Statistical Optimization: Lecture 1

Introduction to Statistics

Zijian Guo

Zhejiang University
Center for data science

March 17, 2026

Outline

AI Era: Statistics and Optimization

Premier of Probability and Statistics

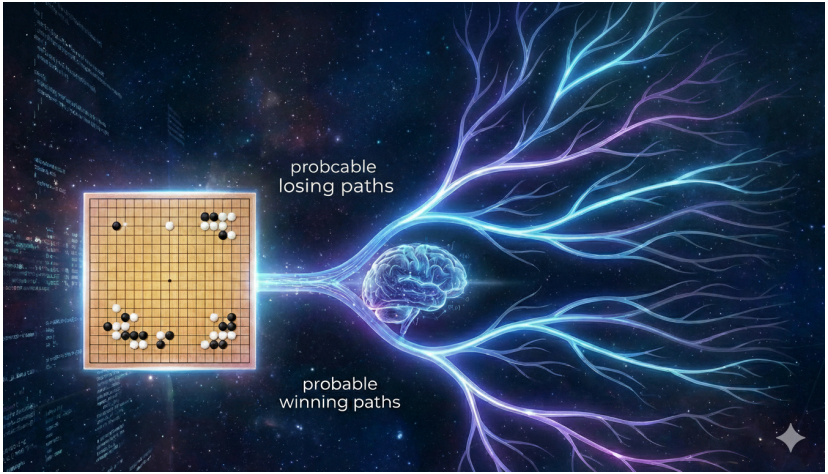
- Probability
- Statistics

Artificial Intelligence

Large Language Models

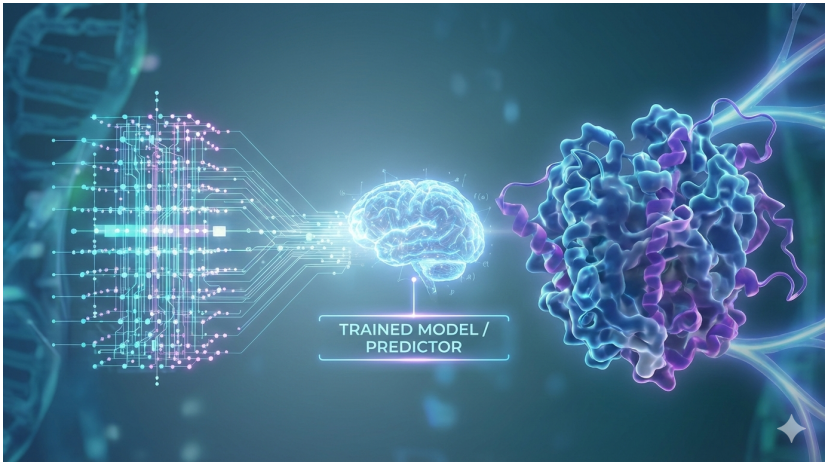
Built on the conditional probability: $P(w_n \mid w_1, \ldots, w_{n-1})$

AlphaGo



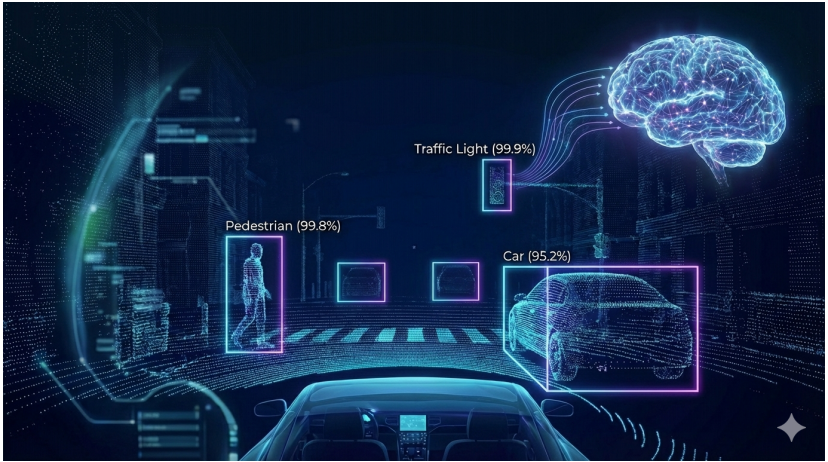
Built on Markov processes

AlphaFold



Built on Deep Neural Networks

Auto Driving



Built on Bayesian statistics $P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$

Data and Model

Data arise in many forms–numbers, text, images, and measurements of variables of interest–across a wide range of scientific studies and experiments.

Statistical model. A statistical model is a mathematical representation of how observed data are generated under certain assumptions, typically described using probability distributions. Formally, we consider a class of distributions

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where distributions are indexed by a parameter θ, and Θ is the *parameter space*.

Data. The data points $X_1, X_2, \cdots, X_n$ are random variables following the distribution P_{θ^*} where $\theta^* \in \Theta$ is unknown.

Statistics and Optimization

Statistics: Learning the information from Data.

- Estimation and Prediction
- Regression and Classification
- Point estimation and Uncertainty Quantification

Optimization: the process of Learning

$$\min_{\theta \in \Theta} f(\theta) \quad \text{and} \quad \widehat{\theta} = \arg\min_{\theta \in \Theta} f(\theta) \tag{1}$$

- Optimality condition: constrained and unconstrained
- Iterative algorithm: first-order and second-order

Outline

AI Era: Statistics and Optimization

Premier of Probability and Statistics

- Probability
- Statistics

Probability and Statistics

Probability: a mathematics language of describing the data.

- Specifying P_{θ^*} with $\theta^* \in \Theta$;
- Specifying the relationship between all data points;
- Characterizing or representing the distribution P_{θ^*}.

Statistics: learning the information from the data, which can be viewed as an inverse process of probability.

- Estimating θ^* or even P_{θ^*};
- Constructing an interval containing θ^*;
- Testing the hypothesis $\theta^* = 0$
- Predicting the value of next observation X_{n+1}

Outline

AI Era: Statistics and Optimization

Premier of Probability and Statistics

- Probability
- Statistics

Random Variable

A *random variable* is a numerical outcome of a random experiment. Equivalently, it is a rule that assigns a real number to each possible outcome of the experiment.

- **Discrete random variable:** X takes values in a finite or countable set (e.g., $\{0, 1\}$ or $\{1, 2, \ldots, 6\}$).
- **Continuous random variable:** X takes values on an interval of $\mathbb{R}$ and is described by a density f_X.

Examples.

- **Coin toss (discrete):** $X = 1$ if Head, $X = 0$ if Tail.
- **Die roll (discrete):** $X \in \{1, 2, \ldots, 6\}$ with $\mathbb{P}(X = i) = 1/6$.
- **Uniform on $[0, 1]$ (continuous):** $X \sim \mathrm{Unif}(0, 1)$ with

$$f_X(x) = \begin{cases} 1, & x \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Probability Mass Function (PMF)

For a discrete random variable taking values in $\mathcal{X}$, we denote its probability mass function as

$$\{P(x)\}_{x\in\mathcal{X}},$$

which is required to satisfy

$$P(x) \geq 0 \qquad \text{for } x \in \mathcal{X},$$

and

$$\sum_{x\in\mathcal{X}} P(x) = 1.$$

Example: Bernoulli

Bernoulli Random Variable. A Bernoulli distribution is a **discrete** probability distribution for a Bernoulli trial, a random experiment with only two outcomes, usually called a “success” and a “failure.” It describes a binary random variable X that can take values 1 (success) and 0 (failure). We denote the probability of success by p with $0 \leq p \leq 1$ and write

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

Example: Binomial

Binomial Random Variable. Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Bernoulli}(p)$, and define

$$S_n = \sum_{i=1}^{n} X_i,$$

the number of successes in n independent trials. Then S_n is **binomial**:

$$S_n \sim \mathrm{Binomial}(n, p),$$

with probability mass function

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Probability Density Function (PDF)

For a continuous random variable, we denote its Probability Density Function $f(x)$ as

$$f(x) \geq 0 \qquad \text{for all } x, \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Gaussian Random Variable. We write the PDF of $X \sim N(\mu, \sigma^2)$ as

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

The normal distribution is a **continuous** probability distribution determined by two parameters: the mean and the variance (or standard deviation). It is symmetric about the mean, which is also the median and the mode.

Cumulative Distribution Function (CDF)

The cumulative distribution function (CDF) $F(x)$ of a random variable X is defined as:

$$F(x) = P(X \leq x).$$

For discrete random variables:

$$F(x) = \sum_{t \leq x} P(t).$$

For continuous random variables:

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

The probability that X lies in the interval $[a, b]$ is:

$$P(a \leq X \leq b) = \int_{a}^{b} f(t)\, dt.$$

Expectation

The expectation of a random variable X, denoted as $E[X]$, is:

$$E[X] = \sum_{x \in \mathcal{X}} xP(x) \quad \text{for discrete variables,}$$

and

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx \quad \text{for continuous variables.}$$

Variance

The variance of a random variable X, denoted as $\text{Var}(X)$, can also be expressed as:

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

For discrete variables:

$$E[(X - E[X])^2] = \sum_{x \in \mathcal{X}} (x - E[X])^2 P(x),$$

and for continuous variables:

$$E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\, dx.$$

Covariance

The covariance between two random variables X and Y, denoted as $\mathrm{Cov}(X, Y)$, is defined as:

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

For discrete variables:

$$E\big[(X - E[X])(Y - E[Y])\big] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - E[X])(y - E[Y])\, P_{X,Y}(x, y),$$

and for continuous variables:

$$E\big[(X - E[X])(Y - E[Y])\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[X])(y - E[Y])\, f_{X,Y}(x, y)\, dx\, dy.$$

Independence

Two random variables X and Y are **independent** if for all Borel sets A and B,

$$\mathbb{P}(X \in A,\ Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B).$$

Joint distribution of independent random variables. If X and Y are independent, then their CDFs satisfy

$$F_{X,Y}(x, y) = F_X(x)\,F_Y(y).$$

Moreover, if X, Y are discrete with PMFs p_X, p_Y, then

$$p_{X,Y}(x, y) = p_X(x)\,p_Y(y).$$

If X, Y are continuous with PDFs f_X, f_Y, then

$$f_{X,Y}(x, y) = f_X(x)\,f_Y(y).$$

Independently and Identically Distributed (i.i.d.)

Random variables $X_1, \ldots, X_n$ are **independently and identically distributed (i.i.d.)** if

$$X_1, \ldots, X_n \text{ are independent} \quad \text{and} \quad F_{X_1} = \cdots = F_{X_n}.$$

Joint distribution of n i.i.d. random variables. If $X_1, \ldots, X_n$ are i.i.d. with common CDF F, then for all $x_1, \ldots, x_n \in \mathbb{R}$,

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

If they are discrete with common PMF p, then

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

If they are continuous with common PDF f, then

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

Outline

AI Era: Statistics and Optimization

Premier of Probability and Statistics

- Probability
- Statistics

Probability and Statistics

Probability: a mathematics language of describing the data.

- Specifying P_{θ^*} with $\theta^* \in \Theta$;
- Specifying the relationship between all data points;
- Characterizing or representing the distribution P_{θ^*}.

Statistics: learning the information from the data, which can be viewed as an inverse process of probability.

- Estimating θ^* or even P_{θ^*};
- Constructing an interval containing θ^*;
- Testing the hypothesis $\theta^* = 0$
- Predicting the value of next observation X_{n+1}

Example: Binomial Model

Probability Model. Let $X_1, X_2, \cdots, X_n$ be i.i.d. Bernoulli Random Variables with the success probability p^*

$$X_i = \begin{cases} 1 & \text{with probability } p^*, \\ 0 & \text{with probability } 1 - p^*. \end{cases}$$

Statistical Inference. The parameter space is therefore $\Theta = \{p : 0 \leq p \leq 1\}$, and the underlying true parameter $p^* \in \Theta$. The statistical objective is to learn p^* based on the observed data $X_1, X_2, \cdots, X_n$.

Example: Normal Model

Probability Model. Suppose $X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu^*, \sigma^2)$.
Statistical Inference. The parameter space in this case is

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma^2 > 0\}.$$

The statistical objective is to learn μ^* and σ^2 based on the observed data $X_1, X_2, \cdots, X_n$.

Example: Regression Model

A regression model estimates the relationship between a dependent variable (response or outcome) and one or more independent variables (predictors, regressors, or covariates). The objective of regression is to understand how the predictors influence the response.

Probability Model. Consider the simple linear regression model

$$Y_i = X_i\beta + \epsilon_i, \qquad X_i \perp \epsilon_i,$$

where Y_i is the response, $X_i \in \mathbb{R}$ is a predictor, β is the parameter of interest that captures the relationship between X and Y, and ϵ_i is random noise.

Statistical Inference. Learn β based on the paired data $\{X_i, Y_i\}_{1 \le i \le n}$.

Example: Multiple Linear Regression

Probability Model. We consider that we have access to multiple predictors $X_{i,1}, X_{i,2}, \cdots, X_{i,p}$ and consider the regression model

$$Y_i = \sum_{j=1}^{p} X_{i,j}\beta_j + \epsilon_i = X_{i,\cdot}^{\intercal}\beta + \epsilon_i, \qquad \{X_{i,j}\}_{1\leq j\leq p} \perp \epsilon_i,$$

Interpretation of β_j: fixing all other covariates $X_{i,1}, \cdots X_{i,j-1}, X_{i,j+1}, \cdots, X_{i,p}$, if we increase $X_{i,p}$ by a unit, then the conditional outcome increases by β_j.

Statistical Inference. Learn $\beta_1, \cdots, \beta_p$ based on i.i.d. data $\{X_{i,1}, \cdots, X_{i,p}, Y_i\}_{1\leq i\leq n}$.