Differences in estimated coefficients/cost function between batch gradient decent and stochastic gradient decent

The differences between estimated coefficients and cost functions obtained with batch gradient descent and stochastic gradient descent can be attributed to the inherent characteristics of these two optimization algorithms. BGD computes the gradient of the cost function with respect to the parameters using the entire dataset in each iteration. It aims to find the global minimum by considering the average gradient over the entire dataset. SGD updates the parameters based on the gradient of the cost function for a single data point chosen randomly in each iteration. It introduces more randomness and noise in parameter updates, making it less likely to converge to the global minimum but often more adaptable to variations in the data, yet it comes at a smaller computational cost. The nature of these optimizations impacts the convergence behavior, with BGD producing smoother updates and SGD exhibiting more oscillations, something which can be observed when plotting the cost function against the number of iterations. Learning rate plays a crucial role, with BGD applying it to the average gradient and SGD applying it to individual data points, influencing the stability and variability of updates. While BGD is computationally expensive, especially for large datasets, SGD offers greater efficiency by processing one data point at a time.

Overall, in BGD, the cost function is generally smoother and exhibits a more regular decrease over iterations. The updates to the parameters are based on the average gradient computed from the entire dataset, leading to a more stable convergence. As a result, the cost function graph tends to show a gradual and continuous decrease, indicating a systematic approach towards the global minimum. Whilst the cost function plot for SGD is characterized by more irregular fluctuations and oscillations. This behavior arises from the fact that SGD updates parameters based on the gradient of individual, randomly selected data points. While this introduces more variability and noise in the updates, it can lead to faster convergence in terms of iterations.

Both algorithms are sensitive to the choice of learning rate, and the cost function plots can be affected accordingly. If the learning rate is too large, both BGD and SGD might exhibit overshooting or divergence, while too small of a learning rate can slow down convergence.