UCB Momentum Q-learning:
Correcting the bias without forgetting

Pierre Ménard [1] **Omar Darwiche Domingues** [2] **Xuedong Shang** [2 3] **Michal Valko** [2 3 4]

Abstract

We propose UCBMQ, Upper Confidence Bound Momentum Q-learning, a new algorithm for reinforcement learning in tabular and possibly stage-dependent, episodic Markov decision process. UCBMQ is based on Q-learning where we add a momentum term and rely on the principle of optimism in face of uncertainty to deal with exploration. Our new technical ingredient of UCBMQ is the use of momentum to correct the bias that Q-learning suffers while, *at the same time*, limiting the impact it has on the the second-order term of the regret. For UCBMQ, we are able to guarantee a regret of at most $\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^4SA)$ where H is the length of an episode, S the number of states, A the number of actions, T the number of episodes and ignoring terms in $\operatorname{poly}\log(SAHT)$. Notably, UCBMQ is the first algorithm that simultaneously matches the lower bound of $\Omega(\sqrt{H^3SAT})$ for large enough T and has a second-order term (with respect to the horizon T) that scales *only linearly* with the number of states S.

1. Introduction

In reinforcement learning (RL), an agent interacts with an environment with the objective of maximizing the sum of collected rewards (Sutton & Barto, 1998). We model the environment as an unknown episodic tabular Markov Decision Process (MDP) with S states, A actions and episodes of length H. After T episodes, we measure the performance of the agent by its cumulative regret which is the difference between the total reward collected by an optimal policy and the total reward collected by the agent during the learning. In order to minimize the regret the agent needs to balance the exploration of the environment and exploitation of the current knowledge to act optimally.

In particular, we study the *non-stationary* setting where rewards and transitions can change within an episode, and for which Jin et al. (2018) and Domingues et al. (2021b) provide a problem-independent lower bound on the regret of order $\Omega(\sqrt{H^3SAT})$ (see also Azar et al. 2017 for stationary transitions).

Following the previous work on the infinite-horizon setting (Jaksch et al., 2010; Fruit et al., 2018; Talebi & Maillard, 2018), a first line of research on episodic MDPs (Azar et al., 2017; Dann et al., 2017; Zanette & Brunskill, 2019) investigate model-based algorithms. The idea is to perform an optimistic value-iteration with an estimated model (i.e. estimated transitions here), and act greedily with respect to the obtained upper bounds on the optimal Q-values.

In particular, Azar et al. (2017) provide an upper bound on the regret of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^3S^2A)$. This bound matches the lower bound for $T \geq H^3S^3A$, where the first-order term, $\sqrt{H^3SAT}$, dominates. However, for $T \leq H^3S^3A$, which is an important regime, the bound is affected by the second order term that scales in S^2 and can be harmful. Indeed, when the number of states is very large (e.g., for continuous states MDPs after discretization), the second order term can dominate the regret bound, which in such case leads to a bound with a potentially sub-optimal rate (see Domingues et al. 2020; Sinclair et al. 2020). Furthermore, in order to obtain a non-trivial upper bound on the regret (i.e., a bound smaller than HT), at least H^3S^2A samples are needed. That means we roughly need H^2S samples per state-action pair while we rather expect to have a meaningful bound with only poly(H) samples per state-action pair. In the current analyses, the S^2 factor in the second-order term comes from the fact that, for model-based algorithms, the estimated transitions and the upper confidence bounds on the optimal value functions are correlated.[1] A union bound over a covering of all possible value functions with a cardinal that scales exponentially

[1]Otto von Guericke University [2]Inria [3]Université de Lille [4]DeepMind Paris. Correspondence to: Pierre Ménard <pierre.menard@ovgu.de>, Omar Darwiche Domingues <omar.darwiche-domingues@inria.fr>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1] It is the same reason why there is an extra factor S in the first order term of the bound of UCRL algorithm by Jaksch et al. (2010). This factor is "pushed" to the second-order term by the improved analysis of Azar et al. (2017).

Algorithm	**Upper bound** (non-stationary case)
UCBVI (Azar et al., 2017)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^3S^2A)$
UBEV (Dann et al., 2017)	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^2S^3A^2)$
EULER (Zanette & Brunskill, 2019)	$\widetilde{\mathcal{O}}\Big(\sqrt{H^3SAT} + H^3S^{3/2}A(\sqrt{S}+\sqrt{H})\Big)$
OptQL (Jin et al., 2018)	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^{9/2}S^{3/2}A^{3/2})$
UCB-Advantage (Zhang et al., 2020b)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^{33/4}S^2A^{3/2}T^{1/4})$
UCBMQ (this paper)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^4SA)$

Table 1. Regret upper bound under unknown episodic, non-stationary, tabular MDPs.

with the number of states S is (implicitly) used to break the correlation. A similar remark also holds for other model-based algorithms like EULER (see Table 1 for details).

A second line of work initiated by Jin et al. (2018) consider model-free algorithms based on Q-learning (Watkins & Dayan, 1992). Interestingly, such an approach does not suffer from the same issue as model-based algorithms. Indeed, the Q-values are estimated in an online fashion (see Section 3.1), and there is no correlation issue anymore as for model-based algorithms. On the other hand, the current estimate of the optimal Q-value for Q-learning-based algorithms relies on the target computed with past estimates of the same quantity (possibly inaccurate), therefore they suffer from a larger bias (see Section 3.1).

In particular, Jin et al. (2018) propose to use a more aggressive learning rate to mitigate that bias by forgetting old estimates, but at the price of increasing the variance. It leads to a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^4SAT})$ with an extra $\sqrt{H}$ in the first-order term with respect to the lower bound.[2] Building on variance reduction techniques, Sidford et al. (2018b) and Zhang et al. (2020b) manage to avoid this extra dependency on the horizon. The idea is to first provide a rough estimate of the optimal value, namely the value reference function, and then leverage the low variance of a reference-advantage decomposition of the optimal Q-value to compensate the forgotten past samples. However, in their current analyses, the initial phase of learning the reference value functions degrades the second order term and brings back a S^2 factor (see Table 1).

In this paper we rather follow another approach. Following the work of Azar et al. (2011) (see also Weng et al. 2020), we propose UCBMQ, which adds a momentum term to the targets in the Q-value updates so as to correct the bias of Q-learning. However, contrary to the generative setting considered by Azar et al. (2011) where all state-action pairs are sampled at each update of the Q-value, we have to deal with two additional challenges in our setting. First, we need to handle the exploration and we do it by introducing optimism. Second, in the absence of the oracle we do not see all state-action pairs at each "episode", but *only the ones encountered along the trajectory*. Consequently, each state-action pair learns at its own pace.

To address the above two challenges, we build a *value function for each state-action pair* that represents the bias of this particular pair, and use it to build a momentum term that is able to correct the bias of previous estimates on the Q-value. Every new sample is thus used to refine the estimate on the Q-value via the target and correct the bias of the past targets via the momentum term at the same time. Moreover, with the *careful* use of a Freedman-Bernstein-type inequality we manage to obtain tight dependence on the horizon without degrading the second-order term.

Using the above techniques, we prove a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^4SA)$ for UCBMQ. This upper bound matches the lower bound up to $\operatorname{poly}\log$ factors in S, A, H, T for $T \geq H^5SA$. This rate improves over the one of previous model-free algorithms and, for $S \geq H$, the one of previous model-based algorithms. In particular, we provide an algorithm that enjoys a second-order term *only* in S instead of S^2. Our results make a step towards resolving an open question that was hinted by Azar et al. (2011) and also recently explicitly raised by Zhang et al. (2020a). Finally, in Section 4, we provide numerical simulations on a grid-world environment to illustrate the benefits of not forgetting the targets in UCBMQ.

We highlight our main contributions:

- We carefully design a momentum term Q-learning in the episodic setting and analyse its benefits for the regret guarantees.
- We propose UCBMQ, with a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT}+H^4SA)$. It is the first algorithm, up to our knowledge, that matches the problem-independent lower bound $\Omega(\sqrt{H^3SAT})$ up to $\operatorname{poly}\log$ terms and

[2] Specifically, with Hoeffding-type bonuses they have an extra H and second-order term of order H^2SA; with Bernstein type bonuses, the discrepancy is only of a factor $\sqrt{H}$, but the second-order term is no longer linear in S, see Table 1.

has a second-order term that is linear in S.

2. Setting

In this paper, we consider a tabular episodic MDP $(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]})$, with $\mathcal{S}$ the set of states, $\mathcal{A}$ the set of actions, H the number of steps in one episode, $p_h(s'|s,a)$ is the probability transition from state s to state s' by taking the action a at step h, and $r_h(s,a) \in [0,1]$ is the bounded deterministic reward received after taking the action a in state s at step h. Note that we consider the general case of rewards and transition functions that are possibly non-stationary, i.e., that may change over the decision steps $h \in [H]$[3] within an episode. We denote by S and A the number of states and actions, respectively.

Policy & value functions. A *deterministic* policy π is a collection of functions $\pi_h : \mathcal{S} \to \mathcal{A}$ for all $h \in [H]$, where every π_h maps each state to a *single* action. The value functions of π, denoted by V_h^π, as well as the optimal value functions, denoted by $V_h^\star$ are given respectively by the Bellman equations (Puterman, 1994):

$$Q_h^\pi(s,a) = r_h(s,a) + p_h V_{h+1}^\pi(s,a) \qquad V_h^\pi(s) = \pi_h Q_h^\pi(s)\,.$$

By convention, $V_{H+1}^\pi \triangleq 0$. Furthermore, $p_h f(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $(\pi_h g)(s) \triangleq \pi_h g(s) \triangleq g(s, \pi_h(s))$ denotes the composition with the policy π at step h. An optimal policy $\pi^\star$ is such that $\pi^\star \in \arg\max_\pi V_1^\pi(s_1)$. The optimal Q-value and value functions are the ones of an optimal policy. Precisely we have $V_h^\star = V_h^{\pi^\star}$ and $Q_h^\star = Q_h^{\pi^\star}$ for all h.

Learning problem. The agent, to which the transitions are *unknown*, interacts with the environment during T episodes of length H, with a *fixed* initial state s_1.[4] Before each episode t the agent selects a policy π^t based only on the past observed transitions up to episode $t-1$. At each step $h \in [H]$ of episode t, the agent observes a state $s_h^t \in \mathcal{S}$, takes an action $\pi_h^t(s_h^t) = a_h^t \in \mathcal{A}$ and makes a transition to a new state s_{h+1}^t according to the probability distribution $p_h(s_h^t, a_h^t)$ and receives a deterministic reward $r_h(s_h^t, a_h^t)$.

Regret. We measure the agent performance through regret, which is the difference between what it could obtain (in expectation) by acting optimally and what it really gets,

$$R^T \triangleq \sum_{t=1}^T V_1^\star(s_1) - V_1^{\pi^t}(s_1)\,.$$

Notation. We denote the number of visits of state-action pair (s,a) by $n_h^t(s,a) = \sum_{k=1}^t \chi_h^t(s,a)$ where $\chi_h^t(s,a)$ is the indicator function $\chi_h^t(s,a) \triangleq \mathbb{1}_{\{(s_h^t, a_h^t)=(s,a)\}}$. We also use the indicator function $\chi_h^t(s) \triangleq \mathbb{1}_{\{s_h^t = s\}}$ to represent the event where state s is visited at step h in episode t. We denote by p_h^t the Dirac distribution at (s_{h+1}^t), i.e., for all functions f defined on $\mathcal{S}$ we have $(p_h^t f)(s,a) = f(s_{h+1}^t)$. In particular, this distribution does not depend on (s,a).

3. UCBMQ algorithm

Before presenting the algorithm we provide an intuition of how it works.

3.1. Intuition

If the agent knows the transition probabilities, it could perform real-time Q-value iteration and obtain a bounded regret (see Efroni et al. 2019). In this case upper bounds on the Q-value functions are updated as follows[5]

$$\overline{Q}_h^n(s,a) = (r_h + p_h \overline{V}_h^{n-1})(s,a)\,, \tag{1}$$

where upper bounds on the optimal value functions are defined by $\overline{V}_h^n(s) = \max_a \overline{Q}_h^n(s,a)$ and initialized to $\overline{V}_h^0(s) = H$. When the model is unknown we can approximate it by averaging successive sample updates as in Q-learning (Watkins & Dayan, 1992),

$$Q_h^n(s,a) = \alpha_n(r_h + p_h^n \overline{V}_h^{n-1})(s,a) + (1-\alpha_n) Q_h^{n-1}(s,a)\,. \tag{2}$$

A usual choice for the learning rate is $\alpha_n = 1/n$ instead of $\alpha_n = 1$ used for real-time Q-value iteration above. Unfolding the previous inequality and using Azuma–Hoeffding inequality to move for the sample expectation p_h^i to the true expectation p_h, we have with high probability

$$\begin{aligned} Q_h^n &\approx r_h(s,a) + \frac{1}{n}\sum_{i=1}^n p_h^i \overline{V}_{h+1}^{i-1}(s,a) \\ &\approx r_h(s,a) + p_h \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \overline{V}_{h+1}^{i-1}\right)}_{:= V_{h,s,a}^n \text{ bias-value function}}(s,a) \pm \underbrace{\sqrt{\frac{H^2}{n}}}_{\text{variance term}}\,, \end{aligned} \tag{3}$$

where the bias-value function of state-action (s,a) encodes the bias of the estimate Q_h^n with respect to the randomness of the $(p_h^i)_{i\geq 1}$. Thus choosing (Hoeffding-type) bonuses of order $\beta^n(s,a) \approx \sqrt{H^2/n}$, we can build upper bounds on the optimal Q-value and the value functions

$$\overline{Q}_h^n(s,a) = Q_h^n(s,a) + \beta_h^n(s,a), \quad \overline{V}_h^n(s) = \max_{a\in\mathcal{A}} \overline{Q}_h^n(s,a)\,.$$

[3] For any integer $n \in \mathbb{N}^\star$, we define $[n] := \{1, \ldots, n\}$.

[4] As explained by Fiechter (1994) and Kaufmann et al. (2021), if the first state is sampled randomly as $s_1 \sim p_0$, we can simply add an artificial first state s_0 such that for any action a, the transition probability is defined as the distribution $p_0(s_0, a) \triangleq p_0$.

[5] We index the quantities by n in this section where n is the number of times the state-action pair (s,a) is visited. In particular this is different from the time t since, in our setting, all the state-action pair are not visited at each episode. See Section 3.2 for precise notation.

However, the bias term in (3) is too large because of the old (and potentially inaccurate) upper bound $\overline{V}^{i}_{h+1}$ that appears in the bias-value function $V^n_{h,s,a}$. Indeed it is not clear how to prove a bound that is not exponential in the horizon H in this case (see Jin et al. 2018). Note that on contrary when the model is known, i.e. using (1), we have a smaller $V^n_{h,s,a} = \overline{V}^{n-1}_{h+1}$ bias provided that the $(\overline{V}^{i}_{h+1})_{i\geq 1}$ are non-increasing.

To overcome this issue, Jin et al. (2018) propose with the `OptQL` algorithm[6] to choose a learning rate of order $\alpha_n \approx H/n$ to keep only the recent upper-bounds $\overline{V}^{i}_{h+1}$ in the bias-value value function. Indeed, proceeding as above, we have

$$\begin{aligned} Q^n_h(s,a) &\approx r_h(s,a) + \frac{H}{n}\sum_{i\geq n-H/n}^{n} p^i_h \overline{V}^{i-1}_{h+1}(s,a) \\ &\approx r_h(s,a) + p_h \underbrace{\left(\frac{H}{n}\sum_{i\geq n-n/H}^{n} \overline{V}^{i-1}_{h+1}\right)}_{:=V^n_{h,s,a} \text{bias-value function}}(s,a) \pm \underbrace{\sqrt{\frac{H^3}{n}}}_{\text{variance term}}\,. \qquad (4) \end{aligned}$$

Because of the aggressive learning rate of order H/n there are only n/H samples in the sum of (4) leading to a high variance. Thus we need to add an extra H factor in the bonus which leads to the sub-optimal regret bound of order $\widetilde{O}(\sqrt{H^5SAT})$. One workaround for this issue is to learn a reference value function (Zhang et al., 2020b), but it is not clear how to obtain a second order term that depends linearly on the size of the state space with this approach.

We consider another approach in this paper. Following the work by Azar et al. (2011), we add a momentum term in the update of the Q-value that corrects the bias at the price of a small vanishing increase of the variance. Precisely for a momentum rate γ_n, we now consider the following update,

$$\begin{aligned} Q^n_h(s,a) = \alpha_n(r_h + p^n_h\overline{V}^{n-1}_{h+1})(s,a) + (1-\alpha_n)Q^{n-1}_h(s,a) \\ + \gamma_n p^n_h(\overline{V}^{n-1}_{h+1} - V^{n-1}_{h,s,a})(s,a)\,, \end{aligned}$$

where we call $V^{n-1}_{h,s,a}$ the bias-value function of state-action (s,a) defined by

$$V^n_{h,s,a}(s') = (\alpha_n+\gamma_n)\overline{V}^{n-1}_{h+1}(s') + (1-\alpha_n-\gamma_n)V^{n-1}_{h,s,a}(s')\,.$$

Note that there is a priori a different bias-value function for each state-action pair. In particular if we force the sequence of upper bounds on the value functions to be non-increasing, it holds that $V^n_{h,s,a} - \overline{V}^n_{h+1} \geq 0$. We choose $\alpha_n \approx 1/n$ to not forget samples as in (4). The momentum rate is $\gamma_n \approx H/n$ to correct the bias that will appear otherwise as in (3). As explained by Azar et al. (2011), this aggressive momentum will be compensated by the fact that $\overline{V}^{n-1}_{h+1} - V^{n-1}_{h,s,a}$ is small when the two quantities converge toward $V^\star_{h+1}$. Thanks to these choices, the bias-value function is the same as in (4),

$$\begin{aligned} V^n_{h,s,a}(s') &\approx \frac{H+1}{n}(V^{n-1}_{h,s,a} - \overline{V}^{n-1}_{h})(s') + \overline{V}^{n-1}_{h}(s') \\ &\approx \frac{H}{n}\sum_{i\geq n-n/H}^{n} \overline{V}^{i-1}_{h+1}(s')\,. \end{aligned}$$

Now we explain why $V^n_{h,s,a}$ is named *bias-value function*. We have, with high probability,

$$\begin{aligned} Q^n_h(s,a) &\approx r_h(s,a) + \frac{1}{n}\sum_{i=1}^{n} p^i_h\Big((H+1)\overline{V}^{i-1}_{h+1} - V^{i-1}_{s,a,h}\Big)(s,a) \\ &\approx r_h(s,a) + p_h \underbrace{\left(\frac{H}{n}\sum_{i\geq n-n/H}^{n} \overline{V}^{i-1}_{h}\right)}_{\approx V^n_{h,s,a} \text{ bias-value function}}(s,a) \pm \underbrace{\sqrt{\frac{H^2}{n}}}_{\text{variance term}} \\ &\pm \underbrace{\sqrt{\frac{H^3}{n}\sum_{i=1}^{n} p_h(V^{n-1}_{h,s,a} - \overline{V}^{n-1}_{h})(s,a)\frac{1}{n}}}_{\text{momentum variance term}}\,. \end{aligned}$$

Note that, we use a *negative* momentum since it allows to put more weight on the recent targets. We thus manage to get the advantages of the two learning rates: use all the samples for small variance and get a bias-value function that only relies on the recent upper-bounds on the optimal value function. This comes only at the cost of an additional momentum variance term that will only influence the dependence on H of the second order term in the regret. Note that here, for sake of simplicity, we used Azuma-Hoeffding inequality which leads to a sub-optimal dependence on the horizon. That is why in the sequel we rather use a Freedman-Bernstein-type inequality (and adapted bonuses) to obtain the optimal dependence on the horizon in the first order term.

Indeed `OptQL` by Jin et al. (2018) with Hoeffding-type bonuses has a regret bound of order $\sqrt{H^5SAT}$ with an extra factor H with respect to the lower bound of $\sqrt{H^3SAT}$ (in particular without second order term in S^2). Using Bernstein-type bonuses allows to waive a $\sqrt{H}$ factor in the first-order term. But there is still an extra $\sqrt{H}$ because of the aggressive learning rate of H/n used to deal with the bias issue as described above[7]. Note that doing so also introduces a second-order term which is not linear in the number of states S, see Table 1. This is because in their analysis they need a coarse upper bound on $\overline{V}^t_h - V^\star_h$ (see Lemma C.7 in the proof of Lemma C.3 then C.6 by Jin et al. (2018), such a coarse upper bound is also used by Azar et al. (2017)) to link the empirical variance to the true one. The key point in our analysis is to avoid such an intermediate coarse upper bound which leads inexorably to an extra

[6] With Hoeffding-type bonuses.

[7] Which will be removed because of the momentum in `UCBMQ`.

factor S. But instead postpone bounding such quantity to the next step error (we rather control $\overline{V}_h^t - V_h^{\pi^{t+1}}$), see Lemma 9 and Lemma 10. Indeed, we control $(p_h^t - p_h)\overline{V}_h$ instead of $(p_h^t - p_h)V_h^\star$ which allows us avoid upper bounding $\overline{V}_h^t - V_h^\star$ to build the upper confidence bound (see Lemma 7 and 8). But we do not know if it is impossible to build an upper confidence bound (that does not depends on S) by only controlling $(p_h^t - p_h)V_h^\star$.

3.2. Algorithm

We initialize the upper bounds on the optimal value functions by $\overline{V}_h^0(s) = H$ for all $(s,h) \in \mathcal{S} \times [H]$. We fix a learning rate $\alpha_h^t(s,a) \geq 0$ a momentum rate $\gamma_h^t(s,a) \geq 0$ such that $\alpha_h^t(s,a) + \gamma_h^t(s,a) \leq 1$. We also consider a bonus function $\beta_h^t(s,a)$. The update of the Q-value for UCBMQ is defined as follows. We update a (biased) estimator of the optimal Q-value function as follow,

$$\begin{aligned} Q_h^t(s,a) &= \alpha_h^t(s,a)\big(r_h(s,a) + p_h^t \overline{V}_{h+1}^{t-1}(s,a)\big) \\ &\quad + \gamma_h^t(s,a) p_h^t (\overline{V}_{h+1}^{t-1} - V_{h,s,a}^{t-1})(s,a) \\ &\quad + \big(1 - \alpha_h^t(s,a)\big) Q_h^{t-1}(s,a)\,, \end{aligned} \tag{5}$$

where the bias-value function for state-action (s,a) is defined by, $V_{h,s,a}^0(s') = H$,

$$V_{h,s,a}^t(s') = \eta_h^t(s,a)\overline{V}_{h+1}^{t-1}(s') + \big(1 - \eta_h^t(s,a)\big) V_{h,s,a}^{t-1}(s')\,, \tag{6}$$

where we define $\eta_h^t(s,a) = \alpha_h^t(s,a) + \gamma_h^t(s,a)$. We name this quantity the bias-value function because we will prove that with high probably $Q_h^t(s,a) \approx r_h(s,a) + p_h V_{h,s,a}^t(s,a)$ in Lemma 9 of Appendix E. Then we build upper-confidence bounds on the Q-values by adding a bonus and on the value functions by taking the maximum of the upper-confidence bounds on the Q-values (clipped to be non-increasing)

$$\overline{Q}_h^t(s,a) = Q_h^t(s,a) + \beta_h^t(s,a)\,,$$

$$\overline{V}_h^t(s) = \operatorname{clip}\big(\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a), 0, \overline{V}_h^{t-1}(s)\big)\,,$$

where the clipping operator is defined as $\operatorname{clip}(x,y,z) = \min(\max(x,y),z)$. We also fix the upper bounds of the value function at step $H+1$ to zero: $\overline{V}_{H+1}^t(s) = 0$. Note that $\overline{Q}_h^t(s,a)$ could be negative because of the momentum but it will still be an upper bound on the optimal Q-value with high probability, see Lemma 1. We also enforce the upper bound on the value function to be non-increasing. We then pick the action greedily with respect to the upper-bounds $\overline{Q}_h^t$. The complete procedure is described in Algorithm 1. We choose (with the convention $0 \times \infty = 0$ and $1/0 = \infty$)

$$\alpha_h^t(s,a) = \chi_h^t(s,a)\frac{1}{n_h^t(s,a)}\,, \tag{7}$$

$$\gamma_h^t(s,a) = \chi_h^t(s,a)\frac{H}{H + n_h^t(s,a)}\frac{n_h^t(s,a) - 1}{n_h^t(s,a)}\,, \tag{8}$$

for the learning rate and the momentum. Note that in particular it holds $\eta_h^t(s,a) = \chi_h^t(s,a)(H+1)/(H+n_h^t(s,a))$ which is the learning rate used by Jin et al. (2018). We can unfold (5) to obtain explicit formulas for the estimate of the Q-value function when $n_h^t(s,a) > 0$:

$$\begin{aligned} Q_h^t(s,a) &= r_h(s,a) + \frac{1}{n_h^t(s,a)}\sum_{k=1}^t \chi_h^k(s,a) p_h^k \overline{V}_h^{k-1}(s,a) \\ &+ \frac{1}{n_h^t(s,a)}\sum_{k=1}^t \chi_h^k(s,a)\mathring{\gamma}_h^k(s,a) p_h^k(\overline{V}_h^{k-1} - V_{h,s,a}^{k-1})(s,a)\,, \end{aligned} \tag{9}$$

where the normalized momentum is definied as

$$\mathring{\gamma}_h^k(s,a) = H\frac{n_h^t(s,a) - 1}{n_h^t(s,a) + H}\,.$$

We use a bonus derived from the Bernstein inequality plus a correction term. Precisely if $n_h^t(s,a) = 0$ then $\beta_h^t(s,a) = H$ otherwise

$$\begin{aligned} \beta_h^t(s,a) &= 2\sqrt{W_h^t(s,a)\frac{\zeta}{n_h^t(s,a)}} + 53H^3\frac{\zeta\log(T)}{n_h^t(s,a)} \\ &\quad + \sum_{k=1}^t \frac{\chi_h^k(s,a)\mathring{\gamma}_h^k(s,a)}{H\log(T) n_h^t(s,a)} p_h^k(V_{h,s,a}^{k-1} - \overline{V}_{h+1}^{k-1})(s,a)\,, \end{aligned}$$

where ζ is some exploration threshold that we specify later and W_h^t is a proxy for the variance term

$$W_h^t(s,a) = \sum_{k=1}^t \frac{\chi_h^k(s,a)}{n_h^t(s,a)} p_h^k\left(\overline{V}_{h+1}^{k-1} - \sum_{l=1}^t \frac{\chi_h^l(s,a)}{n_h^t(s,a)} p_h^l \overline{V}_{h+1}^{l-1}\right)^2(s,a)\,.$$

Note that the third term in the bonus will not compensate the momentum because it is $1/(H\log(T))$ times smaller than the momentum term.

3.3. Regret bound

We assume in this section that $T \geq 3$. We fix $\delta \in (0,1)$ and the exploration threshold

$$\zeta = \log(32eHSA(2T+1)/\delta)\,. \tag{10}$$

We can now state the main result of the paper which is proved in Appendix E. We sketch the proof in Section 3.4.

Theorem 1. *For UCBMQ, with probability at least $1 - \delta$*

$$R^T \leq C_1(\delta,T)\sqrt{H^3SAT} + C_2(\delta,T)H^4SA$$

where $C_1(\delta,T) = 126e^{127}\log(T)\sqrt{\zeta}$ *and* $C_2(\delta,T) = 3527e^{127}\log(T)^2\zeta$.

Algorithm 1 UCBMQ

```
Initialize: For all $(s,a,h)$, $V^0_{h,s,a} = \overline{V}^0_h = H$ and $Q^0_h = 0$
for $t \in [T]$ do
    for $h \in [H]$ do
        Play $a^t_h \in \arg\max \overline{Q}^{t-1}_h(s^t_h, a)$
        Observe $s^t_{h+1} \sim p_h(s^t_h, a^t_h)$
    end for
    for all $s, a, h$ do
        Update $Q^t_h(s,a)$ using Equation 5
        Update $V^t_{h,s,a}$ for all $s'$ using Equation 6
        $\overline{Q}^t_h(s,a) = Q^t_h(s,a) + \beta^t_h(s,a)$
        $\overline{V}^t_h(s) = \mathrm{clip}\big(\max_{a\in\mathcal{A}} \overline{Q}^t_h(s,a), 0, \overline{V}^{t-1}_h(s)\big)$
    end for
end for
```

Note that we did not try to optimize the constants C_1, C_2. The regret of UCBMQ is thus of order $\widetilde{\mathcal{O}}\big(\sqrt{H^3SAT} + H^4SA\big)$ matching the lower bound of $\widetilde{\mathcal{O}}\big(\sqrt{H^3SAT}\big)$ by Domingues et al. (2021b) for $T \geq H^5SA$.

Computational complexity. Note that the update of the upper bounds on the Q-values and value functions and the bias-value functions can be performed online. Indeed at step h and episode t, the learning rate $\alpha^t_h(s,a)$ and the momentum rate $\gamma^t_h(s,a)$ equal to zero if $(s,a) \neq (s^t_h, a^t_h)$. Thus the time complexity of UCBMQ is of order $\mathcal{O}(H(S+A)T)$ for T episodes. This complexity is smaller than the one of model-based algorithms, $\mathcal{O}(HSAT)$ at best (see Efroni et al. 2019), but is larger than $\mathcal{O}(HAT)$, the one of model-free algorithms (Jin et al., 2018; Zhang et al., 2020b). The space complexity is $O(HS^2A)$ since we need to store all the bias-value functions, which is the same as the one of model-based algorithms.

Model-free or model-based algorithm. UCBMQ does not estimate the probability transitions but rather estimates directly the Q-values/values. Therefore UCBMQ can be viewed as a model-free algorithm. On the other hand, the space complexity of UCBMQ is the same as the size of the model HS^2A. Thus from a space complexity point of view (see e.g. the definition of model-free algorithms by Jin et al. 2018), UCBMQ is a model-based algorithm.

Comparison with variance reduction methods. Building on variance reduction techniques, Sidford et al. (2018a;b) and Zhang et al. (2020b) propose model-free algorithms that match the problem-independent lower bound for large enough T (see Table 1). They use, for some reference value function V^{ref}, the following advantage decomposition of the optimal Q function,

$$Q^\star(s,a) = r_h(s,a) + p_h V^{\mathrm{ref}}_{h+1}(s,a) + p_h(V^\star_{h+1} - V^{\mathrm{ref}}_{h+1})(s,a).$$

To derive their algorithm, they estimate the two expectations above differently. The expectation $p_h V^{\mathrm{ref}}_{h+1}(s,a)$ is estimated using *all the samples*, and the expectation $p_h(V^\star_{h+1} - V^{\mathrm{ref}}_{h+1})$ is estimated using only the *last* $1/H$*-fraction* of the samples. The key point is to learn a reference value function V^{ref} that is close enough to V^* to compensate the smaller number of samples. However, learning such V^{ref}, which is done by using similar update as (4), requires a certain number of episodes, and increases the second term in their analysis. Interestingly, our update (5) could be seen as an advantage decomposition: considering (9), the bias-value function $V^t_{h,s,a}$ acts as a reference value function. However, contrary to the approach of Zhang et al. (2020b), $V^t_{h,s,a}$ is updated continuously as (6), instead of being fixed after a “burn-in” phase.

3.4. Proof sketch of Theorem 1

We first prove that $\overline{Q}^t$ and $\overline{V}^t$ are indeed upper confidence bounds on the optimal Q-values and the optimal value functions respectively.

Lemma 1. *On the event $\mathcal{E}$ that holds with probability $1-\delta$ (see Section C.1), $\forall t \in \mathbb{N}, \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ (also for $h = H+1$ for the value function), we have*

$$\overline{Q}^t_h(s,a) \geq Q^\star_h(s,a) \quad \textit{and} \quad \overline{V}^t_h(s) \geq V^\star_h(s)\,.$$

Step 1: Upper-bound $(\overline{Q}^t_h - Q^{\pi^{t+1}}_h)(s,a)$. We first upper-bound the difference $(\overline{Q}^t_h - Q^{\pi^{t+1}}_h)(s,a)$ for a certain state-action pair (s,a). Considering the rewriting (9) we can apply a Freedman-Bernstein-type inequality (see Appendix C.1) to replace the sample expectation by the true expectation (see Lemma 9),

$$\big|Q^t_h(s,a) - r_h(s,a) - p_h V^t_{h,s,a}(s,a)\big| \leq b^t_h(s,a)\,,$$

where we define, for $\widetilde{n}^t_h(s,a) = n^t_h(s,a) \wedge 1$,

$$b^t_h(s,a) = \sqrt{\frac{4}{\widetilde{n}^t_h(s,a)}\sum_{k=1}^t \chi^k_h(s,a)\mathrm{Var}_{p_h}(V^{\pi^k}_{h+1})(s,a)\frac{\zeta}{\widetilde{n}^t_h(s,a)}}$$
$$+\sum_{k=1}^t \frac{2\chi^k_h(s,a)}{H\log(T)\widetilde{n}^t_h(s,a)} p_h(\overline{V}^{k-1}_{h+1} - V^{\pi^k}_{h+1})(s,a) + 24H^3\frac{\log(T)\zeta}{\widetilde{n}^t_h(s,a)}.$$

In Lemma 10, we upper bound the bonus $\beta^t_h(s,a)$ with high probability, with a quantity of the same order as $b^t_h(s,a)$. Combining these two bounds we obtain

$$(\overline{Q}^t_h - Q^{\pi^{t+1}}_h)(s,a) \leq p_h(V^t_{h,s,a} - V^{\pi^{t+1}}_{h+1})(s,a) + 6b^t_h(s,a). \tag{11}$$

Step 2: Upper-bound the local optimistic regret. Next, we upper-bound the local optimistic regret of state-action (s,a) at step h defined by

$$\widetilde{R}_h^T(s,a) = \sum_{t=0}^{T-1} \chi_h^{t+1}(s,a)(\overline{Q}_h^t - Q_h^{\pi^{t+1}})(s,a).$$

We decompose the first term that appears in (11) by introducing the optimal value function

$$\begin{aligned} p_h(V_{h,s,a}^t - V_{h+1}^{\pi^{t+1}})(s,a) &= p_h(V_{h,s,a}^t - V_{h+1}^\star)(s,a) \\ &\quad + p_h(V_{h+1}^\star - V_{h+1}^{\pi^{t+1}})(s,a). \end{aligned}$$

Then, using Lemma 13 from Appendix F.2 yields

$$\begin{aligned} &\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(V_{h,s,a}^t - V_{h+1}^\star)(s,a) \\ &\le H + \sum_{k=1}^{T-1}\left(\sum_{t=k}^{T-1} \chi_h^{t+1}(s,a)\eta_h^{t,k}(s,a)\right) p_h(\overline{V}_{h+1}^{k-1} - V_{h+1}^\star)(s,a) \\ &\le H + \left(1+\frac{1}{H}\right)\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(\overline{V}_{h+1}^{t-1} - V_{h+1}^\star)(s,a), \end{aligned}$$

we get $V_{h,s,a}^t(s') = \sum_{k=1}^t \widetilde{\eta}_h^{t,k}(s,a)\overline{V}_{h+1}^{k-1}(s')$ by unfolding (6), see (15) in Appendix B. Combining this inequality with the previous decomposition and using that $V_{h+1}^\star \ge V_{h+1}^{\pi^{k+1}}$, we get

$$\begin{aligned} &\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(V_{h,s,a}^t - V_{h+1}^{\pi^{t+1}})(s,a) \\ &\le \sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(V_{h+1}^\star - V_{h+1}^{\pi^{t+1}})(s,a) + H \\ &\quad + \left(1+\frac{1}{H}\right)\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(\overline{V}_{h+1}^{t-1} - V_{h+1}^\star)(s,a) \\ &\le H + \left(1+\frac{1}{H}\right)\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^{t+1}})(s,a). \end{aligned}$$

We can proceed similarly to upper-bound the bonus term using this time Lemma 12, 14 from Appendix F.2, see (27), (28) and (29) in Appendix E, and get the upper bound on the optimistic local regret,

$$\begin{aligned} \widetilde{R}_h^T(s,a) &\le 44\log(T)\sqrt{\zeta \sum_{t=0}^{T-1} \chi_h^{t+1}(s,a)\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)} \\ &\quad + \left(1+\frac{41}{H}\right)\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a) p_h(\overline{V}_{h+1}^{t} - V_{h+1}^{\pi^{t+1}})(s,a) \\ &\quad + 1041 H^3 \log(T)^2 \zeta\,. \end{aligned}$$

Step 3: From visit to reach probability. We denote by $\bar{p}_h^t(s,a)$ respectively $\bar{p}_h^t(s)$ the probability to reach state-action (s,a) respectively state s at step h under the policy π^t. We replace the indicator function χ_h^t by its expectation $\bar{p}_h^t$. Using again an Freedman-Bernstein-type inequality (see Appendix C.2), from the upper bound on the optimistic local regret above we obtain

$$\begin{aligned} \widetilde{R}_h^T(s,a) &\le 63\log(T)\sqrt{\zeta \sum_{t=0}^{T-1} \bar{p}_h^{t+1}(s,a)\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)} \\ &\quad + \left(1+\frac{83}{H}\right)\sum_{t=0}^{T-1} \bar{p}_h^{t+1}(s,a) p_h(\overline{V}_{h+1}^{t} - V_{h+1}^{\pi^{t+1}})(s,a) \\ &\quad + 1754 H^3 \log(T)^2 \zeta\,. \end{aligned} \tag{12}$$

Step 4: Upper-bound the step h optimistic regret. We define the regret at step h by

$$\widetilde{R}_h^T = \sum_{s\in\mathcal{S}}\sum_{t=0}^{T-1} \bar{p}_h^{t+1}(s)(\overline{V}_h^{t-1} - V_h^{\pi^{t+1}})(s)\,.$$

Note that we used the probability to reach the state s rather than the indicator function $\chi_h^t(s)$ above. Using again a Freedman-Bernstein-type inequality (see Appendix C.2) to upper-bound the reach probability by the indicator function and the definition of $\overline{V}_h^k$, we have for $s \in \mathcal{S}$

$$\begin{aligned} &\sum_{t=0}^{T-1} \bar{p}_h^{t+1}(s)(\overline{V}_h^t - V_h^{\pi^{t+1}})(s) \\ &\le \left(1+\frac{1}{H}\right)\sum_{t=0}^{T-1} \chi_h^{t+1}(s)(\overline{V}_h^t(s) - V_h^{\pi^{t+1}})(s) + 19H^2\zeta \\ &\le \left(1+\frac{1}{H}\right)\sum_{t=0}^{T-1} \chi_h^{t+1}(s)\pi_h^{t+1}(\overline{Q}_h^k - Q_h^{\pi^{t+1}})(s) + 19H^2\zeta\,. \end{aligned}$$

Combining this inequality with (12) then the fact the policies π^t are deterministic and Cauchy-Schwarz inequality yield the upper-bound the step h optimistic regret

$$\begin{aligned} \widetilde{R}_h^T &\le \left(1+\frac{1}{H}\right)\sum_{s,a}\sum_{t=0}^{T-1} \chi_h^{t+1}(s,a)(\overline{Q}_h^k - Q_h^{\pi^{t+1}})(s,a) + 19H^2 S\zeta \\ &= \left(1+\frac{1}{H}\right)\sum_{s,a} \widetilde{R}_h^T(s,a) + 19H^2 S\zeta \\ &\le 126\log(T)\sqrt{\zeta S A \sum_{s,a}\sum_{t=0}^{T-1} \bar{p}_h^{t+1}(s,a)\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)} \\ &\quad + \left(1+\frac{167}{H}\right)\widetilde{R}_{h+1}^T + 3527 H^3 S A \log(T)^2\zeta\,, \end{aligned} \tag{13}$$

where in the last inequality we used that

$$\sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} \bar{p}_h^{t+1}(s,a)p_h(s'|s,a) = \sum_{s'\in\mathcal{S}} \bar{p}_{h+1}^{t+1}(s')\,.$$

Step 5: Upper-bound the regret. We upper-bound the Step 1 regret $\widetilde{R}_1$. By successively unfolding (13) with the fact that $\widetilde{R}^T_{h+1} = 0$, using the Cauchy-Schwarz inequality and the law of total variance (Lemma 11 in Appendix F.1),

$$
\begin{aligned}
\widetilde{R}^T_1 \leq & \sum_{h=1}^{H} C_1(\delta,T)\sqrt{SA\sum_{s,a}\sum_{t=0}^{T-1}\bar{p}_h^{t+1}(s,a)\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)} \\
& + C_2(\delta,T)H^3SA \\
\leq & \ C_1(\delta,T)\sqrt{SAH\sum_{s,a,h}\sum_{t=0}^{T-1}\bar{p}_h^{t+1}(s,a)\mathrm{Var}_{p_h}(V_{h+1}^{\pi^{t+1}})(s,a)} \\
& + C_2(\delta,T)H^4SA \\
\leq & \ C_1(\delta,T)\sqrt{H^3SAT} + C_2(\delta,T)H^4SA\,.
\end{aligned}
$$

It remains to relate the opstimistic regret with the regret. Thanks to Lemma 1 we have

$$
V_1^\star(s_1) - V_h^{\pi^{t+1}}(s_1) \leq \overline{V}_1^t(s_1) - V_1^{\pi^{t+1}}(s_1)\,,
$$

which allows us to conclude

$$
R^T \leq \widetilde{R}^T_1 \leq C_1(\delta,T)\sqrt{SAH^3T} + C_2(\delta,T)SAH^4\,.
$$

4. Experiments

In this section, we present a numerical simulation to illustrate the benefits of not forgetting the targets in UCBMQ. We compare UCBMQ to the following baselines: (i) UCBVI (Azar et al., 2017); (ii) OptQL (Jin et al., 2018), and (iii) Greedy-UCBVI, a version of UCBVI using real–time dynamic programming (Efroni et al., 2019). We use a grid-world environment with 50 states $(i,j) \in [10] \times [5]$ and 4 actions (left, right, up and down). When taking an action, the agent moves in the corresponding direction with probability $1-\varepsilon$, and moves to a neighbor state at random with probability ε. The starting position is $(1,1)$. The reward equals to 1 at the state $(10,5)$ and is zero elsewhere.[8]

Using different exploration bonuses (e.g., by changing the multiplicative constants) can result in drastically different regrets empirically. In order to fairly compare the algorithmic ideas of UCBMQ to the baselines, we use the *same exploration bonus* for all the algorithms, given by:

$$
\beta_h^t(s,a) = \min\left(\sqrt{\frac{1}{n_h^t(s,a)}} + \frac{H-h+1}{n_h^t(s,a)}, H-h+1\right).
$$

Although the confidence intervals required by the algorithms are not always satisfied with this bonus, they hold for $n_h^t(s,a) = 0$ (resulting in $\beta_h^t(s,a) = H-h+1$), so that this choice does not hurt the initial exploration. When $n_h^t(s,a) > 0$, the bonus behaves as a simplified version of the Bernstein-type bonuses used in different algorithms.

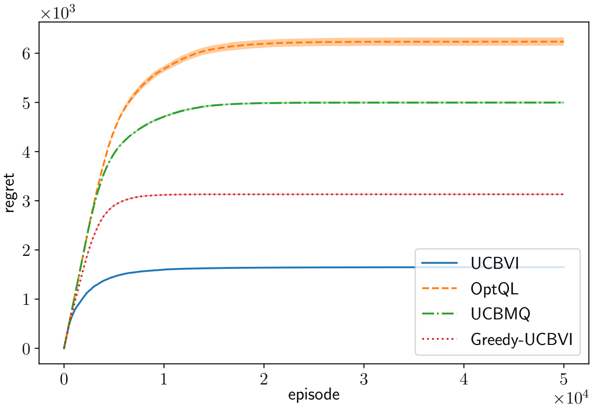

Figure 1. Regret of UCBMQ compared to baselines, for $H = 100$ and transition noise $\varepsilon = 0.15$. Average over 8 runs.

In Figure 1, we observe that UCBMQ outperforms OptQL in our experiments, whereas the only differences in the implementations of the two algorithms are the learning rates and the momentum term used by UCBMQ (since the bonuses were kept identical). This illustrates the potential gain in sample efficiency enabled by not forgetting the targets.

We also observe that, in this simulation, UCBMQ has a larger regret than UCBVI and Greedy-UCBVI, which are model-based algorithms using empirical estimates of the transitions probabilities and planning. It is not surprising since explicitly using a model and backward induction allows new information to be more quickly propagated to the value function computed by the algorithms. UCBVI performs *full planning* after each episode. Greedy-UCBVI does *1-step* planning, propagating information more quickly than UCBMQ, but more slowly than UCBVI, which explains the results in Figure 1. However, current regret bounds for model-based algorithms, such as UCBVI, still feature a second order term scaling with S^2 (see Table 1): an interesting open question is whether a bound scaling linearly with S can be obtained when a transition model is used.

5. Conclusion

We studied regret minimization in tabular, non-stationary, episodic MDPs. For this settings, we provided an algorithm a regret bound that is optimal in a problem-independent sense for a large enough number of episodes T and such that the *second-order term in the regret bound scales only linearly with the number of states S*. Our result rises following interesting open questions for a further research.

Problem-independent optimal regret. We conjecture that the optimal problem-independent regret is $\mathcal{O}(\sqrt{H^3SAT} + H^2SA)$. This conjecture is coherent with the one of Wang et al. (2020) for PAC problem-

[8] The code to reproduce the experiments is available on GitHub, and uses the rlberry library (Domingues et al., 2021a).

independent optimal sample complexity if we do not assume that the sum of the rewards along any trajectory is smaller than 1. In particular, it is not clear how to obtain a better dependency on the horizon H in the second-order term, while being only linear in S. For UCBMQ we have an extra H factor in the second-order term in comparison to the regret bound of UCBVI. This is due to the momentum rate γ which scales with H (Equation 8). This scaling seems necessary to refrain from getting an extra H factor at the first-order term and it is unclear how to avoid it. Note that if our conjecture for the optimal problem-independent regret is true, the regret bounds for the model-based algorithms (e.g., UCBVI, see Table 1) would be sub-optimal in H in the second-order term.

Dependency on S for model-based algorithms. Even if UCBMQ could be considered as a model-based algorithm (Section 3.2) it relies on the model-free Q-learning algorithm. This is the main reason behind obtaining a linear dependence on the size of the state space S in the second-order term. As explained in Section 1, it is not clear to obtain similar bounds for model-based algorithm but experimentally they perform better, see Section 4. Interestingly with access to a generative model, Szita & Szepesvári (2010) managed to get rid of the extra factor S for PAC-MDP complexity (Kakade, 2003).

Computational complexity. As we need to maintain a separate bias-value function for each state-action pairs, UCBMQ has a larger complexity both in time and space than the algorithms of Jin et al. (2018) and Zhang et al. (2020b). It is not clear how to obtain an algorithm with the same guarantees as UCBMQ while having a space complexity of $\mathcal{O}(HSA)$ and a time complexity of $\mathcal{O}(HT)$.

Acknowledgements

The research presented was supported by European CHIST-ERA project DELTA. Pierre Ménard is supported by the SFI Sachsen-Anhalt for the project RE-BCI ZS/2019/10/102024 by the Investitionsbank Sachsen-Anhalt.

References

Azar, Mohammad Gheshlaghi, Munos, Remi, Ghavamzadeh, Mohammad, and Kappen, Hilbert J. Speedy Q-learning. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pp. 2411–2419, 2011.

Azar, Mohammad Gheshlaghi, Osband, Ian, and Munos, Rémi. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 405–433, 2017.

Dann, Christoph, Lattimore, Tor, and Brunskill, Emma. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 5714–5724, 2017.

Domingues, Omar D., Ménard, Pierre, Pirotta, Matteo, Kaufmann, Emilie, and Valko, Michal. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020.

Domingues, Omar Darwiche, Flet-Berliac, Yannis, Leurent, Edouard, Ménard, Pierre, Shang, Xuedong, and Valko, Michal. rlberry - A Reinforcement Learning Library for Research and Education. https://github.com/rlberry-py/rlberry, 2021a.

Domingues, Omar Darwiche, Ménard, Pierre, Kaufmann, Emilie, and Valko, Michal. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, 2021b.

Efroni, Yonathan, Merlis, Nadav, Ghavamzadeh, Mohammad, and Mannor, Shie. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 12224–12234, 2019.

Fiechter, Claude Nicolas. Efficient reinforcement learning. In *Proceedings of the 7th Annual Conference on Learning Theory (CoLT)*, pp. 88–97, 1994.

Fruit, Ronan, Pirotta, Matteo, Lazaric, Alessandro, and Ortner, Ronald. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1578–1586, 2018.

Jaksch, Thomas, Ortner, Ronald, and Auer, Peter. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, 2010.

Jin, Chi, Allen-Zhu, Zeyuan, Bubeck, Sebastien, and Jordan, Michael I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 4863–4873, 2018.

Kakade, Sham. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

Kaufmann, Emilie, Ménard, Pierre, Domingues, Omar Darwiche, Jonsson, Anders, Leurent, Edouard, and Valko, Michal. Adaptive reward-free exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, 2021.

Ménard, Pierre, Domingues, Omar Darwiche, Jonsson, Anders, Kaufmann, Emilie, Leurent, Edouard, and Valko, Michal. Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*, 2021.

Puterman, Martin L. *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc, 1994.

Sidford, Aaron, Wang, Mengdi, Wu, Xian, Yang, Lin F., and Ye, Yinyu. Near-Optimal time and sample complexities for solving discounted Markov decision process with a generative model. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018a.

Sidford, Aaron, Wang, Mengdi, Wu, Xian, and Ye, Yinyu. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018b.

Sinclair, Sean R., Wang, Tianyu, Jain, Gauri, Banerjee, Siddhartha, and Yu, Christina Lee. Adaptive discretization for model-based reinforcement learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Sutton, Richard S. and Barto, Andrew G. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Szita, István and Szepesvári, Csaba. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 1031–1038, 2010.

Talebi, Mohammad Sadegh and Maillard, Odalric Ambrym. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, 2018.

Wang, Ruosong, Du, Simon S., Yang, Lin F., and Kakade, Sham M. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, pp. 1–19, 2020.

Watkins, Chris J. and Dayan, Peter. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Weng, Bowen, Xiong, Huaqing, Zhao, Lin, Liang, Yingbin, and Zhang, Wei. Momentum Q-learning with finite-sample convergence guarantee. *arXiv preprint arXiv:2007.15418*, 2020.

Zanette, Andrea and Brunskill, Emma. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 12676–12684, 2019.

Zhang, Zihan, Ji, Xiangyang, and Du, Simon S. Is reinforcement learning more difficult than bandits? A near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, pp. 1–28, 2020a.

Zhang, Zihan, Zhou, Yuan, and Ji, Xiangyang. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020b.