Homework 3

1. Trading off Resources in Neural Net Training

1.1 Effect of batch size

1.1.1 Batch size vs. learning rate

Recall that $g_B(w) = \frac{1}{B}\sum g_i(w)$, where B denotes the batch size. As batch size increases, the minibatch gradient noise will decrease and therefore increases the optimal learning rate. Thus, higher the batch size B, lower the increased learning rate.

1.1.2 Training steps vs. batch size

(a)

For the three points (A,B,C) on Figure 1, point C has the most efficient batch size (in terms of best resource and training time trade-off). If the batch size is higher, the resources it needs to use will be higher as well, with a decrease in training time, vice versa. Therefore, A is great in terms of resource but not training time, and B is great in terms of training time but not resources, while C can balance between resource usage and training time.

(b)

Point A: Regime: noise dominated
Point B: Regime: curvature dominated.

1.1.3 Batch size, Optimizer, Normalization, Learning Rate

(a) II-, III-, IV
(b) I, II+, III+

1.2 Model size, dataset size and compute

(a)

(1) Model A has more parameters. This is because it takes longer for A to reduce loss.
(2) Model B has been training for more iterations at the intersection "X". This is because total compute = number of training steps × number of parameters × constants. At "X", A has more parameter thus B has more training steps.

(b)

If I have an urgent deadline coming up, I will prefer employing model B rather than model A because it takes longer for A to reduce loss than B. Therefore it requires more wallclock time for A to reach “X” (assuming we have the same desired test loss using the same total compute).

2. Generalization Error of Linear Regression

2.1 Bias-variance decomposition

2.1.1

Let $n > d$.
Then, we have $\hat{w} = (X^T X)^{-1} X^T t$
Let $\theta = (\tilde{x}, \epsilon, w_*)$.
Then, we know that

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

Then, as hinted, by using bias-variance decomposition, we can get:

$$\begin{aligned} R(\hat{w}) &= B(\hat{w}) + V(\hat{w}) \\ &= E_\theta[|E_\theta(w_*^T \tilde{x}) - E_\theta(\hat{w}^T \tilde{x})|]^2 + E_\theta[|\hat{w}^T \tilde{x} - E_\theta(\hat{w}^T \tilde{x})|^2] \end{aligned}$$

Meanwhile, we have $E_\theta(w_*^T\tilde{x}) = E_\theta(w_*^T)E_\theta(\tilde{x}) = w_*^T 0 = 0$ (since $x_i \sim N(0, I_d)$).

In addition, $E_\theta(\hat{w}^T\tilde{x}) = E_\theta([(X^TX)^{-1}X^Tt]^T\tilde{x})$ since $\hat{w} = (X^TX)^{-1}X^Tt$.

We also know that $t = Xw_* + \varepsilon$.

Therefore,

$$\begin{aligned}
E_\theta(\hat{w}^T\tilde{x}) &= E_\theta([(X^TX)^{-1}X^T(w_*^Tx + \epsilon)]^T\tilde{x}) \\
&= E_\theta((w_*^Tx + \epsilon)^T[(X^TX)^{-1}X^T]^T\tilde{x}) \\
&= E_\theta(\tilde{x})E_\theta[(w_*^Tx + \epsilon)^T[(X^TX)^{-1}X^T]^T] \\
&= 0 \cdot E_\theta[(w_*^Tx + \epsilon)^T[(X^TX)^{-1}X^T]^T] \\
&= 0
\end{aligned}$$

Therefore, we know that

$$\begin{aligned}
R(\hat{w}) &= E_\theta[|E_\theta(w_*^T\tilde{x}) - E_\theta(\hat{w}^T\tilde{x})|]^2 + E_\theta[|\hat{w}^T\tilde{x} - E_\theta(\hat{w}^T\tilde{x})|^2] \\
&= 0 + E_\theta[|\hat{w}^T\tilde{x} - E_\theta(\hat{w}^T\tilde{x})|^2] \\
&= E_\theta[|\hat{w}^T\tilde{x} - E_\theta(\hat{w}^T\tilde{x})|^2] \\
&= E_\theta(\hat{w}^T\tilde{x})^2 && \text{since } E_\theta(\hat{w}^T\tilde{x}) = 0 \\
&= E_\theta[\tilde{x}^T\hat{w}^T\hat{w}\tilde{x}] \\
&= E_\theta[Tr(\tilde{x}^T\hat{w}^T\hat{w}\tilde{x})] \\
&= E_\theta[Tr(\hat{w}^T\tilde{x}\tilde{x}^T\hat{w})] \\
&= E_\theta[Tr(\hat{w}\hat{w}^T\tilde{x}\tilde{x}^T)] && \text{since } \mathrm{Tr(ABCD)} = \mathrm{Tr(DABC)} \text{ by cyclic property of trace} \\
&= Tr(E_\theta(\hat{w}\hat{w}^T)E_\theta(\tilde{x}\tilde{x}^T))
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
E_\theta(\hat{w}\hat{w}^T) &= E_\theta([(X^TX)^{-1}X^Tt][(X^TX)^{-1}X^Tt]^T) \\
&= E_\theta([(X^TX)^{-1}X^T(Xw_* + \varepsilon)][(X^TX)^{-1}X^T(Xw_* + \varepsilon)]^T) && \text{since } t = Xw_* + \varepsilon \\
&= E_\theta([A(Xw_* + \varepsilon)][A(Xw_* + \varepsilon)]^T) && \text{let } A = (X^TX)^{-1}X^T \\
&= E_\theta([A(Xw_* + \varepsilon)][(Xw_* + \varepsilon)^TA^T]) \\
&= E_\theta([A(Xw_* + \varepsilon)][(w_*^TX^T + \varepsilon^T)A^T]) \\
&= E_\theta([A(Xw_* + \varepsilon)][(w_*^TX^TA^T + \varepsilon^TA^T)]) \\
&= E_\theta(A(Xw_* + \varepsilon)(w_*^TX^TA^T) + A(Xw_* + \varepsilon)(\varepsilon^TA^T) \\
&= E_\theta(AXw_*w_*^TX^TA^T + A\varepsilon w_*^TX^TA^T + AXw_*\varepsilon^TA^T + A\varepsilon\varepsilon^TA^T) \\
&= E_\theta(AXw_*w_*^TX^TA^T + A\varepsilon\varepsilon^TA^T) \\
&= E_\theta(w_*w_*^T)E_\theta(AXX^TA^T) + E_\theta(\varepsilon\varepsilon^T)E_\theta(AA^T) \\
&= E_\theta(w_*w_*^T)E_\theta(A)I_dE_\theta(A^T) + E_\theta(\varepsilon\varepsilon^T)E_\theta(AA^T) \\
&= E_\theta(w_*w_*^T)E_\theta(AA^T)I_d + E_\theta(\varepsilon\varepsilon^T)E_\theta(AA^T) \\
&= E_\theta(AA^T)[E_\theta(w_*w_*^T)I_d + E_\theta(\varepsilon\varepsilon^T)E_\theta([(X^TX)^T]^{-1})] \\
&= E_\theta(AA^T)[\frac{1}{d}I_d^2 + \sigma^2[(X^TX)^{-1}]E_\theta(XX^T)^{-1}]
\end{aligned}$$

And, $E_\theta(\tilde{x}\tilde{x}^T) = I_d$ is given.

...

(sry i really tried my best 😟 something went wrong)

2.2 Deriving the exact expressions

2.2.1

As instructed, I will be ignoring the regume $n-1 <= d <= n+1$

1. Let $n > d$
   $E(R(\hat{w})) = E(\sigma^2 Tr((X^T X)^{-1})) = \sigma^2 E(Tr((X^T X)^{-1})$
   Given that $[X]_{ij} \sim N(0,1)$ and $X \in R^{n \times d}$, we can conclude that $E(R(\hat{w})) = \sigma^2 \frac{d}{n-d-1}$
2. Let $n < d$
   $E(R(\hat{w})) = \frac{1}{d}(d-n) + \sigma^2 E(Tr(XX^T)^{-1})$
   Given that $[X]_{ij} \sim N(0,1)$, we can conclude that $E(R(\hat{w})) = \frac{1}{d}(d-n) + \sigma^2 \frac{n}{d-n-1}$

2.2.2

(1) Under what conditions (on n, d, σ) is it possible for the model to achieve perfect generalization, i.e., the expected risk $E[R(\hat{w})] = 0$?

For n>d (underparametrized case), $E[R(\hat{w})] = 0$ iff $\sigma^2 \frac{d}{n-d-1} = 0$.(i.e. $d = 0$ or $\sigma = 0$). d=0 means that there is no feature at all. $\sigma =0) means that there is no noise at all. Thus, either of these two cases can lead the model to perfect generalization.

For n<d , $E[R(\hat{w})] = 0$ iff $\frac{1}{d}(d-n) + \sigma^2 \frac{n}{d-n-1} = 0$.
$\implies E[R(\hat{w})] = 0$ iff $\sigma^2 = -\frac{d-n}{d} \cdot \frac{d-n-1}{n}$
$\implies E[R(\hat{w})] = 0$ iff $\sigma^2 = \frac{(n-d)(d-n-1)}{dn}$
Therefore, we can conclude that if the noise variance is $\frac{(n-d)(d-n-1)}{dn}$, then the model can have perfect generalization

(2) Does adding more training examples always help generalization? Why?
Unfortunately, adding more training examples cannot always help generalization. For example, if the additional data point is noisy, it will make the model confused of what's really significant and therefore might even hurt the test performance.

2.3 Ridge regularization

2.3.2

Heuristically argue whether the regularization strength λ should increase or decrease with the training set size n and noise level σ. Provide your arguments in no more than two sentences.

The regularization strength λ should increase if the training set size n and noise level σ decrease. Vice versa. This is because when the training size and noise level are both large, decreasing λ can help reduce the variance of the model and thus avoiding overfitting while making sure no important data being lost. When the training size and noise level are both small, still decreasing λ can lead to increase in bias and thus result in underfitting.

2.3.4

1. The test loss of the unregularized estimator in Section 2.2.1 is dependent on the X only wheras the ridge-regularized test loss is not only dependent on the training data, but also λ, which makes the latter a more flexible choice. We can always tune it as needed.
2. Yes, adding more training data will always lead to better test performance as the model adjusts according to its size and number of features.