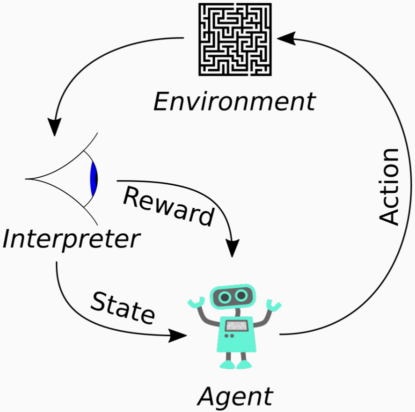
A Unified Algorithm for Stochastic Path Problems

Christoph Dann*, Chen-Yu Wei† and **Julian Zimmert***

August 30, 2023

*Google Research, †University of Southern California

Introduction

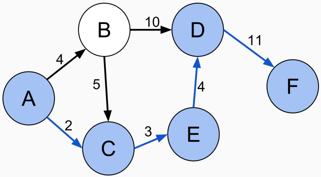


Setting

- Episodic
- Tabular
- No Discount factor
- **Goal-state** (no fixed horizon)

Special case: Stochastic shortest path

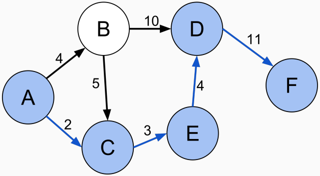
Only suffer costs: $r_t \in [-1, 0]$. Find the goal at the smallest expected cost.



- Rosenberg et al. (2020)
- Cohenet al. (2021)
- Tarbouriech et al. (2020,2021)
- Vial et al. (2022)
- Chen et al. (2021, 2022)
- Chenand Luo (2021, 2022)
- Jafarnia-Jahromi et al. (2021)
- Min et al. (2022)
- Yin et al. (2022)

Special case: Stochastic shortest path

Only suffer costs: $r_t \in [-1, 0]$. Find the goal at the smallest expected cost.

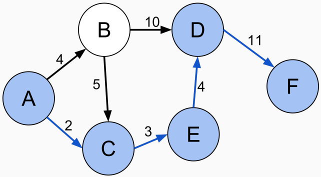


- Rosenberg et al. (2020)
- Cohenet al. (2021)
- Tarbouriech et al. (2020,2021)
- Vial et al. (2022)
- Chen et al. (2021, 2022)
- Chenand Luo (2021, 2022)
- Jafarnia-Jahromi et al. (2021)
- Min et al. (2022)
- Yin et al. (2022)

- **What about general rewards?**
- Can we reduce the problem to SSP?

Special case: Stochastic shortest path

Only suffer costs: $r_t \in [-1, 0]$. Find the goal at the smallest expected cost.



- Rosenberg et al. (2020)
- Cohenet al. (2021)
- Tarbouriech et al. (2020,2021)
- Vial et al. (2022)
- Chen et al. (2021, 2022)
- Chenand Luo (2021, 2022)
- Jafarnia-Jahromi et al. (2021)
- Min et al. (2022)
- Yin et al. (2022)

- **What about general rewards?**
- Can we reduce the problem to SSP?
- **No because of Random stopping time**

Stochastic Path

Algorithm 1: Stochastic path protocol

Input: State space $\mathcal{S} \cup \{g\}$, Action set $\mathcal{A}$

Optional: Problem parameters $B^\star$

for k=1,. . . , K **do**

    $t \leftarrow 1$

    $s_1^k \sim \mu_0$

    while $s_t^k \neq g$ **do**

        Take action $a_t^k \in \mathcal{A}$

        Receive $r_t^k \leftarrow r(s_t^k, a_t^k)$, $r_t^k \in [-1, 1]$

        Observe $s_{t+1}^k \sim P(s_t^k, a_t^k)$

        $t \leftarrow t + 1$

Goal: Minimize

$$\text{Reg} := \max_{\pi \in \Pi} \mathbb{E}[V^\pi(s_1)K - \sum_{k=1}^{K} \sum_{t=1}^{\tau} r_t^k] .$$

Definitions and assumptions

- Π^{HD} history dependent deterministic policy.
- **Assumption:** All policies in Π^{HD} are proper.
- Π^{SD} Stationary deterministic policy.
- $V^{\pi}(s) = \mathbb{E}^{\pi}[\sum_{t=1}^{\tau} r(s_t, a_t) \,|\, s_1 = s]$
- $\pi^{\star} \in \Pi^{\mathsf{SD}}$ such that $\forall \pi \in \Pi^{\mathsf{HD}} : \; V^{\pi^*}(s) \geq V^{\pi}(s)$.

Main algorithm

Algorithm 2: VI-SP

input: $B \geq 1$, $0 < \delta < 1$.
Initialize: $t \leftarrow 0$, $s_1 \sim \nu_0$, $V(g) \leftarrow 0$.
$\forall s \in \mathcal{S}$: $n(s,a,s') = n(s,a) \leftarrow 0, \quad Q(s,a) \leftarrow B, \quad V(s) \leftarrow B$.
for $k = 1, \ldots, K$ **do**
  while *true* **do**
    $t \leftarrow t + 1$
    Play $a_t = \operatorname{argmax}_a Q(s_t, a)$, receive $r(s_t, a_t)$, transit to s'_t.
    Update: $n_t \triangleq n(s_t, a_t) \leftarrow n(s_t, a_t) + 1$, $n(s_t, a_t, s'_t) \leftarrow n(s_t, a_t, s'_t) + 1$.
    Define $\bar{P}_t(s') \triangleq \frac{n(s_t, a_t, s')}{n_t}$ $\forall s'$.
    Define $b_t \triangleq \max\left\{c_1 \sqrt{\frac{\mathbb{V}(\bar{P}_t, V)\iota_t}{n_t}}, \frac{c_2 B \iota_t}{n_t}\right\}$, where $\iota_t = \ln(SA/\delta) + \ln\ln(Bn_t)$.
    $Q(s_t, a_t) \leftarrow \min\left\{r(s_t, a_t) + \bar{P}_t V + b_t, Q(s_t, a_t)\right\}$
    $V(s_t) \leftarrow \max_a Q(s_t, a)$.
    if $s'_t \neq g$ **then** then $s_{t+1} \leftarrow s'_t$;
    else $s_{t+1} \sim \nu_0$ and **break**;

Main result

- $B_\star \triangleq \max_s |V^{\pi^*}(s)|$
- $R \triangleq \sup_{\pi \in \Pi^{\mathrm{HD}}} \sqrt{\mathbb{E}^{\pi}_{s_1 \sim \nu_0}\left[\left(\sum_{i=1}^{\tau} r(s_i, a_i)\right)^2\right]}$
- $R_{\max} \triangleq \max_s \sup_{\pi \in \Pi^{\mathrm{HD}}} \sqrt{\mathbb{E}^{\pi}\left[\left(\sum_{i=1}^{\tau} r(s_i, a_i)\right)^2 \,\middle|\, s_1 = s\right]}$

Theorem

If VI-SP is run with $B \geq B_\star$, then with probability at least $1 - \delta$:

$$\mathrm{Reg} = \widetilde{\mathcal{O}}\left(R\sqrt{SAK} + R_{\max} SA + BS^2 A\right).$$

Main result

- $B_\star \triangleq \max_s |V^{\pi^*}(s)|$
- $R \triangleq \sup_{\pi \in \Pi^{\text{HD}}} \sqrt{\mathbb{E}^{\pi}_{s_1 \sim \nu_0}\left[\left(\sum_{i=1}^{\tau} r(s_i, a_i)\right)^2\right]}$
- $R_{\max} \triangleq \max_s \sup_{\pi \in \Pi^{\text{HD}}} \sqrt{\mathbb{E}^{\pi}\left[\left(\sum_{i=1}^{\tau} r(s_i, a_i)\right)^2 \,\middle|\, s_1 = s\right]}$

Theorem

If VI-SP is run with $B \geq B_\star$, then with probability at least $1 - \delta$:

$$\text{Reg} = \widetilde{\mathcal{O}}\left(R\sqrt{SAK} + R_{\max} SA + BS^2 A\right).$$

SLP/SSP

Let $V_\star = |E_{s_1 \sim \nu_0}[V^{\pi^\star}(s_1)]|$

Lemma

If $r \geq 0$, then $R = \widetilde{\mathcal{O}}(\sqrt{V_\star B_\star})$, $R_{\max} = \widetilde{\mathcal{O}}(B_\star)$.

With known $B_\star$ SLP regret is bounded by

$$\widetilde{\mathcal{O}}\left(\sqrt{V_\star B_\star SAK} + B_\star S^2 A\right)$$

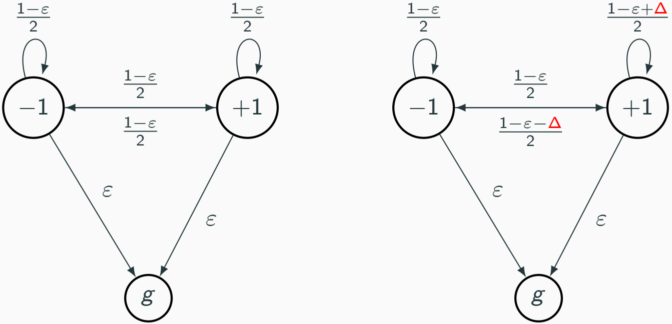
- Same regret in SSP (with more careful analysis)
- Matches Tarbouriech et al. (2021), Chen et al. (2021)

Can we improve the general case?

Lower bounds general case

Theorem

For any $u \geq 2$, and $K \geq \Omega(SA)$, we can construct a set of SP instances such that $R \leq u$, $\sqrt{B_\star \cdot V_\star} \ll u$ for all instances, and there exists a distribution over these instances such that the expected regret of any algorithm is at least $\Omega(u\sqrt{SAK})$.

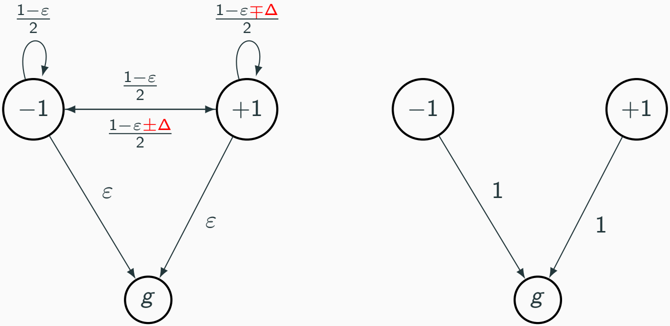


In these instances $R = \mathcal{O}(R_\star) = \mathcal{O}(u)$, can we replace R by $R_\star$?

Lower bound general case

Theorem

Let $u \geq 2$ be arbitrarily chosen, and let $K \geq \Omega(SA)$. For any algorithm that obtains a expected regret bound of $\tilde{O}(u\sqrt{SAK})$ for all problem instances with $R_\star = R_{\max} \leq u$, there exists a problem instance with $R_\star = O(1)$ and $R_{\max} \leq u$ but the expected regret is at least $\tilde{\Omega}(u\sqrt{SAK})$.



Removing knowledge of $B_\star$

Simple idea: Use $B = \sqrt{K/S^3A}$.
Either $B \geq B_\star$ and $\text{Reg} = \widetilde{\mathcal{O}}(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$, or

$B < B_\star$ and $\text{Reg} = \mathcal{O}(V_\star K) \leq \mathcal{O}(V_\star B_\star^2 S^3 A)$.

Removing knowledge of $B_\star$

Simple idea: Use $B = \sqrt{K/S^3A}$.
Either $B \geq B_\star$ and $\text{Reg} = \widetilde{\mathcal{O}}(\sqrt{V_\star B_\star SAK} + B_\star S^2 A)$, or

$B < B_\star$ and $\text{Reg} = \mathcal{O}(V_\star K) \leq \mathcal{O}(V_\star B_\star^2 S^3 A)$.

We give up on all instances where the additive term matters.
Can we do better?

Removing knowledge of $B_\star$ SSP

Idea: We can initialize $Q = 0$, only require B for confidence intervals. Use doubling based on $\max |Q_t|$.

Optimal regret without knowledge of $B_\star$!

Removing knowledge of $B_\star$, SLP

Assume we know $V_\star$.
Simple idea: Use $B = V_\star\sqrt{K}$.
Either $B \geq B_\star$ and $\text{Reg} = \widetilde{\mathcal{O}}(\sqrt{V_\star B_\star K})$, or

$B < B_\star$ and $\text{Reg} = \mathcal{O}(V_\star K) \leq \mathcal{O}\min\{\frac{B_\star}{V_\star}B_\star, B_\star\sqrt{K}\})$.

Removing knowledge of $B_\star$, SLP

Assume we know $V_\star$.
Simple idea: Use $B = V_\star\sqrt{K}$.
Either $B \geq B_\star$ and $\text{Reg} = \widetilde{\mathcal{O}}(\sqrt{V_\star B_\star K})$, or

$B < B_\star$ and $\text{Reg} = \mathcal{O}(V_\star K) \leq \mathcal{O}\min\{\frac{B_\star}{V_\star}B_\star, B_\star\sqrt{K}\})$.

We provide an algorithm to estimate $V_\star$!

Either:	$\text{Reg} = \widetilde{\mathcal{O}}(\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star}S^3A))$
Or:	$\text{Reg} = \widetilde{\mathcal{O}}(B_\star S\sqrt{AK} + B_\star S^2A)\})$

Lower bounds adaptivity

We can not do better in general.

Theorem

Any algorithm with an asymptotic upper bound of

$$\widetilde{\mathcal{O}}\left(B_\star^\alpha V_\star^{1-\alpha}\sqrt{SAK}\right) + o\left(B_\star^2\right) \,,$$

satisfies at least $\alpha \geq 1$ *and any algorithm with an upper bound of*

$$\widetilde{\mathcal{O}}\left(\sqrt{V_\star B_\star SAK} + \left(\frac{B_\star}{V_\star}\right)^2 poly(V_\star, S, A)\right)$$

requires the constant term to be at least $\tilde{\Omega}\left(\frac{B_\star^2 SA}{V_\star}\right)$.

Summary

Setting	Scale $B_\star$	Reg_K in $\tilde{O}(\cdot)$	
SP	known	$R\sqrt{SAK} + R_{\max}SA + B_\star S^2 A$	Theorem 2
		$R\sqrt{SAK}$ **(lower bound)**	Thm 3, Thm 4
SLP	known	$\sqrt{V_\star B_\star SAK} + B_\star S^2 A$	Theorem 6
	unknown	$B_\star S\sqrt{AK}$ or $\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star} S^3 A$	Theorem 8
		$B_\star\sqrt{SAK}$ or $\sqrt{V_\star B_\star SAK} + \frac{B_\star^2}{V_\star} SA$ **(lower bound)**	Corollary 10
SSP	known	$\sqrt{V_\star B_\star SAK} + B_\star S^2 A$	[1],[2]
	unknown	$\sqrt{V_\star B_\star SAK} + B_\star^3 S^3 A$	[1],[2]
		$\sqrt{V_\star B_\star SAK} + B_\star S^2 A$	Theorem 11

[1] Tarbouriech et al.(2021)
[2] Chen et al.(2021)

Questions?