1.2.1 $x_1=[2,1]$, $t_1=2$. $\quad J_{reg}=J+\lambda R \quad J=\frac{1}{2n}\|Xw-t\|_2^2+\frac{\lambda}{2}w^Tw$.

$w^0=\begin{bmatrix}0\\0\end{bmatrix}$

$dJ/dw=\frac{2}{n}X^T(X\hat{w}-t)+\lambda w$

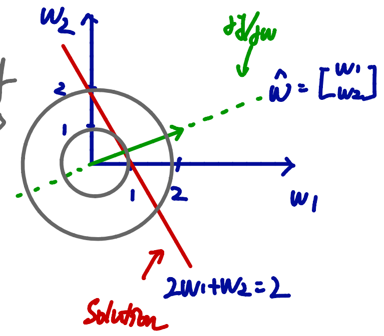
$w^1\leftarrow(1-\alpha\lambda)\begin{bmatrix}0\\0\end{bmatrix}-\alpha\left[\frac{2}{1}\begin{bmatrix}2\\1\end{bmatrix}\left([2,1]\begin{bmatrix}0\\0\end{bmatrix}-2\right)+\lambda\begin{bmatrix}0\\0\end{bmatrix}\right]$

$=4\alpha\begin{bmatrix}2\\1\end{bmatrix}$

$w^2\leftarrow(1-\alpha\lambda)4\alpha\begin{bmatrix}2\\1\end{bmatrix}-\alpha\left[2\begin{bmatrix}2\\1\end{bmatrix}\left([2\ 1]4\alpha\begin{bmatrix}2\\1\end{bmatrix}-2\right)+4\alpha\lambda\begin{bmatrix}2\\1\end{bmatrix}\right]$

$\hat{w}=ro\left(\begin{bmatrix}2\\1\end{bmatrix}\right)$. direction of grad stays in the row space of $\begin{bmatrix}2\\1\end{bmatrix}$.

$=a\begin{bmatrix}2\\1\end{bmatrix}$, $a\in\mathbb{R}$



$\frac{\lambda}{2}w^Tw=\frac{\cancel{\lambda}}{2}a[2\,1]a\begin{bmatrix}2\\1\end{bmatrix}=\frac{5}{2}\lambda a^2=\frac{\lambda}{2}[w_1\ w_2]\begin{bmatrix}w_1\\w_2\end{bmatrix}=\frac{\lambda}{2}(w_1^2+w_2^2)$

$w^Tx_1=t_1 \Rightarrow [2,1]\begin{bmatrix}w_1\\w_2\end{bmatrix}=2$, $\begin{cases}2w_1+w_2=2 \Rightarrow \text{solution dir.}\\ w_1^2+w_2^2=5a^2 \Rightarrow \text{circle}\end{cases}$

The weight decay will not change the direction of gradient descent, but the solution will not be on the solution space: $2w_1+w_2=2$. since the weight decay term perturb the solution off this line.

1.3. Yes. as we showed before, the weigh decay term guides/adds to update the weight in direction of row space of weight, while other terms shifts the offset. without weight decay terms, solution resides on any lines parallel to solution, while weight decay term penalize the weight to smaller value such that it generalize to the solution in the end.

2.1.1. $h_{weight}(x;D)=\left(\frac{1}{m}\sum_{i=1}^{m}w_i^T\right)x=\frac{1}{m}\left(\sum_{i=1}^{m}w_i^Tx\right)=h_{pred}(x;D)$.

2.2.2. $Var[\bar{h}(x;D)]=Var\left[\frac{1}{k}\sum_{i=1}^{k}h(x;D_i)\right]=\frac{1}{k^2}Var\left[\sum_{i=1}^{k}h(x;D_i)\right]=1/k^2\sum_{i=1}^{k}Var(h(x;D_i))=1/k^2\cdot k\sigma^2=1/k\,\sigma^2$

2.3.1. NO. $E[\bar{h}(x;D)|x]=E[\frac{1}{k}\sum h(x;D_i)|x]=\frac{1}{k}\sum E[h(x;D_i)|x]=E[h(x;D)|x]$

$\Rightarrow E[|E[\bar{h}(x;D)|x]-y_*(x)|^2]=E[|E[h(x;D)|x]-y_*(x)|^2]$, $\Rightarrow$ not changed.

2.3.3. Variance $=\left(\rho+\frac{1-\rho}{k}\right)\sigma^2$ if $k\uparrow \Rightarrow Var\downarrow$. then the variance decreases

$\rho=0 \Rightarrow Cov(h(x;D_j),h(x;D_i))=0\ \ \forall i\neq j$

$\Leftrightarrow$ all ensemble members are not related to each other i.e. mutually independent.

Then. $Var=\sigma^2/k$

$\rho=1 \Rightarrow Cov(h(x;D_j),h(x;D_i))=\sigma_i\sigma_j\ \ \forall i\neq j$

$\Leftrightarrow$ The correlations between each ensemble members are very strong. i.e. perfect positive relationship.

Then $Var=\sigma^2$ same as without bagging.

we see that the more correlations between data, the less effect bagging has.

3.1.2. Predicting Y only with X_2.

$\hat{y}^{(i)} = w_2 x_2^{(i)}$

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

$E(G(0,2\sigma^2)(G(0,2\sigma^2+1))$

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

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

$\operatorname{Var} Y = EY^2 - E^2 Y$

$EY^2 = E^2 Y + \operatorname{Var} Y$

$= 0 + 2\sigma^2 = 2\sigma^2$

$$= EY^2 - 2w_2 EYX_2 + w_2^2 EX_2^2$$

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

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

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

$EYX_2 = E(Y(Y + \underline{\text{Gaussian}(0,1)}))$ $\to G(0,1)$ in short

$= EY^2 + EYG(0,1)$. $Y \perp\!\!\!\perp G(0,1)$

$= EY^2 + EY\,E\,G(0,1)$

$= 2\sigma^2$

$EX_2^2 = E^2 X_2 + \operatorname{Var} X_2$

$= 0 + 2\sigma^2 + 1$

$= 2\sigma^2 + 1$

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

3.1.3. $\hat{y} = w_1 x_1 + w_2 x_2$

$\hat{Y} = w_1 X_1 + w_2 X_2$

$$J = E[(Y - \hat{Y})^2] = E[(Y - w_1X_1 - w_2X_2)^2]$$

$$= E[Y^2 - 2w_1X_1Y - 2w_2X_2Y + w_1^2X_1^2 + w_2^2X_2^2 + 2w_1w_2X_1X_2]$$

$EX_1^2 = E^2X_1 + \operatorname{Var}(X_1)$

$= 0 + \sigma^2 = \sigma^2$

$$= EY^2 - 2w_1E(X_1Y) - 2w_2E(X_2Y) + w_1^2E(X_1^2) + w_2^2E(X_2^2) + 2w_1w_2E(X_1X_2)$$

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

$$= \sigma^2(2 - 2w_1 - 4w_2 + w_1^2 + 2w_2^2 + 2w_1w_2) + w_2^2$$

$E(X_1X_2) = E(X_1(X_1 + G(0,\sigma^2) + G(0,1))$

$= E(X_1^2) + E(X_1)E(G(0,\sigma^2)) + E(X_1)E(G(0,1))$

$= \sigma^2$ $\quad X_1 \perp\!\!\!\perp G(0,\sigma^2), X_1 \perp\!\!\!\perp G(0,1)$

$E(X_1Y) = E(X_1(X_1 + G(0,\sigma^2)))$

$= EX_1^2 + EX_1E(G(0,\sigma^2))$

$= EX_1^2 = \sigma^2$

$$\partial J/\partial w_1 = -2\sigma^2 + 2w_1\sigma^2 + 2w_2\sigma^2 = 0$$

$$\partial J/\partial w_2 = -4\sigma^2 + 4w_2\sigma^2 + 2w_1\sigma^2 + 2w_2 = 0$$

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

$\Rightarrow$ as $\sigma^2\downarrow$, $\Rightarrow \begin{cases} w_1 \uparrow \\ w_2 \downarrow, \end{cases}$ we have causual effect that Y is more related to X_1 than X_2.

Therefore the model will generalize more to X_1 in test, which could potentially destroy the generalization in all.

3.3. $\hat{Y} = 2(m_1w_1X_1 + m_2w_2X_2)$

$$E[J] = E[(Y - \hat{Y})^2] = E[(Y - 2m_1w_1X_1 - 2m_2w_2X_2)^2]$$

$E(m_1) = E(m_2) = 1/2$

$E(m_1^2) = \operatorname{Var}(m_1) + E(m_1)^2 = (\tfrac{1}{2})^2 + (\tfrac{1}{2})^2 = \tfrac{1}{2}$

$$= E[Y^2 + 4m_1^2w_1^2X_1^2 + 4m_2^2w_2^2X_2^2 - 4m_1w_1X_1Y - 4m_2w_2X_2Y + 8m_1m_2w_1w_2X_1X_2]$$

$$= EY^2 + 4Em_1^2w_1^2EX_1^2 + 4Em_2^2w_2^2EX_2^2 - 4Em_1w_1EX_1Y - 4Em_2w_2EX_2Y + 8Em_1m_2w_1w_2EX_1X_2$$

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

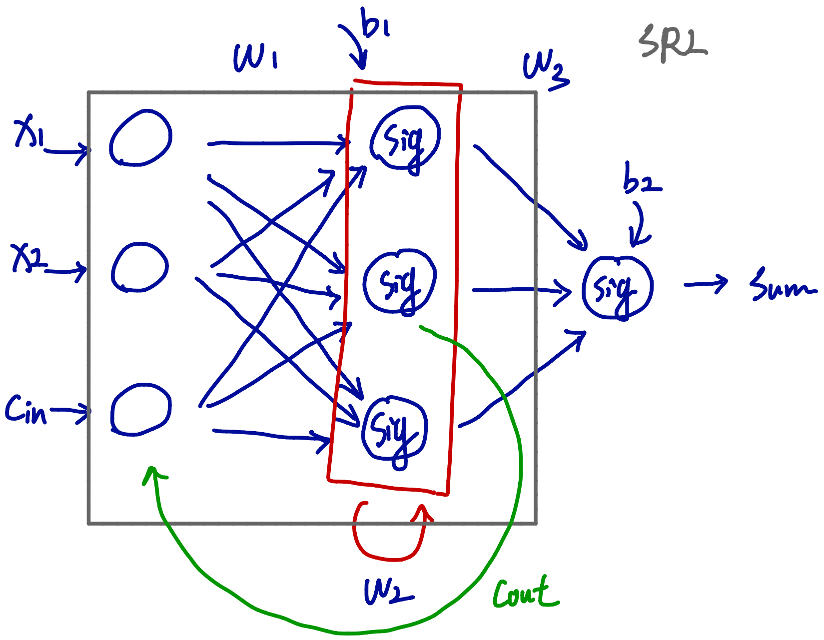
$$\partial EJ/\partial w_1 = 4w_1\sigma^2 - 2\sigma^2 + 2w_2\sigma^2 = 0$$

$$\partial EJ/\partial w_2 = 4(2\sigma^2+1)w_2 - 4\sigma^2 + 2w_1\sigma^2 = 0$$

$$\Rightarrow \begin{cases} w_1 = \dfrac{2\sigma^2+2}{7\sigma^2+4} \\ w_2 = \dfrac{3\sigma^2}{7\sigma^2+4} \end{cases}$$

Applying dropout changes would definitely help on the generalization, since Y is more evenly influenced by w_1 and w_2, and less affected by σ^2.

A "Simple Recurrent Layer" component (SRL):



$$W_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \quad b_1 = \begin{bmatrix} -0.5 \\ -1.5 \\ -2.5 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad W_2 h = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \end{bmatrix} = \begin{bmatrix} h_{12} \\ h_{12} \\ h_{12} \end{bmatrix}$$

$\Rightarrow$ each time, h_{12} output is working as a carry-on bit.

$$W_3 = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \qquad b_2 = -0.1$$

$\mathrm{sig}(x_1+x_2+Cin-0.5)$

$\mathrm{sig}(x_1+x_2+Cin-1.5)$

$\mathrm{sig}(x_1+x_2+Cin-2.5)$

x_1	x_2	Cin	Sumup	Sum	Cout	$\sim h_{11}$	$\sim h_{12}$	$\sim h_{13}$
0	0	0	0	0	0	0	0	0
0	1	0	1	1	0	0	0	1
1	0	0	1	1	0	0	0	1
0	0	1	1	1	0	0	0	1
0	1	1	2	0	1	0	1	1
1	1	0	2	0	1	0	1	1
1	0	1	2	0	1	0	1	1
1	1	1	3.	1	1	1	1	1