CSC413 Homework 3

Name: Yulin Wang

Student #: 1003942326

1.1.1 Batch size vs. learning rate

As batch size increases, the gradient noise will decrease.

followed by a larger optimal learning rate.

If B is small initially, then the increase in the optimal learning rate will be more prominent.

1.1.2 Training steps vs. batch size

(a) Point C has the most efficient batch size.

For the batch sizes smaller than C, increasing batch size could

reduce the training time without much change in total compute.

For the batch sizes greater than C, the reduction in training time is minimal,

and would cost more compute.

(b) Point A: Regime: noise dominated

Point B: Regime: curvature dominated

1.1.3 Batch size, Optimizer, Normalization, Learning Rate

(a) options I and IV

(b) options II+ and III-

1.2 Model size, dataset size and compute

(a) (i) Model A has more parameters.

Since Mode A converges to a smaller total loss in Figure 3.

(ii) At "X", Model B has been training for more iterations.

Since Model B has fewer parameters then it needs to train more iterations to achieve the same total loss as model A.

(b) I will choose model A.

Since model A could converge to a smaller total loss than model B, when given the same total compute.

2.1 Bias-variance decomposition

2.1.1

$$R(\hat{w}) = E(W_*^T X - \hat{W}^T X)^2$$

$$= E[(W_*^T X - \hat{W}^T X)^T (W_*^T X - \hat{W}^T X)]$$

$$= E[X^T (W_* - \hat{W})(W_*^T - \hat{W}^T) X]$$

$$= E[tr[X^T (W_* - \hat{W})(W_*^T - \hat{W}^T) X]]$$

$$= tr[E(XX^T) E[(W_* - \hat{W})(W_*^T - \hat{W}^T)]]$$

$$= tr[\underset{①}{E(XX^T)} E(\underset{②}{W_* W_*^T} - \underset{③}{W_* \hat{W}^T} - \underset{④}{\hat{W} W_*^T} + \underset{⑤}{\hat{W}\hat{W}^T})]$$

Since $\hat{W} = (X^T X)^{-1} X^T (XW_* + \varepsilon) = W_* + (X^T X)^{-1} X^T \varepsilon$ when $n > d$.

Then ① $E(XX^T) = I_d$; ② $E(W_* W_*^T) = \frac{1}{d} I_d$

③ $E(W_* \hat{W}^T) = E[W_* (W_* + (X^T X)^{-1} X^T \varepsilon)^T]$

$= E[W_* W_*^T + W_* \varepsilon^T X (X^T X)^{-1}]$; $E(\varepsilon) = E(\varepsilon^T) = 0$

$= \frac{1}{d} I_d + 0 = \frac{1}{d} I_d$

④ $E[\hat{W} W_*^T] = E[(W_* + \hat{W})^T] = \frac{1}{d} I_d$

⑤ $E[\hat{W}\hat{W}^T] = E\left[(W_* + (X^TX)^{-1}X^T\xi)(W_* + (X^TX)^{-1}X^T\xi)^T\right]$

$= E\left[(W_* + (X^TX)^{-1}X^T\xi)(W_*^T + \xi^TX(X^TX)^{-1})\right]$

$= E[W_* W_*^T] + E[(X^TX)^{-1}X^T\xi W_*^T] + E[(X^TX)^{-1}X^T\xi W_*^T] + E\left[(X^TX)^{-1}X^T\xi\xi^TX(X^TX)^{-1}\right]$

$= \frac{1}{d} I_d + 0 + 0 + E\left[(X^TX)^{-1}X^T\xi\xi^TX(X^TX)^{-1}\right]$

Thus, $R(\hat{W}) = tr[E(X^TX)\, E(W_* W_*^T - W_*\hat{W}^T - \hat{W}W_*^T + \hat{W}\hat{W}^T)]$

$= tr\left[I_d\left(\frac{1}{d}I_d - \frac{1}{d}I_d - \frac{1}{d}I_d + \frac{1}{d}I_d + E\left[(X^TX)^{-1}X^T\xi\xi^TX(X^TX)^{-1}\right]\right)\right]$

$= 0 + tr\, E\left[(X^TX)^{-1}X^T\xi\xi^TX(X^TX)^{-1}\right]$; $E[\xi\xi^T] = \sigma^2 I_n$

$= 0 + \sigma^2 \cdot tr[(X^TX)^{-1}]$

2.2 Deriving the exact expressions

2.2.1

① under-parameterized case $n > d$

$E[R(\hat{W})] = \frac{d\sigma^2}{n-d-1}$

② over-parameterized case $n < d$; by 2.1.2

$tr[I_d - X^T(XX^T)^{-1}X] = d - n$; $tr[(XX^T)^{-1}] = \frac{n}{d-n-1}$

$\Rightarrow E[R(\hat{W})] = E\left[\frac{1}{d} tr[I_d - X^T(XX^T)^{-1}X] + \sigma^2 tr[(XX^T)^{-1}]\right]$

$\Rightarrow E[R(\hat{W})] = \frac{d-n}{d} + \frac{n\sigma^2}{d-n-1}$

2.2.2 Double descent

(1) ① under-parameterized case $n > d$

Set $E[R(\hat{W})] = 0 \Rightarrow \frac{d\sigma^2}{n-d-1} = 0 \Rightarrow \sigma = 0 \Rightarrow$ variance equals to 0.

② over-parameterized case $n < d$

Set $E[R(\hat{W})] = 0 \Rightarrow \frac{d-n}{d} + \frac{n\sigma^2}{d-n-1} = 0 \Rightarrow n = d$ and ($n = 0$ or $\sigma = 0$)

since $n < d \Rightarrow$ contradicts $\Rightarrow$ impossible

(2) NO, it does NOT always help generalization.

As the above formulas show, when adding more training examples might increase the variance.

There exists a trade-off between them.

2.3 Ridge regularization

2.3.2

① When n increases, the model would be easier to have overfitting issue,

should increase λ.

② When σ increases, the model becomes less likely to have overfitting issue,

should decrease λ.

2.3.4

① From 2.2.1 $\begin{cases} \text{under: } E[R(\hat{w})] = \dfrac{d\sigma^2}{n-d-1} \\ \text{over: } E[R(\hat{w})] = \dfrac{d-n}{d} + \dfrac{n\sigma^2}{d-n-1} \end{cases}$

ridge-regularized: $E[R(\hat{w}_\lambda)] = \sigma^2 \cdot \dfrac{-(1-\gamma+\lambda)+\sqrt{(1-\gamma+\lambda)^2+4\gamma\lambda}}{2\lambda}$

Instead of seperated d and n, the ridge-regularized estimator generizes them by using $\gamma = \frac{d}{n}$.

From the formula, as $\gamma = \frac{d}{n}$ increases, $E[R(\hat{w})]$ increases, which makes it easier to find effects by conditions.

② YES, adding more training data always leads to better test performance, given $\lambda = \sigma^2\gamma$.