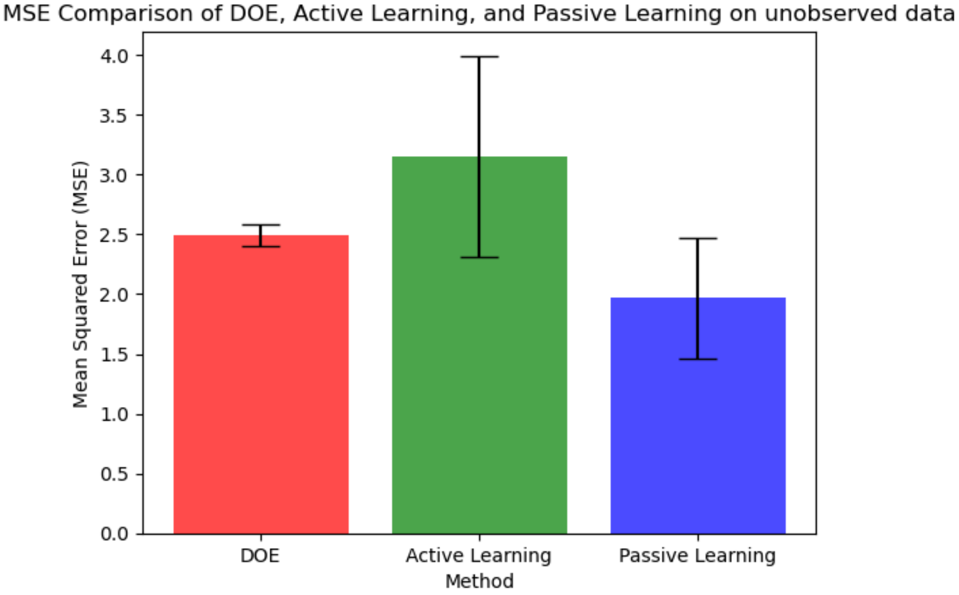
02750 HW 3
Anushka Sinha
March 25, 24

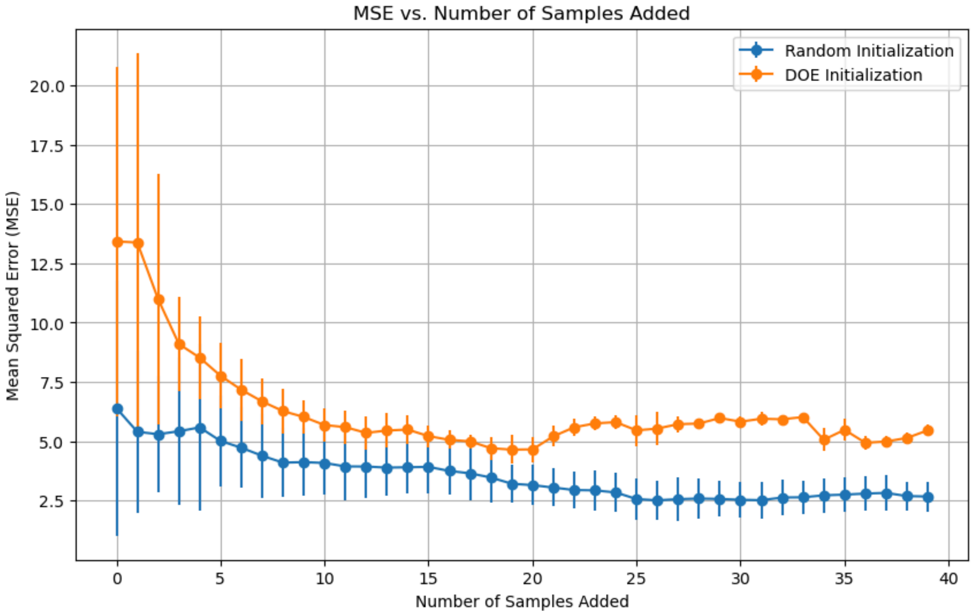
Exercise 1 Active Learning vs Design of Experiments

a. **Describe in a few sentences how you approximate the solution of the DOE strategy and provide corresponding details for that approach. Discuss whether the results matched your expectations and explain your reasoning.**



Since brute force approach of checking every possible set of 30 instances was not possible due to time complexity, one million random sets of 30 instances were generated. The solution of the DOE strategy was approximated using the D-optimality criteria indicated by the equation $max \overset{x_i i \ = k}{\underset{x_i i = 1}{}} |X'X|$ which selects k instances that maximize the determinant of $X'X$. Therefore, the determinant $X'X$ was calculated for all the random sets of 30 instances, and the set of 30 that resulted in the maximum determinant was selected. These 30 instances were used as the training set to train a random forest regressor and to test the rest 70 instances. Similarly, 30 instances were picked using the active learning approach where in each iteration the instance with maximum variance was added to the training set to train a random forest regressor and then test on the remaining 70 instances. Finally, in the passive learning approach, a random instance was chosen to be added to the training set to collect 30 instances and train a random forest regressor to test on the remaining 70 instances. I expected DOE to perform better than the active learning method which is what the plot above shows as the mse for DOE is lower than the active learning method. Since DOE chooses the k-instances that maximize the determinant of the information matrix, it also leads to the confidence ellipsoid having smaller axes. In this way, DOE makes the confidence ellipsoid region more compact which makes it easier to find the true parameter values resulting in a lower mse score than the active learning method.

b. Discuss whether the results matched your expectations and explain your reasoning.



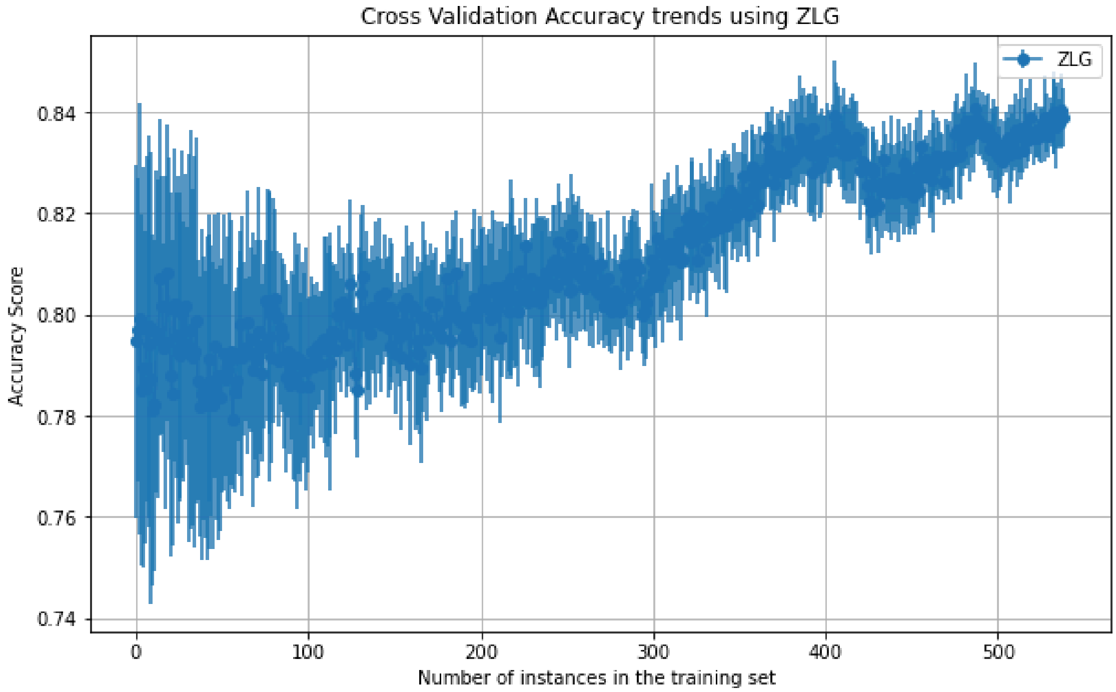
The plot above shows the comparison of mse obtained by using DOE initialization vs random initialization of the 10 instances combined with active learning approach. As the plot shows, the DOE initialization performed worse than the random initialization. This matches my expectations because the DOE implementation was non-deterministic so even if we expected to improve the performance of the model, there is a chance that it might not perform as expected for the specific rounds we run it. There is also a chance that choosing just 10 samples (10% of the instances) is not enough to reduce the range of the confidence ellipsoid region which contains the true values of the parameter which could lead to the model not performing better even after choosing the 10 instances that maximize the determinant of $X'X$.

Exercise 2 Implementing a Type II algorithm

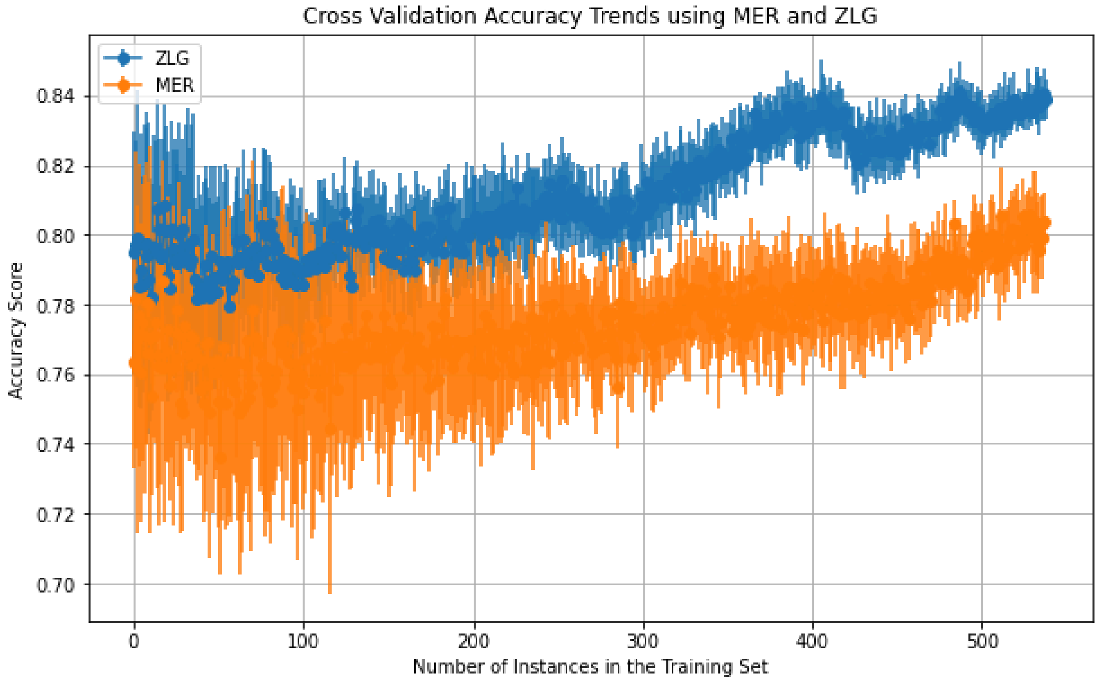
a. Implement ONE of the following three Type II algorithms (ZLG, DH, or PLAL)
Implemented the algorithm: Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions (ZLG algorithm) as described in Zhu et al. (2003),

Implemented ZLG algorithm.

The plot below presents the cross validation accuracy score obtained using ZLG against the number of instances in the training set.



b. **Comment on the results in part 2a compare against any of the Type I algorithms implemented in previous homework assignments. Discuss whether the results matched your expectations and explain your reasoning.**



The plot above compares the cross-validation accuracy score obtained using the ZLG algorithm to the cross-validation accuracy score obtained using the minimization expected risk algorithm

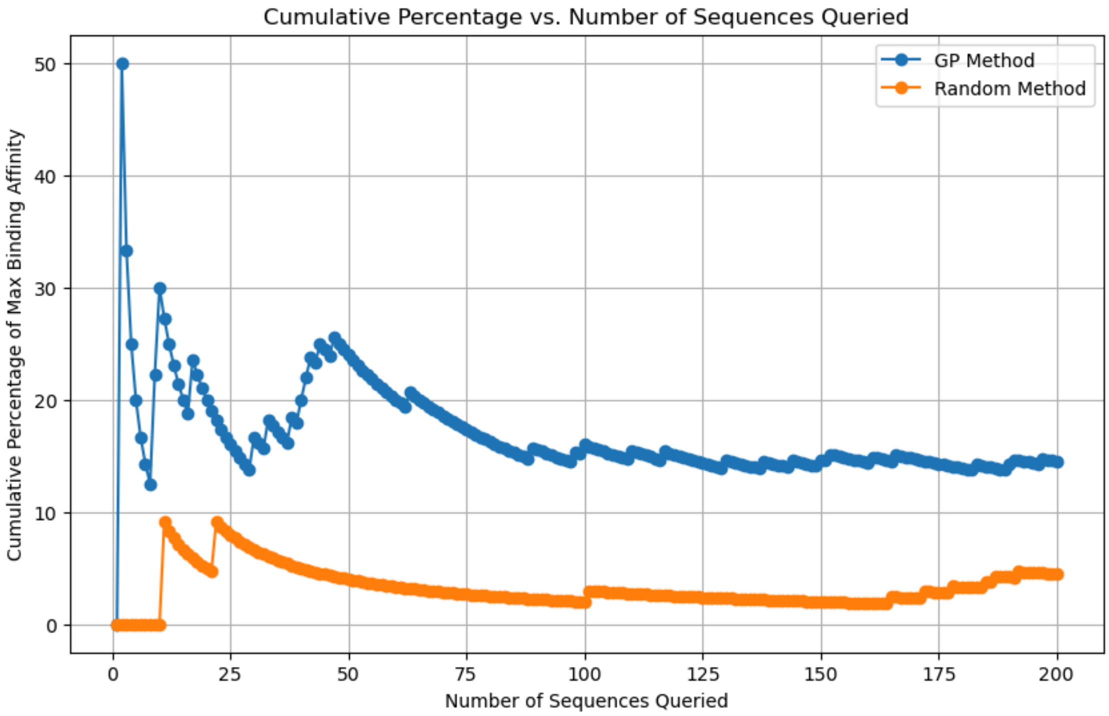
(MER) on the same dataset. According to the plot above, ZLG performs better than the type I MER algorithm which matches my expectation. I expected the type II algorithm, in this case ZLG algorithm, to perform better than the type I algorithm because the type II algorithm exploits the natural clusters found in the dataset which acts as an additional layer of ensuring that the training data includes diverse instances. Since the type II algorithm selects instances from different clusters, it maximizes the information gain in each iteration and improves the generalization ability of the model.

Exercise 3 Sequential Bayesian optimization

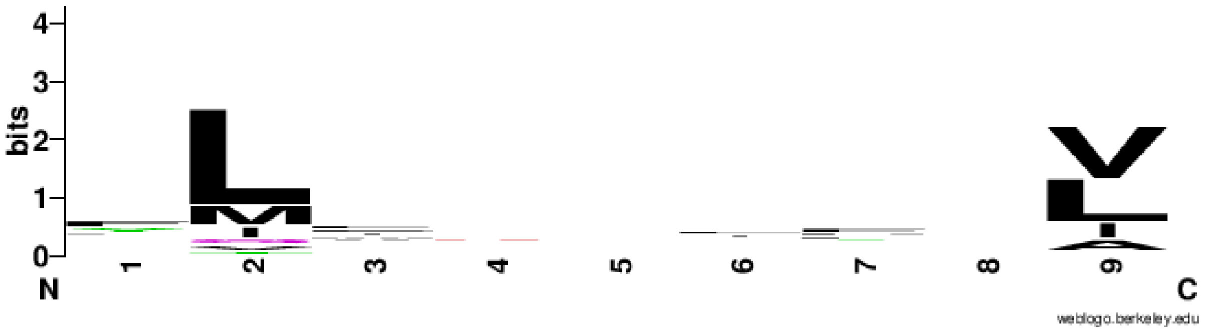
a. **Implement a Bayesian optimizer with Gaussian Process as regressor and use any selection/acquisition function that combines exploitation and exploration. Make sure to concisely describe (i.e., no more than a few sentences), how you defined your selection function**
   The upper confidence bound (UCB) denoted by the equation, $a(x,\ \lambda) = \mu(x) + \lambda\sigma(x)$, was used as the selection function to implement a bayesian optimizer with a gaussian process as a regressor. UCB trades off between exploration and exploitation. UCB is a weighted sum of $\mu(x)$ which is the predicted mean of the Gaussian Process at point x and $\sigma(x)$ which is the standard deviations of the predictions, where λ is a parameter that could be used to tune the tradeoff between exploration and exploitation. In each iteration of Bayesian optimizer with Gaussian Process as a regressor, the instance which maximizes the UCB is added to the training set.

b. **Implement a random query strategy for randomly selecting a sample to query from the data**
   Refer to code. The random query strategy was implemented to select a sample to query from the data randomly using np.random.randint which randomly picks the index of the instances in the dataset.

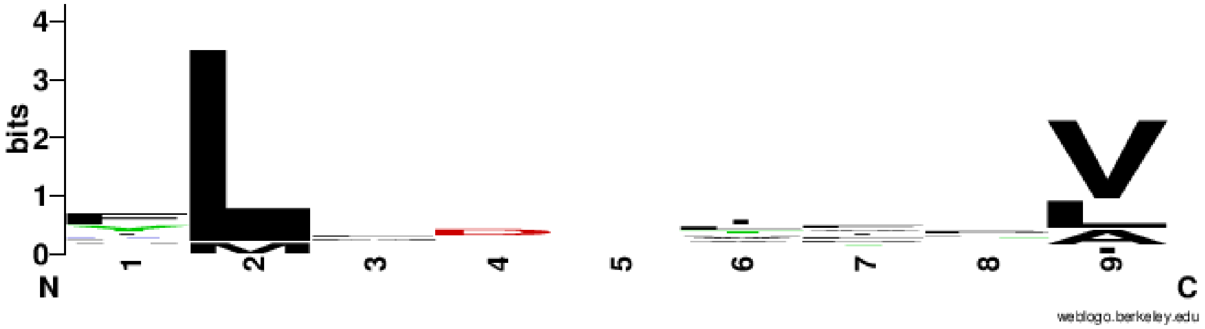
c. **Plot the cumulative percentage of sequences with maximum binding affinity with respect to number of sequences queried for 3a and 3b**
   The plot below represents the cumulative percentage of sequences with maximum binding affinity with respect to the number of sequences queried for both the GP method and random query strategy.



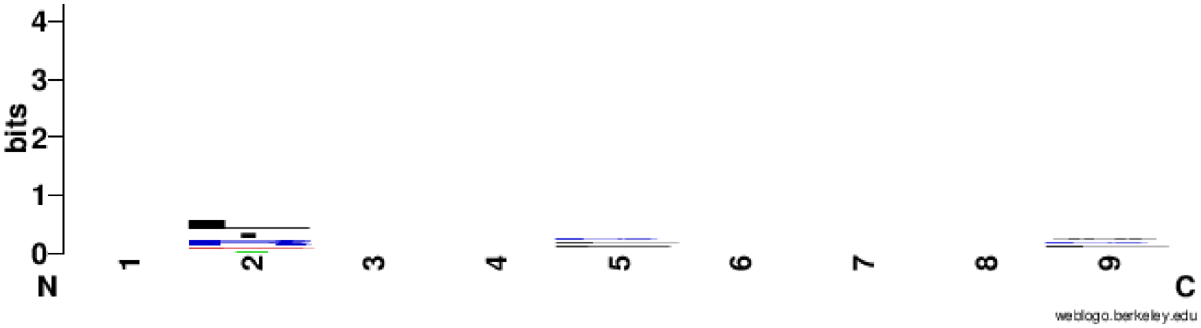
d. Create sequence logo based on sequence found with each querying strategy.



Seqlogo for all sequences with affinity = 9.0



Seqlogo for the sequences with affinity = 9.0 using the GP method



Seqlogo for the sequences with affinity = 9.0 using random query strategy

e. **Compare the approaches in parts 3a and 3b. Discuss whether the results matched your expectations and explain your reasoning**

The plot presented in 3c represents the cumulative percentage of sequences with max binding affinity found using GP method and random query strategy against the number of total sequences queried by both methods. According to the plot, it is clear that the GP method does better than random query strategy in querying sequences that have max binding affinity (affinity = 9.0) for the same number of total sequences queried. This better performance is reflected in the seqlogo of the sequences with max binding affinity found using the GP method which accurately finds the conservation of amino acid “L” and “V” at their respective position which is essential for achieving maximum binding affinity whereas the seqlogo of the sequences with max binding affinity found using random query strategy shows slight conservation of the amino acid “L” but nothing for the amino acid “V”. These results match my expectations because Bayesian optimization using Gaussian Process uses a selection function, in this case UCB, to strategically select the optimal instance to query next by balancing the need for exploration and exploitation. Since random query strategy just randomly selects the next instance to query, it does not allow the regressor to focus on the promising areas in the search space as it does not take advantage of the previous knowledge of the model or the uncertainty in the predictions.