No-Regret Exploration in Goal-Oriented Reinforcement Learning

Jean Tarbouriech[†§], **Evrard Garcelon**[†], **Michal Valko**[§], **Matteo Pirotta**[†], **Alessandro Lazaric**[†]

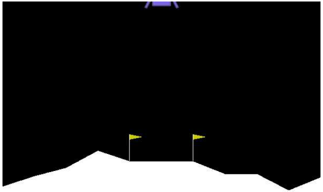
[†]**Facebook AI Research and** [§]**SequeL team, Inria Lille**

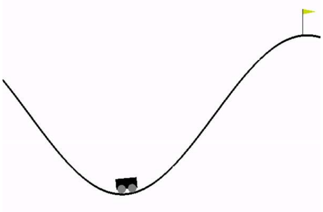
International Conference of Machine Learning (ICML), 2020

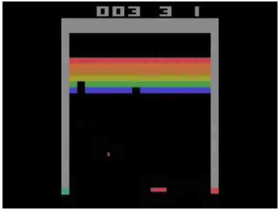
FACEBOOK AI

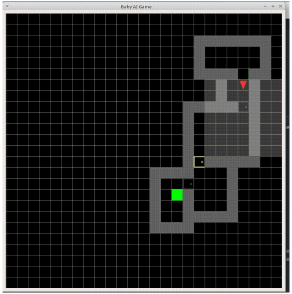
Stochastic Shortest Path (SSP)

[Bertsekas, 2012]









Many popular RL problems are *goal-oriented* tasks:
Minimize the cumulative *cost to reach the goal*

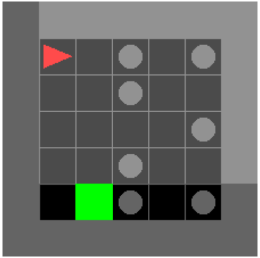
This paper
First study of *exploration-exploitation dilemma*
in goal-oriented RL

(image sources 1, 2, 3, 4)

SSP-Markov Decision Process

[Bertsekas, 2012]

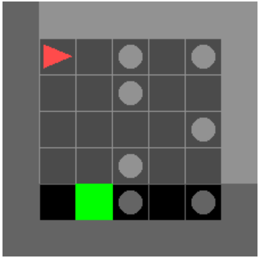
- State space $\mathcal{S}' = \mathcal{S} \cup \{\overline{s}\}$
  - Starting state $s_0 \in \mathcal{S}$ ( )
  - Goal state $\overline{s}$ ( )
- Action space $\mathcal{A} = \left\{ \textit{Up}, \textit{Down}, \textit{Left}, \textit{Right} \right\}$
- Transition $p(s'|s,a)$
  - Goal is absorbing $p(\overline{s}|\overline{s},a) = 1$
- Cost function $c(s,a)$
  - Empty state $c(s,a) = 1$ ( )
  - Easy state $c(s,a) = 0.1$ ( )
  - Goal state $c(\overline{s},a) = 0$



SSP-Markov Decision Process

[Bertsekas, 2012]

- State space $\mathcal{S}' = \mathcal{S} \cup \{\overline{s}\}$
  - Starting state $s_0 \in \mathcal{S}$ ()
  - Goal state $\overline{s}$ ()
- Action space $\mathcal{A} = \left\{ \textit{Up}, \textit{Down}, \textit{Left}, \textit{Right} \right\}$
- Transition $p(s'|s,a)$
  - Goal is absorbing $p(\overline{s}|\overline{s},a) = 1$
- Cost function $c(s,a)$
  - Empty state $c(s,a) = 1$ ()
  - Easy state $c(s,a) = 0.1$ ()
  - Goal state $c(\overline{s},a) = 0$



Discounted and *finite-horizon MDPs* are sub-classes of *SSP-MDPs* [e.g., Bertsekas, 2012]

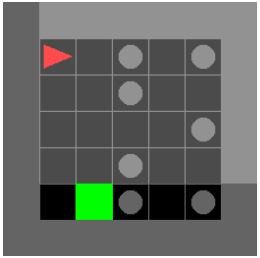
SSP-MDP
[Bertsekas, 2012]

- Policy $\pi : \mathcal{S} \to \mathcal{A}$
- Goal-reaching time

$$\tau_\pi(s) := \min \left\{ t \geq 0 : s_{t+1} = \overline{s} \,|\, s_1 = s, \pi \right\}$$

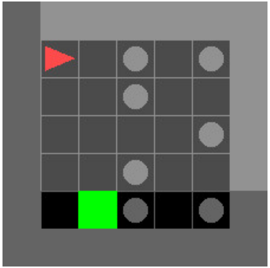
- Value function

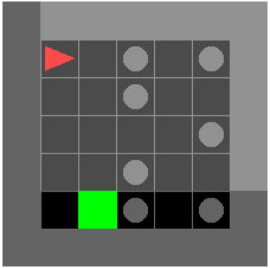
$$V^\pi(s) := \mathbb{E}\left[\sum_{t=1}^{\tau_\pi(s)} c(s_t, \pi(s_t)) \,\middle|\, s_1 = s \right]$$

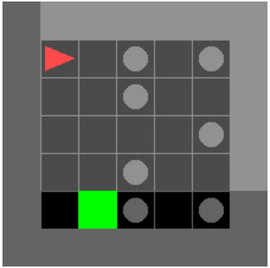


SSP-MDP

Policy categorization





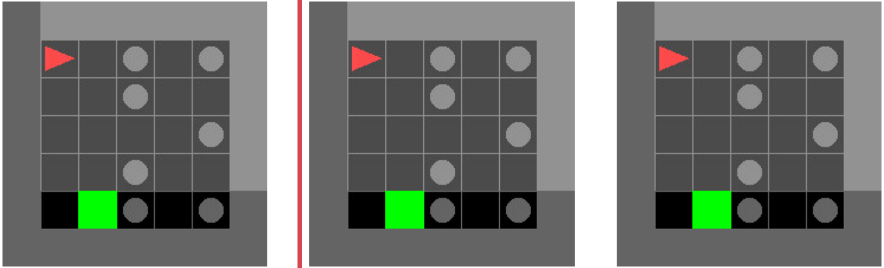


Time to goal $\tau_\pi(s_0)$	$+\infty$	6	9
Cumulative cost	$+\infty$	5	4.4

Proper policies

SSP-MDP

Policy categorization



Time to goal $\tau_\pi(s_0)$	$+\infty$	6	9
Cumulative cost	$+\infty$	5	4.4

Proper policies

$\Rightarrow$ Objective: *reach the goal* while *minimizing the cumulative cost*

Assumptions

Assumption

① *There exist known constants* $0 < c_{\min} \leq c_{\max}$ *such that* $c(s,a) \in [c_{\min}, c_{\max}]$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$.

② *There exists at least one proper policy (i.e., that reaches the goal* $\overline{s}$ *with probability 1 from any state in* $\mathcal{S}$*).*

Assumptions

Assumption

① *There exist known constants* $0 < c_{\min} \leq c_{\max}$ *such that* $c(s,a) \in [c_{\min}, c_{\max}]$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*.*

② *There exists at least one proper policy (i.e., that reaches the goal* $\overline{s}$ *with probability 1 from any state in* $\mathcal{S}$*).*

Lemma (SSP problem is well-posed, see [Bertsekas, 2012])

Under Asm. 1 & 2, there exists an optimal policy that is proper, stationary and deterministic

$$\pi^\star \in \arg\min_{\pi} V^\pi$$

Assumptions

Assumption

① *There exist known constants* $0 < c_{\min} \leq c_{\max}$ *such that* $c(s,a) \in [c_{\min}, c_{\max}]$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$.

② *There exists at least one proper policy (i.e., that reaches the goal* $\overline{s}$ *with probability 1 from any state in* $\mathcal{S}$*).*

Lemma (SSP problem is well-posed, see [Bertsekas, 2012])

Under Asm. 1 & 2, there exists an optimal policy that is proper, stationary and deterministic

$$\pi^\star \in \arg\min_\pi V^\pi$$

Lemma

Under Asm. 1 & 2, we have $\|V^\star\|_\infty \leq c_{\max} D$, *where we introduce the SSP-diameter* D

$$D := \max_{s \in \mathcal{S}} \underbrace{\min_\pi \mathbb{E}\left[\tau_\pi(s)\right]}_{\text{shortest path from } s \text{ to the goal } \overline{s}} < +\infty$$

Learning Problem

Input: $\mathcal{S}, \overline{s}, \mathcal{A}, c_{\min}, c_{\max}$, **no prior knowledge of** p

for *episodes* $k = 1, 2, \dots, K$ **do**

  Set $h = 0$ and initial state $s_{k,0} = s_0$

  while $s_{k,h} \neq \overline{s}$ **do**

    Execute $a_{k,h} = \pi_{k,h}(s_{k,h})$

    Observe cost $c_{k,h} = c(s_{k,h}, a_{k,h})$ and next state $s_{k,h+1} \sim p(\cdot | s_{k,h}, a_{k,h})$

    Update policy $\pi_{k,h}$

    Set $h = h + 1$

  end

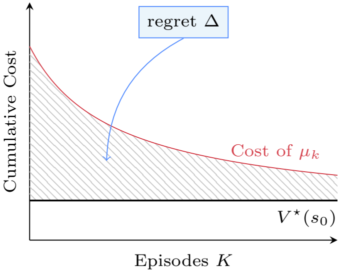
end

Learning Problem

If executing a non-proper policy π_k, episodes may never terminate

Input: $\mathcal{S}, \overline{s}, \mathcal{A}, c_{\min}, c_{\max}$, **no prior knowledge of p**
for *episodes* $k = 1, 2, \ldots, K$ **do**
  Set $h = 0$ and initial state $s_{k,0} = s_0$
  while $s_{k,h} \neq \overline{s}$ **do**
    Execute $a_{k,h} = \pi_{k,h}(s_{k,h})$
    Observe cost $c_{k,h} = c(s_{k,h}, a_{k,h})$ and next state $s_{k,h+1} \sim p(\cdot|s_{k,h}, a_{k,h})$
    Update policy $\pi_{k,h}$
    Set $h = h + 1$
  end
end

SSP Regret



length of episode k

$$\Delta(\mathfrak{A}, K) := \sum_{k=1}^{K} \left[\left(\sum_{h=1}^{\tau_k(s_0)} c\big(s_{k,h} \, , \, \mu_k(s_{k,h}) \big) \right) - V^\star(s_0) \right]$$

h-th state visited during episode k

possibly non-stationary policy followed during episode k

In finite horizon we consider the expected performance of μ_k: $\sum_{k=1}^{K} \left[\boldsymbol{V^{\mu_k}(s_0)} - V^\star(s_0) \right]$

UC-SSP: Upper-Confidence SSP

Input: $\mathcal{S}, \overline{s}, \mathcal{A}, c_{\min}, c_{\max}$
for *episodes* $k = 1, 2, \ldots, K$ **do**
  ① Compute an *optimistic cost-weighted* SSP policy $\widetilde{\pi}_k$
  ② *Execute* policy $\widetilde{\pi}_k$ for up to H_k *steps*
  if $\overline{s}$ *is not reached* **then**
    Reach the goal *as fast as possible*,
    by performing ① + ② with unit costs $c(s, a) = 1$, $c(\overline{s}, a) = 0$
  end
end

UC-SSP: Upper-Confidence SSP

1) How to compute the policy $\widetilde{\pi}_k$?

Input: $\mathcal{S}, \overline{s}, \mathcal{A}, c_{\min}, c_{\max}$
for *episodes* $k = 1, 2, \ldots, K$ **do**
    ① Compute an *optimistic cost-weighted* SSP policy $\widetilde{\pi}_k$
    ② *Execute* policy $\widetilde{\pi}_k$ for up to H_k *steps*
    if $\overline{s}$ *is not reached* **then**
        Reach the goal *as fast as possible*,
        by performing ① + ② with unit costs $c(s, a) = 1$, $c(\overline{s}, a) = 0$
    end
end

2) How to select the horizon H_k?

1) How to compute the policy $\widetilde{\pi}_k$?

We introduce an *Extended Value Iteration* scheme tailored to SSP problems.

Objective: select a policy $\widetilde{\pi}_k$ with **lowest optimistic value** $\widetilde{V}_k$.

Lemma

With high probability, for any episode k, we have for any $s \in \mathcal{S}$,

$$\widetilde{V}_k(s) \leq V^\star(s)$$

2) How to select the horizon H_k?

Denote by $\widetilde{\tau}_k$ the *optimistic* goal-reaching time of the policy $\widetilde{\pi}_k$.

The horizon H_k is selected such that

$$\max_{s \in \mathcal{S}} \mathbb{P}\Big(\widetilde{\tau}_k(s) \geq H_k\Big)$$

is small enough.

Lemma

With high probability, for any episode k,

$$H_k \leq \left\lceil 6 \frac{c_{\max}}{c_{\min}} \, D \, \log(2\sqrt{K}) \right\rceil$$

SSP-diameter $D := \max_{s \in \mathcal{S}} \min_{\pi} \mathbb{E}\left[\tau_\pi(s)\right]$

Regret Guarantee of UC-SSP

Theorem

For any tabular SSP-MDP with $c_{\min} > 0$, the regret of UC-SSP *can be bounded with high probability as follows:*

$$\Delta(\text{UC-SSP}, K) = \widetilde{O}\left(c_{\max} DS \sqrt{\frac{c_{\max}}{c_{\min}} ADK} \;+\; c_{\max} S^2 AD^2 \right)$$

- Dominant $\sqrt{K}$-order optimal term
- Small constant “burn-in” term

UC-SSP is the first no-regret learning algorithm for SSP

Extensions

$c_{\min} = 0$

- We offset all the costs by a *small additive perturbation* [Bertsekas and Yu, 2013]
- We directly obtain a $\widetilde{O}(K^{2/3})$ *regret*
- Later work [Cohen et al., 2020] (at this ICML) devise an algorithm with Bernstein inequalities with a $\widetilde{O}(\sqrt{K})$ regret when $c_{\min} = 0$

$D = +\infty$

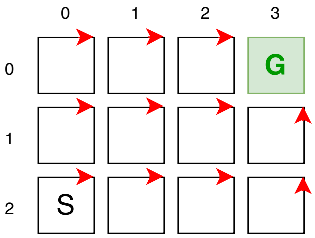
- The SSP-MDP is non-communicating
- We *truncate the SSP Bellman operator* to avoid divergence at *dead-end* states
- The regret's dependency on D is replaced by a known upper bound of $V^\star(s_0)$

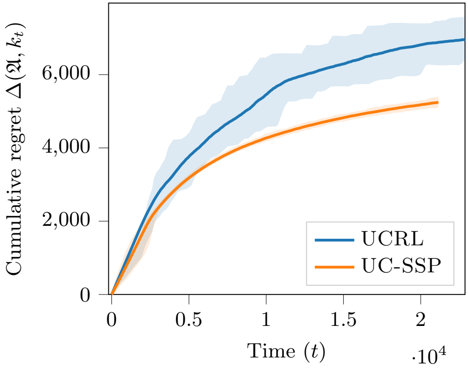
Experimental validation

(see paper for additional experiments)

If $c(s,a) = 1$ for all $s \neq \overline{s}$ and all a (i.e., uniform cost), the SSP problem is equivalent to an infinite-horizon undiscounted problem.

- UCRL2 [Jaksch et al., 2010] achieves sub-linear SSP-regret*
- However UC-SSP achieves better performance





* UCRL2 is an algorithm for regret minimization in average reward MDPs

Conclusion

Summary

- Most of the theoretical literature on exploration focused on finite-horizon and average-reward
- **SSP** is a *more general and practical* setting
- We propose the *first exploration-exploitation* algorithm for SSP

Future work

- Model-free exploration in SSP
- Linear function approximation

Details are in our paper:

No-Regret Exploration in Goal-Oriented Reinforcement Learning

https://arxiv.org/pdf/1912.03517

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, Alessandro Lazaric

Thank you

FACEBOOK AI

Dimitri Bertsekas. *Dynamic programming and optimal control*, volume 2. 2012.

Dimitri P Bertsekas and Huizhen Yu. Stochastic shortest path problems under weak conditions. *Lab. for Information and Decision Systems Report LIDS-P-2909, MIT*, 2013.

Alon Cohen, Haim Kaplan, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret bounds for stochastic shortest path. *CoRR*, abs/2002.09869, 2020.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.