Policy Gradient Methods

Thursday, June 25, 2020 2:45 PM

Parameterized Policy:

A policy dependent on a parameter vector (θ) instead of action-value estimates

General Parameterized Policy Gradient Ascent:

$\vec{\theta}_{t+1} = \vec{\theta}_t + \alpha \nabla J(\theta_t)$

Where $\nabla J(\theta)$ is a stochastic estimate whose expectation approximates the gradient of the performance measure with respect to the parameter vector $\vec{\theta}$

Policy Gradient Method:

A methods that uses gradient ascent to maximize a parameterized policy

The policy can be parameterized in any way that makes $\nabla \pi \left(a \middle| s, \vec{\theta}\right)$ exist and be finite for all states, actions, and parameter vectors

Actor Critic Method:

A method that learns approximations to both policy and value functions

Actor refers to the learned policy and critic refers to the learned value function (usually a state-value function)

Soft-max Policy Parameterization:

$\pi(a|s,\theta) \equiv \frac{e^{h(s,a,\theta)}}{\sum_b e^{h(s,b,\theta)}}$

This is generally called soft-max in action preferences

Note that the preferences themselves $\left(h(s,a,\theta)\right)$ can be arbitrary computed, including by a neural net

Parameterizing policies in this way allows the policy to approach a deterministic policy

Policy approximating methods can find stochastic optimal policies, which action-value methods cannot

Policy gradient methods have the advantage of the action probabilities changing smoothly instead of a potentially arbitrarily small change in action values making a new action the greedy action

Episodic Performance:

$J(\theta) \equiv v_{\pi_\theta}(s_0)$

Where v_{π_θ} is the true value function for π_θ, the policy determined by θ

In policy gradient methods, performance depends on both the action selections and distribution of states in which the selections are made (which states are reached), which are both affected by the policy parameter

Given a state the effect of the policy parameter on the action selected, and therefore the reward is straightforward to compute, but the effect of the policy on the state distribution is a function of the environment

Policy Gradient Theorem:

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi \left(a \middle| s, \vec{\theta}\right) = E_\pi \left[\sum_a q_\pi(S_t, a) \nabla \left(a \middle| S_t, \vec{\theta}\right) \right]$$

Where $\mu(s)$ is the on-policy distribution under π

This being only proportional in the episodic case doesn't matter since the constant gets absorbed in to the step-size parameter (learning rate)

The constant to make it an equality for the episodic case:

$$\sum_{s'} \eta(s')$$

This multiplies on the right outside of the sums

Including Baseline:

$$\nabla J\left(\vec{\theta}\right) \propto \sum_s \mu(s) \sum_a \left(q_\pi(s,a) - b(s)\right) \nabla \pi\left(a \middle| s, \vec{\theta}\right)$$

Where $b(s)$ may be any function, even a random variable, as long as it doesn't vary with a

This is valid since the sum can be moved inside the gradient and the sum over the conditional probabilities $\sum_a \pi\left(a \middle| s, \vec{\theta}\right)$ is 1

Definitions for the Continuing Case:

$v_\pi(s) \equiv E_\pi\left[G_t \middle| S_t = s\right]$

$q_\pi(s,a) \equiv E_\pi\left[G_t \middle| S_t = s, A_t = a\right]$

$$G_t \equiv \sum_i R_{t+1} - r(\pi)$$

Where G_t is the differential return as defined earlier and $r(\pi)$ is as defined below (here called average reward performance)

All-Actions Method:

$$\vec{\theta}_{t+1} \equiv \vec{\theta}_t + \alpha \sum_a \hat{q}(S_t, a, \vec{w}) \nabla \pi\left(a \middle| S_t, \vec{\theta}\right)$$

Where $\hat{q}$ is a learned approximation to q_π

REINFORCE:

$$\vec{\theta}_{t+1} \equiv \vec{\theta}_t + \alpha G_t \frac{\nabla \pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}{\pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}$$

In words this is that the increment is proportional to the return multiplied by the gradient of the probability of taking the action that was actually taken divided by the probability of taking the action that was actually taken

Recall that the gradient is in the direction in parameter space that most increases the probability of repeating action on future visits to state S_t

Note that this uses the full return so this is a MC method well defined for only the episodic case

In discounted situations the second term will be multiplied by γ^t

Alternate Form:

$$\vec{\theta}_{t+1} \equiv \vec{\theta}_t + \alpha G_t \nabla \ln\left(A_t \middle| S_t, \vec{\theta}_t\right)$$

With Baseline:

$$\vec{\theta}_{t+1} \equiv \vec{\theta}_t + \alpha \left(G_t - b\left(S_t\right)\right) \frac{\nabla \pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}{\pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}$$

This is a strict generalization since could be chosen to be 0

This can reduce the variance and result in faster learning

One common choice for the baseline is a state value estimate $\hat{v}(S_t, \vec{w})$ where $\vec{w}$ is a weight vector learned by any method

One-Step Actor-Critic:

$$\vec{\theta}_{t+1} \equiv \vec{\theta}_t + \alpha \delta_t \frac{\nabla \pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}{\pi\left(A_t \middle| S_t, \vec{\theta}\right)}$$

Where $\delta_t = R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \vec{w}\right) - \hat{v}\left(S_t, \vec{w}\right)$ is the one-step TD error

This is usually paired with semi-gradient TD(0) to form a complete algorithm

Alternate Form:

$$\vec{\theta}_{t+1} \equiv \vec{\theta}_t + \alpha \left(G_{t:t+1} - \hat{v}\left(S_t, \vec{w}\right)\right) \frac{\nabla \pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}{\pi\left(A_t \middle| S_t, \vec{\theta}_t\right)}$$

To generalize the algorithm to the n-step and λ-return methods just substitute $G_{t:t+n}$ or G_t^λ respectively

Average Reward Performance:

$$J(\theta) \equiv \sum_s \mu(s) \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) r = r(\pi)$$

Where $\mu(s) \equiv \lim_{t\to\infty} \Pr\{S_t = s | A_{0:t} \sim \pi\}$ is the steady state distribution, which is assumed to be independent of S_0 (an ergodicity assumption)

Gaussian (Normal) Probability Density:

$$p(x) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where μ is the mean and σ is the standard deviation

Gaussian Policy Parameterization:

$$\pi\left(a\middle|s,\vec{\theta}\right) \equiv \frac{1}{\sigma\left(s,\vec{\theta}\right)\sqrt{2\pi}} e^{-\frac{\left(a-\mu\left(s,\vec{\theta}\right)\right)^2}{2\sigma\left(s,\vec{\theta}\right)^2}}$$

Standard Form of Mean and Standard Deviation:

$$\mu\left(s,\vec{\theta}\right) \equiv \vec{\theta}_\mu^T \vec{x}_\mu(s)$$

$$\sigma\left(s,\vec{\theta}\right) \equiv e^{\vec{\theta}_\sigma^T \vec{x}_\sigma(s)}$$

In words this is making the mean a linear function and the standard deviation the exponential of a linear function (which takes care of the requirement that it is always positive)