

What We Are Trying to Do

- Learn parameters to give us the best performance
- Given data, find the best b for that data
- Define what is performance

$$b^* = \arg \min_b \frac{1}{N} \sum_i^N \ell(y_i, \sigma(z_i))$$

Diagram illustrating the Empirical Risk Minimization formula with annotations:

- b^* : optimal parameters (indicated by an upward arrow)
- $\arg \min_b$: minimization over parameters b
- $\frac{1}{N}$: average (indicated by an upward arrow)
- \sum_i^N : sum over N samples
- ℓ : loss function (indicated by a downward arrow)
- y_i : true label (indicated by an upward arrow)
- $\sigma(z_i)$: guess (indicated by a downward arrow)

Empirical Risk Minimization

- A loss function defines a penalty for poor predictions
- Want to minimize average loss

$$b^* = \arg \min_b \frac{1}{N} \sum_i^N \ell(y_i, \sigma(z_i))$$

Diagram illustrating the loss function formula with annotations:

- b^* : optimal parameters (indicated by an upward arrow)
- $\ell(y_i, \sigma(z_i))$: loss function (indicated by a downward arrow from "loss function")
- y_i : true label (indicated by an upward arrow from "true label")
- $\sigma(z_i)$: guess (indicated by a downward arrow from "guess")

Loss Function

- Define $\sigma(z_i)$ as the predicted probability

- y_i is our true label

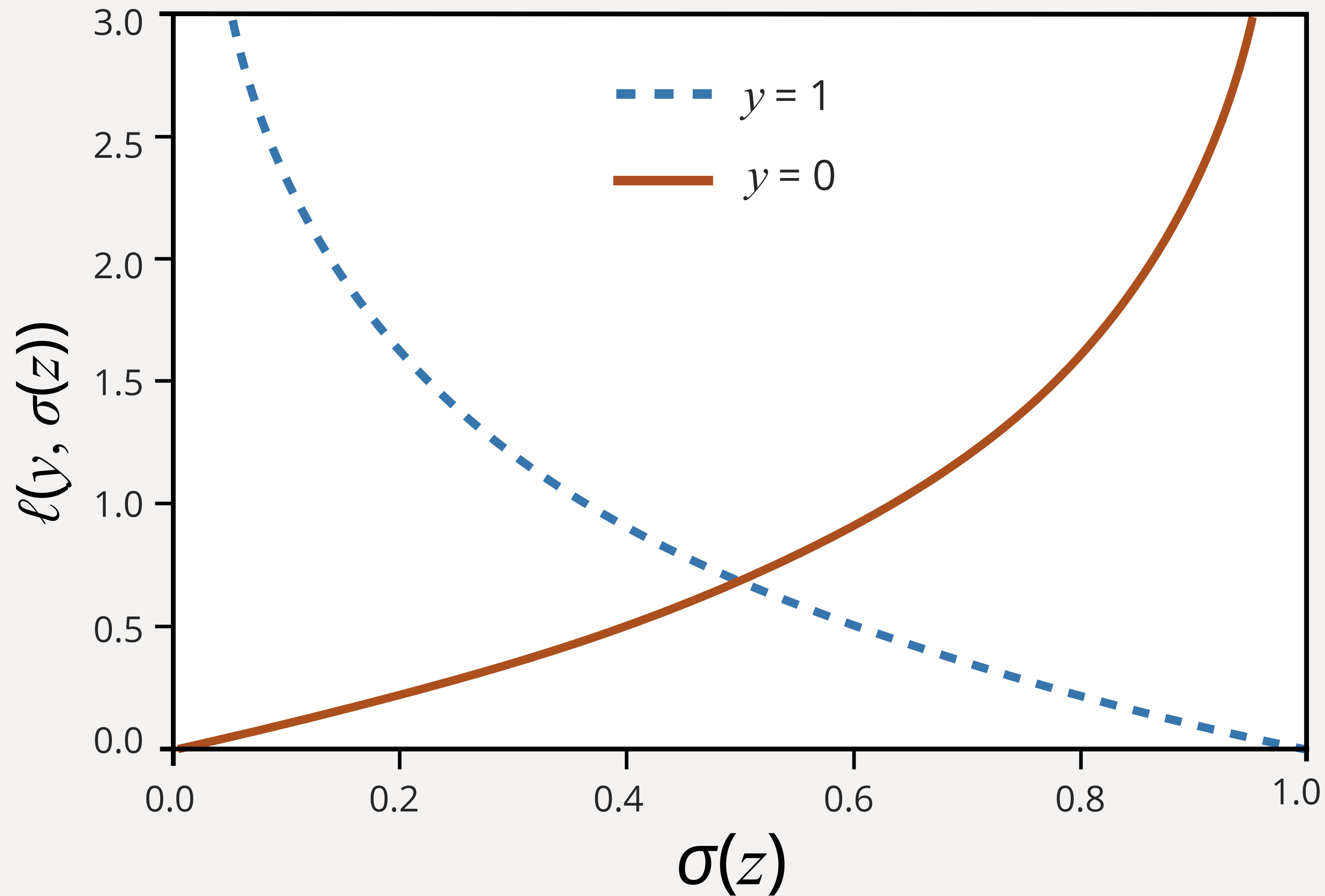
- Viewed as the negative log-likelihood:

$$\ell(y_i, \sigma(z_i)) = -\log p(y_i | \sigma(z_i))$$

- Specific mathematical form:

$$\ell(y, \sigma(z)) = -y \log \sigma(z) - (1 - y) \log(1 - \sigma(z))$$

$$\ell(y, \sigma(z)) = -y \log \sigma(z) - (1 - y) \log(1 - \sigma(z))$$



Predicting Probability of One

- When we give 100% of a 1, we pay no penalty
- If we are overconfident or wrong, we pay an increasing penalty

$$b^* = \arg \min_b \frac{1}{N} \sum_i^N \ell(y_i, \sigma(z_i))$$

Binary Classification Optimization

- Goal is to minimize the average loss

$$\ell(y, \sigma(z)) = -y \log \sigma(z) - (1 - y) \log(1 - \sigma(z))$$

Binary Classification Optimization

- Goal is to minimize the average loss
- Binary (0/1) problems can use the logistic or cross-entropy loss

Learned Model Parameters

