Generalized Optimistic Q-Learning with Provable Efficiency

Grigory Neustroev
Delft University of Technology
Delft, the Netherlands
g.neustroev@tudelft.nl

Mathijs M. de Weerdt
Delft University of Technology
Delft, the Netherlands
m.m.deweerdt@tudelft.nl

ABSTRACT

Reinforcement learning (RL), like any on-line learning method, inevitably faces the exploration-exploitation dilemma. When a learning algorithm requires as few data samples as possible, it is called sample efficient. The design of sample-efficient algorithms is an important area of research. Interestingly, all currently known provably efficient model-free RL algorithms utilize the same well-known principle of optimism in the face of uncertainty. We unite these existing algorithms into a single general model-free optimistic RL framework. We show how this facilitates the design of new optimistic model-free RL algorithms by simplifying the analysis of their efficiency. Finally, we propose one such new algorithm and demonstrate its performance in an experimental study.

KEYWORDS

Reinforcement learning; model-free learning; sample efficiency

ACM Reference Format:
Grigory Neustroev and Mathijs M. de Weerdt. 2020. Generalized Optimistic Q-Learning with Provable Efficiency. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

Reinforcement learning (RL) [24] is a popular framework for sequential decision-making problems in an unknown environment, applicable to a wide range of problems. In general, RL methods fall into two categories: model-based and model-free. Model-based approaches build an approximate model of the environment and use it to reason about optimality of actions. Model-free approaches, in contrast, estimate optimality of actions directly. To find the best possible course of actions, RL requires many repeated trials, which is effective but costly. Therefore, one of the important challenges in RL is the design of *sample-efficient* algorithms, that is, algorithms utilizing as much information from each interaction as possible. Sample efficiency of model-based RL has been studied extensively, and several methods were proven to be sample efficient [4, 15].

Even though most RL breakthroughs—from seminal Q-learning [27] to state-of-the-art deep Q-networks [11, 17]—are of the model-free paradigm, theory on sample efficiency of model-free RL remains limited. Only recently some dispersed results have appeared for a few model-free methods. For proper understanding of the potential of model-free RL, and thus of the design of optimal RL algorithms, we need to better understand the relation between the efficiency of these methods and various components of their design.

The first provably efficient model-free RL algorithm was introduced by Jin et al. [14]. It is called upper confidence bound Q-learning and comes in two forms: with Hoeffding-style bonus (UCB-H), and with Bernstein-style bonus (UCB-B). Its conception sparked interest in sample complexity of model-free RL; as a result, several similar methods have been proposed, namely, infinite-horizon UCB (∞-UCB) Q-learning [26], optimistic pessimistically-initialized Q-learning (OPIQ) [21], and UCB2-based methods in the context of problems with limited adaptivity [5]. All of these methods attribute their success to the use of the same learning rate [14].

Another factor that allows these (both model-based and model-free) algorithms to achieve sample efficiency is their use of *optimism in the face of uncertainty* [25], which postulates that a learning agent should assume that its actions lead to the best realistically possible outcomes. In practice, this principle is implemented in two ways: *optimistic initialization*—unencountered state-action pairs are assumed to have the best outcomes [24, Chapter 2.6], and *action selection based on UCBs*—each previously encountered state-action pair is assumed to yield a reward that is as good as is statistically plausible [24, Chapter 2.7]. While there exist other techniques to improve the efficiency of learning, such as variance reduction methods [8], posterior sampling [2, 19], or use of randomized value functions [18], this research aims to better understand the effect of optimism.

The main contribution of this work is a generalized theory on optimistic Q-learning which unifies the existing algorithms. In the context of model-based methods, there already exists a generalization known as optimistic initial model (OIM) [25]. Instead, we focus on model-free methods because they have better space complexity and can be adapted to deep learning, which is arguably the most promising direction of future work, while being provably efficient.

We also perform a generalized theoretical analysis of sample efficiency. In order to establish efficiency of an algorithm, two related techniques are used. Some authors provide PAC-bounds on the *time* required to achieve near-optimal performance [15, 22, 23, 26]. We employ another approach and establish efficiency by showing that the *regret* of the algorithm—the total loss of reward incurred while learning—grows sub-linearly with respect to the number of interactions [5, 14, 21]. The two approaches are similar; in fact, it is known that one implies the other, and *vice versa* [14, 19].

To summarize, in this work, we study the effects of optimism on the regret of model-free RL algorithms. We start with examining the existing sample-efficient Q-learning methods and identifying their common features. Then we propose a generalized model of optimistic Q-learning, which encompasses these methods. Next, we perform a theoretical regret analysis and derive a regret bound for the generalized model, which allows us to identify the sources of regret. We show how these general results can be used to facilitate the design of new optimistic model-free algorithms by proposing one such algorithm, and evaluate its performance experimentally.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

2 BACKGROUND

This section introduces the underlying model and our notation.

2.1 Non-Stationary Markov Decision Processes

We use episodic non-stationary Markov decision process (NS-MDP) as an underlying model because the total regret is a well-defined value in episodic learning [26] but is not as clearly defined in other settings. An episodic NS-MDP is defined as a tuple $\mathbf{M} \triangleq \langle \mathbb{S}, \mathbb{A}, \mathbb{A}_h, p_h, r_h, \gamma, H, K \rangle$. In this setting, the agent interacts with the environment for K episodes, each consisting of H time steps for the total number of $T \triangleq HK$ interactions. We denote the sets of all episodes and steps of each episode as $\mathbb{K} \triangleq \{1, \ldots, K\}$ and $\mathbb{H} \triangleq \{1, \ldots, H\}$. At each time step h, an agent observes the state of the environment $s_h \in \mathbb{S}$ and chooses one of the available actions $a_h \in \mathbb{A}_h(s_h) \subseteq \mathbb{A}$. The environment transitions to a new state s_{h+1} with probability $p_h(s_{h+1}|x_h)$; the agent observes this transition and receives a reward $r_h(x_h)$. We use $x_h \triangleq (s_h, a_h)$ for state-action pairs and $\mathbb{X}_h \triangleq \mathbb{S} \times \mathbb{A}_h$ for the set of all state-action pairs that can be encountered in time step h. We denote the space of all possible state-action pairs as $\mathbb{X} \triangleq \bigcup_{h\in\mathbb{H}} \mathbb{X}_h \subseteq \mathbb{S} \times \mathbb{A}$, and its size as $X \leq SA$.

Possible courses of actions are known as *policies* $\pi \triangleq \{\pi_h\}_{h\in\mathbb{H}}$, where $\pi_h : \mathbb{S} \longrightarrow \mathbb{A}$ maps states to admissible actions $\pi_h(s) \in \mathbb{A}_h(s)$. Given the state s at time step h, each policy has a *value* $V_h^\pi(s)$ that can be found using the Bellman policy equations:

$$V_h^\pi(s) = Q_h^\pi\big(s, \pi_h(s)\big), \qquad V_{H+1}^\pi(s) = 0, \tag{1}$$

$$Q_h^\pi(x) = [r_h + \gamma \mathcal{P}_h V_{h+1}^\pi](x),$$

$$[\mathcal{P}_h f](x) \triangleq \textstyle\sum_{s'\in\mathbb{S}} p_h(s'|x) f(s') \qquad \forall f : \mathbb{S} \to \mathbb{R}. \tag{2}$$

The agent needs to learn an optimal policy, that is, a policy $\pi^\star$ with the highest possible values $V_h^{\pi^\star}(s) = V_h^\star(s) \triangleq \max_\pi V_h^\pi(s)$. The optimal values $V_h^\star(s)$ satisfy the Bellman optimality equations

$$V_h^\star(s) = [\mathcal{M}_h Q_h^\star](s), \qquad V_{H+1}^\star(s) = 0, \tag{3}$$

$$Q_h^\star(x) = [r_h + \gamma \mathcal{P}_h V_{h+1}^\star](x), \tag{4}$$

$$\text{where} \quad [\mathcal{M}_h g](s) \triangleq \max_{a\in\mathbb{A}_h(s)} g(s,a) \qquad \forall g : \mathbb{X}_h \to \mathbb{R}.$$

In each episode k, the agent follows some policy π^k. The (expected) *total regret* R of such agent in an episodic NS-MDP $\mathbf{M}$ is defined as

$$R \triangleq \textstyle\sum_{k=1}^K R^k = \sum_{k=1}^K \big(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\big).$$

Finally, in this paper we assume that the rewards and values are bounded, but the bounds may vary between steps, that is, $r_h(x) \in [r_h^-, r_h^+]$ and $V_h^\pi(x) \in [V_h^-, V_h^+]$ for all $x \in \mathbb{X}$ and π. For simplicity, we use deterministic rewards; however, our results can be extended to randomized rewards. We denote the reward bounds of the whole episode as $r^\pm(H)$, that is, $r^-(H) \leq \min_{h\in\mathbb{H}} r_h^-$ and $r^+(H) \geq \max_{h\in\mathbb{H}} r_h^-$. We denote the reward span of a step as $r_h^\triangle \triangleq r_h^+ - r_h^-$, and of an episode as $r^\triangle(H) \triangleq r^+(H) - r^-(H)$. We define the value bounds $V^\pm(H)$ and spans $V_h^\triangle$ and $V^\triangle(H)$ similarly.

2.2 Reinforcement Learning

In RL the transition and reward functions of an MDP are not known, so the Bellman equation (4) cannot be applied directly. Instead, the optimal *Q-values* are learned through interactions with the environment. The initial Q-values $Q_h^0(x)$ are chosen arbitrarily, and at each episode $k+1$ they are gradually updated from the previous Q-values $Q_h^k(x)$. In Q-learning [27], the update rule is:

$$Q_h^{k+1}(x) = \begin{cases} (1-\alpha_t) Q_h^k(x) + \alpha_t U_h^k(x, s_{h+1}) & \text{if } x = x_h^{k+1}, \\ Q_h^k(x) & \text{otherwise,} \end{cases} \tag{5}$$

where $U_h^k(x,s) \triangleq r_h(x) + \gamma[\mathcal{M}_{h+1} Q_{h+1}^k](s)$ is the *update*. To easier relate these values to the optimal Q-values $Q_h^\star(x)$, we use the empirical transition operator $\hat{\mathcal{P}}_h^k$ for each $k \in \mathbb{K}$ and $h \in \mathbb{H}$:

$$[\hat{\mathcal{P}}_h^k f](x) \triangleq f(s_{h+1}^k) \quad \text{if } h < H, \quad \text{and} \quad [\hat{\mathcal{P}}_H^k f](x) \triangleq 0. \tag{6}$$

Using this operator, the update term can be written similarly to the Bellman equations (3) and (4):

$$U_h^k(x, s_{h+1}^k) \triangleq [r_h + \gamma \hat{\mathcal{P}}_h^k V_{h+1}^k](x) \quad \text{with} \quad V_h^k(s_h^k) \triangleq [\mathcal{M}_h Q_h^k](s_h^k).$$

The function α_t is called the *learning rate*. We use t as a shorthand for the *visitation function* $\#_h^k(x)$, which gives the number of times the state-action pair x has been visited in time step h of the first k episodes. The learning rate is used to balance the newly acquired information $U_h^k(x,s)$ with the old experiences $Q_h^k(x)$. For an appropriate choice of the learning rate, the sequence $\{Q_h^k(x)\}_{k=1}^\infty$ converges to $Q_h^\star(x)$ w.p. one, if the state-action space $\mathbb{X}$ is finite $|\mathbb{X}| < \infty$ and the rewards function r is bounded [13]. In particular, the conditions on the learning rate are:

$$\textstyle\sum_{t=1}^\infty \alpha_t(x) = \infty \quad \text{and} \quad \sum_{t=1}^\infty \alpha_t^2(x) < \infty \quad \text{for all } x \in \mathbb{X}. \tag{7}$$

The first condition ensures that the updates remain large enough to affect Q-values, while the second condition guarantees that the variance of the resulting iterative stochastic process remains bounded (i.e., that it converges).

Using the notation of [14], we introduce the following values, which we call the *cumulative learning rates*:

$$\alpha_t^0 = \textstyle\prod_{j=1}^t (1-\alpha_j), \quad \text{and} \quad \alpha_t^i = \alpha_i \prod_{j=i+1}^t (1-\alpha_j). \tag{8}$$

For $t = 0$, we define $\alpha_t^0 \triangleq 1$ and $\sum_{i=1}^t \alpha_t^i \triangleq 0$. If a state-action pair $x = (s,a)$ was previously visited in time step h of episodes $k_1, \ldots, k_t < k$, then by the update equation (5) on k_i we can write

$$Q_h^k(x) = \alpha_t^0 Q_h^0(x) + \textstyle\sum_{i=1}^t \alpha_t^i U_h^{k_i}(x, s_{h+1}^{k_i}). \tag{9}$$

3 OPTIMISM IN Q-LEARNING

This section presents our main contribution. We start with an overview of optimism in model-free RL methods. Then we propose a generalized framework of optimistic RL. Next, we formulate the conditions under which the total regret of optimistic Q-learning can be bounded and present an intuitive interpretation of the bound.

3.1 Representation of Optimism

As briefly mentioned in Section 1, the principle of optimism in the face of uncertainty is usually applied in two ways: optimistic initialization, and use of UCBs in action selection. We looked at UCB-H [14], UCB-B [14], ∞-UCB [26], and OPIQ [21] to see how they incorporate these two aspects of optimism.

For initialization, all of the methods use $Q_h^0(x) = V^+(H) = V^+$ except for OPIQ. The latter uses $Q_h^0(x) = V^-$, but additionally augments Q-values with a *bonus for optimism* $v(t)$, depending on the visitation counter t. These *augmented Q-values* $\bar{Q}_h(x) \triangleq Q_h(x) + v(t)$

overestimate the true Q-values (i.e., they are optimistic) and are used for action selection. The particular choice of this bonus is $v(t) = C/(t+1)^M$, where $C \geq V^{\triangle}$ and M is a sufficiently large number. It ensures that the augmented Q-values $\bar{Q}^0_h(x)$ of unvisited state-action pairs are optimistic:

$$\bar{Q}^0_h(x) = Q^0_h(x) + v(t) \geq V^- + V^{\triangle}/1^M = V^+.$$

If $t > 0$, however, the bonus for optimism becomes close to zero as $\lim_{M\to\infty} C/(t+1)^M = 0$ and the effect of the augmentation vanishes fast. This bonus for optimism is motivated by deep learning models, where it is hard to ensure optimistic initialization, but an addition of an extra summand is easier to implement [21]. As deep learning represents an interesting area of study, we choose to keep the bonus for optimism in our model and allow arbitrary initialization. We allow this bonus for optimism $v_h(t)$ to differ with time step h, and therefore define the *augmented Q-values* and *augmented values* as

$$\bar{Q}_h(x) \triangleq Q_h(x) + v_h(t), \quad \bar{V}_h(s) \triangleq \min\{V^+_h, [\mathcal{M}_h\bar{Q}_h](s)\}. \quad (10)$$

For exploration, all of the models store UCB Q-values and explore greedily based on them. Compared to regular Q-learning, these Q-values include an additional confidence bonus $b(t)$ in their updates $U^k_h(x,s) \triangleq r_h(x) + \gamma[\mathcal{M}_{h+1}Q^k_{h+1}](s) + b(t)$. The goal of this bonus is to ensure that the learned Q-values $Q^k_h(x)$ are the UCB-estimates of the optimal Q-values $Q^\star_h(x)$. The exact form of the bonus depends on which concentration inequalities are used in the method's design. These concentration inequalities provide probabilistic bounds on the total regret, and the bonuses are carefully crafted to ensure that the resulting bounds hold with high probability $1-\delta$. Instead of designing bonuses to guarantee the probability that regret bound holds, we do the reverse, that is, we allow arbitrary bonuses $b_h(t)$, and see how they affect the probability δ.

Additionally, we introduce a *cumulative confidence bonus* $\beta_h(t)$:

$$\beta_h(t) \triangleq \textstyle\sum_{i=1}^t \alpha^i_t b_h(i).$$

We choose the cumulative bonus form as it simplifies presentation of the theoretical results. For example, it allows us to define the *total cumulative bonus*

$$\vartheta_h(t) \triangleq \beta_h(t) + v_h(t),$$

which represents all of the optimistic bias of an algorithm and which plays an important role in our analysis.

Summarizing the aforementioned, a generalization of the UCB-based methods should include two kind of bonuses: a bonus for optimism $v_h(t)$ and a confidence bonus $b_h(t)$ (or its cumulative form $\beta_h(t)$), and use the augmented Q-values $\bar{Q}_h(x)$.

3.2 Generalized Optimistic Q-Learning

Following the discussion of Section 3.1, the existing sample-efficient optimistic Q-learning methods differ with respect to three hyperparameters: initial Q-values Q^0_h, bonus for optimism $v_h(t)$, and cumulative confidence bonus $\beta_h(t)$. We unify these methods into a single algorithm, which we name *Generalized optimistic Q-learning*. It is presented in Algorithm 1. Table 1 summarizes how the existing methods fit into the generalized optimistic Q-learning framework.

Algorithm 1 has two extra hyperparameters, a learning rate α_t and an exploration rate ϵ. It is shown in [14] that the learning rate $\alpha_t = (H+1)/(H+t)$ offers significant improvements in performance compared to previously considered rates $\alpha_t = t^{-1}$ and $t^{-\omega}$, where $0.5 < \omega \leq 1$ is a constant. Therefore, it is possible that other learning rates may offer similar, or even better improvements.

We want generalized optimistic Q-learning to be as general as (reasonably) possible, so we include the exploration rate ϵ as a parameter. This allows us to represent several other methods in our framework as well, as shown at the top of Table 1. In our theoretical study, however, we assume greedy action selection, that is, $\epsilon = 0$, as is the case for all variants of UCB, and we leave the analysis of regret for $\epsilon > 0$ as an interesting future direction.

Following the discussion of Section 3.1, we would like to point out that the update equation (5) of Algorithm 1 uses a slightly different update term (see step 9) by adding a bonus term $v_h(t)$:

$$U^k_h(x,s) \triangleq r_h(x) + b_h(\#^k_h(x)) + \gamma\bar{V}^k_{h+1}(s), \quad \text{where} \quad (11)$$

$$\bar{V}^k_{h+1}(s) \triangleq \min\{V^+_{h+1}, [\mathcal{M}_{h+1}\bar{Q}^k_{h+1}](s)\} \quad \text{and} \quad (12)$$

$$\bar{Q}^k_h(x) \triangleq Q^k_h(x) + v_h(\#^k_h(x)). \quad (13)$$

New optimistic model-free RL algorithms can be expressed by Algorithm 1 with different hyperparameter combinations. Below we present a novel algorithm, which is designed using this framework.

Example 3.1 (UCB-H with generalized learning rate, UCB-H$^+$). UCB-H$^+$ follows the flow of Algorithm 1 with the hyperparameters presented in the last row of Table 1. In particular, UCB-H$^+$ utilizes a new learning rate

$$\alpha_t \triangleq \tfrac{\lambda H+1}{\lambda H+t^\omega}, \quad \text{where } \lambda \geq 0 \text{ and } \tfrac{1}{2} < \omega \leq 1. \quad (14)$$

The learning rate of UCB-H$^+$ generalizes the previously used learning rates, complies with the learning rate conditions (7), and is motivated by two observations. Firstly, for the discounted problems the learning rate $t^{-\omega}$ outperforms $1/t$, and the best performance is achieved for $\omega \approx 0.8$ [3, 9]. Secondly, switching from $\alpha_t = 1/t$ to $(H+1)/(H+t)$ allowed Jin et al. to bound the regret blow-up with respect to H and achieve efficiency [14]. We would like to note that our generalized framework does not rely on this particular learning rate, instead, this example serves as an illustration of its use.

The generality of our framework complicates the theoretical analysis of Algorithm 1. To achieve interesting, interpretable results, we need to impose at least some conditions on the hyperparameters of the model. We would like to point out that none of these conditions are particularly restrictive, and they (sometimes trivially) hold for all of the existing optimistic methods, albeit not being explicitly mentioned. At the same time, these conditions encompass a broader class of models, including the aforementioned UCB-H$^+$.

3.2.1 Conditions on the learning rate. We start with conditions on the learning rate α_t. By inspection of various proofs involving the learning rates presented in Table 1, we identified that their successful application can be attributed to the following condition.

Condition 1. *The learning rate satisfies* $\alpha_1 = 1$.

Intuitively, Condition 1 means that when a state-action pair is visited for the first time, the update equation becomes $Q^k = (1-\alpha_1)Q^0 + \alpha_1 U = U$, and the initial value Q^0 becomes "forgotten", being replaced by a UCB-based update U. Thus, under a condition $\alpha_1 = 1$ the initialization affects the optimistic view of unencountered state-action pairs only.

Algorithm 1: Generalized optimistic Q-Learning

```
Data: episodic NS-MDP $\mathbf{M}$, initial Q-values $Q_h^0$, bonuses $\upsilon_h(t)$ and $\beta_h(t)$, learning rate $\alpha_t$, and exploration rate $\epsilon$.
Initialize Q-table $Q_h(x) \leftarrow Q_h^0$ and visitation counter $\#_h(x) \leftarrow 0$ for all $h \in \mathbb{H}, x \in \mathbb{X}_h$;
for episode $k \leftarrow 1, \ldots, K$ do
    observe initial state $s_1$;
    for step $h \leftarrow 1, \ldots, H$ do
        take action $a_h \leftarrow \textsc{Greedy}_\epsilon(\bar{Q}_h, s_h)$;    ▹ where $\bar{Q}_h(x) \triangleq Q_h(x) + \upsilon_h(t)$
        receive reward $r_h$, observe next state $s_{h+1}$, and let $x_h = (s_h, a_h)$ denote the current state-action pair;
        increment visitation counter $t = \#_h(x_h)$ by 1;
        compute confidence bonus $b_h(t) \leftarrow \alpha_t^{-1}\beta_h(t) + (1 - \alpha_t^{-1})\beta_h(t-1)$;
        compute update $U_h(x_h, s_{h+1}) \leftarrow r_h(x_h) + b_h(t) + \gamma \bar{V}_{h+1}(s_{h+1})$;    ▹ where $\bar{V}_h(s) \triangleq \min\{V_h^+, [\mathcal{M}_h \bar{Q}_h](s)\}$
        update Q-table $Q_h(x_h) \leftarrow (1 - \alpha_t)Q_h(x_h) + \alpha_t U_h(x_h, s_{h+1}, t)$;
```

Q-learning variant		Q_h^0	α_t	ϵ	$\upsilon_h(t)$	$\beta_h(t)$	regret
Regular	[9, 27]	any	$t^{-\omega}$	ϵ	0	0	$\Omega_{H,X}(T)$
Optimistic	[10]	V^+/α_T^0	$t^{-\omega}$	ϵ	0	0	?
Speedy	[3]	any	t^{-1}	ϵ	0	$\sum_{i=1}^t \mathcal{P}_i(Q_i - Q_{i-1})$	$\tilde{\mathrm{O}}_{H,X}(T^{2/3})$
UCB-H	[14]	V^+	$\frac{H+1}{H+t}$	0	0	$c_1 H \sum_{i=1}^t \alpha_i \sqrt{H\iota/i}$	$\tilde{\mathrm{O}}(H^2\sqrt{TX})$
UCB-B	[14]	V^+	$\frac{H+1}{H+t}$	0	0	$\frac{1}{2}\min\{c_1(\sqrt{H(W_t + H)\iota/t} + \sqrt{H^7 X}\iota/t), c_2\sqrt{H^3\iota/t}\}$	$\tilde{\mathrm{O}}(H\sqrt{HTX})$
∞-UCB	[26]	V^+	$\frac{H+1}{H+t}$	0	0	$c_1(1-\gamma)^{-1}\sum_{i=1}^t \alpha_i \sqrt{H\iota/i}$	$\tilde{\mathrm{O}}_H(\sqrt{TX})$
OPIQ	[21]	V^-	$\frac{H+1}{H+t}$	0	$\frac{C}{(t+1)^M}$	$c_1 H \sum_{i=1}^t \alpha_i \sqrt{H\iota/i}$	$\tilde{\mathrm{O}}(H^2\sqrt{TX})$
UCB-H$^+$	[this]	V_h^+	$\frac{\lambda H+1}{\lambda H+t^\omega}$	0	0	$c\gamma V_{h+1}^\triangle \sqrt{(\lambda H + t^\omega)^{-1}(\lambda H + 1)\iota}$	$\tilde{\mathrm{O}}\big(\mu\sqrt{H^{\omega-1}T^{2-\omega}X^\omega}\big)$

Table 1: Different Q-Learning algorithms as generalized optimistic Q-learning. Below the line are provably efficient methods.

Iterative approximation of optimal Q-values via equation (9) leads to a scaling factor of $\sum_{i=1}^t \alpha_t^i$. As the learning process is stochastic, we want to ensure that its variance remains bounded similarly to equations (7). Moreover, as UCB depends on this variance, we need to be able to quantify it in order to compare the bonus terms we use to the actual confidence bounds. This observation leads us to the following condition.

Condition 2. *There exists a function* $0 \le \zeta(t) \le 1$ *such that*

$$\textstyle\sum_{i=1}^t (\alpha_t^i)^2 \le \zeta^2(t).$$

Next, to quantify the total regret, we need to be able to express its propagation from one time step to another; we see from Corollary 4.8 that the total regret inflates by a factor of $\gamma\eta(H, K)$ with each step, where $\eta(H, K)$ satisfies the following condition.

Condition 3. *There exists a function* $\eta(H, K) \ge 1$ *such that*

$$\textstyle\sum_{n=t}^K \alpha_n^t \le \eta(H, K).$$

Knowing the learning rate, it is possible to express η analytically.[1] For example, Jin et al. show that $\sum_{n=t}^\infty \alpha_n^t \le 1 + 1/H = \eta(H)$ in their analysis [14], which implies Condition 3. However, without any assumptions on the form of the learning rate, we have to fall back to η as a generalized term.

Function η serves as a "scaling factor" for the total regret, but there are other scale parameters, for example, the discounting factor, the lower r_h^- and the upper r_h^+ reward functions affect the total regret scale as well. We want to be able to quantify their effect and combine all of the scale parameters together as follows.

Condition 4. *Let* $V_h^\uparrow$ *denote the asymptotically dominant term between the upper value function* V_h^+ *and the value span* $V_h^\triangle$*, that is,*

$$V_h^\uparrow \triangleq \begin{cases} V_h^\triangle & \text{if } V_h^+ = \mathrm{O}(V_h^\triangle), \\ V_h^+ & \text{otherwise,} \end{cases}$$

and similarly for the reward bound $r^\uparrow(H)$ *and the value bound* $V^\uparrow(H)$*. Then there exists a function* $\mu(H, K, \gamma)$ *such that*

$$\textstyle\sum_{h=1}^H (\gamma\eta)^{h-1} V_h^\uparrow = \mathrm{O}\big(\mu(H, K, \gamma)\big). \qquad (15)$$

We call the function μ of Condition 4 the *magnitude function*, because it quantifies the asymptotic behavior of the total regret blowup in all H time steps. Intuitively, regret of each time step is at most $V^\triangle = \mathrm{O}(V^\uparrow)$, which means that the total regret grows at most at a rate of $\sum_{h=1}^H (\gamma\eta)^{h-1} V_h^\uparrow$ as H grows.

All of the existing UCB-based methods utilize the same learning rate $\alpha_t = (H+1)/(H+t)$ as showed in Table 1. It is easy to check that this learning rate satisfies Conditions 1–4. In particular, $\zeta(t) = 2H/t$ and $\eta(H) = 1 + 1/H$ are proposed by Jin et al. and used by other authors [14, 21, 26]. Due to the fact that $(1+1/H) < e$, the magnitude function equal to $\mu(H) = V^\uparrow = H$ is used.

[1] We omit the arguments of η and other functions introduced later for brevity of notation, if it does not lead to ambiguity.

3.2.2 Conditions on the bonuses. All of the remaining conditions are rather intuitive. The first one addresses the initialization and was already discussed in Section 3.1. We require that the initial values are not too high or too low, and that the augmented initial values $\bar{Q}_h^0$ used in action selection are optimistic.

Condition 5. *The initial values Q_h^0 belong to intervals $[V_h^-, V_h^+]$, and the bonus for optimism $\upsilon_h(t)$ is such that $Q_h^0 + \upsilon_h(0) \geq V_h^+$.*

Finally, we present two conditions (6 and 7) on the bonuses.

Condition 6. *The total bonus function is non-negative and non-increasing in t, $\vartheta_h(t) \geq \vartheta_h(t+1) \geq 0$ for all $t \in \mathbb{N}$.*

As t represents the number of visitations of a state-action pair, we want the bonus to decrease as it grows, that is, as we collect more samples and build higher confidence. Non-negativity ensures that the bonuses are optimistic.

Condition 7. *There exists a function $\theta(t)$ such that*

$$\textstyle\sum_{n=1}^{t} \vartheta_h(n) \leq \mathrm{O}\big(V_h^{\uparrow}\theta(t)\big).$$

This condition is used to quantify the effect of the total bonus $\vartheta_h(t)$ on the regret by a function $\theta(t)$, similarly to how the magnitude function μ quantifies the other effects. We call this function the *bonus scaling* function.

The existing methods satisfy Conditions 5 and 6 trivially. Condition 7 depends on the particular bonus design, and also holds for all of the methods. For example, UCB-H and OPIQ both use $\theta(t) = \sqrt{Ht\iota}$ as the bonus scaling function, although implicitly.

3.3 The Total Regret Bound

Finally, we are ready to give a high-probability bound on the total regret, which is our main theoretical contribution. The total regret is bounded by the sum of three different terms, each amplified by the *magnitude* function μ of Condition 4. These terms are:

- the *size* of the state-action space X,
- the total effect of the *bonuses* $B \triangleq X\theta(K/X)$, which depends on the bonus scaling function of Condition 7, and
- the total effect of the *estimation error* $E \triangleq c\sqrt{K\iota}$, where $\iota \triangleq \ln(TX/\delta)$ is the logarithmic term.

The state-action space size X represents the effect of the *optimistic initialization*, as the number of initial values is proportionate to X. The bonus effect B relates to *optimistic action selection.*

The third factor E is caused by replacing the unknown transition operator (2) with its empirical counterpart (6). The constant c depends on how much uncertainty there is in the transitions, and is formally introduced later. An important property is that for deterministic problems $c = 0$, and the estimation term disappears. The probability δ used in the estimation error term E depends on our confidence in the total regret bound, that is, the bound holds w.p. at least $1 - 2\delta$. It depends on the choice of the cumulative confidence bonus $\beta_h(t)$ as follows:

$$\delta = \begin{cases} 2KX \sum_{h \in \mathbb{H}} \exp\left(-\frac{1}{2}\left(\frac{\beta_h(t)}{\gamma c V_{h+1}^{\triangle}\zeta(t)}\right)^2\right) & \text{if } c > 0, \\ 0 & \text{if } c = 0, \end{cases} \tag{16}$$

Theorem 3.2 formalizes these results.

Theorem 3.2. *Let Conditions 1–7 hold. Then for some constant $0 \leq c \leq 1$, w.p. at least $1 - 2\delta$ the total regret of generalized optimistic Q-learning with no exploration (i.e., when $\epsilon = 0$) is bounded by*

$$R(\mathbf{M}, \alpha, \vartheta) = \mathrm{O}\big(\mu(X + B + E)\big), \tag{17}$$

If there are no random transitions in the NS-MDP, the learning process becomes fully deterministic as well (we assume no random exploration). This leads us to the following corollary.

Corollary 3.3. *If the transitions of the underlying NS-MDP $\mathbf{M}$ are deterministic, the total effect of the estimation error is equal to zero, $E = 0$. Moreover, the bound of Theorem 3.2 holds w.p. 1.*

4 PROOF OF THEOREM 3.2

We prove Theorem 3.2 by using a recurrent decomposition of the regret of a time step h in terms of the next time step $h+1$. We bound the regret of each time step using the differences between augmented Q-values $\bar{Q}_h(x)$ of generalized optimistic Q-learning and the optimal Q-values $Q_h^{\star}(x)$, provided by Lemma 4.5. To derive these bounds, we employ some properties of the learning rate.

4.1 Properties of the Learning Rate

We prove two lemmas, both relying on Condition 1 only.

Lemma 4.1. *If $\alpha_1 = 1$, then*

- *$\alpha_t^0 = 0$ and $\sum_{i=1}^{t} \alpha_t^i = 1$ for $t \geq 1$;*
- *$\sum_{i=0}^{t} \alpha_t^i = 1$ for any $t \geq 0$.*

Proof. By definition, $\alpha_t^0 = (1 - \alpha_1) \cdot \prod_{j=2}^{t}(1 - \alpha_j) = 0$.

We prove that $\sum_{i=1}^{t} \alpha_t^i = 1$ by induction. For $t = 1$, $\sum_{i=1}^{t} \alpha_t^i = \alpha_1 = 1$. Assume that $\sum_{i=1}^{t} \alpha_t^i = 1$. Then using the definition of α_t^i,

$$\begin{aligned} \textstyle\sum_{i=1}^{t+1}\alpha_{t+1}^i &= \textstyle\sum_{i=1}^{t}\alpha_i\prod_{j=i+1}^{t+1}(1-\alpha_j) + \alpha_{t+1} \\ &= \Big(\textstyle\sum_{i=1}^{t}\alpha_i\prod_{j=i+1}^{t}(1-\alpha_j)\Big)(1-\alpha_{t+1}) + \alpha_{t+1} \end{aligned}$$

where the expression in the first brackets is equal to $\sum_{i=1}^{t} \alpha_t^i = 1$ by the induction hypothesis, and therefore $\sum_{i=1}^{t+1} \alpha_{t+1}^i = 1$.

The second statement follows trivially from the first for $t \geq 1$ and from the definition of α_t^i for $t = 0$. □

Lemma 4.1 allows us to write $Q_h^{\star}(x) = \sum_{i=1}^{t} \alpha_t^i Q_h^{\star}(x)$ similarly to the decomposition (9) of $Q_h^k(x)$ in order to relate them to each other.

We also prove the following relation between the confidence bonus $b(t)$ and the cumulative confidence bonus $\beta(t)$, justifying our choice of the bonus in step 8 of Algorithm 1.

Lemma 4.2. *If $b(t) \triangleq \alpha_t^{-1}\beta(t) + (1-\alpha_t^{-1})\beta(t-1)$ for some function $\beta(t)$, and either $\alpha_1 = 1$ or $\beta(0) = 0$, then $\sum_{i=1}^{t} \alpha_t^i b(i) = \beta(t)$.*

Proof. By induction. For $t = 1$, $\sum_{i=1}^{1} \alpha_1^i b(i) = \alpha_1 b(1) = \beta(1) + (\alpha_1 - 1)\beta(0) = \beta(1)$. Assume $\sum_{i=1}^{t} \alpha_t^i b(i) = \beta(t)$ for some t. Then

$$\begin{aligned} \textstyle\sum_{i=1}^{t+1}\alpha_{t+1}^i b(i) &= \textstyle\sum_{i=1}^{t}\alpha_{t+1}^i b(i) + \alpha_{t+1}b(t+1) = (1-\alpha_{t+1})\beta(t) \\ &\quad + \alpha_{t+1}b(t+1) = (1-\alpha_{t+1})\beta(t) + \beta(t+1) \\ &\quad + \alpha_{t+1}(1-\alpha_{t+1}^{-1})\beta(t) = \beta(t+1). \end{aligned}$$ □

4.2 Bounds on Q-Value Differences

First, we show that the augmented Q-values $\bar{Q}_h(x)$ are related to the augmented values $\bar{V}_{h+1}(s)$ of previous episodes as follows.

Lemma 4.3 (recursion on $\bar{Q}$, generalization of Lemma 4.2 of [14]). *For any step $h \in \mathbb{H}$, state-action pair $x = (s, a) \in \mathbb{X}_h$ and episode $k \in \mathbb{K}$, let $t \triangleq \#_h^k(x)$ and suppose that for state s action a was previously taken in time step h of episodes $k_1, \dots, k_t < k$. Then under Condition 1*

$$[\bar{Q}_h^k - Q_h^\star](x) = \alpha_t^0[Q_h^0 - Q_h^\star](x) + \textstyle\sum_{i=1}^t \alpha_t^i \Big(\gamma[\bar{V}_{h+1}^{k_i} - V_{h+1}^\star](s_{h+1}^{k_i}) + \gamma\big[(\hat{\mathcal{P}}_h^{k_i} - \mathcal{P}_h)V_{h+1}^\star\big](x)\Big) + \vartheta_h(t). \tag{18}$$

Proof sketch. Similarly to the proof of Lemma 4.2 of [14], we use equations (13) and (9) to express $\bar{Q}_h^k(x)$ in terms of the initial values Q_h^0. Then we apply Lemma 4.1 and the Bellman optimality equation (4) to do a similar decomposition for $Q_h^\star(x)$. □

Next, we introduce the parameter c that quantifies the difference between the empirical transition operator (6) and the true transition operator (2), both of which appear in the equation (18).

Proposition 4.4. *Let $f(x) : \mathbb{X}_{h+1} \to [a, b]$. There exists a constant $0 \le c \le 1$ such that $c(a - b) \le \big[(\hat{\mathcal{P}}_h^k - \mathcal{P}_h)f\big](x) \le c(b - a)$.*

Remark 4.1. Note that while the case $c = 1$ holds trivially for any problem, a smaller constant possibly exists. For example, if the transitions of an NS-MDP **M** are not random, operators $\hat{\mathcal{P}}_h^k$ and $\mathcal{P}_h$ coincide and $c = 0$ provides a sharper bound.

Using Proposition 4.4 and Lemma 4.3, we bound the difference between the augmented Q-values $\bar{Q}_h^k(x)$ and the optimal Q-values $Q_h^\star(x)$. The bound consists of four summands, three of which correspond to the three factors of the total regret discussed in Section 3.3. The fourth term, $\gamma\Delta_h\zeta(t)$, disappears from the regret bound because it is asymptotically dominated by the total bonus $\vartheta_h(t)$.

Lemma 4.5 (bound on $\bar{Q}^k - Q^\star$, generalization of Lemma 3 of [21]). *Let Conditions 1, 2, 5, and 6 hold. Given constants $\delta_h > 0$ such that $\beta_h(t) \ge \gamma\Delta_h\zeta(t)$, where $\Delta_h \triangleq cV_{h+1}^\triangle\sqrt{2\ln(2/\delta_h)}$, and c is a constant from Proposition 4.4, the following holds with probability at least $1 - \delta$, where $\delta \triangleq KX\sum_{h\in\mathbb{H}}\delta_h$:*

$$0 \le [\bar{Q}_h^k - Q_h^\star](x) \le \alpha_t^0(Q_h^0 - V_h^-) + \gamma\textstyle\sum_{i=1}^t \alpha_t^i[\bar{V}_{h+1}^{k_i} - V_{h+1}^\star](s_{h+1}^{k_i}) + \vartheta_h(t) + \gamma\Delta_h\zeta(t). \tag{19}$$

Proof sketch. Let $Y_t^i(x) \triangleq \alpha_t^i\big[(\hat{\mathcal{P}}_h^{k_i} - \mathcal{P}_h)V_{h+1}^\star\big](x)$. Note that $\big|Y_t^i(x)\big| \le \alpha_t^i cV_{h+1}^\triangle$. Follow the argument of the proof of Lemma 4.3 of [14], we apply the Azuma–Hoeffding inequality [16, Theorem 3.13] to see that w.p. at least $1 - \delta$

$$\big|\textstyle\sum_{i=1}^t Y_t^i(x)\big| \le \sqrt{2\textstyle\sum_{i=1}^t(\alpha_\tau^i cV_{h+1}^\triangle)^2\ln(2/\delta_h)} \le \Delta_h\zeta(t), \tag{20}$$

for all $x \in \mathbb{X}$, $h \in \mathbb{H}$, and $k \in \mathbb{K}$. The r.h.s. of inequality (19) follows from Lemma 4.3 and the fact that $Q_h^\star(x) \ge V_h^-$. The l.h.s. proof follows the existing proof [21] using equation (20). □

A direct consequence of Lemma 4.5 is that for an arbitrary chosen bonus function we can lower-bound the probability that inequalities (19) hold (note that sometimes the bound can be zero though).

Corollary 4.6. *Under Conditions 1, 2, 5, and 6, for an arbitrary chosen cumulative confidence bonus function $\beta_h(t)$, inequalities (19) hold w.p. at least $1 - \delta$, where δ is given by (16) for c introduced in Proposition 4.4.*

Proof. The special case $c = 0$ trivially follows from Condition 6 and Lemma 4.5. Otherwise δ can be obtained by solving $\beta_h(t) = c\gamma V_{h+1}^\triangle\zeta(t)\sqrt{2\ln(2/\delta_h)}$ for δ_h. □

4.3 Properties of the Total Regret

We are now ready to provide an upper bound on total regret of generalized optimistic Q-learning using the results of the previous sections. We start by introducing the following proposition, generalizing the arguments used in the literature [14, 21].

Proposition 4.7 (recursion on total regret bound). *Denote*

$$\psi_h^k \triangleq [\bar{V}_h^k - V_h^{\pi_k}](s_h^k), \quad \xi_h^k \triangleq \big[(\hat{\mathcal{P}}_h^k - \mathcal{P}_h)(\bar{V}_{h+1}^k - V_{h+1}^\star)\big](x_h^k).$$

Let Conditions 1–3, 5 and 6 hold. Using notation of Lemma 4.5, the following two statements hold w.p. at least $1 - \delta$:

(1) *the total regret R is upper-bounded by $R \le \sum_{k=1}^K \psi_1^k$.*
(2) *for any $h \in \mathbb{H}$ and $k \in \mathbb{K}$, ψ_h^k is upper-bounded by*

$$\psi_h^k \le \gamma\eta\psi_{h+1}^k + \Psi_h^k(t), \quad \text{where} \tag{21}$$

$$\Psi_h^k(t) \triangleq \alpha_t^0(Q_h^0 - V_h^-) + \vartheta_h(t) + \gamma\big(\Delta_h\zeta(t) + \xi_h^k\big). \tag{22}$$

Next, applying the bounds (21) iteratively on $h = 1, 2, \dots, H + 1$ and noticing that $\psi_{H+1}^k = 0$ by equations (3) and (1), we bound R.

Corollary 4.8. *Under Conditions 1–3, 5 and 6 w.p. at least $1 - \delta$ the total regret is upper-bounded by*

$$R \le \textstyle\sum_{k=1}^K\sum_{h=1}^H(\gamma\eta)^{h-1}\Psi_h^k(t), \tag{23}$$

where δ and $\Psi_h^k(t)$ are given by equations (16) and (22).

Finally, we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. We study the right-hand side of inequality (23) by rewriting it as

$$R(K) \le \rho_K\big(\alpha_t^0(Q_h^0 - Q_h^-)\big) + \rho_K\big(\vartheta_h(t)\big) + \gamma\rho_K\big(\Delta_h\zeta(t)\big) + \gamma\rho_K(\xi_h^k),$$

where $\rho_K\big(g_h^k(t)\big) \triangleq \sum_{h=1}^H(\gamma\eta)^{h-1}\sum_{k=1}^K g_h^k(t)$.

For the first element $\rho_K\big(\alpha_t^0(Q_h^0 - Q_h^-)\big)$, by changing the summation order and using the fact that $Q_h^0 - Q_h^- \le V_h^\triangle$ we write

$$\rho_K\big(\alpha_t^0(Q_h^0 - Q_h^-)\big) \le \textstyle\sum_{k=1}^K\sum_{h=1}^H(\gamma\eta)^{h-1}\alpha_t^0 V_h^\triangle.$$

In this sum $\alpha_t^0 = \mathbb{I}[t = 0]$ by Lemma 4.1 and $\alpha_0^0 = 1$. In this sum, $\mathbb{I}\big[\#_h^k(x_h^k) = 0\big] \ne 0$ means that x has never been visited in step h before episode k, and the number of such state-action pairs is $\mathrm{O}(X)$ independent of K and H; therefore, we have $\mathrm{O}(X)$ summands $(\gamma\eta)^{h-1}V_h^\triangle$, and each of them is $\mathrm{O}(\mu)$, so $\rho_K(\alpha_t^0 V_h^\triangle) = \mathrm{O}(\mu X)$.

For $\rho_K(\xi_h^k)$ we use the fact that $\{\xi_h^k\}_{k\in\mathbb{K}}$ is a martingale difference sequence [14, proof of Theorem 1]. Note that $V_{h+1}^- \le V_{h+1}^\star(x) \le \bar{V}_{h+1}^k(x) \le V_{h+1}^+$, therefore $[\bar{V}_{h+1}^k - V_{h+1}^\star](x) \in [0, V_{h+1}^\triangle]$. Using these bounds, Proposition 4.4, an argument similar to the proof of Lemma 4.5, and Azuma–Hoeffding inequality, we see that w.p. at least $1 - \delta$

$$\left|\sum_{k=1}^{K} \xi_h^k\right| \leq \sqrt{2\sum_{k=1}^{K}(cV_{h+1}^{+})^2 \ln \tfrac{2HX}{\delta}} = \mathrm{O}\left(cV_{h+1}^{+}\sqrt{K \ln \tfrac{HX}{\delta}}\right),$$

for all $h \in \mathbb{H}$ and $x \in \mathbb{X}$. Note that $\ln(HX)/\delta = \mathrm{O}(\iota)$, therefore

$$\rho_K(\xi_h^k) = \mathrm{O}\left(c\sum_{h=1}^{H}(\gamma\eta)^{h-1}V_{h+1}^{+}\sqrt{K\iota}\right) = \mathrm{O}\big(c\mu\sqrt{K\iota}\big) = \mathrm{O}\big(\mu E\big).$$

Finally, for the last two terms we notice that $\vartheta_h(t) \geq \gamma\Delta_h\zeta(t) \geq 0$ and thus $\vartheta_h(t)$ is the asymptotically dominant term, that is, $\Delta_h\zeta(t) = \mathrm{O}\big(\vartheta_h(t)\big)$. We write

$$\rho_K\big(\vartheta_h(t)\big) = \sum_{h=1}^{H}(\gamma\eta)^{h-1}\sum_{k=1}^{K}\vartheta_h\big(\#_h^k(x_h^k)\big).$$

First, we consider the inner sum $\Sigma_h^{\vartheta} \triangleq \sum_{k=1}^{K} \vartheta_h\big(\#_h^k(x_h^k)\big)$. Instead of summing in order of episodes $k \in \mathbb{K}$, we can sum the total bonuses $\vartheta_h\big(\#_h^k(x_h^k)\big)$ separately for each state-action pair $x \in \mathbb{X}_h$ first, and add all visitations $n = 1, \ldots, \#_h^K(x)$ of x in all episodes. This yields

$$\Sigma_h^{\vartheta} = \sum_{x\in\mathbb{X}_h}\sum_{n=1}^{\#_h^K(x)}\vartheta_h(n) \quad \text{where } \sum_{x\in\mathbb{X}_h}\#_h^K(x) = K.$$

Because $\vartheta_h(t)$ is decreasing in t by Condition 6, Σ_h^{ϑ} is maximized when as many state-action pairs x are visited, which happens when $\#_h^K(x) = K/X$ for all $x \in \mathbb{X}$:

$$\Sigma_h^{\vartheta} \leq \sum_{x\in\mathbb{X}}\sum_{n=1}^{K/X}\vartheta_h(n) = X\sum_{n=1}^{K/X}\vartheta_h(n) = \mathrm{O}\big(V_h^{\uparrow}X\theta(K/X)\big),$$

where $\theta(t)$ is defined in Condition 7. Thus, $\rho_K\big(\vartheta_h(t)\big) = \mathrm{O}(\mu B)$.

Adding the three factors together, the bound (17) holds with probability at least $1 - 2\delta$. □

5 DESIGNING A NEW UCB-BASED METHOD

In this section, we apply Theorem 3.2 to prove efficiency of UCB-H$^+$ presented in Example 3.1. We show how the proposed generalized learning rate (14) satisfies the required condition, and how the bonus design is based on it. We only consider the case $\lambda > 0$, as inclusion of H is required to achieve sub-linear regret [14], but similar analysis can be performed for $\lambda = 0$, yielding worse bounds.

Conditions on the Learning Rate. First, we want to ensure that the generalized learning rate (14) satisfies the Conditions 1–4. Condition 1 holds trivially. We now show that so do the other ones.

Proposition 5.1. *$t^{\omega} + j \geq (t + j)^{\omega}$ for any $t \in \mathbb{N}_0$ and $j \in \mathbb{N}_0$.*

Lemma 5.2. *For the generalized learning rate given by equation* (14), *Condition 2 holds with $\zeta(t) = \sqrt{(\lambda H + 1)/(\lambda H + t^{\omega})}$.*

Proof. Notice that $\sum_{i=1}^{t}(\alpha_t^i)^2 \leq \max_{i=1}^{t} \alpha_t^i \cdot \sum_{i=1}^{t} \alpha_t^i$, which by Lemma 4.1 is equal to $\max_{i=1}^{t} \alpha_t^i$. By definition,

$$\begin{aligned}\alpha_t^i &= \tfrac{\lambda H+1}{\lambda H+i^{\omega}}\left(\tfrac{(i+1)^{\omega}-1}{\lambda H+(i+1)^{\omega}} \cdot \tfrac{(i+2)^{\omega}-1}{\lambda H+(i+2)^{\omega}} \cdots \tfrac{t^{\omega}-1}{\lambda H+t^{\omega}}\right)\\ &= \tfrac{\lambda H+1}{\lambda H+t^{\omega}}\left(\tfrac{(i+1)^{\omega}-1}{\lambda H+i^{\omega}} \cdot \tfrac{(i+2)^{\omega}-1}{\lambda H+(i+1)^{\omega}} \cdots \tfrac{t^{\omega}-1}{\lambda H+(t-1)^{\omega}}\right).\end{aligned}$$

By Proposition 5.1 for $j = 1$ each fraction in the brackets is less than 1, so $\alpha_t^i \leq (\lambda H + 1)/(\lambda H + t^{\omega}) \triangleq \zeta^2(t)$. □

Proposition 5.3 (c.f. equation B.1 of [14]). *For any $m \geq k$,*

$$\tfrac{m}{k} = 1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} \tfrac{m-k+j-1}{m+j}.$$

Lemma 5.4. *For the learning rate given by equation* (14), *Condition 3 holds with $\eta(H) = 1 + (\lambda H)^{-1}$ if $\lambda > 0$.*

Proof. By Proposition 5.3 with $m = \lambda H + t^{\omega}$ and $k = \lambda H$,

$$\begin{aligned}\sum_{n=t}^{K}\alpha_n^t \leq \sum_{n=t}^{\infty}\alpha_n^t &= \alpha_t\left(1 + \sum_{i=1}^{\infty}\prod_{j=1}^{i}(1 - \alpha_{t+j})\right)\\ &\leq \tfrac{\lambda H+1}{\lambda H+t^{\omega}}\left(1 + \sum_{i=1}^{\infty}\prod_{j=1}^{i}\tfrac{t^{\omega}+j-1}{\lambda H+t^{\omega}+j}\right)\\ &= \tfrac{\lambda H+1}{\lambda H+t^{\omega}}\,\tfrac{\lambda H+t^{\omega}}{\lambda H} = 1 + \tfrac{1}{\lambda H},\end{aligned}$$

where the second inequality holds by Proposition 5.1, because

$$1 - \alpha_{t+j} = \tfrac{(t+j)^{\omega}-1}{\lambda H+(t+j)^{\omega}} \leq \tfrac{t^{\omega}+j-1}{\lambda H+t^{\omega}+j}. \qquad \square$$

Lemma 5.5. *For the generalized learning rate Condition 4 holds with $\mu(H, V^{\uparrow}, \gamma) = HV^{\uparrow}$ if $\gamma = 1$ and $V^{\uparrow}/(1 - \gamma)$ otherwise.*

We omit the proof of Lemma 5.5. It is straightforward as the sum in the definition (15) can easily be computed directly.

Conditions on the Bonuses. Lemmas 4.5 and 5.2 explain our choice of the bonuses, namely,

$$\beta_h(t) \triangleq c\gamma V_{h+1}^{\triangle}\sqrt{8\tfrac{\lambda H+1}{\lambda H+t^{\omega}}\ln\tfrac{2TX}{\delta}} \quad \text{and} \quad \upsilon_h(t) = 0 \qquad (24)$$

for the constant c of Proposition 4.4. By Corollary 4.6, Lemma 4.5 holds w.p. at least $1 - \delta$ for this cumulative bonus for any δ. Conditions 5 and 6 both hold trivially.

Lemma 5.6. *For the bonuses given by equation* (24), *Condition 7 holds with $\theta(t) = \sqrt{Ht^{2-\omega}\iota}$.*

Proof. Note that $\sum_{n=1}^{t}(\lambda H + n^{\omega})^{-1/2} \leq \sum_{n=1}^{t} n^{-\omega/2} = \mathrm{H}_t^{(\omega/2)}$, where $\mathrm{H}_n^{(r)}$ denotes the generalized harmonic number of n of order r. By Euler–Maclaurin sum [1, formula 3.6.28], for a given $r \neq 1$, $\mathrm{H}_n^{(r)} = \zeta(r) + (1 - r)^{-1}n^{1-r} + \mathrm{o}(n^{1-r}) = \mathrm{O}(n^{1-r})$. Thus

$$\sum_{n=1}^{t}\beta_h(n) = \mathrm{O}\left(\sqrt{H\iota}\sum_{n=1}^{t}\tfrac{1}{\sqrt{\lambda H+n^{\omega}}}\right) = \mathrm{O}\big(\sqrt{H\iota}\cdot t^{1-\omega/2}\big). \qquad \square$$

Regret Bound. Combining the aforementioned results, we prove the efficiency of UCB-H$^+$.

Theorem 5.7. *For any $\delta > 0$ w.p. at least $1 - \delta$ the total regret of UCB-H$^+$ with $\lambda > 0$ is bounded by $\mathrm{O}\big(\mu\sqrt{H^{\omega-1}T^{2-\omega}X^{\omega}\iota}\big)$, where the magnitude μ is given by Lemma 5.5.*

Proof. Using $\theta = \sqrt{Ht^{2-\omega}\iota}$, we write the sum in Theorem 3.2 as $X + B + E = X + \sqrt{H^{\omega-1}T^{2-\omega}X^{\omega}\iota} + c\sqrt{K\iota}$. The last term is trivially dominated by the second one, so it can be omitted. Now we show that the first term is also dominated in the total regret bound.

Assume $T \leq \sqrt{H^{1+\omega}T^{2-\omega}X^{\omega}\iota}$. The total regret is bounded by

$$\sum_{k=1}^{K}\psi_1^k \leq V^{+}K \leq V^{+}T/H = \mathrm{O}(V^{\uparrow}\sqrt{H^{\omega-1}T^{2-\omega}X^{\omega}\iota}),$$

which is dominated by the second term multiplied by μ. The opposite assumption implies that $T > H^{1+1/\omega}X\iota^{1/\omega}$, and

$$\sqrt{H^{\omega}T^{2-\omega}X^{\omega}\iota} > \sqrt{H^{\omega}\big(H^{1+1/\omega}X\iota^{1/\omega}\big)^{2-\omega}X^{\omega}\iota} \geq \sqrt{H^3X^2} > HX.$$

In either case $\mu\sqrt{H^{\omega-1}T^{2-\omega}X^{\omega}\iota}$ is the dominant term. □

6 EXPERIMENTS

To illustrate the performance of UCB-H$^+$, we consider two problems, one stochastic and one deterministic. The latter, while being less interesting in context of RL, allows us to alleviate the regret caused by the estimation error, highlighting the effect of optimism.

We start with a classical problem known as *the automobile replacement problem* [12]. This problem is based on real data and is considered as a benchmark by different authors [6, 9, 20]. In the replacement problem, the agent operates an automobile, which can be in one of the 40 states (from brand new one to ten years old, quantified quarterly). At the beginning of each quarter the agent chooses to either keep the automobile, or to replace it with a different one, which can be in any of the 40 available states. The detailed description of the problem, including transition probabilities and rewards, can be found in the original paper [12].

We consider a two-year plan (i.e., $H = 8$ steps), and $K = 62{,}500$ episodes, each starting with state $s = 1$ (i.e., a brand new car). Therefore, the problem size is equal to $HX = 13{,}120$, and the total duration of the learning is $T = 5 \times 10^5$ time steps. We assume no discounting $\gamma = 1$, and use the same values $\delta = c = 10^{-3}$ for UCB-based algorithms. As a baseline for comparison, we use regular Q-learning optimistically initialized with $Q^0_h(x) = V^+$ with an exponentially decaying exploration rate $\epsilon = 0.9999^{k-1}$ and the same learning rate $\alpha_t = (H + 1)/(H + t)$ as UCB-H. For UCB-H$^+$ we use $\omega = 0.8$ and $\lambda = 1$ as the learning rate parameters.

The experiment was repeated 50 times. The results are presented in Figure 1. The thin horizontal line represents the optimal value $V^\star$, the vertical bars show 95%-confidence intervals on mean estimates, and the ribbons show the interquartile range. Data is smoothed using a moving average with a bandwidth of $0.05K$.

This experiment shows that the total regret of Q-learning, equal to the area between the line and the optimal line above it, is 1525 ± 2 thousand dollars on average. While the plot lines may seem close to each other, UCB-H was able to achieve a regret of $1037{\pm}2$ thousands, showing a 32% reduction over the naïve approach. Finally, UCB-H$^+$ incurred a regret of 907 ± 3 thousand dollars, enjoying a reduction of 41% compared to Q-learning and 13% when compared to UCB-H. Interestingly, UCB-H$^+$ has only a slightly higher variance, which we expect to increase as the exponent ω approaches 0.5 (with $\omega = 0.5$ preventing convergence by violating conditions (7)).

Our second experiment is based on the 8×8 Frozen lake problem of OpenAI Gym [7]. The agent navigates a grid world searching for a goal state. The world has holes, stepping into one terminates the current episode. All states give no rewards, except for the goal with a reward of 1. We consider $K = 10^4$ episodes of up to $H = 16$ time steps. The problem size is $HX = 16 \times 64 \times 4 = 4{,}096$, and the total duration of the learning is $T = 1{,}6 \times 10^5$ time steps. Because this problem is simpler, we can use a faster decaying exploration rate $\epsilon = 0.99^{k-1}$ for Q-learning. For UCB-H and UCB-H$^+$ we use $c = 0$ as per Remark 4.1. The rest of the parameters remain the same.

The results are presented in Figure 2. Interestingly, UCB-H suffered from the largest regret of 5503, while Q-learning and UCB-H$^+$ achieved the regret of ≈ 4900 and 3144 respectively. UCB-H$^+$ offers a 43% improvement over UCB-H. As mentioned earlier, this problem has no stochasticity in transitions, and thus the last term of the regret is zero. Moreover, all algorithms use the same initialization, therefore, the only reasons for the performance difference is the choice of the learning rate and the optimism representation.

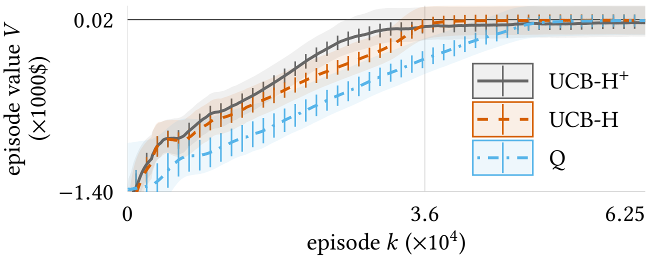

Figure 1: Replacement problem. UCB-H$^+$ offers a 41% total regret improvement over Q-learning and 13% over UCB-H.

7 CONCLUSIONS

This paper presents generalized optimistic Q-learning, a novel framework for optimistic model-free reinforcement learning that incorporates many existing methods, such as Q-learning, UCB-H, and OPIQ. We showed that under some mild conditions the total regret of optimistic model-free methods is driven by three distinct terms multiplied by the magnitude of the problem. These terms are: the size of the state-action space, the total effect of the bonuses, and the total effect of the estimation error.

To the extent of our knowledge, this is the first study of RL performance that does not rely on a particular form of the learning rate. This high level of abstraction facilitates transfer of our results to new algorithms within the generalized optimistic Q-learning framework. As an example, we present one such algorithm, UCB-H$^+$, prove its efficiency in terms of regret, and illustrate its performance in experiments. Our analysis shows that the regret is driven by the bonuses and the learning rate, therefore, their choice is a promising direction for the design of more efficient optimistic RL algorithms.

Future work includes further relaxations of the conditions used, and extensions of generalized optimistic Q-learning to other settings such as infinite-horizon non-episodic learning, deep reinforcement learning, and models with continuous state and/or action space. The algorithm UCB-H$^+$ can be extended to the continuous setting as well. One of the possible ways to do this is to employ deep Q-networks and pseudo-visitation counters similarly to [21].

ACKNOWLEDGMENTS

This research received funding from the Netherlands Organization for Scientific Research (NWO).

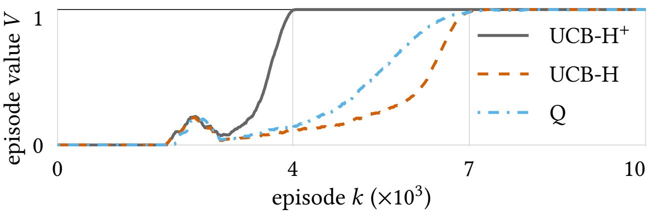

Figure 2: Frozen lake. UCB-H$^+$ offers a 43% total regret improvement over UCB-H.

REFERENCES

[1] Milton Abramowitz and Irene A. Stegun. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* US Government Printing Office, Washington, DC, USA. 1046 pages.

[2] Shipra Agrawal and Randy Jia. 2017. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 1184–1194.

[3] Mohammad Gheshlaghi Azar, Rémi Munos, Mohammad Ghavamzadeh, and Hilbert J. Kappen. 2011. Speedy Q-learning. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 2411–2419.

[4] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. 2017. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, D. Precup and Y. W. Teh (Eds.), Vol. 70. JMLR.org, Sydney, NSW, Australia, 263–272.

[5] Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. 2019. Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 8002–8011.

[6] Richard E. Bellman and Stuart E. Dreyfus. 2016. *Applied dynamic programming.* Princeton University Press, Princeton, NJ, USA. 390 pages.

[7] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI gym. (2016). arXiv:cs.LG/1606.01540

[8] Adithya M. Devraj and Sean Meyn. 2017. Zap Q-learning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 2235–2244.

[9] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. 2006. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research* 7 (Dec. 2006), 1079–1105.

[10] Eyal Even-Dar and Yishay Mansour. 2002. Convergence of optimistic and incremental Q-learning. In *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.). The MIT Press, Cambridge, MA, USA, 1499–1506.

[11] Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. 2018. Rainbow: combining improvements in deep reinforcement learning. In *The Thirty-Second AAAI Conference on Artificial Intelligence.* AAAI Press, New Orleans, LA, USA, 3215–3222.

[12] Ronald A. Howard. 1960. *Dynamic programming and Markov processes.* Technology Press of the Massachusetts Institute of Technology and Wiley, New York, NY, USA. 136 pages.

[13] Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. 1994. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector (Eds.). Morgan-Kaufmann, Burlington, MA, USA, 703–710.

[14] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. 2018. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, NY, USA, 4863–4873.

[15] Sham Kakade, Mengdi Wang, and Lin F. Yang. 2018. Variance reduction methods for sublinear reinforcement learning. (2018). arXiv:cs.AI/1802.09184

[16] Colin McDiarmid. 1998. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics. Algorithms and Combinatorics*, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed (Eds.). Vol. 16. Springer, Berlin, Heidelberg, Germany, 195–248.

[17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing Atari with deep reinforcement learning. (2013). arXiv:cs.LG/1312.5602

[18] Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. 2019. Deep exploration via randomized value functions. *Journal of Machine Learning Research* 20, 124 (2019), 1–62.

[19] Ian Osband and Benjamin Van Roy. 2017. Why is posterior sampling better than optimism for reinforcement learning?. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, D. Precup and Y. W. Teh (Eds.), Vol. 70. JMLR.org, Sydney, NSW, Australia, 2701–2710.

[20] Martin L. Puterman. 1994. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, Inc., Hoboken, NJ, USA. 672 pages.

[21] Tabish Rashid, Bei Peng, Wendelin Boehmer, and Shimon Whiteson. 2020. Optimistic exploration even with a pessimistic initialisation. In *International Conference on Learning Representations.* OpenReview.net, Addis Ababa, Ethiopia, Article 588, 28 pages. https://openreview.net/forum?id=r1xGP6VYwH

[22] Alexander L. Strehl, Lihong Li, and Michael L. Littman. 2009. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research* 10 (Dec. 2009), 2413–2444.

[23] Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. 2006. PAC model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06).* Association for Computing Machinery, New York, NY, USA, 881–888.

[24] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: an introduction* (2nd ed.). The MIT Press, Cambridge, MA, USA. 552 pages.

[25] István Szita and András Lőrincz. 2008. The many faces of optimism: a unifying approach. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08).* Association for Computing Machinery, New York, NY, USA, 1048–1055.

[26] Yuanhao Wang, Kefan Dong, Xiaoyu Chen, and Liwei Wang. 2020. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *International Conference on Learning Representations.* OpenReview.net, Addis Ababa, Ethiopia, Article 509, 18 pages. https://openreview.net/forum?id=BkglSTNFDB

[27] Christopher John Cornish Hellaby Watkins. 1989. *Learning from delayed rewards.* PhD Thesis. King's College, Cambridge, UK.