Neural Networks (CSC372)

Unit 3: Model Building through Regression (5 Hrs.)

Reference: Simon Haykin (3rd Edition)

By Kiran Bagale,

Department of IT, 2025

Table of Contents

01

Introduction to Regression Models

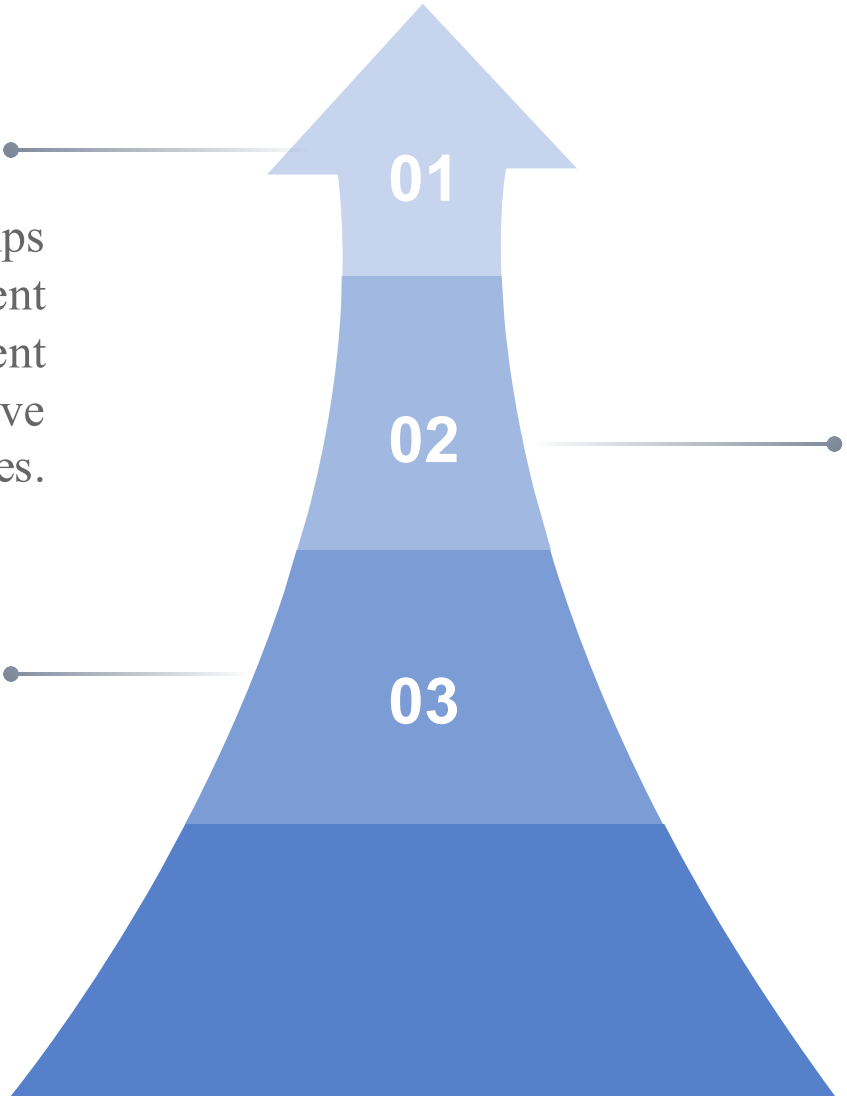
Defining Regression

Regression Explained

Regression models explore relationships between variables, predicting a dependent variable (response) based on independent variables (regressors). They use an additive error term to account for uncertainties.

Dependent vs. Independent

The response variable is the one being predicted; regressors are used to explain or predict the response's behavior.



Error Term Significance

The additive error term accounts for uncertainties in the formulated dependence between variables.

Linear vs. Nonlinear Regression

Classification of Models

There are two classes of regression models: linear and nonlinear.

Linear Model Traits

Linear regression models use a linear function, which simplifies statistical analysis.

Nonlinear Model Challenges

Nonlinear models employ a nonlinear function, increasing analytical complexity.

Bayesian Approach to Parameter Estimation

Bayesian Theory Application

01

Bayesian Theory Explained

Bayesian theory uses prior beliefs and evidence to update probabilities, allowing for continual model refinement.

02

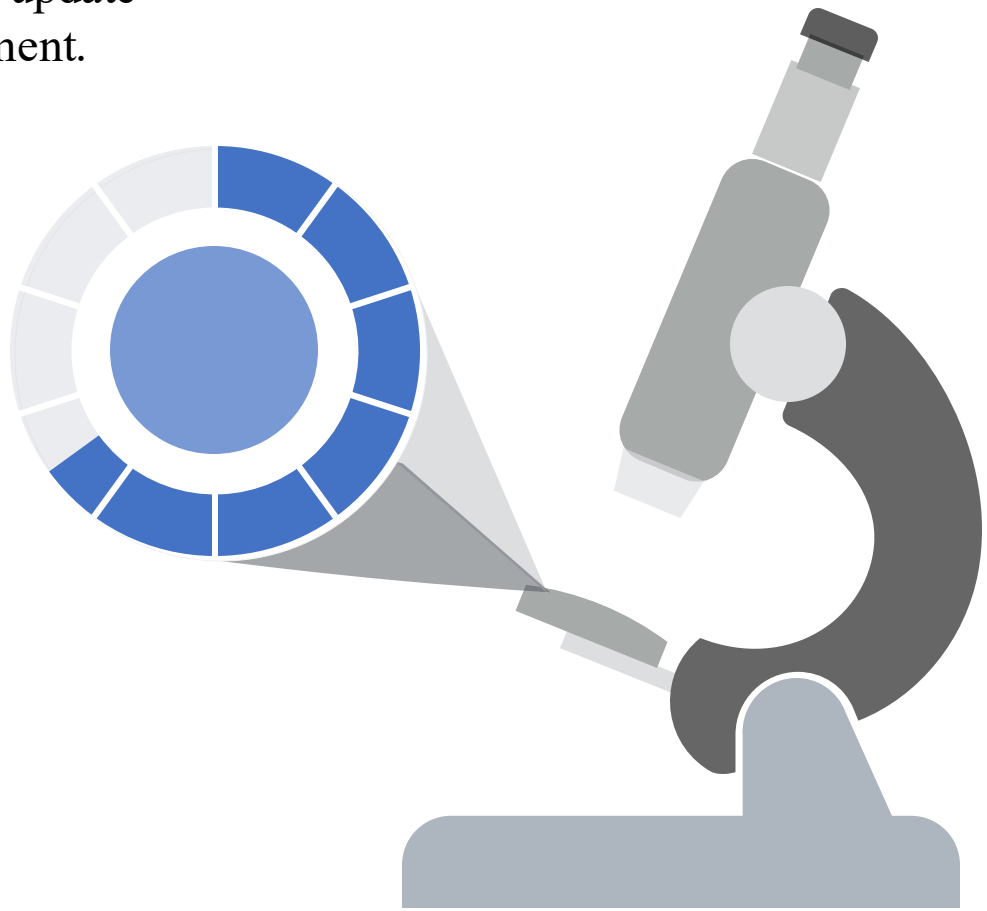
Practical Use cases

Bayesian methods provide robust parameter estimates, especially when data is limited or noisy; therefore, it is a powerful tool for real-world applications.

03

Maximizing Posterior Probability

Bayesian theory can be used to derive the ***maximum a posteriori*** **(MAP)** estimate of the parameter vector in a linear regression model.



Parameter Vector Estimation

01

Parameter Vector Role

The parameter vector is crucial for defining the specifics of the linear regression model.

02

MAP Estimation Benefits

MAP estimation combines prior knowledge with observed data to find the most probable parameter values.

03

Model Accuracy Improvement

Refined estimates improve the model's accuracy in predicting outcomes.

02 LINEAR REGRESSION MODEL: PRELIMINARY CONSIDERATIONS

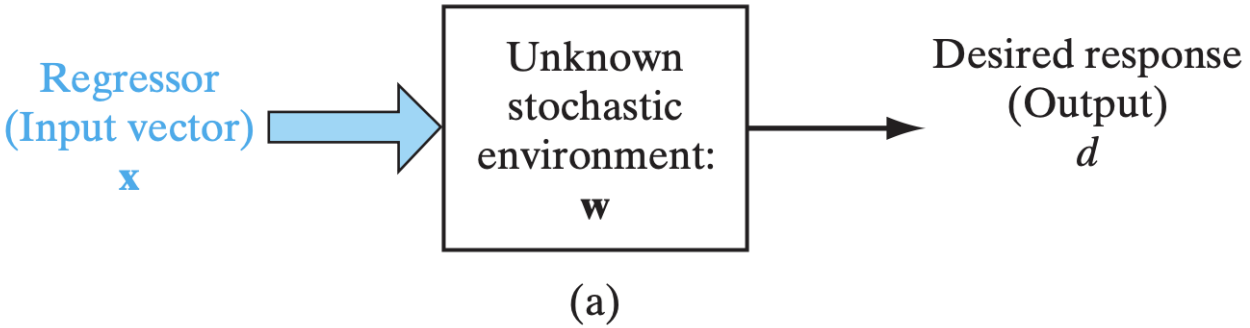


Fig: (a) Unknown stationary stochastic environment.

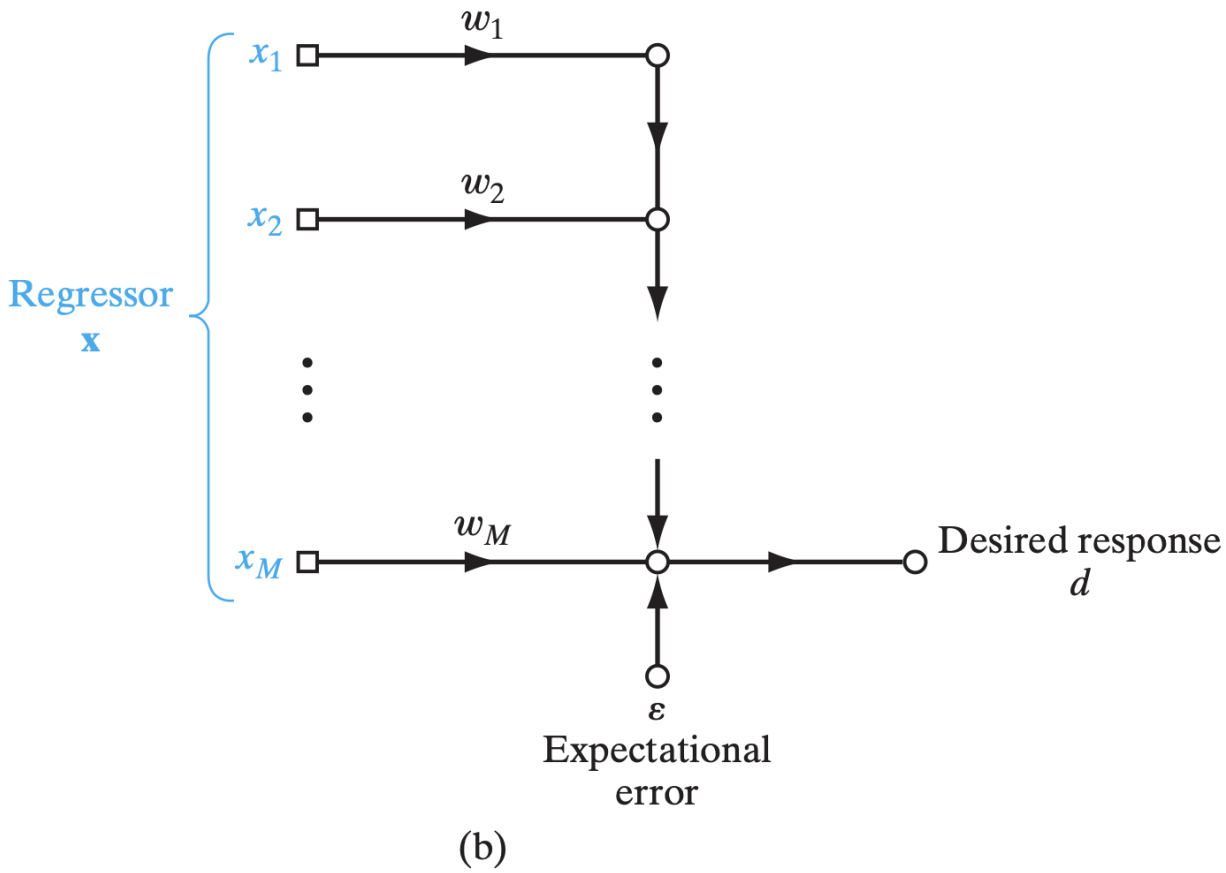


Fig: (b) Linear regression model of the environment.

- Considerng an unknown stochastic environment as focus of attention (as in Fig (a)),
  - Set of inputs, constituting the regressor,

$$\boldsymbol{x} = [x_1, x_2, \dots, xM]^T \dots\dots\dots\dots (1)$$

Output of the environment,

$$d = \sum_{j=1}^{M} w_j x_j + \varepsilon \dots\dots\dots\dots (2)$$

where, $w_1, w_2, \dots w_M$ – set of fixed unknown parameters (meaning environment is stationary)

ε - is the exceptional error of the model

Using the matrix notation, we may write the equation (2)in compact form as,

$$d = \boldsymbol{w}^T \boldsymbol{x} + \varepsilon \dots\dots\dots (3)$$

Where the regressor **x** is defined in terms of its elements in equation (1)

Parameter vector w is definde by,

$\boldsymbol{w} = [w1, w2, \dots, wM]^T \dots\dots\dots (4)$ same dimention as **x** and M is the model order.

- In a stochastic (random) environment, the input **x**, output d, and error ε are just specific examples (samples) of the random variables **X**, D, and E.
- Given this setup, the main goal is:

  *Estimate the unknown parameter vector **w** using the joint statistics of **X** (input) and D (desired output).*
- By "joint statistics," we mean:
  - The correlation matrix of X
  - The variance of D
  - The cross-correlation between X and D
- Also, it's assumed that both **X** and D have a mean of zero.

03 Maximum A Posteriori (MAP) estimation of the parameter vector

- **Bayesian paradigm** addresses **uncertainty** in parameter estimation.
- Focus: Estimating the **parameter vector w** in linear regression.
- Key assumptions:
  1. Regressor **X** is independent of **w**.
  2. Information about **W** comes from desired response D.
- We examine the **joint probability density function (PDF)** of **W** and D, conditional on **X**.
- Denoted as:

$$p_{W,D|X}(w,d|x)$$

- From probability theory:

$$p_{W,D|X}(w,d|x) = p_{W|D,X}(w|d,x).PD_{(d)} \ldots\ldots (5)$$

$$p_{W,D|X}(w,d|x) = pD_{|W,X}(d|w,x).pW(w) \ldots\ldots (6)$$

These two expressions form the foundation for deriving Bayes' theorem in the context of parameter estimation, leading us to identify four crucial density functions that characterize the complete Bayesian framework.

$$p_{W|D,X}(w,d|x) = \frac{p_{D|W,X}(d|w,x) \times pW(w)}{p_D(d)} \ldots\ldots(7)\ special\ form\ of\ Bayes\ Rule$$

here, $p_D(d) \neq 0$

Four Essential Density Functions

1. Observation Density

$$p_{D|W,X}(d|w,x)$$

Conditional probability density function referring to the "observation" of environmental response **d** due to regressor **x**, given parameter vector **w**.

2. Prior

$$p_W(w) = \pi(w)$$

Probability density function referring to information about parameter vector **w**, **prior** to any observations made on the environment.

3. Posterior Density

$$p_{W|D,X}(w|d,x) = \pi(w|d,x)$$

Conditional probability density function referring to parameter vector **w "after"** observation of the environment has been completed.

4. Evidence

$$p_{D|X}(d|x)$$

Probability density function referring to the "information" contained in response **d** for statistical analysis.

Observation Model: The conditioning response-regressor pair (x,d) embodies the response d of the environment due to regressor x.

Likelihood Function & Bayes' Rule

Likelihood Function

The observation density is commonly reformulated as the **likelihood function**:

$$\ell(w|d,x) = p_{D|W,X}(d|w,x)$$

Bayes' Rule for Parameter Estimation

Mathematical Form:

$$\pi(w|d,x) = [\ell(w|d,x) \times \pi(w)] / p_{D|X}(d|x)$$

Proportional Form (Key Result):

$$\pi(w|d,x) \propto \ell(w|d,x) \times \pi(w)$$

"The posterior density is proportional to the product of the likelihood function and the prior"

Note: The evidence $p_{D|X}(d|x)$ acts merely as a normalizing constant for parameter estimation purposes.

Maximum Likelihood vs. Maximum A Posteriori

Maximum Likelihood (ML) Estimator

$$w_{ML} = \arg max_w \ell(w|d,x)$$

Based solely on the likelihood function

Maximum A Posteriori (MAP) Estimator

$$w_{MAP} = \arg max_w \pi(w|d,x)$$

Based on the complete posterior density

Aspect	ML Estimator	MAP Estimator
Information Used	Observation model only	All conceivable information
Prior Knowledge	Ignores prior π(w)	Incorporates prior π(w)
Solution Uniqueness	May lead to non-unique solution	Enforces uniqueness and stability
Computational Demand	Less demanding	More computationally demanding

Why MAP Estimator is More Profound

- **Two Important Reasons:**
  - **Complete Information Utilization:** The Bayesian paradigm for parameter estimation, rooted in Bayes' theorem and exemplified by the MAP estimator, exploits *all conceivable information* about the parameter vector w.
  - **Stability and Uniqueness:** The ML estimator relies solely on the observation model (d,x) and may lead to non-unique solutions. The prior $\pi(w)$ enforces uniqueness and stability on the solution.
- **The Challenge**
  - The main challenge in applying MAP estimation is **how to come up with an appropriate prior**, which makes MAP more computationally demanding than ML.
- **Relationship to Bayesian Paradigm**
- **ML vs. Bayesian Framework**
  - The ML estimator lies on the **fringe** of the Bayesian paradigm, while MAP estimation represents the **complete** Bayesian approach to parameter estimation.

Computational Considerations

- **Logarithmic Transformation: Why Use Logarithms?**
  - More convenient from computational perspective
  - Logarithm is monotonically increasing function
  - Preserves the location of maximum
  - Transforms products into sums (numerical stability)
- **Modified MAP Estimator**

$$w_{MAP} = \arg max_w \, log(\pi(w|d,x))$$

where "log" denotes the natural logarithm

- **Similar Application to ML**

$$w_{ML} = \arg max_w \, log(\ell(w|d,x))$$

- **Summary**

  The Bayesian paradigm provides a comprehensive framework for parameter estimation that addresses uncertainty, incorporates prior knowledge, and ensures stable, unique solutions through the MAP estimator.

Parameter Estimation in a Gaussian Environment

- Let $\mathbf{x}_i$ and d_i denote the regressor applied to the environment and the resulting response, respectively, on the ith trial of an experiment performed on the environment.
- Let the experiment be repeated a total of N times.

$$\boldsymbol{Training\ Sample}: \tau = \{\boldsymbol{x}_i, d_i\}_{i=1}^{N}$$

$$\boldsymbol{Goal}: Estimate\ parameter\ vector\ \boldsymbol{w}$$

Three Fundamental Assumptions

Assumption 1
Statistical Independence
The N examples are **independent and identically distributed (iid)**

Assumption 2
Gaussianity
Environment is **Gaussian distributed** with zero mean error

$$p(\varepsilon_i) = (1/\sqrt{2\pi\sigma^2}) \exp(-\varepsilon_i{}^2/2\sigma^2)$$

Assumption 3
Stationarity
Parameter vector **w** is **fixed but unknown** throughout N trials

$$p(w_k) = (1/\sqrt{2\pi\sigma^2{}_w}) \exp(-w_k{}^2/2\sigma^2{}_w)$$

Mathematical Framework

Linear Regression Model

$$d_i = \boldsymbol{w}^T \boldsymbol{x}_i + \varepsilon_i, i = 1, 2, \ldots, N$$

Likelihood Function

$$l(\boldsymbol{w}|\boldsymbol{d}, \boldsymbol{x}) = \prod^{N}_{i=1} (1/\sqrt{2\pi\sigma^2}) \exp[-(d_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2/2\sigma^2]$$

Prior Distribution

$$\pi(\boldsymbol{w}) = \prod^{M}_{k=1} (1/\sqrt{2\pi{\sigma^2}_w}) \exp(-{w_k}^2/2{\sigma^2}_w)$$

Key Point:

The posterior combines both **observation data** (likelihood) and **prior knowledge** about parameters

MAP Estimation Result

Posterior Density

$$\pi(\boldsymbol{w}|\boldsymbol{d},\boldsymbol{x}) \propto \exp[-1/2\sigma^2 \sum^{N}_{i=1}(d_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 - 1/2\sigma^2{}_w ||\boldsymbol{w}||^2]$$

MAP Estimate

$$\hat{w}_{map}(N) = [\hat{R}_{xx}(N) + \lambda \boldsymbol{I}]^{-1} \hat{\boldsymbol{r}}_{\boldsymbol{dx}}(N)$$

Where:

$$\hat{\boldsymbol{R}}_{xx}(N) = (1/N)\sum^{N}_{i=1}\boldsymbol{x}_i\boldsymbol{x}_i{}^T \ (correlation\ matrix\ of\ regressors)$$

$$\hat{\boldsymbol{r}}_{\boldsymbol{dx}}(N) = (1/N)\sum^{N}_{i=1}\boldsymbol{x}_i d_i \ (cross - correlation\ vector)$$

$$\lambda = \sigma^2/\sigma^2{}_w \ (regularization\ parameter)$$

$$\boldsymbol{I} = identity\ matrix$$

Maximum Likelihood vs MAP

Aspect	Maximum Likelihood (ML)	Maximum A Posteriori (MAP)
Formula	$\hat{w}_{ml} = \hat{R}^{-1}_{xx}\hat{r}_{dx}$	$\hat{w}_{map} = [\hat{R}_{xx} + \lambda I]^{-1}\hat{r}_{dx}$
Prior Knowledge	No prior used	Incorporates prior $\pi(w)$
Bias	Unbiased estimator	Biased estimator
Stability	Can be unstable	More stable (regularized)
Large $\sigma^2 w$ Limit	Same result	Reduces to ML estimate

Key Tradeoff:
Stability vs Bias - MAP improves stability through regularization but introduces bias

Summary & Key Takeaways

What We Achieved

- Derived MAP estimator for Gaussian environment
- Combined likelihood with prior knowledge
- Obtained closed-form solution with regularization
- Analyzed bias-stability tradeoff

Practical Implications

- Regularization prevents overfitting
- Prior knowledge improves generalization
- Parameter λ controls regularization strength
- Reduces to ML when prior is uninformative

$$\boldsymbol{Final\ Result}: \hat{w}_{map} = [\hat{R}_{xx} + \boldsymbol{\lambda I}]^{-1}\hat{r}_{dx}$$

"In improving stability through regularization, we accept bias as a reasonable tradeoff"

04 Relationship Between Regularized Least-Squares Estimation and Map Estimation

A Mathematical Framework for Parameter Estimation

- We explore an alternative approach to estimating the parameter vector **w** by focusing on a cost function $e_0(w)$ that measures squared expectational errors.

Key Concept: *Instead of using maximum likelihood estimation directly, we define a cost function based on prediction errors summed over N experimental trials.*

- **The Foundation**
  - Parameter vector estimation through cost function minimization
  - Squared expectational errors as the basis for optimization
  - Connection between different estimation approaches

Basic Cost Function

The cost function $\mathbf{e_0(w)}$ is defined as the sum of squared errors over N experimental trials:

$$e_0(w) = \Sigma_{i=1}^{N} \varepsilon_i^2(w)$$

Where:

$\varepsilon_i(w) = prediction\ error\ for\ trial\ i$

$N = number\ of\ experimental\ trials$

$w = parameter\ vector\ to\ be\ estimated$

Key Points:

- Uncertainty stems from unknown vector **w**
- Based on experimental training data
- Foundation for least-squares approach

Error Term Expansion

From the regression model, we can express the error term as:

$$\boldsymbol{\varepsilon_i(w) = d_i - w^T x_i, i = 1, 2, \ldots, N}$$

Substituting this into our cost function:

$$\boldsymbol{e_0(w) = \frac{1}{2} \Sigma_{i=1}^{N} (d_i - w^T x_i)^2}$$

Important: ***This formulation relies solely on the training sample and forms the basis for ordinary least-squares estimation.***

Problems with Basic Least-Squares

Critical Issue: Minimizing $e_0(w)$ yields the ordinary least-squares estimator, which is identical to the maximum-likelihood estimator but may lack uniqueness and stability.

Why This Matters:

- **Non-uniqueness:** Multiple solutions may exist
- **Instability:** Small changes in data can cause large changes in estimates
- **Overfitting:** Model may memorize training data rather than learn patterns
- **Poor generalization:** Performance on new data may be compromised

Solution: *We need to add constraints to ensure well-posed estimation problems.*

Regularized Cost Function

To overcome stability issues, we expand the cost function by adding a regularization term:

$$\boldsymbol{e(w)} = \boldsymbol{e_0(w)} + \boldsymbol{\lambda/2}\,||\boldsymbol{w}||^2$$

$$= \tfrac{1}{2}\,\boldsymbol{\Sigma_i}^{=1N}\,(\boldsymbol{d_i} - \boldsymbol{w^T x_i})^2 + \frac{\boldsymbol{\lambda}}{\boldsymbol{2}}\,||\boldsymbol{w}||^2$$

Components:

$\boldsymbol{e_0(w)}$: *Original data fitting term*

$\boldsymbol{\lambda/2}\,||\boldsymbol{w}||^2$: *Regularization term*

$\boldsymbol{\lambda}$: *Regularization parameter*

Benefits:

- Ensures solution uniqueness
- Improves numerical stability
- Controls model complexity

Structural Regularization

Definition: *The inclusion of the squared Euclidean norm* $||w||^2$ *is called structural regularization.*

The Regularization Term: $\boldsymbol{\lambda/2\ ||w||^2}$

- **Penalizes large parameter values**
- **Encourages simpler models**
- **Prevents overfitting to training data**
- **Promotes smooth solutions**

Mathematical Intuition

The regularization term acts as a "complexity penalty" that:

- Shrinks parameter estimates toward zero
- Balances model fit against model complexity
- Provides a form of prior knowledge about parameter distribution

The Regularization Parameter λ

$\lambda \to 0$

Complete confidence in the observation model exemplified by the training sample.

Result: Approaches ordinary least-squares

$\lambda \to \infty$

No confidence in the observation model.

Result: Forces parameters toward zero

Practical Choice

- In practice, the regularization parameter λ is chosen somewhere between these two limiting cases, typically through: Cross-validation
- Information criteria (AIC, BIC)
- Domain expertise

Connection to MAP Estimation

Key Result: For a prescribed value of the regularization parameter λ, the solution obtained by minimizing the regularized cost function is identical to the MAP (Maximum A Posteriori) estimate.

The Equivalence

- **Regularized Least-Squares (RLS)** minimizes: $\boldsymbol{e(w)} = \frac{1}{2} \boldsymbol{\Sigma_{i=1}^{N}} (\boldsymbol{d_i} - \boldsymbol{w^T x_i})^2 + \boldsymbol{\lambda/2} \, \boldsymbol{||w||^2}$
- **MAP estimation** maximizes posterior probability with Gaussian priors
- **Mathematical equivalence** under specific assumptions about noise and prior distributions

This connection provides a bridge between frequentist (RLS) and Bayesian (MAP) approaches to parameter estimation.

Summary and Applications

Key Takeaways

- Regularization solves uniqueness and stability problems in least-squares estimation
- The regularization parameter λ controls the bias-variance tradeoff
- RLS solution is mathematically equivalent to MAP estimation
- Provides a principled approach to preventing overfitting

Applications

Machine Learning

- Ridge regression
- Neural network training
- Feature selection

Signal Processing

- System identification
- Adaptive filtering
- Image reconstruction

Bottom Line: Regularized least-squares provides a robust, theoretically grounded approach to parameter estimation with strong connections to Bayesian methods.

05 Computer Experiment: Pattern Classification

- Least Squares Classification vs Perceptron Algorithm
- **Experiment Overview**
- **Objective**
  - Compare least squares and perceptron algorithms for binary classification
  - Analyze performance on double-moon synthetic dataset
  - Study effect of class separation distance on classification accuracy
- **Key Parameters**
  - Dataset: Double-moon patterns
  - Separation distances: d = 1, d = -4
  - Radius: 10, Width: 6
  - Two classes: Upper moon (×) vs Lower moon (+)

Double-Moon Dataset

Dataset Characteristics

- **Upper Moon (Class 1)**: Blue × markers
- **Lower Moon (Class 2)**: Black + markers

•**Geometric Properties**:

- Radius: 10 units
- Width: 6 units
- Variable separation distance (d)

Why Double-Moon?

- Tests algorithm's ability to handle curved class boundaries
- Challenges linear classifiers with non-linear data distribution
- Real-world analogy: Many datasets have crescent-shaped clusters

06 The Minimum-Description-Length Principle,

07 Finite Sample-Size Considerations

08 The instrumental- Variables Method