$$\mathbb{P}\left(\left\|\hat{Q}_{k-1}(x_a, \cdot) - Q(x_a, \cdot)\right\|_1 \geq \epsilon\right) \leq 2^S e^{-N_{k-1}(x_a)\epsilon^2/2}.$$

Following the same approach as for Equation 22 with $\epsilon = \sqrt{2\log(SnK\tau_k 2^S)/N_{k-1}(x_a)}$, we use the union-bound to show that:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{E}^H_{k-1} \wedge \exists a, x_a, k' \leq k-1 : \left\|\hat{Q}_{k'}(x_a, \cdot) - Q(x_a, \cdot)\right\|_1 \geq \sqrt{\frac{2\log(SnK\tau_k 2^S)}{N_{k'}(x_a)}}\right)\\
&\leq \tau_k nS 2^S e^{-N_{k'}(x_a)\frac{2\log(SnK\tau_k 2^S)}{2N_{k'}(x_a)}} = 1/K.
\end{aligned}$$

By definition of $L_{k-1} = \sqrt{2\log(2SnK\tau_k)}$ and since $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, we have

$$\begin{aligned}
\sqrt{2\log(SnK\tau_k 2^S)} &= \sqrt{2\log(2SnK\tau_k) + 2(S-1)\log 2}\\
&\leq L_{k-1} + \sqrt{2(S-1)\log 2} \leq L_{k-1} + 1.5\sqrt{S}.
\end{aligned}$$

Hence:

$$\mathbb{P}\left(\mathcal{E}^H_{k-1} \wedge \exists a, x_a, k' \leq k-1 : \left\|\hat{Q}_{k'}(x_a, \cdot) - Q(x_a, \cdot)\right\|_1 \geq \frac{L_{k-1} + 1.5\sqrt{S}}{\sqrt{N_{k'}(x_a)}}\right) \leq 1/K.$$

As done for $\mathcal{E}^r_{k-1}$, we have $\neg\mathcal{E}^Q_{k-1} = (\mathcal{E}^H_{k-1} \wedge \neg\mathcal{E}^Q_{k-1}) \vee (\neg\mathcal{E}^H_{k-1} \wedge \neg\mathcal{E}^Q_{k-1})$. With the same process, we get $\mathbb{P}\left(\neg\mathcal{E}^Q_{k-1}\right) \leq 2/K$.

For event $\mathcal{E}^V_{k-1}$, we have that $\hat{r}_{k-1} + \hat{P}_{k-1}V^{\pi_*}_M$ is the empirical mean of $r + PV^{\pi_*}_M$. This is because $V^{\pi_*}_M$ is deterministic and $\hat{r}_{k-1}$ and $\hat{P}_{k-1}$ are empirical mean of $r$ and $P$ respectively. Using Hoeffding's inequality and following the same approach above, we have $\mathbb{P}\left(\neg\mathcal{E}^V_{k-1}\right) \leq 2/K$. □

Note that Lemma 1 is about the statistical properties of the observations $\mathcal{O}_{k-1}$ in the observation space. These properties are true for any learning algorithms. In fact, we will combine different events of this lemma to bound the regret of our algorithm accordingly.

### A.2.2 Concentration Gap

At episode $k$, our algorithms believe that the unknown MDP $M$ is the MDP $M_k$. For bayesian algorithms, $M_k$ is sampled from posterior distribution while for optimistic algorithms, $M_k$ is chosen with respect to optimism principle. The algorithms follow the policy $\pi_k$ that is optimal for $M_k$. Recall that $W^{\pi_k}_{M,1:H_k}(\boldsymbol{x})$ is the expected reward of the MDP $M$ under policy $\pi_k$, starts in state $\boldsymbol{x}$ and lasts for $H_k$ time steps and the expected cumulative discounted reward in $M$ starting from state $\boldsymbol{x}$ under policy $\pi_k$ is $V^{\pi_k}_M(\boldsymbol{x})=\mathbb{E}[W^{\pi_k}_{M,1:H_k}(\boldsymbol{x})]$ where $H_k \sim \text{Geom}(1-\beta)$ is the horizon of episode $k$.

**Lemma 2.** *For episode $k$, let $B_k \in \mathbb{R}^+$ be an upper bound[3] of $W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{x})$, i.e., a constant $B_k$ such that for any $\boldsymbol{x} \in \mathcal{X}$, $W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{x}) \leq B_k$. We have,*

$$\begin{aligned}
&\mathbb{E}\left[\Delta^{conc}_k | \mathcal{O}_{k-1}, H_k, M_k, M\right] = \mathbb{E}\left[W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{X}_{t_k}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{X}_{t_k}) | \mathcal{O}_{k-1}, H_k, M_k, M\right] \\
&\leq \mathbb{E}\left[\sum_{t=t_k}^{t_{k+1}-1} |r_k(X_{t,A_t}) - r(X_{t,A_t})| + B_k \left\|Q_k(X_{t,A_t},\cdot) - Q(X_{t,A_t},\cdot)\right\|_1 | \mathcal{O}_{k-1}, H_k, M_k, M\right]
\end{aligned} \tag{23}$$

*Proof.* From Equation 8 with $a = \pi_k(\boldsymbol{x})$,

$$W^{\pi_k}_{M,1:H_k}(\boldsymbol{x}) = r(x_a) + \sum_{\boldsymbol{y}} P^{\pi_k}(\boldsymbol{x},\boldsymbol{y}) W^{\pi_k}_{M,2:H_k}(\boldsymbol{y}) \tag{24}$$

where $P^{\pi_k}$ is the state transition dynamic of the system when following the policy $\pi_k$. Comparing the sampled MDP $M_k$ with the original $M$ and using Equation 24, one has

$$\begin{aligned}
W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{x}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{x}) &= r_k(x_a) - r(x_a) \\
&\quad + \sum_{\boldsymbol{y}} P^{\pi_k}_k(\boldsymbol{x},\boldsymbol{y}) W^{\pi_k}_{M_k,2:H_k}(\boldsymbol{y}) - \sum_{\boldsymbol{y}} P^{\pi_k}(\boldsymbol{x},\boldsymbol{y}) W^{\pi_k}_{M,2:H_k}(\boldsymbol{y}).
\end{aligned}$$

Note that in the above equation, the last term is of the form $P^{\pi_k}_k W^{\pi_k}_{M_k} - P^{\pi_k} W^{\pi_k}_M$, which is equal to $(P^{\pi_k}_k - P^{\pi_k}) W^{\pi_k}_{M_k} + P^{\pi_k}(W^{\pi_k}_{M_k} - W^{\pi_k}_M)$. Moreover, $W^{\pi_k}_{M_k}$ is less than $B_k$. Plugging this to the above equation shows that:

$$\begin{aligned}
&W^{\pi_k}_{M_k,1:H_k}(\boldsymbol{x}) - W^{\pi_k}_{M,1:H_k}(\boldsymbol{x}) \\
&\quad \leq |r_k(x_a) - r(x_a)| + B_k \sum_{\boldsymbol{y}} |P^{\pi_k}_k(\boldsymbol{x},\boldsymbol{y}) - P^{\pi_k}(\boldsymbol{x},\boldsymbol{y})| \\
&\qquad + \sum_{\boldsymbol{y}} P^{\pi_k}(\boldsymbol{x},\boldsymbol{y})(W^{\pi_k}_{M_k,2:H_k}(\boldsymbol{y}) - W^{\pi_k}_{M,2:H_k}(\boldsymbol{y})) \\
&\quad = |r_k(x_a) - r(x_a)| + B_k \left\|P^{\pi_k}_k(\boldsymbol{x},\cdot) - P^{\pi_k}(\boldsymbol{x},\cdot)\right\|_1 + D^{M_k,M}_{H_k}(\boldsymbol{x}) \\
&\qquad + W^{\pi_k}_{M_k,2:H_k}(\boldsymbol{X}_1) - W^{\pi_k}_{M,2:H_k}(\boldsymbol{X}_1)
\end{aligned}$$

where $D^{M_k,M}_{H_k}(\boldsymbol{x}) := \sum_{\boldsymbol{y}} P^{\pi_k}(\boldsymbol{x},\boldsymbol{y})(W^{\pi_k}_{M_k,2:H_k}(\boldsymbol{y}) - W^{\pi_k}_{M,2:H_k}(\boldsymbol{y})) - (W^{\pi_k}_{M_k,2:H_k}(\boldsymbol{X}_1) - W^{\pi_k}_{M,2:H_k}(\boldsymbol{X}_1))$. Note that in the equation above, $D^{M_k,M}_{H_k}(\boldsymbol{x})$ is a martingale difference with $\boldsymbol{X}_1 \sim P^{\pi_k}(\boldsymbol{x},\cdot)$. Hence, the expected value of the martingale difference sequence is zero. As only arm $a$ makes a transition, we have $\left\|P^{\pi_k}_k(\boldsymbol{x},\cdot) - P^{\pi_k}(\boldsymbol{x},\cdot)\right\|_1 = \left\|Q_k(x_a,\cdot) - Q(x_a,\cdot)\right\|_1$. Hence, a direct induction shows that Equation 23 holds. □

### A.2.3 Bound on the double sum

Recall that for $k \leq K$, any $a \in [n]$ and any $x_a \in \mathcal{S}^a$, $N_{k-1}(x_a) = \sum_{t=1}^{t_k-1} \mathbb{I}_{\{X_{t,A_t}=x_a\}}$ is the number of times so far that an arm $a$ was activated in state $x_a$ (at episode 1, we have $N_0(x_a) = 0$) and $\{H_k\}_{k \leq K}$ be the sequence of episode horizons.

[3] We will use $B_k = H_k$ for MB-PSRL and MB-UCRL2 and $B_k = H_k L_{k-1}/(2(1-\beta))$ for MB-UCBVI.

**Lemma 3.** *For any learning algorithms, we have*

$$\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\frac{1}{\sqrt{\max\{1,N_{k-1}(X_{t,A_t})\}}}\leq Sn\max_{k\leq K}H_k+2\sqrt{SnK\max_{k\leq K}H_k}$$

*Proof.* Let $\tilde{N}_t(x_a)$ be the number of times that arm $a$ has been activated before time $t$ while being in state $x_a$. By definition, $\tilde{N}_{t_k}(x_a)=N_{k-1}(x_a)$. Moreover, if $t\in\{t_k,\dots,t_{k+1}-1\}$, then $\tilde{N}_t(x_a)\leq N_{k-1}(x_a)+H_k$. This shows that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\frac{1}{\sqrt{\max\{1,N_{k-1}(X_{t,A_t})\}}}&\leq\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\frac{1}{\sqrt{\max\{1,\tilde{N}_t(X_{t,A_t})-H_k\}}}\\&\leq\sum_{t=1}^{t_{K+1}-1}\frac{1}{\sqrt{\max\{1,\tilde{N}_t(X_{t,A_t})-\max_k H_k\}}}.\end{aligned}$$

The above sum can be reordered to group terms by state: The above sum equals

$$\begin{aligned}\sum_{a,x_a}\sum_{m=1}^{\tilde{N}_{t_{K+1}}(x_a)}\frac{1}{\sqrt{\max\{1,m-\max_k H_k\}}}&\leq\sum_{a,x_a}\left[\max_k H_k+\sum_{m=1}^{\max\{1,\tilde{N}_{t_{K+1}}(x_a)-\max_k H_k\}}\frac{1}{\sqrt{m}}\right],\\&\leq Sn\max_k H_k+\sum_{a,x_a}\sum_{m=1}^{\tilde{N}_{t_{K+1}}(x_a)}\frac{1}{\sqrt{m}},\\&\leq Sn\max_k H_k+2\sum_{a,x_a}\sqrt{\tilde{N}_{t_{K+1}}(x_a)},\end{aligned}$$

where the last inequality holds because $\sum_{m=1}^{t_{K+1}}1/\sqrt{m}\leq\int_1^{t_{K+1}}1/\sqrt{x}dx\leq 2\sqrt{t_{K+1}}$.

Now, by Cauchy-Schwartz inequality, and because $\sum_{a,x_a}\tilde{N}_{t_{K+1}}(x_a)=t_{K+1}-1=\sum_{k=1}^{K}H_k$, we have:

$$\sum_{a,x_a}\sqrt{\tilde{N}_{t_{K+1}}(x_a)}\leq\Big(\sum_{a,x_a}\tilde{N}_{t_{K+1}}(x_a)\Big)^{1/2}\Big(\sum_{a,x_a}1\Big)^{1/2}=\sqrt{Sn\sum_{k=1}^{K}H_k}\leq\sqrt{SnK\max_{k\leq K}H_k}.$$

□

#### A.2.4 Bound on the expectation of $\mathbb{E}\left[\max_{k\leq K}H_k\right]$

**Lemma 4.** *Let $\alpha\in[1,2.5]$. Then,*

$$\mathbb{E}\left[\max_{k\leq K}(H_k)^\alpha\right]\leq 5+5\left(\frac{\log K}{1-\beta}\right)^\alpha. \tag{25}$$

*Proof.* By definition, we have

$$\begin{aligned}\mathbb{E}\left[\max_{k\leq K}(H_k)^\alpha\right]&=\sum_{i=1}^{\infty}\mathbb{P}\left(\max_{k\leq K}(H_k)^\alpha\geq i\right)\\&\leq\sum_{i=1}^{\infty}\min(1,K\mathbb{P}\left((H_k)^\alpha\geq i\right))\\&=\sum_{i=1}^{\infty}\min(1,K\beta^{i^{1/\alpha}}),\end{aligned}$$

where the inequality comes from the union bound and the last equality is because the random variables $H_k$ are geometrically distributed.

Let $A = \min\{i : K\beta^{i^{1/\alpha}} \leq 1\}$. Decomposing the above sum by group of size $A$, we have

$$\begin{aligned}\sum_{i=1}^{\infty} \min(1, K\beta^{i^{1/\alpha}}) &= \sum_{j=0}^{\infty} \sum_{i=Aj+1}^{A(j+1)} \min(1, K\beta^{i^{1/\alpha}}) \\ &\leq \sum_{j=0}^{\infty} A \min(1, K\beta^{(Aj)^{1/\alpha}}) \\ &= A + A\sum_{j=1}^{\infty} K(\beta^{A^{1/\alpha}})^{j^{1/\alpha}}, \end{aligned} \tag{26}$$

where the inequality holds because $\beta^{i^{1/\alpha}}$ is decreasing in $i$.

By definition of $A$, we have $\beta^{A^{1/\alpha}} \leq 1/K$. This implies that the second term of Equation 26 is smaller than $\sum_{j=1}^{\infty} K(1/K)^{j^{1/\alpha}} = \sum_{j=1}^{\infty} K^{1-j^{1/\alpha}}$. As $\alpha \leq 2.5$, if $K \geq 5$, this is smaller than $\sum_{j=1}^{\infty} 5^{1-j^{1/2.5}} \approx 3.92 < 4$. This shows that for $K \geq 5$, we have:

$$\mathbb{E}\left[\max_{k \leq K} (H_k)^{\alpha}\right] \leq 5A,$$

where $A = \lceil (-\log K/\log \beta)^{\alpha} \rceil \leq 1 + (\log K/(1-\beta))^{\alpha}$.

As for the case where $K \leq 4$, we have $\mathbb{E}\left[\max_{k \leq K} (H_k)^{\alpha}\right] \leq K\mathbb{E}\left[H_1^{\alpha}\right] \leq \frac{K}{(1-\beta)^{\alpha}}$. This term is smaller than Equation 25 for $K \leq 4$. □

### A.3 Detailed analysis of MB-PSRL

We decompose the analysis of PSRL in three steps:

- We define the high-probability event $\mathcal{E}_{k-1}^{\text{PSRL}}$.
- We analyze $\sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{model} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right]$ (which equals 0 here because of posterior sampling).
- We analyze $\sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{conc} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right]$.

We will use the same proof structure for MB-UCRL2 and MB-UCBVI.

Before doing the proof, we start by a first lemma that that essentially formalizes the fact that the distribution of $M$ given $\mathcal{O}_{k-1}$ is the same as the distribution of the sampled MDP $M_k$ conditioned on $\mathcal{O}_{k-1}$.

**Lemma 5.** *Assume that the MDP $M$ is drawn according to the prior $\phi$ and that $M_k$ is draw according to the posterior $\phi(\cdot \mid \mathcal{O}_{k-1})$. Then, for any $\mathcal{O}_{k-1}$-measurable function $g$, one has:*

$$\mathbb{E}\left[g(M)\right] = \mathbb{E}\left[g(M_k)\right]. \tag{27}$$

*Proof.* At the start of each episode $k$, MB-PSRL computes the posterior distribution of $M$ conditioned on the observations $\mathcal{O}_{k-1}$, and draws $M_k$ from it. This implies that $M$ and $M_k$ are identically distributed conditioned on $\mathcal{O}_{k-1}$. Consequently, if $g$ is a $\mathcal{O}_{k-1}$-measurable function, one has:

$$\mathbb{E}\left[g(M) \mid \mathcal{O}_{k-1}\right] = \mathbb{E}\left[g(M_k) \mid \mathcal{O}_{k-1}\right].$$

Equation 27 then follows from the tower rule. □

### A.3.1 Definition of the high-probability event $\mathcal{E}_{k-1}^{\text{PSRL}}$

**Lemma 6.** *At episode $k$, the event*

$$\mathcal{E}_{k-1}^{PSRL} = \Bigg\{ \forall a \in [n], x_a \in \mathcal{S}^a, k' \leq k-1 \colon |r_{k'+1}(x_a) - r(x_a)| \leq \frac{L_{k-1}}{\sqrt{\max\{1, N_{k'}(x_a)\}}}, \\ \|Q_{k'+1}(x_a, \cdot) - Q(x_a, \cdot)\|_1 \leq \frac{2L_{k-1} + 3\sqrt{S}}{\sqrt{\max\{1, N_{k'}(x_a)\}}}, \text{ and } H_{k'} \leq \frac{\log(K(k-1))}{1-\beta} \Bigg\}$$

*is $\mathcal{O}_{k-1}$-measurable and true with probability at least $1 - 9/K$.*

*Proof.* Recall that for MB-PSRL, at the beginning of episode $k$, we sample a MDP $M_k$. We define the two events that are the analogue of the events Equation 19 and Equation 20 of Lemma 1 but replacing the true MDP $M$ by the sampled MDP $M_k$:

$$\tilde{\mathcal{E}}_{k-1}^{r} := \left\{ \forall a \in [n], x_a \in \mathcal{S}^a, k' \leq k-1 \colon |\hat{r}_{k'}(x_a) - r_{k'+1}(x_a)| \leq \frac{L_{k-1}}{2\sqrt{\max\{1, N_{k'}(x_a)\}}} \right\}$$

$$\tilde{\mathcal{E}}_{k-1}^{Q} := \left\{ \forall a \in [n], x_a \in \mathcal{S}^a, k' \leq k-1 \colon \left\| \hat{Q}_{k'}(x_a, \cdot) - Q_{k'+1}(x_a, \cdot) \right\|_1 \leq \frac{L_{k-1} + 1.5\sqrt{S}}{\sqrt{\max\{1, N_{k'}(x_a)\}}} \right\}$$

These events are $\mathcal{O}_{k-1}$-measurable. Hence, Lemma 5, combined with Lemma 1 implies that $\mathbb{P}\left(\neg\tilde{\mathcal{E}}_{k-1}^{r}\right) = \mathbb{P}\left(\neg\mathcal{E}_{k-1}^{r}\right) \leq 2/K$ and $\mathbb{P}\left(\neg\tilde{\mathcal{E}}_{k-1}^{Q}\right) = \mathbb{P}\left(\neg\mathcal{E}_{k-1}^{Q}\right) \leq 2/K$. Since the complement of $\mathcal{E}_{k-1}^{\text{PSRL}}$ is the union of $\neg\mathcal{E}_{k-1}^{r}, \neg\tilde{\mathcal{E}}_{k-1}^{r}, \neg\mathcal{E}_{k-1}^{Q}, \neg\tilde{\mathcal{E}}_{k-1}^{Q}$ and $\neg\mathcal{E}_{k-1}^{H}$, the union bound implies that $\mathbb{P}\left(\mathcal{E}_{k-1}^{\text{PSRL}}\right) \geq 1 - 9/K$. □

### A.3.2 Analysis of $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right]$ for MB-PSRL.

Lemma 5 implies that for MB-PSRL, $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] = 0$ because $\mathcal{E}_{k-1}^{\text{PSRL}}$, $\pi_k$ and $M_k$ are $\mathcal{O}_{k-1}$-measurable.

### A.3.3 Analysis of $\mathbb{E}\left[\Delta_k^{conc}\right]$ for MB-PSRL.

Following Equation 13, the bayesian regret can be written as:

$$\begin{aligned} \text{BayReg}(K, \text{MB-PSRL}, \phi) &= \sum_{k=1}^{K} \mathbb{E}\left[\Delta_k\right] \leq \sum_{k=1}^{K} \mathbb{E}\left[H_k\right] \mathbb{P}\left(\neg\mathcal{E}_{k-1}^{\text{PSRL}}\right) + \mathbb{E}\left[\Delta_k \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] \\ &\leq \frac{9}{(1-\beta)} + \sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] + \mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] \end{aligned} \tag{28}$$

where the last inequality holds due to Lemma 6. By the previous section, the second term of Equation 28 is zero. As all rewards are bounded by 1, $W_{M_k,1:H_k}^{\pi_k}(\boldsymbol{X}_{t_k}) \leq H_k$. Hence, by applying Lemma 2 with the upper bound $B_k = H_k$, and because $\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}$ is deterministic given $\mathcal{O}_{k-1}$, we have

$$\begin{aligned} \mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}} \mid \mathcal{O}_{k-1}, H_k, M_k, M\right]\right] \\ &\leq \mathbb{E}\Bigg[\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}} \sum_{t=t_k}^{t_{k+1}-1} |r_k(X_{t,A_t}) - r(X_{t,A_t})| \\ &\qquad + H_k \left\|Q_k(X_{t,A_t}, \cdot) - Q(X_{t,A_t}, \cdot)\right\|_1 \Bigg]. \end{aligned} \tag{29}$$

Let $\mathcal{R}_k := \sum_{t=t_k}^{t_{k+1}-1} |r_k(X_{t,A_t}) - r(X_{t,A_t})| + H_k \left\| Q_k(X_{t,A_t},\cdot) - Q(X_{t,A_t},\cdot) \right\|_1$. By using the definition of $\mathcal{E}_{k-1}^{\text{PSRL}}$, we have:

$$\begin{aligned}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\mathcal{R}_k &\leq \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}} \sum_{t=t_k}^{t_{k+1}-1} \frac{L_{k-1}+(2L_{k-1}+3\sqrt{S})H_k}{\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}} \\ &\leq \sum_{t=t_k}^{t_{k+1}-1} \frac{L_{k-1}+(2L_{k-1}+3\sqrt{S})H_k}{\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}} \end{aligned} \tag{30}$$

Hence, summing over all $K$ episodes gives us:

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\mathcal{R}_k &\leq \big(L_K + (2L_K+3\sqrt{S}) \max_{k\leq K} H_k\big) \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}} \\ &\leq 3(L_K+\sqrt{S}) \max_{k\leq K} H_k \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}}, \end{aligned} \tag{31}$$

where the first inequality holds because $L_k \leq L_K$ and $\max_{k\leq K} H_k \geq 1$. Note that the last inequality leads to a slightly worst bound but simplifies the expression. By Lemma 3, we get

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\mathcal{R}_k &\leq 3(L_K+\sqrt{S}) \max_{k\leq K} H_k (Sn \max_{k\leq K} H_k + 2\sqrt{SnK \max_{k\leq K} H_k}) \\ &= 3(L_K+\sqrt{S})(Sn \max_{k\leq K} (H_k)^2 + 2\sqrt{SnK} \max_{k\leq K} (H_k)^{3/2}) \end{aligned}$$

Then,

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{conc} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] &\leq 3(L_K+\sqrt{S})\left(Sn\mathbb{E}\left[\max_{k\leq K}(H_k)^2\right] + 2\sqrt{SnK}\mathbb{E}\left[\max_{k\leq K}(H_k)^{3/2}\right]\right) \\ &\leq 3(L_K+\sqrt{S})\left(Sn\left(5+5\left(\frac{\log K}{1-\beta}\right)\right)^2 + \sqrt{SnK}\left(5+5\left(\frac{\log K}{1-\beta}\right)\right)^{3/2}\right) \end{aligned} \tag{32}$$

where the last inequality is true due to Lemma 4. With $L_K = \sqrt{2\log\frac{4SnK^2\log K}{1-\beta}}$, this implies that there exists a constant $C$ independent of all problem's parameters such that:

$$\text{BayReg}(K, \text{MB-PSRL}, \phi) \leq C\left(\sqrt{S} + \log\left(\frac{SnK\log K}{1-\beta}\right)\right)\left(Sn\left(\frac{\log K}{1-\beta}\right)^2 + \sqrt{SnK}\left(\frac{\log K}{1-\beta}\right)^{3/2}\right).$$

#### A.3.4 Remark on the dependence on $S$

Our bound is linear in $S$, the state size of each arm, because our proof follows the approach used in Osband et al. (2013). Using another proof methodology, it is argued in Osband & Van Roy (2017) that the regret of PSRL grows as the square root of the state space size and not linearly. In our paper, we choose to use the more conservative approach of Osband et al. (2013) because we believe that the proof used in Osband & Van Roy (2017) is not correct (in particular the use of a deterministic $v$ in Equation (16) of the proof of Lemma 3 in Appendix A in the arXiv version of Osband & Van Roy (2017) seems incompatible with the use of Lemma 4 of the same paper). In fact, when considering the worst case realization of $v$, the concentration bound in Equation (16) of the paper is equivalent to the (scaled) L1 norm of transition concentration. We are not alone to point out this error. Effectively, Agrawal & Jia (2017) used Lemma C.1 and Lemma C.3 (equivalence of Lemma 3 of Osband & Van Roy (2017)) to get a bound in square root of the state space size. But both lemmas are erroneous as mentioned in the latest arXiv version of Agrawal & Jia (2017). The validity of Lemma 3 is also questioned on page 87 of Fruit (2019). While it is informal, the recent work of Qian et al. (2020) also theoretically contradicts the lemma.

### A.4 Case of MB-UCRL2

The proof follows the same steps as for MB-PSRL. While the high probability event is simpler, the additional complexity is to show that $\sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{model}\right] \leq 0$ by using the optimism principle.

#### A.4.1 Definition of the high probability event

**Lemma 7.** *At episode $k$, the event*

$$\mathcal{E}_{k-1}^{UCRL2} = \Bigg\{ \forall a \in [n], x_a \in \mathcal{S}^a, k' \leq k-1 \colon |\hat{r}_{k'}(x_a) - r(x_a)| \leq \frac{L_{k-1}}{2\sqrt{\max\{1, N_{k'}(x_a)\}}},$$
$$\left\| \hat{Q}_{k'}(x_a, \cdot) - Q(x_a, \cdot) \right\|_1 \leq \frac{L_{k-1} + 1.5\sqrt{S}}{\sqrt{\max\{1, N_{k'}(x_a)\}}}, \text{ and } H_{k'} \leq \frac{\log(K(k-1))}{1-\beta} \Bigg\}$$

*is $\mathcal{O}_{k-1}$-measurable and true with probability at least $1 - 5/K$.*

*Proof.* The complement of $\mathcal{E}_{k-1}^{\text{UCRL2}}$ is the union of $\neg\mathcal{E}_{k-1}^{r}$, $\neg\mathcal{E}_{k-1}^{Q}$ and $\neg\mathcal{E}_{k-1}^{H}$. We conclude the proof by using the union bound and $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^{r}\right) \leq 2/K$, $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^{Q}\right) \leq 2/K$ and $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^{H}\right) \leq 1/K$. □

#### A.4.2 Analysis of $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right]$ – Optimism of MB-UCRL2

Recall that $\pi_*$ is the optimal policy of the unknown MDP $M$ and that $\pi_k$ is the policy used in episode $k$. $\pi_k$ is optimal for the optimistic MDP that is chosen from the plausible MDP set $\mathbb{M}_k$:

$$\pi_k \in \arg\max_{\pi} \max_{M' \in \mathbb{M}_k} V_{M'}^{\pi}.$$

For each episode $k$, the plausible MDP set $\mathbb{M}_k$ is defined by

$$\mathbb{M}_k = \Bigg\{ (r', Q') : \forall a, x_a, |r'(x_a) - \hat{r}_{k-1}(x_a)| \leq \frac{L_{k-1}}{2\sqrt{\max\{1, N_{k-1}(x_a)\}}}, \text{ and}$$
$$\left\| Q'(x_a, \cdot) - \hat{Q}_{k-1}(x_a, .) \right\|_1 \leq \frac{L_{k-1} + 1.5\sqrt{S}}{\sqrt{\max\{1, N_{k-1}(x_a)\}}} \Bigg\}. \tag{33}$$

As Auer et al. (2008), we argue that there exists a MDP $M_k \in \mathbb{M}_k$ such that $\pi_k$ is an optimal policy for $M_k$. Moreover, under event $\mathcal{E}_{k-1}^{\text{UCRL2}}$, one has $M \in \mathbb{M}_k$, which implies that $\max_{\pi} \max_{M' \in \mathbb{M}_k} V_{M'}^{\pi}(\boldsymbol{x}) \geq V_M^{\pi_*}(\boldsymbol{x})$. By Equation 14, we get $\mathbb{E}\left[\Delta_k^{model}\right] \leq 0$. If $\mathcal{E}_{k-1}^{\text{UCRL2}}$ does not hold, we simply have $\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCRL2}}\}} = 0$. We conclude that: $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right] \leq 0$.

#### A.4.3 Analysis of $\mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right]$ for MB-UCRL2

Following Equation 13, the expected regret can be written as:

$$\mathbb{E}\left[\text{Reg}(K, \text{MB-UCRL2}, M)\right] = \sum_{k=1}^{K} \mathbb{E}\left[\Delta_k\right] \leq \sum_{k=1}^{K} \mathbb{E}\left[H_k\right] \mathbb{P}\left(\neg\mathcal{E}_{k-1}^{\text{UCRL2}}\right) + \mathbb{E}\left[\Delta_k \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right]$$
$$\leq \frac{5}{1-\beta} + \sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\neg\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right] + \mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\neg\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right] \tag{34}$$

where the last inequality holds due to Lemma 7. By the previous section, the second term of Equation 34 is non-positive. In the following, we therefore analyze the last term whose analysis is then similar to the one for MB-PSRL. Indeed, with $B_k = H_k$ and definition of $\mathcal{E}_{k-1}^{\text{UCRL2}}$, the use of Lemma 2 shows that one has

$$\mathbb{E}\left[\Delta_k^{\text{conc}}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCRL2}}\}}\right] \leq \mathbb{E}\left[\frac{1}{2} \sum_{t=t_k}^{t_{k+1}-1} \frac{L_{k-1} + (2L_{k-1} + 3\sqrt{S})H_k}{\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}}\right].$$

Up to a factor 1/2, the expression inside the expectation is the same as Equation 30 of MB-PSRL. Hence, one can use Lemma 3 the same way to show that

$$\sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{conc}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{PSRL}}\}}\right] \leq \frac{3}{2}(L_K+\sqrt{S})\left(Sn\mathbb{E}\left[\max_{k\leq K} H_k^2\right] + 2\sqrt{SnK}\mathbb{E}\left[\max_{k\leq K} H_k^{3/2}\right]\right).$$

Up to a factor 1/2, the right term of the above equation is equal to the right term of Equation 32. Following the same process done for the later, we can conclude that there exists a constant $C'$ independent of all problem's parameters such that:

$$\text{Reg}(K, \text{MB-UCRL2}, M) \leq C'\left(\sqrt{S}+\log\left(\frac{SnK\log K}{1-\beta}\right)\right)\left(Sn\left(\frac{\log K}{1-\beta}\right)^2 + \sqrt{SnK}\left(\frac{\log K}{1-\beta}\right)^{3/2}\right)$$

### A.5 Case of MB-UCBVI

We start by defining the high probability event. Then, we prove the optimistic property of MB-UCBVI. Finally, we bound its expected regret.

#### A.5.1 Definition of the high-probability event

**Lemma 8.** *The event*

$$\mathcal{E}_{k-1}^{UCBVI} = \Bigg\{\forall a\in[n], \boldsymbol{x}\in\mathcal{X}, k' \leq k-1 \colon |\hat{r}_{k'}(x_a) - r(x_a)| \leq \frac{L_{k-1}}{2\sqrt{\max\{1, N_{k'}(x_a)\}}},$$
$$\left\|\hat{Q}_{k'}(x_a,\cdot) - Q(x_a,\cdot)\right\|_1 \leq \frac{L_{k-1}+1.5\sqrt{S}}{\sqrt{\max\{1, N_{k'}(x_a)\}}}, H_{k'} \leq \frac{\log(K(k-1))}{1-\beta},$$
$$\text{and } \left|\hat{r}_{k'}(x_a) - r(x_a) + \beta\sum_{\boldsymbol{y}}(\hat{P}_{k'}^a(\boldsymbol{x},\boldsymbol{y}) - P^a(\boldsymbol{x},\boldsymbol{y}))V_M^{\pi_*}(\boldsymbol{y})\right| \leq \frac{L_{k-1}}{2(1-\beta)\sqrt{\max\{1, N_{k'}(x_a)\}}}\Bigg\}$$

*is $\mathcal{O}_{k-1}$-measurable and true with probability at least $1-7/K$.*

*Proof.* The complement of $\mathcal{E}_{k-1}^{\text{UCBVI}}$ is the union of $\neg\mathcal{E}_{k-1}^r$, $\neg\mathcal{E}_{k-1}^Q$, $\neg\mathcal{E}_{k-1}^H$ and $\neg\mathcal{E}_{k-1}^V$. We conclude the proof by using the union bound and $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^r\right) \leq 2/K$, $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^Q\right) \leq 2/K$, $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^V\right) \leq 2/K$ and $\mathbb{P}\left(\neg\mathcal{E}_{k-1}^H\right) \leq 1/K$ □

#### A.5.2 Analysis of $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right]$ – Optimism of MB-UCBVI

The following lemma guarantees that $\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right] \leq 0$. Indeed, as $\mathcal{E}_{k-1}^{\text{UCBVI}}$ is $\mathcal{O}_{k-1}$-measurable, one has

$$\begin{aligned}\mathbb{E}\left[\Delta_k^{model}\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\Delta_k^{model} \mid \mathcal{O}_{k-1}\right]\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right] \\ &= \mathbb{E}\left[(V_M^{\pi_*}(\boldsymbol{X}_{t_k}) - V_{M_k}^{\pi_k}(\boldsymbol{X}_{t_k}))\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right] \leq 0.\end{aligned}$$

**Lemma 9.** *If $\mathcal{E}_{k-1}^{UCBVI}$ is true, then, for any $\boldsymbol{x}\in\mathcal{X}$, we have*

$$V_{M_k}^{\pi_k}(\boldsymbol{x}) \geq V_M^{\pi_*}(\boldsymbol{x})$$

*Proof.* Recall that at episode $k$, we define the optimistic MDP of MB-UCBVI by $M_k$ in which the parameters of any arm $a\in[n]$ are $(\hat{\boldsymbol{r}}_{k-1}^a + b_{k-1}^a, \hat{Q}_{k-1}^a)$ with $b_{k-1}(x_a) = \frac{L_{k-1}}{2(1-\beta)\sqrt{\max\{1, N_{k-1}(x_a)\}}}$ for any $x_a \in \mathcal{S}^a$. The

Gittins index policy $\pi_k$ is optimal for MDP $M_k$. For any state $\boldsymbol{x}$, let $a = \pi_k(\boldsymbol{x})$ and $a_* = \pi_*(\boldsymbol{x})$. Then,

$$\begin{aligned}
V_{M_k}^{\pi_k}(\boldsymbol{x}) - V_M^{\pi_*}(\boldsymbol{x}) &= b_{k-1}(x_a) + \hat{r}_{k-1}(x_a) + \beta \sum_{\boldsymbol{y}} \hat{P}_{k-1}^{a}(\boldsymbol{x},\boldsymbol{y}) V_{M_k}^{\pi_k}(\boldsymbol{y}) - V_M^{\pi_*}(\boldsymbol{x}) \\
&\geq b_{k-1}(x_{a_*}) + \hat{r}_{k-1}(x_{a_*}) + \beta \sum_{\boldsymbol{y}} \hat{P}_{k-1}^{a_*}(\boldsymbol{x},\boldsymbol{y}) V_{M_k}^{\pi_k}(\boldsymbol{y}) \\
&\quad - r(x_{a_*}) - \beta \sum_{\boldsymbol{y}} P^{a_*}(\boldsymbol{x},\boldsymbol{y}) V_M^{\pi_*}(\boldsymbol{y}) \\
&= b_{k-1}(x_{a_*}) + \hat{r}_{k-1}(x_{a_*}) - r(x_{a_*}) + \beta \sum_{\boldsymbol{y}} (\hat{P}_{k-1}^{a_*}(\boldsymbol{x},\boldsymbol{y}) - P^{a_*}(\boldsymbol{x},\boldsymbol{y})) V_M^{\pi_*}(\boldsymbol{y}) \\
&\quad + \beta \sum_{\boldsymbol{y}} \hat{P}_{k-1}^{a_*}(\boldsymbol{x},\boldsymbol{y}) (V_{M_k}^{\pi_k}(\boldsymbol{y}) - V_M^{\pi_*}(\boldsymbol{y}))
\end{aligned}$$

In matrix form, we have

$$V_{M_k}^{\pi_k} - V_M^{\pi_*} \geq b_{k-1}^{\pi_*} + \hat{r}_{k-1}^{\pi_*} - r^{\pi_*} + \beta(\hat{P}_{k-1}^{\pi_*} - P^{\pi_*}) V_M^{\pi_*} + \beta \hat{P}_{k-1}^{\pi_*} (V_{M_k}^{\pi_k} - V_M^{\pi_*})$$

Under event $\mathcal{E}_{k-1}^{\text{UCBVI}}$, $b_{k-1}^{\pi_*} + \hat{r}_{k-1}^{\pi_*} - r^{\pi_*} + \beta(\hat{P}_{k-1}^{\pi_*} - P^{\pi_*}) V_M^{\pi_*} \geq 0$. This implies that:

$$(I - \beta \hat{P}_{k-1}^{\pi_*})(V_{M_k}^{\pi_k} - V_M^{\pi_*}) \geq 0.$$

As $(I - \beta \hat{P}_{k-1}^{\pi_*})^{-1} = I + (I - \beta \hat{P}_{k-1}^{\pi_*}) + (I - \beta \hat{P}_{k-1}^{\pi_*})^2 + \dots$ is a matrix whose coefficients are all non-negative, this implies that $V_{M_k}^{\pi_k} - V_M^{\pi_*} \geq 0$. □

#### A.5.3 Analysis of $\mathbb{E}\left[\Delta_k^{conc} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right]$ for MB-UCBVI

Following Equation 13, the expected regret can be written similarly to Equation 34 for MB-UCRL2, one can write that

$$\mathbb{E}\left[\text{Reg}(K, \text{MB-UCBVI}, M)\right] \leq \frac{7}{1-\beta} + \sum_{k=1}^{K} \mathbb{E}\left[\Delta_k^{model} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right] + \mathbb{E}\left[\Delta_k^{conc} \mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}}\right].$$

The same as MB-UCRL2, the second term is non-positive. We are therefore left with the last term. Using Lemma 2 with $B_k = \frac{H_k L_{k-1}}{2(1-\beta)}$ and the definition of $M_k$ for MB-UCBVI, we have:

$$\begin{aligned}
\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}} \Delta_k^{\text{conc}}\right] &\leq \sum_{k=1}^{K} \mathbb{E}\Bigg[\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}} \sum_{t=t_k}^{t_{k+1}-1} |b_{k-1}(X_{t,A_t}) + \hat{r}_{k-1}(X_{t,A_t}) - r(X_{t,A_t})| \\
&\qquad + \frac{H_k L_{k-1}}{2(1-\beta)} \left\| \hat{Q}_{k-1}(X_{t,A_t}, \cdot) - Q(X_{t,A_t}, \cdot) \right\|_1 \Bigg] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \frac{(2-\beta)L_{k-1} + H_k L_{k-1}(L_{k-1} + 1.5\sqrt{S})}{2(1-\beta)\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}}\right] \\
&\leq \mathbb{E}\left[\frac{2L_K(L_K + \sqrt{S}) \max_{k \leq K} H_k}{1-\beta} \sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \frac{1}{\sqrt{\max\{1, N_{k-1}(X_{t,A_t})\}}}\right] \\
&\leq \mathbb{E}\left[\frac{2L_K(L_K + \sqrt{S}) \max_{k \leq K} H_k}{1-\beta} \left(Sn \max_{k \leq K} H_k + 2\sqrt{SnK \max_{k \leq K} H_k}\right)\right]
\end{aligned}$$

where the second inequality holds due to the definition of $\mathcal{E}_{k-1}^{\text{UCBVI}}$ and the last one holds due to Lemma 3. With $L_K = \sqrt{2 \log \frac{4SnK^2 \log K}{1-\beta}}$, we have

$$\sum_{k=1}^{K} \mathbb{E}\left[\mathbb{I}_{\{\mathcal{E}_{k-1}^{\text{UCBVI}}\}} \Delta_k^{\text{conc}}\right] \leq \frac{2(1+\sqrt{S})}{1-\beta} 2 \log\left(\frac{4SnK^2 \log K}{1-\beta}\right) \left(Sn\mathbb{E}\left[\max_{k \leq K} H_k^2\right] + 2\sqrt{SnK}\mathbb{E}\left[\max_{k \leq K} H_k^{3/2}\right]\right)$$

The last term of the right side above can be analyzed exactly the same as what is done for Equation 32 using Lemma 4. This concludes the proof.

## B Proof of Theorem 2

To prove the lower bound, we consider a specific markovian bandit problem that is composed of $S$ independent *stochastic bandit problems*. This allows us to reuse the existing minimax lower bound for stochastic bandit problems. This existing result can be stated as follows: let $\mathcal{L}^{\text{stoc.pb}}$ be a learning algorithm for the stochastic bandit problem. It is shown in Theorem 3.1 of Bubeck et al. (2012) that for any number of arms $n$ and any number of time steps $\tau$, there exists parameters for a stochastic bandit problem $M^{\text{stoc.pb}}$ with $n$ arms such that the regret of the learning algorithm over $\tau$ time steps is at least $(1/20)\sqrt{n\tau}$.

$$\text{Reg}^{\text{stoc.pb}}(\tau, \mathcal{L}^{\text{stoc.pb}}, M^{\text{stoc.pb}}) \geq \frac{1}{20}\sqrt{n\tau}. \tag{35}$$

This lower bound (Theorem 3.1 of Bubeck et al. (2012)) is constructed by considering $n$ stochastic bandit problems $M^{\text{stoc.pb},j}$ for $j \in [n]$ with parameters that depend on $\tau$ and $n$. In the problem $M^{\text{stoc.pb},j}$, all arms have a reward $\gamma(\tau, n)$ except arm $j$ that has a reward $\gamma'(\tau, n) > \gamma(\tau, n)$. It is shown in Theorem 3.1 of Bubeck et al. (2012) that a learning algorithm cannot perform uniformly well on all problems because it is impossible to distinguish them *a priori*. More precisely, in the proof of Lemma 3.2 of Bubeck et al. (2012), it is shown that if the best arm is chosen at random, then the expected (bayesian) regret of any learning algorithm is at least $(1/20)\sqrt{n\tau}$.

As for our problem, let $K$ be a number of episodes, $\beta$ a discount factor, $n$ a number of arms, $S$ a number of states per arm and set $\tau = K/(2S(1-\beta))$. We consider a random markovian bandit model $M$ constructed as follows. Each arm $a$ has $S$ states with the state space $\mathcal{S}^a = \{1_a, 2_a, \dots, S_a\}$. The transition matrix $Q_a$ is the identity matrix. For each state $i \in \{1 \dots S\}$, we choose the best arm $a_i^*$ uniformly at random among the $n$ arms, independently for each $i$. The rewards of a state $i_a$ are *i.i.d.* Bernoulli rewards with mean $\gamma(\tau, n)$ if $a \neq a_i^*$ and $\gamma'(\tau, n)$ if $a = a_i^*$. The initial distribution $\rho$ couples the initial states of all arms for all $i \in \{1 \dots S\}$,

$$\mathbb{P}\left(\forall a \in [n] : x_{0,a} = i_a\right) = \frac{1}{S}.$$

In this case, the markovian bandit problem becomes a combination of $S$ independent stochastic bandit problems with $n$ arms each. We denote by $M_i^{\text{stoc.pb}}$ the random stochastic bandit problem for the initial state $\boldsymbol{i} = (i_a)_{a \in [n]}$. As the $a_i^*$ are chosen independently, a learning algorithm $\mathcal{L}$ cannot use the information for $M_i^{\text{stoc.pb}}$ to perform better on $M_j^{\text{stoc.pb}}$, $j \neq i$.

Let $\phi$ be the distribution of the random markovian bandit model $M$ defined above and let $T_i$ be the number of time steps spent in state $\boldsymbol{i}$ by the learning algorithm $\mathcal{L}$.

$$\begin{aligned}
\text{BayReg}(K, \mathcal{L}, \phi) &\geq \sum_{i=1}^{S} \mathbb{E}\left[\text{Reg}^{\text{stoc.pb}}(T_i, \mathcal{L}_i^{\text{stoc.pb}}, M_i^{\text{stoc.pb}})\right] \\
&\geq \sum_{i=1}^{S} \mathbb{E}\left[\text{Reg}^{\text{stoc.pb}}(\tau, \mathcal{L}_i^{\text{stoc.pb}}, M_i^{\text{stoc.pb}})\mathbb{I}_{\{T_i \geq \tau\}}\right] && (36) \\
&\geq \frac{S}{20}\sqrt{n\tau}\,\mathbb{P}\left(T_i \geq \tau\right) && (37) \\
&= \frac{1}{20\sqrt{2}}\mathbb{P}\left(T_i \geq \tau\right)\sqrt{\frac{SnK}{1-\beta}}, && (38)
\end{aligned}$$

where Equation 36 is true because the expected regret is non-decreasing function of the number of episodes, Equation 37 comes from Equation 35 and Equation 38 from the definition of $\tau$.

We show in the Lemma 10 below that $\mathbb{P}\left(T_i \leq K/(2S(1-\beta))\right) \leq 8S/K$. This shows that for $K \geq 16S$, one has $\mathbb{P}\left(T_i \geq \tau\right) \geq 1/2$. This concludes the proof as $40\sqrt{2} \leq 60$.

**Lemma 10.** *Recall that $T_i$ is the number of time steps that the MDP is in state $\boldsymbol{i}$ for the MDP model above. Let $G_k$ be a sequence of* i.i.d. *Bernoulli random variable of mean $1/S$ and let $H_k$ be an independent* i.i.d. *sequence of geometric random variable of parameter $1-\beta$. Then:*

*(i)* $T_i \sim \sum_{k=1}^{K} G_k H_k$,

*(ii)* $\mathbb{E}[T_i] = K/(S(1-\beta))$,

*(iii)* $\mathbb{P}(T_i \geq \mathbb{E}[T_i]/2) \geq 1 - 8S/K$.

*Proof.* Let $G_k$ be a random variable that equals 1 if the initial state $\boldsymbol{i}$ is chosen at the beginning of episode $k$ and recall that $H_k$ is the episode length. By definition, the variables $G_k$ and $H_k$ are independent and follow respectively Bernoulli and geometric distribution. This shows *(i)*.

Let $W_k = G_k H_k$. As the $W_k$ are *i.i.d.* and $G_k$ and $H_k$ are independent, we have:

$$\mathbb{E}[T_i] = K\mathbb{E}[H_1 G_1] = \frac{K}{S(1-\beta)}.$$

This shows (ii).

Moreover, $\text{var}[T_i] = K\text{var}[H_1 G_1]$. Hence, by using Chebyshev's inequality, one has:

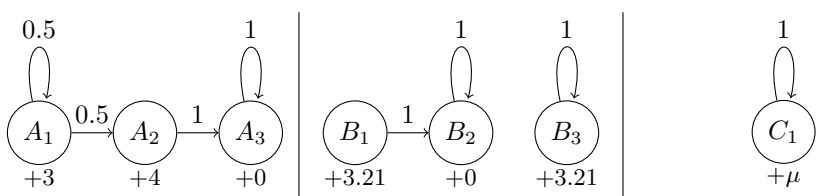
$$\begin{aligned}\mathbb{P}\left(T_i \leq \frac{\mathbb{E}[T_i]}{2}\right) &\leq \mathbb{P}\left(|T_i - \mathbb{E}[T_i]| \geq \frac{\mathbb{E}[T_i]}{2}\right)\\ &\leq \frac{4\text{var}[T_i]}{(\mathbb{E}[T_i])^2}\\ &= \frac{4}{K}\frac{\text{var}[H_1 G_1]}{(\mathbb{E}[H_1 G_1])^2}.\end{aligned}$$

Concerning the variance, the second moment of a geometric random variable of parameter $1-\beta$ is $(1+\beta)/(1-\beta)^2$. This shows that $\mathbb{E}\left[(H_1 G_1)^2\right] = (1+\beta)/(S(1-\beta)^2) \leq 2S(\mathbb{E}[H_1 G_1])^2$. This implies:

$$\text{var}[H_1 G_1] \leq (2S-1)(\mathbb{E}[H_1 G_1])^2 \leq 2S(\mathbb{E}[H_1 G_1])^2.$$

This implies (iii). □

## C Proof of Theorem 3




(a) $\hat{Q}_a$ and $\hat{r}_a = r_a$. (b) $\hat{Q}_b = Q_b$ and $\hat{r}_b = r_b$. (c) $\hat{Q}_c = Q_c$ and $\hat{r}_c = r_c$.

Figure 2: Counterexample for OFU indices: $\hat{\mathcal{B}}_a$, $\hat{\mathcal{B}}_b = \mathcal{B}_b$, $\hat{\mathcal{B}}_c = \mathcal{B}_c$.

In this proof, we reason by contradiction and assume that there exists a procedure that computes local indices such that the obtained policy is such that for any estimate $\hat{\mathcal{B}}$ and any initial condition $\rho$, then if $M \in \mathbb{M}(\hat{\mathcal{B}})$, one has

$$\sup_{M \in \mathbb{M}(\hat{\mathcal{B}})} V_M^{\pi^{I^{\hat{\mathcal{B}}}}}(\rho) \geq \sup_{\pi} V_M^{\pi}(\rho). \tag{39}$$

In the remaining of this section, we set the discount factor to $\beta = 0.5$. For a given state $x_a$, we denote by $I(x_a)$ the local index of state $x_a$ computed by this hypothetically optimal algorithm.

We first consider a markovian bandit problem with two arms $\{b, c\}$. We consider that these two arms are perfectly estimated (*i.e.*, $\epsilon_b^r(x_b) = \epsilon_b^Q(x_b) = \epsilon_c^r(x_c) = \epsilon_c^Q(x_c) = 0$). The Markov chains for these arms are depicted in Figure 2. Their transitions matrices and rewards are

$$Q_b = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } r_b = [3.21, 0, 3.21]; \qquad Q_c = [1] \text{ and } r_c = [\mu].$$

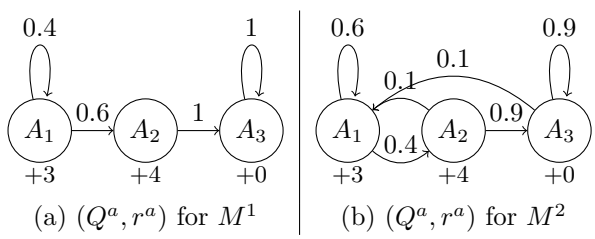
As the markovian bandit are perfectly known, the indices $I(B_1)$, $I(B_2)$, $I(B_3)$ and $I(C_1)$ must be such that the obtained priority policy is optimal for the true MDP, that is: states $B_1$ and $B_3$ should have priority over $C_1$ (*i.e.*, $I(B_1) > I(C_1)$ and $I(B_3) > I(C_1)$) if and only if $\mu < 3.21$, and state $B_2$ should have priority over $C_1$ (*i.e.*, $I(B_2) > I(C_1)$) if and only if $\mu < 0$. This implies that the local indices defined by our hypothetically optimal algorithm must satisfy

$$I(B_1) = I(B_3) > I(B_2).$$

Now, we consider markovian bandit problems with two arms $\{a, b\}$, where Arm $b$ is as before. For Arm $a$, we consider a confidence set $\hat{\mathcal{B}}_a = (\hat{Q}_a, \hat{r}_a, \epsilon_a^r, \epsilon_a^Q)$ where $(\hat{Q}_a, \hat{r}_a)$ are depicted in Figure 2(a) and where $\epsilon_a^r(x_a) = 0$ and $\epsilon_a^Q(x_a) = 0.2$:

$$\hat{Q}_a = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \hat{\boldsymbol{r}}_a = \boldsymbol{r}_a = [3, 4, 0] \qquad \epsilon_a^Q = [0.1, 0.1, 0.1] \text{ and } \epsilon_a^r = [0, 0, 0].$$

We consider two possible instances of the "true" markovian bandit problem, denoted $M^1$ and $M^2$. For $M^1$, the transition matrix and reward function of the first arm are depicted in Figure 3(a). For $M^2$, they are depicted in Figure 3(b). In both cases, $(Q_b, r_b)$ are as in Figure 2(b). It should be clear that $M^1 \in \mathbb{M}$ and $M^2 \in \mathbb{M}$.




(a) $(Q^a, r^a)$ for $M^1$ (b) $(Q^a, r^a)$ for $M^2$

Figure 3: The two instances of $\mathcal{B}_a^1$ and $\mathcal{B}_a^2$

If there exist indices that can be computed locally, then the indices for an arm should not depend on the confidence that one has on the other arms. The indices $I(A_1)$, $I(A_2)$ and $I(A_3)$ must satisfy the following facts:

- $I(A_3) \in (I(B_2), I(B_3))$ because for all markovian bandit $M \in \mathbb{M}$, state $A_3$ should have priority over state $B_2$ and should not have priority over state $B_3$ (because of the discount factor $\beta = 1/2$).
- $I(A_2) > I(B_1) = I(B_3)$ because for all markovian bandit $M \in \mathbb{M}$, state $A_2$ will give a higher instantaneous reward than state $B_1$ or $B_3$. It should therefore have a higher priority.

This leaves two possibilities for $I(A_1)$:

- If $I(A_1) > I(B_1) = I(B_3)$, then state $A_1$ has priority over both $B_1$ and $B_3$. We denote the corresponding priority policy $\pi^1$.

- If $I(A_1) < I(B_1) = I(B_3)$, then state $B_1$ and $B_3$ have a higher priority than state $A_1$. We denote the corresponding priority policy by $\pi^2$.

We use a numerical implementation of extended value iteration (available in the supplementary material) to find that:

$$\sup_{M\in\mathbb{M}} V_M^{\pi^2}(A_1, B_3) \approx 6.42 < \sup_{\pi} V_{M^1}^{\pi}(A_1, B_3) \approx 6.47 \tag{40}$$
$$\sup_{M\in\mathbb{M}} V_M^{\pi^1}(A_1, B_1) \approx 5.96 < \sup_{\pi} V_{M^2}^{\pi}(A_1, B_1) \approx 6.00$$

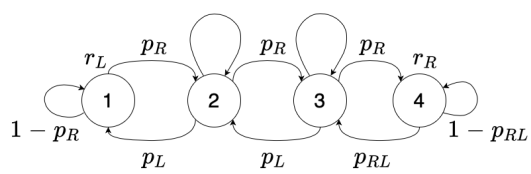
This implies that there does not exist any definition of indices such that Equation 11 holds regardless of $M$ and $\boldsymbol{x}$.

## D Description of the Algorithms and Choice of Hyperparameter

In this section, we provide a detailed description of the simulation environment used in the paper. We first describe the Markov chain used in our example. Then, we describe all algorithms that we compare in the paper. For each algorithm, we give some details about our choice of hyperparameters. Last, we also describe the experimental methodology that we used in our simulations.

### D.1 Description of the example

We design an environment with 3 arms, all following a Markov chain represented in Table 1. This Markov chain is obtained by applying the optimal policy on the river swim MDP of Filippi et al. (2010). In each chain, there are 2 rewarding states: state 1 with low mean reward $r_L$, and state 4) with high mean reward $r_R$, both with Bernoulli distributions. At the beginning of each episode, all chains start in their state 1. Each chain is parametrized by the values of $p_L, p_R, p_{RL}, r_L, r_R$ that are given in Table 1 along with the corresponding Gittins indices of each chain.




| | | | | | Gittins index for each state | | | |
|---|---|---|---|---|---|---|---|---|
| $p_L$ | $p_R$ | $p_{RL}$ | $r_L$ | $r_R$ | 1 | 2 | 3 | 4 |
| 0.1 | 0.2 | 0.3 | 0.2 | 1.0 | 0.276 | 0.2894 | 0.392 | 1.0 |
| 0.1 | 0.5 | 0.7 | 0.35 | 0.7 | 0.35 | 0.256 | 0.2892 | 0.7 |
| 0.1 | 0.4 | 0.5 | 0.4 | 0.65 | 0.4 | 0.250 | 0.286 | 0.65 |

Table 1: The random walk chain with 4 states. In state 4, the chain has an average reward $r_R$. For state 2 and 3, the chain gives zero reward. In state 1, the mean reward is $r_L$. This chain is obtained by applying the optimal policy on the 4-state river swim MDP of Filippi et al. (2010). The table contains the parameters that we used, along with Gittins indices of all states when the discount factor is $\beta = 0.99$.

### D.2 MB-PSRL

MB-PSRL, the adaption from PSRL, puts prior distribution on the parameters $(\boldsymbol{r}^a, Q^a)$ of each Arm $a$, draws a sample from the posterior distribution and uses it to compute the Gittins indices at the start of each episode. We implement two posterior updates for the mean reward vector $\boldsymbol{r}^a$: Beta and Gaussian-Gamma. The second posterior, Gaussian-Gamma, will be used in prior choice sensitivity tests. For the transition matrix $Q^a$, we implemented Dirichlet posterior update because Dirichlet distribution is the only natural conjugate prior for categorical distribution. Beta, Gaussian-Gamma and Dirichlet distributions can be easily sampled using the numpy package of Python. This greatly contributes to the computational efficiency of MB-PSRL.

We give more details on this prior distribution and their conjugate posterior in the subsections below.

#### D.2.1 Bayesian Updates: Conjugate Prior and Posterior Distributions

MB-PSRL is a bayesian learning algorithm. As such, it samples reward vectors and transition matrices at the start each episode. We would like to emphasize that neither the definition of the algorithm nor its performance guarantees that we prove in Theorem 1 depend on a specific form of the prior distribution $\phi$. Yet, in practice, some prior distributions are more preferable because their conjugate distributions are easy to implement. In the following, we give concrete examples on how to update the conjugate distribution given the observations.

For $a \in [n]$ and $x_a \in \mathcal{S}^a$, let $N_{k-1}(x_a)$ be the number of activations of arm $a$ while in state $x_a$ up to episode $k$. For this state $x_a$, the number of samples of the reward and of transitions from $x_a$ are equal to $N_{k-1}(x_a)$. To ease the exposition, we drop the label $a$ and assume that we are given:

- $N_{k-1}(x)$ *i.i.d.* samples $\{Y_1, \dots, Y_{N_{k-1}(x)}\}$ of next states to which the arm transitioned from $x$.
- $N_{k-1}(x)$ *i.i.d.* samples $\{R_1, \dots, R_{N_{k-1}(x)}\}$ of random immediate rewards earned while the arm was activated in state $x$

Each $Y_i$ is such that $\mathbb{P}(Y_i = y) = Q(x, y)$ and each $R_i$ is such that $\mathbb{E}[R_i] = r(x)$. In what follows, we describe natural priors that can be used to estimate the transition matrix and the reward vector.

#### D.2.2 Transition Matrix

If no information is known about the arm, the natural prior distribution is to consider the lines $Q(x, \cdot)$ of the matrix as independent multivariate random variables uniformly distributed among all non-negative vectors of length $S$ that sum to 1. This corresponds to a Dirichlet distribution of parameters $\alpha = (1, \dots, 1)$. For a given $x$, the variables $\{Y_1, \dots, Y_{N_{k-1}(x)}\}$ are generated according to a categorical distribution $Q(x, \cdot)$. The Dirichlet distribution is self-conjugate with respect to the likelihood of a categorical distribution. So, the posterior distribution $\phi(Q(x, \cdot)|Y_1, \dots, Y_{N_{k-1}(x)})$ is a Dirichlet distribution with parameters $\mathbf{c} = (c_1 \dots c_S)$ where $c_y = 1 + \sum_{i=1}^{N_{k-1}(x)} \mathbb{I}_{\{Y_i = y\}}$.

#### D.2.3 Reward Distribution

As for the reward vector, the choice of a good prior depends on the distribution of rewards. We consider two classical examples: Bernoulli and Gaussian.

**Bernoulli distribution** A classical case is to assume that the reward distribution of a state $x$ is Bernoulli with mean value $r(x)$. A classical prior in this case is to consider that $\{r(x)\}_{\{x \in \mathcal{S}\}}$ are *i.i.d.* random variables following a uniform distribution whose support is $[0, 1]$. The posterior distribution of $r(x)$ at time $t$ is the distribution of $r(x)$ conditional to the reward observations from state $x$ gathered up to time $t$. The posterior distribution $\phi(r(x) \mid R_1, \dots, R_{N_{k-1}(x)})$ is then a Beta distribution with parameters $(1 + \sum_{i=1}^{N_{k-1}(x)} \mathbb{I}_{\{R_i = 1\}}, 1 + \sum_{i=1}^{N_{k-1}(x)} \mathbb{I}_{\{R_i = 0\}})$. Recall that the Beta distribution is a special case of the Dirichlet distribution in the same way as the Bernoulli distribution is a special case of the Categorical distribution.

**Gaussian distribution** We now consider the case of Gaussian rewards and we assume that the immediate rewards earned in state $x$ are *i.i.d.* Gaussian random variables of mean and variance $(r(x), \sigma^2(x))$. A natural prior for Gaussian rewards is to consider that $\{(r(x), \frac{1}{\sigma^2(x)})\}_{\{x \in \mathcal{S}\}}$ are *i.i.d.* bivariate random variables where the marginal distribution of each $\frac{1}{\sigma^2(x)}$ is a Gamma distribution (it is a natural belief since the empirical variance of Gaussian has a chi-square distribution which is a special case of Gamma distribution). Conditioned on $\frac{1}{\sigma^2(x)}$, $r(x)$ follows a Gaussian distribution of variance $\sigma^2(x)$. We say that $(r(x), \frac{1}{\sigma^2(x)})$ has a Gaussian-Gamma distribution, which is self-conjugate with respect to a Gaussian likelihood (*i.e.,* the likelihood of Gaussian rewards). So, given the reward observations, the marginal distribution of $\frac{1}{\sigma^2(x)}$ is still a Gamma distribution. $r(x)$ has Gaussian distribution conditioned on the reward observations and $\frac{1}{\sigma^2(x)}$.

Indeed, let $\hat{r}(x) = \frac{1}{N_{k-1}(x)} \sum_{i=1}^{N_{k-1}(x)} R_i$ and $\hat{\sigma}^2(x) = \frac{1}{N_{k-1}(x)} \sum_{i=1}^{N_{k-1}(x)} (R_i - \hat{r}(x))^2$ be the empirical mean and empirical variance of $R_i$. Then it can be shown that the posterior distribution of $\frac{1}{\sigma^2(x)}$ and $r(x)$ are:

$$\frac{1}{\sigma^2(x)} \mid R_1, \ldots, R_{N_{k-1}(x)} \sim \text{Gamma}\left(\frac{N_{k-1}(x)+1}{2}, \frac{1}{2} + \frac{N_{k-1}(x)\hat{\sigma}^2(x)}{2} + \frac{N_{k-1}(x)\hat{r}^2(x)}{2(N_{k-1}(x)+1)}\right)$$
$$r(x) \mid \frac{1}{\sigma^2(x)}, R_1, \ldots, R_{N_{k-1}(x)} \sim \mathcal{N}\left(\frac{N_{k-1}(x)\hat{r}(x)}{N_{k-1}(x)+1}, \frac{\sigma^2(x)}{N_{k-1}(x)+1}\right).$$

For more details about the analysis of conjugate prior and posterior presented above as well as more conjugate distributions, we refer the reader to Fink (1997); Murphy (2007).

Notice that a reward that has a Gaussian distribution violates the property that all rewards are in $[0, 1]$. This could invalidate the bound on the regret of our algorithm proven in Theorem 1. Actually, it is possible to correct the proof to cover the Gaussian case by replacing the Hoeffding's inequality used in Lemma 1 by a similar inequality, also valid for sub-Gaussian random variables, see Vershynin (2018). In the experimental section (see E.3), we also show that a bad choice for the prior distribution of the reward (assuming a Gaussian distribution while the rewards are actually Bernoulli) does not alter too much the performance of the learning algorithm.

### D.3 Experimental Methodology

In our numerical experiment, we did 3 scenarios to evaluate the algorithms (scenario 2 and 3 are given in Appendix E). In each scenario, we choose the discount factor $\beta = 0.99$ (which is classical) and we compute the regret over $K = 3000$ episodes. The number of simulations varies over scenario depending on how the regret is computed. For each run, we draw a sequence of horizons $\{H_k\}_{k\in[3000]}$ from a geometric distribution of parameter 0.01 and we run all algorithms for this sequence of time-horizons to remove a source of noise in the comparisons.

For a given sequence of policies $\pi_k$, following Equation 4, the expected regret is $\mathbb{E}\left[\sum_{k=1}^{K} \Delta_k(\boldsymbol{X}_{t_k})\right]$ where $\Delta_k(\boldsymbol{X}_{t_k})$ is the expected regret over episode $k$. To reduce the variance in the numerical experiment, we compute $\Delta_k(\boldsymbol{X}_{t_k}) = V_M^{\pi_*}(\boldsymbol{X}_{t_k}) - V_M^{\pi_k}(\boldsymbol{X}_{t_k})$. For a given markovian bandit problem and state $\boldsymbol{x}$, the value $V_M^{\pi_*}(\boldsymbol{x})$ can be computed by using the retirement evaluation presented in Page 272 of Whittle (1996). It seems, however, that the same methodology is not applicable to compute the value function of an index policy that is not the Gittins policy. This means that while the policy $\pi_k$ is easily computable, we do not know of an efficient algorithm to compute its value $V_M^{\pi_k}(\boldsymbol{x})$. Hence, in our simulations, we will use two methods to compute the regret, depending on the problem size:

1. (Exact method) Let $(r^\pi, P^\pi)$ be the reward vector and transition matrix under policy $\pi$ (i.e. $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{E}, r^\pi(\boldsymbol{x}) = r(\boldsymbol{x}, \pi(\boldsymbol{x})), P^\pi(\boldsymbol{x}, \boldsymbol{y}) = P^{\pi(\boldsymbol{x})}(\boldsymbol{x}, \boldsymbol{y})$). Using the Bellman equation, the value function under policy $\pi$ is computed by
$$V_M^\pi = (\mathbf{1} - \beta P^\pi)^{-1} r^\pi. \tag{41}$$
The matrix inversion can be done efficiently with the numpy package of Python. However, this takes $S^{2n} + 2S^n$ of memory storage. Hence, when the number of states and arms are too large, the exact computation method cannot be performed.

2. (Monte Carlo method) In Scenario 2, the model has $n = 9$ arms with $S = 11$ states each, which makes the exact method inapplicable. In this case, it is still possible to compute the optimal policy and to apply Gittins index based algorithms but computing their value is intractable. In such a case, to measure the performance, we do 240 simulations for each algorithm and try to approximate $\Delta_k$ by
$$\hat{\Delta}_k = \frac{1}{\#\text{replicas}} \sum_{j=1}^{\#\text{replicas}} \sum_{t=0}^{H_k^{(j)}-1} \left[r(X^{*,(j)}_{t,A_t^{*,(j)}}) - r(X^{(j)}_{t,A_t^{(j)}})\right], \tag{42}$$

where $H_k^{(j)}$ is the horizon of the $k$th episode of the $j$th simulation and $\{X^{*,(j)}_{t,A^{*,(j)}_t}\}$ and $\{X^{(j)}_{t,A^{(j)}_t}\}$ are the trajectories of the oracle and the agent respectively. The term oracle refers to the agent that knows the optimal policy.

Note that the expectation of Equation 42 is equal to the value given in Equation 41 but Equation 42 has a high variance. Hence, when applicable (Scenario 1 and 3) we use Equation 41 to compute the expected regret.

## E Additional Numerical Experiments

### E.1 Scenario 1: Small Dimensional Example (Random Walk chain)

This scenario is explained in Appendix D.1 and the main numerical results are presented in Section 7. Here, we provide the result with error bars with respect to the random seed. The error bar size equals twice the standard deviation over 80 samples (each sample is a simulation with a given random seed and the random seeds are different for different simulations).

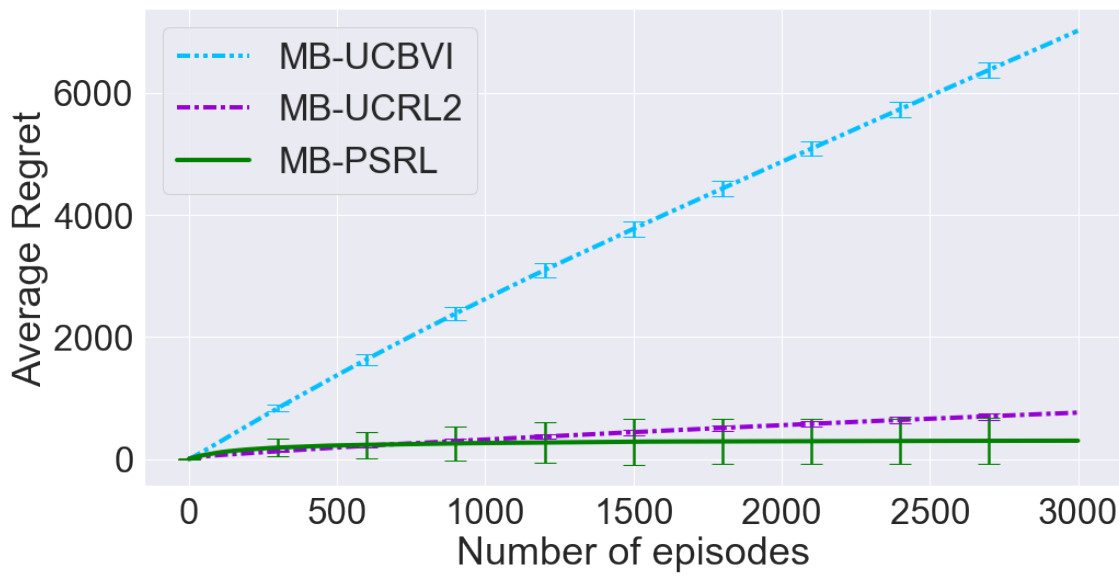



Figure 4: Average cumulative regret in function of the number of episodes. Result from 80 simulations in a markovian bandit problem with three 4-state random walk chains given in Table 1. The horizontal axis is the number of episodes. The size of the error bar equals twice the standard deviation over 80 simulations.

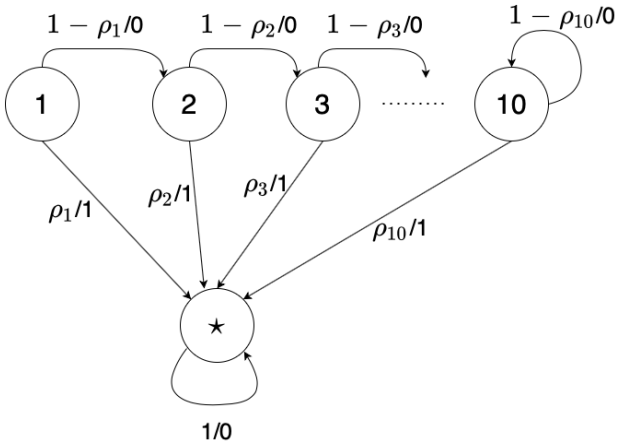
### E.2 Scenario 2: Higher Dimensional Example (Task Scheduling)

We now study an example that is too large to apply MB-UCRL2 . Hence, here we only compare MB-PSRL and MB-UCBVI.

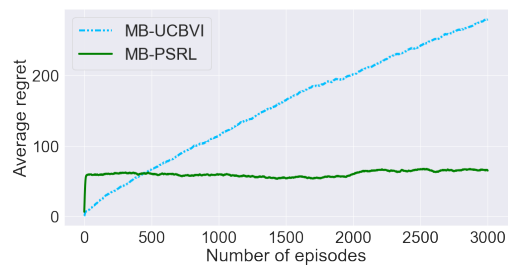
We implement the environment proposed on page 19 of Duff (1995) that was used as a benchmark for the algorithm in the cited paper. Each chain represents a task that needs to be executed, and is represented in Figure 5(a). Each task has 11 states (including finished state $\star$ that is absorbing). For a given chain $a \in \{1, \dots, 9\}$ and a state $i \in \{1, \dots, 10\}$, the probability that a task $a$ ends at state $i$ is $\rho_i^{(a)} = \mathbb{P}(\tau^{(a)} = i \mid \tau^{(a)} \geq i)$ where $\tau^{(a)}$ is the execution time of task $a$. We choose the same values of the parameters as in Duff (1995): $\rho_1^{(a)} = 0.1a$ for $a \in \{1, \dots, 9\}$, $\lambda = 0.8$, $\beta = 0.99$ and for $i \geq 2$,

$$\mathbb{P}\{x_a = i\} = \{1 - [1 - \rho_1^{(a)}]\lambda^{i-1}\}[1 - \rho_1^{(a)}]^{i-1}\lambda^{\frac{(i-1)(i-2)}{2}}.$$

Hence, the hazard rate $\rho_i^{(a)}$ is increasing with $i$. The reward in this scenario is deterministic: the agent receives 1 if the task is finished (*i.e.,* under the transition from any state $i$ to state $\star$) and 0 otherwise (*i.e.,* any other transitions including the one from state $\star$ to itself). For MB-PSRL, we use a uniform prior for the expected rewards and consider that the rewards are Bernoulli distributed.




(a) In state $i$, the task is finished with probability $\rho_i$ or transitions to state $i+1$ with probability $1-\rho_i$. For $i = 1, \ldots, 10$, the transition from state $i$ to state $\star$ provides 1 as the immediate reward. Otherwise, the agent always receives 0 reward.




(b) Average cumulative regret over 240 simulations.

Figure 5: Task Scheduling with 11 states including the absorbing state (finished state).

The average regret of the two algorithms is displayed in Figure 5(b). As before, MB-PSRL outperforms MB-UCBVI. Note that we also studied the time to run one simulation for 3000 episodes. This time is around 1 min for MB-PSRL and MB-UCBVI.

### E.3 Scenario 3: Bayesian Regret and Sensitivity to the Prior

In this section, we study how robust the two implementations of PSRL are, namely MB-PSRL and vanilla PSRL (to simplify, we will just call the later PSRL), to a choice of prior distributions. As explained in Appendix D.2.3, the natural conjugate prior for Bernoulli reward is the Beta distribution. In this section, we simulate MB-PSRL and PSRL in which the rewards are Bernoulli but the conjugate prior used for the rewards are Gaussian-Gamma which is incorrect for Bernoulli random reward. In other words, MB-PSRL and PSRL have Gaussian-Gamma prior belief while the real rewards are Bernoulli random variables.

To conduct our experiments, we use a markovian bandit problem with three 4-state random walk chains represented in Table 1. We draw 16 models by generating 16 pairs of $(r_L, r_R)$ from $U[0,1]$, 16 pairs of $(p_L, p_R)$ from Dirichlet(3,(1,1,1)) and 16 values of $p_{RL}$ from Dirichlet(2, (1,1)) for each chain. Each model is an unknown MDP that will be learned by MB-PSRL or PSRL. For each of these 16 models, we simulate MB-PSRL and PSRL 5 times with correct priors and 5 times with incorrect priors. The result can be found in Figure 6 which suggests that MB-PSRL performs better when the prior is correct and is relatively robust to the choice of priors in term of bayesian regret. This figure also shows that PSRL seems more sensitive to the choice of prior distribution. Also note that for both MB-PSRL and PSRL, some trajectories deviate a lot from the mean, under correct priors but even more so with incorrect priors. This illustrates the general fact that learning can go wrong, but with a small probability.

## F Experimental environment

The code of all experiments is given in a separated zip file that contains all necessary material to reproduce the simulations and the figures.

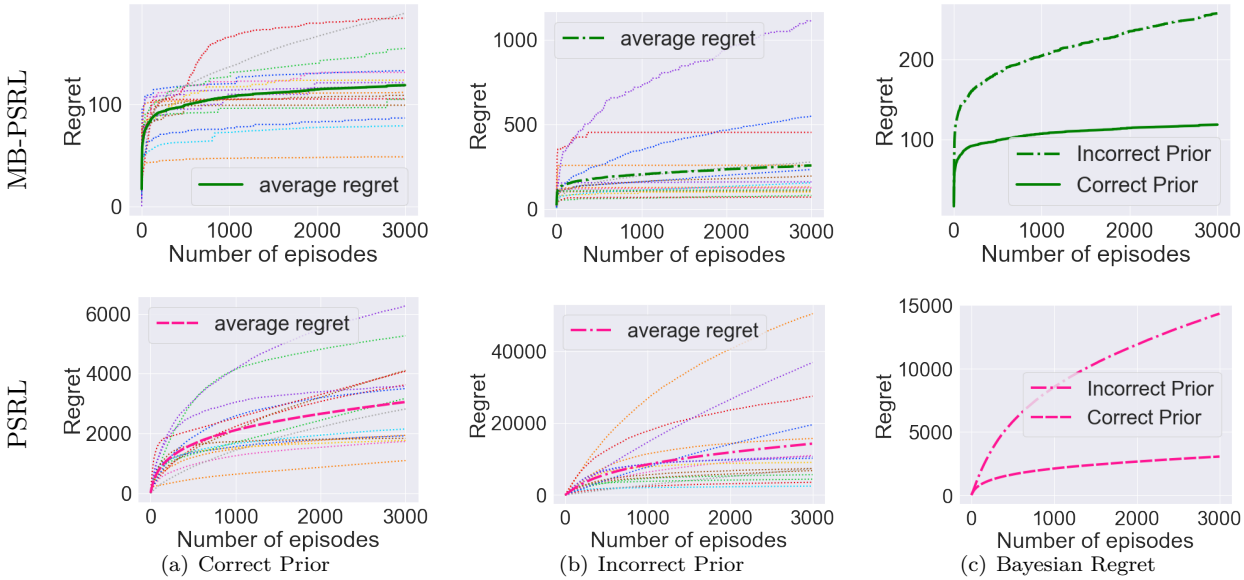



Figure 6: Bayesian regret of MB-PSRL and vanilla PSRL in 3 4-state Random Walk chains. For each chain, we draw 16 random models and run the algorithms for 5 simulations in each model (there are 80 simulations in total). In panels (a) and (b), we plot 16 dotted lines that correspond to the average cumulative regret over 5 simulations in the 16 samples. The solid and dash-dot lines are the average regret each over 80 simulations (the estimated bayesian regret). Figure 6(a) shows the performance when reward prior is well chosen (namely, $U([0,1])$). Figure 6(b) is when the reward prior is incorrectly chosen (namely Gaussian-Gamma distribution). Figure 6(c) compares the bayesian regret of the correct prior with the incorrect one (dash-dot line). In both case, the prior of next state transition is well chosen (namely, Dirichlet distribution). Y-axis range changes for each figure.

Our experiments were run on HPC platform with 1 node of 16 cores of Xeon E5. The experiments were made using Python 3 and Nix and submitted as supplementary material and will be made publicly available with the full release of the paper. The package requirement are detailed in README.md. Using only *1 core* of Xeon E5, the Table 2 gives some orders of duration taken by each experiment (with discount factor $\beta = 0.99$, and 3000 episodes per simulation). We would like to draw two remarks. First, the duration reported in Figure 1(b) is the time for policy computation (algorithm's parameters update and policy computation). The duration reported in Table 2 include this plus the computation time for oracle (because we track the regret), the state transition time along the trajectories of oracle and of each algorithm, resetting time... This explains why the duration reported in Table 2 cannot be compared to the duration reported in Figure 1(b). Second, the duration shown in Table 2 are meant to be a rough estimation of the computation time (we only ran the simulation once and the average duration might fluctuate).

| Experiment | MB-PSRL | PSRL | MB-UCRL2 | MB-UCBVI | Total |
|---|---|---|---|---|---|
| Scenario 1 | 40 min | - | 3days | 50 min | 3days |
| Scenario 2 | 200 min | - | - | 200 min | 400 min |
| Scenario 3 | 90 min | 260 min | - | - | 350 min |

Table 2: Approximative execution time for simulating each algorithm and tracking its regret in each scenario. This time includes the time given in Figure 1(b) and the computation time needed by oracle (because we track the regret), the state transition time along the trajectories of oracle and each algorithm, etc. In each scenario, we set the discount factor $\beta = 0.99$ and run the algorithms for 3000 episodes per simulation.