Research internship proposal:
Improved strategies for Undiscounted Reinforcement Learning

ODALRIC-AMBRYM MAILLARD

Inria Lille - Nord europe, Équipe Inria SequeL
odalric.maillard@inria.fr
odalricambrymmaillard.neowordpress.fr

Abstract

Reinforcement Learning (RL) theory uses contraction of Bellman operators has a fundamental building block. Contraction in an MDP is naturally enforced by artificially setting a discount factor $\gamma < 1$. However, it has been overlooked in the RL literature (unlike in control) that contraction may also happen when $\gamma = 1$. This gives rise to the notion of intrinsic contraction coefficient: We want to shed light on this fundamental object, study its interplay with the discount factor γ, and explore some fruitful related concepts such as coalescence times. We indeed believe this can greatly help the design of the next generation of reinforcement learning strategies. To achieve such a goal, we suggest to revisit the UCRL strategy designed for regret minimization in undiscounted reinforcement, and to implement and study a series of successive refinements that can be provided to improve its basic version to a novel state-of-the-art, both from the perspective of numerical performance and regret minimization guarantees.

Key words: Reinforcement Learning, Contraction, Hypothesis testing, Regret minimization.

Markov Decision Processes, Reinforcement Learning and discount In its usual formulation, a Markov Decision Process (MDP) is specified by a tuple $(\mathcal{S}, \mathcal{A}, I, R, P, \gamma)$ such that $I \in \mathcal{P}(\mathcal{S})$, $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ and $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{P}(\mathbb{R})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on a set $\mathcal{X}$. $\mathcal{S}$ is called the state space of the MDP, $\mathcal{A}$ is the action space and I is called the initial state distribution. The two most important objects are the transition function P that assigns to each state-action pair a distribution of states, and the reward function R that assigns a real-value to each state-action-state tuple. The process specifies S_0 to be a random-variable with law I, then for any decision $A_0 \in \mathcal{A}$, S_1 has law $P(S_0, A_0)$, and R_1 has law $R(S_0, A_0, S_1)$. More generally:

$$S_0 \sim I, \qquad \forall k \in \mathbb{N}_\star, \quad S_k \sim P(S_{k-1}, A_{k-1}), \quad R_k \sim R(S_{k-1}, A_{k-1}, S_k),$$

where for each $k \in \mathbb{N}$, A_k is an $\mathcal{A}$-valued random variable generated by an external process, and assumed to be adapted to the filtration $\mathcal{F}(S_0, A_0, S_1, R_1, A_1, \dots, S_{k-1}, R_{k-1}, A_{k-1}, S_k)$. Hence in an MDP, the states and rewards are observed quantities that are available to a decision maker.

Finally, $\gamma \in [0, 1]$ is a constant called the discount factor. This constant has nothing to do with the process generating the observations, but is only used in order to specify the objective function that the decision maker tries to optimize. One of the most usual goal is to maximize the cumulative sum or rewards discounted by the factor γ, namely, $\sum_{k \in \mathbb{N}_\star} \gamma^{k-1} R_k$, either in expectation, or sometimes in high probability. We note that when the rewards R_k, $k \in \mathbb{N}_\star$ are random variables bounded almost surely by B (that is, $|R_k| \leqslant B$) and $\gamma < 1$, then the cumulative sum is always finite and bounded by $B/(1-\gamma)$. For a given policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, that is a function that specifies which action to play for each state, the value function captures the expected discounted sum:

$$V^\pi(s) = \mathbb{E}\Big[\sum_{k \in \mathbb{N}_\star} \gamma^{k-1} R_k \Big| S_0 = s\Big] \text{where} \quad A_k \sim \pi(S_k) \text{ for each } k \in \mathbb{N}\,.$$

A general goal is then to find an optimal policy π, that maximizes $V^\pi(s)$ for a given s or when possible $V^\pi(s)$ simultaneously for all $s \in \mathcal{S}$. When such a policy exists (under some mild assumptions), we denote is $\star$.

Markov Decision Processes have been studied from two different perspective: Control Theory (CT), when I, P, R are considered to be perfectly known to the decision maker, and Reinforcement Learning (RL), that considers I, P, R are unknown to the decision maker, but instead belong to some known set, but possibly large set, such as the set of reward functions that generate bounded rewards in $[0, 1]$. Both consider the Bellman fixed-point equations as fundamental starting points. Indeed, it can be shown under some mild conditions on the MDP, that V^π satisfies (written here in a somewhat general form) the following fixed-point equation:

$$\text{(Bellman) } V^\pi = T^\pi[V^\pi] \text{ where } T^\pi[V] : s \mapsto \int_{s'} \int_a [\mu(R(s,a,s')) + \gamma V(s')] dP(s,a)(s') d\pi(s)(a)\,,$$

where we introduced for any $\mathbb{R}$-valued distribution Q its mean $\mu(Q) = \mathbb{E}_{X \sim Q}[X]$. On the other hand, $V^\star$ satisfies:

$$\text{(Bellman optimality) } V^\star = T^\star[V^\star] \text{ where } T^\star[V] : s \mapsto \sup_{\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})} \int_{s'} \int_a [\mu(R(s,a,s')) + \gamma V(s')] dP(s,a)(s') d\pi(s)(a).$$

The γ discount factor: the illusion of simplicity The use of the discount factor in RL has mostly been for historical reasons and simplicity. The reason is probably because the main focus is on the problem of join estimation of P and R, while optimizing the decision strategy, which is especially difficulty when P and R are unknown. On the other hand, it is easy to show that when $\gamma < 1$, these two operators are contracting for the infinite norm $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$, which ensures the uniqueness of a fixed point. We also argue that imposing a contraction γ is artificial, plus asks the difficult question of tuning γ. For instance, it has been reported (see Jiang et al. (2015)) that for a given MDP problem with given discount factor γ, it may be more profitable to solve the same MDP but with another discount factor γ', although there does not seem to be a universal way of choosing γ' based on γ, which looks natural once we understand the most important object is the intrinsic contraction coefficient of the MDP, which has no reason to be related to γ at all.

Contraction of undiscounted Bellman operator A novel trend of research is however focusing on reinforcement learning in the undiscounted setup when $\gamma = 1$ following the seminal paper of Jaksch et al. (2010), that introduced the popular UCRL (Upper Confidence bound Reinforcement Learning) strategy and proved finite regret guarantees for this strategy. In the undiscounted setup, the Bellman equation takes another form called the Poisson equation and does not necessarily admits a unique solution.

The undiscounted case has been studied in the control theory (see e.g. Puterman (2014)), and it can be shown that indeed these operators can be contracting even when $\gamma = 1$, when measuring contraction in the semi-norm span operator, defined by $\mathbb{S} : f \to \max_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} f(x)$. One of the interesting results of control theory is that the contraction coefficient then strongly depends on the shared probability mass when following a policy from two different states for some steps. The more common mass, the stronger the contraction. Hence, studying the "coalescence" probabilities of a policy (or different policies) can lead to better understanding of the optimal Bellman operator and the optimal policy, thus opening an interesting path for building improvement reinforcement learning algorithms.

Caveats of UCRL and improvements Somewhat surprisingly, in Jaksch et al. (2010), the authors do not much study the contraction properties of these operators from a reinforcement learning stand point (when the dynamics is unknown and must be learned). Also, the UCRL strategy was introduced as a simple, not optimized strategy. As such, it suffers from many shortcomings that people have started to fix.

One of this shortcoming is the use of poor confidence intervals for the estimation of the transition and reward distributions, another one is the to question the metric used when comparing transition distributions. For instance in Filippi et al. (2011), the authors use a Kullback-Leibler divergence between distributions, which improves other the initial strategy, as recently shown in Talebi and Maillard (2018). In Maillard et al. (2014), the authors consider yet another discrepancy measure, leading to massive reductino in sampling complexity. Finally, yet another shortcoming is the criterion used by the strategy to decide when recompute its estimates (it proceeds in internal episodes, where an episode heuristically ends when a count of a visited state-action pair has "changed a lot"). Recent findings suggest that this stopping criterion can be replaced with a sound hypothesis testing procedure, leading to a considerable improvement observed in numerical experiments.

Research questions In this project, we suggest to study the questions raised by the intrinsic contraction properties of the Bellman operator in an undiscounted MDP, from the reinforcement learning standpoint, as well as the different improvements that can be made for the UCRL strategy. We will focus on 1) better characterizing this coefficient, and 2) how this can be used to modify standard algorithms such as value iteration or extended value iteration, 3) revisit the initial UCRL, and provide a detailed, step by step improvement of the strategy and its analysis using state-of-the-art statistical techniques, We hope this study also sheds light on 4) when to choose a discount factor γ' to solve an γ or undiscounted MDP in an RL context, 5) the coalescence times of policies in an MDP, their estimation and optimization, and 6) provide a improved version of UCRL whose aim is to become the novel state-of-the-art strategy.

Potential impact Investigating these questions is of utmost importance for the development of the next generation of reinforcement learning algorithms. Indeed, despite the long history of the field and the excitement recently brought by the community around RL, questions regarding the contraction coefficients have been largely overlooked. Also, in increasingly many applications the discount factor is set extremely close to 1, which definitely call for a novel look at the undiscounted setup. Hence, it is natural that an increasingly large number of recent papers have tried to address, however somewhat unsuccessfully, the challenges of regret minimization in average-reward regret minimization using standard approaches. Standing on this large number of unsuccessful trials, we find that tackling this problem without studying intrinsic contraction coefficients is no longer sustainable. Further, we can greatly improve the vanilla UCRL strategy with novel statistical tools. For this reason, we also believe that focusing now on the above fundamental research questions from a theoretical standpoint, will lead to a better understanding of the precise regret that is indeed achievable in undiscounted reinforcement learning, which in turns can have a massive impact on the reinforcement learning community.

References

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimally sensing a single channel without prior information: The tiling algorithm and regret bounds. *Selected Topics in Signal Processing, IEEE Journal of*, 5(1):68–76, 2011.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *The Journal of Machine Learning Research*, 11:1563–1600, 2010.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

Odalric-Ambrym Maillard, Timothy A. Mann, and Shie Mannor. How hard is my MDP? "the distribution-norm to the rescue". In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1835–1843, 2014.

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

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