ECE-GY 9223 Reinforcement Learning Spring 2019

Lecture 1 — 8th, Feb, 2019

Prof. Quanyan Zhu *Scribe: Guanze Peng*

1 Overview

In the last lecture, we are introduced the background of reinforcement learning.

In this lecture, we learn the dynamic programming for the optimizing the infinite-horizon discounted problem. We give the basic proof of the convergence of the algorithm.

2 Linear Vector Space

Cauchy Sequence Let X be a metric space, and let $\{x_n\}$ be a sequence of points in X. We say that $\{x_n\}$ is a Cauchy sequence if for every $\epsilon > 0$, there exists a $N \in \mathbb{N}$ so that $\forall i, j > N$, $d(x_i, x_j) < \epsilon$.

Comlepeteness of Complete Space A normed linear space X is said to be complete if every Cauchy sequence on X has a limit, and the limit is in X.

Banach Space Banach Space is a complete normed linear vector space.

3 Infinite-Horizon Discounted Problems

We assume the state is finite, $i \in I = \{1, 2, ..., n\}$. Let $u \in U$ be the control and $\alpha \in [0, 1)$ be the discounting factor (Note that when $\alpha = 1$, the problem becomes the shortest path problem). $g\left(i, u, j\right)$ stands for the incurred cost under control u when the system transits from state i to j. Introduce $P_{ij}(u)$ to be the probability of transition from state i to j under control u. Thus, the cost function under policy μ and initial state i is given by

$$J_\mu(i) = \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{i} g\left(i_k, u_k, i_{k+1}\right) | i_0 = i\right].$$

where the policy $\mu : I \to U$, i.e., $u_k = \mu(i_k)$. The optimal cost function is given by

$$J^*(i) = \min_\mu J_\mu u(i).$$

In order to make the problem well defined, we need $g(i, u, j)$ be a bounded function for all i, u and j.

Theorem 3.1 The optimal cost J^* satisfies the equation

$$J^*(i) = \min_u \mathbb{E}\left[g(i,u,j) + \alpha J^*(j)\right] = \min_u \sum_{j=1}^n P_{ij}(u)\left(g(i,u,j) + \alpha J^*(j)\right), \quad \forall i. \tag{1}$$

Proof To give the equality in (1), we prove "$\geq$" first and then "$\leq$".

"$\geq$": Let μ be an arbitrary policy. Under this policy, the system produces action u at $t = 0$.

$$\begin{aligned} J_\mu(i) &= \sum_{j=1}^n P_{ij}(u)\left(g(i,u,j) + \alpha \tilde{J}(j)\right), \\ \tilde{J}(i) &\geq J^*(i) \\ &\geq \sum_{j=1}^n P_{ij}(u)\left(g(i,u,j) + \alpha J^*(j)\right) \\ &\geq \min_u \sum_{j=1}^n P_{ij}(u)\left(g(i,u,j) + \alpha J^*(j)\right). \end{aligned}$$

Pick $\mu = \mu^*$, , which is the optimal policy, then

$$J_{\mu^*}(i) = J^*(i) \geq \sum_{j=1}^n \min_u P_{ij}(u)\left(g(i,u,j) + \alpha J^*(j)\right).$$

"$\leq$": Suppose μ_0 is the optimal policy solve (1). Let μ_0 produce u_0 at time $t = 0$. If the next state is j, use a new policy μ_j, satisfying,

$$J_{\mu_j}(j) \leq J^*(j) + \varepsilon.$$

Under the constructed policy,

$$\begin{aligned} J_\mu(i) &= \sum_{j=1}^n P_{ij}(u_0)\left(g(i,u_0,j) + \alpha \tilde{J}_{\mu_j}(j)\right) \\ &\leq \sum_{j=1}^n P_{ij}(u_0)\left(g(i,u_0,j) + \alpha J^*(j) + \alpha\epsilon\right), \quad \forall u_0. \end{aligned}$$

We then have

$$J^*(j) \leq J_\mu(j) \leq \min_{u_0} \sum_{j=1}^n P_{ij}(u_0)\left(g(i,u_0,j) + \alpha J^*(j) + \alpha\epsilon\right)$$

Define $\epsilon' > 0$, satisfying

$$J^*(j) \leq J_\mu(j) - \epsilon'.$$

Pick ϵ so that

$$J^*(j) \leq J_\mu(j) - \epsilon' \leq J^*(i) - \alpha\epsilon \leq \min_{u_0} \sum_{j=1}^n P_{ij}(u_0)\left(g(i,u_0,j) + \alpha J^*(j)\right).$$

□

Definition Let S be a subset of a normed space X and let T be a transformation mapping from S to S. Then T is said to be a contraction mapping, if there exists an $\alpha \in (0,1)$, such that

$$\|T(x_1) - T(x_2)\| \leq \alpha \|x_1 - x_2\|, \quad \forall x_1, x_2 \in S.$$

Before we give the most important theorem of this lecture, we introduce two operators T and T_μ on the cost function vector $J = [J(1), ..., J(n)]'$.

$$(TJ)(i) = TJ(i) = \min_{u \in U} \sum_{j=1}^{n} P_{ij}(u) \left(g(i,u,j) + \alpha J(j) \right).$$

- Take arbitrary J and T produces the optimal cost-to-go.
- $T : B(I) \to B(I)$, where $B(I)$ is the space of all the bounded functions with domain of non-negative integers.

$$T_\mu J(i) = \sum_{j=1}^{n} P_{ij}(\mu(i)) \left(g(i,\mu(i),j) + \alpha J(j) \right).$$

- T_μ produces cost-to-go under policy μ.
- $T : B(I) \to B(I)$.
- $T_\mu J = g_\mu + \alpha P_\mu J$.

Given a policy μ, evaluate the policy.

$$J_\mu(i) = \lim_{N \to \infty} \mathbb{E} \left[\sum_{k=0}^{\infty} \alpha^k g(i_k, \mu_k, i_{k+1}) | i_0 = i \right].$$

(1) R.h.s. is well defined for $i = 1, ..., n$.

(2)

$$\begin{aligned} J_\mu(i) &= \lim_{N \to \infty} \mathbb{E} \left[g(i,\mu(i),j) + \sum_{k=1}^{N} \alpha^k g(i_k, \mu_k, i_{k+1}) | i_1 = j \right] \\ &= g_\mu + \alpha \sum_{j=1}^{N} P_{ij}(\mu(i)) J_\mu(j). \end{aligned}$$

Then $J_\mu = T_\mu J_\mu$ is to evaluate the performance of a policy μ.

Theorem 3.2 There exists a unique $\bar{J}_\mu$ which solves $T_\mu = T_\mu J_\mu$.

Proof

$$J_\mu = g_\mu + \alpha P_\mu J_\mu, \quad (I - \alpha P_\mu) J_\mu = g_\mu.$$

Since $(I - \alpha P_\mu)$ is non-singular, then $J_\mu = (I - \alpha P_\mu)^{-1} g_\mu$.

Theorem 3.3 (Contraction Mapping Theorem)

(1) If T is a contraction mapping on a closed subset of a Banach space, there is a unique $x_0 \in S$ satistying $x_0 = T(x_0)$.

(2) x_0 can be obtained by the method of successive approximation $x_{n+1} = T(x_n)$.

Proof

"Existence" Select an arbitrary $x_1 \in S$ and generate a sequence $\{x_n\}$ by $x_{n+1} = T(x_n)$. By contraction,

$$\|x_{n+1} - x_n\| = \|T(x_{n+1}) - T(x_n)\| \le \alpha \|x_{n+1} - x_n\|.$$

and

$$\begin{aligned} \|x_{n+p} - x_n\| &= \|x_{n+p} - x_{n+p-1} + x_{n+p-1} - ... + x_{n+1} - x_n\| \\ &\le \|x_{n+p} - x_{n+p-1}\| + ... + \|x_{n+1} - x_n\| \\ &\le \left(\alpha^{n+p-2} + ... + \alpha^{n-1}\right) \|x_2 - x_1\| \\ &\le \frac{\alpha^{n-1}}{1-\alpha} \|x_2 - x_1\|. \end{aligned}$$

Since $\{x_n\}$ is Cauchy sequence and S is closed subset of a complete space, there exists $x_0 \in S$ such that $\lim_{n \to \infty} x_n = x_0$.
Now we show that $x_0 = T(x_0)$.

$$\begin{aligned} \|x_0 - T(x_0)\| &= \|x_0 - x_n\| + \|x_n - T(x_0)\| \\ &\le \|x_0 - x_n\| + \|x_n - T(x_0)\| \\ &= \|x_0 - x_n\| + \|T(x_{n-1}) - T(x_0)\| \\ &\le \|x_0 - x_n\| + \alpha \|x_{n-1} - x_0\|. \end{aligned}$$

Let n go to infinity on both sides, we have $x_0 = T(x_0)$.

"Uniqueness" Suppose the solution is not unique and x_0, y_0 are both fixed points.

$$\|x_0 - y_0\| = \|T(x_0) - T(y_0)\| \le \alpha \|x_0 - y_0\|.$$

Apperantly, $\alpha = 0$ or 1. So, $x_0 = y_0$.

□