

CSC413: Homework 3

Tianyu Du (1003801647)

2020/03/04 at 20:57:44

1 Weight Decay

1.1 Under-parameterized Model [0pt]

Solution. Given

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

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

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:

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

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

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

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

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

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

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

■

1.2 Over-parameterized Model

1.2.1 Warmup: Visualizing Weight Decay [1pt]

Solution. Given

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

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

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

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

$$\frac{d}{d\hat{\mathbf{w}}} \mathcal{J}(\hat{\mathbf{w}}) = \frac{1}{n} (X\hat{\mathbf{w}} - \mathbf{t})^T X + \lambda \hat{\mathbf{w}}^T \quad (1.11)$$

$$= ((2, 1) \cdot (w_1, w_2) - 2)(2, 1) + \lambda(w_1, w_2) \quad (1.12)$$

$$= (2w_1 + w_2 - 2)(2, 1) + (\lambda w_1, \lambda w_2) = 0 \quad (1.13)$$

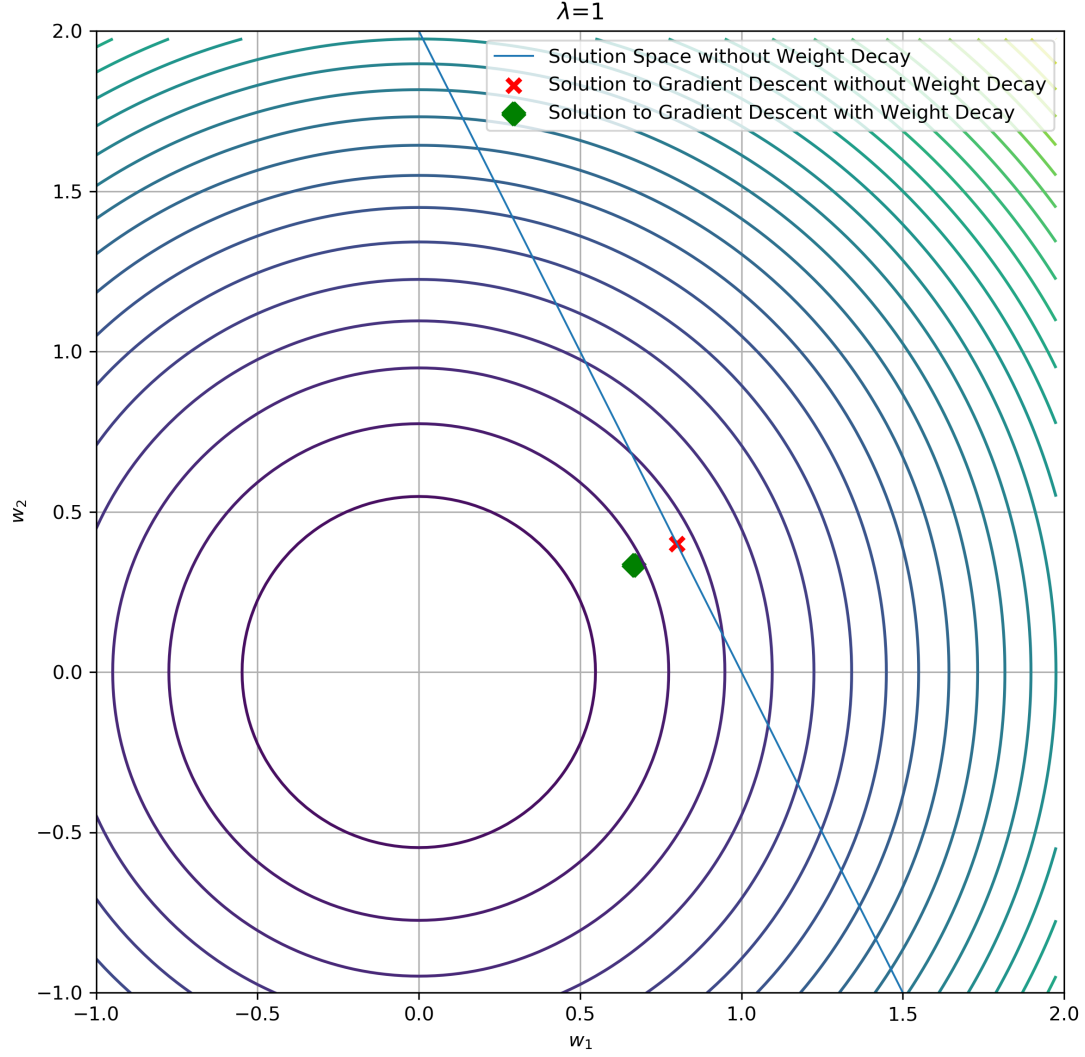
$$(1.14)$$

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

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

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

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:

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

■

1.3 Adaptive optimizer and Weight Decay [1pt]

Solution. ■

2 Ensembles and Bias-variance Decomposition

2.1 Weight Average or Prediction Average?

2.1.1 [1pt]

Solution. Without loss of generality, assume the bias is zero. This is equivalent to inserting a column of ones to the X , so that $X \in \mathbb{R}^{n \times (d+1)}$, and we can ignore the bias. Suppose there are K different models indexed using $j \in \{1, 2, \dots, K\}$:

$$h_j(\mathbf{x}) = \mathbf{w}_j(\mathcal{D}_j)\mathbf{x} \quad (2.1)$$

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

$$\bar{h}(\mathbf{x}) = \bar{\mathbf{w}}\mathbf{x} \quad (2.2)$$

$$= \sum_{j=1}^K \mathbf{w}_j(\mathcal{D}_j)\mathbf{x} \quad (2.3)$$

And the prediction average at dataset point \mathbf{x} is simply

$$\sum_{j=1}^K h_j(\mathbf{x}) \quad (2.4)$$

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

$$\mathbb{E} [\bar{h}(\mathbf{x}; \mathcal{D}) | \mathbf{x}] = \mathbb{E} \left[\frac{1}{k} \sum_{i=1}^k h(\mathbf{x}; \mathcal{D}_i) | \mathbf{x} \right] \quad (2.5)$$

$$= \frac{1}{k} \sum_{i=1}^k \mathbb{E} [h(\mathbf{x}; \mathcal{D}_i) | \mathbf{x}] \quad (2.6)$$

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

$$\mathbb{E} [h(\mathbf{x}; \mathcal{D}_i) | \mathbf{x}] = \mathbb{E} [h(\mathbf{x}; \mathcal{D}_j) | \mathbf{x}] \quad \forall i, j \in \{1, 2, \dots, k\} \quad (2.7)$$

Hence,

$$\mathbb{E} [\bar{h}(\mathbf{x}; \mathcal{D}) | \mathbf{x}] = \frac{1}{k} \sum_{i=1}^k \mathbb{E} [h(\mathbf{x}; \mathcal{D}_i) | \mathbf{x}] = \mathbb{E} [h(\mathbf{x}; \mathcal{D}_i) | \mathbf{x}] \quad \forall i \in \{1, 2, \dots, k\} \quad (2.8)$$

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.

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

Therefore,

$$bias = \mathbb{E} \left[\left| \mathbb{E} [\bar{h}(\mathbf{x}; \mathcal{D}) | \mathbf{x}] - y_*(\mathbf{x}) \right|^2 \right] = \mathbb{E} \left[\left| \mathbb{E} [h(\mathbf{x}; \mathcal{D}) | \mathbf{x}] - y_*(\mathbf{x}) \right|^2 \right] \quad (2.10)$$

■

2.2.2 Variance with bagging [1pt]

Solution. Suppose

$$\mathbb{E} \left[|h(\mathbf{x}; \mathcal{D}) - \mathbb{E} [h(x; \mathcal{D}) | \mathbf{x}]|^2 \right] = \sigma^2 \quad (2.11)$$

For the bagging model,

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

$$= \mathbb{E} \left[\left| \frac{1}{k} \sum_{i=1}^k h(\mathbf{x}; \mathcal{D}_i) - \mathbb{E} [\bar{h}(x; \mathcal{D}) | \mathbf{x}] \right|^2 \right] \quad (2.13)$$

$$= \mathbb{E} \left[\left(\frac{1}{k} \sum_{i=1}^k h(\mathbf{x}; \mathcal{D}_i) - \mathbb{E} [\bar{h}(x; \mathcal{D}) | \mathbf{x}] \right)^2 \right] \quad (2.14)$$

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

$$= \mathbb{E} \left[\left(\frac{1}{k} \sum_{i=1}^k \{h(\mathbf{x}; \mathcal{D}_i) - \mathbb{E} [h(x; \mathcal{D}_i) | \mathbf{x}]\} \right)^2 \right] \quad (2.15)$$

$$= \frac{1}{k^2} \mathbb{E} \left[\left(\sum_{i=1}^k \{h(\mathbf{x}; \mathcal{D}_i) - \mathbb{E} [h(x; \mathcal{D}_i) | \mathbf{x}]\} \right)^2 \right] \quad (\dagger) \quad (2.16)$$

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

$$\mathbb{E} [(h(\mathbf{x}; \mathcal{D}_i) - \mathbb{E} [h(x; \mathcal{D}_i) | \mathbf{x}]) (h(\mathbf{x}; \mathcal{D}_j) - \mathbb{E} [h(x; \mathcal{D}_j) | \mathbf{x}])] = Cov(h_i, h_j) = 0 \quad (2.17)$$

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

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

$$= \frac{1}{k^2} \sum_{i=1}^k \mathbb{E} \left[(h(\mathbf{x}; \mathcal{D}_i) - \mathbb{E}[h(x; \mathcal{D}_i) | \mathbf{x}])^2 \right] \quad (2.19)$$

$$= \frac{1}{k^2} \sum_{i=1}^k \text{Var}(h_i) \quad (2.20)$$

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

$$\text{Var}(\bar{h}) = \frac{\sigma^2}{k} \quad (2.21)$$

■

2.3 Bagging - General Case

2.3.1 Bias under Correlation [1pt]

2.3.2 Variance under Correlation [0pt]

2.3.3 Intuitions on Bagging [1pt]

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:

$$\mathcal{J} = \mathbb{E}_{(x_2, y) \sim (X_2, Y)} [(y - \hat{y})^2] \quad (3.1)$$

$$= \mathbb{E}_{(x_2, y) \sim (X_2, Y)} [(y - w_2 x_2)^2] \quad (3.2)$$

$$= \mathbb{E}_{(x_2, y) \sim (X_2, Y)} [(y - w_2(y + \text{Gaussian}(0, 1)))^2] \quad (3.3)$$

$$= \mathbb{E}_{(x_2, y) \sim (X_2, Y)} [((1 - w_2)y - w_2 \text{Gaussian}(0, 1))^2] \quad (3.4)$$

$$= \mathbb{E}_{(x_2, y) \sim (X_2, Y)} [((1 - w_2)y)^2] + w_2^2 \mathbb{E}_{(x_2, y) \sim (X_2, Y)} [\text{Gaussian}(0, 1)^2] \quad (3.5)$$

$$= (1 - w_2)^2 \mathbb{E}_{y \sim Y} [y^2] + w_2^2 \quad (3.6)$$

Taking the gradient and solve the first order condition:

$$\nabla_{w_2}(1 - w_2)^2 \mathbb{E}_{y \sim Y}[y^2] + w_2^2 = 0 \quad (3.7)$$

$$\implies -2(1 - w_2) \mathbb{E}_{y \sim Y}[y^2] + 2w_2 = 0 \quad (3.8)$$

$$\implies \mathbb{E}_{y \sim Y}[y^2] - w_2 \mathbb{E}_{y \sim Y}[y^2] - w_2 = 0 \quad (3.9)$$

$$\implies \mathbb{E}_{y \sim Y}[y^2] + w_2(1 - \mathbb{E}_{y \sim Y}[y^2]) = 0 \quad (3.10)$$

$$\implies w_2 = \frac{\mathbb{E}_{y \sim Y}[y^2]}{\mathbb{E}_{y \sim Y}[y^2] + 1} \quad (3.11)$$

The expectation of y^2 is

$$\mathbb{E}_{y \sim Y}[y^2] = \mathbb{E}_{x_1 \sim X_1}(x_1 + \text{Gaussian}(0, \sigma^2))^2 \quad (3.12)$$

$$= 2\sigma^2 \quad (3.13)$$

$$\implies w_2 = \frac{2\sigma^2}{2\sigma^2 + 1} \quad (3.14)$$

■

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

$$X_1 \leftarrow G_1 \quad (3.15)$$

$$Y \leftarrow X_1 + G_2 \quad (3.16)$$

$$X_2 \leftarrow Y + G_3 \quad (3.17)$$

So that,

$$\mathcal{J} = \mathbb{E}_{(x_1, x_2, y) \sim (X_1, X_2, Y)}[(y - \hat{y})^2] \quad (3.18)$$

$$= \mathbb{E}[G_1 + G_2 - w_1 G_1 - w_2(G_1 + G_2 + G_3)]^2 \quad (3.19)$$

$$= \mathbb{E}[(1 - w_1 - w_2)G_1 + (1 - w_2)G_2 - w_2 G_3]^2 \quad (3.20)$$

$$= (1 - w_1 - w_2)^2 \sigma^2 + (1 - w_2)^2 \sigma^2 + w_2^2 \sigma^2 \quad (3.21)$$

For w_1 :

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

For w_2 :

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

Solving two equations:

$$w_1 = \frac{1}{\sigma^2 + 1} \quad (3.24)$$

$$w_2 = \frac{\sigma^2}{\sigma^2 + 1} \quad (3.25)$$

3.1.4 Different σ s [0pt]

Solution.

3.2 Dropout as Data-Dependent $L2$ Regularization

3.2.1 Expectation and variance of predictions [0pt]

Solution. Let

$$\tilde{y} = 2(m_1 w_1 x_1 + m_2 w_2 x_2) \quad (3.26)$$

Then

$$\mathbb{E}[\tilde{y}] = \mathbb{E}[2(m_1 w_1 x_1 + m_2 w_2 x_2)] \quad (3.27)$$

3.3 Effect on Dropout [1pt]

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

$$\mathbb{E}[\tilde{\mathcal{L}}] = \mathbb{E}[(\tilde{y} - y)^2] \quad (3.28)$$

$$= \mathbb{E}[(\mathbb{E}_m[\tilde{y}] - y)^2] + \text{Var}(\tilde{y}) \quad (3.29)$$

$$= \mathbb{E}[(\hat{y} - y)^2] + \text{Var}(2(m_1 w_1 x_1 + m_2 w_2 x_2)) \quad (3.30)$$

$$= \mathbb{E}[(\hat{y} - y)^2] + 4\text{Var}(m_1 w_1 x_1 + m_2 w_2 x_2) \quad (3.31)$$

$$= \mathbb{E}[(\hat{y} - y)^2] + 4\text{Var}(m_1 w_1 x_1) + 4\text{Var}(m_2 w_2 x_2) \quad (3.32)$$

$$= \mathbb{E}[(\hat{y} - y)^2] + \text{Var}(x_1)w_1^2 + \text{Var}(x_2)w_2^2 \quad (3.33)$$

$$= \mathbb{E}[(\hat{y} - y)^2] + \sigma^2 w_1^2 + (2\sigma^2 + 1)w_2^2 \quad (3.34)$$

Therefore, adding the dropout is equivalent to adding a regularization term in which the level of penalty (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 $\sigma = \frac{1}{1+\exp(-z)}$, and $\mathbf{x}_t = (x_1^t, x_2^t)$ 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:

$$\hat{y}_t = \sigma(\mathbf{w}_{hy} \mathbf{h}_t + b_y) \quad (4.1)$$

$$\mathbf{h}_t = \sigma(\mathbf{w}_{xh} \mathbf{x}_t + \mathbf{w}_{hh} \mathbf{h}_{t-1} + \mathbf{b}_h) \quad (4.2)$$

with the following parameters:

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

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

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

Justification:

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

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,

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

For the output layer,

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

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. ■