On-Policy Control with Approximation

Sunday, June 21, 2020 6:02 PM

General Action-Value Gradient Descent:

$\vec{w}_{t+1} \equiv \vec{w}_t + \alpha\big[U_t - \hat{q}\big(S_t, A_t, \vec{w}_t\big)\big]\nabla\hat{q}\big(S_t, A_t, \vec{w}_t\big)$

Episodic Semi-Gradient Sarsa:

$\vec{w}_{t+1} \equiv \vec{w}_t + \alpha\big[R_{t+1} + \gamma\hat{q}\big(S_{t+1}, A_{t+1}, \vec{w}_t\big) - \hat{q}\big(S_t, A_t, \vec{w}_t\big)\big]\nabla\hat{q}\big(S_t, A_t, \vec{w}_t\big)$

The rest of the algorithm includes calculating $\hat{q}\big(S_{t+1}, a, \vec{w}_t\big)$ for each possible action and finding the greedy action, policy improvement occurs through setting the policy to a soft approximation of the greedy policy

This converges in the same way and with the same error bounds as TD(0)

Semi-Gradient n-Step Sarsa:

$G_{t:t+n} \equiv \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n \hat{q}\big(S_{t+n}, A_{t+n}, \vec{w}_{t+n-1}\big)$

$= R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{q}\big(S_{t+n}, A_{t+n}, \vec{w}_{t+n-1}\big)$

$\vec{w}_{t+1} \equiv \vec{w}_t + \alpha\big[G_{t:t+n} + \gamma\hat{q}\big(S_{t+1}, A_{t+1}, \vec{w}_t\big) - \hat{q}\big(S_t, A_t, \vec{w}_t\big)\big]\nabla\hat{q}\big(S_t, A_t, \vec{w}_t\big)$

Average Reward/Reward Rate:

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

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

Assumption for the Second and Third Forms:

$\mu_\pi(s) \equiv \lim_{t\to\infty} \Pr\{S_t = s \big| A_{0:t-1}\sim\pi\}$

The assumption includes this existing and being independent of S_0

This is equivalent to the MDP this represents being ergodic, meaning that the starting state and any early decision made by the agent can only have a temporary effect so in the long run the expectation of being in a state depends only on the policy and MDP transition probabilities

Technically ergodicity is sufficient but not necessary for the limit to exist

Note that this is the average reward obtained per time step when following policy with no discounting (there is no discounting in the average reward setting, which must be applied to continuing problems)

All policies that obtain the maximal value of $r(\pi)$ are considered optimal

Steady State Distribution:

$\mu_\pi(s') = \sum_s \mu_\pi(s) \sum_a \pi(a|s) p(s'|s, a)$

Differential Return:

$G_t \equiv \sum_{i=0}^{\infty} R_{t+i} - r(\pi) = R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots$

This is defining return based on the difference between the reward obtained and the average reward

Value functions defined using this definition of return are called differential value functions

To get differential Bellman equations remove all occurrences of the discount rate γ and replace all rewards with the difference between the reward and the average reward

Most algorithms transfer through simple substitution of the newly defined values

Differential TD Error:

$\delta_t \equiv R_{t+1} - \bar{R}_t + \hat{v}\big(S_{t+1}, \vec{w}_t\big) - \hat{v}\big(S_t, \vec{w}_t\big)$

$$\delta_t \equiv R_{t+1} - \bar{R}_t + \hat{q}(S_{t+1}, A_{t+1}, \vec{w}_t) - \hat{q}(S_t, A_t, \vec{w}_t)$$

Where $\bar{R}_t$ is an estimate at time t of the average reward $r(\pi)$

Discounting has no effect in function approximation methods since it affects the average reward only through a constant

Average Reward with Discounting:

$$r(\pi) = \frac{r(\pi)}{1-\gamma}$$

n-Step Differential Return

$$G_{t:t+n} \equiv \sum_{k=0}^{n-1} R_{t+k+1} + \hat{q}(S_{t+n}, A_{t+n}, \vec{w}_{t+n-1}) = R_{t+1} + R_{t+2} + \cdots + R_{t+n} + \hat{q}(S_{t+n}, A_{t+n}, \vec{w}_{t+n-1})$$

n-Step TD Differential Error:

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

Semi-gradient Sarsa can be used with this definition