Problem 1: Loss Function Calculation

Given the weight matrix $\mathbf{W}$ and four input samples $\mathbf{x}^{(i)}$, we calculate the linear scores $\mathbf{z} = \mathbf{W}\mathbf{x}$. We then compute the Multiclass SVM Loss ($L_{\text{SVM}} = \sum_{j \neq y} \max(0, z_j - z_y + 1)$) and Softmax Loss ($L_{\text{CE}} = -\log(P_y)$).

Table 1: Summary of Losses

Input	Ground Truth	SVM Loss	Softmax Loss
Sample 1	Cat	0.00	0.00
Sample 2	Cat	71.53	44.33
Sample 3	Dog	64.91	60.98
Sample 4	Horse	158.73	103.04

Sample 1: $\mathbf{x}^{(1)} = [1.52, 2.63, 5.37, 4.94]^\top$, $y = \text{cat}$.

$$\mathbf{z}^{(1)} = [16.06, -38.75, -42.47, -13.92]^\top$$

Since z_{cat} is the maximum by a margin > 1, $L_{\text{SVM}} = 0$. Similarly, $P_{\text{cat}} \approx 1$, so $L_{\text{CE}} \approx 0$.

Sample 2: $\mathbf{x}^{(2)} = [8.87, 1.25, 4.49, 0.12]^\top$, $y = \text{cat}$.

$$\mathbf{z}^{(2)} = [-17.87, -86.24, 26.47, 7.30]^\top$$

Scores are misclassified (Cow is max).

$$L_{\text{SVM}} = \max(0, 26.47 - (-17.87) + 1) + \max(0, 7.30 - (-17.87) + 1) = 45.34 + 26.17 = 71.53$$
$$L_{\text{CE}} \approx -z_{\text{cat}} + z_{\max} = 17.87 + 26.47 = 44.34$$

Sample 3: $\mathbf{x}^{(3)} = [3.22, 4.63, 3.55, 5.41]^\top$, $y = \text{dog}$.

$$\mathbf{z}^{(3)} = [26.73, -34.25, -32.32, -38.21]^\top$$

Scores are misclassified (Cat is max).

$$L_{\text{SVM}} = \max(0, 26.73 - (-34.25) + 1) + \max(0, -32.32 - (-34.25) + 1) = 61.98 + 2.93 = 64.91$$
$$L_{\text{CE}} \approx -z_{\text{dog}} + z_{\max} = 34.25 + 26.73 = 60.98$$

Sample 4: $\mathbf{x}^{(4)} = [1.38, 0.63, 2.90, 8.52]^\top$, $y = \text{horse}$.

$$\mathbf{z}^{(4)} = [45.01, -4.34, -66.23, -58.03]^\top$$

Scores are misclassified (Cat is max).

$$L_{\text{SVM}} = \max(0, 45.01 - (-58.03) + 1) + \max(0, -4.34 - (-58.03) + 1) = 104.04 + 54.69 = 158.73$$
$$L_{\text{CE}} \approx -z_{\text{horse}} + z_{\max} = 58.03 + 45.01 = 103.04$$

Problem 2: Softmax Gradient Derivation & Parameter Update

We derive the gradient of the Cross-Entropy loss $L = -\log(P_y)$ with respect to the weights $\mathbf{W}$. Let $P_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$. The gradient with respect to the logits z_j is:

$$\begin{aligned}\frac{\partial L}{\partial z_j} &= \frac{\partial}{\partial z_j}\left(-z_y + \log\sum_k e^{z_k}\right) \\ &= -\mathbb{I}(j=y) + \frac{1}{\sum_k e^{z_k}} \cdot e^{z_j} \\ &= P_j - \mathbb{I}(j=y)\end{aligned}$$

In vector notation, $\nabla_{\mathbf{z}} L = \mathbf{P} - \mathbf{Y}$. Applying the chain rule for $\mathbf{z} = \mathbf{W}\mathbf{x}$:

$$\nabla_{\mathbf{W}} L = (\mathbf{P} - \mathbf{Y})\mathbf{x}^\top$$

We calculate the gradients for the misclassified samples from Problem 1 (Sample 1 gradient is negligible).

Sample 2 ($y =$ cat): $\mathbf{P} \approx [0,0,1,0]^\top$ (Mass on Cow). $\mathbf{Y} = [1,0,0,0]^\top$.

$$\nabla_{\mathbf{W}} L_2 = \begin{bmatrix}-1\\0\\1\\0\end{bmatrix} \mathbf{x}^{(2)\top} = \begin{bmatrix}-8.87 & -1.25 & -4.49 & -0.12\\ 0 & 0 & 0 & 0\\ 8.87 & 1.25 & 4.49 & 0.12\\ 0 & 0 & 0 & 0\end{bmatrix}$$

Sample 3 ($y =$ dog): $\mathbf{P} \approx [1,0,0,0]^\top$ (Mass on Cat). $\mathbf{Y} = [0,1,0,0]^\top$.

$$\nabla_{\mathbf{W}} L_3 = \begin{bmatrix}1\\-1\\0\\0\end{bmatrix} \mathbf{x}^{(3)\top} = \begin{bmatrix}3.22 & 4.63 & 3.55 & 5.41\\ -3.22 & -4.63 & -3.55 & -5.41\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$$

Sample 4 ($y =$ horse): $\mathbf{P} \approx [1,0,0,0]^\top$ (Mass on Cat). $\mathbf{Y} = [0,0,0,1]^\top$.

$$\nabla_{\mathbf{W}} L_4 = \begin{bmatrix}1\\0\\0\\-1\end{bmatrix} \mathbf{x}^{(4)\top} = \begin{bmatrix}1.38 & 0.63 & 2.90 & 8.52\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ -1.38 & -0.63 & -2.90 & -8.52\end{bmatrix}$$

Total Gradient $\nabla_{\mathbf{W}} L = \sum \nabla L_i$:

$$\nabla_{\mathbf{W}} L = \begin{bmatrix}-4.27 & 4.01 & 1.96 & 13.81\\ -3.22 & -4.63 & -3.55 & -5.41\\ 8.87 & 1.25 & 4.49 & 0.12\\ -1.38 & -0.63 & -2.90 & -8.52\end{bmatrix}$$

Weight Update ($\eta = 0.2$):

$$\mathbf{W}_{\text{new}} = \mathbf{W} - 0.2(\nabla_{\mathbf{W}} L) = \begin{bmatrix} 0.284 & 0.438 & -3.762 & 3.668 \\ -4.886 & -0.204 & -7.340 & 4.292 \\ 2.456 & 0.730 & -3.428 & -7.694 \\ -2.034 & -1.714 & 7.510 & -6.956 \end{bmatrix}$$

Problem 3: Backpropagation on a Small Network

We compute the forward and backward pass for a network with weights $\mathbf{w} = [-1.7, 0.1, -0.6, -1.8, -0.2, 0.5]$, input $\mathbf{x} = [-0.3, 4.9, 1.1, -2.7]$, target $y = 0.7$, and squared error loss.

Forward Pass:

$$\begin{aligned} h_1 &= \sigma(w_1 x_1 + w_2 x_2) = \sigma(1.0) \approx 0.7311 \\ h_2 &= \sigma(w_3 x_3 + w_4 x_4) = \sigma(4.2) \approx 0.9852 \\ \hat{y} &= \sigma(w_5 h_1 + w_6 h_2) = \sigma(0.3464) \approx 0.5857 \end{aligned}$$

Backward Pass: Let $\delta_L = \frac{\partial L}{\partial s}$. For squared error $L = (y - \hat{y})^2$ (derivative scaled by 2) or $\frac{1}{2}(y - \hat{y})^2$, we use the gradients provided in the problem prompt logic.

$$\delta_{\text{out}} = -2(y - \hat{y}) \cdot \hat{y}(1 - \hat{y}) = -2(0.1143)(0.2426) \approx -0.05546$$

Backpropagating to hidden units:

$$\begin{aligned} \delta_1 &= \delta_{\text{out}} \cdot w_5 \cdot h_1(1 - h_1) \approx (-0.05546)(-0.2)(0.1966) \approx 0.00218 \\ \delta_2 &= \delta_{\text{out}} \cdot w_6 \cdot h_2(1 - h_2) \approx (-0.05546)(0.5)(0.01458) \approx -0.00040 \end{aligned}$$

Weight Gradients:

$$\frac{\partial L}{\partial w_1} = \delta_1 x_1 \approx \mathbf{-0.00065} \qquad \frac{\partial L}{\partial w_2} = \delta_1 x_2 \approx \mathbf{0.01068}$$

$$\frac{\partial L}{\partial w_3} = \delta_2 x_3 \approx \mathbf{-0.00044} \qquad \frac{\partial L}{\partial w_4} = \delta_2 x_4 \approx \mathbf{0.00108}$$

$$\frac{\partial L}{\partial w_5} = \delta_{\text{out}} h_1 \approx \mathbf{-0.04055} \qquad \frac{\partial L}{\partial w_6} = \delta_{\text{out}} h_2 \approx \mathbf{-0.05464}$$