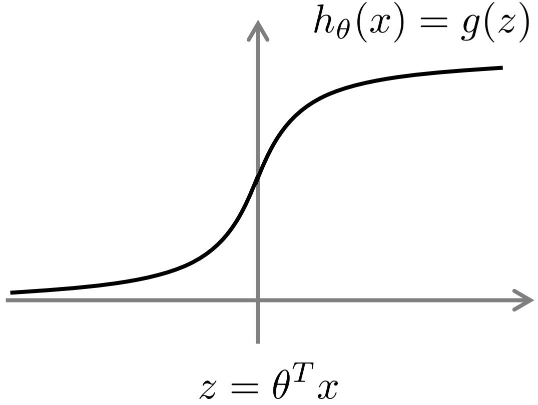
Linear Support Vector Machines (SVM)

Part I

Logistic Regression Review

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$



If $y = 1$, we want $h_\theta(x) \approx 1$, $\theta^T x \gg 0$

If $y = 0$, we want $h_\theta(x) \approx 0$, $\theta^T x \ll 0$

Binary Classification

$$x^1 \quad x^2 \quad x^3 \quad \ldots\ldots$$
$$\hat{y}^1 \quad \hat{y}^2 \quad \hat{y}^3$$

$$\hat{y}^n = 1, 0$$

- Step 1: Function set (Model)

$$g(x) = \begin{cases} f(x) > 0 & \text{Output = 1} \\ f(x) < 0 & \text{Output = 0} \end{cases}$$

- Step 2: Loss function:

$$L(f) = \sum_n \cancel{\delta(g(x^n) \neq \hat{y}^n)}$$
$$l(f(x^n), \hat{y}^n)$$

The number of times **g** yields incorrect results on training data.

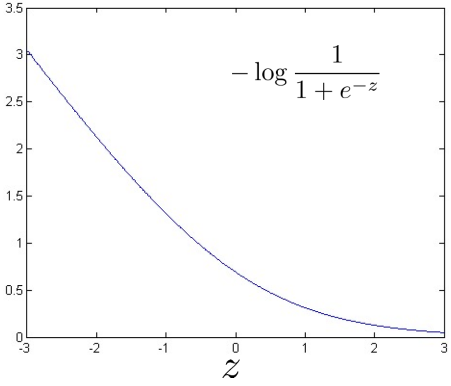
- Step 3: Gradient descent: Training by this model is difficult.

Logistic Regression Review (Log Loss)

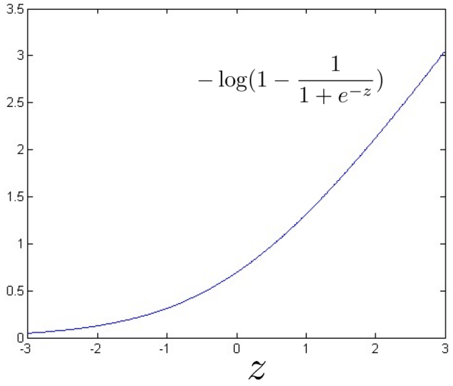
Loss of an example: $-(y \log h_\theta(x) + (1-y)\log(1-h_\theta(x)))$

$$= -y \log \frac{1}{1+e^{-\theta^T x}} - (1-y)\log(1 - \frac{1}{1+e^{-\theta^T x}})$$

If $y = 1$ (want $\theta^T x \gg 0$):



If $y = 0$ (want $\theta^T x \ll 0$):



Loss Function for $\hat{y}^n$ = 1: Part I

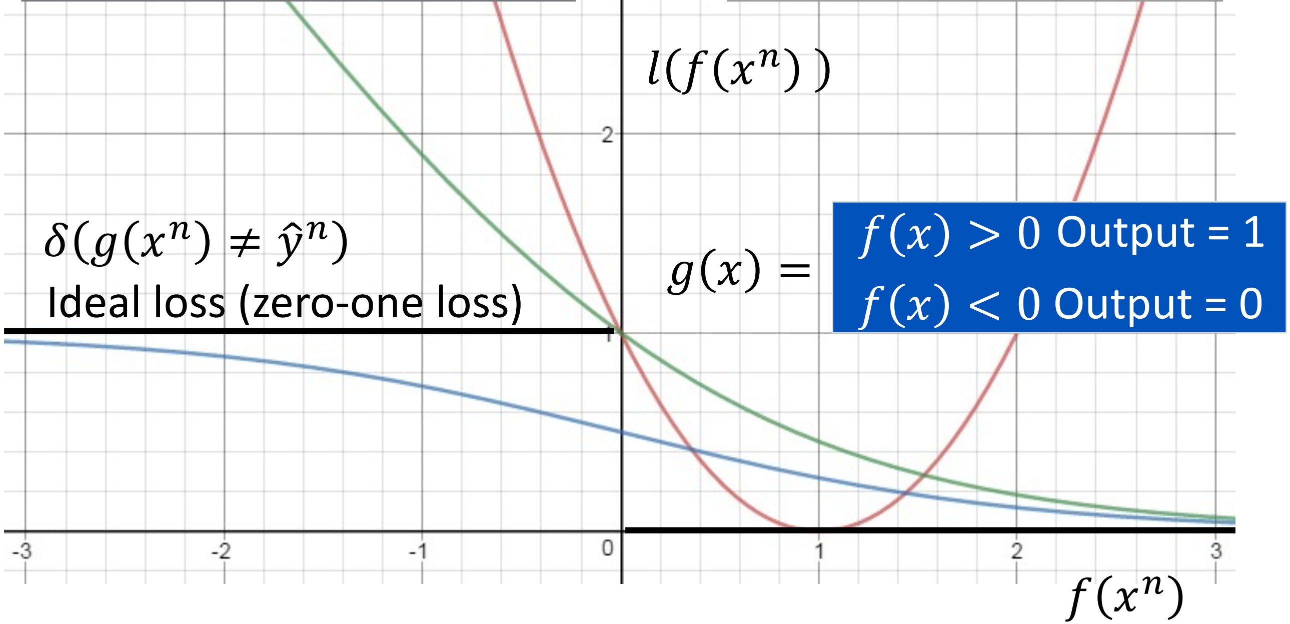
Ideal loss:

$$L(f) = \sum_n \delta(g(x^n) \neq \hat{y}^n)$$

Approximation:

$$L(f) = \sum_n l(f(x^n), \hat{y}^n)$$

Larger value, smaller loss



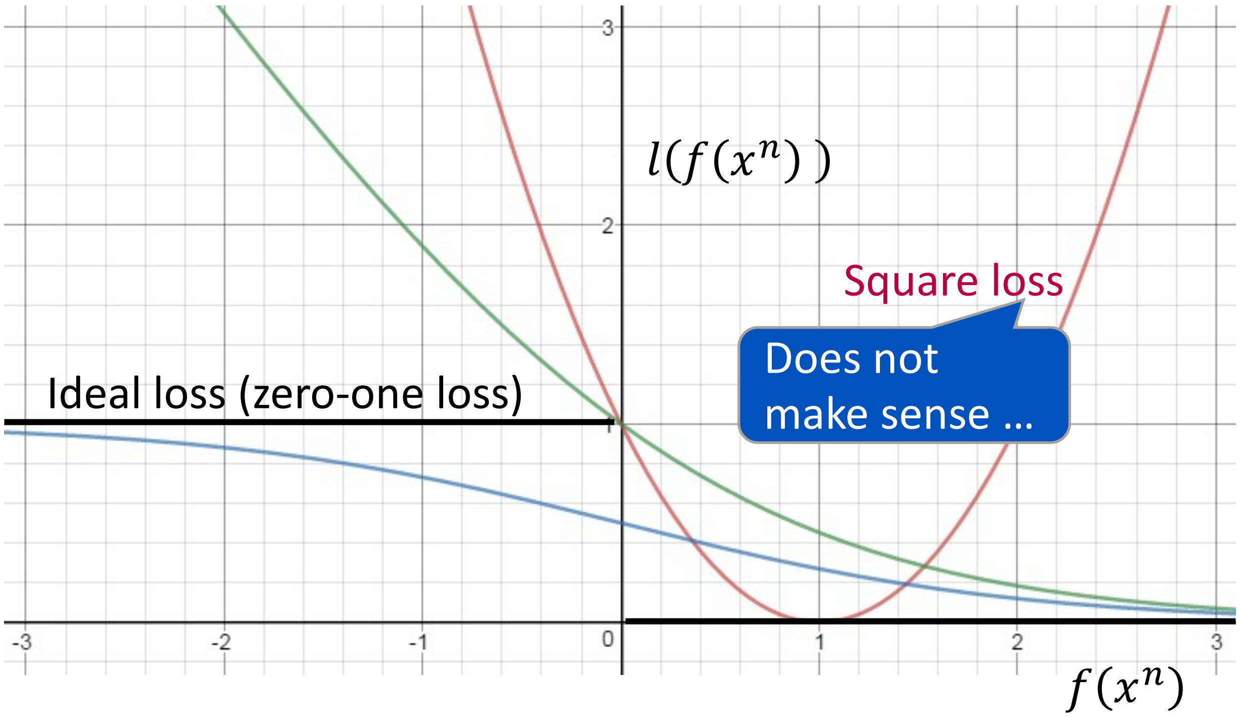
Loss Function for $\hat{y}^n = 1$: Part II

Square Loss:

If $\hat{y}^n = 1$, $f(x)$ close to 1

$$l(f(x^n), \hat{y}^n) = (f(x^n) - 1)^2$$

Larger value, smaller loss

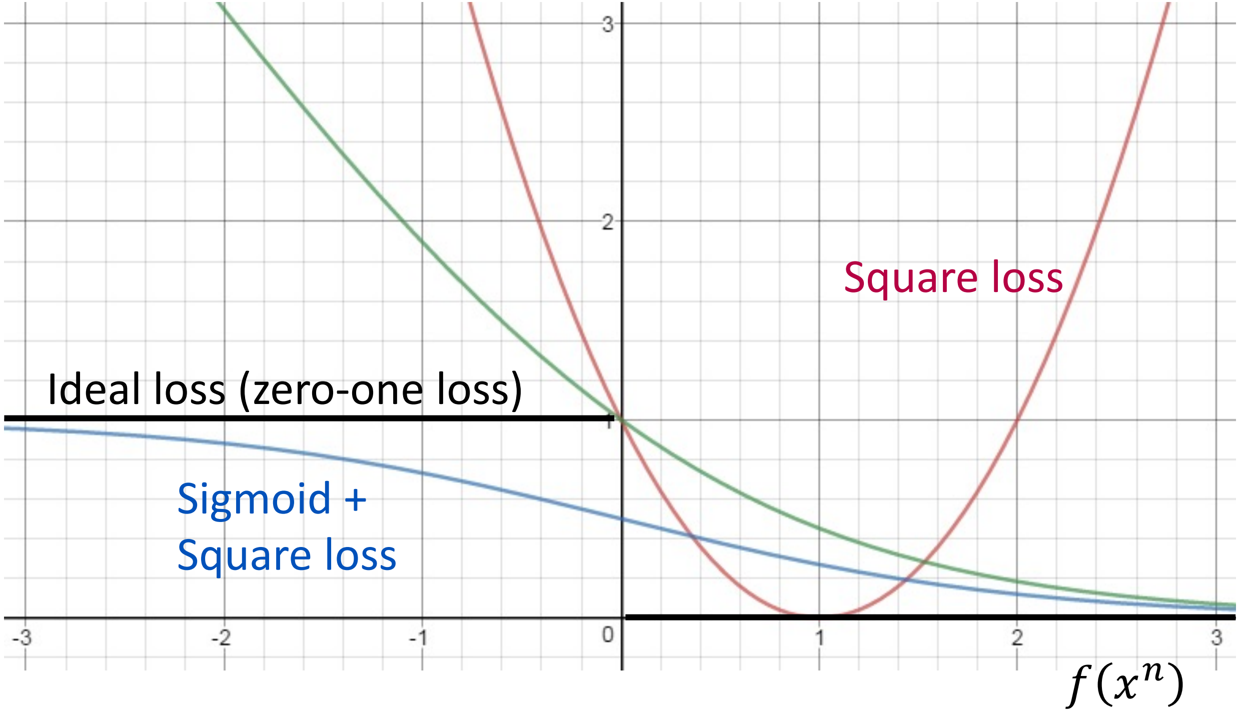


Loss Function for $\hat{y}^n = 1$: Part III

Sigmoid + Square Loss: If $\hat{y}^n = 1$, $\sigma(f(x))$ close to 1

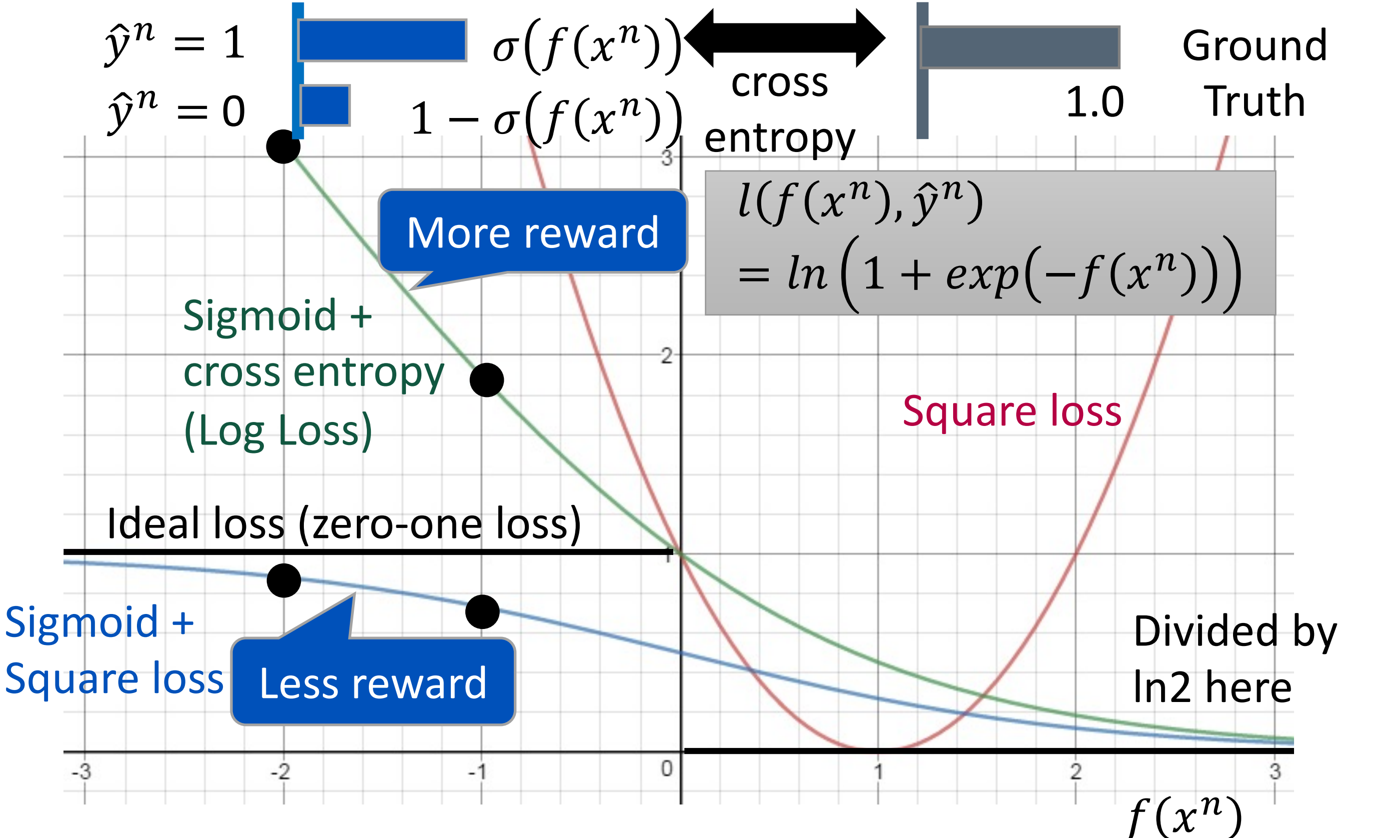
$$l(f(x^n)) = \left(\sigma(f(x^n)) - 1\right)^2$$

Larger value, smaller loss



Loss Function for $\hat{y}^n = 1$: Part IV

Sigmoid + cross entropy (logistic regression)



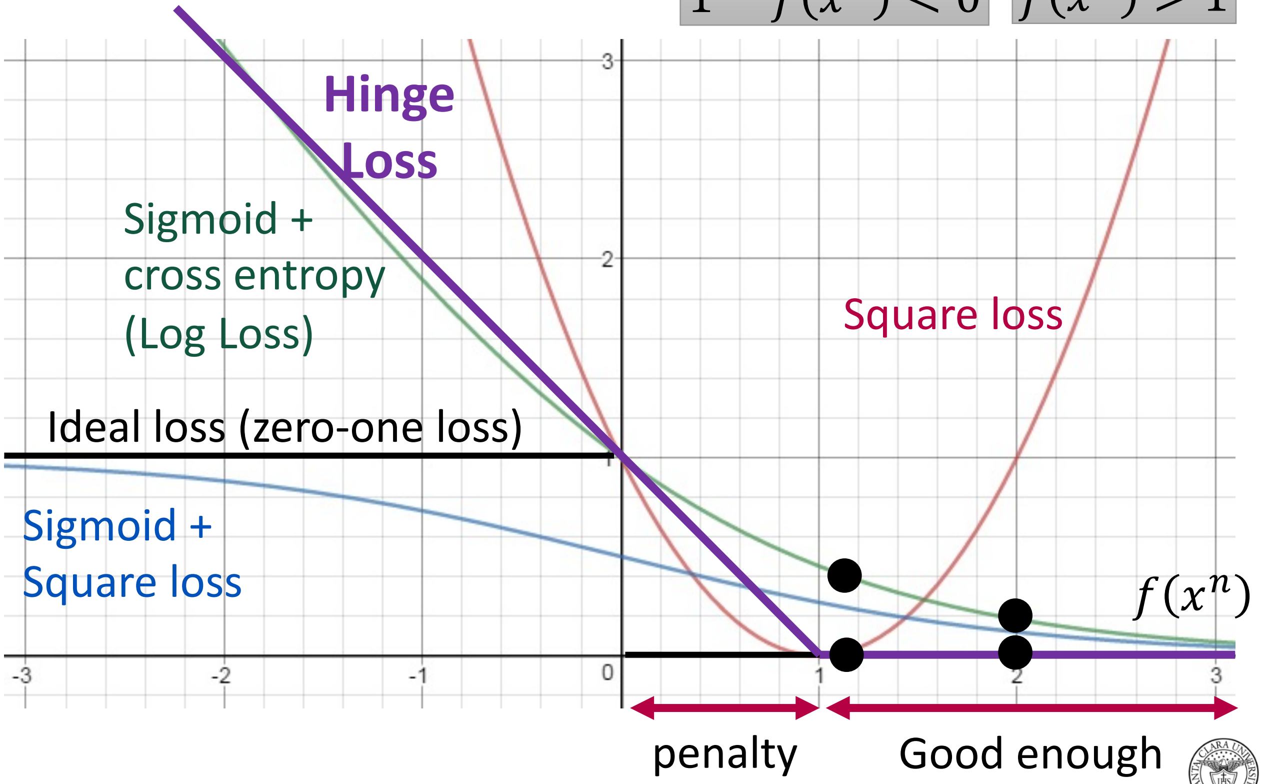
Larger value, smaller loss

Loss Function for $\hat{y}^n = 1$: Part V

$l(f(x^n), \hat{y}^n) = max(0, 1 - f(x^n))$

$1 - f(x^n) < 0$ $\quad$ $f(x^n) > 1$

Larger value, smaller loss



Logistic Regression Versus SVM With Regularization

Logistic regression (log loss):

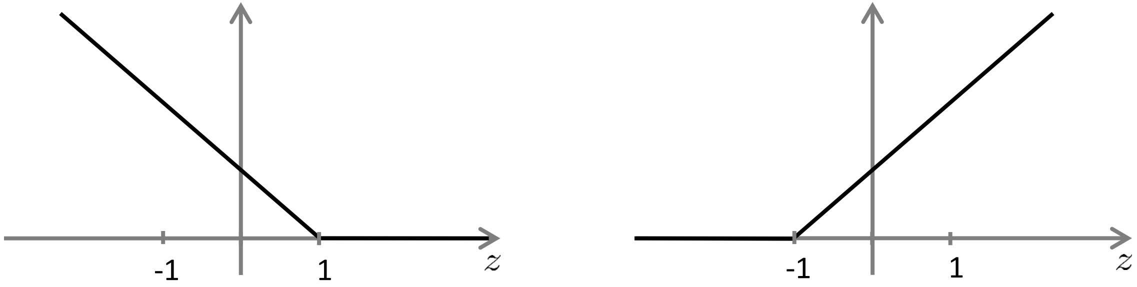
$$\min_{\theta} \frac{1}{m} \left[\sum_{i=1}^{m} y^{(i)} \left(-\log h_\theta(x^{(i)}) \right) + (1 - y^{(i)}) \left((-\log(1 - h_\theta(x^{(i)})) \right) \right] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$

Support vector machine (hinge loss):

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

SVM (Hard Margin and Soft Margin)

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$



If C is very large,

If $y = 1$, we want $\theta^T x \geq 1$ (not just ≥ 0)

If $y = 0$, we want $\theta^T x \leq -1$ (not just < 0)