NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 15: Multiclass SVM Loss Function

CONCEPTS COVERED

Concepts Covered:

- ❑ Linear Machine
- ❑ Multiclass Support Vector Machine
- ❑ Multiclass SVM Loss Function
- ❑ Optimization

Multiclass Problem: Linear Machine

$$f : R^D \rightarrow R^K$$

$$f(X_i, W, b) = WX_i + b = s$$

$$\begin{bmatrix} W_{11} & W_{12} & W_{13} & .. & W_{1D} \\ W_{21} & W_{22} & W_{23} & .. & W_{2D} \\ .. & .. & .. & .. & .. \\ W_{K1} & W_{K2} & W_{K3} & .. & W_{KD} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ .. \\ X_D \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ .. \\ b_K \end{bmatrix} = \begin{bmatrix} s_1 \\ s_2 \\ .. \\ s_K \end{bmatrix}$$

Multiclass SVM

$$\left.\begin{aligned} s_j &= f(X_i, W)_j \\ &= WX_i \end{aligned}\right\} \rightarrow Score\ \ for\ \ j^{th}\ Class\ of\ \ i^{th}\ Vector\ (X_i, y_i)$$

$$s_{y_i} = f(X_i, W)_{y_i} \quad \rightarrow \text{should be maximum}$$

$$s_{y_i} - s_j \geq \Delta$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$$

Loss Function: An Example

$$For\ some\ (X_i, y_i)\ where\ y_i = 2$$

$$s = (10\ \ 30\ \ -20\ \ 25)^t \qquad \Delta = 10$$

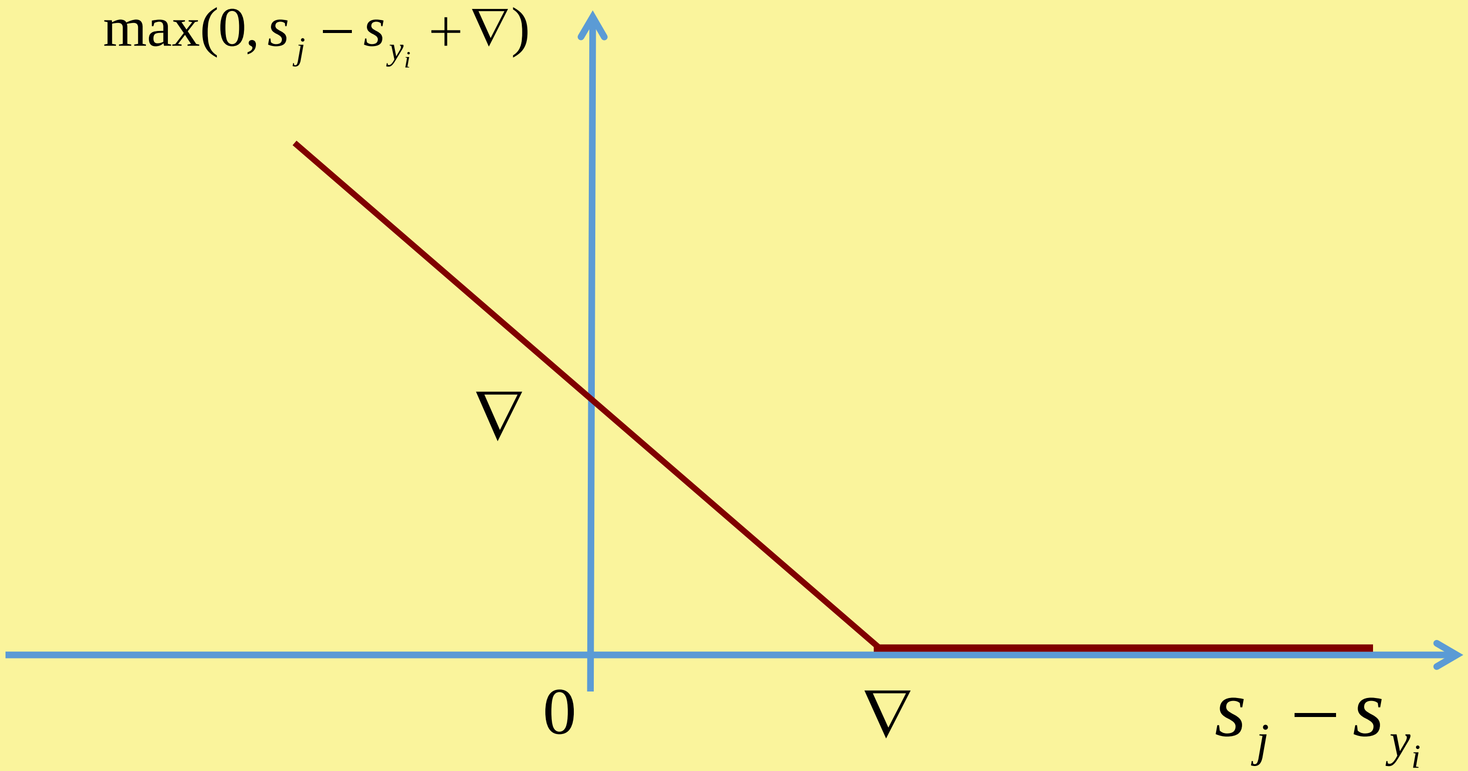
$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$$

$$= \max(0,10 - 30 + 10) + \max(0,-20 - 30 + 10) + \max(0,25 - 30 + 10)$$

$$= 0 + 0 + 15$$

$$= 15$$

Hinge Loss



Regularization

$$s_j - s_{y_i} = W_j^t X_i - W_{y_i}^t X_i$$

$$\text{Scaling } W \text{ by } \lambda : W \leftarrow \lambda W$$

$$\Downarrow$$

$$s_j - s_{y_i} \leftarrow \lambda\left(s_j - s_{y_i}\right)$$

Regularization

Include a regularization term *R(W)*

$$R(W) = \lambda \sum_k \sum_l W_{kl}^2$$

$$L = \frac{1}{N} \sum_i L_i + \lambda R(W)$$

$$L = \frac{1}{N} \sum_i \sum_{j \neq y_i} [\max(0,\ f(X_i, W)_j - f(X_i, W)_{y_i} \ \ + \nabla) + \lambda \sum_k \sum_l W_{kl}^2$$

Choice of Hyper Parameter

$$L = \frac{1}{N}\sum_{i}\sum_{j\neq y_i}[\max(0,\ f(X_i,W)_j - f(X_i,W)_{y_i} \ +\nabla) + \lambda\sum_{k}\sum_{l}W_{kl}^2$$

∇ and λ control the same tradeoff $\Rightarrow \nabla = 1$

$$L = \frac{1}{N}\sum_{i}\sum_{j\neq y_i}[\max(0,\ f(X_i,W)_j - f(X_i,W)_{y_i} +1) + \lambda\sum_{k}\sum_{l}W_{kl}^2$$

$$\text{Binary SVM} \Rightarrow L_i = C\max(0, 1 - y_i W^t X_i) + R(W)$$

Loss Function Visualization

Visualizing Loss Function

$$\text{Consider 3 Classes} \Rightarrow W = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix}$$

$$3 \;\; 1-\text{dimensional points} \Rightarrow (X_1,1), (X_2,2) \text{ and } (X_3,3)$$

Visualizing Loss Function

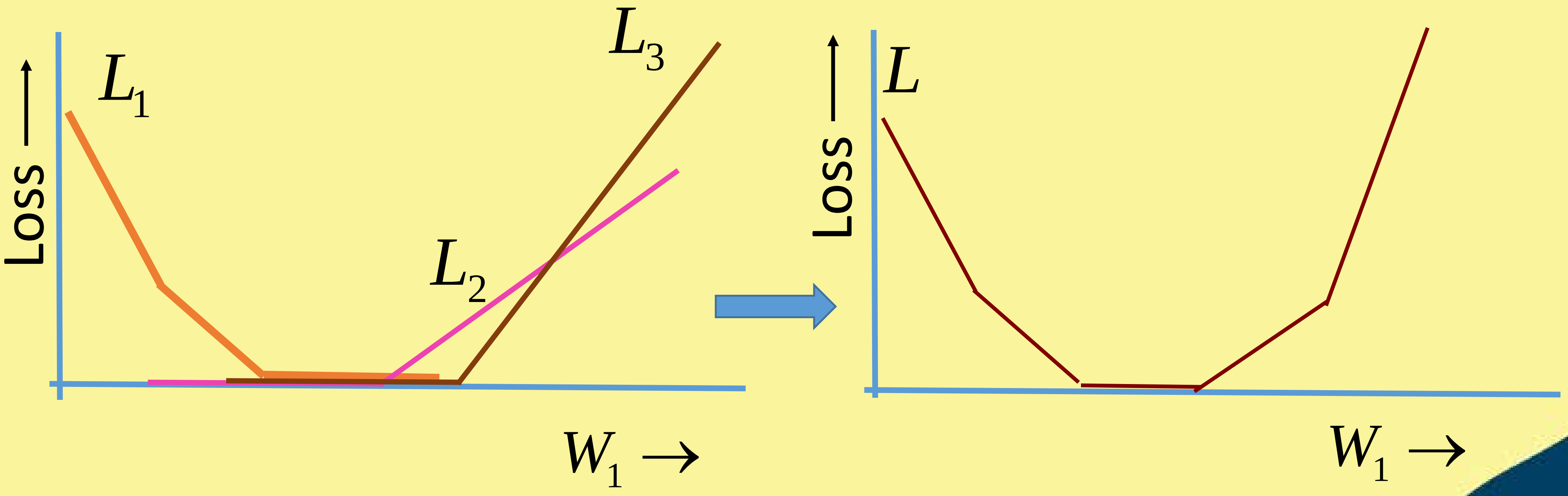
$$L_1 = \max(0, W_2^t X_1 - W_1^t X_1 + 1) + \max(0, W_3^t X_1 - W_1^t X_1 + 1)$$

$$L_2 = \max(0, W_1^t X_2 - W_2^t X_2 + 1) + \max(0, W_3^t X_2 - W_2^t X_2 + 1)$$

$$L_3 = \max(0, W_1^t X_3 - W_3^t X_3 + 1) + \max(0, W_2^t X_3 - W_3^t X_3 + 1)$$

$$L = \frac{1}{3}(L_1 + L_2 + L_3)$$

Visualizing Loss Function

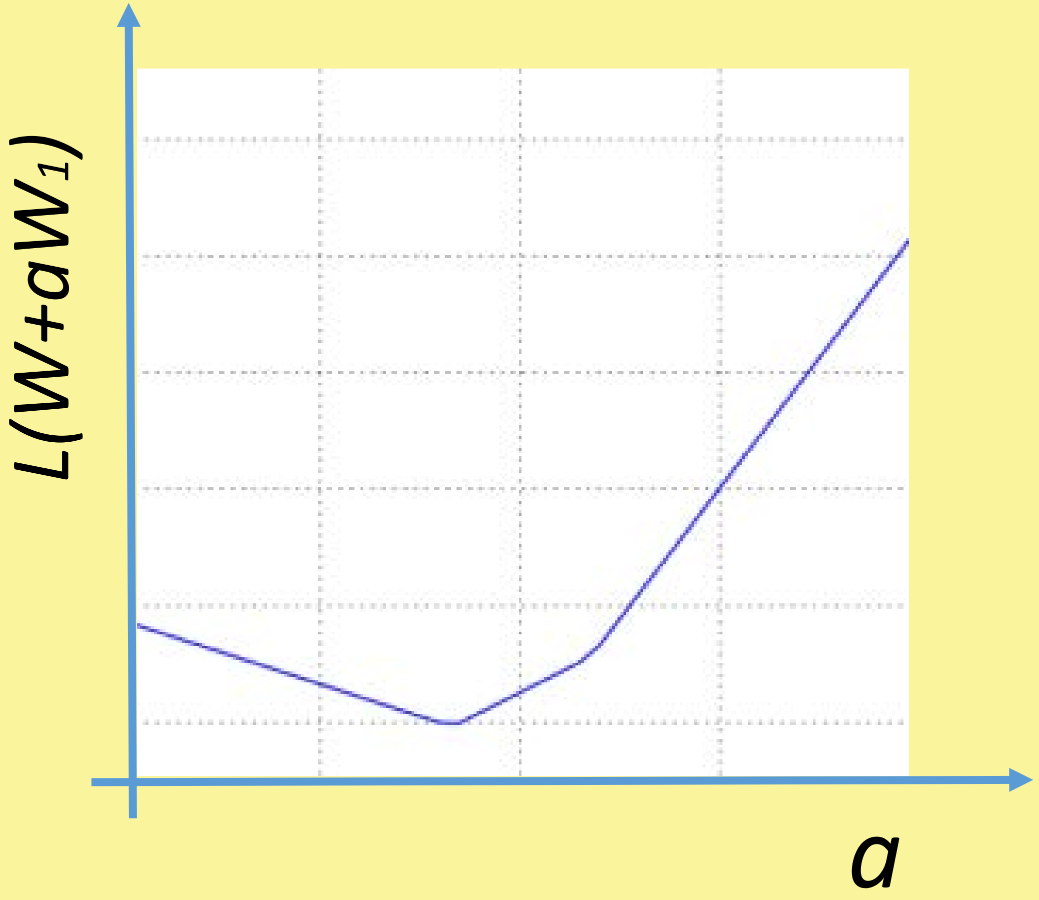


Visualizing Loss Function

- Take a random W (a single point in space)
- Take a random direction W_1
- Record the Loss along W_1

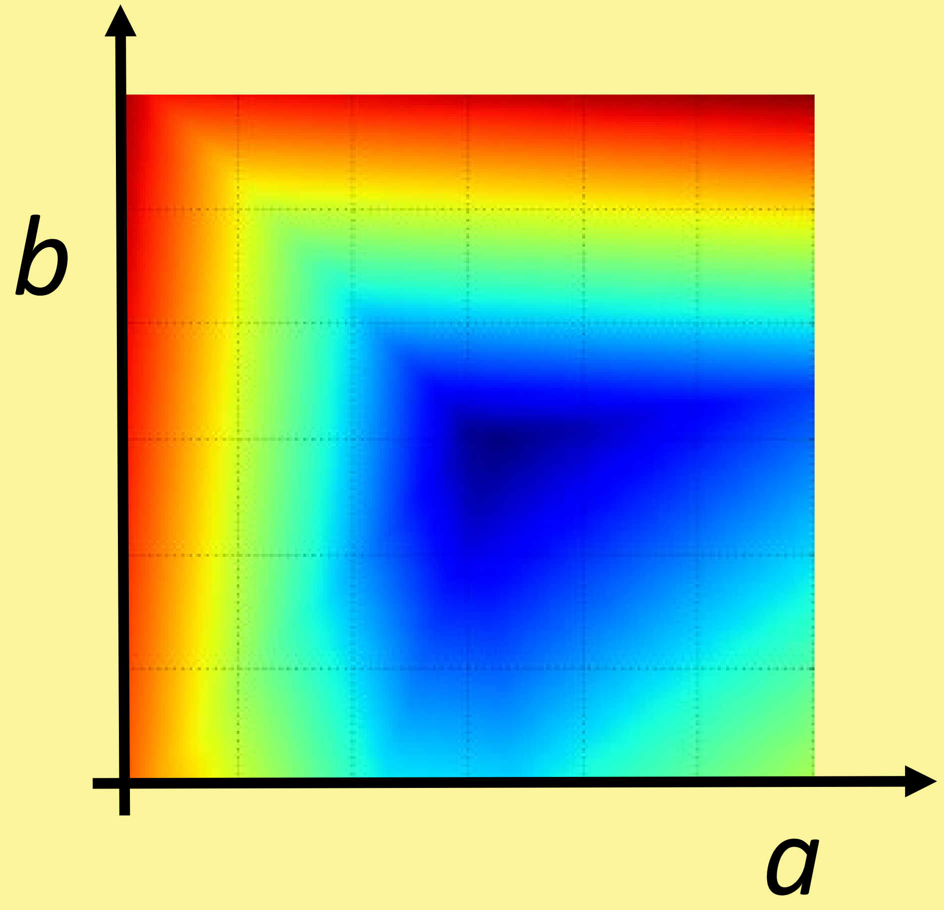
$$\Rightarrow L(W + aW_1)$$

Visualizing Loss Function



Source - http://cs231n.github.io

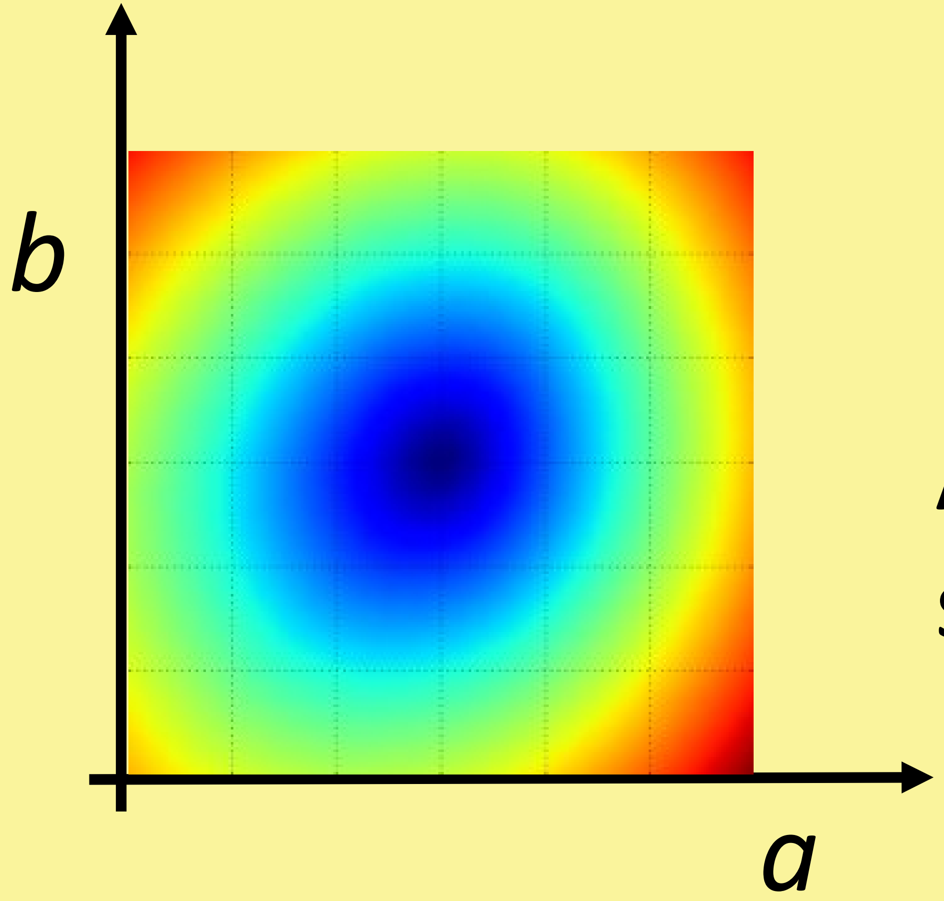
Visualizing Loss Function



$$L(W + aW_1 + bW_2)$$

Source - http://cs231n.github.io

Visualizing Loss Function



$$L(W + aW_1 + bW_2)$$

Averaged over multiple samples

Optimizing Loss Function

$$L = \frac{1}{N}\sum_{i}\sum_{j \neq y_i}\left[\max(0, W_j^t X_i - W_{y_i}^t X_i + \nabla\right] + \lambda \sum_k \sum_l W_{kl}^2$$

$$\nabla_{W_{y_i}} = -\frac{1}{N}\sum_{i}\sum_{j \neq y_i}[X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \eta W_{y_i}$$

$$\nabla_{W_j} = \frac{1}{N}\sum_{i}\sum_{j \neq y_i}[X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \xi W_j$$

Optimizing Loss Function

$$\nabla_{W_{y_i}} = -\frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \eta W_{y_i} \qquad \nabla_{W_j} = \frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \xi W_j$$

Gradient descent

$$W_{y_i}(k+1) \leftarrow (1-\eta) W_{y_i}(k) + \frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)]$$

$$W_j(k+1) = (1-\xi) W_j(k) - \frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)]$$

NPTEL ONLINE CERTIFICATION COURSES

Thank you