Lecture 19

Monday, October 26, 2020

CS 59000-RL

MDP

- stochastic approximation
- Q-learning
- Infinit horizon MDP (undiscounted)

For off policy vs on policy } check out
Q-learning SARSA } Rich, Andrews
Sutton Barto
An Intro to RL

Let's see when Q-learning may converge:
we use tools from stochastic approximation Theory
Robbins-Munro 1951, Dvoretzky 1956

Theorem (Jaakkola et al. 1993)

A random iterative process

$$\longrightarrow \Delta_{t+1}(x) = \left(1-\alpha_t(x)\right)\Delta_t(x) + \beta_t(x)\, F_t(x)$$

converges to zero a.s. if for all $x \in \Omega$

the following condition holds:

Monday, October 26, 2020

1) The Ω is finite.

2) $\sum_t \alpha_t(x) = \infty$, $\sum_t \alpha_t^2(x) < \infty$

$\sum_t \beta_t(x) = \infty$, $\sum_t \beta_t^2(x) < \infty$

$$E\left[\beta_t(x) \mid \mathcal{F}_t\right] \leq E\left[\alpha_t(x) \mid \mathcal{F}_t\right] \text{ a.s.}$$

3) $E\left[\underbrace{F_t(x)} \mid \mathcal{F}_t\right] \leq \gamma\, E\left[\|\Delta_t\| \mid \mathcal{F}_t\right] \quad 0 \leq \gamma < 1$

4) $\mathrm{Var}\left[F_t(x) \mid \mathcal{F}_t\right] \leq C\left(1 + \underbrace{E\left[\|\Delta_t\| \mid \mathcal{F}_t\right]}\right)^2$

for some constant C.

where $\{\Delta_t, \Delta_{t-1}, \ldots, F_{t-1}, F_{t-2}, \ldots, \alpha_{t-1}, \alpha_{t-2}, \ldots, \beta_{t-1}, \beta_{t-2}, \ldots\}$

is $\mathcal{F}_t$ measurable.

Theorem: The Q learning algorithm given by

$$Q_{t+1}(x,a) = (1-\alpha_t(x,a))\, Q_t(x,a) + \alpha_t(x,a)\left[r(x,a) + \gamma \max_{a'} Q_t(x',a')\right]$$

$r(x,a) \sim R(x,a)$
$x' \sim P(\cdot \mid x,a)$

converges to the optimal Q^* if:

1) The state and action spaces are finite.

2) $\sum_t \alpha_t(x,a) = \infty$ and $\sum_t \alpha_t^2(x,a) < \infty$ a.s.

3) $Var[r(x,a) \mid x, a]$ is bounded.

---

proof: Let set $\Delta_t(x,a) = Q_t(x,a) - Q^*(x,a)$

$$\rightarrow \quad F_t(x,a) = r_t(x,a) + \gamma \max_{a'} Q_t(x',a') - \underbrace{Q^*(x,a)}_{E\left[r(x,a) + \max_{a'} Q^*(x',a')\right]}$$

$$\beta_t = \alpha_t$$

- $\mathcal{X} \times \mathcal{A} = \Omega$ is finite. :)

  ↳ The first condition

- Since $\alpha_t = \beta_t \Rightarrow$ The second condition is satisfies. :)

- For the third one.

$$\max_a \left| E\left[F_t(x,a) \mid \mathcal{F}_t\right]\right| = \max_a \left| E\left[\cancel{r_t(x,a)} + \gamma \max_{a'} Q_t(x',a') - \cancel{r(x,a)} - \gamma \max_{a'} Q^*(x',a')\right]\right|$$

$$= \gamma \max_a \left| \sum_{x'} P(x' \mid x,a) \left(\underbrace{\max_{a'} Q_t(x',a')} - \underbrace{\max_{a'} Q^*(x',a')}\right)\right|$$

Monday, October 26, 2020

$$\leq \gamma \max_a \sum_{x'} \underbrace{p(x'|x,a)} \underbrace{\max_{a'} \left| Q_t(x',a') - Q^*(x',a') \right|}_{\|Q_t - Q^*\|_\infty}$$

$$\leq \gamma \, \| Q_t - Q^* \|_\infty = \gamma \, \| \Delta_t \|_\infty$$

we showed

$$\| E[F_t(\cdot,\cdot)] \mathcal{F}_t \|_\infty = \max_{x,a} \| E[F_t(x,a) | \mathcal{F}_t]$$

$$\leq \gamma \, \| \Delta_t \|_\infty$$

Now the force one:

$$\mathrm{Var}[F_t(x,a) | \mathcal{F}_t] =$$

$$E\Big[\big(\underbrace{r_t(x,a)} + \gamma \max_{a'} Q_t(x',a') - Q^*(x,a)$$

$$\underbrace{- \bar{r}(x,a)} - \gamma \sum_{x'} p(x'|x,a) \max_{a'} Q_t(x',a') + Q^*(x,a)\big)^2 \Big| \mathcal{F}_t\Big]$$

$$= E\left[\left(r_t(x,a) - \bar{r}(x,a)\right)^2\right]$$

$$+ \gamma^2 E\left[\left(\max_{a'} Q_t(x',a') - \sum_{x'} p(x'|x,a) \max_{a'} Q_t(x',a')\right)^2 \Big| \mathcal{F}_t\right]$$

Monday, October 26, 2020

$$\leq Var[r(x,a)] + \gamma^2 E\left[\left(\max_{a'} Q_t(x',a')\right)^2\right]$$

$$\leq Var[r(x,a) + \gamma^2 E\left[\max_{a'}(Q^*(x',a'))^2\right] + \gamma^2 \|\Delta_t\|_\infty^2$$

$$\leq \max_{x,a} Var[r(x,a)] + \gamma^2 E\left[\max_{a'}(Q^*(x',a'))^2\right] + \gamma^2 \|\Delta_t\|_\infty^2$$

$$\leq C\left(1 + \|\Delta_t\|_\infty^2\right)$$

which satisfies the fourth condition.

Infinite horizon undiscounted MDP

$$M : (X, A, P, P_1, R)$$

Expected return under a policy π

$$E^\pi\left[\sum_{t>0} r_t\right] \to \text{expected cumulative reward.}$$

$\sum_{t>0} r_t$ if the sum exist, might not be integrable - But some time we like.

For strongly connected MDPs, define long term average expected return of a memory-less π stating from:

$$\rho_x^\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} E^\pi\left[r(x_t, A_t) \mid X_1 = x\right]$$

(Assumption, the limit exists)

Monday, October 26, 2020

Define $\bar{\rho}^{\pi} = \max_{x} \rho^{\pi}_{x}$, since M is strongly connected and π is memoryless, for optimal policy

$$\rho^* = \max_{\pi \to \text{memoryless}} \bar{\rho}^{\pi}$$

$$\rho^* = \rho^{\pi^*}_{x} \quad \text{for } x \in \mathcal{X}$$

Back to matrix notation:

$$\rho^{\pi} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_{\pi}^{t-1} r_{\pi}$$

Define $\bar{P}_{\pi} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_{\pi}^{t-1}$ as the stationary transition matrix, $\Rightarrow \rho^{\pi} = \bar{P}_{\pi} r_{\pi}$

Starting from a state $X_1 = x$ does not change the average expected return, but has some gain:

Value function:

First define $V_{\pi}^{(T)} = \sum_{t=1}^{T} P_{\pi}^{t-1} (r_{\pi} - \rho^{\pi})$

The value function is the Cesaro sum of $P_{\pi}^{t}(r_{\pi} - \rho^{\pi})$

$$V_{\pi} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} V_{\pi}^{t}$$

Monday, October 26, 2020

$V_\pi(x) - V_\pi(x')$ represents long term advantage of starting from x v.s. x'.

Lemma: $V_\pi = \left((I - P_\pi - \bar{P}_\pi)^{-1} - \bar{P}_\pi \right) r_\pi$

proof: HW.

Monday, October 26, 2020

Monday, October 26, 2020

Monday, October 26, 2020

Monday, October 26, 2020

Monday, October 26, 2020

Monday, October 26, 2020