Lecture 17: Stochastic Shortest Path Problems

Course: Reinforcement Learning Theory
Instructor: Lei Ying
Department of EECS
University of Michigan, Ann Arbor

Stochastic shortest path problem

$$\min \sum_{k=0}^{\infty} E[c(x_k, u_k) | x_0].$$

- Let us assume the states are $\{1, 2, \ldots, N\}$
- Also, there is a distinct state called "0" such that
  - (a) if $x_k = 0$, then $x_l = 0 \quad \forall l \geq k$
  - (b) $c(x_k, u_k) = 0$ if $x_k = 0$
- Hence, as the name SSP suggests: find the cheapest path to "0"
- c is assumed to be deterministic, but works for random c as well.

Stochastic shortest path problem

$$\max \sum_{k=0}^{\infty} E[r(x_k, u_k) | x_0],$$

where $r(x_k, u_k) = -c(x_k, u_k)$.

- Bellman Equation:

$$J(i) = \underbrace{\max_u r(i, u) + \sum_{j \neq i} P_{ij}(u) J(j)}_{T(J)}$$

  with $J(0) = 0$

- We will prove that T is a contraction mapping in a weighted ∞-norm. Many of the earlier results for Q-learning continue to hold with the ∞-norm replaced by the weighted ∞-norm.

Stochastic shortest path problem

Fixed policy: consider a fixed policy μ. Then,

$$J_\mu(i) = r(i, \mu(i)) + \sum P_{ij}(\mu(i)) J_\mu(j) = r(i) + \sum P_{ij} J_\mu(j),$$

we dropped the dependence on u since the policy is fixed. We want to show there exist $w_i > 0, \beta \in [0, 1)$ such that

$$\max_i w_i |T_\mu Y_1(i) - T_\mu Y_2(i)| \leq \beta \max_i w_i |Y_1(i) - Y_2(i)|$$

$$\max_i w_i \left| r(i) + \sum_j P_{ij} Y_1(j) - r(i) - \sum_j P_{ij} Y_2(j) \right| \leq \beta \max_i w_i |Y_1(i) - Y_2(i)|$$

$$\Longleftrightarrow$$

$$\max_i w_i \left| \sum_j P_{ij} (Y_1(j) - Y_2(j)) \right| \leq \beta \max_i w_i |Y_1(i) - Y_2(i)|$$

Stochastic shortest path problem

Start from the LHS:

$$\max_i w_i \left| \sum_j \frac{P_{ij}}{w_j} w_j (Y_1(j) - Y_2(j)) \right|$$

$$\leq \max_i w_i \sum_j \frac{P_{ij}}{w_j} \underbrace{\max_j w_j |J_1(j) - J_2(j)|}_{\|Y_1 - Y_2\|_{\infty,w}}$$

$$= \boxed{\max_i w_i \sum_j \frac{P_{ij}}{w_j}} \|Y_1 - Y_2\|_{\infty,w}$$

- We want to show that the quantity in red box is $\leq \beta$ for an appropriate choice of β.

Stochastic shortest path problem

- We want

$$\sum_j \frac{P_{ij}}{w_j} \leq \frac{\beta}{w_i} \quad \forall i$$

- Consider an SSP where all rewards are 1 expect in state 0, where the reward is 0. Then the Bellman equation for this problem is

$$\begin{aligned} J(i) &= 1 + \max_u \sum_j P_{ij}(u) J(j) \\ &\geq 1 + \sum_j P_{ij}(\mu(i)) J(j) \quad \forall \mu \end{aligned}$$

$$\forall j \neq 0, J(j) \geq 1$$

Stochastic shortest path problem

$$\sum_j P_{ij}(\mu(i))J(j) \leq J(i) - 1 \leq J(i)\frac{J(i)-1}{J(i)}$$

- Define

$$\beta = \max_i \frac{J(i)-1}{J(i)} < 1$$

Then

$$\sum_j P_{ij}(\mu(i))J(j) \leq \beta J(i) \implies \sum_j \frac{P_{ij}(\mu(i))}{w_j} \leq \frac{\beta}{w_i}$$

Stochastic shortest path problem

- Assumption: μ is such that there is a finite time t such that

$$\max_i P(x_n \neq 0 | x_0 = i) < 1$$

We used this assumption to ensure that $J_\mu(i) < \infty \quad \forall i$. To reason about the optimal policy, we also need

- Assumption: All stationary policies are proper, i.e. satisfying the previous assumption.
- We showed that

$$w_i \left| T_\mu(Y_1)(i) - T_\mu(Y_2)(i) \right| \leq \beta \max_j w_j \left| Y_1(j) - Y_2(j) \right|$$

$$T(Y_1)(i) \leq \max_\mu T_\mu(Y_1)(i) \leq \max_\mu T_\mu(Y_2)(i) + \beta \frac{1}{w_i} \max_j w_j |Y_1(j) - Y_2(j)|$$

Stochastic shortest path problem

Thus,

$$T(Y_1)(i) \leq T(Y_2)(i) + \beta \frac{1}{w_i} \max_j w_j |Y_1(j) - Y_2(j)|$$

$$w_i \left(T(Y_1)(i) - T(Y_2)(i) \right) \leq \beta \max_j w_j |Y_1(j) - Y_2(j)|$$

Switching the roles of Y_1 and Y_2, we have

$$w_i \left(T(Y_2)(i) - T(Y_1)(i) \right) \leq \beta \max_j w_j |Y_1(j) - Y_2(j)|$$

$$w_i \left| T(Y_1)(i) - T(Y_2)(i) \right| \leq \beta \max_j w_j |Y_1(j) - Y_2(j)|$$

$$\max_i w_i \left| T(Y_1)(i) - T(Y_2)(i) \right| \leq \beta \max_j w_j |Y_1(j) - Y_2(j)|$$

Therefore we conclude that, T is a contraction mapping.

Reference

- This lecture is based on R. Srikant's lecture notes on *Stochastic Shortest Path Problems* available at `https://sites.google.com/illinois.edu/mdps-and-rl/lectures?authuser=1`

Acknowledgements: I would like to thank Alex Zhao for helping prepare the slides, and Honghao Wei and Zixian Yang for correcting typos/mistakes.