CME241 Notes

Saakaar Bhatnagar

Policy Gradient Method

1 Summary

- Value function approximation does not help deal with large/continuous action spaces
- Obtaining a parametrized expression for the policy directly without necessary dealing with the value function can deal with this
- The PG algorithm enables this by replacing the policy improvement step by essentially a gradient ascent step
- The cost function maximized is the expected returns
- The gradients of parameter vector θ is calculated based on the above function
- Since the gradient contains an expression for the true action value-function, we replace it by an approximation
- These approximations might suffer from bias and variance issues
- We deal with variance issues by replacing Q^{π} by the advantage function A^{π}
- We deal with the bias problem by using the compatible function approximation theorem

2 Score Calculations

2.1 Discrete Action Space

Using a softmax prediction to calculate probabilities:

$$\pi(a, s; \theta) = \frac{\exp\{\theta^T \phi(s, a)\}}{\sum_b \exp\{\theta^T \phi(s, b)\}}$$

Therefore,

$$log\pi = \theta^T \phi(s, a) - log \sum_b \exp\{\theta^T \phi(s, b)\}$$

Differentiate by θ

$$\nabla_\theta log\pi(a, s; \theta) = \phi(a, s) - \sum_b \pi(s, b; \theta)\phi(s, b)$$

2.2 Continuous Action Space

In this case we set the mean from the linear span, and can use either a fixed or a parametrized variance

Therefore,

$$f(a) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-(a - \theta^T \phi(s))^2}{2\sigma^2} \right\}$$

Hence,

$$log\pi = -0.5\frac{(a - \theta^T \phi(s))^2}{\sigma^2} - log\sigma\sqrt{2\pi}$$

Differentiate w.r.t θ,

$$\nabla_\theta log\pi = \frac{a - \theta^T \phi(s)}{\sigma^2}\phi(s)$$

3 Compatible Function Approximation Theorem

Helps reduce the bias present in the updates for θ, by changing the objective function we minimize for w. It is based on meeting two conditions:

$$\nabla_\theta log\pi(a, s; \theta) = \nabla_w Q(a, s; w)$$

and that the function we minimize to obtain w is:

$$\epsilon = \int_S \rho^\pi(s) \int_A \pi(a, s; \theta)(Q^\pi - Q)^2.da.ds$$

Where Q is a function of w.If we differentiate the second condition by w,

$$0 = \int_S \rho^\pi(s) \int_A \pi(a, s; \theta)(Q^\pi - Q)\nabla_w Q.da.ds$$

Now if we substitute the first condition and rearrange:

$$\int_S \rho^\pi(s) \int_A \pi(a, s; \theta)(Q^\pi)\nabla_\theta log\pi.da.ds = \int_S \rho^\pi(s) \int_A \pi(a, s; \theta)(Q)\nabla_\theta log\pi.da.ds$$

But since

$$\pi\nabla_\theta log\pi = \nabla_\theta \pi$$

the LHS is equal to gradient of cost function required,or

$$\nabla_\theta J(\theta) = \int_S \rho^\pi(s) \int_A \nabla_\theta \pi(a, s; \theta)Q(s, a; w).da.ds$$

Therefore RHS is as well, and we have shown that using an approximation for Q will give us the exact same result as true policy gradient, removing the error due to bias.