03/11/14

Review:

* Markov decision process

$$MDP = \{S, A, P(s'|s,a), R(s), \gamma\}$$

* Value functions

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, s_0 = s\right]$$

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \,\middle|\, s_0 = s, a_0 = a\right]$$

* Bellman optimality Equation.

$$V^*(s) = R(s) + \gamma \max_a \sum_{s'} P(s'|s,a)\, V^*(s')$$

* Algorithm: Value iteration ← for Computing $V^*(s)$ directly.

→ initialize $V_0(s) = 0 \quad \forall s$

→ iterate $V_{K+1}(s) = R(s) + \gamma \max_a \left[\sum_{s'} P(s'|s,a)\, V_K(s')\right]$

To prove: The algorithm Converges

→ $V^*(s)$ is a fixed point of iteration

→ There are no other fixed points.

→ The algorithm always reaches $V^*(s)$

Before going to the proof of Converges, we will prove a lemma

Lemma: for any functions $f(a)$ and $g(a)$

$$\left| \max_a f(a) - \max_a g(a) \right| \leq \max_a |f(a) - g(a)|$$

Proof of lemma:

$$f(a) - \max_{a'} g(a') \leq f(a) - g(a), \text{ for all } a$$

$$\max_a f(a) - \max_{a'} g(a') \leq \max_a \left(f(a) - g(a) \right), \text{ max over } a$$

$$\leq \max_a |f(a) - g(a)| \longrightarrow (1)$$

By symmetry (exchanging f & g)

$$\max_a g(a) - \max_a f(a) \leq \max_a |g(a) - f(a)| \longrightarrow (2)$$

($|g(a) - f(a)|$ is same as $|f(a) - g(a)|$)

The L.H.S of eq (1) or eq (2), must correspond to absolute value of $\max_a f(a) - \max_a g(a)$

Thus, we get

$$\left| \max_a f(a) - \max_a g(a) \right| \leq \max_a |f(a) - g(a)|$$

Using this lemma, we will prove

Theorem: Value iteration converges.

$$\Rightarrow \lim_{k \to \infty} [V_k(s)] \to V^*(s) \quad \text{for all states, } s$$

Proof: let $\Delta_k = \max_s \left| V_k(s) - V^*(s) \right|$

$$\Delta_{k+1} = \max_s \left| V_{k+1}(s) - V^*(s) \right|$$

$$= \max_s \left| \left[R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V_k(s') \right] - \left[R(s) + \gamma \max_a \sum_{s'} P(s'|s,a) V^*(s') \right] \right|$$

$$= \gamma \max_s \left| \underbrace{\max_a \sum_{s'} P(s'|s,a) V_k(s')}_{f(a)} - \underbrace{\max_a \sum_{s'} P(s'|s,a) V^*(s')}_{g(a)} \right|$$

using the lemma ↘

$$\leq \gamma \max_s \max_a \left| \sum_{s'} P(s'|s,a) \left[V_k(s') - V^*(s') \right] \right|$$

$$\leq \gamma \max_s \max_a \left| \left[\sum_{s'} P(s'|s,a) \right] \left| \max_{s''} \left| V_k(s'') - V^*(s'') \right| \right| \right|$$

($\sum_{s'} P(s'|s,a) = 1$)

$$= \gamma \max_s \max_a \left| \Delta_k \right|$$

$$= \gamma \Delta_k \qquad \left[\Delta_k \text{ does } \underline{\text{not}} \text{ depend on } a \text{ or } s \right]$$

Thus $\Delta_{k+1} \leq \gamma \Delta_k$, where $0 \leq \gamma < 1$

By induction, we can show

$$\boxed{\Delta_k \leq \gamma^k \Delta_0}$$

Thus $\Delta_k \longrightarrow 0$ as $k \longrightarrow \infty$

Assume all the rewards are bounded.

$$\Delta_0 = \max_s \left| V_0(s) - V^*(s) \right| = \max_s \left| V^*(s) \right| \quad [\because V_0(s) = 0 \ \forall s]$$

$$\leq \max_s |R(s)| \left[1 + \gamma + \gamma^2 + \cdots \right]$$

$$= \max_s |R(s)| \left(\frac{1}{1-\gamma} \right)$$

Thus $\boxed{\Delta_k \leq \frac{\gamma^k}{1-\gamma} \max_s |R(s)|} \rightarrow 0$ as $k \rightarrow \infty$

The above inequality suggests that the convergence rate
→ • depends on γ
→ more iterations are required as $\gamma \rightarrow 1$

FINAL COVERS UP TO HERE !!

REINFORCEMENT LEARNING (not for final)

* what if $P(s'|s,a)$ and $R(s)$ are not known?
Can we learn $\pi^*(s)$ or $V^*(s)$ from experience?

1) Model-based (indirect) approach

Explore world, estimate model $\hat{P}(s'|s,a) \approx P(s'|s,a)$, (true MDP)

Compute $\hat{\pi}^*(s)$ or $v^*(s)$ from $\hat{P}(s'|s,a)$

* Cons: → to store $P(s'|s,a)$ is $O(n^2)$ for n. states

→ only care about $\pi^*(s)$ or $v^*(s)$ which are $O(n)$

→ Is it really necessary to estimate a model?

* Pros: → model $P(s'|s,a)$ useful for task transfer, whose rewards $R(s)$ or discount factor γ changes, but $P(s'|s,a)$ stay the same.

2) Direct Approach: learn $\pi^*(s)$, $v^*(s)$ w/o building a model

Stochastic Approximation Theory!

* To estimate mean of random variable, x from samples $x_0, x_1, x_2 \ldots x_{T-1}$?

1) obvious - sample average

$$\mu = \frac{1}{T}\left[x_0 + x_1 + \ldots + x_{T-1}\right] \rightarrow E[x] \text{ as } T \rightarrow \infty$$

Converges by law of large numbers

2) Incremental update:

→ Initialize $\mu_0 = 0$

→ update $\mu_t = (1-\alpha_t)\,\mu_{t-1} + \alpha_t x_t$ for $0 < \alpha_t < 1$

This can be also written as

$$\mu_t = \mu_{t-1} + \alpha_t (x_t - \mu_{t-1})$$

↳ temporal difference

↳ This algorithm is called Temporal Difference algorithm.

Theorem: $\mu \to E[x]$ as $t \to \infty$ if

i) α_t doesn't decay too fast → $\sum_{t=1}^{\infty} \alpha_t = \infty$

ii) α_t doesn't decay too slow → $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ is finite

One possible choice of α_t:

choose $\alpha_t = \frac{1}{t}$ satisfies the conditions.

This actually gives the same estimate as the running average.