Congratulations! You passed!

Grade received 97.92% **To pass** 80% or higher

Go to next item

1. Which of the following is TRUE about coarse coding? (Select all that apply) **1 / 1 point**

- [x] In coarse coding, generalization occurs between states that have features with overlapping receptive fields.

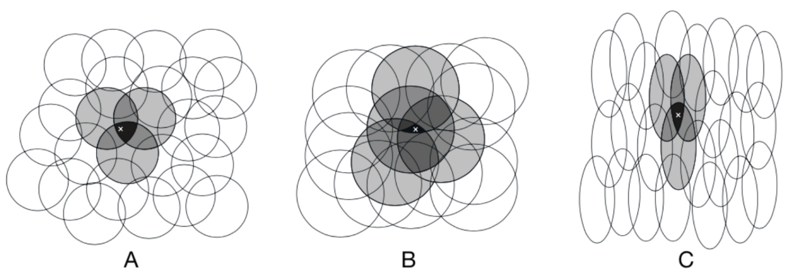
  Correct
  Correct.

- [x] In coarse coding, generalization between states depend on the size and shape of the receptive fields.

  Correct
  Correct.

- [ ] When using features with large receptive fields, the function approximator cannot make discriminations that are finer than the width of the receptive fields.
- [ ] When training at one state, the learned value function will be updated over all states within the intersection of the receptive fields.

2. Consider a continuous two-dimensional state space. Assuming linear function approximation with the coarse-codings in either A, B or C, which of the following is TRUE? (Select all that apply) **1 / 1 point**



- [ ] Generalization is broader in case A as compared to case B.
- [x] In case B, when updating the state marked by an ‘x’ , the value function will be affected for a larger number of states as compared to case A.

  Correct
  Correct. In case B, the receptive fields of the features are larger and include a larger number of states.

- [x] In case C, each update results in more generalization along the vertical dimension, as compared to horizontal dimension.

  Correct
  Correct. Updates to the state marked by the ‘x’ change the values for more states further away in the vertical dimension, as indicated by the greyed areas.

- [ ] In case C, each update results in more generalization along the horizontal dimension, as compared to vertical dimension.

3. Which of the following is TRUE about tile coding? (Select all that apply) **1 / 1 point**

- [x] Tile coding is a form of coarse coding with tiles as its receptive fields.

  Correct
  Correct.

- [x] Tile coding with one tiling is an instance of state aggregation.

  Correct
  Correct.

- [ ] The size of the feature vector equals the number of tilings: each feature corresponds to one tiling.
- [x] With only one tiling, generalization occurs only between states within the same tile.

- [x] With only one tiling, generalization occurs only between states within the same tile.

Correct
Correct.

4. When tile coding is used for feature construction, the number of active or non-zero features

1 / 1 point

- ○ is the number of tiles.
- ◉ is the number of tilings.
- ○ is the number of tilings multiplied by the number of tiles.
- ○ depends on the state.

Correct
Correct.

5. Which of the following is TRUE about neural networks (NNs) ? (Select all that apply)

1 / 1 point

- [x] A NN is feedforward if there are no paths within the network by which a unit's output can influence its input.

Correct
Correct.

- [x] Hidden layers are layers that are neither input nor output layers.

Correct
Correct.

- [ ] The output of the units in NNs are typically a linear function of their input signals.
- [x] NNs are parameterized functions that enable the agent to learn a nonlinear value function of state.

Correct
Correct.

- [x] The nonlinear functions applied to the weighted sum of the input signals are called the activation function.

Correct
Correct.

6. Which of the following is the rectified linear activation function?

1 / 1 point

- ○ $f(x) = \frac{1}{1+e^{-x}}$
- ○ $f(x) = 1$ if $x > 0$ and 0 otherwise
- ◉ $f(x) = max(0, x)$
- ○ $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

Correct
Correct.

7. Which of the following is TRUE about neural networks (NNs)?

1 / 1 point

- ○ A NN with a single hidden layer can represent a smaller class of functions compared to a NN with two hidden layers.
- ◉ The universal approximation property of one-hidden-layer NNs is not true when linear activation functions are used for the hidden layer.
- ○ Given the universal approximation property of one-hidden-layer NNs, there is no benefit to including more layers in the network.

Correct
Correct.

8. Which of the following is TRUE about backpropagation? (Select all that apply)

0.75 / 1 point

- [ ] Backpropagation corresponds to updating the parameters of a neural network using gradient descent.
- [x] Backpropagation involves computing the partial derivatives of an objective function with respect to the weights of the network.

Correct
Correct.

- [ ] The forward pass in backpropagation updates the weights of the network using the partial derivatives computed by the backward passes.
- [x] Backpropagation computes partial derivatives starting from the last layer in the network, to save computation.

Correct
Correct. Because of the nested structure in the neural network, the partial derivatives for earlier layers have some shared components with layers near the output. Starting from the output layer means we can cache some of these shared computations, and avoid needlessly recomputing them.

You didn't select all the correct answers

9. Training neural networks (NNs) with backpropagation can be challenging because (Select all that apply)

1 / 1 point

- [x] the loss surface might have flat regions, or poor local minima, meaning gradient descent gets stuck at poor solutions.

Correct
Correct.

- [x] the initialization can have a big impact on how much progress the gradient updates can make and on the quality of the final solution.

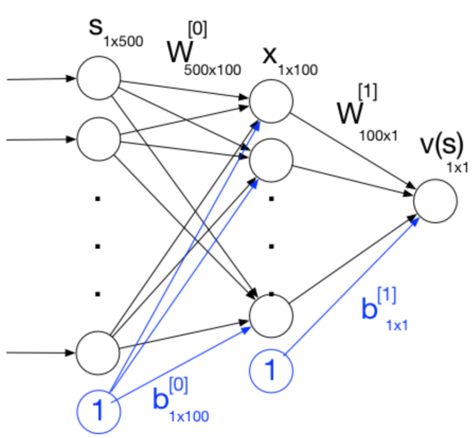
Correct
Correct.

- [ ] neural networks cannot accurately represent most functions, so the loss stays large.
- [x] learning can be slow due to the vanishing gradient problem, where if the partial derivatives for later nodes in the network are zero or near zero then this causes earlier nodes in the network to have small or near zero gradient updates.

Correct
Correct.

10. Consider the following network:

1 / 1 point



where for a given input s, value of s is computed by:

$\psi = sW^{[0]} + b^{[0]}$

$x = max(0, \psi)$

$v = xW^{[1]} + b^{[1]}$

What is the partial derivative of $v(s)$ with respect to $W_{ij}^{[0]}$?

- ( ) s_i
- (•) $W_j^{[1]} s_i$ if $x_j > 0$ and 0 otherwise
- ( ) x_j
- ( ) $W_j^{[1]} x_j$ if $x_j > 0$ and 0 otherwise

Correct

Correct.

11. Which of the following is TRUE? (Select all that apply)

1 / 1 point

- [ ] When using stochastic gradient descent, we often completely eliminate the error for each example.
- [x] The difference between stochastic gradient descent methods and batch gradient descent methods is that in the former the weights get updated using one random example whereas in the latter they get updated based on batches of data.

Correct
Correct.

- [x] Adagrad, Adam, and AMSGrad are stochastic gradient descent algorithms with adaptive step-sizes.

Correct
Correct.

- [x] Setting the step-size parameter for stochastic gradient descent can be challenging because a small step-size makes learning slow and a large step-size can result in divergence.

Correct
Correct.

12. Which of the following is TRUE about artificial neural networks (ANNs)? (Select all that apply)

1 / 1 point

- [ ] It is best to initialize the weights of a NN to large numbers so that the input signal does not get too small as it passes through the network.
- [ ] It is best to initialize the weights of a NN to small numbers so that the input signal does not grow rapidly as it passes through the network.
- [x] If possible, it would be best to initialize the weights of an NN near the global optimum.

Correct
Correct. Then the solution is just a few gradient descent steps away!

- [x] A reasonable way to initialize the NN is with random weights, with each weight sampled from a normal distribution with the variance scaled by the number of inputs to the layer for that weight.

Correct
Correct. This is the initialization strategy we discussed. It is by no means optimal, and how to improve the initialization is the subject of ongoing research.