Documentation-Linear Regression

1 Problem Formulation

1.1 Input Space

The input space consists of feature vectors $\mathbf{X} \in \mathbb{R}^{\mathbf{d}}$,

where each data point contains **d** numerical features.

1.2 Output Space

The output space is continuous and real-valued: $\mathbf{Y} \in \mathbb{R}$

1.3 Data Distribution

Assume data samples $(\mathbf{x_i}, \mathbf{y_i})$ are drawn independently and identically distributed (i.i.d.) from an unknown joint distribution **P(X, Y)**.

1.4 Learning Objective (Expected Risk Minimization)

The goal is to learn a function **f(x)** that minimizes the expected prediction error:

$$\mathbf{R(f) = E_{(X,Y)} [\, L(Y, f(X)) \,]}$$

Since the true distribution is unknown, we approximate this using empirical risk minimization.

2 Model Specification

2.1 Hypothesis Function

Linear Regression assumes a linear relationship between input features and output:

$$\mathbf{f(x) = w^T x + b}$$

where:

- $\mathbf{w} \in \mathbb{R}^{\mathbf{d}}$ are model coefficients
- $\mathbf{b} \in \mathbb{R}$ is the bias term

2.2 Parameter Space

$\boldsymbol{\Theta = \{ (w, b) \mid w \in \mathbb{R}^d, b \in \mathbb{R} \}}$

2.3 Structural Assumptions

- Linearity between features and target
- Additive noise model
- Independence of observations

3 Loss Function

3.1 Explicit Mathematical Form

The most common loss function is Mean Squared Error (MSE): $\boldsymbol{L(y, \hat{y}) = (y - \hat{y})^2}$

3.2 Why This Loss? (Statistical Reasoning)

- Penalizes larger errors more heavily
- Differentiable and mathematically tractable
- Corresponds to Maximum Likelihood Estimation under Gaussian noise assumption

3.3 Convexity Properties

The squared error loss is convex with respect to model parameters, ensuring a unique global minimum.

4 Objective Function

4.1 Empirical Risk Expression

$\boldsymbol{J(w, b) = (1/n) \sum_{i=1}^{n} (y_i - (w^T x_i + b))^2}$

4.2 Regularized Formulation (if used)

- Ridge (L2) Regularization: $\boldsymbol{J(w) = (1/n) \sum (y_i - w^T x_i)^2 + \lambda \|w\|^2}$
- Lasso (L1) Regularization: $\boldsymbol{J(w) = (1/n) \sum (y_i - w^T x_i)^2 + \lambda \|w\|_1}$

5 Optimization Method

5.1 Closed-form or iterative?

Closed-form solution via Normal Equation: $\mathbf{w = (X^TX)^{-1}X^Ty}$

5.2 Gradient expression

Iterative solution via Gradient Descent

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

5.3 Convergence guarantees

Since the objective is convex, gradient descent converges to a global minimum with a proper learning rate.

5.4 Computational complexity

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

6 Statistical Interpretation

6.1 MLE / MAP connection

Linear regression is equivalent to Maximum Likelihood Estimation (MLE) assuming:

$\mathbf{Y = w^TX + \epsilon}$

Where $\epsilon \sim \mathbf{N(0, \sigma^2)}$

6.2 Noise model assumption

Errors are assumed to be normally distributed with constant variance.

6.3 Probabilistic meaning of outputs

The model predicts the conditional mean:

$\mathbf{E[Y \mid X]}$

7 Regularization & Generalization

7.1 Bias–variance tradeoff

- High bias → Underfitting
- High variance → Overfitting
- Regularization balances both

7.2 Overfitting behavior

Occurs when model complexity is high or when multicollinearity exists.

7.3 Capacity control mechanism

Regularization limits coefficient magnitude to prevent overfitting.

8 Theoretical Properties

8.1 Convexity / Global optimality

The optimization problem is convex → guarantees global optimum.

8.2 Consistency (if applicable)

Under standard assumptions, the estimator is statistically consistent as sample size increases.

8.3 Stability considerations

Sensitive to outliers due to squared loss.

9 Computational Complexity

9.1 Training time complexity

- Normal Equation: **$O(d^3)$**
- Gradient Descent: **$O(nd \times \text{iterations})$**

9.2 Inference time complexity

$O(d)$

9.3 Memory complexity

O(nd)

🔟 Limitations

10.1 When it fails

- Non-linear relationships
- High multicollinearity
- Heteroscedastic data

10.2 Assumption violations

- Non-normal residuals
- Correlated features
- Dependent observations

10.3 Sensitivity issues (outliers, scaling, etc.)

- Sensitive to outliers
- Requires feature scaling for numerical stability