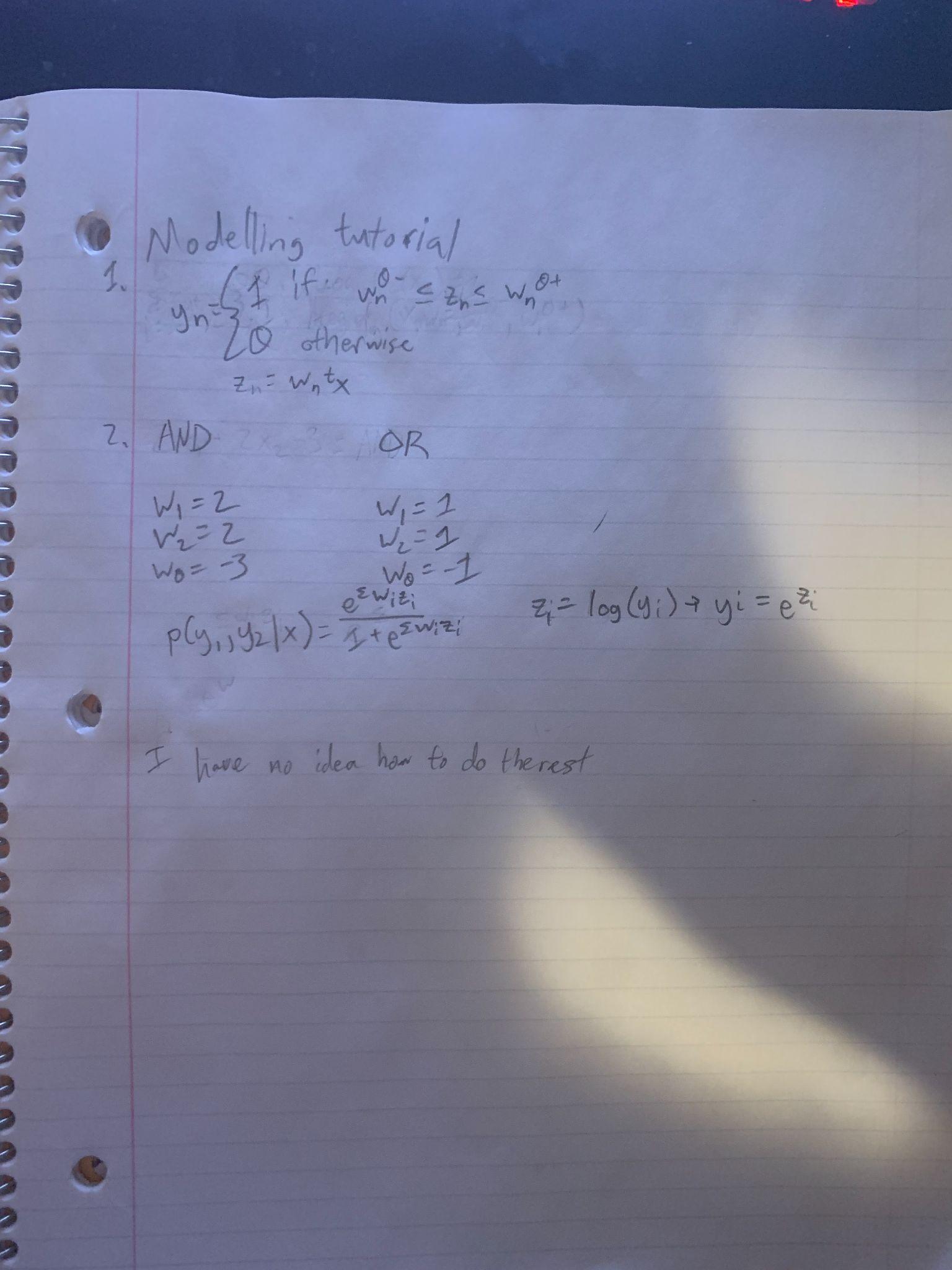
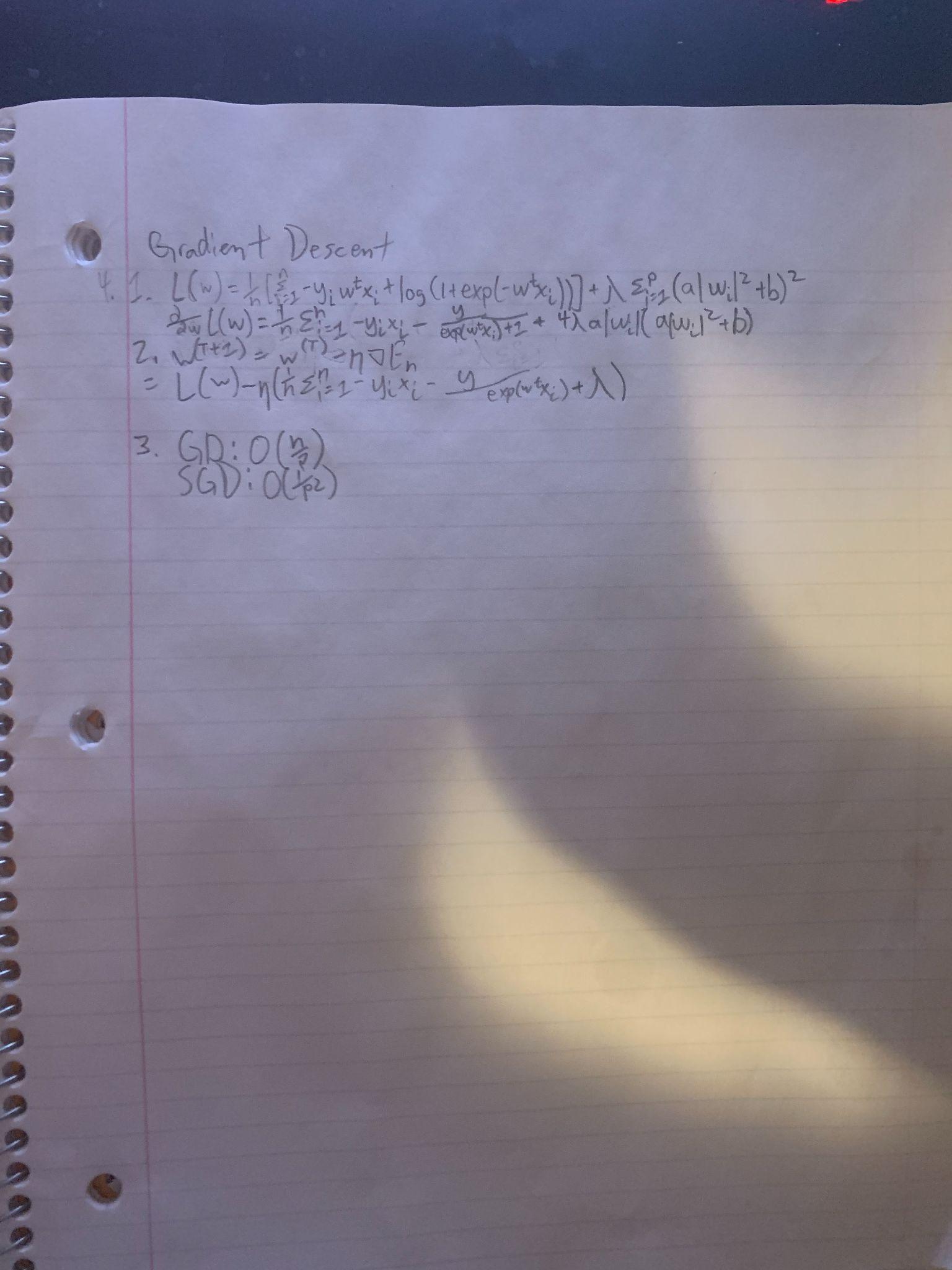
* 1. Where and

where

* 1. The mean of all the errors estimates is bias,
  2. The standard deviation of all the errors estimates is variance
  3. Noise, average difference between test and training set

1. For a squared-loss function, the error is squared. Since outliers have a lot of error, their cost would be squared which leads to an undesirable model. If points are randomly distributed, I think it’s considered variance since random is best represented by variance. While if points are clustered, it would be bias since there would clearly be a factor that’s not variance, meaning it could only be bias.
2. Regularization is needed to control the complexity of the model. If we don’t regularize then we wouldn’t be able to fit our model on data with outliers and on data that’s too complex. Two different regularization methods are ridge and lasso regression. Both are methods that add a cost function, reducing complexity and overfitting.
   1. In HW1, we were given knowledge that it was a Gaussian distribution. With this we were able to use a predefined gaussian function, requiring only the distribution parameters to be plugged in. This works because gaussian distribution is derived from statistics with logical likelihoods for prediction.
   2. If we don’t know the distribution, using a non-parametric model would be best, like KNN. Since these models wouldn’t assume any distribution
   3. The soft margin parameter, , and the Gaussian kernel parameter, , affect bias. Increasing the soft margin parameter causes the margin to get smaller and increases bias.Increasing the gaussian parameter increases bias.Decreasing both would increase variance, lowering bias.
      1. Least-squares regression:
      2. Logistic regression:
      3. Generative gaussian,
      4. SVM,
      5. SVM and Logistic Regression are better compared to others with less training data and large number of features. The others are worse at it.
      6. Performance would be negatively affected by the new data since the model is optimized for the original data
      7. I would scale the coefficients so it could achieve the original likelihood
      8. would need to be adjusted
      9. The weights won’t matter because this is a new dataset
      10. An ensemble would be better because it would reduce bias
      11. The form would matter in case one, but it wouldn’t matter in case two.
3. B, Forming an ensemble would be the obvious choice since it’s aggregating the results of all the models. Simply choosing our model, A would have too much bias. While setting our model to the one with least error would lead to too much overfitting. I’m assuming the training and test data is “normal” where neither are too small. And train, test split is “normal” too, where train is larger than the test.



References:

<https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff>

<https://codesachin.wordpress.com/2015/08/30/cross-validation-and-the-bias-variance-tradeoff-for-dummies/>

<https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>

<https://en.wikipedia.org/wiki/Normal_distribution>

<https://en.wikipedia.org/wiki/Support-vector_machine>

<https://www.cs.toronto.edu/~jlucas/teaching/csc411/lectures/lec9_handout.pdf>

<https://www.robots.ox.ac.uk/~az/lectures/ml/lect3.pdf>

<https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>

<https://stats.stackexchange.com/questions/399318/invariance-of-results-when-scaling-explanatory-variables-in-logistic-regression>