# Linear Regression

---

## 1 Problem Formulation

**Input space:**

X⊆Rd\mathcal{X} \subseteq \mathbb{R}^dX⊆Rd

Feature vector x=(x1,x2,...,xd)x = (x_1, x_2, ..., x_d)x=(x1,x2,...,xd)

**Output space:**

Y⊆R\mathcal{Y} \subseteq \mathbb{R}Y⊆R

Continuous target variable

**Data distribution:**
Samples (xi,yi)(x_i, y_i)(xi,yi) drawn i.i.d. from unknown distribution P(X,Y)P(X,Y)P(X,Y)

**Learning Objective (Expected Risk Minimization):**

R(w)=E(X,Y)[(Y−fw(X))2]R(w) = \mathbb{E}_{(X,Y)}[(Y - f_w(X))^2]R(w)=E(X,Y)[(Y−fw(X))2]

Goal:

min wR(w)\min_w R(w)wminR(w)

---

## 2 Model Specification

**Hypothesis function:**

fw(x)=wTx+bf_w(x) = w^T x + bfw(x)=wTx+b

**Parameter space:**

w∈Rd,b∈Rw \in \mathbb{R}^d, \quad b \in \mathbb{R}w∈Rd,b∈R

**Structural assumptions:**

- Linearity in parameters

- Additive noise
- Homoscedasticity
- No multicollinearity (ideal case)

---

## 3 Loss Function

**Explicit form (Squared Error Loss):**

$$L(y, \hat{y}) = (y - \hat{y})^2$$

**Why this loss?**

- Corresponds to Gaussian noise assumption
- Maximum Likelihood Estimation under normal errors

**Convexity:**

- Convex
- Differentiable
- Global minimum exists

---

## 4 Objective Function

**Empirical Risk:**

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} (y_i - w^T x_i)^2$$

**Regularized (Ridge example):**

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda ||w||^2$$

---

## 5 Optimization Method

**Closed-form solution (Normal Equation):**

$$w = (X^TX)^{-1}X^Ty$$

**Gradient:**

$$\nabla J(w) = -\frac{2}{n} X^T (y - Xw)$$

**Convergence:**

- Convex ⇒ global optimum guaranteed

**Computational complexity:**

- Closed-form: $O(d^3)$
- Gradient Descent: $O(nd)$ per iteration

---

# 6 Statistical Interpretation

**MLE connection:**
Assume:

$$Y = w^T X + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\sigma^2)$$

Then minimizing MSE = maximizing likelihood.

**Noise model:** Gaussian

**Probabilistic meaning:**
Predicts conditional mean:

$$\mathbb{E}[Y|X]$$

---

# 7 Regularization & Generalization

**Bias–Variance Tradeoff:**

- No regularization → low bias, high variance
- With Ridge → slightly higher bias, lower variance

**Overfitting behavior:**

- High-degree polynomial features can overfit

**Capacity control:**

- L2 (Ridge)
- L1 (Lasso)

---

## 8️⃣ Theoretical Properties

- Convex optimization
- Unique global minimum (if XTXX^TXXTX invertible)
- Consistent estimator (under assumptions)

---

## 9️⃣ Computational Complexity

Training:

- O(nd2)O(nd^2)O(nd2) or O(d3)O(d^3)O(d3)

Inference:

- O(d)O(d)O(d) per prediction

Memory:

- O(nd)O(nd)O(nd)

---

## 🔟 Limitations

- Sensitive to outliers
- Assumes linear relationship
- Requires homoscedasticity
- Multicollinearity issues