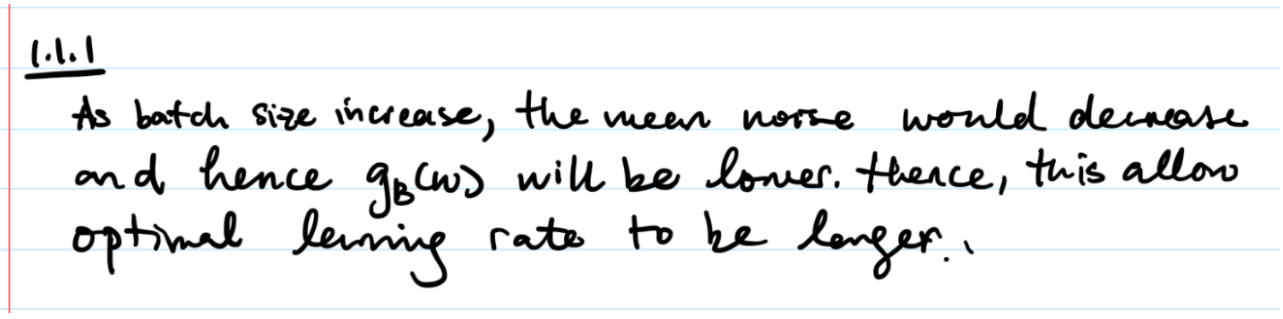
CSC413 Homework 3

Name: Kwan Kiu Choy
Student Id: 1005005879
Utorld: choykwan

1.1.1



1.1.2

1.1.2

a) C is the most efficient batch size. It is big enough such that the mean noise is reduced and the training time to be significantly reduced. B is not a good option as the training time did not decrease too much even though B is much larger than C. This means that the increase in computation for processing a larger batch is not worthy.

b) Point A: noise dominated
Point B: curvature dominated

1.1.3

1.1.3

a) I, III+, II-, IV

b) II+, III-

1.2

1.2

a) A contains more parameters because

① It takes longer time for the model to reach a converged test error

② Once the test error starts to drop, the test error will be lower than the model with less parameters since the biases in A is smaller

ii) B would have gone through more updates At X, A and B have the same total compute. Since we know that B has less parameters than A, then B would have more training steps than A.

b) If we have unlimited time, i.e: no deadline, it would be desirable to train it with A together with regularization to prevent overfitting because the biases would be smaller and at the same time preventing overfitting. However, if this is to be done in a time sensitive manner, then it would be more desirable to use B, since B takes less time to reach an accuracy of X. (i.e: The total compute would be smaller than that of A).

2.1.1

2.1.1

$$\mathbb{E}(w_*^T \tilde{x} - \hat{w}^T \tilde{x})^2$$

$$\hat{w} = (X^TX)^{-1}X^T(Xw_* + \epsilon) = \cancel{(X^TX)^{-1}X^TXw_*} + (X^TX)^{-1}X^T\epsilon$$

$$= w_* + (X^TX)^{-1}X^T\epsilon$$

$$\hat{w}^T = w_*^T + \epsilon^T X(X^TX)^{-1} \longrightarrow (*)$$

$$(w_*^T\tilde{x} - \hat{w}^T\tilde{x})^2$$

$$= w_*^T\tilde{x}\, w_*^T\tilde{x} - w_*^T\tilde{x}\,\hat{w}^T\tilde{x} - \hat{w}^T\tilde{x}\, w_*^T\tilde{x} + \hat{w}^T\tilde{x}\,\hat{w}^T\tilde{x}$$

$$= \cancel{w_*^T\tilde{x}\, w_*^T\tilde{x}} - w_*^T\tilde{x}\, w_*^T\tilde{x} - w_*^T\tilde{x}\epsilon^T X(X^TX)^{-1}\tilde{x} - \hat{w}^T\tilde{x}\, w_*^T\tilde{x} + \hat{w}^T\tilde{x}\,\hat{w}^T\tilde{x}$$

$$= -w_*^T\tilde{x}\epsilon^T X(X^TX)^{-1}\tilde{x} - w_*^T\tilde{x}\, w_*^T\tilde{x} - \epsilon^T X(X^TX)^{-1}\tilde{x}\, w_*^T\tilde{x} + \hat{w}^T\tilde{x}\,\hat{w}^T\tilde{x}$$

$$= \cancel{-w_*^T\tilde{x}\epsilon^T X(X^TX)^{-1}\tilde{x}} - \cancel{w_*^T\tilde{x}\, w_*^T\tilde{x}} = \cancel{\epsilon^T X(X^TX)^{-1}\tilde{x}\, w_*^T\tilde{x}} + \cancel{w_*^T\tilde{x}\, w_*^T\tilde{x}} + \cancel{w_*^T\tilde{x}\epsilon^T X(X^TX)^{-1}\tilde{x}} + \cancel{\epsilon^T X(X^TX)^{-1}\tilde{x}\, w_*^T\tilde{x}} + \epsilon^T X(X^TX)^{-1}\tilde{x}\,\epsilon^T X(X^TX)^{-1}\tilde{x}$$

$$\mathbb{E}(\epsilon^T X(X^TX)^{-1}\tilde{x}\,\epsilon^T X(X^TX)^{-1}\tilde{x})$$

$$= \mathbb{E}(\epsilon^T X(X^TX)^{-1}\tilde{x})^2$$

$$= Tr(\mathbb{E}(\epsilon^T X(X^TX)^{-1}\tilde{x})^2)$$

$$= Tr(\mathbb{E}(\tilde{x}^T(X^TX)^{-1}X^T\epsilon\,\epsilon^T X(X^TX)^{-1}\tilde{x}))$$

By making use of the cyclic property of trace,

$$= Tr(\mathbb{E}(\tilde{x}^T\tilde{x})\ \mathbb{E}(\epsilon\epsilon^T)(X^TX)^{-1}X^TX(X^TX)^{-1})$$

$$= Tr(Id\ \sigma^2(X^TX)^{-1})$$

$$= \sigma^2 Tr((X^TX)^{-1})$$

2.2.1

2.2.1

Underparameterized case: $(n > d)$

$$\mathbb{E}(\sigma^2 \mathrm{Tr}(X^T X)^{-1})$$

$$= \sigma^2 \mathbb{E}(\mathrm{Tr}(X^T X)^{-1})$$

$$= \sigma^2 \frac{d}{n-d-1} //$$

Overparameterized case: $(d > n)$

$$\mathbb{E}\left(\frac{1}{d}\mathrm{Tr}(I_d - X^T(XX^T)^{-1}X) + \sigma^2 \mathrm{Tr}((XX^T)^{-1})\right)$$

$$= \frac{1}{d}(d-n) + \sigma^2 \mathbb{E}(\mathrm{Tr}((XX^T)^{-1}))$$

$$= \frac{1}{d}(d-n) + \sigma^2 \frac{n}{d-n-1} //$$

2.2.2

2.2.2

(1) For underparameterized model, perfect generalization can be reached if $\sigma = 0$.
For overparameterized model, perfect generalization can be reached if $\sigma = 0$ and $d - n = 0$.

(2) For underparametrized model, the denominator of the fraction $(\frac{d}{n-d-1})$ will be longer and hence the value of $\mathbb{E}[R(\hat{w})]$ will be smaller and closer to 0. For overparameterized model, increasing the training data would not always help generalizations as the variance terms would increase as its numerator = number of training data.

2.3.2

2.3.2

As training data n decrease and noise level σ increase, λ should increase. This is because a larger λ would small dataset with high noise perform more hence help generalization.

2.3.4

2.3.4

As n approaches d, the test error started to increase for the unregularized estimator, while the test error decreases as training set size increase in regions around n=d for the regularized estimator. Under ridge regularization adding more training data would always lead to a better test performance.