Planning in Markov Decision Processes with Gap-Dependent Sample Complexity

Anders Jonsson [1], **Emilie Kaufmann** [2,3,4], **Pierre Ménard** [4], **Omar Darwiche Domingues** [4], **Edouard Leurent** [4,5] **& Michal Valko** [6]

1. Universitat Pompeu Fabra
2. CNRS
3. Université de Lille, CRIStAL
4. Inria Lille, Scool team
5. Renault
6. Deepmind Paris

NEURAL INFORMATION PROCESSING SYSTEMS

- Monte-Carlo planning: recommend action in a given state s_1.
- Monte-Carlo Tree Search (MCTS): sample trajectories using a *forward model* that simulates actions in the *current state*.

Contribution

- A new trajectory-based MCTS algorithm, `MDP-GapE`.
- Easy to implement, performs well in practice.
- Sample complexity bounds for the fixed confidence setting.
- Bounds depend on the *sub-optimality gaps* of actions in s_1.
  ➜ `MDP-GapE` does *not* explore trajectories uniformly.

Setting

A discounted, episodic MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r \rangle$, transitions $p = \{p_h\}_{h\geq 1}$ and rewards $r = \{r_h\}_{h\geq 1}$ bounded in $[0, 1]$ with

- disount factor $\gamma \in (0, 1]$, horizon $H \in \mathbb{N}^*$,
- number of actions $K = |\mathcal{A}|$, finite branching factor B.

The optimal action-value function $Q = \{Q_h\}_{h\geq 1}$ is defined as

$$Q_h(s, a) = r_h(s, a) + \gamma \sum_{s'} p_h(s'|s, a) \max_{a'} Q_{h+1}(s', a').$$

- Optimal action in step 1: $a^\star = \underset{a\in\mathcal{A}}{\text{argmax}}\ Q_1(s_1, a)$.

Fixed confidence planning

Given ε and δ, **output an action $\hat{a}^\tau$** that satisfies

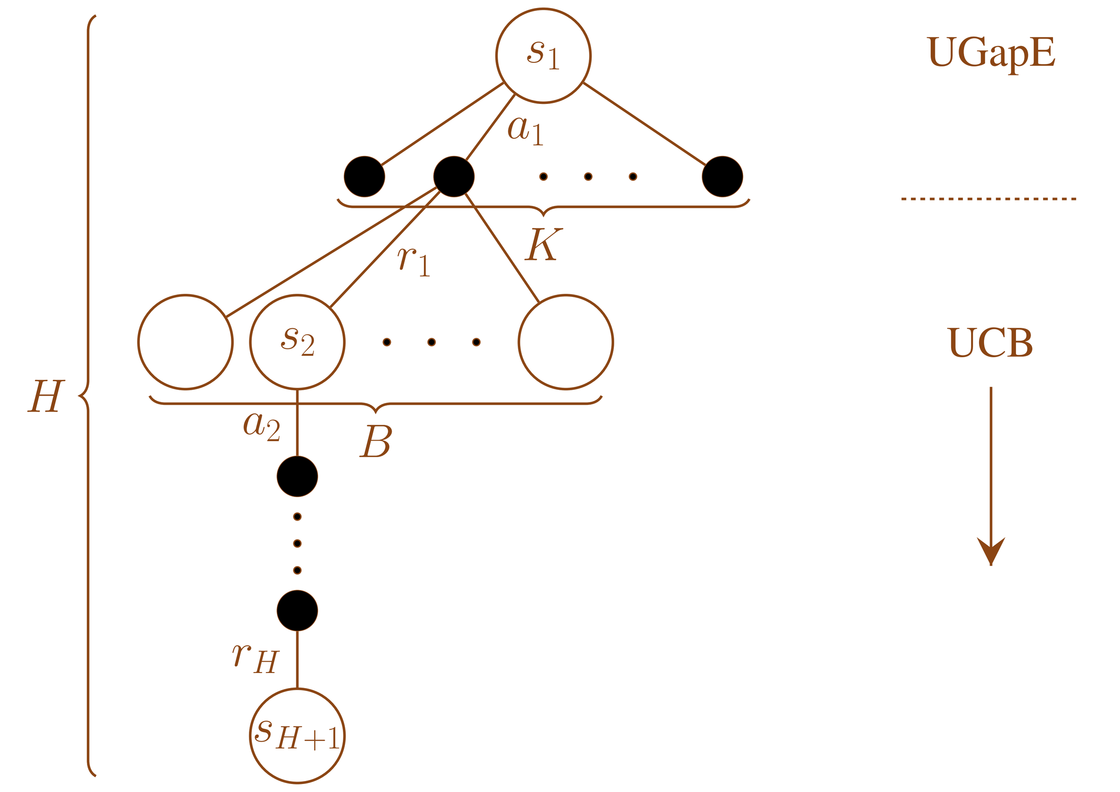
$$\mathbb{P}\left(Q_1(s_1, a^\star) - Q_1(s_1, \hat{a}^\tau) < \varepsilon\right) \geq 1 - \delta,$$

after generating the **smallest possible number of episodes** τ.

Algorithm	Setting	Sample complexity
Sparse Sampling [19]	Fixed confidence	$H^5(BK)^H/\varepsilon^2$
OLOP [2]	Fixed budget	$\varepsilon^{-\max\left(2,\frac{\log\kappa}{\log(1/\gamma)}\right)}$
OP [3]	Anytime	$\varepsilon^{-\frac{\log\kappa}{\log(1/\gamma)}}$
BRUE [8]	Anytime	$H^4(BK)^H/\Delta^2$
StOP [28]	Fixed confidence	$\varepsilon^{-\left(2+\frac{\log\kappa}{\log(1/\gamma)}+o(1)\right)}$
TrailBlazer [13]	Fixed confidence	$\varepsilon^{-\max\left(2,\frac{\log(B\kappa)}{\log(1/\gamma)}+o(1)\right)}$
SmoothCruiser [14]	Fixed confidence	ε^{-4}
`MDP-GapE` (ours)	Fixed confidence	$\sum_{a_1\in\mathcal{A}} \frac{H^2(BK)^{H-1}B}{(\Delta_1(s_1,a_1)\vee\Delta\vee\varepsilon)^2}$

Number of observed transitions n needed by existing algorithms to guarantee $Q_1(s_1, a^\star) - Q_1(s_1, \hat{a}^n) < \varepsilon$.

The `MDP-GapE` Algorithm



Based on data from the first t episodes, build

- Confidence bounds $[\ell_h^{t,\delta}(s, a), u_h^{t,\delta}(s, a)]$ on the rewards $r_h(s, a)$.
- Confidence sets $\mathcal{C}_h^{t,\delta}(s, a)$ on the probability vectors $p_h(\cdot|s, a)$.

Confidence bounds on action values

Define confidence bounds on the action value $Q_h(s, a)$:

$$U_h^{t,\delta}(s, a) = u_h^{t,\delta}(s, a) + \gamma \max_{p\in\mathcal{C}_h^{t,\delta}(s,a)} \sum_{s'} p(s'|s, a) \max_{a'} U_{h+1}^{t,\delta}(s', a'),$$

$$L_h^{t,\delta}(s, a) = \ell_h^{t,\delta}(s, a) + \gamma \min_{p\in\mathcal{C}_h^{t,\delta}(s,a)} \sum_{s'} p(s'|s, a) \max_{a'} L_{h+1}^{t,\delta}(s', a').$$

Lemma 1. *For each confidence level $\delta \in [0, 1]$, w.p. at least $1 - \delta$,*

$$Q_h(s, a) \in [L_h^{t,\delta}(s, a), U_h^{t,\delta}(s, a)]$$

for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, episode t and depth h.

- The policy $\pi^{t+1} = \{\pi_h^{t+1}\}_{h\geq 1}$ used in episode $t + 1$ is

$$\pi_1^{t+1}(s_1) = \underset{b\in\{b^t,c^t\}}{\text{argmax}} \left[U_1^{t,\delta}(s_1, b) - L_1^{t,\delta}(s_1, b)\right], \quad \text{(UGapE)}$$

$$\pi_h^{t+1}(s_h) = \underset{a\in\mathcal{A}}{\text{argmax}}\ U_h^{t,\delta}(s_h, a), \quad h > 1, \quad \text{(UCB)}$$

for a specific choice of best action b^t and challenger c^t in state s_1:

$$b^t = \underset{b}{\text{argmin}} \left[\max_{a\neq b} U_1^{t,\delta}(s_1, a) - L_1^{t,\delta}(s_1, b)\right], \quad c^t = \underset{c\neq b^t}{\text{argmax}}\ U_1^{t,\delta}(s_1, c).$$

Output

Stopping rule $\tau = \inf\{t \in \mathbb{N} : U_1^{t,\delta}(s_1, c^t) - L_1^{t,\delta}(s_1, b^t) < \varepsilon\}$.
After stopping, the algorithm outputs action $\hat{a}^\tau = b^\tau$.

Lemma [Correctness]

Since the confidence bounds hold w.p. $1-\delta$, the stopping rule implies

$$Q_1(s_1, a^\star) - Q_1(s_1, \hat{a}^\tau) \leq U_1^{\tau,\delta}(s_1, c^\tau) - L_1^{\tau,\delta}(s_1, b^\tau) < \varepsilon.$$

- Define the sub-optimality gaps as

$$\Delta = \min_{a\neq a^\star} [Q_1(s_1, a^\star) - Q_1(s_1, a)],$$

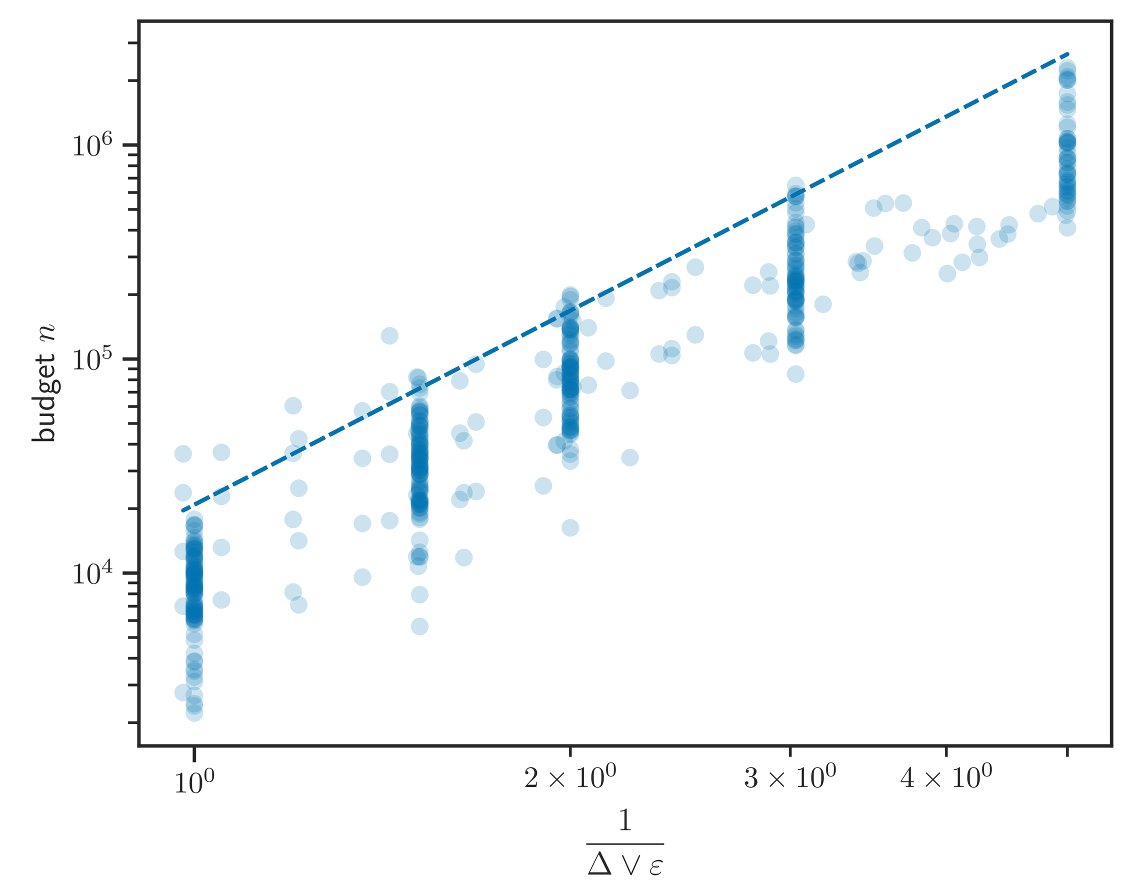
$$\Delta_h(s, a) = \max_b Q_h(s, b) - Q_h(s, a), \quad 1 \leq h \leq H.$$

Theorem [Sample Complexity]

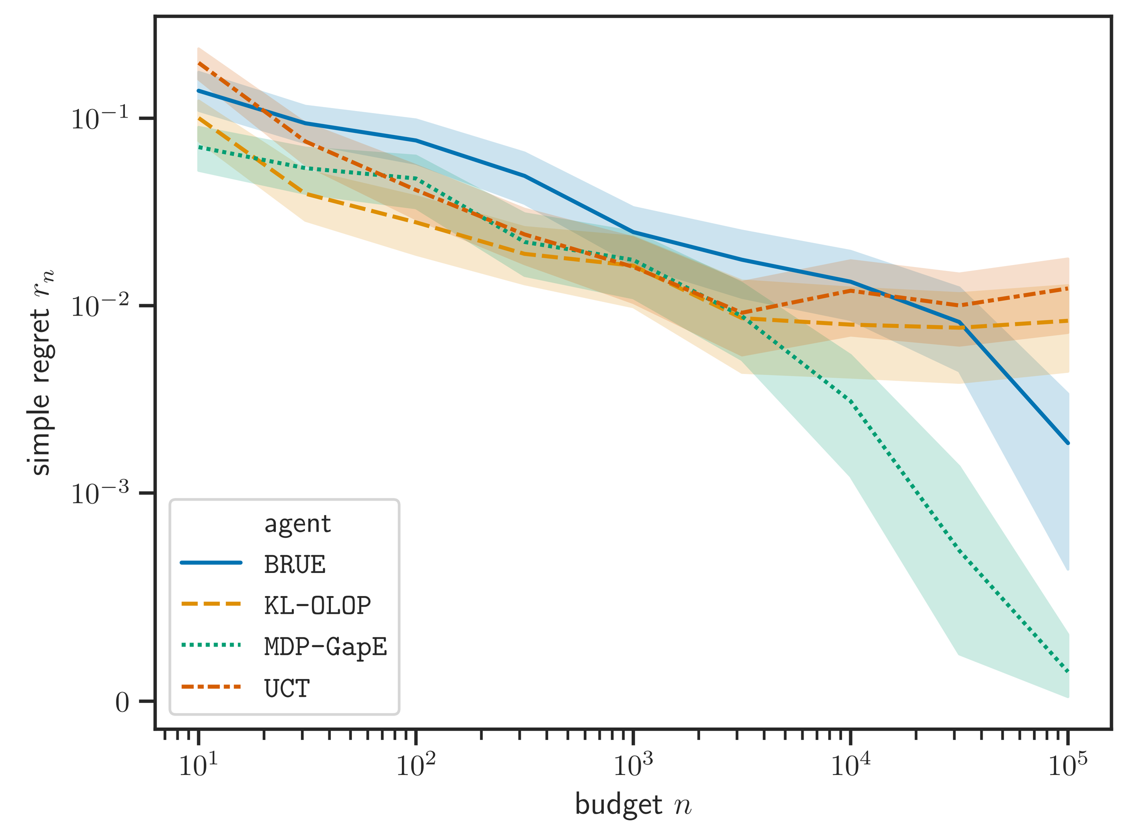
The number of episodes used by MDP-GapE satisfies

$$\tau = \mathcal{O}\left(\sum_a \frac{(BK)^{H-1}}{(\Delta_1(s_1, a) \vee \Delta \vee \varepsilon)^2} \left[\log\frac{1}{\delta} + BH \log(BK)\right]\right)$$

with probability at least $1 - \delta$.



Empirical scaling of sample complexity $n = O(1/\varepsilon^{3.0})$.



Fixed-budget comparison to other algorithms.