Assignment 1 for CMPUT 652 RL with Robots

October 4, 2019

Total 110 points. Complete within the given space using LaTeX or handwritten notes. Show the derivations or the steps for obtaining partial points. On the other hand, mistakes, missteps or bad reasoning behind a correct answer will result in points deducted. Submit at armahmood@ualberta.ca

1 **Consider the supervised regression problem of predicting an output Y from a vector input $\mathbf{x}$ using a linear function with a vector parameter $\mathbf{w}$. We found in the lecture that the Mean Squared Error (MSE) $\mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right]$ is a reasonable objective for this problem. Derive the Least-Squares (LS) method, defined below, from MSE (5 points).**

$$\mathbf{w}_{t+1}^{\mathrm{LS}} = \left(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}_k Y_k\right). \tag{1}$$

$$\begin{aligned}
\mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= \mathrm{E}\left[Y^2 + \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x} Y\right] \\
&= \mathrm{E}\left[Y^2\right] + \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right] \mathbf{w} - 2\mathbf{w}^\top \mathrm{E}\left[\mathbf{x} Y\right] \\
&\text{taking the gradient w.r.t weight} \\
\nabla_{\mathbf{w}} \mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= 2\,\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right] \mathbf{w} - 2\,\mathrm{E}\left[\mathbf{x} Y\right] = 0 \\
\mathbf{w}^* &= \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]^{-1} \mathrm{E}\left[\mathbf{x} Y\right] \\
&\text{use t samples to approximate the result from least squared method} \\
\mathbf{w}_{t+1}^{\mathrm{LS}} &= \left(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}_k Y_k\right)
\end{aligned}$$

2 **Continuing with the supervised regression problem, derive the Stochastic Gradient Descent (SGD) method, defined below, from MSE (5 points).**

$$\mathbf{w}_{t+1} = \mathbf{w}_t + 2\alpha_t \left(Y_t - \mathbf{x}_t^\top \mathbf{w}_t\right) \mathbf{x}_t. \tag{2}$$

$$\begin{aligned}
\mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= \mathrm{E}\left[Y^2 + \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x} Y\right] \\
&= \mathrm{E}\left[Y^2\right] + \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right] \mathbf{w} - 2\mathbf{w}^\top \mathrm{E}\left[\mathbf{x} Y\right] \\
&\text{taking the gradient w.r.t weight} \\
\nabla_{\mathbf{w}} \mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= 2\,\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right] \mathbf{w} - 2\,\mathrm{E}\left[\mathbf{x} Y\right] = 0 \\
&\text{approximate with a single sample} \\
&\approx 2\left(\mathbf{x}_t \mathbf{x}_t^\top \mathbf{w} - \mathbf{x}_t Y_t\right) \\
&\approx -2\left(Y_t - \mathbf{x}_t^\top \mathbf{w}_t\right) \mathbf{x}_t \\
&\text{use a as step size} \\
\mathbf{w}_{t+1} &= \mathbf{w}_t + 2\alpha_t \left(Y_t - \mathbf{x}_t^\top \mathbf{w}_t\right) \mathbf{x}_t
\end{aligned}$$

3 **According to the strong law of large numbers, sample mean is a consistent estimator of the expected value (See Lemma 3, pg 18, of Mahmood 2017). Use this fact to argue that the sample MSE $\frac{1}{t}\sum_{k=1}^{t}\left(Y_k - \mathbf{x}_k^\top \mathbf{w}\right)^2$ objective is a consistent estimator of MSE (5 points).**

4 **Similarly, argue that LS is a consistent estimator of the solution to MSE (5 points).**

5 **Establish a relationship between MSE and the following: $\mathrm{E}_{\mathbf{x}}\left[\left(\mathrm{E}_{Y|\mathbf{x}}[Y]-\mathbf{x}^\top\mathbf{w}\right)^2\right]$. Note that MSE can be written as $\mathrm{E}\left[(Y-\mathbf{x}^\top\mathbf{w})^2\right]=\mathrm{E}_{\mathbf{x}}\,\mathrm{E}_{Y|\mathbf{x}}\left[(Y-\mathbf{x}^\top\mathbf{w})^2\right]$. Also, the expectations E_X and $\mathrm{E}_{Y|X}$ can be elaborated as $\mathrm{E}_X\,f(X)=\int_{\mathcal{X}}p^*(x)f(x)dx$ and $E_{Y|X}f(Y)=\int_{\mathcal{Y}}p^*(y|x)f(y)dy$ (5 points).**

$$
\begin{aligned}
\mathrm{E}\left[(Y-\mathbf{x}^\top\mathbf{w})^2\right] &= \mathrm{E}_{\mathbf{x}}\,\mathrm{E}_{Y|\mathbf{x}}\left[(Y-\mathbf{x}^\top\mathbf{w})^2\right] \\
&= \int_{\mathcal{X}} p^*(x)\int_{\mathcal{Y}} p^*(y|x)(y-x^\top\mathbf{w})^2 dydx \\
&= \int_{\mathcal{X}} p^*(x)\int_{\mathcal{Y}} p^*(y|x)y^2-2yx^\top\mathbf{w}+\mathbf{w}^\top xx^\top\mathbf{w}dydx \\
&= \int_{\mathcal{X}} p^*(x)\left[\mathrm{E}_{Y|\mathbf{x}}\left[Y^2\right]-2\,\mathrm{E}_{Y|\mathbf{x}}[Y]\,\mathbf{x}^\top\mathbf{w}+\mathbf{w}^\top xx^\top\mathbf{w}\right]dx \\
&= \int_{\mathcal{X}} p^*(x)\left[\mathrm{E}_{Y|\mathbf{x}}[Y]^2-2\,\mathrm{E}_{Y|\mathbf{x}}[Y]\,\mathbf{x}^\top\mathbf{w}+\mathbf{w}^\top xx^\top\mathbf{w}\right]dx \;Y \text{ is independent given } \mathbf{x} \\
&= \int_{\mathcal{X}} p^*(x)\left(\mathrm{E}_{Y|\mathbf{x}}[Y]-\mathbf{x}^\top\mathbf{w}\right)^2 dx \\
&= \mathrm{E}_{\mathbf{x}}\left[\left(\mathrm{E}_{Y|\mathbf{x}}[Y]-\mathbf{x}^\top\mathbf{w}\right)^2\right]
\end{aligned}
$$

6 **Objective functions serve many purposes. They bring specificity to the problem formulation, provide a criteria for evaluation and can be used as a basis for deriving new algorithms. Accordingly, the same problem may have different associated objectives. For example, the sample MSE is widely used to evaluate the performance of a learning method, but not all methods ensue from MSE. Consider the following objective $\mathrm{MSE}_{\mathrm{L}\,2}$, which is a slight modification to MSE: $\mathrm{MSE}_{\mathrm{L}\,2}(\mathbf{w})=\mathrm{E}\left[\left(Y-\mathbf{x}^\top\mathbf{w}\right)^2+\lambda\mathbf{w}^\top\mathbf{w}\right]$. Derive the following method know as ridge regression from $\mathrm{MSE}_{\mathrm{L}\,2}$ (5 points).**

$$
\mathbf{w}_{t+1}^{\mathrm{RR}}=\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k\mathbf{x}_k^\top+\lambda\mathbf{I}\right)^{-1}\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_kY_k\right),\ \text{where } \mathbf{I} \text{ is the identity matrix.} \tag{3}
$$

$$
\begin{aligned}
&\mathrm{MSE}_{\mathrm{L}\,2}(\mathbf{w})=\mathrm{E}\left[Y^2-2Y\mathbf{x}^\top\mathbf{w}+\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}+\lambda\mathbf{w}^\top\mathbf{w}\right] \\
&\rightarrow \nabla_{\mathbf{w}}\,\mathrm{MSE}_{\mathrm{L}\,2}(\mathbf{w})=\mathrm{E}\left[-2\mathbf{x}Y+2\mathbf{x}\mathbf{x}^\top\mathbf{w}+2\lambda\mathbf{w}\right] \\
&\text{set } \mathrm{E}\left[-2\mathbf{x}Y+2\mathbf{x}\mathbf{x}^\top\mathbf{w}+2\lambda\mathbf{w}\right]=0 \\
&\rightarrow \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top+\lambda\mathbf{I}\right]\mathbf{w}=\mathrm{E}\left[\mathbf{x}Y\right] \\
&\rightarrow \mathbf{w}=\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top+\lambda\mathbf{I}\right]^{-1}\mathrm{E}\left[\mathbf{x}Y\right] \\
&\text{approximate with samples} \\
&\rightarrow \mathbf{w}_{t+1}^{\mathrm{RR}}=\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k\mathbf{x}_k^\top+\lambda\mathbf{I}\right)^{-1}\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_kY_k\right)
\end{aligned}
$$

7 **Relate the MSE objective for supervised regression to the cross-entropy objective:** $\mathrm{CE}(\mathbf{w}) = -\int_{\mathcal{X}} p^*(x) \int_{\mathcal{Y}} p^*(y|x) \log p_{\mathbf{w}}(y|x) dy dx$**, where** p^* **is the true distribution of** Y **and** $\mathbf{x}$**, and** $p_{\mathbf{w}}$ **is the estimated distribution. Use the fact that** $\mathrm{MSE}(\mathbf{w}) = \mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] = \int_{\mathcal{X}} p^*(x) \int_{\mathcal{Y}} p^*(y|x)(y - x^\top \mathbf{w})^2 dy dx$**. (5 points).**

$$\rightarrow p_{\mathbf{w}}(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - x^\top \mathbf{w})^2}{2\sigma^2}\right)$$

$$\rightarrow \log p_{\mathbf{w}}(y|x) = -\frac{(y - x^\top \mathbf{w})^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma$$

substitute

$$\rightarrow \mathrm{CE}(\mathbf{w}) = -\int_{\mathcal{X}} p^*(x) \int_{\mathcal{Y}} p^*(y|x) \left(-\frac{(y - x^\top \mathbf{w})^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma\right) dy dx$$

$$\rightarrow \mathrm{CE}(\mathbf{w}) = \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \int_{\mathcal{X}} p^*(x) \int_{\mathcal{Y}} p^*(y|x) \left(-(y - x^\top \mathbf{w})^2\right) dy dx$$

The only difference is constant term

8 **We discussed how supervised and reinforcement learning problems are fundamentally different. Show how that difference impacts the choice of policy optimization objective in terms of cross entropy in the case of bandits. More specifically, formulate the policy optimization problem with a cross entropy objective, discuss the difference between this objective and the cross entropy objective for supervised regression defined above, and relate this difference to the fundamental difference between supervised and reinforcement learning problems (5 points).**

By definition, we could have:

$$\mathrm{CE}(\pi_\theta, \pi^*) = -\int \pi_\theta(a) \log \pi^*(a)\, da$$

substitute with a Boltzman distribution

$$\pi^*(a) = Z^{-1} e^{q(a)\tau^{-1}}$$

$$\log \pi^*(a) = \tau^{-1} q(a) - \log Z$$

substitue, the objective will be:

$$-\arg\min_\theta \int \pi_\theta(a) \log \pi^*(a)\, da$$

$$= -\arg\min_\theta \int \pi_\theta(a)(\tau^{-1} q(a) - \log Z)\, da$$

$$= -\arg\min_\theta \int \pi_\theta(a) q(a)\, da.$$

The difference is ...

9 **The expected policy gradient update for contextual bandit problem can be given as follows: $\mathrm{E}\left[-\log \pi_\theta\left(A|X\right) R\right]$, where X is a given context, π_θ is the parameterized policy distribution maintained by the agent, A is the action chosen by the agent according to π_θ, and R is the subsequent reward. Show that this update can be derived by calculating the gradient of the cross entropy objective: $\mathrm{CE}(\theta) = -\int_{\mathcal{X}} p(x) \int_{\mathcal{A}} \pi_\theta(a|x) \log \pi^*(a|x) dadx$, where p, which is not a function of θ, is the distribution of context X. Also, π^* is the ideal policy distribution defined as the Boltzmann distribution $\pi^*(a|x) = Z^{-1} e^{q(x,a)\tau^{-1}}$, where $q(x,a) = \mathrm{E}\left[R|A=a, X=x\right]$ is the contextual action value. (10 points).**

Bolzman distribution again

$$\pi^*(a|x) = Z^{-1} e^{q(x,a)\tau^{-1}}$$

$$\log \pi^*(a|x) = \tau^{-1} q(x,a) - \log Z$$

$$\mathrm{CE}(\theta) = -\int_{\mathcal{X}} p(x) \int_{\mathcal{A}} \pi_\theta(a|x) \left(\tau^{-1} q(x,a) - \log Z\right) dadx$$

$$\mathrm{CE}(\theta) = \log Z - \tau^{-1} \int_{\mathcal{X}} p(x) \int_{\mathcal{A}} \pi_\theta(a|x) q(x,a) dadx$$

$$\nabla_\theta \mathrm{CE}(\theta) = -\tau^{-1} \int_{\mathcal{X}} p(x) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|x) q(x,a) dadx$$

$$\nabla_\theta \mathrm{CE}(\theta) = -\tau^{-1} \int_{\mathcal{X}} p(x) \int_{\mathcal{A}} \pi_\theta(a|x) \frac{\nabla_\theta \pi_\theta(a|x)}{\pi_\theta(a|x)} \nabla_\theta \pi_\theta(a|x) q(x,a) dadx$$

$$\nabla_\theta \mathrm{CE}(\theta) = -\tau^{-1} \int_{\mathcal{X}} p(x) \int_{\mathcal{A}} \pi_\theta(a|x) \nabla_\theta \log \pi_\theta(a|x) q(x,a) dadx$$

$$\nabla_\theta \mathrm{CE}(\theta) = -\tau^{-1} \int_{\mathcal{X}} p(x) \, \mathrm{E}\left[\nabla_\theta \log \pi_\theta(A|x) q(x,A)\right] dx$$

$$\nabla_\theta \mathrm{CE}(\theta) = -\tau^{-1} \, \mathrm{E}\left[\nabla_\theta \log \pi_\theta(A|X) q(X,A)\right]$$

$$\nabla_\theta \mathrm{CE}(\theta) = \tau^{-1} \, \mathrm{E}\left[-\nabla_\theta \log \pi_\theta(A|X) R\right]$$

ignore the constant term, they now have identical term

10 **We wrote a PyTorch code for a discrete action bandit problem. Consider a continuous action bandit problem, where actions are distributed normally: $A \sim N(\mu, \sigma^2)$, and the reward is a quadratic function of the action: $R = -(A-10)^2$. What is the optimal policy in terms of mean μ and standard deviation σ (5 points)? Write a PyTorch code for learning the mean and the standard deviation using policy gradient. Attach the code here together with the curve for estimated mean and standard deviation over time for 100K time steps. The code should run without error in colab and bring the plots. (15 points, total 20).**

11 **For discrete MDP, Derive the following Bellman equation for action value function from its definition: $q_\pi(s,a) = \mathrm{E}\left[G_t|S_t = s, A_t = a, A_k \sim \pi, \forall k > t\right]$, where S_t is the state at t, A_t is the action at t, G_t is the return after t, π is the policy, γ is the discount factor, $r(s,a) = \sum_r r \sum_{s'} p(s', r|s, a)$ and $p_\pi(s', a'|s, a) = \sum_r p(s', r|s, a)\pi(a'|s')$ (5 points).**

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s',a'} p_\pi(s', a'|s, a) q_\pi(s', a'). \tag{4}$$

$$\begin{aligned}
q_\pi(s,a) &= \mathrm{E}\left[G_t|S_t = s, A_t = a, A_k \sim \pi, \forall k > t\right] \\
&= \mathrm{E}_\pi\left[G_t|S_t = s, A_t = a\right] \\
&= \mathrm{E}_\pi\left[R_t + \gamma G_{t+1}|S_t = s, A_t = a\right] \\
&= \mathrm{E}\left[R_t|S_t = s, A_t = a\right] + \gamma\, \mathrm{E}_\pi\left[G_{t+1}|S_t = s, A_t = a\right] \\
&= \sum_r r \sum_{s'} p(s', r|s, a) + \gamma\, \mathrm{E}\left[\mathrm{E}_\pi\left[G_{t+1}|S_{t+1}, A_{t+1}\right]|S_t = s, A_t = a\right] \\
&= r(s,a) + \gamma\, \mathrm{E}\left[q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a\right] \\
&= r(s,a) + \gamma \sum_{s',a'} \sum_r p(s', r|s, a)\pi(a'|s') q_\pi(s', a') \\
&= r(s,a) + \gamma \sum_{s',a'} p_\pi(s', a'|s, a) q_\pi(s', a')
\end{aligned}$$

12 **Write the above Bellman equation in matrix form (5 points).**

$$\begin{aligned}
v_\pi(s) &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\big(r + \gamma v_\pi(s')\big) \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) r + \gamma \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) v_\pi(s') \\
&= \sum_a \pi(a|s) r(s,a) + \gamma \sum_{s'} \Big[\sum_a \pi(a|s) p(s'|s, a)\Big] v_\pi(s') \\
&= r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s) v_\pi(s')
\end{aligned}$$

$$\begin{bmatrix} v_\pi(s_0) \\ v_\pi(s_1) \\ \vdots \end{bmatrix} = \begin{bmatrix} r_\pi(s_0) \\ r_\pi(s_1) \\ \vdots \end{bmatrix} + \gamma \begin{bmatrix} p_\pi(s_0|s_0) & p_\pi(s_1|s_0) & \dots \\ p_\pi(s_0|s_1) & p_\pi(s_1|s_1) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} v_\pi(s_0) \\ v_\pi(s_1) \\ \vdots \end{bmatrix}$$

$$\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi \mathbf{v}_\pi$$

13 **Consider a policy π that is not optimal, that is $v_\pi(s) < v_{\pi^*}(s), \exists s$. Also consider the greedy-policy operator g that gives a policy gq which is greedy with respect to a given function q, that is, $(gq)(s) = \arg\max_b q(s,b)$, assuming there is no tie. Show that, $v_{gq_\pi}(s) > v_\pi(s), \exists s$ (5 points).**

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a|s) q_\pi(s,a) \\
v_{gq_\pi}(s) &= \sum_a gq_\pi(a|s) q_\pi(s,a) \\
&= q_\pi(s, \arg\max_b q_\pi(s,b)) \\
&= \max_a q_\pi(s,a) \\
v_{gq_\pi}(s) = \max_a q_\pi(s,a) &\geq v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a), \forall s
\end{aligned}
$$

The condition of equality under all states is that optimal policy, contradict

$$v_{gq_\pi}(s) > v_\pi(s), \exists s$$

14 **As we have seen, the action value iteration method can be written as: $\mathbf{q}_{k+1} = \mathbf{q}_k - \alpha\left(\left(\mathbf{I} - \gamma \mathbf{P}_{g\mathbf{q}_k}\right)\mathbf{q}_k - \mathbf{r}\right)$, where $\mathbf{q}$ is the estimate of action value in vector form, and the vector $\mathbf{r}$ contains $[\mathbf{r}]_{sa} = r(s,a)$ for different s, a. Also, the matrix $\mathbf{P}_{g\mathbf{q}_k}$ contains state-action-transition probabilities, $[\mathbf{P}_{g\mathbf{q}_k}]_{sa,s'a'} = p(s'|s,a)(g\mathbf{q}_k)(a'|s')$, where next actions are taken under policy $g\mathbf{q}_k$, which is the greedy policy with respect to $\mathbf{q}_k$. Find the condition under which the action-value iteration method converges, and convince that the condition is satisfied in general. (5 points).**

15 **Write the following state-value iteration method in matrix form:** $v_{k+1}(s) = v_k(s) + \alpha \left(\max_b \left[r(s,b) + \gamma \sum_{s'} p(s'|s,b) v_k(s')\right] - v_k(s)\right)$ **(5 points).**

16 **Consider the two-state MDP example from lecture, where $\gamma = 0.9$, reward $r(s,a)$ is 1 for state 1 and action 1, and zero for all other state action pairs, and the state transition probabilities $p(s'|s,a)$ are as follows: $p(0|0,0) = p(1|0,1) = p(0|1,0) = p(1|1,1) = 1$, and zero for any other transition. (5 points). Now, consider a policy π which takes action 0 at state 0 with probability 0.25 but takes the same action at state 1 with probability 0.75. Find $r_\pi(s) = \sum_a \pi(a|s) r(s,a)$, $p_\pi(s'|s) = \sum_a \pi(a|s) p(s'|s,a)$, and $v_\pi(s)$ for all s and s'. Mention the method you used for finding v_π or show calculation. (5 points).**

17 **For the same MDP, find q_π, mention the method you used for finding q_π or show calculation, and describe gq_π, that is the greedy policy over q_π for the policy π described in the previous question (5 points).**

18 **Consider the nonlinear function approximator for supervised regression $Y \approx \theta^\top \phi(\mathbf{W}\mathbf{x})$, where Y is the target scalar output, $\mathbf{x}$ is the vector input, θ is the output weight vector, $\mathbf{W}$ is the input weight matrix, and ϕ is the element-wise activation function for the hidden layer. Consider the activation function to be sigmoid. Show the backpropagation update for input weights for this case in matrix form. (5 points).**