香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Introduction to Data Science

Lecture 14 Statistics & Optimization

Zicheng Wang

Motivation Example

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$ (MLE)

- Drug 1: $\hat{p}_1 = 70\%$.
- Drug 2: $\hat{p}_2 = 68\%$.

Which drug do you think is more effective?

Motivation Example

Whether a drug can cure a disease: $\hat{p} = \frac{\Sigma_i X_i}{n}$ (MLE)

- Drug 1: $\hat{p}_1 = 70\%$. 10 experiments.
- Drug 2: $\hat{p}_2 = 68\%$. 1000000 experiments.

Which drug do you think is more effective?

Confidence Statements

- Fortune Teller

"I believe the cure rate is 70%"

point

- Scientist

With probability 90%, the cure rate is within [40%, 100%]"

interval

- Point estimation involves using a single value to estimate the model parameter.
- Interval estimation provides the probability that the true model parameter lies within an interval of values.
- Why interval? We are not 100% sure that the estimator is exactly the parameter.
- Interval estimation helps quantify the uncertainty of our estimation.

The Value of Interval Estimation

- Assessing the reliability of treatment effects

- Suppose that the Drug Administration will only approve a drug whose cure rate is higher than 55% with probability 90%.
  - Drug 1: Point estimation 70%
  - Drug 2: Point estimation 68%

The Value of Interval Estimation

- Assessing the reliability of treatment effects

- Suppose that the Drug Administration will only approve a drug whose cure rate is higher than 55% with probability 90%.
  - Drug 1: with probability 85%, the cure rate is higher than 55%
  - Drug 2: with probability 99.99999%, the cure rate is higher than 55%

The Value of Interval Estimation

- A blind-box seller tells you: with a probability higher than 50%, you will get a Harry-Potter

- You bought “a few” of them,
  - Point estimation: 49%. Did the seller lie to you?

The Value of Interval Estimation

- A blind-box seller tells you: with a probability higher than 50%, you will get a Harry-Potter

- You bought “a few” of them,
  - Point estimation: 49%. Did the seller lie to you?
  - Interval estimation: with probability 99%, the probability is higher than 50%.

Focus on estimating the mean

A random variable: X with variance $\boldsymbol{\sigma}^2$

Data: $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_n}$

Target: estimate the mean of the random variable.

Focus on estimating the mean

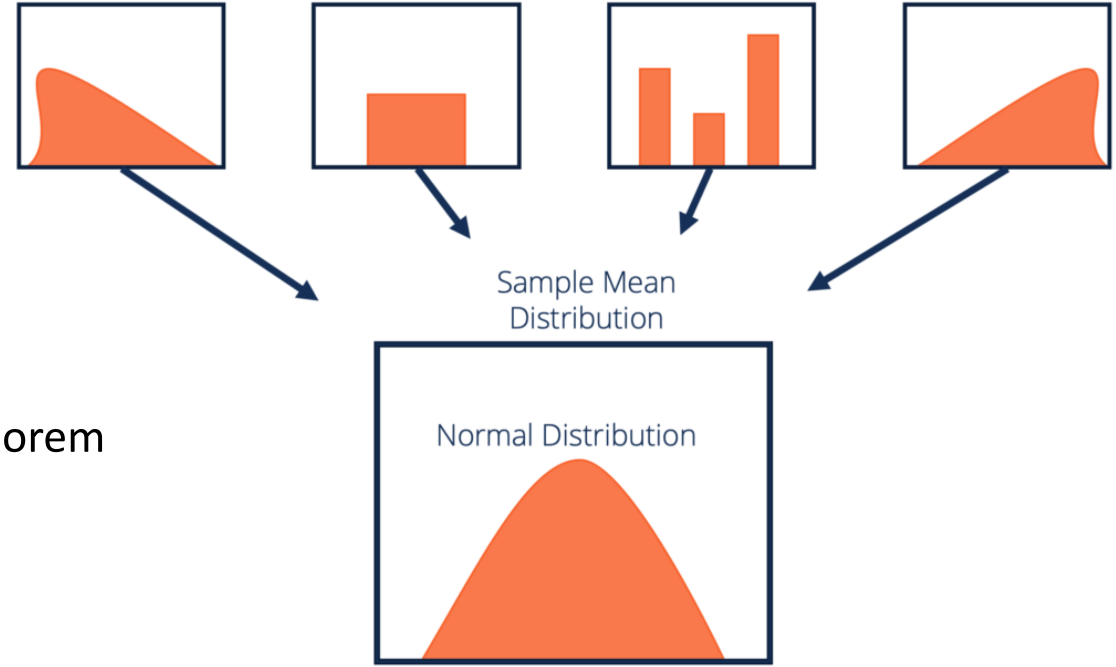
A random variable: X with variance $\boldsymbol{\sigma}^2$

Data: $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_n}$

Target: estimate the mean of the random variable.

Let $\overline{\mathbf{X}} = \frac{\mathbf{X_1}+\mathbf{X_2}+\cdots\mathbf{X_n}}{\mathbf{n}}$ be the sample mean, $\frac{\sqrt{\mathrm{n}}\,(\overline{\mathrm{X}}-\mu)}{\boldsymbol{\sigma}} \sim N(0,1)$

Why proposing this model?



Central Limit Theorem

No matter what the true distribution is, the distribution of the **sample mean** will be very close to the **normal distribution**, as long as the sample size is **large (30-sample size rule of thumb)**.

Interval Estimation

- Consider an interval: T = $[\bar{X} - \frac{b\,\sigma}{\sqrt{n}}, \bar{X} + \frac{a\,\sigma}{\sqrt{n}}]$, where $a, b > 0$ are constants. ($\bar{X}$ is within this interval)

- What's the probability that the true model parameter lies in T?

Interval Estimation

- Consider an interval: T = $[\bar{X} - \frac{b\,\sigma}{\sqrt{n}}, \bar{X} + \frac{a\,\sigma}{\sqrt{n}}]$, where $a, b > 0$ are constants.

- What's the probability that the true model parameter lies in T? We consider the following event.

$$\bar{X} - \frac{b\,\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{a\,\sigma}{\sqrt{n}}$$

Interval Estimation

- Consider an interval: T = $[\bar{X} - \frac{b\,\sigma}{\sqrt{n}}, \bar{X} + \frac{a\,\sigma}{\sqrt{n}}]$, where $a, b > 0$ are constants.

- What's the probability that the true model parameter lies in T? We consider the following event.

$$\bar{X} - \frac{b\,\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{a\,\sigma}{\sqrt{n}}$$

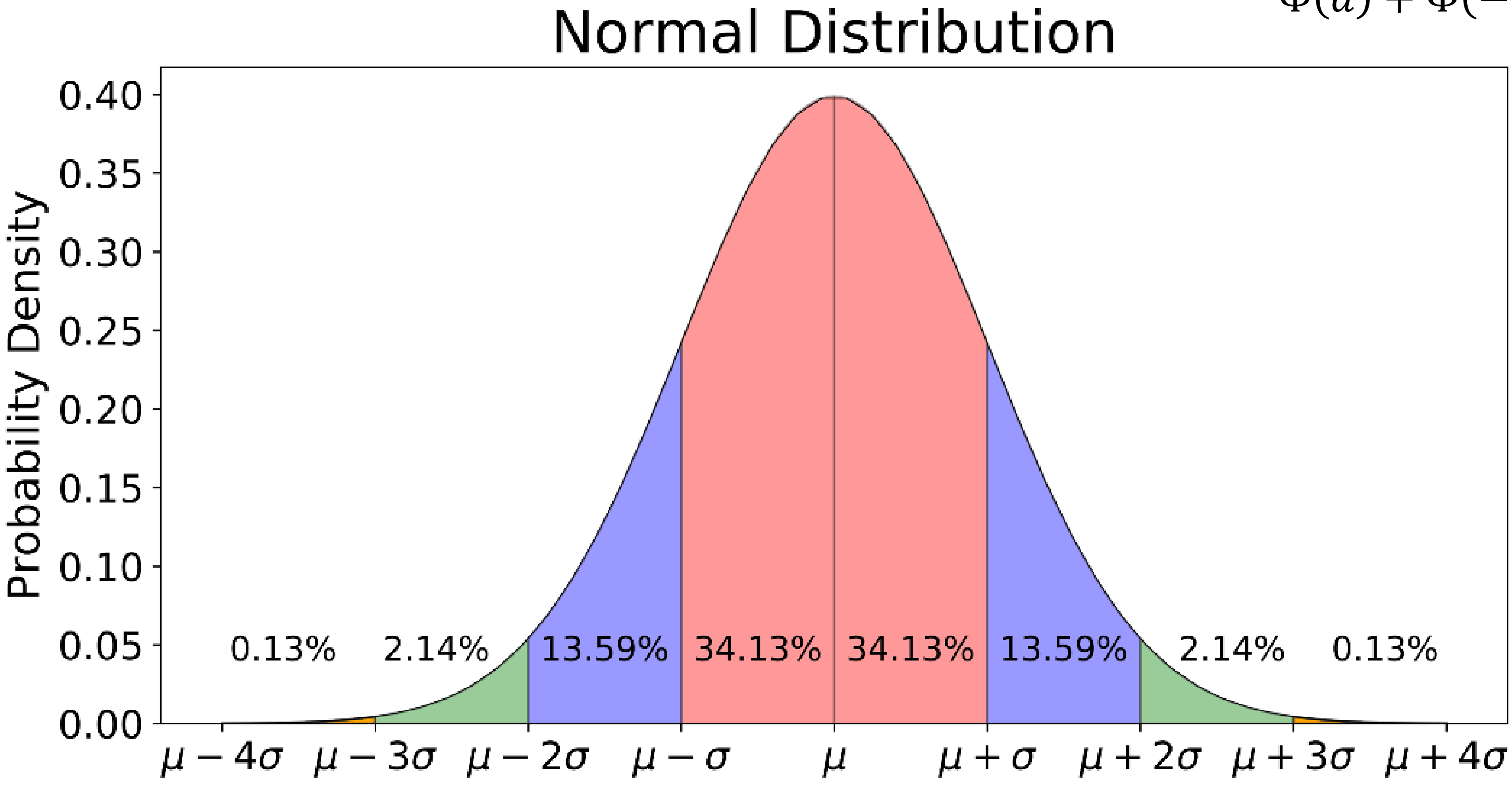
$$-a \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\boldsymbol{\sigma}} \leq b$$

Interval Estimation

- What's the probability that $-a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b$.

- Recall $\frac{\sqrt{\mathrm{n}}\,(\bar{\mathrm{X}}-\mu)}{\boldsymbol{\sigma}} \sim \mathrm{N}(0,1)$
  - P$\left(-a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b\right)$ = $\Phi(b) - \Phi(-a)$
  - $\Phi(x)$: CDF of a standard normal distribution $N\,(0,1)$.

$$\Phi(a) + \Phi(-a) = 1$$


Normal Distribution
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
$\mu - 4\sigma$
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$
$\mu + 4\sigma$

Interval Estimation

- What's the probability that $-a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b$.

- Recall $\frac{\sqrt{\mathrm{n}}\,(\bar{\mathrm{X}}-\mu)}{\boldsymbol{\sigma}} \sim \mathrm{N}(0,1)$
  - P($-a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b$) = $\Phi(b) + \Phi(a) - 1$ (symmetry of standard normal)

  - $\Phi(x)$: CDF of a standard normal distribution $N\,(0,1)$.

Interval Estimation

- What's the probability that $-a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b$.

- Recall $\frac{\sqrt{\mathrm{n}}\,(\bar{\mathrm{X}}-\mu)}{\boldsymbol{\sigma}} \sim \mathrm{N}(0,1)$
  - P$(-a \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\boldsymbol{\sigma}} \leq b)$ = $\Phi(b) + \Phi(a) - 1$
  - $\Phi(x)$: CDF of a standard normal distribution $N\ (0,1)$.

- Formally, we say $[\bar{X} - \frac{b\,\sigma}{\sqrt{n}}, \bar{X} + \frac{a\,\sigma}{\sqrt{n}}]$ is a $\Phi(b) + \Phi(a) - 1$ confidence interval.

Interval Estimation

- When estimating the mean, we usually let $a = b$

- **Define $\boldsymbol{z_{\alpha/2}}$ such that $\boldsymbol{P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha}$, where $\boldsymbol{Z \sim N(0,1)}$**
- Let b = $\boldsymbol{z_{\alpha/2}}$, a = $\boldsymbol{z_{\alpha/2}}$. Then $\Phi(b) - \Phi(-a)$ = 1- α.

- Then, $[\boldsymbol{\bar{X}} - \boldsymbol{z_{\alpha/2}} \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}, \boldsymbol{\bar{X}} + \boldsymbol{z_{\alpha/2}} \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}]$ is a **1- $\boldsymbol{\alpha}$ confidence interval.**

Some observations

1-α Confidence Interval (CI):

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

The length of the confidence interval is affected by several factors

- As the sample size n **increases**, the length of CI **decreases**
- As the variance σ^2 **increases**, the length of CI **increases**
- As the confidence level increases (α **decreases**), the length of CI **increases**.

If σ is unknown?

- We may use MLE to find a model, then use this MLE model's variance to approximate σ^2.

- Drug 1: $\hat{p}_1 = 70\%$.
- $\sigma \approx \sqrt{\hat{p}_1(1-\hat{p}_1)} = 0.458$

- Drug 2: $\hat{p}_2 = 68\%$.
- $\sigma \approx \sqrt{\hat{p}_2(1-\hat{p}_2)} =0.466$

Drug 1

- Drug 1: $\hat{p}_1 = 70\%$. 10 experiments.
- What's the probability that the cure rate is larger than 55%?

$$[\bar{X} - \frac{b\,\sigma}{\sqrt{n}}, \bar{X} + \frac{a\,\sigma}{\sqrt{n}}]$$

  - $\bar{X}$ = 0.7, n = 10, σ = 0.458
  - a = ∞
  - $\bar{X} - \frac{b\,\sigma}{\sqrt{n}}$ = 55%, then b = 1.04
  - Then the probability is $\Phi(b) + \Phi(a) - 1 = \Phi(1.04)$=85%.

Drug 2

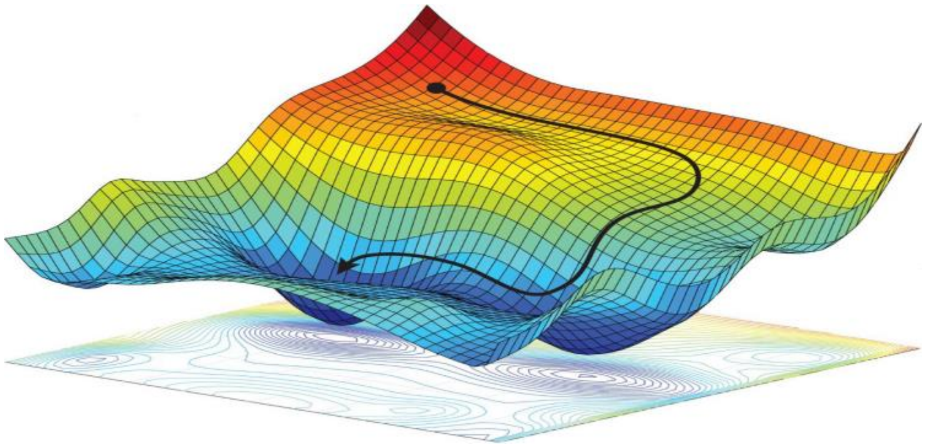
- Drug 2: $\hat{p}_2 = 68\%$. 1000000 experiments.

- What's the probability that the cure rate is larger than 55%?

$$[\bar{X} - \frac{b\,\sigma}{\sqrt{n}}, \bar{X} + \frac{a\,\sigma}{\sqrt{n}}]$$

  - $\bar{X}$ = 0.68, n = 1000000, σ = 0.466
  - a = ∞
  - $\bar{X} - \frac{b\,\sigma}{\sqrt{n}}$ = 55%, then b ≈ 278.97
  - Then the probability is $\Phi(b) + \Phi(a) - 1$ ≈ 1.

Optimization Basics
Introduction



Why study optimization

- Optimization is the foundation to Data Science
  - Modeling (together with statistics and machine learning models)
  - Solution methods
- Optimization can help
  - Deepen your understanding of probability/statistics/ML approaches.
  - Interpret the algorithm – know why you got the result
  - Develop optimal (or sub-optimal) algorithms.

Mathematical optimization (alternatively spelled *optimisation*) or **mathematical programming** is the selection of a best element, with regard to some criterion, from some set of available alternatives.

Given: a function $f : A \to \mathbb{R}$ from some set A to the real numbers
Sought: an element $\mathbf{x}_0 \in A$ such that $f(\mathbf{x}_0) \leq f(\mathbf{x})$ for all $\mathbf{x} \in A$ ("minimization") or such that $f(\mathbf{x}_0) \geq f(\mathbf{x})$ for all $\mathbf{x} \in A$ ("maximization").

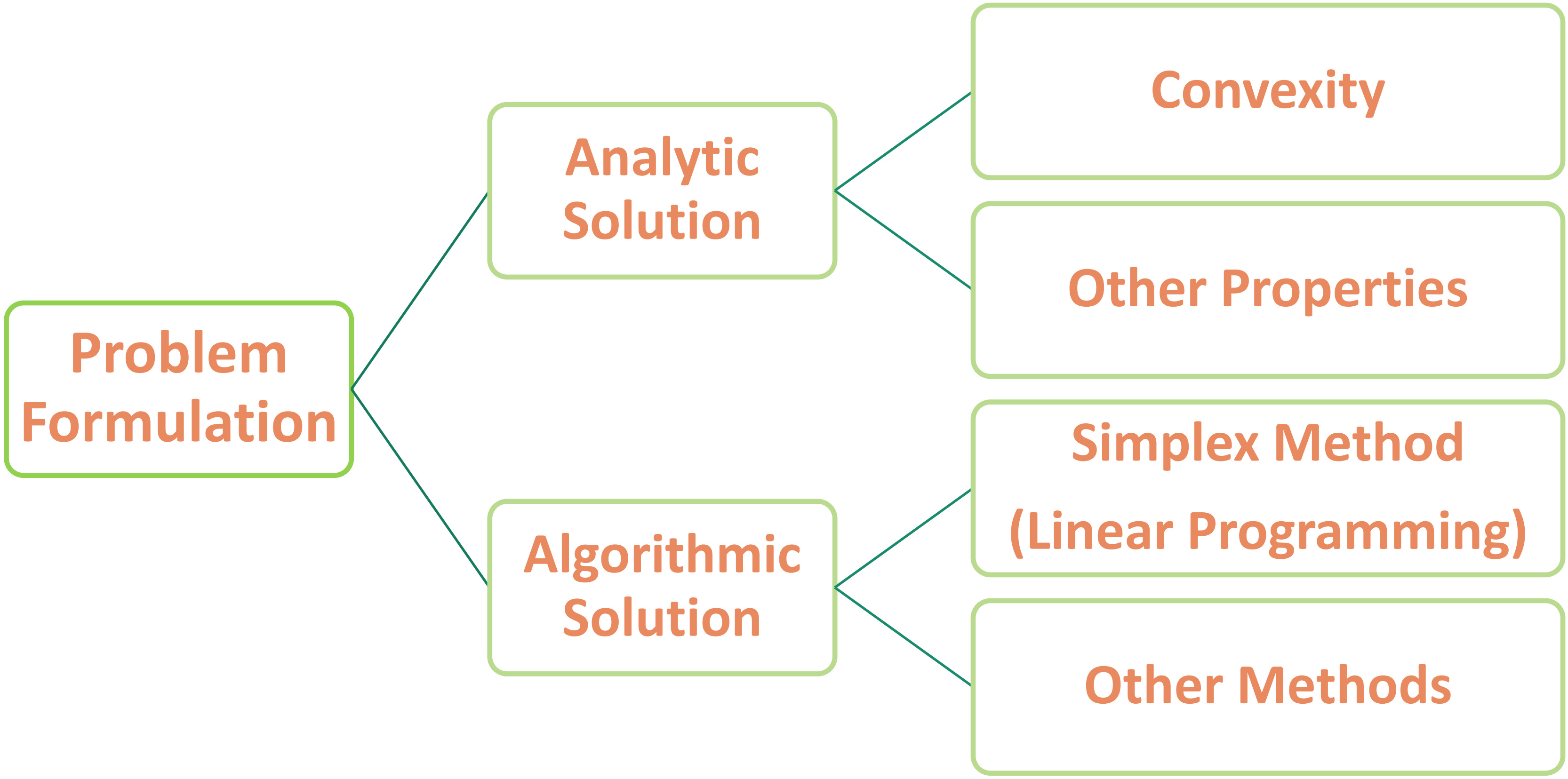
Mathematical optimization (alternatively spelled *optimisation*) or **mathematical programming** is the selection of a best element, with regard to some criterion, from some set of available alternatives.

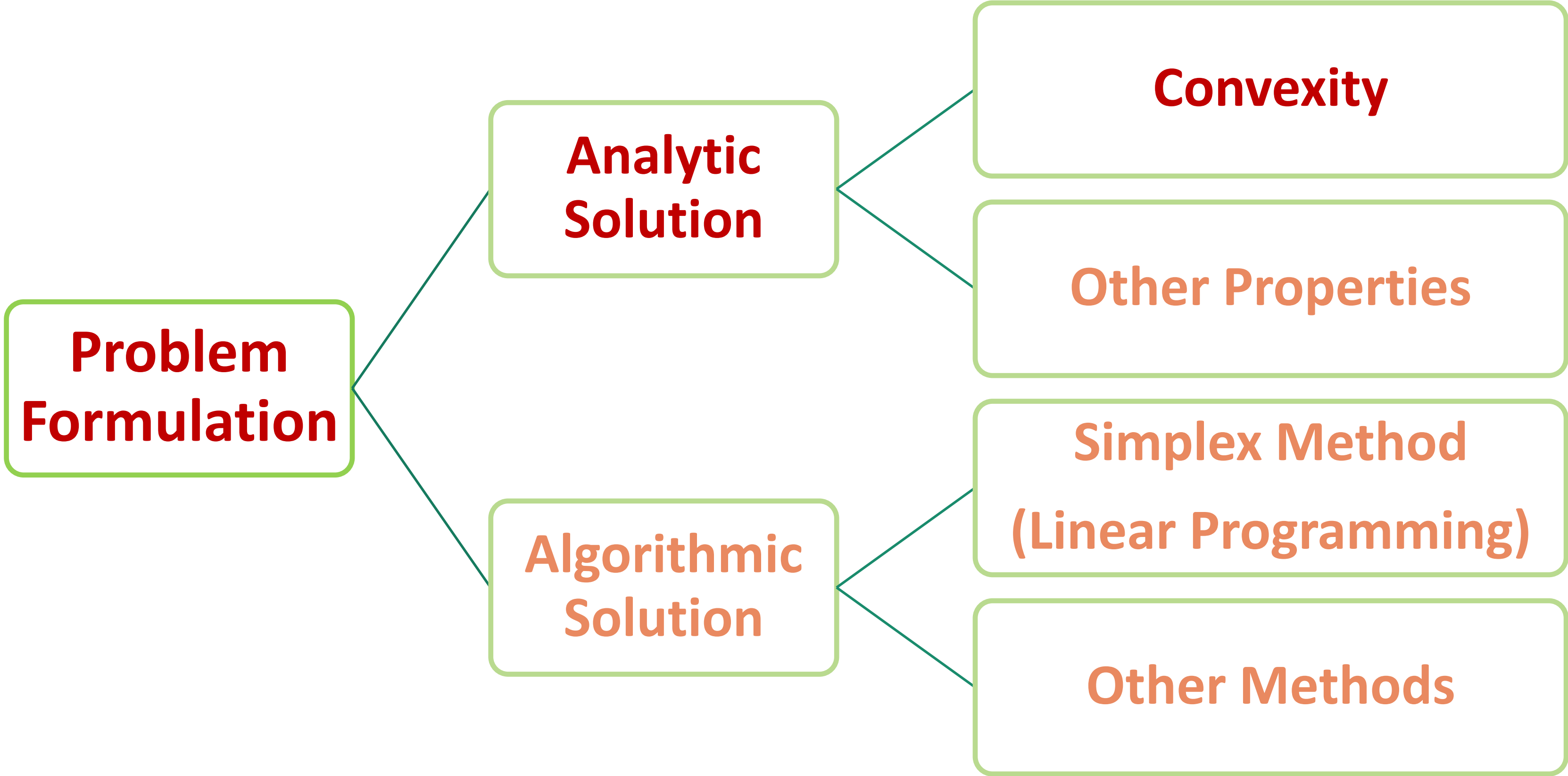
Objective Function

Constraints

Given: a function f: $A \to \mathbb{R}$ from some set A to the real numbers
Sought: an element $\mathbf{x}_0 \in A$ such that $f(\mathbf{x}_0) \leq f(\mathbf{x})$ for all $\boxed{\mathbf{x}} \in A$ ("minimization") or such that $f(\mathbf{x}_0) \geq f(\mathbf{x})$ for all $\mathbf{x} \in A$ ("maximization").

Decision Variables


Problem Formulation
Analytic Solution
Algorithmic Solution
Convexity
Other Properties
Simplex Method (Linear Programming)
Other Methods


Problem Formulation
Analytic Solution
Algorithmic Solution
Convexity
Other Properties
Simplex Method (Linear Programming)
Other Methods

Problem Formulation

The focus of this lecture is how to formulate the problem.

Motivating Example

- Suppose you want to start your own blind box business.

- Let D denote the one season (three months) random demand, with CDF $F(\cdot)$, and mean $\mu = E[D]$.

- At the beginning of each season, you place an order Q to Pop Mart, with a cost c for each blind box.

- Each blind box can be sold at a price of $p > c$.

- At the end of each season, unsold blind boxes are salvaged, and you get $s < c$ for each salvaged box.

Motivating Example

- For simplicity, let's assume that D is a continuous random variable and you can also place a continuous order Q.

- You want to choose the optimal order quantity Q so as to maximize your expected profit.

- How should you formulate the problem?

Optimization problem in standard form

$$
\begin{array}{ll}
\text{minimize} & f_0(x) \\
\text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\
& h_i(x) = 0, \quad i = 1, \ldots, p
\end{array}
$$

- $x \in \mathbf{R}^n$ is the optimization variable
- $f_0 : \mathbf{R}^n \to \mathbf{R}$ is the objective or cost function
- $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1, \ldots, m$, are the inequality constraint functions
- $h_i : \mathbf{R}^n \to \mathbf{R}$ are the equality constraint functions

Problem Formulation

- First we observe that if the realized demand $D > Q$, then your profit is $(p - c)Q$. Otherwise, your profit is $(p - c)D + (s - c)(Q - D)$.

- Let's define $(Q - D)^+ = \max(Q - D, 0)$.

- Given D, your profit is $p \cdot \min(Q, D) + s(Q - D)^+ - cQ$.

- The objective function (i.e., the expected profit) is then given by $f(Q) = pE[\min(Q, D)] + sE[(Q - D)^+] - cQ$

Problem Formulation

- For this problem, we have one inequality constraint: $Q \geq 0$.

- Hence, the optimization problem is as follows

$$\begin{aligned} &\text{maximize} \quad && pE[\min(Q, D)] + sE[(Q-D)^+] - cQ \\ &\text{subject to} \quad && Q \geq 0 \end{aligned}$$

- In standard form, we have

$$\begin{aligned} &\text{minimize} \quad && -(pE[\min(Q, D)] + sE[(Q-D)^+] - cQ) \\ &\text{subject to} \quad && Q \geq 0 \end{aligned}$$

Try Yourself

- Suppose you want to start your own blind box business.

- Let D denote the one season (three months) random demand, which follows a uniform distribution in $[10,100]$.

- At the beginning of each season, you place an order Q to Pop Mart, with a cost 10 Yuan for each blind box.

- Each blind box can be sold at a price of 20 Yuan.

- At the end of each season, unsold blind boxes are salvaged, and you get 3 Yuan for each salvaged box.