Reinforcement Learning — Homework 2

Chapters 9, 10, 13 Sutton & Barto 2nd edition + extra literature

Deadline: December 3rd, 2018.

1 Instructions

This is the first assignment for Reinforcement Learning. This assignment covers a refresher of earlier material and some RL basics. Please note the following:

- You are expected to hand in your solutions in LaTeX;
- Pre-pend the name of your TA to the file name you hand in and remember to put your name on the submission;
- The deadline for this first assignment is (3rd of December 2018).

2 Gradient Descent Methods

1. Why is the Monte Carlo target, G_t, an unbiased estimate of $v_\pi(S_t)$?
2. Why does using the Dynamic Programming target

$$\sum_{a,s',r} \pi(a|S_t)p(s',r|S_t,a)[r + \gamma\hat{v}(s',\mathbf{w}_t)] \tag{1}$$

   result in a semi-gradient method?
3. Despite not being unbiased, semi-gradient methods that use bootstrapping have certain advantages w.r.t. Monte Carlo approaches. Why, for example, would you prefer bootstrapping in the Mountain Car problem?

3 Basis functions

1. Tabular methods can be seen as a special case of linear function approximation. Show that this is the case and give the corresponding feature vectors.
2. You want to design the feature vectors for a state space with $s = [x, y]$. You expect that x and y interact in some unknown way. How would you design a polynomial feature vector for s?
3. What happens to the size of the polynomial feature vector if the number of variables in your state space increases?

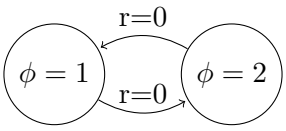

Figure 1: Two-state MDP

4. You are working on a problem with a state space consisting of two dimensions. You expect that one of them will have a larger impact on the performance than the other. How might you encode this prior knowledge in your basis function?

5. You can view coarse coding as a special case of Radial Basis Functions. Why?

4 Geometry of linear value-function approximation

Consider the MDP given in Figure 1. It consists of two states with one action, that transition into one another with reward 0. The features for both states are $\phi = 1$ for state s_0 and $\phi = 2$ for state s_1. We will now predict the value of the states using $v_w = w \cdot \phi$.

1. We can write the Bellman error vector as

$$\bar{\delta}_w = B^\pi v_w - v_w, \tag{2}$$

   where B^π is the Bellman operator. What is the Bellman error vector after initialization with $w = 1$ and using $\gamma = 1$?

2. What is the Mean Squared Bellman Error?

3. What w results in the value functions that is closest (in least-squares sense) to the target values $B^\pi v_w$?

4. Plot v_w, B^π_w and $\Pi B^\pi v_w$. Explain what is happening. (hint: Refer to Figure 11.3 in the book).

5 Neural Networks

1. Consider the state distribution, $\mu(s)$. How does it depend on the parameters of the value function approximator?

2. How does this differ from standard (un-)supervised learning problems?

3. What does this mean for the weighting of the errors (such as in e.g. Eq. 9.1)?

4. The DQN paper [1] relies, amongst others, on the use of an earlier idea called *experience replay* [2]. What does this trick do that is important for the algorithm?

5. An other important trick is the use of a separate target network that is frozen for periods of time. What does this trick do that is important for the algorithm?

6 Policy Gradient

6.1 REINFORCE

In the lecture, we have seen that introducing a constant baseline b does not introduce a bias to our policy gradient.

$$\nabla J = \mathbb{E}_\tau \Big[\Big(G(\tau) - b \Big) \nabla \log p(\tau) \Big] \tag{3}$$

We now want to consider the variance when introducing a baseline.

1. Derive the optimal constant baseline that minimizes the variance of the policy gradient. Interpret your result.

2. Consider the simple example from the lecture in a bandit setting (i.e. no states):

$$r = a + 2 \tag{4}$$
$$a \sim \mathcal{N}(\theta, 1) \tag{5}$$
$$\nabla_\theta \log \pi(a) = a - \theta \tag{6}$$

   Can you argue what should be the optimal constant baseline in this case? You can use your result from 1.

3. Now, consider a baseline that is not constant, but dependent on the state $b(s_t)$. We want to establish that in this case, the policy gradient remains unbiased. Show that

$$\mathbb{E}_\tau \Big[\sum_{t=1}^{T} \nabla \log \pi(a_t|s_t) b(s_t) \Big] = 0. \tag{7}$$

   Hint: you can use the linearity of expectation or the law of iterated expectation to "decouple" parts of the full trajectory τ.

6.2 Compatible Function Approximation Theorem

Consider the actor-critic method that optimizes the expected return J using the gradient

$$\nabla J = \mathbb{E}_\tau \left[\sum_{t=1}^{T} \nabla \log \pi(a_t|s_t) q_\pi(s_t, a_t) \right] \tag{8}$$

where π is our parameterized policy (actor) and q_π is the true state-action value under policy π. In the lecture we have seen that we can approximate q_π with a parameterized value function $\hat{q}_w$ which is unbiased if it fulfills the condition

$$\nabla_w \hat{q}_w = \nabla_\theta \log \pi_\theta(s, a) \tag{9}$$
$$\text{e.g. } \hat{q}_w = w^T \nabla_\theta \log \pi_\theta(s, a) \tag{10}$$

1. Show that $\mathbb{E}_a \Big[\hat{q}_w(s, a) \Big] = 0, \ \forall s \in S$. Briefly interpret this result.

2. Evaluate the expectation $\mathbb{E}_a \Big[q_\pi(s, a) - v_\pi(s) \Big] \ \forall s \in S$. The term $A(s, a) = q_\pi(s, a) - v_\pi(s)$ is also known as the advantage function.

3. What do you conclude from the results in 1. and 2.

4. Consider the following policy

$$\pi_\theta(s,a) = \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}} \tag{11}$$

corresponding to a softmax with linear parametrization with respect to the state-action features ϕ_{sa}. Give an expression for the parametrization of $\hat{q}_w$ that satisfies the compatible function approximation theorem.

6.3 Natural Gradient

In this section we parameterize our policy with a univariate Gaussian probability density over real-valued actions. We consider a stateless bandit scenario where you can assume that an episode solely consists of one action and one reward. Mean and variance are learned.

$$\pi(a|\theta) = \frac{1}{\sigma(\theta_\sigma)\sqrt{2\pi}} \exp\left(- \frac{(a - \mu(\theta_\mu))^2}{2\sigma(\theta_\sigma)^2} \right) \tag{12}$$

where we approximate the mean with a single parameter and the standard deviation with the exponential of a parameter

$$\mu(\theta_\mu) = \theta_\mu \tag{13}$$
$$\sigma(\theta_\sigma) = \exp\left(\theta_\sigma\right) \tag{14}$$

1. Calculate $\nabla \log \pi(a|\theta)$ w.r.t. θ_π and θ_σ.

2. Why would we want use the natural policy gradient as opposed to the "vanilla" policy gradient?

3. By applying the natural policy gradient we want to the solve the contraint optimization problem

$$\theta^* - \theta_0 = \max_{d\theta} J(\theta_0 + d\theta) \;\text{ s.t. }\; d\theta^T F_\theta d\theta = c \tag{15}$$

where we limit our gradient change to c in KL-divergence between the old and updated policy. The fisher information matrix F and the natural policy gradient update step is then given by

$$F_\theta = \mathbb{E}_\tau \Big[\nabla_{d\theta} \log \pi(a|\theta_0 + d\theta) \nabla_{d\theta} \log \pi(a|\theta_0 + d\theta)^T \Big] \tag{16}$$
$$\theta^* - \theta_0 \propto F^{-1} \nabla_\theta J(\theta_0). \tag{17}$$

Calculate the Fisher information matrix F_θ for our Gaussian policy.

4. Explain the effect of the Fisher information matrix on gradient updates of θ_μ when using the natural gradient.

5. Consider a slightly different parametrization of the standard deviation as

$$\sigma(\theta_\sigma) = \theta_\sigma \tag{18}$$

where we simply learn σ directly. We will ignore that the variance can technically become negative in this case. Calculate $\nabla \log \pi(a|\theta)$ and F_θ w.r.t. θ_σ for this alternative parametrization.

6. Compare how the Fisher information matrix influences the gradient of θ_σ considering both parametrizations of σ. Explain your observations.

7. Suppose our current parameters are $\mu = 0$ and $\sigma = 4$. We now observe an episode with $a = 8$ and $G(\tau) = 1$. Perform a gradient update on θ_σ with learning rate $\alpha = 0.01$ using the vanilla policy gradient as well as using the natural policy gradient for both parametrizations, i.e. a total of four independent updates. Report the updated values of σ and discuss your observations.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[2] Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3-4):293–321, 1992.