

## Homework 1 Solutions

1. **Hard-Coding a Network.** The idea is that each of the hidden units in the first layer will respond to a violation of one of the inequalities. The output unit will check that there are no violations, by checking that the hidden units are all off.

$$\mathbf{W}^{(1)} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \mathbf{b}^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{W}^{(2)} = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix} \text{ and } \mathbf{b}^{(2)} = \frac{1}{2}.$$

2. **Backprop.**

- Computation graph:

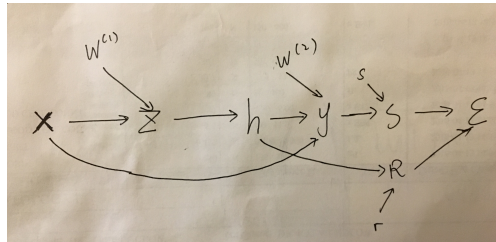


Figure 1: The computation graph for Problem 2. Showing parameters (e.g.  $\mathbf{r}$ ,  $\mathbf{s}$ , weights and biases) is optional.

- Backprop equations:

$$\begin{aligned} \bar{\mathcal{J}} &= 1 \\ \bar{\mathcal{S}} &= \bar{\mathcal{J}} \\ \bar{\mathcal{R}} &= \bar{\mathcal{J}} \\ \bar{\mathbf{y}} &= \bar{\mathcal{S}} \frac{\partial \mathcal{S}}{\partial \mathbf{y}} \\ &= \bar{\mathcal{S}}(\mathbf{y} - \mathbf{s}) \\ \bar{\mathbf{h}} &= \bar{\mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{h}} + \bar{\mathcal{R}} \frac{\partial \mathcal{R}}{\partial \mathbf{h}} \\ &= [\mathbf{W}^{(2)}]^\top \bar{\mathbf{y}} + \mathbf{r} \\ \bar{\mathbf{z}} &= \bar{\mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \\ &= \bar{\mathbf{h}} \circ \sigma'(\mathbf{z}) \\ \bar{\mathbf{x}} &= \bar{\mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \bar{\mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \\ &= [\mathbf{W}^{(1)}]^\top \bar{\mathbf{z}} + \bar{\mathbf{y}} \end{aligned}$$

3. **Sparsifying Activation Function.** There are two ways to approach this problem. First, you could write out the backprop equations. Second, you could use the fact that  $\partial\mathcal{J}/\partial w$  represents the effect on  $\mathcal{J}$  of an infinitesimal change to  $w$ , and argue whether this effect is zero. Here, we'll denote a generic activation function with  $\phi$ , ReLU with  $r$ , and the input to an activation function with  $z$ .

- $\frac{\partial\mathcal{J}}{\partial w_1}$ : YES.
  - Justification 1:  $\frac{\partial\mathcal{J}}{\partial w_1} = \bar{y} \frac{\partial y}{\partial w_1} = \bar{y} \phi'(z) h_1 = 0$  (given  $h_1=0$ ).
  - Justification 2: Since  $y = \phi(w_1 h_3)$  and  $h_3 = 0$ , changing  $w_1$  has no effect on the predictions.
- $\frac{\partial\mathcal{J}}{\partial w_2}$ : YES.
  - Justification 1:  $\frac{\partial\mathcal{J}}{\partial w_2} = \bar{h}_1 \frac{\partial h_1}{\partial w_2} = \bar{h}_1 r'(z_1) h_3 = 0$ , which is zero because  $r'(-1) = 0$ .
  - Justification 2: Changing  $w_2$  by an infinitesimal amount has no effect, because it only affects the input to  $h_1$ , which is in the flat region of the ReLU.
- $\frac{\partial\mathcal{J}}{\partial w_3}$ : NO. Changing  $w_3$  by a small amount can change  $h_3$ , which changes  $h_2$ , which changes  $y$ . Both  $h_3$  and  $h_2$  may be positive. (This argument can also be spelled out explicitly by writing out the backprop rules for each of these steps.)

## Marking Rubrics

### 1. Hard-Coding a Network.

- works only for  $\mathbb{Z}$  but not for  $\mathbb{R}$ : -1 mark
- doesn't work when some of the inputs are equal: -0.5 mark
- doesn't work for some other issue: up to 0.5 mark from 2

### 2. Backprop.

(a)

- a missing edge: -0.25 mark
- showing parameters (e.g.  $\mathbf{r}$ ,  $\mathbf{s}$ , weights and biases) is optional, no down-mark

(b)

- mistakes in the order of tensors in dot products or mismatched dimensions: -0.75 mark
- missing some parts of gradients: -1.5 mark

### 3. Sparsifying Activation Function.

- one wrong answer: -1 mark
- mistake in the reasoning: -1 mark
- all answers correct but no justification: -2 marks