Quiz 2

● Graded

Student

Alaa Alghwiri

Total Points

5 / 5 pts

Question 1

Gradient Descent Going Past Optimal

1 / 1 pt

✓ **+ 1 pt** Correct

+ 0 pts Incorrect

Question 2

Step size in Gradient Descent

1 / 1 pt

✓ **+ 1 pt** Correct

+ 0 pts Incorrect

Question 3

SGD Convergence

1 / 1 pt

✓ **+ 1 pt** Correct

+ 0 pts Incorrect

Question 4

Step size Linear Function Approximation

1 / 1 pt

✓ **+ 1 pt** Correct

+ 0 pts Incorrect

Question 5

Bias from Optimization

1 / 1 pt

✓ **+ 1 pt** Correct

+ 0 pts Incorrect

Q1 Gradient Descent Going Past Optimal

1 Point

In gradient descent, we can never step past the optimum point, or the sequence of weights w^k will diverge to $\pm\infty$.

○ true

● false

Q2 Step size in Gradient Descent

1 Point

Large step sizes get closer to the optimal weights in stochastic gradient descent while small step sizes stay further away.

○ true

● false

Q3 SGD Convergence

1 Point

In which case, will stochastic gradient descent converge to (locally) optimal weights w^*.

☐ With linear function approximation, a small constant step size, and running for 1 billion iterations.

☐ Any function approximator (linear or nonlinear), small constant step size, averaging of the weights vectors, and 1 billion iterations

☑ Linear regression, decaying step size (proportional to $1/k$), infinite number of iterations.

☐ linear function approximation, convex objective function, constant step sizes, averaging weight vectors, and $n < \infty$ iterations.

☑ linear function approximation, convex objective function, constant step sizes, averaging weight vectors, and infinite iterations.

Q4 Step size Linear Function Approximation

1 Point

For linear regression when using SGD, the amount of change in the output in the function for one iteration is $|f(x_i, w^{k+1}) - f(x_i, w^k)| = \eta|f(x_i, w^k) - y_i| \|x_i\|_2^2$. Which of the following cases below are the correct choices to change η.

□ If every y_i is scaled by some constant $\alpha > 0$, i.e., $\hat{y}_i = y_i\alpha$, then we should use the step size $\eta\alpha$, e.g., if all targets y_i are scaled up by $1,000$ then we should take steps that are $1,000$ times larger.

☑ Assume that initially, $\forall i$, $\|x_i\|_2^2 = 10$ and $\eta = 0.01$ is a good choice. If we add 10 features with a maximum absolute value of 1, then we should set $\eta = \frac{0.1}{20}$ for this new feature vector.

☑ If we rescale each feature vector by a constant $\alpha > 0$, then we should rescale η to be η/α.

☑ If we remove features from the features vector, we will likely need to increase η.

Q5 Bias from Optimization

1 Point

Which of the following are reasons we cannot evaluate how good the model is on the same data we used to find the best weights?

☑ The optimization process adjusts the weights specifically to perform well on the training data, leading to an overestimation of the model's true performance on unseen data.

☑ The model may memorize the training data (overfit), which means it performs well on the training data but poorly on new, unseen data.

☑ In the case of regression, the model might just be predicting the noise and not capturing the true signal underneath. Thus it will look like it does better on the training data than it will on new data.

☑ Using the same data for both training and evaluation leads to a biased estimate of the model's performance, as it doesn't reflect how the model will generalize to new data.

□ It’s impossible to get an estimate of the true before because it requires having infinite data.