Principles of Machine Learning

Lecture 2-1: Probability and Distributions

Sharif University of Technology
Dept. of Aerospace Engineering

March 2, 2025

Table of Contents

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Chapter Overview: Probability and Statistics

- **Understanding Uncertainty**
  - **Degree of Belief:** How strongly we believe an event will happen.
  - **Relative Frequency:** How often an event occurs out of a number of trials.
- **Quantifying Uncertainty:**
  - **Data:** Collecting and analyzing data.
  - **Model:** Building models to represent data.
  - **Prediction Uncertainty:** Estimating uncertainty in model predictions.

Chapter Overview: Probability and Statistics

- **Understanding Uncertainty**
  - **Degree of Belief:** How strongly we believe an event will happen.
  - **Relative Frequency:** How often an event occurs out of a number of trials.
- **Quantifying Uncertainty:**
  - **Data:** Collecting and analyzing data.
  - **Model:** Building models to represent data.
  - **Prediction Uncertainty:** Estimating uncertainty in model predictions.
- **Core Concepts:**
  - **Random Variables:** Variables whose values are determined by chance.
  - **Probability Distributions:** Mathematical functions that describe the likelihood of different outcomes.
- **Building Knowledge:**
  - **Probabilistic Modeling:** Creating models that incorporate uncertainty.
  - **Graphical Models:** Using graphs to represent relationships between variables.
  - **Model Selection:** Choosing the best model based on criteria like accuracy and simplicity.

Key Definitions

Definition: Random Variable

A **Random Variable** is a mapping that assigns each outcome of a random experiment to a numerical value representing a specific property or characteristic.

Definition: Probability Distribution

A **Probability Distribution** is a function that assigns a probability to each possible outcome of a random variable, indicating the likelihood of each outcome.

Applications in ML

- Data Uncertainty
- Model Uncertainty
- Prediction Uncertainty
- Basis for Advanced Topics

Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Mathematical Structure of Probability

- **Need for formal structure:** Systematic framework to quantify uncertainty
- **Key observations:**
  - Individual outcomes are unpredictable (e.g., single coin toss)
  - Regular patterns emerge in aggregate (e.g., 50% heads in 1000 tosses)
- **Core components:**
  - Sample space (complete outcome catalog)
  - Event space (meaningful outcome combinations)
  - Probability measure (quantified uncertainty)
- **Why it matters:**
  - Extends Boolean logic to uncertain reasoning
  - Provides foundation for statistical inference

From Everyday Reasoning to Probability

Classical Logic Limitations

- Binary truth values (True/False)
- No gradation for uncertainty
- Example: Friend's tardiness
  - Strict logic: Either late or not
  - Reality: Multiple plausible scenarios

Probabilistic Reasoning

- Continuous plausibility scale [0,1]
- Evidence-based belief updates
- Example: Tardiness hypotheses
  - On time: 20%
  - Traffic delay: 75%
  - Alien abduction: 5%

Probability: Mathematics of plausible reasoning under uncertainty

Axiomatic Foundations (Cox-Jaynes)

Theorem (Cox-Jaynes Theorem)

Any system of plausible reasoning satisfying:

1. ***Representation:*** *Plausibilities as real numbers*
2. ***Consistency:***
   - *Non-contradiction*
   - *Honesty*
   - *Reproducibility*
3. ***Continuity:*** *Small evidence changes $\Rightarrow$ small plausibility changes*

must obey probability axioms (up to isomorphism).

Deep Insight

- Probability theory is *unique* extension of Boolean logic
- Subjectivity vs Objectivity: Personal beliefs vs physical frequencies

Probability Space Triad

Sample Space (Ω) : Elementary outcomes
(e.g., $\Omega = \{\text{H}, \text{T}\}$ for coin toss)

Event Space $(\mathcal{F})$: Space of potential results of the experiment (collection of subsets of Ω)
(e.g., $\mathcal{F} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, \Omega\}$)

Probability Measure (P) : $P : \mathcal{F} \to [0, 1]$ with

- $P(\Omega) = 1$ (Certainty)
- $P(\bigcup_i A_i) = \sum_i P(A_i)$ for disjoint A_i

Kolmogorov's Axioms: Operational Perspective

Fundamental Probability Rules

1. **Non-negativity:** $P(A) \geq 0 \ \forall A \in \mathcal{F}$
2. **Normalization:** $P(\Omega) = 1$
3. **Countable Additivity:** For disjoint $\{A_i\}_{i=1}^{\infty}$,

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

Example

Die Roll Example

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- For fair die: $P(\{i\}) = 1/6$
- Event $A = \{2, 4, 6\}$: $P(A) = 3 \times 1/6 = 1/2$

Putting It All Together: Coin Toss Space

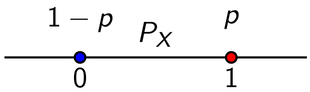
Components

- $\Omega = \{\mathrm{H}, \mathrm{T}\}$
- $\mathcal{F} = 2^{\Omega}$ (power set)
- $P(\{\mathrm{H}\}) = p$
- $P(\{\mathrm{T}\}) = 1 - p$

Random Variable

- $X(\mathrm{H}) = 1$
- $X(\mathrm{T}) = 0$

Event	Description	Probability
$\emptyset$	Impossible	0
$\{\mathrm{H}\}$	Heads	p
$\{\mathrm{T}\}$	Tails	$1 - p$
Ω	Certain	1



Example

Two-Coin Draw (Biased Coins)

- **Sample Space** Ω: $\{(\$, \$), (\$, £), (£, \$), (£, £)\}$
- **Random Variable** X: Counts number of \$ drawn
- **Mapping:**
  - $X((\$, \$)) = 2$
  - $X((\$, £)) = 1$
  - $X((£, \$)) = 1$
  - $X((£, £)) = 0$
- **PMF:**
  - $P(X = 2) = 0.09$
  - $P(X = 1) = 0.42$
  - $P(X = 0) = 0.49$

Statistics vs. Probability in ML

- **Probability:** Models random processes; quantifies uncertainty.
- **Statistics:** Infers underlying processes from observed data.
- **Machine Learning:** Integrates both for model selection and generalization.

Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Discrete vs. Continuous Distributions

- **Discrete:** Target space is countable (finite or countably infinite)
- **Continuous:** Target space is uncountably infinite (e.g., intervals in $\mathbb{R}$)
- **Nomenclature:**
  - Discrete $\Rightarrow$ Probability Mass Function (PMF)
  - Continuous $\Rightarrow$ Probability Density Function (PDF)
- **Cumulative Distribution Function (CDF):** Used for continuous RVs, but also defined for discrete

Discrete Probabilities

- **PMF:** $p(x) = P(X = x)$
- **Joint Probability:** $p(x, y)$
- **Marginal Probability:** $p(x) = \sum_y p(x, y)$
- **Conditional Probability:** $p(y \mid x) = \frac{p(x,y)}{p(x)}$
- **Applications:** Categorical features, labels, finite mixture models

Definition: Probability Mass Function (PMF)

Definition: PMF

- Let X be a discrete random variable with target space T.
- The *probability mass function* $p(x)$ assigns

$$p(x) = P(X = x), \quad x \in T.$$

- Satisfies $\sum_{x \in T} p(x) = 1$ and $p(x) \geq 0$.

Example: Joint Discrete PMF

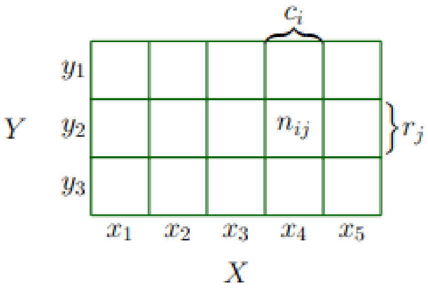
Example (Example: Bivariate Discrete Distribution)

- Two discrete random variables X and Y with states $\{x_1, \ldots, x_5\}$ and $\{y_1, y_2, y_3\}$.
- Joint probability:

$$p(x_i, y_j) = \frac{n_{ij}}{N}, \quad \text{where } n_{ij} \text{ is the count of events with } (x_i, y_j).$$

- **Marginal:** $p(x_i) = \sum_j p(x_i, y_j), \quad p(y_j) = \sum_i p(x_i, y_j)$.
- **Conditional:** $p(y_j \mid x_i) = \frac{p(x_i, y_j)}{p(x_i)}$.

Discrete Bivariate PMF



- Visualization of a discrete bivariate probability mass function.
- Random variables X and Y with a joint PMF $p(x, y)$.

Continuous Probabilities

- **Target space:** Intervals in $\mathbb{R}$ or $\mathbb{R}^D$
- **Probability of exact value:** Zero $(P(X = x) = 0)$
- **Use integrals:** $P(a \leq X \leq b) = \int_a^b f(x)\, dx$
- **Applications:** Real-valued features, Gaussian distributions

Definition: Probability Density Function (PDF)

Definition: PDF

- A function $f : \mathbb{R}^D \to \mathbb{R}$ is a *PDF* if

$$f(x) \geq 0 \quad \text{and} \quad \int_{\mathbb{R}^D} f(x)\, dx = 1.$$

- $\forall a, b \in \mathbb{R}$ (with $a < b$):

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx.$$

Definition: Cumulative Distribution Function (CDF)

Definition: CDF

- For a real-valued random variable $X \in \mathbb{R}^D$, the *CDF* is

$$F_X(x) = P(X_1 \leq x_1, \ldots, X_D \leq x_D).$$

- Can be written as an integral of $f(x)$ when f exists:

$$F_X(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z)\, dz.$$

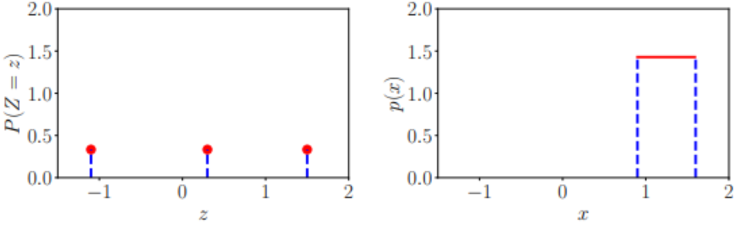
Example: Uniform Distributions

Example (Discrete vs. Continuous Uniform)

- **Discrete Uniform:** Finite states $\{z_1, z_2, z_3\}$, each with $p(z_i) = \frac{1}{3}$.
- **Continuous Uniform:** Interval $[a, b]$ with

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

- Note: PDF can exceed 1 if interval is very small, but integrates to 1.



(a) Discrete distribution (b) Continuous distribution

- Note the difference in how probabilities/densities are visualized.

- **Discrete PMF:** $\sum_{x \in T} p(x) = 1$; each $p(x) \in [0, 1]$.
- **Continuous PDF:** $\int f(x)\, dx = 1$; values of $f(x)$ can exceed 1.
- **Probability of exact point:**
  - Discrete: $P(X = x)$ can be nonzero.
  - Continuous: $P(X = x) = 0$.
- **Common Notational Overlaps:**
  - $p(x)$ used for both PMF and PDF
  - $P(X \leq x)$ also called distribution

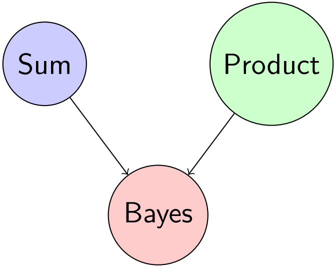
Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

The Three Fundamental Rules

Probability Toolkit for Reasoning

Rule	**What it Does**
Sum Rule	Simplifies complex scenarios
Product Rule	Connects joint & conditional probabilities
Bayes' Theorem	Updates beliefs with new evidence



Sum Rule: Seeing the Big Picture

Weather Example

	Rainy	Sunny
Walk	0.2	0.3
Drive	0.1	0.4

Probability of walking: $0.2 + 0.3 = 0.5$

The Rule in Action

- **Discrete:** $p(\text{walk}) = \sum_{\text{weather}} p(\text{walk}, \text{weather})$
- **Continuous:** $p(\text{height}) = \int p(\text{height}, \text{weight}) d\text{weight}$

Key Idea

"Zoom out" by adding up details you don't need

Product Rule: Connecting Events

Cookie Jar

2 chocolate   3 oatmeal

Probability of first chocolate, then oatmeal:
$\frac{2}{5} \times \frac{3}{4} = \frac{3}{10}$

The Mathematics

$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$

- Joint = Conditional Œ Marginal
- Works for dependent events

Bayes' Theorem: Learning from Data

Medical Testing

- 1% disease prevalence
- 90% test accuracy if sick
- 5% false positive rate
  $\Rightarrow p(\text{Sick}|+) = \frac{0.9\times 0.01}{0.9\times 0.01 + 0.05\times 0.99} \approx 15\%$

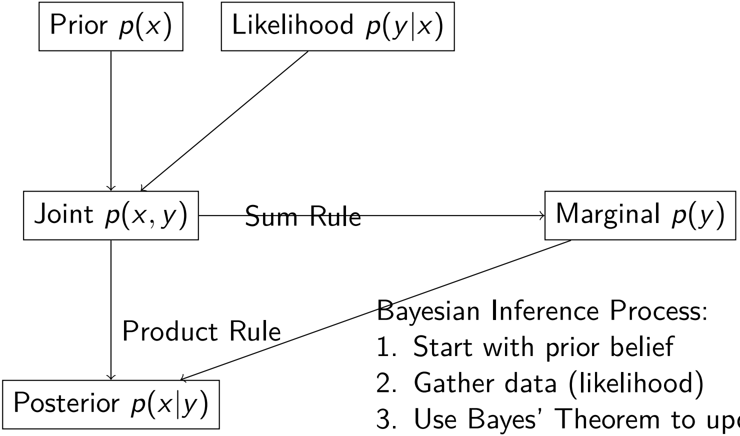
The Formula

$$\underbrace{p(\text{Hypothesis}|\text{Evidence})}_{\text{What we want}} = \frac{\overbrace{p(\text{Evidence}|\text{Hypothesis})}^{\text{Test accuracy}}\ \overbrace{p(\text{Hypothesis})}^{\text{Prior}}}{\underbrace{p(\text{Evidence})}_{\text{Normalizer}}}$$

Why It Matters

Updates beliefs systematically using evidence

Putting It All Together



Bayesian Inference Process:

1. Start with prior belief
2. Gather data (likelihood)
3. Use Bayes' Theorem to update beliefs

Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Expected Value: Definition

Definition (Expected Value)

For a random variable X and a function g,

$$E[g(X)] = \begin{cases} \sum_{x \in \mathcal{X}} g(x)\, p(x) & \text{(discrete)} \\ \int_{\mathcal{X}} g(x)\, p(x)\, dx & \text{(continuous)} \end{cases}$$

Linearity of Expectation

- For functions $g(x)$ and $h(x)$ with scalars a, b:

$$E[a\,g(x) + b\,h(x)] = a\,E[g(x)] + b\,E[h(x)]$$

- Key property used in variance and covariance derivations.

Mean of a Random Variable

Definition (Mean)

The mean of X is given by

$$E[X] = \begin{cases} \sum_{x \in \mathcal{X}} x\, p(x) & \text{(discrete)} \\ \int_{\mathcal{X}} x\, p(x)\, dx & \text{(continuous)} \end{cases}$$

For multivariate $X \in \mathbb{R}^D$, $E[X]$ is computed element-wise.

Other Averages: Median and Mode

- **Median:** The middle value where 50% of the data is below.
- **Mode:** The most frequently occurring value or peak in the density.
- Useful when distributions are skewed or multimodal.

Bimodal Distribution Example (Figure 6.4)

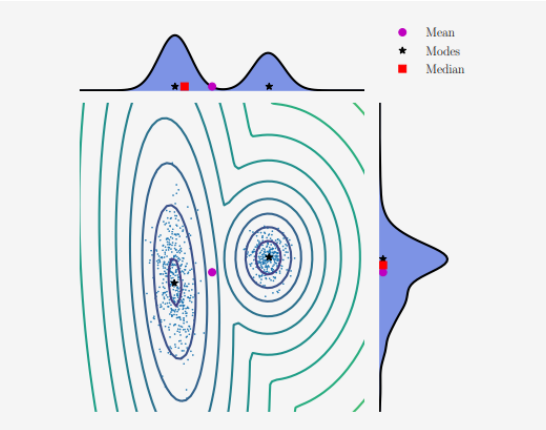
Example (Bimodal Distribution)

Consider a mixture model:

$$p(x) = 0.4\,\mathcal{N}(x \mid \mu_1, \Sigma_1) + 0.6\,\mathcal{N}(x \mid \mu_2, \Sigma_2)$$

Note: The joint distribution is bimodal, though one marginal may be unimodal.

Bimodal Example



- 2D mixture of Gaussians.
- Two distinct modes in the joint distribution.
- Marginal distributions can have different characteristics.

Covariance: Univariate Case

Definition (Covariance (Univariate))

For random variables X and Y,

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

Equivalently,

$$\mathrm{Cov}[X, Y] = E[XY] - E[X]\,E[Y]$$

Variance: A Special Case of Covariance

- Variance is defined as:

$$\mathrm{Var}[X] = \mathrm{Cov}[X, X] = E\left[(X - E[X])^2\right]$$

- Standard deviation:

$$\sigma(X) = \sqrt{\mathrm{Var}[X]}$$

Covariance: Multivariate Case

Definition (Covariance (Multivariate))

For $X \in \mathbb{R}^D$ and $Y \in \mathbb{R}^E$,

$$\mathrm{Cov}[X, Y] = E\left[X\, Y^\top\right] - E[X]\, E[Y]^\top$$

The result is a $D \times E$ matrix.

Variance for Multivariate Variables

Definition (Variance (Covariance Matrix))

For a multivariate random variable $X \in \mathbb{R}^D$ with mean μ,

$$V[X] = \mathrm{Cov}[X, X] = E\left[(X - \mu)(X - \mu)^\top\right]$$

- Diagonals: individual variances.
- Off-diagonals: cross-covariances.

Definition (Empirical Mean & Covariance)

Given data $x_1, \ldots, x_N \in \mathbb{R}^D$:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \Sigma = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^\top.$$

Note: The unbiased covariance uses $1/(N-1)$.

Three Expressions for Variance

- Standard definition:

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

- Raw-score formula:

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

- Pairwise differences:

$$\frac{1}{N^2}\sum_{i,j}(x_i - x_j)^2 = 2\left[\frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2\right]$$

Affine Transformations of Random Variables

- For $y = Ax + b$:

$$E[y] = A\,E[x] + b$$

- Variance under affine transformation:

$$\mathrm{Var}(y) = A\,\mathrm{Var}(x)\,A^\top$$

- Essential in linear models and dimensionality reduction.

Statistical Independence

Definition (Statistical Independence)

Random variables X and Y are independent if

$$p(x, y) = p(x)\, p(y)$$

Equivalently, $p(x \mid y) = p(x)$.

Conditional Independence

Definition (Conditional Independence)

X and Y are conditionally independent given Z if

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z) \quad \forall z.$$

Notation: $X \perp\!\!\!\perp Y \mid Z$.

Inner Products of Random Variables

- Define inner product as:

$$\langle X, Y \rangle := \mathrm{Cov}[X, Y]$$

- Length:

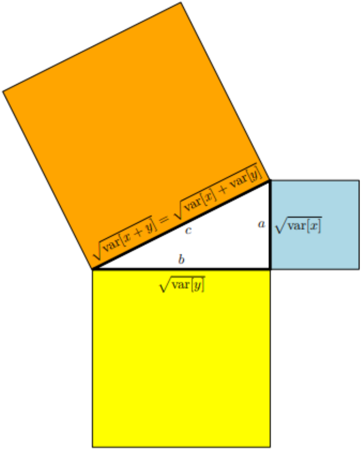
$$\|X\| = \sqrt{\mathrm{Var}[X]} = \sigma(X)$$

- Angle between X and Y:

$$\cos\theta = \frac{\mathrm{Cov}[X, Y]}{\sigma(X)\,\sigma(Y)}$$

- Interpretation: Correlation as the cosine of the angle.

Figure 6.6: Covariance as Inner Product



- Visualizes inner-product structure using covariance.
- Zero covariance corresponds to orthogonality.

Summary of Key Mathematical Relationships

- Sum Rule: $p(x) = \sum_y p(x, y)$ or $\int p(x, y)dy$
- Product Rule: $p(x, y) = p(x \mid y)p(y)$
- Bayes' Theorem: $p(x \mid y) = \frac{p(y|x)p(x)}{p(y)}$
- Linearity: $E[aX + bY] = aE[X] + bE[Y]$
- Variance: $V[X] = E[X^2] - (E[X])^2$
- Affine: $E[Ax + b] = A\,E[X] + b,\ V[Ax + b] = A\,V[X]A^\top$

Implications in Machine Learning

- I.I.D. assumption simplifies model training.
- Covariance and correlation are key in feature analysis.
- Statistical independence assumptions underpin many algorithms.
- Basis for probabilistic modeling and inference.

Closing Remarks on Summary Statistics & Independence

- Summary statistics capture essential aspects of distributions.
- Independence (and conditional independence) simplify joint models.
- Geometric interpretations (inner products, angles) provide intuition.
- These mathematical relationships are foundational in ML.

Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Gaussian Distribution Overview

- Most well-studied continuous distribution
- Also called the Normal distribution
- Arises from the Central Limit Theorem
- Widely used in ML, signal processing, control, statistics

Key Properties

- Fully characterized by mean and covariance
- Closed-form expressions for marginals and conditionals
- Linear transformations preserve Gaussianity
- Computationally convenient in inference tasks

Univariate Gaussian Density

- Density formula:

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Standard case: $\mu = 0$, $\sigma^2 = 1$

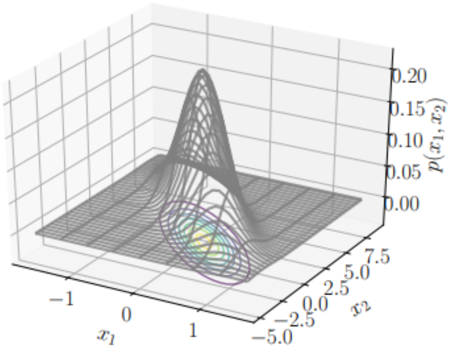
Multivariate Gaussian Density

- For $x \in \mathbb{R}^D$:

$$p(x \mid \mu, \Sigma) = (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

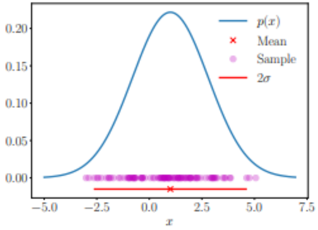
- Denoted as $N(x \mid \mu, \Sigma)$

Bivariate Gaussian Mesh

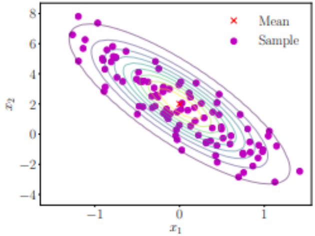


- Mesh plot of a bivariate Gaussian
- Contour lines illustrate elliptical density shapes

Gaussian Samples



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.



(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

- Left: Univariate Gaussian with samples
- Right: Bivariate Gaussian with overlaid samples

Marginals of Joint Gaussian

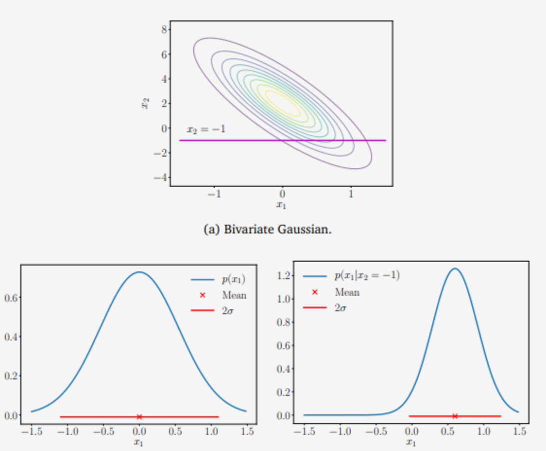
- Joint Gaussian:

$$p(x, y) = N\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\right)$$

- Marginal of x:

$$p(x) = N(x \mid \mu_x, \Sigma_{xx})$$

Joint, Marginal, and Conditional



(a) Bivariate Gaussian.

(b) Marginal distribution.

(c) Conditional distribution.

- (a) Joint bivariate Gaussian
- (b) Marginal of joint Gaussian is Gaussian
- (c) Conditional distribution is also Gaussian

Product of Gaussian Densities

Definition (Product of Gaussians)

The product of two Gaussian densities is proportional to a Gaussian:

$$N(x \mid a, A)\, N(x \mid b, B) = c\, N(x \mid c, C)$$

with

$$C = (A^{-1} + B^{-1})^{-1}, \quad c = C\,(A^{-1}a + B^{-1}b),$$

and scaling constant

$$c = (2\pi)^{-D/2}|A + B|^{-1/2} \exp\left(-\frac{1}{2}(a - b)^\top (A + B)^{-1}(a - b)\right).$$

Scaling Constant for Product

- Expressible as a Gaussian:

$$c = N(a \mid b, A + B) = N(b \mid a, A + B)$$

- Compact notation for Gaussian products

Sums of Independent Gaussians

- If $X \sim N(\mu_x, \Sigma_x)$ and $Y \sim N(\mu_y, \Sigma_y)$ are independent,

$$X + Y \sim N(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$$

- Follows from linearity of expectation and variance additivity

Linear Transformations

- For $Y = AX + b$ and $X \sim N(\mu, \Sigma)$,

$$Y \sim N(A\mu + b,\ A\Sigma A^\top)$$

- Affine transformations preserve Gaussianity

Example: Weighted Sum

Example (Weighted Sum)

For independent Gaussian random variables:

$$p(ax + by) = N(a\mu_x + b\mu_y,\ a^2\Sigma_x + b^2\Sigma_y)$$

Gaussian Applications in ML

- Likelihoods and priors in linear regression
- Mixture models for density estimation
- Gaussian processes for regression and classification
- Kalman filters in signal processing and control

Outline

1. Introduction
2. Construction of Probability Space
3. Discrete and Continuous Probabilities
4. Sum Rule, Product Rule, and Bayes Theorem
5. Summary Statistics and Independence
6. Gaussian Distribution
7. Conjugacy & Exponential Family

Why Special Distributions Matter

The Challenge:

- Bayesian updating often changes distribution forms
- Want tractable math for:
  - Posterior calculations
  - Predictive distributions

The Solution:

- Conjugate families that:
  - Keep same distribution type after updating
  - Maintain fixed number of parameters
- Exponential families provide foundation

Key Distributions for Binary Outcomes

The Bernoulli-Beta Family

	Likelihood	**Conjugate Prior**
Single trial	Bernoulli	Beta
Multiple trials	Binomial	Beta

Coin Flip Analogy

- Bernoulli: Single flip result (H/T)
- Binomial: Count of heads in 10 flips
- Beta: Describes our belief about fairness

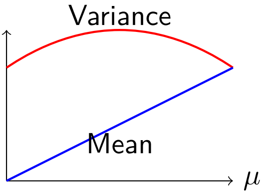
Bernoulli Distribution Demystified

The Coin Flip Distribution

$$p(x \mid \mu) = \underbrace{\mu^x}_{\text{Success if } x=1} \underbrace{(1-\mu)^{1-x}}_{\text{Failure if } x=0}$$

Key Properties:

- $\mathbb{E}[X] = \mu$
- $\text{Var}[X] = \mu(1-\mu)$
- Maximum entropy for binary



From Single Flips to Multiple Trials: Binomial

Counting Successes

$$p(m \mid N, \mu) = \underbrace{\binom{N}{m}}_{\text{Count arrangements}} \underbrace{\mu^m}_{m \text{ successes}} \underbrace{(1-\mu)^{N-m}}_{N-m \text{ failures}}$$

Dice Example

Probability of rolling exactly 3 sixes in 10 rolls:

$$\binom{10}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7 \approx 0.155$$

Exponential Family: The Unified Framework

Canonical Form

$$p(x \mid \eta) = h(x) \exp\left(\eta^\top T(x) - A(\eta)\right)$$

- η: Natural parameters
- $T(x)$: Sufficient statistics
- $A(\eta)$: Log-normalizer

Why It Matters:

- Guarantees conjugacy
- Enables efficient computation
- Unifies discrete/continuous

Examples:

- Bernoulli
- Gaussian
- Poisson
- Beta

Conjugacy in Action: Beta-Bernoulli

Prior: $\mathrm{Beta}(\alpha, \beta)$

$$p(\mu) \propto \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

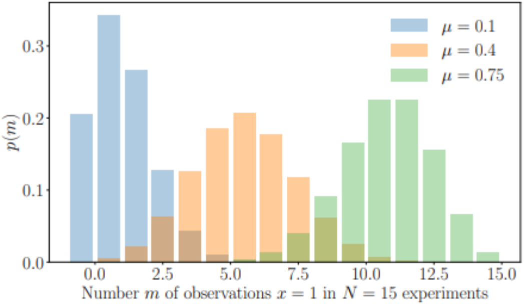
Likelihood: $\mathrm{Bern}(x \mid \mu)$

$$\mu^{x}(1-\mu)^{1-x}$$

Posterior: $\mathrm{Beta}(\alpha + x, \beta + 1 - x)$

$$p(\mu \mid x) \propto \mu^{\alpha+x-1}(1-\mu)^{\beta+(1-x)-1}$$

Binomial Distribution



- Illustrates probability mass vs. number of successes.
- Typical for coin-flip experiments.

Example: Beta Distribution

Example (Beta Distribution)

For $\mu \in [0, 1]$ with parameters $\alpha, \beta > 0$,

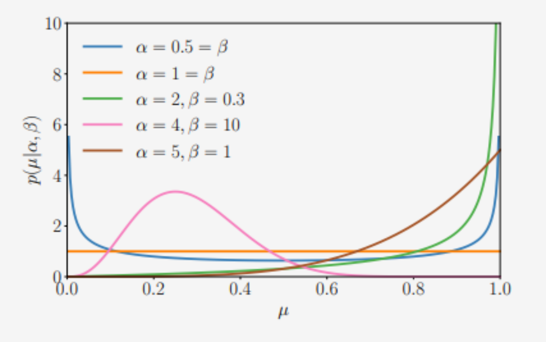
$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1 - \mu)^{\beta-1}.$$

Also,

$$E[\mu] = \frac{\alpha}{\alpha + \beta}, \quad V[\mu] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

- Used to model uncertainty in a probability parameter.

Beta Distribution



- Shows effects of varying α and β.
- Special cases: Uniform ($\alpha = \beta = 1$), bimodal ($\alpha, \beta < 1$), unimodal ($\alpha, \beta > 1$).

Intuition: Beta Parameters

- α: Shifts mass toward 1.
- β: Shifts mass toward 0.
- Special cases yield uniform, bimodal, or symmetric unimodal shapes.

Conjugacy: Motivation

- Prior knowledge should be updated analytically.
- Desire for the posterior to be in the same family as the prior.
- Simplifies computation in Bayesian inference.

Definition: Conjugate Prior

Definition (Conjugate Prior)

A prior $p(\theta)$ is *conjugate* to a likelihood $p(x \mid \theta)$ if the posterior $p(\theta \mid x)$ is in the same family as $p(\theta)$.

Example: Beta-Binomial Conjugacy

Example (Beta-Binomial Conjugacy)

For $x \sim \text{Bin}(N, \mu)$,

$$p(x \mid N, \mu) = \binom{N}{x} \mu^x (1 - \mu)^{N-x}.$$

With prior $\mu \sim \text{Beta}(\alpha, \beta)$,

$$p(\mu \mid \alpha, \beta) \propto \mu^{\alpha-1}(1 - \mu)^{\beta-1}.$$

Then the posterior is

$$p(\mu \mid x, N, \alpha, \beta) \propto \mu^{x+\alpha-1}(1 - \mu)^{N-x+\beta-1},$$

i.e., $\mu \mid x \sim \text{Beta}(x + \alpha, N - x + \beta)$.

Beta-Binomial Posterior Update

- Posterior parameters: $\alpha' = x + \alpha, \quad \beta' = N - x + \beta.$
- Conjugacy simplifies parameter updates.

Example: Beta-Bernoulli Conjugacy

Example (Beta-Bernoulli Conjugacy)

For $x \in \{0, 1\}$ with

$$p(x \mid \theta) = \theta^x (1 - \theta)^{1-x},$$

and prior $\theta \sim \text{Beta}(\alpha, \beta)$,

$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1} (1 - \theta)^{\beta-1}.$$

Then,

$$p(\theta \mid x, \alpha, \beta) \propto \theta^{\alpha+x-1} (1 - \theta)^{\beta+(1-x)-1},$$

i.e., $\theta \mid x \sim \text{Beta}(\alpha + x, \beta + 1 - x)$.

Conjugate Priors in ML

- Conjugacy yields closed-form posteriors.
- Common pairs: Beta-Binomial/Bernoulli, Gaussian-Gaussian, Gamma-Poisson.
- Facilitates iterative updates with new data.

Sufficient Statistics: Motivation

- Statistics that capture all information about parameters.
- Enable data reduction without loss of inferential power.
- Underpin conjugacy and exponential families.

Theorem: Fisher-Neyman Factorization

Theorem (Fisher-Neyman)

Let X have density $p(x \mid \theta)$. Then a statistic $\phi(x)$ is sufficient for θ if and only if

$$p(x \mid \theta) = h(x)\, g_\theta(\phi(x)),$$

where $h(x)$ is independent of θ.

Sufficient Statistics in ML

- Finite-dimensional summaries even for infinite data.
- Key for efficient maximum likelihood estimation.
- Basis for the exponential family formulation.

Definition: Exponential Family

Definition (Exponential Family)

A family of distributions is in the exponential family if it can be written as

$$p(x \mid \theta) = h(x)\, \exp\Big(\langle \theta, \phi(x) \rangle - A(\theta)\Big),$$

where:

- $\phi(x)$: vector of sufficient statistics,
- θ: natural parameters,
- $A(\theta)$: log-partition function.

Exponential Family: Features

- Finite-dimensional sufficient statistics.
- Conjugate priors are easy to derive.
- Log-likelihood is concaveefficient optimization.
- Unifies many common distributions (e.g., Gaussian, Bernoulli, Poisson).

Natural Parameters and Sigmoid

- In the Bernoulli, relate μ and θ via:

$$\mu = \frac{1}{1 + \exp(-\theta)}.$$

- Sigmoid (logistic) function: maps $\theta \in \mathbb{R}$ to $\mu \in (0, 1)$.
- Crucial for logistic regression and neural network activations.

Exponential Families & Conjugacy

- Every exponential family member has a conjugate prior.
- Posterior update involves only sufficient statistics.
- Simplifies Bayesian inference and parameter estimation.

Summary: Conjugacy & Exponential Family

- Conjugate priors yield posteriors of the same form.
- Sufficient statistics capture all necessary data information.
- Exponential families unify many common distributions.
- These properties enable efficient inference in ML.

Closing Remarks

- Understanding these concepts aids in selecting proper models.
- Conjugacy and exponential families simplify Bayesian updates.
- Fundamental for many advanced ML algorithms.