Optimization Modeling — Part 3/4

Recalls on probability theory

Parin Chaipunya
KMUTT
└Mathematics @ Faculty of Science
└The Joint Graduate School of Energy and Environments

Areas of research:

- Multi-agent optimization: Bilevel programs, Game theory
- Optimization modeling: mainly focused on energy and environmental applications

A coordinated course with Metropolitan Electricity Authority (MEA) on
STOCHASTIC OPTIMIZATION

by Parin CHAIPUNYA (KMUTT, Thailand) and Michel DE LARA (ENPC, France)

17 Nov – 18 Dec 2025

การไฟฟ้านครหลวง
Metropolitan Electricity Authority

KMUTT

CERMICS
École nationale des ponts et chaussées

parin.cha@kmutt.ac.th
parinchaipunya.com
github.com/parinchaipunya

Overview

The main objective of this lecture is to make a quick summary on the portion of probability theory that is needful for stochastic optimization modeling.

Table of contents

Section 1

Basic probability

Basic concepts

Probability theory is a branch of mathematics that is used to handle **uncertainty** and describe **likeliness**.

A probability space

To systematically describe likeliness, we need few ingredients.

Notation	**Terminology**	**Description**
Ω	Sample space	The set of all possible outcomes.
$\mathcal{F} \subset 2^{\Omega}$	Information set	Elements of $\mathcal{F}$ are subsets of Ω (called **events**) that one could actually observe and assign a probability.
$\mathbb{P} : \mathcal{F} \to [0, 1]$	Probability measure	A function that assigns a probability to an event.

Baye's approach.
The value $\mathbb{P}(A)$ is interpreted as the **degree of belief** that an occuring outcome will belong to an event A.

Frequentist's approach.
If we repeat a random experiment and record the outcomes, then the **likeliness** that an outcome belongs to an event A converges to $\mathbb{P}(A)$.

A probability space

To rigorously formulate a probability space, we need the following assumptions.

- $\mathcal{F}$ is a σ-field, *i.e.* must satisfy:
  - $\Omega \in \mathcal{F}$.
  - If $\{A_i\}_{i \in I} \subset \mathcal{F}$ is countable, then $\bigcup_{i \in I} A_i \in \mathcal{F}$.
  - If $A \in \mathcal{F}$, then $A^{\complement} \in \mathcal{F}$.
- $\mathbb{P}$ is a measure, *i.e.* must satisfy:
  - $\mathbb{P}(\Omega) = 1$.
  - If $\{A_i\}_{i \in I}$ are disjoint*, then $\mathbb{P}\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} \mathbb{P}(A_i)$.

The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

In many cases, Ω is a finite set and $\mathcal{F} = 2^{\Omega}$, *i.e.* every thing is an event and be assigned with a probability.

In some complex cases, we make a sequence of decisions and at each decision we have a different information set (different σ-field).

*$A_i \cap A_j = \emptyset$ whenever $i \neq j$.

Some intuitions

We start with a very simple situation to familiarize you with the concept.
Consider a case of **rolling a fair die**.
The sample space is obviously $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Question. What is the information set ?
Answer. It depends on the situation.

- If you observe directly the die, then $\mathcal{F} = 2^{\Omega}$ (everything is an event).
- If your friend doesn't let you see the die but tells you whether the die comes out Odd or Even, then $\mathcal{F} = \{\emptyset, \text{Odd}, \text{Even}, \Omega\}$ with $\text{Odd} = \{1, 3, 5\}$ and $\text{Even} = \{2, 4, 6\}$.
- If your friend doesn't let you see the die but tells you whether the die comes out High or Low, then $\mathcal{F} = \{\emptyset, \text{High}, \text{Low}, \Omega\}$ with $\text{High} = \{4, 5, 6\}$ and $\text{Low} = \{1, 2, 3\}$.
- Something weird from your friend's clue
  like $\mathcal{F} = \{\emptyset, \{1, 2\}, \{1, 2, 3\}, \{3\}, \{3, 4, 5, 6\}, \{4, 5, 6\}, \{1, 2, 4, 5, 6\}, \Omega\}$.

Partitions

A practical case of $\mathcal{F}$ is perhaps generated from a partition.

Definition

A **partition** over Ω is a collection $\mathfrak{P} = \{A_i\}_{i \in I}$ of subsets (called **cells**) of Ω such that
- $\bigcup_{i \in I} A_i = \Omega$
- $A_i \cap A_j = \emptyset$ whenever $i \neq j$.

One can then consider the information set $\mathcal{F} = \sigma(\mathfrak{P})$, where $\sigma(\cdot)$ is *implicityly* given from the following result.

Theorem

For any collection $\mathcal{E} \subset 2^{\Omega}$, there exists a unique smallest σ-field, denoted by $\sigma(\mathcal{E})$, over Ω containing $\mathcal{E}$.

Intuitively, the elements of $\mathcal{F} = \sigma(\mathfrak{P})$ are the complements, and countable unions and intersections of cells.

Facts about a probability space

Always take $(\Omega, \mathcal{F}, \mathbb{P})$ as a probability space.

Additional properties

- $\emptyset \in \mathcal{F}$.
- If $\{A_i\} \subset \mathcal{F}$ is countable, then $\bigcap_{i=1}^{\infty} \in \mathcal{F}$.
- If $A, B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$.
- If $A, B \in \mathcal{F}$, then $A \setminus B \in \mathcal{F}$.
- $\mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A)$ for any $A \in \mathcal{F}$.
- $\mathbb{P}(\emptyset) = 0$.
- If $A, B \in \mathcal{F}$, then $A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$.
- If $A, B \in \mathcal{F}$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Practical situations

In practice, the sample space Ω is usually finite (or quite simple if it is not).
It can be as simple as $\Omega = \{\text{Hot}, \text{Rain}, \text{Cool}\}$.

When Ω is infinite, like $\Omega = [0, 50]$, we can also play a trick to limit our observations to a partition

$$\mathfrak{P} = \{ \underbrace{[0, 20)}_{\text{cold}}, \underbrace{[20, 30)}_{\text{cool}}, \underbrace{[30, 38)}_{\text{ok}}, \underbrace{[38, 50]}_{\text{hot}} \}.$$

Section 2

Random variables

Subsection 1

Random variables

Uncertain $\neq$ unknown

We cannot specify the value of a RV, not because it is unknown, but because it is **uncertain**.

Random variables

A **random variable** (or **RV**) is a function $X : \Omega \to \mathbb{R}$ such that

$$\{X \leq x\} = \{w \in \Omega \mid X(w) \leq x\} \in \mathcal{F} \quad \text{for all } x \in \mathbb{R}.$$

Intuitively, $X(w)$ is a **focused quantitative interpretation** of an outcome w.

The condition $\{X \leq x\} \in \mathcal{F}$ ensures that we can evaluate its probability. In fact, this defines the **law of** X, which is

$$\mathbb{P}_X(B) = \mathbb{P} \circ X^{-1}(B) = \mathbb{P}(\{X \in B\}) = \mathbb{P}(\{w \in \Omega \mid X(w) \in B\}), \quad \text{for } B \in \mathcal{B}(\mathbb{R})^{\dagger}.$$

With this law, we have a probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$.

[†] $\mathcal{B}(\mathbb{R})$ is the **Borel field**, which is a natural information set over $\mathbb{R}$.

Focused quantitative interpretation

To make a clear picture, consider rolling to fair dice consecutively. One has

$$\Omega = \left\{ \begin{array}{cc} (1,1), \ldots, (1,6) \\ (2,1), \ldots, (2,6) \\ \vdots \qquad\quad \vdots \\ (6,1), \ldots, (6,6) \end{array} \right\} .$$

Let's take $\mathcal{F} = 2^{\Omega}$ and $\mathbb{P}$ the usual counting probability.

Then, focusing on the outcome of the first roll, we would define a RV $X_1 : \Omega \to \mathbb{R}$ to reflect this by

$$X_1(w_1, w_2) = w_1.$$

We can do the same for the second roll with a RV $X_2 : \Omega \to \mathbb{R}$, defined by $X_2(w_1, w_2) = w_2$.

Discrete RVs

A function $X : \Omega \to \mathbb{R}$ may take only countably many values $x_1, x_2, \ldots$.
In this case, X is a RV $\iff \{X = x_i\} = X^{-1}(x_i) \in \mathcal{F}$ for all x_i's, and we say that it is a **discrete RV**.

Note that a discrete RV creates a partition $\mathfrak{P} = \{X^{-1}(x_i)\}_i$.

When X is a discrete RV, its law is simplified to

$$\mathbb{P}_X(x_i) = \mathbb{P} \circ X^{-1}(x_i) = \mathbb{P}(\{X = x_i\}) = \mathbb{P}(X = x_i).$$

What can it be ?

A RV can represent many things in practice.

Let us take $\Omega = \{\text{Clear}, \text{Light cloud}, \text{Heavy cloud}, \text{Rainy}\}$ and $\mathcal{F} = 2^{\Omega}$.
Then we may consider the following different RVs.

- **PV output (kW).** Define $X : \Omega \to \mathbb{R}$ by

$$X(\text{Clear}) = 50, \quad X(\text{Light cloud}) = 40, \quad X(\text{Heavy cloud}) = 25, \quad X(\text{Rainy}) = 10.$$

- **Cooling demand (kW).** Define $W : \Omega \to \mathbb{R}$ by

$$W(\text{Clear}) = 85, \quad W(\text{Light cloud}) = 80, \quad W(\text{Heavy cloud}) = 70, \quad W(\text{Rainy}) = 55.$$

Subsection 2

Operations and properties

Operations on random variables

We may also do algebra with RVs, preserving their measurability.

Properties

- If X and Y are RVs, then $Z = X + Y$ is a RV.
- If X is a RV and $\lambda \in \mathbb{R}$, is $Z = \lambda X$ a RV.
- If $X_1, \ldots, X_n$ are RVs and $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, then $Z = \lambda_1 X_1 + \cdots + \lambda_n X_n$ is a RV.
- If X and Y are RVs, is $Z = XY$ a RV.
- If $g : \mathbb{R} \to \mathbb{R}$ is a continuous function and X is a RV, then $g \circ X$ is a RV.

Examples

We may think of many reasons why one would like to do operations with RVs.

- In the case of rolling two dice consecutively, $Z = X_1 + X_2$ is the sum of two faces.
- A company invests in N assets with investment λ_i in an asset $i \in \{1, \ldots, N\}$. If X_i is a RV representing the value of asset i, then $Z = \lambda_1 X_1 + \cdots + \lambda_N X_N$ represents the company's worth of the investment.
- A company that owns N charging stations has a random demand represented by RVs W_i at station i. The total demand is represented by $Z = W_1 + \cdots + W_N$.

Subsection 3

Expectation

Expectation

The expectation of a RV is understood as its average.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a RV.
The **expectation** of X is defined by

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

provided that the integral is well-defined.

If X is discrete with values $x_1, x_2, \ldots$, then its **expectation** is simplified to

$$\mathbb{E}[X] = \sum_i \underbrace{x_i \mathbb{P}(X = x_i)}_{\text{outcome weighted by its probability}},$$

provided that the summation is well-defined.

Absolutely continuous case

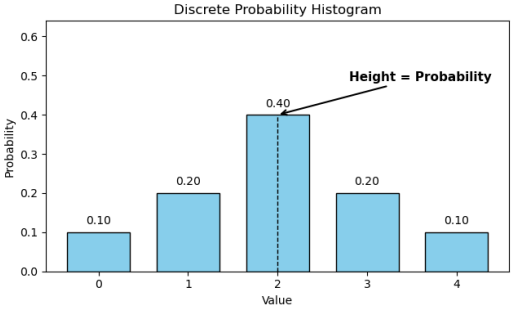
More often than not, a non-discrete RV X is **absolutely continuous**, which means it permets a **density function** $f : \mathbb{R} \to \mathbb{R}$ such that

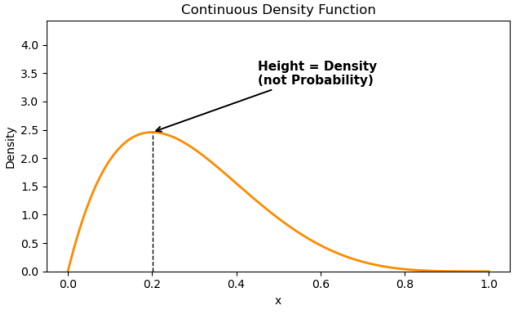
$$\mathbb{P}(X \leq b) = \int_{-\infty}^{b} f(x)\, dx,$$

$$\mathbb{P}(a < X \leq b) = \int_{a}^{b} f(x)\, dx,$$

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

Discrete vs absolutely continuous distributions





Expectation

Take $\Omega = \{\text{Clear}, \text{Light cloud}, \text{Heavy cloud}, \text{Rainy}\}$ and $\mathcal{F} = 2^{\Omega}$. Suppose that we have

$$\mathbb{P}(\text{Clear}) = 0.2, \quad \mathbb{P}(\text{Light cloud}) = 0.4, \quad \mathbb{P}(\text{Heavy cloud}) = 0.3, \quad \mathbb{P}(\text{Rainy}) = 0.1\,.$$

Consider again these RVs and their expectations.

- **PV output (kW).** Define $X : \Omega \to \mathbb{R}$ by

$$X(\text{Clear}) = 50, \quad X(\text{Light cloud}) = 40, \quad X(\text{Heavy cloud}) = 25, \quad X(\text{Rainy}) = 10.$$

  Then we have

$$\mathbb{E}[X] = 50 \times 0.2 \;+\; 40 \times 0.4 \;+\; 25 \times 0.3 \;+\; 10 \times 0.1 = 34.5\,.$$

- **Cooling demand (kW).** Define $W : \Omega \to \mathbb{R}$ by

$$W(\text{Clear}) = 85, \quad W(\text{Light cloud}) = 80, \quad W(\text{Heavy cloud}) = 70, \quad W(\text{Rainy}) = 55.$$

  Then we have

$$\mathbb{E}[W] = 85 \times 0.2 \;+\; 80 \times 0.4 \;+\; 70 \times 0.3 \;+\; 55 \times 0.1 = 75.5\,.$$

Properties of an expection

Properties

The following properties hold.

(a) The expectation is linear, *i.e.*

$$\mathbb{E}[\lambda_1 X_1 + \cdots + \lambda_n X_n] = \lambda_1 \mathbb{E}[X_1] + \ldots \lambda_n \mathbb{E}[X_n]$$

for any $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, where $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are RVs.

(b) The expectation is monotone, *i.e.* for any RVs $X_1, X_2 : \Omega \to \mathbb{R}$ such that $X_1 \leq X_2$, then

$$\mathbb{E}[X_1] \leq \mathbb{E}[X_2].$$

(c) The expectation satisfies the triangle inequality, *i.e.*

$$|\mathbb{E}[X]| \leq \mathbb{E}[\,|X|\,]$$

for a RV $X : \Omega \to \mathbb{R}$.

Properties of an expectation

For a function $g : \mathbb{R} \to \mathbb{R}$ and a RV $X : \Omega \to \mathbb{R}$ such that $g(X)$ is also a RV, we generally have

$$\mathbb{E}[g(X)] \neq g(\mathbb{E}[X]).$$

However, if g is convex, we have the following **Jensen's inequality**

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)].$$

Cost at an averaged demand vs Averaged demand over different scenarios

Consider that g is a convex cost function depedning on a random demand X. Then the Jensen's inequality

$$\underbrace{g(\mathbb{E}[X])}_{\text{cost computed for an averaged demand}} \leq \underbrace{\mathbb{E}[g(X)]}_{\text{averaged cost over different demand scenarios}} .$$

says that the **cost computed for an averaged demand is wrongly optimistic** and not as prepared for uncertainty as the other case.

Planning reserves

If g is the convex reserve needed to cover a random demand X, then the Jensen's inequality

$$\underbrace{g(\mathbb{E}[X])}_{\text{reserve computed for an averaged demand}} \leq \underbrace{\mathbb{E}[g(X)]}_{\text{averaged reserve over different demand scenarios}} .$$

says that using only the averaged demand may cause an underestimated reserve.

The two examples suggest that the **whole distribution** should be taken into account, and not just the expectation.

Subsection 4

Indicator function

Converting an event into a random variable

Take any A. We define the **indicator function** of A as a function $\mathbf{1}_A : \Omega \to \{0, 1\}$ by

$$\mathbf{1}_A(w) = \begin{cases} 1 & w \in A \\ 0 & w \notin A. \end{cases}$$

This effectively converts an event A into a RV $\mathbf{1}_A$.

One may observe that

$$\mathbb{E}[\mathbf{1}_A] = 1 \times \mathbb{P}(\mathbf{1}_A = 1) + 0 \times \mathbb{P}(\mathbf{1}_A = 0) = 1 \times \mathbb{P}(A) = \mathbb{P}(A).$$

Piecewise definition of a random variable from a partition

Suppose that $\mathfrak{P} = \{A_i\}_{i \in I}$ is a countable partition of events on Ω.
We want to define a RV X so that each cell A_i is assigned to a value x_i. This X can be expressed by

$$X = \sum_{i \in I} x_i \mathbf{1}_{A_i}$$

If we take $w \in A_{i_0}$, then

$$X(w) = \sum_{i \in I} x_i \mathbf{1}_{A_i}(w) = \sum_{i \neq i_0} x_i \underbrace{\mathbf{1}_{A_i}(w)}_{=0} + x_{i_0} \underbrace{\mathbf{1}_{A_{i_0}}(w)}_{=1} = x_{i_0}.$$

Section 3

Law of large numbers

Subsection 1

Independence

Independent events

Two events are independent if one doesn't influence the outcome of the other one.

Two RVs X_1 and X_2 are called **independent** if

$$\mathbb{P}(X_1 = x_1,\ X_2 = x_2) = \mathbb{P}(X_1 = x_1) \times \mathbb{P}(X_2 = x_2),$$

for all possible values of x_1 and x_2.

Independent events

Examples

Roll two fair dice consecutively. Take X_1, X_2 the RVs showing respectively the outcome of the first and second roll. Also let $Z = X_1 + X_2$.

- The outcomes of the two dice are independent:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2) = \frac{1}{36}, \quad \mathbb{P}(X_1 = x_1) \times \mathbb{P}(X_2 = x_2) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36},$$

  for any possible x_1 and x_2.

- The outcome of the first die and the sum of the two dice are not independent:

$$\mathbb{P}(X_1 = 4, Z = 3) = 0$$
$$\mathbb{P}(X_1 = 4) \times \mathbb{P}(Z = 3) = \frac{1}{6} \times \frac{2}{36} = \frac{1}{108}$$

Independent and identically distributed RVs

The intuition of having an i.i.d. sequence is that a same situation is repeated in the manner that each repetitions are independent of the others.

For instance, we make an assumption that demand is realized every day and the demand of today is independent of the past days.

A sequence $(X_i)_i$ of RVs is said to be **independent and identically distributed** (briefly **i.i.d.**) if

- all X_i's produce the same possible values $x_1, x_2, \dots$
- any finite selection of X_i's are independent:

$$\mathbb{P}(X_{i_1} = x_{i_1},\, X_{i_2} = x_{i_2},\, \dots,\, X_{i_k} = x_{i_k}) = \mathbb{P}(X_{i_1} = x_{i_1}) \times \mathbb{P}(X_{i_2} = x_{i_2}) \times \cdots \times \mathbb{P}(X_{i_k} = x_{i_k})$$

for all possible values of x_{i_ℓ}'s

- all X_i's have the same law, *i.e.* $\mathbb{P}_{X_i} = \mathbb{P}_{X_j}$ for all i, j.

Subsection 2

Law of large numbers

Sample means

Definition 1
For RVs $X_1, \ldots, X_n$ that are i.i.d., their **sample mean** is defined as

$$M_n := \frac{X_1 + \cdots + X_n}{n}.$$

Since X_i's are r.v.s, the sample mean itself is also a r.v.

A sample mean is understood as an **average of the random experiments till present**.

Since all X_i's are i.i.d., we write their common expectation with μ. Notice that

$$\mathbb{E}[M_n] = \frac{1}{n}\Big(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\Big) = \frac{1}{n}\Big(n\mu\Big) = \mu.$$

Law of large numbers

Consider a sequence $(X_i)_i$ of i.i.d. RVs with common expectation $\mu < +\infty$. The **law of large numbers** (briefly **LLN**) says that the sample mean M_n converges to μ itself.

Law of large numbers

Let us fix a marketing strategy u, which then induces RVs

$$u \mapsto X_i^u,$$

with common expectation $\mathbb{E}[X_i^u] = \mu(u)$.
Let's think of X_i^u as the net profit of sales of day i, which is random due to demand, subject to this marketing choice u.

Each day, the actual sales $X_i^u(w)$ is realized, and sometimes be positive (making money) and other times negative (losing money).

However, the LLN says that, in the long run, these gains and losses average out to $\mu(u)$ (the expected net profit).

Now, the **stochastic optimization** problem comes in and ask the following:

How to choose u so that $\mu(u)$ is **maximized** ?

-» Continue to **Part 4**.