Towards Minimax Regret for Stochastic Shortest Path with Adversarial Costs

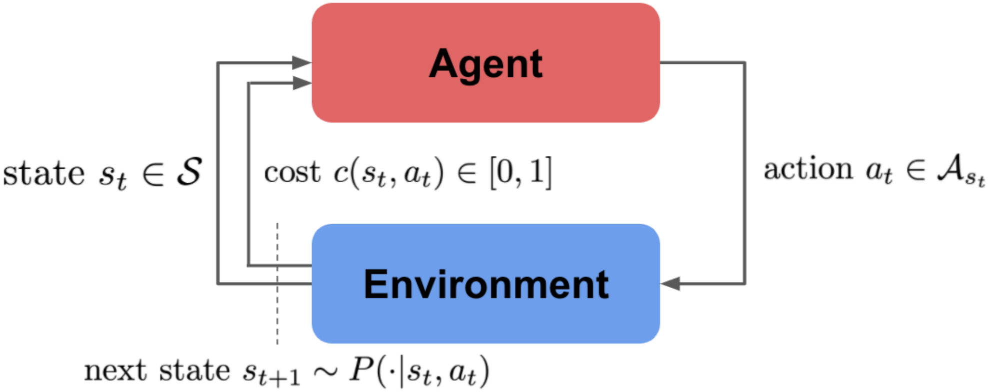
Presenter: Liyu Chen

Liyu Chen Haipeng Luo Chen-Yu Wei

University of Southern California

September 12, 2021

Problem Formulation: Markov Decision Process (MDP)



We assume finite state space $\mathcal{S}$ and action space $\mathcal{A} = \{\mathcal{A}_s\}_{s \in \mathcal{S}}$.

Motivation

Many MDP models have been studied:

- Infinite horizon average reward model (Bartlett & Tewari, 2009; Jaksch et al., 2010)
- Infinite horizon discounted model (Even-Dar et al., 2003; Strehl et al., 2006)
- Finite horizon model (Osband and Van Roy, 2016; Azar et al., 2017; Jin et al., 2018)

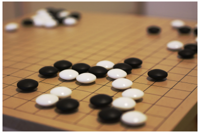
Motivation

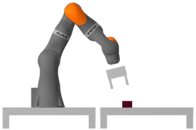
Many MDP models have been studied:

- Infinite horizon average reward model (Bartlett & Tewari, 2009; Jaksch et al., 2010)
- Infinite horizon discounted model (Even-Dar et al., 2003; Strehl et al., 2006)
- Finite horizon model (Osband and Van Roy, 2016; Azar et al., 2017; Jin et al., 2018)

However, there are many real-world applications not modelled well by the above:

- Games (such as Go)
- Car navigation
- Robotic manipulation





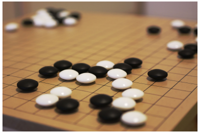
Motivation

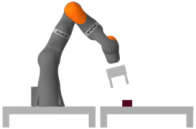
Many MDP models have been studied:

- Infinite horizon average reward model (Bartlett & Tewari, 2009; Jaksch et al., 2010)
- Infinite horizon discounted model (Even-Dar et al., 2003; Strehl et al., 2006)
- Finite horizon model (Osband and Van Roy, 2016; Azar et al., 2017; Jin et al., 2018)

However, there are many real-world applications not modelled well by the above:

- Games (such as Go)
- Car navigation
- Robotic manipulation





For these, Stochastic Shortest Path (SSP) is a better model.

- Episodic MDP with a goal state.

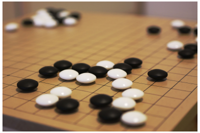
Motivation

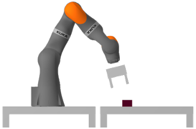
Many MDP models have been studied:

- Infinite horizon average reward model (Bartlett & Tewari, 2009; Jaksch et al., 2010)
- Infinite horizon discounted model (Even-Dar et al., 2003; Strehl et al., 2006)
- Finite horizon model (Osband and Van Roy, 2016; Azar et al., 2017; Jin et al., 2018)

However, there are many real-world applications not modelled well by the above:

- Games (such as Go)
- Car navigation
- Robotic manipulation





For these, Stochastic Shortest Path (SSP) is a better model.

- Episodic MDP with a goal state.
- Challenges: variable episode length, possibly unbounded cost, etc.

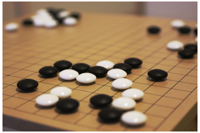
Motivation

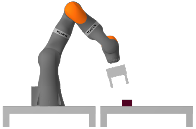
Many MDP models have been studied:

- Infinite horizon average reward model (Bartlett & Tewari, 2009; Jaksch et al., 2010)
- Infinite horizon discounted model (Even-Dar et al., 2003; Strehl et al., 2006)
- Finite horizon model (Osband and Van Roy, 2016; Azar et al., 2017; Jin et al., 2018)

However, there are many real-world applications not modelled well by the above:

- Games (such as Go)
- Car navigation
- Robotic manipulation





For these, Stochastic Shortest Path (SSP) is a better model.

- Episodic MDP with a goal state.
- Challenges: variable episode length, possibly unbounded cost, etc.
- Not well studied yet.

Related Works

S: # states, A: # actions, D: SSP-diameter, K: # episodes,
$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

- SSP with stochastic cost:
  - UC-SSP (Tarbouriech et al., 2020): $\tilde{\mathcal{O}}\left(DS\sqrt{\frac{D}{c_{\min}}AK}\right)$

Related Works

S: # states, A: # actions, D: SSP-diameter, K: # episodes,
$T_\star$: expected hitting time of optimal policy, c_{min}: minimum cost

- SSP with stochastic cost:
  - UC-SSP (Tarbouriech et al., 2020): $\tilde{\mathcal{O}}\left(DS\sqrt{\frac{D}{c_{\text{min}}}AK}\right)$
  - Bernstein-SSP (Cohen et al., 2020): $\tilde{\mathcal{O}}\left(DS\sqrt{AK}\right)$

Related Works

S: # states, A: # actions, D: SSP-diameter, K: # episodes,
$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

- SSP with stochastic cost:
  - UC-SSP (Tarbouriech et al., 2020): $\tilde{\mathcal{O}}\left(DS\sqrt{\frac{D}{c_{\min}}AK}\right)$
  - Bernstein-SSP (Cohen et al., 2020): $\tilde{\mathcal{O}}\left(DS\sqrt{AK}\right)$
- SSP with adversarial cost (full information):
  - SSP-O-REPS (Rosenberg and Mansour, 2020): $\tilde{\mathcal{O}}\left(\frac{D}{c_{\min}}\sqrt{K}\right)$ or $\tilde{\mathcal{O}}\left(\sqrt{DT_\star}K^{3/4}\right)$ with known transition

Our Results

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes
$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

	Minimax Regret (**this talk**)	(Rosenberg and Mansour, 2020)
Full information	$\Theta(\sqrt{DT_\star K})$	$\tilde{\mathcal{O}}\left(\frac{D}{c_{\min}}\sqrt{K}\right)$ or $\tilde{\mathcal{O}}\left(\sqrt{DT_\star}K^{\frac{3}{4}}\right)$
Bandit feedback	$\Theta(\sqrt{DT_\star SAK})$	N/A

Our contributions: we develop **efficient minimax optimal** algorithms for both full information and bandit feedback setting with known transition.

Follow-up Work for Unknown Transition

S: # of states, A: # of actions, D: SSP-diameter, K: # of episodes
$T_\star$: expected hitting time of optimal policy, $c_{\min}$: minimum cost

	Follow-up	(Rosenberg and Mansour, 2020)	Lower bounds
Full information	$\tilde{\mathcal{O}}\left(\sqrt{S^2ADT_\star K}\right)$	$\tilde{\mathcal{O}}\left(\frac{DS}{c_{\min}}\sqrt{AK}\right)$ or $\tilde{\mathcal{O}}\left(\sqrt{S^2AT_\star^2}K^{3/4} + D^2\sqrt{K}\right)$	$\Omega(\sqrt{DT_\star K} + D\sqrt{SAK})$
Bandit feedback	$\tilde{\mathcal{O}}\left(\sqrt{S^3A^2DT_\star K}\right)$	N/A	$\Omega(\sqrt{SADT_\star K} + D\sqrt{SAK})$

Paper: https://arxiv.org/abs/2102.05284.

Highlights

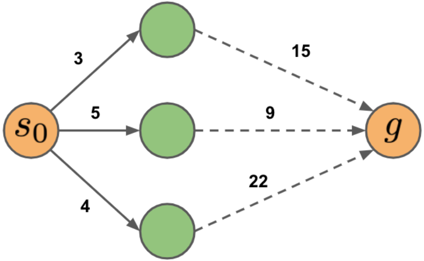
All algorithms are based on Online Mirror Descent (OMD).

Many new ideas are required to achieve desired results.

- A new multi-scale expert algorithm
- A reduction from a general SSP to its loop-free version
- Skewed occupancy measure
- Log-barrier regularizer
- An increasing learning rate schedule
- A negative bias injected to the cost function

Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^{K}$



Problem Formulation

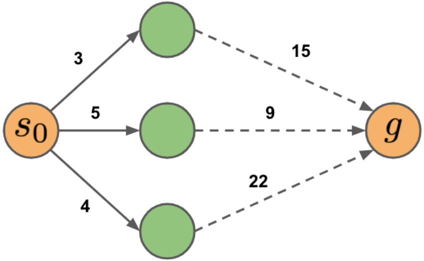
SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Learning Protocol

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

  environment chooses c_k adaptively (based on learner's algorithm and history)

end



Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Learning Protocol

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

environment chooses c_k adaptively (based on learner's algorithm and history)

learner starts in state $s_k^1 = s_0, i \leftarrow 1$

while $s_k^i \neq g$ **do**

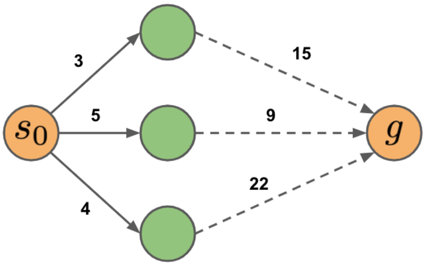
learner chooses action $a_k^i \in \mathcal{A}_{s_k^i}$

learner observes states $s_k^{i+1} \sim P(\cdot | s_k^i, a_k^i)$

$i \leftarrow i + 1$

end

end



Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Learning Protocol

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

  environment chooses c_k adaptively (based on learner's algorithm and history)

  learner starts in state $s_k^1 = s_0, i \leftarrow 1$

  while $s_k^i \neq g$ **do**

    learner chooses action $a_k^i \in \mathcal{A}_{s_k^i}$

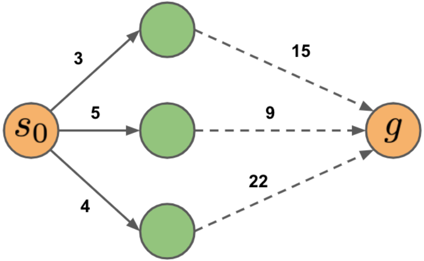
    learner observes states $s_k^{i+1} \sim P(\cdot|s_k^i, a_k^i)$

    $i \leftarrow i + 1$

  end

  learner observes c_k (full information) or $\{c(s_k^i, a_k^i)\}_{i=1}^{I_k}$ (bandit feedback) and suffer cost $\sum_{i=1}^{I_k} c(s_k^i, a_k^i)$

end



Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^{K}$

Notations:

- Valid state-action pairs $\Gamma = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$

Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Notations:

- Valid state-action pairs $\Gamma = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$
- Policy π: maps $s \in \mathcal{S}$ to distribution over $\mathcal{A}_s$
  - Proper: reaches g with probability 1

Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Notations:

- Valid state-action pairs $\Gamma = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$
- Policy π: maps $s \in \mathcal{S}$ to distribution over $\mathcal{A}_s$
  - Proper: reaches g with probability 1
- Cost-to-go function $J^\pi(s) = \mathbb{E}[\sum_{i=1}^I c(s^i, a^i)|P, \pi, s^1 = s]$

Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Notations:

- Valid state-action pairs $\Gamma = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$
- Policy π: maps $s \in \mathcal{S}$ to distribution over $\mathcal{A}_s$
  - Proper: reaches g with probability 1
- Cost-to-go function $J^\pi(s) = \mathbb{E}[\sum_{i=1}^I c(s^i, a^i) | P, \pi, s^1 = s]$
- Expected hitting time $T^\pi(s) = \mathbb{E}[I | P, \pi, s^1 = s]$

Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Notations:

- Valid state-action pairs $\Gamma = \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$
- Policy π: maps $s \in \mathcal{S}$ to distribution over $\mathcal{A}_s$
    - Proper: reaches g with probability 1
- Cost-to-go function $J^\pi(s) = \mathbb{E}[\sum_{i=1}^{I} c(s^i, a^i) | P, \pi, s^1 = s]$
- Expected hitting time $T^\pi(s) = \mathbb{E}[I | P, \pi, s^1 = s]$
- $D = \max_s \min_{\pi \in \Pi_{\text{proper}}} T^\pi(s)$, $T_\star = T^{\pi^\star}(s_0)$.

Problem Formulation

SSP Model: MDP $M = (\mathcal{S}, \mathcal{A}, s_0, g, P)$ + cost functions $\{c_k\}_{k=1}^K$

Notations:

- Valid state-action pairs $\Gamma = \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$
- Policy π: maps $s \in \mathcal{S}$ to distribution over $\mathcal{A}_s$
  - Proper: reaches g with probability 1
- Cost-to-go function $J^\pi(s) = \mathbb{E}[\sum_{i=1}^I c(s^i, a^i) | P, \pi, s^1 = s]$
- Expected hitting time $T^\pi(s) = \mathbb{E}[I | P, \pi, s^1 = s]$
- $D = \max_s \min_{\pi \in \Pi_{\text{proper}}} T^\pi(s)$, $T_\star = T^{\pi^\star}(s_0)$.

Objective: minimize regret w.r.t. the best stationary proper policy in hindsight

$$R_K = \sum_{k=1}^{K} \left(\sum_{i=1}^{I_k} c_k(s_k^i, a_k^i) - J_k^{\pi^\star}(s_0) \right),$$

where $\pi^\star = \operatorname{argmin}_{\pi \in \Pi_{\text{proper}}} \sum_{k=1}^K J_k^{\pi^\star}(s_0)$

Occupancy Measure

A proper policy π induces an occupancy measure $q_\pi \in \mathbb{R}^{\Gamma}_{\geq 0}$ with

$$q_\pi(s, a) = \mathbb{E}\left[\sum_{i=1}^{I} \mathbb{I}\{s^i = s, a^i = a\} \middle| P, \pi, s^1 = s_0\right],$$

which is the expected number of visits to (s, a) when executing π.

Occupancy Measure

A proper policy π induces an occupancy measure $q_\pi \in \mathbb{R}^{\Gamma}_{\geq 0}$ with

$$q_\pi(s, a) = \mathbb{E}\left[\sum_{i=1}^{I} \mathbb{I}\{s^i = s, a^i = a\} \middle| P, \pi, s^1 = s_0\right],$$

which is the expected number of visits to (s, a) when executing π.

- One-to-one correspondence: $\pi_q(a|s) \propto q(s, a)$

Occupancy Measure

A proper policy π induces an occupancy measure $q_\pi \in \mathbb{R}^\Gamma_{\geq 0}$ with

$$q_\pi(s,a) = \mathbb{E}\left[\sum_{i=1}^{I} \mathbb{I}\{s^i = s, a^i = a\} \middle| P, \pi, s^1 = s_0\right],$$

which is the expected number of visits to (s,a) when executing π.

- One-to-one correspondence: $\pi_q(a|s) \propto q(s,a)$
- $J_k^\pi(s_0) = \langle q_\pi, c_k \rangle$, $T^\pi(s_0) = \sum_{(s,a)} q_\pi(s,a)$

Occupancy Measure

A proper policy π induces an occupancy measure $q_\pi \in \mathbb{R}^{\Gamma}_{\geq 0}$ with

$$q_\pi(s,a) = \mathbb{E}\left[\sum_{i=1}^{I} \mathbb{I}\{s^i = s, a^i = a\} \middle| P, \pi, s^1 = s_0\right],$$

which is the expected number of visits to (s,a) when executing π.

- One-to-one correspondence: $\pi_q(a|s) \propto q(s,a)$
- $J_k^\pi(s_0) = \langle q_\pi, c_k \rangle$, $T^\pi(s_0) = \sum_{(s,a)} q_\pi(s,a)$
- $\mathbb{E}[R_K] = \mathbb{E}\left[\sum_{k=1}^K \langle q_{\pi_k} - q_{\pi^\star}, c_k \rangle\right]$

Occupancy Measure

A proper policy π induces an occupancy measure $q_\pi \in \mathbb{R}^{\Gamma}_{\geq 0}$ with

$$q_\pi(s,a) = \mathbb{E}\left[\sum_{i=1}^{I} \mathbb{I}\{s^i = s, a^i = a\} \middle| P, \pi, s^1 = s_0\right],$$

which is the expected number of visits to (s,a) when executing π.

- One-to-one correspondence: $\pi_q(a|s) \propto q(s,a)$
- $J_k^\pi(s_0) = \langle q_\pi, c_k \rangle$, $T^\pi(s_0) = \sum_{(s,a)} q_\pi(s,a)$
- $\mathbb{E}[R_K] = \mathbb{E}\left[\sum_{k=1}^{K} \langle q_{\pi_k} - q_{\pi^\star}, c_k \rangle\right]$

Converting into online linear optimization. Apply Online Mirror Descent (OMD)!

Occupancy Measure

Define the decision set of occupancy measures:

$$\Delta(T) = \Bigg\{ q \in \mathbb{R}^{\Gamma}_{\geq 0} : \sum_{(s,a)\in\Gamma} q(s,a) \leq T,$$

$$\sum_{a\in\mathcal{A}_s} q(s,a) - \sum_{(s',a')\in\Gamma} P(s|s',a')q(s',a') = \mathbb{I}\{s = s_0\},\ \forall s \in \mathcal{S} \Bigg\}$$

T is an upper bound on expected hitting time.

Full information, Expected Regret

Key challenge: achieve optimal bound without knowing $T_\star$

Solution: a new multi-scale expert algorithm as meta learner

Full Information, Expected Regret

Algorithm 1 SSP-O-REPS (Rosenberg and Mansour, 2020)

Input: upper bound on expected hitting time T.

Define: regularizer $\psi(q) = \frac{1}{\eta}\sum_{(s,a)} q(s,a) \ln q(s,a)$ and $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{T\ln(SAT)}{DK}}\right\}$.

Initialization: $q_1 = \mathrm{argmin}_{q\in\Delta(T)}\, \psi(q)$.

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

  Execute π_{q_k}, receive c_k, and update $q_{k+1} = \mathrm{argmin}_{q\in\Delta(T)} \langle q, c_k\rangle + D_\psi(q, q_k)$.

end

Full Information, Expected Regret

Algorithm 1 SSP-O-REPS (Rosenberg and Mansour, 2020)

Input: upper bound on expected hitting time T.

Define: regularizer $\psi(q) = \frac{1}{\eta} \sum_{(s,a)} q(s,a) \ln q(s,a)$ and $\eta = \min \left\{ \frac{1}{2}, \sqrt{\frac{T \ln(SAT)}{DK}} \right\}$.

Initialization: $q_1 = \operatorname{argmin}_{q \in \Delta(T)} \psi(q)$.

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

  Execute π_{q_k}, receive c_k, and update $q_{k+1} = \operatorname{argmin}_{q \in \Delta(T)} \langle q, c_k \rangle + D_\psi(q, q_k)$.

end

Rosenberg and Mansour (2020) proves $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(T\sqrt{K}\right)$.

Full Information, Expected Regret

Algorithm 1 SSP-O-REPS (Rosenberg and Mansour, 2020)

Input: upper bound on expected hitting time T.

Define: regularizer $\psi(q) = \frac{1}{\eta} \sum_{(s,a)} q(s,a) \ln q(s,a)$ and $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{T \ln(SAT)}{DK}}\right\}$.

Initialization: $q_1 = \operatorname{argmin}_{q \in \Delta(T)} \psi(q)$.

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

Execute π_{q_k}, receive c_k, and update $q_{k+1} = \operatorname{argmin}_{q \in \Delta(T)} \langle q, c_k \rangle + D_\psi(q, q_k)$.

end

Rosenberg and Mansour (2020) proves $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(T\sqrt{K}\right)$.

We improve their analysis by the fact $\sum_{k=1}^{K} J_k^{\pi^\star}(s_0) \leq DK$:

Theorem

Algorithm 1 ensures $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(\sqrt{DTK}\right)$ *as long as* $T \geq T_\star$.

Full Information, Expected Regret

Algorithm 1 SSP-O-REPS (Rosenberg and Mansour, 2020)

Input: upper bound on expected hitting time T.

Define: regularizer $\psi(q) = \frac{1}{\eta}\sum_{(s,a)} q(s,a)\ln q(s,a)$ and $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{T\ln(SAT)}{DK}}\right\}$.

Initialization: $q_1 = \operatorname{argmin}_{q\in\Delta(T)} \psi(q)$.

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

Execute π_{q_k}, receive c_k, and update $q_{k+1} = \operatorname{argmin}_{q\in\Delta(T)} \langle q, c_k\rangle + D_\psi(q, q_k)$.

end

Rosenberg and Mansour (2020) proves $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(T\sqrt{K}\right)$.

We improve their analysis by the fact $\sum_{k=1}^K J_k^{\pi^\star}(s_0) \le DK$:

Theorem

Algorithm 1 ensures $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(\sqrt{DTK}\right)$ as long as $T \ge T_\star$. (Problem: need to know $T_\star$)

Full Information, Expected Regret

Question: how to deal with unknown $T_\star$?

Full Information, Expected Regret

Question: how to deal with unknown $T_\star$?

Solution: run multiple O-REPS-SSP instances with different T and learn the best.

- Maintain $N \approx \log_2 K$ SSP-O-REPS instances, where the j-th instance sets $T \approx 2^j$.
- Each instance is an action of the meta-algorithm. Define meta-loss $\ell_k(j) = \left\langle q_k^j, c_k \right\rangle$.

Full Information, Expected Regret

Question: how to deal with unknown $T_\star$?

Solution: run multiple O-REPS-SSP instances with different T and learn the best.

- Maintain $N \approx \log_2 K$ SSP-O-REPS instances, where the j-th instance sets $T \approx 2^j$.
- Each instance is an action of the meta-algorithm. Define meta-loss $\ell_k(j) = \left\langle q_k^j, c_k \right\rangle$.
    - **Issue:** losses have different scales, $\ell_k(j) \leq b(j) \approx 2^j$.

Full Information, Expected Regret

Question: how to deal with unknown $T_\star$?

Solution: run multiple O-REPS-SSP instances with different T and learn the best.

- Maintain $N \approx \log_2 K$ SSP-O-REPS instances, where the j-th instance sets $T \approx 2^j$.
- Each instance is an action of the meta-algorithm. Define meta-loss $\ell_k(j) = \left\langle q_k^j, c_k \right\rangle$.
  - **Issue:** losses have different scales, $\ell_k(j) \le b(j) \approx 2^j$.
  - **Natural first attempt:** apply multi-scale expert algorithm (Bubeck et al., 2017):

$$\psi(p) = \sum_{j=1}^{N} \frac{1}{\eta_j} p(j) \ln p(j), \qquad p_1(j) \propto \eta_j$$

Full Information, Expected Regret

Question: how to deal with unknown $T_\star$?

Solution: run multiple O-REPS-SSP instances with different T and learn the best.

- Maintain $N \approx \log_2 K$ SSP-O-REPS instances, where the j-th instance sets $T \approx 2^j$.
- Each instance is an action of the meta-algorithm. Define meta-loss $\ell_k(j) = \left\langle q_k^j, c_k \right\rangle$.
  - **Issue:** losses have different scales, $\ell_k(j) \leq b(j) \approx 2^j$.
  - **Natural first attempt:** apply multi-scale expert algorithm (Bubeck et al., 2017):

$$\psi(p) = \sum_{j=1}^{N} \frac{1}{\eta_j} p(j) \ln p(j), \qquad p_1(j) \propto \eta_j$$

- However, known multi-scale algorithms only ensure $\tilde{\mathcal{O}}\left(b(j^\star)\sqrt{K}\right)$ regret, not optimal.

Full Information, Expected Regret

Question: how to deal with unknown $T_\star$?

Solution: run multiple O-REPS-SSP instances with different T and learn the best.

- Maintain $N \approx \log_2 K$ SSP-O-REPS instances, where the j-th instance sets $T \approx 2^j$.
- Each instance is an action of the meta-algorithm. Define meta-loss $\ell_k(j) = \left\langle q_k^j, c_k \right\rangle$.
  - **Issue:** losses have different scales, $\ell_k(j) \leq b(j) \approx 2^j$.
  - **Natural first attempt:** apply multi-scale expert algorithm (Bubeck et al., 2017):

$$\psi(p) = \sum_{j=1}^{N} \frac{1}{\eta_j} p(j) \ln p(j), \qquad p_1(j) \propto \eta_j$$

- However, known multi-scale algorithms only ensure $\tilde{\mathcal{O}}\left(b(j^\star)\sqrt{K}\right)$ regret, not optimal.
- **Our solution:** inspired by other works for adaptive regret bound (Steinhardt and Liang, 2014; Wei and Luo, 2018), we change $\ell_k(j)$ to $\ell_k(j) + 4\eta_j \ell_k^2(j)$ (penalizing long horizon policy), which gives $\tilde{\mathcal{O}}\left(\sqrt{b(j^\star)\mathbb{E}[\sum_{k=1}^{K} \ell_k(j^\star)]}\right)$ regret.

Full Information, Expected Regret

Algorithm 2 Adaptive SSP-O-REPS with Multi-scale Experts

Define: $\Omega = \left\{p \in \mathbb{R}^N_{\geq 0} : \sum_{j=1}^N p(j) = 1\right\}$ and $\psi(p) = \sum_{j=1}^N \frac{1}{\eta_j} p(j) \ln p(j)$.

Initialize: $p_1 \in \Omega$ such that $p_1(j) \propto \eta_j$.

Initialize: N instances of SSP-O-REPS, where the j-th instance uses parameter $T = b(j)$.

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

- For each $j \in [N]$, obtain occupancy measure q_k^j from SSP-O-REPS instance j.
- Sample $j_k \sim p_k$, execute π_k induced by $q_k^{j_k}$, receive c_k, and feed c_k to all instances.
- Compute ℓ_k and a_k: $\ell_k(j) = \langle q_k^j, c_k \rangle$, $a_k(j) = 4\eta_j \ell_k^2(j)$, $\forall j \in [N]$.
- Update $p_{k+1} = \mathrm{argmin}_{p \in \Omega} \langle p, \ell_k + a_k \rangle + D_\psi(p, p_k)$.

end

Full Information, Expected Regret

Algorithm 2 Adaptive SSP-O-REPS with Multi-scale Experts

Define: $\Omega = \left\{p \in \mathbb{R}^N_{\geq 0} : \sum_{j=1}^N p(j) = 1\right\}$ and $\psi(p) = \sum_{j=1}^N \frac{1}{\eta_j} p(j) \ln p(j)$.
Initialize: $p_1 \in \Omega$ such that $p_1(j) \propto \eta_j$.
Initialize: N instances of SSP-O-REPS, where the j-th instance uses parameter $T = b(j)$.
for $k = 1, \dots, K$ **do**

For each $j \in [N]$, obtain occupancy measure q_k^j from SSP-O-REPS instance j.
Sample $j_k \sim p_k$, execute π_k induced by $q_k^{j_k}$, receive c_k, and feed c_k to all instances.
Compute ℓ_k and a_k: $\ell_k(j) = \langle q_k^j, c_k \rangle$, $a_k(j) = 4\eta_j \ell_k^2(j)$, $\forall j \in [N]$.
Update $p_{k+1} = \operatorname{argmin}_{p \in \Omega} \langle p, \ell_k + a_k \rangle + D_\psi(p, p_k)$.

end

Theorem

Algorithm 2 ensures $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(\sqrt{DT_\star K}\right)$ without knowing $T_\star$ (which is optimal).

Outline

Full Information, High Probability Bound

Key challenge: control the variance of learner's cost

Solution: loop-free reduction $+$ skewed occupancy measure

Full Information, High Probability Bound

$$R_K = \sum_{k=1}^{K} \langle N_k - q_{\pi^\star}, c_k \rangle = \underbrace{\sum_{k=1}^{K} \langle N_k - q_k, c_k \rangle}_{Deviation} + \underbrace{\sum_{k=1}^{K} \langle q_k - q_{\pi^\star}, c_k \rangle}_{REG},$$

where $N_k(s,a) = \sum_{i=1}^{I_k} \mathbb{I}\{s_k^i = s, a_k^i = a\}$.

Full Information, High Probability Bound

$$R_K = \sum_{k=1}^{K} \langle N_k - q_{\pi^\star}, c_k \rangle = \underbrace{\sum_{k=1}^{K} \langle N_k - q_k, c_k \rangle}_{Deviation} + \underbrace{\sum_{k=1}^{K} \langle q_k - q_{\pi^\star}, c_k \rangle}_{REG},$$

where $N_k(s,a) = \sum_{i=1}^{I_k} \mathbb{I}\{s_k^i = s, a_k^i = a\}$.

Issue: there is no good upper bound on $\langle N_k, c_k \rangle$.

Full Information, High Probability Bound

$$R_K = \sum_{k=1}^{K} \langle N_k - q_{\pi^\star}, c_k \rangle = \underbrace{\sum_{k=1}^{K} \langle N_k - q_k, c_k \rangle}_{\textit{Deviation}} + \underbrace{\sum_{k=1}^{K} \langle q_k - q_{\pi^\star}, c_k \rangle}_{\textit{REG}},$$

where $N_k(s,a) = \sum_{i=1}^{I_k} \mathbb{I}\{s_k^i = s, a_k^i = a\}$.

Issue: there is no good upper bound on $\langle N_k, c_k \rangle$.

Lemma (Quantifying Deviation in SSP)

Consider executing a stationary policy π in episode k. Then $\mathbb{E}_k[\langle N_k, c_k \rangle^2] \le 2 \langle q_\pi, J_k^\pi \rangle$.

Full Information, High Probability Bound

Lemma (Quantifying Deviation in SSP)

Consider executing a stationary policy π *in episode* k. *Then* $\mathbb{E}[\langle N_k, c_k\rangle^2] \le 2\,\langle q_\pi, J_k^\pi\rangle$.

Observation 1: for the optimal policy $\pi^\star$:

$$\sum_{k=1}^{K} \left\langle q_{\pi^\star}, J_k^{\pi^\star} \right\rangle = \sum_{s\in\mathcal{S}} q_{\pi^\star}(s) \sum_{k=1}^{K} J_k^{\pi^\star}(s) \le DK \sum_{s\in\mathcal{S}} q_{\pi^\star}(s) = DT_\star K.$$

Full Information, High Probability Bound

Lemma (Quantifying Deviation in SSP)

Consider executing a stationary policy π *in episode* k. *Then* $\mathbb{E}[\langle N_k, c_k \rangle^2] \leq 2 \langle q_\pi, J_k^\pi \rangle$.

Observation 1: for the optimal policy $\pi^\star$:

$$\sum_{k=1}^{K} \left\langle q_{\pi^\star}, J_k^{\pi^\star} \right\rangle = \sum_{s \in \mathcal{S}} q_{\pi^\star}(s) \sum_{k=1}^{K} J_k^{\pi^\star}(s) \leq DK \sum_{s \in \mathcal{S}} q_{\pi^\star}(s) = DT_\star K.$$

It is thus tempting to enforce $\sum_{k=1}^{K} \langle q_{\pi_k}, J_k^{\pi_k} \rangle \leq DT_\star K$. But how?

- It depends on all cost functions $c_1, \ldots, c_K$.
- Non-convex w.r.t. occupancy measure.

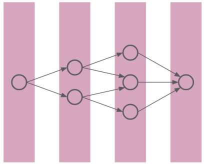
Full Information, High Probability Bound

Observation 2: the variance upper bound takes a much simpler form in a loop-free MDP.

Full Information, High Probability Bound

Observation 2: the variance upper bound takes a much simpler form in a loop-free MDP.

Loop-free layered structure:



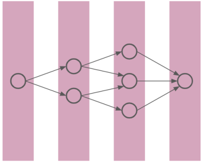
- State space is of the form $\mathcal{S} \times [H]$.
- Transition from (s, h) to (s', h') is only possible if $h' = h + 1$ (except transition to the goal state).

Full Information, High Probability Bound

Observation 2: the variance upper bound takes a much simpler form in a loop-free MDP.

Loop-free layered structure:

- State space is of the form $\mathcal{S} \times [H]$.
- Transition from (s, h) to (s', h') is only possible if $h' = h + 1$ (except transition to the goal state).



Lemma (Quantifying Deviation in loop-free MDP)

If M has a loop-free layered structure, then

$$\langle q_\pi, J_k^\pi \rangle = \sum_{(s,a)} \sum_{h=1}^{H} h \cdot q_\pi(s, a, h) c_k(s, a, h) = \left\langle q_\pi, \vec{h} \circ c_k \right\rangle,$$

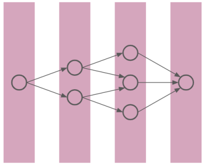
where we define $\vec{h} \circ f(s, a, h) = h \cdot f(s, a, h)$. For simplicity, we write $q_\pi((s, h), a)$ as $q_\pi(s, a, h)$, and $c_k((s, h), a)$ as $c_k(s, a, h)$.

Full Information, High Probability Bound

Observation 2: the variance upper bound takes a much simpler form in a loop-free MDP.

Loop-free layered structure:

- State space is of the form $\mathcal{S} \times [H]$.
- Transition from (s, h) to (s', h') is only possible if $h' = h + 1$ (except transition to the goal state).



Lemma (Quantifying Deviation in loop-free MDP)

If M has a loop-free layered structure, then

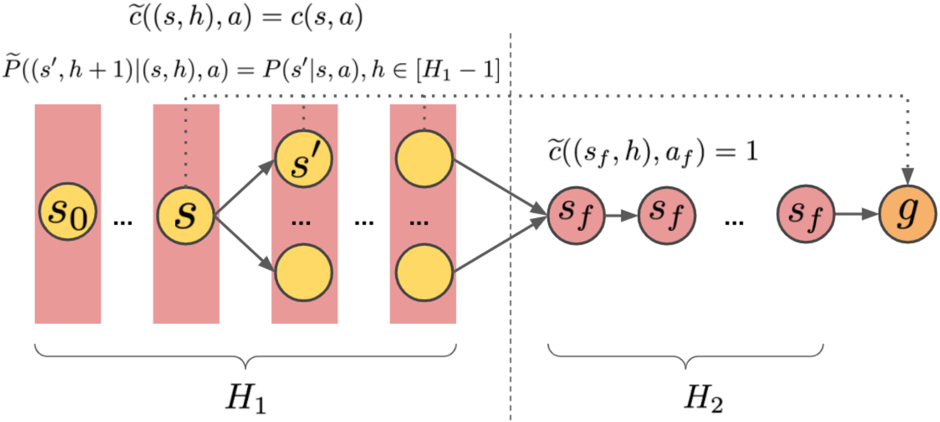
$$\langle q_\pi, J_k^\pi \rangle = \sum_{(s,a)} \sum_{h=1}^{H} h \cdot q_\pi(s,a,h) c_k(s,a,h) = \left\langle q_\pi, \vec{h} \circ c_k \right\rangle,$$

where we define $\vec{h} \circ f(s,a,h) = h \cdot f(s,a,h)$. For simplicity, we write $q_\pi((s,h),a)$ as $q_\pi(s,a,h)$, and $c_k((s,h),a)$ as $c_k(s,a,h)$.

This inspires us to approximate the SSP instance by a loop-free MDP.

First Idea: Loop-free Reduction

Construct $\widetilde{M}$ from M: duplicate each state by attaching a time step h for H_1 steps, and then connect all states to some dummy state that lasts for another H_2 steps.



For simplicity, write $q(s,a,h) = q((s,h),a)$, $c(s,a,h) = \widetilde{c}((s,h),a)$, and define $H = H_1 + H_2$.

First Idea: Loop-free Reduction

Given $\widetilde{\pi}$ in $\widetilde{M}$, define non-stationary policy $\sigma(\widetilde{\pi})$ in M which

1. follows $\widetilde{\pi}(\cdot|(s,h))$ at state s for time step $h \leq H_1$
2. then execute the fast policy π^f until reaching g

First Idea: Loop-free Reduction

Given $\widetilde{\pi}$ in $\widetilde{M}$, define non-stationary policy $\sigma(\widetilde{\pi})$ in M which

1. follows $\widetilde{\pi}(\cdot|(s,h))$ at state s for time step $h \leq H_1$
2. then execute the fast policy π^f until reaching g

Lemma

Suppose $H_1 \gtrsim \max_s T^{\pi^\star}(s), H_2 \gtrsim D$. *Let* $\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K$ *be policies for* $\widetilde{M}$ *with occupancy measure* $q_1, \ldots, q_K$. *Then the regret of executing* $\sigma(\widetilde{\pi}_1), \ldots, \sigma(\widetilde{\pi}_K)$ *in* M *satisfies for any* $\lambda \in (0, 2/H]$, *with probability* $1-\delta$,

$$R_K \leq \sum_{k=1}^{K} \left\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k \right\rangle + \tilde{\mathcal{O}}(1)$$

First Idea: Loop-free Reduction

Given $\widetilde{\pi}$ in $\widetilde{M}$, define non-stationary policy $\sigma(\widetilde{\pi})$ in M which

1. follows $\widetilde{\pi}(\cdot|(s,h))$ at state s for time step $h \le H_1$
2. then execute the fast policy π^f until reaching g

Lemma

Suppose $H_1 \gtrsim \max_s T^{\pi^\star}(s), H_2 \gtrsim D$. Let $\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K$ be policies for $\widetilde{M}$ with occupancy measure $q_1, \ldots, q_K$. Then the regret of executing $\sigma(\widetilde{\pi}_1), \ldots, \sigma(\widetilde{\pi}_K)$ in M satisfies for any $\lambda \in (0, 2/H]$, with probability $1-\delta$,

$$R_K \le \sum_{k=1}^{K} \left\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k \right\rangle + \tilde{\mathcal{O}}(1)$$

Note: applying standard loop-free algorithms does not solve our problem!

First Idea: Loop-free Reduction

Given $\widetilde{\pi}$ in $\widetilde{M}$, define non-stationary policy $\sigma(\widetilde{\pi})$ in M which

1. follows $\widetilde{\pi}(\cdot|(s,h))$ at state s for time step $h \le H_1$
2. then execute the fast policy π^f until reaching g

Lemma

Suppose $H_1 \gtrsim \max_s T^{\pi^\star}(s)$, $H_2 \gtrsim D$. Let $\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K$ be policies for $\widetilde{M}$ with occupancy measure $q_1, \ldots, q_K$. Then the regret of executing $\sigma(\widetilde{\pi}_1), \ldots, \sigma(\widetilde{\pi}_K)$ in M satisfies for any $\lambda \in (0, 2/H]$, with probability $1-\delta$,

$$R_K \le \sum_{k=1}^K \left\langle \widetilde{N}_k - q_{\widetilde{\pi}^\star}, c_k \right\rangle + \tilde{\mathcal{O}}(1) \le \underbrace{\sum_{k=1}^K \langle q_k - q_{\widetilde{\pi}^\star}, c_k \rangle}_{\textsc{Reg}} + \lambda \underbrace{\sum_{k=1}^K \left\langle q_k, \vec{h} \circ c_k \right\rangle}_{\textsc{Var}} + \frac{2\ln(2/\delta)}{\lambda} + \tilde{\mathcal{O}}(1).$$

Note: applying standard loop-free algorithms does not solve our problem!

Second Idea: Skewed Occupancy Measure Space

$$R_K \lesssim \sum_{k=1}^{K} \langle q_k - q_{\widetilde{\pi}^\star}, c_k \rangle + \lambda \underbrace{\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle}_{\textsc{Var}} + \frac{1}{\lambda}.$$

Second Idea: Skewed Occupancy Measure Space

$$R_K \lesssim \sum_{k=1}^{K} \langle q_k - q_{\widetilde{\pi}^\star}, c_k \rangle + \lambda \underbrace{\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle}_{\textsc{Var}} + \frac{1}{\lambda}.$$

It is still hard to enforce $\textsc{Var} \leq DT_\star K$. Instead, we have the following observation:

$$\begin{aligned} R_K &\lesssim \sum_{k=1}^{K} \left\langle (q_k + \lambda \vec{h} \circ q_k) - (q_{\widetilde{\pi}^\star} + \lambda \vec{h} \circ q_{\widetilde{\pi}^\star}), c_k \right\rangle + \lambda \sum_{k=1}^{K} \left\langle q_{\widetilde{\pi}^\star}, \vec{h} \circ c_k \right\rangle + \frac{1}{\lambda} \\ &\lesssim \sum_{k=1}^{K} \langle \phi_{\pi_k} - \phi_{\widetilde{\pi}^\star}, c_k \rangle + \lambda D T_\star K + \frac{1}{\lambda}. \qquad (\phi_\pi = q_\pi + \lambda \vec{h} \circ q_\pi) \end{aligned}$$

Second Idea: Skewed Occupancy Measure Space

$$R_K \lesssim \sum_{k=1}^{K} \langle q_k - q_{\widetilde{\pi}^\star}, c_k \rangle + \lambda \underbrace{\sum_{k=1}^{K} \left\langle q_k, \vec{h} \circ c_k \right\rangle}_{\textsc{Var}} + \frac{1}{\lambda}.$$

It is still hard to enforce $\textsc{Var} \leq DT_\star K$. Instead, we have the following observation:

$$\begin{aligned} R_K &\lesssim \sum_{k=1}^{K} \left\langle (q_k + \lambda \vec{h} \circ q_k) - (q_{\widetilde{\pi}^\star} + \lambda \vec{h} \circ q_{\widetilde{\pi}^\star}), c_k \right\rangle + \lambda \sum_{k=1}^{K} \left\langle q_{\widetilde{\pi}^\star}, \vec{h} \circ c_k \right\rangle + \frac{1}{\lambda} \\ &\lesssim \sum_{k=1}^{K} \langle \phi_{\pi_k} - \phi_{\widetilde{\pi}^\star}, c_k \rangle + \lambda D T_\star K + \frac{1}{\lambda}. \qquad (\phi_\pi = q_\pi + \lambda \vec{h} \circ q_\pi) \end{aligned}$$

It thus motivates us to perform OMD over a skewed occupancy measure space:

$$\Omega = \left\{ \phi = q + \lambda \vec{h} \circ q : q \in \widetilde{\Delta}(T_\star) \right\}.$$

Full Information, High Probability Bound

Algorithm 3 SSP-O-REPS with Loop-free Reduction and Skewed Occupancy Measure

Parameters: $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{T_\star}{DK}}\right\}, \lambda = \sqrt{\frac{\ln(1/\delta)}{DT_\star K}}, H_2 = \lceil 4D \ln \frac{4K}{\delta} \rceil$

Define: $H = H_1 + H_2$, regularizer $\psi(\phi) = \frac{1}{\eta}\sum_{h=1}^{H}\sum_{(s,a)\in\widetilde{\Gamma}} \phi(s,a,h)\ln\phi(s,a,h)$

Initialization: $\phi_1 = q_1 + \lambda\vec{h}\circ q_1 = \mathrm{argmin}_{\phi\in\Omega}\,\psi(\phi)$.

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

  Execute $\sigma(\widetilde{\pi}_k)$ where $\widetilde{\pi}_k$ is such that $\widetilde{\pi}_k(a|(s,h)) \propto q_k(s,a,h)$, and receive c_k.

  Update $\phi_{k+1} = q_{k+1} + \lambda\vec{h}\circ q_{k+1} = \mathrm{argmin}_{\phi\in\Omega}\,\langle\phi, c_k\rangle + D_\psi(\phi,\phi_k)$.

end

Full Information, High Probability Bound

Algorithm 3 SSP-O-REPS with Loop-free Reduction and Skewed Occupancy Measure

Parameters: $\eta = \min\left\{\frac{1}{2}, \sqrt{\frac{T_\star}{DK}}\right\}, \lambda = \sqrt{\frac{\ln(1/\delta)}{DT_\star K}}, H_2 = \lceil 4D \ln \frac{4K}{\delta} \rceil$

Define: $H = H_1 + H_2$, regularizer $\psi(\phi) = \frac{1}{\eta} \sum_{h=1}^{H} \sum_{(s,a) \in \widetilde{\Gamma}} \phi(s,a,h) \ln \phi(s,a,h)$

Initialization: $\phi_1 = q_1 + \lambda \vec{h} \circ q_1 = \mathrm{argmin}_{\phi \in \Omega} \psi(\phi)$.

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

  Execute $\sigma(\widetilde{\pi}_k)$ where $\widetilde{\pi}_k$ is such that $\widetilde{\pi}_k(a|(s,h)) \propto q_k(s,a,h)$, and receive c_k.

  Update $\phi_{k+1} = q_{k+1} + \lambda \vec{h} \circ q_{k+1} = \mathrm{argmin}_{\phi \in \Omega} \langle \phi, c_k \rangle + D_\psi(\phi, \phi_k)$.

end

Theorem

Algorithm 3 ensures that $R_K = \tilde{\mathcal{O}}\left(\sqrt{DT_\star K}\right)$ *with high probability.*

Open Problem: How to achieve the same without knowing $T_\star$?

Bandit Feedback, Expected Bound

Key challenge: large variance of unbiased cost estimators

Solution: log-barrier regularizer + skewed occupancy measure

Bandit Feedback, Expected Bound

- Standard technique: construct an importance-weighted unbiased cost estimator. The natural estimator is $\widehat{c}_k(s, a) = \frac{N_k(s,a)c_k(s,a)}{q_k(s,a)}$.

Bandit Feedback, Expected Bound

- Standard technique: construct an importance-weighted unbiased cost estimator. The natural estimator is $\widehat{c}_k(s,a) = \frac{N_k(s,a)c_k(s,a)}{q_k(s,a)}$.
- With the entropy regularizer, the stability term of OMD is $\sum_{(s,a)} q_k(s,a)\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \sum_{(s,a)} \frac{\mathbb{E}_k[N_k^2(s,a)]c_k(s,a)}{q_k(s,a)}$, which could be prohibitively large.

Bandit Feedback, Expected Bound

- Standard technique: construct an importance-weighted unbiased cost estimator. The natural estimator is $\widehat{c}_k(s,a) = \frac{N_k(s,a)c_k(s,a)}{q_k(s,a)}$.
- With the entropy regularizer, the stability term of OMD is $\sum_{(s,a)} q_k(s,a)\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \sum_{(s,a)} \frac{\mathbb{E}_k[N_k^2(s,a)]c_k(s,a)}{q_k(s,a)}$, which could be prohibitively large.

We resolve these problems with **Log-barrier** regularizer $\psi(\phi) = -\sum_{(s,a)} \ln(\sum_{h=1}^{H} \phi(s,a,h))$.

Bandit Feedback, Expected Bound

- Standard technique: construct an importance-weighted unbiased cost estimator. The natural estimator is $\widehat{c}_k(s,a) = \frac{N_k(s,a)c_k(s,a)}{q_k(s,a)}$.
- With the entropy regularizer, the stability term of OMD is $\sum_{(s,a)} q_k(s,a)\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \sum_{(s,a)} \frac{\mathbb{E}_k[N_k^2(s,a)]c_k(s,a)}{q_k(s,a)}$, which could be prohibitively large.

We resolve these problems with **Log-barrier** regularizer $\psi(\phi) = -\sum_{(s,a)} \ln(\sum_{h=1}^H \phi(s,a,h))$.

- It leads to a smaller stability term:

$$\sum_{(s,a)} q_k^2(s,a)\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \sum_{(s,a)} \mathbb{E}_k[N_k^2(s,a)]c_k^2(s,a) \leq \mathbb{E}_k[\langle N_k, c_k \rangle^2].$$

Exactly the variance of actual cost and can be handled by skewed occupancy measure!

Bandit Feedback, Expected Bound

- Standard technique: construct an importance-weighted unbiased cost estimator. The natural estimator is $\widehat{c}_k(s,a) = \frac{N_k(s,a)c_k(s,a)}{q_k(s,a)}$.
- With the entropy regularizer, the stability term of OMD is $\sum_{(s,a)} q_k(s,a)\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \sum_{(s,a)} \frac{\mathbb{E}_k[N_k^2(s,a)]c_k(s,a)}{q_k(s,a)}$, which could be prohibitively large.

We resolve these problems with **Log-barrier** regularizer $\psi(\phi) = -\sum_{(s,a)} \ln(\sum_{h=1}^H \phi(s,a,h))$.

- It leads to a smaller stability term:

$$\sum_{(s,a)} q_k^2(s,a)\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \sum_{(s,a)} \mathbb{E}_k[N_k^2(s,a)]c_k^2(s,a) \le \mathbb{E}_k[\langle N_k, c_k\rangle^2].$$

  Exactly the variance of actual cost and can be handled by skewed occupancy measure!
- Summing over H inside to avoid H dependency (leveraging the fact $c(s,a,h) = c(s,a)$).

Bandit Feedback, Expected Bound

Algorithm 4 Log-barrier Policy Search for SSP

Define: regularizer $\psi(\phi) = -\frac{1}{\eta}\sum_{(s,a)\in\widetilde{\Gamma}} \ln \phi(s,a)$ where $\phi(s,a) = \sum_{h=1}^{H} \phi(s,a,h)$

Initialization: $\phi_1 = q_1 + \lambda\vec{h} \circ q_1 = \mathrm{argmin}_{\phi\in\Omega}\,\psi(\phi)$.

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

Execute $\sigma(\widetilde{\pi}_k)$ where $\widetilde{\pi}_k$ is such that $\widetilde{\pi}_k(a|(s,h)) \propto q_k(s,a,h)$.

Construct cost estimator $\widehat{c}_k \in \mathbb{R}_{\geq 0}^{\widetilde{\Gamma}}$ such that $\widehat{c}_k(s,a) = \frac{\widetilde{N}_k(s,a)c_k(s,a)}{q_k(s,a)}$.

Update $\phi_{k+1} = q_{k+1} + \lambda\vec{h} \circ q_{k+1} = \mathrm{argmin}_{\phi\in\Omega}\,\langle\phi, \widehat{c}_k\rangle + D_\psi(\phi, \phi_k)$.

end

Bandit Feedback, Expected Bound

Algorithm 4 Log-barrier Policy Search for SSP

Define: regularizer $\psi(\phi) = -\frac{1}{\eta}\sum_{(s,a)\in\widetilde{\Gamma}} \ln \phi(s,a)$ where $\phi(s,a) = \sum_{h=1}^{H} \phi(s,a,h)$

Initialization: $\phi_1 = q_1 + \lambda\vec{h} \circ q_1 = \operatorname{argmin}_{\phi\in\Omega} \psi(\phi)$.

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

  Execute $\sigma(\widetilde{\pi}_k)$ where $\widetilde{\pi}_k$ is such that $\widetilde{\pi}_k(a|(s,h)) \propto q_k(s,a,h)$.

  Construct cost estimator $\widehat{c}_k \in \mathbb{R}^{\widetilde{\Gamma}}_{\geq 0}$ such that $\widehat{c}_k(s,a) = \frac{\widetilde{N}_k(s,a)c_k(s,a)}{q_k(s,a)}$.

  Update $\phi_{k+1} = q_{k+1} + \lambda\vec{h} \circ q_{k+1} = \operatorname{argmin}_{\phi\in\Omega} \langle \phi, \widehat{c}_k \rangle + D_\psi(\phi, \phi_k)$.

end

Theorem

Algorithm 4 ensures $\mathbb{E}[R_K] = \tilde{\mathcal{O}}\left(\sqrt{DT_\star SAK}\right)$ *(which is optimal).*

Bandit Feedback, High Probability Bound

Key challenge: large variance of the cost estimators for $\pi^\star$

Solution: skewed occupancy measure + increasing learning rate + negative bias injected to cost function (positive bias + negative bias)

Bandit Feedback, High Probability Bound

Question: how to obtain a high probability bound?

Bandit Feedback, High Probability Bound

Question: how to obtain a high probability bound?

- The key is to bound the deviation of $\pi^\star$: $\sum_{k=1}^{K} \langle q_{\pi^\star}, \widehat{c}_k - c_k \rangle$.

Bandit Feedback, High Probability Bound

Question: how to obtain a high probability bound?

- The key is to bound the deviation of $\pi^\star$: $\sum_{k=1}^{K} \langle q_{\pi^\star}, \widehat{c}_k - c_k \rangle$.
- By $\mathbb{E}_k[\widetilde{N}_k^2(s,a)] \le \sum_h h \cdot q_k(s,a,h)$ in the loop-free setting:

$$\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \frac{\mathbb{E}_k[\widetilde{N}_k^2(s,a)]c_k^2(s,a)}{q_k^2(s,a)} \le \frac{\sum_h h \cdot q_k(s,a,h)c_k(s,a)}{q_k^2(s,a)} \le \rho_K(s,a)b_k(s,a),$$

where $\rho_K(s,a) = \max_k \frac{1}{q_k(s,a)}$ and $b_k(s,a) = \frac{\sum_h h q_k(s,a,h)c_k(s,a)}{q_k(s,a)}$.

Bandit Feedback, High Probability Bound

Question: how to obtain a high probability bound?

- The key is to bound the deviation of $\pi^\star$: $\sum_{k=1}^{K} \langle q_{\pi^\star}, \widehat{c}_k - c_k \rangle$.
- By $\mathbb{E}_k[\widetilde{N}_k^2(s,a)] \le \sum_h h \cdot q_k(s,a,h)$ in the loop-free setting:

$$\mathbb{E}_k[\widehat{c}_k^2(s,a)] = \frac{\mathbb{E}_k[\widetilde{N}_k^2(s,a)] c_k^2(s,a)}{q_k^2(s,a)} \le \frac{\sum_h h \cdot q_k(s,a,h) c_k(s,a)}{q_k^2(s,a)} \le \rho_K(s,a) b_k(s,a),$$

where $\rho_K(s,a) = \max_k \frac{1}{q_k(s,a)}$ and $b_k(s,a) = \frac{\sum_h h q_k(s,a,h) c_k(s,a)}{q_k(s,a)}$.

- By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a) \sqrt{\rho_K(s,a) \sum_{k=1}^{K} b_k(s,a)} \le \frac{1}{\eta} \langle q_{\widetilde{\pi}^\star}, \rho_K \rangle + \eta \sum_{k=1}^{K} \langle q_{\widetilde{\pi}^\star}, b_k \rangle,$$

Bandit Feedback, High Probability Bound

By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a) \sqrt{\rho_K(s,a) \sum_{k=1}^{K} b_k(s,a)} \le \frac{1}{\eta} \langle q_{\widetilde{\pi}^\star}, \rho_K \rangle + \eta \sum_{k=1}^{K} \langle q_{\widetilde{\pi}^\star}, b_k \rangle ,$$

- The first term $\frac{1}{\eta} \langle q_{\widetilde{\pi}^\star}, \rho_K \rangle$ appears in (Lee et al., 2020a) already and can be handled by an increasing learning rate schedule.

Bandit Feedback, High Probability Bound

By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a)\sqrt{\rho_K(s,a)\sum_{k=1}^{K} b_k(s,a)} \leq \frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K\right\rangle + \eta\sum_{k=1}^{K}\left\langle q_{\widetilde{\pi}^\star}, b_k\right\rangle,$$

- The first term $\frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K\right\rangle$ appears in (Lee et al., 2020a) already and can be handled by an increasing learning rate schedule.
- To handle the second term, we inject a negative bias: replacing $\widehat{c}_k$ by $\widehat{c}_k - \eta b_k$.

Bandit Feedback, High Probability Bound

By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a)\sqrt{\rho_K(s,a)\sum_{k=1}^{K} b_k(s,a)} \le \frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K\right\rangle + \eta\sum_{k=1}^{K}\left\langle q_{\widetilde{\pi}^\star}, b_k\right\rangle,$$

- The first term $\frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K\right\rangle$ appears in (Lee et al., 2020a) already and can be handled by an increasing learning rate schedule.
- To handle the second term, we inject a negative bias: replacing $\widehat{c}_k$ by $\widehat{c}_k - \eta b_k$.
    - Gives a negative term $-\eta\sum_{k=1}^{K}\left\langle q_{\widetilde{\pi}^\star}, b_k\right\rangle$. Cancel out the second term.

Bandit Feedback, High Probability Bound

By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a)\sqrt{\rho_K(s,a)\sum_{k=1}^{K} b_k(s,a)} \leq \frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K \right\rangle + \eta \sum_{k=1}^{K} \left\langle q_{\widetilde{\pi}^\star}, b_k \right\rangle,$$

- The first term $\frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K \right\rangle$ appears in (Lee et al., 2020a) already and can be handled by an increasing learning rate schedule.
- To handle the second term, we inject a negative bias: replacing $\widehat{c}_k$ by $\widehat{c}_k - \eta b_k$.
  - Gives a negative term $-\eta\sum_{k=1}^{K}\left\langle q_{\widetilde{\pi}^\star}, b_k \right\rangle$. Cancel out the second term.
  - Incurs a bias $\eta\sum_{k=1}^{K}\left\langle q_k, b_k \right\rangle = \eta\sum_{k=1}^{K}\left\langle q_k, \vec{h}\circ c_k \right\rangle$. Again handled by the skewed occupancy measure.

Bandit Feedback, High Probability Bound

By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a)\sqrt{\rho_K(s,a)\sum_{k=1}^K b_k(s,a)} \le \frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K\right\rangle + \eta\sum_{k=1}^K \left\langle q_{\widetilde{\pi}^\star}, b_k\right\rangle,$$

- The first term $\frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K\right\rangle$ appears in (Lee et al., 2020a) already and can be handled by an increasing learning rate schedule.
- To handle the second term, we inject a negative bias: replacing $\widehat{c}_k$ by $\widehat{c}_k - \eta b_k$.
  - Gives a negative term $-\eta\sum_{k=1}^K \left\langle q_{\widetilde{\pi}^\star}, b_k\right\rangle$. Cancel out the second term.
  - Incurs a bias $\eta\sum_{k=1}^K \left\langle q_k, b_k\right\rangle = \eta\sum_{k=1}^K \left\langle q_k, \vec{h}\circ c_k\right\rangle$. Again handled by the skewed occupancy measure.
- Since c_k is unknown, we use $\widehat{b}_k$ instead of b_k with $\widehat{b}_k(s,a) = \frac{\sum_h q_k(s,a,h)\widehat{c}_k(s,a)}{q_k(s,a)}$.

Bandit Feedback, High Probability Bound

By Freedman's inequality, the deviation is bounded by

$$\sum_{(s,a)} q_{\widetilde{\pi}^\star}(s,a)\sqrt{\rho_K(s,a)\sum_{k=1}^{K} b_k(s,a)} \le \frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K \right\rangle + \eta \sum_{k=1}^{K}\left\langle q_{\widetilde{\pi}^\star}, b_k \right\rangle,$$

- The first term $\frac{1}{\eta}\left\langle q_{\widetilde{\pi}^\star}, \rho_K \right\rangle$ appears in (Lee et al., 2020a) already and can be handled by an increasing learning rate schedule.
- To handle the second term, we inject a negative bias: replacing $\widehat{c}_k$ by $\widehat{c}_k - \eta b_k$.
  - Gives a negative term $-\eta \sum_{k=1}^{K}\left\langle q_{\widetilde{\pi}^\star}, b_k \right\rangle$. Cancel out the second term.
  - Incurs a bias $\eta \sum_{k=1}^{K}\left\langle q_k, b_k \right\rangle = \eta \sum_{k=1}^{K}\left\langle q_k, \vec{h} \circ c_k \right\rangle$. Again handled by the skewed occupancy measure.
- Since c_k is unknown, we use $\widehat{b}_k$ instead of b_k with $\widehat{b}_k(s,a) = \frac{\sum_h q_k(s,a,h)\widehat{c}_k(s,a)}{q_k(s,a)}$.
- We apply both positive (skewed occupancy measure) and negative bias (increasing learning rate, $-\eta\widehat{b}_k$)!

Bandit Feedback, High Probability Bound

Algorithm 5 Log-barrier Policy Search for SSP (High Probability)

Initialization: for all $(s, a) \in \widetilde{\Gamma}, \eta_1(s, a) = \eta, \rho_1(s, a) = 2T$.

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

  Execute $\sigma(\widetilde{\pi}_k)$ where $\widetilde{\pi}_k$ is such that $\widetilde{\pi}_k(a|(s, h)) \propto q_k(s, a, h)$.

  $\phi_{k+1} = q_{k+1} + \lambda\vec{h} \circ q_{k+1} = \text{argmin}_{\phi \in \Omega} \left\langle \phi, \widehat{c}_k - \gamma\widehat{b}_k \right\rangle + D_{\psi_k}(\phi, \phi_k)$.

  for $\forall(s, a) \in \widetilde{\Gamma}$ **do**

    if $\frac{1}{\phi_{k+1}(s,a)} > \rho_k(s, a)$ **then** $\rho_{k+1}(s, a) = \frac{2}{\phi_{k+1}(s,a)}, \eta_{k+1}(s, a) = \beta\eta_k(s, a)$;

    else $\rho_{k+1}(s, a) = \rho_k(s, a), \eta_{k+1}(s, a) = \eta_k(s, a)$;

  end

end

Bandit Feedback, High Probability Bound

Algorithm 5 Log-barrier Policy Search for SSP (High Probability)

Initialization: for all $(s,a) \in \widetilde{\Gamma}, \eta_1(s,a) = \eta, \rho_1(s,a) = 2T$.

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

Execute $\sigma(\widetilde{\pi}_k)$ where $\widetilde{\pi}_k$ is such that $\widetilde{\pi}_k(a|(s,h)) \propto q_k(s,a,h)$.

$\phi_{k+1} = q_{k+1} + \lambda \vec{h} \circ q_{k+1} = \text{argmin}_{\phi \in \Omega} \left\langle \phi, \widehat{c}_k - \gamma \widehat{b}_k \right\rangle + D_{\psi_k}(\phi, \phi_k)$.

for $\forall (s,a) \in \widetilde{\Gamma}$ **do**

if $\frac{1}{\phi_{k+1}(s,a)} > \rho_k(s,a)$ **then** $\rho_{k+1}(s,a) = \frac{2}{\phi_{k+1}(s,a)}, \eta_{k+1}(s,a) = \beta \eta_k(s,a)$;

else $\rho_{k+1}(s,a) = \rho_k(s,a), \eta_{k+1}(s,a) = \eta_k(s,a)$;

end

end

Theorem

Algorithm 5 ensures $R_K = \tilde{\mathcal{O}}\left(\sqrt{DT_\star SAK}\right)$ *with high probability.*

Open Problems

- How to achieve high probability bound without knowing $T_\star$?
- Minimax optimal algorithms for the unknown transition setting.
    - The bounds in our follow-up work are not optimal yet.

Thank You!

References

- Sébastien Bubeck, Nikhil R Devanur, Zhiyi Huang, and Rad Niazadeh. Online auctions and multiscale online learning. In Proceedings of the 2017 ACM Conference on Economics and Computation, pages 497–514, 2017
- Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. Advances in Neural Information Processing Systems, 33, 2020a.
- 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 Proceedings of the 37th International Conference on Machine Learning, pages 8210–8219, 2020.
- Jean Tarbouriech, Evrard Garcelon, Michal Valko, Matteo Pirotta, and Alessandro Lazaric. Noregret exploration in goal-oriented reinforcement learning. In International Conference on Machine Learning, pages 9428–9437. PMLR, 2020.

References

- Jaksch, Thomas, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning." Journal of Machine Learning Research 11.4 (2010).
- Bartlett, P. L. and Tewari, A. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 35–42. AUAI Press, 2009.
- Even-Dar, Eyal, Yishay Mansour, and Peter Bartlett. "Learning Rates for Q-learning." Journal of machine learning Research 5.1 (2003).
- Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In Proceedings of the 23rd international conference on Machine learning, pages 881–888. ACM, 2006.
- Azar, Mohammad Gheshlaghi, Ian Osband, and Rémi Munos. "Minimax regret bounds for reinforcement learning." International Conference on Machine Learning. PMLR, 2017.
- Osband, Ian and Van Roy, Benjamin. On lower bounds for regret in reinforcement learning. stat, 1050:9, 2016a.

References

- Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In Advances in neural information processing systems, pages 4863–4873, 2018.
- Wei, Chen-Yu, and Haipeng Luo. "More adaptive algorithms for adversarial bandits." Conference On Learning Theory. PMLR, 2018.
- Jacob Steinhardt and Percy Liang. Adaptivity and optimism: An improved exponentiated gradient algorithm. In International Conference on Machine Learning, pages 1593–1601, 2014.

Backup Slides

Outline

Lower Bound

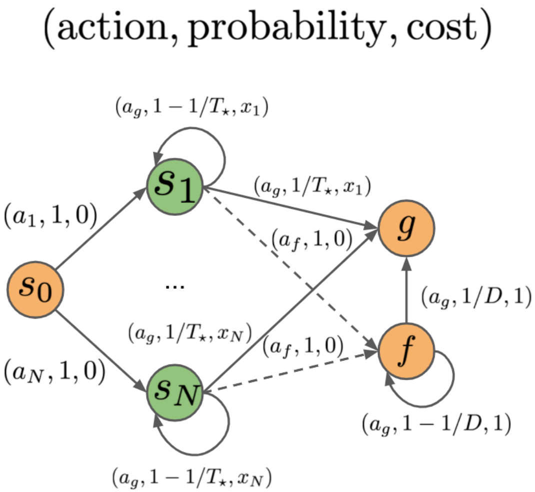
Lower Bound

Main idea: an analogy to an expert / MAB problem with loss scale $T_\star$ and total losses DK.

Lower Bound

Main idea: an analogy to an expert / MAB problem with loss scale $T_\star$ and total losses DK.

- Uniformly sample a good state $j^\star \in [N]$ and fixed throughout the K episodes
- In each episode:
  - $x_{j^\star} \sim \text{Bernoulli}(\frac{D}{2T_\star})$
  - $x_j \sim \text{Bernoulli}(\frac{D}{2T_\star} + \epsilon)$ for any $j \neq j^\star$

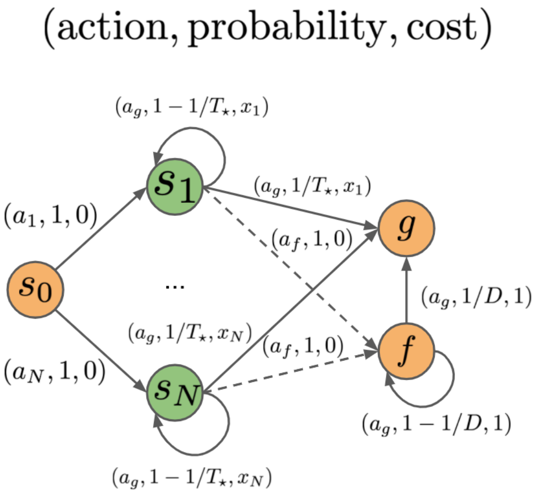


Lower Bound

Main idea: an analogy to an expert / MAB problem with loss scale $T_\star$ and total losses DK.

- Uniformly sample a good state $j^\star \in [N]$ and fixed throughout the K episodes
- In each episode:
  - $x_{j^\star} \sim \text{Bernoulli}(\frac{D}{2T_\star})$
  - $x_j \sim \text{Bernoulli}(\frac{D}{2T_\star} + \epsilon)$ for any $j \neq j^\star$

Full information: $\Omega(\sqrt{DT_\star K})$
Bandit feedback: $\Omega(\sqrt{DT_\star SAK})$

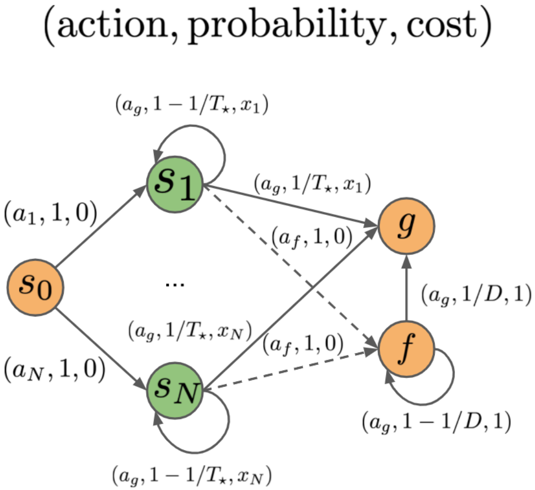


Lower Bound

Main idea: an analogy to an expert / MAB problem with loss scale $T_\star$ and total losses DK.

- Uniformly sample a good state $j^\star \in [N]$ and fixed throughout the K episodes
- In each episode:
  - $x_{j^\star} \sim \text{Bernoulli}(\frac{D}{2T_\star})$
  - $x_j \sim \text{Bernoulli}(\frac{D}{2T_\star} + \epsilon)$ for any $j \neq j^\star$

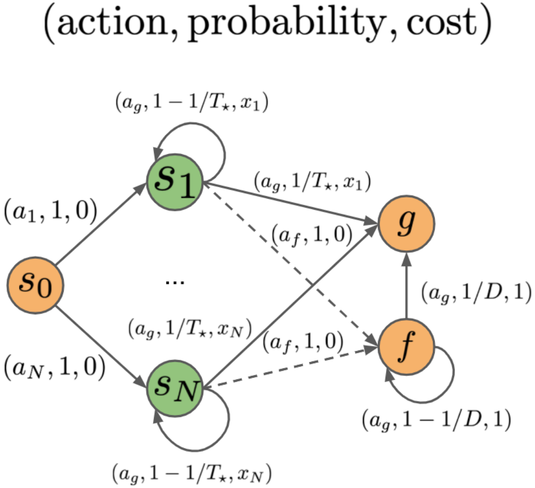
Full information: $\Omega(\sqrt{DT_\star K})$
Bandit feedback: $\Omega(\sqrt{DT_\star SAK})$
Stochastic cost (Cohen et al., 2020): $\Omega(D\sqrt{SAK})$



Lower Bound

Main idea: an analogy to an expert / MAB problem with loss scale $T_\star$ and total losses DK.

- Uniformly sample a good state $j^\star \in [N]$ and fixed throughout the K episodes
- In each episode:
  - $x_{j^\star} \sim \text{Bernoulli}(\frac{D}{2T_\star})$
  - $x_j \sim \text{Bernoulli}(\frac{D}{2T_\star} + \epsilon)$ for any $j \neq j^\star$



Full information: $\Omega(\sqrt{DT_\star K})$
Bandit feedback: $\Omega(\sqrt{DT_\star SAK})$
Stochastic cost (Cohen et al., 2020): $\Omega(D\sqrt{SAK})$

Our setting is harder due to the larger variance of costs (with $T_\star$ dependency).