Exploration in
Goal-Oriented Reinforcement Learning

Part 1

Matteo Pirotta

Meta FAIR

The 15th European Workshop on Reinforcement Learning (EWRL 2022)
Milan, Italy
September 19, 2022

Meta

Collaborators

- Jean Tarbouriech (META FAIR)
- Alessandro Lazaric (META FAIR)
- Evrard Garcelon (META FAIR)
- Andrea Tirinzoni (META FAIR)
- Michal Valko (Deepmind)
- Simon Du (University of Washington)
- Runlong Zhou (Tsinghua University)
- Liyu Chen (University of Southern California)

Why Talking about Goal-Oriented Problems?

- Growing interest, especially in deep RL community
  - Many applications are goal-oriented (reward driven)
  - Goal-conditioning RL (generalization)
  - Unsupervised RL (generalization)
- Impressive results in complex domains

Goal-Oriented Reinforcement Learning

Holds the promise to model and learn goal-oriented behavior

Learn to **reach the goal** state with **minimum total expected cost**

Markov decision Process (MDP)

- State space $\mathcal{S}$
- Action space $\mathcal{A}$
- Transition probabilities $p(s'|s, a)$
- Cost function $c(s, a) \in [0, 1]$ (= negative reward)
- Goal state $g \in \mathcal{S}$

Policy Value

Find the policy $\pi : \mathcal{S} \to \mathcal{A}$ that minimizes the expected cumulative cost

$$\min_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=1}^{\infty} \omega(t) c(s_t, a_t) \right]$$

	Finite Horizon	Infinite-Horizon Discounted	**Goal-Oriented** (a.k.a. stochastic shortest path)
Weights $\omega(t)$	$\mathbb{1}[t \leq H]$	γ^{t-1}	$\mathbb{1}[\boldsymbol{t \leq \tau_{\pi}}]$
Intrinsic Horizon	H	$\frac{1}{1-\gamma}$	$\tau_{\pi} := \inf \{t \geq 1 : s_t = g, \pi\}$

- H and γ fixed and known in advance
- τ_{π} is a **random variable**. It may be ∞ for many policies

Stochastic Shortest Path (SSP)

- SSP strictly generalizes the finite-horizon and discounted models [Guillot and Stauffer, 2020]
- SSP captures tasks with varying and unknown horizon

Stochastic Shortest Path (SSP): value functions

- Goal-reaching (hitting) time

$$\tau_\pi(s \to g) := \inf \{t \geq 1 : s_1 = s, s_t = g, \pi\}$$

- Value Functions of policy π

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

$$Q^\pi(s, a, g) := \mathbb{E}\left[\sum_{t=1}^{\pi_\pi(s \to g)} c(s_t, a_t) | s_1 = s, a_1 = a, a_t = \pi(s_t)\right]$$

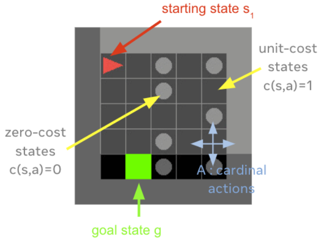
* $p(g|g, a) = 1, c(g, a) = 0$

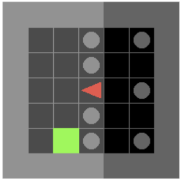
Example

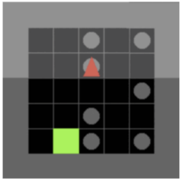
improper policies

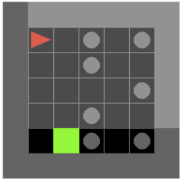
$\mathbb{E}[\tau_\pi(s_1 \to g)]$	$+\infty$	$+\infty$	5	7
$V^\pi(s_1 \to g) := \mathbb{E}\left[\sum_{t=1}^{\tau_\pi(s \to g)} c(s_t, \pi(s_t)) \vert s_1 = s\right]$	$+\infty$	2	5	3

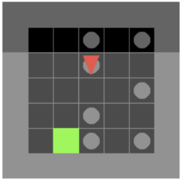
A policy is ***proper*** if it reaches the goal g with probability 1 from any state











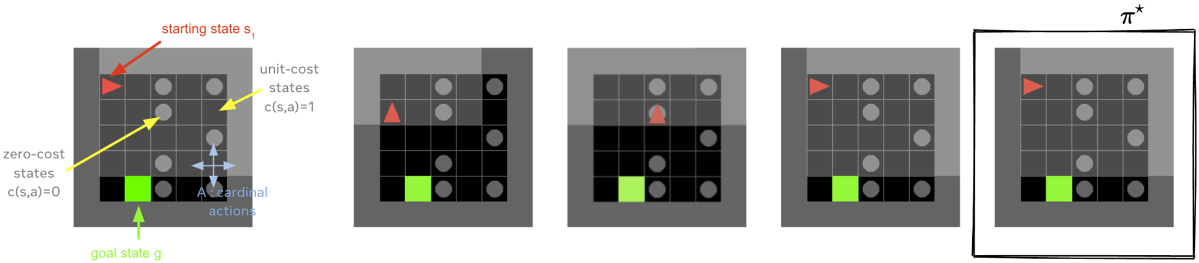
Optimal policy

Trade-off between two objectives

$$\pi^\star = \arg\min_\pi V^\pi(s \to g)$$

$$\text{s.t.} \quad \max_s \mathbb{E}[\tau_\pi(s \to g)] < \infty$$

- Object 1: minimize the cumulative cost
- Object 2: reach the goal*



*assumed to be reachable

Important Quantities in SSP

[Tarbouriech et al., 2021c, Cohen et al., 2021]

- Minimum cost $c_{\min} = \min_{s \neq g, a} c(s, a)$
- Value function bound $B_\star = \max_s \{V^\star(s)\}$
- Hitting time bound $T_\star = \max_s \{T^{\pi_\star}(s)\}$
- Diameter $D = \max_s \min_\pi T^\pi(s)$

Are they related?

$$B_\star \leq D \leq T_\star \leq \frac{B_\star}{c_{\min}}$$

*assuming $c(s, a) \in [0, 1]$

Summary

- A policy is proper if it reaches g with probability 1 starting from any state
- Assumption: there exists at least one proper policy
- We denote by $\pi^\star$ the optimal proper policy

$$T^\pi(s) = \mathbb{E}[\tau_\pi(s \to g)] \qquad \pi^\star \in \arg\min_{\pi:\|T^\pi\|_\infty < \infty} V^\pi$$

- Important quantities

$$B_\star = \max_s\{V^\star(s)\} \qquad T_\star = \max_s\{T^{\pi_\star}(s)\} \qquad B_\star \le T_\star \le \frac{B_\star}{c_{\min}}$$

*assuming $c(s,a) \in [0,1]$

Planning in SSP

Bellman Equations

- For any stationary policy we define the Bellman operator

$$L^{\pi}V(s) = c(s, \pi(s)) + \sum_{y} p(y|s, \pi(s))V^{\pi}(y)$$

- Optimal Bellman Operator

$$LV(s) = \min_{a} \left\{ c(s, a) + \sum_{y} p(y|s, a)V^{\pi}(y) \right\}$$

The fixed point equations are generally expected to hold in MDP models. Yet this may not be the case in SSP [Bertsekas and Tsitsiklis, 1991]

Classical SSP Assumptions

If

1. There exists at least one proper policy (gurанteed when $c_{\min} < 0$)
2. For every improper policy there is at least one state s such that $V^{\pi}(s) = +\infty$

Then

- The optimal value function is the unique solution of $V^{\star} = LV^{\star}$
- A stationary policy is optimal if and only if $L^{\pi}V^{\star} = LV^{\star}$
- The method of value iteration converges to $V^{\star}$ from every initial vector
- The method of policy iteration yields an optimal proper policies starting from a proper policy
- The optimal value function and policy can be computed using linear programming

Value Iteration

Input: p and c
Set $V_0 = 0$
for $k = 1, 2, \ldots$ **do**
| $V_k = LV_{k-1}$

Still not easy to define a termination condition

- L may not be a contraction w.r.t. any norm
- If **all the stationary policies are proper**, L is a contraction in a weighted sup-norm

Stochastic Shortest Path (SSP)

Planning in SSP studied since the 1990s [Bertsekas and Tsitsiklis, 1991]

Online learning in SSP has been studied only recently

Online Learning Problem

- Transitions P and costs c are unknown
- Episode k starts at s_1 and ends if and only if goal g is reached
- We compete against the optimal proper policy

$$\pi^\star = \arg\min_{\pi \text{ proper}} V^\pi = \arg\min_{\pi : \|T^\pi\|_\infty < \infty} V^\pi$$

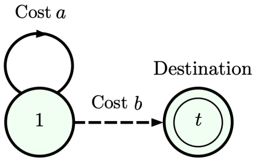


*figure from [Bertsekas and Yu, 2013]

Online Learning Problem

Input: $\mathcal{S}, g, \mathcal{A}$, **no prior knowledge of p and c**
for *episodes* $k = 1, 2, \ldots, K$ **do**
Set $t = 0$ and initial state $s_t = s_1$
while $s_{k,h} \neq g$ **do**
Execute $a_t = \pi_t(s_t)$
Observe cost c_t and next state $s_{t+1} \sim P(\cdot|s_t, a_t)$
Update policy π_{t+1}
Set $t = t + 1$

Question: how do we evaluate the performance of an algorithm?

❶ Sample-Complexity

How many samples are sufficient to compute a near-optimal policy w.h.p.?

❶ Sample-Complexity

How many samples are sufficient to compute a near-optimal policy w.h.p.?

Let $\mathcal{T}$ be the random stopping time by when an algorithm terminates and returns a policy $\widehat{\pi}$. An algorithm is (ε, δ)-correct algorithm with sample complexity $N(\mathfrak{A})$ if

$$\mathbb{P}\left[\mathcal{T} \leq N(\mathfrak{A}), \;\; \|V^{\pi_t} - \min_{\pi:\text{proper}} V^{\pi}\|_\infty \leq \varepsilon\right] \geq 1 - \delta$$

and $N(\mathfrak{A}) \lesssim \text{poly}\left(\frac{1}{\varepsilon}, \log(1/\delta), B_\star, T_\star, S, A\right)$.

* $\lesssim$ hides possibly constants and logarithmic factors

❷ Regret

How much suboptimal is the total cost of the algorithm compared to executing the optimal policy?

❷ Regret

How much suboptimal is the total cost of the algorithm compared to executing the optimal policy?

Let I_k be the length of episode k and

$$R_K := \sum_{k=1}^{K} \left[\left(\sum_{h=1}^{I_k} c_{k,h} - \min_{\pi:\text{proper}} V^{\pi}(s_1) \right) \right]$$

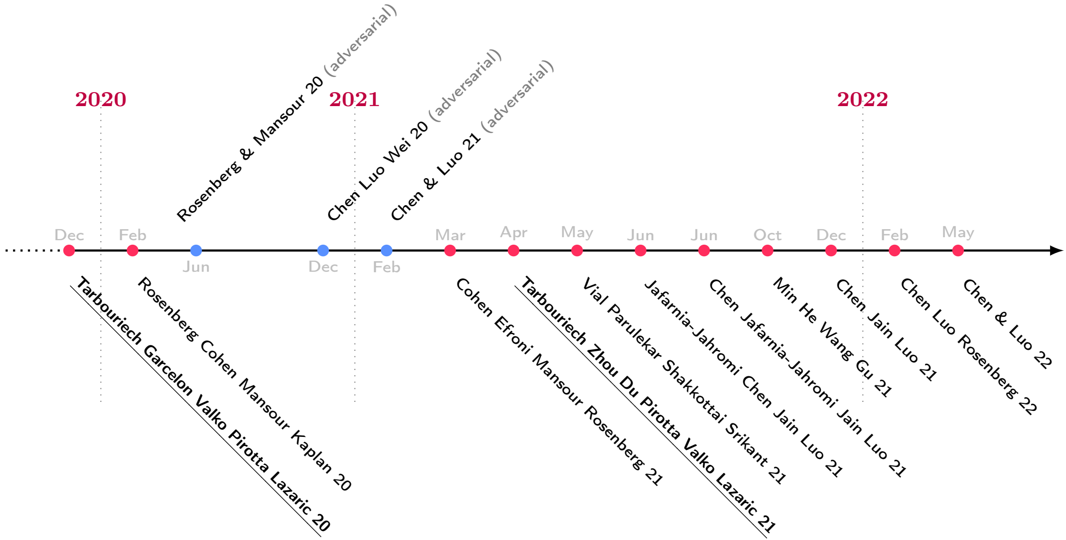
Then an algorithm has sublinear regret if

$$R_K \leq \text{poly}(S, A, B_\star, T_\star, \log(1/\delta)) \cdot K^{\alpha}, \quad 0 < \alpha < 1$$

- If $\exists k, I^k = \infty$, then we define $R_K = \infty$
- The algorithm may execute a non-stationary policy π_k in episode k

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

Regret Minimization in SSP



*we consider SSP with loops (i.e., episodes last as long as the goal is reached)

What is the best performance we can achieve?

Minimax Lower Bound

Theorem ([Rosenberg et al., 2020])

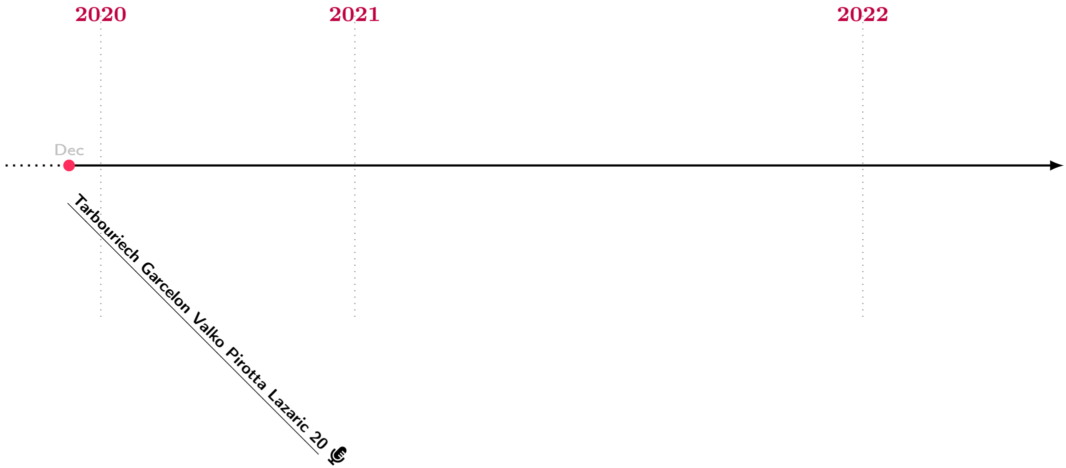
There exists a SSP-MDP with S states, A actions and $B_\star = \max_s \{V^\star(s)\} \geq 1$, any algorithm $\mathfrak{A}$ at any episode K suffers a regret of at least

$$\Omega\left(B_\star \sqrt{SAK}\right)$$

* if $B_\star < 1$ the lower bound is $\Omega(\sqrt{B_\star SAK})$ [Cohen et al., 2021]

Regret Upper-Bounds

The start...


2020
2021
2022
Dec
Tarbouriech Garcelon Valko Pirotta Lazaric 20

UC-SSP: Upper-Confidence SSP

The first algorithm for regret minimization in SSP

Input: $\mathcal{S}, g, \mathcal{A}$

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

① Compute an optimistic cost-weighted SSP policy π_k

② Execute policy π_k for up to H_k steps

if g *is not reached* **then**

Reach the goal **as fast as possible**,

by performing ① + ② with unit costs $c(s,a) = 1$, $c(g,a) = 0$

UC-SSP: Upper-Confidence SSP

The first algorithm for regret minimization in SSP

1) How to compute the policy π_k?

Input: $\mathcal{S}, g, \mathcal{A}$

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

① Compute an optimistic cost-weighted SSP policy π_k

② Execute policy π_k for up to H_k steps

if g *is not reached* **then**

Reach the goal **as fast as possible**,

by performing ① + ② with unit costs $c(s, a) = 1$, $c(g, a) = 0$

2) How to select the horizon H_k?

1) How to compute the policy π_k?

Optimism: select a policy π_k with **lowest optimistic value** V_k.

Lemma

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

$$V_k(s) \leq V^\star(s)$$

1) How to compute the policy π_k?

UC-SSP uses model-optimism for SSP based on Hoeffding inequality

Recipe for Model Optimism

1. Build confidence set around empirical transitions and rewards

$$D(p_h(\cdot|s,a), \widehat{p}_h(\cdot|s,a)) \leq \beta^p_{hk}(s,a)$$
$$|r_h(s,a), \widehat{r}_h(s,a)| \leq \beta^r_{hk}(s,a)$$

and, with high probability

$$p_h(s,a) \in B^p_{hk}(s,a), \quad r_h(s,a) \in B^r_{hk}(s,a)$$

2. **Jointly** optimize over models and policies

$$(M_k, \pi_k) \in \underset{M=(p,r)\in(B^p,B^r),\pi}{\arg\min} \left\{ V^{\pi}_{1,M} \right\}$$

2) How to select the horizon H_k?

Denote by τ_k the *optimistic* goal-reaching time of the policy π_k.

The horizon H_k is selected such that

$$\max_{s \in \mathcal{S}} \mathbb{P}\Big(\tau_k(s) \geq H_k\Big) \text{ is small enough}$$

Regret Guarantee of UC-SSP

Theorem

For any tabular SSP-MDP the regret of UC-SSP can be bounded with high probability as follows:

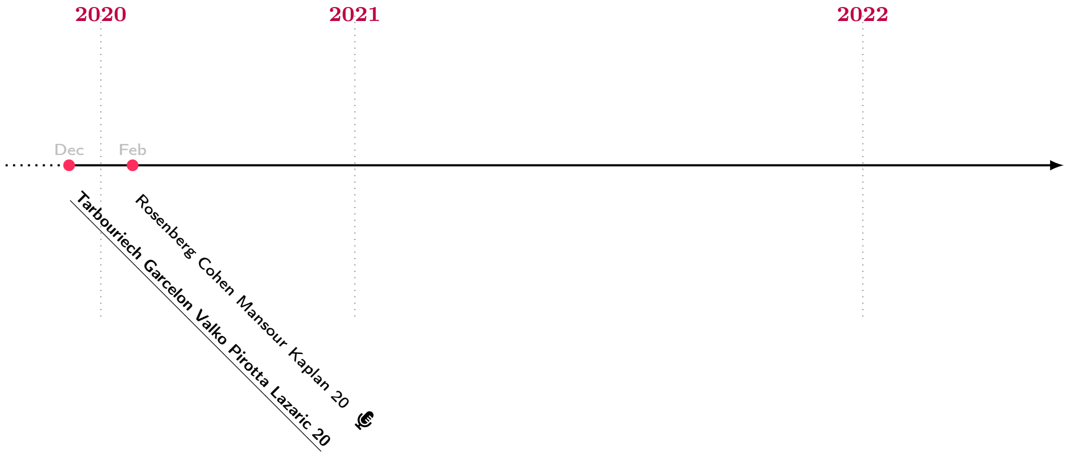
$$R_K \leq \widetilde{O}_K\left(\sqrt{\frac{K}{c_{\min}}}\right) \quad \text{or} \quad R_K \leq \widetilde{O}_K\left(K^{2/3}\right)$$

- Does not require prior knowledge about $B_\star$ or $T_\star$
- Offset all the costs by a small perturbation to deal with the case $c_{\min} = 0$

$$c'(s,a) = \max\{c(s,a), \eta\} \quad \text{or} \quad c'(s,a) = c(s,a) + \eta$$

$$\eta = \frac{1}{poly(K)}$$

A First Improvement...



*we consider SSP with loops (i.e., episodes last as long as the goal is reached)

UCRL2-based Algorithm

[Rosenberg et al., 2020]

Input: $\mathcal{S}, g, \mathcal{A}$
for *episodes* $k = 1, 2, \ldots, K$ **do**
  $s_t = s_1$
  while g *is not reached* **do**
    if *some quantity is "doubled"* **then**
      Compute optimistic policy π_t
    Execute policy π_t

- Simple algorithm based on the principle of model optimism (based on UCRL2)
  - Leverages Bernstein-like confidence intervals for the model
- Very smart and refined analysis
- Use cost perturbation to deal with $c_{\min} = 0$

* *condition* is the usual doubling condition of UCRL2 [Jaksch et al., 2010], an algorithm for regret minimization in average reward

UCRL2-based Algorithm: Regret Guarantee

Theorem

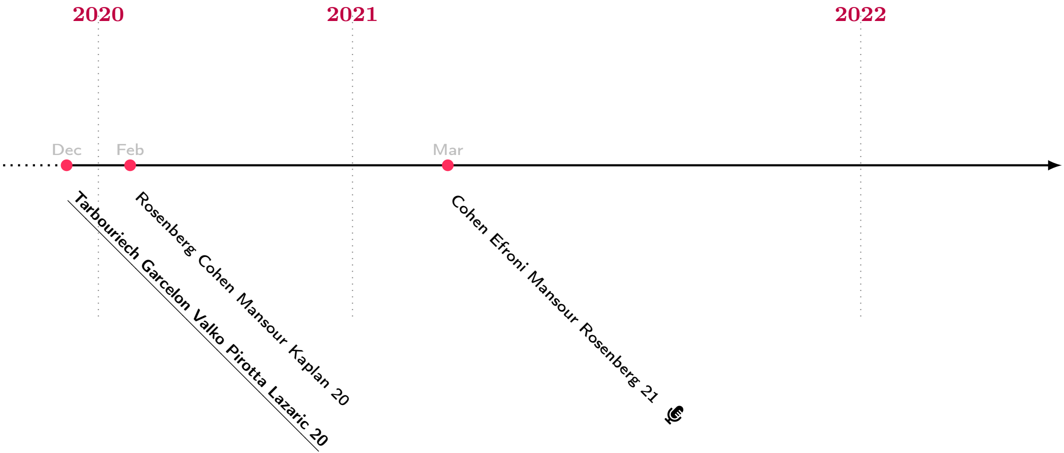
For any tabular SSP-MDP the regret of [Rosenberg et al., 2020] can be bounded with high probability as follows:

$$R_K \leq \widetilde{O}\left(B_\star S\sqrt{AK}\right)$$

where $B_\star$ is provided as prior knowledge to the algorithm

- $\sqrt{K}$ also in the case of $c_{\min} = 0$
- Requires prior knowledge about $B_\star$ (otherwise worse bound $B_\star^{3/2}$)
- Not yet minimax optimal

A Minimax Algorithm...



*we consider SSP with loops (i.e., episodes last as long as the goal is reached)

A Novel Reduction from SSP to Finite-Horizon*

[Cohen et al., 2021]

Input: $\mathcal{S}, g, \mathcal{A}, B_\star, T_\star$, an algorithm $\mathfrak{A}_{\mathrm{FH}}$ for regret min. in finite-horizon MDPs

Set horizon $H = O(T_\star \log(K))$

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

  $s_t = s_1$

  while g *is not reached* **do**

    Run one episode of $\mathfrak{A}_{\mathrm{FH}}$ from the current state

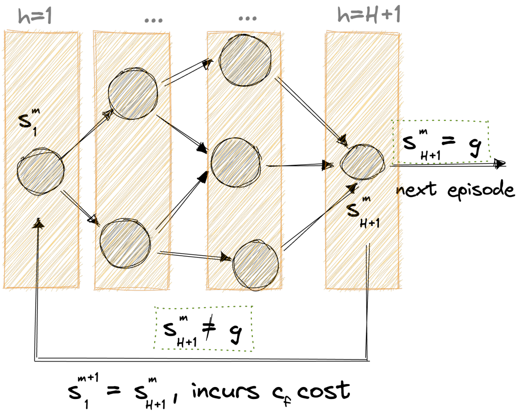
    if g *was reached* **then**

      Pad trajectory to be of length H and feed it to $\mathfrak{A}_{\mathrm{FH}}$

    else

      Give an additional terminal cost of $O(B_\star)$

      Feed trajectory and terminal cost to $\mathfrak{A}_{\mathrm{FH}}$



* [Chen and Luo, 2021] and [Chen et al., 2021b] use a different reduction to finite-horizon in adversarial SSP.

The Finite-Horizon Model

- The finite-horizon MDP M_H with horizon H
  - Same transitions P and costs c as in the SSP

$$\widehat{c}(s,a) = c(s,a)\mathbb{I}(s \neq g), \qquad \widehat{P}(s'|s,a) = \begin{cases} P(s'|s,a) & s \neq g \\ 1 & s = g, s' = g \end{cases}$$

  - Additional terminal cost $c_f(s) = O(B_\star \mathbb{1}\{s \neq g\})$

- The value function

$$V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} \widehat{c}(s_{h'}, a_{h'}) + c_f(s_{H+1}) \middle| a_{h'} = \pi_{h'}(s_{h'})\right]$$

for $H = \widetilde{O}(T_\star)$, $V^\star(s) \approx V_1^\star(s) = \arg\min_{\pi=(\pi_h)} V_h^\pi$

Properties of The Finite-Horizon Algorithm

- Since we estimate P and c, the FH algorithm should be
  - Model-based (i.e, keeps estimates of P and c)
  - Greedy w.r.t. an estimated Q-function
  - Optimistic
  - Fast enough. After a certain number of visits, the error in estimated P and c should decrease at a proper rate

A "Novel" Finite-Horizon Algorithm

- They proposed ULCVI a value optimistic algorithm for finite-horizon
  $\implies$ Maintains both an optimistic and a pessimistic estimate of the Q-function

Recipe for Value Optimism

1. Compute exploration bonus $b_{hk}(s,a)$
2. Solve optimistic Bellman equation

$$Q_{hk}(s,a) = c_{hk}(s,a) - b_{hk}(s,a) + \widehat{p}_{hk}(s,a)V_{h+1,k}$$

i.e., value iteration on $\overline{M}_k = (\mathcal{S}, \mathcal{A}, \widehat{c}_{hk} - b_{hk}, \widehat{p}_{hk}, H)$

upper confidence bounds directly on the optimal value function $V^\star$

* By leveraging primal and dual LP formulation of the MDP formalism, "*Every model-optimistic algorithm can be written as a value-optimistic algorithm*" [Neu and Pike-Burke, 2020]. b_h is based on the conjugate of the divergence D used for model uncertainty.

A "Novel" Finite-Horizon Algorithm

- They proposed ULCVI a Value Optimistic algorithm for finite-horizon
  $\Longrightarrow$ Maintains both an optimistic and a pessimistic estimate of the Q-function

- They proved a horizon-free[1] regret bound

$$R_{K,FH} = \sum_{k=1}^{K} V_1^{\pi_k}(s_1) - V_1^{\star}(s_1) = \widetilde{O}(B_\star \sqrt{SAK})$$

when $B_\star = \max_{s,h}\{V_h^\star(s)\}$ is known to the algorithm

[1] An algorithm for online finite-horizon MDPs with (expected) total reward bounded by B is (nearly) horizon-free if its regret depends only logarithmically on the horizon H (and polynomially in B)

Reduction from SSP to FH: Regret Guarantee

Theorem

For any tabular SSP-MDP the regret of [Cohen et al., 2021] using ULCVI (with $H = \widetilde{O}(T_\star \log(K))$) can be bounded with high probability as follows:

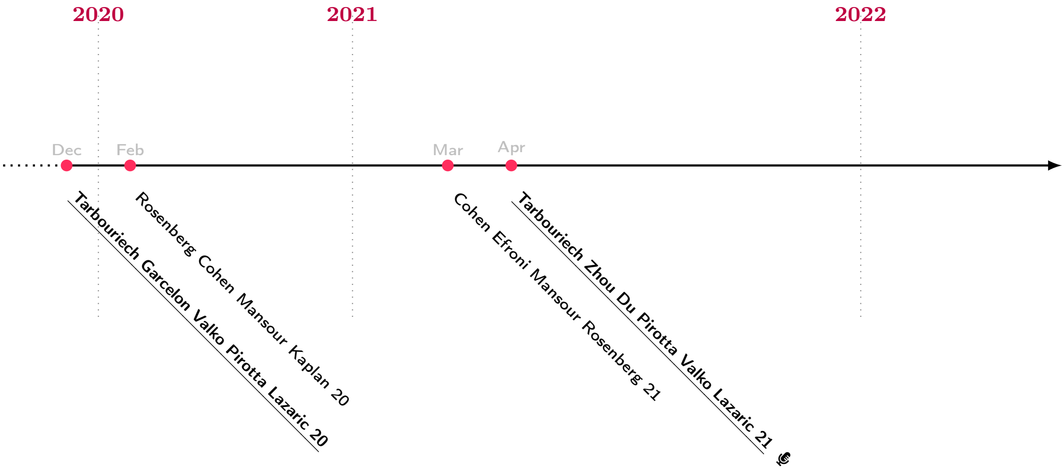
$$R_K \leq \widetilde{O}\left(B_\star \sqrt{SAK}\right)$$

where $B_\star, T_\star$ are provided as prior knowledge to the algorithm

- Minimax optimal
- It runs a non-stationary policy
- Requires prior knowledge of $T_\star$ and $B_\star$[2]

[2] $B_\star$ can be estimated in $T_\star^2 S^2 A$ episodes

Towards a "Better" Minimax Algorithm...



*we consider SSP with loops (i.e., episodes last as long as the goal is reached)

Three desired properties

for a learning algorithm in online SSP

① **Minimax**

$\implies$ regret $\widetilde{O}(B_\star \sqrt{SAK})$

② **Parameter-free**

$\implies$ no knowledge of $B_\star$ and $T_\star$

③ **Horizon-free**

$\implies$ regret depends only logarithmically on $T_\star$.

Three desired properties

for a learning algorithm in online SSP

① **Minimax**

$\Longrightarrow$ regret $\widetilde{O}(B_\star \sqrt{SAK})$

② **Parameter-free**

$\Longrightarrow$ no knowledge of $B_\star$ and $T_\star$

③ **Horizon-free**

$\Longrightarrow$ regret depends only logarithmically on $T_\star$.

While $B_\star \leq T_\star$ always holds, the gap may be *arbitrarily large*

Lower bound: the regret depends on $B_\star$, but a priori not on $T_\star$, even as a lower-order term (see [Rosenberg et al., 2020, Cohen et al., 2021])

Where do we stand...

Algorithm	Approach	Regret	Minimax	Parameters	Horizon-Free
[Tarbouriech et al., 2020a]	Model optim.	$\widetilde{O}_K(\sqrt{K/c_{\min}})$ or $\widetilde{O}_K(K^{2/3})$	No	None	No
[Rosenberg et al., 2020]	Model optim.	$\widetilde{O}\left(B_\star^{3/2} S\sqrt{AK} + T_\star B_\star S^2 A\right)$	No	None	No
		$\widetilde{O}\left(B_\star S\sqrt{AK} + T_\star^{3/2} S^2 A\right)$	No	$B_\star$	No
[Cohen et al., 2021]	Value optim. on finite-horizon reduction	$\widetilde{O}\left(B_\star \sqrt{SAK} + T_\star^4 S^2 A\right)$	Yes	$B_\star$, $T_\star$	No

Lower Bound: $\Omega(B_\star \sqrt{SAK})$

EB-SSP Algorithm
[Tarbouriech et al., 2021c]

Key ingredients:

- Model-based, value optimistic on the non-truncated SSP
- Carefully skews the empirical transitions + perturbs the empirical costs with an exploration bonus
- Induces an optimistic SSP problem whose associated value iteration scheme is guaranteed to converge
- Does not need to known $T_\star$, and uses an adaptive proxy B for unknown $B_\star$

EB-SSP: Algorithmic Idea

Set $C = 0$, $t = 1$
for *episode* $k = 1, \dots, K$ **do**
  while $s_t \neq g$ **do**
    if *some quantity is "doubled"* **then**
      Compute Q_t using VISCO and $\widetilde{B}$
    if $\|Q_t\|_\infty > \widetilde{B}$ *or* $C > \widetilde{B}$ **then**
      Set $\widetilde{B} = 2\widetilde{B}, C = 0$
      Compute Q_t using VISCO and $\widetilde{B}$
    Execute $a_t = \arg\max_a Q_t(s_t, a)$, observe c_t and s_{t+1}
    Set $C = C + c_t$ and $t = t + 1$
  $s_{t+1} = s_1$

EB-SSP: Value Optimism

1. Empirical transitions $\widehat{P}_{s,a,s'}$, empirical costs $\widehat{c}(s,a)$, visit counters $n(s,a)$
2. Slightly goal-skewed empirical transitions $\widetilde{P}$:

$$\widetilde{P}_{s,a,s'} := \frac{n(s,a)}{n(s,a)+1}\widehat{P}_{s,a,s'} + \frac{\mathbb{I}[s'=g]}{n(s,a)+1}$$

Transition model	P	$\widehat{P}$	$\widetilde{P}$
Number of proper policies	At least one	Possibly none	All

3. Refined bonus $b(V,s,a)$

Value Optimism on SSP

Algorithm 1: VISCO: Value Iteration with Slight Goal Optimism

Input: Precision ε

Set $V^{(0)} = 0$

while $\|V^{(i+1)} - V^{(i)}\|_\infty > \varepsilon$ **do**

$$V^{(i+1)} = \max\left\{ \min_{a\in\mathcal{A}} \left\{\widehat{c}(s,a) + \widetilde{P}_{s,a} V - b(V,s,a) \right\}, 0\right\}$$

1. Optimistic
2. Convergence in a finite number of iterations

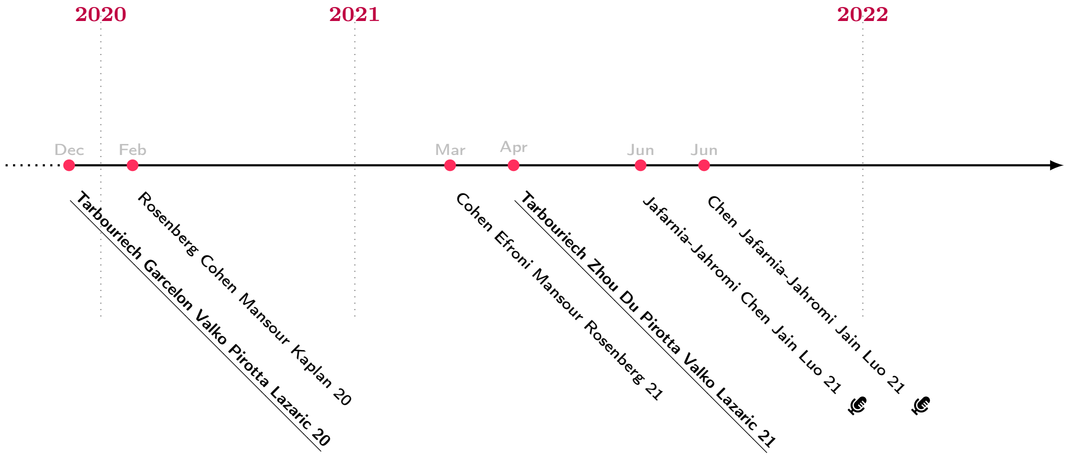
EB-SSP: Regret Guarantees

Algorithm	Approach	Regret	Minimax	Parameters	Horizon-Free
[Tarbouriech et al., 2021c]	Value optim. on non-truncated SSP	$\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star S^2 A\right)$	**Yes**	$B_\star$, $T_\star$	**Yes**
		$\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star S^2 A + \frac{T_\star}{\text{poly}(K)}\right)$	**Yes**	$B_\star$	No*
		$\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star^3 S^3 A\right)$	**Yes**	$T_\star$	**Yes**
		$\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star^3 S^3 A + \frac{T_\star}{\text{poly}(K)}\right)$	**Yes**	**None**	No*

Lower Bound: $\Omega(B_\star\sqrt{SAK})$

* We can show that a $T_\star$ dependence is unavoidable without prior knowledge [Chen et al., 2022]

Other approaches...



*we consider SSP with loops (i.e., episodes last as long as the goal is reached)

Posterior Sampling for SSP

[Jafarnia-Jahromi et al., 2021]

- Keep a Bayesian posterior for the unknown MDP (i.e., model-based)
- A sample from the posterior is used as an estimate of the unknown MDP
- Act greedily on the sampled MDP

Pros and Cons

- 👍 Does not require knowledge of $B_\star$ or $T_\star$, only of the prior μ_1
- 👎 Bayesian regret
- 👎 Not minimax optimal

Implicit Reduction to Finite-Horizon

[Chen et al., 2021a]

- Generic template leveraging an implicit reduction to finite horizon

Algorithm 1 A General Algorithmic Template for SSP

Initialize: $t \leftarrow 0$, $s_1 \leftarrow s_{\text{init}}$, $Q(s, a) \leftarrow 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
for $k = 1, \ldots, K$ **do**
  repeat
    Increment time step $t \overset{+}{\leftarrow} 1$.
    Take action $a_t = \operatorname{argmin}_a Q(s_t, a)$, suffer cost c_t, transit to and observe s'_t.
    Update Q (so that it satisfies Property 1 and Property 2).
    if $s'_t \neq g$ **then** $s_{t+1} \leftarrow s'_t$; **else** $s_{t+1} \leftarrow s_{\text{init}}$, **break**.
  Record $T \leftarrow t$ (that is, the total number of steps).

Property 1: optimism   Property 2: recursive decomposition of estimation error

* Image from [Chen et al., 2021a].

Implicit Reduction to Finite-Horizon

[Chen et al., 2021a]

This template can be instantiated with both model-free and model-based approaches

Algorithm	Approach	Regret	Minimax	Parameters	Horizon-Free
[Chen et al., 2021a]	Model-Free	$\widetilde{O}\left(B_\star\sqrt{SAK} + \frac{B_\star^5 S^2 A}{c_{\min}}\right)$	$\sim$	$B_\star, c_{\min} > 0$	No
		$\widetilde{O}\left(K^{4/5}\right)$	No	$B_\star$	No
	Model-Based	$\widetilde{O}\left(B_\star\sqrt{SAK} + B_\star S^2 A\right)$	Yes	$B_\star$	No

* Can be made parameter-free by leveraging the idea in [Tarbouriech et al., 2021c].

Summary

- Different algorithmic approaches
  - SSP planning + fast policy
  - SSP planning (model optimism, value optimism)
  - Reduction to finite horizon
- Both model-based and model-free algorithms exists
- Minimax optimality only with model-based, and it is possible with a parameter free algorithm

Sample-Complexity

How many samples are sufficient to compute a near-optimal policy w.h.p.?

Sample-Complexity

How many samples are sufficient to compute a near-optimal policy w.h.p.?

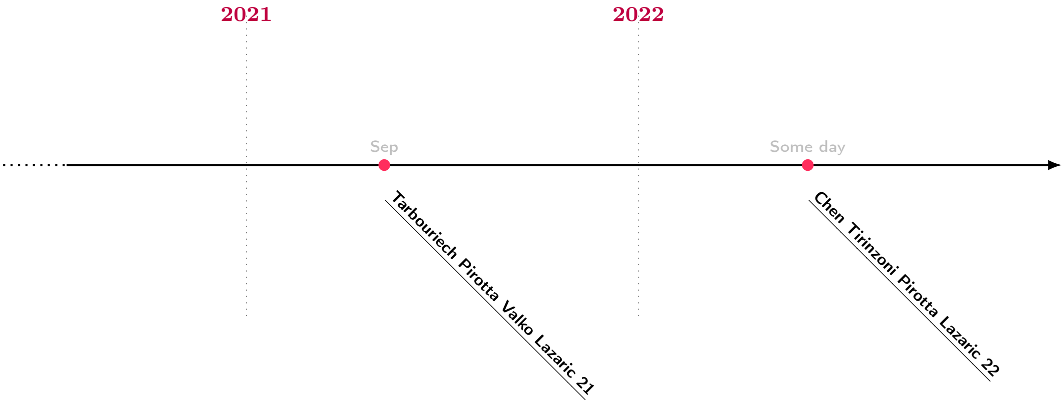
Two standard settings

- **Generative Model**
  We can query the transition model and cost function in any (s, a) pair

- **Online** (a.k.a. best policy identification)
  We need to interact online with the model, no teleporting

⚠ Only the sample complexity with generative model has been studied in the literature

Sample-Complexity in SSP



*work in preparation

Possible direction: Regret-to-PAC conversion?

[Tarbouriech et al., 2021b]

Finite-horizon regret: $\sum_{k=1}^{K} \boldsymbol{V^{\pi_k}(s_1)} - KV^{\star}(s_1)$

Regret bound can be converted to a PAC guarantee by selecting as a candidate optimal solution any policy chosen at random out of all episodes [e.g. Jin et al., 2018]

Possible direction: Regret-to-PAC conversion?

[Tarbouriech et al., 2021b]

Finite-horizon regret: $\sum_{k=1}^{K} \boldsymbol{V^{\pi_k}(s_1)} - KV^\star(s_1)$

Regret bound can be converted to a PAC guarantee by selecting as a candidate optimal solution any policy chosen at random out of all episodes [e.g. Jin et al., 2018]

Challenge in SSP: the regret is defined as:

$$R_K = \Big[\sum_{k=1}^{K} \underbrace{\sum_{\boldsymbol{h=1}}^{\boldsymbol{I_K}} \boldsymbol{c}\big(\boldsymbol{s_{k,h}}, \boldsymbol{\pi_k}(\boldsymbol{s_{k,h}})\big)}_{\text{empirical costs over episode } k} \Big] - KV^\star(s_1)$$

A priori no guarantee on $V^{\pi_k}(s_1)$, which may even be $+\infty$...

Learning Objective

Question:

How many calls to the generative model are sufficient
to compute a near-optimal policy w.h.p.?

Definition

An algorithm is (ε, δ)-correct with sample complexity n, if after n calls to the generative model it returns a policy π that verifies $\|V^\pi - V^\star\|_\infty \leq \varepsilon$ w.p. at least $1 - \delta$.

* We assume there exists a proper policy.

What is the best performance we can achieve?

Learning Without Prior Knowledge

[Chen Tirinzoni Pirotta Lazaric 22]

Theorem

There exists an MDP such that any (ε, δ)-correct algorithm requires

$$\widetilde{\Omega}\left(\frac{B_\star}{c_{\min}} \frac{B_\star^2 S A}{\varepsilon^2}\right)$$

samples.

- Same dependence on S, A and ε as in discounted and finite-horizon case
- $B_\star^2$ connected to the range of the optimal policy
  In discounted setting $(1-\gamma)^{-1}$ bounds V^π for any π
- $B_\star / c_{\min}$ is a bound to the hitting time of the optimal policy ($T_\star \leq \dfrac{B_\star}{c_{\min}}$)

Learning without Prior Knowledge

[Chen et al., 2022]

Theorem

There exists an MDP such that any (ε, δ)-correct algorithm requires

$$\widetilde{\Omega}\left(\frac{B_\star}{c_{\min}} \frac{B_\star^2 SA}{\varepsilon^2}\right)$$

samples.

- $c_{\min} > 0 \implies$ it is possible to adapt to the structure of the problem without prior knowledge (either $B_\star$ or $T_\star$)
- $c_{\min} = 0 \implies$ the problem is not learnable without prior knowledge
  This is in contrast with regret minimization where the regret is bounded in any setting

Learning without Prior Knowledge

[Chen et al., 2022]

Theorem

There exists an MDP such that any (ε, δ)-correct algorithm requires

$$\widetilde{\Omega}\left(\frac{B_\star}{c_{\min}} \frac{B_\star^2 S A}{\varepsilon^2}\right)$$

samples.

- $c_{\min} > 0 \implies$ it is possible to adapt to the structure of the problem without prior knowledge (either $B_\star$ or $T_\star$)
- $c_{\min} = 0 \implies$ the problem is not learnable without prior knowledge
  This is in contrast with regret minimization where the regret is bounded in any setting

Sample complexity in SSPs is strictly harder than in the finite-horizon and discounted case

Learning with Prior Knowledge

[Chen et al., 2022]

Theorem

For any $T \geq T_\star$, there exists an MDP such that any (ε, δ)-correct algorithm knowing T requires

$$\widetilde{\Omega}\left(\min\left\{\frac{B_\star}{c_{\min}}, T\right\}\frac{B_\star^2 SA}{\varepsilon^2}\right)$$

samples.

- T allows the algorithm to focus only on policies such that $\max_s T^\pi(s) \leq T$
- When $T < \dfrac{B_\star}{c_{\min}}$ the algorithm benefits from prior knowledge $\implies$ pruning of policies is effective
- When $T \geq \dfrac{B_\star}{c_{\min}}$ there is no benefit from the prior knowledge

Learning under Restricted Optimality

[Chen et al., 2022]

- If T is too small, the objective may change

$$\pi_T^\star(s) \in \underset{\pi:\|T^\pi\|_\infty \leq T}{\arg\min}\, V^\pi(s), \qquad V_T^\star(s) = V_T^{\pi_\star}(s), \qquad B_{\star,T} = \max_s V_T^\star(s)$$

$\implies$ If $T < T_\star$ then $\pi_T^\star(s) \neq \pi_\infty^\star(s)$

- No reason to talk about (ε, δ)-correctness but rather of (ε, δ, T)-correctness

Learning under Restricted Optimality

[Chen et al., 2022]

Theorem

For any $T < T_\star$, there exists an MDP with $c_{\min} = 0$ such that any (ε, δ, T)-correct algorithm requires

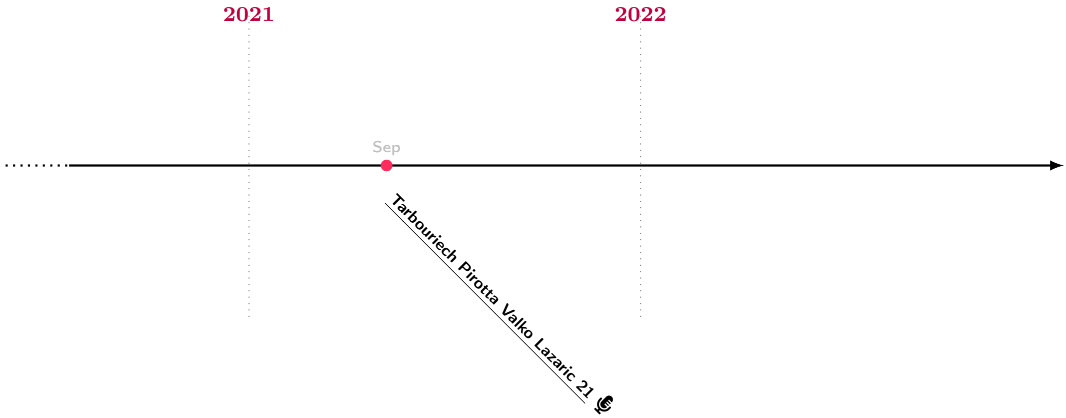
$$\widetilde{\Omega}\left(\frac{B_{\star,T}^2 TSA}{\varepsilon^2}\right)$$

samples.

- This shows a clear dependence on the range of the value function $B_{\star,T}$ and the hitting time T of the optimal policy
- Case $c_{\min} > 0$ is an open problem

Sample-Complexity Upper-Bounds

The start...



An Optimistic Algorithm

[Tarbouriech et al., 2021b]

Input: $c_{\min} > 0$, accuracy ε, precision δ, allocation function ϕ

Set $\widetilde{B} = 1/2$

while continue **do**

  $\widetilde{B} = 2\widetilde{B}$

  Get $\phi(\widetilde{B}, c_{\min})$ samples for each (s, a)

  Compute $\widetilde{v}$, $\widetilde{\pi}$ using an optimistic value iteration

  if $\|\widetilde{v}\|_\infty \le \widetilde{B}$ **then**

    continue = False

An Optimistic Algorithm: Regret Guarantees

Theorem ($c_{\min} > 0$)

For any accuracy $\varepsilon \in (0,1]$, confidence $\delta \in (0,1)$, and cost function c in $[c_{\min}, 1]$ with $c_{\min} > 0$, the algorithm in [Tarbouriech et al., 2021b] is (ε, δ)-correct with a sample complexity bounded as

$$\widetilde{O}\left(\frac{B_\star^3 \Gamma S A}{c_{\min} \varepsilon^2}\right)$$

- Not minimax optimal, off by a factor $\Gamma = \max_{s,a} \|P(\cdot|s,a)\|_0 \leq S$
- Require knowledge of $c_{\min} > 0$

And when $c_{\min} = 0$?

- Target a restricted optimality

$$\pi_{\star,\theta} = \operatorname*{arg\,min}_{\pi : \|T^\pi\|_\infty \leq \theta D} V^\pi$$

where $D = \max_s \min_\pi T^\pi(s)$ is the SSP diameter [Tarbouriech et al., 2020a]

- An algorithm is $(\varepsilon, \delta, \theta)$-correct with sample complexity n, if after n calls to the generative model it returns a policy π that verifies $\|V^\pi - V^{\pi_{\star,\theta}}\|_\infty \leq \varepsilon$ w.p. at least $1 - \delta$.

⚠ $(\varepsilon, \delta, \theta)$-correctness is different than $(\varepsilon, \delta, T = \theta D)$-correct since D is unknown

An Optimistic Algorithm for $c_{\min} = 0$

[Tarbouriech et al., 2021b]

Input: $\theta \in [1, \infty)$, accuracy ε, precision δ, allocation function ϕ
Estimate $\widetilde{D} \geq D$
Set cost perturbation $\nu = \dfrac{\varepsilon}{2\theta\widetilde{D}}$
Set $\widetilde{B} = 1/2$
while True **do**
    $\widetilde{B} = 2\widetilde{B}$
    Get $\phi(\widetilde{B}, c_{\min})$ samples for each (s, a)
    Compute $\widetilde{v}$, $\widetilde{\pi}$ using an optimistic value iteration with perturbed costs
    if $\|\widetilde{v}\|_\infty \leq \widetilde{B}$ **then**
        break

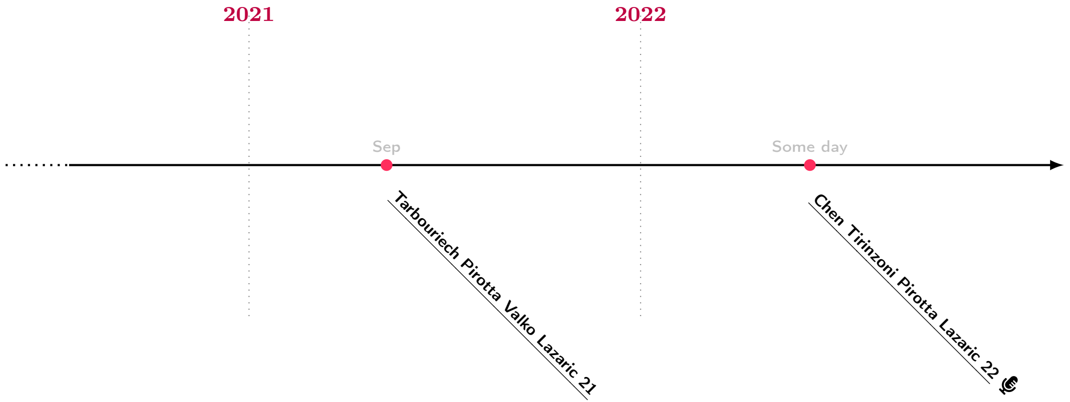
An Optimistic Algorithm for $c_{\min} = 0$: Regret Guarantees

Theorem ($c_{\min} = 0$)

For any accuracy $\varepsilon \in (0,1]$, $\theta \geq 1$, confidence $\delta \in (0,1)$, and cost function c in $[0,1]$, the algorithm in [Tarbouriech et al., 2021b] is $(\varepsilon, \delta, \theta)$-correct with a sample complexity bounded as

$$\widetilde{O}\left(\frac{\theta D B_\star^3 \Gamma S A}{c_{\min} \varepsilon^3}\right)$$

A Minimax Algorithm...



*work in preparation

A Minimax Algorithm

Input: $T \in [1, \infty]$, accuracy ε, precision δ, allocation functions ϕ, ϕ'
Set $\widetilde{B} = 2$
while True **do**
  Set $H = \min\{\widetilde{B}/c_{\min}, T\}$
  Get $\phi(\widetilde{B}, H)$ samples for each (s, a)
  Compute $\widetilde{v}$, $\widetilde{\pi}$ using finite-horizon reduction with horizon H and final cost $B\mathbb{I}\{s \neq g\}$
  if $\|\widetilde{v}\|_\infty \lesssim \widetilde{B}$ **then**
    break
  $\widetilde{B} = 2\widetilde{B}$
Recompute policy using ϕ' samples

Regret Guarantees

Theorem

For any accuracy $\varepsilon \in (0,1]$, $T \geq 1$, *confidence* $\delta \in (0,1)$, *and cost function* c *in* $[0,1]$, *the algorithm by [Chen, Tirinzoni, Pirotta, Lazaric] is* (ε, δ, T)*-correct with a sample complexity bounded as*

$$\widetilde{O}\left(\min\left\{ T, \frac{B_\star}{c_{\min}} \right\} \frac{B_{\star,T}^2 SA}{\varepsilon^2} \right)$$

- Minimax optimal for (ε, δ)-correctness with and without prior knowledge
- Minimax optimal for (ε, δ, T)-correctness when $c_{\min} = 0$

Sample-Complexity with Generative Model

Summary

Performance	Lower Bound	[Chen, Tirinzoni, Pirotta, Lazaric 22] finite-horizon reduction	[Tarbouriech, Pirotta, Valko, Lazaric, 21]* optimistic SSP planning
(ε, δ)	$\min\left\{\frac{B_\star}{c_{\min}}, T\right\} \frac{B_\star^2 SA}{\varepsilon^2}$	$\min\left\{\frac{B_\star}{c_{\min}}, T\right\} \frac{B_\star^2 SA}{\varepsilon^2}$	$\frac{B_\star^3 \Gamma SA}{c_{\min}\varepsilon^2}$
(ε, δ, T)	$\frac{TB_{\star,T}^2 SA}{\varepsilon^2}$ when $c_{\min} = 0$ unknown when $c_{\min} > 0$	$\min\left\{\frac{B_\star}{c_{\min}}, T\right\} \frac{B_{\star,T}^2 SA}{\varepsilon^2}$	$\frac{TB_{\star,T}^3 \Gamma SA}{\varepsilon^3}$

* as mentioned $(\varepsilon, \delta, \theta)$ and (ε, δ, T)-correctness are not exaclty equivalent. This is simplified comparison.

Best Policy Identification

How many interactions with the environment are sufficient to identify a near-optimal policy w.h.p.?

Input: accuracy ε, precision δ
while True **do**
    $s_t = s_1$
    while $s_t \neq g$ **do**
        $a_t = \pi_t(s_t)$
        Observe cost c_t and next state s_{t+1}
        Update policy π_{t+1}
        if *condition* **then**
            Stop
        $t = t + 1$

Definition (BPI)

An algorithm is (ε, δ)-correct with sample complexity n, if

1. it stops after n interactions
   $\mathbb{P}(\tau_n) = 1$
2. it returns w.h.p. a policy that is ε-accurate
   $\mathbb{P}(\|V^{\pi_n^\star} - V^\star\|_\infty \leq \varepsilon) \geq 1 - \delta$

Best Policy Identification: the generic case

[Chen, Tirinzoni, Pirotta, Lazaric, 22]

Theorem

There exists a SSP-MDP where any (ε, δ)*-correct requires* $\Omega\left(\frac{A^S}{\varepsilon}\right)$ *samples to perform BPI, even with the knowledge of* $B_\star$*,* $T_\star$ *and* $c_{\min}$*.*

Message and follow ups

- BPI is "impossible" in the general case
- However, under certain structural assumptions (e.g., reset action) it is possible to perform BPI

Discussion

- SSP is provably harder than other settings
- Trade off between performance ($B_\star$) and ($T_\star$) time is critical
- As well as properness plays a critical role

- Regret minimization is “simpler” than sample-complexity
  - Learnable in all the settings
  - No need to commit to a specific policy
  - Robust to imprecise prior knowledge

Discussion

Other SSP-related problems

- **Multi-Goal Exploration** [Tarbouriech et al., 2021a, 2022]
- **Autonomous Exploration** [Lim and Auer, 2012, Tarbouriech et al., 2020b, Cai et al., 2022]

Dimitri P Bertsekas and John N Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.

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.

Haoyuan Cai, Tengyu Ma, and Simon S. Du. Near-optimal algorithms for autonomous exploration and multi-goal stochastic shortest path. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 2434–2456. PMLR, 2022.

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: the adversarial cost and unknown transition case. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 1651–1660. PMLR, 2021.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. In *NeurIPS*, pages 10849–10861, 2021a.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. In *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 1180–1215. PMLR, 2021b.

Liyu Chen, Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. Reaching goals is hard: Settling the sample complexity of the stochastic shortest path. *CoRR*, abs/2210.04946, 2022.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. In *NeurIPS*, pages 28350–28361, 2021.

Matthieu Guillot and Gautier Stauffer. The stochastic shortest path problem: a polyhedral combinatorics perspective. *European Journal of Operational Research*, 285(1):148–158, 2020.

Mehdi Jafarnia-Jahromi, Liyu Chen, Rahul Jain, and Haipeng Luo. Online learning for stochastic shortest path model via posterior sampling. *CoRR*, abs/2106.05335, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is q-learning provably efficient? In *NeurIPS*, pages 4868–4878, 2018.

Shiau Hong Lim and Peter Auer. Autonomous exploration for navigating in mdps. In *COLT*, volume 23 of *JMLR Proceedings*, pages 40.1–40.24. JMLR.org, 2012.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *NeurIPS*, 2020.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 8210–8219. PMLR, 2020.

Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. No-regret exploration in goal-oriented reinforcement learning. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9428–9437. PMLR, 2020a.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Improved sample complexity for incremental autonomous exploration in mdps. In *NeurIPS*, 2020b.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. A provably efficient sample collection strategy for reinforcement learning. In *NeurIPS*, pages 7611–7624, 2021a.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Sample complexity bounds for stochastic shortest path with a generative model. In *ALT*, volume 132 of *Proceedings of Machine Learning Research*, pages 1157–1178. PMLR, 2021b.

Jean Tarbouriech, Runlong Zhou, Simon S. Du, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Stochastic shortest path: Minimax, parameter-free and towards horizon-free regret. In *NeurIPS*, pages 6843–6855, 2021c.

Jean Tarbouriech, Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Adaptive multi-goal exploration. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pages 7349–7383. PMLR, 2022.