# Probabilistic and Statistical Modeling

Vasilis Gkolemis

ATHENA RC — HUA

June 2025

# Contents

# Helping Material

- **Primer on Probabilistic Modeling**
  https://www.inf.ed.ac.uk/teaching/courses/pmr/22-23/assets/notes/probabilistic-modelling-primer.pdf

## Session 1 – Recap

- **What we covered:**
  - **Probabilistic Modeling:** Model the world using probabilities
  - **Probabilistic Reasoning (Inference):** Use knowns to infer unknowns
  - **Bayesian Analysis:** Modeling and Reasoning with Bayes' rule.
  - **Core Rules of Probability:** The sum, product and Bayes' rule.
  - *Example: Alzheimer's diagnostic test.*

- **What's still to explore:**
  - Our example was simple
    - $X$: the test result — a $1D$ random variable in $\{0, 1\}$
    - $Y$: the disease status — a $1D$ random variable in $\{0, 1\}$
  - Real-world problems are more complex
    - Involve high-dimensional random variables
    - Involve complex relationships between variables
  - How can we model these complexities?
    - Session 2 extended our probabilistic toolbox.
    - Session 3 will show how to use it in practice.

# Session 2 – Recap

- **What we covered:**
  - **Multivariate Random Variables and Distributions:**
    - PDFs, PMFs and CDFs
    - Key properties: expectation and variance.
    - How to sample from these distributions.
    - Key-distributions: Bernoulli, Normal, Poisson.
  - We now have powerful tools to model complexity!

- **What's still to explore:**
  - A glue to connect our probabilistic tools for performing analysis.
  - A *principled* and *unified* way to:
    - Model complex relationships between variables
    - Infer unknowns from knowns
    - Make predictions about future observations
  - The **Bayesian framework** is (among others) a powerful glue for this.

# Session 3 – Overview

- **What we'll explore:**
  - The Bayesian Framework with each key components:
    - ⋆ **Prior Distribution:** Our belief before seeing the data.
    - ⋆ **Likelihood:** How compatible is the observed data is with different parameter values.
    - ⋆ **Posterior Distribution:** Our updated beliefs after observing the data.
    - ⋆ **Predictive Distribution:** Make predictions about new, unseen data.
  - How to use the Bayesian framework for predictive tasks.

- **Be confident. You already know important stuff:**
  - Session 1:
    - ⋆ Intuition about Bayesian modeling $\rightarrow$ Alzheimer's test case
    - ⋆ Core probability rules: sum, product, and Bayes' rule
  - Session 2:
    - ⋆ Multivariate random variables and distributions
    - ⋆ Key properties: expectation and variance
    - ⋆ How to sample from these distributions

# Models

- The term "model" has multiple meanings, see e.g. https://en.wikipedia.org/wiki/Model
- Let's distinguish between three types of models:
  - probabilistic model
  - (parametric) statistical model
  - Bayesian model
- Note: the three types are often confounded, and often just called probabilistic or statistical model, or just "model".
- Introduction to Probabilistic Modelling → for further reading.

# Probabilistic model

- From first lecture:

  *A probabilistic model is an abstraction of reality that uses probability theory to quantify the chance of uncertain events.*

- Example from the first lecture: cognitive impairment test
  - Sensitivity of 0.8 and specificity of 0.95 (Scharre, 2010)
  - Probabilistic model for presence of impairment ($x = 1$) and detection by the test ($y = 1$):
    - $P(x = 1) = 0.11$ (prior)
    - $P(y = 1 \mid x = 1) = 0.8$ (sensitivity)
    - $P(y = 0 \mid x = 0) = 0.95$ (specificity)

# Probabilistic model

- More technically:

  *probabilistic model* $\equiv$ *probability distribution (pmf/pdf).*

- Probabilistic model was written in terms of the probability $P$.
- In terms of the pmf it is:
  - $p_x(1) = 0.11$
  - $p_{y|x}(1 \mid 1) = 0.8$
  - $p_{y|x}(0 \mid 0) = 0.95$
- Commonly written as:
  - $p(x = 1) = 0.11$
  - $p(y = 1 \mid x = 1) = 0.8$
  - $p(y = 0 \mid x = 0) = 0.95$
- where the notation for probability measure $P$ and pmf $p$ are confounded.

# Statistical model

- If we substitute the numbers with parameters, we obtain a (parametric) statistical model:
  - $p(x = 1) = \theta_1$
  - $p(y = 1 \mid x = 1) = \theta_2$
  - $p(y = 0 \mid x = 0) = \theta_3$
- For each value of the $\theta_i$, we obtain a different pmf.
- Dependency highlighted by writing:
  - $p(x = 1; \theta_1) = \theta_1$
  - $p(y = 1 \mid x = 1; \theta_2) = \theta_2$
  - $p(y = 0 \mid x = 0; \theta_3) = \theta_3$
- $p(x, y; \theta)$ where $\theta = (\theta_1, \theta_2, \theta_3)$ is a vector of parameters.
- or $p(x, y \mid \theta)$, for highlighting that $\theta$ is considered a random variable.
- A statistical model corresponds to a set of probabilistic models, here indexed by the parameters $\theta$: $\{p(x; \theta)\}_\theta$

# What is Bayesian modeling?

*A Bayesian model turns a statistical model into a probabilistic one by treating parameters $\theta$ as random variables.*

**Goal:** Learn what we believe about $\theta$ after seeing data — and use that to make predictions.

- A Bayesian model is a probabilistic model $p(x, y, \theta)$
- In supervised settings, we consider $x$ as observed, so we care about $p(y, \theta \mid x)$.

# Bayesian Modeling in Steps

**We don't know the full joint distribution $p(x, y, \theta)$.**

- If we did, every analysis would be trivial.

**What we do have:**

- Observed data $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ — i.i.d. samples from $p(x, y)$

**What we assume:**

- $p(y \mid x, \theta)$ — how data is generated given a specific parameter $\theta$
- $p(\theta)$ — our beliefs about the parameters before seeing data

**What we want to learn:**

- $p(\theta \mid \mathcal{D})$ — what we believe about the parameters after seeing data
- The predictive distribution $p(y \mid x, \mathcal{D})$ — predictions that account for parameter uncertainty
- Possibly others: e.g. marginal likelihood $p(\mathcal{D})$

# Bayesian Modeling for Supervised Tasks

- **Supervised learning:** We observe a dataset of input–output pairs $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ drawn from an unknown joint distribution $p(x, y)$.
- **Bayesian perspective:** Use this data to learn the relationship $x \mapsto y$ while capturing uncertainty in the model parameters.
- **Hypothesis:**
  - (1): assume a parametric family for $p(y \mid x, \theta)$, such as linear regression, neural networks, etc. (Parametric modeling assumption)
  - (2): assume a prior belief over the parameters $p(\theta)$ (prior assumption)

# Example: Linear Regression

- **Hypothesis:** The outcome depends linearly on some input plus noise.
- **Example:** Predict house price from size.
  - $Y$ — house price
  - $X$ — house size
  - Model: $Y = wX + b + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $p(Y \mid X; \theta) = \mathcal{N}(Y \mid wX + b, \sigma^2)$
- **Parameters:** $\theta = (w, b)$ – 2 variables.

# Example: Linear Regression

- **Hypothesis:** The outcome depends linearly on some input plus noise.
- Generalizes to any number of input features
- **Example:** Predict house price from size, number of rooms, and
  - $Y$ — house price
  - $X = (X_1, X_2, \ldots, X_d)$ — house features (size, number of rooms, etc.)
  - Model: $Y = wX + b + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $p(Y \mid X; \theta) = \mathcal{N}(Y \mid wX + b, \sigma^2)$
- **Parameters:** $\theta = (w, b)$: $d + 1$ variables.

# Example: Non-linear Regression

- **Hypothesis:** The relationship between input and output is complex and nonlinear plus noise.
- **Example:** Predict bike rentals from weather data.
  - $Y$ — number of bikes rented per hour
  - $X$ — weather features (temperature, humidity, etc.)
  - Model $Y = f_\theta(X) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
  - $f_\theta(X)$ is a non-linear function parameterized by $\theta$.
- $p(Y \mid X; \theta) = \mathcal{N}(Y \mid f_\theta(X), \sigma^2)$
- what is $f_\theta(X)$?
  - A neural network with weights $\theta$, normally of thousands of variables.
  - A random forest with decision trees where the structure and parameters are defined by $\theta$, normally of hundreds of variables.

# The Flexibility—and Cost—of the Bayesian Framework

- **Bayesian framework is flexible:** We can assume any model for $p(y \mid x, \theta)$.
  - Example: $y = f_\theta(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $f_\theta(x)$ can be a neural network, random forest, etc.
  - The output $y$ can follow any distribution — not just Normal: skewed, heavy-tailed, discrete (e.g., Poisson, Bernoulli), etc.
- **But flexibility comes at a cost**
  - the more complex $p(y \mid x, \theta)$:
    - Complex $\rightarrow$ a high-dimensional $\theta$.
    - Complex $\rightarrow$ a complex $f_\theta(x)$
  - the harder it is to perform inference.