Statistics

Full Course

hbar.university

Module 1 — Probability Foundations

Probability Space

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.1

Formal Definition

A **probability space** is a triple:

$$(\Omega, \mathcal{F}, \mathbb{P})$$

where:

- Ω is the sample space (set of possible outcomes),
- $\mathcal{F} \subseteq 2^{\Omega}$ is a sigma-algebra of events,
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure satisfying:

1. $\mathbb{P}(\Omega) = 1$
2. Countable additivity:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

for disjoint events A_i.

Intuition

A probability space formalizes uncertainty.

- Ω = what could happen

- $\mathcal{F}$ = what we are allowed to talk about
- $\mathbb{P}$ = how belief mass is distributed

It is the minimal mathematical container for randomness.

Structural Role

Probability spaces are the foundation of all statistical reasoning.

Every statistical model ultimately assumes:

- Data arises from a probability space.
- Parameters index probability measures.
- Inference operates over families of such spaces.

Without this structure, estimation and testing have no formal meaning.

Mathematical Development

From the axioms, we derive:

- $\mathbb{P}(\emptyset) = 0$
- Monotonicity:
$$A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$$
- Complement rule:
$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

These properties make probability a measure-theoretic object.

Worked Example

Coin Toss

Let:

$$\Omega = \{H, T\}$$

$$\mathcal{F} = 2^{\Omega}$$

$$\mathbb{P}(H) = \mathbb{P}(T) = 0.5$$

This defines a valid probability space.

Cross-Links

- **Linear Algebra:** Measure as linear functional.
- **AI:** Data-generating distribution.
- **Economics:** Uncertainty over states of the world.
- **Quantum:** Born rule defines probability measures over measurement outcomes.

Structural Compression

Invariant:
Total probability mass equals 1.

Mechanism:
Additive measure over disjoint events.

Cross-System Echo:
Conservation principles in physics; normalization in machine learning.

Exercises

1. Prove that $\mathbb{P}(\emptyset) = 0$.
2. Show that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
3. Construct a probability space for a fair six-sided die.

Events and Sigma-Algebras

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.2

Formal Definition

An **event** is a subset of the sample space Ω that we assign a probability to.

A **sigma-algebra** $\mathcal{F} \subseteq 2^{\Omega}$ is a collection of subsets of Ω such that:

1. $\Omega \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
3. If $A_1, A_2, \cdots \in \mathcal{F}$, then
$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

From (2) and (3) it follows also that $\mathcal{F}$ is closed under countable intersections.

Intuition

A sigma-algebra is "what questions you are allowed to ask."

Sometimes, the full power set 2^{Ω} is too large or too pathological, so we restrict to a consistent family of events where probability behaves well.

Think:

- Ω: reality-space
- $\mathcal{F}$: the set of measurable questions
- $\mathbb{P}$: the answer-weighting function

Structural Role

Sigma-algebras are the "language layer" of probability.

They ensure:

- probabilities can be consistently assigned,
- limits of events remain events (countable operations),
- random variables can be defined rigorously (measurability depends on $\mathcal{F}$).

This is the gateway to random variables and integration (expectation).

Mathematical Development

Consequences

If $A_1, A_2, \cdots \in \mathcal{F}$, then:

- Countable intersections:

$$\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$$

  because

$$\bigcap_{i=1}^{\infty} A_i = \left(\bigcup_{i=1}^{\infty} A_i^c \right)^c$$

  and sigma-algebras are closed under complements and countable unions.

- Finite unions and intersections are special cases of the countable rules.

Worked Examples

Example 1: Full power set (finite Ω)

If Ω is finite, then $\mathcal{F} = 2^{\Omega}$ is always a sigma-algebra.

Example 2: Trivial sigma-algebra

$$\mathcal{F} = \{\emptyset, \Omega\}$$

This corresponds to having only one meaningful question: "Did anything happen at all?"

Cross-Links

- **Statistics:** Models restrict which events are relevant (measurable) under a sampling process.
- **AI:** Feature representations define what distinctions the model can express (a "learned sigma-algebra" intuition).
- **Economics:** Information sets correspond to restricted event families (what an agent can distinguish).
- **Quantum:** Measurement choice defines the event structure (what outcomes are distinguishable).

Structural Compression

Invariant:
Closure under complement + countable union keeps probability consistent under limits.

Mechanism:
"Allowed questions" must be stable under repeated refinement.

Cross-System Echo:
In programming: closed interfaces; in physics: observable sets depend on measurement.

Exercises

1. Verify that $\{\emptyset, \Omega\}$ is a sigma-algebra.
2. If $\mathcal{F}$ is a sigma-algebra and $A, B \in \mathcal{F}$, prove that $A \setminus B \in \mathcal{F}$.
3. Let $\Omega = \{1, 2, 3, 4\}$. List all events in the sigma-algebra generated by the partition $\{\{1, 2\}, \{3, 4\}\}$.

Random Variables

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.3

Formal Definition

A **random variable** is a measurable function:

$$X : \Omega \to \mathbb{R}$$

such that for every Borel set $B \subseteq \mathbb{R}$,

$$X^{-1}(B) \in \mathcal{F}.$$

This measurability condition ensures that probabilities of events of the form $\{X \in B\}$ are well-defined.

Intuition

A random variable is not "a random number."

It is a function that translates outcomes into numbers.

$$\text{Outcome} \longrightarrow \text{Number}$$

It allows us to measure, summarize, and analyze uncertainty quantitatively.

Structural Role

Random variables connect abstract probability spaces to numerical structure.

They enable:

- Expectation
- Variance
- Distribution functions
- Statistical modeling

They are the bridge from pure measure theory to statistics.

Mathematical Development

Induced Distribution

The random variable X induces a probability measure on $\mathbb{R}$:

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B)$$

This is called the **pushforward measure** or distribution of X.

Discrete Example

If:

$$\Omega = \{H, T\}$$

Define:

$$X(H) = 1, \quad X(T) = 0.$$

Then:

$$\mathbb{P}(X = 1) = 0.5, \quad \mathbb{P}(X = 0) = 0.5.$$

Cross-Links

- **Linear Algebra:** Random vectors are vector-valued random variables.
- **AI:** Data features are realizations of random variables.
- **Economics:** Utility functions act on random states.
- **Quantum:** Observables are operator-valued analogues of random variables.

Structural Compression

Invariant:
Random variables preserve measurable structure.

Mechanism:
Function pushes probability measure forward.

Cross-System Echo:
In ML, feature maps transform input space into numeric representation.

Exercises

1. Prove that if X is measurable, then $aX + b$ is measurable.
2. Construct a random variable on a die roll representing parity.
3. Show how a CDF can be defined from the induced distribution.

Distribution Functions

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.4

Formal Definition

The **cumulative distribution function (CDF)** of a random variable X is the function:

$$F_X(x) = \mathbb{P}(X \leq x).$$

The CDF satisfies:

1. F_X is non-decreasing.
2. $\lim_{x \to -\infty} F_X(x) = 0$.
3. $\lim_{x \to \infty} F_X(x) = 1$.
4. F_X is right-continuous.

Intuition

The CDF accumulates probability from left to right.

At any real number x, it tells us:

> "How much probability mass lies at or below this value?"

It compresses the full distribution of X into a single monotonic function.

Structural Role

The distribution function:

- Fully characterizes the law of a random variable.
- Enables computation of probabilities:

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

- Bridges discrete and continuous cases into a unified framework.

All density and mass functions are derived from the CDF.

Mathematical Development

Discrete Case

If X takes values x_i with probabilities p_i, then:

$$F_X(x) = \sum_{x_i \leq x} p_i.$$

The CDF has jump discontinuities at mass points.

Continuous Case

If X has density $f_X(x)$, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

The density is the derivative of the CDF:

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Worked Example

Fair Die

Let $X \in \{1, 2, 3, 4, 5, 6\}$ with equal probability.

Then:

$$F_X(3) = \frac{3}{6} = 0.5.$$

The CDF increases in steps of $\frac{1}{6}$.

Cross-Links

- **Linear Algebra:** CDFs generalize to multivariate distributions via joint probability measures.
- **AI:** Model outputs often represent CDF or probability mass functions.
- **Economics:** Risk analysis depends on distribution tails.
- **Quantum:** Measurement outcome probabilities form cumulative distributions over eigenvalues.

Structural Compression

Invariant:
CDF is monotone and bounded between 0 and 1.

Mechanism:
Accumulation of probability mass via integration or summation.

Cross-System Echo:
In calculus, integration accumulates area; in learning, loss aggregates over data.

Exercises

1. Prove that every CDF is right-continuous.
2. Show that a CDF uniquely determines a probability measure on $\mathbb{R}$.
3. Sketch the CDF of a Bernoulli(p) random variable.

Expectation

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.5

Formal Definition

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The **expectation** of X is defined as:

Discrete Case

$$\mathbb{E}[X] = \sum_{x} x\, \mathbb{P}(X = x),$$

provided the sum converges absolutely.

Continuous Case

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

provided the integral is finite.

General (Measure-Theoretic Form)

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\, d\mathbb{P}(\omega).$$

Intuition

Expectation is not "what usually happens."

It is the **probability-weighted average** of all possible outcomes.

It is the center of mass of the distribution.

If probability is mass, then expectation is balance.

Structural Role

Expectation is the fundamental linear operator of probability theory.

It enables:

- Definition of variance
- Risk evaluation in estimation

- Decision theory
- Optimization of statistical procedures
- Law of Large Numbers

Almost all statistical reasoning is built on properties of expectation.

Mathematical Development

Linearity

For random variables X, Y and constants a, b:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Linearity holds without independence.

Indicator Trick

For an event A, define:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Then:

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

This connects expectation directly to probability.

Worked Example

Fair Die

$$\mathbb{E}[X] = \frac{1+2+3+4+5+6}{6} = 3.5.$$

Note: 3.5 is not an achievable outcome. Expectation need not lie in the support of X.

Cross-Links

- **Linear Algebra:** Expectation is a linear functional.
- **AI:** Loss functions are expectations over data distributions.
- **Economics:** Expected utility theory.
- **Quantum:** Expected value of an observable corresponds to $\langle\psi|A|\psi\rangle$.

Structural Compression

Invariant:
Expectation is linear.

Mechanism:
Integration against probability measure.

Cross-System Echo:
In physics, expectation resembles averaging over states; in ML, empirical averages approximate expectations.

Exercises

1. Prove linearity of expectation.
2. Show that $\mathbb{E}[c] = c$ for constant c.
3. Let $X \sim \text{Bernoulli}(p)$. Compute $\mathbb{E}[X]$.
4. Show that expectation of an indicator recovers probability.

Variance and Covariance

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.6

Formal Definition

Let X be a random variable with finite expectation.

Variance

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

Equivalently,

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Covariance

For random variables X, Y:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big].$$

Intuition

Expectation measures center.

Variance measures spread.

Covariance measures joint variation.

If expectation is “balance,”
variance is “dispersion energy.”

Covariance tells us whether two variables move together.

Structural Role

Variance and covariance introduce geometry into probability.

They enable:

- Measurement of uncertainty magnitude
- Correlation analysis
- Construction of covariance matrices

- Linear regression
- Multivariate statistics

They transform probability space into an inner-product structure.

Mathematical Development

Key Properties

1. Non-negativity:
$$\mathrm{Var}(X) \geq 0.$$
2. Scaling:
$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X).$$
3. Variance of Sum:
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

If X and Y are independent:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Correlation

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Correlation is dimensionless and bounded:

$$-1 \leq \rho(X,Y) \leq 1.$$

Worked Example

Bernoulli(p)

If $X \sim \mathrm{Bernoulli}(p)$,

$$\mathbb{E}[X] = p,$$

$$\mathrm{Var}(X) = p(1-p).$$

Maximum variance occurs at $p = 0.5$.

Cross-Links

- **Linear Algebra:** Covariance defines a positive semidefinite matrix.
- **AI:** Covariance matrices appear in PCA and Gaussian models.
- **Economics:** Portfolio variance measures risk.
- **Quantum:** Variance of observables relates to uncertainty principles.

Structural Compression

Invariant:
Variance is always non-negative.

Mechanism:
Quadratic deviation around mean.

Cross-System Echo:
In physics, variance mirrors energy around equilibrium; in ML, variance reflects model sensitivity.

Exercises

1. Prove that variance is non-negative.
2. Show the formula $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
3. Compute variance of a fair die.
4. Prove that if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

Conditional Probability and Conditional Expectation

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.7

Formal Definition

Conditional Probability

For events $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Conditional Expectation (Discrete Case)

If Y is a discrete random variable,

$$\mathbb{E}[X \mid Y = y] = \sum_x x \, \mathbb{P}(X = x \mid Y = y).$$

General Definition (Measure-Theoretic Form)

The **conditional expectation** of X given a sigma-algebra $\mathcal{G} \subseteq \mathcal{F}$ is a random variable $\mathbb{E}[X \mid \mathcal{G}]$ such that:

1. $\mathbb{E}[X \mid \mathcal{G}]$ is $\mathcal{G}$-measurable.
2. For all $G \in \mathcal{G}$,

$$\int_G \mathbb{E}[X \mid \mathcal{G}] \, d\mathbb{P} = \int_G X \, d\mathbb{P}.$$

Intuition

Conditional probability updates belief given information.

Conditional expectation updates averages given information.

If expectation is balance, conditional expectation is balance **after learning something**.

It represents the best prediction of X given available information.

Structural Role

Conditional expectation is the backbone of:

- Bayesian inference
- Regression
- Martingale theory
- Filtering
- Sequential decision-making

It formalizes how information changes predictions.

Mathematical Development

Law of Total Probability

If $\{B_i\}$ partitions Ω,

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Law of Total Expectation

$$\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big].$$

This is sometimes called the **tower property**.

Best Prediction Property

Conditional expectation minimizes mean squared error:

$$\mathbb{E}[X \mid Y] = \arg\min_g \mathbb{E}\big[(X - g(Y))^2\big].$$

This connects probability directly to regression.

Worked Example

Let $X \sim \text{Bernoulli}(p)$ and suppose we observe an event B.

Then:

$$\mathbb{P}(X = 1 \mid B) = \frac{\mathbb{P}(B \mid X = 1)\mathbb{P}(X = 1)}{\mathbb{P}(B)}.$$

This is Bayes' rule.

Cross-Links

- **Linear Algebra:** Conditional expectation is an orthogonal projection in L^2.
- **AI:** Bayesian updating and predictive modeling.
- **Economics:** Expectations conditional on information sets.
- **Quantum:** State update after measurement.

Structural Compression

Invariant:
Conditioning preserves total expectation (tower property).

Mechanism:
Projection onto information sigma-algebra.

Cross-System Echo:
In ML, regression projects outcomes onto feature space.

Exercises

1. Prove the law of total probability.
2. Prove the tower property.
3. Show that conditional expectation minimizes mean squared error.
4. Derive Bayes' rule from the definition of conditional probability.

Law of Large Numbers

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.8

Formal Statement

Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with finite expectation $\mu = \mathbb{E}[X_i]$.

Define the sample mean:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Weak Law of Large Numbers (WLLN)

$$\bar{X}_n \xrightarrow{P} \mu.$$

That is, the sample mean converges in probability to the true expectation.

Strong Law of Large Numbers (SLLN)

$$\bar{X}_n \xrightarrow{a.s.} \mu.$$

The convergence occurs almost surely.

Intuition

Averages stabilize.

As we collect more data, random fluctuations cancel out.

The sample mean becomes a reliable estimator of the true mean.

Probability turns into empirical regularity.

Structural Role

The Law of Large Numbers justifies:

- Statistical estimation
- Empirical averages
- Risk minimization
- Machine learning training procedures

Without LLN, statistics would not work.

Mathematical Insight

Variance of the sample mean:

$$\mathrm{Var}(\bar{X}_n) = \frac{\mathrm{Var}(X)}{n}.$$

As $n \to \infty$, variance shrinks to zero.

Concentration drives convergence.

Worked Example

Flip a fair coin.

Let $X_i = 1$ for heads, 0 for tails.

Then:

$$\bar{X}_n \to 0.5.$$

The empirical frequency converges to true probability.

Cross-Links

- **AI:** Empirical risk converges to expected risk.
- **Economics:** Large populations stabilize aggregate behavior.
- **Quantum:** Repeated measurements converge to expectation values.
- **Linear Algebra:** Averaging as projection onto constant vector.

Structural Compression

Invariant:
Sample mean approaches expectation.

Mechanism:
Variance shrinks as $1/n$.

Cross-System Echo:
In ML, more data reduces estimation noise.

Exercises

1. Prove convergence in probability using Chebyshev's inequality.
2. Compute variance of $\bar{X}_n$.
3. Explain difference between weak and strong LLN.

Central Limit Theorem

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Probability Foundations

Node:
1.9

Formal Statement

Let $X_1, X_2, \ldots$ be i.i.d. random variables with:

$$\mathbb{E}[X_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2 < \infty.$$

Define:

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Then:

$$Z_n \xrightarrow{d} \mathcal{N}(0, 1).$$

Intuition

No matter the original distribution (under mild conditions), normalized averages become approximately normal.

The Gaussian distribution emerges universally.

Noise aggregates into structure.

Structural Role

The CLT enables:

- Approximate confidence intervals
- Hypothesis testing
- Asymptotic inference
- Statistical modeling at scale

It connects finite samples to Gaussian structure.

Mathematical Insight

The CLT explains why:

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

for large n.

This approximation powers classical statistics.

Worked Example

Let $X_i \sim \text{Bernoulli}(p)$.

Then:

$$\sqrt{n}(\bar{X}_n - p)$$

converges to:

$$\mathcal{N}(0, p(1-p)).$$

Cross-Links

- **AI:** SGD noise approximates Gaussian under large batch regimes.
- **Economics:** Aggregate shocks often modeled as Gaussian.
- **Quantum:** Gaussian approximations in large systems.
- **Linear Algebra:** Multivariate CLT yields multivariate normal with covariance matrix.

Structural Compression

Invariant:
Normal distribution emerges under aggregation.

Mechanism:
Independent fluctuations accumulate additively.

Cross-System Echo:
Universal Gaussian noise in physics and learning systems.

Exercises

1. State Lindeberg condition for CLT.
2. Show how CLT justifies normal approximation to binomial.
3. Explain difference between LLN and CLT.

Module 2 — Statistical Models

Statistical Models

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.1

Formal Definition

A **statistical model** is a family of probability distributions:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

- Θ is the parameter space.
- Each θ indexes a probability distribution.
- Data is assumed to arise from one unknown member of this family.

Intuition

Probability theory assumes the distribution is known.

Statistics assumes:

> The distribution belongs to a structured family, but we do not know which one.

The parameter θ encodes the unknown structure of reality.

Inference means learning θ from data.

Structural Role

Statistical models:

- Bridge probability and data.
- Introduce parameters.
- Define the object of estimation.
- Enable likelihood construction.

Without models, data cannot inform unknown structure.

Mathematical Development

Parametric Model

Finite-dimensional parameter space:

$$\theta \in \mathbb{R}^k.$$

Example: Normal model

$$X_i \sim \mathcal{N}(\mu, \sigma^2).$$

$$\theta = (\mu, \sigma^2).$$

Nonparametric Model

Infinite-dimensional parameter space.

Example: All continuous distributions on $\mathbb{R}$.

Identifiability

A model is **identifiable** if:

$$\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2.$$

Otherwise, parameters cannot be uniquely recovered.

Worked Example

Bernoulli Model

$$X_i \sim \text{Bernoulli}(p), \quad p \in [0, 1].$$

Parameter p fully determines the distribution.

Data informs p.

Cross-Links

- **Linear Algebra:** Parameter vectors as elements of $\mathbb{R}^k$.
- **AI:** Model class in machine learning.
- **Economics:** Structural models of agents.
- **Quantum:** Quantum states parameterize measurement distributions.

Structural Compression

Invariant:
Model = structured family of distributions.

Mechanism:
Unknown parameter indexes probability measure.

Cross-System Echo:
In ML, hypothesis class defines learning space.

Exercises

1. Define a Poisson statistical model.
2. Give an example of a non-identifiable model.
3. Explain difference between parametric and nonparametric models.

Sampling and the Data-Generating Process

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.2

Formal Definition

Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model.

A dataset:

$$X_1, \ldots, X_n$$

is assumed to be generated according to:

$$X_i \sim \mathbb{P}_\theta,$$

often with the additional assumption that the observations are **independent and identically distributed (i.i.d.)**.

Intuition

A statistical model alone is not enough.

We also assume a mechanism:

> Reality produces data from one unknown distribution in the model family.

This mechanism is called the **data-generating process (DGP).**

Statistics attempts to reverse-engineer the DGP.

Structural Role

Sampling assumptions determine:

- Joint distribution of the data
- Likelihood structure
- Variance scaling
- Asymptotic behavior

For i.i.d. sampling:

$$\mathbb{P}_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i).$$

This factorization is the backbone of likelihood theory.

Mathematical Development

Joint Density (Continuous Case)

If density exists:

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Non-i.i.d. Case

More generally:

$$f_\theta(x_1, \ldots, x_n) \neq \prod_{i=1}^{n} f_\theta(x_i).$$

Examples include:

- Time series
- Markov chains
- Dependent sampling

Worked Example

Normal Model

If:

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

then joint density:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

Cross-Links

- **AI:** Training data assumed i.i.d. in ERM.
- **Economics:** Structural models assume sampling from populations.
- **Quantum:** Repeated measurement outcomes under identical preparation.
- **Linear Algebra:** Vector of observations as random vector.

Structural Compression

Invariant:
Joint distribution determined by sampling structure.

Mechanism:
Independence factorizes probability.

Cross-System Echo:
In ML, mini-batch sampling approximates full distribution.

Exercises

1. Derive joint density for Bernoulli model.
2. Explain how dependent sampling changes likelihood.
3. Give example where i.i.d. assumption fails.

Likelihood Function

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.3

Formal Definition

Let:

$$X_1, \ldots, X_n$$

be observed data from a statistical model:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}.$$

The **likelihood function** is defined as:

$$L(\theta; x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n),$$

viewed as a function of θ, with the data fixed.

Under the i.i.d. assumption:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Intuition

Probability asks:

> Given θ, how likely is the data?

Likelihood asks:

> Given the data, how plausible is θ?

The same formula. Different perspective.

Likelihood reverses the direction of reasoning.

Structural Role

Likelihood is the engine of:

- Maximum likelihood estimation (MLE)
- Likelihood ratio tests
- Asymptotic theory
- Information measures
- Bayesian updating (via Bayes' rule)

It is the bridge between model and inference.

Mathematical Development

Key Distinction

Probability:

$$f_\theta(x)$$

is a function of data x.

Likelihood:

$$L(\theta; x)$$

is a function of parameter θ.

The functional role changes.

Example: Bernoulli Model

If $X_i \sim \text{Bernoulli}(p)$,

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}.$$

This simplifies to:

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

Likelihood depends only on the number of successes.

Worked Example

Let:

$$\sum x_i = k.$$

Then:

$$L(p) = p^k(1-p)^{n-k}.$$

The parameter p that maximizes this will be the MLE.

Cross-Links

- **Linear Algebra:** Log-likelihood often leads to quadratic forms.
- **AI:** Empirical risk minimization equals negative log-likelihood minimization.
- **Economics:** Structural estimation uses likelihood maximization.
- **Quantum:** Likelihood used in quantum state tomography.

Structural Compression

Invariant:
Likelihood is the joint density viewed as function of parameter.

Mechanism:
Fix data, vary parameter.

Cross-System Echo:
In ML, model parameters are optimized against fixed dataset.

Exercises

1. Derive likelihood for Poisson model.
2. Show that likelihood depends only on sufficient statistics in Bernoulli case.
3. Explain difference between probability and likelihood.

Log-Likelihood and the Score Function

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.4

Formal Definition

Given the likelihood function:

$$L(\theta; x_1, \ldots, x_n),$$

the **log-likelihood** is defined as:

$$\ell(\theta) = \log L(\theta).$$

Under i.i.d. sampling:

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

Score Function

The **score function** is the gradient of the log-likelihood:

$$S(\theta) = \frac{\partial}{\partial \theta} \ell(\theta).$$

For vector parameters:

$$S(\theta) = \nabla_\theta \ell(\theta).$$

Intuition

Taking logs turns products into sums.

Optimization becomes tractable.

The score measures:

> How sensitive the likelihood is to changes in the parameter.

If the score is zero, we are at a stationary point.

Structural Role

Log-likelihood enables:

- Maximum Likelihood Estimation (MLE)
- Derivation of estimating equations
- Asymptotic normality
- Fisher Information
- Numerical optimization

The score is the foundation of modern inference.

Mathematical Development

Additivity

Because of independence:

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta).$$

The score becomes:

$$S(\theta) = \sum_{i=1}^{n} S_i(\theta).$$

This additive structure drives asymptotic theory.

Example: Bernoulli Model

If:

$$L(p) = p^k (1-p)^{n-k},$$

then:

$$\ell(p) = k \log p + (n-k) \log(1-p).$$

Score:

$$\frac{d}{dp}\ell(p) = \frac{k}{p} - \frac{n-k}{1-p}.$$

Setting score to zero gives:

$$\hat{p} = \frac{k}{n}.$$

Cross-Links

- **Linear Algebra:** Score is gradient; Hessian gives curvature.
- **AI:** Training neural networks minimizes negative log-likelihood.
- **Economics:** Structural models solved via likelihood maximization.
- **Quantum:** Log-likelihood used in state reconstruction algorithms.

Structural Compression

Invariant:
Log preserves maxima of likelihood.

Mechanism:
Product $\rightarrow$ sum $\rightarrow$ gradient optimization.

Cross-System Echo:
In ML, loss gradients guide parameter updates.

Exercises

1. Derive log-likelihood for Poisson model.
2. Compute score for Normal mean with known variance.
3. Show that maximizing likelihood equals maximizing log-likelihood.
4. Explain why logs simplify asymptotic analysis.

Fisher Information

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.5

Formal Definition

Let $\ell(\theta)$ be the log-likelihood for a single observation.

The **Fisher Information** is defined as:

$$\mathcal{I}(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right].$$

Equivalently, under regularity conditions:

$$\mathcal{I}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right].$$

For n i.i.d. observations:

$$\mathcal{I}_n(\theta) = n\mathcal{I}(\theta).$$

Intuition

Fisher Information measures how sharply peaked the likelihood is.

- Flat likelihood $\rightarrow$ low information
- Sharp likelihood $\rightarrow$ high information

It quantifies how much the data tells us about the parameter.

It is curvature of the log-likelihood in expectation.

Structural Role

Fisher Information underlies:

- Cramér–Rao lower bound
- Asymptotic normality of MLE
- Efficiency
- Information geometry
- Hypothesis testing

It introduces metric structure on parameter space.

Mathematical Development

Score Variance Form

Since:

$$S(\theta) = \frac{\partial}{\partial \theta} \log f_\theta(X),$$

and

$$\mathbb{E}[S(\theta)] = 0,$$

we have:

$$\mathcal{I}(\theta) = \mathrm{Var}(S(\theta)).$$

Thus, information equals variance of the score.

Example: Bernoulli(p)

$$\log f_p(x) = x \log p + (1 - x) \log(1 - p).$$

Score:

$$S(p) = \frac{x}{p} - \frac{1 - x}{1 - p}.$$

Then:

$$\mathcal{I}(p) = \frac{1}{p(1 - p)}.$$

Information is highest near boundaries.

Cross-Links

- **Linear Algebra:** Information matrix is positive semidefinite.
- **AI:** Natural gradient uses Fisher Information metric.
- **Economics:** Efficiency bounds in estimation.
- **Quantum:** Quantum Fisher Information generalizes classical case.

Structural Compression

Invariant:
Information ≥ 0.

Mechanism:
Curvature of expected log-likelihood.

Cross-System Echo:
In optimization, curvature determines step size and stability.

Exercises

1. Prove equivalence of the two definitions of Fisher Information.
2. Compute Fisher Information for Normal mean with known variance.
3. Explain why information scales linearly with sample size.
4. Show that score has zero expectation under regularity conditions.

Exponential Families

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Statistical Models and Likelihood

Node:
2.6

Formal Definition

A statistical model belongs to the **exponential family** if its density can be written as:

$$f_\theta(x) = h(x) \exp\left(\eta(\theta)^\top T(x) - A(\theta)\right),$$

where:

- $T(x)$ is the sufficient statistic,
- $\eta(\theta)$ is the natural parameter,
- $A(\theta)$ is the log-partition function,
- $h(x)$ is the base measure.

Intuition

Exponential families separate:

- Data structure $T(x)$
- Parameter structure $\eta(\theta)$

All parameter dependence flows through a low-dimensional summary $T(x)$.

They are maximally structured statistical models.

Structural Role

Exponential families:

- Guarantee existence of sufficient statistics
- Simplify likelihood analysis
- Enable conjugate priors in Bayesian inference
- Underlie generalized linear models
- Connect to information geometry

They are the algebraic backbone of classical statistics.

Mathematical Development

Log-Likelihood Form

For i.i.d. data:

$$\ell(\theta) = \eta(\theta)^\top \sum_{i=1}^{n} T(x_i) - nA(\theta) + \sum_{i=1}^{n} \log h(x_i).$$

Parameter dependence appears only through:

$$\sum_{i=1}^{n} T(x_i).$$

This implies sufficiency.

Mean and Variance Structure

The log-partition function satisfies:

$$\nabla_\theta A(\theta) = \mathbb{E}_\theta[T(X)].$$

Second derivative:

$$\nabla^2_\theta A(\theta) = \mathrm{Var}_\theta(T(X)).$$

Thus, curvature encodes variance.

Examples

Bernoulli

$$f_p(x) = \exp\left(x \log \frac{p}{1-p} + \log(1-p)\right).$$

Natural parameter:

$$\eta = \log \frac{p}{1-p}.$$

Normal (Known Variance)

$$f_\mu(x) = \exp\left(\frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) h(x).$$

Cross-Links

- **Linear Algebra:** Natural parameter space as vector space.
- **AI:** Softmax, logistic regression are exponential family models.
- **Economics:** Logit models in discrete choice.
- **Quantum:** Gibbs states resemble exponential family structure.

Structural Compression

Invariant:
Parameter enters linearly in sufficient statistic.

Mechanism:
Log-partition function controls normalization and variance.

Cross-System Echo:
In physics, Boltzmann distributions are exponential family models.

Exercises

1. Show Poisson distribution is exponential family.
2. Derive sufficient statistic for Normal mean.
3. Prove that expectation of sufficient statistic equals gradient of $A(\theta)$.
4. Explain why exponential families are closed under i.i.d. sampling.

Module 3 — Point Estimation

Estimators, Risk, and Mean Squared Error

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.1

Formal Definition

Let $X_1, \ldots, X_n$ be data generated under a model $\{\mathbb{P}_\theta : \theta \in \Theta\}$.

An **estimator** of a parameter θ (or a function $g(\theta)$) is a statistic:

$$\hat{\theta} = T(X_1, \ldots, X_n).$$

A **loss function** quantifies error of an estimate a when the true parameter is θ:

$$L(\theta, a) \geq 0.$$

A **decision rule** (estimator is a special case) is a map $\delta(X_1, \ldots, X_n) \in \mathcal{A}$.

The **risk function** is the expected loss:

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X_1, \ldots, X_n))]\,.$$

Intuition

An estimator is a rule that turns data into a guess.

Risk is "how bad this rule is on average" if the world is θ.

Statistics is not just computing $\hat{\theta}$; it is designing rules with controlled risk.

Structural Role

Risk is the evaluation metric that makes estimation a well-posed optimization problem.

Everything that follows—bias/variance, optimal estimators, Cramér–Rao bounds—talks about controlling or lower-bounding risk.

This node installs the objective function for inference.

Mathematical Development

Squared Error Loss and MSE

For estimating a scalar $g(\theta)$, a canonical loss is squared error:

$$L(\theta, a) = (a - g(\theta))^2.$$

Then the risk is the **mean squared error (MSE)**:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta\left[(\hat{\theta} - g(\theta))^2\right].$$

Bias–Variance Decomposition

Let $b_\theta = \mathbb{E}_\theta[\hat{\theta}] - g(\theta)$ be the **bias**. Then:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + b_\theta^2.$$

This decomposition is the first fundamental tradeoff in estimation.

Worked Example

Estimating a Bernoulli Mean

Let $X_i \sim \mathrm{Bernoulli}(p)$, $g(\theta) = p$, and estimator $\hat{p} = \bar{X}_n$.

Then:

$$\mathbb{E}[\hat{p}] = p \quad \text{(unbiased)},$$

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n},$$

so:

$$\mathrm{MSE}(\hat{p}) = \frac{p(1-p)}{n}.$$

Cross-Links

- **AI:** Expected loss over a data distribution is risk; ERM approximates it with empirical risk.
- **Economics:** Decision theory and expected utility are risk frameworks with different loss functions.
- **Linear Algebra:** Variance/covariance are quadratic forms; risk often becomes a quadratic objective.
- **Quantum:** Estimation risk bounds connect to (quantum) Fisher information.

Structural Compression

Invariant:
Performance of an estimator is defined by expected loss under $\mathbb{P}_\theta$.

Mechanism:
Choose a loss → take expectation → obtain risk as the optimization target.

Cross-System Echo:
In ML, "training" is minimizing an empirical proxy for risk.

Exercises

1. Prove the bias–variance decomposition for squared error loss.
2. For $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, compute the MSE of $\bar{X}_n$ as an estimator of μ.
3. Give an example of a loss function other than squared error (e.g., absolute loss) and define the corresponding risk.

Method of Moments and Maximum Likelihood Estimation (MLE)

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.2

Formal Definition

Let $X_1, \ldots, X_n$ be data generated under a model $\{f_\theta : \theta \in \Theta\}$.

Method of Moments (MoM)

Choose k population moments $m_j(\theta) = \mathbb{E}_\theta[X^j]$ (or other moment-type quantities), and set them equal to the corresponding sample moments:

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

Solve the system:

$$m_j(\theta) = \hat{m}_j, \quad j = 1, \ldots, k$$

for θ. The solution (if it exists) is the MoM estimator $\hat{\theta}_{\text{MoM}}$.

Maximum Likelihood Estimation (MLE)

Define the likelihood:

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

and log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i).$$

An MLE is any maximizer:

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \ell(\theta).$$

Equivalently (when differentiable), solve the score equation:

$$\nabla_\theta \ell(\theta) = 0$$

and select maxima.

Intuition

MoM: "Match what the model predicts on average to what the data shows on average."
Fast, direct, often algebraic.

MLE: "Choose the parameter under which the observed data would be most plausible."
Optimization-based, often statistically efficient, and the central object of likelihood theory.

Structural Role

These are the two canonical construction principles for estimators:

- MoM provides quick, often closed-form estimators and a gateway to estimating equations.
- MLE is the backbone of classical inference (Wald/Score/LR tests, asymptotics, efficiency, Fisher information).

Understanding their contrast sets up: - consistency and asymptotic normality (Domain 7), - Fisher-information bounds (Node 3.4/3.5), - computational optimization (Domain 9).

Mathematical Development

MoM Example Pattern

If the model implies $\mathbb{E}_\theta[X] = g(\theta)$, then MoM sets:

$$g(\hat{\theta}_{\text{MoM}}) = \bar{X}_n.$$

If g is invertible:

$$\hat{\theta}_{\text{MoM}} = g^{-1}(\bar{X}_n).$$

MLE Example Pattern

MLE uses curvature of $\ell(\theta)$: - gradient (score) gives stationary points, - Hessian indicates maxima/minima, - Fisher information predicts asymptotic variance.

Practical Comparison (Canonical)

- **MoM**: simple, may be less efficient, depends on which moments you pick.
- **MLE**: often efficient under regularity, invariant under reparameterization, but may require numerical optimization and can be sensitive to model misspecification.

Worked Examples

Example 1: Bernoulli(p)

MoM using first moment:

$$\mathbb{E}[X] = p, \quad \hat{m}_1 = \bar{X}_n \Rightarrow \hat{p}_{\text{MoM}} = \bar{X}_n.$$

MLE:

$$\ell(p) = \sum_{i=1}^{n} \left(x_i \log p + (1 - x_i) \log(1 - p) \right).$$

Setting derivative to zero yields:

$$\hat{p}_{\text{MLE}} = \bar{X}_n.$$

Here MoM = MLE.

Example 2: Exponential(λ)

Density: $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

MoM:

$$\mathbb{E}[X] = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_{\text{MoM}} = \frac{1}{\bar{X}_n}.$$

MLE:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i \Rightarrow \hat{\lambda}_{\text{MLE}} = \frac{n}{\sum x_i} = \frac{1}{\bar{X}_n}.$$

Again MoM = MLE.

(They diverge in many multi-parameter or constrained models.)

Cross-Links

- **AI:** MLE corresponds to minimizing negative log-likelihood; MoM resembles matching feature statistics (e.g., method-of-moments in latent variable models).
- **Economics:** Structural estimation often uses MLE; MoM generalizes to GMM (bridge to econometrics).
- **Linear Algebra:** MoM often reduces to solving linear systems; MLE uses gradients/Hessians (quadratic forms).
- **Quantum:** Tomography can be posed as MLE; moment-matching appears via expectation-value constraints.

Structural Compression

Invariant:
Both methods define estimators by enforcing model–data agreement.

Mechanism:
MoM matches expectations; MLE maximizes plausibility via likelihood curvature.

Cross-System Echo:
In control systems: match moments (constraints) vs optimize a global objective.

Exercises

1. Compute MoM and MLE for Poisson(λ) and compare.
2. Give an example where MoM and MLE differ (hint: use a multi-parameter model or a constrained parameter space).
3. Explain why maximizing $\ell(\theta)$ is equivalent to maximizing $L(\theta)$.
4. For Normal(μ, σ^2), derive the MLEs for μ and σ^2.

Properties of Estimators: Bias, Consistency, and Efficiency

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.3

Formal Definitions

Let $\hat{\theta}_n$ be an estimator of θ.

1. Bias

$$\text{Bias}_\theta(\hat{\theta}_n) = \mathbb{E}_\theta[\hat{\theta}_n] - \theta.$$

The estimator is **unbiased** if this equals 0 for all θ.

---

2. Consistency

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \to \infty.$$

That is, the estimator converges in probability to the true parameter.

(Strong consistency replaces convergence in probability with almost sure convergence.)

---

3. Efficiency (Introductory Notion)

Among unbiased estimators, an estimator is **more efficient** if it has smaller variance.

In asymptotic settings:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)).$$

An estimator is asymptotically efficient if it achieves the minimal possible variance (to be formalized via Cramér–Rao bound).

Intuition

Bias measures systematic error.
Consistency measures long-run correctness.
Efficiency measures precision.

These are orthogonal dimensions:

- An estimator may be biased but consistent.
- It may be unbiased but inefficient.

- It may be consistent but converge slowly.

Inference requires balancing all three.

Structural Role

This node defines evaluation criteria for estimators.

It prepares:

- Cramér–Rao lower bound (finite-sample efficiency)
- Asymptotic normality
- Optimality theory
- Decision-theoretic comparisons

Without these properties, estimators are just formulas.

Mathematical Development

Bias–Variance Perspective

Recall:

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) + \mathrm{Bias}_\theta(\hat{\theta})^2.$$

Thus:

- Unbiasedness controls the second term.
- Efficiency controls the first term.

Consistency via Law of Large Numbers

If:

$$\hat{\theta}_n = g(\bar{X}_n),$$

and:

$$\bar{X}_n \xrightarrow{P} \mu(\theta),$$

then by continuous mapping theorem:

$$\hat{\theta}_n \xrightarrow{P} g(\mu(\theta)).$$

Thus, many estimators are consistent via LLN.

Worked Examples

Example 1: Sample Mean

If $X_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathbb{E}[\bar{X}_n] = \mu,$$

$$\mathrm{Var}_\theta(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Thus:

- Unbiased
- Consistent (variance $\to 0$)
- Variance decreases at rate $1/n$

Example 2: Biased but Consistent Estimator

Let:

$$\tilde{\theta}_n = \frac{n-1}{n}\bar{X}_n.$$

Bias:

$$\mathbb{E}[\tilde{\theta}_n] - \mu = -\frac{1}{n}\mu.$$

Bias $\to 0$ as $n \to \infty$.

Thus biased but consistent.

Cross-Links

- **AI:** Bias–variance tradeoff mirrors underfitting vs overfitting.
- **Economics:** Consistency crucial for structural parameter estimation.
- **Linear Algebra:** Variance often expressible as quadratic forms.
- **Quantum:** Efficiency bounds tied to (quantum) Fisher information.

Structural Compression

Invariant:
Estimators are evaluated by expectation and variance under the model.

Mechanism:
Long-run convergence (LLN/CLT) governs consistency and asymptotics.

Cross-System Echo:
In optimization, convergence guarantees mirror statistical consistency.

Exercises

1. Prove that if $\text{Var}(\hat{\theta}_n) \to 0$ and estimator is unbiased, then it is consistent.
2. Give an example of a consistent but inefficient estimator.
3. Show that MLE for Bernoulli is unbiased and consistent.
4. Explain difference between finite-sample efficiency and asymptotic efficiency.

The Cramér–Rao Lower Bound

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.4

Regularity Conditions (Important)

Assume:

1. The model $f_\theta(x)$ is differentiable in θ.
2. Differentiation under the integral sign is valid.
3. The support of f_θ does not depend on θ.

These are required for the identities below.

Theorem (Cramér–Rao Inequality)

Let $\hat{\theta}$ be an unbiased estimator of θ.

Then:

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta)},$$

where:

$$\mathcal{I}_n(\theta) = n\mathcal{I}(\theta) = n\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right].$$

Interpretation

No unbiased estimator can have variance smaller than the reciprocal of Fisher information.

Fisher Information sets a fundamental precision limit.

More curvature → smaller lower bound → more precise estimation possible.

Proof Sketch (Core Geometry)

Let:

$$S(\theta) = \frac{\partial}{\partial\theta}\log L(\theta).$$

Then:

$$\mathbb{E}_\theta[S(\theta)] = 0.$$

Differentiate unbiasedness:

$$\mathbb{E}_\theta[\hat{\theta}] = \theta.$$

Differentiating both sides w.r.t. θ:

$$\frac{d}{d\theta}\mathbb{E}_\theta[\hat{\theta}] = 1.$$

Using interchange of derivative and expectation:

$$\mathbb{E}_\theta\left[\hat{\theta}S(\theta)\right] = 1.$$

Now apply Cauchy–Schwarz:

$$1^2 \leq \mathrm{Var}_\theta(\hat{\theta}) \cdot \mathrm{Var}_\theta(S(\theta)).$$

But:

$$\mathrm{Var}_\theta(S(\theta)) = \mathcal{I}_n(\theta).$$

Thus:

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta)}.$$

Equality Condition

Equality holds iff:

$$\hat{\theta} = \theta + c(\theta)S(\theta).$$

In exponential families, the MLE often achieves the bound.

Example: Normal Mean (Known Variance)

If $X_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathcal{I}_n(\mu) = \frac{n}{\sigma^2}.$$

Thus:

$$\mathrm{Var}(\hat{\mu}) \geq \frac{\sigma^2}{n}.$$

The sample mean satisfies:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

It achieves the bound → efficient.

Cross-Links

- **Linear Algebra:** Bound derived from Cauchy–Schwarz in L^2 space.
- **AI:** Fisher Information underlies natural gradient methods.
- **Economics:** Efficiency bounds in structural estimation.
- **Quantum:** Quantum Cramér–Rao bound generalizes this inequality.

Structural Compression

Invariant:
Variance $\geq$ inverse information.

Mechanism:
Geometry of score function via Cauchy–Schwarz.

Cross-System Echo:
In physics, uncertainty principles bound precision.

Exercises

1. Prove the unbiasedness differentiation step carefully.
2. Compute Fisher information for Poisson and derive CRLB.
3. Show sample mean achieves CRLB for Normal mean.
4. Explain why biased estimators can beat the bound.

Sufficiency, Rao–Blackwell, and Lehmann–Scheffé

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.5

Core Definitions

Statistic

A **statistic** is any measurable function of the data:

$$T = T(X_1, \ldots, X_n).$$

Sufficiency (Factorization Theorem)

A statistic T is **sufficient** for θ if the likelihood can be factored as:

$$f_\theta(x_1, \ldots, x_n) = g_\theta(T(x_1, \ldots, x_n))\, h(x_1, \ldots, x_n),$$

where h does not depend on θ.

Equivalently: conditional distribution of the data given T does not depend on θ.

Intuition

A sufficient statistic is a *lossless compression of the data* for the purpose of estimating θ.

Once you know T, the remaining details of the sample contain no additional information about θ.

This is why exponential families matter: they often admit low-dimensional sufficient statistics.

Structural Role

This node gives you the canonical pipeline:

1. Find a sufficient statistic T.
2. Improve any unbiased estimator by conditioning on T (Rao–Blackwell).
3. If T is complete, the improved estimator is uniquely optimal (Lehmann–Scheffé).

This is "optimality by structure," not by brute-force search.

Theorem 1: Rao–Blackwell

Let $\hat{\theta}$ be an estimator of $g(\theta)$ with finite variance, and let T be sufficient for θ.

Define the Rao–Blackwellized estimator:

$$\hat{\theta}^\star = \mathbb{E}_\theta[\hat{\theta} \mid T].$$

Then:

1. If $\hat{\theta}$ is unbiased for $g(\theta)$, so is $\hat{\theta}^\star$.
2. Variance improves (or stays equal):

$$\mathrm{Var}_\theta(\hat{\theta}^\star) \leq \mathrm{Var}_\theta(\hat{\theta}).$$

Theorem 2: Completeness and Lehmann–Scheffé

A statistic T is **complete** if:

$$\mathbb{E}_\theta[a(T)] = 0 \text{ for all } \theta \quad \Rightarrow \quad \mathbb{P}(a(T) = 0) = 1.$$

Lehmann–Scheffé Theorem

If:

- T is sufficient and complete for θ,
- $\hat{\theta}$ is any unbiased estimator of $g(\theta)$,

then:

$$\hat{\theta}^\star = \mathbb{E}[\hat{\theta} \mid T]$$

is the **unique** uniformly minimum-variance unbiased estimator (UMVU) of $g(\theta)$.

Worked Example: Bernoulli(p)

If $X_i \sim \text{Bernoulli}(p)$,

the likelihood is:

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

Thus:

$$T = \sum_{i=1}^{n} X_i$$

is sufficient for p.

An unbiased estimator of p is $\bar{X}$, which is already a function of T, so Rao–Blackwell does not change it.

This illustrates a key point:

Once you are already using a sufficient statistic, there is no "information left" to gain.

Cross-Links

- **AI:** Sufficient representations = no loss of task-relevant information (representation learning analogue).
- **Linear Algebra:** Conditioning is projection in L^2 (variance reduction via orthogonal projection).
- **Economics:** Sufficient statistics appear in exponential family and discrete choice models.
- **Quantum:** Informationally complete measurements vs sufficient summaries for parameter estimation.

Structural Compression

Invariant:
Conditioning on sufficient information cannot increase variance.

Mechanism:
Rao–Blackwell is variance reduction via conditional expectation (projection).

Cross-System Echo:
In ML, better representations reduce sample complexity and variance.

Exercises

1. Prove the variance reduction statement using the law of total variance.
2. Show that if $\hat{\theta}$ is a function of T, then Rao–Blackwell leaves it unchanged.
3. For $X_i \sim \text{Poisson}(\lambda)$, show that $T = \sum X_i$ is sufficient for λ.
4. Explain (conceptually) why completeness yields uniqueness in Lehmann–Scheffé.

Invariance and Equivariance in Estimation

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Point Estimation

Node:
3.6

Core Principle

Estimation procedures should behave naturally under transformations of the parameter.

If we reparameterize the model, the estimator should transform accordingly.

This is called **invariance** (or more precisely, equivariance).

Definition: Invariance of the MLE

Let $\hat{\theta}_{\text{MLE}}$ be the MLE of θ.

If $\phi = g(\theta)$ is a one-to-one transformation, then:

$$\hat{\phi}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}}).$$

That is:

> The MLE of a transformed parameter equals the transformation of the MLE.

This is called the **invariance property of the MLE**.

Why It Holds

The likelihood for ϕ is:

$$L_\phi(\phi) = L_\theta(g^{-1}(\phi)).$$

Maximizing over ϕ is equivalent to maximizing over θ.

Thus the argmax transforms accordingly.

No recalculation needed.

Intuition

If you estimate:

- variance σ^2,
- then want standard deviation σ,

you do not re-estimate. You transform the estimate.

MLE respects the structure of the parameter space.

Other estimators (e.g., unbiased estimators) generally do **not** satisfy this property.

Structural Role

This explains why MLE is "natural":

- It respects reparameterization.
- It behaves coherently under model transformations.
- It is defined through the likelihood geometry rather than moment constraints.

This prepares for:

- Asymptotic normality
- Information geometry
- Delta method
- Likelihood-based testing

Worked Example

Example: Normal Variance

Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with known μ.

MLE for variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum (X_i - \mu)^2.$$

Now consider $\sigma = \sqrt{\sigma^2}$.

By invariance:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

No separate optimization needed.

Contrast: Unbiased Estimators

Unbiased estimators do **not** generally preserve invariance.

Example:

If $\hat{\theta}$ is unbiased for θ,

$$\mathbb{E}[\hat{\theta}] = \theta,$$

it does **not** imply:

$$\mathbb{E}[g(\hat{\theta})] = g(\theta).$$

Thus unbiasedness is not invariant under nonlinear transformations.

This is one reason unbiasedness is not always the dominant principle in modern inference.

Equivariance Under Group Transformations

More generally, if a model is invariant under a transformation group G, estimators can be required to satisfy:

$$\delta(g \cdot x) = g \cdot \delta(x).$$

Such estimators are called **equivariant**.

This becomes important in:

- Location-scale families
- Decision theory
- Invariant testing

Cross-Links

- **Linear Algebra:** Coordinate changes preserve geometric objects.
- **AI:** Parameterization choice affects optimization but not the optimum distribution.
- **Economics:** Reparameterizations in structural models.
- **Quantum:** Different parameterizations of states must yield consistent estimators.

Structural Compression

Invariant:
MLE commutes with smooth one-to-one transformations.

Mechanism:
Argmax of likelihood preserved under reparameterization.

Cross-System Echo:
In geometry, intrinsic properties do not depend on coordinates.

Exercises

1. Prove the invariance property formally.
2. Give an example where unbiasedness fails under transformation.
3. Show that MLE for $\log \theta$ equals $\log(\hat{\theta}_{\text{MLE}})$.
4. Consider a location family $f(x - \theta)$. What form must an equivariant estimator take?

Module 4 — Interval Estimation

Confidence Sets

Position in Knowledge Graph

System:
Statistics

Structural Domain:
Interval Estimation

Node:
4.1

Statistical Setting

Let

$$X = (X_1, \dots, X_n) \sim P_\theta,$$

where:

- $\theta \in \Theta \subseteq \mathbb{R}^k$,
- $\{P_\theta : \theta \in \Theta\}$ is a statistical model,
- inference is based on the observed data X.

In Domain 3, an estimator was a measurable function

$$\hat{\theta} : \mathcal{X} \to \Theta.$$

Interval estimation replaces point-valued estimators by set-valued estimators.

Definition — Confidence Set

A **confidence set** with level $1 - \alpha$, $0 < \alpha < 1$, is a measurable mapping

$$C : \mathcal{X} \to 2^\Theta$$

such that for every $\theta \in \Theta$,

$$\mathbb{P}_\theta\big(\theta \in C(X)\big) = 1 - \alpha.$$

Measurability requires that for each fixed θ,

$$\{X : \theta \in C(X)\}$$

is a measurable event under the sampling σ-algebra.

The probability is taken with respect to P_θ.
The parameter θ is fixed; randomness arises from the data.

Coverage Probability

For each $\theta \in \Theta$, define

$$\gamma(\theta) = \mathbb{P}_\theta\big(\theta \in C(X)\big).$$

Exact confidence requires

$$\gamma(\theta) = 1 - \alpha \quad \forall \theta \in \Theta.$$

If instead

$$\gamma(\theta) \geq 1 - \alpha \quad \forall \theta \in \Theta,$$

the procedure is conservative.

Finite-Sample and Asymptotic Confidence

Finite-Sample Validity

$$\mathbb{P}_\theta\big(\theta \in C(X)\big) = 1 - \alpha$$

holds exactly for each n.

Asymptotic Validity

A sequence $C_n(X)$ satisfies

$$\lim_{n\to\infty} \mathbb{P}_\theta\big(\theta \in C_n(X)\big) = 1 - \alpha.$$

Asymptotic confidence relies on limit distribution theory (LLN, CLT, Slutsky, delta method).

One-Dimensional Case

If $\theta \in \mathbb{R}$, a confidence set reduces to an interval

$$C(X) = [L(X), U(X)],$$

with

$$\mathbb{P}_\theta\big(L(X) \leq \theta \leq U(X)\big) = 1 - \alpha.$$

Here $L(X)$ and $U(X)$ are statistics.

Structural Role

Point estimation provides a single estimate.
Confidence sets quantify uncertainty around the parameter.

They formalize long-run coverage control and connect directly to hypothesis testing via duality (developed in Module 5).

Structural Compression

Invariant:
$\mathbb{P}_\theta\big(\theta \in C(X)\big) = 1 - \alpha$.

Mechanism:
Construct a random subset of parameter space whose sampling distribution controls coverage.

Cross-System Echo:
Confidence sets generalize error control: risk bounds in estimation and Type I error control in testing are parallel structures.

Pivotal Quantities

Statistical Setting

Let

$$X = (X_1, \dots, X_n) \sim P_\theta,$$

with parameter $\theta \in \Theta \subseteq \mathbb{R}^k$.

In Node 4.1, confidence sets were defined abstractly via coverage probability.
We now construct such sets using **pivotal quantities**.

Definition — Pivot

A statistic

$$T(X, \theta)$$

is called a **pivotal quantity** if its sampling distribution does not depend on the parameter θ.

Formally,

$$\mathcal{L}_\theta\big(T(X,\theta)\big)$$

is the same for all $\theta \in \Theta$.

Construction Principle

Suppose:

1. $T(X, \theta)$ is a pivot.
2. Its distribution function F_T is known.
3. Constants a_α, b_α satisfy

$$\mathbb{P}(a_\alpha \le T \le b_\alpha) = 1 - \alpha.$$

Then:

$$a_\alpha \le T(X, \theta) \le b_\alpha$$

can be algebraically inverted to produce a confidence set for θ.

Coverage follows because the distribution of T does not depend on θ.

Example — Normal Mean (Known Variance)

Let

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad \sigma^2 \text{ known.}$$

Define

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Then

$$Z \sim \mathcal{N}(0, 1),$$

independently of μ.

Thus Z is a pivot.

Since

$$\mathbb{P}\big(-z_{\alpha/2} \le Z \le z_{\alpha/2}\big) = 1 - \alpha,$$

we obtain

$$\mathbb{P}\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

This yields the classical confidence interval.

Finite-Sample Exactness

When a true pivot exists and its distribution is known exactly, the resulting confidence set has finite-sample coverage:

$$\mathbb{P}_\theta(\theta \in C(X)) = 1 - \alpha.$$

Exact pivots typically arise in:

- Normal models
- Chi-square and t-distributions
- Exponential family models with known variance structure

Structural Role

Pivots provide the **finite-sample engine** of confidence theory.

They convert:

Distributional identity

into

Coverage control.

They are the constructive counterpart to the abstract definition in 4.1.

Structural Compression

Invariant

Distribution of $T(X, \theta)$ does not depend on θ.

Mechanism

Probability statement for pivot $\rightarrow$ algebraic inversion $\rightarrow$ confidence set.

Cross-System Echo

Symmetry removal: parameter dependence is normalized away, leaving a fixed reference distribution.

Likelihood-Based Confidence Sets

Statistical Setting

Let

$$X \sim P_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}^k,$$

with likelihood function

$$L(\theta) = L(\theta; X).$$

In Node 4.2, confidence sets were constructed from pivots.
We now construct them directly from the likelihood function.

Likelihood Ratio Statistic

Define the likelihood ratio:

$$\Lambda(\theta) = \frac{L(\theta)}{L(\hat{\theta})},$$

where $\hat{\theta}$ is the MLE.

Equivalently, in log-scale,

$$2 \log \Lambda(\theta) = 2\big(\ell(\theta) - \ell(\hat{\theta})\big).$$

Since $\ell(\hat{\theta}) \geq \ell(\theta)$, the statistic is nonpositive.

Define instead

$$W(\theta) = 2\big(\ell(\hat{\theta}) - \ell(\theta)\big).$$

Likelihood-Based Confidence Set

For a scalar parameter, define

$$C(X) = \{\theta \in \Theta : W(\theta) \leq c_\alpha\},$$

where c_α is chosen such that

$$\mathbb{P}_\theta\big(W(\theta) \leq c_\alpha\big) \approx 1 - \alpha.$$

Under regularity conditions and for large n,

$$W(\theta) \xrightarrow{d} \chi^2_1.$$

Thus one may choose

$$c_\alpha = \chi^2_{1,\,1-\alpha}.$$

Interpretation

Likelihood-based confidence sets consist of parameter values whose log-likelihood is sufficiently close to the maximum.

They are defined by:

$$\ell(\theta) \geq \ell(\hat{\theta}) - \frac{1}{2} c_\alpha.$$

Geometrically, they are level sets of the likelihood function.

Finite vs Asymptotic Validity

Unlike pivots, likelihood-based intervals are typically:

- Asymptotically valid,
- Not exactly finite-sample valid (except in special cases).

Their justification relies on:

- Asymptotic normality of MLE,
- Quadratic approximation of log-likelihood,
- Fisher information curvature.

These results are formalized in Domain 7.

Structural Role

Likelihood-based intervals:

- Are invariant under reparameterization,
- Extend naturally to multiparameter settings,
- Connect directly to likelihood ratio tests (Module 5),
- Reflect local information geometry.

They generalize pivot constructions when exact pivots do not exist.

Structural Compression

Invariant

Confidence set defined by likelihood level:

$$\ell(\theta) \geq \ell(\hat{\theta}) - \frac{1}{2} c_\alpha.$$

Mechanism

Likelihood curvature + asymptotic χ^2 distribution of LR statistic.

Cross-System Echo

Level-set geometry parallels quadratic approximation in optimization and second-order Taylor expansions.

Wald, Score, and Likelihood Ratio Intervals

Statistical Setting

Let

$$X_1, \ldots, X_n \sim P_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}.$$

Assume:

1. The model satisfies regularity conditions for MLE asymptotics.
2. The MLE $\hat{\theta}_n$ exists and is consistent.
3. Fisher information $I(\theta)$ is positive and finite.
4. Asymptotic normality holds:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\big(0, I(\theta)^{-1}\big).$$

All asymptotic statements are in distribution under P_θ.

---

1. Wald Interval

From asymptotic normality,

$$\frac{\hat{\theta}_n - \theta}{\sqrt{I(\hat{\theta}_n)^{-1}/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Thus,

$$C_n^W = \left[\hat{\theta}_n \pm z_{\alpha/2}\sqrt{\frac{1}{nI(\hat{\theta}_n)}}\right].$$

Coverage

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n^W) = 1 - \alpha.$$

Wald intervals depend directly on the estimator scale.

---

2. Score (Rao) Interval

Let

$$U_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$$

be the score function, and

$$I_n(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell_n(\theta).$$

Under regularity,

$$\frac{U_n(\theta)}{\sqrt{I_n(\theta)}} \xrightarrow{d} \mathcal{N}(0,1).$$

The score confidence set is

$$C_n^S = \left\{ \theta : \frac{U_n(\theta)^2}{I_n(\theta)} \leq z_{\alpha/2}^2 \right\}.$$

Coverage

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n^S) = 1 - \alpha.$$

Score intervals evaluate curvature at the candidate parameter, not at the estimator.

---

3. Likelihood Ratio (LR) Interval

Define the likelihood ratio statistic:

$$\Lambda_n(\theta) = 2\big(\ell_n(\hat{\theta}_n) - \ell_n(\theta)\big).$$

Under standard regularity conditions,

$$\Lambda_n(\theta) \xrightarrow{d} \chi_1^2.$$

The LR confidence set is

$$C_n^{LR} = \left\{ \theta : \Lambda_n(\theta) \leq \chi_{1,1-\alpha}^2 \right\}.$$

Coverage

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n^{LR}) = 1 - \alpha.$$

LR intervals are invariant under smooth reparameterization.

Comparative Structure

Method	Center	Information Evaluated At	Invariance
Wald	$\hat{\theta}$	Estimated parameter	Not invariant
Score	Solution set	Candidate θ	Partially invariant
LR	MLE	Global likelihood	Fully invariant

All three coincide to first order in $n^{-1/2}$, but differ in finite samples.

Structural Role

These are the three canonical asymptotic constructions:

- Wald: estimator-based geometry.
- Score: local curvature testing geometry.
- LR: global likelihood geometry.

They form the asymptotic backbone of modern inference and connect directly to hypothesis testing in Domain 5.

Structural Compression

Invariant

Asymptotic reference distribution (Normal or Chi-square) independent of θ.

Mechanism

Approximate pivotal statistic $\rightarrow$ inequality inversion $\rightarrow$ confidence set.

Cross-System Echo

Local quadratic expansion of log-likelihood (LAN structure) governs both interval estimation and hypothesis testing.

Asymptotic Confidence Theory

Statistical Setting

Let

$$X_1, \dots, X_n \sim P_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}^k.$$

Let $C_n(X)$ be a sequence of measurable confidence sets.

All probability statements are taken under P_θ for fixed θ.

Definition — Asymptotic Level

The sequence C_n has **asymptotic level** $1-\alpha$ if for every fixed $\theta \in \Theta$,

$$\lim_{n\to\infty} \mathbb{P}_\theta\big(\theta \in C_n(X)\big) = 1-\alpha.$$

This is pointwise convergence in θ.

No uniformity over Θ is implied.

Uniform Asymptotic Coverage

If

$$\sup_{\theta\in\Theta} |\mathbb{P}_\theta(\theta \in C_n(X)) - (1-\alpha)| \to 0,$$

then coverage holds uniformly over Θ.

Uniformity is strictly stronger than pointwise convergence.

Construction via Asymptotic Pivots

Suppose there exists a statistic $T_n(X,\theta)$ such that

$$T_n(X,\theta) \xrightarrow{d} T,$$

where the limiting distribution of T does not depend on θ.

Let constants a_α, b_α satisfy

$$\mathbb{P}(a_\alpha \le T \le b_\alpha) = 1-\alpha.$$

Define

$$C_n(X) = \{\theta : a_\alpha \le T_n(X,\theta) \le b_\alpha\}.$$

If convergence in distribution holds under P_θ, then

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in C_n(X)) = 1 - \alpha.$$

This follows from the Portmanteau theorem when boundary probabilities vanish.

First-Order Expansion

In regular parametric models with scalar parameter,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}).$$

Replacing $I(\theta)$ by a consistent estimator yields

$$\frac{\hat{\theta}_n - \theta}{\sqrt{I(\hat{\theta}_n)^{-1}/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Intervals constructed from this expansion achieve asymptotic level $1 - \alpha$.

Higher-Order Considerations

Asymptotic coverage error can be expressed as

$$\mathbb{P}_\theta(\theta \in C_n) = 1 - \alpha + O(n^{-1/2}).$$

Higher-order corrections (e.g., Bartlett adjustments, Edgeworth expansions) reduce coverage error to smaller orders.

Such refinements require smoothness conditions and higher-order moment existence.

Structural Role

Finite-sample pivots rarely exist outside special models.

Asymptotic confidence theory generalizes interval construction by replacing exact distributional identities with limit laws.

It transfers the geometry of asymptotic normality into coverage control.

Structural Compression

Invariant

Limiting distribution independent of θ.

Mechanism

Asymptotic pivot → probability bound in limit → inverted inequality.

Cross-System Echo

Large-sample normality underlies both interval estimation and hypothesis testing; confidence level and test size share the same asymptotic reference law.

Bootstrap Confidence Intervals

Statistical Setting

Let

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P,$$

where P is unknown. Let $\theta = \theta(P)$ be a functional of P, and let $\hat{\theta}_n = \theta(\hat{P}_n)$ be the plug-in estimator, where $\hat{P}_n$ is the empirical distribution.

The bootstrap replaces the unknown sampling distribution of $\hat{\theta}_n - \theta$ with the distribution of $\hat{\theta}_n^* - \hat{\theta}_n$ under resampling from $\hat{P}_n$.

Definition — Bootstrap Distribution

A **bootstrap sample** is

$$X_1^*, \dots, X_n^* \overset{\text{i.i.d.}}{\sim} \hat{P}_n,$$

i.e., a sample of size n drawn with replacement from $\{X_1, \dots, X_n\}$.

Let $\hat{\theta}_n^* = \theta(\hat{P}_n^*)$ be the estimator computed on the bootstrap sample.

The **bootstrap distribution** of $\hat{\theta}_n^* - \hat{\theta}_n$ approximates the sampling distribution of $\hat{\theta}_n - \theta$.

Construction — Percentile Interval

Let $q_\alpha^*(\cdot)$ denote quantiles of the bootstrap distribution of $\hat{\theta}_n^*$.

The **percentile bootstrap interval** is

$$C_n^{\text{perc}}(X) = \left[q_{\alpha/2}^*,\ q_{1-\alpha/2}^*\right].$$

Construction — Basic Bootstrap Interval

The **basic bootstrap interval** (reverse percentile) inverts the bootstrap distribution of $\hat{\theta}_n^* - \hat{\theta}_n$:

$$C_n^{\text{basic}}(X) = \left[2\hat{\theta}_n - q_{1-\alpha/2}^*,\ 2\hat{\theta}_n - q_{\alpha/2}^*\right].$$

This corrects for asymmetry in the sampling distribution of $\hat{\theta}_n - \theta$.

Construction — Bootstrap-t Interval

Let $\hat{\sigma}_n$ be a consistent estimator of the standard deviation of $\hat{\theta}_n$. Define the studentized pivot

$$Z_n = \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n}.$$

Let $Z_n^* = (\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$ be the bootstrap analogue.

The **bootstrap-t interval** is

$$C_n^t(X) = \Big[\hat{\theta}_n - q_{1-\alpha/2}^{*,t}\,\hat{\sigma}_n,\ \hat{\theta}_n - q_{\alpha/2}^{*,t}\,\hat{\sigma}_n\Big],$$

where $q_p^{*,t}$ are quantiles of the bootstrap distribution of Z_n^*.

The bootstrap-t achieves second-order accuracy: coverage error is $O(n^{-1})$ rather than $O(n^{-1/2})$.

Consistency

Under regularity conditions, the bootstrap distribution of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ converges weakly to the same limit as $\sqrt{n}(\hat{\theta}_n - \theta)$, in probability over the original sample.

Formally, if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} L$, then

$$\sup_t \Big|\mathbb{P}^*\big(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \le t\big) - \mathbb{P}\big(\sqrt{n}(\hat{\theta}_n - \theta) \le t\big)\Big| \xrightarrow{p} 0,$$

where $\mathbb{P}^*$ denotes probability under resampling conditional on $X_1, \ldots, X_n$.

Structural Role

Pivotal constructions (Node 4.2) require knowledge of the sampling distribution of $T_n(X, \theta)$.

The bootstrap removes this requirement by estimating the sampling distribution nonparametrically from the data.

It provides asymptotically valid confidence intervals for functionals $\theta(P)$ where no closed-form pivot exists.

Structural Compression

Invariant

The empirical distribution $\hat{P}_n$ consistently estimates P; the bootstrap distribution of $\hat{\theta}_n^* - \hat{\theta}_n$ consistently estimates the sampling distribution of $\hat{\theta}_n - \theta$.

Mechanism

Replace the unknown P with $\hat{P}_n$ in the sampling distribution calculation; simulate the resulting distribution by resampling.

Cross-System Echo

The bootstrap is a nonparametric Monte Carlo approximation to the sampling distribution; it connects to MCMC (Module 9) in that both substitute simulation for analytical distribution computation. In AI, bootstrap resampling underlies bagging and ensemble variance estimation.

Module 5 — Hypothesis Testing

Hypothesis Testing Framework

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.1

This node establishes the formal language in which all of frequentist hypothesis testing is expressed. It defines tests, errors, power, level, and size as mathematical objects, setting the stage for optimal test construction (Neyman–Pearson, UMP, LR) and calibration (p-values, confidence duality). Every subsequent node in Module 5 inherits the definitions and constraints introduced here.

Formal Definition

Statistical Model

Let

$$X = (X_1, \ldots, X_n) \sim P_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}^k,$$

where $\{P_\theta : \theta \in \Theta\}$ is a parametric family dominated by a sigma-finite measure μ, with densities p_θ.

Hypotheses

A **null hypothesis** is a subset $\Theta_0 \subseteq \Theta$. An **alternative hypothesis** is a subset $\Theta_1 \subseteq \Theta$ with $\Theta_0 \cap \Theta_1 = \emptyset$. The testing problem is written

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

A hypothesis is **simple** if it specifies a unique distribution ($|\Theta_0| = 1$) and **composite** otherwise.

Test Function

A **test function** (or **randomized test**) is a measurable function

$$\phi : \mathcal{X} \to [0, 1],$$

where $\phi(x)$ is the probability of rejecting H_0 upon observing x. A **non-randomized test** takes values in $\{0, 1\}$ only.

Critical Region

For a non-randomized test, the **critical region** (rejection region) is

$$R = \{x \in \mathcal{X} : \phi(x) = 1\},$$

and the **acceptance region** is R^c.

Intuition

Hypothesis testing formalizes a binary decision under uncertainty: given data, decide whether the evidence is strong enough to reject a default claim H_0. The framework does not ask "Is H_0 true?" in any absolute sense. Instead, it asks: "Under the assumption that H_0 holds, is the observed data sufficiently surprising?"

The test function ϕ is the decision rule. The power function β_ϕ is its operating characteristic, describing how the test behaves across the entire parameter space. The level constraint bounds the worst-case false rejection rate, converting the problem into constrained optimization: maximize power subject to a budget on Type I error.

Structural Role

The test function ϕ and power function β_ϕ encode the complete frequentist decision structure. All subsequent constructions in Module 5 are special cases of this framework:

- Neyman–Pearson (Node 5.2): optimal ϕ for simple-vs-simple problems.
- UMP tests (Node 5.3): optimal ϕ for one-sided composite problems under MLR.
- LR tests (Node 5.4): general-purpose ϕ for composite problems.
- p-values (Node 5.5): continuous calibration of ϕ against H_0.
- Test–confidence duality (Node 5.6): geometric reinterpretation of families of tests.

The framework requires the probabilistic foundations of Module 1, the distributional theory of Module 2, and the estimation concepts of Module 4 (particularly sufficient statistics).

Mathematical Development

Error Types

For a test function ϕ:

Type I error at $\theta \in \Theta_0$:

$$\alpha(\theta) = \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{reject } H_0).$$

Type II error at $\theta \in \Theta_1$:

$$\beta^{\text{II}}(\theta) = 1 - \mathbb{E}_\theta[\phi(X)] = \mathbb{P}_\theta(\text{fail to reject } H_0).$$

Power Function

The **power function** of ϕ is

$$\beta_\phi(\theta) = \mathbb{E}_\theta[\phi(X)], \quad \theta \in \Theta.$$

Ideal behavior: $\beta_\phi(\theta)$ small on Θ_0, large on Θ_1.

Level and Size

A test ϕ has **level** α if

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)] \leq \alpha.$$

The **size** of ϕ is

$$\alpha^* = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)].$$

An exact level α test achieves $\alpha^* = \alpha$.

The Power Function Characterizes the Test

Proposition. Two tests ϕ_1 and ϕ_2 have the same power function if and only if $\phi_1 = \phi_2$ almost everywhere with respect to every P_θ, $\theta \in \Theta$.

Proof. If $\beta_{\phi_1}(\theta) = \beta_{\phi_2}(\theta)$ for all $\theta \in \Theta$, then $\mathbb{E}_\theta[\phi_1(X) - \phi_2(X)] = 0$ for all θ. Since $\phi_1 - \phi_2$ is bounded and $\{P_\theta\}$ is a family of distributions, this means $\int (\phi_1 - \phi_2)\, dP_\theta = 0$ for all θ. If the family $\{P_\theta\}$ is complete (Node 2.6), then $\phi_1 = \phi_2$ a.e. μ. More generally, $\phi_1 = \phi_2$ a.e. with respect to every P_θ. The converse is immediate. □

Unbiased Tests

A test ϕ is **unbiased** at level α if

$$\mathbb{E}_\theta[\phi(X)] \leq \alpha \quad \forall \theta \in \Theta_0, \qquad \mathbb{E}_\theta[\phi(X)] \geq \alpha \quad \forall \theta \in \Theta_1.$$

Unbiasedness ensures the test is at least as likely to reject under the alternative as under the null.

Impossibility of Simultaneous Minimization

Proposition. Unless P_θ is identical for all $\theta \in \Theta_0 \cup \Theta_1$, no test can simultaneously achieve $\beta_\phi(\theta) = 0$ for all $\theta \in \Theta_0$ and $\beta_\phi(\theta) = 1$ for all $\theta \in \Theta_1$ when P_{θ_0} and P_{θ_1} have overlapping support.

Proof. Suppose P_{θ_0} and P_{θ_1} are mutually absolutely continuous for some $\theta_0 \in \Theta_0$, $\theta_1 \in \Theta_1$. If $\phi(x) = 1$ on a set R, then $P_{\theta_0}(R) = 0$ requires R to be a P_{θ_0}-null set. By mutual absolute continuity, R is also P_{θ_1}-null, so $\beta_\phi(\theta_1) = 0$, a contradiction. □

This impossibility drives the entire theory: since perfect separation is unattainable, one must manage the tradeoff between error types.

Worked Examples

Example 1: Normal Mean, Known Variance

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$.

The test statistic is $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$. Under H_0, $Z \sim N(0, 1)$.

The level α non-randomized test is $\phi(x) = \mathbf{1}(Z > z_\alpha)$, where $z_\alpha = \Phi^{-1}(1 - \alpha)$.

The power function is

$$\beta_\phi(\mu) = \mathbb{P}_\mu(Z > z_\alpha) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).$$

At $\mu = \mu_0$: $\beta_\phi(\mu_0) = \alpha$ (size equals level). As $\mu \to \infty$: $\beta_\phi(\mu) \to 1$. As $n \to \infty$ for any fixed $\mu > \mu_0$: $\beta_\phi(\mu) \to 1$ (consistency).

Example 2: Randomized Test for Discrete Data

Let $X \sim \text{Binomial}(10, p)$. Test $H_0 : p = 0.5$ vs $H_1 : p > 0.5$ at level $\alpha = 0.05$.

Under H_0, $\mathbb{P}_{0.5}(X \geq 9) = \binom{10}{9}(0.5)^{10} + (0.5)^{10} = 11/1024 \approx 0.0107$, and $\mathbb{P}_{0.5}(X \geq 8) \approx 0.0547$.

Since $0.0107 < 0.05 < 0.0547$, no non-randomized test achieves exact level 0.05.

The randomized test sets $\phi(x) = 1$ for $x \geq 9$, $\phi(8) = \gamma$, and $\phi(x) = 0$ for $x \leq 7$, where

$$\gamma = \frac{0.05 - 0.0107}{0.0547 - 0.0107} = \frac{0.0393}{0.0440} \approx 0.893.$$

This achieves $\mathbb{E}_{0.5}[\phi(X)] = 0.05$ exactly.

Cross-Links

- **AI:** Type I and Type II errors map directly to false positive rate and false negative rate in binary classification. The ROC curve is the graph of power vs size as the threshold varies.
- **Economics:** Hypothesis tests underpin structural break detection and policy evaluation (e.g., testing whether a treatment effect is zero).
- **Linear Algebra:** The rejection region R is a subset of $\mathbb{R}^n$; in Gaussian models, R is a half-space or cone determined by projections.
- **Quantum:** Quantum hypothesis testing distinguishes quantum states ρ_0 vs ρ_1; the test is a POVM element $0 \leq E \leq I$, generalizing $\phi : \mathcal{X} \to [0, 1]$ to the operator setting.

Structural Compression

Invariant: The fundamental tradeoff: Type I and Type II errors cannot simultaneously be reduced without additional information (larger n, stronger separation of P_{θ_0} and P_{θ_1}).

Mechanism: The level constraint $\sup_{\Theta_0} \mathbb{E}_\theta[\phi] \leq \alpha$ fixes a budget on false rejection; optimality is defined by maximizing power within that budget.

Cross-System Echo: Test functions correspond to decision rules in statistical decision theory; the power function is the risk function under 0-1 loss. In signal detection theory, the same tradeoff appears as the receiver operating characteristic. In machine learning, precision-recall tradeoffs mirror the Type I / Type II structure.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Write the power function of the test that rejects $H_0 : \lambda = \lambda_0$ in favor of $H_1 : \lambda < \lambda_0$ when $\bar{X} > c$, and find c in terms of α and the gamma distribution.
2. Prove that if ϕ is an unbiased level α test, then $\beta_\phi(\theta_0) = \alpha$ for every boundary point θ_0 of Θ_0 that is also in the closure of Θ_1.
3. Consider $X \sim \text{Poisson}(\lambda)$. Construct a randomized test of $H_0 : \lambda = 1$ vs $H_1 : \lambda = 3$ at exact level $\alpha = 0.05$. Compute the power.
4. Prove that for a one-parameter exponential family, the power function $\beta_\phi(\theta)$ of any non-trivial level α test is a continuous function of θ.
5. Let ϕ_1 and ϕ_2 be two level α tests. Show that $\phi_3 = \lambda\phi_1 + (1 - \lambda)\phi_2$ for $\lambda \in [0, 1]$ is also a level α test. What does this say about the set of level α tests?

6. For testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X \sim N(\mu, 1)$ (single observation), plot the power function of the test $\phi(x) = \mathbf{1}(|x| > z_{\alpha/2})$ for $\alpha = 0.05$. Show that $\beta_\phi(\mu) \to 1$ as $|\mu| \to \infty$ and that $\beta_\phi(0) = 0.05$.

7. Argue, structurally, why the level constraint $\sup_{\Theta_0} \mathbb{E}_\theta[\phi] \leq \alpha$ is a linear constraint on ϕ, and explain how this convex structure enables the existence of optimal tests via constrained optimization (connecting to the Neyman–Pearson framework of Node 5.2).

Neyman–Pearson Theory

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.2

The Neyman–Pearson Lemma is the foundational optimality result in hypothesis testing. It characterizes the most powerful test for simple-vs-simple problems and establishes the likelihood ratio as the fundamental statistic for hypothesis testing. All optimal testing theory (UMP tests, LR tests) descends from this result. It requires the testing framework of Node 5.1 and the distributional theory (densities, likelihood) of Nodes 2.3–2.4.

Formal Definition

Setting

Let P_0 and P_1 be two distributions with densities p_0 and p_1 with respect to a common dominating measure μ. Consider the simple-vs-simple testing problem:

$$H_0 : P = P_0 \quad \text{vs} \quad H_1 : P = P_1.$$

Most Powerful Test

Among all level α tests, a test ϕ^* is **most powerful (MP)** if

$$\mathbb{E}_1[\phi^*(X)] \geq \mathbb{E}_1[\phi(X)]$$

for every test ϕ satisfying $\mathbb{E}_0[\phi(X)] \leq \alpha$.

Intuition

The Neyman–Pearson Lemma answers the question: given a budget α on false alarm probability, how should one allocate rejection probability across the sample space to maximize detection?

The answer is greedy allocation. Rank every point x by how much more likely it is under P_1 than under P_0, i.e., by the likelihood ratio $p_1(x)/p_0(x)$. Reject starting from the points with the highest ratio, spending the budget α on the points that contribute the most power per unit of size.

This is an isoperimetric principle: among all sets of prescribed P_0-measure α, the one with the largest P_1-measure is the upper level set of the likelihood ratio.

Structural Role

The Neyman–Pearson Lemma characterizes optimal tests for simple hypotheses. The likelihood ratio p_1/p_0 is the sufficient statistic for the simple-vs-simple problem: all relevant information for this test is contained in this ratio.

Extensions to composite hypotheses require additional structure: - Monotone likelihood ratio yields UMP tests for one-sided problems (Node 5.3). - Asymptotic approximations yield LR tests for general composite problems (Node 5.4). - The p-value (Node 5.5) calibrates the NP test statistic on the probability scale.

Mathematical Development

Theorem – Neyman–Pearson Lemma

There exists a most powerful level α test of $H_0 : P_0$ vs $H_1 : P_1$, and it has the form

$$\phi^*(x) = \begin{cases} 1 & \text{if } p_1(x) > k\, p_0(x), \\ \gamma & \text{if } p_1(x) = k\, p_0(x), \\ 0 & \text{if } p_1(x) < k\, p_0(x), \end{cases}$$

where $k \geq 0$ and $\gamma \in [0, 1]$ are chosen to satisfy $\mathbb{E}_0[\phi^*(X)] = \alpha$.

Moreover, ϕ^* is unique almost everywhere with respect to μ, except possibly on the set $\{p_1 = k\, p_0\}$.

Proof

Let ϕ be any level α test, i.e., $\mathbb{E}_0[\phi(X)] \leq \alpha$. Define $\Delta(x) = \phi^*(x) - \phi(x)$. We must show $\mathbb{E}_1[\Delta(X)] \geq 0$.

Consider the decomposition of the sample space into three regions:

- On $S_+ = \{x : p_1(x) > k\, p_0(x)\}$: $\phi^*(x) = 1 \geq \phi(x)$, so $\Delta(x) \geq 0$, and $p_1(x) - k\, p_0(x) > 0$.
- On $S_0 = \{x : p_1(x) = k\, p_0(x)\}$: $p_1(x) - k\, p_0(x) = 0$.
- On $S_- = \{x : p_1(x) < k\, p_0(x)\}$: $\phi^*(x) = 0 \leq \phi(x)$, so $\Delta(x) \leq 0$, and $p_1(x) - k\, p_0(x) < 0$.

Therefore $\Delta(x)(p_1(x) - k\, p_0(x)) \geq 0$ everywhere, giving

$$\int \Delta(x)\, p_1(x)\, d\mu(x) \geq k \int \Delta(x)\, p_0(x)\, d\mu(x).$$

The right-hand side is $k(\mathbb{E}_0[\phi^*] - \mathbb{E}_0[\phi]) = k(\alpha - \mathbb{E}_0[\phi]) \geq 0$, since $\mathbb{E}_0[\phi] \leq \alpha$.

Hence $\mathbb{E}_1[\phi^*] \geq \mathbb{E}_1[\phi]$. $\square$

Existence of k and γ

Define $F_0(t) = P_0(p_1(X)/p_0(X) \leq t)$, the CDF of the likelihood ratio under P_0. Set k to be the $(1-\alpha)$-quantile of $p_1(X)/p_0(X)$ under P_0:

$$k = \inf\{t : F_0(t) \geq 1 - \alpha\}.$$

If $P_0(p_1(X)/p_0(X) = k) = 0$, then γ is irrelevant and the test is non-randomized. Otherwise, choose

$$\gamma = \frac{\alpha - P_0(p_1(X)/p_0(X) > k)}{P_0(p_1(X)/p_0(X) = k)}.$$

Corollary: Power Is at Least α

If $P_0 \neq P_1$, then the power of the NP test satisfies $\mathbb{E}_1[\phi^*] \geq \alpha$, with equality only when $p_1 = k\,p_0$ a.e. In particular, the NP test is unbiased.

Proof. From the proof above, $\mathbb{E}_1[\phi^*] \geq k \cdot 0 + \mathbb{E}_1[\phi]$ for the trivial test $\phi \equiv \alpha$, which has $\mathbb{E}_1[\phi] = \alpha$. Since $\int \Delta(x)(p_1(x) - kp_0(x))\,d\mu \geq 0$ with the choice $\phi \equiv \alpha$, and equality requires $p_1 = kp_0$ a.e. on sets where $\Delta \neq 0$. $\square$

Sufficiency of the Likelihood Ratio

The NP test depends on the data only through the likelihood ratio $\Lambda(x) = p_1(x)/p_0(x)$. By the factorization theorem (Node 2.4), $\Lambda(X)$ is a sufficient statistic for the family $\{P_0, P_1\}$. Thus the NP test achieves optimality by reducing the data to a one-dimensional sufficient statistic.

Worked Examples

Example 1: Normal Mean Shift

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$. Test $H_0 : \mu = 0$ vs $H_1 : \mu = 1$ at level α.

The likelihood ratio is

$$\frac{p_1(x)}{p_0(x)} = \prod_{i=1}^{n} \frac{e^{-(x_i-1)^2/2}}{e^{-x_i^2/2}} = \exp\left(\sum_{i=1}^{n} x_i - \frac{n}{2}\right).$$

This is monotone increasing in $\bar{x} = n^{-1}\sum x_i$. The NP test rejects when $\bar{X} > c$, where c satisfies $P_0(\bar{X} > c) = \alpha$.

Under H_0, $\bar{X} \sim N(0, 1/n)$, so $c = z_\alpha/\sqrt{n}$.

The power is

$$\beta(\mu = 1) = P_1(\bar{X} > z_\alpha/\sqrt{n}) = 1 - \Phi(z_\alpha - \sqrt{n}).$$

For $n = 25$, $\alpha = 0.05$: $c = 1.645/5 = 0.329$, and power $= 1 - \Phi(1.645 - 5) = 1 - \Phi(-3.355) \approx 0.9996$.

Example 2: Exponential Rate

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$ with density $\lambda e^{-\lambda x}$. Test $H_0 : \lambda = 1$ vs $H_1 : \lambda = 2$ at level $\alpha = 0.05$.

The likelihood ratio is

$$\frac{p_1(x)}{p_0(x)} = \frac{2^n e^{-2\sum x_i}}{e^{-\sum x_i}} = 2^n \exp\left(-\sum_{i=1}^{n} x_i\right).$$

This is decreasing in $T = \sum x_i$, so the NP test rejects for small T.

Under H_0, $2T = 2\sum X_i \sim \chi^2_{2n}$. For $n = 5$, reject when $2T < \chi^2_{10,0.95}$. From tables, $\chi^2_{10,0.95} = 3.940$, so reject when $T < 1.970$.

Power: Under H_1, $4T = 4\sum X_i \sim \chi^2_{10}$, so $P_1(T < 1.970) = P(\chi^2_{10} < 7.880) \approx 0.36$.

Cross-Links

- **AI:** The likelihood ratio is the Bayes-optimal classifier under equal priors and 0-1 loss. The NP lemma formalizes the optimality of log-odds thresholding in binary classification.
- **Economics:** In mechanism design, the NP structure appears in optimal auction theory (Myerson): allocate the good to the bidder with the highest virtual value, analogous to rejecting for the highest likelihood ratio.
- **Linear Algebra:** In Gaussian models, the NP critical region is a half-space in $\mathbb{R}^n$, determined by the projection of the data onto the direction $\mu_1 - \mu_0$.
- **Quantum:** The Helstrom bound on quantum state discrimination is the quantum analogue of the NP bound: the optimal POVM element projects onto the positive part of $\rho_1 - k\rho_0$.

Structural Compression

Invariant: Among level α tests of P_0 vs P_1, the likelihood ratio test is most powerful.

Mechanism: Reject when $p_1(x)/p_0(x) > k$: the decision boundary is a level set of the likelihood ratio, allocating rejection probability greedily to the highest-evidence points.

Cross-System Echo: The NP lemma is an instance of the general isoperimetric principle: among all sets of fixed measure under one distribution, the one with maximal measure under another is the upper level set of the Radon–Nikodym derivative. This principle recurs in information theory (channel coding), optimal transport (Monge sets), and quantum information (Helstrom measurement).

Exercises

1. Prove that the Neyman–Pearson test is unique almost everywhere when the distribution of $p_1(X)/p_0(X)$ under P_0 is continuous.
2. Show that the power $\mathbb{E}_1[\phi^*(X)]$ is a non-decreasing function of α, and strictly increasing when P_0 and P_1 are mutually absolutely continuous.
3. Let $X \sim \text{Poisson}(\lambda)$. Derive the NP test of $H_0 : \lambda = 1$ vs $H_1 : \lambda = 3$ at level $\alpha = 0.05$. Compute the threshold k, the randomization probability γ, and the exact power.
4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Show that the NP test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu = \mu_1$ (with $\mu_1 > \mu_0$) rejects for large $\bar{X}$, and that the critical value does not depend on μ_1.
5. Prove that if ϕ^* is the NP test at level α and ϕ^* has power β^*, then the NP test of $H_0 : P_1$ vs $H_1 : P_0$ at level $1 - \beta^*$ has power $1 - \alpha$.
6. Let $P_0 = N(0, 1)$ and $P_1 = N(0, \sigma^2)$ with $\sigma^2 > 1$. Show that the NP test rejects for large $|X|$, and compute the threshold.
7. Argue, structurally, why the Neyman–Pearson Lemma is an instance of Lagrangian optimization: identify the objective, the constraint, and the role of the multiplier k, and explain why the complementary slackness condition corresponds to the randomization on the boundary $\{p_1 = kp_0\}$.

Uniformly Most Powerful Tests

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.3

This node extends the Neyman–Pearson optimality result from simple to composite hypotheses. When the parametric family possesses a monotone likelihood ratio, a single test can be simultaneously most powerful against every point alternative on one side, yielding a uniformly most powerful (UMP) test. The key structural result is the Karlin–Rubin theorem, which reduces composite one-sided testing to the simple-vs-simple NP framework via the MLR property. This node requires Node 5.2 (NP Lemma) and the exponential family theory of Node 2.6.

Formal Definition

UMP Test

Consider a composite testing problem:

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1.$$

A level α test ϕ^* is **uniformly most powerful (UMP)** if

$$\mathbb{E}_\theta[\phi^*(X)] \geq \mathbb{E}_\theta[\phi(X)] \quad \forall \theta \in \Theta_1$$

for every level α test ϕ.

A UMP test is simultaneously most powerful against every point in Θ_1.

Monotone Likelihood Ratio

A family $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ with densities $\{p_\theta\}$ has **monotone likelihood ratio (MLR)** in a statistic $T(X)$ if for every $\theta_1 < \theta_2$, the ratio

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$$

is a non-decreasing function of $T(x)$ (on the set where $p_{\theta_1}(x) > 0$).

Intuition

The NP Lemma gives the best test for a single point alternative. The challenge with composite alternatives is that different point alternatives might require different critical regions.

The MLR property resolves this conflict: it guarantees that the likelihood ratio between any two parameter values is monotone in the same statistic T. Consequently, the NP test for every pair (θ_0, θ_1) with $\theta_1 > \theta_0$ takes the same form — reject when T is large. A single threshold test on T is therefore optimal against the entire composite alternative simultaneously.

The MLR condition is a coherence requirement: as θ increases, the distribution of T shifts in a consistent direction.

Structural Role

UMP tests occupy a narrow but important niche: they exist for one-sided hypotheses in MLR families, which includes all one-parameter exponential families with increasing natural parameter.

For two-sided alternatives ($H_1 : \theta \neq \theta_0$), UMP tests generally do not exist. This non-existence motivates the shift to: - Unbiased UMP (UMPU) tests for two-sided problems in exponential families. - Likelihood ratio tests (Node 5.4) as a general-purpose fallback.

The Karlin–Rubin theorem connects the exponential family structure (Node 2.6) to the testing framework (Node 5.1) through the NP optimality criterion (Node 5.2).

Mathematical Development

Theorem – Karlin–Rubin

Suppose the family $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ has MLR in $T(X)$. For testing

$$H_0 : \theta \leq \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0,$$

the test

$$\phi^*(x) = \begin{cases} 1 & T(x) > c, \\ \gamma & T(x) = c, \\ 0 & T(x) < c, \end{cases}$$

where c and γ are chosen so that $\mathbb{E}_{\theta_0}[\phi^*(X)] = \alpha$, is UMP level α.

Proof

Step 1: Most powerful at each $\theta_1 > \theta_0$.

Fix any $\theta_1 > \theta_0$. By the Neyman–Pearson Lemma, the most powerful level α test of $H_0' : \theta = \theta_0$ vs $H_1' : \theta = \theta_1$ rejects for large values of

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}.$$

By the MLR property, this ratio is non-decreasing in $T(x)$. Therefore the NP critical region $\{p_{\theta_1}/p_{\theta_0} > k\}$ is equivalent to $\{T(x) > c'\}$ for some threshold c'.

The level constraint $\mathbb{E}_{\theta_0}[\phi] = \alpha$ uniquely determines the threshold (and randomization constant) when the distribution of T under θ_0 is specified. Since this threshold depends only on θ_0 and α, not on θ_1, the same test ϕ^* is most powerful against every $\theta_1 > \theta_0$.

Step 2: Level α over all of Θ_0.

We must verify $\mathbb{E}_\theta[\phi^*(X)] \leq \alpha$ for all $\theta \leq \theta_0$, not just at θ_0.

The power function $\beta(\theta) = \mathbb{E}_\theta[\phi^*(X)]$ is non-decreasing (proved below). Since $\beta(\theta_0) = \alpha$, we have $\beta(\theta) \leq \alpha$ for all $\theta \leq \theta_0$. $\square$

Proposition: Monotonicity of the Power Function

If $\{P_\theta\}$ has MLR in T and ϕ^* is a threshold test on T, then $\beta(\theta) = \mathbb{E}_\theta[\phi^*(X)]$ is non-decreasing in θ.

Proof. Let $\theta_1 < \theta_2$. The test ϕ^* is the NP test of P_{θ_1} vs P_{θ_2} at some level $\beta(\theta_1)$. By the NP corollary (the NP test is unbiased), $\beta(\theta_2) \geq \beta(\theta_1)$. □

MLR in Exponential Families

For a one-parameter exponential family with canonical density

$$p_\theta(x) = h(x) \exp\{\eta(\theta)T(x) - A(\theta)\},$$

if $\eta(\theta)$ is strictly increasing in θ, then for $\theta_1 < \theta_2$:

$$\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} = \exp\{[\eta(\theta_2) - \eta(\theta_1)]T(x) - [A(\theta_2) - A(\theta_1)]\},$$

which is strictly increasing in $T(x)$ since $\eta(\theta_2) - \eta(\theta_1) > 0$. Thus exponential families with increasing natural parameter have MLR in the natural sufficient statistic T.

Non-Existence of UMP for Two-Sided Alternatives

Proposition. For a one-parameter exponential family, there is no UMP level α test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ (when $0 < \alpha < 1$).

Proof sketch. The UMP test of $\theta = \theta_0$ vs $\theta > \theta_0$ rejects for large T. The UMP test of $\theta = \theta_0$ vs $\theta < \theta_0$ rejects for small T. These are different tests, so no single test is simultaneously most powerful against alternatives on both sides. □

Unbiased UMP (UMPU) Tests

For two-sided problems in exponential families, one restricts to unbiased tests and seeks the UMP within that class. A level α test is **unbiased** if $\beta(\theta) \geq \alpha$ for all $\theta \in \Theta_1$.

For testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ in a one-parameter exponential family, the UMPU test rejects when $T \leq c_1$ or $T \geq c_2$, with randomization at the boundaries, where c_1, c_2 are determined by

$$\mathbb{E}_{\theta_0}[\phi(X)] = \alpha, \qquad \mathbb{E}_{\theta_0}[T(X)\phi(X)] = \alpha\, \mathbb{E}_{\theta_0}[T(X)].$$

Worked Examples

Example 1: Normal Mean (Known Variance)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, σ^2 known. Test $H_0 : \mu \leq \mu_0$ vs $H_1 : \mu > \mu_0$.

This is a one-parameter exponential family with $T = \sum X_i$ and natural parameter $\eta = \mu/\sigma^2$, which is increasing in μ. By the Karlin–Rubin theorem, the UMP level α test rejects when $\bar{X} > c$.

Under μ_0: $\bar{X} \sim N(\mu_0, \sigma^2/n)$, so $c = \mu_0 + z_\alpha \sigma/\sqrt{n}$.

Power function:

$$\beta(\mu) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).$$

This is strictly increasing in μ, confirming monotonicity.

Example 2: Bernoulli Proportion

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. Test $H_0 : p \leq 1/2$ vs $H_1 : p > 1/2$ at level $\alpha = 0.05$, with $n = 20$.

The sufficient statistic is $T = \sum X_i \sim \text{Binomial}(20, p)$. The Bernoulli family has MLR in T since $\eta(p) = \log(p/(1-p))$ is strictly increasing.

Under $p = 1/2$: - $P_{1/2}(T \geq 15) = \sum_{k=15}^{20} \binom{20}{k}(1/2)^{20} \approx 0.0207$. - $P_{1/2}(T \geq 14) \approx 0.0577$.

Non-randomized: reject when $T \geq 15$, with size $0.0207 < 0.05$.

Randomized UMP test: $\phi(x) = 1$ for $T \geq 15$, $\phi(x) = \gamma$ for $T = 14$, where

$$\gamma = \frac{0.05 - 0.0207}{P_{1/2}(T = 14)} = \frac{0.0293}{0.0370} \approx 0.792.$$

Power at $p = 0.7$: $\beta(0.7) = P_{0.7}(T \geq 15) + 0.792 \cdot P_{0.7}(T = 14) \approx 0.584 + 0.792 \cdot 0.192 \approx 0.736$.

Cross-Links

- **AI:** Monotone decision boundaries in UMP tests correspond to threshold classifiers on a score function. In calibrated binary classifiers under log-loss, the optimal decision rule is monotone in the predicted probability.
- **Economics:** One-sided tests arise in policy evaluation (e.g., testing whether a program has a positive effect). The UMP property ensures no other level-α test has higher power to detect the posited direction of effect.
- **Linear Algebra:** The sufficient statistic T is a linear functional of the data in exponential families; the UMP critical region is a half-space in the sufficient statistic space.
- **Quantum:** For quantum state families with a monotone structure (e.g., thermal states parameterized by temperature), the optimal measurement for one-sided discrimination is the quantum analogue of the UMP test.

Structural Compression

Invariant: MLR reduces the composite testing problem to a sequence of simple-vs-simple problems all solved by the same critical region.

Mechanism: The statistic T orders the family monotonically; the threshold test on T is simultaneously Neyman–Pearson optimal against every point alternative in H_1, because the NP critical region is the same for every $\theta_1 \in \Theta_1$.

Cross-System Echo: The monotone likelihood ratio is a stochastic ordering condition: T under P_{θ_2} is stochastically larger than under P_{θ_1} when $\theta_2 > \theta_1$. This ordering principle recurs in reliability theory (increasing failure rate families), decision theory (monotone decision procedures), and auction theory (revenue monotonicity under regular distributions).

Exercises

1. Show that the Bernoulli(θ) family has MLR in $T = \sum_{i=1}^{n} X_i$, and derive the UMP test for $H_0 : \theta \leq 1/2$ vs $H_1 : \theta > 1/2$ at level $\alpha = 0.10$ with $n = 10$.
2. Prove that the power function of the Karlin–Rubin test is non-decreasing in θ, using only the NP Lemma and the MLR property.

3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Verify the MLR property in $T = \sum X_i$ and find the UMP test of $H_0 : \lambda \leq 1$ vs $H_1 : \lambda > 1$ for $n = 10$ and $\alpha = 0.05$.

4. Prove that for a one-parameter exponential family with strictly increasing $\eta(\theta)$, there is no UMP test of $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

5. For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha_0, \beta)$ with α_0 known, show that the family has MLR in $\sum X_i$ and state the UMP test for $H_0 : \beta \leq \beta_0$ vs $H_1 : \beta > \beta_0$.

6. Construct the UMPU test of $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ for $X \sim N(\mu, 1)$ (single observation). Verify that it coincides with the two-sided z-test.

7. Argue, structurally, why the MLR property is the precise condition under which the NP lemma's pointwise optimality lifts to uniform optimality, and explain what breaks for families that fail to satisfy MLR (connecting to the non-existence result for two-sided alternatives).

Likelihood Ratio Tests

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.4

The likelihood ratio (LR) test is the general-purpose construction for composite hypothesis testing. Where Neyman–Pearson handles simple-vs-simple and UMP handles one-sided MLR families, LR tests apply to arbitrary parametric constraints without requiring special structural conditions. Wilks' theorem provides the asymptotic reference distribution, making LR tests practically computable in large samples. This node requires the testing framework (Node 5.1), density and likelihood theory (Node 2.3), and MLE theory (Node 3.2).

Formal Definition

Generalized Likelihood Ratio Statistic

Let $X_1, \ldots, X_n \sim p_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Consider

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta \setminus \Theta_0,$$

where $\Theta_0 \subset \Theta$ is defined by r smooth equality constraints, so $\dim(\Theta_0) = k - r$.

The **generalized likelihood ratio statistic** is

$$\Lambda_n = \frac{\sup_{\theta \in \Theta_0} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta; X)} = \frac{L(\hat{\theta}_0; X)}{L(\hat{\theta}_n; X)},$$

where $\hat{\theta}_n$ is the unrestricted MLE and $\hat{\theta}_0$ is the MLE restricted to Θ_0.

Note $\Lambda_n \in (0, 1]$; the test rejects H_0 for small values of Λ_n.

Log-Likelihood Form

The standard form is

$$-2 \log \Lambda_n = 2 \left[\ell(\hat{\theta}_n) - \ell(\hat{\theta}_0) \right],$$

which is non-negative. The test rejects H_0 for large values of $-2 \log \Lambda_n$.

Intuition

The LR statistic measures how much the data favors the unrestricted model over the restricted model. If the constraint $\theta \in \Theta_0$ costs little in likelihood, Λ_n is near 1 and there is no evidence against H_0. If the constraint forces a substantial loss in fit, Λ_n is small and the data contradicts H_0.

The factor of -2 in the log transform is not arbitrary: it is chosen so that the asymptotic distribution under H_0 is exactly chi-squared, without additional scaling. This normalization arises from the quadratic approximation to the log-likelihood near its maximum, with the Fisher information providing the natural metric.

Structural Role

LR tests provide a universal construction for composite hypotheses. Their advantages:

1. **Generality.** Applicable to any parametric model with smooth constraints.
2. **Invariance.** The LR statistic is invariant under reparameterization of θ.
3. **Asymptotic universality.** Wilks' theorem gives a single reference distribution (χ^2_r) regardless of the specific model.

The LR test, Wald test, and Score (Rao) test form the "holy trinity" of asymptotic tests (Node 4.4). All three converge to the same first-order asymptotic limit, but they differ in finite samples and in computational convenience.

Mathematical Development

Theorem – Wilks (1938)

Regularity conditions. Assume: (i) Θ is an open subset of $\mathbb{R}^k$; (ii) $\Theta_0 = \{\theta \in \Theta : g(\theta) = 0\}$ where $g : \mathbb{R}^k \to \mathbb{R}^r$ is continuously differentiable with rank r Jacobian; (iii) the log-likelihood is three times continuously differentiable with non-singular Fisher information $I(\theta)$; (iv) the MLE $\hat{\theta}_n$ is consistent and asymptotically normal.

Statement. Under H_0 (with true parameter $\theta_0 \in \Theta_0$),

$$-2\log\Lambda_n \xrightarrow{d} \chi^2_r \quad \text{as } n \to \infty,$$

where $r = \dim(\Theta) - \dim(\Theta_0)$.

Proof Sketch

Step 1: Taylor expansion. Expand $\ell(\theta)$ around $\hat{\theta}_n$ to second order:

$$\ell(\theta) \approx \ell(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \left[-\nabla^2 \ell(\hat{\theta}_n)\right](\theta - \hat{\theta}_n).$$

Define the observed information $\hat{J}_n = -\nabla^2\ell(\hat{\theta}_n)/n$. By the law of large numbers, $\hat{J}_n \xrightarrow{p} I(\theta_0)$.

Step 2: Quadratic approximation. Using the expansion at the restricted MLE:

$$-2\log\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\hat{\theta}_0)] \approx n(\hat{\theta}_n - \hat{\theta}_0)^\top \hat{J}_n (\hat{\theta}_n - \hat{\theta}_0).$$

Step 3: Asymptotic normality. Under H_0, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$.

Step 4: Constrained projection. The restricted MLE $\hat{\theta}_0$ is the projection of $\hat{\theta}_n$ onto Θ_0 in the metric defined by $I(\theta_0)$. The difference $\hat{\theta}_n - \hat{\theta}_0$ lies in the r-dimensional normal space to Θ_0 at θ_0.

Step 5: Chi-squared conclusion. The quadratic form $n(\hat{\theta}_n - \hat{\theta}_0)^\top I(\theta_0)(\hat{\theta}_n - \hat{\theta}_0)$ is asymptotically a sum of squares of r independent standard normals, hence χ^2_r. □

Level α LR Test

Reject H_0 when

$$-2\log\Lambda_n > \chi^2_{r,\alpha},$$

where $\chi^2_{r,\alpha}$ is the upper α quantile of χ^2_r. This test has asymptotic level α by Wilks' theorem.

Relationship to Wald and Score Tests

Under the same regularity conditions:

Wald statistic:

$$W_n = n(\hat{\theta}_n - \theta_0)^\top \hat{I}_n(\hat{\theta}_n - \theta_0) \xrightarrow{d} \chi^2_r.$$

Score (Rao) statistic:

$$S_n = \frac{1}{n}\nabla\ell(\hat{\theta}_0)^\top I(\hat{\theta}_0)^{-1}\nabla\ell(\hat{\theta}_0) \xrightarrow{d} \chi^2_r.$$

All three statistics satisfy $W_n = -2\log\Lambda_n + o_p(1) = S_n + o_p(1)$ under H_0, so they are first-order asymptotically equivalent.

Bartlett Correction

The LR statistic admits a Bartlett correction: there exists a constant c (depending on the model) such that

$$(1 - c/n)\cdot(-2\log\Lambda_n)$$

has a χ^2_r distribution accurate to order $O(n^{-2})$, improving finite-sample performance. This correction is specific to the LR statistic and does not apply to the Wald or Score statistics.

Worked Examples

Example 1: Testing a Normal Mean (Known Variance)

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$. Test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

Here $k = 1$, $r = 1$, $\hat{\mu}_n = \bar{X}$, $\hat{\mu}_0 = \mu_0$.

$$\ell(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum(X_i - \mu)^2.$$

$$-2\log\Lambda_n = 2[\ell(\bar{X}) - \ell(\mu_0)] = n(\bar{X} - \mu_0)^2.$$

Under H_0, $\sqrt{n}(\bar{X} - \mu_0) \sim N(0,1)$, so $-2\log\Lambda_n \sim \chi^2_1$ exactly.

Reject when $n(\bar{X} - \mu_0)^2 > \chi^2_{1,\alpha} = z^2_{\alpha/2}$, which is equivalent to the two-sided z-test $|\bar{X} - \mu_0| > z_{\alpha/2}/\sqrt{n}$.

Example 2: Testing Equality of Two Normal Means

Let $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_2, \sigma^2)$, σ^2 known. Test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.

Parameters: $\theta = (\mu_1, \mu_2) \in \mathbb{R}^2$, constraint $g(\theta) = \mu_1 - \mu_2 = 0$, so $r = 1$.

Unrestricted MLEs: $\hat{\mu}_1 = \bar{X}$, $\hat{\mu}_2 = \bar{Y}$. Restricted MLE: $\hat{\mu}_0 = (m\bar{X} + n\bar{Y})/(m+n)$.

$$-2\log\Lambda_n = \frac{mn}{(m+n)\sigma^2}(\bar{X} - \bar{Y})^2.$$

Under H_0, $\bar{X} - \bar{Y} \sim N(0, \sigma^2(1/m + 1/n))$, so $-2\log\Lambda_n \sim \chi^2_1$ exactly.

Example 3: Testing Independence in a Multinomial

Consider a 2×2 contingency table with cell counts $(n_{11}, n_{12}, n_{21}, n_{22})$ from a multinomial with $N = \sum n_{ij}$ and cell probabilities p_{ij}. Test $H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ (independence).

Here $k = 3$ (full multinomial has 3 free parameters), $\dim(\Theta_0) = 2$ (row and column marginals determine Θ_0), so $r = 1$.

The LR statistic is

$$-2\log\Lambda_n = 2\sum_{i,j} n_{ij}\log\frac{n_{ij}}{\hat{e}_{ij}},$$

where $\hat{e}_{ij} = n_{i\cdot}n_{\cdot j}/N$ are the expected counts under independence.

Under H_0, $-2\log\Lambda_n \xrightarrow{d} \chi^2_1$ as $N \to \infty$.

Cross-Links

- **AI:** Model comparison in machine learning (e.g., nested neural network architectures, feature selection) uses likelihood ratio principles. The deviance in generalized linear models is exactly $-2\log\Lambda_n$.
- **Economics:** LR tests are standard for testing restrictions in econometric models (e.g., testing coefficient restrictions in regression, testing overidentifying restrictions in GMM).
- **Linear Algebra:** The quadratic form $n(\hat{\theta}_n - \hat{\theta}_0)^\top I(\theta_0)(\hat{\theta}_n - \hat{\theta}_0)$ is a Mahalanobis distance; its chi-squared distribution follows from the spectral decomposition of the information matrix.
- **Quantum:** Quantum likelihood ratio tests for composite hypotheses on quantum states follow the same structure; the quantum Fisher information matrix replaces the classical one, and the SLD (symmetric logarithmic derivative) plays the role of the score.

Structural Compression

Invariant: The LR statistic compresses all information about the constraint $\theta \in \Theta_0$ into a single scalar via the log-likelihood difference.

Mechanism: Second-order Taylor expansion of the log-likelihood under regularity reduces $-2\log\Lambda_n$ to a chi-squared quadratic form in r asymptotically normal directions, with r equal to the codimension of the constraint.

Cross-System Echo: The chi-squared reference distribution is shared by Wald and Score statistics; all three converge to the same first-order asymptotic limit. The $-2\log\Lambda$ form reappears in information criteria (AIC $= -2\ell + 2k$ penalizes by model dimension), in deviance analysis for GLMs, and in the quantum Chernoff bound for asymptotic state discrimination.

Exercises

1. Derive the LR statistic for testing $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$ in the model $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with μ known. Verify the asymptotic chi-squared distribution.

2. Show that the LR statistic is invariant under reparameterization: if $\eta = g(\theta)$ is a diffeomorphism, the LR statistic computed in the η-parameterization equals the one in the θ-parameterization.

3. For the multinomial goodness-of-fit test ($H_0 : p = p_0$ with p_0 fully specified), compute $-2 \log \Lambda_n$ and verify that $r = k - 1$ where k is the number of categories.

4. Prove that for a one-parameter model with simple null $H_0 : \theta = \theta_0$, the LR statistic equals the square of the signed likelihood root: $-2 \log \Lambda_n = [\text{sign}(\hat{\theta}_n - \theta_0)\sqrt{-2 \log \Lambda_n}]^2$, and the signed root is asymptotically $N(0, 1)$.

5. Compute the LR test for $H_0 : \lambda_1 = \lambda_2$ vs $H_1 : \lambda_1 \neq \lambda_2$ given independent samples $X_1, \ldots, X_m \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_1)$ and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda_2)$.

6. State the Bartlett correction for the LR test of $H_0 : \mu = 0$ in a $N(\mu, 1)$ model and verify that the corrected statistic has improved distributional accuracy.

7. Argue, structurally, why the LR test is the natural successor to the NP lemma for composite hypotheses: identify what structural property is lost when moving from simple to composite testing, and explain how the asymptotic chi-squared approximation substitutes for the exact NP optimality.

p-Values

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.5

The p-value is the continuous calibration of a test statistic against the null distribution. It encodes an entire family of hypothesis tests into a single random variable, connecting the binary decision framework of Node 5.1 to the continuous probability scale. This node requires the testing framework (Node 5.1) and the NP theory (Node 5.2) that motivates specific test statistics. It directly enables the test–confidence duality (Node 5.6) through the relationship between p-values and confidence set membership.

Formal Definition

p-Value

Let $T(X)$ be a test statistic for $H_0 : \theta \in \Theta_0$, with large values of T constituting evidence against H_0. The **p-value** is the statistic

$$p(X) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big(T(X') \geq T(X)\big),$$

where $X' \sim P_\theta$ is an independent copy drawn under H_0.

The p-value is the probability, under the least favorable null distribution, of observing a test statistic at least as extreme as the observed value.

Valid p-Value

A statistic $p(X)$ taking values in $[0, 1]$ is a **valid p-value** for H_0 if

$$\mathbb{P}_\theta\big(p(X) \leq u\big) \leq u \quad \forall u \in [0, 1],\ \forall \theta \in \Theta_0.$$

This is precisely the condition that $p(X)$ is stochastically no smaller than Uniform$(0, 1)$ under every null distribution.

Intuition

The p-value answers: "If the null hypothesis were true, how surprising is the observed data?" A small p-value means the observed test statistic is in the tail of the null distribution, providing evidence against H_0.

Crucially, the p-value is NOT: - The probability that H_0 is true. - The probability that the data arose by chance. - A measure of effect size or practical significance.

The p-value is a frequentist object: it measures the extremity of the observed statistic relative to the null reference distribution. Its uniform distribution under H_0 is the key structural property that makes it a valid calibration device.

Structural Role

The p-value occupies a central position in applied statistics as the standard summary of evidence against H_0. It mediates between:

- The test statistic $T(X)$, which depends on the specific model and test construction.
- The binary decision (reject/not reject), which depends on the chosen level α.

By reporting the p-value rather than just the binary decision, the analyst communicates the strength of evidence on a universal scale.

The p-value connects directly to confidence sets (Node 5.6): the set of parameter values for which the p-value exceeds α forms a $(1-\alpha)$ confidence set.

Mathematical Development

Uniformity Under Simple Null

Theorem. If $H_0 : P = P_0$ is simple and $T(X)$ has a continuous distribution under P_0, then

$$p(X) = 1 - F_0(T(X)) \sim \text{Uniform}(0,1) \quad \text{under } P_0,$$

where F_0 is the CDF of T under P_0.

Proof. By the probability integral transform, if T has continuous CDF F_0, then $U = F_0(T) \sim \text{Uniform}(0,1)$ under P_0. Therefore $p(X) = 1 - U \sim \text{Uniform}(0,1)$. □

Validity Under Composite Null

Theorem. For a composite null, the p-value defined via the supremum is valid:

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(X) \leq u) \leq u \quad \forall u \in [0,1].$$

Proof. For any $\theta \in \Theta_0$:

$$\mathbb{P}_\theta(p(X) \leq u) = \mathbb{P}_\theta\left(\sup_{\theta' \in \Theta_0} \mathbb{P}_{\theta'}(T(X') \geq T(X)) \leq u\right) \leq \mathbb{P}_\theta(\mathbb{P}_\theta(T(X') \geq T(X)) \leq u)\,.$$

If T has a continuous distribution under θ, the inner probability is $1 - F_\theta(T(X))$, and by the probability integral transform, $\mathbb{P}_\theta(1 - F_\theta(T(X)) \leq u) = u$. For non-continuous T, the inequality $\leq u$ holds by the properties of CDFs. □

Connection to Level α Tests

Proposition. Rejecting H_0 when $p(X) \leq \alpha$ defines a level α test for every $\alpha \in (0,1]$ simultaneously.

Proof. By validity, $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(X) \leq \alpha) \leq \alpha$. □

The p-value therefore carries strictly more information than any single fixed-α binary decision: it encodes the entire family of level-α tests as α ranges over $(0,1]$.

Relationship to Confidence Sets

For each $\theta_0 \in \Theta$, let $p_{\theta_0}(X)$ denote the p-value for testing $H_0 : \theta = \theta_0$. Define

$$C_\alpha(X) = \{\theta_0 \in \Theta : p_{\theta_0}(X) > \alpha\}.$$

Then $C_\alpha(X)$ is a $(1-\alpha)$ confidence set. This is the test inversion construction formalized in Node 5.6.

Lindley's Paradox (Bayesian Interpretation)

The frequentist p-value can conflict sharply with Bayesian posterior probabilities.

Setup. Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X \sim N(\mu, 1/n)$. Suppose the prior assigns probability $\pi_0 = 1/2$ to H_0 and, under H_1, $\mu \sim N(0, \tau^2)$.

The Bayes factor in favor of H_0 is

$$\mathrm{BF}_{01} = \sqrt{1 + n\tau^2} \cdot \exp\left(-\frac{n\bar{x}^2}{2} \cdot \frac{n\tau^2}{1 + n\tau^2}\right).$$

For fixed $\bar{x} = z_\alpha/\sqrt{n}$ (i.e., the p-value is exactly α), as $n \to \infty$: the frequentist p-value remains constant at α, but $\mathrm{BF}_{01} \to \infty$, meaning the Bayesian posterior increasingly favors H_0.

This is **Lindley's paradox**: a fixed p-value becomes increasingly compatible with H_0 from the Bayesian perspective as sample size grows. The paradox arises because the p-value calibrates against the null alone, while the Bayes factor accounts for the prior predictive distribution under both hypotheses.

Multiple Testing

When m hypotheses $H_{0,1}, \ldots, H_{0,m}$ are tested simultaneously with p-values $p_1, \ldots, p_m$, the probability of at least one false rejection (familywise error rate, FWER) can far exceed α.

Bonferroni correction. Reject $H_{0,i}$ if $p_i \leq \alpha/m$. Then FWER $\leq \alpha$ by the union bound. This is conservative: it controls FWER but sacrifices power.

Benjamini–Hochberg (BH) procedure. Controls the **false discovery rate (FDR)** $= \mathbb{E}[V/R]$ where V = number of false rejections, R = total rejections ($V/R := 0$ if $R = 0$).

Algorithm: 1. Order p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$. 2. Find the largest k such that $p_{(k)} \leq k\alpha/m$. 3. Reject all $H_{0,(i)}$ for $i \leq k$.

Theorem (Benjamini–Hochberg, 1995). If the p-values corresponding to the true nulls are independent, then the BH procedure controls FDR $\leq \alpha \cdot m_0/m \leq \alpha$, where m_0 is the number of true nulls.

Worked Examples

Example 1: z-Test p-Value

Let $X_1, \ldots, X_{25} \overset{\text{iid}}{\sim} N(\mu, 4)$. Test $H_0 : \mu = 10$ vs $H_1 : \mu > 10$. Observed $\bar{x} = 10.92$.

Test statistic: $Z = \frac{\bar{x}-10}{2/\sqrt{25}} = \frac{0.92}{0.4} = 2.30$.

p-value: $p = P(Z' \geq 2.30) = 1 - \Phi(2.30) = 0.0107$.

At $\alpha = 0.05$: reject H_0. At $\alpha = 0.01$: do not reject. The p-value captures both decisions simultaneously.

Example 2: Multiple Testing with BH

A genomics study tests $m = 1000$ genes. Suppose 50 genes have true effects. Observed p-values include the ordered values $p_{(1)} = 0.00003$, $p_{(2)} = 0.00012$, ...

Apply BH at FDR level $\alpha = 0.05$. The BH threshold for the k-th ordered p-value is $k \cdot 0.05/1000 = k/20000$.

- $p_{(1)} = 0.00003 \leq 1/20000 = 0.00005$: yes.
- $p_{(2)} = 0.00012 \leq 2/20000 = 0.0001$: no.

So only the first gene is rejected at this threshold. In contrast, Bonferroni at $\alpha = 0.05$: threshold $= 0.05/1000 = 0.00005$, rejecting the same gene.

Now suppose $p_{(2)} = 0.00008$. Then $0.00008 \leq 0.0001$: yes. BH rejects both, while Bonferroni (threshold 0.00005) rejects only the first. This illustrates BH's greater power under many true alternatives.

Example 3: Discrete p-Value (Binomial)

Let $X \sim \text{Binomial}(10, p)$. Test $H_0 : p = 0.5$ vs $H_1 : p > 0.5$. Observed $X = 8$.

$$p\text{-value} = P_{0.5}(X \geq 8) = \sum_{k=8}^{10} \binom{10}{k} (0.5)^{10} = \frac{45 + 10 + 1}{1024} = \frac{56}{1024} \approx 0.0547.$$

At $\alpha = 0.05$: do not reject (barely). Note that the p-value is conservative (super-uniform) due to discreteness: $P_{0.5}(p(X) \leq u) \leq u$ with strict inequality for most u.

Cross-Links

- **AI:** p-values are used in A/B testing for model selection and feature evaluation. The multiple testing problem arises in neural architecture search, where many configurations are compared simultaneously.
- **Economics:** p-values are the standard reporting currency in empirical economics. The "p-hacking" debate (selective reporting of significant results) has driven reforms toward pre-registration and FDR control.
- **Linear Algebra:** For Gaussian test statistics, the p-value is a monotone function of the Mahalanobis distance from the null, connecting null distribution geometry to the probability scale.
- **Quantum:** In quantum state certification, a p-value can be defined for the null hypothesis that a prepared state equals a target state, using fidelity-based test statistics.

Structural Compression

Invariant: Under H_0, $p(X)$ is stochastically no smaller than $\text{Uniform}(0, 1)$; under a simple null with continuous T, it is exactly uniform.

Mechanism: The p-value is the survival function of T under P_0 evaluated at the observed $T(X)$; uniformity follows from the probability integral transform. The supremum over Θ_0 ensures validity for composite nulls.

Cross-System Echo: The p-value transforms a problem-specific test statistic into a universal probability scale, analogous to how quantile normalization maps arbitrary distributions to a reference distribution. In information theory, the p-value under the null is the "surprise" measured in probability units rather than bits.

Exercises

1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with σ^2 known. Derive the p-value for the two-sided test $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ and verify it is Uniform$(0, 1)$ under H_0.
2. Prove that if $p(X)$ is a valid p-value and $g : [0, 1] \to [0, 1]$ is non-decreasing with $g(u) \leq u$ for all u, then $g(p(X))$ is also a valid p-value. Give an example of such a g.
3. Compute the p-value for the LR test of $H_0 : \lambda = 1$ vs $H_1 : \lambda \neq 1$ given $X_1, \dots, X_{20} \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$ with $\sum x_i = 28$, using both the exact distribution and the chi-squared approximation.
4. Demonstrate Lindley's paradox numerically: for $n = 100, 1000, 10000$, compute the sample mean $\bar{x}$ that gives a p-value of exactly 0.05, and compute the Bayes factor BF_{01} with prior $\mu \sim N(0, 1)$ under H_1.
5. Apply the Bonferroni and Benjamini–Hochberg procedures to the p-values $\{0.001, 0.013, 0.029, 0.032, 0.15, 0.45, 0.78\}$ at level $\alpha = 0.05$. Identify which hypotheses are rejected under each method.
6. Prove that for a discrete test statistic, the p-value is super-uniform: $\mathbb{P}_{\theta_0}(p(X) \leq u) \leq u$ with strict inequality for some u. Explain why randomized p-values can restore exact uniformity.
7. Argue, structurally, why the p-value is sufficient for performing a level-α test at any α, and explain how this relates to the p-value being a "minimal sufficient statistic" for the family of level-α rejection decisions indexed by $\alpha \in (0, 1]$.

Duality: Tests and Confidence Sets

Position in Knowledge Graph

System: Statistics

Structural Domain: Hypothesis Testing

Node: 5.6

This node formalizes the exact correspondence between hypothesis tests and confidence sets, unifying the two principal branches of frequentist inference. The duality is not merely an analogy: it is a bijection between families of level-α tests and level-$(1-\alpha)$ confidence sets. This result requires the testing framework (Node 5.1), estimation and confidence set theory (Nodes 4.1, 4.4), and draws on p-values (Node 5.5) for the continuous version of the correspondence. It serves as the capstone of Module 5, connecting testing to the broader inferential landscape.

Formal Definition

Setting

Let $X \sim P_\theta$, $\theta \in \Theta$. For each $\theta_0 \in \Theta$, suppose $\phi_{\theta_0}(X)$ is a level α test of

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0.$$

Test Inversion

The **test-inversion confidence set** is the set-valued function

$$C(X) = \big\{\theta_0 \in \Theta : \phi_{\theta_0}(X) = 0\big\},$$

the set of all parameter values not rejected by their respective level α tests.

Confidence Set Inversion

Given a confidence set $C(X)$ with level $1-\alpha$, define for each θ_0:

$$\phi_{\theta_0}(X) = \mathbf{1}\big(\theta_0 \notin C(X)\big).$$

Intuition

A confidence set collects all parameter values that are "compatible" with the observed data. A hypothesis test declares a specific parameter value "incompatible" when the data falls in the rejection region.

The duality states: a parameter value is in the confidence set if and only if the test at that value does not reject. This is not a coincidence of definitions — it is a structural identity. The coverage probability of the confidence set equals one minus the size of the test, because both quantities measure the same event: the true parameter is not falsely excluded.

The practical consequence is that any method for constructing tests immediately yields a method for constructing confidence sets, and vice versa. UMP tests yield shortest confidence sets; LR tests yield likelihood-based confidence regions; Wald tests yield Wald intervals.

Structural Role

Test–confidence duality unifies the two principal branches of frequentist inference. Confidence sets are not merely interval estimators; they are the geometric encoding of an entire family of hypothesis tests.

This duality explains why Wald, Score, and LR confidence intervals (Node 4.4) correspond exactly to inverting the Wald, Score, and LR test statistics: each test generates its confidence set by accumulating accepted parameter values.

The duality also connects to p-values (Node 5.5): the confidence set at level $1-\alpha$ consists of all θ_0 with p-value exceeding α.

Mathematical Development

Theorem – Test Inversion Produces a Confidence Set

Statement. If ϕ_{θ_0} is a level α test of $H_0 : \theta = \theta_0$ for each $\theta_0 \in \Theta$, then

$$C(X) = \{\theta_0 \in \Theta : \phi_{\theta_0}(X) = 0\}$$

satisfies

$$\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha \quad \forall \theta \in \Theta.$$

Proof. Fix any $\theta \in \Theta$. Then

$$\mathbb{P}_\theta(\theta \in C(X)) = \mathbb{P}_\theta(\phi_\theta(X) = 0) = 1 - \mathbb{E}_\theta[\phi_\theta(X)].$$

Since ϕ_θ is a level α test of $H_0 : \theta = \theta$ (the true parameter), we have $\mathbb{E}_\theta[\phi_\theta(X)] \leq \alpha$. Therefore $\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha$. $\square$

Theorem – Confidence Set Inversion Produces a Family of Tests

Statement. If $C(X)$ satisfies $\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha$ for all θ, then for each θ_0,

$$\phi_{\theta_0}(X) = \mathbf{1}(\theta_0 \notin C(X))$$

is a level α test of $H_0 : \theta = \theta_0$.

Proof. For any $\theta_0 \in \Theta$:

$$\mathbb{E}_{\theta_0}[\phi_{\theta_0}(X)] = \mathbb{P}_{\theta_0}(\theta_0 \notin C(X)) = 1 - \mathbb{P}_{\theta_0}(\theta_0 \in C(X)) \leq 1 - (1 - \alpha) = \alpha. \quad \square$$

The Bijection

The two constructions are inverses. Starting from a family of tests $\{\phi_{\theta_0}\}$, forming $C(X)$, and then recovering tests via $\tilde{\phi}_{\theta_0}(X) = \mathbf{1}(\theta_0 \notin C(X))$ yields $\tilde{\phi}_{\theta_0} = 1 - (1 - \phi_{\theta_0}) = \phi_{\theta_0}$ for non-randomized tests.

For randomized tests, the correspondence generalizes: define $C_\alpha(X) = \{\theta_0 : \phi_{\theta_0}(X) < 1\}$ and the randomized inclusion event $\mathbb{P}(\theta_0 \in C(X)|X) = 1 - \phi_{\theta_0}(X)$.

Property Transfer Table

The duality transfers structural properties between tests and confidence sets:

Test Property	Confidence Set Property
Level α	Coverage $\geq 1-\alpha$
Exact level α	Exact coverage $1-\alpha$
Unbiased test	Unbiased confidence set
UMP test	Uniformly shortest confidence set
LR test	Likelihood-based confidence region
Equivariant test	Equivariant confidence set

UMP Tests Yield Shortest Confidence Sets

Proposition. If $\phi^*_{\theta_0}$ is UMP level α for each θ_0, then $C^*(X) = \{\theta_0 : \phi^*_{\theta_0}(X) = 0\}$ has the smallest expected Lebesgue measure among all level $(1-\alpha)$ confidence sets derived from level α tests.

Proof sketch. Let $C(X)$ be any other $(1-\alpha)$ confidence set derived from level α tests $\{\phi_{\theta_0}\}$. The expected length is $\mathbb{E}_\theta[\lambda(C(X))] = \int \mathbb{P}_\theta(\theta_0 \in C(X))\, d\theta_0 = \int (1 - \mathbb{E}_\theta[\phi_{\theta_0}(X)])\, d\theta_0$. Since $\phi^*_{\theta_0}$ is UMP, $\mathbb{E}_\theta[\phi^*_{\theta_0}(X)] \geq \mathbb{E}_\theta[\phi_{\theta_0}(X)]$ for $\theta_0 \neq \theta$, giving $\mathbb{E}_\theta[\lambda(C^*(X))] \leq \mathbb{E}_\theta[\lambda(C(X))]$. $\square$

p-Value Formulation

The duality admits a clean p-value formulation. Let $p_{\theta_0}(X)$ be the p-value for testing $H_0 : \theta = \theta_0$. Then

$$C_\alpha(X) = \{\theta_0 : p_{\theta_0}(X) > \alpha\}$$

is a $(1-\alpha)$ confidence set. The confidence set is the upper level set of the p-value function viewed as a function of θ_0.

Worked Examples

Example 1: Normal Mean Confidence Interval via Test Inversion

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, σ^2 known. For each μ_0, the level α two-sided test of $H_0 : \mu = \mu_0$ rejects when

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}.$$

The acceptance region (non-rejection) is $|\bar{X} - \mu_0| \leq z_{\alpha/2}\sigma/\sqrt{n}$, i.e., $\mu_0 \in [\bar{X} - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}]$.

Therefore

$$C(X) = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right]$$

is the standard $(1-\alpha)$ confidence interval, recovered by inverting the z-test.

Example 2: LR Confidence Region for Two Parameters

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, both unknown. The LR test of $H_0 : (\mu, \sigma^2) = (\mu_0, \sigma_0^2)$ rejects when $-2 \log \Lambda_n > \chi^2_{2,\alpha}$.

The LR confidence region is

$$C(X) = \left\{ (\mu_0, \sigma_0^2) : -2 \log \Lambda_n(\mu_0, \sigma_0^2) \leq \chi^2_{2,\alpha} \right\}.$$

Explicitly:

$$-2 \log \Lambda_n = n \log \frac{\sigma_0^2}{\hat{\sigma}^2} + \frac{n(\bar{X} - \mu_0)^2 + n\hat{\sigma}^2}{\sigma_0^2} - n,$$

where $\hat{\sigma}^2 = n^{-1} \sum (X_i - \bar{X})^2$. The confidence region is the set of (μ_0, σ_0^2) for which this expression does not exceed $\chi^2_{2,\alpha}$.

Example 3: One-Sided Test to One-Sided Confidence Bound

For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$, the UMP test of $H_0 : \mu \leq \mu_0$ at level α rejects when $\bar{X} > \mu_0 + z_\alpha / \sqrt{n}$.

Inverting: the set of μ_0 not rejected is $\mu_0 \geq \bar{X} - z_\alpha / \sqrt{n}$. This yields the one-sided $(1 - \alpha)$ confidence bound

$$\mu \geq \bar{X} - z_\alpha / \sqrt{n}$$

or equivalently, the confidence set $C(X) = [\bar{X} - z_\alpha / \sqrt{n}, \infty)$.

Cross-Links

- **AI:** In conformal prediction, prediction sets are constructed by inverting a family of tests based on nonconformity scores, directly applying the test–confidence duality in a distribution-free setting.
- **Economics:** Confidence intervals for treatment effects in causal inference are obtained by inverting tests of sharp null hypotheses (Fisher's exact test inversion), a direct application of this duality.
- **Linear Algebra:** The confidence ellipsoid $\{(\mu_0) : (\bar{X} - \mu_0)^\top \Sigma^{-1} (\bar{X} - \mu_0) \leq c\}$ is the sublevel set of a quadratic form; its boundary is the acceptance boundary of the corresponding chi-squared test.
- **Quantum:** Quantum confidence regions for state tomography are constructed by inverting likelihood ratio tests on the density matrix, with the Hilbert–Schmidt metric replacing the Euclidean metric.

Structural Compression

Invariant: Level α size control in testing and $(1 - \alpha)$ coverage in confidence sets are the same structural constraint expressed in dual coordinates.

Mechanism: The acceptance region of the test at θ_0 becomes the event $\{\theta_0 \in C(X)\}$; inclusion probability equals one minus test size. The correspondence is a bijection for non-randomized tests and extends to randomized tests via probabilistic inclusion.

Cross-System Echo: In decision theory, this duality is the frequentist instance of the general duality between loss functions and risk sets. In convex optimization, the primal (confidence set as intersection of half-spaces) and dual (test as separation hyperplane) perspectives mirror the test–confidence correspondence. In Bayesian inference, credible sets are the posterior analogue: coverage is guaranteed by posterior mass rather than sampling frequency, but the structural role is identical.

Exercises

1. Derive the $(1-\alpha)$ confidence interval for the parameter p of a Bernoulli distribution by inverting the exact binomial test. This is the Clopper–Pearson interval.
2. Prove that if $C(X)$ is a $(1-\alpha)$ confidence set and $C'(X) \subseteq C(X)$ is a strict subset satisfying $\mathbb{P}_\theta(\theta \in C'(X)) \geq 1-\alpha$ for all θ, then the family of tests derived from $C'(X)$ has power no less than the family derived from $C(X)$.
3. For $X \sim N(\mu, 1)$, construct the confidence set by inverting the one-sided UMP test of $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$. Compare the result to the two-sided confidence interval.
4. Show that the Wald confidence interval $\hat{\theta} \pm z_{\alpha/2} \cdot \mathrm{se}(\hat{\theta})$ is obtained by inverting the Wald test $|\hat{\theta} - \theta_0|/\mathrm{se}(\hat{\theta}) > z_{\alpha/2}$.
5. Prove that inverting a family of unbiased level α tests yields an unbiased confidence set, i.e., $\mathbb{P}_\theta(\theta' \in C(X)) \leq 1-\alpha$ for all $\theta' \neq \theta$.
6. Consider the Score test for $H_0 : \theta = \theta_0$ in a one-parameter model. Write the confidence set obtained by inverting this test and compare it to the LR and Wald confidence sets in a specific example (e.g., Poisson mean).
7. Argue, structurally, why the test–confidence duality is an instance of a broader principle in mathematics: the duality between points and hyperplanes (or between membership and separation), and identify how this geometric perspective clarifies the relationship between coverage probability and rejection probability.

Module 6 — Bayesian Inference

Priors and Posteriors

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.1

The prior-to-posterior update is the foundational operation of Bayesian inference. Every subsequent node in this module — conjugate families, point estimation, credible sets, predictive distributions, model comparison, and asymptotic reconciliation — is a derived consequence of the posterior distribution defined here. This node depends on the parametric model framework (Node 2.1), the likelihood function (Node 2.3), and the law of total probability (Node 1.7).

Formal Definition

Let $X = (X_1, \ldots, X_n)$ be a random sample with joint density or mass function $p(x \mid \theta)$ for $\theta \in \Theta \subseteq \mathbb{R}^k$.

Prior Distribution

A **prior distribution** $\pi(\theta)$ is a probability measure on $(\Theta, \mathcal{B}(\Theta))$ specified before observing X. It encodes the state of knowledge about θ prior to data collection. A prior is **proper** if $\int_\Theta \pi(\theta)\, d\theta = 1$, and **improper** if this integral diverges.

Posterior Distribution

Given observed data $X = x$, the **posterior distribution** is

$$\pi(\theta \mid x) = \frac{p(x \mid \theta)\, \pi(\theta)}{m(x)},$$

where the **marginal likelihood** (or evidence) is

$$m(x) = \int_\Theta p(x \mid \theta)\, \pi(\theta)\, d\theta.$$

The posterior exists as a proper distribution whenever $0 < m(x) < \infty$.

Jeffreys Prior

The **Jeffreys prior** is defined as

$$\pi^J(\theta) \propto \sqrt{\det I(\theta)},$$

where $I(\theta)$ is the Fisher information matrix. This prior is invariant under reparametrization: if $\phi = g(\theta)$ is a smooth bijection, then $\pi^J(\phi) \propto \sqrt{\det I_\phi(\phi)}$, where I_ϕ is the Fisher information in the ϕ-parametrization.

Intuition

The posterior is a compromise between what we believed before seeing data (the prior) and what the data tell us (the likelihood). In regions where the likelihood is large, the posterior concentrates mass; in regions where the prior assigns little mass, the posterior is suppressed even if the likelihood is moderate.

The marginal likelihood $m(x)$ serves purely as a normalizing constant for the posterior but carries independent significance as the model's overall predictive score for the observed data (Node 6.6).

Improper priors are useful default specifications — they express maximal pre-data ignorance within a parametric family — but require verification that the resulting posterior is proper.

The Jeffreys prior is the canonical "objective" prior: it assigns prior mass in a way that respects the geometry of the statistical model, treating reparametrizations as coordinate changes on the same underlying manifold.

Structural Role

The prior-to-posterior update is the complete description of Bayesian learning. All Bayesian procedures — point estimates (Node 6.3), credible sets (Node 6.4), posterior predictive distributions (Node 6.5), and Bayes factors (Node 6.6) — are functionals of the posterior.

This node requires: parametric models (Node 2.1), the likelihood principle (Node 2.3), and marginalization via the law of total probability (Node 1.7).

This node enables: every subsequent node in Module 6. The posterior is also the bridge to asymptotic reconciliation with frequentist inference via the Bernstein–von Mises theorem (Node 6.7).

Mathematical Development

Derivation of Bayes' Theorem for Densities

Starting from the joint density of (X, θ):

$$p(x, \theta) = p(x \mid \theta)\, \pi(\theta) = \pi(\theta \mid x)\, m(x).$$

Solving for $\pi(\theta \mid x)$ yields Bayes' theorem. The key requirement is $m(x) > 0$, which holds whenever the data are compatible with at least one parameter value that receives prior mass.

Sequential Updating

For i.i.d. data, the posterior after n observations can be computed sequentially. Let $\pi_0(\theta) = \pi(\theta)$ and define

$$\pi_k(\theta) \propto p(x_k \mid \theta)\, \pi_{k-1}(\theta), \quad k = 1, \ldots, n.$$

Then $\pi_n(\theta) = \pi(\theta \mid x_1, \ldots, x_n)$. The order of updating does not matter:

$$\pi(\theta \mid x_1, \ldots, x_n) \propto \pi(\theta) \prod_{i=1}^{n} p(x_i \mid \theta).$$

Sufficiency and the Posterior

If $T(X)$ is a sufficient statistic for θ, then by the factorization theorem $p(x \mid \theta) = g(T(x), \theta)\, h(x)$, so

$$\pi(\theta \mid x) = \frac{g(T(x), \theta)\, h(x)\, \pi(\theta)}{\int g(T(x), \theta)\, h(x)\, \pi(\theta)\, d\theta} = \frac{g(T(x), \theta)\, \pi(\theta)}{\int g(T(x), \theta)\, \pi(\theta)\, d\theta} = \pi(\theta \mid T(x)).$$

The posterior depends on the data only through the sufficient statistic.

Proper vs Improper Priors

An improper prior $\pi(\theta)$ with $\int_\Theta \pi(\theta)\, d\theta = \infty$ can still yield a proper posterior, provided $m(x) = \int p(x \mid \theta)\, \pi(\theta)\, d\theta < \infty$ for almost all x. For example, the flat prior $\pi(\mu) = 1$ on $\mathbb{R}$ for the normal mean yields proper posterior $\mu \mid x \sim \mathcal{N}(\bar{x}, \sigma^2/n)$.

Derivation of the Jeffreys Prior

Consider a one-parameter model with Fisher information $I(\theta) = -\mathbb{E}_\theta[\ell''(\theta)]$, where $\ell(\theta) = \log p(X \mid \theta)$. Under reparametrization $\phi = g(\theta)$, the Fisher information transforms as

$$I_\phi(\phi) = I(\theta) \left(\frac{d\theta}{d\phi} \right)^2.$$

The prior density transforms as $\pi_\phi(\phi) = \pi(\theta) \left| \frac{d\theta}{d\phi} \right|$. Setting $\pi(\theta) \propto \sqrt{I(\theta)}$ gives

$$\pi_\phi(\phi) \propto \sqrt{I(\theta)} \left| \frac{d\theta}{d\phi} \right| = \sqrt{I(\theta) \left(\frac{d\theta}{d\phi} \right)^2} = \sqrt{I_\phi(\phi)}.$$

Hence the Jeffreys prior is invariant under reparametrization.

Posterior Concentration

As $n \to \infty$, the posterior concentrates around the true parameter value θ_0. Informally, the log-posterior is

$$\log \pi(\theta \mid x) = \sum_{i=1}^{n} \log p(x_i \mid \theta) + \log \pi(\theta) - \log m(x).$$

The likelihood sum grows as $O(n)$ while $\log \pi(\theta)$ is $O(1)$, so the prior is overwhelmed. Under regularity conditions, this is made precise by the Bernstein–von Mises theorem (Node 6.7).

Worked Examples

Example 1: Normal Mean with Flat Prior

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Use the improper flat prior $\pi(\mu) \propto 1$.

The likelihood is

$$p(x \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) = \exp\Big(-\frac{n}{2\sigma^2} (\mu - \bar{x})^2 \Big) \cdot C(x),$$

where $C(x)$ does not depend on μ. Thus

$$\pi(\mu \mid x) \propto \exp\Big(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\Big),$$

which is the kernel of $\mathcal{N}(\bar{x}, \sigma^2/n)$. The posterior is proper despite the improper prior.

Example 2: Jeffreys Prior for the Bernoulli Model

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$, $\theta \in (0,1)$. The Fisher information is

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

The Jeffreys prior is $\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is $\text{Beta}(1/2, 1/2)$. With $s = \sum x_i$ successes, the posterior is

$$\pi(\theta \mid x) = \text{Beta}\big(\tfrac{1}{2} + s,\ \tfrac{1}{2} + n - s\big).$$

The posterior mean is $(s + 1/2)/(n+1)$, which shrinks the MLE s/n toward $1/2$ by an amount that vanishes as $n \to \infty$.

Cross-Links

- **Linear Algebra:** The Fisher information matrix $I(\theta)$ is a Riemannian metric on the parameter manifold; the Jeffreys prior is the volume element of this metric, connecting Bayesian inference to differential geometry on statistical manifolds.
- **AI:** Variational inference approximates the posterior $\pi(\theta \mid x)$ by optimizing over a tractable family when the marginal likelihood $m(x)$ is intractable. The prior-to-posterior update is the exact operation that variational and MCMC methods approximate.
- **Economics:** Bayesian agents update beliefs via Bayes' rule; rational expectations equilibria assume all agents compute posteriors from public signals. The prior encodes heterogeneous beliefs across agents.
- **Quantum Mechanics:** State update upon measurement (the Luders rule) is the non-commutative analogue of Bayes' theorem. The prior is a density matrix; the posterior is the post-measurement state.

Structural Compression

Invariant: $\pi(\theta \mid x) \propto p(x \mid \theta)\,\pi(\theta)$. The posterior is proportional to likelihood times prior.

Mechanism: Bayes' theorem normalizes the product $p(x \mid \theta)\,\pi(\theta)$ over Θ to produce a proper probability measure. The likelihood reweights prior mass according to data compatibility. The normalizing constant $m(x)$ ensures coherence.

Cross-System Echo: The posterior update is the universal belief revision rule: it appears as Bayesian filtering in signal processing, as the E-step kernel in EM algorithms, as density matrix conditioning in quantum information, and as belief propagation in graphical models.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$. Derive the posterior under the Jeffreys prior $\pi(\lambda) \propto 1/\lambda$. Verify the posterior is proper for $n \geq 1$.
2. Prove that the posterior distribution depends on the data only through a sufficient statistic by applying the factorization theorem.
3. Show that the Jeffreys prior for the normal variance σ^2 (with known mean) is $\pi(\sigma^2) \propto 1/\sigma^2$. Derive the resulting posterior and identify its distributional family.
4. Compute the posterior for θ in the model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, \theta)$ under the prior $\pi(\theta) \propto \theta^{-1}$. Show that the MLE and the posterior mode differ.
5. Demonstrate sequential updating: let X_1, X_2, X_3 be i.i.d. Bernoulli(θ) with $\pi(\theta) = \text{Beta}(2, 2)$. Compute the posterior after each observation for the sequence $(1, 0, 1)$ and verify the final posterior matches the batch computation.
6. Prove that if $\pi(\theta)$ is a proper prior and $p(x \mid \theta)$ is a bounded likelihood, then the posterior $\pi(\theta \mid x)$ is proper for all x.
7. Argue, structurally, why the prior is asymptotically irrelevant in regular parametric models: identify the rate at which the likelihood dominates the prior, and explain why this argument fails in nonparametric settings where Θ is infinite-dimensional.

Conjugate Families

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.2

Conjugate priors provide the rare setting in which Bayesian updating is analytically tractable, yielding closed-form posteriors that reveal the exact mechanism by which data modify beliefs. This node depends on the prior-to-posterior update (Node 6.1) and the exponential family structure (Node 2.6). It enables tractable computation of Bayesian point estimates (Node 6.3) and credible sets (Node 6.4), and supplies the canonical examples used throughout Module 6.

Formal Definition

Let $\{p(x \mid \theta) : \theta \in \Theta\}$ be a parametric likelihood family and let $\mathcal{F} = \{\pi(\theta \mid \eta) : \eta \in H\}$ be a family of prior distributions indexed by hyperparameters $\eta \in H$.

Conjugate prior family. The family $\mathcal{F}$ is **conjugate** to the likelihood $p(x \mid \theta)$ if for every $\eta \in H$ and every observation x,

$$\pi(\theta \mid \eta) \in \mathcal{F} \implies \pi(\theta \mid x) \in \mathcal{F}.$$

That is, the posterior belongs to the same parametric family as the prior, differing only in the value of the hyperparameter.

Exponential Family Conjugacy

For a likelihood in the canonical exponential family,

$$p(x \mid \theta) = h(x) \exp\big\{\eta(\theta)^\top T(x) - A(\theta)\big\},$$

the conjugate prior takes the form

$$\pi(\theta \mid \lambda_0, n_0) \propto \exp\big\{\lambda_0^\top \eta(\theta) - n_0\, A(\theta)\big\},$$

where (λ_0, n_0) are hyperparameters. After n i.i.d. observations, the posterior hyperparameters update as

$$\lambda_0 \mapsto \lambda_n = \lambda_0 + \sum_{i=1}^{n} T(x_i), \qquad n_0 \mapsto n_n = n_0 + n.$$

Intuition

Conjugacy means the prior and posterior speak the same parametric language. The data do not change the functional form of our belief — they only shift the hyperparameters. The hyperparameter n_0 acts as a prior pseudo-sample size: it quantifies the strength of prior information in units commensurate with data. The ratio λ_0/n_0 is a prior guess at the sufficient statistic per observation. As data accumulate, the posterior hyperparameters drift toward data-driven values and the prior pseudo-sample size becomes negligible relative to the actual sample size.

Conjugacy is the exception rather than the rule. Most interesting models — mixture models, hierarchical models, regression with non-conjugate priors — require computational approximation (MCMC, variational inference). But conjugate families remain the essential test bed for understanding Bayesian mechanics.

Structural Role

Conjugate priors make the posterior analytically tractable by closing the family under Bayesian updating. The hyperparameter update equations reveal the precise mechanism by which data accumulate: additive updates to the natural parameter and sample-size components.

This node requires the exponential family structure (Node 2.6) and the prior-to-posterior update (Node 6.1). It enables closed-form computation in Bayesian point estimation (Node 6.3), credible sets (Node 6.4), and posterior predictive distributions (Node 6.5). When conjugacy is unavailable, computation requires MCMC or variational methods (Module 9).

Mathematical Development

Proof of Beta-Binomial Conjugacy

Let $X_1, \ldots, X_n \mid \theta \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ and $\pi(\theta) = \text{Beta}(\alpha_0, \beta_0)$.

The prior density is

$$\pi(\theta) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\,\Gamma(\beta_0)}\,\theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1}, \quad \theta \in (0,1).$$

The likelihood for $s = \sum_{i=1}^{n} x_i$ successes in n trials is

$$p(x \mid \theta) = \theta^s (1-\theta)^{n-s}.$$

By Bayes' theorem:

$$\pi(\theta \mid x) \propto \theta^s(1-\theta)^{n-s} \cdot \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} = \theta^{\alpha_0 + s - 1}(1-\theta)^{\beta_0 + n - s - 1}.$$

This is the kernel of $\text{Beta}(\alpha_0 + s,\ \beta_0 + n - s)$. Since the kernel uniquely determines the distribution, the posterior is Beta, proving conjugacy. □

Normal-Normal Conjugacy (Known Variance)

Let $X_i \mid \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with σ^2 known, and $\pi(\mu) = \mathcal{N}(\mu_0, \tau_0^2)$.

Define the prior precision $\kappa_0 = 1/\tau_0^2$ and data precision $\kappa_{\text{data}} = n/\sigma^2$. The posterior is

$$\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2),$$

where

$$\tau_n^2 = \frac{1}{\kappa_0 + \kappa_{\text{data}}}, \qquad \mu_n = \tau_n^2(\kappa_0 \mu_0 + \kappa_{\text{data}} \bar{x})\,.$$

The posterior mean is a precision-weighted average of the prior mean and sample mean. As $n \to \infty$, $\mu_n \to \bar{x}$ and $\tau_n^2 \to \sigma^2/n$.

Gamma-Poisson Conjugacy

Let $X_i \mid \lambda \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ and $\pi(\lambda) = \text{Gamma}(\alpha_0, \beta_0)$ with density $\pi(\lambda) \propto \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda}$.

The likelihood is $p(x \mid \lambda) \propto \lambda^{\sum x_i} e^{-n\lambda}$. Thus

$$\pi(\lambda \mid x) \propto \lambda^{\alpha_0 + \sum x_i - 1} e^{-(\beta_0 + n)\lambda},$$

which is $\text{Gamma}(\alpha_0 + \sum x_i,\ \beta_0 + n)$. The posterior mean is $(\alpha_0 + \sum x_i)/(\beta_0 + n)$.

Dirichlet-Multinomial Conjugacy

Let $X \mid \theta \sim \text{Multinomial}(n, \theta_1, \ldots, \theta_K)$ with $\pi(\theta) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$. The posterior is

$$\theta \mid x \sim \text{Dirichlet}(\alpha_1 + x_1, \ldots, \alpha_K + x_K).$$

Each hyperparameter α_k acts as a prior pseudo-count for category k. The posterior mean for component k is $(\alpha_k + x_k)/(\sum_j \alpha_j + n)$.

Conjugate Table Summary

Likelihood	Conjugate Prior	Posterior	Update Rule
Bernoulli(θ)	Beta(α_0, β_0)	Beta($\alpha_0 + s, \beta_0 + n - s$)	add successes, failures
$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\mu_0, \tau_0^2)$	$\mathcal{N}(\mu_n, \tau_n^2)$	precision-weighted mean
Poisson(λ)	Gamma(α_0, β_0)	Gamma($\alpha_0 + \sum x_i, \beta_0 + n$)	add counts, observations
Multinomial(n, θ)	Dirichlet(α)	Dirichlet($\alpha + x$)	add category counts
$\mathcal{N}(\mu, \sigma^2)$ (both unknown)	NIG($\mu_0, \kappa_0, \alpha_0, \beta_0$)	NIG($\mu_n, \kappa_n, \alpha_n, \beta_n$)	see standard references

Interpretation of Hyperparameters

In every conjugate pair, the hyperparameters decompose into a **pseudo-sample size** n_0 and a **pseudo-sufficient statistic** λ_0. For Beta-Binomial: $n_0 = \alpha_0 + \beta_0$ (prior trials), $\lambda_0/n_0 = \alpha_0/(\alpha_0 + \beta_0)$ (prior success rate). For Gamma-Poisson: $n_0 = \beta_0$ (prior observations), $\lambda_0/n_0 = \alpha_0/\beta_0$ (prior rate).

Worked Examples

Example 1: Beta-Binomial with Clinical Trial Data

A drug trial tests 20 patients; 14 respond. The prior on the response rate is $\theta \sim \text{Beta}(2, 2)$, encoding a mild belief toward 1/2.

Posterior: $\text{Beta}(2 + 14, 2 + 6) = \text{Beta}(16, 8)$.

Posterior mean: $16/24 = 2/3 \approx 0.667$.

Prior mean: $2/4 = 0.5$. MLE: $14/20 = 0.7$.

The posterior mean lies between the prior mean and the MLE, closer to the MLE because the data (pseudo-sample size 20) dominate the prior (pseudo-sample size 4).

Posterior variance: $(16 \cdot 8)/(24^2 \cdot 25) = 128/14400 \approx 0.0089$. Posterior standard deviation: ≈ 0.094.

Example 2: Normal-Normal with Sensor Calibration

A sensor measures a voltage with known noise $\sigma^2 = 4$. Prior belief: $\mu \sim \mathcal{N}(10, 1)$, i.e., $\mu_0 = 10, \tau_0^2 = 1, \kappa_0 = 1$.

After $n = 5$ readings with $\bar{x} = 11.2$: data precision $\kappa_{\text{data}} = 5/4 = 1.25$.

Posterior precision: $\kappa_0 + \kappa_{\text{data}} = 1 + 1.25 = 2.25$.

Posterior variance: $\tau_n^2 = 1/2.25 \approx 0.444$.

Posterior mean: $\mu_n = (1/2.25)(1 \cdot 10 + 1.25 \cdot 11.2) = (1/2.25)(10 + 14) = 24/2.25 \approx 10.667$.

The posterior shifts from the prior mean 10 toward the sample mean 11.2, with the shift proportional to the relative data precision.

Cross-Links

- **Linear Algebra:** The precision-weighted update in Normal-Normal conjugacy is a special case of the Woodbury matrix identity applied to precision matrices. In the multivariate case, the posterior precision matrix is the sum of the prior precision and the data precision: $\Sigma_n^{-1} = \Sigma_0^{-1} + n\Sigma^{-1}$.
- **AI:** Conjugate priors underlie Bayesian online learning: each mini-batch update is a hyperparameter shift. Latent Dirichlet Allocation uses Dirichlet-Multinomial conjugacy as its core inferential engine.
- **Economics:** Beta-Bernoulli conjugacy models Bayesian learning by agents who observe binary signals (buy/sell, default/no-default) and update beliefs about an unknown probability.
- **Quantum Mechanics:** The Gaussian state update under homodyne measurement in quantum optics is formally identical to Normal-Normal conjugacy, with the Wigner function playing the role of the prior density.

Structural Compression

Invariant: The posterior hyperparameters are additive in the sufficient statistic of the data: $\eta_n = \eta_0 + T_n$, where T_n is the vector of accumulated sufficient statistics.

Mechanism: Exponential family structure ensures the product of likelihood and conjugate prior remains in the same exponential family. The natural parameter of the posterior is a linear function of the natural parameter of the prior and the sufficient statistic of the data.

Cross-System Echo: Conjugate updating is the Bayesian instance of Kalman filtering (linear-Gaussian conjugacy is the Kalman filter in one dimension). In information geometry, conjugate priors lie on the same exponential family manifold as the likelihood, and the posterior update is a geodesic step on this manifold.

Exercises

1. Prove Gamma-Poisson conjugacy: show that if $X_i \mid \lambda \sim \text{Poisson}(\lambda)$ and $\pi(\lambda) = \text{Gamma}(\alpha_0, \beta_0)$, then the posterior is $\text{Gamma}(\alpha_0 + \sum x_i, \beta_0 + n)$.
2. For the Normal-Normal model with known variance, show that the posterior mean can be written as $\mu_n = w \cdot \mu_0 + (1 - w) \cdot \bar{x}$ where $w = \tau_0^{-2}/(\tau_0^{-2} + n\sigma^{-2})$. Interpret w as a function of n and verify $w \to 0$ as $n \to \infty$.
3. Derive the conjugate prior for the exponential distribution $p(x \mid \lambda) = \lambda e^{-\lambda x}$. Identify the conjugate family, state the posterior hyperparameter update, and compute the posterior mean.

4. In the Dirichlet-Multinomial model with $K = 3$ and prior Dirichlet$(1, 1, 1)$, compute the posterior after observing counts $(x_1, x_2, x_3) = (10, 5, 3)$. Find the posterior mean and the posterior mode of each component.

5. Show that for the Beta-Binomial model, the prior predictive probability of s successes in n trials is the Beta-Binomial distribution: $p(s) = \binom{n}{s} \frac{B(\alpha_0+s, \beta_0+n-s)}{B(\alpha_0, \beta_0)}$.

6. Prove that the Normal-Normal posterior variance τ_n^2 is always smaller than both the prior variance τ_0^2 and the sampling variance σ^2/n. When are the two bounds approximately equal?

7. Argue, structurally, why conjugacy is restricted to exponential families: identify the algebraic property of the likelihood that allows the prior-posterior product to remain in the same family, and explain why non-exponential families (e.g., Cauchy) do not admit conjugate priors of finite parametric dimension.

Bayesian Point Estimation

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.3

Bayesian point estimation selects a single summary of the posterior distribution by minimizing posterior expected loss. This node connects the posterior (Node 6.1) to the decision-theoretic framework (Node 3.1), replacing the frequentist risk with posterior risk. It provides the estimators — posterior mean, MAP, posterior median — that serve as inputs to prediction (Node 6.5) and model comparison (Node 6.6).

Formal Definition

Loss Function and Posterior Risk

A **loss function** is a measurable map $L : \Theta \times \mathcal{A} \to [0, \infty)$, where $\mathcal{A}$ is the action space. The **posterior expected loss** (posterior risk) of action a given data x is

$$\rho(a, x) = \int_{\Theta} L(\theta, a)\, \pi(\theta \mid x)\, d\theta.$$

Bayes Estimator

The **Bayes estimator** under loss L is

$$\hat{\theta}^{\text{Bayes}}(x) = \arg\min_{a \in \mathcal{A}} \ \rho(a, x) = \arg\min_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a)\, \pi(\theta \mid x)\, d\theta.$$

Bayes Risk

The **Bayes risk** of an estimator $\hat{\theta}$ under prior π is the prior expectation of the frequentist risk:

$$r(\pi, \hat{\theta}) = \int_{\Theta} R(\theta, \hat{\theta})\, \pi(\theta)\, d\theta = \int_{\Theta} \int_{\mathcal{X}} L(\theta, \hat{\theta}(x))\, p(x \mid \theta)\, dx\, \pi(\theta)\, d\theta.$$

The Bayes estimator minimizes the Bayes risk over all estimators.

Intuition

The Bayesian framework converts estimation into an optimization problem: given the posterior, choose the single value that incurs the smallest expected loss. Different loss functions extract different features of the posterior.

Squared error loss penalizes large deviations quadratically and yields the posterior mean — the center of mass. Absolute error loss penalizes proportionally and yields the posterior median — the balance point. Zero-one loss (for discrete parameters) or a limiting peaked loss yields the posterior mode — the single most probable value.

The Bayes estimator is always admissible (cannot be dominated in risk by any other estimator at every parameter value), provided the prior assigns positive mass to all open sets. This is a deep connection between Bayesian and frequentist optimality.

Structural Role

Bayesian point estimation reframes optimality from the frequentist minimax or unbiasedness criteria to posterior loss minimization. The choice of loss function determines which posterior functional is optimal.

This node requires: the posterior distribution (Node 6.1) and the decision-theoretic framework (Node 3.1). It enables: posterior predictive distributions (Node 6.5) via plug-in vs integrated prediction, and connects to credible sets (Node 6.4) through the relationship between the MAP estimator and the HPD region.

Mathematical Development

Proof: Posterior Mean Minimizes Squared Error Loss

Theorem. Under squared error loss $L(\theta, a) = (\theta - a)^2$, the Bayes estimator is the posterior mean $\hat{\theta}^{\text{Bayes}}(x) = \mathbb{E}[\theta \mid x]$.

Proof. The posterior expected loss is

$$\rho(a, x) = \int_{\Theta} (\theta - a)^2 \, \pi(\theta \mid x) \, d\theta.$$

Expand the square:

$$\rho(a, x) = \int_{\Theta} \theta^2 \, \pi(\theta \mid x) \, d\theta - 2a \int_{\Theta} \theta \, \pi(\theta \mid x) \, d\theta + a^2.$$

Differentiate with respect to a:

$$\frac{\partial}{\partial a} \rho(a, x) = -2 \, \mathbb{E}[\theta \mid x] + 2a.$$

Setting this to zero gives $a^* = \mathbb{E}[\theta \mid x]$. The second derivative is $2 > 0$, confirming a minimum. $\square$

Proof: Posterior Median Minimizes Absolute Error Loss

Theorem. Under absolute error loss $L(\theta, a) = |\theta - a|$, the Bayes estimator is the posterior median.

Proof. Write

$$\rho(a, x) = \int_{-\infty}^{a} (a - \theta) \, \pi(\theta \mid x) \, d\theta + \int_{a}^{\infty} (\theta - a) \, \pi(\theta \mid x) \, d\theta.$$

Differentiate under the integral sign (using Leibniz's rule):

$$\frac{\partial}{\partial a} \rho(a, x) = \int_{-\infty}^{a} \pi(\theta \mid x) \, d\theta - \int_{a}^{\infty} \pi(\theta \mid x) \, d\theta = 2F_{\pi}(a \mid x) - 1,$$

where $F_{\pi}(\cdot \mid x)$ is the posterior CDF. Setting this to zero gives $F_{\pi}(a^* \mid x) = 1/2$, i.e., a^* is the posterior median. $\square$

MAP Estimator and Its Decision-Theoretic Justification

For a continuous parameter space, the MAP estimator is

$$\hat{\theta}^{\text{MAP}}(x) = \arg\max_{\theta \in \Theta} \ \pi(\theta \mid x) = \arg\max_{\theta \in \Theta} \ \big[\log p(x \mid \theta) + \log \pi(\theta)\big].$$

The MAP is the Bayes estimator under the limiting loss $L_\varepsilon(\theta, a) = \mathbf{1}(|\theta - a| > \varepsilon)$ as $\varepsilon \to 0$, provided the posterior is unimodal and continuous.

Connection to Regularized Maximum Likelihood

Under a Gaussian prior $\pi(\theta) = \mathcal{N}(0, \sigma_\pi^2 I)$, the MAP estimator solves

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \ \left[\sum_{i=1}^{n} \log p(x_i \mid \theta) - \frac{\|\theta\|^2}{2\sigma_\pi^2}\right],$$

which is **ridge regression** (L2-regularized MLE) with penalty $\lambda = 1/(2\sigma_\pi^2)$.

Under a Laplace prior $\pi(\theta_j) \propto \exp(-|\theta_j|/b)$, the MAP estimator becomes the **LASSO** (L1-regularized MLE).

Admissibility of Bayes Estimators

Theorem. If $\pi(\theta) > 0$ for all $\theta \in \Theta$ and the Bayes risk $r(\pi, \hat{\theta}^{\text{Bayes}}) < \infty$, then $\hat{\theta}^{\text{Bayes}}$ is admissible.

Proof sketch. Suppose $\hat{\theta}'$ dominates $\hat{\theta}^{\text{Bayes}}$: $R(\theta, \hat{\theta}') \le R(\theta, \hat{\theta}^{\text{Bayes}})$ for all θ, with strict inequality on a set of positive prior measure. Then $r(\pi, \hat{\theta}') < r(\pi, \hat{\theta}^{\text{Bayes}})$, contradicting the minimality of the Bayes risk. □

Posterior Mean as Shrinkage Estimator

In the Normal-Normal conjugate setting with prior $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$ and data $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the posterior mean is

$$\hat{\mu}^{\text{Bayes}} = w\,\mu_0 + (1 - w)\,\bar{X}, \quad w = \frac{\sigma^2/n}{\tau_0^2 + \sigma^2/n}.$$

This shrinks the MLE $\bar{X}$ toward the prior mean μ_0. The shrinkage factor w decreases with n and with prior variance τ_0^2, vanishing as either grows.

Worked Examples

Example 1: Posterior Mean for Beta-Binomial

Let $X_1, \ldots, X_{20} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\theta)$ with $s = 12$ successes. Prior: $\theta \sim \text{Beta}(3, 3)$.

Posterior: $\text{Beta}(15, 11)$.

Posterior mean (Bayes estimator under squared error loss):

$$\hat{\theta}^{\text{Bayes}} = \frac{15}{15 + 11} = \frac{15}{26} \approx 0.577.$$

MLE: $12/20 = 0.6$. Prior mean: $3/6 = 0.5$.

Posterior mode (MAP estimator): $(15 - 1)/(15 + 11 - 2) = 14/24 \approx 0.583$.

Posterior median: approximately 0.577 (close to the mean since Beta(15,11) is nearly symmetric).

Example 2: Posterior Mean under Normal-Normal with Asymmetric Information

Prior: $\mu \sim \mathcal{N}(0, 100)$ (vague). Data: $n = 10$, $\bar{x} = 3.5$, $\sigma^2 = 4$.

Prior precision: $\kappa_0 = 1/100 = 0.01$. Data precision: $\kappa_{\text{data}} = 10/4 = 2.5$.

Posterior mean:

$$\hat{\mu}^{\text{Bayes}} = \frac{0.01 \cdot 0 + 2.5 \cdot 3.5}{0.01 + 2.5} = \frac{8.75}{2.51} \approx 3.486.$$

The vague prior barely influences the estimate: shrinkage weight $w = 0.01/2.51 \approx 0.004$.

Posterior variance: $1/2.51 \approx 0.398$. Compare to sampling variance $\sigma^2/n = 0.4$. The prior contributes almost no precision.

Cross-Links

- **Linear Algebra:** The posterior mean in multivariate normal models is a solution to a linear system involving precision matrices: $\hat{\mu}^{\text{Bayes}} = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n\Sigma^{-1}\bar{x})$. This is a Tikhonov-regularized least squares problem.
- **AI:** MAP estimation under structured priors produces regularized learning: L2 prior gives weight decay, L1 prior gives sparsity, and the full posterior mean gives Bayesian model averaging. Deep learning approximations to the posterior mean include MC dropout and deep ensembles.
- **Economics:** In rational expectations models, agents form point estimates of unknown parameters (productivity shocks, inflation rates) by computing posterior means under squared error loss, producing optimal forecasts given their information sets.
- **Quantum Mechanics:** Quantum state estimation uses Bayesian point estimators to reconstruct density matrices from measurement data. The posterior mean over density matrices (under the Hilbert-Schmidt metric) is the Bayes estimator under Frobenius loss.

Structural Compression

Invariant: The Bayes estimator minimizes posterior expected loss. The specific functional of the posterior — mean, median, or mode — is uniquely determined by the loss geometry.

Mechanism: Differentiate the posterior expected loss with respect to the action. The first-order condition selects the posterior summary matching the loss function. Squared error yields the expectation, absolute error yields the median, zero-one loss yields the mode.

Cross-System Echo: MAP estimation under Gaussian prior is ridge regression; under Laplace prior, it is the LASSO. The Bayes estimator is the admissible estimator corresponding to the prior π, bridging Bayesian and frequentist admissibility theory.

Exercises

1. Prove that the posterior mean minimizes posterior expected loss under weighted squared error loss $L(\theta, a) = w(\theta)(\theta - a)^2$, and find the resulting estimator in terms of the posterior.
2. For the Gamma-Poisson model with prior $\text{Gamma}(\alpha_0, \beta_0)$ and data $\sum x_i = 45$, $n = 10$, compute the posterior mean, posterior mode, and posterior median. Compare all three to the MLE.

3. Show that the MAP estimator is not invariant under reparametrization: if $\hat{\theta}^{\text{MAP}}$ is the MAP for θ, then $g(\hat{\theta}^{\text{MAP}})$ is generally not the MAP for $\phi = g(\theta)$. Give an explicit example.

4. Prove that the Bayes estimator under squared error loss has smaller Bayes risk than the MLE, for the Beta-Binomial model with prior $\text{Beta}(\alpha, \alpha)$. Under what conditions is the improvement largest?

5. Derive the posterior mean for the multivariate normal model: $X_i \sim \mathcal{N}_k(\mu, \Sigma)$, prior $\mu \sim \mathcal{N}_k(\mu_0, \Sigma_0)$, with Σ known. Express the answer in terms of precision matrices.

6. In the James-Stein setting ($k \geq 3$, $X \sim \mathcal{N}_k(\theta, I)$), show that the posterior mean under the prior $\theta \sim \mathcal{N}_k(0, \tau^2 I)$ produces a shrinkage estimator of the form $(1 - c/\|X\|^2)X$ for appropriate c, and connect this to the inadmissibility of the MLE.

7. Argue, structurally, why the choice of loss function cannot be resolved within the Bayesian framework itself: explain what additional information or judgment is required, and how the decision-theoretic framework (Node 3.1) addresses this gap.

Credible Sets

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.4

Credible sets are the Bayesian analogue of confidence sets, providing posterior probability statements about parameter regions. This node depends on the posterior distribution (Node 6.1) and the frequentist confidence set framework (Node 4.1), which it reinterprets from the Bayesian perspective. It enables posterior predictive model checking (Node 6.5) and connects to asymptotic frequentist-Bayesian reconciliation via the Bernstein–von Mises theorem (Node 6.7).

Formal Definition

Let $\pi(\theta \mid x)$ be the posterior distribution on $\Theta \subseteq \mathbb{R}^k$.

Credible Set

A set $C(x) \subseteq \Theta$ is a $(1-\alpha)$**-credible set** if

$$\int_{C(x)} \pi(\theta \mid x)\, d\theta \geq 1 - \alpha.$$

The probability is with respect to the posterior. The parameter θ is the random quantity; the data x are fixed.

Highest Posterior Density (HPD) Region

The **HPD region** at level $1-\alpha$ is

$$C^{\text{HPD}}(x) = \big\{\theta \in \Theta : \pi(\theta \mid x) \geq k_\alpha\big\},$$

where k_α is the largest constant such that

$$\int_{C^{\text{HPD}}(x)} \pi(\theta \mid x)\, d\theta = 1 - \alpha.$$

Equal-Tailed Credible Interval

For scalar θ, the **equal-tailed** $(1-\alpha)$**-credible interval** is

$$C^{\text{ET}}(x) = \big[Q_{\alpha/2},\ Q_{1-\alpha/2}\big],$$

where $Q_p = Q_p(\pi(\cdot \mid x))$ denotes the p-quantile of the posterior distribution.

Intuition

A credible set directly answers: given the observed data and the model, what region contains θ with posterior probability at least $1-\alpha$?

The HPD region collects the most probable parameter values first, so it is the shortest interval (smallest volume region) achieving the desired posterior coverage. For symmetric unimodal posteriors, the HPD and equal-tailed intervals coincide. For skewed posteriors, the HPD region is shorter but asymmetric, while the equal-tailed interval is symmetric in probability but longer.

The credible set interpretation is conceptually simpler than the frequentist confidence set: it conditions on the observed data and makes a probability statement about the parameter. The price is dependence on the prior.

Structural Role

Credible sets are the canonical Bayesian uncertainty quantification tool for parameters. They require the posterior (Node 6.1) and parallel the frequentist confidence sets (Node 4.1).

The HPD region connects to Bayesian point estimation (Node 6.3): the MAP estimator is always contained in the HPD region, and as $\alpha \to 1$ the HPD region shrinks to the MAP.

The Bernstein–von Mises theorem (Node 6.7) establishes that in regular parametric models, $(1-\alpha)$-credible sets have asymptotic frequentist coverage $1-\alpha$, reconciling the two frameworks.

Mathematical Development

Proof: HPD is the Shortest Credible Interval

Theorem. Among all $(1-\alpha)$-credible sets in $\mathbb{R}$, the HPD region has the smallest Lebesgue measure.

Proof. Let C be any $(1-\alpha)$-credible set with $\int_C \pi(\theta \mid x)\, d\theta \geq 1-\alpha$. Let C^{HPD} be the HPD region with threshold k_α.

Consider the symmetric difference. For any $\theta_1 \in C^{\text{HPD}} \setminus C$ and $\theta_2 \in C \setminus C^{\text{HPD}}$, we have $\pi(\theta_1 \mid x) \geq k_\alpha > \pi(\theta_2 \mid x)$.

The posterior mass of $C \setminus C^{\text{HPD}}$ must be at least that of $C^{\text{HPD}} \setminus C$ (since both C and C^{HPD} achieve coverage $1-\alpha$). But every point in $C \setminus C^{\text{HPD}}$ has posterior density strictly less than every point in $C^{\text{HPD}} \setminus C$. Therefore

$$\lambda(C \setminus C^{\text{HPD}}) \geq \frac{\int_{C^{\text{HPD}} \setminus C} \pi(\theta \mid x)\, d\theta}{k'_\alpha} \geq \lambda(C^{\text{HPD}} \setminus C),$$

where $k'_\alpha \leq k_\alpha$ bounds the density on $C \setminus C^{\text{HPD}}$ from above. Thus $\lambda(C) = \lambda(C \cap C^{\text{HPD}}) + \lambda(C \setminus C^{\text{HPD}}) \geq \lambda(C^{\text{HPD}})$. $\square$

Formal Distinction from Confidence Sets

A frequentist $(1-\alpha)$-confidence set $C_{\text{freq}}(X)$ satisfies

$$\mathbb{P}_\theta\big(\theta \in C_{\text{freq}}(X)\big) \geq 1-\alpha \quad \text{for all } \theta \in \Theta.$$

Here θ is fixed and X varies. The probability is over repetitions of the experiment.

A Bayesian $(1-\alpha)$-credible set $C(x)$ satisfies

$$\pi\big(\theta \in C(x) \mid X = x\big) \geq 1 - \alpha.$$

Here x is fixed and θ varies under the posterior. The probability is conditional on the observed data.

These are fundamentally different probability statements. A credible set with good posterior coverage need not have good frequentist coverage, and vice versa.

When Credible Sets Have Frequentist Coverage

Under the Bernstein–von Mises theorem (Node 6.7), for regular parametric models with fixed dimension:

$$\pi(\theta \mid X_1, \ldots, X_n) \approx \mathcal{N}\left(\hat{\theta}_n,\ \frac{I(\theta_0)^{-1}}{n}\right)$$

in total variation. The HPD credible region based on this posterior is asymptotically

$$C_n^{\mathrm{HPD}} \approx \big\{\theta : n(\theta - \hat{\theta}_n)^\top I(\theta_0)(\theta - \hat{\theta}_n) \leq \chi^2_{k,1-\alpha}\big\},$$

which coincides with the Wald confidence ellipsoid. Therefore $\mathbb{P}_{\theta_0}(\theta_0 \in C_n^{\mathrm{HPD}}) \to 1 - \alpha$.

HPD for Common Posterior Families

Normal posterior: If $\theta \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$, the HPD interval is symmetric about μ_n:

$$C^{\mathrm{HPD}} = \big[\mu_n - z_{1-\alpha/2}\,\sigma_n,\ \mu_n + z_{1-\alpha/2}\,\sigma_n\big].$$

This coincides with the equal-tailed interval by symmetry.

Beta posterior: If $\theta \mid x \sim \mathrm{Beta}(\alpha_n, \beta_n)$ with $\alpha_n \neq \beta_n$, the HPD interval is shorter than the equal-tailed interval. It must be computed numerically by finding k_α such that the density level set has posterior mass $1 - \alpha$.

Gamma posterior: Similarly asymmetric; the HPD interval shifts toward smaller values relative to the equal-tailed interval due to right skewness.

Multivariate HPD Regions

For $\theta \in \mathbb{R}^k$, the HPD region is a level set of the joint posterior density. For a multivariate normal posterior $\theta \mid x \sim \mathcal{N}_k(\mu_n, \Sigma_n)$, the HPD region is the ellipsoid

$$C^{\mathrm{HPD}} = \big\{\theta : (\theta - \mu_n)^\top \Sigma_n^{-1} (\theta - \mu_n) \leq \chi^2_{k,1-\alpha}\big\}.$$

Worked Examples

Example 1: HPD Interval for a Beta Posterior

Data: $n = 30$ Bernoulli trials, $s = 21$ successes. Prior: $\mathrm{Beta}(1, 1)$ (uniform).

Posterior: $\mathrm{Beta}(22, 10)$. Posterior mean: $22/32 = 0.6875$. Posterior mode: $21/30 = 0.7$.

Equal-tailed 95% interval: Using Beta quantiles, $[Q_{0.025}, Q_{0.975}] \approx [0.518, 0.832]$. Length: 0.314.

HPD 95% interval: The density level set $\{\theta : f(\theta) \geq k\}$ must be found numerically. For $\mathrm{Beta}(22, 10)$: $C^{\mathrm{HPD}} \approx [0.530, 0.840]$. Length: 0.310.

The HPD interval is shifted slightly toward the mode and is shorter than the equal-tailed interval due to the right skewness of Beta$(22, 10)$.

Example 2: Normal Posterior Credible Interval vs Confidence Interval

Data: $n = 25$, $\bar{x} = 5.2$, $\sigma^2 = 9$ (known). Prior: $\mu \sim \mathcal{N}(4, 4)$.

Posterior: $\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$ with

$$\tau_n^2 = \frac{1}{1/4 + 25/9} = \frac{1}{0.25 + 2.778} = \frac{1}{3.028} \approx 0.330,$$

$$\mu_n = 0.330 \cdot (0.25 \cdot 4 + 2.778 \cdot 5.2) = 0.330 \cdot (1 + 14.444) = 0.330 \cdot 15.444 \approx 5.097.$$

95% credible interval: $5.097 \pm 1.96\sqrt{0.330} = 5.097 \pm 1.125 = [3.972, 6.222]$.

95% confidence interval (frequentist, no prior): $5.2 \pm 1.96 \cdot 3/5 = 5.2 \pm 1.176 = [4.024, 6.376]$.

The credible interval is shorter (width 2.25 vs 2.35) and shifted toward the prior mean of 4. As $n \to \infty$, the two intervals converge.

Cross-Links

- **Linear Algebra:** Multivariate HPD regions are ellipsoids defined by the posterior precision matrix. The eigenvectors of the precision matrix give the principal axes of the credible ellipsoid, connecting to PCA-like decompositions of posterior uncertainty.
- **AI:** Bayesian neural networks produce posterior distributions over weights; credible sets for predictions quantify epistemic uncertainty. Approximate HPD regions are computed via MCMC samples or variational posteriors.
- **Economics:** Bayesian credible intervals for structural parameters (elasticities, policy multipliers) directly quantify parameter uncertainty conditional on observed data, which is the natural object for policy decisions.
- **Quantum Mechanics:** Credible regions for quantum states (density matrices) arise in quantum state tomography. The HPD region over the space of density matrices provides the smallest set of states consistent with measurement data at a given posterior probability level.

Structural Compression

Invariant: $\pi(\theta \in C(x) \mid X = x) \geq 1 - \alpha$. The posterior mass of the credible set is at least $1 - \alpha$.

Mechanism: Integrate the posterior density over the candidate region $C(x)$. The HPD region selects the level set of the posterior density, which minimizes volume subject to the coverage constraint. This is a constrained optimization: minimize Lebesgue measure subject to posterior mass $\geq 1 - \alpha$.

Cross-System Echo: Credible sets are the Bayesian analogue of confidence sets. The Bernstein–von Mises theorem (Node 6.7) establishes their asymptotic equivalence in regular models, making credible sets the bridge between Bayesian and frequentist uncertainty quantification.

Exercises

1. For the posterior $\theta \mid x \sim \text{Gamma}(10, 2)$, compute the equal-tailed 95% credible interval and explain why the HPD interval will be shorter.
2. Prove that for a unimodal symmetric posterior, the HPD interval and the equal-tailed interval coincide.
3. Let $\theta \mid x \sim \text{Beta}(5, 15)$. Compute the posterior mean, posterior mode, and the equal-tailed 90% credible interval. Without computing it exactly, argue whether the HPD interval will shift left or right relative to the equal-tailed interval.
4. Construct an example where a 95% credible interval has frequentist coverage far from 95% for a specific θ_0. Specify the model, prior, and true parameter.
5. For the multivariate normal posterior $\theta \mid x \sim \mathcal{N}_2(\mu_n, \Sigma_n)$ with $\mu_n = (3, 1)^\top$ and $\Sigma_n = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, derive the equation of the 95% HPD ellipse and find its semi-axes.
6. Prove that the MAP estimator is always contained in the HPD region $C^{\text{HPD}}(x)$ for every $\alpha \in (0, 1)$, and that the HPD region shrinks to $\{\hat{\theta}^{\text{MAP}}\}$ as $\alpha \to 1$.
7. Argue, structurally, why the distinction between credible sets and confidence sets is sharpest in small samples with informative priors, and vanishes asymptotically under the conditions of the Bernstein–von Mises theorem: identify the role of the prior, the likelihood dominance rate, and the resulting posterior shape.

Posterior Predictive Distribution

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.5

The posterior predictive distribution is the Bayesian answer to prediction: it integrates parameter uncertainty into the distribution of future observations. This node depends on the posterior (Node 6.1), conjugate families for tractable computation (Node 6.2), and marginalization (Node 1.7). It enables model comparison (Node 6.6) through the marginal likelihood, which is the prior predictive evaluated at the observed data, and provides the foundation for posterior predictive model checking.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p(x \mid \theta)$ with posterior $\pi(\theta \mid x_{1:n})$. Let $\tilde{X}$ denote a future observation from the same model.

Posterior Predictive Distribution

The **posterior predictive distribution** of $\tilde{X}$ given observed data $x_{1:n}$ is

$$p(\tilde{x} \mid x_{1:n}) = \int_\Theta p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n})\, d\theta.$$

Prior Predictive Distribution

Before data are observed, the **prior predictive distribution** is

$$p(\tilde{x}) = \int_\Theta p(\tilde{x} \mid \theta)\, \pi(\theta)\, d\theta.$$

Marginal Likelihood

The marginal likelihood of the observed data is the prior predictive evaluated at $x_{1:n}$:

$$m(x_{1:n}) = \int_\Theta p(x_{1:n} \mid \theta)\, \pi(\theta)\, d\theta = p(x_{1:n}).$$

This quantity appears in the denominator of Bayes' theorem (Node 6.1) and in Bayes factors (Node 6.6).

Intuition

The posterior predictive averages the sampling model over all parameter values, weighted by their posterior plausibility. It does not commit to a single estimate of θ; instead, it propagates the full posterior uncertainty into the prediction.

Compared to plug-in prediction $p(\tilde{x} \mid \hat{\theta})$, which conditions on a single estimate, the posterior predictive is typically more dispersed. This extra spread reflects parameter uncertainty, which the plug-in approach ignores. The posterior predictive is the honest assessment of predictive uncertainty given the model and data.

The prior predictive distribution has a dual role: it describes the marginal behavior of the data before any conditioning, and its evaluation at the observed data yields the marginal likelihood — the key ingredient in Bayesian model comparison.

Structural Role

The posterior predictive distribution requires the posterior (Node 6.1) and marginalization (Node 1.7). Conjugate families (Node 6.2) provide closed-form solutions in standard cases.

This node enables: Bayes factors (Node 6.6) through the marginal likelihood, and posterior predictive checks for model adequacy. It connects to the decision-theoretic framework: under squared error loss, the optimal point prediction of $\tilde{X}$ is $\mathbb{E}[\tilde{X} \mid x_{1:n}]$, computed from the posterior predictive.

Mathematical Development

Derivation as Marginalization

The joint distribution of $(\tilde{X}, \theta)$ given data is

$$p(\tilde{x}, \theta \mid x_{1:n}) = p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n}),$$

using the conditional independence $\tilde{X} \perp X_{1:n} \mid \theta$. Marginalizing over θ:

$$p(\tilde{x} \mid x_{1:n}) = \int_\Theta p(\tilde{x}, \theta \mid x_{1:n})\, d\theta = \int_\Theta p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n})\, d\theta.$$

Posterior Predictive Moments

The mean of the posterior predictive is

$$\mathbb{E}[\tilde{X} \mid x_{1:n}] = \int \tilde{x}\, p(\tilde{x} \mid x_{1:n})\, d\tilde{x} = \mathbb{E}_{\theta \mid x}\big[\mathbb{E}[\tilde{X} \mid \theta]\big] = \mathbb{E}[\mu(\theta) \mid x_{1:n}],$$

where $\mu(\theta) = \mathbb{E}[\tilde{X} \mid \theta]$. By the law of total variance:

$$\mathrm{Var}(\tilde{X} \mid x_{1:n}) = \mathbb{E}_{\theta \mid x}\big[\mathrm{Var}(\tilde{X} \mid \theta)\big] + \mathrm{Var}_{\theta \mid x}\big[\mathbb{E}[\tilde{X} \mid \theta]\big].$$

The first term is the average sampling variance (aleatoric uncertainty). The second term is the posterior variance of the conditional mean (epistemic uncertainty). The posterior predictive variance is always at least as large as the plug-in predictive variance $\mathrm{Var}(\tilde{X} \mid \hat{\theta})$.

Conjugate Posterior Predictives

Beta-Binomial: Let $X_i \mid \theta \sim \mathrm{Bernoulli}(\theta)$, $\pi(\theta) = \mathrm{Beta}(\alpha_0, \beta_0)$. After s successes in n trials, the posterior is $\mathrm{Beta}(\alpha_n, \beta_n)$ with $\alpha_n = \alpha_0 + s$, $\beta_n = \beta_0 + n - s$. The posterior predictive for a single new trial is

$$p(\tilde{X} = 1 \mid x_{1:n}) = \mathbb{E}[\theta \mid x_{1:n}] = \frac{\alpha_n}{\alpha_n + \beta_n}.$$

For m future trials, the number of successes $\tilde{S}$ follows a Beta-Binomial distribution:

$$p(\tilde{S} = k \mid x_{1:n}) = \binom{m}{k} \frac{B(\alpha_n + k,\ \beta_n + m - k)}{B(\alpha_n, \beta_n)}, \quad k = 0, 1, \ldots, m.$$

Normal-Normal: Let $X_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, and posterior $\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$. The posterior predictive is

$$\tilde{X} \mid x_{1:n} \sim \mathcal{N}(\mu_n,\ \sigma^2 + \tau_n^2).$$

The predictive variance $\sigma^2 + \tau_n^2$ exceeds the sampling variance σ^2 by the posterior uncertainty τ_n^2. As $n \to \infty$, $\tau_n^2 \to 0$ and the predictive distribution approaches the plug-in distribution.

Gamma-Poisson: Let $X_i \mid \lambda \sim \text{Poisson}(\lambda)$, posterior $\lambda \mid x \sim \text{Gamma}(\alpha_n, \beta_n)$. The posterior predictive for a single new observation is Negative Binomial:

$$\tilde{X} \mid x_{1:n} \sim \text{NegBin}\left(\alpha_n,\ \frac{\beta_n}{\beta_n + 1}\right).$$

Posterior Predictive Checks

A **posterior predictive check** assesses model adequacy by comparing observed data summaries to their distribution under the posterior predictive. Define a test statistic $T(x)$. Generate replicated data $x^{\text{rep}} \sim p(x^{\text{rep}} \mid x_{1:n})$ by:

1. Draw $\theta^{(j)} \sim \pi(\theta \mid x_{1:n})$.
2. Draw $x^{\text{rep},(j)} \sim p(x \mid \theta^{(j)})$.

The **posterior predictive p-value** is

$$p_B = \Pr\big(T(x^{\text{rep}}) \geq T(x_{\text{obs}}) \mid x_{1:n}\big).$$

Values near 0 or 1 indicate model inadequacy: the observed summary is extreme relative to what the fitted model predicts.

Marginal Likelihood via Sequential Prediction

The marginal likelihood decomposes as a product of one-step-ahead predictive densities:

$$m(x_{1:n}) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}),$$

where $p(x_1 \mid x_0) = p(x_1)$ is the prior predictive. This decomposition is the **prequential** form of the marginal likelihood and connects to information-theoretic model evaluation.

Worked Examples

Example 1: Normal Posterior Predictive

Data: $n = 16$, $\bar{x} = 50$, $\sigma^2 = 64$ (known). Prior: $\mu \sim \mathcal{N}(45, 16)$.

Posterior: $\mu \mid x \sim \mathcal{N}(\mu_n, \tau_n^2)$.

$$\tau_n^2 = \frac{1}{1/16 + 16/64} = \frac{1}{0.0625 + 0.25} = \frac{1}{0.3125} = 3.2.$$

$$\mu_n = 3.2 \cdot (0.0625 \cdot 45 + 0.25 \cdot 50) = 3.2 \cdot (2.8125 + 12.5) = 3.2 \cdot 15.3125 = 49.0.$$

Posterior predictive: $\tilde{X} \mid x \sim \mathcal{N}(49.0,\ 64 + 3.2) = \mathcal{N}(49.0,\ 67.2)$.

Predictive standard deviation: $\sqrt{67.2} \approx 8.20$.

A 95% posterior predictive interval for a new observation: $49.0 \pm 1.96 \cdot 8.20 = [32.93,\ 65.07]$.

Compare to plug-in prediction at MLE $\hat{\mu} = 50$: $\mathcal{N}(50, 64)$, giving interval $[34.32, 65.68]$. The Bayesian interval is centered slightly differently and accounts for parameter uncertainty.

Example 2: Beta-Binomial Posterior Predictive

Data: $n = 50$, $s = 30$. Prior: Beta$(1, 1)$.

Posterior: Beta$(31, 21)$. Posterior mean: $31/52 \approx 0.596$.

Posterior predictive probability of success on the next trial: $31/52 \approx 0.596$.

For $m = 10$ future trials, the posterior predictive distribution of the number of successes is Beta-Binomial$(10, 31, 21)$. The predictive mean is $10 \cdot 31/52 \approx 5.96$. The predictive variance is

$$\mathrm{Var}(\tilde{S}) = m \cdot \frac{\alpha_n \beta_n (\alpha_n + \beta_n + m)}{(\alpha_n + \beta_n)^2 (\alpha_n + \beta_n + 1)} = 10 \cdot \frac{31 \cdot 21 \cdot 62}{52^2 \cdot 53} \approx 2.81.$$

Compare to the binomial variance at the plug-in $\hat{\theta} = 0.6$: $10 \cdot 0.6 \cdot 0.4 = 2.4$. The posterior predictive variance is larger, reflecting parameter uncertainty.

Cross-Links

- **Linear Algebra:** In Gaussian linear models, the posterior predictive at a new input $\tilde{x}$ is a quadratic form involving the posterior covariance of the coefficient vector, connecting matrix algebra of precision updates to predictive uncertainty.
- **AI:** Gaussian process regression computes the posterior predictive distribution at new inputs in closed form — it is the canonical example of an exact posterior predictive under a nonparametric prior. Bayesian neural networks approximate the posterior predictive via MC dropout or ensembles.
- **Economics:** The posterior predictive distribution is the natural object for Bayesian forecasting: given observed economic data, it provides the full distribution of future GDP, inflation, or asset returns, incorporating parameter uncertainty from the structural model.
- **Quantum Mechanics:** In quantum Bayesian inference (QBism), the predictive distribution for a future measurement outcome given past measurement results is obtained by marginalizing over the posterior distribution of the quantum state, exactly paralleling the posterior predictive.

Structural Compression

Invariant: $p(\tilde{x} \mid x_{1:n}) = \int p(\tilde{x} \mid \theta)\, \pi(\theta \mid x_{1:n})\, d\theta$. Prediction marginalizes over posterior uncertainty in the parameter.

Mechanism: The posterior reweights the parameter space according to data compatibility. The predictive integrates the sampling model against this reweighted measure, producing a distribution over new observations that is strictly more dispersed than the plug-in predictive, with the excess variance equal to the posterior variance of the conditional mean.

Cross-System Echo: The posterior predictive is the statistical instance of Bayesian model averaging. In ensemble methods, the averaged prediction across models implements the same marginalization. The marginal likelihood decomposition into sequential one-step predictives connects to online learning and prequential analysis.

Exercises

1. Derive the posterior predictive distribution for the Gamma-Poisson model: show that the posterior predictive for a single new observation is Negative Binomial, and identify its parameters in terms of the posterior hyperparameters.
2. For the Normal-Normal model, prove that the posterior predictive variance equals $\sigma^2 + \tau_n^2$ and decompose this into aleatoric and epistemic components. Show that the epistemic component vanishes as $n \to \infty$.
3. Compute the marginal likelihood $m(x_{1:n})$ for the Beta-Binomial model in closed form using the Beta function. Verify that it equals the product of sequential one-step predictive densities.
4. Design a posterior predictive check for the Poisson model: specify a test statistic $T(x)$ that would detect overdispersion, and describe the procedure for computing the posterior predictive p-value via simulation.
5. Show that the posterior predictive mean $\mathbb{E}[\tilde{X} \mid x_{1:n}]$ equals the posterior mean of $\mathbb{E}[\tilde{X} \mid \theta]$. Under what conditions does the posterior predictive median equal the posterior median of θ?
6. For the Normal-Normal model with prior $\mu \sim \mathcal{N}(0, \tau_0^2)$ and known σ^2, compute the marginal likelihood $m(x_{1:n})$ in closed form and show it is proportional to $\exp\left(-\frac{n\bar{x}^2}{2(\sigma^2 + n\tau_0^2)}\right)$.
7. Argue, structurally, why the posterior predictive distribution is the unique coherent predictive distribution given the model and data: identify the assumptions (conditional independence, coherent updating) that force the marginalization over the posterior, and explain why plug-in prediction violates coherence.

Bayes Factors

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.6

Bayes factors are the Bayesian instrument for comparing models or hypotheses. They quantify the relative evidence the data provide for one model over another, on the scale of marginal likelihoods. This node depends on the posterior and marginal likelihood (Node 6.1), Bayesian estimation (Node 6.3), and the likelihood function (Node 2.3). It connects to the Neyman-Pearson framework (Node 5.2) as a special case and enables the Bernstein–von Mises reconciliation (Node 6.7) by establishing the role of the marginal likelihood in model evaluation.

Formal Definition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two statistical models with parameter spaces Θ_0 and Θ_1, likelihoods $p_0(x \mid \theta_0)$ and $p_1(x \mid \theta_1)$, and priors $\pi_0(\theta_0)$ and $\pi_1(\theta_1)$.

Marginal Likelihood

The **marginal likelihood** under model $\mathcal{M}_j$ is

$$m_j(x) = \int_{\Theta_j} p_j(x \mid \theta_j)\, \pi_j(\theta_j)\, d\theta_j.$$

Bayes Factor

The **Bayes factor** in favor of $\mathcal{M}_1$ against $\mathcal{M}_0$ is

$$\mathrm{BF}_{10}(x) = \frac{m_1(x)}{m_0(x)}.$$

Posterior Model Odds

Given prior model probabilities $\pi(\mathcal{M}_j)$, the posterior odds satisfy

$$\frac{\pi(\mathcal{M}_1 \mid x)}{\pi(\mathcal{M}_0 \mid x)} = \mathrm{BF}_{10}(x) \cdot \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_0)}.$$

The Bayes factor transforms prior odds into posterior odds. It is the data's contribution to model comparison, separated from the prior model weights.

Intuition

The Bayes factor is a likelihood ratio averaged over parameter uncertainty. Unlike the frequentist likelihood ratio, which evaluates each model at its best-fitting parameter value, the Bayes factor evaluates each model at its average performance over the prior. This builds in an automatic complexity penalty: a model with a diffuse prior over many parameter values will spread its marginal likelihood thinly, while a model that concentrates prior mass near the data-generating region will score well.

The Bayes factor does not require choosing a significance level, does not depend on the sampling distribution of a test statistic, and can provide evidence in favor of the null hypothesis — not merely a failure to reject.

Structural Role

The Bayes factor is the coherent Bayesian update factor for model probabilities. It requires marginal likelihoods (Node 6.1) and connects to the likelihood ratio test (Node 5.2) as a special case for simple hypotheses.

It depends on prior specification over parameters within each model. Improper priors within models render the Bayes factor undefined (the marginal likelihood is defined only up to an arbitrary constant), so model comparison requires proper priors or careful calibration.

This node enables: the Bernstein–von Mises theorem (Node 6.7), which provides asymptotic approximations to the marginal likelihood and hence to Bayes factors via the BIC.

Mathematical Development

Relation to Neyman-Pearson Likelihood Ratio

For simple hypotheses $H_0 : P = P_0$, $H_1 : P = P_1$ (no free parameters):

$$\mathrm{BF}_{10}(x) = \frac{p_1(x)}{p_0(x)},$$

which is exactly the likelihood ratio of the Neyman-Pearson lemma (Node 5.2). The Bayes factor generalizes the likelihood ratio to composite hypotheses by integrating over parameters.

Jeffreys' Interpretive Scale

Harold Jeffreys proposed the following scale: $\log_{10} \mathrm{BF}_{10} \in (0, 1/2]$ is "barely worth mentioning," $(1/2, 1]$ is "substantial," $(1, 3/2]$ is "strong," $(3/2, 2]$ is "very strong," and > 2 is "decisive." This is a guideline, not a formal decision rule.

Savage-Dickey Density Ratio

For nested models where $\mathcal{M}_0$ is $\mathcal{M}_1$ with the restriction $\psi = \psi_0$ (a subvector of the parameter), the Bayes factor simplifies to

$$\mathrm{BF}_{01}(x) = \frac{\pi_1(\psi_0 \mid x)}{\pi_1(\psi_0)},$$

where $\pi_1(\psi_0 \mid x)$ is the marginal posterior density of ψ under $\mathcal{M}_1$ evaluated at ψ_0, and $\pi_1(\psi_0)$ is the corresponding prior density. This avoids computing the marginal likelihoods directly.

Derivation. Write $\theta_1 = (\psi, \lambda)$ where λ is the nuisance parameter. Under $\mathcal{M}_0$, $\psi = \psi_0$ is fixed and

$$m_0(x) = \int p(x \mid \psi_0, \lambda)\, \pi_0(\lambda)\, d\lambda.$$

Under $\mathcal{M}_1$, the marginal posterior of ψ evaluated at ψ_0 is

$$\pi_1(\psi_0 \mid x) = \frac{\int p(x \mid \psi_0, \lambda)\, \pi_1(\lambda \mid \psi_0)\, d\lambda \cdot \pi_1(\psi_0)}{m_1(x)}.$$

If the conditional prior $\pi_1(\lambda \mid \psi_0) = \pi_0(\lambda)$ (the priors on nuisance parameters agree), then

$$\pi_1(\psi_0 \mid x) = \frac{m_0(x)\, \pi_1(\psi_0)}{m_1(x)},$$

so $\mathrm{BF}_{01} = m_0/m_1 = \pi_1(\psi_0 \mid x)/\pi_1(\psi_0)$. $\square$

BIC Approximation to the Bayes Factor

The Schwarz (BIC) approximation to the log marginal likelihood is

$$\log m_j(x) \approx \ell_j(\hat{\theta}_j) - \frac{d_j}{2} \log n,$$

where $\ell_j(\hat{\theta}_j)$ is the maximized log-likelihood, $d_j = \dim(\Theta_j)$, and n is the sample size.

Derivation. Apply the Laplace approximation to the marginal likelihood integral. The log-likelihood near the MLE is

$$\ell(\theta) \approx \ell(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^\top \hat{J} (\theta - \hat{\theta}),$$

where $\hat{J} = -\nabla^2 \ell(\hat{\theta})$. The marginal likelihood becomes

$$m(x) \approx p(x \mid \hat{\theta})\, \pi(\hat{\theta})\, (2\pi)^{d/2}\, |\hat{J}|^{-1/2}.$$

Taking logarithms: $\log m(x) \approx \ell(\hat{\theta}) + \log \pi(\hat{\theta}) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\hat{J}|$. Since $\hat{J} \approx n\, I(\theta_0)$, $\log|\hat{J}| \approx d \log n + \log|I(\theta_0)|$. Dropping $O(1)$ terms (which include the prior and the Fisher information determinant) leaves

$$\log m(x) \approx \ell(\hat{\theta}) - \frac{d}{2} \log n + O(1).$$

The BIC approximation to the log Bayes factor is therefore

$$\log \mathrm{BF}_{10} \approx \big[\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)\big] - \frac{d_1 - d_0}{2} \log n.$$

Lindley's Paradox

Consider testing $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$ with $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ and prior $\mu \sim \mathcal{N}(0, \tau^2)$ under H_1.

The Bayes factor in favor of H_0 is

$$\mathrm{BF}_{01} = \sqrt{1 + n\tau^2}\, \exp\!\left(-\frac{n\bar{x}^2}{2} \cdot \frac{n\tau^2}{1 + n\tau^2} \right).$$

Fix $\bar{x} = z/\sqrt{n}$ (so that the z-statistic is fixed at some value that would reject H_0 in a frequentist test). As $n \to \infty$ with z fixed, $\mathrm{BF}_{01} \to \infty$: the Bayes factor increasingly favors H_0, even though the frequentist test rejects. This is **Lindley's paradox**: at any fixed significance level, there exist sample sizes for which the data reject H_0 but the Bayes factor favors H_0.

The resolution: the frequentist test and the Bayes factor answer different questions. The p-value measures tail probability; the Bayes factor measures average likelihood.

Worked Examples

Example 1: Two Poisson Rates

$\mathcal{M}_0$: $X_1, \ldots, X_n \sim \text{Poisson}(3)$ (simple).

$\mathcal{M}_1$: $X_i \sim \text{Poisson}(\lambda)$, $\lambda \sim \text{Gamma}(3, 1)$.

Data: $n = 20$, $\sum x_i = 72$, $\bar{x} = 3.6$.

$m_0 = \prod_{i=1}^{20} \frac{3^{x_i} e^{-3}}{x_i!} = \frac{3^{72} e^{-60}}{\prod x_i!}$.

$m_1 = \int_0^\infty \frac{\lambda^{72} e^{-20\lambda}}{\prod x_i!} \cdot \frac{\lambda^2 e^{-\lambda}}{\Gamma(3)}\, d\lambda = \frac{1}{\prod x_i! \cdot \Gamma(3)} \int_0^\infty \lambda^{74} e^{-21\lambda}\, d\lambda = \frac{\Gamma(75)}{\prod x_i! \cdot \Gamma(3) \cdot 21^{75}}$.

The Bayes factor $\text{BF}_{10} = m_1/m_0$. Taking the log ratio:

$$\log \text{BF}_{10} = \log \Gamma(75) - \log \Gamma(3) - 75 \log 21 - 72 \log 3 + 60.$$

Numerical evaluation gives $\log \text{BF}_{10} \approx -1.8$, so $\text{BF}_{10} \approx 0.17$. The data moderately favor $\mathcal{M}_0$ (the simple Poisson(3) model) over the Gamma-Poisson model, since the observed rate 3.6 is close to 3.

Example 2: Savage-Dickey for a Normal Mean

Test $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. Under H_1: $X_i \sim \mathcal{N}(\mu, 1)$, $\mu \sim \mathcal{N}(0, \tau^2)$ with $\tau^2 = 10$.

Data: $n = 25$, $\bar{x} = 0.8$.

Posterior under H_1: $\mu \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$ with $\sigma_n^2 = (1/10 + 25)^{-1} = 1/25.1 \approx 0.0398$, $\mu_n = 0.0398 \cdot (0 + 25 \cdot 0.8) = 0.796$.

Prior density at 0: $\pi_1(0) = (2\pi \cdot 10)^{-1/2} = 0.0399/\sqrt{2\pi} \approx 0.1262$.

Posterior density at 0: $\pi_1(0 \mid x) = (2\pi \cdot 0.0398)^{-1/2} \exp(-0.796^2/(2 \cdot 0.0398)) = 3.989 \cdot \exp(-7.96) \approx 3.989 \cdot 0.000349 \approx 0.00139$.

Savage-Dickey: $\text{BF}_{01} = 0.00139/0.1262 \approx 0.011$.

Thus $\text{BF}_{10} \approx 91$: strong evidence against H_0.

Cross-Links

- **Linear Algebra:** The Laplace approximation to the marginal likelihood involves the determinant of the observed information matrix $|\hat{J}|$, connecting model comparison to the spectral properties of the Hessian of the log-likelihood.
- **AI:** Bayesian model selection via Bayes factors underlies model comparison in Gaussian process regression (choosing kernel hyperparameters) and Bayesian neural architecture search. The marginal likelihood is the "type-II likelihood" maximized in empirical Bayes.
- **Economics:** Bayesian model comparison using Bayes factors is standard in macroeconomics for comparing DSGE models. The marginal likelihood penalizes overfit, addressing the same concern as information criteria.

- **Quantum Mechanics:** Quantum hypothesis testing (discriminating between quantum states) uses a quantum analogue of the Bayes factor, where the marginal likelihoods are replaced by traces of the state operators against measurement outcomes.

Structural Compression

Invariant: $\mathrm{BF}_{10}(x) = m_1(x)/m_0(x)$. The ratio of marginal likelihoods is the coherent Bayesian update factor for model probabilities.

Mechanism: Each marginal likelihood averages the data likelihood over prior uncertainty in the parameters. The ratio compares these averages, automatically penalizing model complexity through prior dispersion: a model that spreads prior mass over regions of low likelihood is penalized.

Cross-System Echo: Bayes factors generalize the Neyman-Pearson likelihood ratio to composite hypotheses. The BIC approximation $2 \log \mathrm{BF} \approx 2\Delta\ell - \Delta d \cdot \log n$ connects Bayesian model comparison to penalized likelihood criteria (AIC, BIC). In coding theory, the marginal likelihood is the inverse of the minimum description length under the prior.

Exercises

1. Derive the Bayes factor for testing $H_0 : \theta = 1/2$ vs $H_1 : \theta \neq 1/2$ in the Binomial model with prior $\mathrm{Beta}(1, 1)$ under H_1. Compute numerically for $n = 100$, $s = 55$.
2. Prove the Savage-Dickey density ratio for a one-dimensional parameter of interest ψ, starting from the marginal likelihood ratio and the definition of the marginal posterior.
3. Derive the BIC approximation to $\log m(x)$ via the Laplace method, keeping all terms through $O(1)$ before dropping to the $O(\log n)$ approximation. Identify precisely which terms are absorbed into the $O(1)$ remainder.
4. Demonstrate Lindley's paradox numerically: for the normal mean test with $\tau^2 = 1$ and a fixed z-statistic of 2.5, compute BF_{01} for $n = 10, 100, 1000, 10000$ and verify the Bayes factor increasingly favors H_0.
5. Show that the Bayes factor is symmetric: $\mathrm{BF}_{01} = 1/\mathrm{BF}_{10}$, and that this property fails for p-values (a p-value against H_0 is not the complement of a p-value against H_1).
6. For two nested normal linear regression models with d_0 and $d_1 > d_0$ predictors, express the BIC-approximated Bayes factor in terms of the residual sums of squares and the sample size. Show that BIC penalizes additional parameters more heavily than AIC when $n \geq 8$.
7. Argue, structurally, why improper priors make the Bayes factor undefined: identify where the arbitrary normalizing constant enters and explain why the prior predictive interpretation of the marginal likelihood fails when the prior is not a probability measure.

Bernstein–von Mises Theorem

Position in Knowledge Graph

System: Statistics

Structural Domain: Bayesian Inference

Node: 6.7

The Bernstein–von Mises theorem is the fundamental asymptotic bridge between the Bayesian and frequentist frameworks. It establishes that in regular parametric models, the posterior distribution is asymptotically Gaussian, centered at the MLE, with variance equal to the inverse Fisher information — regardless of the prior. This node depends on the posterior (Node 6.1), the asymptotic theory of the MLE (Node 3.3), and the Fisher information (Node 7.2). It is the terminal node of Module 6: it closes the loop by showing that Bayesian and frequentist inference converge in large samples.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$ where $\theta_0 \in \Theta \subseteq \mathbb{R}^k$ is the true parameter value. Let $\hat{\theta}_n$ denote the MLE and $I(\theta_0)$ the Fisher information matrix at θ_0.

Regularity Conditions

1. **Identifiability:** $\theta \neq \theta' \implies p_\theta \neq p_{\theta'}$.
2. **Smoothness:** The log-likelihood $\ell(\theta) = \log p(x \mid \theta)$ is twice continuously differentiable in a neighborhood of θ_0.
3. **Fisher information:** $I(\theta_0) = -\mathbb{E}_{\theta_0}[\nabla^2 \ell(\theta_0)]$ is finite and positive definite.
4. **Prior positivity:** π is continuous and positive at θ_0: $\pi(\theta_0) > 0$.
5. **Consistency:** There exists a consistent sequence of estimators (e.g., the MLE $\hat{\theta}_n \to \theta_0$ in probability).
6. **Local asymptotic normality (LAN):** The log-likelihood ratio admits the expansion $\ell_n(\theta_0 + h/\sqrt{n}) - \ell_n(\theta_0) = h^\top \Delta_n - \frac{1}{2} h^\top I(\theta_0) h + o_P(1)$, where $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$.

Theorem Statement

Under the regularity conditions above:

$$\left\| \pi\Big(\sqrt{n}(\theta - \hat{\theta}_n) \in \cdot \;\Big|\; X_1, \ldots, X_n\Big) - \mathcal{N}\big(0,\ I(\theta_0)^{-1}\big) \right\|_{\text{TV}} \xrightarrow{P^n_{\theta_0}} 0.$$

Equivalently, the posterior distribution of θ given the data satisfies

$$\pi(\theta \mid X_1, \ldots, X_n) \approx \mathcal{N}\left(\hat{\theta}_n,\ \frac{I(\theta_0)^{-1}}{n}\right)$$

in the sense of total variation convergence under the true data-generating distribution.

Intuition

As data accumulate, the likelihood dominates the prior. The log-posterior is approximately

$$\log \pi(\theta \mid x) \approx \ell_n(\theta) + \log \pi(\theta) - \text{const},$$

where $\ell_n(\theta) = \sum_{i=1}^{n} \log p(x_i \mid \theta)$ grows as $O(n)$ while $\log \pi(\theta)$ remains $O(1)$. A Taylor expansion of ℓ_n around the MLE produces a quadratic in θ, which is the kernel of a Gaussian.

The theorem says the prior washes out: any two investigators with different priors (both positive and continuous at θ_0) will have posteriors that converge to the same Gaussian as $n \to \infty$. This is the Bayesian analogue of the frequentist result that the MLE is asymptotically normal.

Structural Role

The Bernstein–von Mises theorem is the asymptotic reconciliation of Bayesian and frequentist inference. It depends on the posterior (Node 6.1), asymptotic MLE theory (Node 3.3), and Fisher information (Node 7.2).

It has consequences for every other node in Module 6: - **Node 6.3:** The posterior mean, MAP, and posterior median all converge to the MLE at rate $O_p(n^{-1/2})$. - **Node 6.4:** Credible sets have asymptotic frequentist coverage. - **Node 6.6:** The BIC approximation to the Bayes factor is derived from the same Laplace expansion.

The theorem fails in nonparametric models, misspecified models, and high-dimensional settings, delineating the boundary of Bayesian-frequentist agreement.

Mathematical Development

Proof Sketch

The argument proceeds in four steps.

Step 1: Posterior concentration. By the consistency of the MLE and the positivity of the prior at θ_0, the posterior mass outside any ball $B(\theta_0, \varepsilon)$ converges to zero in probability:

$$\pi\big(\|\theta - \theta_0\| > \varepsilon \mid X_{1:n}\big) \xrightarrow{P} 0.$$

This follows from the exponential growth of the likelihood ratio at the true value relative to distant parameter values.

Step 2: Log-posterior expansion. Expand the log-likelihood around the MLE:

$$\ell_n(\theta) = \ell_n(\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{J}_n (\theta - \hat{\theta}_n) + R_n(\theta),$$

where $\hat{J}_n = -\nabla^2 \ell_n(\hat{\theta}_n)$ is the observed information and $R_n(\theta)$ is a remainder. By the law of large numbers, $\hat{J}_n / n \to I(\theta_0)$ in probability.

Step 3: Prior contribution. On the concentration set $\|\theta - \hat{\theta}_n\| = O(n^{-1/2})$, the prior satisfies $\log \pi(\theta) = \log \pi(\hat{\theta}_n) + O(n^{-1/2})$. Since $\log \pi$ is $O(1)$ and the quadratic term is $O(n)$, the prior contributes a negligible additive correction.

Step 4: Gaussian identification. Combining Steps 2 and 3:

$$\log \pi(\theta \mid x) = -\frac{1}{2}(\theta - \hat{\theta}_n)^\top \hat{J}_n (\theta - \hat{\theta}_n) + o_P(1)$$

on the set where $\sqrt{n}(\theta - \hat{\theta}_n)$ is bounded. This is the log-density of $\mathcal{N}(\hat{\theta}_n, \hat{J}_n^{-1})$. Since $\hat{J}_n / n \to I(\theta_0)$, the posterior is asymptotically $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$ in total variation. $\square$

Consequence: Posterior Mean and MLE Agreement

Since the posterior is approximately $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, its mean is

$$\mathbb{E}[\theta \mid X_{1:n}] = \hat{\theta}_n + O_P(n^{-1}).$$

The posterior mean and the MLE agree to first order. Their difference is $O_P(n^{-1})$, a second-order correction driven by the prior and the curvature of the likelihood.

Consequence: Frequentist Coverage of Credible Sets

Let $C_n(x)$ be any $(1-\alpha)$-credible set derived from the posterior. By the total variation convergence:

$$\pi(\theta \in C_n \mid x) \to 1-\alpha \implies \Phi\left(\frac{\sqrt{n}\, I(\theta_0)^{1/2}(\theta_0 - \hat{\theta}_n)}{\text{boundary}}\right) \to 1-\alpha.$$

Since $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$ under P_{θ_0}, the event $\theta_0 \in C_n(X)$ has probability converging to $1-\alpha$:

$$P_{\theta_0}\big(\theta_0 \in C_n(X)\big) \to 1-\alpha.$$

Asymptotically, Bayesian credible sets are frequentist confidence sets.

Consequence: HPD and Wald Interval Agreement

The HPD credible set based on the limiting Gaussian posterior is the ellipsoid

$$\big\{\theta : n(\theta - \hat{\theta}_n)^\top I(\theta_0)(\theta - \hat{\theta}_n) \le \chi^2_{k,1-\alpha}\big\},$$

which coincides with the Wald confidence ellipsoid based on the MLE.

Failure Modes

Nonparametric models. When Θ is infinite-dimensional (e.g., density estimation), the posterior can concentrate at rate slower than $1/\sqrt{n}$, and the Gaussian approximation fails. The Diaconis-Freedman theorem provides examples where the posterior concentrates on the wrong value.

Misspecified models. If the true distribution $P_0 \notin \{P_\theta : \theta \in \Theta\}$, the posterior concentrates around the KL-minimizer $\theta^* = \arg\min_\theta \mathrm{KL}(P_0 \| P_\theta)$, but the variance is no longer $I(\theta^*)^{-1}/n$ — a sandwich-type correction is needed.

High-dimensional models. When $k = k_n \to \infty$ with n, the prior contribution does not vanish relative to the likelihood, and the Bernstein–von Mises conclusion can fail even for parametric models.

Boundary parameters. If θ_0 lies on the boundary of Θ, the posterior may not be asymptotically normal (it may concentrate on a half-space).

Worked Examples

Example 1: Normal Mean

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with prior $\mu \sim \mathcal{N}(0, \tau^2)$. The exact posterior is $\mu \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$ with

$$\sigma_n^2 = \frac{1}{1/\tau^2 + n} = \frac{\tau^2}{1 + n\tau^2}, \qquad \mu_n = \sigma_n^2 \cdot n\bar{x} = \frac{n\tau^2}{1 + n\tau^2}\bar{x}.$$

As $n \to \infty$: $\sigma_n^2 \to 1/n$ and $\mu_n \to \bar{x} = \hat{\mu}_n$. The Fisher information is $I(\mu) = 1$, so the BvM limit is $\mathcal{N}(\hat{\mu}_n, 1/n)$. The exact posterior converges to this limit, confirming the theorem. The convergence rate of the prior correction is $O(1/n)$: the term $1/\tau^2$ in the posterior precision becomes negligible relative to n.

Example 2: Bernoulli Model and Coverage Verification

Let $X_i \sim \text{Bernoulli}(\theta_0)$ with $\theta_0 = 0.3$. Prior: Beta$(2, 2)$. For large n, the posterior is approximately

$$\theta \mid x \dot{\sim} \mathcal{N}\left(\hat{\theta}_n,\ \frac{\hat{\theta}_n(1 - \hat{\theta}_n)}{n}\right),$$

since $I(\theta) = 1/[\theta(1 - \theta)]$.

For $n = 500$: if $s = 155$ successes, then $\hat{\theta}_n = 0.31$. The asymptotic posterior variance is $0.31 \cdot 0.69/500 \approx 0.000428$. The 95% HPD interval is approximately $0.31 \pm 1.96\sqrt{0.000428} = 0.31 \pm 0.041 = [0.269, 0.351]$.

The exact posterior is Beta$(157, 347)$ with mean $157/504 \approx 0.3115$ and variance $157 \cdot 347/(504^2 \cdot 505) \approx 0.000429$. The Gaussian approximation is excellent: the prior contributes only 4 pseudo-observations out of 504 total.

The frequentist coverage: $\theta_0 = 0.3 \in [0.269, 0.351]$, and repeated sampling confirms coverage near 95%.

Cross-Links

- **Linear Algebra:** The quadratic expansion of the log-likelihood around the MLE is governed by the Hessian matrix $\hat{J}_n$. The Bernstein–von Mises theorem requires this matrix to converge (after scaling by $1/n$) to the positive definite Fisher information matrix, a spectral condition.
- **AI:** The Bernstein–von Mises theorem motivates Laplace approximations for Bayesian deep learning: approximate the posterior by a Gaussian centered at the MAP estimate with covariance given by the inverse Hessian. The theorem's failure in high dimensions explains why this approximation is unreliable for overparameterized networks.
- **Economics:** The theorem justifies the common econometric practice of interpreting Bayesian credible intervals as confidence intervals in large samples. It also explains why prior choice matters less in large macro panels than in small cross-sections.
- **Quantum Mechanics:** Quantum local asymptotic normality (quantum LAN) extends the Bernstein–von Mises theorem to quantum statistical models, showing that the posterior over quantum states converges to a Gaussian shift experiment in the limit of many copies of the state.

Structural Compression

Invariant: In regular parametric models, the large-sample posterior is asymptotically $\mathcal{N}(\hat{\theta}_n, I(\theta_0)^{-1}/n)$, independent of the prior.

Mechanism: A second-order Taylor expansion of the log-likelihood around the MLE produces an $O(n)$ quadratic that dominates the $O(1)$ prior contribution. The resulting Gaussian shape is determined entirely by the Fisher information geometry. The prior survives only in the $O(n^{-1})$ correction to the posterior mean.

Cross-System Echo: The Bernstein–von Mises theorem is the statistical manifestation of the Laplace approximation to integrals dominated by a sharp peak. In physics, the saddle-point approximation to partition

functions is the same phenomenon. In AI, the Laplace approximation to the neural network posterior and the correspondence between MAP and MLE in the large-data regime are direct applications of this asymptotic equivalence.

Exercises

1. Verify the Bernstein–von Mises conclusion for the Poisson model: let $X_i \sim \text{Poisson}(\lambda_0)$ with prior $\text{Gamma}(\alpha, \beta)$. Show that the exact posterior is Gamma and that it converges in total variation to $\mathcal{N}(\hat{\lambda}_n, \hat{\lambda}_n/n)$ as $n \to \infty$.
2. Show that the Bernstein–von Mises conclusion implies the posterior variance satisfies $\text{Var}(\theta \mid X) \to I(\theta_0)^{-1}/n$ in probability. Derive the rate of convergence of the difference.
3. Construct an explicit example where the Bernstein–von Mises theorem fails: take a uniform prior on a misspecified model where the true density is not in the parametric family, and show the posterior concentrates at the KL-minimizer but with incorrect variance.
4. For the normal location model with known variance and prior $\mathcal{N}(\mu_0, \tau_0^2)$, compute the exact total variation distance between the posterior and the BvM Gaussian limit as a function of n, and verify it is $O(n^{-1})$.
5. Prove that the second-order correction to the posterior mean is $\mathbb{E}[\theta \mid x] - \hat{\theta}_n = O_P(n^{-1})$ by expanding the posterior mean in powers of $1/n$ using the exact posterior for a conjugate exponential family.
6. Explain why the Bernstein–von Mises theorem fails in the model $X_i \sim \text{Uniform}(0, \theta)$: identify which regularity condition is violated, and describe the actual asymptotic behavior of the posterior.
7. Argue, structurally, why the Bernstein–von Mises theorem establishes a hierarchy between Bayesian and frequentist inference: in the regular parametric setting, the Bayesian approach subsumes the frequentist one (credible sets are asymptotically confidence sets, posterior mean agrees with MLE), but the theorem's failure modes in nonparametric, misspecified, and high-dimensional settings show where this subsumption breaks down and the two frameworks genuinely diverge.

Module 7 — Asymptotic Theory

Modes of Convergence

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.1 — Modes of Convergence.

Modes of convergence form the foundational language for all large-sample results in statistics. Every asymptotic statement — consistency, asymptotic normality, convergence of test statistics — is expressed through one of the convergence modes defined here. This node depends on the measure-theoretic probability framework (Node 1.8) and moment theory (Node 1.9), and enables the consistency (Node 7.2) and asymptotic normality (Node 7.3) results that follow.

Formal Definition

Let $(T_n)_{n\geq 1}$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking values in $\mathbb{R}^k$.

Convergence in probability. $T_n \xrightarrow{p} T$ if for every $\varepsilon > 0$,

$$\mathbb{P}(|T_n - T| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Almost sure convergence. $T_n \xrightarrow{a.s.} T$ if

$$\mathbb{P}\Big(\lim_{n\to\infty} T_n = T\Big) = 1.$$

Equivalently, for every $\varepsilon > 0$, $\mathbb{P}(\sup_{m\geq n} |T_m - T| > \varepsilon) \to 0$ as $n \to \infty$.

Convergence in distribution. $T_n \xrightarrow{d} T$ if for every continuity point t of the CDF F_T,

$$F_{T_n}(t) \to F_T(t) \quad \text{as } n \to \infty.$$

Equivalently, $\mathbb{E}[f(T_n)] \to \mathbb{E}[f(T)]$ for every bounded continuous function f (the Portmanteau characterization).

L^p convergence. For $p \geq 1$, $T_n \xrightarrow{L^p} T$ if $\mathbb{E}[|T_n - T|^p] \to 0$. The case $p = 2$ is mean-square convergence.

Intuition

Almost sure convergence is path-level: on almost every realization ω, the sequence $T_n(\omega)$ literally converges. Convergence in probability is weaker: the probability of a large deviation vanishes, but exceptional paths may misbehave infinitely often. Convergence in distribution is weakest: it says only that the laws of T_n approach the law of T, without requiring the random variables to be defined on the same space. L^p convergence controls the average magnitude of the deviation, and its strength depends on p.

The distinction matters because different statistical guarantees require different modes. Consistency of an estimator is convergence in probability; the central limit theorem gives convergence in distribution; the strong law of large numbers gives almost sure convergence.

Structural Role

Modes of convergence are the formal language for all large-sample statements in statistics. Consistency (Node 7.2) is convergence in probability of an estimator to the true parameter. Asymptotic normality (Node 7.3) is convergence in distribution of a standardized estimator to a Gaussian. Slutsky's theorem and the continuous mapping theorem, stated below, are the primary tools for combining and transforming convergent sequences to establish distributional limits of composite statistics. The hierarchy of modes determines which conclusions transfer between settings.

Mathematical Development

Theorem (Hierarchy of modes).

$$T_n \xrightarrow{a.s.} T \implies T_n \xrightarrow{p} T \implies T_n \xrightarrow{d} T.$$

$$T_n \xrightarrow{L^p} T \implies T_n \xrightarrow{p} T.$$

Proof that a.s. implies in probability. Fix $\varepsilon > 0$. Define $A_n = \{|T_n - T| > \varepsilon\}$. Almost sure convergence means $\mathbb{P}(\limsup_n A_n) = 0$. Since $\mathbb{P}(A_n) \le \mathbb{P}(\bigcup_{m \ge n} A_m)$ and $\mathbb{P}(\bigcup_{m \ge n} A_m) \downarrow \mathbb{P}(\limsup_n A_n) = 0$, we have $\mathbb{P}(A_n) \to 0$. □

Proof that L^p implies in probability. By Markov's inequality, $\mathbb{P}(|T_n - T| > \varepsilon) \le \varepsilon^{-p}\mathbb{E}[|T_n - T|^p] \to 0$. □

Proof that in probability implies in distribution. Fix a continuity point t of F_T. For any $\delta > 0$,

$$F_{T_n}(t) = \mathbb{P}(T_n \le t) \le \mathbb{P}(T \le t + \delta) + \mathbb{P}(|T_n - T| > \delta).$$

Similarly, $F_{T_n}(t) \ge \mathbb{P}(T \le t - \delta) - \mathbb{P}(|T_n - T| > \delta)$. Taking $n \to \infty$ and then $\delta \to 0$ using continuity of F_T at t yields $F_{T_n}(t) \to F_T(t)$. □

Converse: convergence in distribution to a constant. If $T_n \xrightarrow{d} c$ for a constant c, then $T_n \xrightarrow{p} c$. This follows because $F_c(t) = \mathbf{1}(t \ge c)$ has a single discontinuity, and pointwise convergence of CDFs at all continuity points forces the mass to concentrate at c.

Theorem (Continuous Mapping Theorem). If $T_n \xrightarrow{d} T$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous $\mathbb{P}_T$-almost everywhere, then $g(T_n) \xrightarrow{d} g(T)$. The same holds with $\xrightarrow{p}$ and $\xrightarrow{a.s.}$ in place of $\xrightarrow{d}$.

Proof sketch for the distributional case. Let D_g be the set of discontinuity points of g. By assumption, $\mathbb{P}(T \in D_g) = 0$. For any bounded continuous f, the composition $f \circ g$ is continuous on $\mathbb{R}^k \setminus D_g$. By the Portmanteau theorem extended to functions continuous a.e., $\mathbb{E}[f(g(T_n))] \to \mathbb{E}[f(g(T))]$. □

Theorem (Slutsky). If $T_n \xrightarrow{d} T$ and $S_n \xrightarrow{p} c$ for a constant c, then:

$$T_n + S_n \xrightarrow{d} T + c, \qquad T_n S_n \xrightarrow{d} cT, \qquad T_n / S_n \xrightarrow{d} T/c \ \ (c \neq 0).$$

Proof. Consider the map $g(t, s) = t + s$. The pair $(T_n, S_n) \xrightarrow{d} (T, c)$ since $S_n \xrightarrow{p} c$ (the joint convergence follows because convergence in probability to a constant makes the marginals asymptotically independent). Applying the continuous mapping theorem to g yields $T_n + S_n \xrightarrow{d} T + c$. The product and ratio cases follow analogously. □

Skorokhod representation theorem. If $T_n \xrightarrow{d} T$, there exist random variables T_n', T' on a common probability space with the same marginal distributions such that $T_n' \xrightarrow{a.s.} T'$. This converts distributional convergence to almost sure convergence on a possibly different probability space, often simplifying proofs.

Borel-Cantelli and almost sure convergence. The first Borel-Cantelli lemma states: if $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_n A_n) = 0$. Applied with $A_n = \{|T_n - T| > \varepsilon\}$: if $\sum_n \mathbb{P}(|T_n - T| > \varepsilon) < \infty$ for every $\varepsilon > 0$, then $T_n \xrightarrow{a.s.} T$. This provides a checkable sufficient condition for almost sure convergence from tail probability bounds.

Subsequence characterization. $T_n \xrightarrow{p} T$ if and only if every subsequence of (T_n) has a further subsequence that converges almost surely to T. This characterization is frequently used in proofs: to show convergence in probability, extract a subsequence via the diagonal argument and upgrade to almost sure convergence.

Dominated convergence and L^p convergence. If $T_n \xrightarrow{p} T$ and $|T_n|^p \le Y$ a.s. for some integrable Y, then $T_n \xrightarrow{L^p} T$. This is the probabilistic version of the dominated convergence theorem, and it provides the bridge from convergence in probability back up to L^p convergence when a dominating bound exists.

Multivariate extension. All four modes extend to $\mathbb{R}^k$-valued random vectors by replacing $|T_n - T|$ with $\|T_n - T\|$ (any norm; all norms on $\mathbb{R}^k$ are equivalent). The Cramer-Wold device characterizes convergence in distribution of random vectors: $T_n \xrightarrow{d} T$ in $\mathbb{R}^k$ if and only if $a^\top T_n \xrightarrow{d} a^\top T$ for every $a \in \mathbb{R}^k$. This reduces multivariate distributional convergence to checking one-dimensional convergence along all projections.

Worked Examples

Example 1. Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}[X_i] = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Set $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. By the weak law of large numbers, $\bar{X}_n \xrightarrow{p} \mu$. By the strong law, $\bar{X}_n \xrightarrow{a.s.} \mu$. Since $\mathbb{E}[(\bar{X}_n - \mu)^2] = \sigma^2/n \to 0$, we also have $\bar{X}_n \xrightarrow{L^2} \mu$. By the CLT, $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} \mathcal{N}(0,1)$. All four modes appear in this single example.

Example 2 (Slutsky application). Suppose $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and we estimate σ^2 by $S_n^2 = (n-1)^{-1}\sum(X_i - \bar{X}_n)^2$. Since $S_n^2 \xrightarrow{p} \sigma^2$, Slutsky gives $\sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{d} \mathcal{N}(0,1)$. This justifies the standard z-type confidence interval with estimated variance.

Example 3 (Continuous mapping). Let $T_n \xrightarrow{d} \mathcal{N}(0,1)$. Then $T_n^2 \xrightarrow{d} \chi^2_1$ by the continuous mapping theorem with $g(t) = t^2$. More generally, if $T_n \xrightarrow{d} \mathcal{N}(0, I_k)$, then $\|T_n\|^2 \xrightarrow{d} \chi^2_k$.

Example 4 (Convergence in distribution but not in probability). Let $X \sim \mathcal{N}(0,1)$ and define $T_n = (-1)^n X$. Then $T_n \sim \mathcal{N}(0,1)$ for every n, so $T_n \xrightarrow{d} \mathcal{N}(0,1)$. But $T_n - T_{n+1} = (-1)^n X - (-1)^{n+1} X = 2(-1)^n X$, which does not converge to 0, so T_n does not converge in probability (or almost surely). Convergence in distribution says nothing about the joint behavior of the sequence.

Example 5 (L^p convergence vs. almost sure). Let $\Omega = [0,1]$ with Lebesgue measure. Define $T_n = n \cdot \mathbf{1}_{[0,1/n]}$. Then $T_n \xrightarrow{a.s.} 0$ (for any $\omega > 0$, eventually $T_n(\omega) = 0$). However, $\mathbb{E}[|T_n|^p] = n^p/n = n^{p-1} \to \infty$ for $p > 1$, so $T_n \not\xrightarrow{L^p} 0$ for $p > 1$. Almost sure convergence does not imply L^p convergence without uniform integrability.

Theorem (Uniform integrability bridge). $T_n \xrightarrow{L^1} T$ if and only if $T_n \xrightarrow{p} T$ and $\{T_n\}$ is uniformly integrable: $\lim_{M\to\infty} \sup_n \mathbb{E}[|T_n| \mathbf{1}_{|T_n| > M}] = 0$. This provides the missing link between convergence in probability and L^p convergence.

Cross-Links

AI. Convergence in distribution underpins the theoretical analysis of stochastic gradient descent: under appropriate step-size schedules, the iterate distribution converges to a stationary distribution, and Slutsky-type arguments justify plug-in variance estimation in online learning.

Economics. Asymptotic results in econometrics (consistency and normality of GMM estimators) are stated entirely in terms of these modes; the distinction between convergence in probability and almost sure convergence determines whether path-dependent economic dynamics can be analyzed.

Linear Algebra. Convergence in L^2 is norm convergence in the Hilbert space $L^2(\Omega, \mathbb{P})$; the hierarchy of modes mirrors the hierarchy of topologies (norm, weak, weak-$*$) on Banach spaces.

Quantum Mechanics. Convergence of density matrices in trace norm is the quantum analogue of L^1 convergence; weak convergence of quantum states corresponds to convergence in distribution of measurement outcomes.

Structural Compression

Invariant. The hierarchy a.s. $\Rightarrow$ in prob. $\Rightarrow$ in dist. is strict; each mode imposes a progressively weaker condition on the probability measure. $L^p \Rightarrow$ in prob. provides an independent chain.

Mechanism. Each mode controls a different aspect of the probability mass: almost sure convergence controls path behavior; convergence in probability controls tail probabilities; convergence in distribution controls only the law. L^p convergence controls moment-weighted deviations. The continuous mapping theorem and Slutsky's theorem transfer convergence through smooth operations.

Cross-System Echo. Convergence in distribution is weak convergence of measures; in functional analysis, this is weak-$*$ convergence of probability measures as linear functionals on $C_b(\mathbb{R})$. The Skorokhod representation theorem bridges distributional and pathwise convergence by constructing a coupling.

Exercises

1. Construct a sequence T_n that converges to 0 in probability but not almost surely. Verify both claims.
2. Let $T_n \xrightarrow{d} \mathcal{N}(0,1)$. Prove that $|T_n| \xrightarrow{d} |Z|$ where $Z \sim \mathcal{N}(0,1)$, and identify the distribution of $|Z|$.
3. Suppose $T_n \xrightarrow{L^2} T$. Prove that $\mathbb{E}[T_n] \to \mathbb{E}[T]$ and $\mathrm{Var}(T_n) \to \mathrm{Var}(T)$.
4. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(0, \theta)$. Show that $X_{(n)} = \max_i X_i$ satisfies $X_{(n)} \xrightarrow{a.s.} \theta$. Find the rate: determine the limit distribution of $n(\theta - X_{(n)})$.
5. Prove that if $T_n \xrightarrow{p} T$ and $T_n \xrightarrow{p} T'$, then $T = T'$ a.s. (uniqueness of limits in probability).
6. Let $T_n \xrightarrow{d} T$ and $S_n \xrightarrow{d} S$. Give a counterexample showing that $T_n + S_n$ need not converge in distribution to $T + S$ without additional conditions.
7. Argue, structurally, why convergence in distribution is the natural mode for the central limit theorem while convergence in probability is the natural mode for the law of large numbers, relating each to the type of statistical conclusion it supports.

Consistency

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.2 — Consistency.

Consistency is the minimal asymptotic requirement for an estimator: as the sample size grows, the estimator must converge to the true parameter. This node instantiates the convergence modes of Node 7.1 in the estimation setting, depends on the likelihood framework (Nodes 3.1, 3.2), and enables the rate-of-convergence results in asymptotic normality (Node 7.3) and the general M-estimation framework (Node 7.5).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be a sequence of estimators.

Weak consistency. $\hat{\theta}_n$ is weakly consistent for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ under P_θ for all $\theta \in \Theta$.

Strong consistency. $\hat{\theta}_n$ is strongly consistent for θ if $\hat{\theta}_n \xrightarrow{a.s.} \theta$ under P_θ for all $\theta \in \Theta$.

Fisher consistency. A functional $T(F)$ is Fisher consistent for θ if $T(F_\theta) = \theta$ for all $\theta \in \Theta$. Fisher consistency is a population-level property; combined with convergence of the plug-in estimator $T(\hat{F}_n)$ to $T(F)$, it yields statistical consistency.

Intuition

Consistency says that with enough data, the estimator gets the answer right. It does not say how quickly or how precisely — that is the role of asymptotic normality (Node 7.3). An inconsistent estimator is fundamentally broken: no amount of data can rescue it. Consistency is therefore a necessary filter, not a measure of quality.

The conceptual mechanism is always the same: the sample criterion function converges to a population criterion function, and the population criterion has a unique optimizer at the truth. Identifiability ensures the truth is distinguishable; the law of large numbers drives the convergence.

Structural Role

Consistency is the first-order asymptotic property of an estimator. It guarantees that the sampling distribution of $\hat{\theta}_n$ concentrates at the true parameter as $n \to \infty$. It is a necessary but not sufficient condition for good statistical performance: it does not control the rate of convergence, the shape of the sampling distribution, or finite-sample behavior. Rate and distributional shape are characterized in Node 7.3. Consistency of M-estimators is studied in the general framework of Node 7.5.

Mathematical Development

Sufficient condition via moments. If $\mathbb{E}_\theta[\hat{\theta}_n] \to \theta$ and $\mathrm{Var}_\theta(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n \xrightarrow{p} \theta$ by Chebyshev's inequality: $\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \leq \mathrm{Var}(\hat{\theta}_n)/\varepsilon^2 \to 0$. This is the simplest route to consistency but requires both vanishing bias and vanishing variance.

Consistency of the MLE: Wald's proof.

Let $M_n(\theta') = n^{-1}\ell(\theta'; X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n \log p_{\theta'}(X_i)$ be the normalized log-likelihood. Define the population criterion $M(\theta') = \mathbb{E}_\theta[\log p_{\theta'}(X)]$.

Step 1: Unique maximum. By the information inequality (Gibbs' inequality / KL divergence nonnegativity),

$$M(\theta') = \mathbb{E}_\theta[\log p_{\theta'}(X)] \leq \mathbb{E}_\theta[\log p_\theta(X)] = M(\theta),$$

with equality if and only if $p_{\theta'} = p_\theta$ P_θ-a.e. Under identifiability ($p_{\theta'} = p_\theta$ a.e. implies $\theta' = \theta$), the maximum of $M(\theta')$ over Θ is uniquely attained at $\theta' = \theta$.

Step 2: Uniform convergence. Assume $\sup_{\theta' \in \Theta} |M_n(\theta') - M(\theta')| \xrightarrow{p} 0$. This holds under a uniform law of large numbers (ULLN), which requires either compactness of Θ and continuity of $\theta' \mapsto \log p_{\theta'}(x)$, or a Glivenko-Cantelli argument for the class $\{\log p_{\theta'} : \theta' \in \Theta\}$.

Step 3: Argmax continuous mapping theorem. If $M_n \to M$ uniformly and M has a unique maximizer θ, then $\hat{\theta}_n = \arg\max M_n \xrightarrow{p} \theta$.

Proof of step 3. Fix $\varepsilon > 0$. By uniqueness and continuity of M, there exists $\delta > 0$ such that $M(\theta') \leq M(\theta) - \delta$ for all $\|\theta' - \theta\| > \varepsilon$. Under uniform convergence, for large n,

$$M_n(\hat{\theta}_n) \geq M_n(\theta) \geq M(\theta) - \delta/3,$$

and $M_n(\hat{\theta}_n) \leq M(\hat{\theta}_n) + \delta/3$. If $\|\hat{\theta}_n - \theta\| > \varepsilon$, then $M(\hat{\theta}_n) \leq M(\theta) - \delta$, giving $M_n(\hat{\theta}_n) \leq M(\theta) - 2\delta/3$, a contradiction. □

Consistency of the method of moments. Let $\mu_k(\theta) = \mathbb{E}_\theta[X^k]$ and $\hat{\mu}_k = n^{-1} \sum X_i^k$. By the WLLN, $\hat{\mu}_k \xrightarrow{p} \mu_k(\theta)$. If the moment map $g : \theta \mapsto (\mu_1(\theta), \ldots, \mu_r(\theta))$ is continuous and injective, the method-of-moments estimator $\hat{\theta}_n = g^{-1}(\hat{\mu}_1, \ldots, \hat{\mu}_r)$ satisfies $\hat{\theta}_n \xrightarrow{p} \theta$ by the continuous mapping theorem.

Uniform law of large numbers (ULLN). A class of functions $\mathcal{F}$ is Glivenko-Cantelli for P if $\sup_{f \in \mathcal{F}} |n^{-1} \sum f(X_i) - \mathbb{E}[f(X)]| \xrightarrow{a.s.} 0$. Sufficient conditions include: (i) $\mathcal{F}$ has finite VC dimension; (ii) $\mathcal{F}$ is uniformly bounded and parametrized by a compact set with equicontinuity. The ULLN is the key technical ingredient linking pointwise LLN convergence to the uniform convergence needed for consistency of extremum estimators.

Rate of convergence and consistency. Consistency alone does not specify a rate. The sample mean converges at rate $n^{-1/2}$ (by the CLT). The MLE for the uniform upper bound θ converges at rate n^{-1} (superefficient). Nonparametric density estimators converge at rate $n^{-2/(4+p)}$ (slower than parametric). These distinctions motivate the refined study of asymptotic normality in Node 7.3.

Consistency under model misspecification. When the true distribution P_0 does not belong to the model $\{P_\theta : \theta \in \Theta\}$, the MLE converges to the pseudo-true parameter

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{KL}(P_0 \| P_\theta) = \arg\max_{\theta \in \Theta} \mathbb{E}_0[\log p_\theta(X)].$$

This is the KL projection of P_0 onto the model. The proof of consistency carries through verbatim — the ULLN and argmax CMT arguments do not require correct specification — but the limit θ^* is no longer the "true" parameter. This observation is the foundation for the sandwich variance under misspecification (Node 7.3).

Consistency of Bayesian estimators. The posterior mode (MAP estimator) is consistent under the same conditions as the MLE, since the prior contributes $O(1)$ to the log-posterior while the likelihood contributes $O(n)$. More generally, the posterior distribution itself concentrates at the true parameter (posterior consistency), provided the prior assigns positive mass to every KL neighborhood of the truth (Schwartz's theorem). This connects to the Bernstein-von Mises theorem (Node 6.7).

Well-separated maximum condition. An alternative to compactness for the argmax CMT: if $M(\theta)$ has a well-separated maximum, meaning

$$\sup_{\|\theta'-\theta\|>\varepsilon} M(\theta') < M(\theta) \quad \forall \varepsilon > 0,$$

then uniform convergence $\sup_{\theta'} |M_n(\theta') - M(\theta')| \xrightarrow{p} 0$ implies $\hat{\theta}_n \xrightarrow{p} \theta$. This condition holds when M is strictly concave or when the KL divergence $\mathrm{KL}(P_\theta \| P_{\theta'}) > 0$ for $\theta' \neq \theta$ (identifiability) and M is continuous.

Consistency and efficiency are independent. The sample mean and the sample median are both consistent for the center of a symmetric distribution. However, their rates differ: under normality, the mean converges at rate σ^2/n while the median converges at rate $\pi\sigma^2/(2n)$. Consistency is a qualitative property (yes/no); efficiency is quantitative (how fast).

Consistency of the bootstrap. The bootstrap is consistent if the bootstrap distribution of $\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ converges to the same limit as $\sqrt{n}(\hat{\theta}_n - \theta)$. This holds under the conditions for asymptotic normality of $\hat{\theta}_n$, providing nonparametric confidence intervals that are asymptotically correct.

Worked Examples

Example 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda)$. The MLE is $\hat{\lambda}_n = 1/\bar{X}_n$. By the SLLN, $\bar{X}_n \xrightarrow{a.s.} 1/\lambda$. By the continuous mapping theorem with $g(x) = 1/x$, $\hat{\lambda}_n \xrightarrow{a.s.} \lambda$. The MLE is strongly consistent.

Example 2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Uniform}(0, \theta)$. The MLE is $\hat{\theta}_n = X_{(n)} = \max_i X_i$. We have $\mathbb{P}(X_{(n)} \leq t) = (t/\theta)^n$ for $0 \leq t \leq \theta$. Thus $\mathbb{P}(\theta - X_{(n)} > \varepsilon) = (1 - \varepsilon/\theta)^n \to 0$ for any $\varepsilon > 0$, so $\hat{\theta}_n \xrightarrow{p} \theta$. In fact $\sum_n \mathbb{P}(\theta - X_{(n)} > \varepsilon) < \infty$, so by Borel-Cantelli, $X_{(n)} \xrightarrow{a.s.} \theta$. Note that $\hat{\theta}_n$ is biased ($\mathbb{E}[X_{(n)}] = n\theta/(n+1)$) but consistent — consistency does not require unbiasedness.

Example 3 (Inconsistency). Let $\hat{\theta}_n = X_1$ regardless of n. Then $\hat{\theta}_n$ is unbiased but $\mathrm{Var}(\hat{\theta}_n) = \mathrm{Var}(X_1) \not\to 0$. This estimator is not consistent: unbiasedness alone is insufficient.

Example 4 (Method of moments). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Gamma}(\alpha, \beta)$ with $\mathbb{E}[X] = \alpha/\beta$ and $\mathrm{Var}(X) = \alpha/\beta^2$. The moment equations are $\hat{\mu}_1 = \hat{\alpha}/\hat{\beta}$ and $\hat{\mu}_2 - \hat{\mu}_1^2 = \hat{\alpha}/\hat{\beta}^2$. Solving: $\hat{\beta} = \hat{\mu}_1/(\hat{\mu}_2 - \hat{\mu}_1^2)$ and $\hat{\alpha} = \hat{\mu}_1^2/(\hat{\mu}_2 - \hat{\mu}_1^2)$. By the WLLN, $\hat{\mu}_k \xrightarrow{p} \mu_k$, and by the continuous mapping theorem, $(\hat{\alpha}, \hat{\beta}) \xrightarrow{p} (\alpha, \beta)$ since the moment map is continuously invertible on $(0, \infty)^2$.

Example 5 (Non-compact parameter space). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ with $\Theta = \mathbb{R}$. The MLE $\hat{\mu} = \bar{X}_n$ is consistent by the WLLN. Here the ULLN holds despite non-compact Θ because the Gaussian log-likelihood has quadratic curvature that prevents the maximizer from escaping. In contrast, for a mixture model with the number of components as a parameter, unbounded Θ can lead to inconsistency of the MLE when the likelihood is unbounded.

Cross-Links

AI. Consistency of empirical risk minimizers (ERM) is the learning-theoretic analogue: under finite VC dimension, the empirical risk minimizer converges to the population risk minimizer. The ULLN over function classes is precisely the Glivenko-Cantelli condition used in both settings.

Economics. Consistency of GMM estimators in econometrics follows the same argmax/argmin continuous mapping theorem structure, with the GMM objective replacing the log-likelihood.

Linear Algebra. The uniform convergence condition can be expressed as operator norm convergence of the empirical measure operator to the population operator in the space of bounded functions on Θ.

Quantum Mechanics. Consistency of quantum state tomography estimators — the reconstructed density matrix converges to the true state — requires analogous identifiability (informationally complete measurements) and ULLN conditions on the measurement statistics.

Structural Compression

Invariant. $\hat{\theta}_n \xrightarrow{p} \theta$: the sampling distribution of $\hat{\theta}_n$ concentrates at θ as $n \to \infty$.

Mechanism. The WLLN drives sample averages to population means; under identifiability, the population criterion is uniquely maximized at the truth; the argmax continuous mapping theorem transfers uniform convergence of the criterion to convergence of the optimizer.

Cross-System Echo. Consistency of the MLE corresponds to the convergence of empirical risk minimizers to the population risk minimizer in statistical learning theory; the structural parallel is: identifiability $\leftrightarrow$ realizability, ULLN $\leftrightarrow$ Glivenko-Cantelli, argmax CMT $\leftrightarrow$ fundamental theorem of statistical learning.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Prove that $(\bar{X}_n, S_n^2)$ is strongly consistent for (μ, σ^2).
2. Show that $\hat{\theta}_n = \bar{X}_n + 1/n$ is consistent for μ even though it is biased for every n.
3. Let $\hat{\theta}_n$ be consistent for θ and let g be continuous at θ. Prove that $g(\hat{\theta}_n)$ is consistent for $g(\theta)$.
4. Give an example of a strongly consistent estimator that is not unbiased for any finite n.
5. Verify the ULLN for the family $\{\log p_\lambda : \lambda > 0\}$ where $p_\lambda(x) = \lambda e^{-\lambda x}$ on $x > 0$, by checking equicontinuity on compact subsets of $(0, \infty)$.
6. Suppose Θ is not compact. Construct an example where the MLE is inconsistent because the argmax escapes to infinity, and explain which step of Wald's proof fails.
7. Argue, structurally, why identifiability is necessary for consistency of the MLE, connecting the information inequality to the KL divergence and explaining what fails in the proof when two distinct parameters yield the same distribution.

Asymptotic Normality

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.3 — Asymptotic Normality.

Asymptotic normality is the second-order asymptotic characterization of an estimator: beyond consistency (Node 7.2), it specifies the rate of convergence and the shape of the limiting sampling distribution. This result depends on the modes of convergence (Node 7.1), consistency (Node 7.2), and moment theory (Node 1.9), and enables the delta method (Node 7.4), M-estimation theory (Node 7.5), and likelihood asymptotics (Node 7.6).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$, $\theta \in \Theta \subseteq \mathbb{R}^k$. A consistent estimator $\hat{\theta}_n$ is **asymptotically normal** with asymptotic variance $V(\theta)$ if

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)) \quad \text{under } P_\theta,$$

where $V(\theta) \in \mathbb{R}^{k \times k}$ is positive definite. The rate $\sqrt{n}$ is the canonical parametric rate, meaning $\hat{\theta}_n - \theta = O_p(n^{-1/2})$.

Intuition

Asymptotic normality says that for large n, the estimator $\hat{\theta}_n$ behaves approximately as $\theta + n^{-1/2}Z$ where $Z \sim \mathcal{N}(0, V(\theta))$. The Gaussian shape arises because the estimator is determined by an average of many small independent contributions, and the CLT applies. The matrix $V(\theta)$ controls the size and shape of the confidence ellipsoid. When $V(\theta) = I(\theta)^{-1}$, the estimator achieves the Cramer-Rao bound and extracts all available information from the data.

Structural Role

Asymptotic normality is the canonical second-order characterization of estimator performance. It enables construction of confidence intervals and hypothesis tests (Modules 4-5), establishes efficiency comparisons via asymptotic relative efficiency, provides the foundation for the delta method (Node 7.4), and underlies the Bernstein-von Mises theorem (Node 6.7) connecting frequentist and Bayesian asymptotics.

Mathematical Development

Asymptotic normality of the MLE.

Under regularity conditions — (R1) the parameter space Θ is open, (R2) the support of p_θ does not depend on θ, (R3) $\log p_\theta(x)$ is three times differentiable in θ with the third derivative bounded in expectation, (R4) the Fisher information $I(\theta) = \mathbb{E}_\theta[\nabla \log p_\theta(X) \nabla \log p_\theta(X)^\top]$ is positive definite — the MLE satisfies

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0,\, I(\theta)^{-1}).$$

Proof. The MLE satisfies the score equation $\nabla_\theta \ell(\hat{\theta}_n) = 0$, where $\ell(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$. Taylor-expand around the true θ:

$$0 = \nabla_\theta \ell(\theta) + \nabla_\theta^2 \ell(\theta^*) \cdot (\hat{\theta}_n - \theta),$$

where θ^* lies between $\hat{\theta}_n$ and θ (componentwise, via the mean value theorem). Rearranging and multiplying by $n^{-1/2}$:

$$\sqrt{n}(\hat{\theta}_n - \theta) = -\left[\frac{1}{n}\nabla_\theta^2 \ell(\theta^*)\right]^{-1} \frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta).$$

CLT for the score. Since $\mathbb{E}_\theta[\nabla \log p_\theta(X_i)] = 0$ (the score identity) and $\mathrm{Cov}_\theta(\nabla \log p_\theta(X_i)) = I(\theta)$, the multivariate CLT gives

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla \log p_\theta(X_i) \xrightarrow{d} \mathcal{N}(0, I(\theta)).$$

WLLN for the Hessian. By the law of large numbers and consistency of $\hat{\theta}_n$ (hence $\theta^* \xrightarrow{p} \theta$),

$$\frac{1}{n}\nabla_\theta^2 \ell(\theta^*) \xrightarrow{p} \mathbb{E}_\theta[\nabla^2 \log p_\theta(X)] = -I(\theta).$$

The last equality uses the second Bartlett identity: $\mathbb{E}[\nabla^2 \log p_\theta] = -I(\theta)$.

Combining via Slutsky. By Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta) = I(\theta)^{-1} \cdot \frac{1}{\sqrt{n}} \nabla_\theta \ell(\theta) + o_p(1) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}). \quad \square$$

Proof for exponential families. Let $p_\theta(x) = h(x) \exp(\eta(\theta)^\top T(x) - A(\eta(\theta)))$ with natural parameter $\eta = \eta(\theta)$. The score is $\nabla_\theta \ell(\theta) = \sum_i [\dot{\eta}(\theta)]^\top (T(X_i) - \nabla_\eta A(\eta))$. The Fisher information is $I(\theta) = \dot{\eta}(\theta)^\top \nabla_\eta^2 A(\eta)\, \dot{\eta}(\theta)$. Since $\nabla^2 \log p_\theta$ does not depend on the data (in the natural parametrization, it equals $-\nabla_\eta^2 A(\eta)$), the WLLN step is exact: the sample Hessian equals the expected Hessian. The regularity conditions (R3) are automatically satisfied, making the exponential family the cleanest setting for asymptotic normality.

Asymptotic relative efficiency. For two asymptotically normal estimators $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ with scalar θ,

$$\mathrm{ARE}(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}) = \frac{V_2(\theta)}{V_1(\theta)}.$$

The Cramer-Rao lower bound (Node 3.4) states $V(\theta) \geq I(\theta)^{-1}$ for all regular estimators. An estimator achieving $V(\theta) = I(\theta)^{-1}$ is **asymptotically efficient**. The MLE is asymptotically efficient under the regularity conditions above.

Sandwich variance under misspecification. When the model is misspecified ($P_0 \notin \{P_\theta : \theta \in \Theta\}$), the MLE converges to the pseudo-true parameter $\theta^* = \arg\min_\theta \mathrm{KL}(P_0 \| P_\theta)$. The asymptotic variance takes the sandwich form:

$$V = A(\theta^*)^{-1} B(\theta^*) A(\theta^*)^{-\top},$$

where $A(\theta^*) = \mathbb{E}_0[-\nabla^2 \log p_{\theta^*}(X)]$ and $B(\theta^*) = \mathbb{E}_0[\nabla \log p_{\theta^*}(X) \nabla \log p_{\theta^*}(X)^\top]$. Under correct specification, the second Bartlett identity gives $A = B = I(\theta)$ and the sandwich reduces to $I(\theta)^{-1}$.

Connection to Fisher information. The Fisher information matrix $I(\theta)$ plays three roles simultaneously: (i) it is the variance of the score, (ii) it is the negative expected Hessian of the log-likelihood, (iii) it is the

inverse of the asymptotic variance of the efficient estimator. These three characterizations coincide under correct specification and regularity.

Asymptotic normality of the sample mean. As a baseline: for i.i.d. X_i with mean μ and finite variance σ^2, the CLT gives $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. This is the prototype for all asymptotic normality results. The MLE generalizes this: the score $\nabla \log p_\theta(X_i)$ plays the role of $X_i - \mu$, and the Fisher information $I(\theta)$ plays the role of σ^2.

Non-regular cases. Asymptotic normality at rate $\sqrt{n}$ requires regularity. For the uniform model Uniform$(0, \theta)$, the MLE $X_{(n)}$ converges at rate n (not $\sqrt{n}$) and the limit distribution is exponential, not Gaussian. The regularity condition that fails is differentiability of the support with respect to θ. For the Laplace location model, the MLE converges at rate $\sqrt{n}$ but the proof requires different techniques (the score has a discontinuity).

Multivariate CLT statement. The multivariate CLT, which underpins the score asymptotics, states: if $X_1, \ldots, X_n$ are i.i.d. in $\mathbb{R}^k$ with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Cov}(X_i) = \Sigma$, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

The proof uses the Cramer-Wold device: show $a^\top \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, a^\top \Sigma a)$ for all $a \in \mathbb{R}^k$ using the univariate CLT, then invoke the equivalence between convergence of all one-dimensional projections and convergence of the joint distribution.

Worked Examples

Example 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. The MLE is $\hat{p} = \bar{X}_n$. The Fisher information is $I(p) = 1/(p(1-p))$. By asymptotic normality, $\sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$. This recovers the CLT for the sample proportion. The asymptotic 95% confidence interval is $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$.

Example 2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both parameters unknown, $\theta = (\mu, \sigma^2)$. The Fisher information matrix is

$$I(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}.$$

The MLE is $(\bar{X}_n, \hat{\sigma}_n^2)$ where $\hat{\sigma}_n^2 = n^{-1}\sum(X_i - \bar{X}_n)^2$. By asymptotic normality,

$$\sqrt{n}\begin{pmatrix} \bar{X}_n - \mu \\ \hat{\sigma}_n^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right).$$

The off-diagonal zeros reflect asymptotic independence of $\bar{X}_n$ and $\hat{\sigma}_n^2$, which is exact under normality.

Example 3 (Comparing estimators via ARE). For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, the sample mean has asymptotic variance σ^2 and the sample median has asymptotic variance $\pi\sigma^2/2$. The ARE of the mean relative to the median is $(\pi\sigma^2/2)/\sigma^2 = \pi/2 \approx 1.57$: the mean is 57% more efficient at the normal. However, for heavy-tailed distributions (e.g., Cauchy), the median is consistent while the mean is not even defined, illustrating that ARE is model-dependent.

Cross-Links

AI. In stochastic gradient descent, the iterate θ_n satisfies an analogous asymptotic normality result under decreasing step sizes: $\sqrt{n}(\theta_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, V)$ where V depends on the noise covariance and the Hessian of the loss, mirroring the sandwich structure.

Economics. Asymptotic normality of GMM estimators with the efficient weighting matrix yields variance $(G^\top S^{-1} G)^{-1}$, the econometric analogue of the inverse Fisher information.

Linear Algebra. The proof uses the spectral properties of the Fisher information matrix: positive definiteness ensures invertibility, and the quadratic form $h^\top I(\theta) h$ defines a Riemannian metric on the parameter space.

Quantum Mechanics. The quantum Cramer-Rao bound gives the minimum variance for estimating a quantum state parameter, with the quantum Fisher information replacing $I(\theta)$; the classical asymptotic normality result holds for quantum measurements in the limit of many identical preparations.

Structural Compression

Invariant. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$: the standardized MLE converges in distribution to a Gaussian with variance equal to the inverse Fisher information.

Mechanism. Score equation linearized via Taylor expansion; CLT applied to the score; WLLN applied to the Hessian; Slutsky yields the Gaussian limit. The Fisher information emerges as both the score variance and the negative expected Hessian.

Cross-System Echo. Asymptotic normality is the statistical manifestation of the central limit theorem for estimating functions. In information geometry, the Fisher information defines the unique invariant Riemannian metric on the statistical manifold, and asymptotic normality states that the MLE is a normal coordinate chart centered at the truth.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. Compute the Fisher information and verify that the MLE $\hat{\lambda} = \bar{X}_n$ achieves the asymptotic variance $I(\lambda)^{-1} = \lambda$.
2. Verify the second Bartlett identity $\mathbb{E}[\nabla^2 \log p_\theta(X)] = -I(\theta)$ for the exponential distribution $p_\lambda(x) = \lambda e^{-\lambda x}$.
3. For the Cauchy location model $p_\theta(x) = \pi^{-1}(1 + (x - \theta)^2)^{-1}$, the MLE exists and is consistent but the Fisher information involves a nontrivial integral. Compute $I(\theta)$ and state the asymptotic distribution of the MLE.
4. Show that the sample median of $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ is asymptotically normal with variance $\pi\sigma^2/(2n)$, and compute its ARE relative to the sample mean.
5. Derive the sandwich variance formula when $X_i \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$ but the working model is $\mathcal{N}(\mu, \sigma^2)$. Identify the pseudo-true parameters.
6. Let $\hat{\theta}_n$ be asymptotically normal with variance $V(\theta)$. Prove that $a^\top \hat{\theta}_n$ is asymptotically normal with variance $a^\top V(\theta) a$ for any fixed $a \in \mathbb{R}^k$.
7. Argue, structurally, why the $\sqrt{n}$ rate is universal for parametric models while nonparametric rates are slower, connecting the rate to the dimension of the parameter space and the CLT.

Delta Method

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.4 — Delta Method.

The delta method transfers asymptotic normality from a parameter to any smooth function of that parameter. It depends on asymptotic normality (Node 7.3) and enables the general M-estimation framework (Node 7.5) and likelihood asymptotics (Node 7.6). It is the primary tool for constructing confidence intervals for derived quantities.

Formal Definition

Let $\hat{\theta}_n$ be asymptotically normal:

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)), \quad \theta \in \Theta \subseteq \mathbb{R}^k.$$

Let $g : \Theta \to \mathbb{R}^m$ be continuously differentiable at θ with Jacobian $\dot{g}(\theta) = \partial g / \partial \theta^\top \in \mathbb{R}^{m \times k}$.

First-order delta method. If $\dot{g}(\theta) \neq 0$ (has full row rank), then

$$\sqrt{n}\,\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0,\ \dot{g}(\theta)\, V(\theta)\, \dot{g}(\theta)^\top\big)\,.$$

Second-order delta method. When $g : \mathbb{R} \to \mathbb{R}$ with $g'(\theta) = 0$ and $g''(\theta) \neq 0$:

$$n\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \frac{g''(\theta)}{2}\, V(\theta)\, \chi^2_1.$$

Intuition

A smooth function of an estimator inherits the estimator's asymptotic Gaussianity because, locally, a smooth function is approximately linear. The Jacobian $\dot{g}(\theta)$ acts as a magnification factor: it stretches or compresses the Gaussian fluctuations of $\hat{\theta}_n$ into Gaussian fluctuations of $g(\hat{\theta}_n)$. When the first derivative vanishes, the linear approximation is trivial and the curvature (second derivative) determines the leading-order behavior, producing a chi-squared rather than Gaussian limit.

Structural Role

The delta method extends asymptotic normality from the parameter θ to any smooth function $g(\theta)$. It is the primary tool for constructing asymptotic confidence intervals and tests for derived quantities — odds ratios, relative risks, variance components, functions of regression coefficients — and is used directly in M-estimation (Node 7.5), variance-stabilizing transformations, and the construction of asymptotic pivots (Node 4.5).

Mathematical Development

Proof of the first-order delta method.

By a first-order Taylor expansion of g around θ:

$$g(\hat{\theta}_n) = g(\theta) + \dot{g}(\theta)(\hat{\theta}_n - \theta) + R_n,$$

where the remainder $R_n = O(\|\hat{\theta}_n - \theta\|^2)$. Since $\hat{\theta}_n - \theta = O_p(n^{-1/2})$, we have $R_n = O_p(n^{-1})$, so $\sqrt{n}R_n = O_p(n^{-1/2}) = o_p(1)$.

Therefore:

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) = \dot{g}(\theta) \cdot \sqrt{n}(\hat{\theta}_n - \theta) + o_p(1).$$

Since $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z \sim \mathcal{N}(0, V(\theta))$ and $\dot{g}(\theta)$ is a fixed matrix, Slutsky's theorem gives

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \dot{g}(\theta)Z \sim \mathcal{N}(0,\, \dot{g}(\theta)V(\theta)\dot{g}(\theta)^\top). \quad \square$$

Scalar case. For $g : \mathbb{R} \to \mathbb{R}$ and scalar θ:

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0,\ [g'(\theta)]^2 V(\theta)\big)\,.$$

The asymptotic variance is $[g'(\theta)]^2$ times the asymptotic variance of $\hat{\theta}_n$.

Proof of the second-order delta method.

When $g'(\theta) = 0$, expand to second order:

$$g(\hat{\theta}_n) - g(\theta) = \frac{1}{2}g''(\theta)(\hat{\theta}_n - \theta)^2 + O_p(n^{-3/2}).$$

Multiplying by n:

$$n\big(g(\hat{\theta}_n) - g(\theta)\big) = \frac{g''(\theta)}{2}\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2 + o_p(1).$$

Since $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta))$, the continuous mapping theorem gives

$$\left[\sqrt{n}(\hat{\theta}_n - \theta)\right]^2 \xrightarrow{d} V(\theta)\chi_1^2,$$

so $n(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \frac{g''(\theta)}{2}V(\theta)\chi_1^2$. $\square$

Multivariate second-order delta method. For $g : \mathbb{R}^k \to \mathbb{R}$ with $\nabla g(\theta) = 0$, let $H = \nabla^2 g(\theta)$ be the Hessian. Then

$$n\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \frac{1}{2}Z^\top H Z, \quad Z \sim \mathcal{N}(0, V(\theta)).$$

This is a quadratic form in a Gaussian, expressible as a weighted sum of independent χ_1^2 variables with weights equal to the eigenvalues of $HV(\theta)$.

Variance-stabilizing transformations. Choose g so that the asymptotic variance of $g(\hat{\theta}_n)$ does not depend on θ. For scalar θ with asymptotic variance $V(\theta)$, solve

$$[g'(\theta)]^2 V(\theta) = c \quad \Longrightarrow \quad g(\theta) = \int \frac{d\theta}{\sqrt{V(\theta)}}.$$

This produces an asymptotically pivotal quantity: $\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, c)$ with c independent of θ.

Application to the MLE. For the MLE with $V(\theta) = I(\theta)^{-1}$:

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0,\ \dot{g}(\theta)\, I(\theta)^{-1}\, \dot{g}(\theta)^\top\big)\,.$$

Worked Examples

Example 1 (Log-odds / logit). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ with MLE $\hat{p} = \bar{X}_n$ and $V(p) = p(1-p)$. Consider the log-odds $g(p) = \log(p/(1-p))$. Then $g'(p) = 1/(p(1-p))$, so

$$[g'(p)]^2 V(p) = \frac{1}{p^2(1-p)^2} \cdot p(1-p) = \frac{1}{p(1-p)}.$$

Thus $\sqrt{n}(\text{logit}(\hat{p}) - \text{logit}(p)) \xrightarrow{d} \mathcal{N}(0, 1/(p(1-p)))$. A confidence interval for $\text{logit}(p)$ is $\text{logit}(\hat{p}) \pm z_{\alpha/2}/\sqrt{n\hat{p}(1-\hat{p})}$.

Example 2 (Fisher z-transform). Let r be the sample correlation coefficient estimating ρ. For bivariate normal data, $\sqrt{n}(r-\rho)$ is asymptotically $\mathcal{N}(0, (1-\rho^2)^2)$. The Fisher z-transform $g(r) = \frac{1}{2}\log\frac{1+r}{1-r} = \text{arctanh}(r)$ satisfies $g'(\rho) = 1/(1-\rho^2)$, giving

$$[g'(\rho)]^2 (1-\rho^2)^2 = 1.$$

This is a variance-stabilizing transformation: $\sqrt{n}(\text{arctanh}(r) - \text{arctanh}(\rho)) \xrightarrow{d} \mathcal{N}(0, 1)$, with asymptotic variance independent of ρ. The standard approximation uses $n-3$ in place of n for finite-sample correction.

Example 3 (Ratio of means). Let (X_i, Y_i) be i.i.d. with $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y \neq 0$, and covariance matrix Σ. The ratio $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$ has gradient $\nabla g = (1/\mu_Y, -\mu_X/\mu_Y^2)^\top$. By the delta method,

$$\sqrt{n}\left(\frac{\bar{X}}{\bar{Y}} - \frac{\mu_X}{\mu_Y}\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma_X^2}{\mu_Y^2} - \frac{2\mu_X\sigma_{XY}}{\mu_Y^3} + \frac{\mu_X^2\sigma_Y^2}{\mu_Y^4}\right).$$

Cross-Links

AI. The delta method underlies the Laplace approximation for posterior predictive uncertainty: the variance of a prediction $f(\theta)$ is approximated by $\nabla f(\hat{\theta})^\top \hat{V} \nabla f(\hat{\theta})$, a direct application of the first-order delta method to neural network outputs.

Economics. In econometrics, the delta method is the standard tool for inference on nonlinear functions of regression coefficients, such as elasticities, marginal effects, and impulse response functions.

Linear Algebra. The delta method is the pushforward of a Gaussian measure under a smooth map: the Jacobian acts as a linear map between tangent spaces, transforming the Gaussian by the standard formula for affine transformations of normals.

Quantum Mechanics. Error propagation in quantum parameter estimation uses the same linearization: the variance of a derived quantity $g(\theta)$ under the quantum Cramer-Rao bound is $[g'(\theta)]^2/(n \cdot F_Q(\theta))$, where F_Q is the quantum Fisher information.

Structural Compression

Invariant. Smooth functions of asymptotically normal estimators are asymptotically normal, with variance determined by the Jacobian of the transformation.

Mechanism. First-order Taylor linearization reduces $g(\hat{\theta}_n) - g(\theta)$ to a linear function of $\hat{\theta}_n - \theta$; the linear map transforms the Gaussian limit by the Jacobian. When the Jacobian vanishes, the second-order term dominates, producing a chi-squared limit.

Cross-System Echo. The delta method is the statistical instance of error propagation in measurement theory. In differential geometry, it is the pushforward of a distribution under a smooth map. The variance-stabilizing transformation is the unique reparametrization that makes the Fisher information metric flat.

Exercises

1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$. Use the delta method to find the asymptotic distribution of $g(\bar{X}_n) = 1/\bar{X}_n$ (the MLE of λ).

2. For the Poisson MLE $\hat{\lambda} = \bar{X}_n$, apply the delta method to find the asymptotic distribution of $\sqrt{\hat{\lambda}}$. Verify that $g(\lambda) = \sqrt{\lambda}$ is a variance-stabilizing transformation.

3. Prove the multivariate delta method: if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V)$ in $\mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable, derive the asymptotic distribution of $g(\hat{\theta}_n)$.

4. Let $\hat{p}_1$ and $\hat{p}_2$ be independent sample proportions from n_1 and n_2 observations. Use the delta method to derive the asymptotic distribution of the log-relative-risk $\log(\hat{p}_1/\hat{p}_2)$.

5. Apply the second-order delta method to $g(\theta) = (\theta - \mu)^2$ with $\hat{\theta}_n = \bar{X}_n$ and $\theta = \mu$, and identify the limiting distribution.

6. Find the variance-stabilizing transformation for the binomial proportion $\hat{p}$ (with $V(p) = p(1-p)$), and show it is $g(p) = \arcsin(\sqrt{p})$.

7. Argue, structurally, why the delta method breaks down for non-differentiable functions g, giving a specific example where $g(\hat{\theta}_n)$ is asymptotically normal but with a different variance formula, and explaining what replaces the Jacobian.

M-Estimation

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.5 — M-Estimation.

M-estimation provides a unified asymptotic framework that encompasses MLE, method of moments, robust estimators, and regression estimators as special cases. This node depends on consistency (Node 7.2), asymptotic normality (Node 7.3), the delta method (Node 7.4), and the likelihood framework (Node 3.2), and enables the likelihood asymptotics of Node 7.6.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_0$.

M-estimator. An M-estimator maximizes a sample objective:

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m(X_i, \theta).$$

Z-estimator. A Z-estimator (estimating equations estimator) $\hat{\theta}_n$ solves

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \hat{\theta}_n) = 0,$$

where $\psi(x, \theta) : \mathcal{X} \times \Theta \to \mathbb{R}^k$ is the estimating function.

If $m(x, \theta)$ is differentiable in θ, the M-estimator solves the Z-estimating equations with $\psi(x, \theta) = \nabla_\theta m(x, \theta)$. The MLE is the M-estimator with $m(x, \theta) = \log p_\theta(x)$.

Intuition

M-estimation abstracts away the specific form of the objective function and asks: under what conditions does a sample optimizer converge to a population optimizer, and what is its limiting distribution? The answer separates into two independent components — consistency (the optimizer of the sample criterion converges to the optimizer of the population criterion) and asymptotic normality (the fluctuations around the limit are Gaussian with a specific covariance). The sandwich covariance formula captures the mismatch between the curvature of the objective and the variability of the estimating function, which coincide under correct model specification but diverge under misspecification.

Structural Role

M-estimation unifies MLE, method of moments, robust estimators, and regression estimators under a single asymptotic framework. It decouples the consistency argument (uniform law of large numbers applied to the criterion) from the normality argument (CLT applied to the estimating function at the solution), making both components verifiable independently. The sandwich covariance provides valid inference under misspecification and is the theoretical basis for robust standard errors in econometrics and biostatistics.

Mathematical Development

Consistency of M-estimators.

Let $M(\theta) = \mathbb{E}_{P_0}[m(X,\theta)]$ and $M_n(\theta) = n^{-1}\sum_{i=1}^n m(X_i,\theta)$.

Conditions: 1. $M(\theta)$ has a unique maximizer θ_0. 2. $\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \xrightarrow{p} 0$ (uniform convergence). 3. Θ is compact (or the maximizer is well-separated).

Conclusion: $\hat\theta_n \xrightarrow{p} \theta_0$.

The proof is the same argmax continuous mapping theorem argument as in Node 7.2: uniform convergence of the criterion plus unique maximization of the limit implies convergence of the maximizers.

Consistency of Z-estimators.

Let $\Psi(\theta) = \mathbb{E}_{P_0}[\psi(X,\theta)]$. If $\Psi(\theta_0) = 0$ has a unique solution, $n^{-1}\sum_i \psi(X_i,\theta) \xrightarrow{p} \Psi(\theta)$ uniformly, and Ψ is continuous, then $\hat\theta_n \xrightarrow{p} \theta_0$.

Asymptotic normality of Z-estimators.

Define the Jacobian and variance matrices:

$$A(\theta_0) = \mathbb{E}[\nabla_\theta \psi(X,\theta_0)], \qquad B(\theta_0) = \mathbb{E}[\psi(X,\theta_0)\psi(X,\theta_0)^\top].$$

Regularity conditions: (i) $A(\theta_0)$ is nonsingular; (ii) ψ is differentiable in θ with the derivative satisfying a ULLN; (iii) $B(\theta_0)$ is finite and positive definite.

Theorem. Under these conditions,

$$\sqrt{n}(\hat\theta_n - \theta_0) \xrightarrow{d} \mathcal{N}\big(0,\ A(\theta_0)^{-1}B(\theta_0)[A(\theta_0)^{-1}]^\top\big)\,.$$

Proof. Taylor-expand the sample estimating equation around θ_0:

$$0 = \frac{1}{n}\sum_{i=1}^n \psi(X_i,\hat\theta_n) = \frac{1}{n}\sum_{i=1}^n \psi(X_i,\theta_0) + \left[\frac{1}{n}\sum_{i=1}^n \nabla_\theta\psi(X_i,\theta_n^*)\right](\hat\theta_n - \theta_0),$$

where θ_n^* lies between $\hat\theta_n$ and θ_0. Multiply through by $\sqrt{n}$:

$$0 = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i,\theta_0) + \left[\frac{1}{n}\sum_{i=1}^n \nabla_\theta\psi(X_i,\theta_n^*)\right]\sqrt{n}(\hat\theta_n - \theta_0).$$

By the CLT, $n^{-1/2}\sum_i \psi(X_i,\theta_0) \xrightarrow{d} \mathcal{N}(0, B(\theta_0))$.

By the WLLN and consistency, $n^{-1}\sum_i \nabla_\theta\psi(X_i,\theta_n^*) \xrightarrow{p} A(\theta_0)$.

Solving:

$$\sqrt{n}(\hat\theta_n - \theta_0) = -A(\theta_0)^{-1}\cdot\frac{1}{\sqrt{n}}\sum_i \psi(X_i,\theta_0) + o_p(1).$$

By Slutsky's theorem, the limit is $-A(\theta_0)^{-1}\cdot\mathcal{N}(0,B) = \mathcal{N}(0, A^{-1}BA^{-\top})$. $\square$

Sandwich variance estimator. A consistent estimator of the asymptotic covariance is

$$\widehat{V}_n = \hat A_n^{-1}\hat B_n[\hat A_n^{-1}]^\top,$$

where $\hat{A}_n = n^{-1}\sum_i \nabla_\theta \psi(X_i, \hat{\theta}_n)$ and $\hat{B}_n = n^{-1}\sum_i \psi(X_i, \hat{\theta}_n)\psi(X_i, \hat{\theta}_n)^\top$.

Information matrix equality. Under correct model specification with $\psi = \nabla_\theta \log p_\theta$, the Bartlett identities give $A(\theta_0) = -I(\theta_0)$ and $B(\theta_0) = I(\theta_0)$, so $A^{-1}BA^{-\top} = I(\theta_0)^{-1}$. The sandwich reduces to the inverse Fisher information.

Connection to robust statistics. Huber's M-estimator uses $\psi(x, \theta) = \psi_c(x - \theta)$ where

$$\psi_c(u) = \begin{cases} u & |u| \le c, \\ c \cdot \operatorname{sign}(u) & |u| > c, \end{cases}$$

which corresponds to $m(x, \theta) = \rho_c(x - \theta)$ with $\rho_c(u) = u^2/2$ for $|u| \le c$ and $c|u| - c^2/2$ for $|u| > c$. This estimator is consistent for the location parameter under symmetric distributions and has bounded influence function, providing robustness against outliers at the cost of efficiency: the ARE relative to the sample mean under normality is $[2\Phi(c) - 1 - 2c\phi(c) + 2c^2(1 - \Phi(c))]^{-1} \cdot (2\Phi(c) - 1)^2$.

Connection to empirical risk minimization. In machine learning, the empirical risk $n^{-1}\sum_i L(Y_i, f_\theta(X_i))$ is an M-estimation objective with $m(x, \theta) = -L(y, f_\theta(x))$. Consistency requires the function class to be Glivenko-Cantelli; asymptotic normality requires finite VC dimension or analogous complexity control. The sandwich variance appears in the asymptotic distribution of the ERM solution.

Worked Examples

Example 1 (Huber estimator). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ with F symmetric about μ. The Huber M-estimator with $c = 1.345$ solves $\sum_i \psi_{1.345}(X_i - \hat{\mu}) = 0$. Under $F = \mathcal{N}(\mu, 1)$: $A = -\mathbb{E}[\psi'_{1.345}(Z)] = -(2\Phi(1.345) - 1) \approx -0.8227$ and $B = \mathbb{E}[\psi_{1.345}(Z)^2] \approx 0.7101$. The asymptotic variance is $B/A^2 \approx 1.049$, giving ARE ≈ 0.953 relative to the sample mean — 95.3% efficiency at the normal with much greater robustness to outliers.

Example 2 (Heteroskedasticity-robust inference). In the linear model $Y_i = x_i^\top \beta + \varepsilon_i$ with $\mathbb{E}[\varepsilon_i | x_i] = 0$ but $\operatorname{Var}(\varepsilon_i | x_i) = \sigma^2(x_i)$ (heteroskedastic), OLS is an M-estimator with $\psi(X_i, \beta) = x_i(Y_i - x_i^\top \beta)$. Here $A = -\mathbb{E}[x_i x_i^\top]$ and $B = \mathbb{E}[\varepsilon_i^2 x_i x_i^\top]$. The sandwich variance $A^{-1}BA^{-\top}$ is the Huber-White heteroskedasticity-consistent covariance estimator, estimated by $(X^\top X)^{-1}(\sum_i \hat{e}_i^2 x_i x_i^\top)(X^\top X)^{-1}$ where $\hat{e}_i$ are OLS residuals.

Cross-Links

AI. Empirical risk minimization is M-estimation; the training loss is the M-objective, and the sandwich variance appears in the asymptotic distribution of learned parameters. Stochastic gradient descent can be analyzed as an approximate Z-estimator solving the sample score equation.

Economics. Generalized Method of Moments (GMM) is a Z-estimation framework with $\psi = g(\theta)^\top W \cdot h(X_i, \theta)$; the sandwich variance is the Eicker-Huber-White robust standard error, standard in applied econometrics.

Linear Algebra. The sandwich form $A^{-1}BA^{-\top}$ is a congruence transformation of B by A^{-1}, the same algebraic operation that governs covariance transformation under linear maps (Node 8.1).

Quantum Mechanics. Quantum state estimation via maximum likelihood on measurement outcomes is an M-estimation problem; the sandwich variance accounts for measurement-model mismatch when the assumed measurement model differs from the true POVM.

Structural Compression

Invariant. Z-estimators solving $\mathbb{E}[\psi(X, \theta_0)] = 0$ are $\sqrt{n}$-consistent and asymptotically normal with sandwich covariance $A^{-1}BA^{-\top}$.

Mechanism. Taylor linearization of the estimating equation at the solution maps the CLT for ψ into a Gaussian limit for $\hat{\theta}_n$ via the inverse Jacobian A^{-1}. The sandwich structure emerges because the variability (B) and the curvature (A) of the estimating equation are distinct objects that coincide only under correct specification.

Cross-System Echo. M-estimation is the frequentist analogue of variational inference (both minimize an objective over parameters). The sandwich variance is the standard covariance correction in generalized estimating equations (GEE) in biostatistics, in heteroskedasticity-robust standard errors in econometrics, and in quasi-likelihood theory.

Exercises

1. Show that the MLE is a special case of M-estimation with $m(x, \theta) = \log p_\theta(x)$, and verify that the sandwich reduces to $I(\theta)^{-1}$ under correct specification.
2. Derive the influence function $\mathrm{IF}(x; T, F) = -A^{-1}\psi(x, \theta_0)$ for a Z-estimator and verify that $\int \mathrm{IF}(x; T, F)\, dF(x) = 0$.
3. For the Huber M-estimator with tuning constant c, compute the breakdown point and the asymptotic efficiency at the normal as functions of c.
4. In the linear regression model with heteroskedastic errors, derive the Huber-White sandwich estimator from the general Z-estimator formula with $\psi_i = x_i(Y_i - x_i^\top \beta)$.
5. Show that the sample median is a Z-estimator with $\psi(x, \theta) = \mathrm{sign}(x - \theta)$. Compute A and B when $X \sim \mathcal{N}(\mu, \sigma^2)$ and derive the asymptotic variance $\pi\sigma^2/(2n)$.
6. Consider empirical risk minimization with the hinge loss $m(x, \theta) = -\max(0, 1 - y \cdot x^\top \theta)$. Explain why the standard M-estimation asymptotics do not apply directly (non-differentiability) and describe modifications needed.
7. Argue, structurally, why the sandwich variance is the correct asymptotic covariance even under model misspecification, while the model-based variance A^{-1} alone is invalid, connecting this to the failure of the information matrix equality.

Likelihood Asymptotics

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 7 — Asymptotic Theory. Node 7.6 — Likelihood Asymptotics.

Likelihood asymptotics studies the large-sample geometry of the log-likelihood surface and the asymptotic behavior of all inference procedures derived from it. This node depends on asymptotic normality (Node 7.3), M-estimation (Node 7.5), and the exponential family and sufficiency theory (Nodes 2.3, 2.5). It is the capstone of Module 7, unifying the Wald, Score, and Likelihood Ratio tests under a single asymptotic framework.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}$, $\theta_0 \in \Theta \subseteq \mathbb{R}^k$. Define the log-likelihood $\ell(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$ and the normalized score $\Delta_n = n^{-1/2}\nabla_\theta \ell(\theta_0)$.

Local Asymptotic Normality (LAN). The model $\{P_\theta^{(n)}\}$ satisfies LAN at θ_0 if for every $h \in \mathbb{R}^k$,

$$\log \frac{L(\theta_0 + h/\sqrt{n})}{L(\theta_0)} = h^\top \Delta_n - \frac{1}{2} h^\top I(\theta_0) h + o_p(1),$$

where $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$.

Wilks' theorem. Under $H_0 : \theta = \theta_0$,

$$\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)] \xrightarrow{d} \chi^2_k.$$

Intuition

LAN says that at the local scale $1/\sqrt{n}$ around the truth, every regular parametric model looks like a Gaussian shift experiment. The log-likelihood ratio is a quadratic function of the local parameter h plus Gaussian noise — the same structure as observing $Z = h + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I^{-1})$. All first-order asymptotic inference is determined by this Gaussian approximation. The three classical test statistics (Wald, Score, LR) are simply three different ways to measure distance from the null in this locally Gaussian world.

Structural Role

LAN is the deepest structural statement about likelihood-based inference. It provides the theoretical foundation for: (i) the asymptotic equivalence of Wald, Score, and LR tests (Nodes 4.4, 5.4); (ii) the Cramer-Rao efficiency bound (Node 3.4); (iii) the Bernstein-von Mises theorem (Node 6.7); and (iv) the asymptotic optimality of the MLE. Every regular parametric model is, locally, a Gaussian shift experiment — this is the fundamental simplification that makes asymptotic inference tractable.

Mathematical Development

Score asymptotics. At the true parameter θ_0, the score contributions $s_i = \nabla_\theta \log p_{\theta_0}(X_i)$ are i.i.d. with $\mathbb{E}[s_i] = 0$ (score identity) and $\mathrm{Cov}(s_i) = I(\theta_0)$ (Fisher information). By the multivariate CLT,

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i \xrightarrow{d} \mathcal{N}(0, I(\theta_0)).$$

Proof of LAN. Taylor-expand the log-likelihood ratio:

$$\ell(\theta_0 + h/\sqrt{n}) - \ell(\theta_0) = \frac{h^\top}{\sqrt{n}} \nabla \ell(\theta_0) + \frac{1}{2n} h^\top \nabla^2 \ell(\theta_n^*) h,$$

where θ_n^* lies between θ_0 and $\theta_0 + h/\sqrt{n}$. The first term is $h^\top \Delta_n$. For the second term, $n^{-1}\nabla^2 \ell(\theta_n^*) \xrightarrow{p} -I(\theta_0)$ by the WLLN and continuity. Thus the second term converges to $-h^\top I(\theta_0) h/2$, establishing LAN. □

Asymptotic equivalence of the three test statistics.

Under $H_0 : \theta = \theta_0$:

Wald statistic:

$$W_n = n(\hat{\theta}_n - \theta_0)^\top I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0).$$

Score (Rao) statistic:

$$S_n = \frac{1}{n} \nabla \ell(\theta_0)^\top I(\theta_0)^{-1} \nabla \ell(\theta_0) = \Delta_n^\top I(\theta_0)^{-1} \Delta_n.$$

Likelihood ratio statistic:

$$\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)].$$

All three converge to χ_k^2 under H_0, and $W_n - \Lambda_n = o_p(1)$, $S_n - \Lambda_n = o_p(1)$.

Proof sketch for Wilks' theorem. By LAN with $h = \sqrt{n}(\hat{\theta}_n - \theta_0)$:

$$\Lambda_n = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)] = 2h^\top \Delta_n - h^\top I(\theta_0) h + o_p(1).$$

Using the asymptotic normality of the MLE, $h = \sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1}\Delta_n + o_p(1)$. Substituting:

$$\Lambda_n = 2\Delta_n^\top I(\theta_0)^{-1}\Delta_n - \Delta_n^\top I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} \Delta_n + o_p(1) = \Delta_n^\top I(\theta_0)^{-1}\Delta_n + o_p(1).$$

Since $\Delta_n \xrightarrow{d} \mathcal{N}(0, I(\theta_0))$, the quadratic form $\Delta_n^\top I(\theta_0)^{-1}\Delta_n \xrightarrow{d} \chi_k^2$. □

Composite null hypothesis. For $H_0 : \theta \in \Theta_0$ where Θ_0 is a smooth $(k-q)$-dimensional submanifold of Θ, the LR statistic satisfies $\Lambda_n \xrightarrow{d} \chi_q^2$ under H_0, where $q = k - \dim(\Theta_0)$ is the number of constraints.

Bartlett correction. The LR statistic admits a refinement: defining $\Lambda_n^* = \Lambda_n/(1 + a/n)$ where a depends on the third and fourth cumulants of the log-likelihood, the corrected statistic satisfies $\mathbb{P}(\Lambda_n^* \le x) = F_{\chi_k^2}(x) + O(n^{-2})$, improving the $O(n^{-1})$ error of the uncorrected statistic. This correction is exact for exponential families.

Profile likelihood. For a partition $\theta = (\psi, \lambda)$ where ψ is the parameter of interest and λ is the nuisance parameter, the profile likelihood is $L_P(\psi) = \max_\lambda L(\psi, \lambda)$. The profile log-likelihood ratio $2[\ell_P(\hat{\psi}) - \ell_P(\psi_0)] \xrightarrow{d} \chi^2_{\dim(\psi)}$ under $H_0 : \psi = \psi_0$, eliminating the nuisance parameter.

Asymptotic efficiency of the MLE. Under LAN, the MLE achieves the asymptotic minimax bound: for any loss function $L(\theta, \hat{\theta}) = w(\hat{\theta} - \theta)$ with w subconvex and symmetric, no regular estimator can have smaller asymptotic risk than the MLE. This is the Hajek-Le Cam convolution theorem: any regular estimator $\hat{\theta}_n$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z + W$ where $Z \sim \mathcal{N}(0, I^{-1})$ and W is independent noise. The MLE has $W = 0$.

Contiguity. Two sequences P_n, Q_n are contiguous if $P_n(A_n) \to 0 \implies Q_n(A_n) \to 0$. Under LAN, $P^{(n)}_{\theta_0+h/\sqrt{n}}$ and $P^{(n)}_{\theta_0}$ are contiguous. Le Cam's third lemma: if (T_n, Λ_n) converges jointly under P_{θ_0}, the distribution of T_n under the contiguous alternative is obtained by tilting the joint limit. This enables power calculations for local alternatives.

Power under local alternatives. Under $\theta = \theta_0 + h/\sqrt{n}$, the LR statistic has a noncentral chi-squared distribution:

$$\Lambda_n \xrightarrow{d} \chi^2_k(\delta), \quad \delta = h^\top I(\theta_0) h.$$

The power of the LR test at level α is $\mathbb{P}(\chi^2_k(\delta) > \chi^2_{k,\alpha})$, which increases with the noncentrality δ. The Wald and Score tests have the same asymptotic power function under local alternatives, confirming first-order equivalence.

Higher-order asymptotics. The first-order theory treats all three test statistics as equivalent. Second-order analysis reveals differences: - The LR statistic admits Bartlett correction (improved χ^2 approximation to order $O(n^{-2})$). - The Wald statistic is not invariant under reparametrization: testing $H_0 : \theta = \theta_0$ and $H_0 : g(\theta) = g(\theta_0)$ can give different results. - The Score statistic requires only the restricted (null) estimate, making it computationally simpler.

In practice, the LR test is generally preferred for its reparametrization invariance and Bartlett correctability.

Asymptotic optimality of the Neyman-Pearson test. For testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_0 + h/\sqrt{n}$, the LR test is asymptotically most powerful among all tests of the same level. This follows from the Neyman-Pearson lemma applied to the locally Gaussian limit experiment.

Worked Examples

Example 1 (LR test for normal mean). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$. Test $H_0 : \mu = 0$. The LR statistic is $\Lambda_n = 2[\ell(\bar{X}) - \ell(0)] = 2 \cdot n[\bar{X}^2/2] = n\bar{X}^2$. Under H_0, $\sqrt{n}\bar{X} \sim \mathcal{N}(0,1)$, so $\Lambda_n \sim \chi^2_1$ exactly. The Wald statistic is $W_n = n\bar{X}^2$ and the Score statistic is $S_n = n\bar{X}^2$ — all three coincide exactly for the normal location model.

Example 2 (LR test for Poisson rate). Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. Test $H_0 : \lambda = \lambda_0$. The MLE is $\hat{\lambda} = \bar{X}$. The LR statistic is $\Lambda_n = 2n[\bar{X} \log(\bar{X}/\lambda_0) - (\bar{X} - \lambda_0)]$. Under H_0, $\Lambda_n \xrightarrow{d} \chi^2_1$. The Wald statistic is $W_n = n(\bar{X} - \lambda_0)^2/\bar{X}$ and the Score statistic is $S_n = n(\bar{X} - \lambda_0)^2/\lambda_0$. Here the three statistics differ at finite n; the Bartlett correction improves the chi-squared approximation for Λ_n.

Cross-Links

AI. The likelihood ratio test underpins model selection criteria (AIC, BIC) that penalize the LR statistic by model complexity. In deep learning, the local quadratic approximation of the loss surface (LAN analogue) is the basis for second-order optimization methods (natural gradient, Fisher-information-based preconditioning).

Economics. The LR, Wald, and Score (Lagrange multiplier) tests are the trinity of econometric testing; the Wald test is most common in practice due to its computational simplicity (requires only the unrestricted estimate).

Linear Algebra. The LAN expansion is a second-order Taylor expansion of a function on a Riemannian manifold, where the Fisher information is the metric tensor. The chi-squared limit of the LR statistic is the squared Riemannian distance from the null.

Quantum Mechanics. Quantum Local Asymptotic Normality (qLAN) extends LAN to quantum state estimation: n copies of a quantum state ρ_θ are asymptotically equivalent to a quantum Gaussian shift experiment, with the quantum Fisher information replacing $I(\theta)$.

Structural Compression

Invariant. At local scale $1/\sqrt{n}$, the log-likelihood ratio is a quadratic Gaussian process; all first-order asymptotic inference is determined by the Fisher information $I(\theta_0)$.

Mechanism. Taylor expansion of the log-likelihood to second order; CLT applied to the score; WLLN applied to the Hessian; the resulting Gaussian quadratic form produces χ^2 limits for all three test statistics. Contiguity transfers these results to local alternatives.

Cross-System Echo. LAN is the statistical instance of a general principle in information geometry: the Fisher information metric is the unique Riemannian metric on the manifold of probability distributions invariant under sufficient statistics. The Hajek-Le Cam theory establishes that LAN models are, asymptotically, Gaussian shift experiments — the simplest possible inference problem.

Exercises

1. Verify the LAN expansion explicitly for the Bernoulli model $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ by computing Δ_n, $I(\theta_0)$, and the remainder term.
2. For the exponential family $p_\eta(x) = h(x)\exp(\eta x - A(\eta))$, show that the LAN expansion is exact to second order (the remainder is identically zero).
3. Prove that the Wald, Score, and LR statistics are asymptotically equivalent under H_0 by expressing each as $\Delta_n^\top I(\theta_0)^{-1}\Delta_n + o_p(1)$.
4. Compute the Bartlett correction factor a for the Poisson model and verify that $\Lambda_n/(1 + a/n)$ has improved chi-squared approximation.
5. For the normal model $\mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$, construct the profile likelihood for μ by maximizing over σ^2, and show that the profile LR statistic for $H_0 : \mu = \mu_0$ equals $n\log(1 + T^2/(n-1))$ where T is the usual t-statistic.
6. Use Le Cam's third lemma to derive the power of the LR test for $H_0 : \mu = 0$ against the local alternative $\mu = h/\sqrt{n}$ in the $\mathcal{N}(\mu, 1)$ model.
7. Argue, structurally, why LAN implies that all regular parametric models are asymptotically equivalent to Gaussian shift experiments, and explain why this makes the Fisher information the only relevant local quantity for asymptotic inference.

Module 8 — Multivariate Linear Models

Random Vectors and Covariance Structure

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.1 — Random Vectors and Covariance Structure.

This node establishes the algebraic and geometric framework for multivariate statistics. The mean vector and covariance matrix are the fundamental objects governing linear operations on random vectors. This node depends on the probability foundations (Nodes 1.3, 1.5, 1.6) and enables the multivariate normal (Node 8.2), linear models (Node 8.3), and PCA (Node 8.7).

Formal Definition

Let $X = (X_1, \ldots, X_p)^\top$ be a p-dimensional random vector on $(\Omega, \mathcal{F}, \mathbb{P})$.

Mean vector. $\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p])^\top \in \mathbb{R}^p$.

Covariance matrix. $\Sigma = \mathrm{Cov}(X) = \mathbb{E}[(X-\mu)(X-\mu)^\top] \in \mathbb{R}^{p\times p}$, with $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

Correlation matrix. $R = D^{-1/2}\Sigma D^{-1/2}$ where $D = \mathrm{diag}(\Sigma_{11}, \ldots, \Sigma_{pp})$, with $R_{ij} = \mathrm{Cor}(X_i, X_j) \in [-1, 1]$ and $R_{ii} = 1$.

Intuition

The mean vector locates the center of mass of the distribution. The covariance matrix captures both the spread (diagonal: variances) and the linear dependence structure (off-diagonal: covariances). Geometrically, the covariance matrix defines an ellipsoid: the set $\{x : (x-\mu)^\top \Sigma^{-1}(x-\mu) \le c\}$ is an ellipsoid whose axes are aligned with the eigenvectors of Σ and whose half-lengths are proportional to the square roots of the eigenvalues. The correlation matrix is the standardized version: it strips out the individual scales and retains only the dependence structure.

Structural Role

The covariance matrix is the central object in multivariate analysis. It encodes the variance and dependence structure, governs linear transformations via the congruence $A\Sigma A^\top$, determines the geometry of confidence ellipsoids, regression projections, and PCA directions. In the Gaussian case, μ and Σ fully determine the distribution. The spectral decomposition of Σ provides the principal component directions (Node 8.7) and the Mahalanobis metric for multivariate inference (Node 8.2).

Mathematical Development

Positive semidefiniteness. For any $a \in \mathbb{R}^p$,

$$a^\top \Sigma a = \mathrm{Var}(a^\top X) \ge 0.$$

Thus $\Sigma \succeq 0$. Equality holds for some $a \neq 0$ if and only if $a^\top X$ is degenerate (a.s. constant), meaning the distribution of X is supported on a proper affine subspace. Σ is positive definite ($\Sigma \succ 0$) if and only if no nontrivial linear combination is degenerate.

Linear transformation rule. For $A \in \mathbb{R}^{q\times p}$ and $b \in \mathbb{R}^q$:

$$\mathbb{E}[AX + b] = A\mu + b, \qquad \mathrm{Cov}(AX + b) = A\Sigma A^\top.$$

Proof. $\mathrm{Cov}(AX + b) = \mathbb{E}[A(X - \mu)(X - \mu)^\top A^\top] = A\,\mathbb{E}[(X - \mu)(X - \mu)^\top]\,A^\top = A\Sigma A^\top$. □

This is the congruence transformation: Σ transforms as a bilinear form under linear maps.

Cross-covariance. For random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, the cross-covariance is $\Sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\top] \in \mathbb{R}^{p\times q}$. For the joint vector $(X^\top, Y^\top)^\top$, the covariance is

$$\mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix},$$

with $\Sigma_{YX} = \Sigma_{XY}^\top$.

Spectral decomposition. Since Σ is symmetric positive semidefinite, the spectral theorem gives

$$\Sigma = Q\Lambda Q^\top = \sum_{j=1}^{p} \lambda_j q_j q_j^\top,$$

where $Q = [q_1, \ldots, q_p]$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. The eigenvalues are the variances along the principal directions q_j: $\mathrm{Var}(q_j^\top X) = \lambda_j$.

Mahalanobis distance. For $\Sigma \succ 0$, the Mahalanobis distance between x and μ is

$$d_M(x, \mu) = \sqrt{(x - \mu)^\top \Sigma^{-1}(x - \mu)}.$$

This is the Euclidean distance after whitening: if $Z = \Sigma^{-1/2}(X - \mu)$, then $\mathrm{Cov}(Z) = I_p$ and $d_M(x, \mu) = \|z\|$. The Mahalanobis distance accounts for the scale and correlation structure, making it invariant under affine transformations of the data.

Sample covariance matrix. For observations $X_1, \ldots, X_n$, the sample mean and covariance are

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top.$$

$\bar{X}$ is unbiased for μ and S is unbiased for Σ. Under finite fourth moments, the multivariate CLT gives $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ and $S \xrightarrow{p} \Sigma$.

Partial correlation. The partial correlation of X_i and X_j given the remaining variables is

$$\rho_{ij\cdot\mathrm{rest}} = -\frac{(\Sigma^{-1})_{ij}}{\sqrt{(\Sigma^{-1})_{ii}(\Sigma^{-1})_{jj}}}.$$

Partial correlations encode the conditional linear dependence structure and are read from the precision matrix Σ^{-1}, not from Σ itself.

Whitening transformation. The matrix $W = \Sigma^{-1/2}$ transforms X into $Z = W(X - \mu)$ with $\mathrm{Cov}(Z) = I_p$. The whitening transformation decorrelates the components and normalizes their variances to 1. It is not unique: any matrix W satisfying $W\Sigma W^\top = I$ produces a whitened vector. The symmetric whitening $W = \Sigma^{-1/2}$ (via the matrix square root) is the unique whitening that is also symmetric positive definite.

Cauchy-Schwarz for covariance. For any two random variables X_i, X_j in the vector,

$$|\mathrm{Cov}(X_i, X_j)|^2 \leq \mathrm{Var}(X_i)\mathrm{Var}(X_j),$$

which is equivalent to $|R_{ij}| \leq 1$. This follows from the Cauchy-Schwarz inequality in $L^2(\Omega, \mathbb{P})$. Equality holds if and only if $X_j = aX_i + b$ a.s. for constants a, b.

Block matrix inversion. For the partitioned covariance, the precision matrix is

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix}.$$

The diagonal blocks are the inverses of the Schur complements, connecting covariance structure to conditional distributions (Node 8.2).

Best linear predictor. The best linear predictor of X_1 given X_2 (minimizing $\mathbb{E}[\|X_1 - a - BX_2\|^2]$) is

$$\hat{X}_1 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

with prediction error covariance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. This holds for any distribution with finite second moments, not just the Gaussian. Under Gaussianity, the best linear predictor coincides with the conditional expectation.

Worked Examples

Example 1. Let $X = (X_1, X_2)^\top$ with $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$. The eigenvalues solve $\det(\Sigma - \lambda I) = (4-\lambda)(3-\lambda) - 4 = \lambda^2 - 7\lambda + 8 = 0$, giving $\lambda_1 = (7 + \sqrt{17})/2 \approx 5.56$ and $\lambda_2 = (7 - \sqrt{17})/2 \approx 1.44$. The correlation is $R_{12} = 2/\sqrt{4 \cdot 3} = 1/\sqrt{3} \approx 0.577$. The Mahalanobis distance from the origin to $(1, 1)^\top$ (assuming $\mu = 0$) is $\sqrt{(1,1)\Sigma^{-1}(1,1)^\top} = \sqrt{(1,1)(3,-2;-2,4)(1,1)^\top/8} = \sqrt{3/8} \approx 0.612$, using $\Sigma^{-1} = (1/8)\begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}$.

Example 2. Let $Y = AX + b$ with $A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ and $\Sigma_X = I_2$. Then $\Sigma_Y = AA^\top = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2I_2$. The transformation rotates and scales: the two components of Y are uncorrelated with equal variance 2, even though the transformation is not orthogonal. If instead $\Sigma_X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, then $\Sigma_Y = \begin{pmatrix} 2(1+\rho) & 0 \\ 0 & 2(1-\rho) \end{pmatrix}$ — the transformation A diagonalizes the covariance when the off-diagonal structure aligns with its rows.

Example 3 (Whitening). Let $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$. The Cholesky decomposition gives $\Sigma = LL^\top$ with $L = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{pmatrix}$. The whitening matrix is $W = L^{-1} = \begin{pmatrix} 1/2 & 0 \\ -1/(2\sqrt{2}) & 1/\sqrt{2} \end{pmatrix}$, and $Z = W(X - \mu)$ has $\mathrm{Cov}(Z) = I_2$. This is the Cholesky whitening, which produces a lower-triangular transformation. The symmetric whitening $\Sigma^{-1/2}$ produces a different but equally valid decorrelation.

Example 4 (Best linear predictor). With $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$ and $\mu = (1, 2)^\top$, the best linear predictor of X_1 given $X_2 = x_2$ is $\hat{X}_1 = 1 + (2/3)(x_2 - 2)$. The prediction error variance is $4 - 4/3 = 8/3$. The coefficient $\Sigma_{12}/\Sigma_{22} = 2/3$ is the regression coefficient of X_1 on X_2.

Cross-Links

AI. The empirical covariance matrix is central to PCA, linear discriminant analysis, Gaussian process kernels, and whitening in preprocessing pipelines. In neural networks, the Fisher information matrix (Node 7.3) is a covariance matrix of score vectors, used in natural gradient methods.

Economics. Portfolio theory (Markowitz) uses the covariance matrix of asset returns to compute the efficient frontier; the minimum-variance portfolio is $w \propto \Sigma^{-1}\mathbf{1}$, a direct function of the covariance structure.

Linear Algebra. The covariance matrix is a symmetric positive semidefinite linear operator on $\mathbb{R}^p$. Its spectral decomposition is the spectral theorem for self-adjoint operators; the congruence transformation $A\Sigma A^\top$ is the action of A on the quadratic form defined by Σ.

Quantum Mechanics. The covariance matrix of quadrature operators $(\hat{q}_1, \hat{p}_1, \ldots, \hat{q}_n, \hat{p}_n)$ characterizes Gaussian quantum states; the Heisenberg uncertainty principle imposes $\Sigma + i\Omega/2 \succeq 0$ where Ω is the symplectic form, a constraint without classical analogue.

Structural Compression

Invariant. $\mathrm{Cov}(AX) = A\Sigma A^\top$: covariance transforms by congruence under linear maps.

Mechanism. The covariance matrix is the second-moment operator; congruence transformation is the second-order analogue of the first-order pushforward $\mathbb{E}[AX] = A\mu$. Positive semidefiniteness is preserved because variance is nonnegative.

Cross-System Echo. The covariance matrix is a Riemannian metric on the space of linear functionals of X; its eigendecomposition provides the natural coordinate system (principal components). In information geometry, the Fisher information matrix is a covariance matrix (of the score) and defines the Riemannian metric on the statistical manifold.

Exercises

1. Prove that every symmetric positive semidefinite matrix Σ is a valid covariance matrix by constructing a random vector with that covariance.
2. Let $X \in \mathbb{R}^p$ and $Y = AX$ for invertible A. Show that the Mahalanobis distance is invariant: $d_M^Y(Ax, A\mu) = d_M^X(x, \mu)$.
3. Prove that the sample covariance matrix S is unbiased for Σ, and explain why the divisor is $n-1$ rather than n.
4. For $X = (X_1, X_2, X_3)^\top$ with $\Sigma^{-1} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{pmatrix}$, compute the partial correlation $\rho_{13\cdot 2}$ and interpret the zero in $(\Sigma^{-1})_{13}$.
5. Show that $\mathrm{tr}(\Sigma) = \sum_j \mathrm{Var}(X_j) = \sum_j \lambda_j$, connecting the total variance to the trace and the eigenvalue sum.
6. Prove that $|\mathrm{Cor}(X_i, X_j)| = 1$ if and only if $X_j = aX_i + b$ a.s. for some constants $a \neq 0, b$.
7. Argue, structurally, why the covariance matrix is the natural inner product on the space of linear functionals of X, and why its spectral decomposition provides the coordinate system that diagonalizes this inner product.

Multivariate Normal Distribution

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.2 — Multivariate Normal Distribution.

The multivariate normal (MVN) is the canonical multivariate distribution, completely determined by its first two moments. It is the reference distribution for all finite-sample inference in linear models. This node depends on the covariance structure (Node 8.1) and distribution theory (Node 1.4), and enables linear models (Node 8.3), inference (Node 8.5), and PCA (Node 8.7).

Formal Definition

A random vector $X \in \mathbb{R}^p$ has the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ if every linear combination is univariate normal:

$$a^\top X \sim \mathcal{N}(a^\top \mu,\ a^\top \Sigma a) \quad \forall\, a \in \mathbb{R}^p.$$

When $\Sigma \succ 0$, the density is

$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\!\left(-\tfrac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

The moment generating function is $M(t) = \exp(t^\top \mu + \tfrac{1}{2} t^\top \Sigma t)$.

Intuition

The MVN generalizes the bell curve to multiple dimensions. The density is constant on ellipsoids centered at μ, with axes aligned along the eigenvectors of Σ and half-lengths proportional to the square roots of the eigenvalues. The key structural property is that the entire dependence structure is captured by Σ: for jointly normal vectors, uncorrelatedness implies independence, and all conditional and marginal distributions remain normal. This makes the MVN the only distribution where second-moment analysis is sufficient for complete probabilistic characterization.

Structural Role

The MVN is the reference distribution for multivariate statistics. Its closure under affine transformations, the normality of marginals and conditionals, and the characterization of quadratic forms underlie all finite-sample distributional results in the normal linear model — t-tests, F-tests, confidence ellipsoids. The Wishart distribution (the sampling distribution of the sample covariance) is derived from the MVN. In asymptotic theory, the MVN appears as the limit distribution in the multivariate CLT.

Mathematical Development

Affine closure. If $X \sim \mathcal{N}(\mu, \Sigma)$ and $Y = AX + b$ with $A \in \mathbb{R}^{q \times p}$, then

$$Y \sim \mathcal{N}(A\mu + b,\ A\Sigma A^\top).$$

Proof. For any $c \in \mathbb{R}^q$, $c^\top Y = c^\top AX + c^\top b = (A^\top c)^\top X + c^\top b$, which is univariate normal since X is MVN. The mean is $c^\top(A\mu + b)$ and variance is $c^\top A\Sigma A^\top c$. $\square$

Marginal distributions. Partition $X = (X_1^\top, X_2^\top)^\top$ with $X_1 \in \mathbb{R}^{p_1}$, $X_2 \in \mathbb{R}^{p_2}$, and correspondingly

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Theorem. $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$.

Proof. $X_1 = (I_{p_1}, 0)X$, so by affine closure, $X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$. $\square$

Conditional distributions. The conditional distribution of $X_1 \mid X_2 = x_2$ is

$$X_1 \mid X_2 = x_2 \;\sim\; \mathcal{N}\big(\mu_{1|2},\, \Sigma_{1|2}\big)\,,$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof. Define $Z = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$. Then $\mathrm{Cov}(Z, X_2) = \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0$. Since (Z, X_2) is jointly normal (as an affine function of X) and uncorrelated, $Z \perp\!\!\!\perp X_2$. Therefore $X_1 \mid X_2 = x_2$ has the same distribution as $Z + \Sigma_{12}\Sigma_{22}^{-1}x_2$, giving the stated mean and variance. $\square$

The conditional variance $\Sigma_{1|2}$ is the Schur complement of Σ_{22} in Σ. It does not depend on the conditioning value x_2 — this is the homoskedasticity of the MVN conditional.

Independence and uncorrelatedness. For jointly normal X_1, X_2:

$$\Sigma_{12} = 0 \iff X_1 \perp\!\!\!\perp X_2.$$

The forward direction follows because uncorrelated jointly normal vectors have independent MGFs: $M_{X_1,X_2}(t_1, t_2) = M_{X_1}(t_1)M_{X_2}(t_2)$.

Quadratic forms. If $X \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, then

$$(X - \mu)^\top \Sigma^{-1}(X - \mu) \sim \chi^2_p.$$

Proof. Let $Z = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_p)$. Then $(X - \mu)^\top\Sigma^{-1}(X - \mu) = Z^\top Z = \sum_{j=1}^p Z_j^2 \sim \chi^2_p$. $\square$

More generally, if $X \sim \mathcal{N}(0, I_n)$ and A is symmetric idempotent with rank r, then $X^\top AX \sim \chi^2_r$. This is the fundamental result connecting projection matrices to chi-squared distributions (used in Node 8.5).

Cochran's theorem. Let $X \sim \mathcal{N}(0, \sigma^2 I_n)$ and let $A_1, \ldots, A_m$ be symmetric matrices with $\sum_j A_j = I_n$ and $\mathrm{rank}(A_j) = r_j$ with $\sum_j r_j = n$. If each A_j is idempotent (equivalently, each $A_j \succeq 0$ and the rank condition holds), then $X^\top A_1 X/\sigma^2, \ldots, X^\top A_m X/\sigma^2$ are independent $\chi^2_{r_j}$ random variables.

Proof sketch. Since A_j is symmetric idempotent with rank r_j, it has eigendecomposition $A_j = U_j U_j^\top$ where $U_j \in \mathbb{R}^{n \times r_j}$ has orthonormal columns. The condition $\sum A_j = I_n$ implies the columns of $U_1, \ldots, U_m$ form an orthonormal basis of $\mathbb{R}^n$. Define $Z_j = U_j^\top X/\sigma \sim \mathcal{N}(0, I_{r_j})$. Then $X^\top A_j X/\sigma^2 = Z_j^\top Z_j \sim \chi^2_{r_j}$, and the Z_j are independent because $U_j^\top U_k = 0$ for $j \neq k$. $\square$

Characterization via MGF. A random vector X is MVN if and only if its MGF has the form $M(t) = \exp(t^\top\mu + t^\top\Sigma t/2)$. This characterization is useful for proving closure properties: if X has this MGF and $Y = AX + b$, then $M_Y(t) = \exp(t^\top(A\mu + b) + t^\top A\Sigma A^\top t/2)$, confirming $Y \sim \mathcal{N}(A\mu + b, A\Sigma A^\top)$.

Singular MVN. When Σ is positive semidefinite but not positive definite (rank $r < p$), X is supported on an r-dimensional affine subspace and has no density with respect to Lebesgue measure on $\mathbb{R}^p$. The distribution is still well-defined: $X = \mu + BZ$ where $Z \sim \mathcal{N}(0, I_r)$ and $B \in \mathbb{R}^{p \times r}$ satisfies $BB^\top = \Sigma$. All linear combination properties, marginal/conditional formulas, and the Portmanteau characterization remain valid.

Stein's lemma. If $X \sim \mathcal{N}(\mu, \Sigma)$ and $g : \mathbb{R}^p \to \mathbb{R}$ is differentiable with $\mathbb{E}[|\nabla g(X)|] < \infty$, then

$$\mathrm{Cov}(X_j, g(X)) = \sum_{k=1}^{p} \Sigma_{jk} \mathbb{E}\left[\frac{\partial g}{\partial x_k}(X)\right].$$

This identity characterizes the normal distribution and is the basis for Stein's method for proving normal approximation. In the scalar case: $\mathrm{Cov}(X, g(X)) = \sigma^2 \mathbb{E}[g'(X)]$.

Multivariate CLT. If $X_1, \ldots, X_n$ are i.i.d. in $\mathbb{R}^p$ with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Cov}(X_i) = \Sigma$, then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. The limit is MVN regardless of the distribution of X_i — the Gaussian character is inherited from the CLT, not assumed. This is the reason the MVN appears throughout asymptotic statistics.

Wishart distribution. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$. The matrix $W = \sum_{i=1}^{n} X_i X_i^\top$ follows the Wishart distribution $W_p(n, \Sigma)$. The density (for $n \geq p$) is

$$f(W) = \frac{|W|^{(n-p-1)/2}}{2^{np/2} |\Sigma|^{n/2} \Gamma_p(n/2)} \exp\!\left(-\tfrac{1}{2}\mathrm{tr}(\Sigma^{-1} W)\right),$$

where Γ_p is the multivariate gamma function. The Wishart is the multivariate generalization of the chi-squared distribution: when $p = 1$, $W_1(n, \sigma^2) = \sigma^2 \chi^2_n$.

Worked Examples

Example 1. Let $X \sim \mathcal{N}\left(\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}4 & 2\\2 & 3\end{pmatrix}\right)$. The conditional distribution of $X_1 \mid X_2 = 5$ is $\mathcal{N}(1 + (2/3)(5 - 2),\ 4 - 4/3) = \mathcal{N}(3,\ 8/3)$. The conditional mean is a linear function of the conditioning value: a unit increase in X_2 shifts the conditional mean of X_1 by $\Sigma_{12}/\Sigma_{22} = 2/3$. The conditional variance $8/3$ is less than the marginal variance 4, reflecting the information gained from observing X_2.

Example 2. Let $Z \sim \mathcal{N}(0, I_3)$. The quadratic form $Q = Z_1^2 + Z_2^2 + Z_3^2 \sim \chi^2_3$. Let H be the projection onto $\mathrm{span}(1,1,1)^\top/\sqrt{3}$, so $H = \mathbf{1}\mathbf{1}^\top/3$. Then $Z^\top H Z / 1 = (\bar{Z})^2 \cdot 3 \sim \chi^2_1$ and $Z^\top (I - H) Z \sim \chi^2_2$, with the two quadratic forms independent by Cochran's theorem. This decomposition is the simplest instance of the ANOVA F-test structure.

Example 3 (Conditional distribution computation). Let $X \sim \mathcal{N}\left(0, \begin{pmatrix}1 & 0.5 & 0.3\\0.5 & 1 & 0.4\\0.3 & 0.4 & 1\end{pmatrix}\right)$. Partition as $X_1 = X_1$ (scalar) and $X_2 = (X_2, X_3)^\top$. Then $\Sigma_{12} = (0.5, 0.3)$, $\Sigma_{22} = \begin{pmatrix}1 & 0.4\\0.4 & 1\end{pmatrix}$, $\Sigma_{22}^{-1} = \frac{1}{0.84}\begin{pmatrix}1 & -0.4\\-0.4 & 1\end{pmatrix}$. The conditional mean of $X_1 \mid X_2 = x_2, X_3 = x_3$ is $\Sigma_{12}\Sigma_{22}^{-1}(x_2, x_3)^\top = \frac{1}{0.84}(0.5 - 0.12)x_2 + (0.3 - 0.2)x_3) = \frac{1}{0.84}(0.38 x_2 + 0.1 x_3)$. The conditional variance is $1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - \frac{1}{0.84}(0.5 \cdot 0.38 + 0.3 \cdot 0.1) \approx 0.738$.

Cross-Links

AI. Gaussian processes use the MVN as the finite-dimensional distribution of function values; the conditional distribution formula gives the posterior predictive distribution. Variational autoencoders use the MVN as the latent space prior and the reparametrization trick exploits affine closure.

Economics. The MVN is the standard assumption in portfolio theory (Markowitz) and CAPM; the mean-variance efficient frontier is derived from the MVN assumption on asset returns.

Linear Algebra. The MVN density is an exponential of a negative definite quadratic form; level sets are ellipsoids determined by the spectral decomposition of Σ^{-1}. The Schur complement in the conditional variance formula is the same Schur complement that appears in block matrix inversion.

Quantum Mechanics. Gaussian quantum states are characterized by the first two moments of the quadrature operators, with the covariance matrix subject to the Heisenberg uncertainty constraint $\Sigma + i\Omega/2 \succeq 0$. The conditional distribution formula has a direct quantum analogue in Gaussian state conditioning (homodyne detection).

Structural Compression

Invariant. The MVN is determined by its first two moments; all higher-order cumulants vanish. Marginals, conditionals, and affine transformations of MVN vectors are MVN.

Mechanism. The density is an exponential of a negative definite quadratic form in x; level sets are ellipsoids aligned with the eigenvectors of Σ^{-1}. The MGF $\exp(t^\top \mu + t^\top \Sigma t/2)$ encodes all moments and makes closure under affine maps immediate.

Cross-System Echo. The MVN is the maximum-entropy distribution with fixed mean and covariance (information-theoretic characterization). In information geometry, the Gaussian family forms a statistical manifold with Fisher metric determined by the precision matrix. The Wishart distribution generalizes the chi-squared to matrix-valued sufficient statistics.

Exercises

1. Let $X \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$. Derive the density by applying the change-of-variables formula to $Z = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_p)$.
2. Prove that if $X \sim \mathcal{N}(\mu, \Sigma)$ and $\Sigma_{12} = 0$, then $X_1 \perp\!\!\!\perp X_2$, using the MGF factorization.
3. Let $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Compute the conditional distribution $X_1 \mid X_2 = x_2$ and show that the regression of X_1 on X_2 has slope ρ and residual variance $1 - \rho^2$.
4. Prove Cochran's theorem for the case $m = 2$: if $A_1 + A_2 = I_n$ with A_1, A_2 symmetric and idempotent, then $X^\top A_1 X$ and $X^\top A_2 X$ are independent chi-squared.
5. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$. Show that $\bar{X}$ and $S = (n-1)^{-1} \sum (X_i - \bar{X})(X_i - \bar{X})^\top$ are independent, and that $(n-1)S \sim W_p(n-1, \Sigma)$.
6. Compute the entropy of $\mathcal{N}(\mu, \Sigma)$ and show it depends only on $|\Sigma|$, confirming the maximum-entropy characterization.
7. Argue, structurally, why the equivalence of uncorrelatedness and independence under joint normality is a special property of the Gaussian and fails for general distributions, connecting this to the role of higher-order cumulants.

Linear Models: Geometry and OLS

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.3 — Linear Models: Geometry and OLS.

The linear model is the most fundamental regression framework in statistics. This node develops OLS as an orthogonal projection, establishes the geometric interpretation, and derives the properties of the hat matrix and residuals. It depends on the covariance structure (Node 8.1) and the MVN (Node 8.2), and enables Gauss-Markov (Node 8.4), inference (Node 8.5), and GLMs (Node 8.6).

Formal Definition

The linear model is

$$Y = X\beta + \varepsilon,$$

where $Y \in \mathbb{R}^n$ is the response, $X \in \mathbb{R}^{n \times p}$ is the design matrix with $\text{rank}(X) = p \leq n$, $\beta \in \mathbb{R}^p$ is the coefficient vector, and $\varepsilon \in \mathbb{R}^n$ satisfies $\mathbb{E}[\varepsilon] = 0$, $\text{Cov}(\varepsilon) = \sigma^2 I_n$.

The **OLS estimator** is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Intuition

The column space $\mathcal{C}(X) = \{X\beta : \beta \in \mathbb{R}^p\}$ is a p-dimensional subspace of $\mathbb{R}^n$. The model says the mean response $\mathbb{E}[Y] = X\beta$ lies in this subspace. OLS finds the point in $\mathcal{C}(X)$ closest to the observed Y in Euclidean distance — this is orthogonal projection. The residuals are the component of Y perpendicular to $\mathcal{C}(X)$, and the fitted values are the projection. This geometric view makes the algebraic properties of OLS transparent.

Structural Role

The linear model separates signal $X\beta \in \mathcal{C}(X)$ from noise $\varepsilon \perp \mathcal{C}(X)$ via orthogonal projection. The hat matrix H is the central object: it encodes both the estimator and the residuals, and its rank determines degrees of freedom. Optimality of OLS under second-moment assumptions is established in Node 8.4. Under normality, OLS is also the MLE, and exact finite-sample inference follows (Node 8.5).

Mathematical Development

Derivation of the normal equations. The OLS criterion is $\min_\beta \|Y - X\beta\|^2$. Expanding:

$$\|Y - X\beta\|^2 = Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta.$$

Taking the gradient and setting to zero:

$$\nabla_\beta \|Y - X\beta\|^2 = -2X^\top Y + 2X^\top X\beta = 0.$$

This gives the **normal equations** $X^\top X\hat{\beta} = X^\top Y$. Since $\text{rank}(X) = p$, $X^\top X$ is invertible and $\hat{\beta} = (X^\top X)^{-1}X^\top Y$.

Geometric interpretation. The normal equations $X^\top(Y - X\hat{\beta}) = 0$ state that the residual $Y - X\hat{\beta}$ is orthogonal to every column of X, hence orthogonal to $\mathcal{C}(X)$. This is the defining property of orthogonal projection.

Hat matrix. The fitted values are

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y = HY,$$

where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix.

Properties of H: - Symmetric: $H^\top = H$. - Idempotent: $H^2 = H$ (projecting twice is the same as projecting once). - $\text{rank}(H) = \text{tr}(H) = p$. - $HX = X$ (columns of X are in the range of H). - Eigenvalues of H are 0 or 1, with exactly p ones.

Residuals. The residuals are

$$\hat{\varepsilon} = Y - \hat{Y} = (I_n - H)Y.$$

The matrix $I_n - H$ is also symmetric and idempotent, with rank $n - p$. Key property: $(I - H)X = 0$, so the residuals are orthogonal to every column of X.

Orthogonal decomposition. The observation vector decomposes as

$$Y = HY + (I - H)Y = \hat{Y} + \hat{\varepsilon},$$

with $\hat{Y} \perp \hat{\varepsilon}$. The Pythagorean theorem gives

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2.$$

After centering (subtracting the mean), this becomes the ANOVA decomposition: TSS = ESS + RSS.

Moments of the OLS estimator.

$$\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1}X^\top \mathbb{E}[Y] = (X^\top X)^{-1}X^\top X\beta = \beta.$$

$$\text{Cov}(\hat{\beta}) = (X^\top X)^{-1}X^\top \text{Cov}(Y)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

OLS is unbiased for any error distribution with mean zero and any σ^2.

Residual sum of squares.

$$\text{RSS} = \|\hat{\varepsilon}\|^2 = Y^\top(I - H)Y.$$

Since $\mathbb{E}[Y^\top(I - H)Y] = \text{tr}((I - H)\sigma^2 I) + \beta^\top X^\top(I - H)X\beta = \sigma^2(n - p)$ (using $(I - H)X = 0$), we get

$$\mathbb{E}[\text{RSS}] = \sigma^2(n - p),$$

so $\hat{\sigma}^2 = \text{RSS}/(n - p)$ is unbiased for σ^2.

Leverage scores. The diagonal elements $h_{ii} = H_{ii}$ are the leverage scores. They satisfy $0 \le h_{ii} \le 1$ and $\sum_i h_{ii} = p$. The leverage h_{ii} measures the influence of the i-th observation on its own fitted value: $\hat{Y}_i = h_{ii}Y_i + \sum_{j \ne i} h_{ij}Y_j$. High-leverage points have h_{ii} close to 1 and dominate their own fit.

Residual properties. Under homoskedasticity, $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(I - H)$, so $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$. The residuals are heteroskedastic even when the errors are not. Standardized residuals $r_i = \hat{\varepsilon}_i / (\hat{\sigma}\sqrt{1 - h_{ii}})$ correct for this.

Connection to the Moore-Penrose pseudoinverse. When X has full column rank, $(X^\top X)^{-1}X^\top = X^+$ is the Moore-Penrose pseudoinverse. The OLS solution is $\hat{\beta} = X^+ Y$, the minimum-norm least-squares solution.

Worked Examples

Example 1 (Simple linear regression). Let $X = (\mathbf{1}, x) \in \mathbb{R}^{n \times 2}$ with $x = (x_1, \ldots, x_n)^\top$. Then

$$X^\top X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad X^\top Y = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Solving gives $\hat{\beta}_1 = \sum (x_i - \bar{x})(Y_i - \bar{Y}) / \sum (x_i - \bar{x})^2$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$. The hat matrix has $h_{ii} = 1/n + (x_i - \bar{x})^2 / \sum (x_j - \bar{x})^2$, showing that leverage increases with distance from $\bar{x}$.

Example 2 (One-way ANOVA). For k groups with n_j observations each, X is a block-diagonal indicator matrix. The hat matrix projects onto group means: $\hat{Y}_{ij} = \bar{Y}_{j\cdot}$. The residual is $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{j\cdot}$. RSS $= \sum_j \sum_i (Y_{ij} - \bar{Y}_{j\cdot})^2$ with $n - k$ degrees of freedom, and ESS $= \sum_j n_j (\bar{Y}_{j\cdot} - \bar{Y})^2$ with $k - 1$ degrees of freedom.

Cross-Links

AI. Linear regression is empirical risk minimization under squared loss. The hat matrix appears in leave-one-out cross-validation via the shortcut $\mathrm{CV} = n^{-1} \sum (\hat{\varepsilon}_i / (1 - h_{ii}))^2$, avoiding n refits. Ridge regression replaces H with $X(X^\top X + \lambda I)^{-1} X^\top$, shrinking the projection.

Economics. The linear model is the workhorse of empirical economics. OLS with heteroskedasticity-robust standard errors (Huber-White) replaces $\sigma^2 (X^\top X)^{-1}$ with the sandwich covariance (Node 7.5). Instrumental variables estimation projects onto a different column space.

Linear Algebra. OLS is the orthogonal projection theorem in $\mathbb{R}^n$ with the standard inner product. The normal equations are the first-order optimality conditions for a convex quadratic. The hat matrix is a rank-p orthogonal projector, decomposing $\mathbb{R}^n = \mathcal{C}(X) \oplus \mathcal{C}(X)^\perp$.

Quantum Mechanics. In quantum state tomography, linear inversion estimators reconstruct the density matrix by projecting measurement outcomes onto the space of valid states, the same projection structure as OLS but on the cone of positive semidefinite matrices.

Structural Compression

Invariant. $\hat{Y} = HY$ is the orthogonal projection of Y onto $\mathcal{C}(X)$; the OLS estimator is the unique solution to the normal equations $X^\top X \hat{\beta} = X^\top Y$.

Mechanism. Minimizing $\|Y - X\beta\|^2$ over β is equivalent to finding the closest point in $\mathcal{C}(X)$ to Y; the projection theorem yields the hat matrix as the unique orthogonal projector. The orthogonal decomposition $Y = \hat{Y} + \hat{\varepsilon}$ separates signal from noise.

Cross-System Echo. OLS is the Moore-Penrose pseudoinverse solution. In numerical linear algebra, the QR decomposition provides a numerically stable computation: $X = QR$ gives $\hat{\beta} = R^{-1}Q^\top Y$ and $H = QQ^\top$. In AI, the leverage scores h_{ii} determine the effective number of parameters and the generalization error via the optimism.

Exercises

1. Derive the OLS estimator by completing the square in $\|Y - X\beta\|^2$ rather than by differentiation.
2. Show that H is the unique symmetric idempotent matrix with $\mathcal{C}(H) = \mathcal{C}(X)$.
3. Prove that $\sum_{i=1}^{n} h_{ii} = p$ using $\mathrm{tr}(H) = \mathrm{tr}(X(X^\top X)^{-1}X^\top)$ and the cyclic property of trace.
4. For simple linear regression with equally spaced $x_i = i$, compute the leverage scores and identify the points with maximum leverage.
5. Show that adding a predictor to the model (increasing p by 1) always decreases RSS, and relate this to the geometry of the projection.
6. Prove the leave-one-out formula: the i-th cross-validation residual is $\hat{\varepsilon}_i^{(-i)} = \hat{\varepsilon}_i/(1 - h_{ii})$ using the Sherman-Morrison-Woodbury formula.
7. Argue, structurally, why the hat matrix is the fundamental object in the linear model — more fundamental than $\hat{\beta}$ — by showing that all inferential quantities (fitted values, residuals, RSS, leverage, degrees of freedom) are determined by H alone.

Gauss-Markov Theorem

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.4 — Gauss-Markov Theorem.

The Gauss-Markov theorem establishes OLS as optimal within the class of linear unbiased estimators under second-moment assumptions. This is a finite-sample result requiring no distributional assumptions beyond the first two moments. This node depends on the linear model geometry (Node 8.3) and estimation theory (Node 3.1), and enables inference in linear models (Node 8.5).

Formal Definition

Let $Y = X\beta + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$, and $\mathrm{rank}(X) = p$.

An estimator $\tilde{\beta} = CY$ is **linear** (in Y) and **unbiased** if $CX = I_p$.

Gauss-Markov Theorem. The OLS estimator $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ is the Best Linear Unbiased Estimator (BLUE): for any linear unbiased estimator $\tilde{\beta} = CY$,

$$\mathrm{Cov}(\tilde{\beta}) - \mathrm{Cov}(\hat{\beta}) \succeq 0 \quad \text{(positive semidefinite)}.$$

Equivalently, for every $a \in \mathbb{R}^p$, $\mathrm{Var}(a^\top \tilde{\beta}) \geq \mathrm{Var}(a^\top \hat{\beta})$.

Intuition

Among all estimators that are linear functions of Y and unbiased for β, OLS has the smallest variance in every direction. The reason is geometric: OLS uses the minimum-norm linear map from Y to $\hat{\beta}$, namely the pseudoinverse $X^+ = (X^\top X)^{-1} X^\top$. Any other linear unbiased estimator adds a component D that is orthogonal to the column space of X (to preserve unbiasedness) but contributes additional variance $\sigma^2 DD^\top \succeq 0$.

Structural Role

Gauss-Markov establishes the optimality of OLS within a restricted class (linear, unbiased) under minimal assumptions (first two moments only). It is a finite-sample result — no asymptotics needed. It does not claim OLS is the best among all estimators: biased estimators (ridge, LASSO) or nonlinear estimators can have lower MSE. Under normality, OLS is additionally the MLE and achieves the Cramer-Rao bound, giving a stronger optimality. When the homoskedasticity assumption fails, GLS replaces OLS as the BLUE.

Mathematical Development

Proof of the Gauss-Markov theorem.

Let $\tilde{\beta} = CY$ be any linear unbiased estimator. Write $C = (X^\top X)^{-1} X^\top + D$ where $D = C - (X^\top X)^{-1} X^\top$.

Unbiasedness constraint: $CX = I_p$ requires $(X^\top X)^{-1} X^\top X + DX = I_p + DX = I_p$, so $DX = 0$.

Covariance computation:

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2 CC^\top = \sigma^2[(X^\top X)^{-1}X^\top + D][(X^\top X)^{-1}X^\top + D]^\top.$$

Expanding, using $DX = 0$ (so $D \cdot X(X^\top X)^{-1} = 0$, meaning the cross terms vanish):

$$\mathrm{Cov}(\tilde{\beta}) = \sigma^2(X^\top X)^{-1} + \sigma^2 DD^\top = \mathrm{Cov}(\hat{\beta}) + \sigma^2 DD^\top.$$

Since $DD^\top \succeq 0$, we have $\mathrm{Cov}(\tilde{\beta}) \succeq \mathrm{Cov}(\hat{\beta})$. $\square$

Why the cross terms vanish: $(X^\top X)^{-1}X^\top D^\top = [(X^\top X)^{-1}(DX)^\top]^\top$. Since $DX = 0$, this is zero. Geometrically, the rows of D lie in $\mathcal{C}(X)^\perp$ while the rows of $(X^\top X)^{-1}X^\top$ lie in $\mathcal{C}(X)$, making them orthogonal.

Scalar estimand version. For any linear functional $\psi = a^\top\beta$, the BLUE is $a^\top\hat{\beta}$ with variance $\sigma^2 a^\top(X^\top X)^{-1}a$. Any other linear unbiased estimator of ψ has variance at least this large.

Assumptions revisited. The Gauss-Markov theorem requires: 1. Linearity: $\mathbb{E}[Y] = X\beta$. 2. Homoskedasticity: $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. 3. Full column rank of X.

It does not require: normality of ε, independence of observations (only uncorrelatedness and equal variance), or any knowledge of σ^2.

When assumptions fail: Generalized Least Squares.

If $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega$ with known $\Omega \succ 0$ and $\Omega \neq I$, OLS is no longer BLUE. The BLUE is the GLS estimator:

$$\hat{\beta}_{\mathrm{GLS}} = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}Y.$$

Derivation. Transform the model by $\Omega^{-1/2}$: let $Y^* = \Omega^{-1/2}Y$, $X^* = \Omega^{-1/2}X$, $\varepsilon^* = \Omega^{-1/2}\varepsilon$. Then $\mathrm{Cov}(\varepsilon^*) = \sigma^2 I_n$, so OLS applied to the transformed model gives

$$\hat{\beta}_{\mathrm{GLS}} = (X^{*\top}X^*)^{-1}X^{*\top}Y^* = (X^\top\Omega^{-1}X)^{-1}X^\top\Omega^{-1}Y.$$

By Gauss-Markov applied to the transformed model, this is BLUE.

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}}) = \sigma^2(X^\top\Omega^{-1}X)^{-1}.$$

Feasible GLS. When Ω is unknown, replace it by a consistent estimator $\hat{\Omega}$. The feasible GLS estimator $\hat{\beta}_{\mathrm{FGLS}} = (X^\top\hat{\Omega}^{-1}X)^{-1}X^\top\hat{\Omega}^{-1}Y$ is asymptotically equivalent to GLS but not exactly BLUE in finite samples.

Beyond BLUE: bias-variance tradeoff. Ridge regression $\hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1}X^\top Y$ is biased but has MSE

$$\mathrm{MSE}(\hat{\beta}_\lambda) = \sigma^2\mathrm{tr}[(X^\top X + \lambda I)^{-2}X^\top X] + \lambda^2\beta^\top(X^\top X + \lambda I)^{-2}\beta.$$

For any $\beta \neq 0$ and $\lambda > 0$ small enough, $\mathrm{MSE}(\hat{\beta}_\lambda) < \mathrm{MSE}(\hat{\beta}) = \sigma^2\mathrm{tr}(X^\top X)^{-1}$. Gauss-Markov does not contradict this because ridge regression is biased; BLUE is optimal only within the unbiased class.

Under normality: stronger optimality. If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, then OLS is the MLE and achieves the Cramer-Rao lower bound among all unbiased estimators (not just linear ones). Under normality, BLUE = UMVUE for $a^\top\beta$.

Weighted least squares as GLS. When $\mathrm{Var}(\varepsilon_i) = \sigma^2/w_i$ with known weights $w_i > 0$, the covariance is $\mathrm{Cov}(\varepsilon) = \sigma^2 W^{-1}$ where $W = \mathrm{diag}(w_1, \ldots, w_n)$. GLS gives

$$\hat{\beta}_{\mathrm{WLS}} = (X^\top WX)^{-1}X^\top WY,$$

which minimizes $\sum_i w_i(Y_i - x_i^\top \beta)^2$. Observations with smaller variance (larger w_i) receive more weight — they are more informative.

Efficiency loss of OLS under heteroskedasticity. When $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega \neq \sigma^2 I$, OLS remains unbiased ($\mathbb{E}[\hat{\beta}] = \beta$) but its covariance is

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) = \sigma^2 (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1},$$

which differs from the naive $\sigma^2(X^\top X)^{-1}$. The difference $\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) - \mathrm{Cov}(\hat{\beta}_{\mathrm{GLS}}) \succeq 0$, quantifying the efficiency loss from ignoring the error structure.

Gauss-Markov for estimable functions. When X does not have full column rank, β is not identifiable, but linear functions $a^\top\beta$ that lie in $\mathcal{C}(X^\top)$ are still estimable. The Gauss-Markov theorem extends: for any estimable $a^\top\beta$, the OLS estimate (via the pseudoinverse) is BLUE.

Worked Examples

Example 1. Consider the model $Y_i = \mu + \varepsilon_i$ with $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$. Here $X = \mathbf{1}_n$, so $\hat{\mu} = \bar{Y}$ with $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$. Any other linear unbiased estimator is $\tilde{\mu} = \sum c_i Y_i$ with $\sum c_i = 1$, and $\mathrm{Var}(\tilde{\mu}) = \sigma^2 \sum c_i^2 \geq \sigma^2/n$ by Cauchy-Schwarz with equality iff $c_i = 1/n$ for all i.

Example 2 (Heteroskedastic errors). Let $Y_i = \beta x_i + \varepsilon_i$ with $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$ (no intercept). Here $\Omega = \mathrm{diag}(x_1^2, \ldots, x_n^2)$. The OLS estimator is $\hat{\beta}_{\mathrm{OLS}} = \sum x_i Y_i / \sum x_i^2$, but the GLS estimator is $\hat{\beta}_{\mathrm{GLS}} = \sum Y_i/x_i / \sum 1 = n^{-1}\sum Y_i/x_i$, which weights each observation inversely proportional to its variance. GLS has smaller variance than OLS in this setting.

Example 3 (Ridge vs. OLS). Consider $X^\top X = I_p$, $n = p$, $\sigma^2 = 1$, $\|\beta\|^2 = B^2$. The MSE of OLS is $\mathrm{tr}(I_p) = p$. The MSE of ridge with λ is $p/(1+\lambda)^2 + \lambda^2 B^2/(1+\lambda)^2$. Setting the derivative to zero: optimal $\lambda^* = p/B^2$, giving MSE $= pB^2/(p+B^2) < p$. Ridge strictly dominates OLS for every $B > 0$ when $p \geq 3$ (Stein's phenomenon). The Gauss-Markov theorem is not violated because ridge is biased.

Cross-Links

AI. Gauss-Markov justifies OLS in the unbiased regime; regularized methods (ridge, LASSO) trade bias for variance reduction, stepping outside the BLUE framework. The bias-variance tradeoff is the central concept in statistical learning theory.

Economics. GLS is the basis for correcting heteroskedasticity (weighted least squares) and serial correlation (Cochrane-Orcutt, Prais-Winsten) in time series econometrics. The Hausman test compares OLS and GLS to detect misspecification.

Linear Algebra. The Gauss-Markov proof uses the decomposition of the linear map $C = X^+ + D$ into the pseudoinverse plus a perturbation orthogonal to the column space. The Loewner order $\succeq$ on covariance matrices is the partial order on symmetric matrices induced by the cone of positive semidefinite matrices.

Quantum Mechanics. The BLUE concept has an analogue in quantum estimation: the locally unbiased estimator with minimum variance under the symmetric logarithmic derivative inner product achieves the quantum Cramer-Rao bound, with the same structure of projecting onto a minimum-variance subspace.

Structural Compression

Invariant. Among all linear unbiased estimators of β, OLS has the smallest covariance matrix in the Loewner order.

Mechanism. Any competing linear unbiased estimator adds a matrix D satisfying $DX = 0$; the cross terms vanish by orthogonality and the covariance increment $\sigma^2 DD^\top \succeq 0$ is nonnegative definite. The constraint $DX = 0$ forces the perturbation into the orthogonal complement of the column space.

Cross-System Echo. Gauss-Markov is the second-moment analogue of the Cramer-Rao bound: both establish lower bounds on estimator variance within a restricted class. Under normality, the two coincide. The BLUE property is a finite-sample optimality result; the Cramer-Rao bound can be either finite-sample (under regularity) or asymptotic.

Exercises

1. Verify the Gauss-Markov theorem for the intercept-only model $Y_i = \mu + \varepsilon_i$ by directly showing $\bar{Y}$ has minimum variance among all $\sum c_i Y_i$ with $\sum c_i = 1$.
2. Derive the GLS estimator by transforming the model with $\Omega^{-1/2}$ and applying OLS to the transformed model.
3. Show that if $\mathrm{Cov}(\varepsilon) = \sigma^2\Omega$ with $\Omega \neq I$, OLS is still unbiased but no longer BLUE, and compute its (incorrect) variance and its true variance.
4. Prove that ridge regression has lower MSE than OLS for λ sufficiently small, by showing $d(\mathrm{MSE}(\hat{\beta}_\lambda))/d\lambda|_{\lambda=0} < 0$ when $\beta \neq 0$.
5. In a model with two predictors and $X^\top X = \begin{pmatrix} n & 0 \\ 0 & n \end{pmatrix}$, show that Gauss-Markov implies $\hat{\beta}_1 = \bar{X}_1$ and $\hat{\beta}_2 = \bar{X}_2$ are individually optimal, and explain why orthogonality of predictors simplifies the result.
6. Construct an example where a biased estimator has strictly lower MSE than the BLUE for every value of β in a bounded parameter space.
7. Argue, structurally, why Gauss-Markov is both powerful and limited: powerful because it requires only second-moment assumptions, limited because the class of linear unbiased estimators excludes the most effective modern methods (regularization, nonlinear methods).

Inference in Linear Models

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.5 — Inference in Linear Models.

This node develops the exact finite-sample distributional theory for inference in the normal linear model: t-tests, F-tests, confidence intervals, and prediction intervals. These results depend on the geometry of OLS (Node 8.3), Gauss-Markov (Node 8.4), the MVN (Node 8.2), and the general testing and estimation frameworks (Nodes 5.1, 4.1). This node enables the extension to GLMs (Node 8.6).

Formal Definition

Let $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\text{rank}(X) = p$. Under normality:

$$\hat{\beta} \sim \mathcal{N}(\beta,\ \sigma^2(X^\top X)^{-1}), \qquad \frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}, \qquad \hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2.$$

Intuition

Under Gaussian errors, the OLS estimator is exactly normal (not just approximately). The residual sum of squares follows a chi-squared distribution because it is a quadratic form in normal variables via the idempotent matrix $I - H$. The independence of $\hat{\beta}$ and $\hat{\sigma}^2$ — which follows from the orthogonality of H and $I - H$ — allows us to form exact pivotal quantities. The ratio of a normal to a chi-squared produces a t-distribution; the ratio of two chi-squareds produces an F-distribution. These are exact for any sample size, not asymptotic approximations.

Structural Role

Under normality, the linear model admits exact finite-sample inference. The t-test for single coefficients and the F-test for general linear restrictions are both instances of likelihood ratio testing (Node 5.4) specialized to the Gaussian linear model, where the chi-squared approximation happens to be exact. Cochran's theorem (Node 8.2) provides the distributional foundation, connecting the rank of projection matrices to chi-squared degrees of freedom.

Mathematical Development

Distribution of $\hat{\beta}$. Since $\hat{\beta} = (X^\top X)^{-1}X^\top Y$ is a linear function of $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$, it is exactly normal: $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^\top X)^{-1})$.

Distribution of RSS. Write $\hat{\varepsilon} = (I - H)Y = (I - H)\varepsilon$ (since $(I - H)X\beta = 0$). Then $\hat{\varepsilon}/\sigma \sim \mathcal{N}(0, I - H)$ and

$$\frac{\text{RSS}}{\sigma^2} = \frac{\varepsilon^\top (I - H)\varepsilon}{\sigma^2}.$$

Since $I - H$ is symmetric idempotent with rank $n - p$, this is a chi-squared with $n - p$ degrees of freedom: $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$.

Independence of $\hat{\beta}$ and RSS. $\hat{\beta}$ depends on Y through HY, and RSS depends on Y through $(I-H)Y$. Since $H(I-H)=0$ and Y is normal, $HY \perp\!\!\!\perp (I-H)Y$. Therefore $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$.

Cochran's theorem application. The decomposition $\varepsilon = H\varepsilon + (I-H)\varepsilon$ corresponds to projection onto $\mathcal{C}(X)$ and $\mathcal{C}(X)^\perp$. By Cochran's theorem, $\varepsilon^\top H\varepsilon/\sigma^2 \sim \chi^2_p$ and $\varepsilon^\top(I-H)\varepsilon/\sigma^2 \sim \chi^2_{n-p}$, independently. Under the alternative $\mathbb{E}[Y] = X\beta \neq 0$, the first quadratic form has a noncentral chi-squared distribution, while RSS remains $\sigma^2\chi^2_{n-p}$.

t-test for a single coefficient. To test $H_0 : \beta_j = \beta_j^0$:

$$T_j = \frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}} \sim t_{n-p} \quad \text{under } H_0.$$

Derivation. $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(X^\top X)^{-1}_{jj})$, so $(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{(X^\top X)^{-1}_{jj}}) \sim \mathcal{N}(0,1)$. Dividing by $\sqrt{\hat{\sigma}^2/\sigma^2}$, where $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ independently, gives a t_{n-p} distribution.

Confidence interval for β_j:

$$\hat{\beta}_j \pm t_{n-p,\,\alpha/2} \cdot \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}.$$

F-test for general linear hypotheses. For $H_0 : R\beta = r$ with $R \in \mathbb{R}^{q\times p}$ of rank q:

$$F = \frac{(R\hat{\beta} - r)^\top [R(X^\top X)^{-1}R^\top]^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q,\,n-p} \quad \text{under } H_0.$$

Derivation. Under H_0, $R\hat{\beta} - r \sim \mathcal{N}(0, \sigma^2 R(X^\top X)^{-1}R^\top)$. The numerator quadratic form divided by σ^2 is χ^2_q. Dividing by the independent $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/(n-p)$ gives $F_{q,n-p}$.

Nested model comparison. To compare the full model $Y = X\beta + \varepsilon$ with a reduced model $Y = X_0\beta_0 + \varepsilon$ where $\mathcal{C}(X_0) \subseteq \mathcal{C}(X)$:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(p - p_0)}{\mathrm{RSS}/(n-p)} \sim F_{p-p_0,\,n-p} \quad \text{under the reduced model.}$$

This is the ANOVA F-test.

Confidence ellipsoid for β:

$$\left\{ b \in \mathbb{R}^p : (b-\hat{\beta})^\top (X^\top X)(b-\hat{\beta}) \leq p\hat{\sigma}^2 F_{p,\,n-p,\,\alpha} \right\}.$$

This is the inversion of the F-test for $H_0 : \beta = b$.

Prediction interval. For a new observation $Y_{\text{new}} = x_0^\top\beta + \varepsilon_{\text{new}}$:

$$\hat{Y}_{\text{new}} = x_0^\top\hat{\beta}, \quad \mathrm{Var}(Y_{\text{new}} - \hat{Y}_{\text{new}}) = \sigma^2(1 + x_0^\top(X^\top X)^{-1}x_0).$$

The $(1-\alpha)$ prediction interval is

$$x_0^\top\hat{\beta} \pm t_{n-p,\,\alpha/2} \cdot \hat{\sigma}\sqrt{1 + x_0^\top(X^\top X)^{-1}x_0}.$$

The extra σ^2 accounts for the irreducible noise in the new observation.

Coefficient of determination. $R^2 = 1 - \text{RSS}/\text{TSS}$ where $\text{TSS} = \|Y - \bar{Y}\mathbf{1}\|^2$. The adjusted $R^2 = 1 - (1-R^2)(n-1)/(n-p)$ penalizes for additional predictors. Under the null model (no predictors beyond the intercept), $R^2 \cdot (n-p)/((1-R^2)(p-1)) \sim F_{p-1,n-p}$.

Simultaneous confidence intervals. The Bonferroni correction produces simultaneous $(1-\alpha)$ intervals for all p coefficients by using $t_{n-p,\alpha/(2p)}$ in place of $t_{n-p,\alpha/2}$. The Scheffe method uses the F-distribution: $\hat{\beta}_j \pm \sqrt{pF_{p,n-p,\alpha}} \cdot \hat{\sigma}\sqrt{(X^\top X)^{-1}_{jj}}$, which provides coverage for all linear combinations $a^\top\beta$ simultaneously.

Power of the F-test. Under the alternative $R\beta \neq r$, the F-statistic has a noncentral F distribution $F_{q,n-p,\delta}$ with noncentrality parameter

$$\delta = \frac{(R\beta - r)^\top [R(X^\top X)^{-1}R^\top]^{-1}(R\beta - r)}{\sigma^2}.$$

The power increases with δ, which depends on the effect size, the sample size (through $X^\top X$), and the error variance. Power analysis uses this noncentrality parameter to determine the required sample size.

Residual diagnostics. Under the model, the standardized residuals $r_i = \hat{\varepsilon}_i/(\hat{\sigma}\sqrt{1-h_{ii}})$ are approximately t_{n-p-1}. The externally studentized residuals $t_i = \hat{\varepsilon}_i/(\hat{\sigma}_{(i)}\sqrt{1-h_{ii}})$, where $\hat{\sigma}_{(i)}$ is estimated from the data excluding observation i, follow t_{n-p-1} exactly. Cook's distance $D_i = r_i^2 h_{ii}/(p(1-h_{ii}))$ combines leverage and residual magnitude to measure influence.

Analysis of variance (ANOVA) decomposition. The total sum of squares decomposes as

$$\text{TSS} = \text{ESS} + \text{RSS}, \quad \text{where ESS} = \|\hat{Y} - \bar{Y}\mathbf{1}\|^2,\ \text{RSS} = \|Y - \hat{Y}\|^2.$$

Under $H_0 : \beta_2 = \cdots = \beta_p = 0$ (only intercept), $\text{ESS}/\sigma^2 \sim \chi^2_{p-1}$ and is independent of $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$, yielding the overall F-test $F = (\text{ESS}/(p-1))/(\text{RSS}/(n-p)) \sim F_{p-1,n-p}$.

Connection to likelihood ratio. Under normality, the LR statistic for $H_0 : R\beta = r$ is

$$\Lambda = n\log(\text{RSS}_0/\text{RSS}),$$

where RSS_0 is the residual sum of squares under the restricted model. The F-statistic is a monotone transformation: $F = ((e^{\Lambda/n} - 1)(n-p))/q$. Thus the F-test is equivalent to the likelihood ratio test, with the F-distribution providing the exact (not asymptotic) null distribution.

Multiple testing. When testing many coefficients simultaneously, the individual t-test p-values require adjustment. The Bonferroni correction controls the family-wise error rate by using α/m as the per-test level. The Benjamini-Hochberg procedure controls the false discovery rate, which is less conservative and more appropriate for exploratory analysis with many predictors.

Worked Examples

Example 1. In simple linear regression with $n = 30$ observations, $\hat{\beta}_1 = 2.5$, $\text{SE}(\hat{\beta}_1) = 0.8$. The t-statistic is $T = 2.5/0.8 = 3.125$ on 28 df. The p-value is $2\mathbb{P}(t_{28} > 3.125) \approx 0.004$. The 95% CI is $2.5 \pm 2.048 \times 0.8 = (0.86, 4.14)$.

Example 2. Test $H_0 : \beta_2 = \beta_3 = 0$ in a model with $p = 5$ predictors and $n = 50$. Here $q = 2$, $R = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$, $r = 0$. If $\text{RSS}_{\text{full}} = 120$ and $\text{RSS}_{\text{reduced}} = 160$, then $F = ((160-120)/2)/(120/45) = 20/2.667 = 7.5$ on $(2, 45)$ df. The critical value $F_{2,45,0.05} \approx 3.20$, so we reject H_0.

Cross-Links

AI. The F-statistic for nested model comparison is the basis for feature selection in linear models. In deep learning, the degrees-of-freedom concept generalizes via effective degrees of freedom $\text{tr}(H_\lambda)$ for regularized estimators, connecting to generalization bounds.

Economics. The t-test and F-test are the primary tools in empirical economics. The Chow test for structural breaks is an F-test comparing pooled and split regressions. Robust inference under heteroskedasticity replaces the exact F with asymptotic chi-squared/F using the sandwich covariance (Node 7.5).

Linear Algebra. The distributional results rest on the spectral properties of idempotent matrices: a symmetric idempotent matrix of rank r has r eigenvalues equal to 1 and the rest zero. Quadratic forms in standard normals with such matrices give chi-squared distributions.

Quantum Mechanics. Hypothesis testing in quantum state discrimination uses analogous chi-squared statistics for goodness-of-fit testing of quantum states, with degrees of freedom determined by the dimension of the state space minus the number of estimated parameters.

Structural Compression

Invariant. Under Gaussian errors, all finite-sample inference in the linear model is determined by $\hat{\beta}$, $\hat{\sigma}^2$, and the hat matrix H; the t and F distributions arise from the chi-squared distribution of RSS/σ^2 and the independence of $\hat{\beta}$ and $\hat{\sigma}^2$.

Mechanism. Orthogonal decomposition $Y = HY + (I - H)Y$ separates signal from noise; independence follows from orthogonality under normality (Cochran's theorem); quadratic forms in standard normals yield chi-squared distributions; ratios of independent chi-squareds yield F; a standard normal divided by the square root of an independent chi-squared yields t.

Cross-System Echo. The F-statistic is the likelihood ratio statistic for the Gaussian linear model; the exact F-distribution replaces the asymptotic χ^2 approximation that holds in general parametric models. In ANOVA, the same F-test compares group means, decomposing total variation into between-group and within-group components.

Exercises

1. Derive the distribution of the t-statistic T_j from the joint distribution of $\hat{\beta}_j$ and $\hat{\sigma}^2$, verifying the t_{n-p} result.
2. Show that the F-test for $H_0 : \beta_j = 0$ is equivalent to T_j^2 by verifying $F_{1,n-p} = t_{n-p}^2$.
3. Prove that $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$ using the fact that HY and $(I - H)Y$ are functions of Y through orthogonal projections and Y is normal.
4. For the prediction interval, explain why the variance $\sigma^2(1 + x_0^\top (X^\top X)^{-1} x_0)$ has two components and identify which component is reducible with more data.
5. In a one-way ANOVA with $k = 3$ groups of size $n_j = 10$, derive the F-statistic as a ratio of between-group and within-group mean squares, and state the null distribution.
6. Show that $R^2 = 0$ if and only if $\hat{\beta}_j = 0$ for all $j \geq 2$ (assuming an intercept is included), and explain why R^2 always increases when adding predictors.
7. Argue, structurally, why the exact distributional results of this node (t-tests, F-tests) require Gaussian errors while the Gauss-Markov theorem (Node 8.4) does not, identifying exactly which properties of the normal distribution are needed and which are not.

Generalized Linear Models

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.6 — Generalized Linear Models.

GLMs extend the normal linear model to responses from any exponential family distribution, replacing the identity link with a general link function. This node depends on the linear model geometry (Node 8.3), exponential family theory (Node 2.6), and M-estimation (Node 7.5). It is the natural generalization of the Gaussian linear model to non-Gaussian responses.

Formal Definition

A generalized linear model specifies three components:

1. **Random component.** $Y_i \mid x_i$ independently follows an exponential family distribution with density

$$f(y_i; \eta_i, \phi) = h(y_i, \phi) \exp\left(\frac{y_i \eta_i - b(\eta_i)}{a(\phi)} \right),$$

where η_i is the canonical parameter, $b(\cdot)$ is the cumulant function, ϕ is the dispersion parameter, and $a(\phi) = \phi / w_i$ for known weights w_i.

2. **Systematic component.** A linear predictor $\eta_i^* = x_i^\top \beta$.

3. **Link function.** A monotone differentiable function g such that $g(\mu_i) = x_i^\top \beta$, where $\mu_i = \mathbb{E}[Y_i \mid x_i] = b'(\eta_i)$.

The **canonical link** sets $g = (b')^{-1}$, so that $\eta_i = x_i^\top \beta$ directly.

Intuition

The normal linear model ties the mean directly to the linear predictor: $\mu_i = x_i^\top \beta$. For binary or count data, this is inappropriate — the mean must be constrained (to $[0, 1]$ or $[0, \infty)$). The link function bridges the gap: it maps the constrained mean space to the unconstrained linear predictor space. The canonical link is special because it makes the sufficient statistics linear in the linear predictor, simplifying the score equations and guaranteeing concavity of the log-likelihood.

Structural Role

GLMs unify regression models for continuous, binary, and count responses under a single framework. The exponential family structure (Node 2.6) determines the variance function $V(\mu) = b''(\eta)$, the canonical link, and the form of the log-likelihood. The IRLS algorithm reduces all GLM fitting to a sequence of weighted least squares problems, specializing to OLS in the Gaussian case. Asymptotic inference follows from M-estimation theory (Node 7.5).

Mathematical Development

Mean-variance relationship. From exponential family theory, $\mu = b'(\eta)$ and $\mathrm{Var}(Y) = a(\phi)b''(\eta) = a(\phi)V(\mu)$, where $V(\mu) = b''((b')^{-1}(\mu))$ is the variance function. The variance function determines the distribution:

Distribution	$b(\eta)$	$V(\mu)$	Canonical link
Normal	$\eta^2/2$	1	Identity: $g(\mu) = \mu$
Bernoulli	$\log(1 + e^\eta)$	$\mu(1-\mu)$	Logit: $g(\mu) = \log \frac{\mu}{1-\mu}$
Poisson	e^η	μ	Log: $g(\mu) = \log \mu$
Gamma	$-\log(-\eta)$	μ^2	Reciprocal: $g(\mu) = 1/\mu$

Log-likelihood and score equations. For the canonical link ($\eta_i = x_i^\top \beta$):

$$\ell(\beta) = \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i x_i^\top \beta - b(x_i^\top \beta) \right] + \text{const.}$$

The score is

$$\nabla_\beta \ell(\beta) = \frac{1}{\phi} \sum_{i=1}^{n} w_i (y_i - \mu_i) x_i = \frac{1}{\phi} X^\top W^{1/2} (y - \mu),$$

where $\mu_i = b'(x_i^\top \beta)$. Setting the score to zero gives the score equations $X^\top (y - \mu(\beta)) = 0$ (up to weights and ϕ), which generalize the normal equations of OLS.

For a general link, the score is

$$\nabla_\beta \ell = \sum_{i=1}^{n} \frac{w_i (y_i - \mu_i)}{a(\phi) V(\mu_i) g'(\mu_i)} x_i.$$

Fisher information. The expected Fisher information is

$$\mathcal{I}(\beta) = \frac{1}{\phi} X^\top W X,$$

where $W = \mathrm{diag}(w_1/(V(\mu_1)[g'(\mu_1)]^2), \ldots, w_n/(V(\mu_n)[g'(\mu_n)]^2))$. For the canonical link, $W = \mathrm{diag}(w_i V(\mu_i))$.

Fisher scoring / IRLS. The MLE is computed iteratively. Fisher scoring updates:

$$\beta^{(t+1)} = \beta^{(t)} + [\mathcal{I}(\beta^{(t)})]^{-1} \nabla_\beta \ell(\beta^{(t)}).$$

This is equivalent to Iteratively Reweighted Least Squares: at each step, solve

$$\hat{\beta}^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)},$$

where the working response is $z_i^{(t)} = \hat{\eta}_i^{(t)} + (y_i - \hat{\mu}_i^{(t)}) g'(\hat{\mu}_i^{(t)})$ and $W^{(t)}$ is the weight matrix evaluated at $\hat{\mu}^{(t)}$.

Convergence. For the canonical link, the log-likelihood is concave (since b is convex), so IRLS converges to the unique global maximum. For non-canonical links, local convergence is guaranteed under smoothness conditions.

Logistic regression (Bernoulli GLM). For binary $Y_i \in \{0,1\}$ with $\mathbb{P}(Y_i = 1|x_i) = \mu_i$, the canonical link is the logit: $\log(\mu_i/(1-\mu_i)) = x_i^\top \beta$, so $\mu_i = 1/(1+e^{-x_i^\top \beta})$. The log-likelihood is $\ell(\beta) = \sum_i [y_i x_i^\top \beta - \log(1+e^{x_i^\top \beta})]$. The variance function is $V(\mu) = \mu(1-\mu)$, and the weights are $w_i = \mu_i(1-\mu_i)$.

Poisson regression. For count $Y_i \in \{0, 1, 2, \ldots\}$ with $\mu_i = \mathbb{E}[Y_i|x_i]$, the canonical link is the log: $\log \mu_i = x_i^\top \beta$. The log-likelihood is $\ell(\beta) = \sum_i [y_i x_i^\top \beta - e^{x_i^\top \beta}] + \text{const}$. The variance equals the mean: $V(\mu) = \mu$, and the weights are $w_i = \mu_i$.

Quasi-likelihood. When only the mean-variance relationship $\mathrm{Var}(Y_i) = \phi V(\mu_i)$ is specified (without a full distributional assumption), the quasi-likelihood score equations are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) x_i}{V(\mu_i) g'(\mu_i)} = 0.$$

These coincide with the GLM score equations, so the same IRLS algorithm applies. The quasi-likelihood estimator is consistent and asymptotically normal with the sandwich covariance, even without specifying the full distribution — only the first two conditional moments matter.

Negative binomial regression. For overdispersed count data where $\mathrm{Var}(Y) > \mu$, the negative binomial GLM uses $V(\mu) = \mu + \mu^2/\kappa$ where κ is the overdispersion parameter. This nests Poisson regression ($\kappa \to \infty$) and accommodates extra-Poisson variation.

Probit model. An alternative to logistic regression for binary data uses the link $g(\mu) = \Phi^{-1}(\mu)$ where Φ is the standard normal CDF. The probit model arises naturally from a latent variable formulation: $Y_i = \mathbf{1}(x_i^\top \beta + \varepsilon_i > 0)$ with $\varepsilon_i \sim \mathcal{N}(0,1)$. Logit and probit give nearly identical fitted probabilities in practice (after rescaling by $\pi/\sqrt{3} \approx 1.81$).

Deviance. The deviance is twice the log-likelihood ratio between the saturated model (one parameter per observation) and the fitted model:

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i(\tilde{\eta}_i - \hat{\eta}_i) - b(\tilde{\eta}_i) + b(\hat{\eta}_i) \right],$$

where $\tilde{\eta}_i$ is the canonical parameter of the saturated model ($\tilde{\mu}_i = y_i$). The deviance plays the role of RSS in the normal linear model. Under regularity conditions, the difference in deviances between nested models follows χ^2_q asymptotically, enabling likelihood ratio tests for model comparison.

Asymptotic inference. By the M-estimation theory of Node 7.5, the MLE satisfies

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0,\ \mathcal{I}(\beta)^{-1}).$$

Wald tests use $(\hat{\beta}_j / \mathrm{SE}(\hat{\beta}_j))^2$; Score tests use $\nabla \ell(\beta_0)^\top \mathcal{I}(\beta_0)^{-1} \nabla \ell(\beta_0)$; LR tests use the deviance difference. All three converge to χ^2 under the null.

Worked Examples

Example 1 (Logistic regression). Data: 100 patients, binary outcome (disease/no disease), predictor $x =$ age. Fitted model: $\mathrm{logit}(\hat{\mu}) = -5.0 + 0.08 \cdot \text{age}$. At age 50: $\hat{\mu} = 1/(1 + e^{-(-5+4)}) = 1/(1+e) \approx 0.269$. The odds ratio per year of age is $e^{0.08} \approx 1.083$: each year increases the odds of disease by 8.3%. The Wald 95% CI for the log-odds-ratio is $0.08 \pm 1.96 \times \mathrm{SE}(\hat{\beta}_1)$.

Example 2 (Poisson regression). Data: counts of insurance claims by age group. Model: $\log \hat{\mu}_i = 1.2 + 0.03 \cdot \text{age}_i$. The rate ratio per year is $e^{0.03} \approx 1.030$. Deviance goodness-of-fit: if the residual deviance

is 45.2 on 48 df, the model fits adequately (p-value from χ^2_{48} is ≈ 0.59). If overdispersion is detected ($D/(n-p) \gg 1$), a quasi-Poisson model with estimated $\phi > 1$ inflates standard errors.

Cross-Links

AI. Logistic regression is the canonical binary classifier and the basis for the output layer of neural networks for classification. The cross-entropy loss in deep learning is the negative Bernoulli GLM log-likelihood. Softmax regression extends logistic regression to multinomial responses.

Economics. Probit and logit models are the primary binary choice models in microeconomics. Poisson regression is used for count data in labor economics (number of patents, doctor visits). The quasi-likelihood framework extends GLMs when the full distributional assumption is relaxed.

Linear Algebra. IRLS reduces each GLM iteration to a weighted least squares problem: $(X^\top W X)^{-1} X^\top W z$. The convergence of IRLS is analyzed via the contraction mapping theorem on the Fisher scoring iteration.

Quantum Mechanics. In quantum detector tomography, the detection probabilities follow a multinomial model that can be expressed as a GLM with a log-link on the POVM elements. The Fisher scoring algorithm applies to reconstruct the detector from count data.

Structural Compression

Invariant. The score equations $X^\top(y - \mu(\beta)) = 0$ generalize the normal equations of OLS to any exponential family response; the canonical link aligns the natural parameter with the linear predictor.

Mechanism. The link function maps the constrained mean space to the unconstrained linear predictor; the exponential family structure provides the variance function, score, and Fisher information. IRLS iteratively constructs a locally quadratic approximation via working responses and weights, converging to the MLE.

Cross-System Echo. Logistic regression is maximum entropy classification (AI connection). The deviance is the likelihood ratio statistic comparing fitted and saturated models (Node 5.4). The quasi-likelihood approach relaxes the full distributional assumption to the first two conditional moments, paralleling the Gauss-Markov approach in the linear model.

Exercises

1. Derive the canonical link for the Bernoulli distribution by writing the PMF in exponential family form and identifying $\eta = \log(\mu/(1-\mu))$.
2. Show that the score equations for logistic regression reduce to $X^\top(y - \mu(\beta)) = 0$ and verify that the log-likelihood is concave.
3. For Poisson regression with canonical link, derive the Fisher information matrix and the IRLS weight matrix.
4. Prove that the deviance for the normal linear model equals RSS/σ^2, recovering the residual sum of squares as a special case.
5. Fit a logistic regression to the data $(x, y) = \{(1,0), (2,0), (3,1), (4,0), (5,1), (6,1), (7,1)\}$ by writing the log-likelihood and performing one Newton-Raphson iteration starting from $\beta = (0, 0)$.
6. Explain the concept of overdispersion in the Poisson GLM, derive the quasi-Poisson variance $\text{Var}(Y_i) = \phi\mu_i$, and describe how standard errors are adjusted.

7. Argue, structurally, why the canonical link occupies a privileged position among all possible link functions, connecting its role to sufficiency, concavity of the log-likelihood, and the information matrix equality.

Principal Component Analysis

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.7 — Principal Component Analysis.

PCA is the canonical method for dimensionality reduction via variance maximization. It connects the spectral theorem from linear algebra to the statistical problem of finding the most informative low-dimensional representation of high-dimensional data. This node depends on the covariance structure (Node 8.1) and the multivariate normal (Node 8.2).

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (\mu, \Sigma)$ with $X_i \in \mathbb{R}^p$. The k-th **principal component direction** $v_k \in \mathbb{R}^p$ solves

$$v_k = \arg \max_{\|v\|=1,\ v \perp v_1, \ldots, v_{k-1}} v^\top \Sigma v.$$

The k-th **principal component score** is $Z_k = v_k^\top (X - \mu)$.

The solution is the eigendecomposition $\Sigma = \sum_{j=1}^{p} \lambda_j v_j v_j^\top$ with $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, where v_k is the k-th eigenvector.

Intuition

PCA rotates the coordinate system to align with the directions of greatest variability. The first principal component captures the most variance; each subsequent component captures the maximum remaining variance orthogonal to all previous ones. The eigenvalues are the variances along these directions. If the data lies approximately in a low-dimensional subspace, the first few components explain most of the total variance and the rest can be discarded with minimal information loss. PCA is a descriptive procedure — it requires no probabilistic model — but it is the optimal linear dimensionality reduction under the mean-squared reconstruction error criterion.

Structural Role

PCA identifies the directions of greatest variance, providing a basis for low-rank representation, visualization, and preprocessing. It is the bridge between the algebraic spectral theorem and statistical data analysis. PCA connects to probabilistic PCA (a latent variable model with Gaussian noise), factor analysis, and the SVD of the data matrix. In high-dimensional statistics, PCA is both a tool and an object of study (spiked covariance models, random matrix theory).

Mathematical Development

PCA as variance maximization.

Theorem. The maximum of $v^\top \Sigma v$ subject to $\|v\| = 1$ is λ_1, attained at $v = v_1$.

Proof. By Lagrange multipliers, the stationary condition is $\Sigma v = \lambda v$, so the critical points are eigenvectors with critical values equal to eigenvalues. The maximum is the largest eigenvalue λ_1.

For the k-th component, the constraint $v \perp v_1, \ldots, v_{k-1}$ restricts to the subspace orthogonal to the first $k-1$ eigenvectors. On this subspace, Σ acts with eigenvalues $\lambda_k, \ldots, \lambda_p$, so the maximum is λ_k at v_k. $\square$

PCA as minimum reconstruction error.

Theorem. The rank-r linear subspace minimizing the expected reconstruction error is $\text{span}(v_1, \ldots, v_r)$:

$$\min_{\dim(S)=r} \mathbb{E}\left[\|X - \mu - \text{proj}_S(X-\mu)\|^2\right] = \sum_{j=r+1}^{p} \lambda_j.$$

Proof. Let $V_r = [v_1, \ldots, v_r]$. The projection is $\text{proj}_S(X-\mu) = V_r V_r^\top (X-\mu)$ and the reconstruction error is

$$\begin{aligned}\mathbb{E}[\|X - \mu - V_r V_r^\top (X-\mu)\|^2] &= \mathbb{E}[\text{tr}((I - V_r V_r^\top)(X-\mu)(X-\mu)^\top (I - V_r V_r^\top))] \\ &= \text{tr}((I - V_r V_r^\top)\Sigma) = \sum_{j=1}^{p} \lambda_j - \sum_{j=1}^{r} \lambda_j = \sum_{j=r+1}^{p} \lambda_j.\end{aligned}$$

For any other rank-r subspace, the Eckart-Young theorem (or the Fan-Poincare minimax principle) shows the error is at least $\sum_{j=r+1}^{p} \lambda_j$. $\square$

Variance explained. The proportion of variance explained by the first r components is

$$\frac{\sum_{j=1}^{r} \lambda_j}{\text{tr}(\Sigma)} = \frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

The scree plot displays λ_j vs. j; a sharp drop ("elbow") suggests the appropriate number of components. The cumulative proportion curve $\sum_{j=1}^{r} \lambda_j / \text{tr}(\Sigma)$ provides a threshold-based selection rule (e.g., retain components until 95% of variance is explained).

Connection to SVD. Let $\tilde{X} \in \mathbb{R}^{n \times p}$ be the centered data matrix with rows $(X_i - \bar{X})^\top$. The singular value decomposition $\tilde{X} = UDV^\top$ (with $U \in \mathbb{R}^{n \times p}$, $D = \text{diag}(d_1, \ldots, d_p)$, $V \in \mathbb{R}^{p \times p}$ orthogonal) gives:

- V: columns are the principal component directions (right singular vectors = eigenvectors of $\hat{\Sigma}$).
- $d_j^2/(n-1) = \hat{\lambda}_j$: the sample eigenvalues.
- UD: the matrix of principal component scores.
- $\hat{\Sigma} = V(D^2/(n-1))V^\top$.

The SVD is computationally preferable to forming $\hat{\Sigma}$ explicitly, avoiding the squaring of condition numbers.

Sample PCA. Replace Σ by $\hat{\Sigma} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top$. The sample principal directions $\hat{v}_k$ are eigenvectors of $\hat{\Sigma}$. When population eigenvalues are distinct ($\lambda_j \neq \lambda_k$ for $j \neq k$), $\hat{v}_k \xrightarrow{p} v_k$ (up to sign) by the continuous mapping theorem applied to the eigendecomposition map.

When eigenvalues coincide ($\lambda_j = \lambda_{j+1}$), the eigenvectors are not uniquely identified — any orthonormal basis of the eigenspace is valid — and the sample eigenvectors converge only to the eigenspace, not to a specific direction.

Asymptotic distribution of eigenvalues. Under normality or finite fourth moments, for distinct eigenvalues,

$$\sqrt{n}(\hat{\lambda}_j - \lambda_j) \xrightarrow{d} \mathcal{N}(0,\ 2\lambda_j^2) \quad \text{(under MVN)}.$$

The asymptotic variance of $\hat{v}_j$ depends on the eigenvalue gaps: $\text{Var}(\hat{v}_j^\top v_k) \approx \lambda_j \lambda_k / (n(\lambda_j - \lambda_k)^2)$ for $k \neq j$. Small gaps produce unreliable eigenvector estimates.

Probabilistic PCA (Tipping-Bishop). Assume the generative model $X = Wz + \mu + \varepsilon$ where $z \sim \mathcal{N}(0, I_r)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, and $W \in \mathbb{R}^{p \times r}$. The marginal is $X \sim \mathcal{N}(\mu, WW^\top + \sigma^2 I)$. The MLE for W satisfies $\hat{W} = V_r(\Lambda_r - \sigma^2 I)^{1/2} R$ where V_r contains the first r eigenvectors, Λ_r the corresponding eigenvalues, and R is an arbitrary orthogonal matrix. As $\sigma^2 \to 0$, this recovers classical PCA.

PCA on the correlation matrix. When variables have different units, PCA is applied to the correlation matrix R rather than Σ. This is equivalent to standardizing each variable to unit variance before computing PCA. The choice between Σ and R is a modeling decision: Σ-PCA preserves the original scale; R-PCA treats all variables equally.

High-dimensional PCA and spiked models. When $p/n \to \gamma \in (0, \infty)$, the sample eigenvalues of $\hat{\Sigma}$ are inconsistent for the population eigenvalues due to the Marchenko-Pastur law. In the spiked covariance model $\Sigma = I_p + \sum_{j=1}^{r} \ell_j v_j v_j^\top$, a phase transition occurs: the j-th sample eigenvalue separates from the bulk only when $\ell_j > \sqrt{\gamma}$ (the BBP transition). Below this threshold, the signal is undetectable by PCA.

Kernel PCA. Classical PCA finds directions in the input space $\mathbb{R}^p$. Kernel PCA maps data to a feature space $\phi(x) \in \mathcal{H}$ via a kernel $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ and performs PCA in $\mathcal{H}$. The principal components are computed from the $n \times n$ kernel matrix $K_{ij} = K(X_i, X_j)$ without explicitly constructing ϕ. This enables nonlinear dimensionality reduction.

Relationship to factor analysis. Factor analysis posits $X = \Lambda F + \varepsilon$ with $F \sim \mathcal{N}(0, I_r)$, $\varepsilon \sim \mathcal{N}(0, \Psi)$ where Ψ is diagonal, giving $\Sigma = \Lambda\Lambda^\top + \Psi$. Unlike PCA, factor analysis separates common variance ($\Lambda\Lambda^\top$) from unique variance (Ψ). PCA is equivalent to factor analysis only when $\Psi = \sigma^2 I$ (the probabilistic PCA model).

Courant-Fischer minimax characterization. The eigenvalues of Σ admit the variational characterization:

$$\lambda_k = \min_{\dim(S) = p-k+1} \max_{v \in S, \|v\|=1} v^\top \Sigma v = \max_{\dim(S) = k} \min_{v \in S, \|v\|=1} v^\top \Sigma v.$$

This provides the theoretical basis for PCA: the top k eigenvalues represent the maximum achievable total variance captured by any k-dimensional subspace.

Sparse PCA. In high dimensions, classical PCA eigenvectors are dense and difficult to interpret. Sparse PCA adds an ℓ_1 penalty:

$$\max_v v^\top \Sigma v - \lambda \|v\|_1 \quad \text{subject to } \|v\|_2 = 1,$$

producing sparse loading vectors. This loses optimality in reconstruction error but gains interpretability and consistency in high-dimensional settings where $p \gg n$.

Johnson-Lindenstrauss vs. PCA. Random projections preserve pairwise distances up to distortion ε in $O(\log n / \varepsilon^2)$ dimensions, regardless of the data's intrinsic structure. PCA optimally preserves variance but requires $O(np^2)$ computation. For very high-dimensional problems, random projections provide a computational alternative with different statistical guarantees.

Worked Examples

Example 1. Let $\Sigma = \begin{pmatrix} 5 & 3 \\ 3 & 2 \end{pmatrix}$. The eigenvalues satisfy $\lambda^2 - 7\lambda + 1 = 0$, giving $\lambda_1 = (7 + 3\sqrt{5})/2 \approx 6.854$, $\lambda_2 = (7 - 3\sqrt{5})/2 \approx 0.146$. The first PC explains $6.854/7 \approx 97.9\%$ of the variance. The eigenvector v_1 satisfies $(5 - \lambda_1)v_{11} + 3v_{12} = 0$, giving $v_1 \propto (3, \lambda_1 - 5)^\top \approx (3, 1.854)^\top$, normalized to unit length. The data is nearly one-dimensional: the principal direction captures almost all variability.

Example 2 (SVD computation). Given centered data matrix $\tilde{X} = \begin{pmatrix} 2 & 1 \\ -1 & 1 \\ -1 & -2 \end{pmatrix}$, compute $\tilde{X}^\top \tilde{X} = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix}$. Eigenvalues: 9 and 3, with eigenvectors $(1,1)^\top/\sqrt{2}$ and $(1,-1)^\top/\sqrt{2}$. Sample variances: $\hat{\lambda}_1 = 9/2 = 4.5$, $\hat{\lambda}_2 = 3/2 = 1.5$. First PC direction: $\hat{v}_1 = (1,1)^\top/\sqrt{2}$, the 45-degree line. Scores: $Z_1 = \tilde{X}\hat{v}_1 = (3/\sqrt{2}, 0, -3/\sqrt{2})^\top$. Proportion of variance: $4.5/6 = 75\%$.

Cross-Links

AI. PCA is the basis for whitening, feature extraction, and linear autoencoders. A linear autoencoder with r hidden units recovers the PCA projection. In NLP, PCA applied to word embedding matrices identifies semantic dimensions. Kernel PCA extends to nonlinear dimensionality reduction via the kernel trick.

Economics. PCA on macroeconomic time series extracts latent factors (diffusion indices) used in forecasting. In finance, PCA on bond yield curves identifies level, slope, and curvature factors that explain >95% of yield variation.

Linear Algebra. PCA is the spectral theorem applied to the covariance matrix. The Eckart-Young theorem establishes the optimality of the truncated SVD for low-rank approximation in Frobenius norm. The Davis-Kahan theorem quantifies how perturbations in Σ affect the eigenvectors.

Quantum Mechanics. PCA on quantum state tomography data decomposes the noise covariance of reconstructed states. The eigenvalues of the density matrix ρ are the quantum analogue of PCA eigenvalues: they determine the effective dimensionality (rank) of the quantum state.

Structural Compression

Invariant. The principal components are the eigenvectors of the covariance matrix; they diagonalize Σ and provide the variance-maximizing (equivalently, reconstruction-error-minimizing) orthonormal basis.

Mechanism. The spectral theorem for symmetric positive semidefinite matrices decomposes Σ into rank-one projections $\lambda_j v_j v_j^\top$; each successive eigenvector maximizes residual variance subject to orthogonality. The SVD of the data matrix provides a computationally stable path to the same decomposition.

Cross-System Echo. PCA is the method of moments estimator in probabilistic PCA (Tipping-Bishop). In information theory, PCA maximizes the mutual information between the data and its linear projection under Gaussianity. The connection between variance maximization and reconstruction error minimization is an instance of the duality between primal and dual optimization.

Exercises

1. Prove that the k-th principal component direction maximizes $v^\top \Sigma v$ subject to $\|v\| = 1$ and $v \perp v_1, \ldots, v_{k-1}$ using Lagrange multipliers.
2. Show that PCA on Σ and PCA on $\hat{\Sigma}$ are equivalent to performing the SVD of the centered data matrix, by relating the eigendecomposition of $\hat{\Sigma}$ to the right singular vectors of $\tilde{X}$.
3. For a 3×3 covariance matrix with eigenvalues $\lambda_1 = 10$, $\lambda_2 = 2$, $\lambda_3 = 0.5$, compute the proportion of variance explained by the first two components and the reconstruction error of the rank-2 approximation.
4. Prove that the total variance $\mathrm{tr}(\Sigma)$ is invariant under orthogonal transformations $Q^\top \Sigma Q$, explaining why PCA redistributes but does not create or destroy variance.

5. In the probabilistic PCA model $X = Wz + \mu + \varepsilon$, derive the marginal distribution of X and show that the MLE for μ is $\bar{X}$.

6. Explain why PCA on the correlation matrix vs. the covariance matrix can give qualitatively different results, and give a concrete example where one is appropriate and the other is not.

7. Argue, structurally, why PCA is simultaneously the optimal linear dimensionality reduction under two different criteria (maximum variance and minimum reconstruction error), and explain why this duality breaks down for nonlinear methods.

Module 9 — Computational Statistics

Numerical Optimization

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.1

Maximum likelihood estimation, M-estimation, and posterior mode computation all require maximizing an objective function $Q(\theta)$ over $\theta \in \Theta \subseteq \mathbb{R}^k$. Closed-form solutions exist only in special cases (exponential families with canonical sufficient statistics). This node develops the principal iterative algorithms for numerical optimization in statistics, their convergence properties, and the structural role of curvature information.

Formal Definition

Let $Q : \Theta \to \mathbb{R}$ be a twice continuously differentiable objective function. A point $\theta^* \in \Theta$ is a **stationary point** if $\nabla Q(\theta^*) = 0$. It is a **local maximizer** if, additionally, the Hessian $\nabla^2 Q(\theta^*)$ is negative definite.

An **iterative optimization algorithm** generates a sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ such that $Q(\theta^{(t+1)}) \geq Q(\theta^{(t)})$ (ascent property) and $\theta^{(t)} \to \theta^*$ under appropriate conditions.

The **gradient ascent** update is

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \nabla Q(\theta^{(t)}),$$

where $\alpha_t > 0$ is the step size. The **Newton–Raphson** update is

$$\theta^{(t+1)} = \theta^{(t)} - \left[\nabla^2 Q(\theta^{(t)})\right]^{-1} \nabla Q(\theta^{(t)}).$$

Fisher scoring replaces the Hessian with its expectation under the model, yielding

$$\theta^{(t+1)} = \theta^{(t)} + \mathcal{I}(\theta^{(t)})^{-1} \nabla \ell(\theta^{(t)}),$$

where $\mathcal{I}(\theta) = \mathbb{E}_\theta[-\nabla^2 \ell(\theta)]$ is the expected Fisher information.

Intuition

Gradient ascent moves uphill along the steepest direction but treats all directions equally, ignoring the curvature of the objective surface. Newton–Raphson fits a local quadratic approximation and jumps directly to its maximum, which is why it converges much faster near the optimum. The price is computing and inverting the Hessian matrix at each step, which is $O(k^3)$ in parameter dimension.

Fisher scoring is the statistically natural variant: it replaces the observed curvature (which fluctuates with the data) by the expected curvature (which depends only on the model), producing updates that are equivariant under reparameterization. For generalized linear models, Fisher scoring reduces to iteratively reweighted least squares (IRLS).

Structural Role

Numerical optimization is the computational substrate of all likelihood-based and M-estimation inference. Newton–Raphson is the canonical algorithm for computing the MLE; Fisher scoring is its information-geometric counterpart. The convergence rates of these algorithms are controlled by the curvature of the log-likelihood, which is itself the Fisher information (Node 2.5). This node is prerequisite to the EM algorithm (Node 9.5), which replaces a single intractable optimization with a sequence of tractable ones.

Mathematical Development

Convergence of Gradient Ascent

Suppose Q is L-smooth (i.e., $\|\nabla^2 Q(\theta)\| \le L$ for all θ) and m-strongly concave (i.e., $\nabla^2 Q(\theta) \preceq -mI$ for all θ, with $m > 0$). With fixed step size $\alpha = 1/L$,

$$\|\theta^{(t+1)} - \theta^*\| \le \left(1 - \frac{m}{L}\right) \|\theta^{(t)} - \theta^*\|.$$

This is **linear (geometric) convergence** with rate $\kappa = 1 - m/L$, where L/m is the **condition number** of the problem. Poor conditioning ($L \gg m$) produces slow convergence.

Line Search

The **Armijo backtracking** line search selects α_t as the largest value in $\{\beta^0, \beta^1, \beta^2, \dots\}$ (for shrinkage factor $0 < \beta < 1$) satisfying

$$Q(\theta^{(t)} + \alpha_t d^{(t)}) \ge Q(\theta^{(t)}) + c\,\alpha_t \nabla Q(\theta^{(t)})^\top d^{(t)},$$

where $d^{(t)}$ is the search direction and $0 < c < 1$. This guarantees sufficient increase at each step.

Newton–Raphson Convergence

Theorem (Local Quadratic Convergence). Let Q be three times continuously differentiable in a neighborhood of θ^* with $\nabla^2 Q(\theta^*)$ nonsingular. Then there exists $\delta > 0$ such that for $\|\theta^{(0)} - \theta^*\| < \delta$,

$$\|\theta^{(t+1)} - \theta^*\| \le C \|\theta^{(t)} - \theta^*\|^2,$$

where $C = \frac{1}{2}\|\nabla^2 Q(\theta^*)^{-1}\| \sup_\theta \|\nabla^3 Q(\theta)\|$.

Proof sketch. Taylor-expand $\nabla Q(\theta^{(t)})$ around θ^*:

$$\nabla Q(\theta^{(t)}) = \nabla^2 Q(\theta^*)(\theta^{(t)} - \theta^*) + O(\|\theta^{(t)} - \theta^*\|^2).$$

The Newton step gives $\theta^{(t+1)} - \theta^* = \theta^{(t)} - \theta^* - [\nabla^2 Q(\theta^{(t)})]^{-1} \nabla Q(\theta^{(t)})$. Substituting the Taylor expansion and using continuity of $\nabla^2 Q$, the linear term cancels, leaving only the quadratic remainder. □

Trust Regions

When Newton–Raphson is initialized far from θ^*, the Hessian may not be negative definite and the Newton step may overshoot. A **trust region** method restricts the step to a ball:

$$\theta^{(t+1)} = \arg\max_{\|\delta\| \le \Delta_t} \left[\nabla Q(\theta^{(t)})^\top \delta + \frac{1}{2} \delta^\top \nabla^2 Q(\theta^{(t)}) \delta \right],$$

where the trust radius Δ_t is adapted based on how well the quadratic model predicts the actual change in Q.

Profile Likelihood

For a partitioned parameter $\theta = (\psi, \lambda)$ where λ is a nuisance parameter, the **profile log-likelihood** is

$$\ell_p(\psi) = \max_{\lambda} \ell(\psi, \lambda) = \ell(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the MLE of λ at fixed ψ. Optimization of ℓ_p over ψ yields the marginal MLE, reducing dimension at the cost of iterative inner optimization. The curvature of ℓ_p provides adjusted standard errors for ψ.

Quasi-Newton Methods

When the Hessian is expensive to compute, **quasi-Newton methods** build an approximation $B_t \approx -\nabla^2 Q(\theta^{(t)})$ from successive gradient evaluations. The **BFGS update** maintains a positive definite approximation:

$$B_{t+1} = B_t + \frac{y_t y_t^\top}{y_t^\top s_t} - \frac{B_t s_t s_t^\top B_t}{s_t^\top B_t s_t},$$

where $s_t = \theta^{(t+1)} - \theta^{(t)}$ and $y_t = \nabla Q(\theta^{(t+1)}) - \nabla Q(\theta^{(t)})$. BFGS achieves **superlinear convergence** – faster than linear gradient descent but slower than quadratic Newton – while requiring only gradient evaluations. The limited-memory variant (L-BFGS) stores only the most recent m pairs (s_t, y_t), making it feasible for high-dimensional problems.

Convergence Criteria

Standard termination criteria include:

1. **Gradient norm:** $\|\nabla Q(\theta^{(t)})\| < \varepsilon_g$.
2. **Relative change:** $|Q(\theta^{(t+1)}) - Q(\theta^{(t)})| / (|Q(\theta^{(t)})| + 1) < \varepsilon_f$.
3. **Parameter change:** $\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon_\theta$.

In practice, all three are monitored simultaneously. For MLE computation, typical tolerances are $\varepsilon \sim 10^{-8}$.

Worked Examples

Example 1: Newton–Raphson for Gamma MLE

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha, \beta)$ with density $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. The score equations are

$$\frac{\partial \ell}{\partial \alpha} = n \log \beta - n\psi(\alpha) + \sum_{i=1}^{n} \log x_i = 0, \quad \frac{\partial \ell}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0,$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function. From the second equation, $\hat{\beta} = n\alpha / \sum x_i$. Substituting into the first gives a single equation in α that requires Newton–Raphson iteration:

$$\alpha^{(t+1)} = \alpha^{(t)} - \frac{n \log(\alpha^{(t)}/\bar{x}) - n\psi(\alpha^{(t)}) + \sum \log x_i}{n/\alpha^{(t)} - n\psi'(\alpha^{(t)})}.$$

For data with $n = 50$, $\bar{x} = 3.2$, $\sum \log x_i = 48.7$, initializing at $\alpha^{(0)} = (\bar{x})^2 / s^2$ (method of moments), convergence to six decimal places is typical in 4–5 iterations.

Example 2: Fisher Scoring in Logistic Regression

For logistic regression with $P(Y_i = 1 \mid x_i) = \mu_i(\beta) = (1 + e^{-x_i^\top \beta})^{-1}$, the score is

$$\nabla \ell(\beta) = X^\top (y - \mu(\beta)),$$

and the expected Fisher information is

$$\mathcal{I}(\beta) = X^\top W(\beta) X, \quad W = \operatorname{diag}(\mu_i(1 - \mu_i)).$$

Fisher scoring (equivalently, IRLS) iterates

$$\beta^{(t+1)} = (X^\top W^{(t)} X)^{-1} X^\top W^{(t)} z^{(t)},$$

where the working response is $z^{(t)} = X\beta^{(t)} + (W^{(t)})^{-1}(y - \mu^{(t)})$. For a dataset with $n = 100$, $p = 3$, convergence is typically achieved in 5–8 iterations.

Cross-Links

- **Linear Algebra:** Newton–Raphson requires solving a $k \times k$ linear system at each step; the condition number of the Hessian controls convergence rate.
- **AI:** Adam and other adaptive optimizers approximate second-order information using diagonal or low-rank Hessian estimates; natural gradient descent uses the Fisher information matrix as the Riemannian metric.
- **Economics:** Structural estimation in econometrics relies on Newton–Raphson for GMM and maximum likelihood of DSGE models where closed-form solutions are unavailable.
- **Quantum:** Variational quantum eigensolvers use gradient descent on a parameterized quantum circuit; the quantum Fisher information matrix governs the geometry of the parameter space.

Structural Compression

Invariant: The MLE is a fixed point of the score equation $\nabla \ell(\theta) = 0$; Newton–Raphson converges quadratically to this fixed point when initialized in a sufficiently small neighborhood.

Mechanism: Second-order Taylor expansion of Q around the current iterate; the Newton step solves the resulting quadratic approximation exactly. The Hessian encodes local curvature, collapsing the iterative process to a single step for exactly quadratic objectives.

Cross-System Echo: Newton–Raphson is the natural gradient method in the information geometry of the statistical model, where the Fisher information matrix defines the Riemannian metric on the parameter manifold. The tension between first-order methods (cheap but slow) and second-order methods (expensive but fast) recurs in every optimization discipline: gradient descent versus Newton in statistics, SGD versus K-FAC in deep learning, steepest descent versus conjugate gradients in numerical linear algebra.

Exercises

1. Derive the Fisher scoring update for a Poisson regression model with canonical log link and verify that it reduces to an IRLS iteration.
2. For the normal location model $X_i \sim \mathcal{N}(\mu, 1)$, show that Newton–Raphson converges in exactly one step from any initialization. Explain why.

3. Prove that gradient ascent with step size $\alpha = 1/L$ on an L-smooth, m-strongly concave function satisfies $Q(\theta^*) - Q(\theta^{(t)}) \leq (L/2)(1 - m/L)^t \|\theta^{(0)} - \theta^*\|^2$.

4. Consider the Cauchy location model $X_i \sim \text{Cauchy}(\theta, 1)$. Write the Newton–Raphson update for the MLE. Explain why convergence can fail without a good initialization.

5. Show that for an exponential family in canonical form, the observed and expected Fisher information coincide, so Newton–Raphson and Fisher scoring produce identical iterates.

6. Implement backtracking line search for the log-likelihood of a gamma model. Describe how the step size adapts across iterations as the algorithm approaches the MLE.

7. Argue, structurally, why the condition number L/m of the Hessian plays the same role in optimization convergence that the variance ratio plays in statistical efficiency: both quantify how much information the available data (or curvature) provides per unit of computational (or sampling) effort.

Monte Carlo Integration

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.2

Monte Carlo integration converts the analytical problem of evaluating an integral into the statistical problem of estimating a mean. It inherits all the tools of statistical inference – the law of large numbers for consistency, the central limit theorem for error quantification, and variance reduction techniques for efficiency. This node develops the method, its theoretical foundations, and the principal variance reduction strategies.

Formal Definition

Let $h : \mathcal{X} \to \mathbb{R}$ be a measurable function and P a probability distribution on $\mathcal{X}$. The target is

$$I = \mathbb{E}_P[h(X)] = \int h(x)\, dP(x).$$

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, the **Monte Carlo estimator** is

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

By the Strong Law of Large Numbers, $\hat{I}_n \overset{a.s.}{\to} I$ provided $\mathbb{E}_P[|h(X)|] < \infty$. By the Central Limit Theorem,

$$\sqrt{n}(\hat{I}_n - I) \overset{d}{\to} \mathcal{N}(0, \sigma_h^2), \quad \sigma_h^2 = \mathrm{Var}_P(h(X)),$$

so the Monte Carlo standard error is $\mathrm{SE}(\hat{I}_n) = \sigma_h / \sqrt{n}$.

Intuition

The convergence rate $n^{-1/2}$ is dimension-free. Deterministic quadrature in d dimensions converges at rate $n^{-r/d}$ for r-smooth integrands, suffering the curse of dimensionality. Monte Carlo does not: the rate is $n^{-1/2}$ whether $d = 1$ or $d = 10{,}000$. The constant σ_h depends on the integrand, but the rate does not depend on the dimension of the sample space. This makes Monte Carlo the default method for high-dimensional integration in Bayesian statistics, statistical physics, and finance.

Structural Role

Monte Carlo integration is the foundation for MCMC (Node 9.3), which extends the idea to non-i.i.d. samples from intractable distributions. It underlies the bootstrap (Node 9.4), which simulates the sampling distribution by resampling. Bayesian predictive inference (Node 6.5) uses Monte Carlo averages to compute posterior predictive quantities. The variance reduction techniques developed here reappear in importance-weighted MCMC and particle filtering.

Mathematical Development

Importance Sampling

When direct sampling from P is impossible or inefficient, let Q be a **proposal distribution** with $P \ll Q$. Define the importance weight $w(x) = p(x)/q(x)$. Then

$$I = \int h(x)\frac{p(x)}{q(x)}q(x)\,dx = \mathbb{E}_Q[h(X)w(X)].$$

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Q$, the **importance sampling estimator** is

$$\hat{I}_n^{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w(X_i).$$

This is unbiased for I with variance

$$\text{Var}_Q(\hat{I}_n^{\text{IS}}) = \frac{1}{n}\text{Var}_Q(h(X)w(X)) = \frac{1}{n}\left[\int \frac{(h(x)p(x))^2}{q(x)}\,dx - I^2\right].$$

Optimal proposal. The variance-minimizing proposal is $q^*(x) \propto |h(x)|p(x)$, which gives zero variance when $h \geq 0$. In practice, q^* is unknown (it requires the normalizing constant of $|h|p$), but it guides the design of good proposals: q should be large where $|h(x)|p(x)$ is large.

Self-Normalized Importance Sampling

When p is known only up to a normalizing constant, $p(x) = \tilde{p}(x)/Z$, define unnormalized weights $\tilde{w}(x) = \tilde{p}(x)/q(x)$. The **self-normalized estimator** is

$$\hat{I}_n^{\text{SN}} = \frac{\sum_{i=1}^n h(X_i)\tilde{w}(X_i)}{\sum_{i=1}^n \tilde{w}(X_i)}.$$

This is biased but consistent, with bias $O(n^{-1})$. Its asymptotic variance depends on the **effective sample size**:

$$\text{ESS} = \frac{\left(\sum_{i=1}^n \tilde{w}_i\right)^2}{\sum_{i=1}^n \tilde{w}_i^2} \approx \frac{n}{1 + \text{Var}_Q(w)}.$$

When $Q = P$, all weights are equal and $\text{ESS} = n$. When Q is a poor match, a few weights dominate and $\text{ESS} \ll n$.

Control Variates

Let $g : \mathcal{X} \to \mathbb{R}$ with known expectation $\mu_g = \mathbb{E}_P[g(X)]$. The **control variate estimator** is

$$\hat{I}_n^{\text{CV}} = \frac{1}{n}\sum_{i=1}^{n}\left[h(X_i) - c(g(X_i) - \mu_g)\right].$$

This is unbiased for any c. The variance is

$$\text{Var}(\hat{I}_n^{\text{CV}}) = \frac{1}{n}\left[\sigma_h^2 - 2c\,\text{Cov}(h, g) + c^2\sigma_g^2\right].$$

The optimal coefficient is $c^* = \mathrm{Cov}(h, g)/\mathrm{Var}(g)$, yielding variance reduction factor $1 - \rho_{hg}^2$, where ρ_{hg} is the correlation between h and g. The closer $|\rho_{hg}|$ is to 1, the greater the reduction.

Antithetic Variates

For symmetric distributions, generate pairs (X_i, X_i') that are negatively correlated (e.g., $X_i' = -X_i$ for a distribution symmetric about 0). The estimator

$$\hat{I}_n^{\mathrm{AV}} = \frac{1}{n} \sum_{i=1}^{n/2} \frac{h(X_i) + h(X_i')}{2}$$

has variance reduced by $\mathrm{Cov}(h(X), h(X'))/\mathrm{Var}(h(X))$, which is negative when h is monotone.

Stratified Sampling

Partition $\mathcal{X} = \bigcup_{j=1}^{J} A_j$ with $P(A_j) = p_j$. Draw n_j samples from $P(\cdot \mid A_j)$ and estimate

$$\hat{I}_n^{\mathrm{strat}} = \sum_{j=1}^{J} p_j \hat{I}_{n_j}^{(j)}, \quad \hat{I}_{n_j}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} h(X_i^{(j)}).$$

With proportional allocation $n_j = np_j$, the variance is

$$\mathrm{Var}(\hat{I}_n^{\mathrm{strat}}) = \frac{1}{n} \sum_{j=1}^{J} p_j \sigma_j^2 \leq \frac{\sigma_h^2}{n},$$

where $\sigma_j^2 = \mathrm{Var}(h(X) \mid X \in A_j)$. The inequality is strict whenever $\mathbb{E}[h(X) \mid A_j]$ varies across strata, because stratification eliminates the between-stratum variance component.

Proof of variance reduction. By the law of total variance,

$$\mathrm{Var}(h(X)) = \mathbb{E}[\mathrm{Var}(h \mid \text{stratum})] + \mathrm{Var}(\mathbb{E}[h \mid \text{stratum}]) = \sum_j p_j \sigma_j^2 + \sum_j p_j (\mu_j - I)^2,$$

where $\mu_j = \mathbb{E}[h(X) \mid A_j]$. Simple Monte Carlo has variance $\mathrm{Var}(h)/n$; stratified sampling has variance $\sum p_j \sigma_j^2 / n$. The difference is $\sum p_j (\mu_j - I)^2 / n \geq 0$. □

Optimal allocation. Neyman allocation sets $n_j \propto p_j \sigma_j$, minimizing the stratified variance to $(\sum_j p_j \sigma_j)^2 / n$, which is further reduced when strata have unequal variability.

CLT for Monte Carlo Estimators

The central limit theorem provides the basis for confidence intervals. For the standard Monte Carlo estimator:

$$P\left(I \in \left[\hat{I}_n - z_{\alpha/2} \frac{\hat{\sigma}_h}{\sqrt{n}},\ \hat{I}_n + z_{\alpha/2} \frac{\hat{\sigma}_h}{\sqrt{n}} \right] \right) \to 1 - \alpha,$$

where $\hat{\sigma}_h^2 = (n-1)^{-1} \sum (h(X_i) - \hat{I}_n)^2$. The confidence interval width is $O(n^{-1/2})$ regardless of dimension, confirming the dimension-free convergence rate. For importance sampling, the CLT applies to $h(X)w(X)$, but the effective variance may be much larger or smaller than σ_h^2 depending on the proposal quality.

Worked Examples

Example 1: Estimating a Posterior Mean

Let $\theta \sim \text{Gamma}(3,1)$ (prior) and $X \mid \theta \sim \text{Poisson}(\theta)$ with $x = 7$ observed. The posterior is $\theta \mid x \sim \text{Gamma}(10,2)$. The posterior mean is $\mathbb{E}[\theta \mid x] = 10/2 = 5$. To verify by Monte Carlo: draw $\theta_1, \ldots, \theta_n \sim \text{Gamma}(10,2)$ and compute $\hat{I}_n = n^{-1}\sum \theta_i$. With $n = 10{,}000$, the standard error is $\sqrt{10/4}/100 = 0.0158$, giving a 95% confidence interval of approximately $(4.969, 5.031)$.

Example 2: Importance Sampling for Rare-Event Probability

Estimate $P(X > 5)$ where $X \sim \mathcal{N}(0,1)$. Direct Monte Carlo requires enormous n since $P(X > 5) \approx 2.87 \times 10^{-7}$. Use importance sampling with proposal $Q = \mathcal{N}(5,1)$:

$$w(x) = \frac{\phi(x)}{\phi(x-5)} = \exp\left(-5x + \frac{25}{2}\right).$$

The estimator $\hat{p} = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i > 5) w(X_i)$ with $X_i \sim \mathcal{N}(5,1)$ concentrates samples in the region of interest. With $n = 1{,}000$, approximately half the draws exceed 5, and the weighted average gives an estimate with coefficient of variation below 0.1, compared to a coefficient of variation exceeding 50 for naive Monte Carlo at the same n.

Cross-Links

- **Linear Algebra:** Stratified sampling exploits the decomposition of total variance into within-stratum and between-stratum components, mirroring the ANOVA decomposition.
- **AI:** Stochastic gradient descent is Monte Carlo estimation of the population risk gradient; mini-batch size controls the variance-computation tradeoff. Policy gradient methods in reinforcement learning use control variates (baselines) to reduce gradient variance.
- **Economics:** Monte Carlo simulation is the standard tool for pricing path-dependent derivatives in finance; importance sampling is used to estimate the probability of extreme portfolio losses (Value-at-Risk).
- **Quantum:** Quantum Monte Carlo methods estimate ground-state properties by sampling from the path integral representation; the sign problem is a variance explosion analogous to importance weight degeneracy.

Structural Compression

Invariant: The Monte Carlo estimator $\hat{I}_n$ is unbiased with variance $\text{Var}(h)/n$, converging at rate $n^{-1/2}$ regardless of the dimension of the sample space.

Mechanism: The SLLN guarantees almost-sure convergence of sample averages to population means; the CLT characterizes the $n^{-1/2}$ Gaussian fluctuations. Variance reduction techniques modify the integrand or sampling scheme to decrease the constant σ_h^2 without changing the rate.

Cross-System Echo: The dimension-free convergence rate of Monte Carlo is structurally identical to the statistical principle that sample means converge at rate $n^{-1/2}$ regardless of the population distribution. In every system where high-dimensional integration appears – posterior computation in statistics, partition function estimation in physics, expected utility computation in economics – Monte Carlo is the universal computational primitive precisely because it circumvents the curse of dimensionality.

Exercises

1. Prove that the importance sampling estimator $\hat{I}_n^{\mathrm{IS}}$ is unbiased and derive its variance in terms of $\mathbb{E}_Q[(h(X)w(X))^2]$.
2. For $h(x) = x^2$ and $P = \mathcal{N}(0, 1)$, compute the optimal importance sampling proposal $q^*(x) \propto x^2\phi(x)$. Identify the distribution and show that the resulting estimator has zero variance.
3. Derive the optimal control variate coefficient $c^* = \mathrm{Cov}(h, g)/\mathrm{Var}(g)$ and show that the variance reduction factor is $1 - \rho_{hg}^2$.
4. Show that stratified sampling with proportional allocation always reduces variance compared to simple random sampling, with equality iff the stratum means are all equal.
5. For self-normalized importance sampling, derive the $O(n^{-1})$ bias using a Taylor expansion of the ratio estimator, and show that the bias vanishes as $n \to \infty$.
6. Estimate $\int_0^1 e^x\,dx$ using Monte Carlo with $n = 1{,}000$ and the control variate $g(x) = x$ (with $\mu_g = 1/2$). Compare the standard error with and without the control variate.
7. Argue, structurally, why the dimension-free convergence rate of Monte Carlo integration represents the same principle as the dimension-free convergence of sample means in the CLT: both are consequences of the fact that averaging independent random variables reduces variance at rate $1/n$, and this rate depends on the number of observations, not the complexity of the underlying space.

Markov Chain Monte Carlo

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.3

MCMC extends Monte Carlo integration (Node 9.2) to settings where direct i.i.d. sampling from the target distribution is infeasible. By constructing a Markov chain whose stationary distribution is the target, MCMC converts the problem of sampling into the problem of simulating a chain long enough for ergodic averages to converge. This node develops the Metropolis–Hastings algorithm, proves that detailed balance implies stationarity, derives Gibbs sampling as a special case, and establishes the convergence diagnostic framework.

Formal Definition

Let $\pi(\theta)$ be a target density on $\Theta \subseteq \mathbb{R}^k$, known up to a normalizing constant. A **Markov chain Monte Carlo** method constructs a Markov chain $(\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \dots)$ with transition kernel $K(\theta, \theta')$ such that:

1. π is a **stationary distribution**: $\int K(\theta, \theta')\pi(\theta)\, d\theta = \pi(\theta')$.
2. The chain is **ergodic** (irreducible, aperiodic, positive recurrent).

Under these conditions, the **ergodic theorem** guarantees: for any h with $\mathbb{E}_\pi[|h|] < \infty$,

$$\frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \overset{a.s.}{\to} \mathbb{E}_\pi[h(\theta)].$$

A sufficient condition for stationarity is **detailed balance**:

$$\pi(\theta)K(\theta, \theta') = \pi(\theta')K(\theta', \theta) \quad \text{for all } \theta, \theta'.$$

Intuition

MCMC trades the requirement of independent samples for the ability to sample from distributions known only up to a normalizing constant. The chain wanders through parameter space, spending time in regions proportional to their probability under π. The art of MCMC is designing chains that explore the target distribution efficiently – avoiding both slow random walks and proposals that are rejected too often. The burn-in period discards early samples that are influenced by the initialization, and convergence diagnostics assess whether the chain has reached stationarity.

Structural Role

MCMC is the computational engine of modern Bayesian statistics. It extends posterior computation beyond conjugate families (Node 6.2) to arbitrary posteriors, including hierarchical models, nonparametric models, and latent variable models. It depends on Monte Carlo integration (Node 9.2) for the ergodic averaging framework and enables variational inference (Node 9.6) as a faster alternative when MCMC mixing is prohibitively slow. The bootstrap (Node 9.4) is a parallel computational strategy that uses resampling rather than Markov chains.

Mathematical Development

Metropolis–Hastings Algorithm

Let $q(\theta' \mid \theta)$ be a proposal kernel. The **Metropolis–Hastings** algorithm generates the chain as follows:

1. Given $\theta^{(t)}$, draw $\theta' \sim q(\cdot \mid \theta^{(t)})$.
2. Compute the acceptance probability:

$$\alpha(\theta^{(t)}, \theta') = \min\left(1,\ \frac{\pi(\theta')\, q(\theta^{(t)} \mid \theta')}{\pi(\theta^{(t)})\, q(\theta' \mid \theta^{(t)})}\right).$$

3. With probability α, set $\theta^{(t+1)} = \theta'$; otherwise, set $\theta^{(t+1)} = \theta^{(t)}$.

Theorem (Detailed Balance). The Metropolis–Hastings chain satisfies detailed balance with respect to π.

Proof. The transition kernel is

$$K(\theta, \theta') = q(\theta' \mid \theta)\alpha(\theta, \theta') + r(\theta)\delta_\theta(\theta'),$$

where $r(\theta) = 1 - \int q(\theta' \mid \theta)\alpha(\theta, \theta')\, d\theta'$ is the rejection probability. For the continuous part, we must show $\pi(\theta)q(\theta' \mid \theta)\alpha(\theta, \theta') = \pi(\theta')q(\theta \mid \theta')\alpha(\theta', \theta)$.

Without loss of generality, assume $\pi(\theta')q(\theta \mid \theta') \le \pi(\theta)q(\theta' \mid \theta)$. Then $\alpha(\theta, \theta') = \pi(\theta')q(\theta \mid \theta')/[\pi(\theta)q(\theta' \mid \theta)]$ and $\alpha(\theta', \theta) = 1$. Substituting:

$$\pi(\theta)q(\theta' \mid \theta) \cdot \frac{\pi(\theta')q(\theta \mid \theta')}{\pi(\theta)q(\theta' \mid \theta)} = \pi(\theta')q(\theta \mid \theta') \cdot 1. \quad \square$$

Since π is the stationary distribution and the chain is typically irreducible and aperiodic (given mild conditions on q), the ergodic theorem applies.

Special Case: Random Walk Metropolis

When $q(\theta' \mid \theta) = g(\theta' - \theta)$ for a symmetric density g, the acceptance ratio simplifies to

$$\alpha(\theta^{(t)}, \theta') = \min\left(1, \frac{\pi(\theta')}{\pi(\theta^{(t)})}\right).$$

The proposal variance controls the tradeoff: too small yields high acceptance but slow exploration; too large yields frequent rejection. The optimal acceptance rate for a k-dimensional Gaussian target is approximately 0.234 (Roberts, Gelman, and Gilks, 1997).

Gibbs Sampling

When $\theta = (\theta_1, \dots, \theta_k)$ and the full conditional distributions $\pi(\theta_j \mid \theta_{-j})$ are available, **Gibbs sampling** cycles through:

For $j = 1, \dots, k$: draw $\theta_j^{(t+1)} \sim \pi(\theta_j \mid \theta_1^{(t+1)}, \dots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \dots, \theta_k^{(t)})$.

Proposition. Gibbs sampling is a special case of Metropolis–Hastings with proposal $q(\theta' \mid \theta) = \pi(\theta'_j \mid \theta_{-j}) \prod_{i \neq j} \delta_{\theta_i}(\theta'_i)$ and acceptance probability identically 1.

Proof. The MH ratio for coordinate j is

$$\frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} = \frac{\pi(\theta'_j, \theta_{-j})\, \pi(\theta_j \mid \theta_{-j})}{\pi(\theta_j, \theta_{-j})\, \pi(\theta'_j \mid \theta_{-j})} = \frac{\pi(\theta_{-j})\pi(\theta'_j \mid \theta_{-j})\, \pi(\theta_j \mid \theta_{-j})}{\pi(\theta_{-j})\pi(\theta_j \mid \theta_{-j})\, \pi(\theta'_j \mid \theta_{-j})} = 1. \quad \square$$

Conjugate priors (Node 6.2) yield tractable full conditionals, making Gibbs the canonical algorithm for hierarchical Bayesian models.

Effective Sample Size

MCMC samples are correlated. The **effective sample size** is

$$\mathrm{ESS} = \frac{T}{1 + 2\sum_{k=1}^{\infty} \rho_k},$$

where $\rho_k = \mathrm{Corr}(h(\theta^{(t)}), h(\theta^{(t+k)}))$ is the lag-k autocorrelation. The Monte Carlo standard error for $\bar{h}_T = T^{-1}\sum_t h(\theta^{(t)})$ is

$$\mathrm{SE}(\bar{h}_T) = \sqrt{\frac{\mathrm{Var}_\pi(h)}{\mathrm{ESS}}}.$$

In practice, the infinite sum is truncated at the first negative autocorrelation or estimated using spectral methods.

Convergence Diagnostics

Gelman–Rubin $\hat{R}$ statistic. Run $M \geq 2$ chains of length T. Let B be the between-chain variance and W the within-chain variance. Define

$$\hat{R} = \sqrt{\frac{(T-1)/T \cdot W + B/T}{W}}.$$

As $T \to \infty$, $\hat{R} \to 1$ if all chains have converged to the stationary distribution. Values $\hat{R} > 1.01$ indicate incomplete convergence.

Trace plots provide visual assessment: a well-mixing chain shows rapid, uncorrelated fluctuations around a stable level. Persistent trends or multimodal behavior indicate convergence failure.

Autocorrelation function (ACF): slow decay of ρ_k indicates poor mixing and low ESS per iteration.

MCMC Central Limit Theorem

Under regularity conditions (geometric ergodicity), the MCMC ergodic average satisfies a CLT:

$$\sqrt{T}(\bar{h}_T - \mathbb{E}_\pi[h]) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\mathrm{MCMC}}),$$

where the **asymptotic variance** is

$$\sigma^2_{\mathrm{MCMC}} = \mathrm{Var}_\pi(h) \cdot \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right) = \mathrm{Var}_\pi(h) \cdot \frac{T}{\mathrm{ESS}}.$$

The factor $1 + 2\sum \rho_k$ is the **integrated autocorrelation time** (IACT). A well-mixing chain has small IACT and large ESS; a poorly mixing chain has large IACT and ESS much smaller than the nominal sample size T.

Burn-in and Thinning

Burn-in discards the first T_0 samples to reduce sensitivity to initialization. Formally, if the chain converges geometrically, the bias from initialization decays as $O(\rho^{T_0})$, where $\rho < 1$ is the spectral gap.

Thinning retains every m-th sample, reducing storage but not improving ESS per unit of computation. Thinning is generally unnecessary for estimation but may be practical for storage-limited applications.

Worked Examples

Example 1: Metropolis for a Beta Posterior

Let $X \sim \text{Binomial}(20, \theta)$ with $x = 13$ observed, and $\theta \sim \text{Beta}(1, 1)$ (uniform prior). The posterior is $\theta \mid x \sim \text{Beta}(14, 8)$. Use random walk Metropolis with proposal $\theta' = \theta^{(t)} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, 0.05^2)$, truncated to $[0, 1]$. The acceptance ratio is $\alpha = \min(1, \theta'^{13}(1-\theta')^7 / [\theta^{(t)}]^{13}(1-\theta^{(t)})^7)$. After 10,000 iterations with 1,000 burn-in, the ergodic mean $\bar{\theta} \approx 0.636$ agrees with the exact posterior mean $14/22 \approx 0.636$, and a histogram of the chain matches the $\text{Beta}(14, 8)$ density.

Example 2: Gibbs Sampler for a Bivariate Normal

Let $(X, Y) \sim \mathcal{N}_2(0, \Sigma)$ with $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The full conditionals are $X \mid Y = y \sim \mathcal{N}(\rho y, 1 - \rho^2)$ and $Y \mid X = x \sim \mathcal{N}(\rho x, 1 - \rho^2)$. The Gibbs sampler alternates between these. For $\rho = 0.9$, the lag-1 autocorrelation of the X-chain is $\rho^2 = 0.81$, so $\text{ESS} \approx T(1 - 0.81)/(1 + 0.81) = T \cdot 0.105$. For $T = 10{,}000$, $\text{ESS} \approx 1{,}050$. High correlation in the target distribution produces high autocorrelation in the chain, reducing effective sample size.

Cross-Links

- **Linear Algebra:** The mixing rate of a Metropolis chain on a discrete state space is governed by the spectral gap of the transition matrix; fast mixing requires the second-largest eigenvalue to be bounded away from 1.
- **AI:** Langevin MCMC (gradient of log-density added to random walk proposal) is used for sampling from energy-based models; stochastic gradient Langevin dynamics combines MCMC with mini-batch gradient estimation for scalable Bayesian deep learning.
- **Economics:** MCMC is used to estimate posterior distributions of structural parameters in DSGE models and Bayesian VARs where the likelihood is available only numerically.
- **Quantum:** Quantum Monte Carlo methods (path-integral Monte Carlo, diffusion Monte Carlo) use MCMC to sample from the Boltzmann distribution of quantum systems; the sign problem is a fundamental barrier analogous to multimodality in classical MCMC.

Structural Compression

Invariant: A Markov chain satisfying detailed balance with respect to π has π as its stationary distribution; ergodic averages converge almost surely to π-expectations.

Mechanism: The Metropolis acceptance probability is constructed so that detailed balance holds by construction. The ratio form $\pi(\theta')/\pi(\theta)$ cancels the normalizing constant, enabling sampling from distributions known only up to proportionality.

Cross-System Echo: MCMC converts integration into simulation, just as Monte Carlo integration converts integration into averaging. The structural pattern – replacing an analytically intractable global computation with an iterative local computation that converges to the global answer – appears throughout computational

science: relaxation methods in PDE solvers, message passing in graphical models, and simulated annealing in combinatorial optimization are all instances of the same principle.

Exercises

1. Prove that detailed balance $\pi(\theta)K(\theta, \theta') = \pi(\theta')K(\theta', \theta)$ implies π is a stationary distribution of K, by integrating both sides over θ.

2. For a random walk Metropolis targeting $\pi(\theta) \propto e^{-\theta^4}$ on $\mathbb{R}$ with Gaussian proposal variance σ^2, describe qualitatively how the acceptance rate and mixing behavior change as σ^2 varies from 0.01 to 100.

3. Derive the full conditional distributions for a Bayesian normal model $X_i \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ with priors $\mu \sim \mathcal{N}(0, \tau^2)$ and $\sigma^2 \sim \text{Inv-Gamma}(a, b)$. Write the Gibbs sampling algorithm.

4. Show that the effective sample size satisfies $\text{ESS} \le T$, with equality iff the chain produces independent samples (all autocorrelations are zero).

5. For the Gibbs sampler on a bivariate normal with correlation ρ, prove that the lag-1 autocorrelation of each coordinate is ρ^2, and derive the ESS as a function of T and ρ.

6. Implement the Gelman–Rubin diagnostic for 4 chains of length 5,000 targeting a $\text{Beta}(2, 5)$ distribution. Report $\hat{R}$ and compare it to the threshold 1.01.

7. Argue, structurally, why the normalizing-constant cancellation in the Metropolis ratio is the same structural principle as the sufficiency reduction in classical statistics: both allow inference to proceed without computing a quantity (the marginal likelihood or the normalizing constant) that depends on the full model but not on the parameter of interest.

Bootstrap Implementation

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.4

The bootstrap translates the theoretical coverage guarantees of Node 4.6 into a concrete computational algorithm. It implements the plug-in principle: the empirical distribution $\hat{P}_n$ replaces the unknown population P, and resampling from $\hat{P}_n$ simulates the sampling variability of estimators. This node specifies the nonparametric and parametric bootstrap algorithms, develops bias correction, establishes consistency conditions, and identifies failure modes.

Formal Definition

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ and let $\hat{\theta}_n = T(\hat{P}_n)$ be a functional of the empirical distribution $\hat{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$.

The **nonparametric bootstrap distribution** of $\hat{\theta}_n$ is the conditional distribution of $\hat{\theta}_n^* = T(\hat{P}_n^*)$, where $\hat{P}_n^*$ is the empirical distribution of a sample drawn with replacement from $\{X_1, \ldots, X_n\}$, conditional on the observed data.

The **bootstrap principle** asserts that the distribution of $\hat{\theta}_n^* - \hat{\theta}_n$ (given the data) approximates the distribution of $\hat{\theta}_n - \theta(P)$ (over the sampling distribution).

Intuition

The bootstrap is the computational realization of a simple idea: if the empirical distribution is close to the true distribution (which Glivenko–Cantelli guarantees), then resampling from the empirical distribution should mimic sampling from the true distribution. The power of the bootstrap is its generality – it requires only the ability to compute $\hat{\theta}_n$ on resampled data, with no need for analytical variance formulas or distributional assumptions. The limitation is that the bootstrap is not magic: it fails when the empirical distribution is a poor approximation to aspects of P that matter for the distribution of $\hat{\theta}_n$.

Structural Role

Bootstrap implementation is the computational companion to the asymptotic theory of Node 4.6. It provides standard errors and confidence intervals for estimators where analytical variance formulas are unavailable or unreliable. The Monte Carlo approximation to the bootstrap distribution uses the framework of Node 9.2. In machine learning, the bootstrap underlies bagging (bootstrap aggregating), which reduces variance in ensemble methods.

Mathematical Development

Nonparametric Bootstrap Algorithm

Given observed data $x_1, \ldots, x_n$:

1. For $b = 1, \ldots, B$:
   - Draw $x_1^{(b)}, \ldots, x_n^{(b)}$ by sampling with replacement from $\{x_1, \ldots, x_n\}$.

- Compute $\hat{\theta}_n^{(b)} = T(\hat{P}_n^{(b)})$.

2. The bootstrap distribution is $\{\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(B)}\}$.

Number of resamples. For standard errors: $B = 200$ suffices. For confidence intervals: $B = 1{,}000$ to $2{,}000$. For tail probabilities: $B \geq 5{,}000$.

Parametric Bootstrap

When the model $\{P_\theta : \theta \in \Theta\}$ is specified:

1. Compute $\hat{\theta}_n$ from the data.
2. For $b = 1, \ldots, B$: draw $X_1^{(b)}, \ldots, X_n^{(b)} \overset{\text{i.i.d.}}{\sim} P_{\hat{\theta}_n}$ and compute $\hat{\theta}_n^{(b)}$.

The parametric bootstrap samples from $P_{\hat{\theta}_n}$ rather than $\hat{P}_n$. It achieves higher accuracy when the model is correctly specified, because $P_{\hat{\theta}_n}$ is a smoother (and typically better) estimate of P than $\hat{P}_n$.

Bootstrap Consistency

Theorem (Bootstrap Consistency). Suppose $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ and σ^2 is a smooth functional of P. Then the bootstrap is consistent:

$$\sup_t \left| P^*\Big(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \leq t\Big) - P\Big(\sqrt{n}(\hat{\theta}_n - \theta) \leq t\Big) \right| \xrightarrow{P} 0,$$

where P^* denotes probability under the bootstrap measure.

Proof sketch. By the continuous mapping theorem applied to the functional T and the convergence $\hat{P}_n \to P$ in the Levy–Prokhorov metric, the bootstrap distribution converges to the same Gaussian limit as the sampling distribution. The key condition is that T is Hadamard differentiable at P, which ensures that the linearization $T(\hat{P}_n^*) - T(\hat{P}_n) \approx T'_P(\hat{P}_n^* - \hat{P}_n)$ is valid. □

Bootstrap Standard Error and Confidence Intervals

The bootstrap standard error is

$$\widehat{\mathrm{SE}}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_n^{(b)} - \bar{\theta}^*\right)^2}, \quad \bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_n^{(b)}.$$

Percentile interval:

$$C^{\text{perc}} = \left[q^*_{\alpha/2}, q^*_{1-\alpha/2}\right],$$

where q_p^* is the p-quantile of $\{\hat{\theta}_n^{(b)}\}$.

Basic (reverse) interval:

$$C^{\text{basic}} = \left[2\hat{\theta}_n - q^*_{1-\alpha/2}, 2\hat{\theta}_n - q^*_{\alpha/2}\right].$$

Bootstrap-t interval:

$$C^t = \left[\hat{\theta}_n - q^{*,t}_{1-\alpha/2}\,\hat{\sigma}_n,\ \hat{\theta}_n - q^{*,t}_{\alpha/2}\,\hat{\sigma}_n\right],$$

where $q_p^{*,t}$ are quantiles of the studentized bootstrap statistics $(\hat{\theta}_n^{(b)} - \hat{\theta}_n)/\hat{\sigma}_n^{(b)}$. The bootstrap-$t$ achieves second-order accuracy (coverage error $O(n^{-1})$) compared to $O(n^{-1/2})$ for the percentile interval. The bootstrap-t requires a nested computation: for each bootstrap resample, an internal standard error $\hat{\sigma}_n^{(b)}$ must be computed, either analytically or by a second-level bootstrap.

Coverage error comparison. Let C_n denote a nominal $(1-\alpha)$ confidence interval. The coverage error is $|P(\theta \in C_n) - (1-\alpha)|$:

- Normal approximation: $O(n^{-1/2})$.
- Percentile bootstrap: $O(n^{-1/2})$ (same order, different constant).
- Bootstrap-t and BCa: $O(n^{-1})$ (second-order accurate).
- Double bootstrap: $O(n^{-2})$ (third-order accurate).

BCa (Bias-Corrected and Accelerated) Method

The **BCa interval** corrects for both bias and skewness:

$$C^{\mathrm{BCa}} = \left[q^*_{\alpha_1}, q^*_{\alpha_2}\right],$$

where the adjusted quantile levels are

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})}\right), \quad \alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-\alpha/2}}{1 - a(z_0 + z_{1-\alpha/2})}\right).$$

Here $z_0 = \Phi^{-1}(\#\{\hat{\theta}_n^{(b)} < \hat{\theta}_n\}/B)$ is the bias correction and a is the acceleration constant, estimated from the jackknife:

$$a = \frac{\sum_{i=1}^n (\bar{\theta}_{(\cdot)} - \hat{\theta}_{(-i)})^3}{6\left[\sum_{i=1}^n (\bar{\theta}_{(\cdot)} - \hat{\theta}_{(-i)})^2\right]^{3/2}},$$

where $\hat{\theta}_{(-i)}$ is the estimate with observation i deleted.

When the Bootstrap Fails

The bootstrap fails when $\hat{P}_n$ does not capture the features of P relevant to the distribution of $\hat{\theta}_n$:

1. **Heavy tails without finite variance.** If $\mathrm{Var}_P(X) = \infty$, the bootstrap variance is finite (since $\hat{P}_n$ has finite support) and underestimates the true variability.
2. **Extremes and boundary estimators.** For $\hat{\theta}_n = \max(X_1, \ldots, X_n)$, the bootstrap distribution concentrates on the observed maximum, failing to capture the correct $n(X_{(n)} - \theta)$ scaling.
3. **Non-smooth functionals.** If T is not Hadamard differentiable (e.g., the mode), the bootstrap may be inconsistent.
4. **Dependent data.** The i.i.d. bootstrap is invalid for time series; block bootstrap or sieve bootstrap are required.

Double Bootstrap

For calibrated confidence intervals with coverage error $O(n^{-2})$, the **double bootstrap** uses nested resampling: for each bootstrap sample $b = 1, \ldots, B_1$, generate B_2 second-level bootstrap resamples to estimate the coverage of the first-level interval, then adjust the quantile accordingly.

The double bootstrap is computationally intensive ($B_1 \times B_2$ total evaluations of T) but achieves second-order accuracy without requiring analytical calculation of the acceleration constant a in the BCa method.

Exact Bootstrap Distribution

The exact nonparametric bootstrap distribution averages over all $\binom{2n-1}{n}$ possible resamples (each corresponding to a multinomial draw of size n from n categories). The Monte Carlo bootstrap with B resamples is an approximation to this exact distribution. The Monte Carlo error (from finite B) is distinct from the bootstrap error (from $\hat{P}_n \neq P$); the former is $O(B^{-1/2})$ and can be made arbitrarily small by increasing B, while the latter is an intrinsic statistical limitation.

Residual Bootstrap for Regression

For the linear model $Y = X\beta + \varepsilon$, the **residual bootstrap** resamples the residuals $\hat{e}_i = Y_i - X_i^\top \hat{\beta}$ and constructs bootstrap responses $Y_i^{(b)} = X_i^\top \hat{\beta} + \hat{e}_{\pi(i)}$, where π is a random permutation with replacement. This preserves the fixed design while resampling the error distribution, and is appropriate when the regression structure is assumed correct but the error distribution is unknown.

Worked Examples

Example 1: Bootstrap Standard Error of the Median

Let $x_1, \ldots, x_{20}$ be a sample from an unknown distribution with observed median $\hat{m} = 3.7$. There is no simple formula for $\mathrm{SE}(\hat{m})$ without distributional assumptions. Generate $B = 2{,}000$ bootstrap resamples, compute the median of each, and take the standard deviation: $\widehat{\mathrm{SE}}_B = 0.42$. The 95% percentile interval is $[2.9, 4.5]$. Contrast this with the normal-based interval $\hat{m} \pm 1.96 \cdot \widehat{\mathrm{SE}}_B = [2.88, 4.52]$, which is similar here but would differ for skewed bootstrap distributions.

Example 2: Parametric vs. Nonparametric Bootstrap for Normal Variance

Let $x_1, \ldots, x_{15} \sim \mathcal{N}(5, 4)$ with observed sample variance $s^2 = 3.8$. The exact sampling distribution gives $(n-1)s^2/\sigma^2 \sim \chi^2_{14}$, yielding a 95% CI for σ^2 of $[2.09, 8.44]$.

Nonparametric bootstrap: resample with replacement, compute $s^{2(b)}$ for $B = 2{,}000$ resamples. The percentile interval is approximately $[1.8, 7.1]$ – shorter than the exact interval because the empirical distribution has lighter tails than the normal.

Parametric bootstrap: draw $X^{(b)} \sim \mathcal{N}(\bar{x}, s^2)$, compute $s^{2(b)}$. The percentile interval is approximately $[2.1, 8.3]$, closely matching the exact interval. The parametric bootstrap is superior here because the normality assumption is correct and the parametric model estimates P more accurately than $\hat{P}_n$ for tail behavior.

Example 3: Bootstrap for Correlation Coefficient

Let $(X_i, Y_i)_{i=1}^{30}$ be bivariate data with sample correlation $r = 0.72$. Fisher's z-transformation gives an approximate confidence interval, but the bootstrap provides a nonparametric alternative. Resample pairs $(X_i^{(b)}, Y_i^{(b)})$ with replacement and compute $r^{(b)}$ for $B = 2{,}000$. The bootstrap distribution of r is left-skewed (bounded above by 1). The BCa interval $[0.48, 0.87]$ accounts for this skewness, while the percentile interval $[0.45, 0.86]$ does not. The acceleration constant $a = -0.03$ reflects the negative skewness induced by the upper bound.

Cross-Links

- **Linear Algebra:** The bootstrap covariance matrix of a vector estimator $\hat{\theta}_n \in \mathbb{R}^k$ is the sample covariance of the bootstrap replicates, providing a nonparametric estimate of the asymptotic covariance matrix that would otherwise require the delta method and Fisher information.
- **AI:** Bagging (bootstrap aggregating) generates an ensemble of predictors trained on bootstrap resamples; the variance reduction from averaging parallels the bootstrap's variance estimation. Random forests extend bagging with feature subsampling.
- **Economics:** The bootstrap is the standard tool for inference in structural econometric models where the asymptotic distribution of estimators (GMM, two-stage least squares) depends on nuisance parameters in complex ways.
- **Quantum:** Quantum state tomography uses bootstrap resampling to quantify uncertainty in reconstructed density matrices when analytical error bars are unavailable.

Structural Compression

Invariant: The empirical distribution $\hat{P}_n$ is a consistent estimate of P; resampling from $\hat{P}_n$ consistently estimates the sampling distribution of $T(\hat{P}_n)$ whenever T is sufficiently smooth (Hadamard differentiable).

Mechanism: Drawing with replacement from the observed sample generates a Monte Carlo approximation to the bootstrap distribution; quantiles of this approximation yield confidence interval endpoints. The BCa correction adjusts for bias and skewness, achieving second-order accuracy.

Cross-System Echo: The bootstrap implements the same structural principle as the plug-in estimator: replace the unknown P with its empirical estimate and propagate uncertainty through the functional T. This principle – that consistent estimation of the input yields consistent estimation of smooth functionals of the input – is the computational form of the continuous mapping theorem and appears in every system where simulation replaces analysis.

Exercises

1. Prove that the bootstrap estimate of the mean, $\bar{X}^* = n^{-1}\sum_{i=1}^{n} X_i^*$, has conditional variance $\hat{\sigma}^2/n$ where $\hat{\sigma}^2 = n^{-1}\sum(X_i - \bar{X})^2$, matching the plug-in variance estimator.

2. For the sample variance S^2, derive the bootstrap distribution analytically when $n = 3$ by enumerating all $3^3 = 27$ bootstrap resamples.

3. Show that the percentile bootstrap interval is transformation-respecting: if C^{perc} is an interval for θ, then $g(C^{\text{perc}})$ is a valid interval for $g(\theta)$ for any monotone g.

4. Explain why the bootstrap fails for $\hat{\theta}_n = X_{(n)} = \max(X_1, \ldots, X_n)$ when P is uniform on $[0, \theta]$. What is the correct rate of convergence, and why does the bootstrap miss it?

5. Derive the bias-correction constant z_0 in the BCa method and show that when $z_0 = 0$ and $a = 0$, the BCa interval reduces to the percentile interval.

6. Compare the parametric bootstrap (assuming normality) and the nonparametric bootstrap for the sample median from a skewed distribution. Generate data from Exp(1), $n = 30$, and compare coverage of 95% intervals over 500 simulation replications.

7. Argue, structurally, why the bootstrap's reliance on the plug-in principle means it inherits both the strengths and limitations of empirical distribution estimation: it succeeds whenever the empirical distribution captures the relevant features of P (bulk behavior, smooth functionals) and fails whenever the relevant features are in the tails or boundaries that the empirical distribution cannot resolve with n observations.

Expectation-Maximization Algorithm

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.5

The EM algorithm is the canonical method for maximum likelihood estimation in latent variable models, where the observed-data log-likelihood involves an intractable integral over latent variables. It converts this intractable optimization into a sequence of tractable complete-data optimizations by alternating between computing expected sufficient statistics (E-step) and maximizing the complete-data likelihood (M-step). This node derives the ELBO decomposition that justifies EM, proves monotonicity of the likelihood sequence, analyzes convergence rate, and establishes the connection to variational inference.

Formal Definition

Let Y denote observed data and Z denote latent (unobserved) variables. The complete-data log-likelihood is $\ell_c(\theta; Y, Z) = \log p(Y, Z \mid \theta)$. The observed-data log-likelihood is

$$\ell(\theta; Y) = \log p(Y \mid \theta) = \log \int p(Y, Z \mid \theta)\, dZ.$$

The **Q-function** at iteration t is

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z \mid Y, \theta^{(t)}}[\log p(Y, Z \mid \theta)] = \int \log p(Y, Z \mid \theta)\, p(Z \mid Y, \theta^{(t)})\, dZ.$$

The **EM algorithm** iterates:

E-step: Compute $Q(\theta \mid \theta^{(t)})$ (or equivalently, the sufficient statistics of $p(Z \mid Y, \theta^{(t)})$).

M-step: Set $\theta^{(t+1)} = \arg\max_\theta Q(\theta \mid \theta^{(t)})$.

Intuition

The observed-data likelihood marginalizes over the latent variables, creating a complex surface with potential local maxima. The EM algorithm avoids directly climbing this surface. Instead, it constructs a simpler surrogate function (the Q-function) that touches the log-likelihood at the current parameter value and lies below it everywhere else. Maximizing this surrogate is guaranteed to increase (or maintain) the log-likelihood. The E-step computes what the latent variables "should" look like given the current parameters; the M-step finds the best parameters given those imputed latent variables.

Structural Role

EM is the bridge between latent variable models and computational tractability. It depends on numerical optimization (Node 9.1) for the M-step, on Bayesian posterior computation (Node 6.1) for the E-step, and on the MLE framework (Node 3.2) for the overall objective. It enables variational inference (Node 9.6) by providing the ELBO framework that VI generalizes. EM is the canonical algorithm for Gaussian mixtures, hidden Markov models, factor analysis, and probabilistic PCA.

Mathematical Development

The ELBO Decomposition

For any distribution $q(Z)$ over the latent variables, we can decompose the observed log-likelihood:

$$\ell(\theta; Y) = \log p(Y \mid \theta) = \log \int p(Y, Z \mid \theta)\, dZ.$$

Introduce $q(Z)$ and apply Jensen's inequality:

$$\ell(\theta; Y) = \log \int \frac{p(Y, Z \mid \theta)}{q(Z)} q(Z)\, dZ \geq \int q(Z) \log \frac{p(Y, Z \mid \theta)}{q(Z)}\, dZ.$$

The right-hand side is the **Evidence Lower BOund (ELBO)**:

$$\mathrm{ELBO}(q, \theta) = \mathbb{E}_q[\log p(Y, Z \mid \theta)] - \mathbb{E}_q[\log q(Z)].$$

The gap between $\ell(\theta; Y)$ and the ELBO is exactly the KL divergence:

$$\ell(\theta; Y) = \mathrm{ELBO}(q, \theta) + \mathrm{KL}(q(Z) \| p(Z \mid Y, \theta)).$$

Proof. Write

$$\mathrm{ELBO}(q, \theta) = \mathbb{E}_q[\log p(Y, Z \mid \theta)] - \mathbb{E}_q[\log q(Z)] = \mathbb{E}_q\left[\log \frac{p(Y, Z \mid \theta)}{q(Z)}\right].$$

Since $p(Y, Z \mid \theta) = p(Z \mid Y, \theta) p(Y \mid \theta)$:

$$\mathrm{ELBO}(q, \theta) = \mathbb{E}_q[\log p(Z \mid Y, \theta)] + \log p(Y \mid \theta) - \mathbb{E}_q[\log q(Z)].$$

Rearranging:

$$\log p(Y \mid \theta) = \mathrm{ELBO}(q, \theta) + \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z \mid Y, \theta)}\right] = \mathrm{ELBO}(q, \theta) + \mathrm{KL}(q \| p(Z \mid Y, \theta)). \quad \square$$

EM as Coordinate Ascent on the ELBO

The EM algorithm maximizes the ELBO by alternating:

E-step (optimize over q at fixed $\theta^{(t)}$): The ELBO is maximized when the KL gap is zero, i.e., $q^{(t)}(Z) = p(Z \mid Y, \theta^{(t)})$. At this q, the ELBO equals $\ell(\theta^{(t)}; Y)$.

M-step (optimize over θ at fixed $q^{(t)}$): Maximize $\mathrm{ELBO}(q^{(t)}, \theta) = \mathbb{E}_{q^{(t)}}[\log p(Y, Z \mid \theta)] + H(q^{(t)})$. Since $H(q^{(t)})$ does not depend on θ, this reduces to maximizing $Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{p(Z \mid Y, \theta^{(t)})}[\log p(Y, Z \mid \theta)]$.

Proof of Monotone Likelihood Ascent

Theorem. $\ell(\theta^{(t+1)}; Y) \geq \ell(\theta^{(t)}; Y)$ for every EM iteration.

Proof. After the E-step, $q^{(t)} = p(Z \mid Y, \theta^{(t)})$, so

$$\ell(\theta^{(t)}; Y) = \mathrm{ELBO}(q^{(t)}, \theta^{(t)}).$$

After the M-step, $\theta^{(t+1)}$ maximizes the ELBO over θ, so

$$\mathrm{ELBO}(q^{(t)}, \theta^{(t+1)}) \geq \mathrm{ELBO}(q^{(t)}, \theta^{(t)}) = \ell(\theta^{(t)}; Y).$$

Since the ELBO is always a lower bound on the log-likelihood:

$$\ell(\theta^{(t+1)}; Y) \geq \mathrm{ELBO}(q^{(t)}, \theta^{(t+1)}) \geq \ell(\theta^{(t)}; Y). \quad \square$$

Convergence Rate

EM converges **linearly** near a local maximum. The rate matrix is

$$M = I_c(\theta^*)^{-1} I_m(\theta^*),$$

where $I_c(\theta^*)$ is the complete-data Fisher information and $I_m(\theta^*) = I_c(\theta^*) - I_{\mathrm{obs}}(\theta^*)$ is the missing information. The convergence rate is the largest eigenvalue of M:

$$\rho = \lambda_{\max}(M) = 1 - \frac{\lambda_{\min}(I_{\mathrm{obs}}(\theta^*))}{\lambda_{\max}(I_c(\theta^*))}.$$

High missingness ($\rho \to 1$) implies slow convergence. When the fraction of missing information is large, the E-step provides little new information and the parameter updates are small.

Convergence to Stationary Points

Under regularity conditions (compactness of Θ, continuity of Q, identifiability), the EM sequence $\{\theta^{(t)}\}$ converges to a stationary point of $\ell(\theta; Y)$. The monotone ascent property, combined with boundedness of ℓ, guarantees that $\ell(\theta^{(t)})$ converges; additional regularity ensures that the parameter sequence converges and the limit satisfies the score equation $\nabla \ell(\theta^*; Y) = 0$. Convergence to a local rather than global maximum is possible; multiple initializations are standard practice.

Worked Examples

Example 1: Gaussian Mixture Model

Let $Y_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$ with latent indicators $Z_i \in \{1, \ldots, K\}$.

E-step. Compute responsibilities:

$$r_{ik}^{(t)} = \frac{\pi_k^{(t)} \phi(Y_i; \mu_k^{(t)}, \sigma_k^{2(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \phi(Y_i; \mu_j^{(t)}, \sigma_j^{2(t)})}.$$

M-step. Update:

$$n_k = \sum_{i=1}^{n} r_{ik}^{(t)}, \quad \mu_k^{(t+1)} = \frac{\sum_i r_{ik}^{(t)} Y_i}{n_k}, \quad \sigma_k^{2(t+1)} = \frac{\sum_i r_{ik}^{(t)} (Y_i - \mu_k^{(t+1)})^2}{n_k}, \quad \pi_k^{(t+1)} = \frac{n_k}{n}.$$

For a mixture of $K = 2$ components with $n = 200$ observations, initializing with k-means, EM typically converges in 20–50 iterations. The log-likelihood increases monotonically, with the largest increases in the first few iterations.

Example 2: Censored Exponential Data

Let $Y_i \sim \text{Exp}(\lambda)$ with right-censoring at c. For censored observations, $Y_i = c$ and the true value $Z_i > c$ is unobserved. The complete-data log-likelihood is $\ell_c(\lambda) = n \log \lambda - \lambda \sum Z_i$.

E-step. For uncensored observations, $\mathbb{E}[Z_i \mid Y_i, \lambda^{(t)}] = Y_i$. For censored observations ($Y_i = c$),

$$\mathbb{E}[Z_i \mid Z_i > c, \lambda^{(t)}] = c + 1/\lambda^{(t)}.$$

M-step. $\lambda^{(t+1)} = n / \sum_i \tilde{Z}_i^{(t)}$, where $\tilde{Z}_i^{(t)}$ uses the imputed expectations.

With $n = 50$, $\lambda = 2$, and censoring at $c = 0.5$ (about 63% censored), EM converges in 10–15 iterations. The high censoring fraction corresponds to large missing information and slower convergence, consistent with the rate formula.

Cross-Links

- **Linear Algebra:** The M-step for Gaussian mixtures involves weighted least squares computations; the convergence rate matrix M is analyzed through its eigenvalue decomposition.
- **AI:** The EM structure underlies training of latent variable models including Gaussian mixture models for clustering, hidden Markov models for sequence labeling, and probabilistic PCA for dimensionality reduction. At the population level, VAE training optimizes the same ELBO that EM maximizes.
- **Economics:** EM is used in econometrics for models with missing data (e.g., sample selection models), regime-switching models (Hamilton filter), and finite mixture models for unobserved heterogeneity.
- **Quantum:** Quantum state tomography with missing measurement settings uses EM to estimate the density matrix; the latent variables correspond to the unobserved measurement outcomes.

Structural Compression

Invariant: Each EM iteration increases or maintains the observed log-likelihood; the algorithm converges to a stationary point of $\ell(\theta; Y)$.

Mechanism: The ELBO decomposition $\ell(\theta; Y) = \text{ELBO}(q, \theta) + \text{KL}(q \| p(Z \mid Y, \theta))$ shows that EM performs coordinate ascent: the E-step sets $q = p(Z \mid Y, \theta^{(t)})$ to close the KL gap, and the M-step maximizes the ELBO over θ. Since the ELBO is a lower bound that touches ℓ at the current iterate, each M-step guarantees ascent.

Cross-System Echo: EM is a special case of the majorize-maximize (MM) principle: construct a tight lower bound on the objective and maximize the bound. The same principle appears in variational inference (Node 9.6), where the bound is looser because the E-step uses an approximate rather than exact posterior, and in convex optimization, where proximal operators play the role of the M-step. The ELBO is the central object connecting EM, variational inference, and variational autoencoders across statistics and machine learning.

Exercises

1. Derive the ELBO decomposition $\ell(\theta; Y) = \text{ELBO}(q, \theta) + \text{KL}(q \| p(Z \mid Y, \theta))$ from first principles, without using Jensen's inequality, by direct algebraic manipulation.

2. For the two-component Gaussian mixture, write the complete E-step and M-step updates including the variance parameters σ_k^2. Verify that the M-step for π_k involves a Lagrange multiplier enforcing $\sum_k \pi_k = 1$.

3. Prove that the EM convergence rate for the normal mean with known variance and a fraction f of missing observations is $\rho = f$. Interpret this in terms of the missing information principle.

4. Show that Gibbs sampling (Node 9.3) applied to the joint posterior $p(\theta, Z \mid Y)$ produces an ergodic chain, while EM applied to the same model produces a deterministic sequence converging to a point estimate. Explain the structural difference.

5. Implement EM for a $K = 3$ Gaussian mixture on simulated data. Plot the log-likelihood as a function of iteration and verify monotonicity. Run from 5 different initializations and compare the local maxima found.

6. For censored exponential data, derive the E-step formula $\mathbb{E}[Z_i \mid Z_i > c, \lambda] = c + 1/\lambda$ from the memoryless property of the exponential distribution.

7. Argue, structurally, why the ELBO decomposition is the fundamental identity connecting EM, variational inference, and the marginal likelihood: in each case, the decomposition $\log p(Y) = \text{ELBO} + \text{KL}$ determines which quantity is being optimized (the ELBO), what the gap measures (approximation quality), and how the algorithm trades exactness for tractability.

Variational Inference

Position in Knowledge Graph

System: Statistics

Structural Domain: Computational Statistics

Node: 9.6

Variational inference converts the problem of computing an intractable posterior distribution into an optimization problem. Where MCMC (Node 9.3) produces asymptotically exact samples at high computational cost, VI produces a fast deterministic approximation whose quality depends on the expressiveness of the chosen variational family. This node develops the ELBO, the mean-field approximation, the CAVI algorithm, the directional asymmetry of KL divergence, and the extension to stochastic variational inference for large datasets.

Formal Definition

Let $\pi(\theta \mid x) = p(\theta \mid x)$ be a posterior distribution on $\Theta \subseteq \mathbb{R}^k$ that is intractable – either the normalizing constant $m(x) = \int p(x, \theta)\, d\theta$ cannot be computed or MCMC mixing is prohibitively slow.

Let $\mathcal{Q}$ be a family of tractable distributions on Θ. **Variational inference** seeks

$$q^* = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\theta) \| p(\theta \mid x)).$$

Since $\mathrm{KL}(q \| p(\cdot \mid x))$ involves the unknown normalizing constant $m(x)$, VI instead maximizes the **Evidence Lower BOund (ELBO)**:

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x, \theta)] - \mathbb{E}_q[\log q(\theta)] = \mathbb{E}_q[\log p(x, \theta)] + H(q),$$

where $H(q) = -\mathbb{E}_q[\log q(\theta)]$ is the entropy of q.

Intuition

Bayesian inference requires integrating over the entire parameter space. MCMC does this by simulating a random walk and averaging. Variational inference replaces the integral with a search: find the distribution in a tractable family that is closest to the true posterior. "Closest" is measured by KL divergence, and the ELBO provides a computable objective that avoids the normalizing constant. The approximation quality depends entirely on whether the variational family $\mathcal{Q}$ can capture the shape of the true posterior – if the posterior is multimodal and $\mathcal{Q}$ contains only unimodal distributions, the approximation will miss important structure.

Structural Role

Variational inference trades exactness for scalability. It depends on the ELBO framework established in Node 9.5 (EM) and serves as the scalable alternative to MCMC (Node 9.3) for large datasets and complex models. EM is the special case where the variational family is unrestricted (the E-step uses the exact posterior over latent variables). Variational inference is the foundation of variational autoencoders (VAEs) and amortized inference in modern machine learning.

Mathematical Development

The ELBO Decomposition

For any $q(\theta)$,

$$\log m(x) = \mathcal{L}(q) + \mathrm{KL}(q(\theta) \| p(\theta \mid x)).$$

Proof. Expanding the KL divergence:

$$\mathrm{KL}(q \| p(\cdot \mid x)) = \mathbb{E}_q\left[\log \frac{q(\theta)}{p(\theta \mid x)}\right] = \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta \mid x)].$$

Since $p(\theta \mid x) = p(x, \theta)/m(x)$:

$$\mathrm{KL}(q \| p(\cdot \mid x)) = \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(x, \theta)] + \log m(x).$$

Rearranging:

$$\log m(x) = \underbrace{\mathbb{E}_q[\log p(x, \theta)] - \mathbb{E}_q[\log q(\theta)]}_{\mathcal{L}(q)} + \mathrm{KL}(q \| p(\cdot \mid x)). \quad \square$$

Since $\mathrm{KL} \geq 0$, $\mathcal{L}(q) \leq \log m(x)$, with equality iff $q = p(\theta \mid x)$. Thus maximizing the ELBO over $\mathcal{Q}$ is equivalent to minimizing $\mathrm{KL}(q \| p(\cdot \mid x))$ over $\mathcal{Q}$.

Alternative Form of the ELBO

The ELBO can be decomposed as

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(x \mid \theta)] - \mathrm{KL}(q(\theta) \| p(\theta)),$$

where $p(\theta)$ is the prior. This form reveals the tradeoff: the first term rewards q for placing mass where the likelihood is large; the second term penalizes q for deviating from the prior.

Mean-Field Approximation

The **mean-field family** assumes full factorization:

$$\mathcal{Q}_{\mathrm{MF}} = \left\{ q(\theta) = \prod_{j=1}^{k} q_j(\theta_j) \right\}.$$

Theorem (Optimal Mean-Field Update). Fixing all factors q_i for $i \neq j$, the optimal q_j satisfies

$$\log q_j^*(\theta_j) = \mathbb{E}_{q_{-j}}[\log p(x, \theta)] + \text{const},$$

where $\mathbb{E}_{q_{-j}}$ denotes expectation over all components except j, and the constant ensures q_j^* normalizes to 1.

Proof. Write the ELBO as

$$\mathcal{L}(q) = \int q_j(\theta_j) \left[\int \prod_{i \neq j} q_i(\theta_i) \log p(x, \theta) \, d\theta_{-j} \right] d\theta_j - \int q_j(\theta_j) \log q_j(\theta_j) \, d\theta_j + C,$$

where C collects terms not depending on q_j. Define $f_j(\theta_j) = \mathbb{E}_{q_{-j}}[\log p(x, \theta)]$. Then

$$\mathcal{L}(q) = \int q_j(\theta_j) f_j(\theta_j)\, d\theta_j - \int q_j(\theta_j) \log q_j(\theta_j)\, d\theta_j + C.$$

This is maximized when $q_j(\theta_j) \propto \exp(f_j(\theta_j))$, since the functional is $-\mathrm{KL}(q_j \| \tilde{q}_j) + \text{const}$ where $\tilde{q}_j \propto \exp(f_j)$. □

Coordinate Ascent Variational Inference (CAVI)

CAVI cycles through coordinates $j = 1, \ldots, k$, updating each q_j according to the optimal update while holding q_{-j} fixed. Each update increases (or maintains) the ELBO, guaranteeing convergence to a local optimum. The algorithm terminates when the relative change in ELBO falls below a threshold.

For exponential family models with conjugate priors, each CAVI update yields a distribution in the same exponential family, and the update reduces to updating the natural parameters:

$$\eta_j^{(t+1)} = \eta_j^{\text{prior}} + \mathbb{E}_{q_{-j}^{(t)}}[t_j(x, \theta_{-j})],$$

where η_j^{prior} are the prior natural parameters and t_j are the sufficient statistics contributed by factor j.

KL Divergence Asymmetry

The choice of KL direction has profound consequences:

KL$(q\|p)$ **(variational inference):** This is **mode-seeking** (or zero-avoiding). Where $p(\theta \mid x) \approx 0$, q must also be approximately 0 (otherwise the KL divergence diverges). Consequently, q tends to concentrate on a single mode of p and underestimates posterior variance.

KL$(p\|q)$ **(expectation propagation):** This is **mean-seeking** (or mass-covering). Where $p(\theta \mid x) > 0$, q must also be positive. Consequently, q spreads to cover all modes, typically overestimating variance.

For a bimodal posterior with modes at $\pm\mu$: - KL$(q\|p)$-optimal Gaussian: concentrates on one mode, $q^* \approx \mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(-\mu, \sigma^2)$. - KL$(p\|q)$-optimal Gaussian: centers at 0 with large variance, $q^* \approx \mathcal{N}(0, \mu^2 + \sigma^2)$.

Relation to EM

EM (Node 9.5) is variational inference with $\mathcal{Q} = \{p(Z \mid Y, \theta) : \theta \in \Theta\}$, the family of exact posteriors over latent variables parameterized by θ. The E-step sets $q(Z) = p(Z \mid Y, \theta^{(t)})$ (exact posterior, KL gap = 0), and the M-step optimizes θ. Variational EM replaces the exact E-step with an approximate one: $q(Z) \in \mathcal{Q}_{\text{approx}}$, enabling inference when the E-step is itself intractable.

Stochastic Variational Inference

For large datasets $x = (x_1, \ldots, x_n)$, CAVI requires a full pass through the data at each update. **Stochastic variational inference (SVI)** exploits the fact that the ELBO decomposes as a sum over data points (for conditionally i.i.d. models):

$$\mathcal{L}(q) = \sum_{i=1}^{n} \mathbb{E}_q[\log p(x_i \mid \theta)] - \mathrm{KL}(q(\theta) \| p(\theta)).$$

SVI replaces the full-data gradient with a noisy mini-batch estimate:

$$\hat{\nabla}\mathcal{L} = \frac{n}{|S|} \sum_{i \in S} \nabla_\lambda \mathbb{E}_q[\log p(x_i \mid \theta)] - \nabla_\lambda \mathrm{KL}(q_\lambda \| p),$$

where S is a random mini-batch and λ parameterizes q. With a Robbins–Monro step size schedule ($\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$), SVI converges to a local optimum of the ELBO.

Worked Examples

Example 1: Mean-Field VI for Normal-Normal Model

Let $x_1, \ldots, x_n \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^{-1})$ with priors $\mu \sim \mathcal{N}(0, \sigma_0^2)$ and $\tau \sim \text{Gamma}(a_0, b_0)$. The mean-field approximation is $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$.

Update for q_μ:

$$\log q_\mu^*(\mu) = \mathbb{E}_{q_\tau}\left[-\frac{\mathbb{E}[\tau]}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\mu^2}{2\sigma_0^2} \right] + \text{const.}$$

This is a quadratic in μ, so $q_\mu^* = \mathcal{N}(\mu_q, \sigma_q^2)$ with

$$\sigma_q^2 = (n\mathbb{E}_{q_\tau}[\tau] + 1/\sigma_0^2)^{-1}, \quad \mu_q = \sigma_q^2 \cdot n\mathbb{E}_{q_\tau}[\tau]\bar{x}.$$

Update for q_τ:

$$q_\tau^* = \text{Gamma}\left(a_0 + n/2,\ b_0 + \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] \right).$$

CAVI alternates these updates until the ELBO converges. With $n = 100$, $\bar{x} = 2.1$, convergence is typically achieved in 10–20 iterations. The mean-field approximation underestimates the posterior covariance between μ and τ (which is zero in the factorized approximation but nonzero in the true posterior).

Example 2: KL Direction Comparison

Consider a posterior $p(\theta \mid x) = 0.5\,\mathcal{N}(-3, 0.5^2) + 0.5\,\mathcal{N}(3, 0.5^2)$ and restrict $\mathcal{Q}$ to Gaussians $q = \mathcal{N}(m, s^2)$.

Minimizing $\text{KL}(q\|p)$: the optimizer selects one mode, yielding $q^* \approx \mathcal{N}(3, 0.5^2)$ (or symmetrically $\mathcal{N}(-3, 0.5^2)$). The KL divergence penalizes placing mass where p is near zero (the region between modes), so q collapses to a single mode.

Minimizing $\text{KL}(p\|q)$: the optimizer must cover both modes, yielding $q^* \approx \mathcal{N}(0, 9.25)$. This captures both modes but vastly overestimates the spread, assigning substantial mass to the region between modes where p is nearly zero.

Cross-Links

- **Linear Algebra:** The CAVI updates for Gaussian mean-field approximations reduce to solving linear systems; the natural gradient of the ELBO with respect to the natural parameters of an exponential family q is the Euclidean gradient in the natural parameterization.
- **AI:** Variational autoencoders (VAEs) use amortized variational inference: an encoder network maps each data point x_i to variational parameters λ_i, optimizing the ELBO jointly over the encoder and decoder. The reparameterization trick enables backpropagation through the sampling step.
- **Economics:** Variational Bayes is increasingly used in macroeconometrics for approximating posterior distributions of high-dimensional VAR models where MCMC is too slow for real-time forecasting.
- **Quantum:** Variational quantum algorithms (VQE, QAOA) optimize a parameterized quantum circuit to minimize an energy functional, mirroring the VI principle of optimizing over a tractable family to approximate an intractable distribution.

Structural Compression

Invariant: $\log m(x) = \mathcal{L}(q) + \mathrm{KL}(q \| p(\theta \mid x))$: maximizing the ELBO over $\mathcal{Q}$ minimizes the KL divergence to the posterior, with the gap measuring the approximation quality.

Mechanism: Jensen's inequality applied to the log marginal likelihood yields the ELBO as a lower bound. The mean-field factorization restricts $\mathcal{Q}$ to product distributions, enabling coordinate-wise closed-form updates (CAVI). Each CAVI step increases the ELBO, converging to a local optimum. The mode-seeking direction $\mathrm{KL}(q\|p)$ produces compact approximations that may miss posterior mass, trading coverage for computational tractability.

Cross-System Echo: The ELBO decomposition is the master identity of approximate inference. In EM, the KL gap is zero by construction (exact E-step). In VI, the gap is nonzero but controlled. In VAEs, the ELBO is the training objective, with the encoder parameterizing q and the decoder parameterizing $p(x \mid \theta)$. In information theory, the ELBO is the negative description length under code q; minimizing KL is equivalent to minimizing the excess bits required to describe the data. The pattern – bounding an intractable quantity by a tractable one and optimizing the bound – is the universal computational strategy for problems where exact solutions are infeasible.

Exercises

1. Derive the ELBO decomposition $\log m(x) = \mathcal{L}(q) + \mathrm{KL}(q\|p(\theta \mid x))$ and verify that the ELBO equals the log marginal likelihood when $q = p(\theta \mid x)$.
2. For the normal-normal model with known variance, derive the mean-field CAVI updates for q_μ and q_τ and show that each update is a closed-form exponential family distribution.
3. Prove that CAVI monotonically increases the ELBO at each coordinate update. State precisely what regularity conditions are needed.
4. For a bimodal target $p = 0.5\mathcal{N}(-\mu, 1) + 0.5\mathcal{N}(\mu, 1)$, compute the $\mathrm{KL}(q\|p)$-optimal and $\mathrm{KL}(p\|q)$-optimal Gaussian approximations analytically as functions of μ. Plot both as μ ranges from 0 to 5.
5. Show that EM is a special case of variational inference where the E-step uses the unrestricted family $\mathcal{Q} = \{q(Z) : q \text{ is a density}\}$, so the KL gap is zero after the E-step.
6. For a model with $n = 10{,}000$ observations, estimate the per-iteration cost of CAVI versus SVI with mini-batch size 100. Explain the tradeoff between noise in the gradient and computational savings.
7. Argue, structurally, why the ELBO is the unifying object across EM, variational inference, and variational autoencoders: in each case, the same decomposition $\log p(x) = \mathrm{ELBO} + \mathrm{KL}$ determines the objective, the approximation gap, and the algorithm, with the only difference being the expressiveness of the variational family and whether the parameters θ are integrated out or optimized.

Module 10 — Model Selection Learning

Bias–Variance Tradeoff

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.1.

The bias–variance tradeoff is the foundational decomposition underlying all model selection procedures. It quantifies the tension between underfitting (high bias) and overfitting (high variance), establishing that no estimator can simultaneously minimize both for finite sample size. Every criterion in this module—AIC, BIC, cross-validation, regularization—is a procedure for navigating this tradeoff.

Formal Definition

Let $\hat{f}_n : \mathcal{X} \to \mathbb{R}$ be an estimator of a regression function f_0, fitted from data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ where $Y_i = f_0(X_i) + \varepsilon_i$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, and $\varepsilon_i \perp\!\!\!\perp \mathcal{D}_n$.

The **prediction risk** (conditional on x) is

$$R(\hat{f}_n, x) = \mathbb{E}_{\mathcal{D}_n, \varepsilon}\big[(Y - \hat{f}_n(x))^2\big].$$

Bias–Variance Decomposition. The prediction risk decomposes as

$$R(\hat{f}_n, x) = \underbrace{\big(\mathbb{E}_{\mathcal{D}_n}[\hat{f}_n(x)] - f_0(x)\big)^2}_{\mathrm{Bias}^2(\hat{f}_n(x))} + \underbrace{\mathrm{Var}_{\mathcal{D}_n}(\hat{f}_n(x))}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Irreducible noise}}.$$

The integrated prediction risk is $R(\hat{f}_n) = \mathbb{E}_X[R(\hat{f}_n, X)]$.

Intuition

Every estimator makes two kinds of errors. Bias is systematic: the estimator consistently misses the truth because the model class is too restrictive. Variance is random: the estimator fluctuates across different training samples because the model class is too flexible. The irreducible noise σ^2 is the floor—no estimator can beat it.

Simple models (e.g., linear regression with few features) have high bias and low variance. Complex models (e.g., high-degree polynomials, deep trees) have low bias and high variance. The optimal model complexity is the one that minimizes their sum, and this optimum shifts rightward as n grows.

Structural Role

The bias–variance tradeoff is the fundamental organizing principle of model selection. It determines the optimal model complexity as a function of sample size n: as $n \to \infty$, variance decreases at rate $O(1/n)$, allowing progressively richer models.

Dependencies: Relies on the mean squared error framework (Node 3.1) and the asymptotic theory of estimators (Node 7.3). **Enables:** All subsequent model selection criteria—AIC (Node 10.2), cross-validation (Node 10.3), regularization (Node 10.4)—are procedures for navigating this decomposition.

Mathematical Development

Full Proof of the Decomposition

Write $Y = f_0(x) + \varepsilon$ and let $\bar{f}(x) = \mathbb{E}_{\mathcal{D}_n}[\hat{f}_n(x)]$. Then

$$\mathbb{E}[(Y - \hat{f}_n(x))^2] = \mathbb{E}[(f_0(x) + \varepsilon - \hat{f}_n(x))^2].$$

Add and subtract $\bar{f}(x)$:

$$= \mathbb{E}\big[(\varepsilon + f_0(x) - \bar{f}(x) + \bar{f}(x) - \hat{f}_n(x))^2\big].$$

Let $A = \varepsilon$, $B = f_0(x) - \bar{f}(x)$, $C = \bar{f}(x) - \hat{f}_n(x)$. Then

$$\mathbb{E}[(A + B + C)^2] = \mathbb{E}[A^2] + B^2 + \mathbb{E}[C^2] + 2B\,\mathbb{E}[C] + 2\,\mathbb{E}[AC] + 2B\,\mathbb{E}[A].$$

Now: $\mathbb{E}[A] = \mathbb{E}[\varepsilon] = 0$, so the last term vanishes. $\mathbb{E}[C] = \bar{f}(x) - \mathbb{E}[\hat{f}_n(x)] = 0$, so the $2B\,\mathbb{E}[C]$ term vanishes. $\mathbb{E}[AC] = \mathbb{E}[\varepsilon(\bar{f}(x) - \hat{f}_n(x))] = 0$ by independence of ε from $\mathcal{D}_n$. Therefore

$$R(\hat{f}_n, x) = \sigma^2 + (f_0(x) - \bar{f}(x))^2 + \mathbb{E}[(\hat{f}_n(x) - \bar{f}(x))^2] = \sigma^2 + \text{Bias}^2 + \text{Variance}.$$

Parametric MSE Decomposition

For a parametric estimator $\hat{\theta}_n$ of $\theta_0 \in \mathbb{R}^k$,

$$\text{MSE}(\hat{\theta}_n) = \|\text{Bias}(\hat{\theta}_n)\|^2 + \text{Tr}\,\text{Cov}(\hat{\theta}_n).$$

The Cramer–Rao bound (Node 3.4) gives $\text{Tr}\,\text{Cov}(\hat{\theta}_n) \geq \text{Tr}\, I(\theta_0)^{-1}$ for unbiased estimators. A biased estimator may achieve lower total MSE when the bias term is small relative to the variance reduction. Ridge regression (Node 10.4) exploits this: it introduces bias $O(\lambda)$ while reducing variance by $O(\lambda^2)$ near the optimal λ.

Approximation vs. Estimation Error

The total excess risk decomposes as

$$R(\hat{f}_n) - R(f_0) = \underbrace{R(f^*_{\mathcal{F}}) - R(f_0)}_{\text{Approximation error}} + \underbrace{R(\hat{f}_n) - R(f^*_{\mathcal{F}})}_{\text{Estimation error}},$$

where $f^*_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} R(f)$. Approximation error decreases as $\mathcal{F}$ grows (richer model classes); estimation error increases with the complexity of $\mathcal{F}$ and decreases with n. For a model class with VC dimension d, classical bounds give estimation error $O(\sqrt{d/n})$.

Double Descent and Interpolation

In overparameterized models ($p > n$), the classical U-shaped risk curve breaks down. The **double descent** phenomenon: as p increases past n, the minimum-norm interpolating estimator (which achieves zero training error) can have decreasing risk. The bias–variance decomposition still holds, but the variance term behaves non-monotonically near $p = n$ (the interpolation threshold) where the condition number of $X^\top X$ diverges. For $p \gg n$, implicit regularization from the minimum-norm constraint reduces variance, producing a second descent in risk.

This does not violate the bias–variance tradeoff; rather, it reveals that the effective complexity of an estimator is not simply the number of parameters but depends on the geometry of the estimator class and the implicit constraints of the optimization procedure.

Optimism and Degrees of Freedom

For a linear smoother $\hat{Y} = HY$ with hat matrix H, the **optimism** (gap between in-sample and out-of-sample risk) is

$$\text{Optimism} = \frac{2\sigma^2}{n}\text{Tr}(H) = \frac{2\sigma^2}{n}\,\text{df},$$

where $\text{df} = \text{Tr}(H)$ is the **effective degrees of freedom**. This connects the bias–variance tradeoff to model selection: AIC (Node 10.2) corrects for exactly this optimism, and Mallows' C_p uses df to penalize complex models.

For OLS, $\text{df} = p$ (number of parameters). For ridge regression, $\text{df}(\lambda) = \sum_j d_j^2/(d_j^2 + \lambda) < p$. For k-NN, $\text{df} = n/k$.

Rates for Specific Estimators

k-Nearest Neighbors. For k-NN regression in $\mathbb{R}^d$ with Lipschitz f_0:

$$\text{Bias}^2 = O(k^{-2/d} \cdot n^{-2/d}), \quad \text{Variance} = O(\sigma^2/k).$$

Optimizing over k gives $k^* \sim n^{2/(2+d)}$ and risk $R - \sigma^2 = O(n^{-2/(2+d)})$, exhibiting the curse of dimensionality.

Polynomial Regression of Degree p. For a degree-p polynomial fit to data from a smooth function on $[0,1]$ with m continuous derivatives:

$$\text{Bias}^2 = O(n^{-2m/(2m+1)}) \text{ (when } p \sim n^{1/(2m+1)}), \quad \text{Variance} = O(p/n).$$

The optimal degree balances the two at the minimax rate $n^{-2m/(2m+1)}$.

Worked Examples

Example 1: Polynomial Regression

Consider $Y_i = \sin(X_i) + \varepsilon_i$ with $X_i \sim \text{Uniform}(0, 2\pi)$, $\varepsilon_i \sim \mathcal{N}(0, 0.25)$, and $n = 50$. Fit polynomials of degree $p = 1, 3, 10, 25$.

Degree 1 (linear): $\bar{f}(x) \approx a + bx$. Since $\sin(x)$ is nonlinear, Bias^2 is large. Two parameters yield very low variance. The fit systematically misses the curvature.

Degree 3: Captures the dominant shape of $\sin(x)$ well. Bias is moderate (the Taylor expansion $\sin(x) \approx x - x^3/6$ is a degree-3 polynomial). Variance with 4 parameters on 50 points is still small.

Degree 10: Bias is near zero (degree-10 polynomial can approximate sin on a compact interval to high accuracy). Variance increases: 11 parameters on 50 points. Moderate overfitting at the boundaries.

Degree 25: Bias is essentially zero. Variance is very large: 26 parameters on 50 points means the fit interpolates noise. Test MSE is much worse than degree 3.

Computing $\widehat{\text{MSE}}$ on an independent test set of size 1000: degree 1 gives ≈ 0.35, degree 3 gives ≈ 0.27, degree 10 gives ≈ 0.30, degree 25 gives ≈ 1.2. The minimum is at an intermediate complexity.

Example 2: k-NN in Dimension $d = 1$

Data: $Y_i = X_i^2 + \varepsilon_i$, $X_i \sim \text{Uniform}(0, 1)$, $\sigma^2 = 0.1$, $n = 200$. Evaluate the prediction at $x = 0.5$.

$k = 1$**:** The predictor is Y_{j^*} where $j^* = \arg\min_j |X_j - 0.5|$. Bias ≈ 0 (nearest neighbor is close to x). Variance $= \sigma^2 = 0.1$.

$k = 50$**:** The predictor averages 50 neighbors. If these span roughly $[0.35, 0.65]$, then $\bar{f}(0.5) \approx \mathbb{E}[X^2 \mid X \in [0.35, 0.65]] \approx 0.253$ versus $f_0(0.5) = 0.25$. Bias ≈ 0.003. Variance $\approx \sigma^2/50 = 0.002$. MSE ≈ 0.002.

$k = 200$**:** All data used. $\bar{f}(0.5) = \mathbb{E}[X^2] = 1/3 \approx 0.333$ versus $f_0(0.5) = 0.25$. Bias$^2 \approx 0.007$. Variance $\approx \sigma^2/200 = 0.0005$. MSE ≈ 0.007.

The optimal k is intermediate: large enough to average away noise, small enough to avoid smoothing bias. Theory gives $k^* \sim n^{2/3} \approx 34$ for $d = 1$.

Cross-Links

AI. The bias–variance tradeoff motivates regularization (weight decay, dropout, early stopping) and architecture search in deep learning. The double descent phenomenon in overparameterized models challenges the classical U-shaped risk curve but can be explained through implicit regularization effects.

Economics. In forecasting and structural estimation, the tradeoff between model parsimony (few parameters, interpretable) and fit (many parameters, flexible) governs model specification. Shrinkage estimators in empirical Bayes (James–Stein) dominate the MLE in dimension ≥ 3 precisely because they exploit this tradeoff.

Linear Algebra. The singular value decomposition of the design matrix $X = U\Sigma V^\top$ provides the geometric view: ridge regression shrinks the components along small singular values (where variance is large) proportionally more, and the bias–variance tradeoff is controlled by the spectrum of $X^\top X$.

Quantum Mechanics. The uncertainty principle $\Delta x \cdot \Delta p \geq \hbar/2$ is structurally analogous: simultaneous localization in two conjugate domains is impossible. In quantum state tomography, the bias–variance tradeoff governs the choice of measurement basis and estimator complexity.

Structural Compression

Invariant

MSE $=$ Bias^2 + Variance $+ \sigma^2$: the three components are orthogonal in the L^2 decomposition of prediction error.

Mechanism

Second-moment expansion of the squared error; all cross terms vanish by independence of the noise ε from the training-data-dependent estimator $\hat{f}_n$ and by the definition of the bias as $\mathbb{E}[\hat{f}_n(x)] - f_0(x)$.

Cross-System Echo

The bias–variance tradeoff is the statistical instance of the approximation-generalization tradeoff in statistical learning theory (Vapnik). In AI, it motivates regularization, dropout, and early stopping as variance-reduction techniques. In economics, the tradeoff between model parsimony and fit appears in forecasting and structural estimation. In physics, the uncertainty principle is the conjugate-variable analogue: precision in one domain costs precision in the other.

Exercises

1. Prove the bias–variance decomposition in the multivariate case: for $\hat{\theta}_n \in \mathbb{R}^k$, show that $\mathbb{E}[\|\hat{\theta}_n - \theta_0\|^2] = \|\mathrm{Bias}(\hat{\theta}_n)\|^2 + \mathrm{Tr}\,\mathrm{Cov}(\hat{\theta}_n)$.
2. For the k-NN estimator in dimension d, derive the rates $\mathrm{Bias}^2 = O(k/n)^{2/d}$ and $\mathrm{Variance} = O(\sigma^2/k)$ assuming f_0 is Lipschitz and X has a bounded density.
3. Consider the James–Stein estimator $\hat{\theta}^{\mathrm{JS}} = (1 - (k-2)\sigma^2/\|\bar{X}\|^2)\bar{X}$ for $\theta \in \mathbb{R}^k$, $k \geq 3$. Show that its MSE is strictly less than the MLE's MSE $k\sigma^2/n$ for every θ, and explain this in terms of the bias–variance tradeoff.
4. For a linear smoother $\hat{Y} = HY$, show that the in-sample optimism is $\mathrm{Optimism} = (2\sigma^2/n)\,\mathrm{Tr}(H)$, where $\mathrm{Tr}(H)$ is the effective degrees of freedom.
5. Suppose you observe $n = 100$ points from $Y = \sin(2\pi X) + \varepsilon$ with $\sigma = 0.5$. Compute the bias, variance, and MSE at $x = 0.5$ for the constant estimator $\hat{f}(x) = \bar{Y}$ and for the local average over the 10 nearest neighbors.
6. In ridge regression with design matrix X having singular values $d_1 \geq \cdots \geq d_p$, express both the bias and variance in terms of $\{d_j\}$, λ, and the true coefficients β_0. Show that the variance is $\sigma^2 \sum_j d_j^2/(d_j^2 + \lambda)^2$.
7. Argue, structurally, why the bias–variance decomposition must hold for any loss function admitting a second-moment expansion, and identify conditions under which it fails (e.g., non-squared losses, dependent noise).

Information Criteria: AIC and BIC

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.2.

Information criteria convert the bias–variance tradeoff (Node 10.1) into explicit model-ranking formulas by penalizing the maximized log-likelihood for the number of free parameters. AIC targets predictive accuracy via KL divergence minimization; BIC approximates the Bayesian marginal likelihood. Together they define the two canonical analytic approaches to model selection, feeding into model averaging (Node 10.6).

Formal Definition

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a collection of candidate parametric models. Model $\mathcal{M}_j$ has k_j free parameters and MLE $\hat{\theta}^{(j)}$ with maximized log-likelihood $\ell_j = \ell(\hat{\theta}^{(j)}; x)$.

Akaike Information Criterion:

$$\mathrm{AIC}_j = -2\ell_j + 2k_j.$$

Bayesian Information Criterion:

$$\mathrm{BIC}_j = -2\ell_j + k_j \log n.$$

The preferred model minimizes the criterion. Both penalize the log-likelihood for model complexity, differing only in the penalty weight per parameter.

Intuition

The maximized log-likelihood ℓ_j always improves as parameters are added, but this improvement partly reflects fitting noise rather than signal. Information criteria correct for this optimism by adding a penalty that grows with the number of parameters. AIC adds a fixed penalty of 2 per parameter, derived from the expected optimism of in-sample log-likelihood. BIC adds a penalty of $\log n$ per parameter, derived from Bayesian model comparison. For $n \geq 8$, BIC penalizes more heavily and selects simpler models.

The choice between them reflects the inferential goal: AIC for prediction, BIC for identifying the true model.

Structural Role

Dependencies: Builds on the bias–variance tradeoff (Node 10.1), the KL divergence (Node 2.3), and asymptotic likelihood theory (Node 7.6). **Enables:** Model averaging (Node 10.6), where BIC approximations to marginal likelihoods provide the posterior model weights.

AIC and BIC operationalize two philosophically distinct goals. AIC is the frequentist criterion: it targets the model that best predicts new data from the same distribution. BIC is the Bayesian criterion: it targets the model that most likely generated the observed data. Neither dominates the other; the appropriate choice depends on context.

Mathematical Development

Derivation of AIC from KL Divergence

Let p_0 be the true density and p_θ the model density. The KL divergence from p_0 to p_θ is

$$D_{\mathrm{KL}}(p_0 \| p_\theta) = \mathbb{E}_{p_0}\left[\log \frac{p_0(X)}{p_\theta(X)}\right] = \text{const} - \mathbb{E}_{p_0}[\log p_\theta(X)].$$

Minimizing KL divergence over θ is equivalent to maximizing $\mathbb{E}_{p_0}[\log p_\theta(X)]$, the expected log-likelihood under new data. Define

$$\Delta_j = \mathbb{E}_{X^{\mathrm{new}}}[\log p_{\hat{\theta}^{(j)}}(X^{\mathrm{new}})] - \frac{1}{n}\ell_j.$$

This is the **optimism**: how much the in-sample log-likelihood overestimates the out-of-sample log-likelihood. Akaike's key result:

Theorem (Akaike, 1973). Under regularity conditions (model $\mathcal{M}_j$ correctly specified, $\hat{\theta}^{(j)}$ consistent and asymptotically normal),

$$\mathbb{E}[\Delta_j] = -\frac{k_j}{n} + o(1/n).$$

Proof sketch. Expand $\ell(\hat{\theta}^{(j)}; X^{\mathrm{new}})$ around $\theta_0^{(j)}$ to second order. The cross term between the estimation error $\hat{\theta}^{(j)} - \theta_0^{(j)}$ and the score at X^{new} has mean zero. The quadratic term contributes $-\frac{1}{2}(\hat{\theta}^{(j)} - \theta_0^{(j)})^\top I(\theta_0^{(j)})(\hat{\theta}^{(j)} - \theta_0^{(j)})$, which has expectation $-k_j/(2n)$ by the asymptotic distribution of the MLE. Meanwhile, the in-sample quadratic term contributes $+k_j/(2n)$. The total optimism is $-k_j/n$, so

$$\mathbb{E}_{X^{\mathrm{new}}}\left[\frac{1}{n}\ell(\hat{\theta}^{(j)}; X^{\mathrm{new}})\right] \approx \frac{1}{n}\ell_j - \frac{k_j}{n}.$$

Multiplying by $-2n$ gives

$$\mathbb{E}[-2\ell(\hat{\theta}^{(j)}; X^{\mathrm{new}})] \approx -2\ell_j + 2k_j = \mathrm{AIC}_j.$$

Derivation of BIC from Laplace Approximation

The marginal likelihood of model $\mathcal{M}_j$ under prior $\pi(\theta^{(j)})$ is

$$m_j(x) = \int \exp(\ell(\theta^{(j)}; x))\, \pi(\theta^{(j)})\, d\theta^{(j)}.$$

Apply the Laplace approximation: expand $\ell(\theta; x)$ around $\hat{\theta}^{(j)}$ to second order,

$$\ell(\theta; x) \approx \ell_j - \frac{1}{2}(\theta - \hat{\theta}^{(j)})^\top \hat{I}_n (\theta - \hat{\theta}^{(j)}),$$

where $\hat{I}_n = -\nabla^2 \ell(\hat{\theta}^{(j)}; x)$. The Gaussian integral gives

$$m_j(x) \approx \exp(\ell_j)\, (2\pi)^{k_j/2}\, |\hat{I}_n|^{-1/2}\, \pi(\hat{\theta}^{(j)}).$$

Taking $-2\log$:

$$-2\log m_j(x) \approx -2\ell_j + \log|\hat{I}_n| - k_j\log(2\pi) - 2\log\pi(\hat{\theta}^{(j)}).$$

Under standard asymptotics, $\hat{I}_n = n\cdot\hat{I}_1 + O(\sqrt{n})$, so $\log|\hat{I}_n| = k_j\log n + O(1)$. All remaining terms are $O(1)$ as $n\to\infty$. Therefore

$$-2\log m_j(x) = -2\ell_j + k_j\log n + O(1) = \mathrm{BIC}_j + O(1).$$

Properties: Consistency and Efficiency

BIC Consistency. If the true model $\mathcal{M}^*$ is among the candidates, then for any overfitting model $\mathcal{M}_j \supsetneq \mathcal{M}^*$,

$$\mathrm{BIC}_j - \mathrm{BIC}^* = \underbrace{(-2\ell_j + 2\ell^*)}_{O_p(1)} + \underbrace{(k_j - k^*)\log n}_{\to\infty} \to +\infty.$$

So BIC selects the true model with probability $\to 1$. AIC does not: its penalty $2(k_j - k^*)$ is finite, so it selects the overfitting model with non-vanishing probability.

AIC Efficiency. When no candidate model is true (all are approximations), AIC asymptotically selects the model minimizing KL divergence to p_0. BIC may select a model that is too simple.

Corrected AIC

For small n relative to k_j, the $O(1/n)$ approximation in Akaike's theorem is inaccurate. The corrected AIC is

$$\mathrm{AICc}_j = -2\ell_j + 2k_j\cdot\frac{n}{n-k_j-1}.$$

This is exact for Gaussian linear models; for other models it is an improved approximation. As $n/k_j\to\infty$, $\mathrm{AICc}_j\to\mathrm{AIC}_j$.

Worked Examples

Example 1: Polynomial Regression Model Selection

Data: $n = 30$ observations from $Y = 2 + 3X - X^2 + \varepsilon$, $\varepsilon\sim\mathcal{N}(0,1)$, $X\sim\mathrm{Uniform}(-2,2)$. Fit polynomials of degree 1, 2, 3, 5.

Each model has $k_j = j+2$ parameters ($j+1$ coefficients plus σ^2). Suppose the maximized log-likelihoods are $\ell_1 = -48.2$, $\ell_2 = -39.1$, $\ell_3 = -38.9$, $\ell_5 = -38.4$.

AIC: $\mathrm{AIC}_1 = 96.4 + 6 = 102.4$; $\mathrm{AIC}_2 = 78.2 + 8 = 86.2$; $\mathrm{AIC}_3 = 77.8 + 10 = 87.8$; $\mathrm{AIC}_5 = 76.8 + 14 = 90.8$. AIC selects degree 2.

BIC ($\log 30 = 3.40$): $\mathrm{BIC}_1 = 96.4 + 10.2 = 106.6$; $\mathrm{BIC}_2 = 78.2 + 13.6 = 91.8$; $\mathrm{BIC}_3 = 77.8 + 17.0 = 94.8$; $\mathrm{BIC}_5 = 76.8 + 23.8 = 100.6$. BIC also selects degree 2.

Both criteria correctly identify the true model order. The degree-3 model gains only $\Delta\ell = 0.2$ over degree 2, insufficient to offset the additional parameter penalty.

Example 2: AIC vs. BIC Disagreement

Data: $n = 500$, true model has 5 predictors. Model A: 5 predictors, $\ell_A = -700.0$, $k_A = 6$. Model B: 8 predictors, $\ell_B = -696.5$, $k_B = 9$.

AIC: $\text{AIC}_A = 1400 + 12 = 1412$; $\text{AIC}_B = 1393 + 18 = 1411$. AIC selects Model B (the larger model) by 1 unit.

BIC ($\log 500 = 6.21$): $\text{BIC}_A = 1400 + 37.3 = 1437.3$; $\text{BIC}_B = 1393 + 55.9 = 1448.9$. BIC selects Model A (the true model).

The three extra parameters in Model B improve the likelihood by 3.5, which exceeds AIC's penalty of 6 but not BIC's penalty of 18.6. This illustrates: AIC favors prediction (Model B is slightly better out-of-sample), BIC favors parsimony (Model A is the true model).

Cross-Links

AI. Neural architecture search and hyperparameter tuning replace analytic criteria with validation-set evaluation, but AIC-like principles (penalizing effective parameters) appear in pruning criteria and the minimum description length principle for model compression.

Economics. AIC and BIC are standard in VAR lag-length selection, ARMA order selection, and structural model comparison. Economists often report both, noting that disagreement signals the prediction-vs-identification tension.

Linear Algebra. The effective degrees of freedom in penalized regression, $\text{df}(\lambda) = \text{Tr}(H_\lambda)$, generalizes the parameter count k and can be substituted into AIC for regularized models. The trace operation connects model complexity to the spectrum of the hat matrix.

Quantum Mechanics. In quantum state tomography, model selection between density matrix parameterizations of different rank uses likelihood ratio tests and information criteria adapted to the positive semidefinite constraint, with the BIC penalty reflecting the dimensionality of the state manifold.

Structural Compression

Invariant

Information criteria decompose model quality into fit ($-2\ell_j$) and a complexity penalty ($2k_j$ or $k_j \log n$); model selection minimizes the sum.

Mechanism

AIC: Akaike's theorem shows the in-sample log-likelihood overestimates the expected out-of-sample log-likelihood by k_j/n in expectation; multiplying by $-2n$ yields the penalty $2k_j$. BIC: the Laplace approximation to the marginal likelihood introduces a $(k_j/2)\log n$ penalty for parameter uncertainty integrated over the prior.

Cross-System Echo

AIC is asymptotically equivalent to LOOCV (Stone, 1977) under squared-error loss for linear models. BIC is the Schwarz approximation to the Bayes factor (Node 6.6) for model comparison. In information theory, both connect to the minimum description length (MDL) principle: the best model compresses the data most efficiently.

Exercises

1. Starting from the KL divergence $D_{\mathrm{KL}}(p_0 \| p_{\hat{\theta}})$, carry out the full Taylor expansion to derive the AIC penalty $2k$. Identify where the asymptotic normality of the MLE is used.

2. Derive the BIC via the Laplace approximation. Show explicitly that the prior density $\pi(\hat{\theta})$ and the term $k \log(2\pi)$ are absorbed into the $O(1)$ remainder.

3. Show that BIC is consistent: if the true model $\mathcal{M}^*$ is among the candidates and has the fewest parameters, then $\mathbb{P}(\hat{j}_{\mathrm{BIC}} = *) \to 1$ as $n \to \infty$.

4. Show that for Gaussian linear regression with known σ^2, AIC reduces to $\mathrm{RSS}/\sigma^2 + 2k$ and BIC reduces to $\mathrm{RSS}/\sigma^2 + k \log n$, both up to additive constants independent of the model.

5. For a dataset with $n = 20$ and two models ($k_1 = 3, \ell_1 = -25.0$; $k_2 = 6, \ell_2 = -21.5$), compute AIC, AICc, and BIC for both models. Which does each criterion prefer?

6. Derive the AICc correction for Gaussian linear regression by computing the exact expected optimism (using the non-central chi-squared distribution of the in-sample RSS) rather than the asymptotic approximation.

7. Argue, structurally, why no single information criterion can simultaneously be consistent (selecting the true model) and efficient (minimizing predictive risk when no model is true), and relate this impossibility to the bias–variance tradeoff.

Cross-Validation

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.3.

Cross-validation is the nonparametric, model-agnostic approach to estimating generalization error. Where information criteria (Node 10.2) rely on parametric assumptions and closed-form penalties, cross-validation directly simulates the train–test split on the observed data. It applies to any estimator and any loss function, making it the universal model selection tool for complex or nonparametric pipelines. It feeds into high-dimensional methods (Node 10.5) as the standard tuning procedure.

Formal Definition

Let $\hat{f}_n(\,\cdot\,;\mathcal{D})$ denote an estimator trained on dataset $\mathcal{D}$. The **generalization error** under loss ℓ is

$$R_n = \mathbb{E}_{(X,Y)\sim P_0}\big[\ell(Y, \hat{f}_n(X;\mathcal{D}_n))\big],$$

where the outer expectation is over a new test point independent of $\mathcal{D}_n$.

K-Fold Cross-Validation. Partition $\mathcal{D}_n$ into K disjoint folds $\mathcal{D}_1, \ldots, \mathcal{D}_K$ of approximately equal size n/K. For each $k = 1, \ldots, K$, train on $\mathcal{D}_{-k} = \mathcal{D}_n \setminus \mathcal{D}_k$ to obtain $\hat{f}^{(-k)}$ and evaluate on fold $\mathcal{D}_k$. The CV estimator is

$$\widehat{R}^{K\text{-CV}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{(X_i,Y_i)\in\mathcal{D}_k} \ell\big(Y_i,\, \hat{f}^{(-k)}(X_i)\big).$$

Leave-One-Out Cross-Validation (LOOCV). Set $K = n$:

$$\widehat{R}^{\text{LOO}} = \frac{1}{n}\sum_{i=1}^{n} \ell\big(Y_i,\, \hat{f}^{(-i)}(X_i)\big).$$

Intuition

Cross-validation answers the question: how well does this estimator perform on data it has not seen? Rather than using a single held-out test set (which wastes data), it rotates through all possible partitions so every observation serves as both training and validation data. The average held-out loss is an estimate of the generalization error.

The key tradeoff in choosing K is between bias (training on fewer points overestimates the error) and variance (correlated fold estimates inflate the variance of the CV estimator).

Structural Role

Dependencies: Builds on the bias–variance tradeoff (Node 10.1) and estimation theory (Node 3.1). **Enables:** Tuning parameter selection for regularized estimators (Node 10.4), evaluation of high-dimensional methods (Node 10.5), and weight selection in stacking (Node 10.6).

Cross-validation is the empirical complement to the analytic criteria of Node 10.2. Where AIC and BIC require parametric structure, CV applies to any estimator—random forests, neural networks, kernel methods—making it indispensable for modern statistical practice.

Mathematical Development

Bias of K-Fold CV

The k-th fold trains on $n(1-1/K)$ observations. Since the generalization error typically decreases with training set size, the CV estimator targets the risk of an estimator trained on $n(1-1/K)$ points, not n points. Therefore:

$$\mathbb{E}[\widehat{R}^{K\text{-CV}}] = R_{n(1-1/K)} \geq R_n.$$

The CV estimator has **upward bias** (pessimistic) for R_n. LOOCV ($K = n$) trains on $n-1$ points, so its bias is minimal: $R_{n-1} - R_n = O(1/n^2)$ under smoothness conditions.

Variance of the CV Estimator

The n leave-one-out residuals $e_i = \ell(Y_i, \hat{f}^{(-i)}(X_i))$ are not independent: each pair of models $\hat{f}^{(-i)}$ and $\hat{f}^{(-j)}$ shares $n-2$ training points. The variance of LOOCV is

$$\mathrm{Var}(\widehat{R}^{\mathrm{LOO}}) = \frac{1}{n^2}\left[n\,\mathrm{Var}(e_i) + n(n-1)\,\mathrm{Cov}(e_i, e_j)\right].$$

When the estimator is unstable (high variance), the covariance term is large and LOOCV has high variance. Smaller K reduces this covariance at the cost of increased bias.

Empirical recommendation: $K = 5$ or $K = 10$ provides a reasonable bias–variance balance. Repeated K-fold CV (averaging over multiple random partitions) reduces the variance from the particular fold assignment.

Stone's Theorem: LOOCV and AIC

Theorem (Stone, 1977). For linear models under squared-error loss, LOOCV and AIC are asymptotically equivalent:

$$\widehat{R}^{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2 \approx \frac{\mathrm{RSS}}{n} + \frac{2\hat{\sigma}^2 k}{n} + o(1/n),$$

where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix and $k = \mathrm{Tr}(H)$.

Proof sketch. Expand $(1 - H_{ii})^{-2} \approx 1 + 2H_{ii} + O(H_{ii}^2)$. Then

$$\widehat{R}^{\mathrm{LOO}} \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2(1 + 2H_{ii}) = \frac{\mathrm{RSS}}{n} + \frac{2}{n}\sum_i e_i^2 H_{ii}.$$

Under the Gaussian linear model, $\mathbb{E}[e_i^2 H_{ii}] \approx \sigma^2 H_{ii}$, so $\sum_i e_i^2 H_{ii} \approx \sigma^2 \mathrm{Tr}(H) = \sigma^2 k$. This gives

$$\widehat{R}^{\mathrm{LOO}} \approx \frac{\mathrm{RSS}}{n} + \frac{2\sigma^2 k}{n} = \frac{1}{n}\left(\mathrm{RSS} + 2\sigma^2 k\right),$$

which is AIC (up to a constant and the factor $1/n$) for the Gaussian linear model.

LOOCV Shortcut for Linear Models

For any linear smoother $\hat{Y} = HY$, the LOOCV loss under squared error is computed without refitting:

$$\widehat{R}^{\mathrm{LOO}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2.$$

Derivation. By the Sherman–Morrison formula, removing observation i from OLS gives

$$\hat{Y}_i^{(-i)} = \hat{Y}_i - \frac{H_{ii}(Y_i - \hat{Y}_i)}{1 - H_{ii}},$$

so the leave-one-out residual is

$$Y_i - \hat{Y}_i^{(-i)} = \frac{Y_i - \hat{Y}_i}{1 - H_{ii}}.$$

This reduces the cost of LOOCV from n model fits to a single fit plus computation of the hat matrix diagonal.

Generalized Cross-Validation (GCV)

For large n, the individual leverages H_{ii} can be replaced by their average $\mathrm{Tr}(H)/n = k/n$:

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - k/n}\right)^2 = \frac{\mathrm{RSS}/n}{(1 - k/n)^2}.$$

GCV is invariant under orthogonal transformations of the response and is computationally cheaper than LOOCV when H_{ii} values are expensive to compute.

Worked Examples

Example 1: LOOCV for Polynomial Selection

Data: $n = 15$ observations from $Y = X^2 + \varepsilon$, $X \sim \mathrm{Uniform}(-1, 1)$, $\varepsilon \sim \mathcal{N}(0, 0.5^2)$. Fit polynomials of degree 1, 2, 4.

For the degree-p polynomial, $k = p + 1$, and $H = X_p(X_p^\top X_p)^{-1}X_p^\top$ where X_p is the $n \times (p+1)$ Vandermonde matrix.

Suppose $\mathrm{RSS}_1 = 4.20$, $\mathrm{RSS}_2 = 3.15$, $\mathrm{RSS}_4 = 2.98$, and the average leverages are $\bar{H}_1 = 2/15$, $\bar{H}_2 = 3/15$, $\bar{H}_4 = 5/15$.

Using GCV: $\mathrm{GCV}_1 = (4.20/15)/(1 - 2/15)^2 = 0.280/0.751 = 0.373$; $\mathrm{GCV}_2 = (3.15/15)/(1 - 3/15)^2 = 0.210/0.640 = 0.328$; $\mathrm{GCV}_4 = (2.98/15)/(1 - 5/15)^2 = 0.199/0.444 = 0.448$.

GCV selects degree 2. The degree-4 model reduces RSS by only 0.17 but the GCV denominator penalizes the higher leverage heavily, reflecting the variance cost of extra parameters.

Example 2: CV for Regularization Parameter Selection

Consider ridge regression with $p = 20$ predictors and $n = 50$. Use 5-fold CV to select $\lambda \in \{0.01, 0.1, 1, 10, 100\}$.

For each λ, the ridge estimator is $\hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1} X^\top Y$. Partition the data into 5 folds of 10 observations each.

Suppose the average held-out MSE values are: $\lambda = 0.01$: 2.85; $\lambda = 0.1$: 2.41; $\lambda = 1$: 2.12; $\lambda = 10$: 2.35; $\lambda = 100$: 3.90.

The minimum is at $\lambda = 1$. The one-standard-error rule: if $\mathrm{SE}(\widehat{\mathrm{MSE}}_{\lambda=1}) = 0.30$, select the largest λ with CV error within $2.12 + 0.30 = 2.42$. This is $\lambda = 10$ (CV error 2.35), yielding a sparser, more regularized model.

Cross-Links

AI. Cross-validation is the standard method for hyperparameter tuning in machine learning. In deep learning, the computational cost of K-fold CV for large models is prohibitive; single validation splits or learning-curve extrapolation are common surrogates. Stacking (Node 10.6) uses CV predictions as meta-features.

Economics. Out-of-sample forecasting evaluation is the time-series analogue of CV. Rolling-window and expanding-window designs respect temporal dependence, replacing random fold assignment with forward-chaining to avoid information leakage.

Linear Algebra. The hat matrix $H = X(X^\top X)^{-1} X^\top$ is the orthogonal projector onto the column space of X. The diagonal entries $H_{ii} \in [1/n, 1]$ are the leverages, and the LOOCV shortcut exploits the algebraic structure of rank-one updates to projectors.

Quantum Mechanics. In quantum process tomography, cross-validation selects the complexity of the process matrix reconstruction (e.g., the rank of the Choi matrix) by holding out subsets of measurement data and evaluating predictive log-likelihood on the held-out measurements.

Structural Compression

Invariant

The CV estimator approximates the generalization error by averaging hold-out losses over rotated train–test splits; as the number of folds increases, bias decreases but variance increases due to correlated fold estimates.

Mechanism

Hold-out data provides an unbiased estimate of test error for the model trained on the complement; averaging over folds reduces variance of the estimate. The LOOCV shortcut for linear smoothers exploits the Sherman–Morrison formula to avoid explicit refitting.

Cross-System Echo

Cross-validation is the standard generalization error estimator in machine learning. LOOCV is algebraically related to the jackknife (a variance estimation method via leave-one-out resampling). Stone's theorem establishes the asymptotic equivalence of LOOCV and AIC for linear models, unifying the model-based and model-free approaches to model selection.

Exercises

1. Derive the LOOCV shortcut for linear regression: starting from $\hat{\beta}^{(-i)} = (X_{-i}^\top X_{-i})^{-1} X_{-i}^\top Y_{-i}$, use the Sherman–Morrison formula to show that $Y_i - X_i^\top \hat{\beta}^{(-i)} = (Y_i - \hat{Y}_i)/(1 - H_{ii})$.

2. Prove that LOOCV is exactly unbiased for the expected prediction error of an estimator trained on $n-1$ observations, i.e., $\mathbb{E}[\widehat{R}^{\mathrm{LOO}}] = R_{n-1}$.

3. For a dataset with $n = 100$ and a linear model with $k = 10$ parameters, compute the GCV score given $\mathrm{RSS} = 85.0$. Then compute the exact LOOCV score if the leverage values are approximately constant at $H_{ii} \approx k/n = 0.1$.

4. Show that the variance of the LOOCV estimator satisfies $\mathrm{Var}(\widehat{R}^{\mathrm{LOO}}) \leq \frac{1}{n}\mathrm{Var}(e_i) + \frac{n-1}{n}|\mathrm{Cov}(e_i, e_j)|$, and explain when the covariance term dominates.

5. Implement Stone's theorem: for a Gaussian linear model, show that $n \cdot \widehat{R}^{\mathrm{LOO}}$ and $\mathrm{RSS} + 2\hat{\sigma}^2 k$ differ by $O_p(1)$, so that model rankings under LOOCV and AIC agree asymptotically.

6. In a time-series context, explain why random K-fold CV is invalid and describe the forward-chaining (time-series CV) procedure. State the bias–variance implications of expanding-window vs. fixed-window designs.

7. Argue, structurally, why cross-validation must have higher variance than AIC for model selection in correctly specified parametric models, and identify the settings where this variance cost is justified by the elimination of model-specification assumptions.

Penalized Likelihood and Regularization

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.4.

Penalized likelihood is the analytical implementation of the bias–variance tradeoff (Node 10.1): instead of selecting a model from a discrete set, it continuously shrinks parameter estimates toward a structured target. Ridge regression (L^2), LASSO (L^1), and the elastic net ($L^1 + L^2$) are the canonical instances. The geometry of the penalty ball determines whether the estimator achieves sparsity. This node provides the foundation for high-dimensional methods (Node 10.5) and informs model averaging through regularization paths (Node 10.6).

Formal Definition

Let $\ell(\beta; X)$ be the log-likelihood for $\beta \in \mathbb{R}^p$. The **penalized likelihood estimator** is

$$\hat{\beta}^\lambda = \arg\max_{\beta \in \mathbb{R}^p} \left[\ell(\beta; X) - \lambda\, \Omega(\beta) \right],$$

where $\Omega(\beta) \geq 0$ is a penalty function and $\lambda \geq 0$ controls the penalty strength. Equivalently, for the Gaussian linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$:

$$\hat{\beta}^\lambda = \arg\min_{\beta} \left[\frac{1}{2n} \|Y - X\beta\|^2 + \lambda\, \Omega(\beta) \right].$$

Intuition

When p is large relative to n, the MLE overfits: it finds coefficients that explain the noise in the training data perfectly but predict poorly on new data. Regularization constrains the estimator by penalizing large or numerous coefficients. The geometry matters: the L^2 ball is smooth and round, shrinking all coefficients proportionally; the L^1 ball has corners at the axes, pushing small coefficients to exactly zero. This geometric difference is why LASSO performs variable selection and ridge does not.

Structural Role

Dependencies: Builds on the bias–variance tradeoff (Node 10.1), maximum likelihood theory (Node 3.2), and the Bayesian perspective (Node 6.1). **Enables:** High-dimensional theory (Node 10.5) uses LASSO as its central estimator; model averaging via stacking (Node 10.6) uses regularization to combine models.

Penalized likelihood unifies frequentist and Bayesian perspectives: the penalty is a prior (Node 6.1), and the penalized MLE is the posterior mode. The tuning parameter λ controls the bias–variance tradeoff continuously rather than discretely.

Mathematical Development

Ridge Regression (L^2 Penalty)

The **ridge estimator** is

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2.$$

Setting the gradient to zero: $-X^\top(Y - X\beta) + \lambda\beta = 0$, so

$$\hat{\beta}^{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Properties via the SVD. Let $X = UDV^\top$ with singular values $d_1 \geq \cdots \geq d_p$. Then

$$\hat{\beta}^{\text{ridge}} = V \operatorname{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right) V^\top \hat{\beta}^{\text{OLS}}.$$

Each coefficient in the rotated basis is shrunk by the factor $d_j^2/(d_j^2 + \lambda) \in [0, 1)$. Components along small singular values (where the data provides little information) are shrunk the most.

Bias–variance decomposition for ridge:

$$\text{Bias}^2 = \lambda^2 \beta_0^\top (X^\top X + \lambda I)^{-2} \beta_0, \quad \text{Variance} = \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2}.$$

As $\lambda \to 0$, bias vanishes and variance equals the OLS variance. As $\lambda \to \infty$, variance vanishes and bias approaches $\|\beta_0\|^2$.

Bayesian interpretation: Ridge is the posterior mean under $\beta \sim \mathcal{N}(0, (\sigma^2/\lambda)I)$.

LASSO (L^1 Penalty)

The **LASSO** estimator is

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

The L^1 penalty is not differentiable at $\beta_j = 0$. The **subdifferential** (KKT conditions) for coordinate j is:

$$-X_j^\top(Y - X\hat{\beta}) + \lambda s_j = 0, \quad s_j \in \begin{cases} \{\text{sign}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j \neq 0, \\ [-1, 1] & \text{if } \hat{\beta}_j = 0. \end{cases}$$

This means $\hat{\beta}_j = 0$ whenever $|X_j^\top(Y - X\hat{\beta}_{-j})| \leq \lambda$: the penalty drives the coefficient to zero if the marginal correlation is too small to justify including the variable.

Soft-thresholding operator. For orthogonal X (i.e., $X^\top X = nI$), the LASSO solution decouples coordinate-wise:

$$\hat{\beta}_j^{\text{LASSO}} = S_{\lambda/n}(\hat{\beta}_j^{\text{OLS}}) = \text{sign}(\hat{\beta}_j^{\text{OLS}})\big(|\hat{\beta}_j^{\text{OLS}}| - \lambda/n\big)_+.$$

Derivation of soft-thresholding. For orthogonal X, the objective separates: $\min_{\beta_j} \frac{n}{2}(\hat{\beta}_j^{\text{OLS}} - \beta_j)^2 + \lambda|\beta_j|$. For $\beta_j > 0$: differentiate to get $-n(\hat{\beta}_j^{\text{OLS}} - \beta_j) + \lambda = 0$, so $\beta_j = \hat{\beta}_j^{\text{OLS}} - \lambda/n$, valid when $\hat{\beta}_j^{\text{OLS}} > \lambda/n$. By symmetry and the KKT conditions for $\beta_j = 0$, the full solution is the soft-thresholding operator.

Geometry of sparsity. The constraint set $\{\beta : \|\beta\|_1 \leq t\}$ is a cross-polytope (diamond in 2D). Its corners lie on the coordinate axes. The contours of the quadratic loss are ellipsoids. The point of tangency between

an ellipsoid and a cross-polytope generically lies at a corner, setting one or more coordinates to zero. By contrast, the L^2 ball $\{\beta : \|\beta\|_2 \le t\}$ is smooth, so the tangency generically occurs at an interior point with all coordinates nonzero.

Bayesian interpretation: LASSO is the posterior mode under independent Laplace priors $\beta_j \sim \text{Laplace}(0, \sigma^2/\lambda)$.

Elastic Net

The **elastic net** combines L^1 and L^2 penalties:

$$\hat{\beta}^{\text{EN}} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \frac{\lambda_2}{2}\|\beta\|^2.$$

The L^2 term makes the objective strictly convex, ensuring a unique solution even when $p > n$. The **grouping effect**: for highly correlated predictors $X_j \approx X_k$, the elastic net assigns similar coefficients, whereas LASSO may select one and zero the other arbitrarily.

Equivalently, the elastic net can be written as a LASSO problem on an augmented dataset:

$$\tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}, \quad \tilde{X} = \frac{1}{\sqrt{1+\lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2}\, I \end{pmatrix},$$

and solving the LASSO with penalty $\lambda_1/\sqrt{1+\lambda_2}$.

Selection of λ

The penalty parameter is not estimated from the likelihood. Standard approaches:

1. **Cross-validation** (Node 10.3): select λ minimizing K-fold CV error. For LASSO, compute the full regularization path via LARS (least angle regression) at cost comparable to a single OLS fit.
2. **Information criteria**: for LASSO, the effective degrees of freedom equals $|\hat{S}|$ (the number of nonzero coefficients), so AIC and BIC can be applied with $k = |\hat{S}|$.
3. **Theoretical bounds** (Node 10.5): set $\lambda = C\sigma\sqrt{\log p/n}$ for oracle-rate guarantees.

Worked Examples

Example 1: Ridge vs. OLS with Collinear Predictors

Data: $n = 20$, $p = 2$, X_1 and X_2 have correlation 0.99. True model: $Y = 3X_1 + 3X_2 + \varepsilon$, $\sigma = 1$.

The OLS estimates have high variance because $X^\top X$ is nearly singular. Suppose $X^\top X = 20\begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$.

Eigenvalues: $d_1^2 = 20(1.99) = 39.8$, $d_2^2 = 20(0.01) = 0.2$.

OLS variance: $\sigma^2 \text{Tr}((X^\top X)^{-1}) = 1 \cdot (1/39.8 + 1/0.2) = 0.025 + 5.0 = 5.025$. The small eigenvalue inflates variance enormously.

Ridge with $\lambda = 1$: variance $= \sum_j d_j^2/(d_j^2+1)^2 = 39.8/40.8^2 + 0.2/1.2^2 = 0.024 + 0.139 = 0.163$. Bias^2 is small: the component along the large eigenvalue is barely shrunk, and the component along the small eigenvalue is shrunk significantly but the true signal along that direction is small. Ridge MSE $\approx 0.163 + \text{bias}^2 \ll 5.025$.

Example 2: LASSO Variable Selection

Data: $n = 100$, $p = 10$. True model: $Y = 2X_1 - X_3 + 0.5X_7 + \varepsilon$, $\sigma = 1$. Predictors are independent standard normal.

At $\lambda = 0.5$ (chosen by CV), the LASSO estimates: $\hat{\beta}_1 = 1.85$, $\hat{\beta}_3 = -0.82$, $\hat{\beta}_7 = 0.31$, and all others exactly zero. The LASSO correctly identifies the support $S = \{1, 3, 7\}$ and provides shrunken but useful estimates of the coefficients.

At $\lambda = 2.0$ (too large): only $\hat{\beta}_1 = 0.94$ survives; X_3 and X_7 are incorrectly zeroed. At $\lambda = 0.01$ (too small): all 10 predictors have nonzero coefficients, overfitting.

The CV error curve is U-shaped in λ, with minimum near $\lambda = 0.5$, confirming the bias–variance tradeoff.

Cross-Links

AI. Weight decay in neural networks is L^2 regularization applied to network weights. Dropout induces an implicit regularization effect similar to an adaptive L^2 penalty. Sparse attention mechanisms and network pruning exploit the L^1/sparsity principle.

Economics. Penalized regression underlies high-dimensional empirical applications including factor model selection, treatment effect estimation with many controls, and forecasting with large predictor sets. The post-LASSO estimator (OLS on the LASSO-selected variables) is standard in empirical work to remove shrinkage bias.

Linear Algebra. The geometry of L^p balls for $p \leq 1$ (non-convex), $p = 1$ (convex, corners), and $p = 2$ (smooth, round) determines the sparsity properties of the solution. The LASSO is a linear program (for fixed λ) and can be solved by coordinate descent or the LARS algorithm, both exploiting the polyhedral geometry.

Quantum Mechanics. Compressed sensing (the signal-processing analogue of LASSO) is used in quantum state tomography to reconstruct low-rank density matrices from a small number of measurements. The sparsity-inducing geometry of the nuclear norm (trace norm) plays the role of the L^1 norm for matrix-valued signals.

Structural Compression

Invariant

Penalized likelihood reduces variance at the cost of bias; the optimal penalty λ^* minimizes total prediction error. The L^1 penalty achieves sparsity (exact zeros); the L^2 penalty achieves shrinkage (approximate zeros).

Mechanism

The penalty $\lambda\Omega(\beta)$ restricts the feasible region. For L^1: the KKT conditions show that the subdifferential at zero absorbs the gradient when the signal is weak, setting the coefficient to exactly zero. For L^2: the smooth penalty applies uniform multiplicative shrinkage via $d_j^2/(d_j^2 + \lambda)$. The soft-thresholding operator is the proximal operator of the L^1 norm.

Cross-System Echo

LASSO is equivalent to basis pursuit (compressed sensing) in signal processing. The sparsity-inducing geometry of the L^1 ball is central to both. In AI, weight decay (ridge) and L^1 penalties are foundational regularization techniques. In economics, penalized regression enables credible estimation in settings where the number of potential controls exceeds the sample size.

Exercises

1. Derive the ridge estimator $\hat{\beta}^{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top Y$ by setting the gradient of the penalized objective to zero. Express the solution in terms of the SVD of X.

2. Derive the soft-thresholding operator for the LASSO in the orthogonal design case. Draw the solution path $\hat{\beta}_j(\lambda)$ as a function of λ for $\hat{\beta}_j^{\text{OLS}} = 2.5$.

3. Write out the KKT conditions for the LASSO in the general (non-orthogonal) case. Show that if $|X_j^\top (Y - X\hat{\beta}_{-j})| < \lambda$, then $\hat{\beta}_j = 0$.

4. Prove the grouping property of the elastic net: for $X_j = X_k$, show that $\hat{\beta}_j^{\text{EN}} = \hat{\beta}_k^{\text{EN}}$, whereas the LASSO may assign different values.

5. For ridge regression, show that the effective degrees of freedom is $\text{df}(\lambda) = \sum_{j=1}^{p} d_j^2/(d_j^2 + \lambda)$, where d_j are the singular values of X.

6. Consider $n = 50$, $p = 5$, $X^\top X = 50I$, $\hat{\beta}^{\text{OLS}} = (3.0, 0.2, -2.5, 0.1, 1.8)$, $\sigma^2 = 1$. Compute the LASSO solution for $\lambda = 5$ using the soft-thresholding formula, and compute the ridge solution for $\lambda = 5$. Compare sparsity.

7. Argue, structurally, why the L^1 penalty produces exact zeros while the L^2 penalty does not, using both the geometric (constraint set tangency) and analytic (subdifferential) perspectives.

High-Dimensional Structure and Sparsity

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.5.

When the number of parameters p exceeds the sample size n, classical statistical theory breaks down: the MLE does not exist, consistent estimation in the ambient dimension is impossible, and standard asymptotics (fixed p, $n \to \infty$) are inapplicable. This node develops the theory that replaces the classical paradigm by imposing structural assumptions—primarily sparsity—and establishes oracle inequalities showing that the LASSO (Node 10.4) achieves minimax optimal rates under these assumptions. The connection to compressed sensing provides the information-theoretic foundation.

Formal Definition

Consider the linear model $Y = X\beta_0 + \varepsilon$ where $X \in \mathbb{R}^{n \times p}$, $p \gg n$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

A vector $\beta_0 \in \mathbb{R}^p$ is **s-sparse** if $\|\beta_0\|_0 = |\{j : \beta_{0j} \neq 0\}| \leq s \ll p$.

The **effective dimension** of the problem is s, not p. The goal is estimation and prediction at rates depending on s and $\log p$, not on p itself.

Intuition

In high dimensions, the parameter space is enormous but the true signal occupies only a low-dimensional subspace. Sparsity says: most coefficients are exactly zero. The statistical challenge is to identify the nonzero coefficients (support recovery) and estimate them accurately, despite having far fewer observations than parameters. The LASSO solves both problems simultaneously under conditions on the design matrix.

The price of high dimensionality is a $\log p$ factor in the estimation rate—the cost of searching through p candidates to find the s active ones.

Structural Role

Dependencies: Builds on penalized likelihood (Node 10.4), cross-validation for tuning (Node 10.3), and asymptotic theory (Node 7.3). **Enables:** This node is a terminal node in Module 10, representing the frontier of modern statistical theory.

High-dimensional statistics replaces the classical assumption $p \ll n$ with the structural assumption $s \ll n$. The restricted eigenvalue condition replaces the non-singularity of $X^\top X$. The oracle inequality characterizes estimation error in terms of intrinsic dimension s rather than ambient dimension p.

Mathematical Development

Why Classical Asymptotics Fail

Classical theory assumes p fixed, $n \to \infty$. The MLE satisfies $\sqrt{n}(\hat{\beta} - \beta_0) \to \mathcal{N}(0, I(\beta_0)^{-1})$. When $p > n$:

1. $X^\top X$ is singular (rank at most n), so $(X^\top X)^{-1}$ does not exist.
2. There are infinitely many β satisfying $X\beta = Y$ exactly (interpolation).
3. The Fisher information matrix is $p \times p$ but has rank $\leq n$.

4. Consistent estimation of β_0 in $\|\cdot\|_2$ is impossible without structural assumptions: the minimax rate over all $\beta_0 \in \mathbb{R}^p$ is $\Theta(p/n)$, which diverges.

Restricted Eigenvalue Condition

The **restricted eigenvalue condition** $\mathrm{RE}(s, c_0)$ holds if there exists $\kappa > 0$ such that

$$\frac{1}{n}\|X\delta\|_2^2 \geq \kappa \|\delta_S\|_2^2$$

for all $\delta \in \mathbb{R}^p$ satisfying $\|\delta_{S^c}\|_1 \leq c_0\|\delta_S\|_1$, where $S = \mathrm{supp}(\beta_0)$ with $|S| \leq s$.

The cone condition $\|\delta_{S^c}\|_1 \leq c_0\|\delta_S\|_1$ restricts attention to perturbations that are approximately sparse—precisely the directions explored by the LASSO solution path. The RE condition ensures that X is well-conditioned on these directions.

Relation to other conditions. The RE condition is weaker than: - The **restricted isometry property** (RIP): $(1-\delta_s)\|\delta\|_2^2 \leq \frac{1}{n}\|X\delta\|_2^2 \leq (1+\delta_s)\|\delta\|_2^2$ for all s-sparse δ. - The **irrepresentability condition**: $\|X_{S^c}^\top X_S (X_S^\top X_S)^{-1}\|_\infty < 1$, required for sign consistency.

All three are satisfied with high probability when X has i.i.d. sub-Gaussian rows and $n \gtrsim s \log p$.

Oracle Inequality for the LASSO

Theorem. Suppose the RE condition holds with constant κ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Set $\lambda = A\sigma\sqrt{\log p/n}$ for a sufficiently large constant A. Then with probability at least $1 - 2/p$:

$$\frac{1}{n}\|X\hat{\beta}^{\mathrm{LASSO}} - X\beta_0\|_2^2 \leq C \cdot \frac{s\sigma^2 \log p}{n},$$

and

$$\|\hat{\beta}^{\mathrm{LASSO}} - \beta_0\|_1 \leq C' \cdot s\sigma\sqrt{\frac{\log p}{n}},$$

for constants C, C' depending only on κ and A.

Proof sketch.

Step 1 (Basic inequality). By optimality of $\hat{\beta}^{\mathrm{LASSO}}$:

$$\frac{1}{2n}\|Y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta_0\|^2 + \lambda\|\beta_0\|_1.$$

Let $\delta = \hat{\beta} - \beta_0$. Expanding:

$$\frac{1}{2n}\|X\delta\|^2 \leq \frac{1}{n}\varepsilon^\top X\delta + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1).$$

Step 2 (Stochastic control). The key event is $\mathcal{E} = \{\max_j |X_j^\top \varepsilon/n| \leq \lambda/2\}$. For Gaussian ε:

$$\mathbb{P}\left(\max_{j=1,\ldots,p}\left|\frac{1}{n}X_j^\top\varepsilon\right| > t\right) \leq 2p\exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Setting $t = \lambda/2 = (A/2)\sigma\sqrt{\log p/n}$ gives $\mathbb{P}(\mathcal{E}^c) \leq 2p^{1-A^2/8} \leq 2/p$ for A large enough.

Step 3 (Cone condition). On event $\mathcal{E}$, the triangle inequality gives $\|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1$. This places δ in the cone required by the RE condition.

Step 4 (Combining). $\frac{1}{n}\|X\delta\|^2 \le 2\lambda\|\delta_S\|_1 \le 2\lambda\sqrt{s}\|\delta_S\|_2$. By the RE condition, $\|\delta_S\|_2 \le \frac{1}{\kappa n}\|X\delta\|^2 \cdot \frac{1}{\|\delta_S\|_2}$. Substituting and simplifying gives the oracle rate.

The $\log p$ Factor

The factor $\log p$ arises from controlling the maximum of p sub-Gaussian random variables:

$$\mathbb{E}\left[\max_{j=1,\dots,p}|Z_j|\right] \le C\sqrt{\log p}$$

for Z_j standard Gaussian. This is tight: $\max_j |Z_j| \sim \sqrt{2\log p}$. The union bound over p variables is the information-theoretic price of searching in high dimensions.

Donoho–Tanner Phase Transition

For the noiseless compressed sensing problem ($Y = X\beta_0$, $\varepsilon = 0$), exact recovery of s-sparse β_0 via L^1 minimization exhibits a sharp phase transition.

Define $\rho = s/n$ (sparsity ratio) and $\delta = n/p$ (measurement ratio). The Donoho–Tanner transition curve $\rho^*(\delta)$ satisfies:

$$\begin{aligned}
&\text{If } \rho < \rho^*(\delta): \quad \hat{\beta}^{L^1} = \beta_0 \text{ with probability } \to 1.\\
&\text{If } \rho > \rho^*(\delta): \quad \hat{\beta}^{L^1} \ne \beta_0 \text{ with probability } \to 1.
\end{aligned}$$

The transition curve is computed from the geometry of random polytopes. For $\delta = 0.5$ (twice as many variables as observations), recovery succeeds when $\rho \lesssim 0.09$, i.e., sparsity is below 9% of the number of measurements.

Restricted Isometry Property (RIP)

A matrix $X \in \mathbb{R}^{n\times p}$ satisfies the **RIP of order** s with constant $\delta_s \in (0,1)$ if

$$(1-\delta_s)\|\beta\|_2^2 \le \frac{1}{n}\|X\beta\|_2^2 \le (1+\delta_s)\|\beta\|_2^2$$

for all s-sparse β. Random matrices with i.i.d. sub-Gaussian entries satisfy RIP with $n = O(s\log(p/s))$ measurements. RIP implies the RE condition and guarantees exact recovery in the noiseless case and stable recovery in the noisy case.

Sign Consistency

Under the irrepresentability condition and a minimum signal strength condition $\min_{j\in S}|\beta_{0j}| \gg \sigma\sqrt{\log p/n}$, the LASSO achieves **sign consistency**:

$$\mathbb{P}\Big(\operatorname{sign}(\hat{\beta}^{\text{LASSO}}) = \operatorname{sign}(\beta_0)\Big) \to 1.$$

Without the irrepresentability condition, the LASSO may include false positives. The adaptive LASSO (with data-dependent weights $w_j = 1/|\hat{\beta}_j^{\text{init}}|^\gamma$ in the penalty) achieves sign consistency under weaker conditions.

Worked Examples

Example 1: Oracle Rate Verification

Setup: $n = 200$, $p = 1000$, $s = 5$, $\sigma = 1$. The oracle rate is $s\sigma^2 \log p/n = 5 \cdot 1 \cdot \log(1000)/200 = 5 \cdot 6.91/200 = 0.173$.

Set $\lambda = 2\sigma\sqrt{\log p/n} = 2\sqrt{6.91/200} = 2 \cdot 0.186 = 0.372$.

In simulation (averaging over 100 replications with $X_{ij} \sim \mathcal{N}(0,1)$ and $\beta_0 = (3, -2, 1.5, -1, 0.8, 0, \ldots, 0)$):

Prediction error: $\frac{1}{n}\|X\hat{\beta}^{\text{LASSO}} - X\beta_0\|^2 \approx 0.15$, consistent with the oracle bound of 0.173 (up to constants).

Support recovery: the LASSO correctly identifies all 5 active variables in 92 of 100 replications. The 8 failures include one false negative (coefficient $\beta_5 = 0.8$ is close to the threshold) and occasional false positives.

Example 2: Phase Transition

Noiseless setting: $p = 500$, $X_{ij} \sim \mathcal{N}(0, 1/n)$. Vary $n \in \{50, 100, 150, 200, 250\}$ and $s \in \{5, 15, 25, 35\}$.

Solve $\min_\beta \|\beta\|_1$ subject to $Y = X\beta$. Record whether $\hat{\beta} = \beta_0$ (exact recovery).

Results: at $n = 100$ ($\delta = 0.2$), recovery succeeds for $s \leq 5$ ($\rho = 0.05$) but fails for $s = 15$ ($\rho = 0.15$). At $n = 250$ ($\delta = 0.5$), recovery succeeds for $s \leq 20$ ($\rho = 0.08$). This matches the Donoho–Tanner prediction: the boundary $\rho^*(\delta)$ is approximately 0.09 at $\delta = 0.5$.

Cross-Links

AI. Sparsity priors underlie feature selection, network pruning, and sparse attention in transformers. The lottery ticket hypothesis posits that overparameterized networks contain sparse subnetworks that train equally well, a neural analogue of sparse recovery.

Economics. High-dimensional methods are standard in treatment effect estimation with many potential confounders (double LASSO), forecasting with large predictor sets, and text-as-data applications where p (vocabulary size) far exceeds n.

Linear Algebra. The RIP is a condition on the spectrum of $X^\top X$ restricted to sparse subspaces. Random matrix theory (Marchenko–Pastur law) describes the bulk eigenvalue distribution of $X^\top X/n$ when $p/n \to \gamma > 0$, explaining why classical OLS fails when $\gamma > 1$.

Quantum Mechanics. Compressed sensing is used in quantum state tomography: a rank-r density matrix in d dimensions has $O(rd)$ free parameters, and nuclear norm minimization recovers it from $O(rd \log^2 d)$ measurements—the matrix analogue of sparse recovery via L^1 minimization.

Structural Compression

Invariant

Under sparsity and the RE condition, the LASSO achieves prediction error $O(s \log p/n)$—the minimax optimal rate for s-sparse regression in $\mathbb{R}^p$.

Mechanism

The L^1 penalty restricts the solution to approximately sparse vectors; the RE condition ensures the design matrix distinguishes sparse directions; the union bound over p variables introduces the $\log p$ factor as the search cost in high dimensions.

Cross-System Echo

The oracle inequality is the statistical analogue of compressed sensing recovery guarantees (RIP). Both reduce high-dimensional recovery to conditions on the measurement matrix. The Donoho–Tanner phase transition is a geometric phenomenon arising from the combinatorics of random polytopes, connecting high-dimensional statistics to convex geometry.

Exercises

1. Show that when $p > n$, the OLS system $X^\top X\beta = X^\top Y$ has infinitely many solutions, and the minimum-norm solution $\hat{\beta}^{\mathrm{MN}} = X^\top (XX^\top)^{-1} Y$ has prediction risk that diverges as $p/n \to \gamma > 1$.
2. Verify the stochastic control step: for $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\|X_j\|^2 = n$, show that $\mathbb{P}(\max_j |X_j^\top \varepsilon/n| > t) \leq 2p \exp(-nt^2/(2\sigma^2))$ using the union bound and sub-Gaussian tail bounds.
3. Prove the cone condition: if $\lambda \geq 2 \max_j |X_j^\top \varepsilon/n|$ and $\hat{\beta}$ is the LASSO solution, then $\|\hat{\beta}_{S^c} - \beta_{0,S^c}\|_1 \leq 3\|\hat{\beta}_S - \beta_{0,S}\|_1$.
4. For $n = 500$, $p = 5000$, $s = 10$, $\sigma = 2$, compute the theoretical LASSO penalty λ and the oracle prediction error bound.
5. Show that if X satisfies the RIP of order $2s$ with constant $\delta_{2s} < \sqrt{2} - 1$, then the LASSO recovers β_0 exactly in the noiseless case.
6. Explain why the irrepresentability condition is needed for sign consistency but not for prediction consistency. Construct an example where the LASSO achieves the oracle prediction rate but includes a false positive.
7. Argue, structurally, why the $\log p$ factor in the oracle rate is unavoidable (information-theoretically necessary) by considering the number of possible support sets $\binom{p}{s}$ and the bits of information each observation provides about the support.

Model Averaging

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.6.

Model selection (Nodes 10.2–10.3) commits to a single model, discarding uncertainty about which model generated the data. Model averaging instead combines predictions across all candidates, weighting each by its posterior probability (Bayesian model averaging) or its predictive performance (stacking). This node unifies the Bayesian predictive framework (Node 6.5), Bayes factors (Node 6.6), and information criteria (Node 10.2) into a coherent prediction strategy that accounts for model uncertainty.

Formal Definition

Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be candidate models with prior probabilities $\pi(\mathcal{M}_j)$, $\sum_j \pi(\mathcal{M}_j) = 1$.

Bayesian Model Averaging (BMA). The posterior model probability is

$$\pi(\mathcal{M}_j \mid x) = \frac{m_j(x)\,\pi(\mathcal{M}_j)}{\sum_{k=1}^{K} m_k(x)\,\pi(\mathcal{M}_k)},$$

where $m_j(x) = \int p(x \mid \theta^{(j)}, \mathcal{M}_j)\,\pi(\theta^{(j)} \mid \mathcal{M}_j)\,d\theta^{(j)}$ is the marginal likelihood.

The **BMA predictive distribution** is

$$p(\tilde{x} \mid x) = \sum_{j=1}^{K} \pi(\mathcal{M}_j \mid x)\,p(\tilde{x} \mid x, \mathcal{M}_j).$$

Intuition

Model selection bets everything on one model. If that model is wrong, predictions inherit systematic bias. Model averaging hedges: it spreads probability mass across multiple models in proportion to how well each fits the data (weighted by the prior). When model uncertainty is substantial—no single model clearly dominates—averaging delivers better calibrated predictions and more honest uncertainty quantification than selection.

The cost is interpretability: a weighted average of models is harder to explain than a single selected model.

Structural Role

Dependencies: Builds on information criteria (Node 10.2) for BIC-approximated weights, posterior predictive distributions (Node 6.5), and Bayes factors (Node 6.6) for model comparison. **Enables:** This is a terminal node in Module 10.

Model averaging operationalizes the Bayesian principle that inference should integrate over all sources of uncertainty, including uncertainty about model structure. It connects the computational tools of Nodes 10.2–10.3 to the Bayesian prediction framework, unifying model selection and prediction.

Mathematical Development

BMA Optimality Under Proper Scoring Rules

Theorem. Among all predictive distributions of the form $q(\tilde{x}) = \sum_j w_j \, p(\tilde{x} \mid x, \mathcal{M}_j)$ with $w_j \geq 0$, $\sum_j w_j = 1$, the BMA predictive distribution (with $w_j = \pi(\mathcal{M}_j \mid x)$) minimizes the posterior expected loss under any proper scoring rule.

Proof. A scoring rule $S(q, \tilde{x})$ is **proper** if $\mathbb{E}_{\tilde{x} \sim p}[S(q, \tilde{x})]$ is minimized at $q = p$. The posterior expected score is

$$\mathbb{E}\big[S(q, \tilde{X}) \mid x\big] = \sum_j \pi(\mathcal{M}_j \mid x) \, \mathbb{E}_{\tilde{X} \sim p_j}\big[S(q, \tilde{X})\big] \, .$$

For the **log score** $S(q, \tilde{x}) = -\log q(\tilde{x})$:

$$\mathbb{E}\big[-\log q(\tilde{X}) \mid x\big] = \sum_j \pi(\mathcal{M}_j \mid x) \, \mathbb{E}_{p_j}\big[-\log q(\tilde{X})\big] \, .$$

This is a mixture of KL divergences $D_{\mathrm{KL}}(p_j \| q)$ plus entropy terms. The unique minimizer over mixture distributions is the mixture with weights $w_j = \pi(\mathcal{M}_j \mid x)$, since for any other weights w',

$$\sum_j \pi_j D_{\mathrm{KL}}(p_j \| q_{w'}) \geq \sum_j \pi_j D_{\mathrm{KL}}(p_j \| q_{\pi}),$$

by the convexity of KL divergence and the definition of the posterior mixture. The result extends to all proper scoring rules by the characterization theorem for proper scores.

BIC Approximation to BMA Weights

Substituting the BIC approximation $\log m_j(x) \approx \ell_j - (k_j/2) \log n$ into the posterior model weights:

$$\pi(\mathcal{M}_j \mid x) \approx \frac{\exp(-\mathrm{BIC}_j/2) \, \pi(\mathcal{M}_j)}{\sum_k \exp(-\mathrm{BIC}_k/2) \, \pi(\mathcal{M}_k)}.$$

Under equal priors $\pi(\mathcal{M}_j) = 1/K$, the weights simplify to

$$w_j^{\mathrm{BIC}} = \frac{\exp(-\mathrm{BIC}_j/2)}{\sum_k \exp(-\mathrm{BIC}_k/2)}.$$

This provides a practical BMA implementation without computing marginal likelihoods directly.

Asymptotic Concentration

If the true model $\mathcal{M}^*$ is among the candidates, the BMA weights concentrate:

$$\pi(\mathcal{M}^* \mid x) \to 1 \quad \text{a.s. as } n \to \infty.$$

This follows from the consistency of the Bayes factor: $\log(m_j(x)/m^*(x)) \to -\infty$ for $j \neq *$ (Node 6.6). In finite samples, however, multiple models may have non-negligible weights, and BMA outperforms selection by exploiting this uncertainty.

Frequentist Model Averaging and Stacking

In the frequentist framework, model averaging combines estimators with data-adaptive weights:

$$\tilde{f}(x) = \sum_{j=1}^{K} w_j \, \hat{f}_j(x), \quad w_j \geq 0, \quad \sum_j w_j = 1.$$

Mallows Model Averaging (MMA). Select weights minimizing an estimate of prediction error:

$$\hat{w}^{\text{MMA}} = \arg \min_{w \geq 0, \sum w_j = 1} \left[\|Y - \sum_j w_j \hat{Y}_j\|^2 + 2\sigma^2 \sum_j w_j k_j \right],$$

where k_j is the degrees of freedom of model j. The penalty $2\sigma^2 \sum_j w_j k_j$ corrects for the optimism of in-sample fit, analogous to Mallows' C_p.

Stacking. Select weights minimizing leave-one-out cross-validated prediction error:

$$\hat{w}^{\text{stack}} = \arg \min_{w \geq 0, \sum w_j = 1} \sum_{i=1}^{n} \left(Y_i - \sum_j w_j \, \hat{f}_j^{(-i)}(X_i) \right)^2,$$

where $\hat{f}_j^{(-i)}$ is model j trained without observation i. This is a constrained least squares problem in K weights, solvable by quadratic programming.

Stacking is model-agnostic, requires no marginal likelihood computation, and does not assume that any candidate model is true. It is asymptotically optimal in the sense of minimizing the KL divergence to the true distribution among all convex combinations of the candidate models (Wolpert, 1992).

Model Averaging vs. Model Selection

Criterion	Selection	Averaging
Computational cost	Lower	Higher
Interpretability	Single model	Weighted combination
Risk when one model dominates	Equivalent	Slightly worse (dilution)
Risk under model uncertainty	Higher (wrong model bias)	Lower (hedging)
Uncertainty quantification	Ignores model uncertainty	Propagates model uncertainty

Asymptotically, if a true model exists, both converge to it. In finite samples with substantial model uncertainty, averaging typically has lower prediction MSE.

Worked Examples

Example 1: BMA with Two Nested Models

Model 1: $Y = \alpha + \varepsilon$ (intercept only), $k_1 = 2$, $\ell_1 = -52.3$. Model 2: $Y = \alpha + \beta X + \varepsilon$, $k_2 = 3$, $\ell_2 = -49.1$. $n = 40$, equal priors.

BIC: $\text{BIC}_1 = 104.6 + 2\log 40 = 104.6 + 7.38 = 111.98$; $\text{BIC}_2 = 98.2 + 3\log 40 = 98.2 + 11.07 = 109.27$.

Weights: $w_1 = \exp(-111.98/2)/(\exp(-111.98/2) + \exp(-109.27/2))$.

$\Delta\text{BIC} = 111.98 - 109.27 = 2.71$. So $w_1/w_2 = \exp(-2.71/2) = \exp(-1.355) = 0.258$. Thus $w_2 = 1/(1 + 0.258) = 0.795$, $w_1 = 0.205$.

BMA prediction at new x^*: $\hat{Y}^{\text{BMA}} = 0.205 \cdot \hat{\alpha}_1 + 0.795 \cdot (\hat{\alpha}_2 + \hat{\beta}_2 x^*)$. Model 2 dominates but Model 1 retains 20.5% weight, reflecting non-negligible uncertainty about the predictor's relevance.

Example 2: Stacking Three Regression Models

Models: (A) linear regression with 3 predictors, (B) polynomial regression degree 3, (C) ridge regression with $\lambda = 5$. $n = 60$.

Compute LOO predictions $\hat{f}_A^{(-i)}(X_i)$, $\hat{f}_B^{(-i)}(X_i)$, $\hat{f}_C^{(-i)}(X_i)$ for $i = 1, \ldots, 60$. Form the 60×3 matrix Z with $Z_{ij} = \hat{f}_j^{(-i)}(X_i)$.

Solve: $\min_{w \geq 0, \sum w_j = 1} \|Y - Zw\|^2$.

Suppose the solution is $w_A = 0.35$, $w_B = 0.10$, $w_C = 0.55$. Ridge regression (C) gets the most weight because it has the best leave-one-out predictions. The polynomial model (B) contributes little—its LOO error is high due to overfitting.

For a new observation x^*: $\hat{Y}^{\text{stack}} = 0.35\hat{f}_A(x^*) + 0.10\hat{f}_B(x^*) + 0.55\hat{f}_C(x^*)$.

Cross-Links

AI. Ensemble methods—random forests, boosting, mixture of experts—are frequentist model averaging procedures. Stacking is equivalent to the super-learner algorithm in targeted learning. In deep learning, model ensembles (averaging predictions from multiple trained networks) consistently improve prediction and calibration.

Economics. BMA is used in growth regressions (which variables determine economic growth?), where the set of plausible predictors far exceeds what any single model can support. Forecast combination (averaging multiple economic forecasts) typically outperforms any single forecaster, a robust empirical finding known as the "forecast combination puzzle."

Linear Algebra. The stacking weights solve a constrained least squares problem: $\min_{w \in \Delta^{K-1}} \|Y - Zw\|^2$, where Z is the matrix of leave-one-out predictions. The simplex constraint makes this a quadratic program solvable by active-set or interior-point methods.

Quantum Mechanics. In quantum state estimation, Bayesian model averaging over different state-space dimensions (ranks of the density matrix) produces predictive distributions that account for uncertainty in the state's rank, analogous to BMA over models of different complexity.

Structural Compression

Invariant

BMA predictions are the posterior-weighted average of predictions across models; each model's contribution is proportional to its marginal likelihood times its prior probability.

Mechanism

Marginal likelihoods (or BIC approximations) reweight prior model probabilities by the data's compatibility with each model. BMA is optimal under proper scoring rules because the posterior mixture minimizes the expected KL divergence to the true data-generating process across models. Stacking achieves the analogous frequentist optimality by minimizing cross-validated prediction error over convex combinations.

Cross-System Echo

BMA is the coherent Bayesian solution to model uncertainty. In AI, ensemble methods and mixture-of-experts architectures operationalize the same principle: multiple models combined outperform any single model when uncertainty is non-trivial. Stacking is the super-learner of targeted learning; forecast combination in economics is the applied instantiation.

Exercises

1. Derive the BMA predictive distribution and show it equals $p(\tilde{x} \mid x) = \sum_j \pi(\mathcal{M}_j \mid x)\, p(\tilde{x} \mid x, \mathcal{M}_j)$ starting from the joint posterior $p(\tilde{x}, j \mid x) = p(\tilde{x} \mid x, \mathcal{M}_j)\, \pi(\mathcal{M}_j \mid x)$ and marginalizing over j.
2. Prove that BMA is optimal under the log scoring rule: show that $\mathbb{E}[-\log q(\tilde{X}) \mid x]$ is minimized over mixture distributions $q = \sum_j w_j p_j$ when $w_j = \pi(\mathcal{M}_j \mid x)$.
3. For three models with BIC values $100.0, 102.5, 108.0$ and equal priors, compute the BMA weights. How does the result change if the prior on Model 1 is doubled?
4. Show that as $n \to \infty$, BMA with the true model among candidates concentrates all weight on the true model (i.e., BMA inherits the consistency of Bayes factors).
5. Formulate the stacking optimization as a quadratic program. Write out the objective and constraints for $K = 3$ models and show that the solution lies on the boundary of the simplex when one model strictly dominates the others in LOO performance.
6. Compare BMA and stacking on a concrete example: $K = 2$ models, one correctly specified and one misspecified, with $n = 50$. Show that BMA concentrates weight on the correct model while stacking may assign positive weight to the misspecified model if it improves out-of-sample prediction in the finite sample.
7. Argue, structurally, why model averaging must outperform model selection in expected prediction error when model uncertainty is substantial, and identify the regime where selection is preferable (one model clearly dominates, interpretability is paramount).

Module 11 — Causality

Potential Outcomes Framework

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.1.

The potential outcomes framework (Rubin causal model) is the foundational language for defining causal quantities in statistics. It formalizes what it means for a treatment to cause an effect by indexing outcomes to hypothetical interventions, making explicit the gap between what is observed and what is needed for causal claims. This node defines the estimands (ATE, ATT, CATE), the key assumptions (SUTVA, ignorability, overlap), and the fundamental problem of causal inference. It enables the graphical approach (Node 11.2), identification theory (Node 11.3), and experimental design (Node 11.4).

Formal Definition

For each unit $i \in \{1, \ldots, n\}$ and treatment value $t \in \mathcal{T}$, the **potential outcome** $Y_i(t)$ is the value of the outcome that unit i would exhibit if assigned to treatment t.

For binary treatment $T_i \in \{0, 1\}$:

- $Y_i(1)$: potential outcome under treatment.
- $Y_i(0)$: potential outcome under control.

The **individual treatment effect** (ITE) is

$$\tau_i = Y_i(1) - Y_i(0).$$

The **observed outcome** is

$$Y_i^{\text{obs}} = T_i\, Y_i(1) + (1 - T_i)\, Y_i(0) = Y_i(T_i).$$

Intuition

Causation requires comparing two states of the world: what happened under treatment versus what would have happened under control. The potential outcomes framework takes this literally—it defines both outcomes for every unit, even though only one is ever observed. The unobserved outcome is the **counterfactual**. The fundamental problem of causal inference is that we can never observe both potential outcomes for the same unit, so individual causal effects are never directly identified from data. All causal inference proceeds by making assumptions that link the observable data to the unobservable counterfactual distribution.

Structural Role

Dependencies: Builds on probability spaces (Node 1.3) and estimation theory (Node 3.1). **Enables:** DAGs (Node 11.2) provide a graphical language for the same causal relationships; identifiability (Node 11.3) gives conditions under which causal effects can be recovered; randomized experiments (Node 11.4) achieve identification by design.

The potential outcomes framework and the DAG/structural equation framework (Node 11.2) are complementary formalizations of causality. Potential outcomes define the estimands; DAGs encode the causal structure. The backdoor criterion (Node 11.3) translates between the two.

Mathematical Development

The Fundamental Problem of Causal Inference

For each unit i, only one of $Y_i(0)$ and $Y_i(1)$ is observed. The counterfactual $Y_i(1 - T_i)$ is missing. Therefore:

$$\tau_i = Y_i(1) - Y_i(0) \text{ is never identified from data for any individual } i.$$

Causal inference redirects attention from individual effects to **average effects**, which can be identified under assumptions.

Causal Estimands

Average Treatment Effect (ATE):

$$\tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)].$$

Average Treatment Effect on the Treated (ATT):

$$\tau_{\text{ATT}} = \mathbb{E}[Y(1) - Y(0) \mid T = 1].$$

Conditional Average Treatment Effect (CATE):

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

The ATE is the population-level average; the ATT restricts to the treated subpopulation; the CATE allows treatment effect heterogeneity across covariate values.

SUTVA: Stable Unit Treatment Value Assumption

SUTVA requires:

1. **No interference:** $Y_i(t_1, \ldots, t_n) = Y_i(t_i)$. Unit i's potential outcome depends only on its own treatment, not on the treatments of others.
2. **No hidden variations of treatment:** There is only one version of each treatment level. If $T_i = t$, then $Y_i^{\text{obs}} = Y_i(t)$ regardless of how treatment t was administered.

SUTVA is required for potential outcomes to be well-defined as individual-level quantities. It fails in settings with network effects (vaccination, social programs), where one unit's treatment affects another's outcome.

Ignorability (Unconfoundedness)

Strong ignorability given covariates X:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X.$$

This states that, conditional on observed covariates, treatment assignment is independent of potential outcomes. It is the assumption that there are no unmeasured confounders.

Proof that ignorability identifies the ATE. Under ignorability:

$$\mathbb{E}[Y(t) \mid X] = \mathbb{E}[Y(t) \mid T = t, X] = \mathbb{E}[Y^{\text{obs}} \mid T = t, X],$$

where the first equality uses independence and the second uses consistency ($Y^{\text{obs}} = Y(T)$). Therefore:

$$\tau(x) = \mathbb{E}[Y \mid T = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x],$$

and integrating over X:

$$\tau_{\text{ATE}} = \mathbb{E}_X[\tau(X)] = \mathbb{E}_X\big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\big].$$

Overlap (Positivity)

The **overlap** condition requires

$$0 < e(x) = \mathbb{P}(T = 1 \mid X = x) < 1 \quad \text{for all } x \text{ in the support of } X.$$

Without overlap, some covariate values have no treated (or no control) units, and $\mathbb{E}[Y \mid T = t, X = x]$ is undefined. Overlap ensures every subpopulation contains both treated and control units.

Selection Bias Decomposition

The naive estimator $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$ generally does not equal the ATE:

$$\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = \tau_{\text{ATE}} + \underbrace{\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]}_{\text{Selection bias}}.$$

Selection bias is the difference in baseline (control) outcomes between treated and control groups. It vanishes under ignorability (or randomization) but is generically nonzero in observational studies.

Derivation. Write $\mathbb{E}[Y \mid T = 1] = \mathbb{E}[Y(1) \mid T = 1]$ and $\mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y(0) \mid T = 0]$. Add and subtract $\mathbb{E}[Y(0) \mid T = 1]$:

$$\mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] = \underbrace{\mathbb{E}[Y(1) - Y(0) \mid T = 1]}_{\tau_{\text{ATT}}} + \underbrace{\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0]}_{\text{Selection bias}}.$$

If additionally $\tau_{\text{ATT}} = \tau_{\text{ATE}}$ (homogeneous effects), the decomposition simplifies to the formula above.

Propensity Score

The **propensity score** $e(x) = \mathbb{P}(T = 1 \mid X = x)$ is a balancing score.

Theorem (Rosenbaum–Rubin, 1983). Under ignorability given X,

$$\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid e(X).$$

Proof. It suffices to show $\mathbb{P}(T = 1 \mid Y(0), Y(1), e(X)) = e(X)$. By the law of iterated expectations:

$$\mathbb{P}(T = 1 \mid Y(0), Y(1), e(X)) = \mathbb{E}[\mathbb{P}(T = 1 \mid Y(0), Y(1), X) \mid Y(0), Y(1), e(X)].$$

By ignorability, $\mathbb{P}(T = 1 \mid Y(0), Y(1), X) = \mathbb{P}(T = 1 \mid X) = e(X)$. Since $e(X)$ is measurable with respect to $\sigma(e(X))$, the outer expectation gives $e(X)$.

This reduces the problem from conditioning on the high-dimensional X to conditioning on the scalar $e(X)$, enabling practical estimation (Node 11.5).

Worked Examples

Example 1: Selection Bias in a Job Training Program

A job training program is offered to unemployed workers. Outcome: earnings Y one year later. Treatment: $T = 1$ if enrolled. Observed: $\bar{Y}_1 = \$22{,}000$, $\bar{Y}_0 = \$18{,}000$.

Naive estimate: $\hat{\tau}^{\text{naive}} = \$4{,}000$.

But workers who enrolled may be more motivated (higher $Y(0)$). Suppose the true selection bias is $\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] = \$2{,}500$. Then the true ATT is $\$4{,}000 - \$2{,}500 = \$1{,}500$.

Under ignorability given education and prior earnings X, the ATE is identified by the adjustment formula. If X does not capture motivation, ignorability fails and the naive estimate is biased.

Example 2: Computing ATE Under Ignorability

Binary treatment, binary covariate $X \in \{0, 1\}$ with $\mathbb{P}(X = 1) = 0.6$.

X	$\mathbb{E}[Y \mid T = 1, X]$	$\mathbb{E}[Y \mid T = 0, X]$	$\tau(X)$
0	5.0	3.0	2.0
1	8.0	5.5	2.5

$\tau_{\text{ATE}} = 0.4 \times 2.0 + 0.6 \times 2.5 = 0.8 + 1.5 = 2.3$.

Naive difference in means (without conditioning): suppose $\mathbb{P}(T = 1 \mid X = 1) = 0.7$, $\mathbb{P}(T = 1 \mid X = 0) = 0.3$. Then treated units are disproportionately $X = 1$.

$\mathbb{E}[Y \mid T = 1] = \frac{0.3 \times 0.4 \times 5.0 + 0.7 \times 0.6 \times 8.0}{0.3 \times 0.4 + 0.7 \times 0.6} = \frac{0.6 + 3.36}{0.12 + 0.42} = \frac{3.96}{0.54} = 7.33$.

$\mathbb{E}[Y \mid T = 0] = \frac{0.7 \times 0.4 \times 3.0 + 0.3 \times 0.6 \times 5.5}{0.7 \times 0.4 + 0.3 \times 0.6} = \frac{0.84 + 0.99}{0.28 + 0.18} = \frac{1.83}{0.46} = 3.98$.

Naive estimate: $7.33 - 3.98 = 3.35$, which overstates the true ATE of 2.3 because treated units have higher X values (confounding).

Cross-Links

AI. Counterfactual reasoning in reinforcement learning is the sequential decision analogue: the agent's potential outcomes under different action sequences are the value functions, and the fundamental problem of causal inference becomes the exploration-exploitation tradeoff.

Economics. The Roy model is a structural potential outcomes model with self-selection: agents choose treatment based on their comparative advantage, generating endogenous selection. The potential outcomes framework underpins the credibility revolution in empirical economics (Angrist, Imbens, Rubin).

Linear Algebra. The potential outcomes for n units form the matrix $\mathbf{Y} = [Y_i(t)]_{n \times |\mathcal{T}|}$, of which only one entry per row is observed. The fundamental problem is a missing data problem with a structured missingness pattern determined by the treatment assignment vector.

Quantum Mechanics. In quantum mechanics, counterfactual reasoning arises in the analysis of Bell inequalities: the potential outcomes framework applied to measurement settings formalizes the assumption of local realism, and Bell's theorem shows that no assignment of potential outcomes can reproduce quantum correlations.

Structural Compression

Invariant

Causal estimands are defined on the joint distribution of potential outcomes $(Y(0), Y(1))$; identification requires assumptions linking this unobserved joint distribution to the observed data distribution $p(Y^{\text{obs}}, T, X)$.

Mechanism

Potential outcomes index all counterfactual worlds; observed data indexes only one world per unit. Ignorability $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X$ equates the observational conditional with the interventional distribution. Overlap ensures every covariate stratum has both treated and control units. The propensity score reduces the conditioning dimension from X to the scalar $e(X)$.

Cross-System Echo

The potential outcomes framework is the frequentist formalization of causality; in the structural equation modeling tradition (DAGs, Node 11.2), the same causal quantities are defined via interventions $do(T = t)$. In economics, the Roy model and the Heckman selection model are structural potential outcomes models with endogenous treatment choice. In AI, off-policy evaluation is the sequential analogue of the fundamental problem of causal inference.

Exercises

1. Show that $\tau_{\text{ATE}} = \tau_{\text{ATT}}$ if and only if $\mathbb{E}[Y(1) - Y(0) \mid T = 1] = \mathbb{E}[Y(1) - Y(0) \mid T = 0]$, i.e., the treatment effect is the same for treated and untreated subpopulations.

2. Prove the selection bias decomposition: $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] = \tau_{\text{ATT}} + (\mathbb{E}[Y(0) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0])$.

3. Under ignorability and overlap, derive the inverse probability weighting (IPW) representation: $\tau_{\text{ATE}} = \mathbb{E}\left[\frac{TY}{e(X)} - \frac{(1-T)Y}{1-e(X)}\right]$.

4. Show that the propensity score is a balancing score: $\mathbb{E}[T \mid X, e(X)] = e(X)$, and use this to prove $\mathbb{E}[h(X) \mid T = 1, e(X)] = \mathbb{E}[h(X) \mid T = 0, e(X)]$ for any function h under ignorability.

5. Construct a numerical example with $n = 4$ units where the observed difference in means equals the ATE but individual treatment effects are heterogeneous, illustrating the ecological fallacy in reverse.

6. Consider a setting where SUTVA is violated (e.g., vaccination with herd immunity). Define the potential outcomes $Y_i(\mathbf{t})$ as a function of the full treatment vector $\mathbf{t} = (t_1, \ldots, t_n)$. How many potential outcomes does each unit have? Why does this make identification dramatically harder?

7. Argue, structurally, why the fundamental problem of causal inference—the impossibility of observing both potential outcomes for the same unit—is not a limitation of experimental design but a definitional feature of individual-level causation, and identify the assumptions that convert this individual-level impossibility into a population-level identification result.

Directed Acyclic Graphs

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.2.

Directed acyclic graphs (DAGs) provide the graphical language for encoding causal assumptions, deriving conditional independence relations, and defining interventions. Where the potential outcomes framework (Node 11.1) specifies the estimands, DAGs encode the structural relationships that make identification possible. The d-separation algorithm reads conditional independences from the graph; the do-operator formalizes interventions via graph surgery. This node enables the identifiability and backdoor criterion (Node 11.3).

Formal Definition

A **directed acyclic graph** $\mathcal{G} = (V, E)$ consists of: - A vertex set $V = \{X_1, \ldots, X_p\}$ representing random variables. - A set E of directed edges $X_i \to X_j$, with no directed cycles.

Parents: $\text{pa}(X_j) = \{X_i : X_i \to X_j \in E\}$. **Children:** $\text{ch}(X_j) = \{X_k : X_j \to X_k \in E\}$. **Descendants:** $\text{desc}(X_j)$ = all variables reachable from X_j via directed paths. **Ancestors:** $\text{an}(X_j)$ = all variables from which X_j is reachable via directed paths. **Non-descendants:** $\text{nd}(X_j) = V \setminus (\{X_j\} \cup \text{desc}(X_j))$.

Intuition

A DAG is a map of causal relationships: an arrow $X \to Y$ means X is a direct cause of Y. The absence of an arrow means no direct causal relationship. The graph structure determines which statistical associations are causal, which are confounded, and which are spurious. The power of the DAG framework is that complex causal reasoning reduces to graphical algorithms (d-separation), and interventions reduce to graph surgery (removing arrows).

Structural Role

Dependencies: Builds on conditional independence (Node 1.7) and joint distributions (Node 2.1). **Enables:** Identifiability and the backdoor criterion (Node 11.3), which uses d-separation to determine when causal effects can be estimated from observational data.

DAGs separate the statistical model (the factorization of p) from the causal model (the structural equations and interventions). Statistical dependence is readable from d-separation; causal effects require the interventional distribution, which depends on the graph structure.

Mathematical Development

Markov Factorization

A distribution P is **Markov** with respect to $\mathcal{G}$ if

$$p(x_1, \ldots, x_p) = \prod_{j=1}^{p} p(x_j \mid \text{pa}(x_j)).$$

Local Markov Property. Each variable is conditionally independent of its non-descendants given its parents:

$$X_j \perp\!\!\!\perp \mathrm{nd}(X_j) \mid \mathrm{pa}(X_j).$$

The local Markov property is equivalent to the factorization for positive distributions.

d-Separation: The Algorithm

For disjoint sets $A, B, C \subseteq V$, determine whether A and B are d-separated by C:

1. Enumerate all paths between any node in A and any node in B (paths may traverse edges in either direction).
2. A path is **blocked** by C if it contains at least one of:
   - A **chain** $X_i \to X_k \to X_j$ with $X_k \in C$.
   - A **fork** $X_i \leftarrow X_k \to X_j$ with $X_k \in C$.
   - A **collider** $X_i \to X_k \leftarrow X_j$ with $X_k \notin C$ and no descendant of X_k in C.
3. A and B are d-separated by C if **every** path between A and B is blocked.

Write $A \perp_{\mathcal{G}} B \mid C$ for d-separation.

Key subtlety: Conditioning on a collider (or its descendant) **opens** the path, creating a spurious association. This is the graphical basis of "collider bias" or "Berkson's paradox."

Global Markov Property

Theorem. If P is Markov with respect to $\mathcal{G}$, then

$$A \perp_{\mathcal{G}} B \mid C \implies A \perp\!\!\!\perp B \mid C \quad \text{under } P.$$

d-separation in the graph implies conditional independence in the distribution.

Faithfulness

A distribution P is **faithful** to $\mathcal{G}$ if the converse holds:

$$A \perp\!\!\!\perp B \mid C \quad \text{under } P \implies A \perp_{\mathcal{G}} B \mid C.$$

Faithfulness means the distribution has no "accidental" conditional independences beyond those implied by the graph. It fails on a Lebesgue-measure-zero set of parameters, so it is generically true but can fail in special cases (e.g., exact parameter cancellations).

Under both the Markov property and faithfulness, the conditional independence structure of P is exactly the set of d-separations in $\mathcal{G}$.

Causal DAGs and Structural Equations

A **causal DAG** augments the graphical structure with **structural equations** (also called structural causal models, SCMs):

$$X_j = f_j(\mathrm{pa}(X_j), U_j), \quad j = 1, \ldots, p,$$

where $U_1, \ldots, U_p$ are mutually independent exogenous noise variables. The structural equations define a **data-generating mechanism**: they specify not just the conditional distribution of X_j given its parents, but the functional relationship that would persist under interventions.

The joint distribution implied by the SCM is Markov with respect to $\mathcal{G}$ (the exogenous noise independence implies the factorization).

The do-Operator

The **intervention** $do(X_j = t)$ is defined by:

1. Replace the structural equation for X_j with $X_j = t$ (deterministic).
2. Leave all other structural equations unchanged.

Graphically: remove all edges into X_j (the **mutilated graph** $\mathcal{G}_{\overline{X_j}}$) and set $X_j = t$.

The **interventional distribution** is

$$p(x_1, \ldots, x_p \mid do(X_j = t)) = \prod_{k \neq j} p(x_k \mid \mathrm{pa}(x_k))\Big|_{x_j = t}.$$

This generally differs from the observational conditional $p(x_1, \ldots, x_p \mid X_j = t)$, because conditioning does not remove the causal influence of X_j's parents.

Observational vs. Interventional Distributions

Example. Consider $U \to T \to Y$ and $U \to Y$ (confounding).

Observational: $p(Y \mid T = t) = \int p(Y \mid T = t, U = u)\, p(U \mid T = t)\, du$. The confounder U has a different distribution for different values of T.

Interventional: $p(Y \mid do(T = t)) = \int p(Y \mid T = t, U = u)\, p(U)\, du$. Under intervention, U retains its marginal distribution because the arrow $U \to T$ is severed.

The difference between these two expressions is the essence of confounding.

Markov Equivalence

Two DAGs are **Markov equivalent** if they imply the same set of conditional independences (same d-separations). Markov equivalent DAGs have: - The same skeleton (undirected edges). - The same **v-structures** (immoralities): patterns $X_i \to X_k \leftarrow X_j$ where X_i and X_j are not adjacent.

The set of all DAGs Markov equivalent to $\mathcal{G}$ is the **Markov equivalence class**, represented by a **completed partially directed acyclic graph** (CPDAG). Observational data alone can identify the Markov equivalence class but not the DAG itself (without additional assumptions or interventional data).

Worked Examples

Example 1: d-Separation in a Confounding Structure

DAG: $X \leftarrow U \to Y$, $X \to Y$. Variables: confounder U, treatment X, outcome Y.

Is $X \perp_{\mathcal{G}} Y \mid \emptyset$**?** Path 1: $X \to Y$ (direct, no collider, not blocked). Answer: No.

Is $X \perp_{\mathcal{G}} Y \mid U$**?** Path 1: $X \to Y$ (chain with no middle node in $C = \{U\}$, so not blocked). Answer: No. Conditioning on U blocks the backdoor path $X \leftarrow U \to Y$ but does not block the direct path $X \to Y$.

Example 2: Collider Bias

DAG: $X \to C \leftarrow Y$. No direct path between X and Y.

Is $X \perp_{\mathcal{G}} Y \mid \emptyset$**?** The only path is $X \to C \leftarrow Y$, which contains a collider at C. Since $C \notin \emptyset$, the path is blocked. Answer: Yes, $X \perp_{\mathcal{G}} Y$.

Is $X \perp_{\mathcal{G}} Y \mid C$**?** Now $C \in \{C\}$, so the collider is conditioned on, which **opens** the path. Answer: No, $X \not\perp_{\mathcal{G}} Y \mid C$.

This is Berkson's paradox: conditioning on a common effect creates a spurious association between its causes. For instance, if X = talent and Y = luck, and C = success (which requires one or both), then among successful people, talent and luck appear negatively correlated.

Cross-Links

AI. DAGs are the foundation of Bayesian networks. The d-separation algorithm underlies message-passing (belief propagation) for efficient inference in graphical models. Causal discovery algorithms (PC, GES, FCI) learn DAG structure from data using conditional independence tests.

Economics. Structural equation models in econometrics are causal DAGs with specific functional forms (typically linear). The distinction between structural and reduced-form equations corresponds to the distinction between causal and observational distributions.

Linear Algebra. The adjacency matrix A of a DAG (with $A_{ij} = 1$ if $X_i \to X_j$) is strictly upper-triangular (after topological sorting), hence nilpotent. The reachability matrix $(I - A)^{-1} = \sum_{k=0}^{p-1} A^k$ encodes all ancestor-descendant relationships.

Quantum Mechanics. Quantum causal models extend DAGs to quantum channels, where the edges represent quantum operations rather than classical conditional distributions. The no-signaling principle in quantum mechanics is the analogue of d-separation: spacelike-separated parties are d-separated by their common past.

Structural Compression

Invariant

d-separation in $\mathcal{G}$ implies conditional independence in any distribution Markov with respect to $\mathcal{G}$; under faithfulness, the implication is an equivalence. The interventional distribution is the observational distribution of the mutilated graph.

Mechanism

The Markov factorization decomposes the joint distribution into local conditional distributions, one per variable. Interventions sever parent-child links by replacing a structural equation with a constant, altering the factorization and producing the interventional distribution. d-separation is an algorithmic procedure for reading off conditional independences from the factorization structure.

Cross-System Echo

DAGs are the graphical language of Bayesian networks in AI; d-separation underlies message-passing algorithms (belief propagation) for inference in graphical models. In econometrics, the DAG framework formalizes the concept of exogeneity and the conditions under which regression estimates have causal interpretations. In quantum mechanics, causal structure is formalized by quantum causal models extending DAGs to quantum channels.

Exercises

1. For the DAG $X_1 \to X_3 \leftarrow X_2$, $X_3 \to X_4$, determine all conditional independence relations implied by d-separation (i.e., enumerate all valid $A \perp_{\mathcal{G}} B \mid C$ statements).
2. Show that the Markov factorization $p(x_1, \ldots, x_p) = \prod_j p(x_j \mid \mathrm{pa}(x_j))$ implies the local Markov property $X_j \perp\!\!\!\perp \mathrm{nd}(X_j) \mid \mathrm{pa}(X_j)$.

3. Construct two DAGs on three variables that are Markov equivalent (same skeleton and v-structures) and one that is not. Identify the distinguishing conditional independence.

4. For the DAG $Z \to T \to Y$, $Z \to Y$, compute the interventional distribution $p(Y \mid do(T = t))$ using the truncated factorization and show it differs from $p(Y \mid T = t)$.

5. Prove that conditioning on a collider opens the path: for $X \to C \leftarrow Y$ with $X \perp\!\!\!\perp Y$ marginally, show that $X \not\perp\!\!\!\perp Y \mid C$ when the structural equations are linear with Gaussian noise.

6. Show that the adjacency matrix of a DAG is nilpotent (i.e., $A^p = 0$), and use this to prove that the factorization $p(x) = \prod_j p(x_j \mid \mathrm{pa}(x_j))$ defines a valid joint distribution (no circular dependencies).

7. Argue, structurally, why observational data alone can identify the Markov equivalence class but not the specific DAG, and explain what additional information (interventions, temporal ordering, or functional form assumptions) is needed to distinguish between Markov-equivalent DAGs.

Identifiability and the Backdoor Criterion

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.3.

Identifiability is the question that precedes all estimation: can the causal effect be expressed as a functional of the observed data distribution? This node develops the backdoor criterion (the most commonly applicable graphical condition), the front-door criterion (for unmeasured confounding settings), and the do-calculus (the complete set of rules for converting interventional distributions to observational ones). It connects the potential outcomes framework (Node 11.1) to the DAG framework (Node 11.2) and enables both experimental (Node 11.4) and observational (Node 11.5) estimation.

Formal Definition

Let $\mathcal{G}$ be a causal DAG over (T, Y, X), where T is treatment, Y is outcome, and X is a vector of observed covariates. The target is

$$\tau = \mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)].$$

A causal quantity $Q = Q(p(\cdot \mid do(\cdot)))$ is **identified** from the observational distribution $p(Y, T, X)$ if it can be uniquely computed from $p(Y, T, X)$ for all structural causal models compatible with $\mathcal{G}$.

Non-identifiability arises when two SCMs generate the same observational distribution but imply different values of Q.

Intuition

The causal effect of T on Y is the change in Y when we physically set T to different values (an intervention). But in observational data, T was not set by the researcher—it was determined by a potentially confounded process. Identifiability asks whether the confounding can be removed by conditioning on observed variables. The backdoor criterion gives a graphical test: if we can block all "backdoor paths" (non-causal paths that create spurious association) by conditioning on observed covariates, the causal effect equals the covariate-adjusted observational contrast.

Structural Role

Dependencies: Requires the potential outcomes framework (Node 11.1) for the estimands and DAGs (Node 11.2) for the graphical language. **Enables:** Experimental design (Node 11.4) as the case where identifiability holds by design; observational estimation methods (Node 11.5) which require identification as a precondition.

Identifiability analysis determines whether causal effects can be estimated at all, prior to any estimation procedure. It is the logical prerequisite for all of Nodes 11.4–11.6.

Mathematical Development

Backdoor Paths

A **backdoor path** from T to Y is any path that begins with an arrow into T (i.e., the first edge on the path is $\cdot \to T$). Backdoor paths represent non-causal associations between T and Y—they transmit confounding.

The Backdoor Criterion

Definition. A set of variables X satisfies the **backdoor criterion** relative to (T, Y) in $\mathcal{G}$ if:

1. No variable in X is a descendant of T.
2. X d-separates T from Y in the graph $\mathcal{G}_{\underline{T}}$ (the graph with all arrows out of T removed), or equivalently, X blocks every backdoor path from T to Y.

Theorem: Backdoor Adjustment Formula

Theorem (Pearl, 1993). If X satisfies the backdoor criterion relative to (T, Y), then

$$p(Y = y \mid do(T = t)) = \int p(Y = y \mid T = t, X = x)\, p(X = x)\, dx.$$

Proof sketch. In the mutilated graph $\mathcal{G}_{\overline{T}}$ (arrows into T removed, representing $do(T = t)$):

$$p(y \mid do(t)) = \int p(y \mid t, x, \mathcal{G}_{\overline{T}})\, p(x \mid \mathcal{G}_{\overline{T}})\, dx.$$

By condition 1, X is not a descendant of T, so removing arrows into T does not affect the distribution of X: $p(x \mid \mathcal{G}_{\overline{T}}) = p(x)$.

By condition 2, X blocks all backdoor paths, so in $\mathcal{G}_{\overline{T}}$, $Y \perp\!\!\!\perp$ (removed parents of T) $\mid T, X$. This means the conditional $p(y \mid t, x)$ is the same in the original and mutilated graphs: $p(y \mid t, x, \mathcal{G}_{\overline{T}}) = p(y \mid t, x)$.

Combining: $p(y \mid do(t)) = \int p(y \mid t, x)\, p(x)\, dx$.

The **ATE** is therefore:

$$\tau = \mathbb{E}_X\big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\big].$$

This is the **g-formula** or **adjustment formula**.

Connection to Ignorability

The backdoor criterion in the DAG framework is equivalent to ignorability in the potential outcomes framework:

$$X \text{ satisfies backdoor} \iff \{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X.$$

The DAG makes the structural reason explicit: X blocks all non-causal paths. The potential outcomes framework states the consequence: treatment is as-good-as-random conditional on X.

The Front-Door Criterion

When no measured set satisfies the backdoor criterion (unmeasured confounding between T and Y), identification may still be possible via a **mediator** M.

Definition. A set M satisfies the **front-door criterion** relative to (T, Y) if:

1. T blocks all backdoor paths from M to Y: there are no unblocked backdoor paths from M to Y that do not go through T.
2. M intercepts all directed paths from T to Y: every causal path from T to Y goes through M.
3. There are no backdoor paths from T to M: the effect of T on M is identified without adjustment.

Front-Door Adjustment Formula:

$$p(Y = y \mid do(T = t)) = \sum_{m} p(M = m \mid T = t) \sum_{t'} p(Y = y \mid M = m, T = t')\, p(T = t').$$

Derivation. Apply the do-calculus in two steps:

Step 1: $p(y \mid do(t)) = \sum_m p(y \mid do(t), do(m))\, p(m \mid do(t))$. Since all $T \to Y$ paths go through M, $p(y \mid do(t), do(m)) = p(y \mid do(m))$.

Step 2: $p(m \mid do(t)) = p(m \mid t)$ (no backdoor paths from T to M).

Step 3: $p(y \mid do(m)) = \sum_{t'} p(y \mid m, t')\, p(t')$ (applying the backdoor adjustment with T blocking backdoor paths from M to Y).

Combining: $p(y \mid do(t)) = \sum_m p(m \mid t) \sum_{t'} p(y \mid m, t')\, p(t')$.

Pearl's do-Calculus

The **do-calculus** consists of three rules for manipulating expressions involving $do(\,\cdot\,)$. Let X, Y, Z, W be disjoint sets of variables in a causal DAG $\mathcal{G}$.

Rule 1 (Insertion/deletion of observations):

$$p(y \mid do(x), z, w) = p(y \mid do(x), w) \quad \text{if } Y \perp_{\mathcal{G}_{\overline{X}}} Z \mid W.$$

Rule 2 (Action/observation exchange):

$$p(y \mid do(x), do(z), w) = p(y \mid do(x), z, w) \quad \text{if } Y \perp_{\mathcal{G}_{\overline{X}\underline{Z}}} Z \mid W.$$

Rule 3 (Insertion/deletion of actions):

$$p(y \mid do(x), do(z), w) = p(y \mid do(x), w) \quad \text{if } Y \perp_{\mathcal{G}_{\overline{X}\,\overline{Z(S)}}} Z \mid W,$$

where $Z(S)$ is the set of nodes in Z that are not ancestors of any node in W in $\mathcal{G}_{\overline{X}}$.

Completeness Theorem (Huang–Valtorta, Shpitser–Pearl, 2006). An interventional distribution $p(y \mid do(x))$ is identifiable from observational data if and only if it can be reduced to an observational expression using the three rules of the do-calculus.

Worked Examples

Example 1: Backdoor Adjustment

DAG: $U \to T$, $U \to Y$, $T \to Y$, $X \to U$, $X \to T$. Observed: T, Y, X. Unobserved: U.

Does X satisfy the backdoor criterion for (T, Y)?

Backdoor paths from T to Y: $T \leftarrow U \to Y$ and $T \leftarrow X \to U \to Y$. The path $T \leftarrow U \to Y$: is it blocked by X? The fork $T \leftarrow U \to Y$ requires $U \in X$ to block, but U is unobserved. The path $T \leftarrow X \to U \to Y$: the fork at X is blocked by conditioning on X.

So X blocks one backdoor path but not $T \leftarrow U \to Y$. The backdoor criterion is **not satisfied** by X alone. We cannot identify the causal effect of T on Y by adjusting for X only; the unmeasured confounder U creates a non-blockable backdoor path.

Example 2: Front-Door Criterion Application

DAG: $U \to T$, $U \to Y$, $T \to M \to Y$. All of T, M, Y observed; U unobserved. All directed paths from T to Y go through M. No backdoor paths from T to M (the only candidate $T \leftarrow U \to \cdots \to M$ does not exist since $U \not\to M$). Backdoor paths from M to Y: $M \leftarrow T \leftarrow U \to Y$, blocked by T.

The front-door criterion is satisfied by M. The causal effect is:

$$\mathbb{E}[Y \mid do(T = t)] = \sum_m P(M = m \mid T = t) \sum_{t'} \mathbb{E}[Y \mid M = m, T = t'] \, P(T = t').$$

Suppose $T, M \in \{0, 1\}$, $P(M = 1 \mid T = 1) = 0.8$, $P(M = 1 \mid T = 0) = 0.2$, $P(T = 1) = 0.5$, and the conditional means of Y are: $\mathbb{E}[Y \mid M = 1, T = 1] = 10$, $\mathbb{E}[Y \mid M = 1, T = 0] = 8$, $\mathbb{E}[Y \mid M = 0, T = 1] = 4$, $\mathbb{E}[Y \mid M = 0, T = 0] = 3$.

$\mathbb{E}[Y \mid do(T = 1)] = 0.8[0.5 \cdot 10 + 0.5 \cdot 8] + 0.2[0.5 \cdot 4 + 0.5 \cdot 3] = 0.8 \cdot 9 + 0.2 \cdot 3.5 = 7.2 + 0.7 = 7.9$.

$\mathbb{E}[Y \mid do(T = 0)] = 0.2 \cdot 9 + 0.8 \cdot 3.5 = 1.8 + 2.8 = 4.6$.

Causal effect: $\tau = 7.9 - 4.6 = 3.3$.

Cross-Links

AI. Causal inference in reinforcement learning requires identifying the effect of policies (interventions on action sequences) from offline data. The do-calculus generalizes to sequential settings where the "treatment" is a sequence of actions, and identification requires blocking backdoor paths at each time step.

Economics. The g-formula (adjustment formula) is the semiparametric identification result underlying the potential outcomes approach in economics. Instrumental variables (Node 11.6) handle cases where the backdoor criterion fails due to unmeasured confounders.

Linear Algebra. The adjustment formula $\mathbb{E}[Y \mid do(T = t)] = \sum_x p(y \mid t, x)p(x)$ is a marginalization operation: a matrix-vector product when X is discrete, with the transition matrix $p(y \mid t, x)$ and the marginal distribution $p(x)$ as the vector.

Quantum Mechanics. In quantum causal models, the do-operator corresponds to a quantum intervention (a completely positive trace-preserving map applied to a quantum system), and the analogue of the backdoor criterion determines when quantum causal effects are identifiable from observed quantum correlations.

Structural Compression

Invariant

A causal effect is identified if and only if all non-causal (backdoor) paths between treatment and outcome can be blocked by observed variables; the adjustment formula then expresses the causal effect as a weighted average of the observational conditional.

Mechanism

The backdoor criterion uses d-separation in the mutilated graph to ensure that $p(y \mid t, x) = p(y \mid do(t), x)$. Marginalizing over X with respect to $p(x)$ (not $p(x \mid t)$) removes the confounding. The front-door criterion chains two identification steps through a mediator. The do-calculus provides the complete algorithmic procedure for testing identifiability in arbitrary DAGs.

Cross-System Echo

The g-formula is the causal analogue of the law of total probability. In AI, offline policy evaluation requires identifying the effect of a target policy from data collected under a behavior policy—the sequential

generalization of backdoor adjustment via importance weighting. In economics, the Frisch–Waugh theorem for instrumental variables is the linear-model specialization of identification under confounding.

Exercises

1. For the DAG $Z \to T \to Y$, $Z \to Y$, verify that Z satisfies the backdoor criterion for (T, Y) and write out the adjustment formula.
2. Prove the backdoor adjustment formula by showing that in the mutilated graph $\mathcal{G}_{\overline{T}}$, $p(y \mid t, x, \mathcal{G}_{\overline{T}}) = p(y \mid t, x)$ and $p(x \mid \mathcal{G}_{\overline{T}}) = p(x)$ when X satisfies the backdoor criterion.
3. Construct a DAG where X contains a descendant of T, and show that conditioning on X biases the estimate of the causal effect (an instance of "bad control").
4. Apply the front-door criterion to the smoking–tar–cancer DAG (smoking $\to$ tar deposits $\to$ cancer, with an unmeasured common cause of smoking and cancer) and derive the front-door adjustment formula.
5. Using the three rules of the do-calculus, reduce $p(y \mid do(t))$ to an observational expression for the DAG $T \to M \to Y$, $U \to T$, $U \to Y$, verifying the front-door formula.
6. Show that for the DAG $T \to Y \leftarrow U \to T$ with U unobserved and no other variables, the causal effect $\mathbb{E}[Y \mid do(T = t)]$ is not identified. Explain why the do-calculus fails.
7. Argue, structurally, why completeness of the do-calculus means that identifiability is a purely graphical question—it depends on the DAG structure and which variables are observed, not on the functional forms or parameter values of the structural equations.

Randomized Experiments

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.4.

Randomized experiments are the gold standard for causal identification: they make ignorability hold by design, eliminating confounding without requiring knowledge of the causal structure. This node develops the theory of completely randomized experiments, including the Neyman framework for ATE estimation, Fisher's exact randomization test, and covariate adjustment via ANCOVA. It operationalizes the potential outcomes framework (Node 11.1) under the strongest possible identification conditions, and sets the benchmark against which observational methods (Node 11.5) are evaluated.

Formal Definition

In a **completely randomized experiment (CRE)**, n units are assigned to treatment ($T_i = 1$) or control ($T_i = 0$) by a known random mechanism, with n_1 treated and $n_0 = n - n_1$ control units. The assignment satisfies:

$$\{Y_i(0), Y_i(1)\}_{i=1}^{n} \perp\!\!\!\perp (T_1, \ldots, T_n),$$

and the assignment mechanism is fully specified: $\mathbb{P}(\mathbf{T} = \mathbf{t}) = \binom{n}{n_1}^{-1}$ for all $\mathbf{t}$ with $\sum t_i = n_1$.

Intuition

Randomization physically severs all arrows into the treatment variable in the causal DAG. No matter how complex the confounding structure, random assignment ensures that treated and control groups are comparable on average—both on observed and unobserved characteristics. The beauty of randomization is that identification follows from the design, not from modeling assumptions. The price is feasibility: many causal questions cannot be answered by experiments (ethics, cost, logistics).

Structural Role

Dependencies: Builds on the potential outcomes framework (Node 11.1) for the estimand definitions and hypothesis testing (Node 5.1) for the inferential framework. **Enables:** Observational methods (Node 11.5) by providing the ideal benchmark against which observational designs are compared.

Randomization removes all backdoor paths by design (Node 11.3). Fisher's test provides exact finite-sample inference; Neyman's framework provides asymptotically valid confidence intervals. ANCOVA leverages pre-treatment covariates for precision gain without risking bias.

Mathematical Development

Identification Under Randomization

Under CRE and SUTVA:

$$\mathbb{E}[Y \mid T = 1] = \mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}[Y(1)],$$

where the last equality uses independence. Similarly $\mathbb{E}[Y \mid T = 0] = \mathbb{E}[Y(0)]$. Therefore:

$$\tau_{\text{ATE}} = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0].$$

This requires no modeling, no covariate adjustment, and no distributional assumptions. It follows entirely from the randomization mechanism.

Neyman's Repeated Sampling Framework

The **difference-in-means** estimator is

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} Y_i.$$

Theorem (Neyman, 1923). Under CRE:

$$\mathbb{E}[\hat{\tau}] = \tau_{\text{ATE}}.$$

Proof. $\mathbb{E}[\bar{Y}_1] = \mathbb{E}\Big[\frac{1}{n_1} \sum_{i:T_i=1} Y_i(1)\Big]$. Under random assignment, each unit i has probability n_1/n of being treated, and $\mathbb{E}[\sum_{i:T_i=1} Y_i(1)] = \frac{n_1}{n} \sum_{i=1}^{n} Y_i(1)$, so $\mathbb{E}[\bar{Y}_1] = \frac{1}{n} \sum_i Y_i(1) = \bar{Y}(1)$. Similarly $\mathbb{E}[\bar{Y}_0] = \bar{Y}(0)$. Therefore $\mathbb{E}[\hat{\tau}] = \bar{Y}(1) - \bar{Y}(0) = \tau_{\text{ATE}}$.

Variance of $\hat{\tau}$:

$$\text{Var}(\hat{\tau}) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} - \frac{S_\tau^2}{n},$$

where $S_t^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(t) - \bar{Y}(t))^2$ and $S_\tau^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tau_i - \bar{\tau})^2$ is the variance of individual treatment effects.

The term S_τ^2/n involves both potential outcomes and is **not estimable** from data (the fundamental problem). The **conservative variance estimator** is

$$\hat{V} = \frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}, \quad s_t^2 = \frac{1}{n_t - 1} \sum_{i:T_i=t} (Y_i - \bar{Y}_t)^2,$$

which overestimates $\text{Var}(\hat{\tau})$ by $S_\tau^2/n \geq 0$. Under constant treatment effects ($\tau_i = \tau$ for all i), $S_\tau^2 = 0$ and the estimator is exact.

Fisher's Randomization Test

Sharp null hypothesis:

$$H_0^F : Y_i(1) = Y_i(0) \quad \text{for all } i = 1, \ldots, n.$$

Under H_0^F, both potential outcomes equal Y_i^{obs}, so the complete data table is known. For any test statistic $T(\mathbf{T}, \mathbf{Y}^{\text{obs}})$, the exact null distribution is obtained by enumerating all $\binom{n}{n_1}$ possible treatment assignments:

$$p\text{-value} = \frac{|\{\mathbf{t} : T(\mathbf{t}, \mathbf{Y}^{\text{obs}}) \geq T_{\text{obs}}\}|}{\binom{n}{n_1}}.$$

Properties: - Exact for any sample size (no asymptotic approximation). - No distributional assumptions on Y. - Valid for any test statistic (difference in means, rank statistics, Kolmogorov–Smirnov, etc.). - Tests the sharp null (no effect for any unit), which is stronger than Neyman's null ($\tau_{\text{ATE}} = 0$).

Computational note. For large n, exact enumeration is infeasible. Monte Carlo approximation: sample B random permutations and compute the proportion exceeding T_{obs}. This gives a randomization p-value with precision $O(1/\sqrt{B})$.

Covariate Adjustment: ANCOVA

Even under randomization, pre-treatment covariates X can improve precision.

Lin's ANCOVA estimator. Fit the regression:

$$Y_i = \alpha + \tau T_i + \gamma^\top X_i + \delta^\top (T_i \cdot (X_i - \bar{X})) + \varepsilon_i,$$

where the interaction $T_i \cdot (X_i - \bar{X})$ allows different slopes in treated and control groups. The coefficient $\hat{\tau}^{\text{Lin}}$ on T_i is the ANCOVA estimator.

Theorem (Lin, 2013). Under randomization, $\hat{\tau}^{\text{Lin}}$ is:

1. Consistent for τ_{ATE} regardless of whether the linear model is correctly specified.
2. Asymptotically at least as efficient as the unadjusted $\hat{\tau}$.
3. Strictly more efficient when X is predictive of Y.

Why adjustment does not bias. Under randomization, $T \perp\!\!\!\perp X$, so including X as a covariate does not introduce confounding. The regression coefficient on T estimates the same causal effect with or without covariates; the improvement is purely in variance.

Precision gain. If R^2 is the fraction of outcome variance explained by X, the ANCOVA variance is approximately $(1 - R^2) \cdot \text{Var}(\hat{\tau})$. For $R^2 = 0.5$, the effective sample size doubles.

Stratified (Block) Randomization

In **stratified randomization**, units are grouped into strata based on X, and randomization occurs within each stratum. This guarantees exact covariate balance and further reduces variance.

The stratified estimator is

$$\hat{\tau}^{\text{strat}} = \sum_s \frac{n_s}{n} \hat{\tau}_s,$$

where $\hat{\tau}_s = \bar{Y}_{1,s} - \bar{Y}_{0,s}$ is the within-stratum difference in means and n_s is the stratum size. This is guaranteed to have $\text{Var}(\hat{\tau}^{\text{strat}}) \leq \text{Var}(\hat{\tau})$.

Worked Examples

Example 1: Neyman Inference for a Drug Trial

$n = 200$ patients randomized equally: $n_1 = n_0 = 100$. Outcome: blood pressure reduction (mmHg).

Observed: $\bar{Y}_1 = 12.3$, $\bar{Y}_0 = 8.7$, $s_1^2 = 25.0$, $s_0^2 = 20.0$.

$\hat{\tau} = 12.3 - 8.7 = 3.6$ mmHg. $\hat{V} = 25/100 + 20/100 = 0.45$. SE $= \sqrt{0.45} = 0.671$.

95% CI: $3.6 \pm 1.96 \times 0.671 = (2.28, 4.92)$. The CI is conservative (it overestimates the true variance by omitting S_τ^2/n).

Test $H_0 : \tau = 0$: $z = 3.6/0.671 = 5.37$, $p < 0.001$. Strong evidence of a treatment effect.

Example 2: Fisher Randomization Test

$n = 8$, $n_1 = 4$. Observed outcomes: treated = {7, 9, 6, 8}, control = {3, 5, 4, 2}.

$T_{\text{obs}} = \bar{Y}_1 - \bar{Y}_0 = 7.5 - 3.5 = 4.0$.

Under H_0^F, all 8 outcomes are fixed. There are $\binom{8}{4} = 70$ possible assignments.

For each permutation, compute $\bar{Y}_1 - \bar{Y}_0$. Count the number of permutations yielding ≥ 4.0.

The maximum possible difference (assigning {7, 8, 9, 6} vs. {2, 3, 4, 5}) is $7.5 - 3.5 = 4.0$, but other permutations achieving this: {9, 8, 7, 6} vs. {5, 4, 3, 2} gives 4.0. After enumeration, suppose 2 out of 70 permutations yield ≥ 4.0.

p-value $= 2/70 = 0.029$. Reject H_0^F at the 5% level.

Cross-Links

AI. A/B testing in product development is randomized experimentation with the ATE as the estimand. Multi-armed bandit algorithms extend the experimental framework to sequential settings where treatment assignment adapts based on interim results, trading off exploration and exploitation.

Economics. Randomized controlled trials (RCTs) are the basis for the credibility revolution in program evaluation (Angrist, Duflo, Banerjee). Field experiments in development economics randomize at the village or school level (cluster randomization), requiring cluster-robust variance estimators.

Linear Algebra. The design matrix of a CRE has a specific structure: T is a binary vector with exactly n_1 ones. The hat matrix for the unadjusted estimator is $H = \text{diag}(T/n_1 + (1-T)/n_0)$, and the ANCOVA hat matrix is the orthogonal projection onto the column space of $[T, X, T \cdot X]$.

Quantum Mechanics. Bell tests are randomized experiments in physics: measurement settings (treatments) are randomly assigned to entangled particles, and the outcome statistics test whether local hidden variable models (the causal DAG with unmeasured common cause) can explain the observed correlations. Randomization of measurement settings is essential to close the "freedom of choice" loophole.

Structural Compression

Invariant

Randomization renders treatment independent of all potential outcomes; the observed difference in means is an unbiased estimator of the ATE requiring no model for the outcome and no knowledge of the confounding structure.

Mechanism

Physical randomization removes all backdoor paths between T and Y in the causal DAG. The known assignment mechanism enables exact permutation inference (Fisher) and conservative variance estimation (Neyman). Covariate adjustment reduces variance without introducing bias because $T \perp\!\!\!\perp X$ by design.

Cross-System Echo

Fisher's randomization test is the prototype of permutation tests in nonparametric statistics. In AI, A/B testing is the standard experimental paradigm for causal evaluation of algorithmic changes. In economics, RCTs underpin the credibility revolution. In physics, Bell tests use randomized measurement settings to test fundamental causal structure of quantum mechanics.

Exercises

1. Prove that the difference-in-means estimator $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$ is unbiased for τ_{ATE} under completely randomized assignment. State explicitly where independence of treatment and potential outcomes is used.

2. Derive the Neyman variance formula $\text{Var}(\hat{\tau}) = S_1^2/n_1 + S_0^2/n_0 - S_\tau^2/n$. Show that omitting S_τ^2/n yields a conservative estimator.

3. For $n = 6$, $n_1 = 3$, outcomes $= \{10, 4, 7, 3, 8, 5\}$ (first 3 treated), compute the Fisher exact p-value for the difference-in-means test statistic under the sharp null. Enumerate all $\binom{6}{3} = 20$ permutations.

4. Show that under constant treatment effects ($\tau_i = \tau$ for all i), the Neyman variance formula simplifies to $\text{Var}(\hat{\tau}) = (S_0^2/n)(1/n_1 + 1/n_0)$, and the conservative estimator is exact.

5. Prove Lin's result: under randomization, the ANCOVA estimator with treatment-covariate interactions is consistent for τ_{ATE} even when the linear model is misspecified. (Hint: use $T \perp\!\!\!\perp X$ and the law of iterated expectations.)

6. Compare the variance of the unadjusted estimator, the ANCOVA estimator, and the stratified estimator for an experiment with $n = 100$, two strata of equal size, and $R^2 = 0.4$. Quantify the efficiency gain from each adjustment.

7. Argue, structurally, why randomization is the only design that achieves identification of the ATE without assumptions about the functional form of the outcome model or the set of confounders, and explain why observational methods (Node 11.5) can never fully replicate this property.

Observational Studies and Treatment Effect Estimation

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.5.

When randomization is infeasible, observational studies must rely on identification assumptions to estimate causal effects. This node develops three major classes of methods: (1) propensity score methods under ignorability (IPW, matching), (2) instrumental variables when ignorability fails, and (3) quasi-experimental designs (difference-in-differences, regression discontinuity) that exploit specific structural features of the data-generating process. Each method identifies a specific estimand for a specific subpopulation under distinct assumptions. The doubly robust estimator unifies propensity score and outcome modeling approaches.

Formal Definition

In observational studies, treatment T_i may depend on covariates X_i and unobserved factors. Under **strong ignorability** $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X$ and **overlap** $0 < e(x) < 1$, where

$$e(x) = \mathbb{P}(T = 1 \mid X = x)$$

is the **propensity score**, the ATE is identified via the backdoor formula (Node 11.3).

When ignorability fails (unmeasured confounding), additional structure is needed: instruments, panel variation, or threshold rules.

Intuition

Observational studies face the core challenge that treated and control groups differ systematically. Propensity score methods attempt to recreate the conditions of a randomized experiment by reweighting or matching observations so that covariate distributions are balanced. Instrumental variables circumvent confounding by finding an external source of variation that affects treatment but not the outcome directly. Regression discontinuity exploits a known threshold rule to create local randomization. Each approach trades off generality (what estimand, what subpopulation) against the strength of required assumptions.

Structural Role

Dependencies: Requires identification theory (Node 11.3) as a logical prerequisite, randomized experiments (Node 11.4) as the benchmark, large-sample theory (Node 7.5), and regression (Node 8.6). **Enables:** Policy evaluation (Node 11.6), which extends these methods to decision rules and sequential settings.

All methods in this node require assumptions that are fundamentally untestable from observational data alone. The choice between methods depends on which assumptions are most credible in the specific application.

Mathematical Development

Propensity Score Methods

Theorem (Rosenbaum–Rubin, 1983). Under ignorability given X, $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid e(X)$.

This reduces conditioning from the potentially high-dimensional X to the scalar $e(X)$, enabling practical estimation.

Inverse Probability Weighting (IPW)

The **Horvitz–Thompson IPW estimator** of the ATE is

$$\hat{\tau}^{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}\right).$$

Derivation. Under ignorability:

$$\mathbb{E}\left[\frac{TY}{e(X)}\right] = \mathbb{E}\left[\frac{TY(1)}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{T}{e(X)} \mid X\right] Y(1)\right] = \mathbb{E}[Y(1)],$$

since $\mathbb{E}[T/e(X) \mid X] = e(X)/e(X) = 1$. Similarly for $\mathbb{E}[(1-T)Y/(1-e(X))] = \mathbb{E}[Y(0)]$.

Hajek (normalized) estimator: $\hat{\tau}^{\mathrm{H}} = \frac{\sum_i T_i Y_i/\hat{e}(X_i)}{\sum_i T_i/\hat{e}(X_i)} - \frac{\sum_i (1-T_i)Y_i/(1-\hat{e}(X_i))}{\sum_i (1-T_i)/(1-\hat{e}(X_i))}$. This is more stable (lower variance) than the Horvitz–Thompson version.

Sensitivity to extreme propensity scores. When $e(x)$ is near 0 or 1, the weights $1/e(x)$ or $1/(1-e(x))$ explode, inflating variance. **Trimming** (excluding units with $e(x) < \epsilon$ or $e(x) > 1-\epsilon$) or **truncation** (capping weights) mitigates this at the cost of changing the estimand.

Outcome Regression

The **outcome regression (OR) estimator** models $\mu_t(x) = \mathbb{E}[Y \mid T = t, X = x]$ and computes

$$\hat{\tau}^{\mathrm{OR}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)].$$

This is consistent under correct specification of $\mu_t(x)$ and does not require modeling the propensity score. However, it is sensitive to extrapolation in regions where treatment and control groups do not overlap in covariate space.

Doubly Robust (DR) Estimation

The **augmented IPW (AIPW)** estimator combines both approaches:

$$\hat{\tau}^{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)}\right].$$

Double robustness. $\hat{\tau}^{\mathrm{DR}}$ is consistent if either $\hat{e}$ or $\hat{\mu}_t$ is correctly specified (but not necessarily both).

Proof sketch. Write $\hat{\tau}^{\mathrm{DR}} = \hat{\tau}^{\mathrm{OR}} + \text{correction}$. If $\hat{\mu}_t$ is correct, the residuals $Y_i - \hat{\mu}_t(X_i)$ have conditional mean zero and the correction vanishes. If $\hat{e}$ is correct, the IPW component identifies the ATE independently. The correction term is the product of estimation errors in both models, which is $o_p(1/\sqrt{n})$ when either model is correct.

Semiparametric efficiency. The DR estimator achieves the semiparametric efficiency bound for the ATE under the nonparametric model. Its influence function is

$$\phi(Y, T, X) = \frac{T(Y - \mu_1(X))}{e(X)} - \frac{(1-T)(Y - \mu_0(X))}{1-e(X)} + \mu_1(X) - \mu_0(X) - \tau,$$

and the efficiency bound is $\mathrm{Var}(\phi)/n$.

Matching

Nearest-neighbor matching constructs a matched control for each treated unit:

$$\hat{\tau}^{\text{match}} = \frac{1}{n_1} \sum_{i:T_i=1} \big(Y_i - Y_{j(i)}\big), \quad j(i) = \arg\min_{j:T_j=0} \|X_i - X_j\|.$$

This estimates the ATT. Matching on the propensity score $e(X)$ is valid under ignorability and avoids the curse of dimensionality in X-space.

Bias of matching. With a fixed number of matches M, the bias is $O(n^{-1/d})$ in d dimensions. Bias correction (Abadie–Imbens, 2011) adjusts for the remaining difference between matched and target covariate values.

Instrumental Variables (IV)

When ignorability fails, an **instrument** Z provides identification if:

1. Z affects T (**relevance**).
2. Z affects Y only through T (**exclusion restriction**).
3. Z is independent of confounders (**independence**).

Wald estimator (binary Z, binary T):

$$\hat{\tau}^{\text{IV}} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{T}_{Z=1} - \bar{T}_{Z=0}} = \frac{\text{Reduced form}}{\text{First stage}}.$$

Two-stage least squares (2SLS). Stage 1: regress T on Z and covariates to get $\hat{T}$. Stage 2: regress Y on $\hat{T}$ and covariates. The coefficient on $\hat{T}$ is $\hat{\tau}^{\text{2SLS}}$.

Under monotonicity ($T_i(1) \geq T_i(0)$ for all i), IV identifies the **LATE** (Local Average Treatment Effect) for compliers:

$$\tau^{\text{LATE}} = \mathbb{E}[Y(1) - Y(0) \mid T(1) > T(0)].$$

Difference-in-Differences (DiD)

With panel data (units observed before and after treatment):

$$\hat{\tau}^{\text{DiD}} = (\bar{Y}_{1,\text{post}} - \bar{Y}_{1,\text{pre}}) - (\bar{Y}_{0,\text{post}} - \bar{Y}_{0,\text{pre}}).$$

Parallel trends assumption: absent treatment, treated and control units would have followed the same time trend: $\mathbb{E}[Y_i(0)_{\text{post}} - Y_i(0)_{\text{pre}} \mid T = 1] = \mathbb{E}[Y_i(0)_{\text{post}} - Y_i(0)_{\text{pre}} \mid T = 0]$.

Under parallel trends, DiD identifies the ATT. The assumption is untestable but can be assessed by checking pre-treatment trend similarity.

Regression Discontinuity (RD)

Treatment is assigned by a running variable R_i crossing a threshold c: $T_i = \mathbf{1}(R_i \geq c)$ (sharp RD).

$$\tau^{\text{RD}} = \lim_{r \downarrow c} \mathbb{E}[Y \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[Y \mid R = r].$$

This identifies the ATE at the threshold $R = c$, under continuity of potential outcome means.

Fuzzy RD. When T is not a deterministic function of R but exhibits a discontinuity in $e(R)$ at c, the fuzzy RD estimand is

$$\tau^{\text{FRD}} = \frac{\lim_{r \downarrow c} \mathbb{E}[Y \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[Y \mid R = r]}{\lim_{r \downarrow c} \mathbb{E}[T \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[T \mid R = r]},$$

which is a local Wald (IV) estimator using the threshold as the instrument.

Worked Examples

Example 1: IPW Estimation

Binary treatment, 3 covariate strata. Data:

Stratum	n	n_1	$\bar{Y}_1$	$\bar{Y}_0$	e
A	100	70	15	10	0.7
B	200	100	12	9	0.5
C	100	20	18	14	0.2

IPW estimate of $\mathbb{E}[Y(1)]$: Weight treated observations by $1/e$. Within each stratum, $\sum T_i Y_i / e = n_1 \bar{Y}_1 / e$. Normalized:

$\hat{\mathbb{E}}[Y(1)] = \frac{70 \cdot 15/0.7 + 100 \cdot 12/0.5 + 20 \cdot 18/0.2}{70/0.7 + 100/0.5 + 20/0.2} = \frac{1500+2400+1800}{100+200+100} = \frac{5700}{400} = 14.25.$

$\hat{\mathbb{E}}[Y(0)] = \frac{30 \cdot 10/0.3 + 100 \cdot 9/0.5 + 80 \cdot 14/0.8}{30/0.3 + 100/0.5 + 80/0.8} = \frac{1000+1800+1400}{100+200+100} = \frac{4200}{400} = 10.5.$

$\hat{\tau}^{\text{IPW}} = 14.25 - 10.5 = 3.75.$

Naive (unadjusted): $\hat{\tau}^{\text{naive}} = \frac{70 \cdot 15 + 100 \cdot 12 + 20 \cdot 18}{190} - \frac{30 \cdot 10 + 100 \cdot 9 + 80 \cdot 14}{210} = 13.05 - 11.33 = 1.72$. The naive estimate is biased because treatment is more likely in stratum A (high e) which has lower outcome.

Example 2: Regression Discontinuity

Test score threshold for a scholarship: $c = 75$. Running variable: test score R. Treatment: scholarship ($T = \mathbf{1}(R \geq 75)$). Outcome: college GPA.

Fit local linear regressions on each side of the cutoff using a bandwidth $h = 5$:

Left of cutoff ($R \in [70, 75)$): $\hat{\mathbb{E}}[Y \mid R = 75^-] = 2.80$. Right of cutoff ($R \in [75, 80]$): $\hat{\mathbb{E}}[Y \mid R = 75^+] = 3.15$.

$\hat{\tau}^{\text{RD}} = 3.15 - 2.80 = 0.35$ GPA points.

This is the causal effect of the scholarship for students at the threshold. Validity requires that students cannot precisely manipulate their test scores to cross the threshold (no sorting), so that the density of R is continuous at c. A McCrary density test checks this.

Cross-Links

AI. Off-policy evaluation in reinforcement learning uses IPW and doubly robust estimators to estimate the value of a target policy from data collected under a different behavior policy. This is the sequential generalization of observational causal inference, where the "treatment" is a sequence of actions.

Economics. IV estimation is the workhorse of empirical economics (Angrist–Imbens Nobel, 2021). DiD is standard in policy evaluation (Card–Krueger minimum wage study). RD is used in education (test score thresholds), healthcare (age cutoffs for eligibility), and political science (close elections).

Linear Algebra. 2SLS can be written as a projection: $\hat{\beta}^{\text{2SLS}} = (X^\top P_Z X)^{-1} X^\top P_Z Y$, where $P_Z = Z(Z^\top Z)^{-1} Z^\top$ is the projection onto the instrument space. The first stage is the projection of T onto the column space of Z.

Quantum Mechanics. In quantum causal inference, unmeasured confounders correspond to entangled quantum states shared between parties. Instrumental variable-type arguments (using random measurement settings as instruments) are used to bound causal effects in the presence of quantum confounding, with Bell inequalities as the key identification constraints.

Structural Compression

Invariant

Under ignorability and overlap, the ATE is identified and the DR estimator achieves the semiparametric efficiency bound. When ignorability fails, IV identifies the LATE for compliers, DiD identifies the ATT under parallel trends, and RD identifies the threshold ATE under continuity.

Mechanism

IPW reweights the observed distribution to remove confounding; outcome regression models the conditional expectation surface; the DR estimator combines both and is consistent under misspecification of either nuisance model. IV exploits exogenous variation; DiD exploits time-series variation; RD exploits threshold rules. Each design replaces ignorability with a different structural assumption.

Cross-System Echo

Doubly robust estimation is a special case of semiparametric efficiency theory (Bickel–Klaassen–Ritov–Wellner). In AI, off-policy evaluation uses these estimators for distribution shift correction. In economics, the hierarchy IV > DiD > RD > selection-on-observables reflects decreasing reliance on untestable assumptions, and the credibility revolution is built on matching research designs to credible identification strategies.

Exercises

1. Derive the IPW estimator by showing that $\mathbb{E}[TY/e(X)] = \mathbb{E}[Y(1)]$ under ignorability and overlap. State where each assumption is used.
2. Show that the DR estimator is consistent when the propensity score model is correctly specified but the outcome model is misspecified. Then show the converse.
3. For the Wald estimator $\hat{\tau}^{\text{IV}} = (\bar{Y}_{Z=1} - \bar{Y}_{Z=0})/(\bar{T}_{Z=1} - \bar{T}_{Z=0})$, derive the asymptotic distribution using the delta method and compute the standard error.
4. State the parallel trends assumption for DiD formally using potential outcomes notation. Construct a numerical example where parallel trends holds and DiD recovers the ATT, and another where it fails.
5. Show that the sharp RD estimand $\tau^{\text{RD}} = \lim_{r \downarrow c} \mathbb{E}[Y(1) \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[Y(0) \mid R = r]$ equals the ATE at the threshold under continuity of $\mathbb{E}[Y(t) \mid R = r]$ at $r = c$.
6. Compare the assumptions and estimands of IPW, IV, DiD, and RD in a table. For each method, give one example where it is the appropriate design and explain why the alternatives are not credible.
7. Argue, structurally, why no observational method can replicate the unconditional guarantees of a randomized experiment, and identify the specific untestable assumption that each method introduces as the price for working without randomization.

Policy Evaluation and Structural Identification

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.6.

This capstone node extends causal inference from estimating treatment effects to evaluating decision rules (policies). It develops inverse propensity weighting for policy evaluation, the doubly robust AIPW estimator, and the semiparametric efficiency bound for the ATE. The connection to structural equation models links causal inference to mechanism design and reinforcement learning. It synthesizes the entire causality module: potential outcomes (Node 11.1), DAGs (Node 11.2), identification (Node 11.3), experiments (Node 11.4), and observational methods (Node 11.5) into a unified framework for evaluating interventions in complex systems.

Formal Definition

A **policy** (or treatment rule) is a mapping $\pi : \mathcal{X} \to \{0, 1\}$ (deterministic) or $\pi : \mathcal{X} \to [0, 1]$ (stochastic) that assigns treatment based on observed covariates.

The **value** of policy π is

$$V(\pi) = \mathbb{E}[Y(\pi(X))] = \mathbb{E}[\pi(X)\, Y(1) + (1 - \pi(X))\, Y(0)].$$

The **optimal policy** maximizes the value:

$$\pi^* = \arg\max_{\pi} V(\pi).$$

Under ignorability and overlap, the value is identified:

$$V(\pi) = \mathbb{E}[\mathbb{E}[Y \mid T = 1, X]\, \pi(X) + \mathbb{E}[Y \mid T = 0, X]\, (1 - \pi(X))]\,.$$

Intuition

Treatment effect estimation asks: what is the average effect of treatment? Policy evaluation asks a more refined question: who should be treated? The optimal policy treats unit i if and only if $\tau(X_i) = \mathbb{E}[Y(1) - Y(0) \mid X = X_i] > 0$. This connects causal inference to statistical decision theory: the policy is a decision rule, and the value function is the expected utility. The challenge is estimating the value of a proposed policy from data collected under a different (possibly non-randomized) treatment assignment.

Structural Role

Dependencies: Builds on observational treatment estimation (Node 11.5) for the estimators, regression (Node 8.3) for outcome modeling, and decision theory (Node 6.3) for the policy optimization framework. **Enables:** This is the terminal node of Module 11 and the statistics course, connecting to reinforcement learning in AI and mechanism design in economics.

Policy evaluation is the bridge between causal inference and decision-making. It transforms estimates of causal effects into actionable rules, closing the loop from data to intervention.

Mathematical Development

Inverse Propensity Weighted Policy Evaluation

Given data $\{(X_i, T_i, Y_i)\}_{i=1}^n$ collected under a **behavior policy** $b(x) = \mathbb{P}(T = 1 \mid X = x)$, the value of a target policy π is estimated by reweighting:

$$\hat{V}^{\text{IPW}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}(T_i = \pi(X_i))}{T_i\, b(X_i) + (1 - T_i)(1 - b(X_i))}\, Y_i.$$

For a stochastic policy:

$$\hat{V}^{\text{IPW}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(X_i)}{b(X_i)}\, T_i\, Y_i + \frac{1 - \pi(X_i)}{1 - b(X_i)}\, (1 - T_i)\, Y_i.$$

This is unbiased for $V(\pi)$ when $b(x)$ is known, and consistent when $b(x)$ is estimated consistently. The variance is high when π and b differ substantially (large importance weights).

Outcome Regression Policy Evaluation

Directly model the outcome:

$$\hat{V}^{\text{OR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\pi(X_i)\, \hat{\mu}_1(X_i) + (1 - \pi(X_i))\, \hat{\mu}_0(X_i)\right].$$

Consistent under correct specification of $\mu_t(x)$. No importance weights, so lower variance, but sensitive to model misspecification.

Doubly Robust Policy Evaluation (AIPW)

The **augmented IPW (AIPW)** estimator for policy value is

$$\hat{V}^{\text{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left[\pi(X_i) \left(\hat{\mu}_1(X_i) + \frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{b}(X_i)}\right) + (1 - \pi(X_i)) \left(\hat{\mu}_0(X_i) + \frac{(1 - T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{b}(X_i)}\right)\right].$$

Properties: 1. Doubly robust: consistent if either $\hat{b}$ or $\hat{\mu}_t$ is correctly specified. 2. Achieves the semiparametric efficiency bound when both are correctly specified. 3. Allows the use of flexible/nonparametric estimators for nuisance functions while maintaining $\sqrt{n}$-convergence for the policy value (via cross-fitting).

Semiparametric Efficiency Bound for the ATE

The ATE $\tau = V(\pi \equiv 1) - V(\pi \equiv 0) = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ is a special case of policy evaluation. The **efficient influence function** for τ under the nonparametric model is

$$\phi(Y, T, X; \tau) = \frac{T(Y - \mu_1(X))}{e(X)} - \frac{(1 - T)(Y - \mu_0(X))}{1 - e(X)} + \mu_1(X) - \mu_0(X) - \tau.$$

The **semiparametric efficiency bound** is

$$\text{Var}(\phi) = \mathbb{E}\left[\frac{\text{Var}(Y \mid T = 1, X)}{e(X)} + \frac{\text{Var}(Y \mid T = 0, X)}{1 - e(X)} + (\tau(X) - \tau)^2\right].$$

Derivation. The influence function is derived from the tangent space of the nonparametric model. The nuisance tangent space is spanned by functions of (T, X); the efficient influence function is the projection of the pathwise derivative of τ onto the orthocomplement of this space in $L^2(P)$.

The three terms in the efficiency bound have clear interpretations: - $\text{Var}(Y \mid T = 1, X)/e(X)$: noise in treated outcomes, amplified when few units are treated. - $\text{Var}(Y \mid T = 0, X)/(1 - e(X))$: noise in control outcomes. - $(\tau(X) - \tau)^2$: treatment effect heterogeneity.

Cross-Fitting and Machine Learning

When $\hat{\mu}_t$ and $\hat{b}$ are estimated by flexible methods (random forests, neural networks), the AIPW estimator requires **cross-fitting** (sample splitting) to avoid Donsker conditions:

1. Split data into K folds.
2. For each fold k, estimate nuisance functions $\hat{\mu}_t^{(-k)}, \hat{b}^{(-k)}$ on the complement.
3. Evaluate the AIPW estimand on fold k using the cross-fitted nuisance estimates.
4. Average across folds.

This **DML (Double/Debiased Machine Learning)** procedure (Chernozhukov et al., 2018) achieves $\sqrt{n}$-consistent, asymptotically normal estimation of τ even when nuisance functions converge at slower rates (e.g., $n^{-1/4}$).

Optimal Policy Learning

The optimal deterministic policy is

$$\pi^*(x) = \mathbf{1}(\tau(x) > 0).$$

Estimation approaches:

1. **Indirect method:** Estimate $\hat{\tau}(x)$ (e.g., via CATE estimation using causal forests or metalearners), then set $\hat{\pi}(x) = \mathbf{1}(\hat{\tau}(x) > 0)$.
2. **Direct method (policy search):** Maximize the estimated policy value over a class of policies Π:

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{V}^{\text{DR}}(\pi).$$

The policy regret $V(\pi^*) - V(\hat{\pi})$ converges at rate $O(n^{-1/2})$ for finite-dimensional policy classes and $O(n^{-1/(2+d)})$ for d-dimensional policy classes under smoothness.

Structural Equations and Policy Counterfactuals

In the **structural equation model (SEM)**:

$$Y = f(T, X, U), \quad T = g(Z, X, V),$$

where U, V are unobserved. A policy π replaces the treatment equation with $T = \pi(X)$. The policy counterfactual is

$$Y^\pi = f(\pi(X), X, U),$$

and the policy value is $V(\pi) = \mathbb{E}[f(\pi(X), X, U)]$.

When the SEM is identified (via instruments, panel structure, or functional form), policy evaluation reduces to computing expectations under the structural functions. This connects to mechanism design: the structural equations describe how the system responds to interventions, and the policy is the designed intervention.

Connection to Reinforcement Learning

In sequential settings, the policy maps state histories to actions: $\pi : \mathcal{H}_t \to \mathcal{A}$. The value function is

$$V(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t R_t \right],$$

where R_t is the reward at time t and γ is a discount factor. **Off-policy evaluation** estimates $V(\pi)$ from data collected under a behavior policy b, using:

- **Importance sampling:** $\hat{V}^{\text{IS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \prod_{t=0}^{T} \frac{\pi(A_{it}|H_{it})}{b(A_{it}|H_{it})} \sum_t \gamma^t R_{it}$. Variance grows exponentially with horizon T.
- **Doubly robust (sequential):** Analogous to the AIPW estimator, using fitted Q-functions as the outcome model and importance ratios as the propensity weights.

This is the sequential generalization of the single-stage causal inference of Nodes 11.1–11.5.

Worked Examples

Example 1: Policy Evaluation via AIPW

Data: $n = 500$, binary treatment. Behavior policy: $b(x) = \text{logit}^{-1}(0.5 + x_1)$. Estimated outcome models: $\hat{\mu}_1(x) = 3 + 2x_1$, $\hat{\mu}_0(x) = 3 + 0.5x_1$.

Target policy: treat if $x_1 > 0$, i.e., $\pi(x) = \mathbf{1}(x_1 > 0)$.

For a specific observation with $X_1 = 0.8$, $T = 1$, $Y = 5.2$:

$\hat{b}(X) = \text{logit}^{-1}(0.5 + 0.8) = \text{logit}^{-1}(1.3) = 0.786$. $\hat{\mu}_1(X) = 3 + 1.6 = 4.6$, $\hat{\mu}_0(X) = 3 + 0.4 = 3.4$. $\pi(X) = 1$ (since $x_1 > 0$).

AIPW contribution: $1 \cdot (4.6 + \frac{1 \cdot (5.2 - 4.6)}{0.786}) + 0 \cdot (\cdots) = 4.6 + 0.763 = 5.363$.

Averaging such terms over all 500 observations gives $\hat{V}^{\text{DR}}(\pi)$.

Example 2: Optimal Policy via CATE Estimation

Estimated CATE: $\hat{\tau}(x) = 1.5 - 3x_1 + 2x_2$. The optimal policy treats when $\hat{\tau}(x) > 0$, i.e., $1.5 - 3x_1 + 2x_2 > 0$, or $x_2 > 1.5x_1 - 0.75$.

For patient profiles: - $(x_1, x_2) = (0.2, 0.5)$: $\hat{\tau} = 1.5 - 0.6 + 1.0 = 1.9 > 0$. Treat. - $(x_1, x_2) = (1.0, 0.0)$: $\hat{\tau} = 1.5 - 3.0 + 0.0 = -1.5 < 0$. Do not treat. - $(x_1, x_2) = (0.5, 0.3)$: $\hat{\tau} = 1.5 - 1.5 + 0.6 = 0.6 > 0$. Treat.

The policy $\hat{\pi}$ personalizes treatment based on predicted benefit, targeting treatment to those who gain most.

Cross-Links

AI. Offline reinforcement learning is the sequential generalization of policy evaluation. Fitted Q-evaluation, importance sampling, and doubly robust methods for sequential decisions all extend the single-stage methods of this node. Contextual bandits are the intermediate case between single-stage and full sequential problems.

Economics. Mechanism design asks: given agents' strategic responses (the structural equations), what policy maximizes social welfare? This is the game-theoretic generalization of optimal policy learning. The welfare analysis of tax policies, subsidies, and regulations uses structural models to evaluate counterfactual policies.

Linear Algebra. The semiparametric efficiency bound involves the inverse of the propensity score as a weighting matrix. In the linear model, the efficient estimator for the ATE reduces to a weighted least squares

problem with weights $w_i = 1/[e(X_i)(1 - e(X_i))]$, and the efficiency bound is the trace of the inverse of the weighted Fisher information matrix.

Quantum Mechanics. Quantum control theory optimizes a policy (a sequence of pulses or measurements) to drive a quantum system to a target state. The value function is the fidelity of the final state, and the "treatment effect" is the change in fidelity from applying vs. not applying a control pulse. Doubly robust estimation has analogues in quantum process tomography with adaptive measurement designs.

Structural Compression

Invariant

Each design (IPW, AIPW, structural models) provides a consistent estimator of the policy value $V(\pi)$ under distinct assumptions; the AIPW estimator achieves the semiparametric efficiency bound and is doubly robust.

Mechanism

IPW reweights observations by the ratio of target-to-behavior policy probabilities; the AIPW estimator adds an outcome-regression correction that reduces variance and provides double robustness. The efficient influence function characterizes the information-theoretic lower bound for estimating the policy value. Cross-fitting enables the use of flexible machine learning estimators for nuisance functions while preserving $\sqrt{n}$-inference for the policy value.

Cross-System Echo

Policy evaluation is the bridge between causal inference and decision theory. In AI, off-policy evaluation in RL is the sequential generalization, with importance sampling replacing IPW and fitted Q-functions replacing outcome regression. In economics, structural estimation of policy counterfactuals is the mechanism-based approach, where the "policy" is a change in the rules of a game and the "value" is social welfare. The doubly robust principle—combine a model of the world with a correction for model error—is universal across these domains.

Exercises

1. Derive the IPW estimator for the policy value $V(\pi)$ starting from $V(\pi) = \mathbb{E}[\pi(X)Y(1)+(1-\pi(X))Y(0)]$ and using ignorability to replace potential outcomes with observables.
2. Show that the AIPW estimator for policy value is doubly robust: consistent if either the propensity score or the outcome model is correctly specified.
3. Derive the efficient influence function for the ATE $\tau = \mathbb{E}[Y(1) - Y(0)]$ by projecting the pathwise derivative onto the orthocomplement of the nuisance tangent space. Verify that the AIPW estimator is the sample analogue.
4. For a binary covariate $X \in \{0, 1\}$ with $\mathbb{P}(X = 1) = 0.6$, $\tau(0) = -1$, $\tau(1) = 3$, compute the value of the optimal policy $\pi^*(x) = \mathbf{1}(\tau(x) > 0)$ and compare it to the value of treating everyone and treating no one.
5. Explain the DML (cross-fitting) procedure for estimating the ATE with machine-learning nuisance estimators. Why is sample splitting necessary? What goes wrong without it?
6. Compare the variance of the IPW, OR, and AIPW estimators for the ATE in a setting where the propensity score is extreme ($e(x)$ near 0 or 1 for some x). Which estimator is most robust to this practical challenge?

7. Argue, structurally, why policy evaluation subsumes treatment effect estimation as a special case, and explain how the connection to reinforcement learning generalizes the single-decision framework of causal inference to sequential decision problems with delayed rewards.