CSC413: Homework 3

Tianyu Du (1003801647)

2020/03/06 at 22:02:57

1 Weight Decay

1.1 Under-parameterized Model [0pt]

Solution. Given

$$\mathcal{J}(\hat{\mathbf{w}}) = \frac{1}{2n}\|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}} \tag{1.1}$$

$$\mathcal{J}(\hat{\mathbf{w}}) = \frac{1}{2n}\|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}||\hat{\mathbf{w}}||_2^2 \tag{1.2}$$

The gradient descent converges when $\frac{d}{d\hat{\mathbf{w}}}\mathcal{J}(\hat{\mathbf{w}}) = 0$. Altogether with the fact that $\frac{d}{d\mathbf{x}}||\mathbf{x}||_2^2 = 2\mathbf{x}^T$, the training converges if and only if:

$$\frac{d}{d\hat{\mathbf{w}}}\mathcal{J}(\hat{\mathbf{w}}) = \frac{1}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X + \lambda\hat{\mathbf{w}}^T \tag{1.3}$$

$$= \frac{1}{n}\hat{\mathbf{w}}^T X^T X - \frac{1}{n}\mathbf{t}^T X + \lambda\hat{\mathbf{w}}^T = 0 \quad (\dagger) \tag{1.4}$$

Note that when $d \leq n$, $rank(X) = d$ implies $X^T X$ is invertible. Suppose $X^T X + n\lambda I$ is invertible as well. Therefore,

$$(\dagger) \implies \left(\hat{\mathbf{w}}^T X^T X + n\lambda\hat{\mathbf{w}}^T\right) = \mathbf{t}^T X \tag{1.5}$$

$$\implies \hat{\mathbf{w}}^T\left(X^T X + n\lambda I\right) = \mathbf{t}^T X \tag{1.6}$$

$$\implies \hat{\mathbf{w}}^T = \mathbf{t}^T X\left(X^T X + n\lambda I\right)^{-1} \tag{1.7}$$

$$\implies \hat{\mathbf{w}} = \left(X^T X + n\lambda I\right)^{-1} X^T\mathbf{t} \tag{1.8}$$

■

1.2 Over-parameterized Model

1.2.1 Warmup: Visualizing Weight Decay [1pt]

Solution. Given

$$\mathbf{x}_1 = (2, 1) \text{ and } t_1 = 2. \tag{1.9}$$

From previous homework, the solution of gradient descent without regularization is

$$\mathbf{w}^* = \left(\frac{4}{5}, \frac{2}{5}\right) \tag{1.10}$$

From the previous part, the gradient descent converges if and only if

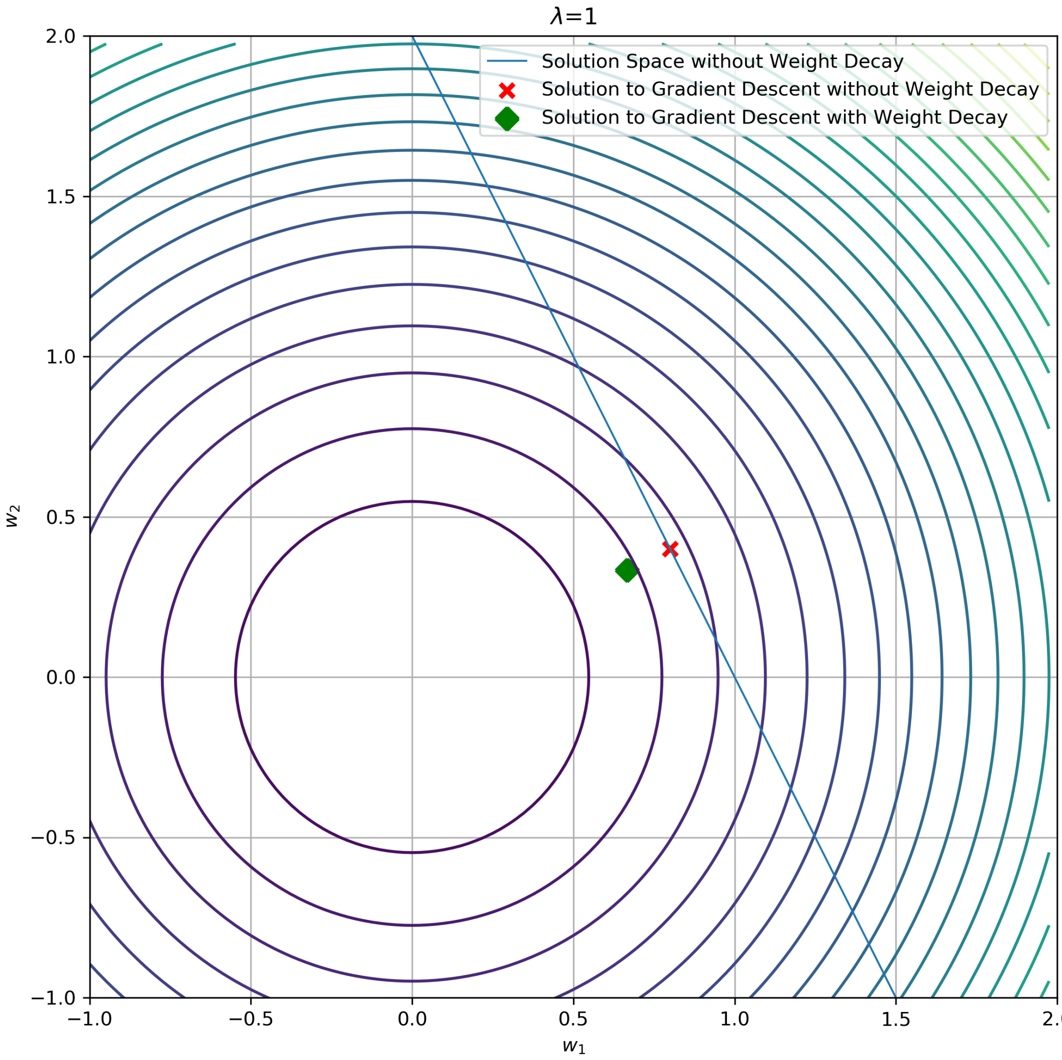
$$\begin{align}
\frac{d}{d\hat{\mathbf{w}}}\mathcal{J}(\hat{\mathbf{w}}) &= \frac{1}{n}(X\hat{\mathbf{w}} - \mathbf{t})^T X + \lambda \hat{\mathbf{w}}^T \tag{1.11}\\
&= ((2,1)\cdot(w_1, w_2) - 2)(2,1) + \lambda(w_1, w_2) \tag{1.12}\\
&= (2w_1 + w_2 - 2)(2,1) + (\lambda w_1, \lambda w_2) = 0 \tag{1.13}\\
& \tag{1.14}
\end{align}$$

Therefore, the solution to gradient descent with weight decay is:

$$\begin{cases} 4w_1 + 2w_2 - 4 + \lambda w_1 & = 0 \\ 2w_1 + w_2 - 2 + \lambda w_2 & = 0 \end{cases} \tag{1.15}$$

$$\implies \begin{cases} w_1^* & = \frac{4}{\lambda+5} \\ w_2^* & = \frac{2}{\lambda+5} \end{cases} \tag{1.16}$$

Figure 1.1: Visualization



■

1.2.2 Gradient Descent and Weight Decay [0pt]

Solution. The solution to gradient descent with weight decay at rate λ has been derived in the previous section:

$$\mathbf{w}^{*\lambda}_{\text{weight decay}} = \left(\frac{4}{\lambda+5}, \frac{2}{\lambda+5}\right) \tag{1.17}$$

■

1.3 Adaptive optimizer and Weight Decay [1pt]

Solution. Note that the original loss function,

$$\mathcal{J}(\hat{\mathbf{w}}) = \frac{1}{2n}\|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 \tag{1.18}$$

is a convex function since its a composite of convex function $||\cdot||_2^2$ and linear function $X\hat{\mathbf{w}} - \mathbf{t}$. Further, note that for any $\theta \in (0, 1)$ and $\mathbf{x} \neq \mathbf{y}$:

$$\begin{aligned}
&\theta||\mathbf{x}||_2^2 + (1-\theta)||\mathbf{y}||_2^2 - ||\theta\mathbf{x} + (1-\theta)\mathbf{y}||_2^2 && (1.19)\\
&= \theta||\mathbf{x}||_2^2 + (1-\theta)||\mathbf{y}||_2^2 - \theta^2||\mathbf{x}||_2^2 - (1-\theta)^2||\mathbf{y}||_2^2 - 2\theta(1-\theta)\langle\mathbf{x}, \mathbf{y}\rangle && (1.20)\\
&= \theta(1-\theta)||\mathbf{x}||_2^2 + \theta(1-\theta)||\mathbf{y}||_2^2 - 2\theta(1-\theta)\langle\mathbf{x}, \mathbf{y}\rangle && (1.21)\\
&= \theta(1-\theta)(||\mathbf{x}||_2^2 - 2\langle\mathbf{x}, \mathbf{y}\rangle + ||\mathbf{y}||_2^2) && (1.22)\\
&= \theta(1-\theta)||\mathbf{x} - \mathbf{y}||_2^2 > 0 && (1.23)
\end{aligned}$$

Therefore, $||\mathbf{w}||_2^2$ is strictly convex. The regularized loss function is a sum of convex and strictly convex functions, so it is strictly convex as well:

$$\mathcal{J}(\hat{\mathbf{w}}) = \frac{1}{2n}\|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 + \frac{\lambda}{2}\hat{\mathbf{w}}^\top\hat{\mathbf{w}} \tag{1.24}$$

Note that a strictly convex objective function has an (unique) global minimal, say $\mathbf{w}^*$. Hence, conventional gradient descent methods, say SGD, will converge to $\mathbf{w}^*$. Consider a SGD starting from $\mathbf{w}_0 = \mathbf{0}$, obviously

$$\mathbf{w}_0 \in Row(X) \tag{1.25}$$

and based on what we've shown in the previous homework,

$$(X^T\mathbf{w} - \mathbf{t})^T X \in Row(X) \tag{1.26}$$

Therefore,

$$\mathbf{w}_1 = \mathbf{w}_0 - \alpha(X^T\mathbf{w}_0 - \mathbf{t})^T X \in Row(X) \tag{1.27}$$

By induction,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha(X^T\mathbf{w}_t - \mathbf{t})^T X \in Row(X) \tag{1.28}$$

Therefore, the weight SGD converges to $\mathbf{w}^* \in Row(X)$. Since the objective function is strictly, so the minimizer is unique. Assuming Adagrad converges to the optimal solution, then Adagrad must converge to $\mathbf{w}^*$, which is in $Row(X)$. Hence, Adagrad with regularization converges to $Row(X)$. ■

2 Ensembles and Bias-variance Decomposition

2.1 Weight Average or Prediction Average?

2.1.1 [1pt]

Solution. Suppose there are K different models indexed using $j \in \{1, 2, \cdots, K\}$:

$$h_j(\mathbf{x}) = \mathbf{w}_j(\mathcal{D}_j)\mathbf{x} + b_j(\mathcal{D}_j) \tag{2.1}$$

where $\mathcal{D}_j$ are i.i.d. realization of datasets. Let $\overline{h}(\mathbf{x})$ denote the weight average ensemble:

$$\begin{align}
\overline{h}(\mathbf{x}) &= \overline{\mathbf{w}}\mathbf{x} + \overline{b} \tag{2.2}\\
&= \frac{1}{K}\left(\sum_{j=1}^{K} \mathbf{w}_j(\mathcal{D}_j)\right)\mathbf{x} + \frac{1}{K}\sum_{j=1}^{K} b_j(\mathcal{D}_j) \tag{2.3}\\
&= \frac{1}{K}\left(\sum_{j=1}^{K} \mathbf{w}_j(\mathcal{D}_j)\mathbf{x}\right) + \frac{1}{K}\sum_{j=1}^{K} b_j(\mathcal{D}_j) \tag{2.4}\\
&= \frac{1}{K}\left(\mathbf{w}_j(\mathcal{D}_j)\mathbf{x} + b_j(\mathcal{D}_j)\right) \tag{2.5}\\
&= \frac{1}{K}\sum_{j=1}^{K} h_j(\mathbf{x}) \tag{2.6}
\end{align}$$

which is the same as the prediction average at query point $\mathbf{x}$. Therefore, the prediction from weight-average and prediction-average ensembles are the same. Hence, the expected generalization error should be the same. ■

2.1.2 [0pt]

Solution. ■

2.2 Bagging - Uncorrelated Models

2.2.1 Bias with bagging [0pt]

Solution. Note that

$$\begin{align}
\mathbb{E}\left[\bar{h}(\mathbf{x};\mathcal{D})|\mathbf{x}\right] &= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] \tag{2.7}\\
&= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] \tag{2.8}
\end{align}$$

Since $\mathcal{D}_i$ are drawn from the identical distribution, so that

$$\mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] = \mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_j\right)|\mathbf{x}\right] \quad \forall i, j \in \{1, 2, \cdots, k\} \tag{2.9}$$

Hence,

$$\mathbb{E}\left[\bar{h}(\mathbf{x};\mathcal{D})|\mathbf{x}\right] = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] = \mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] \quad \forall i \in \{1,2,\cdots,k\} \tag{2.10}$$

Since each data point in $\mathcal{D}_i$ is uniformly sampled with replaced from $\mathcal{D} \sim p_{\text{data}}$, therefore, $\mathcal{D}_i \sim p_{\text{data}}$ as well.

$$\mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}\right)|\mathbf{x}\right] = \mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] \quad \forall i \in \{1,2,\cdots,k\} \tag{2.11}$$

Therefore,

$$bias = \mathbb{E}\left[\left|\mathbb{E}\left[\bar{h}(\mathbf{x};\mathcal{D})|\mathbf{x}\right] - y_*(\mathbf{x})\right|^2\right] = \mathbb{E}\left[\left|\mathbb{E}\left[h(\mathbf{x};\mathcal{D})|\mathbf{x}\right] - y_*(\mathbf{x})\right|^2\right] \tag{2.12}$$

■

2.2.2 Variance with bagging [1pt]

Solution. Suppose

$$\mathbb{E}\left[\left|h(\mathbf{x};\mathcal{D}) - \mathbb{E}\left[h(x;\mathcal{D})|\mathbf{x}\right]\right|^2\right] = \sigma^2 \tag{2.13}$$

For the bagging model,

$$\begin{aligned}
Var(\overline{h}) &= \mathbb{E}\left[\left|\overline{h}(\mathbf{x};\mathcal{D}) - \mathbb{E}\left[\overline{h}(x;\mathcal{D})|\mathbf{x}\right]\right|^2\right] &(2.14)\\
&= \mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[\overline{h}(x;\mathcal{D})|\mathbf{x}\right]\right|^2\right] &(2.15)\\
&= \mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[\overline{h}(x;\mathcal{D})|\mathbf{x}\right]\right)^2\right] &(2.16)
\end{aligned}$$

Since $\mathbb{E}\left[\overline{h}(x;\mathcal{D})|\mathbf{x}\right]$ is constant for all realizations of datasets $\mathcal{D}_i$ and equals $\mathbb{E}\left[h(x;\mathcal{D})|\mathbf{x}\right]$ by linearity of expectation.

$$\begin{aligned}
&= \mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^{k}\left\{h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[h(x;\mathcal{D}_i)|\mathbf{x}\right]\right\}\right)^2\right] &(2.17)\\
&= \frac{1}{k^2}\mathbb{E}\left[\left(\sum_{i=1}^{k}\left\{h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[h(x;\mathcal{D}_i)|\mathbf{x}\right]\right\}\right)^2\right] \quad (\dagger) &(2.18)
\end{aligned}$$

Because datasets $\mathcal{D}_i$ are drawn independently,

$$\mathbb{E}\left[\left(h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[h(x;\mathcal{D}_i)|\mathbf{x}\right]\right)\left(h\left(\mathbf{x};\mathcal{D}_j\right) - \mathbb{E}\left[h(x;\mathcal{D}_j)|\mathbf{x}\right]\right)\right] = Cov(h_i, h_j) = 0 \tag{2.19}$$

Hence, after expanding the squared sum in (†),

$$(\dagger) = \frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^{k}\left(h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[h(x;\mathcal{D}_i)|\mathbf{x}\right]\right)^2\right] \tag{2.20}$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}\left[\left(h\left(\mathbf{x};\mathcal{D}_i\right) - \mathbb{E}\left[h(x;\mathcal{D}_i)|\mathbf{x}\right]\right)^2\right] \tag{2.21}$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}Var(h_i) \tag{2.22}$$

Since $\mathcal{D}_i$ are i.i.d. from the dataset, $Var(h_i) = Var(h)$ for every i, therefore,

$$Var(\bar{h}) = \frac{\sigma^2}{k} \tag{2.23}$$

■

2.3 Bagging - General Case

2.3.1 Bias under Correlation [1pt]

Solution. The bias does not change and is independent from ρ. While deriving the bias, we firstly exchanged the expectation and summation using the linearity of summation operator:

$$\mathbb{E}\left[\bar{h}(\mathbf{x};\mathcal{D})|\mathbf{x}\right] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] \tag{2.24}$$

$$= \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] \tag{2.25}$$

The linearity of expectation holds regardless of the correlation. Then we used the fact that $\mathcal{D}_i$ are identically distributed to show

$$\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}_i\right)|\mathbf{x}\right] = \mathbb{E}\left[h\left(\mathbf{x};\mathcal{D}\right)|\mathbf{x}\right] \tag{2.26}$$

this step did not assume independent distribution. The entire proof (please refer to 2.2.1 for a more detailed derivation) did not use any assumption on distributional independence, hence the original proof is still valid in the general case. ■

2.3.2 Variance under Correlation [0pt]

Solution.

$$
\begin{aligned}
Var(\bar{h}) &\equiv Cov\left(h(\mathbf{x};\mathcal{D}_i), h(\mathbf{x};\mathcal{D}_i)\right) && (2.27)\\
&= Cov\left(\frac{1}{k}\sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right), \frac{1}{k}\sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right)\right) && (2.28)\\
&= \frac{1}{k^2} Cov\left(\sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right), \sum_{i=1}^{k} h\left(\mathbf{x};\mathcal{D}_i\right)\right) && (2.29)\\
&= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} Cov\left(h\left(\mathbf{x};\mathcal{D}_j\right), h\left(\mathbf{x};\mathcal{D}_j\right)\right) && (2.30)\\
&= \frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k} Cov\left(h\left(\mathbf{x};\mathcal{D}_j\right), h\left(\mathbf{x};\mathcal{D}_j\right)\right) && (2.31)\\
&= \frac{1}{k^2}(k\sigma^2 + (k^2 - k)\rho\sigma^2) && (2.32)\\
&= \left(\frac{1}{k} + \rho - \frac{\rho}{k}\right)\sigma^2 && (2.33)\\
&= \left(\rho + \frac{1-\rho}{k}\right)\sigma^2 && (2.34)
\end{aligned}
$$

■

2.3.3 Intuitions on Bagging [1pt]

Proof. Solution When $\rho = 1$, that is, all bootstrapped datasets are perfectly correlated. In fact, all datasets are identical, the variance is independent from k, and increasing number of estimators, k, does not help reduce the variance.

However, For any $\rho < 1$, increasing number of estimators in the bagging, k, helps reduce the variance. In particular, when $\rho = 0$, which is the uncorrelated dataset case, the effect is most significant: the variance shrinks linearly in k. ■

3 Generalization and Dropout

3.1 Regression Coefficients

3.1.1 Regression from X_1 [0pt]

Solution. ■

3.1.2 Regression from X_2 [1pt]

Solution. Since we are using X_2 only, equivalently, we can set the weight of X_1 to zero:

$$\begin{aligned}
\mathcal{J} &= \mathbb{E}_{(x_2,y)\sim(X_2,Y)}[(y-\hat{y})^2] && (3.1)\\
&= \mathbb{E}_{(x_2,y)\sim(X_2,Y)}[(y-w_2x_2)^2] && (3.2)\\
&= \mathbb{E}_{(x_2,y)\sim(X_2,Y)}[(y-w_2(y+Gaussian(0,1)))^2] && (3.3)\\
&= \mathbb{E}_{(x_2,y)\sim(X_2,Y)}[((1-w_2)y-w_2Gaussian(0,1))^2] && (3.4)\\
&= \mathbb{E}_{(x_2,y)\sim(X_2,Y)}[((1-w_2)y)^2]+w_2^2\mathbb{E}_{(x_2,y)\sim(X_2,Y)}[Gaussian(0,1)^2] && (3.5)\\
&= (1-w_2)^2\mathbb{E}_{y\sim Y}[y^2]+w_2^2 && (3.6)
\end{aligned}$$

Taking the gradient and solve the first order condition:

$$\begin{aligned}
&\nabla_{w_2}(1-w_2)^2\mathbb{E}_{y\sim Y}[y^2]+w_2^2=0 && (3.7)\\
\implies\ &-2(1-w_2)\mathbb{E}_{y\sim Y}[y^2]+2w_2=0 && (3.8)\\
\implies\ &\mathbb{E}_{y\sim Y}[y^2]-w_2\mathbb{E}_{y\sim Y}[y^2]-w_2=0 && (3.9)\\
\implies\ &\mathbb{E}_{y\sim Y}[y^2]+w_2(1-\mathbb{E}_{y\sim Y}[y^2])=0 && (3.10)\\
\implies\ &w_2=\frac{\mathbb{E}_{y\sim Y}[y^2]}{\mathbb{E}_{y\sim Y}[y^2]+1} && (3.11)
\end{aligned}$$

The expectation of y^2 is

$$\begin{aligned}
\mathbb{E}_{y\sim Y}[y^2] &= \mathbb{E}_{x_1\sim X_1}(x_1+Gaussian(0,\sigma^2))^2 && (3.12)\\
&= 2\sigma^2 && (3.13)\\
\implies\ w_2 &= \frac{2\sigma^2}{2\sigma^2+1} && (3.14)
\end{aligned}$$

■

3.1.3 Regression from (X_1, X_2) [1pt]

Solution. Let G_1, G_2, G_3 denote the three Gaussian distributions respectively, so that

$$\begin{aligned}
X_1 &\leftarrow G_1 && (3.15)\\
Y &\leftarrow X_1+G_2 && (3.16)\\
X_2 &\leftarrow Y+G_3 && (3.17)
\end{aligned}$$

So that,

$$\begin{aligned}
\mathcal{J} &= \mathbb{E}_{(x_1,x_2,y)\sim(X_1,X_2,Y)}[(y-\hat{y})^2] && (3.18)\\
&= \mathbb{E}[G_1+G_2-w_1G_1-w_2(G_1+G_2+G_3)]^2 && (3.19)\\
&= \mathbb{E}[(1-w_1-w_2)G_1+(1-w_2)G_2-w_2G_3]^2 && (3.20)\\
&= (1-w_1-w_2)^2\sigma^2+(1-w_2)^2\sigma^2+w_2^2 && (3.21)
\end{aligned}$$

For w_1:

$$\frac{\partial}{\partial w_1}\mathcal{J} = -2(1 - w_1 - w_2)\sigma^2 = 0 \tag{3.22}$$

For w_2:

$$\frac{\partial}{\partial w_2}\mathcal{J} = -2(1 - w_1 - w_2)\sigma^2 - 2(1 - w_2)\sigma^2 + 2w_2 = 0 \tag{3.23}$$

Solving two equations:

$$w_1 = \frac{1}{\sigma^2 + 1} \tag{3.24}$$

$$w_2 = \frac{\sigma^2}{\sigma^2 + 1} \tag{3.25}$$

The solution does not generalize well if σ changes since both w_1 and w_2 depend on and are sensitive to σ. ■

3.1.4 Different σs [0pt]

Solution. The expected loss can be re-written using law of total expectation as

$$\mathcal{L}^2 = \frac{1}{2}\mathbb{E}_{(x_1,x_2,y)\sim(X_1,X_2,Y)}[(y - \hat{y})^2|\sigma = \sigma_1] + \frac{1}{2}\mathbb{E}_{(x_1,x_2,y)\sim(X_1,X_2,Y)}[(y - \hat{y})^2|\sigma = \sigma_2] \tag{3.26}$$

Therefore,

$$\sigma_*^2 = \frac{\sigma_1^2 + \sigma_2^2}{2} \tag{3.27}$$

$$\text{and } w_1 = \frac{1}{\sigma_*^2 + 1} \tag{3.28}$$

$$w_2 = \frac{\sigma_*^2}{\sigma_*^2 + 1} \tag{3.29}$$

■

3.2 Dropout as Data-Dependent $L2$ Regularization

3.2.1 Expectation and variance of predictions [0pt]

Solution. Let

$$\tilde{y} = 2\left(m_1 w_1 x_1 + m_2 w_2 x_2\right) \tag{3.30}$$

Then

$$\mathbb{E}\left[\tilde{y}\right] = \mathbb{E}\left[2\left(m_1 w_1 x_1 + m_2 w_2 x_2\right)\right] \tag{3.31}$$

■

3.3 Effect on Dropout [1pt]

Solution. Using bias-variance decomposition of the generalization error while assuming zero irreducible error:

$$\begin{aligned}
\mathbb{E}[\tilde{\mathcal{L}}] &= \mathbb{E}[(\tilde{y} - y)^2] && (3.32)\\
&= \mathbb{E}\left[(\mathbb{E}_m[\tilde{y}] - y)^2\right] + Var(\hat{y}) && (3.33)\\
&= \mathbb{E}\left[(\hat{y} - y)^2\right] + Var(2\,(m_1 w_1 x_1 + m_2 w_2 x_2)) && (3.34)\\
&= \mathbb{E}\left[(\hat{y} - y)^2\right] + 4Var\,(m_1 w_1 x_1 + m_2 w_2 x_2) && (3.35)\\
&= \mathbb{E}\left[(\hat{y} - y)^2\right] + 4Var\,(m_1 w_1 x_1) + 4Var\,(m_2 w_2 x_2) && (3.36)\\
&= \mathbb{E}\left[(\hat{y} - y)^2\right] + Var(x_1) w_1^2 + Var(x_2) w_2^2 && (3.37)\\
&= \mathbb{E}\left[(\hat{y} - y)^2\right] + \sigma^2 w_1^2 + (2\sigma^2 + 1) w_2^2 && (3.38)\\
&= (1 - w_1 - w_2)^2 \sigma^2 + (1 - w_2)^2 \sigma^2 + w_2^2 + \sigma^2 w_1^2 + (2\sigma^2 + 1) w_2^2 && (3.39)
\end{aligned}$$

Solving the first order condition $\nabla_{\mathbf{w}} \mathcal{L} = 0$ gives

$$\begin{cases} w_1 &= \frac{2+2\sigma^2}{4+7\sigma^2} \\ w_2 &= \frac{3\sigma^2}{4+7\sigma^2} \end{cases} \tag{3.40}$$

The squared-norm $\mathbf{w}_{\text{dropped out}}$ is

$$||\mathbf{w}_{\text{dropped out}}||_2^2 = \frac{13\sigma^4 + 8\sigma^2 + 4}{(7\sigma^2 + 4)^2} \tag{3.41}$$

For the original solution:

$$||\mathbf{w}_{\text{regular}}||_2^2 = \frac{\sigma^4 + 1}{(\sigma^2 + 1)^2} \tag{3.42}$$

The ratio of two norms is

$$r = \frac{||\mathbf{w}_{\text{dropped out}}||_2^2}{||\mathbf{w}_{\text{regular}}||_2^2} = \frac{(\sigma^2 + 1)^2 (13\sigma^4 + 8\sigma^2 + 4)}{(7\sigma^2 + 4)^2 (\sigma^4 + 1)} \tag{3.43}$$

For $\sigma \geq 0$, $r(\sigma)$ attains its maximal value of 0.417 at $\sigma \approx 1.11$. Hence, adding the dropout always reduce the norm of solution $\mathbf{w}^*$, but does not necessarily reduce every entry in $\mathbf{w}^*$. Therefore, adding the dropout is equivalent to adding a regularization term in which the level of plenty (λ_j) for each w_j depends on the variance of x_j. In this case, w_1 is more regularized. And such regularization would help the model achieve a better generalization error. ■

4 Hard-Coding Recurrent Neural Networks [1pt]

Solution. Let $\sigma = \frac{1}{1+\exp(-100\times z)}$, so that the sigmoid function behaves like an hard threshold indicator function $\mathbb{1}\{x \geq 0\}$. In the following part of my answer, I am considering σ as a threshold function. Let $\mathbf{x}_t = (x_1^t, x_2^t)$ denotes the input feature at time t. Note that when weights are sufficient large in scale, σ behaves like hard threshold function. Consider the following recurrent network:

$$\begin{aligned}
\hat{y}_t &= \sigma(\mathbf{w}_{hy} \mathbf{h}_t + b_y) && (4.1)\\
\mathbf{h}_t &= \sigma\left(\mathbf{w}_{xh} \mathbf{x}_t + \mathbf{w}_{hh} \mathbf{h}_{t-1} + \mathbf{b}_h\right) && (4.2)
\end{aligned}$$

with the following parameters:

$$\mathbf{w}_{xh} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \mathbf{w}_{hh} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \mathbf{b}_h = \begin{pmatrix} -0.5 \\ -1.5 \\ -2.5 \end{pmatrix} \tag{4.3}$$

$$\mathbf{w}_{hy} = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix} \quad b_y = -0.5 \tag{4.4}$$

$$\mathbf{h}_t = \begin{pmatrix} \mathbb{1}\{x_1^t + x_2^t + h_{t-1} \geq 1\} \\ \mathbb{1}\{x_1^t + x_2^t + h_{t-1} \geq 2\} \\ \mathbb{1}\{x_1^t + x_2^t + h_{t-1} \geq 3\} \end{pmatrix} \tag{4.5}$$

Justification:

$$\mathbf{w}_{xh}\mathbf{x}_t = \begin{pmatrix} x_1^t + x_2^t \\ x_1^t + x_2^t \\ x_1^t + x_2^t \end{pmatrix} \quad \mathbf{w}_{hh}\mathbf{h}_{t-1} = \begin{pmatrix} \mathbb{1}\{x_1^{t-1} + x_2^{t-1} + h_{t-2} \geq 2\} \\ \mathbb{1}\{x_1^{t-1} + x_2^{t-1} + h_{t-2} \geq 2\} \\ \mathbb{1}\{x_1^{t-1} + x_2^{t-1} + h_{t-2} \geq 2\} \end{pmatrix} \tag{4.6}$$

Let c_t denote the carry from the previous significant figure. Therefore, elements in $\mathbf{w}_{hh}\mathbf{h}_{t-1}$ are one only if $c_t = 1$. Then,

$$\mathbf{h}_t = \sigma\left(\mathbf{w}_{xh}\mathbf{x}_t + \mathbf{w}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h\right) = \begin{pmatrix} \mathbb{1}\{x_1^t + x_2^t + c_t \geq 1\} \\ \mathbb{1}\{x_1^t + x_2^t + c_t \geq 2\} \\ \mathbb{1}\{x_1^t + x_2^t + c_t \geq 3\} \end{pmatrix} \tag{4.7}$$

For the output layer,

$$\hat{y}_t = \sigma\left(\mathbf{w}_{xh}\mathbf{x}_t + \mathbf{w}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h\right) = \mathbb{1}\{x_1^t + x_2^t + c_t \geq 1\} \vee \mathbb{1}\{x_1^t + x_2^t + c_t \geq 3\} \tag{4.8}$$

Therefore, let $c \in \{0, 1\}$ denote the carry, $\hat{y}$ whenever $x_1 + x_2 + c$ is one or three, and $\hat{y} = 0$ otherwise. ∎