Reinforcement Learning (E1 277)

- Assignment 03 -

1. For a set $\mathcal{S}$, let $\Delta_{\mathcal{S}}$ be the set of probability distributions on it. Consider the MDP $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where $\mathcal{S}$ denotes a finite state space, $\mathcal{A}$ denotes a finite action space, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition kernel, $r : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is the reward function, while $\gamma \in [0, 1)$ is the discount factor. Let μ be a stationary policy and V^μ, its value function. Also, let $T^\mu : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ be the operator given by

$$T^\mu V(s) = \mathbb{E}[r(s_0, s_1) + \gamma V(s_1)|s_0 = s], \qquad \forall s \in \mathcal{S}. \tag{1}$$

Show that T^μ is a $\gamma-$contraction. Further, show that V^μ is its unique fixed point.

2. Consider the TD(0) algorithm with linear function approximation for estimating the value of a policy π. Let the underlying Markov chain under the given policy π have just two states 0 and 1 with transition probabilities $p_{0,0} = 0$, $p_{0,1} = 1$, $p_{1,0} = 0.5$, and $p_{1,1} = 0.5$, respectively. Let the features associated with the two states 0 and 1 be $\Phi(0) = [1, 0]^T$ and $\Phi(1) = [0, 1]^T$, respectively. Also, suppose the discount factor $\gamma = 0.8$. Further, let the single-stage reward $r : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ satisfy $r(0, 0) = 0, r(0, 1 = 5$, $r(1, 0) = 1$, and $r(1, 1) = 3$; the first coordinate denotes the current state and second coordinate the next state.

   (a) Write now the TD(0) recursion with linear function approximation for this example.

   (b) Write down its limiting ODE and verify if it has a unique globally asymptotically stable equilibrium point.

3. Consider the setup given in Problem 1. Let μ be the policy whose value we wish to evaluate and let Φ be the given feature matrix. For $k, t \geq 0$, let

$$d_t(s_k, s_{k+1}) = r(s_k, s_{k+1}) + (\gamma\phi(s_{k+1}) - \phi(s_k))^\top w_t, \tag{2}$$

where s_k denotes the state at time k, while w_t denotes the solution estimate of the policy evaluation problem at time t. Then, an alternative way to evaluate w_{t+1} is to explicitly solve the optimization problem

$$\arg\min_{w \in \mathbb{R}^d} \left\{ \delta\|w_t - w\|^2 + \sum_{m=0}^{t} \left(\phi^\top(s_m)w - \phi^\top(s_m)w_t - \sum_{k=m}^{t} (\gamma\lambda)^{k-m} d_t(s_k, s_{k+1}) \right)^2 \right\}. \tag{3}$$

Accordingly, the update rule can be written down as

$$w_{t+1} = w_t + \alpha_t B_t^{-1}(A_t w_t + b_t),$$

where $(\alpha_t)_{t \geq 0}$ is some stepsize sequence.

(a) Identify A_t, B_t, and b_t and write down iterative schemes for computing the same.

(b) Can you write down a sufficient condition on δ for the inverse of B_t to exist? Further, can you think of a computationally efficient way of computing B_t^{-1} in each iteration?

(c) For $\lambda = 0$ and $\lambda = 1$, give a brief description of what you think the optimization problem is attempting to do.

4. Consider the TD(0) algorithm with linear function approximation. Suppose that $\mathbb{E}\|w_0\|^2 < \infty$. Using this, show that $\mathbb{E}\|M_t\|^2 < \infty$ for all $t \geq 1$.