Cross-Entropy: The Standard Classification Loss

Binary Cross-Entropy

$L = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$

$y = 1: -\log(\hat{y})$
$y = 0: -\log(1-\hat{y})$

High penalty for wrong prediction

Gradient Behavior

$|\nabla L|$ (Cross-Entropy)
$|\nabla L|$ (MSE)

Cross-entropy: Strong gradient, fast learning

MSE: Weak gradient (saturates)
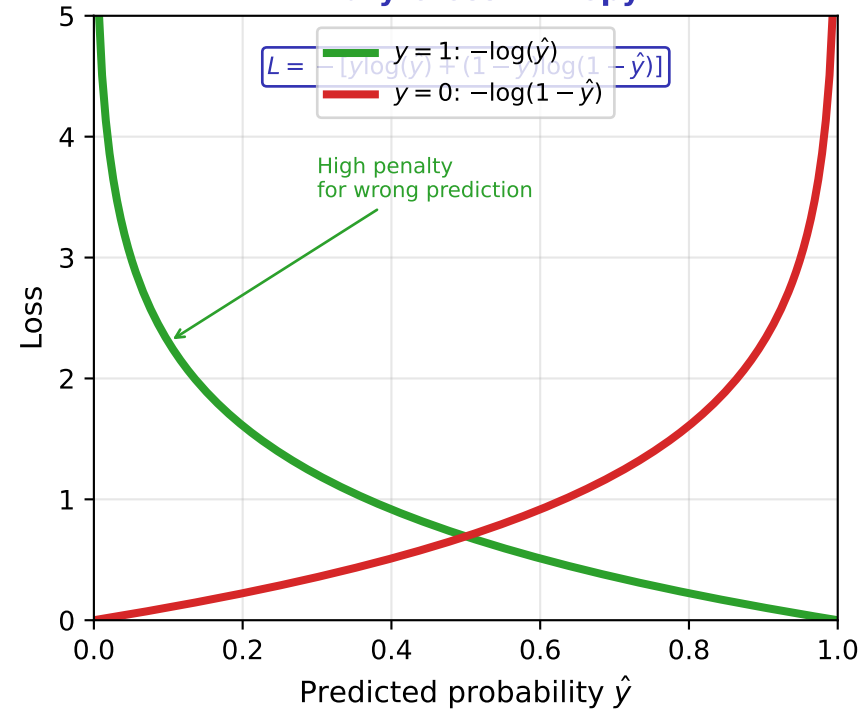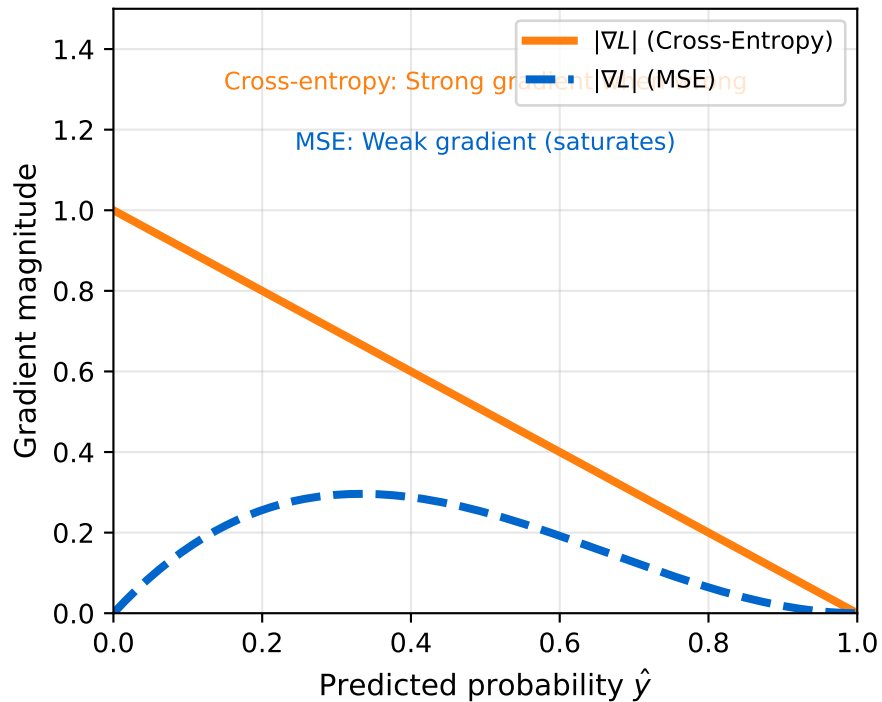
Cross-Entropy Properties

Probability-based: Measures information difference

Strong gradients: Fast learning from mistakes

No saturation: Always learns from errors

Natural for classification: Matches softmax output

Why use Cross-Entropy over MSE?

- Stronger gradients for confident wrong predictions
- Information-theoretic foundation
- Standard for classification tasks