Stochastic Shortest Path:

Minimax, Parameter-Free and Towards Horizon-Free Regret

Jean Tarbouriech (FAIR & Inria Scool)

June 29, 2021

RL Theory Virtual Seminar

Collaborators

Runlong Zhou
Tsinghua University

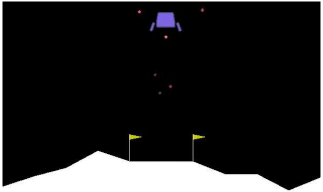
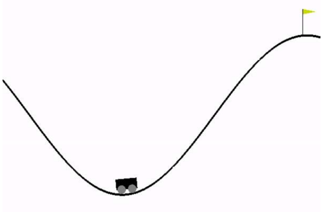
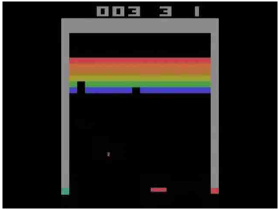
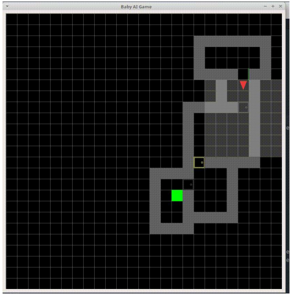
Simon S. Du
Univ. Washington

Matteo Pirotta
FAIR

Michal Valko
DeepMind

Alessandro Lazaric
FAIR

Goal-Oriented RL

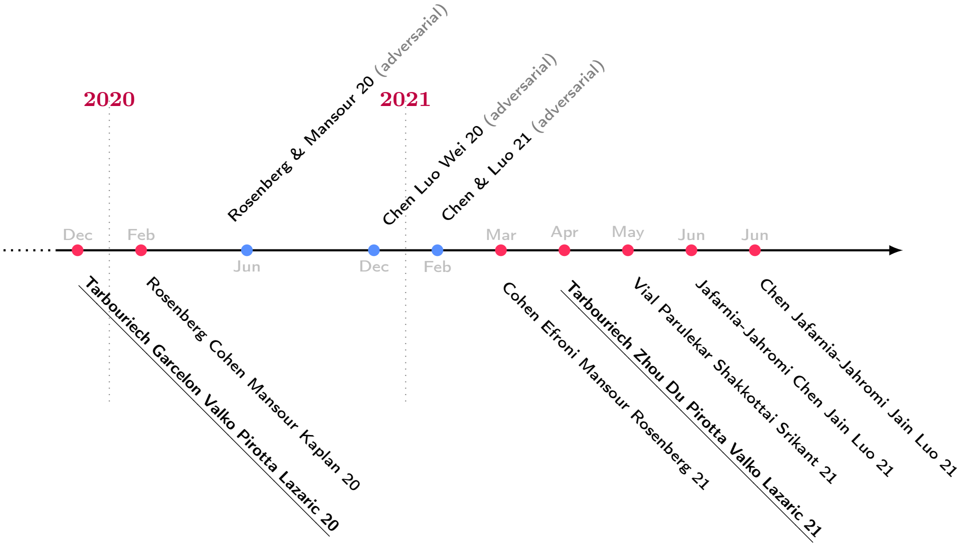





Many popular RL problems are *goal-oriented* tasks:
Minimize the cumulative *cost to reach the goal*

Also coined as the *stochastic shortest path* problem
[Bertsekas, 1995]

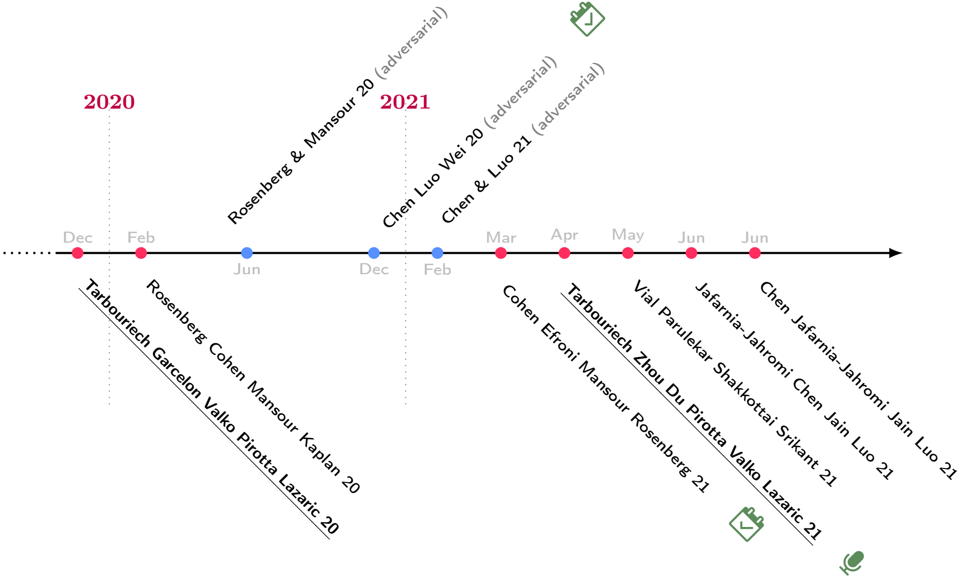
Online learning in SSP has only been studied recently

Regret Minimization in SSP



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

Regret Minimization in SSP



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

SSP-MDP

An SSP-MDP is a tuple $M = \langle\, \mathcal{S},\, \mathcal{A},\, P,\, c,\, s_{\mathsf{init}},\, g\, \rangle$

- State space $\mathcal{S} \cup \{g\}$
  - Goal state g
  - Initial state (distribution) $s_{\mathsf{init}} \in \mathcal{S}$
- Action space $\mathcal{A}$
- Transition probabilities $P(s'|s,a)$
- Cost function $c(s,a) \in [0,1]$

SSP-MDP

An SSP-MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_{\text{init}}, g \rangle$

- State space $\mathcal{S} \cup \{g\}$
  - Goal state g
  - Initial state (distribution) $s_{\text{init}} \in \mathcal{S}$
- Action space $\mathcal{A}$
- Transition probabilities $P(s'|s, a)$
- Cost function $c(s, a) \in [0, 1]$

▶ Specificity: the agent ends its interaction with the MDP once (if) it reaches the goal state g

SSP-MDP

An SSP-MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_{\text{init}}, g \rangle$

- State space $\mathcal{S} \cup \{g\}$
  - Goal state g
  - Initial state (distribution) $s_{\text{init}} \in \mathcal{S}$
- Action space $\mathcal{A}$
- Transition probabilities $P(s'|s, a)$
- Cost function $c(s, a) \in [0, 1]$

▶ Specificity: the agent ends its interaction with the MDP once (if) it reaches the goal state g

absorbing
$P(g|g, a) = 1$

zero-cost
$c(g, a) = 0$

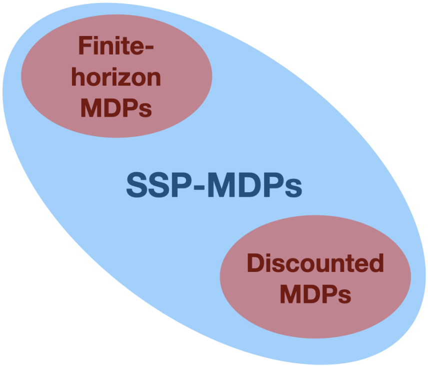
SSP-MDP

An SSP-MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, P, c, s_{\text{init}}, g \rangle$

- State space $\mathcal{S} \cup \{g\}$
  - Goal state g
  - Initial state (distribution) $s_{\text{init}} \in \mathcal{S}$
- Action space $\mathcal{A}$
- Transition probabilities $P(s'|s,a)$
- Cost function $c(s,a) \in [0,1]$

▶ Specificity: the agent ends its interaction with the MDP once (if) it reaches the goal state g

absorbing	zero-cost
$P(g\|g,a) = 1$	$c(g,a) = 0$



- Policy $\pi : \mathcal{S} \to \mathcal{A}$
- *Time-to-goal*:

$$T^\pi(s) := \mathbb{E}\left[\sum_{t=1}^{+\infty} \mathbb{I}\{s_t \neq g\} \,\middle|\, s_1 = s\right]$$

- *Value function* (a.k.a. cost-to-go):

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

- *Q-function*:

$$Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=1}^{+\infty} c(s_t, \pi(s_t))\mathbb{I}\{s_t \neq g\} \,\middle|\, s_1 = s, \pi(s_1) = a\right]$$

- Policy $\pi : \mathcal{S} \to \mathcal{A}$
- *Time-to-goal*:

$$T^\pi(s) := \mathbb{E}\left[\sum_{t=1}^{+\infty} \mathbb{I}\{s_t \neq g\} \,\middle|\, s_1 = s\right]$$

- *Value function* (a.k.a. cost-to-go):

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

- *Q-function*:

$$Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=1}^{+\infty} c(s_t, \pi(s_t))\mathbb{I}\{s_t \neq g\} \,\middle|\, s_1 = s, \pi(s_1) = a\right]$$

⚠ We may have $T^\pi = \infty,\ V^\pi = \infty,\ Q^\pi = \infty$ for many policies π

- A policy is *proper* if it reaches g with probability 1 starting from any state in $\mathcal{S}$
- Assumption: there exists at least one proper policy
- We denote by $\pi^\star$ the *optimal proper policy*, i.e.,

$$\pi^\star \in \operatorname*{arg\,min}_{\pi:\ T^\pi < \infty} V^\pi$$

- A policy is *proper* if it reaches g with probability 1 starting from any state in $\mathcal{S}$
- Assumption: there exists at least one proper policy
- We denote by $\pi^\star$ the *optimal proper policy*, i.e.,

$$\pi^\star \in \underset{\pi:\ T^\pi < \infty}{\arg\min}\, V^\pi$$

- Important quantities:

$$\boldsymbol{B}_\star := \max_{s \in \mathcal{S}} V^{\pi^\star}(s) \quad ; \qquad \boldsymbol{T}_\star := \max_{s \in \mathcal{S}} T^{\pi^\star}(s)$$

Online Learning in SSP

- P and c are **unknown** to the agent
- K episodes, an episode ends if (and only if) the goal is reached

Each episode:

- Agent starts at $s_1 = s_{\mathsf{init}}$
- While $s_t \neq g$:
  - Agent selects action $a_t \in \mathcal{A}$
  - Agent incurs cost $c_t \sim c(s_t, a_t)$
  - Environment draws next state $s_{t+1} \sim P(\cdot|s_t, a_t)$

Online Learning in SSP

- P and c are **unknown** to the agent
- K episodes, an episode ends if (and only if) the goal is reached
- Objective: Minimize the regret:

$$R_K := \sum_{k=1}^{K} \sum_{h=1}^{I^k} c_h^k - K V^{\pi^\star}(s_{\text{init}})$$

- If $\exists k, I^k = \infty$, then we define $R_K = \infty$

Each episode:

- Agent starts at $s_1 = s_{\text{init}}$
- While $s_t \neq g$:
  - Agent selects action $a_t \in \mathcal{A}$
  - Agent incurs cost $c_t \sim c(s_t, a_t)$
  - Environment draws next state $s_{t+1} \sim P(\cdot | s_t, a_t)$

Online Learning in SSP

- P and c are **unknown** to the agent
- K episodes, an episode ends if (and only if) the goal is reached
- Objective: Minimize the regret:

$$R_K := \sum_{k=1}^{K} \sum_{h=1}^{I^k} c_h^k - K\, V^{\pi^\star}(s_{\text{init}})$$

- If $\exists k, I^k = \infty$, then we define $R_K = \infty$

Each episode:

- Agent starts at $s_1 = s_{\text{init}}$
- While $s_t \neq g$:
  - Agent selects action $a_t \in \mathcal{A}$
  - Agent incurs cost $c_t \sim c(s_t, a_t)$
  - Environment draws next state $s_{t+1} \sim P(\cdot|s_t, a_t)$

Two differences with finite-horizon regret:

- We evaluate the *empirical* (not expected) performance of the agent
- We compete against the optimal *proper* policy $\pi^\star$

Three desired properties

for a learning algorithm in online SSP

Three desired properties

for a learning algorithm in online SSP

① First desired property: Minimax

Regret lower bound: $\Omega(B_\star \sqrt{SAK})$ [Rosenberg et al., 2020]

An algorithm for online SSP is (nearly) minimax optimal if its regret is bounded by $\widetilde{O}(B_\star \sqrt{SAK})$, up to logarithmic factors and lower-order terms.

Three desired properties

for a learning algorithm in online SSP

① First desired property: Minimax

Regret lower bound: $\Omega(B_\star \sqrt{SAK})$ [Rosenberg et al., 2020]

An algorithm for online SSP is (nearly) minimax optimal if its regret is bounded by $\widetilde{O}(B_\star \sqrt{SAK})$, up to logarithmic factors and lower-order terms.

② Second desired property: Parameter-free

SSP-specific quantities: $B_\star$ and $T_\star$

An algorithm for online SSP is parameter-free if it relies neither on $B_\star$ nor $T_\star$ prior knowledge.

Three desired properties

for a learning algorithm in online SSP

③ Third desired property: Horizon-free

- Core challenge in SSP: trade off between minimizing costs and quickly reaching the goal
- Harder when the instantaneous costs are small
- i.e., when there is a mismatch between $B_\star$ and $T_\star$

Three desired properties

for a learning algorithm in online SSP

③ Third desired property: Horizon-free

- Core challenge in SSP: trade off between minimizing costs and quickly reaching the goal
- Harder when the instantaneous costs are small
- i.e., when there is a mismatch between $B_\star$ and $T_\star$

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

Three desired properties

for a learning algorithm in online SSP

③ Third desired property: Horizon-free

- Core challenge in SSP: trade off between minimizing costs and quickly reaching the goal
- Harder when the instantaneous costs are small
- i.e., when there is a mismatch between $B_\star$ and $T_\star$

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

An algorithm for online SSP is (nearly) horizon-free if its regret depends only logarithmically on $T_\star$.

More on the horizon-free property

[Wang et al., 2020, Zhang et al., 2020, 2021]

An algorithm for online finite-horizon MDPs with total reward bounded by 1
is (nearly) horizon-free if its regret depends
only logarithmically on the horizon H.

number of time steps by which *any* policy terminates

More on the horizon-free property

[Wang et al., 2020, Zhang et al., 2020, 2021]

An algorithm for online finite-horizon MDPs with total reward bounded by 1
is (nearly) horizon-free if its regret depends
only logarithmically on the horizon H.

number of time steps by which *any* policy terminates

The extension to SSP:

An algorithm for online SSP
is (nearly) horizon-free if its regret depends
only logarithmically on $T_\star$.

expected number of time steps by which the *optimal* policy terminates

More on the horizon-free property

[Wang et al., 2020, Zhang et al., 2020, 2021]

An algorithm for online finite-horizon MDPs with total reward bounded by 1
is (nearly) horizon-free if its regret depends
only logarithmically on the horizon H.

number of time steps by which *any* policy terminates

The extension to SSP:

An algorithm for online SSP
is (nearly) horizon-free if its regret depends
only logarithmically on $T_\star$.

expected number of time steps by which the *optimal* policy terminates

Remarks:

- ⚠ We do **not** make any extra assumption on the SSP model to uncover horizon-free properties.
- Benefit of bounded total reward assumption: can model *sparse spiky reward* [Kakade, 2003, Jiang and Agarwal, 2018]: to the extreme, this scenario is captured by SSP.

Our Results

1. New algorithm for online SSP: `EB-SSP` (`Exploration Bonus for SSP`)
2. First algorithm to achieve the **minimax** regret rate of $\widetilde{O}(B_\star \sqrt{SAK})$ while simultaneously being **parameter-free**
3. First algorithm to achieve **horizon-free** regret in various cases:
   - positive costs,
   - general costs with no almost-sure zero-cost cycles,
   - general costs when an order-accurate estimate of $T_\star$ is available

Our Results w.r.t. Related Work

Algorithm	Approach	Regret	Minimax	Parameters	Horizon-Free
Lower Bound		$\Omega(B_\star\sqrt{SAK})$	-	-	-

Our Results w.r.t. Related Work

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

Our Results w.r.t. Related Work

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

Our Results w.r.t. Related Work

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		$\Omega(B_\star\sqrt{SAK})$	-	-	-

Our Results w.r.t. Related Work

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
This work	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		$\Omega(B_\star\sqrt{SAK})$	-	-	-

Our Results w.r.t. Related Work

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
This work	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		$\Omega(B_\star\sqrt{SAK})$	-	-	-

Our Results w.r.t. Related Work

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
This work	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		$\Omega(B_\star\sqrt{SAK})$	-	-	-

Our Results w.r.t. Related Work

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
This work	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		$\Omega(B_\star\sqrt{SAK})$	-	-	-

Additional Related Work

- SSP with adversarially changing costs [Rosenberg and Mansour, 2020, Chen et al., 2020, Chen and Luo, 2021]
- Sample complexity of SSP with a generative model [Tarbouriech et al., 2021]
- Multi-goal exploration [Lim and Auer, 2012, Tarbouriech et al., 2020b]

Additional Related Work

- SSP with adversarially changing costs [Rosenberg and Mansour, 2020, Chen et al., 2020, Chen and Luo, 2021]
- Sample complexity of SSP with a generative model [Tarbouriech et al., 2021]
- Multi-goal exploration [Lim and Auer, 2012, Tarbouriech et al., 2020b]

Later work:

- SSP with linear function approximation [Vial et al., 2021]
- SSP via posterior sampling [Jafarnia-Jahromi et al., 2021]
- Template for regret minimization in SSP [Chen et al., 2021]
  - Model-based instantiation: matches our regret bound
  - Model-free instantiation: achieves minimax rate under positive costs
  - One-step planning (i.e., sparse computational updates)

EB-SSP Algorithm

Exploration Bonus for SSP

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 Algorithm

- Initialize $Q(s,a)=0$ for all (s,a)
- Sequentially select action $a_t \in \arg\min_{a\in\mathcal{A}} Q(s_t,a)$
- If trigger condition:
  - Compute new $Q(s,a)$ values for all (s,a)

EB-SSP Algorithm

- Initialize $Q(s,a) = 0$ for all (s,a)
- Sequentially select action $a_t \in \arg\min_{a \in \mathcal{A}} Q(s_t, a)$
- If trigger condition:
  - Compute new $Q(s,a)$ values for all (s,a)

▶ Standard "doubling condition": when the visit to a state-action pair doubles [Jaksch et al., 2010, Zhang et al., 2020]

EB-SSP Algorithm

- Initialize $Q(s,a) = 0$ for all (s,a)
- Sequentially select action $a_t \in \arg\min_{a \in \mathcal{A}} Q(s_t, a)$
- If trigger condition:
  - Compute new $Q(s,a)$ values for all (s,a)

▶ Standard "doubling condition": when the visit to a state-action pair doubles [Jaksch et al., 2010, Zhang et al., 2020]

▶ New procedure called `VISGO` — `Value Iteration with Slight Goal Optimism`

VISGO planning procedure

- Input: $\epsilon_{\mathsf{VI}} > 0$ precision level
- Start with optimistic values $V^{(0)} = 0$
- While $\|V^{(i+1)} - V^{(i)}\|_\infty > \epsilon_{\mathsf{VI}}$:
  - Iteratively compute $V^{(i+1)} = \widetilde{\mathcal{L}} V^{(i)}$ for an operator $\widetilde{\mathcal{L}}$
- Output: the values $V^{(i+1)}$ (and Q-values $Q^{(i+1)}$)

VISGO planning procedure

- Input: $\epsilon_{\mathsf{VI}} > 0$ precision level
- Start with optimistic values $V^{(0)} = 0$
- While $\|V^{(i+1)} - V^{(i)}\|_\infty > \epsilon_{\mathsf{VI}}$:
  - Iteratively compute $V^{(i+1)} = \widetilde{\mathcal{L}} V^{(i)}$ for an operator $\widetilde{\mathcal{L}}$
- Output: the values $V^{(i+1)}$ (and Q-values $Q^{(i+1)}$)

How to define $\widetilde{\mathcal{L}}$?

How to define $\widetilde{\mathcal{L}}$ in VISGO?

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 $\widetilde{P}$:

slight goal skewing

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

How to define $\widetilde{\mathcal{L}}$ in VISGO?

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 $\widetilde{P}$:

slight goal skewing

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

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

How to define $\widetilde{\mathcal{L}}$ in VISGO?

③ Bonus function b:

$$b(V,s,a) := \max\left\{c_1\sqrt{\frac{\mathbb{V}(\widetilde{P}_{s,a},V)\iota_{s,a}}{n^+(s,a)}},\ c_2\frac{B\iota_{s,a}}{n^+(s,a)}\right\} + c_3\sqrt{\frac{\widehat{c}(s,a)\iota_{s,a}}{n^+(s,a)}} + c_4\frac{B\sqrt{S\iota_{s,a}}}{n^+(s,a)},$$

given proxy $B > 0$, specific constants $c_1, c_2, c_3, c_4 > 0$ and logarithmic term $\iota_{s,a}$

How to define $\widetilde{\mathcal{L}}$ in VISGO?

③ Bonus function b:

$$b(V, s, a) := \max\left\{c_1\sqrt{\frac{\mathbb{V}(\widetilde{P}_{s,a}, V)\iota_{s,a}}{n^+(s,a)}},\ c_2\frac{B\iota_{s,a}}{n^+(s,a)}\right\} + c_3\sqrt{\frac{\widehat{c}(s,a)\iota_{s,a}}{n^+(s,a)}} + c_4\frac{B\sqrt{S\iota_{s,a}}}{n^+(s,a)},$$

given proxy $B > 0$, specific constants $c_1, c_2, c_3, c_4 > 0$ and logarithmic term $\iota_{s,a}$

④ Operator $\widetilde{\mathcal{L}}$:

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

$\boxed{\text{2 sources of optimism}}$

Theorem (Intermediate regret bound)

Assume that

1. $B \geq B_\star \geq 1$,
2. *the value function of any improper policy has at least one unbounded component*
   - *the optimal policy is proper and satisfies the Bellman optimality equations [Bertsekas and Tsitsiklis, 1991]*

Then w.p. $1 - \delta$,

$$R_K = O\Big(B_\star\sqrt{SAK}\log\left(\frac{B_\star SAT_K}{\delta}\right) + BS^2A\log^2\left(\frac{B_\star SAT_K}{\delta}\right)\Big),$$

with T_K *the accumulated time over the* K *episodes.*

Proof part 1: `VISGO` properties

Lemma

As long as $B \geq B_\star$*:*

(1) ***Optimism:*** $Q^{(i)}(s,a) \leq Q^{\pi^\star}(s,a)$, *for any iteration* $i \geq 0$

(2) ***Finite-time near-convergence:*** `VISGO` *terminates within a finite (polynomially bounded) number of iteration steps*

Proof part 1: `VISGO` properties

Lemma

As long as $B \geq B_\star$:

(1) ***Optimism:*** $Q^{(i)}(s,a) \leq Q^{\pi^\star}(s,a)$, *for any iteration* $i \geq 0$

(2) ***Finite-time near-convergence:*** `VISGO` *terminates within a finite (polynomially bounded) number of iteration steps*

Proof idea.

(1) We derive a *monotonicity* property for $\widetilde{\mathcal{L}}$

Achieved by carefully tuning the constants c_1, c_2, c_3, c_4 in the bonus

Similar argument to analysis of MVP [Zhang et al., 2020]

$\square$

Proof part 1: `VISGO` properties

Lemma

As long as $B \geq B_\star$:

(1) ***Optimism:*** $Q^{(i)}(s,a) \leq Q^{\pi^\star}(s,a)$, *for any iteration* $i \geq 0$

(2) ***Finite-time near-convergence:*** `VISGO` *terminates within a finite (polynomially bounded) number of iteration steps*

Proof idea.

(1) We derive a *monotonicity* property for $\widetilde{\mathcal{L}}$
Achieved by carefully tuning the constants c_1, c_2, c_3, c_4 in the bonus
Similar argument to analysis of MVP [Zhang et al., 2020]

(2) We derive a *contraction* property for $\widetilde{\mathcal{L}}$
Contraction modulus $\rho \leq 1 - \nu^2 < 1$, where $\nu := \min_{s,a} \widetilde{P}_{s,a,g} > 0$

SSP-specific requirement

□

Proof part 2: Regret Decomposition

➔ *First, a bit of notation:*

- Recall that for now we consider $B \geq B_\star \geq 1$
  - The two `VISGO` properties (optimism and convergence) hold
- Let V_t be the `VISGO` value at time t
- Define the normalized value $\overline{V}_t := V_t / B_\star \in [0, 1]$
- Let C_K (resp. T_K) be the **cumulative cost** (resp. **time**) over the K episodes

➔ *Up next, high-level idea in 1 slide:*

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the
Bellman error
(V_t approximates
fixed point of $\widetilde{\mathcal{L}}$)

$$R_K = C_K - KV^{\pi^\star}(s_{\mathsf{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t, a_t)}}$$

bonus expression

$$R_K = C_K - KV^{\pi^\star}(s_{\mathsf{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$R_K = C_K - KV^{\pi^\star}(s_{\mathsf{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$\lesssim \sqrt{SA} \sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)}$$

pigeonhole principle

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$\lesssim \sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)} \lesssim B_\star\sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, \overline{V}_t)}$$

pigeonhole principle, value normalization

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$\lesssim \sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)} \lesssim B_\star\sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, \overline{V}_t)}$$

pigeonhole principle, value normalization

$$\lesssim B_\star\sqrt{SA}\left(\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^2) + \left(\frac{C_K}{B_\star}\right)^2\right)^{1/4}$$

law of total variance...

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$\lesssim \sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)} \lesssim B_\star\sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, \overline{V}_t)}$$

pigeonhole principle, value normalization

$$\lesssim B_\star\sqrt{SA}\left(\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^2) + \left(\frac{C_K}{B_\star}\right)^2\right)^{1/4}$$

law of total variance...

$$\lesssim \ldots \lesssim B_\star\sqrt{SA}\left(\underbrace{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^{2^d})}_{\leq T_K \ (\forall d)} + \left(\frac{C_K}{B_\star}\right)^{2^{d-1}}\right)^{2^{-d}}$$

...recursively...

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t, a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t, a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t, a_t)}}$$

bonus expression, $\widetilde{P} / \widehat{P} / P$ relation

$$\lesssim \sqrt{SA} \sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)} \lesssim B_\star \sqrt{SA} \sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, \overline{V}_t)}$$

pigeonhole principle, value normalization

$$\lesssim B_\star \sqrt{SA} \left(\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^2) + \left(\frac{C_K}{B_\star} \right)^2 \right)^{1/4}$$

law of total variance...

$$\lesssim \ldots \lesssim B_\star \sqrt{SA} \left(\underbrace{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^{2^d})}_{\leq T_K \ (\forall d)} + \left(\frac{C_K}{B_\star} \right)^{2^{d-1}} \right)^{2^{-d}}$$

...recursively...

$$\lesssim \sqrt{B_\star SA\, C_K} \log T_K$$

... expand up to higher order $d = \log T_K$

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$\lesssim \sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)} \lesssim B_\star\sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, \overline{V}_t)}$$

pigeonhole principle, value normalization

$$\lesssim B_\star\sqrt{SA}\left(\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^2) + \left(\frac{C_K}{B_\star}\right)^2\right)^{1/4}$$

law of total variance...

$$\lesssim \ldots \lesssim B_\star\sqrt{SA}\left(\underbrace{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^{2^d})}_{\leq T_K \ (\forall d)} + \left(\frac{C_K}{B_\star}\right)^{2^{d-1}}\right)^{2^{-d}}$$

...recursively...

$$\lesssim \sqrt{B_\star SA\, C_K}\log T_K$$

... expand up to higher order $d = \log T_K$

$\implies$ Solve a quadratic inequality in C_K and plug it back into the regret

$$\implies R_K \lesssim B_\star\sqrt{SAK}\log T_K$$

$$R_K = C_K - KV^{\pi^\star}(s_{\text{init}}) \lesssim \sum_{t=1}^{T_K} b_t(s_t, a_t) + \text{additional terms}$$

bounding the Bellman error (V_t approximates fixed point of $\widetilde{\mathcal{L}}$)

$$\lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widetilde{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(\widehat{P}_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}} \lesssim \sum_{t=1}^{T_K} \sqrt{\frac{\mathbb{V}(P_{s_t,a_t}, V_t)}{n_t(s_t,a_t)}}$$

bonus expression, $\widetilde{P}/\widehat{P}/P$ relation

$$\lesssim \sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, V_t)} \lesssim B_\star\sqrt{SA}\sqrt{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, \overline{V}_t)}$$

pigeonhole principle, value normalization

$$\lesssim B_\star\sqrt{SA}\left(\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^2) + \left(\frac{C_K}{B_\star}\right)^2\right)^{1/4}$$

law of total variance...

$$\lesssim \ldots \lesssim B_\star\sqrt{SA}\left(\underbrace{\sum_{t=1}^{T_K} \mathbb{V}(P_{s_t,a_t}, (\overline{V}_t)^{2^d})}_{\leq T_K \ (\forall d)} + \left(\frac{C_K}{B_\star}\right)^{2^{d-1}}\right)^{2^{-d}}$$

...recursively...

$$\lesssim \sqrt{B_\star SA\, C_K} \log T_K$$

... expand up to higher order $d = \log T_K$

$\Longrightarrow$ Solve a quadratic inequality in C_K and plug it back into the regret

$\Longrightarrow$ $R_K \lesssim B_\star\sqrt{SAK}\log T_K$

	Finite-horizon [Zhang et al., 2020]	SSP [this work]
Terms appearing in recursions	$\sum_{t=1}^{HK} r_t$	$\sum_{t=1}^{T_K} c_t = C_K$
How they are handled	*bounded* by K by *assumption* (total reward ≤ 1)	*eliminated* by quad. ineq. in C_K thanks to regret def.

Theorem (Intermediate regret bound)

Assume that

1. $B \geq B_\star \geq 1$,
2. *the value function of any improper policy has at least one unbounded component.*

Then w.p. $1-\delta$,

$$R_K = O\Big(B_\star\sqrt{SAK}\log\Big(\frac{B_\star SAT_K}{\delta}\Big) + BS^2A\log^2\Big(\frac{B_\star SAT_K}{\delta}\Big)\Big).$$

Relies on condition 2 and depends on T_K:

- Circumvented with *cost perturbation*: $c_\eta(s,a) \leftarrow \max\{c(s,a),\eta\}$
- If costs are lower bounded by $\eta > 0$, then condition 2 holds and $T_K \leq \dfrac{C_K}{\eta}$
- Regret $\lesssim$ "Regret in cost-perturbed MDP" $+\eta T_\star K$
- There remains to tune the cost perturbation:

$$\eta \leftarrow \begin{cases} \frac{1}{\text{poly}(K)} & \\ \frac{1}{X\cdot\text{poly}(K)} & \text{if loose prior knowledge } X \approx T_\star \text{ is available} \end{cases}$$

Theorem (Intermediate regret bound)

Assume that

1. $B \geq B_\star \geq 1$,
2. *the value function of any improper policy has at least one unbounded component.*

Then w.p. $1-\delta$,

$$R_K = O\Big(B_\star\sqrt{SAK}\log\left(\frac{B_\star SAT_K}{\delta}\right) + BS^2A\log^2\left(\frac{B_\star SAT_K}{\delta}\right)\Big).$$

Relies on condition 2 and depends on T_K:

- Circumvented with *cost perturbation*: $c_\eta(s,a) \leftarrow \max\{c(s,a),\eta\}$
- If costs are lower bounded by $\eta > 0$, then condition 2 holds and $T_K \leq \frac{C_K}{\eta}$
- Regret $\lesssim$ "Regret in cost-perturbed MDP" $+\eta T_\star K$
- There remains to tune the cost perturbation:

$$\eta \leftarrow \begin{cases} \frac{1}{\text{poly}(K)} \\ \frac{1}{X\cdot\text{poly}(K)} & \text{if loose prior knowledge } X \approx T_\star \text{ is available} \end{cases}$$

Relies on B being properly tuned: ▶ *Parameter-free scheme* to adaptively tune B

Unknown $B_\star$

▶ Unknown range of the optimal value function

- Exploration bonus requires a bound on $B_\star := \|V^{\pi^\star}\|_\infty$

Setting[1]	Finite-horizon	Finite-horizon w/ bounded total reward	Discounted	**SSP**
Bound on $\|V^{\pi^\star}\|_\infty$	H	1	$1/(1-\gamma)$	**?**

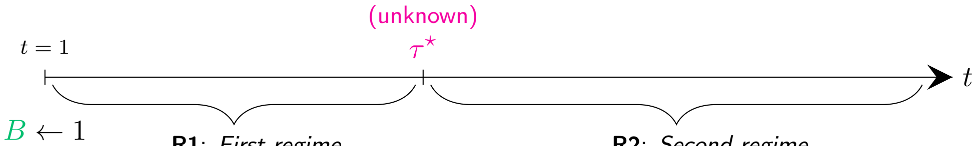
- If $B < B_\star$, optimism and convergence of `VISGO` may not hold

- It may be impossible to estimate $B_\star$ online (some states may be unreachable)

[1] In average reward: open question of [Qian et al., 2019]: *Is it possible to design an exploration bonus strategy without prior knowledge of the "optimal range"?*

Parameter-Free EB-SSP

Parameter-Free EB-SSP



R1: *First regime*

- Underestimate $B < B_\star$

R2: *Second regime*

- Overestimate $B \geq B_\star$

Parameter-Free EB-SSP

(unknown)

$t = 1$ $\quad\quad$ $\tau^\star$ $\quad\quad$ t

$B \leftarrow 1$

R1: *First regime*

- Underestimate $B < B_\star$
  - *How to bound regret?*

Regret(**R1**) $\leq$ Costs(**R1**) $\leq$?

R2: *Second regime*

- Overestimate $B \geq B_\star$
  - *How to bound regret?*

Regret(**R2**) $\lesssim B_\star \sqrt{SAK} + BS^2A$

Parameter-Free EB-SSP

(unknown)

$t = 1$ $\quad$ $\tau^\star$ $\quad$ t

$B \leftarrow 1$

R1: *First regime*

- Underestimate $B < B_\star$
  - *How to bound regret?*

Regret(**R1**) $\leq$ Costs(**R1**) $\leq$?

R2: *Second regime*

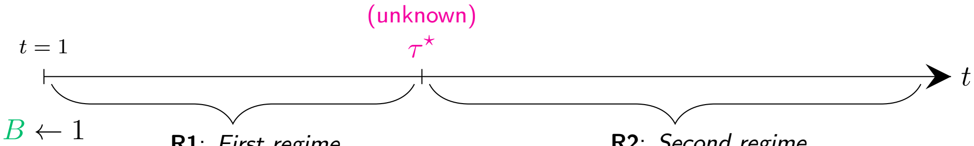
- Overestimate $B \geq B_\star$
  - *How to bound regret?*

Regret(**R2**) $\lesssim B_\star \sqrt{SAK} + BS^2 A$

- *Inter-episode increment of B:*

- *Intra-episode increments of B:*

Parameter-Free EB-SSP



R1: *First regime*
- Underestimate $B < B_\star$
  - *How to bound regret?*

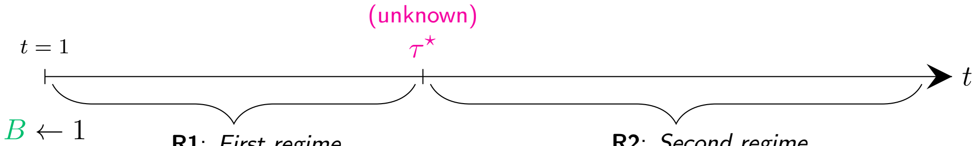
Regret(**R1**) $\leq$ Costs(**R1**) $\leq$?

R2: *Second regime*
- Overestimate $B \geq B_\star$
- *How to bound regret?*

Regret(**R2**) $\lesssim B_\star \sqrt{SAK} + BS^2 A$

- *Inter-episode increment of B:*
  Whenever a new episode k begins, set $B \leftarrow \max\{B, \sqrt{k}/(S^{3/2}A^{1/2})\}$
  - For large enough k, **R2** is reached. But risk of getting stuck in an episode in **R1**...
- *Intra-episode increments of B:*

Parameter-Free EB-SSP



R1: *First regime*

- Underestimate $B < B_\star$
  - *How to bound regret?*

Regret(**R1**) $\leq$ Costs(**R1**) $\leq$?

R2: *Second regime*

- Overestimate $B \geq B_\star$
- *How to bound regret?*

Regret(**R2**) $\lesssim B_\star\sqrt{SAK} + BS^2A$

- *Inter-episode increment of B:*

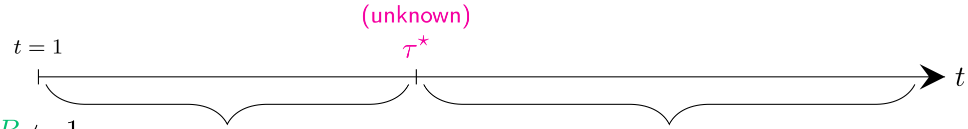
  Whenever a new episode k begins, set $B \leftarrow \max\{B, \sqrt{k}/(S^{3/2}A^{1/2})\}$

  - For large enough k, **R2** is reached. But risk of getting stuck in an episode in **R1**...

- *Intra-episode increments of B:*

  i) *Track* range of each `VISGO` iterate: if $\|V^{(i)}\|_\infty > B$, then double $B \leftarrow 2B$

Parameter-Free EB-SSP



$B \leftarrow 1$

R1: *First regime*
- Underestimate $B < B_\star$
- *How to bound regret?*

Regret(**R1**) $\leq$ Costs(**R1**) $\leq$?

R2: *Second regime*
- Overestimate $B \geq B_\star$
- *How to bound regret?*

Regret(**R2**) $\lesssim B_\star\sqrt{SAK} + BS^2A$

- *Inter-episode increment of B:*

  Whenever a new episode k begins, set $B \leftarrow \max\{B, \sqrt{k}/(S^{3/2}A^{1/2})\}$
  - For large enough k, **R2** is reached. But risk of getting stuck in an episode in **R1**...

- *Intra-episode increments of B:*

  i) *Track* range of each `VISGO` iterate: if $\|V^{(i)}\|_\infty > B$, then double $B \leftarrow 2B$

  ii) *Track* cumulative cost C: if $C \geq \phi$, then double $B \leftarrow 2B$
  - Cost threshold $\phi \approx kB + B\sqrt{SAk} + BS^2A$
  - Violated at most $\alpha = O(\log B_\star)$ times in **R1**

Parameter-Free EB-SSP

(unknown)

$t = 1$ $\quad \tau^\star$ $\quad t$

$B \leftarrow 1$

R1: *First regime*

- Underestimate $B < B_\star$
  - *How to bound regret?*

Regret(**R1**) $\leq$ Costs(**R1**) $\lesssim \alpha \cdot \phi$

R2: *Second regime*

- Overestimate $B \geq B_\star$
- *How to bound regret?*

Regret(**R2**) $\lesssim B_\star\sqrt{SAK} + BS^2A$

- *Inter-episode increment of B:*

  Whenever a new episode k begins, set $B \leftarrow \max\{B, \sqrt{k}/(S^{3/2}A^{1/2})\}$

  - For large enough k, **R2** is reached. But risk of getting stuck in an episode in **R1**...

- *Intra-episode increments of B:*

  i) *Track* range of each `VISGO` iterate: if $\|V^{(i)}\|_\infty > B$, then double $B \leftarrow 2B$

  ii) *Track* cumulative cost C: if $C \geq \phi$, then double $B \leftarrow 2B$

  - Cost threshold $\phi \approx kB + B\sqrt{SAk} + BS^2A$
  - Violated at most $\alpha = O(\log B_\star)$ times in **R1**

Regret of Parameter-Free EB-SSP

Theorem

The regret of parameter-free EB-SSP *can be bounded w.p.* $1-\delta$ *by*

$$R_K = O\left(R_K^\star \log\left(\frac{B_\star SAT_K}{\delta}\right) + B_\star^3 S^3 A \log^3\left(\frac{B_\star SAT_K}{\delta}\right)\right),$$

where $R_K^\star$ *bounds the regret of* EB-SSP *in the case of known* $B_\star$*.*

Regret of Parameter-Free EB-SSP

Theorem

The regret of parameter-free `EB-SSP` *can be bounded w.p.* $1-\delta$ *by*

$$R_K = O\left(R_K^\star \log\left(\frac{B_\star SAT_K}{\delta}\right) + B_\star^3 S^3 A \log^3\left(\frac{B_\star SAT_K}{\delta}\right)\right),$$

where $R_K^\star$ *bounds the regret of* `EB-SSP` *in the case of known* $B_\star$*.*

- *We can circumvent the knowledge of* $B_\star$ *up to logarithmic and lower-order terms.*
- Only algorithmic change to `EB-SSP`:
  - *dual tracking* of the cumulative costs and `VISGO` iterates,
  - *careful increment* of the proxy B in the bonus.

Conclusion and Outlook

Summary

- `EB-SSP` is the first algorithm in online SSP to
  1) achieve the **minimax** regret rate of $\widetilde{O}(B_\star\sqrt{SAK})$ while simultaneously being **parameter-free**
  2) achieve **horizon-free** regret in various cases (e.g., positive costs, or general costs with an order-accurate estimate of $T_\star$ available)

Future directions

- Open question: simultaneously minimax, parameter-free and horizon-free?
- Tight sample complexity bounds for SSP
- SSP beyond tabular & model-based [Vial et al., 2021, Chen et al., 2021]

Beyond the theory?

- On the question of when to reset in goal-oriented deep RL

Details are in our paper:

Stochastic Shortest Path: Minimax, Parameter-Free and Towards Horizon-Free Regret

https://arxiv.org/abs/2104.11186

Jean Tarbouriech*, Runlong Zhou*, Simon S. Du, Matteo Pirotta, Michal Valko, Alessandro Lazaric

Thank you

Dimitri Bertsekas. *Dynamic programming and optimal control*, volume 2. 1995.

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

Liyu Chen and Haipeng Luo. Finding the stochastic shortest path with low regret: The adversarial cost and unknown transition case. *arXiv preprint arXiv:2102.05284*, 2021.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Minimax regret for stochastic shortest path with adversarial costs and known transition. *arXiv preprint arXiv:2012.04053*, 2020.

Liyu Chen, Mehdi Jafarnia-Jahromi, Rahul Jain, and Haipeng Luo. Implicit finite-horizon approximation and efficient optimal algorithms for stochastic shortest path. *arXiv preprint arXiv:2106.08377*, 2021.

Alon Cohen, Yonathan Efroni, Yishay Mansour, and Aviv Rosenberg. Minimax regret for stochastic shortest path. *arXiv preprint arXiv:2103.13056*, 2021.

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

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

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398. PMLR, 2018.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Shiau Hong Lim and Peter Auer. Autonomous exploration for navigating in mdps. In *Conference on Learning Theory*, pages 40–1. JMLR Workshop and Conference Proceedings, 2012.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward mdps. In *Advances in Neural Information Processing Systems*, pages 4891–4900, 2019.

Aviv Rosenberg and Yishay Mansour. Stochastic shortest path with adversarially changing costs. *arXiv preprint arXiv:2006.11561*, 2020.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, 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 *International Conference on Machine Learning*, pages 9428–9437. PMLR, 2020a.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Improved sample complexity for incremental autonomous exploration in mdps. In *Advances in Neural Information Processing Systems*, volume 33, pages 11273–11284, 2020b.

Jean Tarbouriech, Matteo Pirotta, Michal Valko, and Alessandro Lazaric. Sample complexity bounds for stochastic shortest path with a generative model. In *Algorithmic Learning Theory*, pages 1157–1178. PMLR, 2021.

Daniel Vial, Advait Parulekar, Sanjay Shakkottai, and R Srikant. Regret bounds for stochastic shortest path problems with linear function approximation. *arXiv preprint arXiv:2105.01593*, 2021.

Ruosong Wang, Simon S. Du, Lin F. Yang, and Sham M. Kakade. Is long horizon RL more difficult than short horizon RL? In *Advances in Neural Information Processing Systems*, 2020.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021.

Extra slides

SSP Model with Positive Costs

Assumption

Costs are lower bounded by an unknown constant $c_{\min} > 0$.

Corollary

Running `EB-SSP` *with* $B = B_\star \geq 1$ *and* $\eta = 0$ *gives w.p.* $1 - \delta$

$$R_K = O\left(B_\star\sqrt{SAK}\log\left(\frac{KB_\star SA}{c_{\min}\delta}\right) + B_\star S^2 A \log^2\left(\frac{KB_\star SA}{c_{\min}\delta}\right)\right).$$

- *(Nearly) minimax and horizon-free*

SSP Model with General Costs

☐ $T_\star$ **Unknown**

Corollary

Running EB-SSP *with* $B = B_\star \geq 1$ *and* $\eta = K^{-n}$ *for* ***any*** *constant* $n > 1$ *gives w.p.* $1 - \delta$

$$R_K = O\Big(nB_\star\sqrt{SAK}L + \frac{T_\star}{K^{n-1}} + \frac{nT_\star\sqrt{SAL}}{K^{n-1/2}} + n^2 B_\star S^2 AL^2\Big), \qquad L := \log KT_\star SA\delta^{-1}.$$

- *(Nearly) minimax and "horizon-vanishing"*

☐ **Order-Accurate Estimate of $T_\star$ Available**

Assumption

Prior knowledge: a quantity X *s.t.* $T_\star/\upsilon \leq X \leq \lambda T_\star^\zeta$ *for some unknown constants* $\upsilon, \lambda, \zeta \geq 1$.

Corollary

Running EB-SSP *with* $B = B_\star \geq 1$ *and* $\eta = (XK)^{-1}$ *gives w.p.* $1 - \delta$

$$R_K = O\left(B_\star\sqrt{SAK}\log\left(\frac{KT_\star SA}{\delta}\right) + B_\star S^2 A\log^2\left(\frac{KT_\star SA}{\delta}\right)\right).$$

- *(Nearly) minimax and horizon-free*

Case $B_\star > 0$

Theorem (Intermediate regret bound)

Assume that

1. $B \geq B_\star$,
2. *the value function of any improper policy has at least one unbounded component*

Then w.p. $1-\delta$,

$$R_K = O\left(\sqrt{(B_\star^2 + B_\star)SAK} \log\left(\frac{\max\{B_\star, 1\}SAT_K}{\delta}\right) + BS^2A\log^2\left(\frac{\max\{B_\star, 1\}SAT_K}{\delta}\right)\right),$$

with T_K *the accumulated time over the* K *episodes.*