Comparing Exploration Methods in Partially Observable Stochastic Games

Jakub Rada

June 14, 2022

Faculty of Electrical Engineering
Czech Technical University in Prague

Motivation and multi-armed bandits

Motivation

- SGs and POSGs have applications in economy, stock markets, network security, biology, machine learning, etc.
- solving algorithms exist, but often require linear programming
  - value iteration, HSVI
- multi-armed bandits can be used as an alternative approach of exploring the search space

Multi-armed bandits

Stochastic bandit algorithms

- Best of N, ϵ-greedy
- Successive Elimination, UCB
- **observable** variants for each of the above - only in SGs

Adversarial bandits

- Exp3

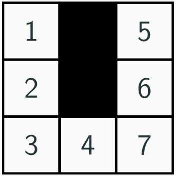
Stochastic Games

SG model

Stochastic game

A *stochastic game* is a tuple $G = (S, A_1, A_2, T, R, \gamma)$, where

- S is a set of states of the game,
- A_1, A_2 are sets of actions available to player 1, resp. player 2,
- $T : S \times A_1 \times A_2 \times S \to [0, 1]$ is a transition function,
- $R : S \times A_1 \times A_2 \to \mathbb{R}$ is a reward function and
- $\gamma \in (0, 1)$ is the discount factor.



Example map of a stochastic game **Tag**

Value iteration for SGs

- based on a value function $V : S \to \mathbb{R}$
- starts from initial V^0
- iterative application of a Bellman operator refines the approximation V^t to V^{t+1}

Stage game u for a state $s \in S$

$$\forall a_1 \in A_1, \forall a_2 \in A_2 :$$
$$u(a_1, a_2) = R(s, a_1, a_2) + \gamma \sum_{s'} T(s'|s, a_1, a_2) \cdot V(s')$$

- stage game can be replaced by some bandit algorithm
- terminates when $|V^t(s) - V^{t+1}(s)| \leq \epsilon \quad \forall s \in S$

Comparison on SG

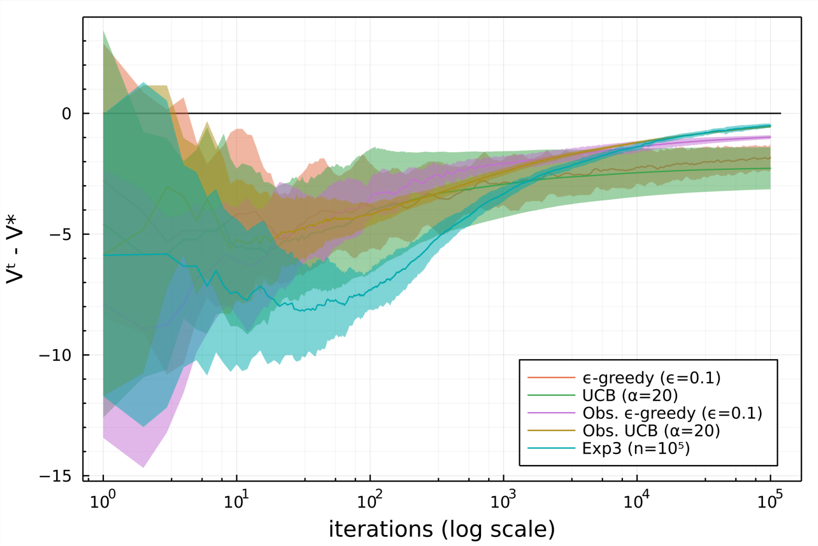


Figure 1: Best bandit algorithms in a state with mixed optimal strategies

One-Sided Partially Observable Stochastic Games

OS-POSG model

One-sided partially observable stochastic game

An *OS-POSG* is a tuple $G = \left(S, A_1, A_2, O, T, R, b^{\text{init}}, \gamma\right)$, where S, A_1, A_2, R, γ are the same as in SGs and

- O is a set of private observations for player 1,
- $T : S \times A_1 \times A_2 \times O \times S \to [0, 1]$ is a transition function,
- $b^{\text{init}} \in \Delta(S)$ is an initial belief over states in S.

●	●	0.0
0.0	0.0	0.0
0.0	0.0	1.0

0.25	●	0.25
0.0	●	0.0
0.25	0.0	0.25

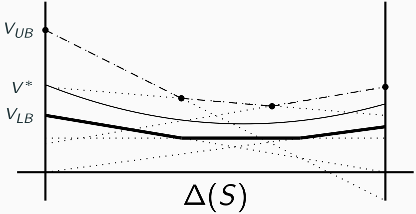
Example initial setting of an OS-POSG **Pursuit-Evasion**

● ... Pursuer units

p ... probability of Evader

HSVI for OS-POSGs

- two bounding value functions: $V_{LB} \leq V^* \leq V_{UB}$
  - values are computed against opponent's *best response*
- iterative application of the Bellman operator pushes the bounds closer together
- terminate when $V_{UB}(b^{\text{init}}) - V_{LB}(b^{\text{init}}) \leq \epsilon \quad \epsilon \in (0, +\infty)$
- the LP performing the update can be replaced by any of the bandits



Comparison on OS-POSG

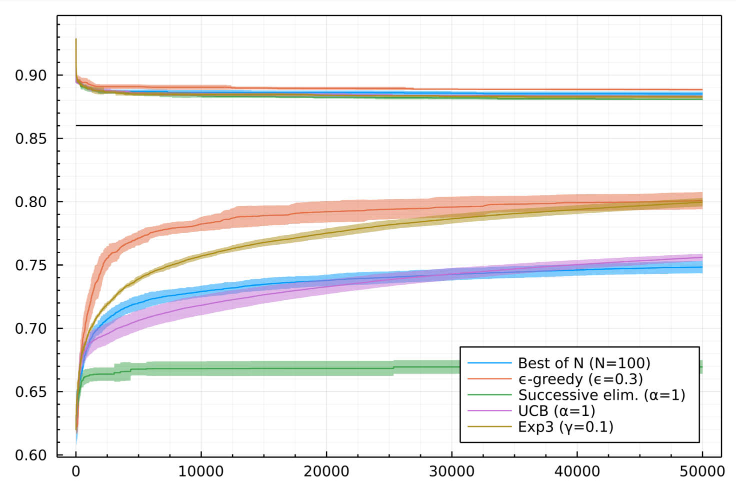


Figure 2: Dependence of the values in b^{init} on number of bound updates. Bounding value functions approach the optimal value.

Conclusion

- successfully integrated and compared multi-armed bandit algorithms into value iteration and HSVI
- **observable** variants consistently better than standard bandits (on SGs)
- Exp3 performs well on both models
- ObservableUCB usually same or better than Exp3 (on SGs)
- high exploration is important in HSVI
- Successive elimination, Best of N and standard UCB rarely get close to the optimal value

Thank you for your attention

"Averaging" factor

A method used in SGs to compute refined value function:

$$V^{t+1}(s) = V^{t}(s) + \delta(t) \cdot (v - V^{t}(s)),$$

where v is the immediate value from current iteration and $\delta(t)$ one of the following functions of time t:

- $\delta(t) = \frac{1}{t}$
- $\delta(t) = \frac{1}{\sqrt{t}}$

Parametrisation and adaptive fitting?

Suggestions

$$V^{t+1}(s) = V^t(s) + \delta(t) \cdot (v - V^t(s))$$

Parametrisation

- $\delta(t, a) = t^{-\frac{1}{a}} \qquad a > 0$
- hyperparameter a can be tuned for each problem separately

Adaptive fitting

- gradually decrease the $\delta(t)$ factor in time
  - at the beggining faster changes in value are beneficial
  - towards the end decrease $\delta(t)$ to prevent fluctuations caused by exploring bandits
  - linear decay, exponential decay from some initial setting
- set $\delta(t)$ based on $d = |v - V^t(s)|$
  - d grows $\rightarrow$ decrease $\delta(t)$ (towards i.e. $\frac{1}{t}$)
  - d gets smaller $\rightarrow$ increase $\delta(t)$ (towards i.e. $\frac{1}{\sqrt{t}}$)

Suggestions (HSVI)

Similar idea for HSVI

- monitor progress of gap $g = V_{UB}(b^{\text{init}}) - V_{LB}(b^{\text{init}})$
- if the decrease of g starts to be slow $\rightarrow$ increase exploration