Tópicos especiais em Inteligência Computacional II: Aprendizado por Reforço Exercício 4: Actor-Critic

Wouter Caarls
wouter@ele.puc-rio.br

May 9, 2019

Actor-critic is a policy search method that reduces variance by using a value function. For this exercise, you may choose one of three policy search methods that differ mostly in the way $Q^{\pi_\theta}(s,a)$ is estimated in the actor update:

REINFORCE uses the directly measured return R_t.

State-value actor critic uses the temporal difference error of an approximated state-value function $\hat{V}(s;\boldsymbol{w})$.

Compatible deterministic actor-critic uses an approximated action-value function $\hat{Q}(s,a) = A(s,a;\boldsymbol{w}) + V(s;\boldsymbol{v})$.

1 Exercise

Once again, the system is a simple underactuated pendulum, the transition function of which is given in the function `pendulum`. You can draw the current state (and action) of the system using `drawip`.

Exercise 1.1 Create two loops, an outer one over episodes and an inner one that simulates the pendulum state within an episode. Keep track of the total reward gathered per episode and plot this learning curve after every episode to see the progress. Use 100 time steps per episode, and the same quadratic rewards as before:

$$r = -5{s'}^2_{\text{pos}} - 0.1{s'}^2_{\text{vel}} - a^2$$

Exercise 1.2 Implement one of the three policy search methods. Use a Gaussian policy and the provided `gaussrbf` function to create the feature vectors. It may be helpful to limit the actor parameter vector θ to values between -3 and 3.

Exercise 1.3 Try to optimize the parameters: actor learning rate, critic learning rate, exploration rate, discount rate, basis functions per dimensions, and basis function support. Start from the `gaussrbf` defaults, $\alpha = 0.2$ and a very low actor learning rate, such as $\beta = 0.01$ or even 0.0001. Report the results of your optimization in tables and graphs.

Exercise 1.4 Characterize the algorithm in terms of learning speed, end performance (reward gained during the last few episodes) and sensitivity to learning parameters. Note that the end performance should be measured without exploration. One way to do this gracefully is to slowly decay the exploration rate, by multiplying it by a value $1 - \nu$ every episode. Support your analysis with quantitative results.