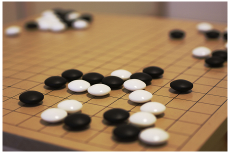
Improved No-Regret Algorithms for Stochastic Shortest Path with Linear MDP

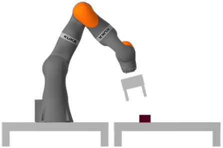
Liyu Chen, Rahul Jain, Haipeng Luo
University of Southern California

July 11, 2022

Motivation

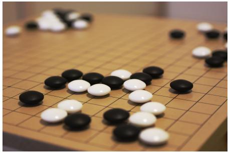
Many real-world applications can be modelled by goal-oriented reinforcement learning.

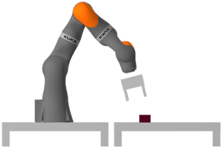




Motivation

Many real-world applications can be modelled by goal-oriented reinforcement learning.



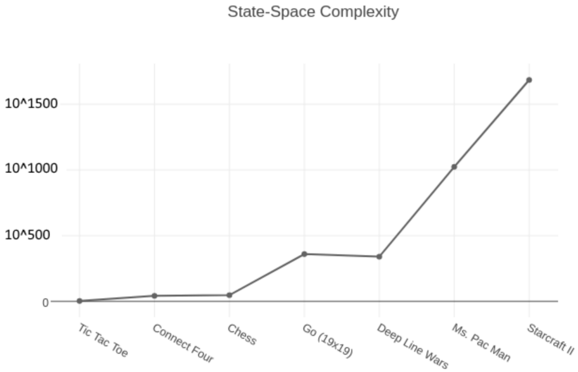


Goal-oriented reinforcement learning can be formulated as Stochastic Shortest Path (SSP) problem.

- Episodic MDP with a goal state.
- The objective is to reach the goal state with minimum cost.

Motivation

In real-world applications, the state-space is often prohibitively large.



Function approximation is necessary in practice.

Image source: https://arxiv.org/pdf/1712.06180.pdf

Our Contributions

We further extend our understanding of SSP with linear function approximation.

	Regret	Remark
(Vial et al., 2021)	$\sqrt{d^3 B_\star^3 K / c_{\min}}$	Inefficient
	$K^{5/6}$ (ignoring other params.)	Efficient
Ours		

d: feature dimension, $c_{\min}$: minimum cost, $\text{gap}_{\min}$: minimum sub-optimality gap
$B_\star$: maximum expected cost of optimal policy over all states
$T_\star$: maximum hitting time of optimal policy over all states, K: #episodes

Our Contributions

We further extend our understanding of SSP with linear function approximation.

	Regret	Remark
(Vial et al., 2021)	$\sqrt{d^3 B_\star^3 K / c_{\min}}$	Inefficient
	$K^{5/6}$ (ignoring other params.)	Efficient
Ours	$\sqrt{d^3 B_\star^2 T_\star K}$	Efficient

d: feature dimension, $c_{\min}$: minimum cost, $\text{gap}_{\min}$: minimum sub-optimality gap
$B_\star$: maximum expected cost of optimal policy over all states
$T_\star$: maximum hitting time of optimal policy over all states, K: #episodes

Our Contributions

We further extend our understanding of SSP with linear function approximation.

	Regret	Remark
(Vial et al., 2021)	$\sqrt{d^3 B_\star^3 K / c_{\min}}$	Inefficient
	$K^{5/6}$ (ignoring other params.)	Efficient
Ours	$\sqrt{d^3 B_\star^2 T_\star K}$	Efficient
	$\frac{d^3 B_\star^4}{c_{\min}^2 \text{gap}_{\min}} \ln^5 \frac{dB_\star K}{c_{\min}}$	Efficient, gap-dependent bound

d: feature dimension, $c_{\min}$: minimum cost, $\text{gap}_{\min}$: minimum sub-optimality gap
$B_\star$: maximum expected cost of optimal policy over all states
$T_\star$: maximum hitting time of optimal policy over all states, K: #episodes

Our Contributions

We further extend our understanding of SSP with linear function approximation.

	Regret	Remark
(Vial et al., 2021)	$\sqrt{d^3 B_\star^3 K / c_{\min}}$	Inefficient
	$K^{5/6}$ (ignoring other params.)	Efficient
Ours	$\sqrt{d^3 B_\star^2 T_\star K}$	Efficient
	$\frac{d^3 B_\star^4}{c_{\min}^2 \mathrm{gap}_{\min}} \ln^5 \frac{dB_\star K}{c_{\min}}$	Efficient, gap-dependent bound
	$\sqrt{d^7 B_\star^2 K}$	Inefficient, horizon-free regret

d: feature dimension, $c_{\min}$: minimum cost, $\mathrm{gap}_{\min}$: minimum sub-optimality gap
$B_\star$: maximum expected cost of optimal policy over all states
$T_\star$: maximum hitting time of optimal policy over all states, K: #episodes

Problem Formulation: SSP

An SSP instance is an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_{\text{init}}, g, c, P)$.

for *episode* $k = 1, \ldots, K$ **do**

learner starts in state $s_1^k = s_{\text{init}} \in \mathcal{S}$, $i \leftarrow 1$

Problem Formulation: SSP

An SSP instance is an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_{\mathrm{init}}, g, c, P)$.

for *episode* $k = 1, \ldots, K$ **do**
- learner starts in state $s_1^k = s_{\mathrm{init}} \in \mathcal{S}, i \leftarrow 1$
- **while** $s_k^i \neq g$ **do**
  - learner chooses action $a_i^k \in \mathcal{A}$, suffer cost $c(s_i^k, a_i^k)$, and observes state $s_{i+1}^k \sim P_{s_i^k, a_i^k}$
  - $i \leftarrow i + 1$

Problem Formulation: SSP

An SSP instance is an MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, s_{\text{init}}, g, c, P)$.

for *episode* $k = 1, \ldots, K$ **do**
- learner starts in state $s_1^k = s_{\text{init}} \in \mathcal{S}$, $i \leftarrow 1$
- **while** $s_k^i \neq g$ **do**
  - learner chooses action $a_i^k \in \mathcal{A}$, suffer cost $c(s_i^k, a_i^k)$, and observes state $s_{i+1}^k \sim P_{s_i^k, a_i^k}$
  - $i \leftarrow i + 1$

$$\text{Regret:} \quad R_K = \sum_{k=1}^{K} \sum_{i=1}^{I_k} c_i^k - \sum_{k=1}^{K} V^\star(s_{\text{init}})$$

Here, $V^\star = V^{\pi^\star}$, $V^\pi(s)$ is the expected cost of policy π starting from s, $\pi^\star = \text{argmin}_{\pi \in \Pi} \sum_{k=1}^{K} V_k^\pi(s_{\text{init}})$, and Π is the set of proper policies which reaches g with probability 1.

Problem Formulation: Linear SSP

Linear SSP

There exist known feature map $\{\phi(s,a)\}_{s,a}$, unknown parameters $\theta^\star \in \mathbb{R}^d$ and $\{\mu(s')\}_{s' \in \mathcal{S} \cup \{g\}} \subseteq \mathbb{R}^d$, such that

$$c(s,a) = \phi(s,a)^\top \theta^\star, \qquad P(s'|s,a) = \phi(s,a)^\top \mu(s').$$

Moreover, we assume $\|\phi(s,a)\|_2 \leq 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\|\theta^\star\|_2 \leq \sqrt{d}$, and $\left\|\int h(s') d\mu(s')\right\|_2 \leq \sqrt{d}\, \|h\|_\infty$ for any $h \in \mathbb{R}^{\mathcal{S}_+}$.

$\sqrt{K}$ Regret Bound

A natural approach is to compute optimistic value functions to guide exploration.

- **Issue:** The value function has circular dependency, which requires computing a fixed point (hard even for discounted MDP).

$\sqrt{K}$ Regret Bound

A natural approach is to compute optimistic value functions to guide exploration.

- **Issue:** The value function has circular dependency, which requires computing a fixed point (hard even for discounted MDP).
- In (Vial et al., 2021), they either 1) perform grid search (inefficient), or 2) find a very inaccurate fixed point with K dependent error (sub-optimal).

$\sqrt{K}$ Regret Bound

A natural approach is to compute optimistic value functions to guide exploration.

- **Issue:** The value function has circular dependency, which requires computing a fixed point (hard even for discounted MDP).
- In (Vial et al., 2021), they either 1) perform grid search (inefficient), or 2) find a very inaccurate fixed point with K dependent error (sub-optimal).

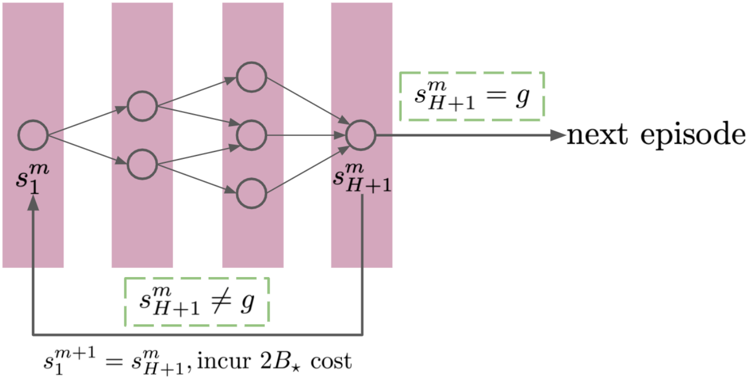
Our Solution:

- Finite-horizon approximation to remove circular dependency!
- Directly run LSVI-UCB (Jin et al., 2020) on the finite-horizon MDP.

Finite-Horizon Approximation

We adopt the finite-horizon approximation scheme in (Cohen et al., 2021).

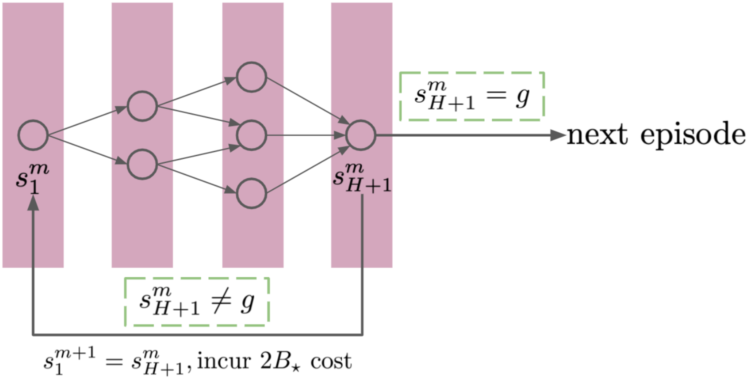
- $\mathcal{M} \to \widetilde{\mathcal{M}}$: each episode in $\mathcal{M}$ is partitioned into one or more intervals in $\widetilde{\mathcal{M}}$.



Finite-Horizon Approximation

We adopt the finite-horizon approximation scheme in (Cohen et al., 2021).

- $\mathcal{M} \to \widetilde{\mathcal{M}}$: each episode in $\mathcal{M}$ is partitioned into one or more intervals in $\widetilde{\mathcal{M}}$.
- $\pi = \widetilde{\pi}$: directly execute $\widetilde{\pi}$ as a non-stationary policy in $\mathcal{M}$.



Technical Challenges & Contributions

Issue: the analysis proposed in (Cohen et al., 2021) assumes a small state-action space.

Technical Challenges & Contributions

Issue: the analysis proposed in (Cohen et al., 2021) assumes a small state-action space.
Our Solution: A new analysis of the finite-horizon approximation.
Intuition: separate the intervals into “good” ones (g is reached) and “bad” ones (g is not reached)

- The large terminal cost implies that each bad interval contributes at least a constant regret.
- Therefore, the number of bad intervals has to be small, and the number of intervals $M = \tilde{\mathcal{O}}(K)$.
- $\tilde{\mathcal{O}}(\sqrt{M})$ in $\widetilde{\mathcal{M}} \implies \tilde{\mathcal{O}}(\sqrt{K})$ in $\mathcal{M}$.

Technical Challenges & Contributions

Highlights:

- Much simpler analysis
- Model agnostic: Does not leverage any modeling assumption on the SSP instance.

Combining with LSVI-UCB gives the first $\tilde{\mathcal{O}}(\sqrt{K})$ regret bound efficiently.

Gap-Dependent Bound

In simpler MDP models, many algorithms are shown to achieve $\mathcal{O}(C \ln K)$ regret, where C is some gap measure.

Gap-Dependent Bound

In simpler MDP models, many algorithms are shown to achieve $\mathcal{O}(C \ln K)$ regret, where C is some gap measure.

- **Gap measure:** $\mathrm{gap}_{\min} = \min_{s,a:\mathrm{gap}(s,a)>0} \mathrm{gap}(s,a)$, where $\mathrm{gap}(s,a) = Q^\star(s,a) - V^\star(s)$.
- **Issue:** after finite-horizon approximation, the gap measure changes to $\mathrm{gap}_h(s,a) = Q^\star_h(s,a) - V^\star_h(s)$.

Gap-Dependent Bound

In simpler MDP models, many algorithms are shown to achieve $\mathcal{O}(C \ln K)$ regret, where C is some gap measure.

- **Gap measure:** $\text{gap}_{\min} = \min_{s,a:\text{gap}(s,a)>0} \text{gap}(s,a)$, where $\text{gap}(s,a) = Q^\star(s,a) - V^\star(s)$.
- **Issue:** after finite-horizon approximation, the gap measure changes to $\text{gap}_h(s,a) = Q_h^\star(s,a) - V_h^\star(s)$.

Our Solution: just need a larger horizon $H = \tilde{\mathcal{O}}(\frac{B_\star}{c_{\min}})$.

Gap-Dependent Bound

High level idea: a two stage analysis.

- For the first $H/2$ layers, we are able to show that $Q_h^\star(s, a) \approx Q^\star(s, a)$, and thus $\mathrm{gap}_h(s, a) \approx \mathrm{gap}(s, a)$.

Gap-Dependent Bound

High level idea: a two stage analysis.

- For the first $H/2$ layers, we are able to show that $Q_h^\star(s,a) \approx Q^\star(s,a)$, and thus $\mathrm{gap}_h(s,a) \approx \mathrm{gap}(s,a)$.
- For the last $H/2$ layers, we further consider two cases:
    - If the learner's policy is near-optimal in the first $H/2$ layers, then the probability of reaching the last $H/2$ layers is negligible.
    - Otherwise, we simply bound the costs by the number of times the learner takes non-near-optimal actions in the first $H/2$ layers, which is of order $\ln K$.

Gap-Dependent Bound

High level idea: a two stage analysis.

- For the first $H/2$ layers, we are able to show that $Q^\star_h(s,a) \approx Q^\star(s,a)$, and thus $\text{gap}_h(s,a) \approx \text{gap}(s,a)$.
- For the last $H/2$ layers, we further consider two cases:
    - If the learner's policy is near-optimal in the first $H/2$ layers, then the probability of reaching the last $H/2$ layers is negligible.
    - Otherwise, we simply bound the costs by the number of times the learner takes non-near-optimal actions in the first $H/2$ layers, which is of order $\ln K$.

Theorem

The algorithm described above ensures $R_K = \tilde{\mathcal{O}}\left(\frac{d^3 B_\star^4}{c_{\min}^2 gap_{\min}} \ln^5 \frac{dB_\star K}{c_{\min}}\right)$.

Horizon-Free Regret

The $T_\star$ or $\frac{1}{c_{\min}}$ dependency is mostly likely unnecessary suggested by the lower bound $\Omega(dB_\star\sqrt{K})$ (Min et al., 2021).

Question: can we obtain horizon-free regret, that is, no polynomial dependency on $T_\star$ or $\frac{1}{c_{\min}}$?

Horizon-Free Regret

The $T_\star$ or $\frac{1}{c_{\min}}$ dependency is mostly likely unnecessary suggested by the lower bound $\Omega(dB_\star\sqrt{K})$ (Min et al., 2021).

Question: can we obtain horizon-free regret, that is, no polynomial dependency on $T_\star$ or $\frac{1}{c_{\min}}$?

Challenges: constructing variance-aware confidence bound is highly non-trivial with linear function approximation, which is known to be the key for obtaining horizon-free regret.

Horizon-Free Regret

Initialize: $t = t' = 1$, $k = 1$, $s_1 = s_{\text{init}}$, $B_1 = 1$.
Define: $V_{w,B}(s) = \min_a [\phi(s,a)^\top w]_{[0,2B]}$, $s'_0 = g$, and $V_t = V_{w_t,B_t}$.

Horizon-Free Regret

Initialize: $t = t' = 1$, $k = 1$, $s_1 = s_{\mathrm{init}}$, $B_1 = 1$.
Define: $V_{w,B}(s) = \min_a [\phi(s,a)^\top w]_{[0,2B]}$, $s'_0 = g$, and $V_t = V_{w_t,B_t}$.
while $k \le K$ **do**
  if $s'_{t-1} = g$ *or some quantity is "doubled" or* $V_{t'}(s_t) = 2B_t$ **then**
    while *True* **do**
      Compute $w_t = \mathrm{argmin}_{w \in \Omega_t(w,B_t)} V_{w,B_t}(s_t)$.
      if $V_t(s_t) > B_t$ **then** $B_t \leftarrow 2B_t$; **else break**.
    Record the most recent update time $t' \leftarrow t$.
  else $(w_t, B_t) = (w_{t-1}, B_{t-1})$.

Horizon-Free Regret

Initialize: $t = t' = 1$, $k = 1$, $s_1 = s_{\rm init}$, $B_1 = 1$.
Define: $V_{w,B}(s) = \min_a [\phi(s,a)^\top w]_{[0,2B]}$, $s'_0 = g$, and $V_t = V_{w_t,B_t}$.
while $k \le K$ **do**
- **if** $s'_{t-1} = g$ *or some quantity is "doubled" or* $V_{t'}(s_t) = 2B_t$ **then**
  - **while** *True* **do**
    - Compute $w_t = \mathrm{argmin}_{w \in \Omega_t(w, B_t)} V_{w,B_t}(s_t)$.
    - **if** $V_t(s_t) > B_t$ **then** $B_t \leftarrow 2B_t$; **else break**.
  - Record the most recent update time $t' \leftarrow t$.
- **else** $(w_t, B_t) = (w_{t-1}, B_{t-1})$.
- Take action $a_t = \mathrm{argmin}_a \, \phi(s_t, a)^\top w_t$, suffer cost $c_t = c(s_t, a_t)$, and transits to s'_t.
- **if** $s'_t = g$ **then** $s_{t+1} = s_{\rm init}$, $k \leftarrow k+1$; **else** $s_{t+1} = s'_t$.
- Increment time step $t \leftarrow t + 1$.

Horizon-free Regret

Technical Highlights

- The construction of transition confidence set is similar to (Zhang et al., 2021), but importantly it computes some fixed point within the decision set.
- Maintain an estimate B_t of $B_\star$, which waives the knowledge of $B_\star$.
- The *overestimate* update condition $V_{t'}(s_t) = 2B_t$ helps remove a $d^{1/4}$ factor.

Horizon-free Regret

Technical Highlights

- The construction of transition confidence set is similar to (Zhang et al., 2021), but importantly it computes some fixed point within the decision set.
- Maintain an estimate B_t of $B_\star$, which waives the knowledge of $B_\star$.
- The *overestimate* update condition $V_{t'}(s_t) = 2B_t$ helps remove a $d^{1/4}$ factor.

Theorem

The algorithm described above ensures $R_K = \tilde{\mathcal{O}}\left(\sqrt{d^7 B_\star^2 K}\right)$.

Conclusion

We further extend our understanding of SSP with linear function approximation.

	Regret	Remark
(Vial et al., 2021)	$\sqrt{d^3 B_\star^3 K / c_{\min}}$	Inefficient
	$K^{5/6}$ (ignoring other params.)	Efficient
Ours	$\sqrt{d^3 B_\star^2 T_\star K}$	Efficient
	$\frac{d^3 B_\star^4}{c_{\min}^2 \mathrm{gap}_{\min}} \ln^5 \frac{dB_\star K}{c_{\min}}$	Efficient, gap-dependent bound
	$\sqrt{d^7 B_\star^2 K}$	Inefficient, horizon-free regret

d: feature dimension, $c_{\min}$: minimum cost, $\mathrm{gap}_{\min}$: minimum sub-optimality gap
$B_\star$: maximum expected cost of optimal policy over all states
$T_\star$: maximum hitting time of optimal policy over all states, K: #episodes