pg_theory

July 26, 2021

1 Policy Gradient Theory

Goal: Learn a parameterized policy that can select actions without consulting a value function. We parameterize the policy according to

$$\pi(a|s,\theta) = \Pr\{A_t = a|S_t = s, \theta_t = \theta\}$$

where the above gives the probability of selecting an action given the state and policy parameter, θ. We are interested in learning the optimal policy parameter and this is based on the gradient of some stochastic estimate of a scalar performance measure $J(\theta)$. Thus leading to a gradient ascent in J given by

$$\theta_{t+1} = \theta_t + \alpha \nabla \mathbb{E}[J(\theta_t)]$$

1.1 The Policy Gradient Theorem

For the episodic case, the performance measure is the value of the start state of the episode.

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

where v_{π_θ} is the true value function for the policy under its current parameter configuration, we would therefore like to change θ such that the value of the performance measure increases. We can write this in terms of the distribution of the policy.

$$J(\theta) = \sum_a \pi(a|s) q(s,a)$$

We are interested in finding the gradient of this measure and adjusting θ to increase its value.

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_a \pi(a|s) q(s,a) = \sum_a \left[q(s,a) \nabla \pi(a|s) + \pi(a|s) \nabla q(s,a) \right]$$

Where we have expressed the state-value function in terms of a weighted average over the policy and then applied the product rule. Next, we express $q(s,a)$ in terms of the next expected next rewards and expected next state values.

$$= \sum_a \left[q(s,a)\nabla\pi(a|s) + \pi(a|s)\nabla \sum_{s',r} p(s',r|s,a)(r + v(s')) \right]$$

Pulling the gradient inside the summation we have

$$= \sum_a \left[q(s,a)\nabla\pi(a|s) + \pi(a|s) \sum_{s'} p(s'|s,a)\nabla v(s') \right]$$

we can again unroll the state-value function

$$= \sum_a \left[q(s,a)\nabla\pi(a|s) + \pi(a|s) \sum_{s'} p(s'|s,a) \sum_{a'} [q(s',a')\nabla\pi(a'|s') + \pi(a'|s') \sum_{s''} p(s''|s',a')\nabla v(s'')] \right]$$

we can recursively unroll yielding the following expression

$$\nabla v_{\pi_\theta}(s_0) = \sum_{x \in S} \sum_{k=0}^{\infty} \Pr(s \to x, k, \pi) \sum_a q(s,a)\nabla\pi(a|x)$$

where $\Pr(s \to x, k, \pi)$ is the probability of transitioning from state s to x in k steps under the policy π. Thus $\sum_{k=0}^{\infty} \Pr(s \to x, k, \pi)$ is simply the the number of timesteps spent, on average, in a state s in a sngle episode, $\eta(s)$. This allows us to write the previous equation in terms of the on-policy distribution in episodic tasks.

$$= \sum_{x \in S} \eta(s) \sum_a q(s,a)\nabla\pi(a|x)$$

'multiplying by one' yields

$$= \sum_{s'} \eta(s') \sum_s \frac{\eta(s)}{\sum_{s'} \eta(s')} \sum_a q(s,a)\nabla\pi(a|x)$$

which simplifies to

$$= \sum_{s'} \eta(s') \sum_s \mu(s) \sum_a q(s,a)\nabla\pi(a|x)$$

and finally

$$\nabla v_{\pi_\theta}(s_0) \propto \sum_s \mu(s) \sum_a q(s,a)\nabla\pi(a|x)$$

The powerful thing about this is that we now have an analytic expression for the gradient of performance with respect to the policy parameter that does not involve the derivative of the state distribution.

1.2 REINFORCE: Monte Carlo Policy Gradient

From the policy gradient theorem we have

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\theta)$$

where $\mu(s)$ is the state distribution when we follow the policy π. Thus we can write this as an expectation

$$\nabla J(\theta) \propto \mathbb{E}_\pi \left[\sum_a q_\pi(S_t,a) \nabla \pi(a|S_t,\theta) \right]$$

We seek an update at time t that involves just A_t. We do this by replacing a sum over the random variable's possible realisations with an expectation under the policy π and then sampling the expectation.

$$\nabla J(\theta) \propto \mathbb{E}_\pi \left[\sum_a \pi(a|S_t,\theta) q_\pi(S_t,a) \frac{\nabla \pi(a|S_t,\theta)}{\pi(a|S_t,\theta)} \right]$$

where we can now replace a with A_t that is sampled from the policy

$$\nabla J(\theta) \propto \mathbb{E}_\pi \left[q_\pi(S_t,A_t) \frac{\nabla \pi(A_t|S_t,\theta)}{\pi(A_t|S_t,\theta)} \right]$$

which reduces to

$$\nabla J(\theta) \propto \mathbb{E}_\pi \left[G_t \nabla \ln \pi(A_t|S_t,\theta) \right]$$

due to the fact that the expected reward G_t is simply the value of following the policy or, $\mathbb{E}[G_t|S_t,A_t] = q_\pi(S_t,A_t)$. Thus we have all we need for the gradient ascent update. The final expression can be sampled on each timestep that has an expectation proportional to the gradient. Thus the parameter updates are given by.

$$\theta_{t+1} = \theta_t + \alpha G_t \nabla \ln \pi(A_t|S_t,\theta)$$

1.3 Gaussian Policy: Parameterization for Continuous Actions

When the action space is continuous which is the case when actions correspond to weights invested in individual assets the policy gradient approach becomes particularly appealing. As a first look at policy gradients for continuous actions, we will consider REINFORCE with a Gaussian policy that is

$$\pi(a|s,\theta) = \frac{1}{\sigma(s,\theta_\sigma)\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\theta_\mu))^2}{2\sigma(s,\theta_\sigma)^2}\right)$$

we see that our policy is parameterised by two components θ_σ and θ_μ where these are used to determine the standard deviation and mean of the policy respectively. Choosing the action then corresponds to calculating the mean and standard deviation given a state and the current parameter configuration and sampling from the distribution. We are interested in the quantity $\ln \pi(A_t|S_t,\theta)$ which for the gaussian distribution is simply

$$\ln \pi(A_t|S_t,\theta) = -\frac{(a-\mu(s,\theta_\mu))^2}{2\sigma(s,\theta_\sigma)^2} - \ln \sigma(s,\theta_\sigma)\sqrt{2\pi}$$

We are free to approximate the mean and standard deviation any way we see fit. One suitable choice is to approximate the mean as a linear function and the standard deviation as the exponential of a linear function (since the scale parameter must be positive).

$$\mu(s,\theta_\mu) = \theta_\mu^T x_\mu(s)$$

and

$$\sigma(s,\theta_\sigma) = \exp\left(\theta_\sigma^T x_\sigma(s)\right)$$

mu: Gradiant Update Next we must determine the gradient update for the mean

$$\nabla_{\theta_\mu} \ln \pi(A_t|S_t,\theta) = \frac{\partial}{\partial \theta_\mu}\left(-\frac{(a-\theta_\mu^T x_\mu(s))^2}{2\exp\left(2\theta_\sigma^T x_\sigma(s)\right)} - \ln \exp\left(\theta_\sigma^T x_\sigma(s)\right))\sqrt{2\pi}\right)$$

$$= x_\mu(s)\frac{(a-\theta_\mu^T x_\mu(s))}{\exp\left(2\theta_\sigma^T x_\sigma(s)\right)}$$

thus for the parameter updates we have

$$\theta_{\mu,t+1} = \theta_{\mu,t} + \alpha G_t x_\mu(s)\frac{(a-\theta_\mu^T x_\mu(s))}{\exp\left(2\theta_\sigma^T x_\sigma(s)\right)}$$

sigma: Gradient Update Sigma can be scheduled or held fixed, alternatively, it may be learned.

$$\nabla_{\theta_\mu} \ln \pi(A_t|S_t,\theta) = \frac{\partial}{\partial \theta_\sigma}\left(-\frac{(a-\theta_\mu^T x_\mu(s))^2}{2\exp\left(2\theta_\sigma^T x_\sigma(s)\right)} - \ln \exp\left(\theta_\sigma^T x_\sigma(s)\right))\sqrt{2\pi}\right)$$

$$= x_\sigma(s)\frac{(a-\theta_\mu^T x_\mu(s))^2}{2\exp\left(2\theta_\sigma^T x_\sigma(s)\right)} - \frac{x_\sigma(s)\exp\left(\theta_\sigma^T x_\sigma(s)\right)}{\exp\left(\theta_\sigma^T x_\sigma(s)\right)}$$

$$= x_\sigma(s) \left(\frac{(a - \theta_\mu^T x_\mu(s))^2}{\exp\left(2\theta_\sigma^T x_\sigma(s)\right)} - 1 \right)$$

thus for the parameter updates we have

$$\theta_{\sigma,t+1} = \theta_{\sigma,t} + \alpha G_t x_\sigma(s) \left(\frac{(a - \theta_\mu^T x_\mu(s))^2}{\exp\left(2\theta_\sigma^T x_\sigma(s)\right)} - 1 \right)$$

1.4 Dirichlet Policy: Parameterization for Continuous Actions

Although a Gaussian policy enables us to sample continuous actions the distribution is unbounded thus requiring manual normalisation of investment weights. The Dirichlet distribution, being defined on an N-1 simplex where there is zero density outside of the simplex, is a natural choice for a situation where short selling is not allowed, where the policy is now defined as (where w now denotes portfolio weights and a is the concentration parameter of the Dirichlet distribution).

$$\pi(w|s,\theta) = \mathrm{Dir}(w, a(s,\theta)) = \frac{1}{\beta(a(s,\theta))} \prod_{n=1}^{N} w_n^{a_n(s,\theta)-1} = \frac{\Gamma\left(\sum_{n=1}^{N} a_n(s,\theta)\right)}{\prod_{n=1}^{N} \Gamma(a_n(s,\theta))} \prod_{n=1}^{N} w_n^{a_n(s,\theta)-1}$$

Note that $a = [a_1, a_2, \ldots, a_N]$ are the concentration parameters where $a_n > 0$ and we parameterise this in order to follow gradient ascent. Following similar steps as to the above we have

$$\ln \pi(A_t|S_t,\theta) = \ln \Gamma\left(\sum_{n=1}^{N} a_n(s,\theta)\right) - \sum_{n=1}^{N} \ln \Gamma(a_n(s,\theta)) + \sum_{n=1}^{N} (a_n(s,\theta) - 1) \ln w_n$$

we can start by considering the concentration parameters to be a linear function of the state and the learnable weights

$$a_n(s,\theta) = \theta^T x(s)$$

Next, we take the derivative

$$\nabla \ln \pi(A_t|S_t,\theta) = \left[\psi\left(\sum_{n=1}^{N} a_n(s,\theta)\right) - \sum_{n=1}^{N} \psi(a_n(s,\theta)) + \sum_{n=1}^{N} \ln w_n \right]^T x(s)$$

where ψ is the digamma function, which is the logarithmic derivative of the gamma function or

$$\psi(x) = \frac{d}{dx} \ln\left(\Gamma(x)\right) = \frac{\Gamma'(x)}{\Gamma(x)} \sim \ln x - \frac{1}{2x}$$

thus for the parameter updates we have

$$\theta_{t+1} = \theta_t + \alpha G_t \left[\psi\left(\sum_{n=1}^{N} a_n(s,\theta)\right) - \sum_{n=1}^{N} \psi(a_n(s,\theta)) + \sum_{n=1}^{N} \ln w_n \right]^T x(s)$$