Noisy linear regression:

$$\tilde{\mathcal{L}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - (x^{(i)} + \delta^{(i)})^T \theta\right)^2$$

a) Since $E[\cdot]$ is a linear operator,

so

$$E_{\delta \sim N}\left[\tilde{\mathcal{L}}(\theta)\right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} E_{\delta \sim N}\left[\left(y^{(i)} - (x^{(i)} + \delta^{(i)})^T \theta\right)^2\right]$$

Hence, if we can compute the term then we are done. Let's compute the term.

Now,

$$\left(y^{(i)} - \left(x^{(i)} + \delta^{(i)}\right)^T \theta\right)^2$$

$$= \left[\left(y^{(i)} - x^{(i)T}\theta\right) - \delta^{(i)T}\theta\right]^2$$

$$= \left(y^{(i)} - x^{(i)T}\theta\right)^2 - 2\left(y^{(i)} - x^{(i)T}\theta\right)\left(\delta^{(i)T}\theta\right) + \left(\delta^{(i)T}\theta\right)^2$$

Since $E[\cdot]$ is a linear operator, so

$$\quad = E_{\delta\sim N}[\quad] - E_{\delta\sim N}[\quad] + E_{\delta\sim N}[\quad]$$

Now let's compute the above 3 quantities:

Since has no δ dependence, so

$$E_{\delta\sim N}[\ ] = \left(y^{(i)} - x^{(i)T}\theta\right)^2$$

Now,

$$E_{\delta\sim N}\left[-2\left(y^{(i)} - x^{(i)T}\theta\right)\left(\delta^{(i)T}\theta\right)\right]$$

$$= -2\left(y^{(i)} - x^{(i)T}\theta\right) E_{\delta\sim N}\left[\delta^{(i)T}\theta\right]$$

From problem statement, $E[\delta^{(i)}] = 0 \in \mathbb{R}^d$

Hence,

$$E_{\delta\sim N}\left[-2\left(y^{(i)} - x^{(i)T}\theta\right)\left(\delta^{(i)T}\theta\right)\right]$$

$$= 0$$

Lastly,

$$\mathbb{E}_{\delta\sim N}\left[\left(\delta^{(i)T}\theta\right)^2\right]$$

$$= \mathbb{E}_{\delta\sim N}\left[\theta^T \delta^{(i)} \delta^{(i)T} \theta\right]$$

$$= \theta^T \, \mathbb{E}_{\delta\sim N}\left[\delta^{(i)} \delta^{(i)T}\right] \theta$$

From the hint we know,

$$\mathbb{E}_{\delta\sim N}\left[\delta\delta^T\right] = \sigma^2 I$$

Hence,

$$\mathbb{E}_{\delta\sim N}\left[\left(\delta^{(i)T}\theta\right)^2\right]$$

$$= \sigma^2 \theta^T\theta = \sigma^2 \|\theta\|_2^2$$

Putting it all together,

$$= \left(y^{(i)} - x^{(i)T}\theta \right)^2 + \sigma^2 \|\theta\|_2^2$$

$\therefore$

$$E_{\delta \sim N}\left[\tilde{L}(\theta) \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - x^{(i)T}\theta \right)^2 + \sigma^2 \|\theta\|_2^2$$

$$= L(\theta) + R$$

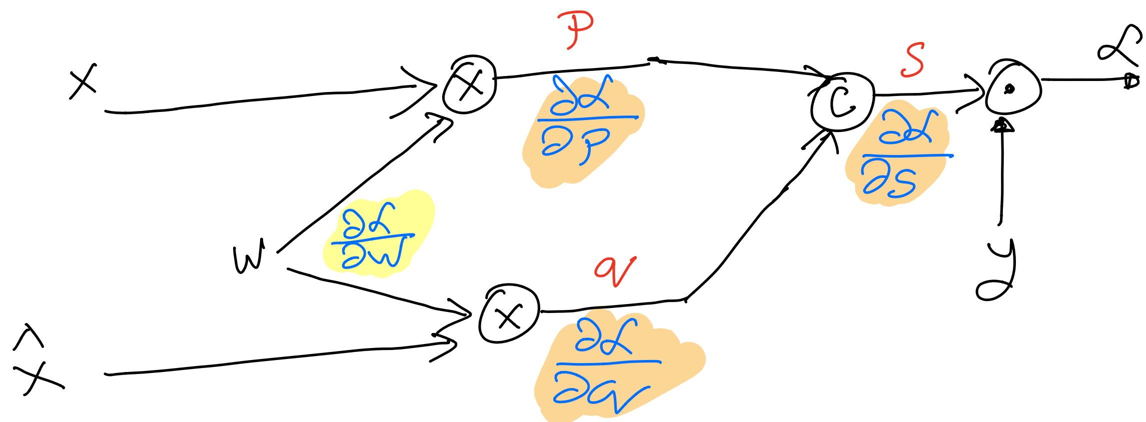
where $R = \sigma^2 \|\theta\|_2^2$

b) From (a), we can clearly observe that noise would have a L-2 regularization effect on the model with regularization strength σ.

c) As the regularization strength $\sigma \to 0$, then we have no regularization and hence the model might overfit the data.

d) As the regularization strength $\sigma \to \infty$, then the objective of the cost function is to minimize the L-2 norm of parameters θ and hence $\theta \to 0$ and the model will underfit the data.

Back propagation:



In the above computational graph, the c operation is defined as follows:

$$c(p,q) = \frac{p^T q}{\|p\|_2 \|q\|_2}$$

Since,

$$\mathcal{L} = s\,y$$

So,

$$\frac{\partial \mathcal{L}}{\partial s} = y$$

Now, by chain rule

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\partial s}{\partial p} \frac{\partial \mathcal{L}}{\partial s} = y \frac{\partial s}{\partial p}$$

$$\frac{\partial \mathcal{L}}{\partial q} = \frac{\partial s}{\partial q} \frac{\partial \mathcal{L}}{\partial s} = y \frac{\partial s}{\partial q}$$

Hence, we need to compute $\frac{\partial s}{\partial p}, \frac{\partial s}{\partial q}$.

From the computational graph, we know

$$S = \frac{p^T q}{\|p\|_2 \|q\|_2}$$

Recall, the quotient rule

$$\frac{\partial}{\partial x}\left[\frac{g(x,y)}{h(x,y)}\right]$$

$$= \frac{\left[h(x,y)\frac{\partial g(x,y)}{\partial x} - g(x,y)\frac{\partial h(x,y)}{\partial x}\right]}{[h(x,y)]^2}$$

Similarly,

$$\frac{\partial}{\partial y}\left[\frac{g(x,y)}{h(x,y)}\right]$$

$$= \frac{\left[h(x,y)\frac{\partial g(x,y)}{\partial y} - g(x,y)\frac{\partial h(x,y)}{\partial y}\right]}{\left[h(x,y)\right]^2}$$

Then using the quotient rule,

$$\frac{\partial s}{\partial p} = \frac{\left[(\|p\|_2\|q\|_2)q - (p^Tq)\left(\frac{\|q\|_2}{\|p\|_2}\right)p\right]}{\|p\|_2^2\,\|q\|_2^2}$$

$$\frac{\partial s}{\partial q} = \frac{\left[\left(\|p\|_2 \|q\|_2 \right) p - (p^T q) \left(\frac{\|p\|_2}{\|q\|_2} \right) q \right]}{\|p\|_2^2 \, \|q\|_2^2}$$

Since there are two paths to W

- One through $\frac{\partial L}{\partial p}$
- One through $\frac{\partial L}{\partial q}$

Hence,

$$\frac{\partial L}{\partial W} = \frac{\partial p}{\partial W} \frac{\partial L}{\partial p} + \frac{\partial q}{\partial W} \frac{\partial L}{\partial q}$$

Now,

$$P = WX$$

So using the outer product rule learned in lecture

$$\frac{\partial P}{\partial W} \frac{\partial \mathcal{L}}{\partial P} = \frac{\partial \mathcal{L}}{\partial P} X^T$$

$$= y \frac{\partial S}{\partial P} X^T$$

Similarly,

$$q = W\hat{X}$$

So, $$\frac{\partial q}{\partial W} \frac{\partial \mathcal{L}}{\partial q} = \frac{\partial \mathcal{L}}{\partial q} \hat{X}^T$$

$$= y \frac{\partial S}{\partial q} \hat{X}^T$$

Putting it all together,

$$\frac{\partial \mathcal{L}}{\partial W} = y \frac{\partial s}{\partial p} x^T + y \frac{\partial s}{\partial q} \hat{x}^T$$