1.1

$$\Phi_{\mathrm{S}}^{*}(x) = Ax$$

2.1

$$\Phi(\mathrm{x_q}) = y_q \text{ for } q = 1, \dots, Q$$

2.2

(7) is a nonsensical formulation of AI since it induces a solution to just memorize the training set. The optimal solution would need $\Phi(\mathrm{x_q}) = y_q$ for all training samples, but it gives no rule for predicting outputs when the input x lies outside the training set. Because the purpose of AI is to generalize beyond the observed data, the unrestricted ERM formulation does not provide a meaningful notion of learning.

3.1

$$H^{*} = YX^{\top}(XX^{\top})^{-1}$$

Loss over training set: 0.4470

Loss over test set: 0.5590

3.2

$$H^{*} = YX^{\top}(XX^{\top})^{-1}$$

Loss over training set: 72.0268

Loss over test set: 89.2371

3.3

$$H^{*} = YX^{\top}(XX^{\top})^{-1}$$

Loss over training set: 0.02

Loss over test set: 1.12

3.4

In Q3.1, the training and test errors are both small because the linear model matches the data and we have plenty of samples. In Q3.2, both errors stay large since a linear model can't capture the nonlinear sign process. In Q3.3, the training error is small but the test error is large because the model is correct but the dataset is too small, so it overfits and fails to generalize.

4.1

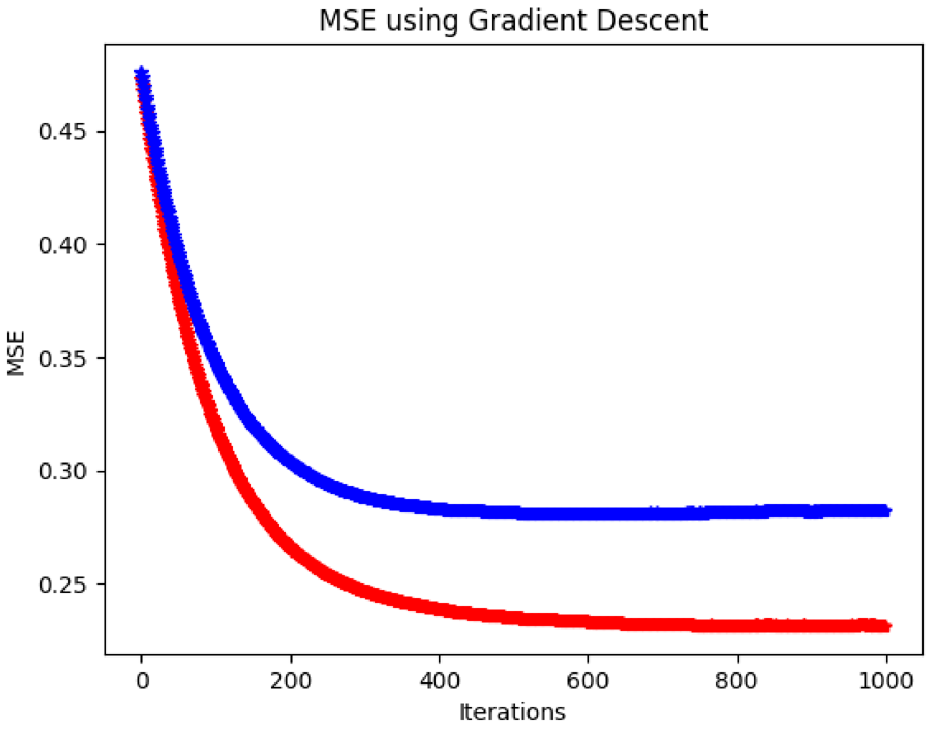
Stochastic Gradient:

$$\widehat{\nabla} L(H) = \frac{1}{Q_t} \sum_{(x_q, y_q) \in T_t} (Hx_q - y_q) x_q^\top$$

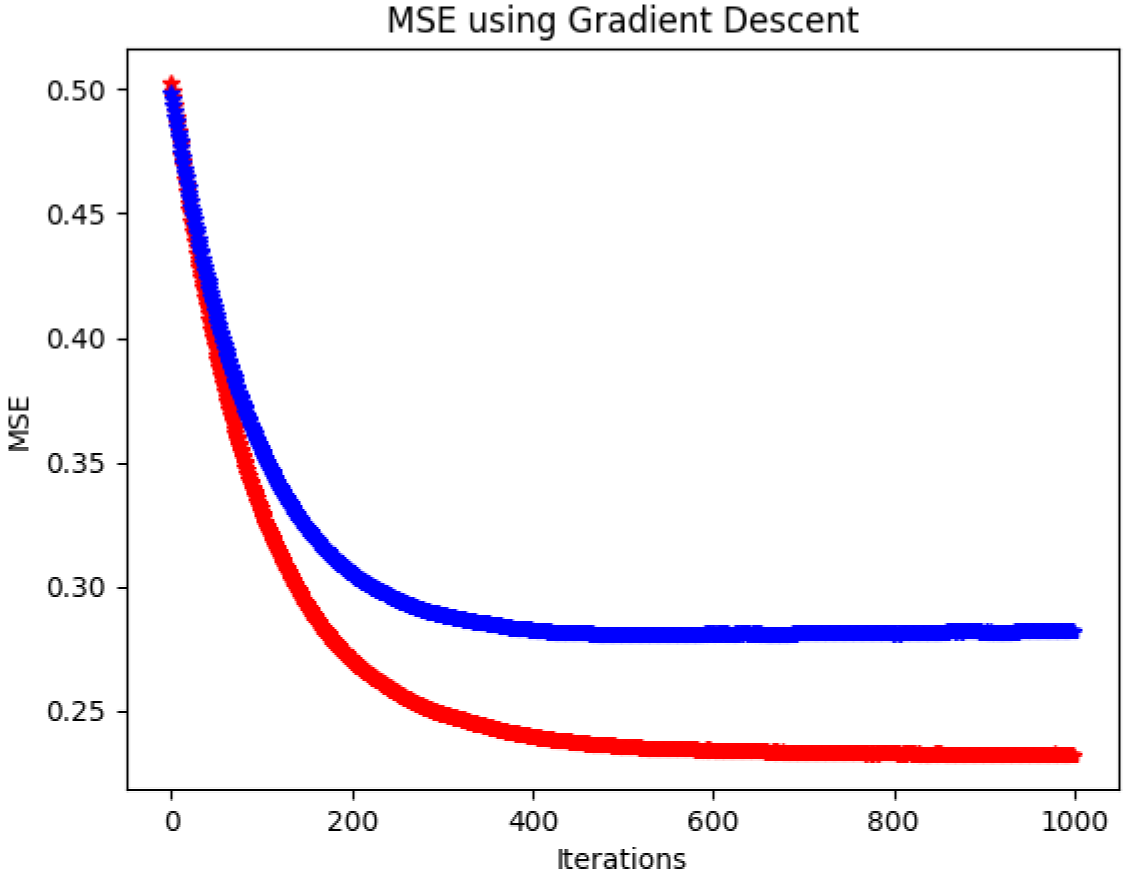
SGD Recursion:

$$H_{t+1} = H_t - \epsilon \cdot \frac{1}{Q_t} \sum_{(x_q, y_q) \in T_t} (Hx_q - y_q) x_q^\top$$

4.2



4.3


MSE using Gradient Descent
MSE
0.50
0.45
0.40
0.35
0.30
0.25
0
200
400
600
800
1000
Iterations