**Homework Set #3 (mini-project)**
**RL techniques for solve MDP with continuous state and continuous action**

This project contains three steps

1. Solving using a well known RL (Reinforcement Learning) problem that contains continuous state and continuous action
2. Solving an average reward infinite horizon MDP of unifilar channel with feedback by discretizing the state and the action.
3. Solving the average reward infinite horizon MDP that you solved in step 2 using the tools from step 1, namely, RL tools and without any discretizing.

## Step 1: RL

In this step the students need to take a well known RL problem and solve it using a well known RL tool. The state and the action must be continuous. Its recommended to an average reward infinite horizon if possible, since in the third and final step the RL tools need to be used for such a problem. However, its also possible to choose a discount problem and then adapt the RL tool for the average reward infinite horizon.

Recommended papers are [1] and [2]. Recommended RL environment gym OpenAi I also got a suggestion using tensorforce.

You may choose other environment: the most important to have continuous state and action.

## Step 2: MDP

In this step you need to solve one of the three infnite-horizon average reward MDP are given below. The DP formulations are given in the form depicted

by Bertsekas in [3], namely, using disturbance and functions. Solve the MDP by discritizing the state and the action space and performing value iteration, namely compute $\frac{1}{n}T^n(V)(s)$ or $T^{n+1}(V)(s) - T^n(V)(s)$ for large enough $n$ (usually $n = 20$).

1. **DP formulation for Trapdoor Channel capacity with feedback [4]**
   Here is the DP formulation as given in [4]:

| | |
|---|---|
| $z_t$ - the state of the DP | $z_t \in [0, 1]$ |
| $z_t = F(z_{t-1}, u_t, w_t)$ | equation (1) |
| $u_t$ - the action at time $t$ | $u_t = (\gamma_t, \delta_t) \in [0, 1 - z_{t-1}] \times [0, z_{t-1}]$ |
| $w_t$ - the disturbance | $w_t \in \{0, 1\}$ |
| $p(w_t = 0 \mid z_{t-1}, u_t)$ | $\frac{1+\delta_t-\gamma_t}{2}$ |
| $g(z_{t-1}, u_t)$ - reward at time $t$ | $H_b(\frac{1}{2} + \frac{\delta_t-\gamma_t}{2}) + \delta_t + \gamma_t - 1$ |

   where $H_b(\alpha)$ is the binary entropy with parameter $\alpha$.

$$z_t = \begin{cases} \frac{2\delta_t}{1+\delta_t-\gamma_t}, & if\, w_t = 0 \\ 1 - \frac{2\gamma_t}{1-\delta_t+\gamma_t}, & if\, w_t = 1 \end{cases} \tag{1}$$

   We consider an objective of maximizing infinite horizon average reward, given a bounded reward function $g : \mathcal{Z} \times \mathcal{U} \to \mathbb{R}$. The infinite horizon average reward for a policy $\pi$ is defined by:

$$\rho_\pi = \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}_\pi\left\{ \sum_{t=0}^{N-1} g\big(z_{t-1}, u_t\big) \right\} \tag{2}$$

   and the optimal average reward is defined by $\rho^* = \sup_\pi \rho_\pi$

   (a) For a bounded function $h : \mathcal{Z} \to \mathbb{R}$ we define the dynamic program operator as:

$$(Th)(z) = g(z, u) + \sum_{w'\in\mathcal{W}} P(w = w'|z, u)h(F(z, u, w')) \tag{3}$$

Prove that

$$(Th)(z) = H_b\left(\frac{1}{2} + \frac{\delta - \gamma}{2}\right) + \delta + \gamma - 1 + \frac{1+\delta-\gamma}{2} h\left(\frac{2\delta}{1+\delta-\gamma}\right) + \frac{1-\delta+\gamma}{2} h\left(1 - \frac{2\gamma}{1-\delta+\gamma}\right)$$

2. **DP formulation for Ising channel [5]**
The DP presented in [5]:

| | |
|---|---|
| $z_t$ - the state of the DP | $z_t \in [0,1]$ |
| $z_t = F(z_{t-1}, u_t, w_t)$ | equation (4) |
| $u_t$ - the action at time $t$ | $u_t = (\gamma_t, \delta_t) \in [0, 1-z_{t-1}] \times [0, z_t]$ |
| $w_t$ - the disturbance | $w_t \in \{0,1\}$ |
| $p(w_t = 0 \mid z_{t-1}, u_t)$ | $\frac{1+\delta_t-\gamma_t}{2}$ |
| $g(z_{t-1}, u_t)$ - reward at time $t$ | $H_b(\frac{1}{2} + \frac{\delta_t-\gamma_t}{2}) + \delta_t + \gamma_t - 1$ |

$$z_t = \begin{cases} 1 + \frac{\delta_t - Z_{t-1}}{1+\delta_t-\gamma_t}, & if\ w_t = 0 \\ \frac{1-z_{t-1}-\gamma_t}{1-\delta_t+\gamma_t}, & if\ w_t = 1 \end{cases} \tag{4}$$

Consider again the objective of infinite horizon average reward problem.

(a) Using the same dynamic program operator defined in (3) derive that:

$$(Th)(z) = H_b\left(\frac{1}{2} + \frac{\delta - \gamma}{2}\right) + \delta + \gamma - 1 + \frac{1+\delta-\gamma}{2} h\left(1 + \frac{\delta - z}{1+\delta-\gamma}\right) + \frac{1-\delta+\gamma}{2} h\left(\frac{1-z-\gamma}{1-\delta+\gamma}\right)$$

3. **DP formulation for binary erasure channel with a no-consecutive-ones input constraint [6]**
The MDP presented in [6]:

| | |
|---|---|
| $z_t$ - the state of the DP | $z_t \in [0,1]$ |
| $z_t = F(z_{t-1}, u_t, w_t)$ | equation (5) |
| $u_t$ - the action at time $t$ | $u_t = \delta_t \in [0, z_t]$ |
| $w_t$ - the disturbance | $w_t \in \{0, ?, 1\}$ |
| $p(w_t = 0 \mid z_{t-1}, u_t)$ | $(1-\delta_t)\bar{\epsilon}$ |
| $p(w_t = 1 \mid z_{t-1}, u_t)$ | $\delta_t \bar{\epsilon}$ |
| $p(w_t = ? \mid z_{t-1}, u_t)$ | $1 - \bar{\epsilon}$ |
| $g(z_{t-1}, u_t)$ - reward at time $t$ | $\bar{\epsilon} H_b(\delta_t)$ |

where $\bar{\epsilon} \in (0,1)$ fixed

$$z_t = \begin{cases} 1, & if\ w_t = 0 \\ 1-\delta_t, & if\ w_t = ? \\ 0, & if w_t = 0 \end{cases} \tag{5}$$

Consider again the objective of infinite horizon average reward problem.

(a) Using the same dynamic program operator defined in (3) derive that:

$$(Th_\epsilon)(z) = \bar{\epsilon} H_b(\delta_t) + (1-\delta)\bar{\epsilon} h_\epsilon(1) + \epsilon h_\epsilon(1-\delta) + \delta\bar{\epsilon} h_\epsilon(0) \tag{6}$$

### Step 3: Solving MDP with the RL tools

In this step you need to solve the MDP from Step 2 using the RL tools from Step 1. The idea is to avoid the quantization of the state and the action, so we can deal with more complicated channels. Namely, MDP with larger state and action space.

## References

[1] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv:1509.02971

[2] Y. Duan, X. Chen, R. Houthooft, J. Schulman, P. Abbeel, Benchmarking Deep Reinforcement Learning for Continuous Control, arXiv:1604.06778

[3] Bertsekas D.P - Dynamic Programming and Optimal Control - Vol1

[4] H. Permuter, P. Cuff, B. V. Roy, and T. Weissman, Capacity of the trapdoor channel with feedback, IEEE Trans. Inf. Theory, vol. 54, no. 7, pp. 31503165, Jul. 2008.http://www.ee.bgu.ac.il/~haimp/trapdoor_channel_it.pdf

[5] O. Elishco and H. Permuter, Capacity and coding for the Ising channel with feedback, IEEE Trans. Inf. Theory, vol. 60, no. 9, pp. 51385149, Sep. 2014.http://www.ee.bgu.ac.il/~haimp/IT_Ising_channel_with_feedback.pdf

[6] O. Sabag, H. H. Permuter, and N. Kashyap, The feedback capacity of the binary symmetric channel with a no-consecutive-ones input constraint, in Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput., Sep. 2015, pp. 160164.http://www.ee.bgu.ac.il/~haimp/IT_no_repeated_ones_erasure_channel.pdf