Class: CSE 546 Name: Hongda Li HW2

A.0: Conceptual Questions

A.0.a

False. a large value of w_i can be caused by overfitting. The error will only increase if it's proven that the features of "Number of Bathrooms" is not having collinearity with other features, or it's the last feature that goes to zero as we increase the lambda value for Ridge Regression.
Another way to think about it is, assume that there is another feature says: "The size of bath room", then in that case, this feature is essentially the same as the removed feature: "Number of Bathrooms", removing it will not have a huge impact on the amount of errors for the model.

A.0.b

The L1 norm is more likely because the derivative of w_j is ± 1, which means that it's not related to the actual value of w_j. Therefore, which ever has the most amount of reduction on the error, that w_j is getting pushed to zero.
Or, use professor's Simon's way of doing it, the L1 Norm $||w||_1$ forms a simplex situated around the origin. If the optimal is not inside of the simplex, then the algorithm will always goes to the vertex of the simplex to optimize it, and going to the vertex of the L1 Ball is setting one of the features to zero.

A.0.c

The regularizer $\sum_i |w|^{0.5}$ promotes sparsity because the region $\sum_i |w|^{0.5} = 1$ is very pointy but this is also a bad regularizer because such region is not convex, potentially given multiple solutions, or, making gradient descent slow.

A.0.d

True, assume it's so large thta it just shoots out of the convex region.

A.0.e

It works by randomly sample a batch, and we know that the sampled samples are likely to be representative of the whole sample, and in that way, the gradient computed is pointing, somewhat, in the correct direction.
The dot product between the gradient given by FIRST iteration GSD and the best Gradient direction is always positive. This is true because the first iteration will always pull the w_j into the range where the data is located in.
This is made reasonable by consider extreme choices of sample x_i that gives maximal, or minimal value of predictor, and this quantity will be bounded. Hence, the first few descent step will always pull the model into that range.

A.0.f

SGD is faster to compute compare to GD because it only uses a few samples but it's less likely to have a clear convergence near the optimal because it's random.

A.1: Convexity and Norms