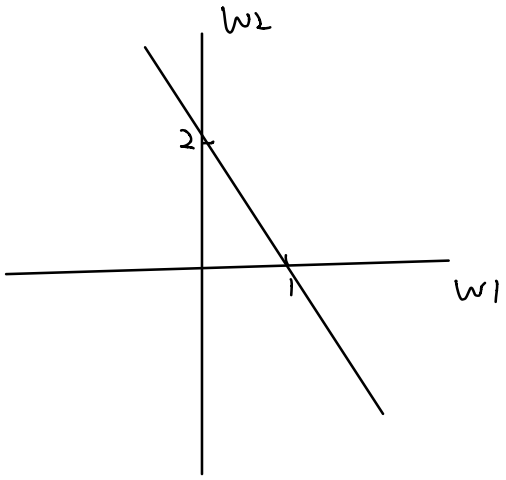
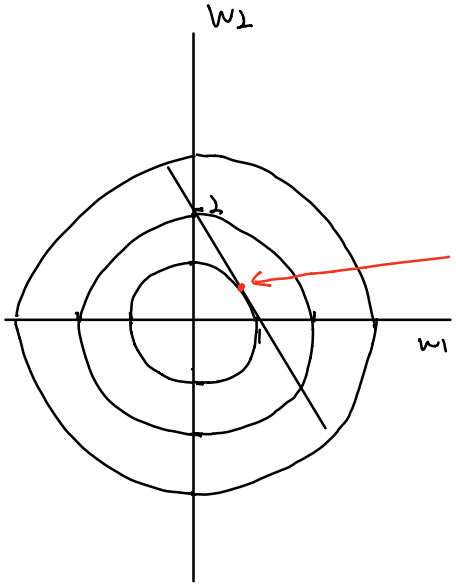
1.2.1

1) $2w_1 + w_2 = 2$

$w_2 = -2w_1 + 2$



2)



gradient descent solutions for both with and without weight decay.

1.3 AdaGrad Update

$$w_{i,t+1} = w_{i,t} - \frac{\eta}{\sqrt{G_{i,t} + \varepsilon}} \nabla_{w_{i,t}} \mathcal{L}(w_{i,t})$$

$$G_{i,t} = G_{i,t-1} + \left(\nabla_{w_{i,t}} \mathcal{L}(w_{i,t})\right)^2$$

Yes, weight decay will help AdaGrad converge to row space.
There are two reasons behind that:

① weight decay term will cancel adaptive gradient "correction" effect which break the linearity of data.

② weight decay aim to have least norm solution which is in the row space of X.

2.1.1

For linear regression problems, have k linear models.

Weight average prediction $\hat{y}_w = \left(\frac{1}{k}\sum_{i=1}^{k} w_i\right) x$

prediction average prediction $\hat{y}_p = \frac{1}{k}\sum_{i=1}^{k} y_i$

$$= \frac{1}{k}\sum_{i=1}^{k} w_i x$$

$$= \frac{x}{k}\sum_{i=1}^{k} w_i$$

$$= \left(\frac{1}{k}\sum_{i=1}^{k} w_i\right) x$$

$$= \hat{y}_w$$

Thus they give the same expected generalization error.

2.2.2

$$\begin{aligned}
\mathrm{Var}(\bar{h}(x;D)) &= \mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k} h(x;D_i)\right) \\
&= \frac{1}{k^2}\mathrm{Var}\left(\sum_{i=1}^{k} h(x;D_i)\right) \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\mathrm{Var}(h(x;D_i)) \quad \text{by assumption} \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\sigma^2 \\
&= \frac{\sigma^2}{k}
\end{aligned}$$

2.3.1

$$\begin{aligned}
\text{bias} &= E\left[\left(E\left[\bar{h}(x;D)\middle|x\right] - y_*(x)\right)^2\right] \\
&= E\left[\left(E\left[\frac{1}{k}\sum_{i=1}^{k} h(x;D_i)\middle|x\right] - y_*(x)\right)^2\right] \\
&= E\left[\left(\frac{1}{k}\sum_{i=1}^{k} E\left[h(x;D_i)\middle|x\right] - y_*(x)\right)^2\right] \\
&= E\left[\left(E\left[h(x;D)\middle|x\right] - y_*(x)\right)^2\right] \\
&= \text{original bias}
\end{aligned}$$

So no change in bias term

2.3.3

i) when k increases, variance $= \left(\rho + \frac{1-\rho}{k}\right)\sigma^2$ decreases

ii) $\rho = 0$ variance $= \frac{\sigma^2}{k}$ It means there is no similarity between different ensemble members, thus the variance of ensemble predictor is equal to $\frac{1}{k}$ variance of single predictor as we proved in 2.2.2

$\rho = 1$ variance $= \sigma^2$ which means all ensemble members are the same so that variane of ensemble predictor is equal to variance a single predictor.

3.1.2

let $n_1 \sim N(0, \sigma^2)$ $n_2 \sim N(0, \sigma^2)$ $n_3 \sim N(0,1)$

$$J = E_{(x_1, x_2, y) \sim (X_1, X_2, Y)}\left[(y^i - \hat{y}^i)^2\right]$$

$$= E_{(x_1, x_2, y) \sim (X_1, X_2, Y)}\left[(y^i - w_2 x_2^i)^2\right]$$

$$= E_{(x_1, x_2, y) \sim (X_1, X_2, Y)}\left[(y^i)^2 - 2y^i w_2 x_2^i + (w_2 x_2^i)^2\right]$$

$$= E[y^{i2}] - 2E[y^i w_2 x_2^i] + E[w_2^2 x_2^{i2}]$$

$$= Var(y^i) - 2w_2 E[y^i x_2^i] + w_2^2 Var(x_2^i)$$

$$= 2\sigma^2 + w_2^2(2\sigma^2+1) - 2w_2 E[x_2^i y^i] \quad *$$

$$E[x_2^i y^i] = E[(n_1+n_2+n_3)(n_1+n_2)]$$
$$= E[n_1^2 + 2n_1n_2 + n_1n_3 + n_2^2 + n_2n_3]$$
$$= E[n_1^2] + E[n_2^2]$$
$$= Var(n_1) + Var(n_2)$$
$$= \sigma^2 + \sigma^2$$
$$= 2\sigma^2$$

$$* = 2\sigma^2 + w_2^2(2\sigma^2+1) - 2w_2 \cdot 2\sigma^2$$
$$= 2\sigma^2 + w_2^2 + 2w_2^2\sigma^2 - 4w_2\sigma^2$$

$$\frac{\partial J}{\partial w_2} = 2w_2 + 2\sigma^2 \cdot 2w_2 - 4\sigma^2$$

let $\frac{\partial J}{\partial w_2} = 0$

$$\Rightarrow 2w_2 + 4\sigma^2 w_2 = 4\sigma^2$$
$$\Rightarrow w_2 = \frac{4\sigma^2}{4\sigma^2+2} = \frac{2\sigma^2}{2\sigma^2+1}$$

3.1.3

i) let $n_1 \sim N(0, \sigma^2)$ $\quad n_2 \sim N(0, \sigma^2)$ $\quad n_3 \sim N(0,1)$

$$J = E[(y - \hat{y})^2]$$

$$= E[y^2 - 2y\hat{y} + \hat{y}^2]$$

$$= E[y^2] - 2E[y\hat{y}] + E[\hat{y}^2]$$

$$E[y^2] = Var[y^2] = 2\sigma^2$$

$$E[y\hat{y}] = E[y(w_1 x_1 + w_2 x_2)]$$

$$= w_1 E[x_1 y] + w_2 E[x_2 y]$$

$$= w_1 E[h_1^2 + h_1 n_2] + w_2 E[(h_1 + h_2 + n_3)(h_1 + n_2)]$$

$$= w_1 E[h_1^2] + w_2 E[h_1^2] + w_2 E[h_2^2]$$

$$= w_1 Var(h_1) + w_2 Var(h_1) + w_2 Var(h_2)$$

$$= w_1 \sigma^2 + w_2 \sigma^2 + w_2 \sigma^2$$

$$= (w_1 + 2w_2)\sigma^2$$

$$E[\hat{y}^2] = E[(w_1 x_1 + w_2 x_2)^2]$$

$$= E[w_1^2 x_1^2 + 2w_1 w_2 x_1 x_2 + w_2^2 x_2^2]$$

$$= w_1^2 E[x_1^2] + 2w_1 w_2 E[x_1 x_2] + w_2^2 E[x_2^2]$$

$$= w_1^2 Var(x_1) + 2w_1 w_2 E[h_1(h_1 + h_2 + h_3)] + w_2^2 Var(x_2)$$

$$= w_1^2\sigma^2 + 2w_1w_2\,\mathrm{Var}(x_1) + w_2^2(2\sigma^2+1)$$
$$= w_1^2\sigma^2 + 2w_1w_2\sigma^2 + w_2^2(2\sigma^2+1)$$

$$\Rightarrow J = 2\sigma^2 - 2(w_1+2w_2)\sigma^2 + w_1^2\sigma^2 + 2w_1w_2\sigma^2 + (2\sigma^2+1)w_2^2$$

$$\frac{\partial J}{\partial w_1} = -2\sigma^2 + 2w_1\sigma^2 + 2w_2\sigma^2$$

let $\frac{\partial J}{\partial w_1} = 0 \Rightarrow w_1 = \frac{2\sigma^2 - 2w_2\sigma^2}{2\sigma^2} = 1 - w_2$

$$\frac{\partial J}{\partial w_2} = -4\sigma^2 + 2w_1\sigma^2 + 2(2\sigma^2+1)w_2$$

let $\frac{\partial J}{\partial w_2} = 0 \Rightarrow w_2 = \frac{4\sigma^2 - 2w_1\sigma^2}{4\sigma^2+2} = \frac{2\sigma^2 - w_1\sigma^2}{2\sigma^2+1}$

$$\begin{cases} w_1 = 1 - w_2 \\ w_2 = \dfrac{2\sigma^2 - w_1\sigma^2}{2\sigma^2+1} \end{cases} \Rightarrow \begin{cases} w_1 = \dfrac{1}{\sigma^2+1} \\ w_2 = \dfrac{\sigma^2}{\sigma^2+1} \end{cases}$$

ii) if σ becomes smaller during test time, then trained $\hat{w}_1$ will be relatively smaller, trained $\hat{w}_2$ is close to true w_2^*. Compared to training samples, all of x_1, Y, x_2 in test set will be smaller at same scale.

Thus $\hat{y} = \hat{w}_1 x_1 + \hat{w}_2 x_2$ will deviate more from target y.
In other words, it won't generalize well in test time.

3.3

let $n_1 \sim N(0, \sigma^2)$ $\quad n_2 \sim N(0, \sigma^2)$ $\quad n_3 \sim N(0,1)$

$E[\hat{y}] = w_1 x_1 + w_2 x_2$ $\quad Var[\hat{y}] = w_1^2 x_1^2 + w_2^2 x_2^2$

$$E[J] = \frac{1}{2N} \sum_{1}^{N} (E[\hat{y}] - y)^2 + \frac{1}{2N} \sum_{i=1}^{N} Var[\hat{y}]$$

$$= \frac{1}{2} E[(E[\hat{y}] - y)^2] + \frac{1}{2} E[Var(\hat{y})]$$

$$= \frac{1}{2} E[(w_1 x_1 + w_2 x_2 - y)^2] + \frac{1}{2} E[w_1^2 x_1^2 + w_2^2 x_2^2]$$

$$E[(w_1 x_1 + w_2 x_2 - y)^2]$$

$$= E[((w_1 + w_2 - 1) x_1 + (w_2 - 1) n_1 + w_2 n_3)^2]$$

$$= (w_1 + w_2 - 1)^2 E[x_1^2] + (w_2 - 1)^2 E[n_1^2] + w_2^2 E[n_3^2]$$

$$= (w_1 + w_2 - 1)^2 \sigma^2 + (w_2 - 1)^2 \sigma^2 + w_2^2$$

$$E[J] = \frac{1}{2}\left[(w_1+w_2-1)^2\sigma^2 + (w_2-1)^2\sigma^2 + w_2^2 + w_1^2\sigma^2 + w_2^2(2\sigma^2+1)\right]$$

$$\frac{\partial E[J]}{\partial w_1} = \frac{1}{2}\left(2(w_1+w_2-1)\sigma^2 + 2w_1\sigma^2\right)$$

let $\frac{\partial E[J]}{\partial w_1} = 0$

$$\Rightarrow (w_1+w_2-1)\sigma^2 + w_1\sigma^2 = 0$$

$$2w_1 + w_2 - 1 = 0$$

$$w_1 = \frac{1-w_2}{2}$$

$$\frac{\partial E[J]}{\partial w_2} = \frac{1}{2}\left(2(w_1+w_2-1)\sigma^2 + 2(w_2-1)\sigma^2 + 2w_2 + 2w_2(2\sigma^2+1)\right]$$

$$= (w_1+w_2-1)\sigma^2 + (w_2-1)\sigma^2 + w_2 + w_2(2\sigma^2+1)$$

$$= (w_1 + 4w_2 - 2)\sigma^2 + 2w_2$$

let it = 0

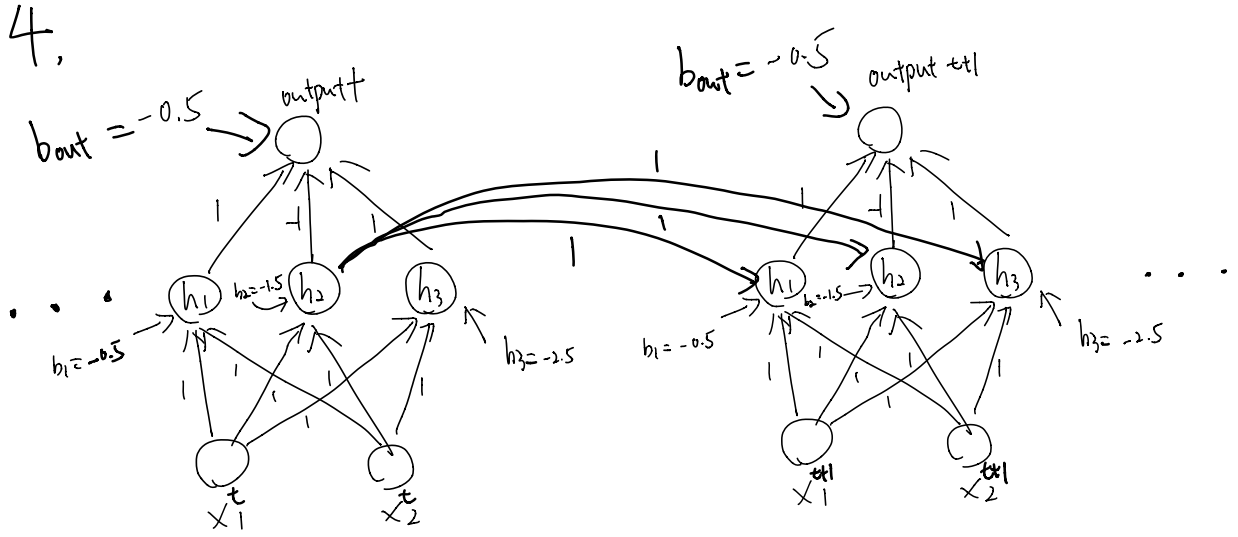
$$\Rightarrow (w_1 - 2)\sigma^2 + 4w_2\sigma^2 + 2w_2 = 0$$

$$\Rightarrow w_2 = \frac{(2 - w_1)\sigma^2}{2 + 4\sigma^2}$$

$$\begin{cases} w_1 = \dfrac{1 - w_2}{2} \\ w_2 = \dfrac{(2 - w_1)\sigma^2}{2 + 4\sigma^2} \end{cases} \Rightarrow \begin{aligned} w_1 &= \frac{2 + 2\sigma^2}{4 + 7\sigma^2} \\ w_2 &= \frac{3\sigma^2}{4 + 7\sigma^2} \end{aligned}$$

This (w_1, w_2) will generalize better than 3.1.3.
Because when σ^2 changes, there are less influence on w_1 and w_2.

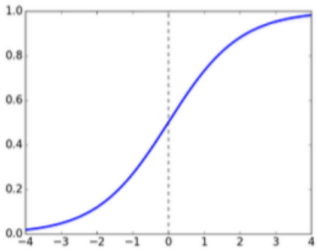
4.



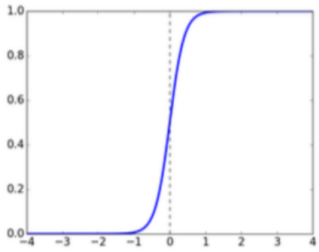
Note

1. h_1 activate if the sum is at least 1.
   h_2 activate if the sum is at least 2.
   h_3 activate if the sum is at least 3.

2. Sigmoid with huge weight will apply all hidden units and output to approximate hard threshold units. It could be illustrated by following.



$y = \sigma(x)$



$y = \sigma(5x)$

3. Initialize first RNN component with zero hidden state.

4. There are $(T+1)$ RNN component, last one for carry-out with $x_1 = x_2 = 0$.