Seminar in Artificial Intelligence

TECHNISCHE UNIVERSITÄT DARMSTADT

Near-Bayesian Exploration in Polynomial Time

Table of Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

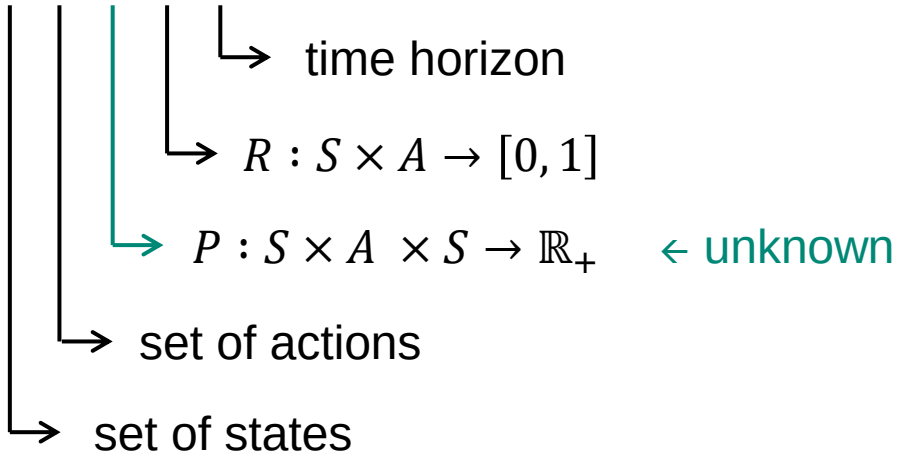
Table of Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

Problem and Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Agent in unknown environment
- Discrete states and actions
- MDP: $\{S, A, P, R, H\}$



Domain Example – two-armed bandit

TECHNISCHE UNIVERSITÄT DARMSTADT

Lever 1
50% chance of winning

Lever 2
60% chance of winning

Table of Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

Value Function 1

$$V_H^{\pi}(s) = R\big(s, \pi(s)\big) + \sum_{s'} P(s'|s,a) V_{H-1}^{\pi}(s')$$

Bellman's equation

transitions of MDP are known

→ can find optimal policy π^* and optimal value function V^*

$$V_H^*(s) = \max_a \left\{ R(s,a) + \sum_{s'} P(s'|s,a) V_{H-1}^*(s') \right\}$$

Problem: P is unknown

Value Function 2

using a *belief state b* – set of Dirichlet distributions

$$b = \{\alpha(s, a, s')\} \qquad \alpha_0(s, a) = \sum_{s'} \alpha(s, a, s')$$

$$P(s'|b, s, a) = \frac{\alpha(s, a, s')}{\alpha_0(s, a)}$$

→ get value function without origin P

$$V_H^*(b, s) = \max_a \left\{ R(s, a) + \sum_{s'} P(s'|b, s, a) V_{H-1}^*(b', s') \right\}$$

Domain Example – two-armed bandit

TECHNISCHE UNIVERSITÄT DARMSTADT

Lever 1
50% chance of winning

pulled 100 times
paid off 52 times
→ **52%**

Lever 2
60% chance of winning

pulled 5 times
paid off 2 times
→ **40%**

Table of Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayesian Exploration Bonus (BEB)

TECHNISCHE UNIVERSITÄT DARMSTADT

Bonus: $\frac{\beta}{1+\alpha_0(s,a)}$

$$V_H^*(b,s) = \max_a \left\{ R(s,a) + \frac{\beta}{1+\alpha_0(s,a)} + \sum_{s'} P(s'|b,s,a) V_{H-1}^*(b',s') \right\}$$

Reward **Bonus** **Estimated mean value of next states**

Domain Example – two-armed bandit

TECHNISCHE UNIVERSITÄT DARMSTADT

Lever 1
50% chance of winning

pulled 100 times
paid off 52 times
→ **52%**

$$R_1 = 0.52 + \frac{\beta}{1 + 102}$$

Lever 2
60% chance of winning

pulled 5 times
paid off 2 times
→ **40%**

$$R_2 = 0.4 + \frac{\beta}{1 + 7}$$

Domain Example – two-armed bandit

TECHNISCHE UNIVERSITÄT DARMSTADT

$$R_1 = 0.52 + \frac{\beta}{1 + 102}$$

$\beta = 0 \rightarrow R_1 = 0.52$

$\beta = 1 \rightarrow R_1 \approx 0.53$

$\beta = 2 \rightarrow R_1 \approx 0.54$

$\beta = 3 \rightarrow R_1 \approx 0.55$

$\beta = 4 \rightarrow R_1 \approx 0.56$

$$R_2 = 0.4 + \frac{\beta}{1 + 7}$$

$\beta = 0 \rightarrow R_2 = 0.4$

$\beta = 1 \rightarrow R_2 = 0.525$

$\beta = 2 \rightarrow R_2 = 0.65$

$\beta = 3 \rightarrow R_2 = 0.775$

$\beta = 4 \rightarrow R_2 = 0.9$

Table of Contents

Complexity

ϵ-close to the optimal Bayesian policy

BEB

$$O\left(\frac{|S||A|H^6}{\epsilon^2}\log\frac{|S||A|}{\delta}\right)$$

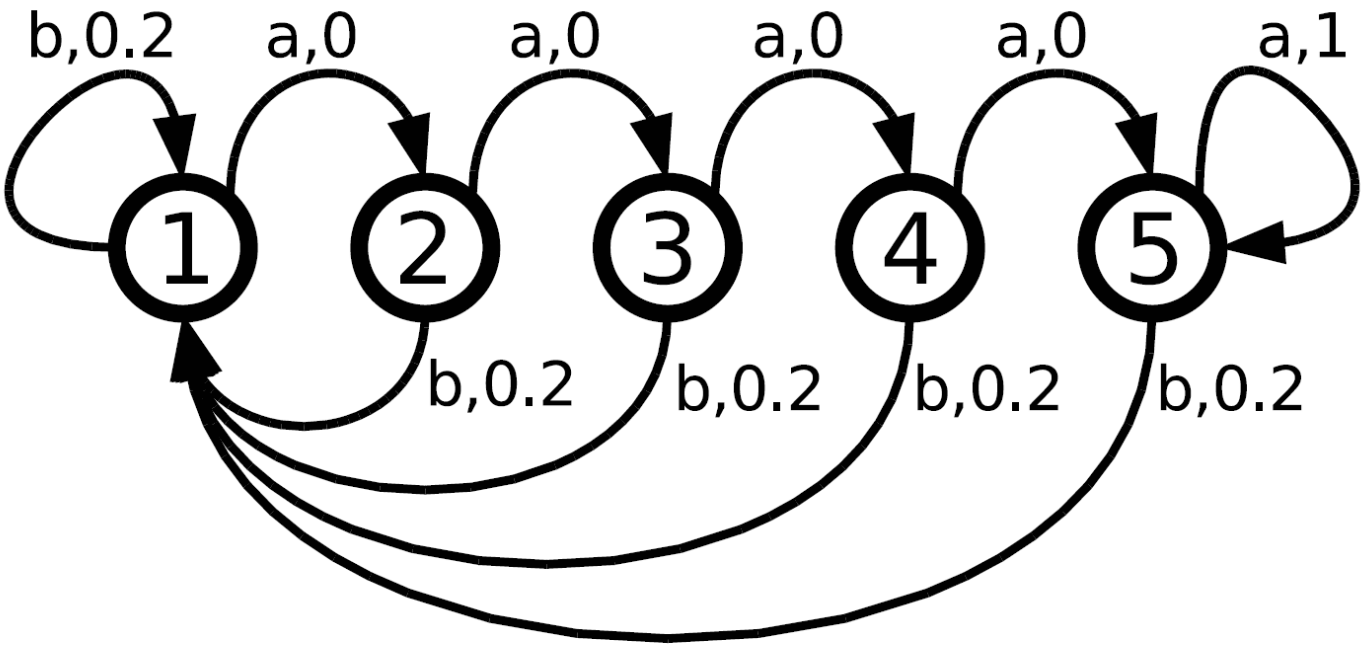
standard PAC-MDP

$$\tilde{O}\left(\frac{|S|^2|A|H^6}{\epsilon^3}\right)$$

$\tilde{O}(\cdot)$ notation suppresses logarithmic factors

Table of Contents

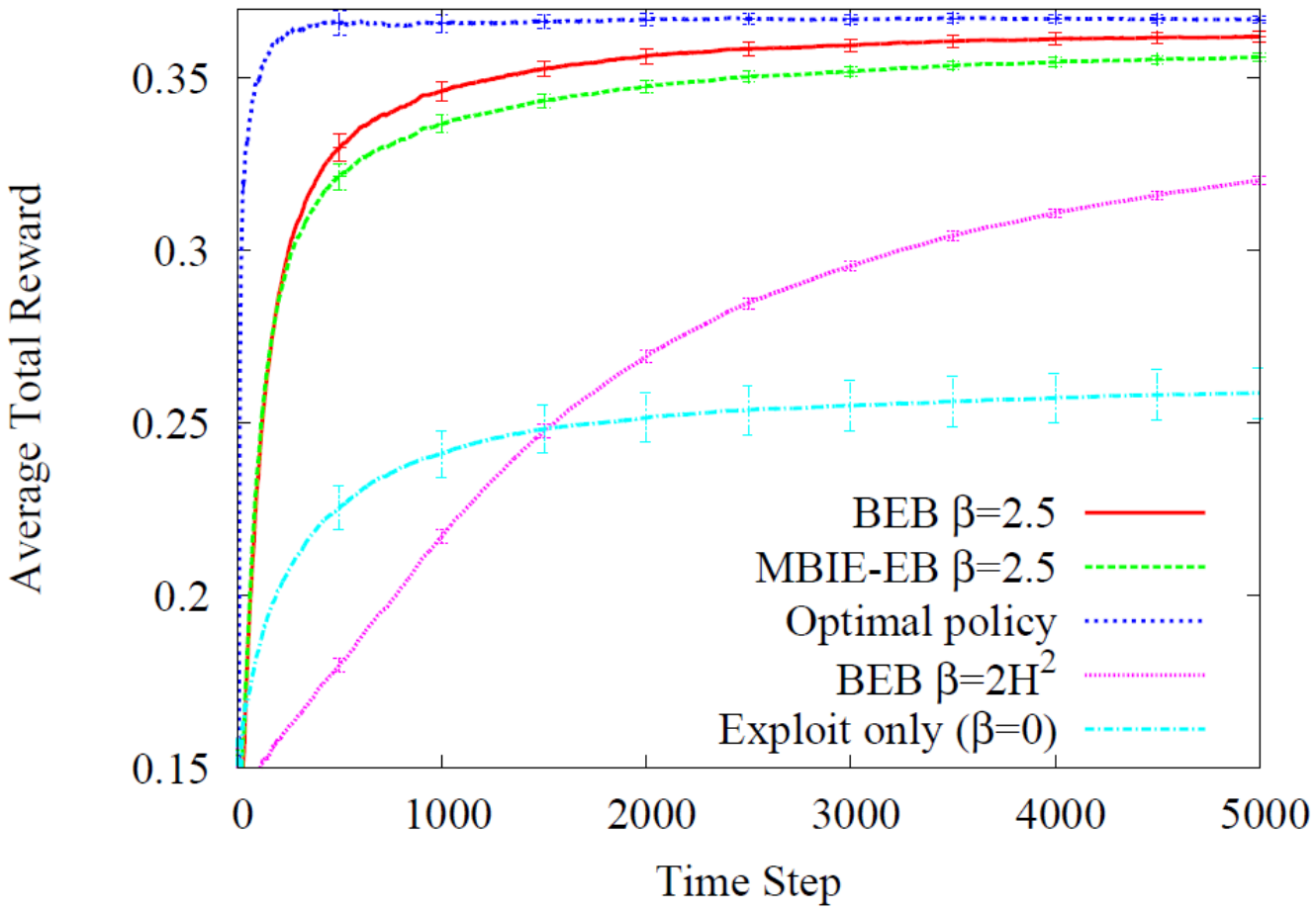
Simulated Domain



Chain domain with five states and two actions.

With probability of 0.2 the agent performs the opposite action as intended.

Simulated Domain – Result 1



Simulated Domain – Result 2

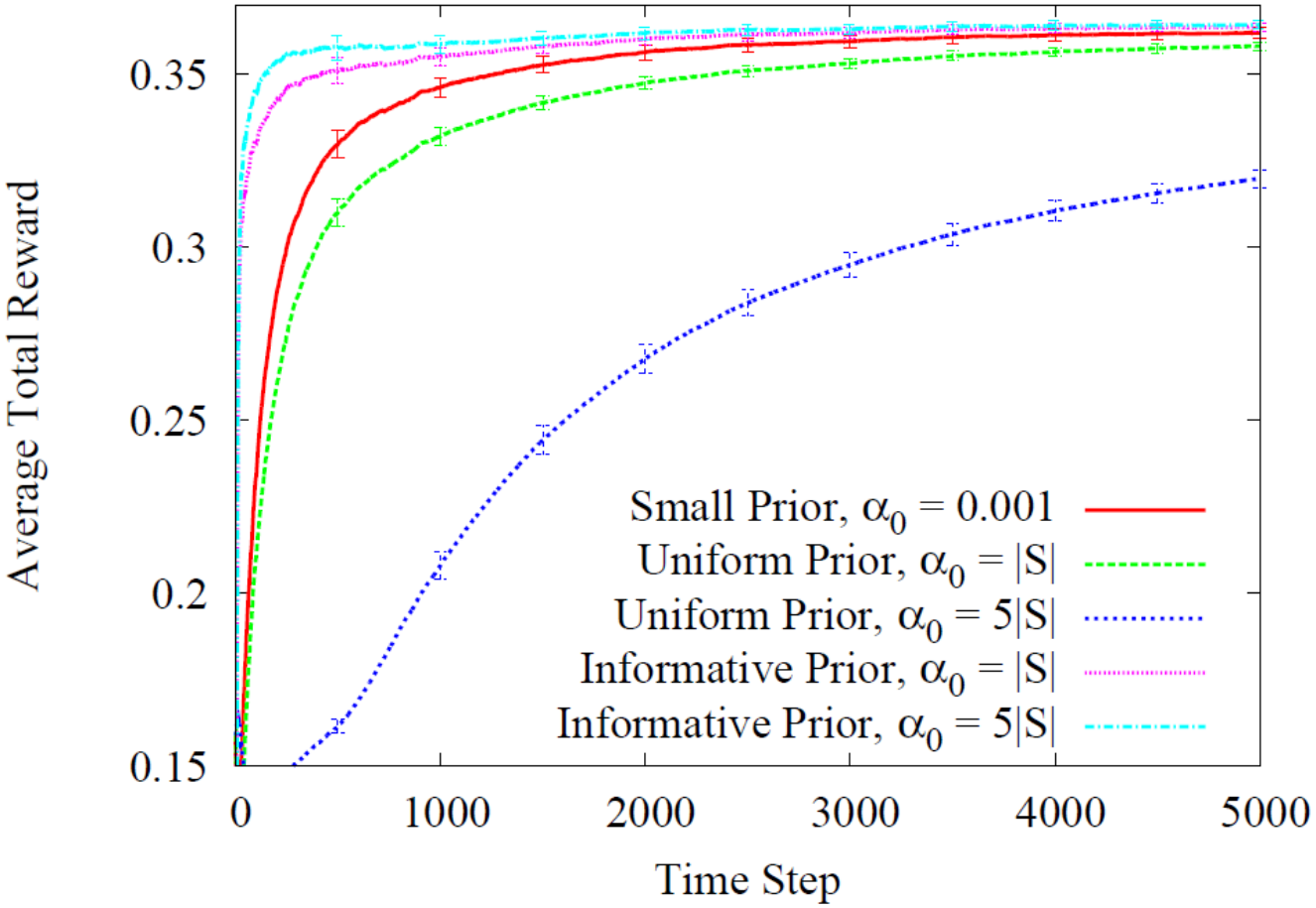


Table of Contents

TECHNISCHE UNIVERSITÄT DARMSTADT

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- ϵ-close to the optimal Bayesian policy
  after a polynomial number of time steps

- Balanced exploration and exploitation

- Better complexity compared to standard PAC-MDP
  (in polynomial time)