1.1.1) As batch size increases, optimal learning rate should increase too because larger batch size leads to more confident gradients (more confident in the direction of the descent) (i.e. less stochastic) (the minibatch gradient noise is proportional to $\frac{1}{B}$ of stochastic noise). so we can use larger learning rate,

$$[g_B(w)] = \text{var}\left[\frac{1}{B}\sum_{i=1}^{B} g_i(w)\right] = \frac{1}{B^2}\sum_{i=1}^{B} \text{Var}[g_i(w)] = \frac{1}{B^2}\sum_{i=1}^{B} \text{var}[\nabla f(w) + \varepsilon_i]$$

$$= \frac{1}{B^2}\sum_{i=1}^{B} \text{var}[\varepsilon_i] = \frac{1}{B^2} B\sigma^2 = \frac{1}{B}\sigma^2$$

1.1.2) a) point C has the most efficient batch size because for batch size smaller than C, increasing batch size significantly decreases the number of steps needed to reach the validation error. For batch size larger than C, this convergence benefits see diminishing returns, but the time per weight update should increase with more batch size.

b) Point A: Regime noise dominated

Point B: Regime curvature dominated.

1.1.3)

a) I, IV

b) II+, III−

1.2

a)

1) Model A has more number of parameters than model B because as total compute (number of training steps) increase, the test loss for B is able to reach much lower than test loss for A, meaning the increased complexity of B makes it able to generalize better to test data.

2) at X, Total compute A = total compute B, and since the number of parameters of A > number of parameters B,
⇒ the number of training steps must be more for B than A to use the same amount of total compute.

b) If I have an urgent deadline coming up, I would use the larger model (A) since it requires less training steps (less wall clock time) to reach the same test loss. On the other hand, if I do not have access to good machines, it might be better to train a smaller model assuming it can reach the level of test loss wanted.

2.1.1)

since $\tilde{x}$, ε, W_* are independent we are able to pull variables out of the expectations over other variables.

$$R(\hat{W}) = E_{\tilde{x},\varepsilon,W_*}\left[\left(W_*^T\tilde{x} - \hat{W}^T\tilde{x}\right)^2\right]$$

$$= E_{\tilde{x},\varepsilon,W_*}\left[(W_*^T\tilde{x})^2 - 2W_*^T\tilde{x}\hat{W}^T\tilde{x} + (\hat{W}^T\tilde{x})^2\right]$$

$$= \underbrace{E_{\tilde{x},\varepsilon,W_*}\left[(W_*^T\tilde{x})^2\right]}_{①} \underbrace{- 2E_{\tilde{x},\varepsilon,W_*}\left[W_*^T\tilde{x}\hat{W}^T\tilde{x}\right]}_{②} + \underbrace{E_{\tilde{x},\varepsilon,W_*}\left[(\hat{W}^T\tilde{x})^2\right]}_{③}$$ by linearity of expectation.

next we evaluate each of the 3 terms seperately:

① $$E_{\tilde{x},\varepsilon,W_*}\left[(W_*^T\tilde{x})^2\right] = E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[(W_*^T\tilde{x})^2\right]\right]\right]$$

$$= E_{\varepsilon}\left[E_{W_*}\left[\left(E_{\tilde{x}}[W_*^T\tilde{x}]\right)^2 + \mathrm{Var}[W_*^T\tilde{x}]\right]\right]$$

$$= E_{\varepsilon}\left[E_{W_*}\left[\left(W_*^T \cancel{E_{\tilde{x}}[\tilde{x}]}^{0}\right)^2 + W_*^T \cancel{\mathrm{Var}[\tilde{x}]}^{I_d} W_*\right]\right]$$

$$= E_{\varepsilon}\left[E_{W_*}\left[W_*^T W_*\right]\right]$$

$$= E_{\varepsilon}\left[E_{W_*}\left[\mathrm{Tr}(W_*^T W_*)\right]\right]$$

$$= E_{\varepsilon}\left[E_{W_*}\left[\mathrm{Tr}(W_* W_*^T)\right]\right]$$ by cyclic property of trace.

$$= E_{\varepsilon}\left[\mathrm{Tr}\left(E_{W_*}\left[W_* W_*^T\right]\right)\right]$$

$$= E_{\varepsilon}\left[\mathrm{Tr}\left(\tfrac{1}{d} I_d\right)\right]$$

$$= E_{\varepsilon}[1]$$

$$= 1$$

② $$-2E_{\tilde{x},\varepsilon,W_*}\left[W_*^T\tilde{x}\hat{W}^T\tilde{x}\right] = -2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[W_*^T\tilde{x}\hat{W}^T\tilde{x}\right]\right]\right]$$

$$= -2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[W_*^T\tilde{x}\, t^T X(X^TX)^{-1}\tilde{x}\right]\right]\right]$$ since $\hat{w} = (X^TX)^{-1}X^Tt \Rightarrow \hat{w}^T = t^TX(X^TX)^{-1}$

$$= -2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[W_*^T\tilde{x}\left(\omega_*^TX^T + \varepsilon^T\right)X(X^TX)^{-1}\tilde{x}\right]\right]\right]$$ since $t = X\omega_* + \varepsilon \Rightarrow t^T = \omega_*^TX^T + \varepsilon^T$

$$= -2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[W_*^T\tilde{x}\,\omega_*^T\cancel{X^TX(X^TX)^{-1}}\tilde{x}\right] + E_{\tilde{x}}\left[W_*^T\tilde{x}\varepsilon^TX(X^TX)^{-1}\tilde{x}\right]\right]\right]$$

$$= -2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[(W_*^T\tilde{x})^2\right]\right]\right] - 2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[\mathrm{tr}\left(W_*^T\tilde{x}\varepsilon^TX(X^TX)^{-1}\tilde{x}\right)\right]\right]\right]$$

$$= -2(1) - 2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[\mathrm{tr}\left(W_*^T\tilde{x}\varepsilon^TX(X^TX)^{-1}\tilde{x}\right)\right]\right]\right]$$ since $E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[(W_*^T\tilde{x})^2\right]\right]\right] = 1$ by ①

$$= -2 - 2E_{\varepsilon}\left[E_{W_*}\left[E_{\tilde{x}}\left[\mathrm{tr}\left(\varepsilon^TX(X^TX)^{-1}\tilde{x}W_*^T\tilde{x}\right)\right]\right]\right]$$ by cyclic property of trace

$$= -2 - 2 \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[tr\left(\mathop{\mathbb{E}}_{\tilde{x}}\left[\varepsilon^T X (X^T X)^{-1} \tilde{x} w_*^T \tilde{x}\right]\right)\right]\right]$$

$$= -2 - 2 \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[tr\left(\varepsilon^T X (X^T X)^{-1} \mathop{\mathbb{E}}_{\tilde{x}}\left[\tilde{x} w_*^T \tilde{x}\right]\right)\right]\right]$$

$$= -2 - 2 \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[tr\left(\varepsilon^T X (X^T X)^{-1} I_d w_*\right)\right]\right] \quad \text{since } \mathop{\mathbb{E}}_{\tilde{x}}\left[\tilde{x} w_*^T \tilde{x}\right] = I_d w_*$$

$$= -2 - 2 \mathop{\mathbb{E}}_{\varepsilon}\left[tr\left(\varepsilon^T X (X^T X)^{-1} I_d \mathop{\mathbb{E}}_{w_*}\left[w_*\right]\right)\right]$$

$$= -2 - 2\, tr\left(\underbrace{\mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon^T\right]}_{0} X (X^T X)^{-1} I_d \mathop{\mathbb{E}}_{w_*}\left[w_*\right]\right)$$

$$= -2 - 0$$

$$= -2$$

(3) $\mathop{\mathbb{E}}_{\tilde{x},\varepsilon,w_*}\left[(\hat{w}^T \tilde{x})^2\right] = \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\mathop{\mathbb{E}}_{\tilde{x}}\left[(t^T X (X^T X)^{-1} \tilde{x})^2\right]\right]\right]$ since $\hat{w}^T = t^T X (X^T X)^{-1}$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\left(\mathop{\mathbb{E}}_{\tilde{x}}\left[t^T X (X^T X)^{-1} \tilde{x}\right]\right)^2 + Var\left[t^T X (X^T X)^{-1} \tilde{x}\right]\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\left(\mathop{\mathbb{E}}_{\tilde{x}}\left[(w_*^T X^T + \varepsilon^T) X (X^T X)^{-1} \tilde{x}\right]\right)^2 + Var\left[(w_*^T X^T + \varepsilon^T) X (X^T X)^{-1} \tilde{x}\right]\right]\right] \quad \text{since } t^T = w_*^T X^T + \varepsilon^T$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\left((w_*^T X^T + \varepsilon^T) X (X^T X)^{-1} \underbrace{\mathop{\mathbb{E}}_{\tilde{x}}\left[\tilde{x}\right]}_{0}\right)^2 + Var\left[(w_*^T X^T + \varepsilon^T) X (X^T X)^{-1} \tilde{x}\right]\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[0 + Var\left[(w_*^T X^T + \varepsilon^T) X (X^T X)^{-1} \tilde{x}\right]\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[Var\left[w_*^T X^T X (X^T X)^{-1} \tilde{x} + \varepsilon^T X (X^T X)^{-1} \tilde{x}\right]\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[Var\left[w_*^T \tilde{x} + \varepsilon^T X (X^T X)^{-1} \tilde{x}\right]\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[Var\left[\left(w_*^T + \varepsilon^T X (X^T X)^{-1}\right) \tilde{x}\right]\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\left(w_*^T + \varepsilon^T X (X^T X)^{-1}\right) Var\left[\tilde{x}\right] \left(w_*^T + \varepsilon^T X (X^T X)^{-1}\right)^T\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\left(w_*^T + \varepsilon^T X (X^T X)^{-1}\right)\left(w_* + (X^T X)^{-1} X^T \varepsilon\right)\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[w_*^T w_* + w_*^T (X^T X)^{-1} X^T \varepsilon + \varepsilon^T X (X^T X)^{-1} w_* + \varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T \varepsilon\right]\right]$$

$$= \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[w_*^T w_*\right]\right] + \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[w_*^T (X^T X)^{-1} X^T \varepsilon\right]\right] + \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\varepsilon^T X (X^T X)^{-1} w_*\right]\right] + \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T \varepsilon\right]\right]$$

by second half of (1)

$$= 1 + \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[w_*^T (X^T X)^{-1} X^T \varepsilon\right]\right] + \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\varepsilon^T X (X^T X)^{-1} w_*\right]\right] + \mathop{\mathbb{E}}_{\varepsilon}\left[\mathop{\mathbb{E}}_{w_*}\left[\varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T \varepsilon\right]\right]$$

$$= 1 + \mathop{\mathbb{E}}_{w_*}\left[w_*^T\right] (X^T X)^{-1} X^T \underbrace{\mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon\right]}_{0} + \underbrace{\mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon^T\right]}_{0} X (X^T X)^{-1} \mathop{\mathbb{E}}_{w_*}\left[w_*\right] + \mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T \varepsilon\right]$$

$$= 1 + \mathop{\mathbb{E}}_{\varepsilon}\left[\varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T \varepsilon\right]$$

$$= 1 + \underset{\varepsilon}{E}\left[tr\left(\varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T \varepsilon\right)\right]$$

$$= 1 + \underset{\varepsilon}{E}\left[tr\left(\varepsilon\varepsilon^T X (X^T X)^{-1} (X^T X)^{-1} X^T\right)\right] \quad \text{by cyclic property of trace}$$

$$= 1 + tr\left(\underset{\varepsilon}{E}\left[\varepsilon\varepsilon^T\right] X (X^T X)^{-1} (X^T X)^{-1} X^T\right)$$

$$= 1 + tr\left(\sigma^2 I_n X (X^T X)^{-1} (X^T X)^{-1} X^T\right)$$

$$= 1 + \sigma^2 tr\left(I_n X (X^T X)^{-1} (X^T X)^{-1} X^T\right)$$

$$= 1 + \sigma^2 tr\left(X^T I_n X (X^T X)^{-1} (X^T X)^{-1}\right) \quad \text{by cyclic property of trace}$$

$$= 1 + \sigma^2 tr\left(X^T X (X^T X)^{-1} (X^T X)^{-1}\right)$$

$$= 1 + \sigma^2 tr\left((X^T X)^{-1}\right)$$

So $R(\hat{W}) = E_{x,\varepsilon,W_*}\left[\left(W_*^T \tilde{x} - \hat{W}^T \tilde{x}\right)^2\right] = ① + ② + ③ = 1 - 2 + 1 + \sigma^2 tr\left((X^T X)^{-1}\right)$

$$= 0 + \sigma^2 tr\left((X^T X)^{-1}\right)$$

as wanted //

2.2.1)

$n > d+1$:

$$E[R(\hat{w})] = E\left[\sigma^2 Tr\left((X^TX)^{-1}\right)\right]$$
$$= \sigma^2 E\left[Tr\left((X^TX)^{-1}\right)\right]$$
$$= \sigma^2 \frac{d}{n-d-1}$$

$n < d-1$:

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

take $G = X^T \in \mathbb{R}^{d \times n}$
$\Rightarrow G^T = X \in \mathbb{R}^{n \times d}$

$$= \frac{1}{d} E[d-n] + \sigma^2 E\left[Tr\left((G^TG)^{-1}\right)\right]$$
$$= \frac{1}{d}(d-n) + \sigma^2 \frac{n}{d-n-1}$$
$$= 1 - \frac{n}{d} + \sigma^2 \frac{n}{d-n-1}$$

2.2.2)

1)

if $n < d-1$, then $\frac{n}{d} < 1$
$\Rightarrow 1 - \frac{n}{d} + \sigma^2 \frac{n}{d-n-1} > 0$, since $1 - \frac{n}{d} > 0$. This means $E[R(\hat{w})] > 0$ for all $n < d-1$.

if $n > d+1$, then $E[R(\hat{w})] = \sigma^2 \frac{d}{n-d-1} = 0$ $\Rightarrow$ ① $n = \infty$, $d < \infty$, $\sigma < \infty$
OR ② $\sigma = 0$
OR ③ $d = 0$.

so $E[R(\hat{w})] = 0$ only if $n > d+1$ and

① $n = \infty$, $d < \infty$, $\sigma < \infty$ $\Rightarrow$ inf training data.
OR ② $\sigma = 0$ $\Rightarrow$ no noise
OR ③ $d = 0$. $\Rightarrow$ meaningless in context

2) No, adding more training examples does not always help generalization. For example, in the overparametrized case, $\frac{\partial E[R(\hat{w})]}{\partial n} = -\frac{1}{d} + \frac{\sigma^2 n}{(d-n-1)^2} + \frac{\sigma^2}{d-n-1}$, and this can be positive for certain σ, n, d (for example, take $\sigma = 1$, $n = 1$, $d = 4$ then $E[R(\hat{w})] = 1 - \frac{1}{4} + 1^2 \frac{1}{4-1-1} = 1.25$; but same σ, d with increased $n = 2$ gives $E[R(\hat{w})] = 1 - \frac{2}{4} + 1^2 \frac{2}{4-2-1} = 2.5$, which is larger.)

2.3.2)

The larger the sample size n, the less prone the model is to overfitting, so we should take smaller λ. On the other hand, larger noise in the dataset could cause the model to overfit on the noise, so we should take larger λ to regularize the model.

2.3.4) Based on Figure 4, the test loss of the unregularized estimator $E[R(\hat{w})]$ in 2.2.1 decreases at first as sample size increases until around $n = 400$ and then increases suddenly until $n = 500$ before dropping down. On the other hand the ridge-regularized $E[R(\hat{w}_\lambda)]$'s test loss decreases stably as sample size n increases. (The two have similar test loss for $n < 400$ and $n > 600$.

Based on Figure 4, under appropriate ridge regularization ($\lambda = \sigma^2\gamma$), adding more training data always leads to better test performance.