Budgeted Reinforcement Learning in Continuous State Space

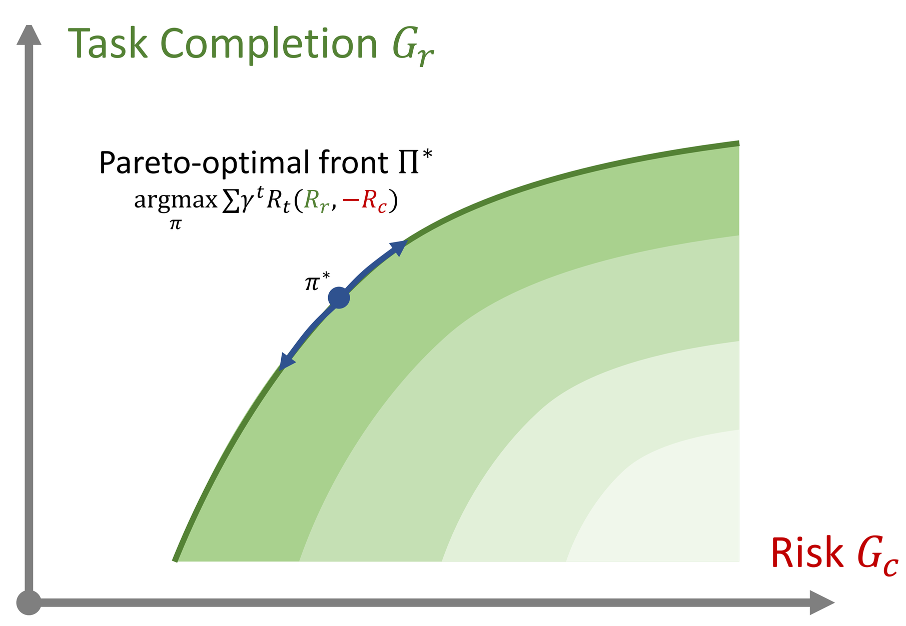
Nicolas Carrara*, Edouard Leurent*, Romain Laroche,
Tanguy Urvoy, Odalric-Ambrym Maillard, Olivier Pietquin

Motivation

Markov Decision Process $(\mathcal{S},\mathcal{A},P,R_r,\gamma)$:

$$\max_\pi \mathbb{E}_\pi \underbrace{\sum_{t=0}^{\infty}\gamma^t R_r(s_t,a_t)}_{G_r^\pi}$$

Single scalar reward for multiple contradictory aspects



Constrained MDP $(\mathcal{S},\mathcal{A},P,R_r,R_c,\gamma,\beta)$

- [Beutler and Ross 1985; Altman 1999]
- Introduce a cost signal R_c and constrained objective

$$\max_{\pi\in\mathcal{M}(\mathcal{A})^{\mathcal{S}}} \mathbb{E}[G_r^\pi|s_0=s] \quad \text{s.t.} \quad \mathbb{E}[G_c^\pi|s_0=s]\leq\beta$$

↳ The cost budget β cannot be changed after training

Budgeted MDP $(\mathcal{S},\mathcal{A},P,R_r,R_c,\gamma,\mathcal{B})$

- [Boutilier and Lu 2016]
- We seek one general policy $\pi(s,\beta)$ that solves every CMDP for any $\beta\in\mathcal{B}$

↳ Can only be solved for finite $\mathcal{S}$ and known P,R_r,R_c.

Setting

Budgeted policies π

- Take a budget β as an additional input
- Output a next budget β'

$$\pi:\underbrace{(s,\beta)}_{\bar s}\to\underbrace{(a,\beta')}_{\bar a}$$

2D signals

1. Rewards $R=(R_r,R_c)$
2. Returns $G^\pi=(G_r^\pi,G_c^\pi)$
3. Values $V^\pi=(V_r^\pi,V_c^\pi)$ and $Q^\pi=(Q_r^\pi,Q_c^\pi)$

Policy Evaluation

The Bellman Expectation equations are preserved, and the Bellman Expectation Operator $\mathcal{T}^\pi$ is a γ-contraction.

Budgeted Optimality

Definition. *In that order, we want to:*

(i) Respect the budget β:

$$\Pi_a(\bar s)\overset{\text{def}}{=}\{\pi\in\Pi: V_c^\pi(s,\beta)\leq\beta\}$$

(ii) Maximise the rewards:

$$V_r^*(\bar s)\overset{\text{def}}{=}\max_{\pi\in\Pi_a(\bar s)}V_r^\pi(\bar s),\quad \Pi_r(\bar s)\overset{\text{def}}{=}\operatorname*{arg\,max}_{\pi\in\Pi_a(\bar s)}V_r^\pi(\bar s)$$

(iii) Minimise the costs:

$$V_c^*(\bar s)\overset{\text{def}}{=}\min_{\pi\in\Pi_r(\bar s)}V_c^\pi(\bar s),\quad \Pi^*(\bar s)\overset{\text{def}}{=}\operatorname*{arg\,min}_{\pi\in\Pi_r(\bar s)}V_c^\pi(\bar s)$$

We define the budgeted action-value function Q^* 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).

Budgeted Dynamic Programming

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

$$Q^*(\bar s,\bar a)=\mathcal{T}Q^*(\bar s,\bar a)\overset{\text{def}}{=}R(\bar s,\bar a)+\gamma\sum_{\bar s'\in\bar{\mathcal{S}}}\overline{P}(\bar s'|\bar s,\bar a)\sum_{\bar a'\in\bar{\mathcal{A}}}\pi_{greedy}(\bar a'|\bar s';Q^*)Q^*(\bar s',\bar a'),\qquad(1)$$

where the greedy policy π_{greedy} is defined by:

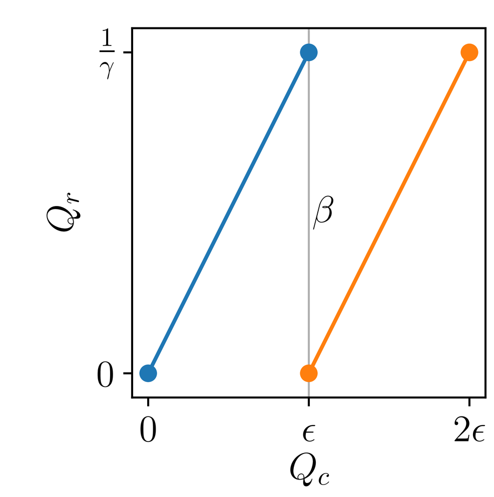
$$\pi_{greedy}(\bar a|\bar s;Q)\in\operatorname*{arg\,min}_{\rho\in\Pi_r^Q}\mathop{\mathbb{E}}_{\bar a\sim\rho}Q_c(\bar s,\bar a),\qquad(2a)$$

$$\text{where}\quad \Pi_r^Q\overset{\text{def}}{=}\operatorname*{arg\,max}_{\rho\in\mathcal{M}(\bar{\mathcal{A}})}\mathop{\mathbb{E}}_{\bar a\sim\rho}Q_r(\bar s,\bar a)\qquad(2b)$$

$$\text{s.t.}\quad \mathop{\mathbb{E}}_{\bar a\sim\rho}Q_c(\bar s,\bar a)\leq\beta\qquad(2c)$$

Algorithm 1: Budgeted Value Iteration

Data: P,R_r,R_c
Result: Q^*
1 $Q_0\leftarrow 0$
2 **repeat**
3 $Q_{k+1}\leftarrow\mathcal{T}Q_k$
4 **until** *convergence*



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

$$\pi_{greedy}(\cdot\,;Q^*)\in\Pi^*(\bar s)$$

In particular, $V^{\pi_{greedy}(\cdot;Q^*)}=V^*$ and $Q^{\pi_{greedy}(\cdot;Q^*)}=Q^*$.

Theorem (Contractivity). *For any BMDP $(\mathcal{S},\mathcal{A},P,R_r,R_c,\gamma)$ with $|\mathcal{A}|\geq 2$, $\mathcal{T}$ is not a contraction.*

$$\forall\varepsilon>0,\exists Q^1,Q^2\in(\mathbb{R}^2)^{\bar{\mathcal{S}}\bar{\mathcal{A}}}:\|\mathcal{T}Q^1-\mathcal{T}Q^2\|_\infty\geq\frac{1}{\varepsilon}\|Q^1-Q^2\|_\infty$$

✗ We cannot guarantee the convergence of $\mathcal{T}^n(Q_0)$ to Q^*.

Theorem (Contractivity on smooth Q-functions). *$\mathcal{T}$ is a contraction when restricted to the subset $\mathcal{L}_\gamma$ of Q-functions such that "Q_r is L-Lipschitz with respect to Q_c", with $L<\frac{1}{\gamma}-1$.*

✓ We guarantee convergence under some (strong) assumptions.
✓ We observe empirical convergence.

Budgeted Reinforcement Learning

We address several limitations of Algorithm 1.

1. The BMDP is unknown
   ↳ Work with a batch of samples $\mathcal{D}=\{(\bar s_i,\bar a_i,r_i,\bar s_i')\}_{i\in[0,N]}$
2. $\mathcal{T}$ contains an expectation $\mathbb{E}_{\bar s'\sim\overline{P}}$ over next states $\bar s'$
   ↳ Replace it with a sampling operator $\hat{\mathcal{T}}$:

$$\hat{\mathcal{T}}Q(\bar s_i,\bar a_i,r_i,\bar s_i')\overset{\text{def}}{=}r_i+\gamma\sum_{\bar a_i'\in\mathcal{A}_i}\pi_{\text{greedy}}(\bar a_i'|\bar s_i';Q)Q(\bar s_i',\bar a_i').$$

3. $\mathcal{S}$ is continuous
   ↳ Employ function approximation Q_θ, and minimise a regression loss

$$\mathcal{L}(Q_\theta,Q_{\text{target}};\mathcal{D})=\sum_{\mathcal{D}}||Q_\theta(\bar s,\bar a)-Q_{\text{target}}(\bar s,\bar a,r,\bar s')||_2^2$$

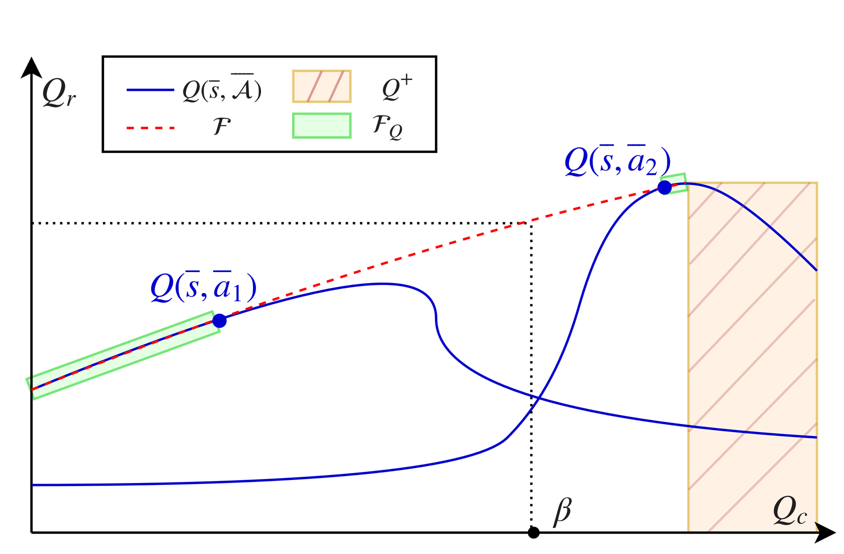
Algorithm 2: Budgeted Fitted-Q Iteration

Data: $\mathcal{D}$
Result: Q^*
1 $Q_{\theta_0}\leftarrow 0$
2 **repeat**
3 $\theta_{k+1}\leftarrow\arg\min_\theta\mathcal{L}(Q_\theta,\hat{\mathcal{T}}Q_{\theta_k};\mathcal{D})$
4 **until** *convergence*

4. How to collect the batch $\mathcal{D}$?
   ↳ We propose a risk-sensitive exploration procedure

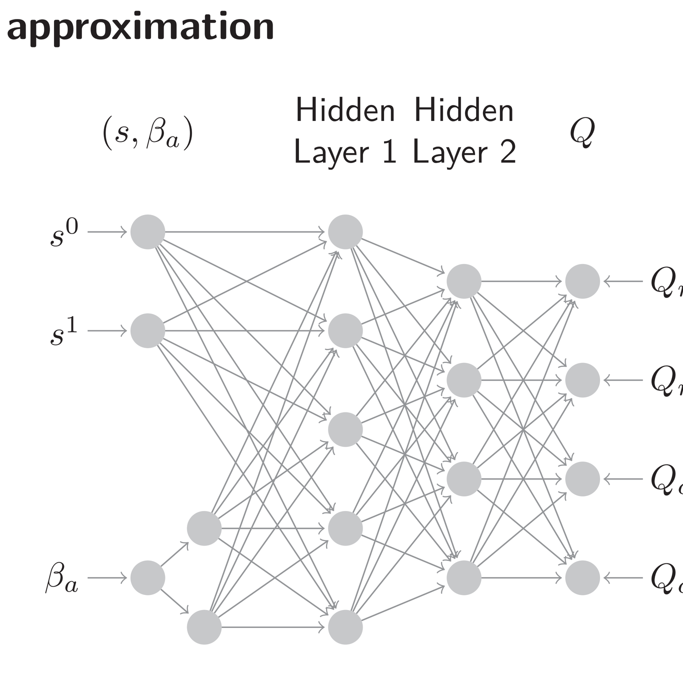
Scalable Implementation

How to compute the greedy policy?



Proposition (Hull policy). *π_{greedy} in (2) can be computed explicitly, as a mixture of two points that lie on the convex hull of Q.*

Function approximation

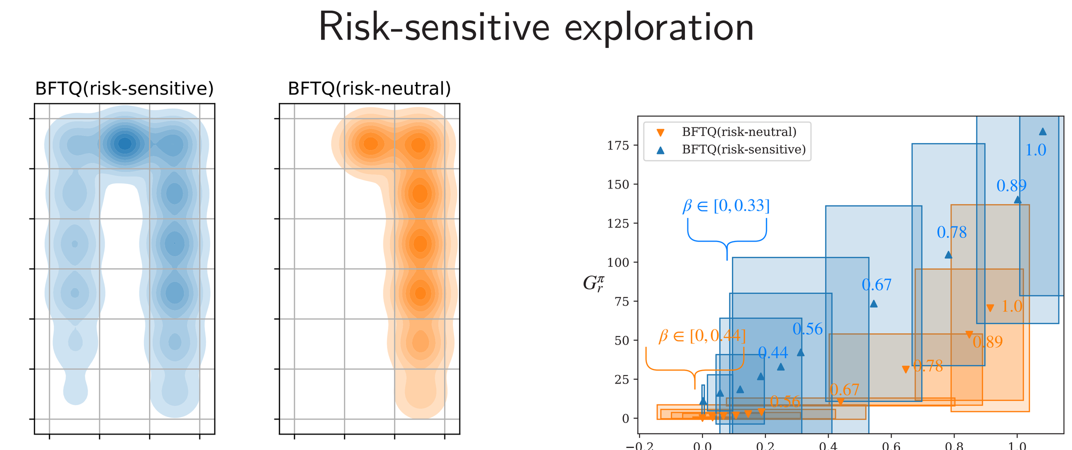


Parallel computing

Experience collection and computation of π_{greedy} can be distributed over several cores.

Experiments

Risk-sensitive exploration



Pareto frontier