On Policy Control with Approximation

- now tackle control problem with action-value function $\hat{q}(s,a,w) \approx q_*(S,a)$, $w \in \mathbb{R}^d$
+ use SARSA, extension of TD(0)
- episodic is easy, continuing requires breakdown of current understanding

Episodic Semi-gradient Control

- action value weight update

$$w_{t+1} = w_t + \alpha\left[U_t - \hat{q}(S_t, A_t, w_t)\right] \nabla \hat{q}(S_t, A_t, w_t)$$

- One step SARSA

$$w_{t+1} \doteq w_t + \alpha\left[R_{t+1} + \gamma \hat{q}(S_{t+1}, A_{t+1}, w_t) - \hat{q}(S_t, A_t, w_t)\right] \nabla \hat{q}(S_t, A_t, w_t)$$

- how to select action?
+ for large discrete spaces or continuous, no clear answer

+ Small spaces, use previous methods

$$A^*_{t+1} = \underset{a}{\operatorname{argmax}}\ \hat{q}(S_{t+1}, a_t, w_t)$$

EXAMPLE 10.1: Mountain Car Task

- difficult continuous task, must go back before climbing hill
- rewards are -1 until goal achieved
- actions +1, 0, -1 for throttle
- movement is simple physics

$$x_{t+1} \doteq \text{bound}\left[x_t + \dot{x}_{t+1}\right]$$

$$\dot{x}_{t+1} \doteq \text{bound}\left[\dot{x}_t + 0.001 A_t - 0.0025 \cos(3x_t)\right]$$

- bound $-1.2 \leq x_{t+1} \leq 0.5$, $-0.07 \leq \dot{x}_{t+1} \leq 0.07$
- if x_{t+1} reaches left bound, $\dot{x}_{t+1}$ resets to 0.

right bound, goal reached

- start $x_t \in [-0.6, -0.4)$ $\dot{x}_t = 0$

- convert $x_t, \dot{x}_t$ from continuous to binary, use 8 tilings for $\frac{1}{8}$ the bounded distance, offset.
- $\hat{q}(s,a,w) = w^T x(s,a) = \sum_{i=1}^{d} w_i \cdot x_i(s,a)$

Semi-gradient n-step Sarsa

- trivial extension of tabular form

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{q}(S_{t+n}, A_{t+n}, w_{t+n-1})$$

for $t+n < T$. $G_{t:t+n} = G_t$ $t+n \geq T$

$$w_{t+n} \doteq w_{t+n-1} + \alpha\left[G_{t:t+n} - \hat{q}(S_t, A_t, w_{t+n-1})\right] \nabla \hat{q}(S_t, A_t, w_{t+n-1})$$

- intermediate n = best

EXERCISE 10.1. No all because it uses full return, which means perfect gradient, not semi-gradient.

Extremely slow, require full episode, poor on Mountain Car

10.2: Expected Sarsa would just use the expected Return $R_t + \gamma \mathbb{E}[\hat{q}(s,a,w) \mid s = S_{t+n}, a = A_{t+n}, w = w_{t+n-1})]$

10.3: large n means current return vs actual return tends to be more disparate

Average Reward: A New Problem Setting for Continuing Tasks

- Immediate reward just as imp. so can't discount
- average reward is used, because discounting struggles in func. approx

- evaluate policy π with rate of avg rate of reward, or just avg reward $r(\pi)$

$$r(\pi) \overset{(1)}{=} \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_t \mid S_0, A_{0:t-1} \sim \pi]$$

$$\overset{(2)}{=} \lim_{t \to \infty} \mathbb{E}[R_t \mid S_0, A_{0:t-1} \sim \pi]$$

$$\overset{(3)}{=} \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_{s',r} p(s', r \mid s, a)\, r$$

- (2) & (3) only hold for steady state distributions $\mu_\pi(s) = \lim_{t \to \infty} \Pr\{S_t = s \mid A_{0:t-1} \sim \pi\}$ and independent of S_0
+ MDP $\boxed{\text{ergodic}}$ $\Rightarrow$ starting state + early decisions have only temporary effect
- returns are now defined as:

$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots$$

+ $\boxed{\text{differential return}}$ $\rightarrow$ $\boxed{\text{differential value functions}}$

× all value and error equation are the same except $r - r(\pi)$
all equations have some offset, converge to differential values with offset

10.4. Q learning is just

$$\delta = R - \bar{R} + \max_a \hat{q}(s', a, w) - \hat{q}(s, A, w)$$

EXERCISE 10.6

$$\bar{r} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} R_t = \frac{1+0}{2} = 0.5$$

$$v_\pi(s) = \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \left(\mathbb{E}[R_{t+1} \mid S_0 = s] - \bar{r}\right)$$

$$R_{t+1} - \bar{r} = \pm 0.5$$

in A, $= +0.5$ at $t = 0, 2, 4, \dots$
-0.5 at $t = 1, 3, 5, \dots$

$$= \lim_{\gamma \to 1} \sum_{t=0}^{\infty} \gamma^t (R_{t+1} - \bar{r})$$

$$= \lim_{\gamma \to 1} \left[\gamma^0(0.5) + \gamma^1(-0.5) + \gamma^2(0.5) \dots\right]$$

$$= 0.5 \left(\lim_{\gamma \to 1} \sum_{t=0}^{\infty} \gamma^t (-1)^t\right) = 0.5\left(\frac{1}{1+\gamma}\right)$$

alternating geometric series $= \frac{0.5}{1+\gamma}$

for A $\Rightarrow \frac{0.5}{1+1} = 0.25$

B $\Rightarrow -0.25$

EXERCISE 10.7 $A \xrightarrow{+0} B \xrightarrow{+0} C \xrightarrow{+1} A \dots$

$$\bar{r} = \frac{1+0+0}{3} = \frac{1}{3}$$

in A $R_{t+1} - \bar{r} = -\frac{1}{3}$

B $R_{t+1} - \bar{r} = -\frac{1}{3}$

C. $R_{t+1} - \bar{r} = \frac{2}{3}$

A:

$$\lim_{\gamma \to 1} \left[\gamma^0\left(\tfrac{1}{3}\right) + \gamma^1\left(\tfrac{1}{3}\right) + \gamma^2\left(-\tfrac{2}{3}\right) + \dots\right]$$

$$\lim_{\gamma \to 1} \sum_{t=0}^{\infty} \left(\frac{1}{3}\gamma^{3t} + \frac{1}{3}\gamma^{3t+1} - \frac{2}{3}\gamma^{3t+2}\right)$$

I think they are all 0?

Deprecating the Discounted Setting

- no clear beginning or end
- no set states
- only distinguishing feat. is reward & actions
- discounting doesn't matter at all when infinite rewards are considered
+ average return $\frac{r(\pi)}{1-\gamma} \approx r(\pi)$ avg reward

× order doesn't matter

- optimizing for average discounted reward same as avg. undiscounted reward
+ can use γ to optimize learning as a solution param for faster learning

- function approximation ≠ GPI, can't guarantee policy improvement with policy eval when state values are approximated
+ local policy improvement guarantee non-existent & an ongoing problem

Differential Semi-gradient n-step Sarsa

- simple generalization

$$G_{t:t+n} \doteq R_{t+1} - \bar{R}_{t+n-1} + \dots + R_{t+n} - \bar{R}_{t+n-1} + \hat{q}(S_{t+n}, A_{t+n}, w_{t+n-1})$$

- error:

$$\delta_t \doteq G_{t:t+n} - \hat{q}(S_t, A_t, w)$$

Summary

- how do you use parametrized function approximation for control?
+ episodic is easy, but continuing case does not easily resolve
- use average reward per time step

+ policies cannot be represented by approximation
× use scalar $r(\pi)$ to rank policies
* create a differential value function, simple concept change to all old ones