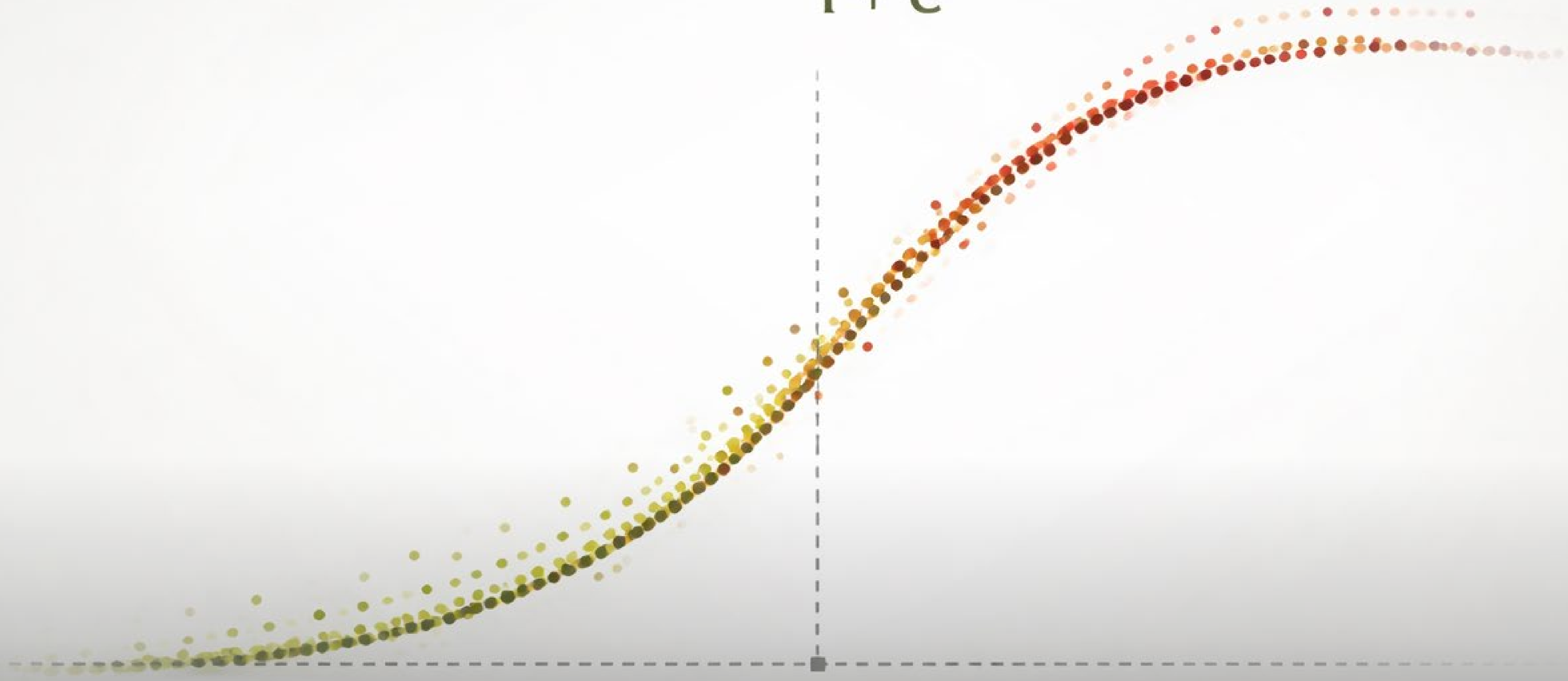
$$\sigma(z) \approx \frac{1}{1+e^{-z}}$$



INTRO TO MACHINE LEARNING

Motivation: Binary Outcomes

- Problem: Predict binary outcome $y \in \{0,1\}$ from features $x \in \mathbb{R}^p$

Why not linear regression?

✗ Unbounded predictions: $\hat{y} \in (-\infty, \infty)$, not $[0,1]$

✗ Inappropriate error model: assumes Gaussian noise, not Bernoulli

Provides a useful starting point with a linear combination of features

From Linear Scores to Probabilities

Linear predictor (score): $z = w^\top x + b$

Treat z as latent score or **log-odds**, not direct prediction

Need a function that:

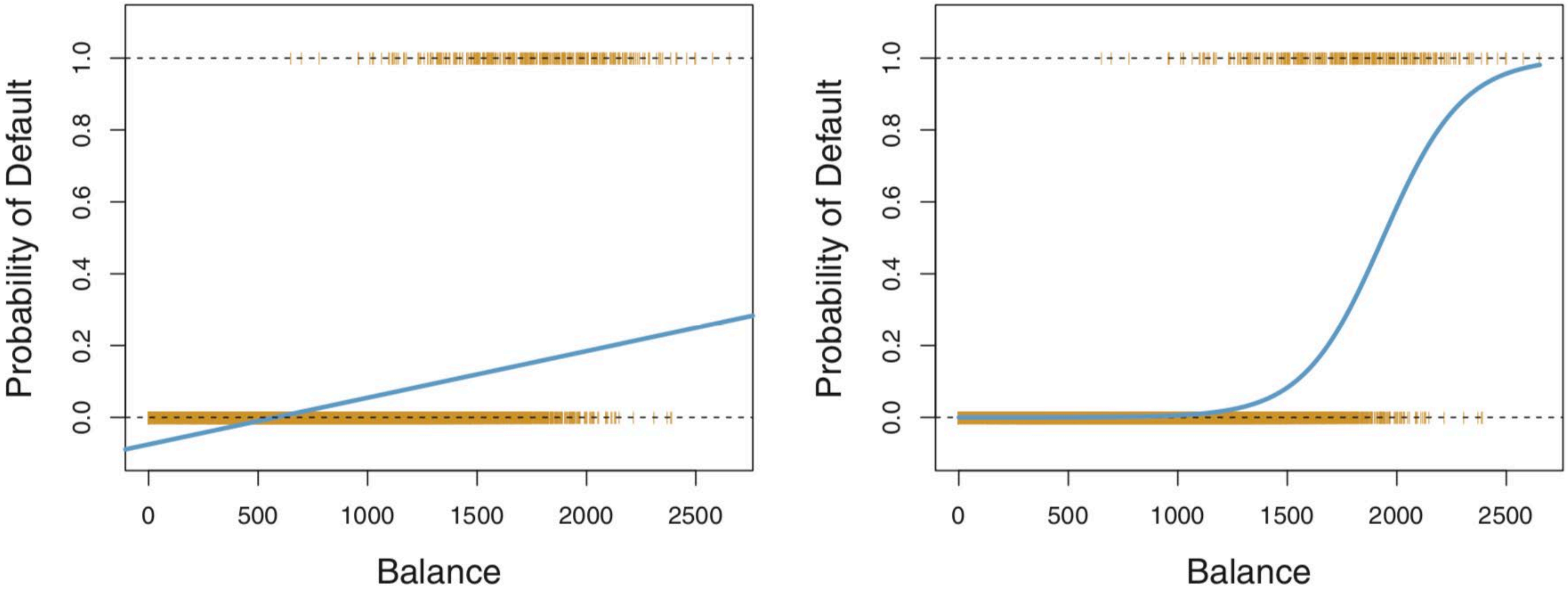
- Maps $\mathbb{R} \to (0,1)$
- Is smooth and monotonic

Solution: Logistic (sigmoid) function: $\sigma(z) = \dfrac{1}{1 + e^{-z}}$

Model: $P(y = 1 \mid x) = \sigma(w^\top x + b)$

Question: What does P(y = 0 | x) look like?

Linear v. Logistic



Picture from *Introduction to Statistic Learning* by James et. al

Statistical Interpretation: Bernoulli Model

Explicitly model the data-generating process:

$$y \mid x \sim \mathrm{Bernoulli}(p), p = \sigma(w^\top x + b)$$

This choice:

✓ Matches the binary nature of the data

✓ Provides a principled likelihood-based objective

Likelihood and Loss Functions

Likelihood for one observation: $P(y \mid x) = p^y(1-p)^{1-y}$

Log-likelihood for n samples:

$$\ell(w,b) = \sum_{i=1}^{n} [y_i \log p_i + (1-y_i)\log(1-p_i)]$$

Loss function (negative log-likelihood):

$$\mathcal{L}(w,b) = -\sum_{i=1}^{n} [y_i \log p_i + (1-y_i)\log(1-p_i)]$$

Optimization Problem

Objective: $$\min_{w,b} \ \mathcal{L}(w, b)$$

Key properties:

✓ Convex objective → unique global minimum

✗ No closed-form solution

Common methods: Gradient Descent, SGD, Newton's Method

Optimization: Gradient Descent

Gradient: $$\nabla_w \mathcal{L} = \sum_{i=1}^{n} (p_i - y_i) x_i$$

Parallels linear regression:

- Linear regression: uses residuals (ŷ - y)
- Logistic regression: uses probabilistic errors (p - y)

Interpretation:

- When $p > y$: feature gradient pushes toward 1
- When $p < y$: feature gradient pulls toward 0
- When $p \approx y$: minimal gradient signal

Classification: Decision Rule

Although logistic regression models probabilities,

classification requires a decision threshold τ:

$$\hat{y} = \begin{cases} 1 & \text{if } P(y = 1 \mid x) \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

- Standard choice: τ = 0.5
- Can be adjusted based on cost considerations

Conceptual Summary

1. Start with linear regression structure
2. Interpret as log-odds (latent score)
3. Apply sigmoid to get probabilities
4. Model with Bernoulli likelihood
5. Optimize cross-entropy loss (convex)
6. Solve via gradient-based optimization