Budgeted Reinforcement Learning in Continuous State Space

Nicolas Carrara, Edouard Leurent, Romain Laroche, Tanguy Urvoy, Odalric-Ambrym Maillard, Olivier Pietquin

Motivation

Markov Decision Process $(S, A, P_r, R, γ)$:

$$\max_{\pi} \sum_{n=0}^{\infty} \gamma^n R(s, a)$$

- Single scalar reward to represent multiple contradictory aspects (e.g., complete task and avoid collisions)
- No control over the spread of the performance distribution

Constrained MDP $(S, A, P_r, R, γ)$

- [Beutler and Ross 1985; Altman 1999]
- Introduce a cost signal $c$
- Constrained objective

$$\max_{\pi} \sum_{n=0}^{\infty} \gamma^n R(s, a) - \sum_{n=0}^{\infty} \gamma^n c(s, a)$$

The cost budget $\beta$ cannot be changed after training

Budgeted MDP
- [Beutler and Lu 2016]
- We seek one general policy $\pi(s, \beta)$ that solves every CMDP for any $\beta$
- Can only be solved for finite $S$ and known $P_r, R, γ$

Budgeted Dynamic Programming

Theorem (Budgeted Bellman Optimality). $Q^*$ verifies:

$$Q^*(s, \pi) = TQ^*(s, \pi) + \gamma \sum_{a, s'} T(s, a, s') \pi_{\text{greedy}}(s', \pi) Q^*(s', \pi)$$

where the greedy policy $\pi_{\text{greedy}}$ is defined by:

$$\pi_{\text{greedy}}(s, \pi) = \arg \max_{a} E_{s' \sim P}|Q(s', s, a)\pi_{\text{greedy}}(s', \pi)|$$

Proposition. $\pi_{\text{greedy}}(\cdot; Q^*)$ is simultaneously optimal in all states $s \in S$.

$$\pi_{\text{greedy}}(\cdot; Q^*) \in \Pi^*$$

In particular, $V_{\text{value}}(Q^*) = V^*$ and $Q_{\text{target}}(Q^*) = Q^*$.

Theorem (Contractivity). For any BMDP $(S, A, P_r, R, γ)$ with $|A| \geq 2$, $T$ is not a contraction.

$$\forall \epsilon > 0, \exists Q^0 \in (R^2)^{S \times A} : \|TQ^0 - TQ\|_{\infty} \leq \frac{1}{2} \|Q^0 - Q\|_{\infty}$$

Despite these theoretical limitations, we observed empirical convergence in our experiments

Budgeted Reinforcement Learning

We address several limitations of Algorithm 1

1. The BMDP is unknown
2. $T$ contains an expectation $\mathbb{E}_{s' \sim P}$ over next states $s'$
3. $S$ is continuous

Algorithm 2: Budgeted Fitted-Q iteration

Data: $P, R, c$
Result: $Q^*$
1. $Q_0 \leftarrow 0$
2. repeat
3. $Q_{k+1} \leftarrow \text{Function approximation}$
4. until convergence

Scalable Implementation

How to compute the greedy policy?

We define the budgeted action-value function $Q_\pi$ similarly

Acknowledgements

This work has been supported by CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, INRIA, and the French Agence Nationale de la Recherche (ANR)

References


Risk-sensitive exploration

Pareto frontier