Probabilistic Artificial Intelligence

Lecturer(s): Prof. Andreas Krause

Author: Saurav Banka

Semester: HS 2025

Last edited: September 23, 2025

Contents

1 Introduction

1.1 Motivation

Uncertainty is all around us. This course discusses how to enable data-driven reasoning and decision-making under uncertainty.

1.2 Probability Basics

Probability is formally defined by a probability space (Ω, F, P):

- Ω: Set of atomic events (e.g., throwing a die).
- $F \subseteq 2^{\Omega}$ is a σ-algebra. Intuitively, it formalizes which types of events can occur.
  - $\Omega \in F$
  - $A \in F \implies \Omega \setminus A \in F$ (intuition: the complement of an event is also an event).
  - $A_1, A_2, \dots \in F \implies \bigcup_i A_i \in F$ (intuition: the countable union of events must also be an event).
- Probability measure $P : F \to [0, 1]$ assigns a value to each event in F.

Axioms (Kolmogorov):

- $P(\Omega) = 1$
- $P(A) \geq 0$ for all $A \in F$
- For all disjoint $A_i \in F$,

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Random variables. A random variable (RV) is a mapping $X : \Omega \to D$ where D is some set of interest. Then,

$$P(X = x) = P\big(\{\omega \in \Omega : X(\omega) = x\}\big),$$

i.e., the probability that the variable X takes on a particular value x.

Continuous case. For continuous values, we introduce a probability density:

$$p(t) = \lim_{\delta \to 0} \frac{P(T \in [t, t + \delta])}{\delta}.$$

Definition 1.1. (Expected value) Given a random variable X on domain D and a function $f : D \to \mathbb{R}$, the expectation is

$$\mathbb{E}[f(X)] = \int_{x \in D} f(x)\, p(x)\, dx.$$

Instead of a random variable, we can specify a random vector $\boldsymbol{X} = [X_i(\omega)]_{i=1}^{N}$. The joint distribution describes relations among all variables.

Definition 1.2. Conditional probability: For events A, B with $P(B) > 0$,

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

Definition 1.3. (Marginalization / Sum Rule)

$$P(X_{[1:n]\setminus i}) = \sum_{x_i} P(X_{1:i-1}, x_i, X_{i+1:n}).$$

Definition 1.4. (Product Rule / Chain Rule)

$$P(X_1, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i \mid X_1, \ldots, X_{i-1}).$$

Theorem 1.1. (Bayes' Rule)

$$P(X \mid Y) = \frac{P(X)P(Y \mid X)}{\sum_x P(X = x)\, P(Y \mid X = x)}.$$

Proof: Can be seen by directly applyinng the product and sum rule to the def of conditional independence.

Definition 1.5. (Independence) Random variables $X_1, \ldots, X_n$ are independent if

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i).$$

Equivalently, $P(X_1 \mid X_2) = P(X_1)$.

Definition 1.6. (Conditional independence) We write $X \perp\!\!\!\perp Y \mid Z$ iff for all x, y, z,

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)\, P(Y = y \mid Z = z),$$

or equivalently,

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z).$$

High-dimensional challenges. Without dependence structure, we face:

- **Representation:** requires $2^N - 1$ parameters for N binary RVs.
- **Marginalization:** computing marginals requires summing over 2^{N-1} terms.
- **Inference:** conditional queries, learning, and prediction are all expensive.

1.3 Multivariate Gaussians

Multivariate Gaussians (MVGs) are tractable for many of these problems.

Definition 1.7. (Covariance Matrix)

For a random vector $\boldsymbol{x} \in \mathbb{R}^d$ with mean $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}]$, the covariance matrix is

$$\boldsymbol{\Sigma} = \mathbb{E}\Big[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\Big].$$

Equivalently,

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)].$$

Definition 1.8. (Multivariate Gaussian) The density of $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\Bigg(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Bigg).$$

Theorem 1.2. In the Gaussian case, uncorrelated variables are independent.

1.3.1 Computing marginals

Marginals of an index set A are obtained by simply selecting the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Formally, $X_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$.

1.3.2 Conditional distributions

Partition $\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_A \\ \boldsymbol{x}_B\end{bmatrix}$ with corresponding mean and covariance

$$\boldsymbol{\mu} = \begin{bmatrix}\mu_A \\ \mu_B\end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}\Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB}\end{bmatrix}.$$

Then,

$$\boldsymbol{x}_A \mid \boldsymbol{x}_B = b \ \sim\ \mathcal{N}\Big(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b-\mu_B),\ \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\Big).$$

Note 1.1. Interpretation.

- The updated mean

$$\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(b-\mu_B)$$

  is the prior mean μ_A shifted by an adjustment term.
  - Σ_{AB} is the *cross-covariance*, describing how uncertainty in A co-varies with B.
  - Σ_{BB}^{-1} is the *precision matrix* of B, which corrects for redundancy among components of B: highly correlated features of B contribute less uniquely to the update.
  - Together, $\Sigma_{AB}\Sigma_{BB}^{-1}$ acts like a regression coefficient matrix mapping observed deviations

$(b - \mu_B)$ into updates of A's mean.

- The updated covariance

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$

is always smaller (in the PSD sense) than Σ_{AA}. Intuitively: conditioning on B reduces uncertainty about A.

1.3.3 Affine transformations of Gaussians

If $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$ and $\boldsymbol{y} = M\boldsymbol{x} + b$ for matrix M and vector $\boldsymbol{b}$, then

$$\boldsymbol{y} \sim \mathcal{N}(A\mu + b,\ M\Sigma M^\mathsf{T}).$$

A natural extension is to use 1-hot vectors as rows of A to express marginals or sums: this shows directly that linear combinations (and in particular sums) of Gaussians are Gaussian. It also allows us to represent *degenerate Gaussians*, where some components have constant mean and zero variance to allow for the covariance matrix to have certain eigenvalues corresponding to 0.

1.3.4 Conditional Linear Gaussians

If $\boldsymbol{y} \mid \boldsymbol{x} \sim \mathcal{N}(A\boldsymbol{x} + \boldsymbol{b}, \Sigma_y)$ with $\boldsymbol{x} \sim \mathcal{N}(\mu_x, \Sigma_x)$, then the joint distribution is Gaussian:

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ A\mu_x + \boldsymbol{b} \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_x A^\mathsf{T} \\ A\Sigma_x & A\Sigma_x A^\mathsf{T} + \Sigma_y \end{bmatrix} \right).$$

This structure is fundamental in Bayesian networks with Gaussian conditional distributions.

Note 1.2. Interpretation.

The conditional distribution $P(\boldsymbol{x} \mid \boldsymbol{y})$ has the form

$$\boldsymbol{x} \mid \boldsymbol{y} \ \sim\ \mathcal{N}(M\boldsymbol{y} + b,\ \Sigma_{x|y}),$$

for some matrix M, vector b, and covariance $\Sigma_{x|y}$.

- The term $M\boldsymbol{y} + b$ shows that the conditional mean is an *affine function* of the observed variable $\boldsymbol{y}$. This means that, once we observe $\boldsymbol{y}$, our best prediction of $\boldsymbol{x}$ is a linear regression on $\boldsymbol{y}$.
- The covariance $\Sigma_{x|y}$ represents the residual uncertainty that cannot be explained by $\boldsymbol{y}$. It acts like *independent Gaussian noise* added to the linear prediction.
- Equivalently: the conditional Gaussian says

$$\boldsymbol{x} = M\boldsymbol{y} + b + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Sigma_{x|y}),$$

which makes explicit the regression + noise interpretation.

This viewpoint underlies Gaussian graphical models and Bayesian linear regression.

2 Bayesian Linear Regression

3 Gaussian Processes

4 Variational Inference

5 Markov Chain Monte Carlo.

6 Bayesian Deep Learning

7 Active Learning and Bayesian Optimization

8 Markov Decision Processes

9 Reinforcement Learning

A Math background: Fourier Transforms

B Useful Math Identities