Note - These are practice questions and in no way contribute to your grade in the course.

Mathematical expressions

For answers that include mathematical expressions, you can type your answer in plain text math as best you can or you can attempt to render LaTeX by including double dollar signs ($$) before and after your expression. Either way is totally fine.

For example, if the answer is the expression for mean squared error of linear regression, $\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)})^2$, you can something like one of the following:

Plain text:

```
1/N sum i=1 to N (y^i - theta^T x^i)^2
```

Rendered:

```
$$\frac{1
```

which renders as $\frac{1}{N}\sum_{i=1}^{N}($

Q2 Multiple Choice

9 Points

Q2.1 Select the best choice

2 Points

Let $V_k(s)$ indicate the value of state s at iteration k in (synchronous) value iteration.

What is the relation ship between $V_{k+1}(s)$ and $\sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V_k(s')]$, for any $a \in$ actions?

Please indicate the most restrictive relationship that applies. For example, if $x < y$ always holds, please use $<$ instead of $\leq$. Selecting ? means it's not possible to assign any true relationship.

Select the best choice

$V_{k+1}(s)$ ☐ $\sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V_k(s')]$

- ○ $=$
- ○ $<$
- ○ $>$
- ○ $\leq$
- ◉ $\geq$
- ○ ?

Q2.2 Select the best choice

2 Points

Let $Q(s,a)$ indicate the estimated Q-value of state-action pair (s,a) at some point during Q-learning. Now your learner gets reward r after taking action a at state s and arrives at state s'. Before updating the Q values based on this experience

$r + \gamma \max_{a'}$

Please indicate the most restrictive relati or example, if $x <$ y always holds, please use $<$ instead of it's not possible to assign any true relationship.

Select the best choice

$Q(s,a)$ ☐ $r + \gamma \max_{a'} Q(s',a')$

- ○ $=$
- ○ $<$
- ○ $>$
- ○ $\leq$
- ○ $\geq$
- ◉ ?

Q2.3 Select all that apply

2 Points

During standard (not approximate) Q-learning, you get reward r after taking action $North$ from state A and arriving at state B. You compute the sample $r + \gamma Q(B, South)$, where $South = \arg\max_a Q(B, a)$.

Which of the following Q-values are updated during this step?

Select all that apply

- [x] Q(A, North)
- [ ] Q(A, South)
- [ ] Q(B, North)
- [ ] Q(B, South)
- [ ] None of the above

Q2.4 True/False

3 Points

In general, for Q-Learning (standard/tabular Q-learning, not approximate Q-learning) to converge to the optimal Q-values, which of the following are true?

True or False: It is necessary that every state-action pair is visited infinitely often.

- (●) True
- ( ) False

EXPLANATION

In order to ensure convergence in general for Q learning, this has to be true. In practice, we generally care about the policy, which converges well before the values do, so it is not necessary to run it infinitely often.

True or False: It is necessary that the discount γ is less than 0.5.

○ True

◉ False

EXPLANATION

The discount factor must be greater than 0 and less than 1, not 0.5.

True or False: It is necessary that actions get chosen according to $\arg\max_a Q(s, a)$.

○ True

◉ False

EXPLANATION

This would actually do rather poorly, because it is purely exploiting based on the Q-values learned thus far, and not exploring other states to try and find a better policy.

Q3 Logistic Regression

16 Points

Q3.1 Conditional Likelihood

4 Points

Given the following dataset, $\mathcal{D}$, and a fixed parameter vector, $\boldsymbol{\theta}$, write an expression for the binary logistic regression conditional likelihood.

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)} = 0), (\mathbf{x}^{(2)}, y^{(2)} = 0), (\mathbf{x}^{(3)}, y^{(3)} = 1), (\mathbf{x}^{(4)}, y^{(4)} = 1)\}$$

- Write your answer in terms of $\boldsymbol{\theta}$, $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$, and $\mathbf{x}^{(4)}$.
- Do not include $y^{(1)}$, $y^{(2)}$, $y^{(3)}$, or $y^{(4)}$ in your answer.
- Don't try to simplify your expression.
- We have provided below the plain text and LaTeX version of the logistic regression hypothesis function, which may help you type up your answers quicker.
- In your answer, you don't have to worry about bold text or parentheses in the superscript. For example `x^1` rather than `$$\mathbf{x}^{(i)}$$`.

Plain text hypothesis function for input $\mathbf{x}$:

```
1/(1 + exp(-theta^T * x))
```

LaTeX hypothesis function for input $\mathbf{x}$:

```
$$\frac{1}{1 + e^{-\theta^T x}}$$
```

Conditional likelihood:

EXPLANATION

$$\left(1 - \frac{1}{1+e^{-\theta^T x^1}}\right)\left(1 - \frac{1}{1+e^{-\theta^T x^2}}\right)\frac{1}{1+e^{-\theta^T x^3}}\frac{1}{1+e^{-\theta^T x^4}}$$

Q3.2 Decision Boundary

4 Points

Write an ex gistic regression with a bias term for two-dimensional input featu $\mathbf{R}$ and parameters b (the intercept parameter).

Assume that the decision boundary occurs when $P(Y = 1 \mid \mathbf{x}, b, w_1, w_2) = P(Y = 0 \mid \mathbf{x}, b, w_1, w_2)$.

Write your answer in terms of x_1, x_2, b, w_1, and w_2.

Decision boundary equation:

EXPLANATION

$$0 = b + w_1 x_1 + w_2 x_2$$

What is the geometric shape defined by this equation?

EXPLANATION

A line.

Q3.3 Decision Boundary

8 Points

We have now feature engineered the two-dimensional input, $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$, mapping it to a new input vector:

$$\mathbf{x} = \begin{bmatrix} 1 \\ {x_1}^2 \\ {x_2}^2 \end{bmatrix}$$

Write an expression for the decision boundary of binary logistic regression with this feature vector and the corresponding parameter vector $\boldsymbol{\theta} = [b, w_1, w_2]^T$.

Assume th $1 \mid x, \boldsymbol{\theta}) = P(Y = 0 \mid x, \boldsymbol{\theta})$.

Write your answer in terms of x_1, x_2, b,

Decision boundary expression:

EXPLANATION

$0 = b + w_1 x_1^2 + w_2 x_2^2$

What is the geometric shape defined by this equation?

EXPLANATION

An ellipse. Probably decent partial credit for circle.

If we add an L2 regularization on $[w_1, w_2]^T$, what happens to **parameters** as we increase the λ that scales this regularization term?

If we add an L2 regularization on $[w_1, w_2]^T$, what happens to the **decision boundary shape** as we increase the λ that scales this regularization term?

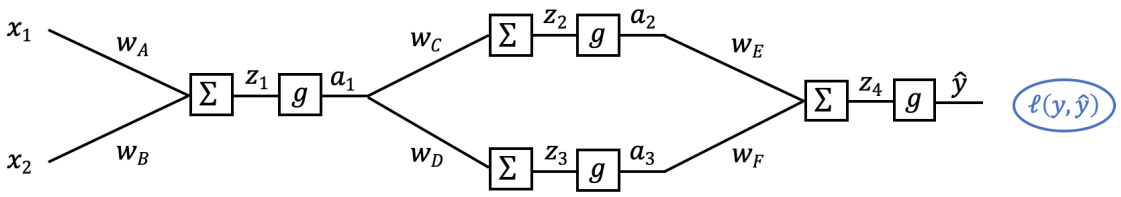
EXPLANATION

The parameters shrink, so the ellipse will get bigger.

Q4 Neural Networks

12 Points

Consider the following neural network for a 2-D input, $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$:



where:

- all g functions are the same arbitrary non-linear activation function with no parameters
- $\ell(y, \hat{y})$ is an arbitrary loss function with no parameters, and:

$z_1 = w_A x_1 + w_B x_2$
$a_1 = g(z_1)$

$$z_2 = w_C a_1$$
$$a_2 = g(z_2)$$

$$z_3 = w_D a_1$$
$$a_3 = g(z_3)$$

$$z_4 = w_E a_2 + w_F a_3$$
$$\hat{y} = g(z_4)$$

Note: There are no bias terms in this network.

Q4.1 Partial derivatives

4 Points

What is the chain of partial derivatives needed to calculate the derivative $\frac{\partial \ell}{\partial w_E}$?

Your answer should be in the form:

Make sure each partial derivative $\frac{\partial ?}{\partial ?}$ in your answer is decomposed further into simpler partial derivatives. **Derivatives.** Be sure to specify the correct subscripts in your answer.

You may write your answer:

In plain text as:
`dl/dwE = d?/d? * d?/d? * ... d?/d?`
$\hat{y}$ can be written as `y_hat`

In LaTeX as:
`$$\frac{d\ell}{dw_E} = \frac{d?}{d?} \frac{d?}{d?} ... \frac{d?}{d?}$$`

Typing d is fine; no need to use `\partial`, ∂
$\hat{y}$ can be written as `$$\hat{y}$$`

where each `?` and the `...` are appropriately replaced.

$\frac{\partial \ell}{\partial w_E} =$

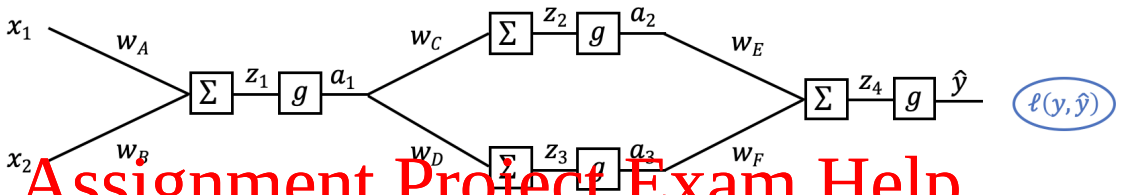
EXPLANATION

$$\frac{\partial \ell}{\partial w_E} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_4} \frac{\partial z_4}{\partial w_E}$$

Q4.2 Partial derivatives

4 Points

The network diagram from above is repeated here for convenience:



What is the the derivative $\frac{\partial \ell}{\partial w_C}$?

Your answer should be in the form:

$$\frac{\partial \ell}{\partial w_C} = \frac{\partial ?}{\partial ?} \frac{\partial ?}{\partial ?}$$

Make sure each partial derivative $\frac{\partial ?}{\partial ?}$ in your answer cannot be decomposed further into simpler partial derivatives. **Do not evaluate the derivatives.** Be sure to specify the correct superscripts in your answer.

$\frac{\partial \ell}{\partial w_C} =$

EXPLANATION

$$\frac{\partial \ell}{\partial w_C} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_4} \frac{\partial z_4}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial w_C}$$

Q4.3 Regularization

4 Points

The gradient descent update step for weight w_C is:

$$w_C \leftarrow w_C - \alpha \frac{\partial \ell}{\partial w_C}$$

where α (alpha) is the learning rate (step size).

Now, we want to change our neural network objective function to add an L2 regularization term on the weights. The new objective is:

$$\ell(y, \hat{y}) + \lambda \frac{1}{2} \|\mathbf{w}\|_2^2$$

where λ (lambda) is the regularization hyperparameter and $\mathbf{w}$ is all of the weights in the neural network stacked into a single vector, $\mathbf{w} = [w_A, w_B, w_C, w_D, w_E, w_F]^T$.

Write the ri ate step for weight w_C given this new objective function. You may use —— in your answer.

Update: $w_C \leftarrow$ ____

EXPLANATION

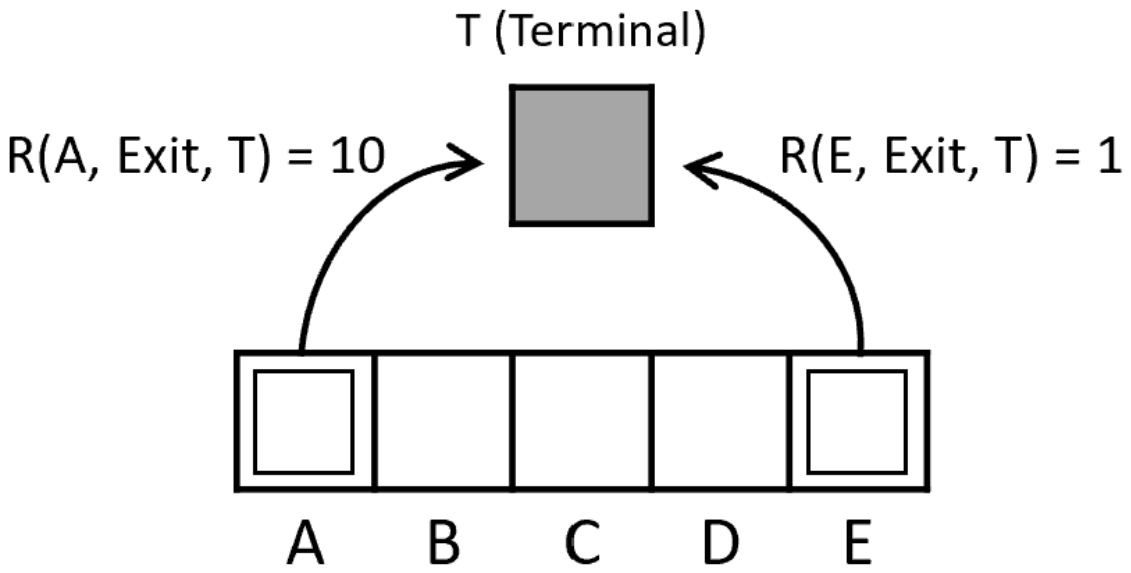
Update for w_C:

$$w_C \leftarrow w_C - \alpha \left(\frac{\partial \ell}{\partial w_C} + \lambda w_C \right)$$

Q5 Value Iteration

8 Points

Consider training a robot to navigate the following grid-based MDP environment.



- There are six states, A, B, C, D, E, and a terminal state T.
- Actions from states B, C, and D are Left and Right.
- The only action from states A and E is Exit, which leads deterministically to the terminal state.

The reward

- R(A, Exit,
- R(E, Exit, T) = 1
- The reward for any other tuple is at 0

Assume the discount factor is just 1.

When taking action Left, with 0.8 probability, the robot will successfully move one space to the left, and with 0.2 probability, the robot will move one space in the opposite direction.

Likewise, when taking action Right, with 0.8 probability, the robot will move one space to the right and with 0.2 probability, it will move one space to the left.

Q5.1

8 Points

Run (synchronous) value iteration on this environment for two iterations. Begin by initializing the value for all states, $V_0(s)$, to zero.

Write the value of each state after the first ($k = 1$) and the second ($k = 2$) iterations. Write your values as a comma-separated list of 6 numerical expressions in the alphabetical order of the states, specifically `V(A), V(B), V(C), V(D), V(E), V(T)`. Each of the six entries may be a number or an expression that evaluates to a number. Do not include any max operations in your response.

There is a space below to type any work that you would like us to consider. Showing work is optional. Correct answers will be given full credit, even if no work is shown.

$V_1(A), V_1(B), V_1(C), V_1(D), V_1(E), V_1(T)$ (values for 6 states):

10,-1,-1,-1,1, 0

$V_2(A), V_2(B), V_2(C), V_2(D), V_2(E), V_2(T)$ (values for 6 states):

10, 0.8, -2, 0.4, 1, 0

What is the optimal policy π_2? Input your answer as a comma-separated list of three actions representing the policy for states, B, C, and D, in that order. Actions may be Left or Right.

$\pi(B), \pi(C), \pi(D)$ based on V_2

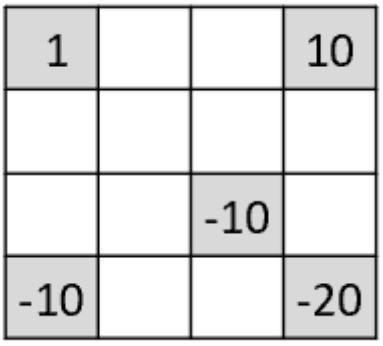
Left, Left, Right

Optional work for this problem:

Q6 MDP Settings and Policies

9 Points

Consider a 4x4 Grid World that follows the same rules as Grid World from lecture.



Specifically:

- The shaded states have only one action, exit, which leads to a terminal state (not shown) and a reward with the corresponding numerical value printed in that state.
- Leaving any other state gives a living reward, $R(s) = r$.
- The agent will travel in the direction of its chosen action with probability $1 - n$ and will travel in one of the two adjacent directions with probability $n/2$ each.
- If the agent travels into a wall, it will remain in the same state.

Match the licies.

Note: We do not expect you to run value iteration to convergence to compute these policies but rather reason about th DP settings.

A)

1	→	←	10
↓	↑	↑	←
↑	←	-10	↑
-10	↑	←	-20

B)

1	←	→	10
↑	↑	↑	↑
↑	↑	-10	↑
-10	↑	←	-20

C)

1	→	→	10
→	↑	↑	↑
↑	↑	-10	↑
-10	↑	←	-20

D)

1	↑	↑	10
←	↑	↑	←
↑	↑	-10	↑
-10	↑	←	-20

E)

1	←	→	10
↑	↑	↑	↑
↑	←	-10	↑
-10	→	←	-20

F)

1	→	↑	10
↓	↑	↑	↑
↑	←	-10	↑
-10	→	←	-20

Q6.1
3 Points

$\gamma = 1.0, n = 0.2, r = 0.1$

- (•) A
- ( ) B
- ( ) C
- ( ) D
- ( ) E
- ( ) F

EXPLANATION

With positive living reward, the agent will try to stay alive as long as possible. With non-zero noise, n, it will avoid negative states if at all possible.

Q6.2

3 Points

$\gamma = 1.0, n$

- ( ) A
- ( ) B
- (•) C
- ( ) D
- ( ) E
- ( ) F

EXPLANATION

With $\gamma = 1$, the policy will travel to the 10.0 state, and with zero noise, n, it doesn't have to worry about slipping sideways into negative states.

Q6.3

3 Points

Figure repeated for convenience:

A)

1	→	←	10
↓	↑	↑	←
↑	←	-10	↑
-10	↑	←	-20

B)

1	←	→	10
↑	↑	↑	↑
↑	↑	-10	↑
-10	↑	←	-20

C)

1	→	→	10
→	↑	↑	↑
↑	↑	-10	↑
-10	↑	←	-20

D)

1	↑	↑	10
←	↑	↑	←
↑	↑	-10	↑
-10	↑	←	-20

E)

1	←	→	10
↑	↑	↑	↑
↑	←	-10	↑
-10	→	←	-20

F)

1	→	↑	10
↓	↑	↑	↑
↑	←	-10	↑
-10	→	←	-20

$\gamma = 0.1, n = 0.2, r = -0.1$

- A
- B
- C
- D
- E
- F

EXPLANATION

With low γ, the policy will prefer the closer 1.0 state, but with non-zero noise, it will avoid negative states if at all possible.