Ex 11-1.

(a) Prediction = policy evaluation

Control = optimal ~~poly~~ policy approximation

(b)

(1) Model-based: model of evironment is known, ie. transition probability is known.

Melel-free: model is not needed.

(2) Both evaluate value function

(3) Model-based: DP.

Medel-free: MC, TD

(c) optimal stat value $u_*(s) = \max_a \sum_{s'} p(s'|s,a)[R(s') + \gamma u_*(s')]$

optimal action value $q_*(s,a) = \sum_{s'} p(s'|s,a)[R(s') + \gamma \max_{a'} q_*(s',a')]$

$$u_*(s) = \max_a q_{\pi_*}(s,a)$$

(d) on-policy: evaluate / optimize ~~*~~ policy

- SARSA. (on-policy TD control)

off-policy: evaluate / optimize target policy. from samples generated by behavior policy

- Q-learning (off-policy TD control)
- MC / TD - prediction / control via importance samply.

(e) $$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

Q-learning: back up best Q-value

from state reached in the observed trasition

SARSA: : waits until an action is actually taken

back up the Q-value for that action

$$Q(S,A) \overset{\alpha}{\leftarrow} R + \gamma Q(S', A')$$

$$Q(S,A) \overset{\alpha}{\leftarrow} R + \gamma \max_{a' \in A} Q(S', a')$$

$$x \overset{\alpha}{\leftarrow} y \equiv x \leftarrow x + \alpha(y - x)$$

(f) Exploration: unknown enviroment, unknown state effect.
⇒ better future, discover new action

Large reward ⇔ ~~Like~~ prefer experience from past.

Exploitation: ~~try new, discover new actions ⇒ better future~~ obtain reward

(g) $$G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

$G_t^0 \Rightarrow TD(0) \Rightarrow TD$

$G_t^1 \Rightarrow$ ∞ MC

} backward

~~$\lambda = 0 \Rightarrow G_t^\lambda = \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$~~

Ex 11-2. (a) DP: $$U(s) \leftarrow \mathbb{E}[R + \gamma U(s') \mid s] = \sum_{s'} p(s'|s)[R + \gamma U(s')]$$

MC: $$U(s) \leftarrow U(s) + \alpha(G_t - U(s))$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

TD: $$U(s) \leftarrow U(s) + \alpha(R + \gamma U(s') - U(s))$$

(b) DP: + low variance
+ ~~high~~ efficient
− model based
− high bias

MC: − high variance
− low efficient
+ mode-free
+ no-bias

TD: + low variance
+ efficient
+ mode-free
− high bias