# Assignment 1 for CMPUT 652 RL with Robots

October 12, 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 x using a linear function with a vector parameter 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).

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

$$\begin{aligned}\mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= \mathrm{E}\left[Y^2 + \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x} Y\right] \\ &= \mathrm{E}\left[Y^2\right] + \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{w} - 2\mathbf{w}^\top \mathrm{E}\left[\mathbf{x}Y\right]\end{aligned}$$

The exact solution can be obtained by setting the derivative equal to 0 and the Least-Square solution can be obtained by substituting the expected values with sample average:

$$\begin{aligned}\nabla_{\mathbf{w}} \mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= 2\,\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{w} - 2\,\mathrm{E}\left[\mathbf{x}Y\right] = 0 \\ \mathbf{w}^* &= \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]^{-1} \mathrm{E}\left[\mathbf{x}Y\right] \\ \mathbf{w}_{t+1}^{\mathrm{LS}} &= \left(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\frac{1}{t}\sum_{k=1}^{t} \mathbf{x}_k Y_k\right)\end{aligned}$$

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

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

$$\begin{aligned}\mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= \mathrm{E}\left[Y^2 + \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w} - 2\mathbf{w}^\top \mathbf{x} Y\right] \\ &= \mathrm{E}\left[Y^2\right] + \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{w} - 2\mathbf{w}^\top \mathrm{E}\left[\mathbf{x}Y\right]\end{aligned}$$

Stochastic Gradient Descent can be obtained from approximate the exact gradient with one sample:

$$\begin{aligned}\nabla_{\mathbf{w}} \mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] &= 2\,\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{w} - 2\,\mathrm{E}\left[\mathbf{x}Y\right] \\ &\approx 2\left(\mathbf{x}_t\mathbf{x}_t^\top \mathbf{w} - \mathbf{x}_t Y_t\right) \\ &\approx -2\left(Y_t - \mathbf{x}_t^\top \mathbf{w}_t\right)\mathbf{x}_t \\ \mathbf{w}_{t+1} &= \mathbf{w}_t + 2\alpha_t\left(Y_t - \mathbf{x}_t^\top \mathbf{w}_t\right)\mathbf{x}_t\end{aligned}$$

## 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).

$$
\begin{aligned}
&\lim_{t\to\infty} \mathrm{E}\left[\mathrm{E}\left[\left(Y - \mathbf{x}^\top \mathbf{w}\right)^2\right] - \frac{1}{t}\sum_{k=1}^{t}\left(Y_k - \mathbf{x}_k^\top \mathbf{w}\right)^2\right] \\
&= \lim_{t\to\infty} \mathrm{E}\left[\mathrm{E}\left[Y^2 - 2\mathbf{w}^\top \mathbf{x} Y + \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w}\right] - \frac{1}{t}\sum_{k=1}^{t}\left(Y_k^2 - 2\mathbf{w}^\top \mathbf{x}_k Y_k + \mathbf{w}^\top \mathbf{x}_k \mathbf{x}_k^\top \mathbf{w}\right)\right] \\
&= \lim_{t\to\infty} \mathrm{E}\left[\mathrm{E}\left[Y^2\right] - \frac{1}{t}\sum_{k=1}^{t} Y_k^2 - 2\mathbf{w}^\top \mathrm{E}\left[\mathbf{x}Y\right] + 2\mathbf{w}^\top \frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k Y_k + \mathbf{w}^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{w} - \mathbf{w}^\top \frac{1}{t}\sum_{k=1}^{t}(\mathbf{x}_k \mathbf{x}_k^\top)\right] \\
&= \lim_{t\to\infty} \mathrm{E}\left[\mathrm{E}\left[Y^2\right] - \frac{1}{t}\sum_{k=1}^{t} Y_k^2\right] + \lim_{t\to\infty} \mathrm{E}\left[2\mathbf{w}^\top \frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k Y_k - 2\mathbf{w}^\top \mathrm{E}\left[\mathbf{x}Y\right]\right] + \lim_{t\to\infty} \mathrm{E}\left[\mathbf{w}^\top \mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{w} - \mathbf{w}^\top \frac{1}{t}\right] \\
&= 0
\end{aligned}
$$

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

From Q1, we can directly write out the form of solutions obtained from MSE and LS:

$$
\begin{aligned}
\lim_{t\to\infty} \mathrm{E}\left[\frac{\mathbf{w}^{\mathrm{MSE}}}{\mathbf{w}_{t+1}^{\mathrm{LS}}}\right] &= \lim_{t\to\infty} \mathrm{E}\left[\frac{\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]^{-1}\mathrm{E}\left[\mathbf{x}Y\right]}{\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k \mathbf{x}_k^\top\right)^{-1}\left(\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k Y_k\right)}\right] \\
&= \lim_{t\to\infty} \mathrm{E}\left[\frac{\mathrm{E}\left[\mathbf{x}\mathbf{x}^\top\right]^{-1}}{\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k \mathbf{x}_k^\top}\right] \lim_{t\to\infty} \mathrm{E}\left[\frac{\mathrm{E}\left[\mathbf{x}Y\right]}{\frac{1}{t}\sum_{k=1}^{t}\mathbf{x}_k Y_k}\right] \\
&= 1
\end{aligned}
$$

## 5

**Establish a relationship between MSE and the following:** $\mathrm{E}_{\mathbf{x}}\left[\left(\mathrm{E}_{Y|\mathbf{x}}\left[Y\right]-\mathbf{x}^\top\mathbf{w}\right)^2\right]$**. 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** $\mathrm{E}_X$ **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).**

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

## 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).**

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

$$
\mathrm{MSE}_{\mathrm{L}\,2}(\mathbf{w})=\mathrm{E}\left[Y^2-2Y\mathbf{x}^\top\mathbf{w}+\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}+\lambda\mathbf{w}^\top\mathbf{w}\right]
$$

$\mathbf{w}$ can be solved by setting the derivative equal to 0 and the required result can be obtained by substituting the expected value with sample average:

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

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

The relation can be built if we assume $p$ here is a Gaussian distribution and $\mathbf{w}$ is the weight of a linear model:

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

If we insert $\log p_{\mathbf{w}}(y|x)$ in the CE objective, it will have the same form as MSE objective except the constant term:

$$\begin{aligned} \mathrm{CE}(\mathbf{w}) &= -\int_{\mathcal{X}} p^*(x) \int_{\mathcal{Y}} p^*(y|x) \left(-\frac{(y - x^\top \mathbf{w})^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma\right) dydx \\ &= \log \sqrt{2\pi}\sigma + \frac{1}{2\sigma^2} \int_{\mathcal{X}} p^*(x) \int_{\mathcal{Y}} p^*(y|x)(y - x^\top \mathbf{w})^2 dydx \end{aligned}$$

## 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).

To build the objective, we can use a Cross Entropy objective of the following form: $\mathrm{CE}(\pi_\theta, \pi^*) = -\int_{\mathcal{A}} \pi_\theta(a) \log \pi^*(a)\, da$, where $\pi_\theta, \pi^*$ are current policy and optimal policy respectively. The relationship can be established with a Boltzman distribution $Z^{-1} e^{q(a)\tau^{-1}}$, where $Z$ is a normalizing constant and $\tau$ is called the *temperature*, which controls the entropy of the distribution. In this case, the objective can be written as:

$$\begin{aligned} -\arg\min_\theta \int_{\mathcal{A}} \pi_\theta(a) \log \pi^*(a)\, da &= -\arg\min_\theta \int_{\mathcal{A}} \pi_\theta(a) \log Z^{-1} e^{q(a)\tau^{-1}}\, da \\ &= -\arg\min_\theta \int_{\mathcal{A}} \pi_\theta(a)(\tau^{-1} q(a) - \log Z)\, da \\ &= -\arg\min_\theta \int_{\mathcal{A}} \pi_\theta(a) q(a)\, da \end{aligned}$$

The fundamental difference between supervised learning and reinforcement learning implied here is where the optimal target(policy) distribution is known. In supervised learning, the optimal distribution is given as the label, however, in reinforcement learning, the reward only tells the agent how good an action instead of whether the action is optimal or not. Such difference makes reinforcement learning inherently difficult than supervised learning.

## 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, $\pi_\theta$ is the parameterized policy distribution maintained by the agent, $A$ is the action chosen by the agent according to $\pi_\theta$, 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 $\theta$, is the distribution of context $X$. Also, $\pi^*$ 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).

Similar to Q8, we insert the Boltzman distribution $\pi^*(a|x) = Z^{-1}e^{q(x,a)\tau^{-1}}$ in to the Cross-Entropy objective:

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

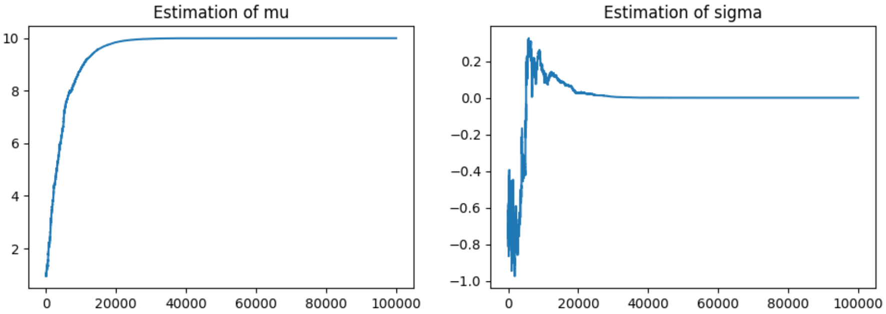
We can obtain the same result if we ignore the constant term:

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

**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 $\mu$ and standard deviation $\sigma$ (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).**

Optimal policy of mean $\mu$ is 10 and standard deviation $\sigma$ is 0.

```
from matplotlib import pyplot as plt
import numpy as np
import torch
from torch.distributions import Normal

T = int(1e5) # total time step
alpha = 1e-4 # learning rate

mu_history = []
sigma_history = []

policy = torch.nn.Sequential(
    torch.nn.Linear(1, 2, bias=False)
)

optimizer = torch.optim.SGD(params=policy.parameters(), lr=alpha)

for t in range(T):
    mu, sigma = policy(torch.ones(1))
    dist = Normal(loc=mu, scale=sigma)
    action = dist.sample()
    reward = - np.square(action - 10)
    mu_history.append(mu.item())
    sigma_history.append(sigma.item())
    loss = -dist.log_prob(action) * reward # policy gradient objective
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()

plt.subplot(1,2,1)
plt.title('Estimation␣of␣mu')
plt.plot(mu_history)
plt.subplot(1,2,2)
plt.title('Estimation␣of␣sigma')
plt.plot(sigma_history)
plt.show()
```



## 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 $t$, $\pi$ is the policy, $\gamma$ 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).

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

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

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

## 13

**Consider a policy $\pi$ 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).**

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

$$v_{gq_\pi}(s) = \max_a q_\pi(s,a) \geq v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a), \forall s \in \mathcal{S}$$

The condition of equality of the above formula is that the policy is optimal, however, given the fact that policy $\pi$ is not optimal, we have:

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

## 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 q is the estimate of action value in vector form, and the vector 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).**

$$\begin{aligned}
\mathbf{q}_{k+1} &= \mathbf{q}_k - \alpha\left(\left(\mathbf{I} - \gamma \mathbf{P}_{g\mathbf{q}_k}\right)\mathbf{q}_k - \mathbf{r}\right) \\
&= \mathbf{q}_k - \alpha \mathbf{I}\mathbf{q_k} + \alpha\gamma \mathbf{P_{gq_k}}\mathbf{q_k} - \alpha \mathbf{r} \\
&= \left(\mathbf{I} - \alpha\mathbf{I} + \alpha\gamma\mathbf{P_{gq_k}}\right)\mathbf{q}_k - \alpha r \\
&\text{denote } A_k \text{ as } \left(\mathbf{I} - \alpha\mathbf{I} + \alpha\gamma\mathbf{P_{gq_k}}\right) \text{ and } b \text{ as } \alpha r \\
&= A_k \mathbf{q}_k - b \\
&= A_k(A_{k-1}\mathbf{q}_{k-1} - b) - b \\
&= A_k A_{k-1}\mathbf{q}_{k-1} - A_k b - b \\
&\ldots \\
&= \prod_{t=0}^{k} A_t \mathbf{q}_0 - \sum_{t=0}^{k}\left[\prod_{j=i}^{k} A_j\right] b - b
\end{aligned}$$

The condition for convergence is $\lim_{k\to+\infty}\prod_{t=0}^{k} A_t \to \mathbf{0}$, which requires $\rho(A_t) = \rho(\mathbf{I} - \alpha\mathbf{I} + \alpha\gamma\mathbf{P_{gq_t}}) < \mathbf{1}, \forall \mathbf{t} \in \mathbf{T}$, where $\rho$ is the spectral radius. Next, we argue that this condition is satisfied in general. The above condition equals to $\rho(\alpha\gamma\mathbf{P}_{g\mathbf{q}_t}) < \alpha, \forall t \in T$ as the eigenvalue of $\mathbf{I}$ is 1. Based on the fact that $\mathbf{P}_{g\mathbf{q}_t}$ is a Markov matrix whose largest eigenvalue is 1, the condition required for $\rho(\alpha\gamma\mathbf{P}_{g\mathbf{q}_t}) < \alpha, \forall t \in T$ is $\gamma < 1$ and $\gamma < 1$ is usually satisfied in general.

## 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).**

$$\begin{aligned} v_{k+1}(s) &= v_k(s) + \alpha\left(\max_b\left[r(s,b) + \gamma\sum_{s'} p(s'|s,b)v_k(s')\right] - v_k(s)\right) \\ &= (1-\alpha)v_k(s) + \alpha\left(\max_b\left[r(s,b) + \gamma\sum_{s'} p(s'|s,b)v_k(s')\right]\right) \\ &= (1-\alpha)v_k(s) + \alpha\max_b\left(q_k(s,b)\right) \\ &= (1-\alpha)v_k(s) + \alpha\mathbf{P}_{gq_k}q_k(s,b) \end{aligned}$$

$$v_{k+1}(s) = (1-\alpha)v_k(s) + \alpha\mathbf{P}_{gq_k}q_k(s,b)$$

$$\begin{bmatrix} v_{k+1}(s_0) \\ v_{k+1}(s_1) \\ \vdots \end{bmatrix} = (1-\alpha)\begin{bmatrix} v_k(s_0) \\ v_k(s_1) \\ \vdots \end{bmatrix} + \alpha\begin{bmatrix} p_{gq_k}(s_0,a_0|s_0) & p_{gq_k}(s_0,a_1|s_0) & \dots \\ p_{gq_k}(s_0,a_0|s_1) & p_{gq_k}(s_0,a_1|s_1) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}\begin{bmatrix} q_k(s_0,a_0) \\ q_k(s_0,a_1) \\ \vdots \end{bmatrix}$$

## 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 $\pi$ 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 $v_\pi$ or show calculation. (5 points).

$$\begin{aligned} r_\pi(0) &= \pi(0|0) * r(0,0) + \pi(1|0) * r(0,1) = 0.25 * 0 + 0.75 * 0 = 0 \\ r_\pi(1) &= \pi(0|1) * r(1,0) + \pi(1|1) * r(1,1) = 0.75 * 0 + 0.25 * 1 = 0.25 \\ p_\pi(0|0) &= \pi(0|0) * p(0|0,0) + \pi(1|0) * p(0|0,1) = 0.25 * 1 + 0.75 * 0 = 0.25 \\ p_\pi(1|0) &= 1 - p_\pi(0|0) = 0.75 \\ p_\pi(0|1) &= \pi(0|1) * p(0|1,0) + \pi(1|1) * p(0|1,1) = 0.75 * 1 + 0.25 * 0 = 0.75 \\ p_\pi(1|1) &= 1 - p_\pi(0|1) = 0.25 \end{aligned}$$

With $p$ and $r$, we can write $v_\pi$ in the matrix form $\mathbf{v}_\pi = \mathbf{r}_\pi + \gamma\mathbf{P}_\pi\mathbf{v}_\pi$ and solve the linear system to obtain $v_\pi$:

$$\begin{bmatrix} v_\pi(0) \\ v_\pi(1) \end{bmatrix} = \begin{bmatrix} r_\pi(0) \\ r_\pi(1) \end{bmatrix} + \gamma\begin{bmatrix} p_\pi(0|0) & p_\pi(1|0) \\ p_\pi(0|1) & p_\pi(1|1) \end{bmatrix}\begin{bmatrix} v_\pi(0) \\ v_\pi(1) \end{bmatrix}$$

Since $\mathbf{I} - \gamma\mathbf{P}_\pi$ is invertible, we can get the exact the solution: $\mathbf{v}_\pi = (\mathbf{I} - \gamma\mathbf{P}_\pi)^{-\mathbf{1}}\mathbf{v}_\pi$, we have $v_\pi(0) \approx 1.16, v_\pi(1) \approx 1.34$.

## 17 For the same MDP, find $q_\pi$, mention the method you used for finding $q_\pi$ or show calculation, and describe $gq_\pi$, that is the greedy policy over $q_\pi$ for the policy $\pi$ described in the previous question (5 points).

Similar to Q16, we can obtain the matrix form of $q_\pi$ and solve the linear system:

$$\begin{aligned} q_\pi(s,a) &= \sum_{s',r} p(s',r|s,a)\left[r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a')\right] \\ &= \sum_{s',r} p(s',r|s,a)r + \gamma \sum_{s',r} p(s',r|s,a) \sum_{a'} \pi(a'|s') q_\pi(s',a') \\ &= \sum_{r} p(r|s,a)r + \gamma \sum_{s'} \sum_{a'} \pi(a'|s') p(s'|s,a) q_\pi(s',a') \\ &= r(s,a) + \sum_{a',s'} p_\pi(s',a'|s,a) q_\pi(s',a') \end{aligned}$$

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

$$\begin{bmatrix} q_\pi(0,0) \\ q_\pi(0,1) \\ \vdots \end{bmatrix} = \begin{bmatrix} r_\pi(0,0) \\ r_\pi(0,1) \\ \vdots \end{bmatrix} + \gamma \begin{bmatrix} p_\pi(0,0|0,0) & p_\pi(0,1|0,0) & \dots \\ p_\pi(0,0|0,1) & p_\pi(0,1|0,1) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} q_\pi(0,0) \\ q_\pi(0,1) \\ \vdots \end{bmatrix}$$

Here $\mathbf{P}_\pi$ is not invertible but it's still a contraction mapping, so we will use fixed point iteration to get the solutions here: $q_\pi(0,0) \approx 1.05, q_\pi(0,1) \approx 1.20, q_\pi(1,0) \approx 1.05, q_\pi(1,1) \approx 2.20$. The greedy policy $gq_\pi$ is $gq_\pi(0|0) = 0, gq_\pi(1|0) = 1, gq_\pi(0|1) = 0, gq_\pi(1|1) = 1$.

## 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, x is the vector input, $\theta$ is the output weight vector, W is the input weight matrix, and $\phi$ 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).

As this is a regression problem, we use Mean Squared Error $\left(Y - \theta^\top \phi(\mathbf{W}\mathbf{x})\right)^2$as the objective:

$$\begin{aligned} \mathbf{W}_{t+1} &= \mathbf{W}_t - \alpha_t \nabla_{\mathbf{W}_t} (Y_t - \theta^\top \phi(\mathbf{W}_t \mathbf{x}_t))^2 \\ &= \mathbf{W}_t + 2\alpha_t (Y_t - \theta^\top \phi(\mathbf{W}_t \mathbf{x}_t)) \nabla_{\mathbf{W}_t} \theta^\top \phi(\mathbf{W}_t \mathbf{x}_t) \\ &= \mathbf{W}_t + 2\alpha_t (Y_t - \theta^\top \phi(\mathbf{W}_t \mathbf{x}_t)) \theta \nabla_{\mathbf{W}_t} \phi(\mathbf{W}_t \mathbf{x}_t) \\ &= \mathbf{W}_t + 2\alpha_t (Y_t - \theta^\top \phi(\mathbf{W}_t \mathbf{x}_t)) \theta \phi(\mathbf{W}_t \mathbf{x}_t) \circ (1 - \phi(\mathbf{W}_t \mathbf{x}_t)) \circ \nabla_{\mathbf{W}_t} \mathbf{W}_t \mathbf{x}_t \\ &= \mathbf{W}_t + 2\alpha_t (Y_t - \theta^\top \phi(\mathbf{W}_t \mathbf{x}_t)) \theta \phi(\mathbf{W}_t \mathbf{x}_t) \circ (1 - \phi(\mathbf{W}_t \mathbf{x}_t)) \circ \mathbf{x}_t^\top \end{aligned}$$

where $\circ$ indicates elementwise multiplication