Instructions: Make sure your questions and answers are on different pages. Do not include your name or any other identifying information. I will know that information from Canvas.

Question 1: What is the difference between logarithmic loss and sigmoid cross entropy loss?

Question 2: What are the trade-offs between using square loss functions vs absolute loss functions in regression problems?

Question 3: How does focal loss account for foreground-background class imbalance in object detection problems?

Answer Question 1: While logarithmic loss and sigmoid cross entropy loss describe the same loss function and objective function, they are derived differently. The probabilities used in the logarithmic loss are obtained using the conditional probability distribution, while the probabilities in sigmoid cross entropy are obtained using the sigmoid of the outputs.

Answer Question 2: Square loss functions enable fast convergence and high accuracy, since the gradient of the loss function is longer when the predicted value differs from the true value by a lot and shorter when the prediction is pretty close. However, this same behavior makes the model more sensitive to outliers since they will have especially strong gradients that will have a strong impact on the parameters and decrease the accuracy for the rest of the data.

Absolute loss functions eliminate the problem of outlier sensitivity, as outliers will have no stronger gradient than any other datapoint. However, this behavior becomes the downfall of absolute loss as well, since the gradient doesn't decrease as the predictions approach the true values it is much harder for the model to converge to parameters that minimize the loss function.

Answer Question 3: Focal loss accounts for foreground-background class imbalance in object detection problems by introducing a modulating term in the form of $-\tau(1 - \tilde{p})^{\gamma}$ to the sigmoid cross entropy loss function. The piecewise equation for p̃ ensures that it is closer to 1 for correct high confidence predictions, such as the easy case when the model correctly identifies a background outline of an object as not being that object, which is the much more prevalent possible classification. The less prevalent predictions, such as those indicating a correct outline, are lower confidence to begin with as there are much less of them to learn on. This lower confidence results in a lower value of p̃, which then causes the modulation function to be higher. A higher modulation function gives more strength to the loss function's gradient and thus the model will be more affected by these less prevalent foreground cases than the more prevalent background cases.