**Motivation**

Markov Decision Process \((S, A, P, R, γ)\):

\[
\max E \sum R(s_t, a_t) G^t
\]

- 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 \((S, A, P, R, γ)\)
- [Beutler and Ross 1985; Altman 1999]
- Introduce a cost signal \(R_c\)
- Constrained objective
  \[
  \max_{\pi \in M(s, γ)} E[G_t | s_0 = s, a_0 = a] \leq \beta
  \]
- The cost budget \(\beta\) cannot be changed after training

**Budgeted Dynamic Programming**

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

\[
Q^*(s, a) = TQ^*(s, a) + γ ∑_{s', r} P(s', r | s, a) \min_{Q(c)} Q^*_r(s', a')\]

(i) Respect the budget \(β\):

\[
H_{\pi}(R) = \min \{π \in \Pi: V^*_\pi(s, β) ≤ β\}
\]

(ii) Maximise the rewards:

\[
V^*_\pi = \max_{\pi \in M(s, γ)} V^*_\pi\]

(iii) Maximise the costs:

\[
V^*_\pi = \min_{\pi \in M(s, γ)} V^*_\pi\]

**References**


**Acknowledgements**

This work was 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).