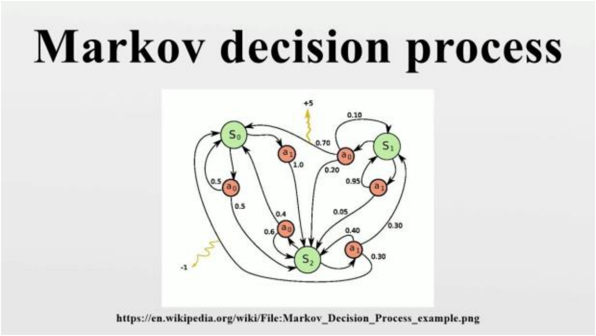
Online Learning in Stochastic Shortest Path

Yishay Mansour

Tel-Aviv University and Google

Joint works with: Alon Cohen, Haim Kaplan and Aviv Rosenberg

Reinforcement Learning



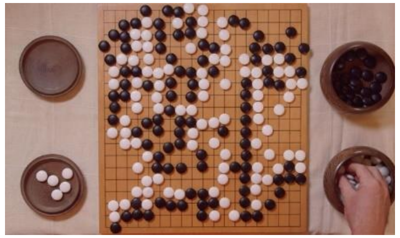
Stochastic Shortest Paths

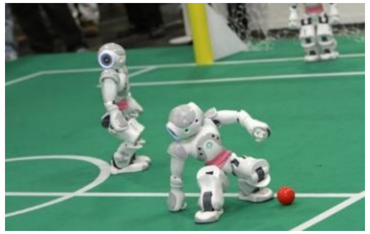
- Basic RL model
  - Episodic

- Dual objective
  - Reach goal state
  - Minimize cost

- Applications:
  - Games
  - Car navigation
  - Robotics

  - Any episodic task





SSP: Model

- MDP with goal state g
- Interaction ends when g is reached
- Dual objectives:
  - Reach goal state
  - Minimize total cost in the process (sum)
- Challenges:
  - The two objectives do not always agree.

SSP generalizes other models

Finite horizon

- Extend states adding the time in the episode
  - |S| H states
- Add a goal state g
- Result: loop-free SSP

Discounted

- Add a goal state g
- From every state s:
  - With probability γ move to the goal state g
- Expected return
  - Exactly the discounted expected return
  - Probability to reach any state s after t steps is γ^t

Online learning SSP: Model

- K episodes
- Have to reach goal state in every episode
- Transition function and cost unknown
- Minimize the regret

- Challenge:
  - A single episode can potentially have infinite cost!
  - Number of time steps of online and opt can be very different

Online learning SSP: Regret

- Fix an optimal policy π^*
- Consider online cost in K episodes
- The expected difference is the regret:

$$E\left[\sum_{i=1}^{K}\sum_{j=1}^{I_k} cost\left(s_j^i, a_j^i\right)\right] - K\, E[cost(\pi^*)]$$

SSP regret: Previous works

- Regret minimization finite horizon MDP:
  - UCRL and variants $\Theta(\sqrt{K})$
    - Note that the regret is always bounded by K
    - Many SSP loop-free works
      - Finite horizon
- Regret minimization SSP:
  - Tarbouriech et al. (ICML 2020)
    - Regret bound $\tilde{O}(K^{2/3})$

SSP regret: our works

Stochastic MDP (ICML 2020)

- Upper bound
  - $\tilde{O}(\sqrt{K})$
- Lower bound:
  - $\Omega(B_* \sqrt{|S||A|K})$

Adversarial MDP (Submitted)

- Upper bound
  - $\tilde{O}(K^{0.75})$

Planning in SSPs (Bertsekas and Tsitsiklis, 1991)

- Proper policy: reaches the goal state from <u>**any**</u> state!

- Assumption:
  - There is a proper policy
  - Any improper policy has infinite cost

- The optimal policy is
  - stationary
  - deterministic
  - proper
  - Can be computed efficiently
    - E.g., Value Iteration.

Making policies proper: $c_{min} > 0$

- Assume strictly positive costs:

$$cost(s,a) \geq c_{min} > 0$$

  - Any improper policy has infinite cost
    - From some state
  - Optimal policy is proper
- Bounded Regret implies:
  - Guarantee that we reach the goal state!

From positive costs to general costs

- Add an ϵ perturbation (bias) to the costs
  - $\text{cost}'(\text{s}, \text{a}) = \max\{cost(s, a), \epsilon\}$

- Perturbation adds a bias:
  - increases the total cost by ϵ per step
  - Optimize later over ϵ to minimize regret

SSP regret: positive costs

Stochastic MDP

- Our upper bound

$$\tilde{O}\left(\sqrt{K} + \frac{1}{\sqrt{c_{min}}}\right) \rightarrow \tilde{O}(\sqrt{K})$$

- Tarbouriech et al.

$$\tilde{O}\left(\sqrt{\frac{K}{c_{min}}}\right) \rightarrow \tilde{O}(K^{2/3})$$

Adversarial MDP

- Upper bound

$$\tilde{O}\left(\frac{\sqrt{K}}{c_{min}}\right) \rightarrow \tilde{O}(K^{0.75})$$

SSP algorithm

- Overview:
  - Keep confidence set for the transitions
    - Similar to UCRL2
    - Assume (w.l.o.g. and for simplicity) that costs are known
  - Compute an optimal optimistic policy
    - When should we re-compute?
  - Keep states known/unknown
    - When all states are known, we have a good model.

SSP algorithm

- Challenge:
  - We cannot allow one policy to run until an episode is completed.
    - It might never complete!
  - This implies that we need to re-compute policies during an episode.

SSP our algorithms

- Simpler
  - Uses Hoeffding bounds
  - Regret matches Tarbouriech et al.
- Re-compute each time you reach an unknown state.

- Advanced
  - Uses Berenstein bounds
  - Gets the improved regret
- Re-compute when the number of visits to some state-action doubles.
  - Similar to UCRL2

Regret Analysis

- Observations:
  - Let B_* be the cost of the optimal policy
    - from the worse state
  - If each state-action visited $M = \Omega\left(\frac{B_*|S|}{c_{min}}\right)$ then:
    - optimal optimistic policy is proper (w.h.p.), its expected cost $O(B_*)$
  - If policy expected cost is $O(B_*)$ then w.h.p it is $O(B_* \log\frac{1}{\delta})$

Regret Analysis

- A state-action is unknown if visited less than $M = \Omega\left(\frac{B_*|S|}{c_{min}}\right)$ times.
- Consider intervals which restart at the end of episode or when we reach an unknown state-action.
- Number of intervals: $\mathrm{I} = K + \tilde{O}\left(\frac{B_*^2|S|^2|A|}{c_{min}}\right)$
- Cost of an interval: $\tilde{O}(B_*)$ w.h.p.

Regret analysis: bounds

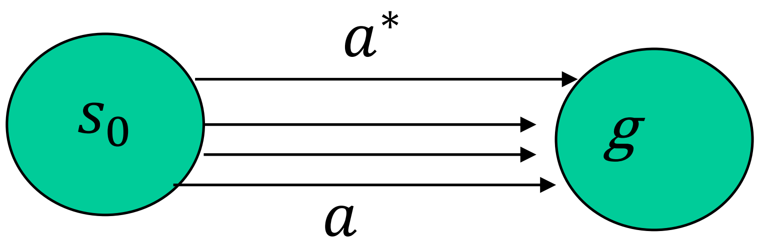
- Using Berenstein: for each interval, variance is $\tilde{O}(B_*^2)$
- Regret scales with the square-root of total variance
  - $REGRET = \tilde{O}\left(B_*|S|\sqrt{A\mathrm{I}}\right) = \tilde{O}\left(B_*|S|\sqrt{AK} + B_*^{1.5}|S|^2\,|A|c_{min}^{-1}\right)$
    - main term optimal up to $\sqrt{|S|}$ factor
- General bound:
  - $REGRET = \tilde{O}\left(B_*^{1.5}|S|\sqrt{AK} + T_*^{1.5}|S|^2\,|A|\right)$
    - T^* the time of the optimal policy

Hoeffding versus Berenstein bound

- Hoeffding bound:
  - Variance per step $\tilde{O}(B_*^2)$
  - Regret is $\tilde{O}\left(B_*\sqrt{T}\right) = \tilde{O}\left(B_*\sqrt{\frac{B_* I}{c_{min}}}\right)$
- Berenstein bound:
  - Variance per episode $\tilde{O}(B_*^2)$
  - Regret is $\tilde{O}\left(B_*\sqrt{I}\right)$

Lower bound

- Yao's principle:
  - Distribution over MDPs
  - Lower bound on regret



- Two states:
- Costs always 1.
- Transitions:
  - $\Pr[g|a^*] = \frac{1}{B_*}$
  - $\Pr[g|a] = \frac{1-\epsilon}{B_*}$
- Optimal policy cost B_*
- Any other action cost $\frac{B_*}{1-\epsilon}$

Lower bound

- Similar in spirit MAB
  - Some technical challenge
- Expected Regret:
  - $\epsilon K B_* \left(\frac{1}{8} - 2\epsilon\sqrt{\frac{2K}{|A|}}\right)$
- MDP:
  - Take |S| such "gadgets"
  - Initial distribution is uniform
  - Visit per gadget $K/|S|$
  - Set $\epsilon = 0.01\sqrt{|A||S|/K}$
    - $\epsilon K B_* = \Omega(B_* \sqrt{|A||S|/K})$
  - Lower bound!

Adversarial SSP

- Model:
  - Fixed unknown transition function
  - Costs change every step.
    - Observed at the end of an episode

- Algorithm:
  - Online Mirror Descent (OMD)
    - selects an occupancy measure
  - Maintains confidence set over transition probabilities
  - Bound the duration to reach goal
    - Bounds the loss in an episode

Summary

- Stochastic Shortest Paths
  - Stochastic model
  - Near optimal bound
  - Adversarial model
    - More work is needed!

Thank you