# Optimistic Q-learning for average reward and episodic reinforcement learning

**Priyank Agrawal** PA2608@COLUMBIA.EDU
*Columbia University*

**Shipra Agrawal** SA3305@COLUMBIA.EDU
*Columbia University*

**Editors:** Nika Haghtalab and Ankur Moitra

## Abstract

Model-free methods for reinforcement learning (RL), particularly Q-learning, have gained popularity in practice because of their simplicity and flexibility, and underlie most successful modern deep RL algorithms. However, the sample complexity and regret bounds for these approaches have often lagged behind their model-based counterparts, especially in the average reward setting. Our work addresses this gap. We present a simple, optimistic Q-learning algorithm for regret minimization in a tabular RL setting that encompasses *both average reward and episodic* settings. Our contributions include new modeling, algorithm design, and regret analysis techniques.

Our first and foremost contribution is a natural modeling assumption that generalizes the episodic and ergodic MDP settings and provides a more practically applicable formulation. We consider the class of MDPs where there is an "upper bound $H$ on the time to visit a frequent state $s_0$", either in expectation or with constant probability. The upper bound $H$ is assumed to hold under all feasible policies (stationary or non-stationary) and is known to the RL agent, although the identity of the frequent state $s_0$ may not be known. This assumption is naturally satisfied by the episodic settings since the terminal state is visited after every set of $H$ steps, and also by the ergodic MDP settings that assume bounded worst-case hitting time $H$ for *all states*. Furthermore, as we demonstrate using several examples from queuing admission control and inventory management, it allows for significantly more modeling flexibility than the existing settings.

A key technical contribution of our work is the introduction of an $\overline{L}$ operator defined as $\overline{L}v = \frac{1}{H}\sum_{h=1}^{H} L^h v$ where $L$ denotes the Bellman operator. Under the given assumption, we show that the $\overline{L}$ operator has a strict contraction (in span) even in the average-reward setting where the discount factor is 1. Our algorithm design builds upon the Q-learning algorithm while replacing the Bellman operator with the novel $\overline{L}$ operator. It uses ideas from episodic Q-learning to estimate and apply this operator iteratively.

Our model-free algorithm improves the existing literature both in simplicity of algorithmic design and regret bounds. Specifically, our algorithm achieves a regret bound of $\tilde{O}(H^5 S\sqrt{AT})$ in the average reward setting, where $S$ and $A$ are the numbers of states and actions, and $T$ is the horizon. A regret bound of $\tilde{O}(H^6 S\sqrt{AT})$ in the episodic setting with fixed episode length $H$ follows as a corollary of this result. Thus, we provide a unified view of algorithm design and regret minimization in episodic and non-episodic settings, which may be of independent interest.[1]

**Keywords:** Q-learning, regret bounds, average reward setting, episodic setting

## Acknowledgments

The research was supported in part by NSF FAI IIS-2147361. The author Priyank Agrawal's work was supported in part by Cheung-Kong Innovation Doctoral Fellowship.

[1] Extended abstract. Full version appears as [arXiv:2407.13743,v3]