Chen Tang's
Knowledge Database

ChenTang@link.cuhk.edu.cn

Last updated on October 13, 2023

Preface

The following is a compendium of my academic notes spanning various domains. I present these notes publicly to share my methodological framework for managing and structuring an individual's knowledge networks.

The inevitability of encountering occasional errors is acknowledged.

This notebook will undergo continuous updates.

Contents

Chapter 1

Mathematics, Statistics & Optimization

1.1 Calculus and Linear Algebra

1.1.1 Keys of Calculus

1.1.2 Keys of Linear Algebra

The properties of Matrix Multiplication:

- Associativity: $(AB)C = A(BC)$;
- Distributivity: $A(C + D) = AC + AD$;
- Identity Multiplication: $I_m A = A, AI_n = A$

Only square matrix has an inverse matrix, and the inverse is unique. If a matrix has an inverse, then it's called **regular/invertible/nonsingular**. The properties of inverses and transposes:

- $(AB)^{-1} = B^{-1}A^{-1}$;
- $(A + B)^{-1} \neq A^{-1} + B^{-1}$;
- $(AB)^\top = B^\top A^\top$;
- $(A + B)^\top = A^\top + B^\top$.

To find the inverse matrix, gaussian elimination can be applied: $[A \mid I] = [I \mid A^{-1}]$.

Solutions of Linear Systems

Reduced Row-Echelon Form Matrix:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 9 \\ 0 & 0 & 0 & 1 & -4 \end{bmatrix} \tag{1.1.2.1}$$

There are *pivot (basic variables)* and *free variable*, and the column of free variable is dependent on pivots. The steps to find solutions of linear systems:

1. Find a particular solution for $Ax = \boldsymbol{b}$ by setting all free variable zero;
2. Find all solutions for $Ax = 0$;
3. Add them up.

There is an iterative method to solve large-scale linear equations: define error $= \|x^{(k+1)} - x_*\|$, then optimize the function $x^{(k+1)} = Cx^{(k)} + d$ and iterate it.

Vector Spaces, Basis and Rank

Definition 1.1.2.1: Vector Space and Subspace

A **Vector Space** $V = (\mathcal{V}, +, \cdot)$ is a set $\mathcal{V}$ with two operations:

1. $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$;
2. $\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$

For $\mathcal{U} \subseteq \mathcal{V}$ and $\mathcal{U} \neq \emptyset$, then $U = (\mathcal{U}, +, \cdot)$ is a vector subspace.

Remark 1.1.1

- $V = (\mathcal{V}, +)$ is an *Abelian group*;
- The subspace needs to satisfy *closure*, $\lambda x \in U, x + y \in U$.

If $v = \sum_{i=1}^{k} \lambda_i x_i$, then v is a linear combination of $(x_1, x_2, \cdots, x_k)$. If **not** all values of a solution are 0, then it's called a non-trivial solution. For $\sum_{i=1}^{k} \lambda_i x_i = 0$, if the non-trivial solution exists, then the vectors are called **linearly dependent**.

Definition 1.1.2.2: Basis and Rank

- **Generating Set**: if all vectors in V can be expressed as a linear combination of $\mathcal{A} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subseteq \mathcal{V}$, then $\mathcal{A}$ is a generating set of V;
- **Span**: The set of linear combinations of $\mathcal{A}$ is its span;
- **Basis**: The minimal generating set (linearly independent) of a vector space V is called its basis;

- **Rank**: the number of linearly independent column vectors in a matrix $\mathcal{A} \in \mathbb{R}m \times n$.

Remark 1.1.2

- $\text{rk}(A) = \text{rk}(A^\top)$;
- A matrix $A \in \mathbb{R}^{n\times n}$ is invertible if and only if $rk(A) = n$;
- The span of a matrix is also called its **image**, $dim(U) = rk(A)$;
- $Ax = b$ only has solution if and only if $rk(A) = rk(A \mid b)$;
- The solution (kernel, null space) to $Ax = 0$ has a dimension of $n - rk(A)$;

Definition 1.1.2.3: Linear Mappings

For vector spaces V, W, a mapping $\Phi : V \longrightarrow W$ is called linear mapping if:

$$\forall x, y \in V \forall \lambda, \psi \in \mathbb{R} : \Phi(\lambda x + \psi \boldsymbol{y}) = \lambda\Phi(x) + \psi\Phi(\boldsymbol{y}) \tag{1.1.2.2}$$

Determinant and Trace

Determinant is used to decide whether a matrix is invertible, denoted by $del(A)$ or $|A|$, its geometric meaning is the signed volume. *Laplace Expansian* can be used to compute the determinant for large matrix.
If $det(A) = 0$, then it's non-invertible. otherwise it's invertible.
Trace is defined as $Tr(A) = \sum_{i=1}^{n} a_{ii}$. Both determinant and trace is applied to only the squared matrix.

Eigenvalue and Postive-Definiteness

Consider matrix $A \in \mathbb{R}^{n\times n}$ as a linear mapping rather than a simple matrix, A can map a n-dimensional space to another n-dimensional space.

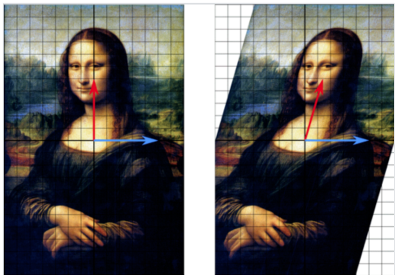


Figure 1.1: A shear mapping example

The Mona-lisa example is a linear mapping on two-dimensional space, but during the mapping, there are some vectors (such as the blue vector) that are only modified in its length rather than direction. This is the idea of the eigenvector.

Definition 1.1.2.4: Eigenvalue and Eigenvector

For a square matrix A, if there exists a scalar λ and a vector $v \in \mathbb{R}^n$, such that $A \cdot v = \lambda \cdot v$. Which would lead us to $(A - \lambda I) \cdot v = \mathbf{0}$.

Remark 1.1.3

The following statements are equivalent:

- λ is an eigenvalue of $A \in \mathbb{R}^{n \times n}$;
- $rk(A - \lambda I) < n$;
- $det(A - \lambda I) = 0$.

The definiteness is similar to the idea of eigenvalue. Consider $X^T MX$ as the inner product of X and MX, where MX can be thought as a transformed version of X. If $X^T MX > 0$ for all X, then we can get $\cos\theta = \frac{X^T \cdot MX}{||X|| \cdot ||MX||} > 0$, whcih means for all vectors, the linear mapping M can map the vector to an angle less than 90 degree.
So to show whether a matrix is postive definite, calculate all the eigenvalues of the matrix. If all values are lager than 0, then it's positive definite/

Remark 1.1.4

- $Tr(A) = \sum_i^n \lambda_i$;
- $\det(A) = \prod_{i=1}^n \lambda_i$.

1.1.3 Norms

Definition 1.1.3.1: Norm

The norm $||||$ is a mapping $V \longrightarrow \mathbb{R}$, which assigns each vector a length that satisfy:

- *Absolutely homogeneous*: $\|\lambda x\| = |\lambda| \|x\|$;
- *Triangle inequality*: $\|x + y\| \leqslant \|x\| + \|y\|$;
- *Positivity*: $||x|| \geq 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \mathbf{0}$.

The L_p norm: $\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$.

Corollary 1.1.3.1: L_p norm

1. **Manhattan Norm**: $\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i|$;
2. **Euclidean Norm**: $\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ (mostly denoted as $||\cdot||$);
3. **Infinity Norm**: $\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$.

Proof 1.1.1: Infinity Norm

By squeeze theorem:

First show that

$$\|x\|_p = (\textstyle\sum_i |x_i|^p)^{\frac{1}{p}} \le (\sum_i \max_i |x_i|^p)^{\frac{1}{p}} = n^{\frac{1}{p}} \max_i |x_i| \longrightarrow \max_i |x_i| = \|x\|_\infty,$$

Next show that $\|x\|_p = (\sum_i |x_i|^p)^{\frac{1}{p}} \ge (\max_i |x_i|^p)^{\frac{1}{p}} = \max_i |x_i| = \|x\|_\infty$.

$\mathbb{Q}.\mathbb{E}.\mathbb{D}.$

Taking derivatives of Norm

Example 1.1.3.1

$X \in \mathbb{R}^m, f(x) : \mathbb{R}^m \to \mathbb{R}^n, g(x) = ||x||_2$, what is $\nabla g(X)$?

Solution:

First $g(X) = \|f(X)\|_2 = \sqrt{\sum_{i=1}^{n} f_i(X)^2}$;

$$\nabla g(X) = \frac{1}{2}\left(\sum_{i=1}^{n} f_i(X)^2\right)^{-\frac{1}{2}} \left(\sum_{i=1}^{n} 2 f_i(X) \nabla f_i(X)\right) = \frac{J_f(X)^T f(X)}{\|f(X)\|_2} \tag{1.1.3.1}$$

1.2 Probability Theory

1.2.1 Probability Distribution

Common Discrete Distribution

Bernoulli(p)
pmf: $P(X = x \mid p) = p^x(1-p)^{1-x}; \quad x = 0, 1; \quad 0 \le p \le 1$
mean: $EX = p$; *variance*: $VarX = p(1-p)$

- A Bernoulli trial (named after James Bernoulli) is an experiment with only two possible outcomes;

- Bernoulli random variable $X = 1$ if "success" occurs and $X = 0$ if "failure" occurs where the probability of a "success" is p.

Binomial(n,p)
pmf: $P(X = x \mid n, p) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = 0, 1, \ldots, n; \quad 0 \le p \le 1$
mean: $EX = np$; *variance*: $np(1-p)$

- A Binomial experiment consists of n independent identical Bernoulli trials;

- $X = \sum_{i=1}^{n} Y_i$, where $Y_1, \cdots, Y_n$ are n identical, independent Bernoulli random variables.

Poisson(λ)
pmf: $P(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}; \quad x = 0, 1, \ldots; \quad 0 \le \lambda < \infty$
mean: $EX = \lambda$; *variance*: $VarX = \lambda$

- A Poisson distribution is typically used to model the probability distribution of the number of occurrences (with λ being the intensity rate) per unit time or per unit area;

- Binomial pmf approximates Poisson pmf. Poisson pmf is also a limiting distribution of a negative binomial distribution;

- A useful result: By Taylor series expansion: $e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$.

Assumption 1.2.1.1: Poisson Process

1. Let $X(\Delta)$ be the number of events that occur during an interval Δ;

2. The events are independent: if $\Delta_1, \cdots, \Delta_n$ are disjoint intervals, then $X(\Delta_1), \cdots, X(\Delta_n)$ are independent;

3. $X(\Delta)$ only depends on the length of Δ;
4. The probability that exactly one event occurs in a small interval of length Δt equals $\lambda\Delta t + o(\Delta t)$;
5. Poisson distribution is not memoryless, but its interval (exponential distribution) is memoryless.

Geometric(p)
pmf: $P(X = x \mid p) = p(1-p)^{x-1}; \quad x = 1, 2, \dots; \quad 0 \le p \le 1$
mean: $\frac{1}{p}$; *variance*: $\frac{1-p}{p^2}$

- The experiment consists of a sequence of independent trials;
- The property of memoryless: $P(X > s \mid X.t) = P(X > s - t)$.

Negative Binomial(r, p)
pmf: $P(X = x \mid r, p) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, \dots; \quad 0 \le p \le 1$
mean: $EX = \frac{r(1-p)}{p}$; *varaince*: $\frac{r(1-p)}{p^2}$

- assume there are many independent and identical experiments, to observe the rth success, X is the number of games to see the failure;

Hypergeometric(N, M, K)

pmf:
$$P(X = x \mid N, M, K) = \frac{\binom{M}{k}\binom{N-M}{K-x}}{\binom{N}{K}}; \quad x = 0, 1, 2, \dots, K;$$
$$M - (N - K) \le x \le M; \quad N, M, K \ge 0$$

mean: $EX = \frac{KM}{N}$; *varaince*: $\frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)}$

Common Continuous Distribution

Uniform(a, b)
pdf:$f(x \mid a, b) = \frac{1}{b-a}$; *mean*:$EX = \frac{b+a}{2}$; *variance*: $VarX = \frac{(b-a)^2}{12}$.
Exponential(β)
pdf: $f(x \mid \beta) = \frac{1}{\beta} e^{-x/\beta}, 0 \le x < \infty, \beta > 0$; *mean*: $EX = \beta$; *variance*: $VarX = \beta^2$.
Gamma(α, β)
pdf:$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad 0 \le x < \infty, \quad \alpha, \beta > 0$; *mean*:$\alpha\beta$; *variance*:$\alpha\beta^2$.

- The *gamma function* is defined as $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$;

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \alpha > 0;$
- $\Gamma(n) = (n-1)!, \quad$ for any integer $n > 0$.

Normal(μ, σ^2)
pdf:$f\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/\left(2\sigma^2\right)}, \quad -\infty < x < \infty$; *mean*:$\mu$; *variance*:$\sigma^2$.

Example 1.2.1.1

for $f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty, \quad \alpha > 0, \quad \beta > 0$:

- (a) Verify $f(x)$ is a pdf;
- (b) Derive the mean and variance of this distribution;
- (c) Prove that the variance does not exist if $\beta \leq 2$.

Solution:

(a)

$$\begin{aligned} \int_\alpha^\infty f(x)dx = \int_\alpha^\infty \frac{\beta \cdot \alpha^\beta}{x^{\beta+1}} dx = -x^{\beta \cdot \alpha^\beta | \alpha} \\ = 0 + \alpha^{-\beta} \cdot \alpha^\beta = 1 \end{aligned} \tag{1.2.1.1}$$

(b)

$$\begin{aligned} Ex = \int_\alpha^\infty xf(x)dx = \int_\alpha^\infty \frac{\beta \cdot \alpha^\beta}{X^\beta} dx \\ = \frac{\beta}{-\beta+1} \cdot \alpha^\beta \cdot x^{-\beta+1}|_\alpha^\infty = \frac{\beta \cdot \alpha}{\beta - 1} \end{aligned} \tag{1.2.1.2}$$

$$\begin{aligned} Ex^2 = \int_\alpha^\infty \frac{\beta \cdot \alpha^\beta}{x^{\beta-1}} dx = \frac{\beta}{-\beta+2} \cdot \alpha^\beta \cdot x^{-\beta+2}|_\alpha^\infty \\ = \frac{\alpha^2\beta}{\beta-2} \end{aligned} \tag{1.2.1.3}$$

$$\operatorname{Var} X = EX^2 - (EX)^2 = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)} \tag{1.2.1.4}$$

(c)
If $\beta < 2$, *then the variance is negative.*

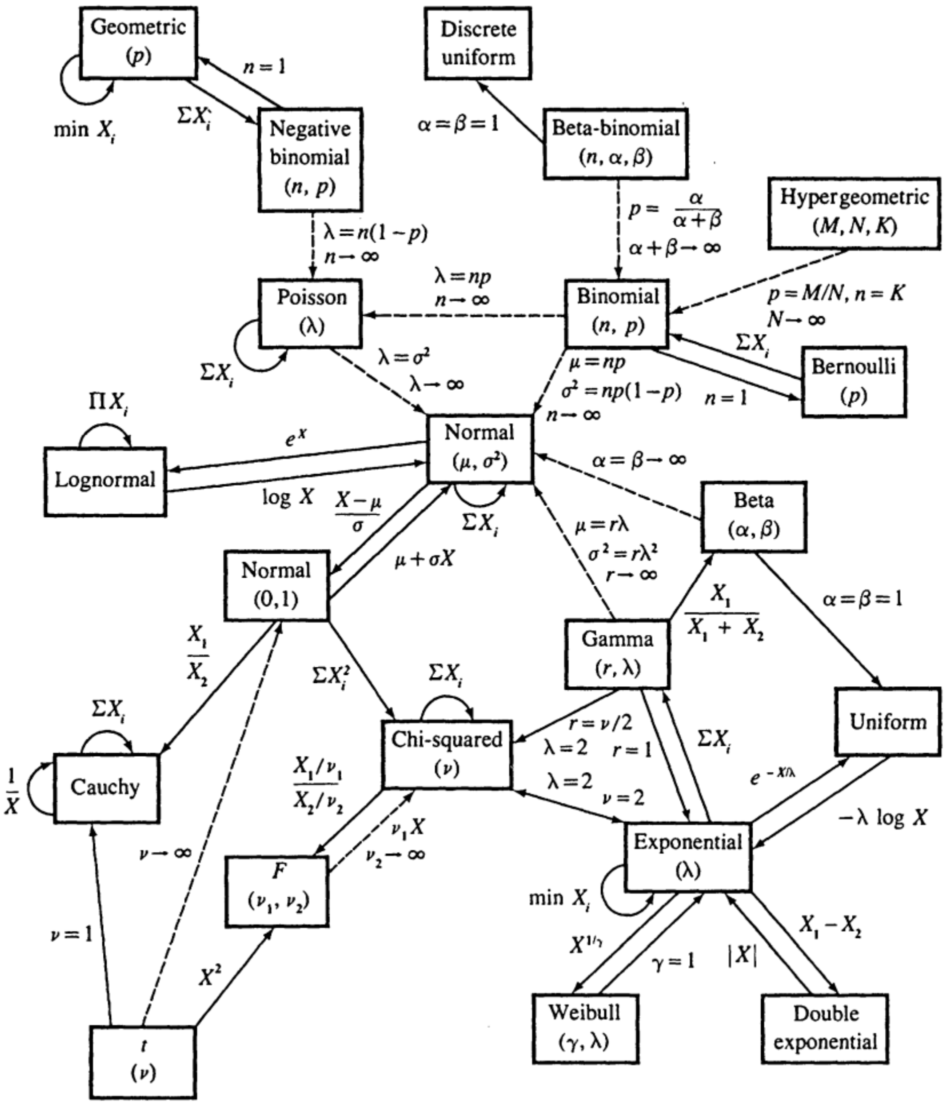


Figure 1.2: Type of distribution

1.3 Statistical Inference

This section is mainly the notes from [Casella and Berger, 2021]

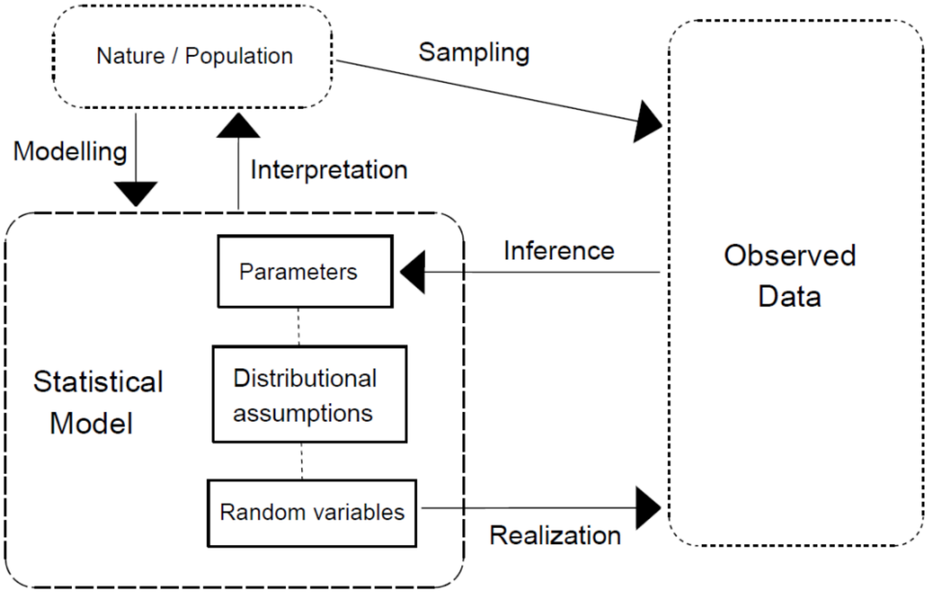


Figure 1.3: The scope of statistical inference

Statistics uses observed data to inference the statistical model.

1.3.1 Exponential Families

Definition 1.3.1.1: Exponential Families

A family of pdfs or pmfs is called an *exponential family* if:

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right) \tag{1.3.1.1}$$

Where $h(x) \leq 0, c(\boldsymbol{\theta}) \leq 0$, and $h(x), t_i(x)$ don't depend on $\boldsymbol{\theta}$. $c(\boldsymbol{\theta}), w_i(\boldsymbol{\theta})$ don't depend on x.

- Continuous: normal, gamma, beta, exponential;
- Discrete: binomial, poisson, nagative binomial;
- $\boldsymbol{\theta} = \theta_1, \theta_2, , \theta_d$, k must $\geq d$;
- If $k = d$, then it's a *full exponential family*, if $k > d$, then it's a *curved exponential family* (For example, most normal distributions are *full exponential family*, but normal distribution satisfy $\mu = \sigma^2$ is a *curved exponential family*).

To verify a pdf is an exponential family, identify the function $h(x), c(\boldsymbol{\theta}), t_i(x), w_i(\boldsymbol{\theta})$, then verify these functions satisfy the condition above.

Example 1.3.1.1

Show that Binomial, Poisson, Exponential and Normal distribution belongs to the exponential families.

Solution:

Binomial:

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x}(1-p)^n \left(\frac{p}{1-p}\right)^x \\ &= \binom{n}{x}(1-p)^n \exp\left(x \log\left(\frac{p}{1-p}\right)\right), \end{aligned} \tag{1.3.1.2}$$

Among which $\binom{n}{x}$ *is* $h(x)$, $(1-p)^n$ *is* $c(\theta)$, x *is* $t_1(x)$, *and* $\log\left(\frac{p}{1-p}\right)$ *is* $w_i(\theta)$. *Note that* $f(x \mid p)$ *is only in exponential families when* $0 < p < 1$.

Poisson:

$$\begin{aligned} & f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{1}{x!} e^{-\lambda} \exp\left(x \log(\lambda)\right) \\ & \textit{then} \\ & h(x) = \frac{1}{x!}, c(\lambda) = e^{-\lambda}, t(x) = x \text{ and} w(\lambda) = \log(\lambda). \end{aligned} \tag{1.3.1.3}$$

Exponential:

$$\begin{aligned} & f(x|\beta) = \frac{1}{\beta} \exp\left(-\frac{x}{\beta}\right) \\ & \textit{then} \\ & h(x) = 1, c(\beta) = \frac{1}{\beta}, t(x) = x \text{ and} w(\beta) = -\frac{1}{\beta}. \end{aligned} \tag{1.3.1.4}$$

Normal:

$$\begin{aligned} & f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\ & \textit{then} \\ & h(x) = 1, c(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right), \\ & t_1(x) = -\frac{x^2}{2}, w_1(\mu, \sigma) = \frac{1}{\sigma^2}, t_2(x) = x \text{ and } w_2(\mu, \sigma) = \frac{\mu}{\sigma^2}. \end{aligned} \tag{1.3.1.5}$$

Example 1.3.1.2

Show the following are exponential families:

- Gamma family with either α, β is unknown or both unknown;
- Beta family with either α, β is unknown or both unknown;
- Negative Binomial family when r is unkown.

Solution:

Gamma α unkown:

$$f(x|\alpha;\beta) = e^{-x/\beta}\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\exp((\alpha-1)\log x) \tag{1.3.1.6}$$

thus $h(x) = e^{-x/\beta}, x > 0; c(\alpha) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}; w_1(\alpha) = \alpha - 1; t_1(x) = \log x.$
Gamma $beta$ unknown:

$$f(x|\beta;\alpha) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{\frac{-x}{\beta}} \tag{1.3.1.7}$$

thus $h(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}, x > 0; c(\beta) = \frac{1}{\beta^{\alpha}}; w_1(\beta) = \frac{1}{\beta}; t_1(x) = -x.$
Gamma both unknown:

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\exp((\alpha-1)\log x - \frac{x}{\beta}) \tag{1.3.1.8}$$

thus $h(x) = I_{\{x>0\}}(x); c(\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}; w_1(\alpha) = \alpha - 1; t_1(x) = \log x; w_2(\alpha,\beta) = -1/\beta; t_2(x) = x.$
Beta α unknown:
$h(x) = (1-x)^{\beta-1}I_{[0,1]}(x), \quad c(\alpha) = \frac{1}{B(\alpha,\beta)}, \quad w_1(x) = \alpha - 1, \quad t_1(x) = \log x$
Beta β unkown:
$h(x) = x^{\alpha-1}I_{[0,1]}(x), c(\beta) = \frac{1}{B(\alpha,\beta)}, w_1(\beta) = \beta - 1, t_1(x) = \log(1-x)$
Beta both unknown:
$h(x) = I_{[0,1]}(x), c(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}, w_1(\alpha) = \alpha - 1, t_1(x) = \log x, w_2(\beta) = \beta - 1, t_2(x) = \log(1-x).$
Negative Binomial:

$$h(x) = \binom{r+x-1}{x}I_{\mathbb{N}}(x), \quad c(p) = \left(\frac{p}{1-p}\right)^{r}, \quad w_1(p) = \log(1-p), \quad t_1(x) = x. \tag{1.3.1.9}$$

Theorem 1.3.1.1: Expectation and Variance of Exponential Families

If X is a random variable that satisfies any distribution from the exponential families, then:

$$\begin{aligned}
&1. \quad \mathrm{E}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial}{\partial \theta_j} \log\left(c(\boldsymbol{\theta})\right); \\
&2. \quad \mathrm{Var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial^2}{\partial \theta_j^2} \log\left(c(\boldsymbol{\theta})\right) - \mathrm{E}\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X)\right).
\end{aligned} \tag{1.3.1.10}$$

Example 1.3.1.3

Derive the mean and variance for binomial and normal distribution using the above theorem.

Solution:

***Binomial**:*

$$h(x) = \binom{n}{x}, c(p) = (1-p)^n, t(x) = x \text{and} w(p) = \log\left(\frac{p}{1-p}\right).$$

Then,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}p} w(p) = \frac{\mathrm{d}}{\mathrm{d}p} \log\left(\frac{p}{1-p}\right) = \frac{1}{p(1-p)}, \\
&\frac{\mathrm{d}^2}{\mathrm{d}p^2} w(p) = -\frac{1}{p^2} + \frac{1}{(1-p)^2} = \frac{2p-1}{p^2(1-p)^2}, \\
&\frac{\mathrm{d}}{\mathrm{d}p} \log\left(c(p)\right) = \frac{\mathrm{d}}{\mathrm{d}p} n \log(1-p) = -\frac{n}{1-p}, \\
&\frac{\mathrm{d}^2}{\mathrm{d}p^2} \log\left(c(p)\right) = -\frac{n}{(1-p)^2}.
\end{aligned} \tag{1.3.1.11}$$

Therefore, from Theorem 3.4.2, we have

$$\begin{aligned}
&\mathrm{E}\left(\frac{1}{p(1-p)} X\right) = \frac{n}{1-p} \Rightarrow \mathrm{E}(X) = np, \\
&\mathrm{Var}\left(\frac{1}{p(1-p)} X\right) = \frac{n}{(1-p)^2} - \mathrm{E}\left(\frac{2p-1}{p^2(1-p)^2} X\right) \Rightarrow \mathrm{Var}(X) = np(1-p)
\end{aligned}$$

***Normal**:*

For Normal Distribution, we have

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

$$t_1(x) = -\frac{x^2}{2}, w_1(\mu, \sigma) = \frac{1}{\sigma^2}, t_2(x) = x \text{and} w_2(\mu, \sigma) = \frac{\mu}{\sigma^2}.$$

Then,

$$\begin{aligned}
&\frac{\partial w_1(\mu,\sigma)}{\partial \mu} = \frac{\partial(1/\sigma^2)}{\partial \mu} = 0,\\
&\frac{\partial w_2(\mu,\sigma)}{\partial \mu} = \frac{\partial(\mu/\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2},\\
&\frac{\partial w_1(\mu,\sigma)}{\partial \sigma} = \frac{\partial(1/\sigma^2)}{\partial \sigma} = -\frac{2}{\sigma^3},\\
&\frac{\partial w_2(\mu,\sigma)}{\partial \sigma} = \frac{\partial(\mu/\sigma^2)}{\partial \sigma} = -\frac{2\mu}{\sigma^3},\\
&\frac{\partial}{\partial \mu} \log\left(c(\mu,\sigma)\right) = \frac{\partial}{\partial \mu}\left(-\frac{\log(2\pi)}{2} - \log(\sigma) - \frac{\mu^2}{2\sigma^2}\right) = -\frac{\mu}{\sigma^2},\\
&\frac{\partial}{\partial \sigma} \log\left(c(\mu,\sigma)\right) = \frac{\partial}{\partial \sigma}\left(-\frac{\log(2\pi)}{2} - \log(\sigma) - \frac{\mu^2}{2\sigma^2}\right) = -\frac{1}{\sigma} + \frac{\mu^2}{\sigma^3}.\\
&\mathrm{E}\left(\frac{1}{\sigma^2}\mathrm{X}\right) = \frac{\mu}{\sigma^2} \text{ and } \mathrm{E}\left(-\frac{2}{\sigma^3}\left(-\frac{\mathrm{X}^2}{2}\right) - \frac{2\mu}{\sigma^3}\mathrm{X}\right) = \frac{1}{\sigma} - \frac{\mu^2}{\sigma^3},
\end{aligned} \tag{1.3.1.12}$$

which implies

$$\mathrm{E}(\mathrm{X}) = \mu, \mathrm{E}(\mathrm{X}^2) = \mu^2 + \sigma^2 \text{and} \mathrm{Var}(\mathrm{X}) = \mathrm{E}(\mathrm{X}^2) - (\mathrm{EX})^2 = \sigma^2$$

Definition 1.3.1.2: The indicator function

$$\mathrm{I}_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

Example 1.3.1.4

Show that $f(x \mid \theta) = \frac{1}{\theta} exp(1 - \frac{X}{\theta})$ is **NOT** an exponential family.

Solution:

$$f(x|\theta) = \frac{1}{\theta} \exp\left(1 - \frac{x}{\theta}\right) I_{[\theta,\infty)}(x) \tag{1.3.1.13}$$

At here $h(x) = I_{[\theta,\infty)}(x)$, *which is not independent with* θ.

Definition 1.3.1.3: Reparameterization of Exponential Families

$$f(x|\boldsymbol{\eta}) = h(x)c^*(\boldsymbol{\eta})\exp\left(\sum_{i=1}^{k}\eta_i t_i(x)\right) \tag{1.3.1.14}$$

Where $h(x)$ and $t_i(x)$ are identical with the original parameterization. $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)$ and $\eta_i = w_i(\theta)$. And to make it a pdf (integrates to 1):

$$c^*(\boldsymbol{\eta}) = \left[\int_{-\infty}^{\infty} h(x)\exp\left(\sum_{i=1}^{k}\eta_i t_i(x)\right)dx\right]^{-1} \tag{1.3.1.15}$$

1.3.2 Scale and Location

step 1: define the standard pdf $f(Z)$, for example: $f(Z) = e^{-Z}, Z \geq 0$;
step 2: find the relationship between X and X: $\sigma Z + \mu = X$, where σ is the scale parameter and μ is the location parameter;
step 3: replace Z using X: $f(x) = \frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ is a distribution transformed from the standard pdf.

Theorem 1.3.2.1: Sacle-Location Family
If f(x) is any pdf, for $\mu, \sigma > 0$, then the function $g(x \mid \mu, \sigma) = \frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ is a valid pdf.

Proof 1.3.1

$$\begin{aligned} &g(x|\mu,\sigma) = \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right) \geq 0 \\ &\int_{-\infty}^{\infty} g(x|\mu,\sigma)dx = \int_{-\infty}^{\infty}\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)dx \xrightarrow{(y=\frac{x-\mu}{\sigma})} \int_{-\infty}^{\infty} f(y)dy = 1. \end{aligned} \tag{1.3.2.1}$$

Q.E.D.

Remark 1.3.1

- For discrete R.V.(pmf), the above theorem doesn't hold. (ignore the $\frac{1}{\sigma}$);
- The σ is the scale parameter, the μ is the location parameter.

Example 1.3.2.1

$$\int_{\mathbf{X}}(x_1 \mid \theta) = \frac{1}{\theta}\exp\{1-\frac{x}{\theta}\}, \quad x \geq \theta \tag{1.3.2.2}$$

$$= \frac{1}{\theta}\exp\{-\frac{x-\theta}{\theta}\}\cdot \mathrm{I}_{[0,\infty)}(x) \tag{1.3.2.3}$$

Is θ a scale parameter or a location parameter?

Solution:

It depends on the standard pdf:
If the standard pdf is $f_Z(z) = \exp\{-z\}\cdot \mathcal{I}_{[0,+\infty)}(z)$*:*
then θ *serves as both parameters because* $f_X(x) = \frac{1}{\theta}\exp\{-\frac{x-\theta}{\theta}\}\cdot I_{[0,+\infty)}(\frac{x-\theta}{\theta})$.
Else if the standard pdf is $f_Z(z) = \exp\{1-z\}\cdot \mathcal{I}_{[0,+\infty)}(z)$*:*
then θ *is only a scale parameter because* $f_X(x) = \frac{1}{\theta}\exp\{1-\frac{x}{\theta}\}\cdot I_{[0,+\infty)}(\frac{x-\theta}{\theta})$.

Theorem 1.3.2.2: For any pdf $f(\cdot)$, and $\sigma, \mu > 0$. Then X is a random variable with pdf $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ **if and only if** there is a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.

Proof 1.3.2: Rigorous Proof

The ncessity of the proof can be found in 1.3.2, here the sufficiency need

Q.E.D.

Proof 1.3.3: Intuitive Proof

Q.E.D.

1.3.3 Data Reduction

The idea of data reduction is to summarize or reduce the data $X_1, X_2, \cdots, X_n$ to get the information of the unknown parameter θ.
There are many sample point $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, which are realizations (observations) of the random variable $\mathbf{X} = (X_1, X_2, \cdots, X_n)$. A **Statistic** $T(\mathbf{X})$ is a form of data reduction, or a summary of the data, $T(\mathbf{x})$ is an observation of $T(\mathbf{X})$. $\mathcal{X}$ is the **sample space**. $\mathcal{T} = \{t : t = T(\mathbf{x}),$ for $\mathbf{x} \in \mathcal{X}\}$ is the image of cX under $T(\mathbf{X})$. $T(\mathbf{X})$ partition the sapmle sapce $\mathcal{X}$ into sets $A_t = \{\mathbf{x} : T(\mathbf{x}) = t, \mathbf{x} \in \mathcal{X}\}$.

Example 1.3.3.1

Give a two-dimensional example for random variable, sample point, Statistic, observation of Statistic, sample space, image and A_t.

Solution:

Assume $\mathbf{X} = X_1, X_2$, *among which* X_1, X_2 *are Bernoulli R.V. with* $p = 0.5$. $\mathbf{x} = (0, 1)$ *is a sample point.*
The sample space $\mathcal{X}$ *is* $(0, 0), (0, 1), (1, 0), (1, 1)$, *set* $T(\mathbf{X})$ *as the statistic, then the image* $\mathcal{T}$ *is* $(0, 1, 2)$.
$A_1(\mathbf{x}) = ((1, 0), (0, 1))$, $A_2(\mathbf{x}) = (1, 1)$, $A_0(\mathbf{x}) = (0, 0)$.

1.3.4 Concentrtion Inequality

☺ General form of concentration inequality:

$$\mathbb{P}(|\overline{X} - \mathbb{E}[\overline{X}]| \leq \varepsilon(n)) \geq 1 - \delta(n) \tag{1.3.4.1}$$

where $\varepsilon(n)$ and $\delta(n)$ converge to 0 when $n \to \infty$.

Lemma 1.3.4.1: Chebyshev's inequality

Let $X_1, \cdots, X_n$ be $i.i.d$, then:

$$\mathbb{P}(|\overline{X} - \mathbb{E}[\overline{X}]| \leq z) \geq 1 - \frac{\sigma^2}{nz^2} \tag{1.3.4.2}$$

Lemma 1.3.4.2: Hoeffding's inequality

If $0 \leq X_i \leq c$, then:

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \leq z) \geq 1 - \exp(-\frac{2nz^2}{c^2}) \tag{1.3.4.3}$$

Remark 1.3.2

A convenient assignment of $z = c\sqrt{\frac{\alpha \log t}{n}}$ $, \alpha > 0, t > 1$, which yields:

$$\mathbb{P}(|\overline{X} - E[\overline{X}]| \leq c\sqrt{\frac{\alpha \log t}{n}}) \geq 1 - 2t^{-2\alpha} \tag{1.3.4.4}$$

If $X_i \in \{0, 1\}$i.i.d., and let $z = \sqrt{\frac{\log n}{n}}$, then we can have:

$$\mathbb{P}(|\overline{X} - p| \leq \sqrt{\frac{\log n}{n}}) \geq 1 - \frac{1}{n^2} \tag{1.3.4.5}$$

Lemma 1.3.4.3: The Chernoff-Hoeffding inequality

If $X_i \sim \mathcal{N}(0, 1)$, then:

$$\mathbb{P}(|\overline{X} - E[\overline{X}]| \leq \sqrt{\frac{\alpha \log t}{n}}) \geq 1 - 2t^{-\alpha/2} \tag{1.3.4.6}$$

Theorem 1.3.4.1: Gaussian Tail Bounds

If $X \sim \mathcal{N}(0, 1)$, then for $x > 0$:

$$\frac{1}{\sqrt{2\pi}}(\frac{1}{x} - \frac{1}{x^3}) \exp(-\frac{x^2}{2}) \leq \mathbb{P}(X \geq x) \leq \frac{1}{\sqrt{2\pi}x} \exp(-\frac{x^2}{2}) \tag{1.3.4.7}$$

Definition 1.3.4.1: Sub-Gaussian RV

A random variable X with mean $\mu = \mathbb{E}(X)$ is **sub-Gaussian** if there exists a positive number σ, such that:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2} \quad \forall \lambda \in \mathbb{R}. \tag{1.3.4.8}$$

For σ^2-sub-gaussian random variable X, for $z > 0$:

$$\mathbb{P}(X - \mathbb{E}[X] \leq z) \geq 1 - \exp(-\frac{z^2}{2\sigma^2}) \tag{1.3.4.9}$$

1.4 Convex Optimization

1.4.1 Linear Programming

The main contents of this section are notes from [Luenberger et al., 1984], [Bertsimas and Tsitsiklis, 1997]. The *standard form* of the linear programming:

$$\begin{aligned} \text{minimize} \quad & c_1x_1 + c_2x_2 + \ldots + c_nx_n \\ \text{subject to} \quad & a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ & a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ & \vdots \quad \vdots \\ & a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \\ \text{and} \quad & x_1 \geqslant 0, x_2 \geqslant 0, \ldots, x_n \geqslant 0, \end{aligned} \tag{1.4.1.1}$$

or the above equations can be concisely written in:

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{Ax} = \mathbf{b} \quad \text{and} \quad \mathbf{x} \geqslant \mathbf{0}. \end{aligned} \tag{1.4.1.2}$$

Convertion to standard form LP

Slack Variable

For this kind of LP formation:

$$\begin{aligned} \text{minimize} \quad & c_1x_1 + c_2x_2 + \ldots + c_nx_n \\ \text{subject to} \quad & a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leq b_1 \\ & a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \leq b_2 \\ & \vdots \quad \vdots \\ & a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \leq b_m \\ \text{and} \quad & x_1 \geqslant 0, x_2 \geqslant 0, \ldots, x_n \geqslant 0, \end{aligned} \tag{1.4.1.3}$$

The above formulation can be transformed into the following standard form:

$$\begin{aligned} \text{minimize} \quad & c_1x_1 + c_2x_2 + \cdots + c_nx_n \\ \text{subject to} \quad & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + y_1 = b_1 \\ & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n + y_2 = b_2 \\ & \vdots \quad \vdots \\ & a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n + y_m = b_m \\ \text{and} \quad & x_1 \geqslant 0, x_2 \geqslant 0, \ldots, x_n \geqslant 0, \\ \text{and} \quad & y_1 \geqslant 0, y_2 \geqslant 0, \ldots, y_m \geqslant 0. \end{aligned} \tag{1.4.1.4}$$

Now the constraint would be modified from **A** to [**A,I**], and the number of unknowns is changed from n to $n+m$.

Surplus Variable

Similar to the slack variable, formulation like:

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \geqslant b_i \tag{1.4.1.5}$$

can be transformed into:

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n - y_i = b_i \tag{1.4.1.6}$$

Free Variable

For some variables without the constraint of the sign, there are two methods to transform it into the standard form. One is to $x_i = u_i - v_i$, where both u_i and v_i are larger or equal to zero. Another method can be used when the following condition holds:

$$a_1x_1 + \cdots + a_ix_i + \cdots + a_nx_n = b_i \tag{1.4.1.7}$$

where x_i is the free variable, thus we can replace x_i by $b_i - \sum_{j \neq i}^{n} a_jx_j$, which can eliminate one variable and one constraint simultaneously.

One important example of the linear programming model is the **maximal flow problem**:

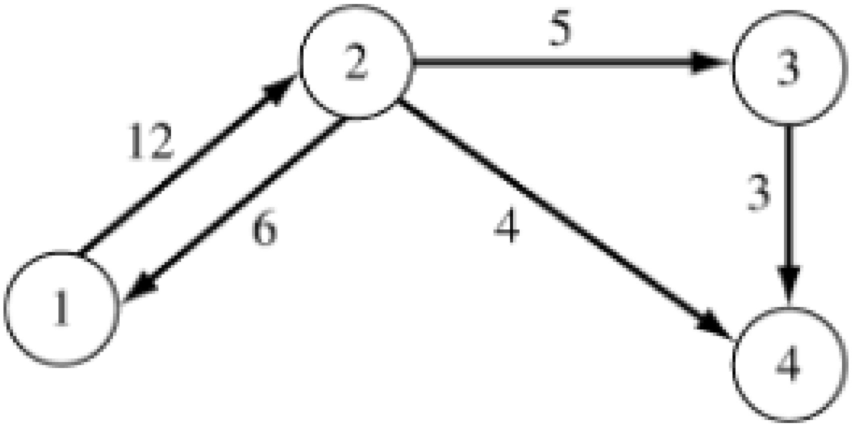


Figure 1.4: A network with capacities

This problem can be formulated into:

$$\begin{aligned}
&\text{minimize} f \\
&\text{subject to} \sum_{j=1}^{n} x_{1j} - \sum_{j=1}^{n} x_{j1} - f = 0 \\
&\sum_{j=1}^{n} x_{ij} - \sum_{j=1}^{n} x_{ji} \quad = 0, \qquad\qquad i \neq 1, m \\
&\sum_{j=1}^{n} x_{mj} - \sum_{j=1}^{n} x_{jm} + f = 0 \\
&0 \leq x_{ij} \leq k_{ij}, \qquad \text{for all } i, j
\end{aligned} \tag{1.4.1.8}$$

Basic Solutions

The system of linear equalities:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.4.1.9}$$

where $\mathbf{A}$ is a $m \times n$ constraint matrix and $\mathbf{x}$ is the $n \times 1$ decision variables.

Assumption 1.4.1.1: Full Rank Assumption

The $m \times n\mathbf{A}$ has $m < n$, and the m rows of $\mathbf{A}$ are linearly independent

. This assumption makes the linear equalities always have at least one basic solution. At least m linearly independent columns → $\mathbf{B}$, then would get:

$$\mathbf{B}\mathbf{x}_{\mathbf{B}} = \mathbf{b} \tag{1.4.1.10}$$

Definition 1.4.1.1: Basic Feasible Solution

$\mathbf{x} = (\mathbf{x}_{\mathbf{B}}, \mathbf{0})$ is the **basic solution**, the components of $\mathbf{x}$ related to the columns of $\mathbf{B}$ are **basic variables**

If the solution also satisfies $\mathbf{x} \geq 0$, then it's called a **basic feasible solution**.

If some of the basic variables have value zero, then it's called a **degenerate basic solution**.

Similar to the definition above, we have **degenerate basic feasible solution**.

Theorem 1.4.1.1: Fundamental Theorem of Linear Programming

Given a linear program in standard form, where $\mathbf{A}$ is an $m \times n$ matrix of rank m:

1. if there is a feasible solution, there is a basic feasible solution;

2. if there is an optimal feasible solution, there is an optimal basic feasible solution.

Proof 1.4.1

(1)

If there is a feasible solution $\mathbb{x}$, the solution satisfy:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{b} \tag{1.4.1.11}$$

Assume there are p of variable $x_i > 0$, then the following equation holds:

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_p\mathbf{a}_n = \mathbf{b} \tag{1.4.1.12}$$

Then there are two cases:

1. if $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$ are linearly independent, then $p \leq m$, which means $\mathbf{x}$ is already a basic solution;

2. otherwise, there would be a non-trivial solution for $y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \cdots + y_p\mathbf{a}_p = \mathbf{0}$, which means $(x_1 - \varepsilon y_1)\mathbf{a}_1 + (x_2 - \varepsilon y_2)\mathbf{a}_2 + \cdots + (x_p - \varepsilon y_p)\mathbf{a}_p = \mathbf{b}$.
   Then for any value of ϵ, $\mathbf{x}-\varepsilon\mathbf{y}$ is a solution but may violate the signal constraint. Mention that there is at least one y_i that is negative or positive, thus there is at least one x_i decreasing when we increase the $\epsilon(\epsilon > 0)$, thus we set $\varepsilon = \min\{x_i/y_i : y_i > 0\}$, which would bring us to a new feasible solution but the number of zeros is larger.
   Through such iteration, we can get at least one basic feasible solution.

(2)

The idea is the same as the first one. In case one it's obvious. in case two, we need to prove for any , $\mathbf{x} - \varepsilon\mathbf{y}$ is still optimal.

Note the new value is $\mathbf{c}^T\mathbf{x} - \varepsilon\mathbf{c}^T\mathbf{y}$. Because ε can be both positive and negative, thus $\mathbf{c}^T\mathbf{y} = 0$, which makes $\mathbf{x} - \varepsilon\mathbf{y}$ still as optimal solution.

Q.E.D.

Remark 1.4.1

This theorem reduce the original problem to the size of $\begin{pmatrix} n \\ m \end{pmatrix} = \frac{n!}{m!(n-m)!}$.

Relationship with the Convex Optimization

Definition 1.4.1.2: Extreme Point

A point $\mathbf{x}$ in a convex set C is an *extreme point* if there are **no** two distinct $\mathbf{x}_1, \mathbf{x}_2 \in C$ such that $\mathbf{x} = \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2$ for some $\alpha, 0 < \alpha < 1$

Theorem 1.4.1.2: Equivalence of extreme points and basic solutions

Let $\mathbf{A}$ be an $m \times n$ matrix with rank m, Let $\mathbf{K}$ denote the *convex polytope* consisting all vector $\mathbf{x}$ satisfying $\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geqslant \mathbf{0}$.

A vector $\mathbf{x}$ is an extreme point of $\mathbf{K}$ if and only if $\mathbf{x}$ is a basic feasible solution.

Proof 1.4.2

(1) BFS → extreme point:

Suppose $\mathbf{x} = (x_1, x_2, \ldots, x_m, 0, 0, \ldots, 0)$ is a BFS, it satisfies $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_m\mathbf{a}_m = \mathbf{b}$. If $\mathbf{x} = \alpha\mathbf{y} + (1-\alpha)\mathbf{z}$, since the value in $\mathbf{y}, \mathbf{z}$ is larger than 0, then we have:

$$y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \cdots + y_m\mathbf{a}_m = \mathbf{b}z_1\mathbf{a}_1 + z_2\mathbf{a}_2 + \cdots + z_m\mathbf{a}_m = \mathbf{b} \tag{1.4.1.13}$$

Because the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ are linearly independent, then we can get $\mathbf{x} = \mathbf{y} = \mathbf{z}$, which means that $\mathbf{z}$ is an extreme point.

(2) Extreme point → BFS:

Assume that $\mathbf{x}$ has k components larger than zero, then we have: $y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \cdots + y_k\mathbf{a}_k = 0$. Assume that $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ are linearly dependent, which would lead to:

$$y_1\mathbf{a}_1 + y_2\mathbf{a}_2 + \cdots + y_k\mathbf{a}_k = 0 \tag{1.4.1.14}$$

Define $\mathbf{y} = (y_1, y_2, \ldots, y_k, 0, 0, \ldots, 0)$, it's obvious to see the following can exist:

$$\mathbf{x} + \epsilon\mathbf{y} \geqslant 0, \quad \mathbf{x} - \epsilon\mathbf{y} \geqslant 0 \tag{1.4.1.15}$$

Contradicts that $\mathbf{x}$ is an extreme point $\longrightarrow$ $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_k$ are linearly independent $\longrightarrow$ $\mathbf{x}$ is a BFS.

Q.E.D.

Corollary 1.4.1.1

1. If the convex set $\mathbf{K}$ is nonempty, there is at least one extreme point;
2. If there is a finite optimal solution to a linear programming problem, there is a finite optimal solution which is an extreme point of the constraint set.;

3. The constraint set **K** possesses at most a finite number of extreme points;
4. If K is bounded, then it's a *convex polyhedron*.

Simplex Method

The standard form linear programming can be transformed to the **canonial form** (reduced row-echelon form/tabular method). The canonical form provides basic variables and non-basic variables. **Pivot equations** can transform a non-basic variable into a basic variable.

$$\begin{cases} \bar{a}'_{ij} = \bar{a}_{ij} - \frac{\bar{a}_{iq}}{\bar{a}_{pq}}\bar{a}_{pj}, i \neq p \\ \bar{a}'_{pj} = \frac{\bar{a}_{pj}}{\bar{a}_{pq}}. \end{cases} \tag{1.4.1.16}$$

1.4.2 Convexity

Definition 1.4.2.1: Differentiability

A mapping $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is said to be *differentiable* at x if there is a function: $DF : \mathbb{R}^n \longrightarrow \mathbb{R}^{m\times n}$ such that:

$$\lim_{h\to 0} \frac{\|F(x+h) - F(x) - DF(x)\cdot h\|}{\|h\|} = 0. \tag{1.4.2.1}$$

The matrix $DF(x) \in \mathbb{R}^{m\times n}$ is the Jacobian Matrix.

If F is a $\mathbb{R}^n \longrightarrow \mathbb{R}$ mapping, then the **gradient** can be expressed as:

$$\nabla f(x) = Df(x)^\top = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix} \tag{1.4.2.2}$$

The **Hessian Matrix** (symmetric matrix) can be expressed by:

$$\nabla^2 f(x) = H_f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1 \partial x_1}(x) & \frac{\partial f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial f}{\partial x_1 \partial x_n}(x) \\ . & . & \cdots & . \\ . & . & . & . \\ . & . & . & . \\ . & . & \cdots & . \\ \frac{\partial f}{\partial x_n \partial x_1}(x) & \frac{\partial f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial f}{\partial x_n \partial x_n}(x) \end{pmatrix} \tag{1.4.2.3}$$

Example 1.4.2.1

Calculate the gradient of $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$, where $A \in \mathbb{R}^{n\times n}$ be symmetric.

Solution:

Thus $\nabla f(x) = Ax + b$ *Use the definition of Differentiability, first calculate* $f(x+h) - f(x)$*:*

$$\begin{aligned}
f(x+h) - f(x) &= \frac{1}{2}(x+h)^T A(x+h) + b^T(x+h) + c - f(x) \\
&= (x+h)^T Ax + (x+h)^T Ah + b^T x + b^T h - \frac{1}{2}x^\top Ax + b^\top x \\
&= \frac{1}{2}h^T Ax + \frac{1}{2}x^T Ah + \frac{1}{2}h^T Ah + b^T h \\
&= \frac{1}{2}(h^T Ax)^T + \frac{1}{2}x^T Ah + \frac{1}{2}h^T Ah + b^T h \\
\\
&= X^T Ah + b^T h
\end{aligned}$$

(1.4.2.4)

Thus $\nabla f(x) = Ax + b$

Remark 1.4.2

FONC (First-Order Necessary Conditions)

If $x^\star$ is a local minimizer of the unconstrained problem, then we must have $\nabla f(x^*) = 0$.

Theorem 1.4.2.1: SONC (Second Order Necessary Condition)

If x^* is a local minimizer of f, then it holds that:

1. $\nabla f(x^*) = 0$;

2. For all $d \in \mathbb{R}^n$:$d^\top \nabla^2 f(x^*) d \geq 0$ ($\nabla^2 f(x^*)$ is positive semidefinite);

Definition 1.4.2.2: Saddle Point

A point satisfying FONC is a **stationary point**, a stationary point with indefinite Hessian matrix is called **saddle point**.

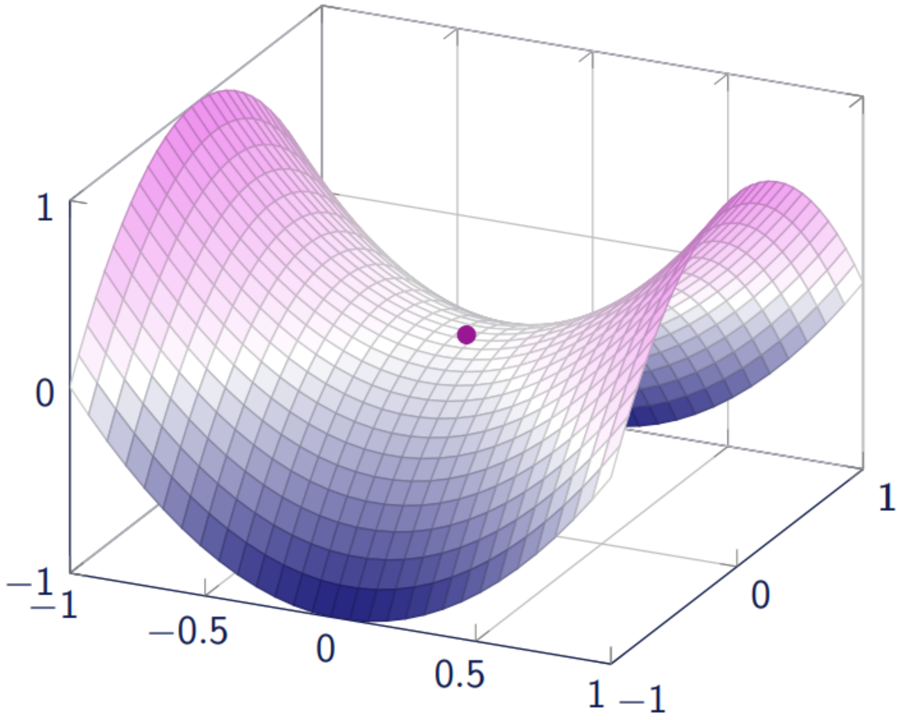


Figure 1.5: An example of saddle point

Theorem 1.4.2.2: SOSC (Second Order Sufficient Conditions)

1. $\nabla f(x^*) = 0$;
2. For all $d \in \mathbb{R}^n$:$d^\top \nabla^2 f(x^*) d > 0$ ($\nabla^2 f(x^*)$ is positive definite);

Then x^* is a *strict local minimum* of f.

Proof 1.4.3

By *taylor expansion*, $f(x^* + td) = f(x^*) + \frac{1}{2}t^2 d^\top \nabla^2 f(x^*) d + o(t^2) > f(x^*)$.
$\mathbb{Q}.\mathbb{E}.\mathbb{D}.$

Definition 1.4.2.3: Coercivity

A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **coercive** if:

$$\lim_{\|x\| \to \infty} f(x) = +\infty \tag{1.4.2.5}$$

What's more, if f is a coercive function, then the level set $L_{\leq \alpha} := \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$ is *compact* and has at least one *global minimizer*.

Definition 1.4.2.4: Convex Sets and Functions

Convex Sets

A set $X \subseteq \mathbb{R}^n$ is **convex** if for any $x, y \in X$, and any $\lambda \in [0, 1]$, we have $\lambda x + (1-\lambda)y \in X$. For example, the half space $H := \{x \in \mathbb{R}^n : a^\top x \leq b\}$ and the closed ball $B_r(a) := \{x \in \mathbb{R}^n : \|x - a\| \leq r\}$ are both convex sets.

Corollary 1.4.2.1: Intersection of Convex Sets

The intersection of convex sets is a convex set. For example, the *Polyhedral Sets*: $\{x \in \mathbb{R}^n : Ax \leq b\}$.

Convex Functions

A function f is said to be *convex* on a *convex sets* X is for every $x, y \in X$ and any $0 \leq \lambda \leq 1$:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \tag{1.4.2.6}$$

Examples are The Euclidean norm $f(x) = \|x\| = \sqrt{x^\top x}$ and affine-linear functions $f(x) = a^\top x + b$.

If a function f has the property $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$, then it's said to be **strongly convex**.

Strongly Convex (with parameter μ)

$$f(\lambda x + (1-\lambda)y) + \frac{\mu\lambda(1-\lambda)}{2}\|y - x\|^2 \leq \lambda f(x) + (1-\lambda) f(y) \tag{1.4.2.7}$$

This is equivalent to $f - \frac{\mu}{2}\|\dot{}\|^2$.

Lemma 1.4.2.1: General Composition

Let $h : X \to \mathbb{R}$ be *convex* and $g : Y \to \mathbb{R}$ be *convex* and **non-decreasing**. Then $f(x) = g(h(x))$ is convex.

Theorem 1.4.2.3: Convexity and Differentiability

f is convex **if and only if**:

$$f(y) - f(x) \geq \nabla f(x)^\top (y - x), \quad \forall\ x, y \in X \tag{1.4.2.8}$$

Or

$$h^\top \nabla^2 f(x) h \geq 0, \quad \forall\ h \in \mathbb{R}^n, \quad \forall\ x \in X \tag{1.4.2.9}$$

Theorem 1.4.2.4: Convexity and Optimality

If f is a convex function and X is a convex set, then for the problem $\min f(x) \quad \text{s.t.} \quad x \in X$, we have:

- Every local minimizer is also a global minimizer;
- If f is strongly convex, then it has at most one global minimizer, which is the stationary point.

1.4.3 Convex Optimization ALgorithms

Chapter 2

Economics and Econometrics

2.1 Development Economics

2.1.1 Models of Development Economics

2.1.2 Clan Culture

Definition 2.1.2.1: Clan Culture
A clan is a consolidated kin group made up of component families that trace their patrilineal descent from a common ancestor.

History of Clans

”Modern” clan originated in the Song Dynasty (860-1279 CE). At that time Neo-Confucian ideology was formed, which provided the theoretical basis as well as clan organization structure design. The characteristics of ”modern” clan culture:

1. The families of a clan lived in the same or several nearby communities;

2. Common properties and organized routine group activities, resource pooling;

3. Compilation of genealogies;

4. Own internal governance structures.

Currently although China has been transitioning for a long time from a traditional society to a modern society, clan culture is still prevalent and has a broad impact on the lives of Chinese people, especially in rural areas.

[Bertrand and Schoar, 2006] shows the positive correlation between the fraction of family control among listed firms and family ties using cross-country level data. [Cheng et al., 2021] uses IV (the minimum distance to two prominent neo-Confucian academies, the Kaoting Academy

(Kaoting Shuyuan) and the Xiangshan Academy (Xiangshan Shuyuan)) to identify that clan culture causes higher firm ownership concentration.

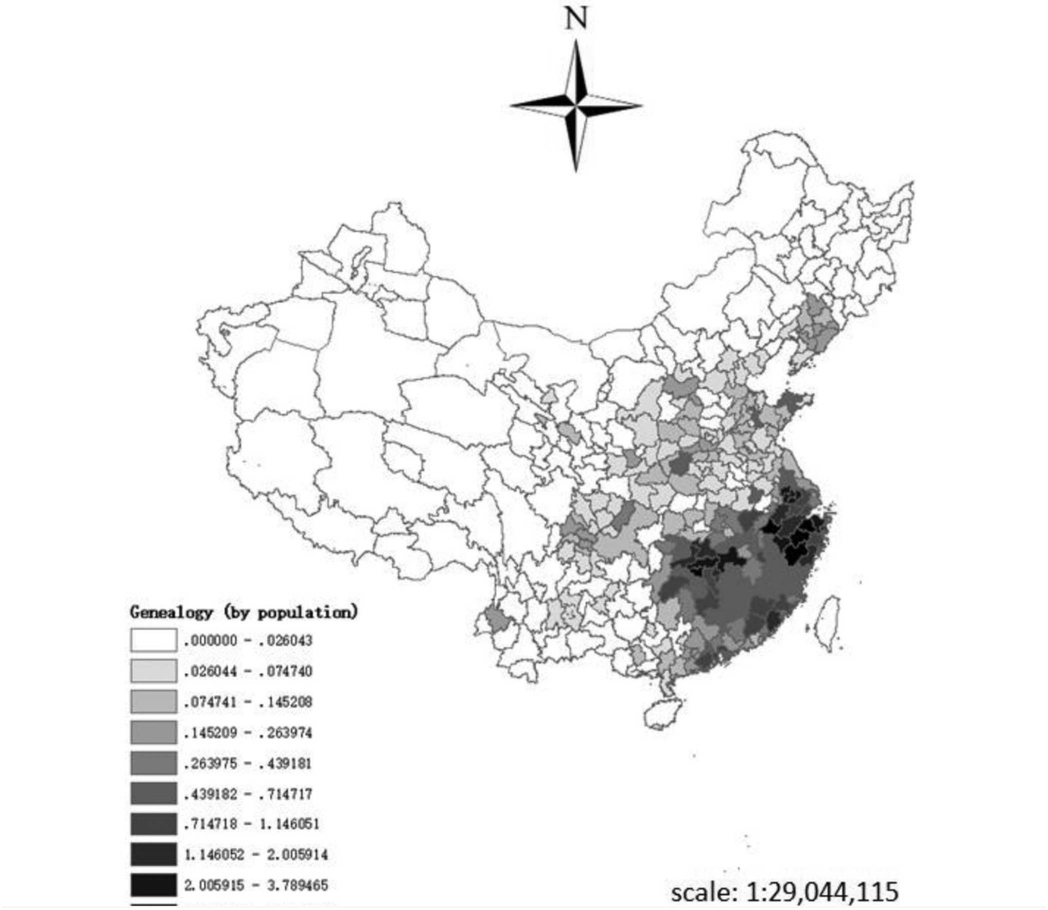


Figure 2.1: Clan Culture Intensity

Potential reasons that culture affects the concentration of family ownership:

1. Clan culture fosters high trust within the family and low trust in outsiders(**short-radius trust attitude**). According to agency-cost-based theories, family ownership can be concentrated in such a situation.

2. *Resources Pooling*: commom property ownership.

3. *Amenity Potential*: other things constant, owners subject to stronger influences of clan culture could have a higher utility.

[Zhang, 2020] estimates the effect of clan on entrepreneurship. He finds that clan leads to a higher occurrence of entrepreneurship by helping overcome financing constraints and escape from local governments' "grabbing hand." [Zhang, 2019] investigates the relationship between the low take-up rate of social pensions and the clan culture intensity. In his article, dummy variable $temple_c$

(whether community c has ancestral temple) is constructed as the proxy variable for the strength of clan culture. Some interesting insights are obtained:

1. Clan culture is positively related to adults raising children for support in their old age;
2. Clan culture is associated with a larger number of children being born and a higher probability of having sons;
3. Clan culture is associated with a higher coresidence rate between old parents and adult sons;
4. Clan culture is associated with a higher likelihood of receiving financial transfers from non-coresident children;
5. Clan culture is associated with a lower likelihood of participating in rural pension programs.

[Cao et al., 2022] There is much research about clan culture outside of Mainland China. [Yang, 2019] found the concave relationship the between the **heterogeneity** of clan family and the provision of public goods. This finding implies that group homogeneity yields not only benefits, but also some possible costs. reCommon Control Variables in Clan Research Identification(Individual Level):

1. *Hukou* Status;

Regional Level:

1. Distance to the sea;

Data resources in Clan Research

Remark 2.1.1

- *Comprehensive Catalogue of the Chinese Genealogy* can be used to construct the strength of clan culture;
-

2.2 Reduced-Form Identification

The main contents of this chapter are the notes of [Angrist and Pischke, 2014], [Angrist and Pischke, 2009].

2.2.1 Randomized Trials

Keywords 2.1

ATT, ATE, Counter-factual World, Potential Outcome

The outcome is Y_i, the potential outcome is Y_{1i}, Y_{0i}:

$$Y_i = \begin{cases} Y_{1i} & if D_i = 1 \\ Y_{0i} & if D_i = 0 \end{cases} Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i \tag{2.2.1.1}$$

Naive Comparison:

$$\begin{aligned} \{\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0]\} = \{&\mathbf{E}[Y_{1i}|D_i = 1] - \mathbf{E}[Y_{0i}|D_i = 1]\} \\ &+ \{\mathbf{E}[Y_{0i}|D_i = 1] - \mathbf{E}[Y_{0i}|D_i = 0]\} \end{aligned} \tag{2.2.1.2}$$

- observed difference: $\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0]$;
- ATT: $\mathbf{E}[Y_{1i}|D_i = 1] - \mathbf{E}[Y_{0i}|D_i = 1]$;
- selection bias: $\mathbf{E}[Y_{0i}|D_i = 1] - \mathbf{E}[Y_{0i}|D_i = 0]$.

The existence of the selection bias is due to the dependence between D_i and the **potential** outcome Y_{1i}, Y_{0i}. Randomization can make $D_i \perp Y_{1i}, Y_{0i}$:

$$E[Y_{1i}|D_i = 1] = E[Y_{1i}|D_i = 0] = E[Y_{1i}]E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0] = E[Y_{0i}] \tag{2.2.1.3}$$

So selection $= \mathbf{E}[Y_{0i}|D_i = 1] - \mathbf{E}[Y_{0i}|D_i = 0] = 0$. 😎 Besides, by randomization, $ATT = ATE$. Randomization example:

- Rand HIE experiment: whether the insurance program makes people healthier;
- STAR: the effects of class size on education;
- OHP: this group experiment is not perfect because the group is not a determinant of whether to receive the treatment, but the treatment group does have a higher probability to get the treatment (**Instrumental Variable** can handle this situation);

Remark 2.2.1

- By randomization, the individual differences still exist;
- Checking for balance is an important step in randomization;
- The most critical idea of randomization is **Other Things Equal**(*ceteris paribus*);
- Randomization was invented by *Ronald Aylmer Fisher* in 1925.

2.2.2 Regression and Matching

Keywords 2.2

- CEF, CEF decomposition, ANOVA;
- regression justification,

Conditional Expectation Function (CEF) is a population concept:

$$\begin{aligned} E[Y_i|X_i = x] &= \int t f_y(t|X_i = x)dt \\ E[Y_i|X_i = x] &= \sum_t tP(Y_i = t|X_i = x) \end{aligned} \tag{2.2.2.1}$$

Lemma 2.2.2.1: The law of iterated expectations

$$E\left[\mathrm{y}_i|\mathrm{X}_i = x\right] = \int t f_y\left(t|\mathrm{X}_i = x\right)dt. \tag{2.2.2.2}$$

Proof 2.2.1

$$\begin{aligned}
E\{E\left[\mathrm{y}_i|\mathrm{X}_i\right]\} = & \int E\left[\mathrm{y}_i|\mathrm{X}_i = u\right] g_x(u)du \\
= & \int \left[\int t f_y\left(t|\mathrm{X}_i = u\right) dt\right] g_x(u)du \\
= = & \int \int t f_y\left(t|\mathrm{X}_i = u\right) g_x(u)dudt \\
= & \int t \left[\int f_y\left(t|\mathrm{X}_i = u\right) g_x(u)du\right] dt = \int t \left[\int f_{xy}\left(u, t\right) du\right] dt \\
= & \int t g_y(t)dt.
\end{aligned} \tag{2.2.2.3}$$

Q.E.D.

♡ 3 important property of CEF:

Theorem 2.2.2.1: CEF Decompostion Property

$$Y_i = E[Y_i|X_i] + \epsilon_i \tag{2.2.2.4}$$

where ϵ_i is mean independent of X_i, and X_i is uncorrelated with any function of X_i.

Proof 2.2.2

Take the expectation of X_i at both sides:

$$\begin{aligned}
E[Y_i|X_i] &= E[E[Y_i|X_i]|X_i] + E[\epsilon_i|X_i] \\
E[\epsilon_i|X_i] &= E[Y_i|X_i] - E[Y_i|X_i] = 0
\end{aligned} \tag{2.2.2.5}$$

$$E[\epsilon_i] = \int_{X_i} f_x(t)E[\epsilon_i|X_i]dt = \int_{X_i} 0dt = 0 = E[epsilon_i|X_i] \tag{2.2.2.6}$$

Q.E.D.

This means that Y_i can be decomposed into 2 parts: explaind by X_i and terms uncorrelated with X_i.

Theorem 2.2.2.2: CEF Prediction Property

CEF is the best estimator of Y_i in the MMSE sense, which means:

$$E\left[\mathrm{y}_i|\mathrm{X}_i\right] = \arg\min_{m(\mathrm{X}_i)} E\left[\left(\mathrm{Y}_i - m\left(\mathrm{X}_i\right)\right)^2\right] \tag{2.2.2.7}$$

Proof 2.2.3

$$\begin{aligned}\left(\mathrm{Y}_i - m\left(\mathrm{X}_i\right)\right)^2 &= \left(\left(\mathrm{Y}_i - E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\right) + \left(E\left[\mathrm{Y}_i|\mathrm{X}_i\right] - m\left(\mathrm{X}_i\right)\right)\right)^2 \\ &= \left(\mathrm{Y}_i - E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\right)^2 + 2\left(E\left[\mathrm{Y}_i|\mathrm{X}_i\right] - m\left(\mathrm{X}_i\right)\right)\left(\mathrm{Y}_i - E\left[\mathrm{Y}_i|\mathrm{X}_i\right]\right) \\ &\quad + \left(E\left[\mathrm{Y}_i|\mathrm{X}_i\right] - m\left(\mathrm{X}_i\right)\right)^2\end{aligned} \tag{2.2.2.8}$$

The formula has the lowest constant value by setting $m\left(\mathrm{X}_i\right) = \mathrm{CEF}$.

Q.E.D.

Theorem 2.2.2.3: ANOVA Theorem

$$\mathrm{V}(Y_i) = E[\mathrm{V}(Y_i|X_i)] + \mathrm{V}(E[Y_i|X_i]) \tag{2.2.2.9}$$

This indicates that the variance of Y_i can be decomposed into two parts:

1. the variance of the CEF;
2. the variance of the residual;

Remark 2.2.2

- The CEF property dosen't rely on any assumption! It has nothing to do with regression right now;
- If X_i is not mean independent of Y_i, then by ANOVA theorem, the variance of the outcome variable controlled by X_i could be smaller;

CEF and (Population) Regression

$$\begin{aligned}&\beta = \arg\min_b E[(Y_i - X_i'b)^2] \\ &1storder : E[X_i(Y_i - X_i'b)] = 0 \\ &solution : \beta = E[X_iX_i']^{-1}E[X_iY_i]\end{aligned} \tag{2.2.2.10}$$

Theorem 2.2.2.4: Regression Anatomy

$$\beta_k = \frac{Cov(Y_i, \tilde{X_{ki}})}{V(\tilde{X_{ki}})} \tag{2.2.2.11}$$

Corollary 2.2.2.1: Bivariate Case

$$\beta = \frac{Cov(Y_i, \tilde{X}_i)}{V(\tilde{X}_i)} \tag{2.2.2.12}$$

Proof 2.2.4

Substitute

$$Y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i \tag{2.2.2.13}$$

$\tilde{x}_{ki}$ is uncorrelated with e_i and other covariates by construction, thus $Cov(\tilde{x}_{ki}, x_{ki}) = Var(\tilde{x}_{ki})$, thus $Cov(Y_i, \tilde{x_{ki}} = \beta_k x_{ki})$.

Q.E.D.

Remark 2.2.3

The regression anatomy shows that each β_k in multi-regression is the bivariate slope after "partialing out" all the other regressors.

☺ Why the population regression coefficient is what we are interested in (Link with CEF):

Theorem 2.2.2.5: Regression Justification

1. Suppose the CEF is linear, then the population regression function is it;
2. In any condition, $\mathrm{X}_i'^{\beta}$ is the best predictor of Y_i in a MMSE sense;
3. The function $X_i'^{\beta}$ provides the MMSE linear approximation to $E[Y_i|X_i]$.

$$\beta = \arg\min_b E\{(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2\} \tag{2.2.2.14}$$

Proof 2.2.5

Suppose $E[\mathrm{Y}_i|\mathrm{X}_i] = \mathrm{X}_i^{\prime\beta^*}$. By regression decomposition theorem:

$$\begin{gathered} E[X_i(Y_i - X_i'\beta^*)] = 0 \\ \beta^* = E[X_iX_i']^{-1}E[X_iY_i] = \tilde{\beta} \end{gathered} \tag{2.2.2.15}$$

$$\begin{aligned} \left(\mathrm{Y}_i - \mathrm{X}_i'b\right)^2 = & \quad \{(\mathrm{y}_i - E[\mathrm{y}_i|\mathrm{X}_i]) + (E[\mathrm{y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)\}^2 \\ = & \quad (\mathrm{y}_i - E[\mathrm{y}_i|\mathrm{X}_i])^2 + (E[\mathrm{y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2 \\ & \quad +2(\mathrm{y}_i - E[\mathrm{y}_i|\mathrm{X}_i])(E[\mathrm{y}_i|\mathrm{X}_i] - \mathrm{X}_i'b). \end{aligned} \tag{2.2.2.16}$$

$\mathbb{Q}.\mathbb{E}.\mathbb{D}.$

Corollary 2.2.2.2

- For saturated model, the population linear regression is the CEF;
- For single dummy variable, the coefficient is the mean probability of receiving treatment;

From Regression to Causality

2.2.3 Asymptotic Analysis

Chapter 3

Data Science

3.1 Cloud Computing

This section is mainly the notes of [Hwang, 2017].

3.1.1 Principles of Cloud Computing System

Keywords 3.1

- HTC, HPC, physical machine, virtual machine, VM clusters;
- IaaS, PaaS, SaaS;
- Amdahl's Law, Gustafson's Law, cloud availability.

Definition 3.1.1.1: HPC and HTC

- **HTC** refers to *highly-throughput computing* system built with parallel and distributed computing technologies;
- **HPC** refers to *highly-performance computing* system in terms of raw speed in batch processing.

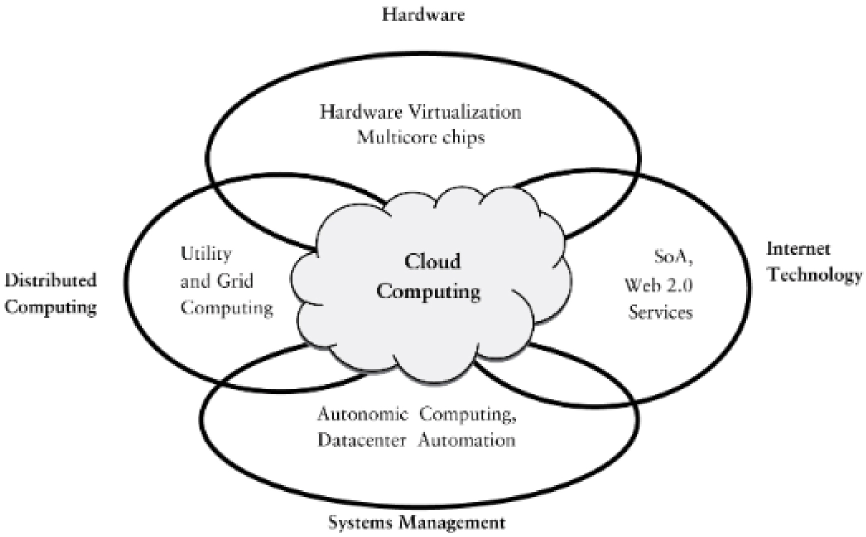


Figure 3.1: Cloud Computing Technology Convergence

Definition 3.1.1.2: Cloud

A cloud is a pool of virtualized computer resources. A cloud can host a variety of different workloads, including batch-style backend jobs and interactive, user-facing applications. Some view the clouds as computing clusters with modest changes in *virtualization*.

Example 3.1.1.1

1. Distinguish between physical machines and virtual machines.
2. What are the fundamental differences between CPU and GPU as building blocks in a modern computer, or datacenter, or a cloud system?
3. What are the fundamental differences between traditional data centers and modern cloud platforms?
4. Differences between HPC and HTC.

Solution:

1

1. *PM runs on hardware, VM runs on virtualized layer such as hypervisor or container;*
2. *The resource of PM is fixed, for VM it can be dynamically adjusted;*

(2)
CPU has several cores and low latency, good for serial processing; GPU has many cores,

high throughput, good for parallel processing.

(3)

One is HTC, built with many servers and the computing services are fixed; Cloud is HTC, built with virtualized clusters and more flexible.

(4)

HPC tasks need large amounts of computing power for short periods, whereas HTC tasks also require large amounts of computing, but for much longer times;

HPC utilizes centralized high-performance processors to deal with a single large task, HTC utilizes less powerful processors from many servers to deal with the task that requires high throughput.

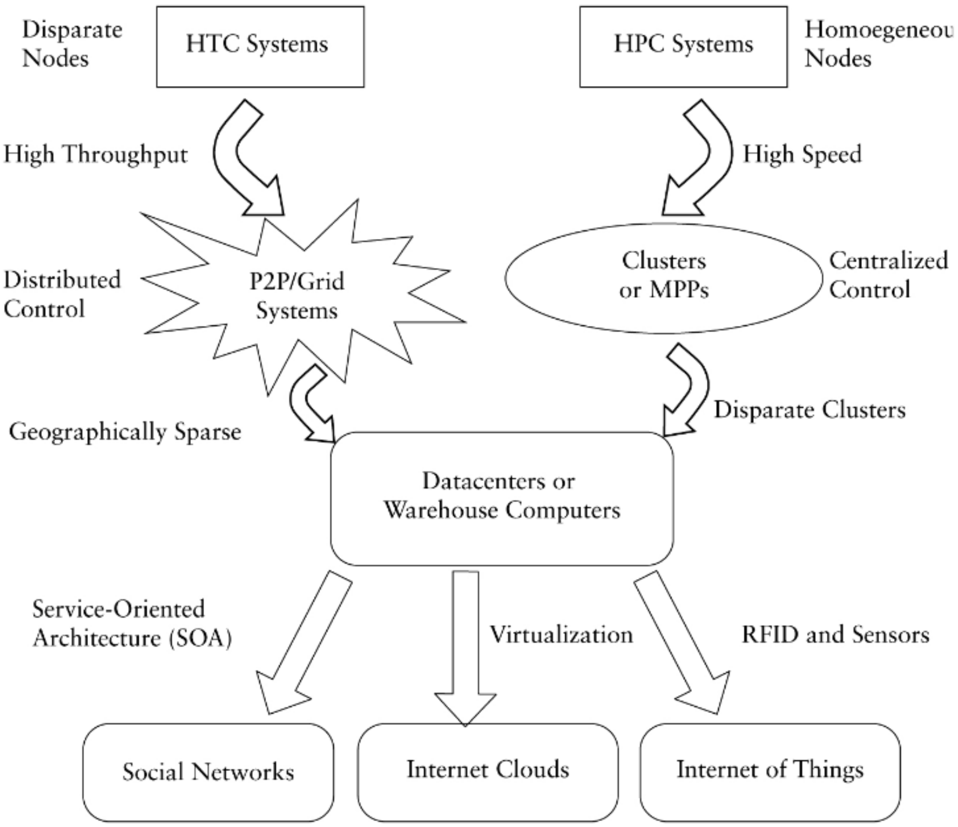
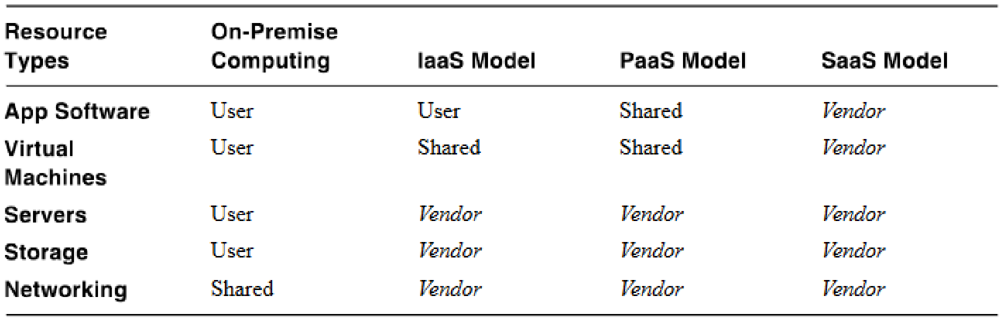


Figure 3.2: From HPC systems and clusters to grids, p2p networks, clouds and IoT

Basic cloud service models are *infrastructure as a service* (IaaS) or infrastructure cloud, *platform as a service* (PaaS) or platform cloud, and *software as a service* (SaaS) or application cloud. Their differences with on-premise computing are listed below:



Resource Types	On-Premise Computing	IaaS Model	PaaS Model	SaaS Model
App Software	User	User	Shared	*Vendor*
Virtual Machines	User	Shared	Shared	*Vendor*
Servers	User	*Vendor*	*Vendor*	*Vendor*
Storage	User	*Vendor*	*Vendor*	*Vendor*
Networking	Shared	*Vendor*	*Vendor*	*Vendor*

Figure 3.3: Comparing three cloud service models with on-premise computing

- **IaaS**: AWS, GoGrid;
- **PaaS**: Google App Engine, Microsoft Azure, Salesforce;
- **SaaS**: CRM, ERP, HR, Hadoop, Google Docs.

A physical cluster is a collection of servers (PMs) interconnected by a physical network. **virtual clusters** are built with multiple VMs installed at PM belong to one or more physical clusters.

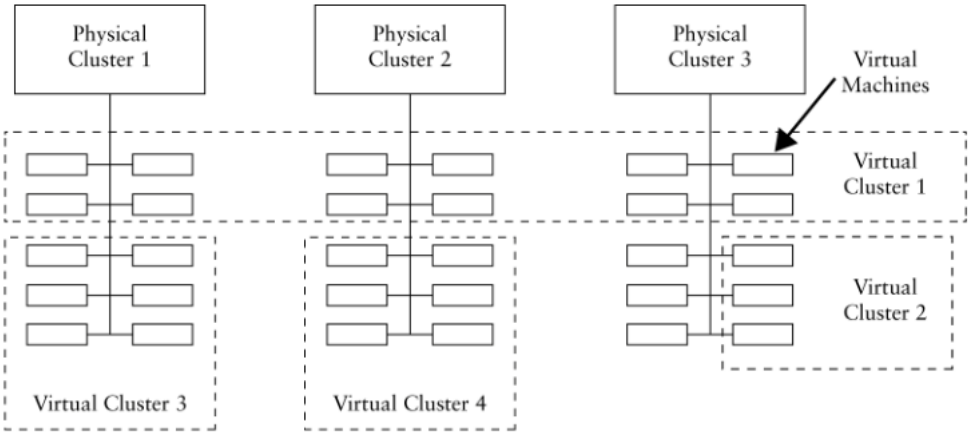


Figure 3.4: physical cluster and virtual cluster

The virtual clusters have the following interesting properties:

1. Multiple VMs running with different OSs can be deployed on the same physical node;
2. A VM runs with a guest OS, which is often different than the host OS that manages the resources in the PM, where the VM is implemented;
3. Using VMs can greatly enhance server utilization and application flexibility;
4. VMs can be colonized (replicated) in multiple servers for fault tolerance and disaster recovery;

5. The size (number of nodes) of a virtual cluster can grow or shrink dynamically;
6. the failure of any physical nodes may disable some VMs installed on the failing nodes but the failure of VMs will not pull down the host system.

Cloud Scalability

There are two fundamental issues on cloud performance: **Scalability** and **Availability**. The total execution time of the program is calculated by $\alpha\mathbb{T} + (1-\alpha)\mathbb{T}/n$.

Theorem 3.1.1.1: Amdahl's Law

The total execution time of the program is calculated by $\alpha\mathbb{T} + (1-\alpha)\mathbb{T}/n$.

$$Speedup\ factor = S = T/[\alpha T + (1-\alpha)T/n] = 1/[\alpha + (1-\alpha)/n] \tag{3.1.1.1}$$

- The communication time and I/O time are excluded in this formula;
- When $n \longrightarrow \infty$, the S can have the upper bound $\frac{1}{\alpha}$, thus α is the *sequential bottleneck* here;
- This speedup is called *fixed-workload speedup*;
- The *cluster efficiency* is defined by $E = S/n = \frac{1}{\alpha n + 1 - \alpha}$; (efficiency means that one more cluster can reduce how much time to process)
- Large sequential bottleneck would lead to many idle servers in the cluster;

Theorem 3.1.1.2: Gustafson's Law

By fixing the parallel execution time at level W:

$$S' = W'/W = [\alpha W + (1-\alpha)nW]/W = \alpha + (1-\alpha)n \tag{3.1.1.2}$$

- The efficiency is obtained by: $E' = S'/n = \alpha/n + (1-\alpha)$, this means one more cluster can increase how much workload;
- For a fixed workload, use Amdahl's Law; for a scaled problem, apply Gustafson's Law.

Cloud Availability

Theorem 3.1.1.3: Cluster Availability

$$\text{Clusater Availability} = MTTF/(MTTF + MTTR) \tag{3.1.1.3}$$

- $MTTF$: mean time to failure;
- $MTTR$: mean time to repair.

Example 3.1.1.2

There is a double-redundancy cluster, the MTTF is 200 units while the MTTR is 5 units, calculate the availability of this system.

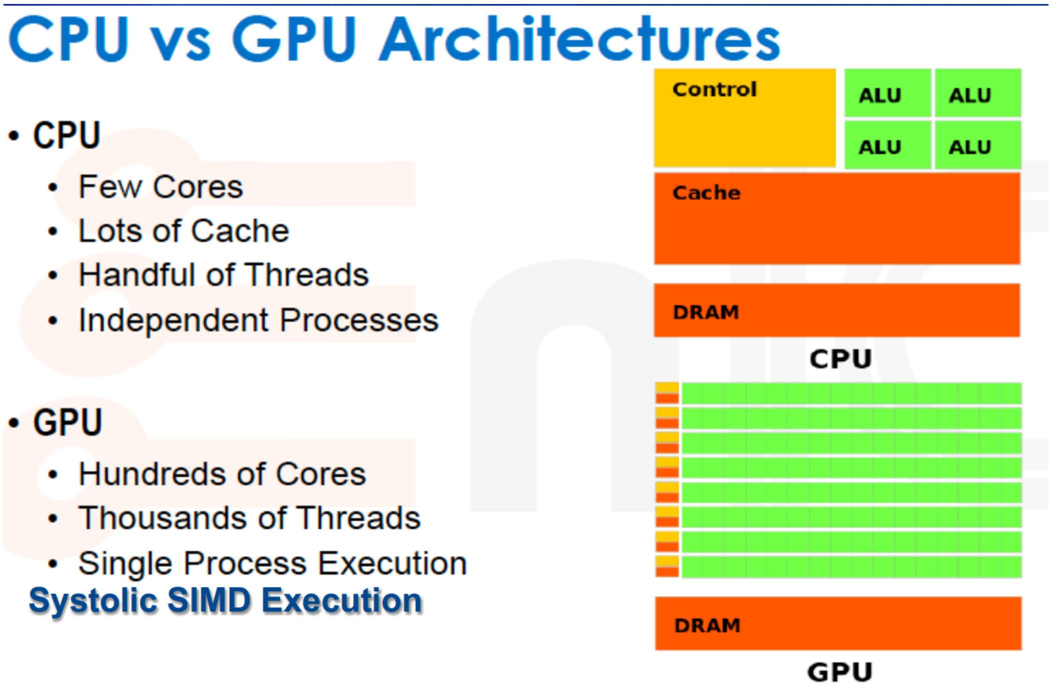
Solution:

The availability of each server is $200/(200 + 5) = 97.5\%$*, the failure rate of the whole system is* $1 - (1 - 97.5\%)^2 = 0.625\%$

Consider the use of a cluster of n homogeneous servers in a system, the system is available when more than k machine is running, then the system availability can be expressed by:

$$\begin{aligned} \mathrm{A} &= \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\ &= \binom{n}{k} p^k (1-p)^{n-k} + \binom{n}{k+1} p^{k+1} (1-p)^{n-k-1} \\ &+ \cdots + \binom{n}{n-1} p^{n-1} (1-p)^1 + \binom{n}{n} p^n (1-p)^0 , \end{aligned} \tag{3.1.1.4}$$

CPU and GPU



CPU vs GPU Architectures

- CPU
  - Few Cores
  - Lots of Cache
  - Handful of Threads
  - Independent Processes
- GPU
  - Hundreds of Cores
  - Thousands of Threads
  - Single Process Execution

Systolic SIMD Execution

Figure 3.5: CPU and GPU Architecture

- CPU has high flexibility for different applications. *Von Neumann bottleneck*: memory access is slow compared to calculation.
- GPU has high throughput, it works well on applications with *massive parallelism*.

3.1.2 Virtual Machines

A VM is essentially built as a software package that can be loaded into a host computer to execute certain user applications. Once the jobs are done, the VM package can be removed from the host computer. The host acts like a “hotel” to accommodate different “guests” at different timeframes.

VMM (virtual machine monitor) can be used to build virtual machine for the guest OS, it can have the following operations:

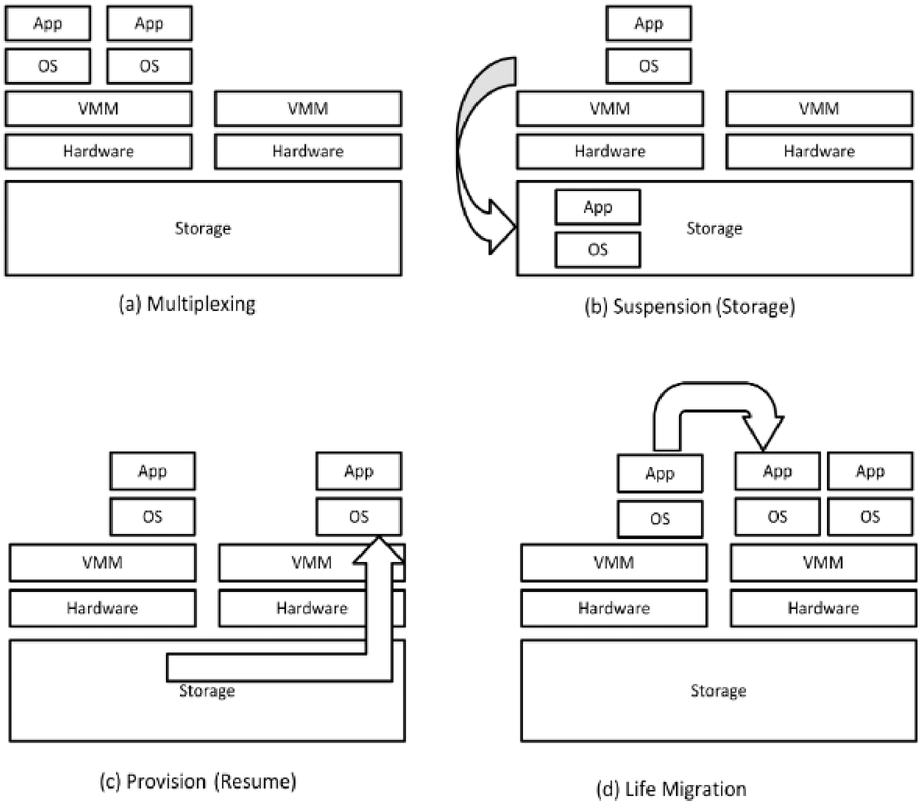


Figure 3.6: VMM Operations

There are five levels of virtualization, among which only 2 are valuable:

- **Hypervisor**: virtualization on top of bare-metal hardware, an example is XEN.
- **Container**: virtualization on operating system level, isolated containers of user app with isolated resources.

Remark 3.1.1

Difference between hypervisor and container

1. *Virtualization Level*: hardware-level virtualization vs. operating system-level virtualization;
2. *Resource Isolation*: strong isolation between VMs vs. lighter resource isolation;
3. *Performance, Startup time, Management*: high vs. low;

Unikernel: combines the advantages of hypervisor and container. Don't rely on a host OS, don't need guest OS for every VM. High performance and start quickly.

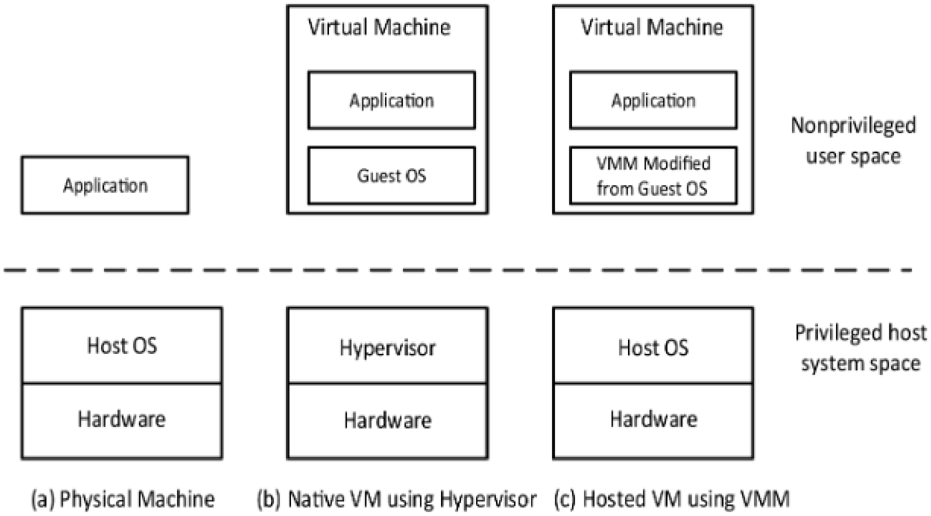


Figure 3.7: virtual machine architecture

3.2 Machine Learning

3.3 Deep Learning

3.4 Causal Inference with Machine Learning

3.4.1 Foundations of Causal Inference

[Peters et al., 2017] provide a overview of causal inference from a foundamental science perspective. [Spirtes, 2010] summarize the development of causal infernce and machine learning interfacce from a computer science perspective.
There are two main causal framework: **Structural Causal Model** and **Potential Outcome Framework**.

Structural Causal Model

SCM (Structural Causal Model) can be devided into two subgroups: *causal graph* and *structural euqation.*

Definition 3.4.1.1

3.5 Reinforcement Learning

This section is mainly the notes of [Thrun and Littman, 2000] and [Agarwal et al., 2019].

3.5.1 Introduction and MDP

Keywords 3.2

- Exploration, Exploitation, Reward, Policy, Value Function;
-

The four elements of reinforcement learning:

1. *Policy*: agent's way to interact with the environment;
2. *Reward Signal*: on each time step, the environment sends to the agent a single number;
3. *Value Function*: specify what is good in the long run, which is the discount value of the rewards;
4. *Model*: mimics behaviors of the environment;

Remark 3.5.1

"Deadly Trials"

1. The balance between **Exploration** and **Exploitation**;
2. Reinforcement learning is difficult to generalize;
3. Delayed consequences may cause RL algorithm to perform poorly.

MDP is a mathematical framework to model *discrete-time* sequential decision process, denoted by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho_0, \gamma)$:

- $\mathcal{S}$: the state space, which is the states for the **entire** environment(MOBA games may can not directly be modeled by MDP);
- $\mathcal{A}$: the action space. $\mathcal{A}$ can depend on $s \in \mathcal{S}$;
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$:the environment transition probability function;
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$: the reward function;
- $\rho_0 \in \Delta(\mathcal{S})$: the initial state distribution;
- $\gamma \in [0, 1]$ is the discount factor.

Remark 3.5.2

- The $\Delta(\dot{})$ may not be deterministic, but some random distribution;
- Among the above tuple, $\mathcal{S}, \mathcal{T}, \mathcal{R}, \rho_0$ can not be modified by the agent, to train a good policy, $\mathcal{A}, \gamma$ is the key;
- RL is more like infants rather than adults;
- The reward function is the way of communicating with the agent *what* to do, not *how* to do;
- The *trajectory* of the MDP sequence: $S_0, A_0, R_1, S_1, A_1, \cdots$.

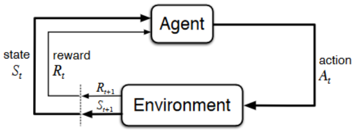


Figure 3.8: The agent–environment interaction in a Markov decision process

Theorem 3.5.1.1: Dynamics of MDP

$$p(s', r|s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r|s, a) = 1, \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s) \tag{3.5.1.1}$$

Corollary 3.5.1.1: Some formula derived from the dynamics theorem

$$p(s'|s, a) \doteq \Pr\{S_t = s' \mid S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathbb{R}} p(s', r|s, a) \tag{3.5.1.2}$$

$$r(s, a) \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathbb{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a) \tag{3.5.1.3}$$

$$r(s, a, s') \doteq \mathbb{E}[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'] = \sum_{r \in \mathbb{R}} r \frac{p(s', r|s, a)}{p(s'|s, a)} \tag{3.5.1.4}$$

Definition 3.5.1.1: Some useful function:

The act value function given policy π:

$$Q^{\pi}(s,a) = \mathbb{E}_{s_t,a_t,r_t,t\geq 0}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a\Big] \tag{3.5.1.5}$$

The expected return at state s given policy π:

$$V^{\pi}(s) = \mathbb{E}_{a\sim\pi(s)}\left[Q^{\pi}(s,a)\right] \tag{3.5.1.6}$$

The **advantage function**:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \tag{3.5.1.7}$$

The temporal-difference error:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{3.5.1.8}$$

If we denote $G_t = R_{t+1} +_{t+1}$, then we have:

Theorem 3.5.1.2: Bellman Equation

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s',r|s,a)\Big[r + \gamma \mathbb{E}_\pi[G_{t+1}|S_{t+1} = s']\Big] \\ &= \sum \pi(a|s) \sum p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big], \quad \text{for all } s \in \mathcal{S}, \end{aligned} \tag{3.5.1.9}$$

A policy π is better π' if and only if $v_\pi(s) \leq v_{\pi'}(S)$for all $s \in \mathcal{S}$. There is always at list one *optimal policy* denoted by π_* (doesn't hold for partially observed MDP). They share the same state-value function, called the *optimal state-value function*: $v_*(s) \doteq \max_\pi v_\pi(s)$. Optimal policy also shares the same *optimal action-value function*: $q_*(s,a) \doteq \max_\pi q_\pi(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a]$. Then we can get the **Bellman optimality function** in an action-value function sense:

Theorem 3.5.1.3: Bellman optimality function

$$\begin{aligned} q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1},a')\Big|S_t = s, A_t = a\Big] \\ &= \sum_{s'.r} p(s',r|s,a)\Big[r + \gamma \max_{a'} q_*(s',a')\Big]. \end{aligned} \tag{3.5.1.10}$$

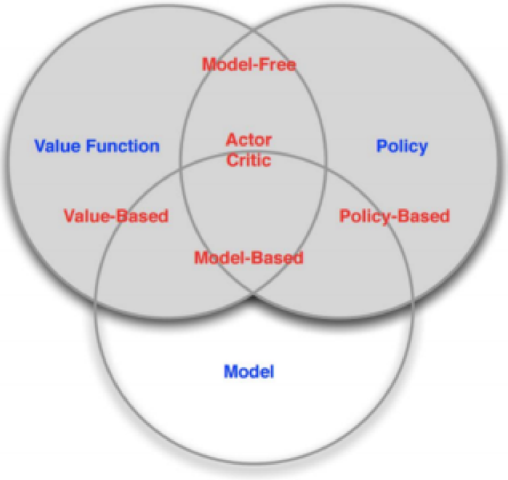


Figure 3.9: Classification of different reinforcement learning agents

Definition 3.5.1.2: The optimal policy π^*

A policy π^* is an optimal policy if for every policy π and every $s \in \mathcal{S}$, we have:

$$V^{\pi^*}(s) \geq V^{\pi}(s) \tag{3.5.1.11}$$

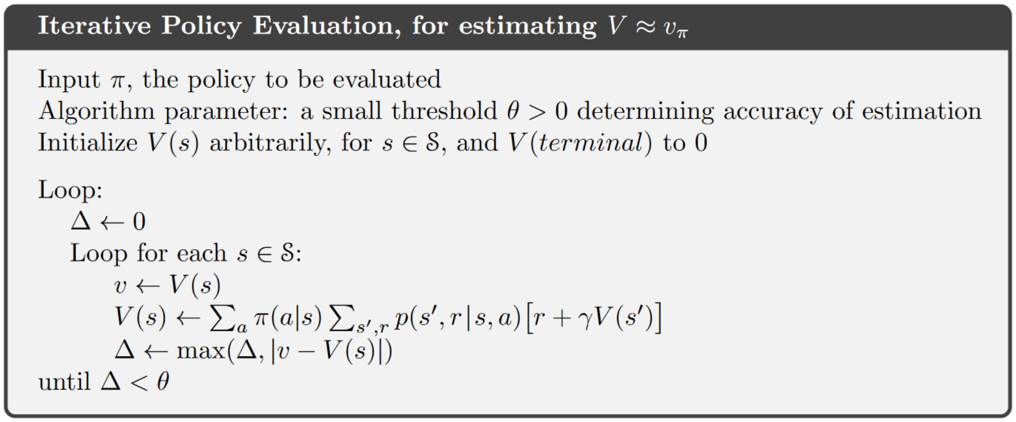
Remark 3.5.3

- If we have the full information of the game (MDP framework), then the optimal policy is always deterministic;
- The optimal policy is always stochastic when there are **minimax** structure (e.g. protection information);
- Whether the optimal policy is stochastic or deterministic has nothing to do with the stochasticity of the game.

3.5.2 Dynamic Programming in RL

Policy Evaluation (Prediction)

$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_k(s')\Big] \end{aligned} \tag{3.5.2.1}$$



Iterative Policy Evaluation, for estimating $V \approx v_\pi$

```
Input $\pi$, the policy to be evaluated
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$ arbitrarily, for $s \in \mathcal{S}$, and $V(terminal)$ to 0

Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
       $v \leftarrow V(s)$
       $V(s) \leftarrow \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$
       $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$
```

Figure 3.10: Iterative Policy Evaluation

Theorem 3.5.2.1: Policy Improvement Theorem

For all $s \in \mathcal{S}$, if:

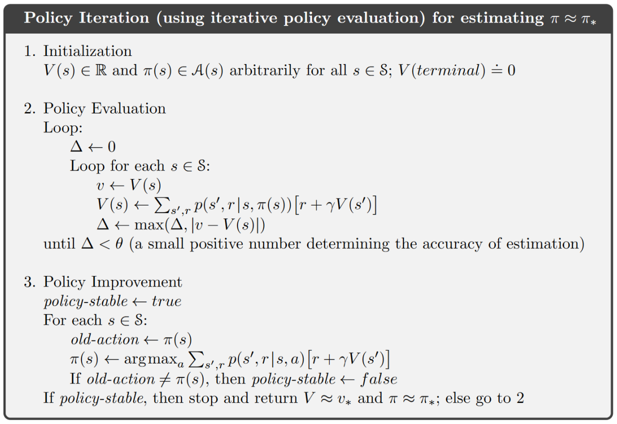
$$q_\pi(s, \pi'(s)) \geq v_\pi(s) \tag{3.5.2.2}$$

Then the policy π' must be better than policy π.

Policy Improvement

$$\begin{aligned} \pi'(s) \doteq &\ \arg\max_a q_\pi(s,a) \\ = &\ \underset{a}{\mathrm{argmax}}\mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \\ = &\ \arg\max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_\pi(s')\Big], \end{aligned} \tag{3.5.2.3}$$

Policy Iteration and Value Iteration



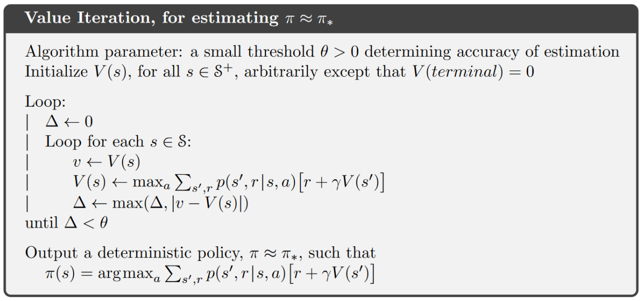
Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

```
1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$; $V(terminal) \doteq 0$

2. Policy Evaluation
   Loop:
       $\Delta \leftarrow 0$
       Loop for each $s \in \mathcal{S}$:
           $v \leftarrow V(s)$
           $V(s) \leftarrow \sum_{s',r} p(s',r|s,\pi(s))\big[r + \gamma V(s')\big]$
           $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ true
   For each $s \in \mathcal{S}$:
       old-action $\leftarrow \pi(s)$
       $\pi(s) \leftarrow \mathrm{argmax}_a \sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]$
       If old-action $\neq \pi(s)$, then policy-stable $\leftarrow$ false
   If policy-stable, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2
```

Figure 3.11: Policy Iteration



Value Iteration, for estimating $\pi \approx \pi_*$

```
Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
|   $\Delta \leftarrow 0$
|   Loop for each $s \in \mathcal{S}$:
|       $v \leftarrow V(s)$
|       $V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]$
|       $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
    $\pi(s) = \mathrm{argmax}_a \sum_{s',r} p(s',r|s,a)\big[r + \gamma V(s')\big]$
```

Figure 3.12: Value Iteration

Differences between VI and PI:

Remark 3.5.4

- In each sweep, VI only updates one step evaluation and one step improvement, PI updates multiple step evaluation and one step improvement;

- PI takes fewer round, but takes more time within each round .

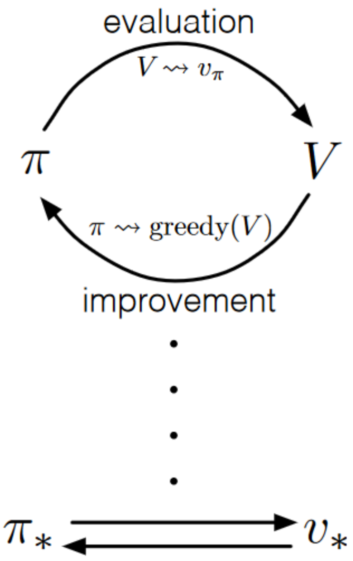

evaluation
$V \rightsquigarrow v_\pi$
π
V
$\pi \rightsquigarrow \text{greedy}(V)$
improvement
π_*
v_*

Figure 3.13: Generalized Policy Iteration

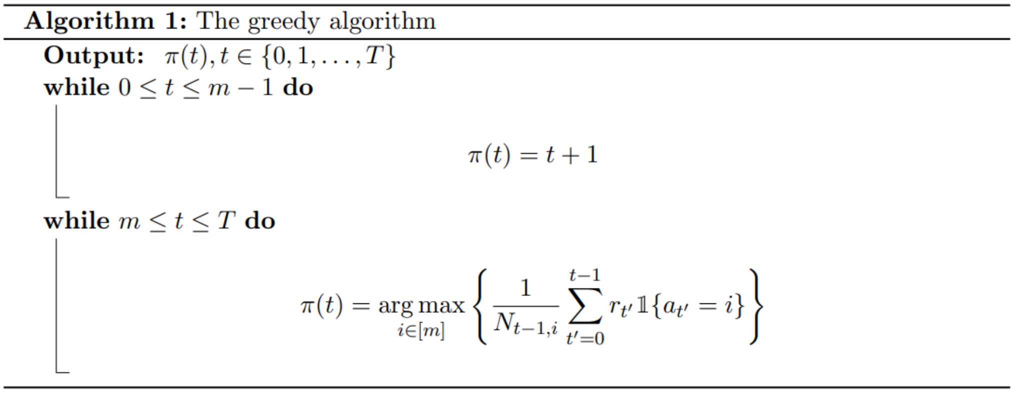
3.5.3 Bandit Algorithms

The regret for bandit games is defined as :

$$\overline{R}_t = \sum_{i=1}^{m} \mathbb{E}[N_{t,i}]\Delta_i \tag{3.5.3.1}$$

Where $N_{t,i} = \sum_{t'=0}^{t} \mathbb{1}\{a_{t'} = i\}$ and $\Delta_i = \mu^* - \mu_i$.

Greedy Algorithms

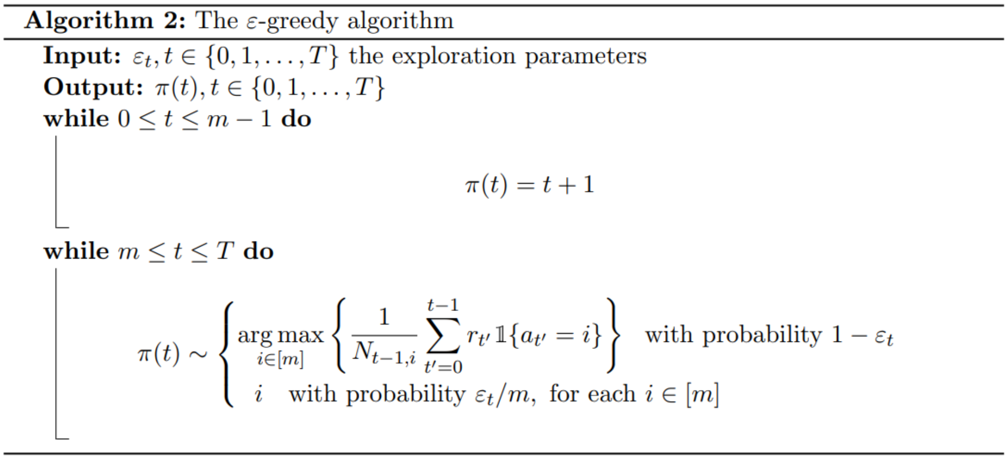


```
Algorithm 1: The greedy algorithm
  Output: $\pi(t), t \in \{0, 1, \dots, T\}$
  while $0 \leq t \leq m - 1$ do
      $\pi(t) = t + 1$
  while $m \leq t \leq T$ do
      $\pi(t) = \arg\max_{i \in [m]} \left\{ \frac{1}{N_{t-1,i}} \sum_{t'=0}^{t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} \right\}$
```

Figure 3.14: The greedy algorithm

This algorithm achieves a regret at most $O(T)$.

The ϵ-greedy algorithm

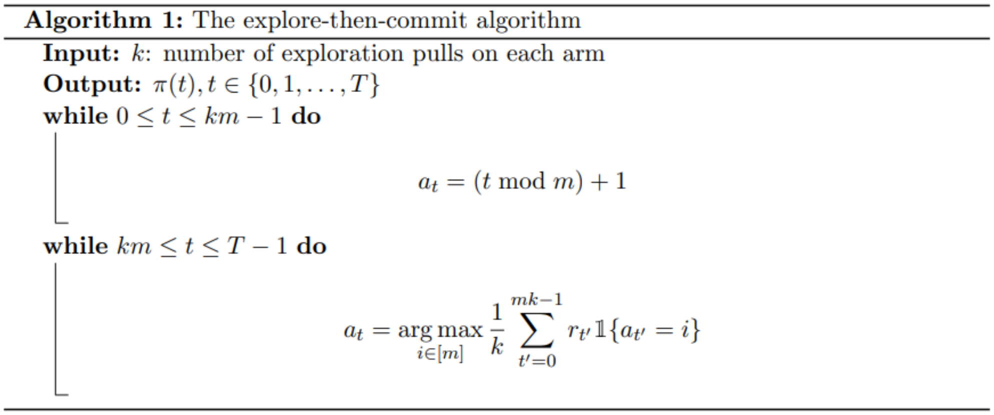


```
Algorithm 2: The $\varepsilon$-greedy algorithm
  Input: $\varepsilon_t, t \in \{0, 1, \dots, T\}$ the exploration parameters
  Output: $\pi(t), t \in \{0, 1, \dots, T\}$
  while $0 \le t \le m - 1$ do
      $\pi(t) = t + 1$
  while $m \le t \le T$ do
      $\pi(t) \sim \begin{cases} \arg\max_{i \in [m]} \left\{ \frac{1}{N_{t-1,i}} \sum_{t'=0}^{t-1} r_{t'} \mathbb{1}\{a_{t'} = i\} \right\} & \text{with probability } 1 - \varepsilon_t \\ i \quad \text{with probability } \varepsilon_t/m, \text{ for each } i \in [m] \end{cases}$
```

Figure 3.15: The epsilon greedy algorithm

The lower bound of ϵ greedy: $\overline{R}_t \geq \frac{1}{m}(\Delta_2 + \cdots + \Delta_m)\varepsilon(T - m)$, where $\epsilon \leq \epsilon_t$ for all t;

The explore-then-commit algorithm (ETC)



```
Algorithm 1: The explore-then-commit algorithm
  Input: $k$: number of exploration pulls on each arm
  Output: $\pi(t), t \in \{0, 1, \dots, T\}$
  while $0 \le t \le km - 1$ do
      $a_t = (t \bmod m) + 1$
  while $km \le t \le T - 1$ do
      $a_t = \arg\max_{i \in [m]} \frac{1}{k} \sum_{t'=0}^{mk-1} r_{t'} \mathbb{1}\{a_{t'} = i\}$
```

Figure 3.16: The ETC algorithm

This algorithm has an upper bound of $O(\Delta^2 logT)$, or $O(T^{\frac{2}{3}})$.

Chapter 4

Operations Management

4.1 Empirical Operations Management

[Roth, 2007] describes the evolution of empirical OM from 1980 to 2007, the author selects 12 profounding papers in this domain. [Brusco et al., 2017] reviewed the clustering methods applied in 6 OM journals.
[Choi et al., 2016], multi-methodological OM is advocated, which includes the empirical methodology.

> **Definition 4.1.0.1:** Multi-methodological OM
> an approach for OM research in which at least two distinct OM research methods are employed nontrivially to meet the research goals.

[Roth and Singhal, 2022] classified 75 papers as empirical among the top 200 cited papers in *POM*, these papers are mainly from 3 topical areas:

1. Responsibility Operations: covers environmental management, sustainability, humanitarian efforts;
2. Supply Chain Management: bullwhip effect, risk management, supply chain finance;
3. Manufacturing Strategy and Quality Management.

Primary data (surveys, experiments, interviews) are used in these studies, followed by secondary data (public database, firm's data). Roth's suggestions:

1. For some topic which is very intuitive and not surprising, focus on the **size** of effect rather than sign;
2. Avoid the confirmation bias and focus on consistency;

3. Focus on endogeneity and causality, using common sense simultaneously.

[Kumar and Tang, 2022] reviewed different domains of OM publications on *POM*, within which a section about empirical OM is covered.
[Mithas et al., 2022] reviewed 411 empirical papers form 2016-2021 on *POM*, with a causal inference and counterfactual perspective.

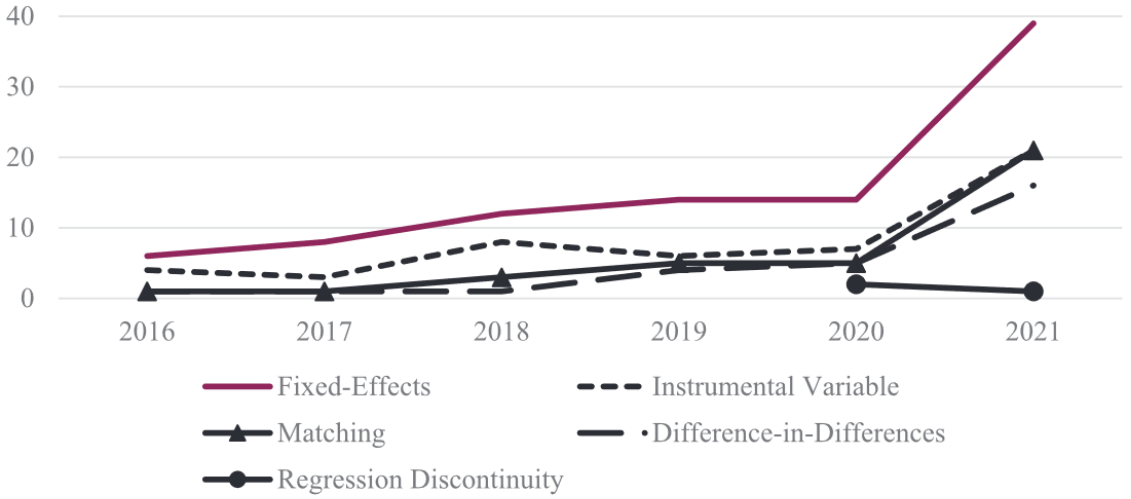


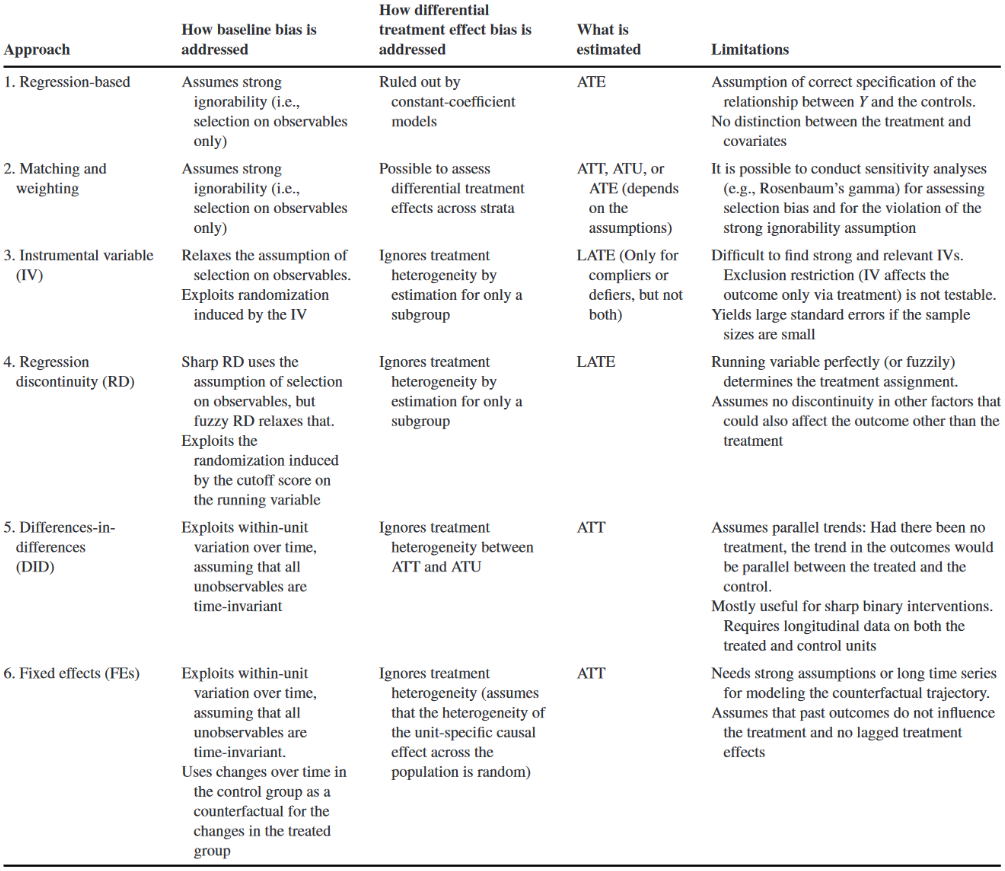
Figure 4.1: identification strategies from 2016 to 2021

Remark 4.1.1

Two challenges in assessing causality:

$$\begin{aligned} T &= ATE + [E(Y(0)|Z=1) - E(Y(0)|Z=0)] \\ &\quad + (1-\pi) * [(ATT - ATU)]\,. \end{aligned} \tag{4.1.0.1}$$

- $[E(Y(0)|Z=1) - E(Y(0)|Z=0)]$ is the **baseline bias**, coming from *OVB* or *simultaneity*;
- $[(ATT - ATU)]$ is the **differential treatment bias**.



Approach	How baseline bias is addressed	How differential treatment effect bias is addressed	What is estimated	Limitations
1. Regression-based	Assumes strong ignorability (i.e., selection on observables only)	Ruled out by constant-coefficient models	ATE	Assumption of correct specification of the relationship between *Y* and the controls. No distinction between the treatment and covariates
2. Matching and weighting	Assumes strong ignorability (i.e., selection on observables only)	Possible to assess differential treatment effects across strata	ATT, ATU, or ATE (depends on the assumptions)	It is possible to conduct sensitivity analyses (e.g., Rosenbaum's gamma) for assessing selection bias and for the violation of the strong ignorability assumption
3. Instrumental variable (IV)	Relaxes the assumption of selection on observables. Exploits randomization induced by the IV	Ignores treatment heterogeneity by estimation for only a subgroup	LATE (Only for compliers or defiers, but not both)	Difficult to find strong and relevant IVs. Exclusion restriction (IV affects the outcome only via treatment) is not testable. Yields large standard errors if the sample sizes are small
4. Regression discontinuity (RD)	Sharp RD uses the assumption of selection on observables, but fuzzy RD relaxes that. Exploits the randomization induced by the cutoff score on the running variable	Ignores treatment heterogeneity by estimation for only a subgroup	LATE	Running variable perfectly (or fuzzily) determines the treatment assignment. Assumes no discontinuity in other factors that could also affect the outcome other than the treatment
5. Differences-in-differences (DID)	Exploits within-unit variation over time, assuming that all unobservables are time-invariant	Ignores treatment heterogeneity between ATT and ATU	ATT	Assumes parallel trends: Had there been no treatment, the trend in the outcomes would be parallel between the treated and the control. Mostly useful for sharp binary interventions. Requires longitudinal data on both the treated and control units
6. Fixed effects (FEs)	Exploits within-unit variation over time, assuming that all unobservables are time-invariant. Uses changes over time in the control group as a counterfactual for the changes in the treated group	Ignores treatment heterogeneity (assumes that the heterogeneity of the unit-specific causal effect across the population is random)	ATT	Needs strong assumptions or long time series for modeling the counterfactual trajectory. Assumes that past outcomes do not influence the treatment and no lagged treatment effects

Figure 4.2: How identification techniques works

[Fisher and Raman, 2022] especially investigate the empirical research in retail operations from traditional ones like forecasting and inventory planning, to new technologies, like radio frequency identification (RFID) and e-commerce.

4.1.1 12 Papers of [Roth, 2007]

[Fisher, 2007] is a conceptual paper, in which a matrix is proposed to navigate conducting research:

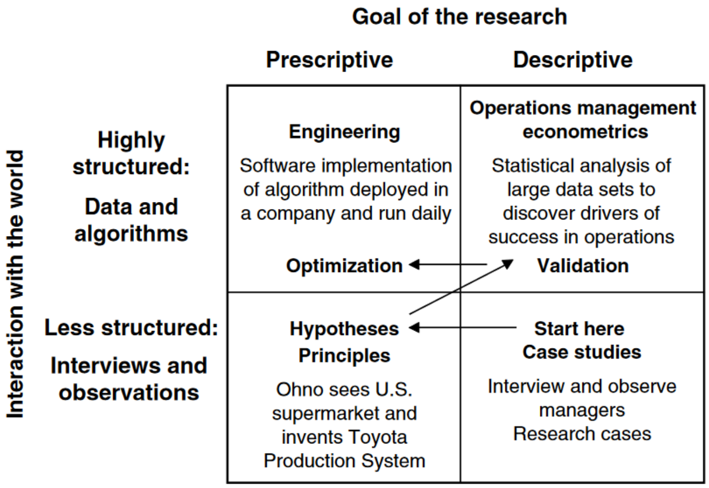

Goal of the research
Prescriptive
Descriptive
Interaction with the world
Highly structured:
Data and algorithms
Less structured:
Interviews and observations
Engineering
Software implementation of algorithm deployed in a company and run daily
Optimization
Operations management econometrics
Statistical analysis of large data sets to discover drivers of success in operations
Validation
Hypotheses
Principles
Ohno sees U.S. supermarket and invents Toyota Production System
Start here
Case studies
Interview and observe managers
Research cases

Figure 4.3: Navigating Matrix Cells

Fisher suggests that a good empirical OM research should include the following:

1. Identifying and verifying important phenomena;

2. Identifying and characterizing important questions on which we can do useful research;

3. Validating models and assumptions that have been made;

4. Establishing the relevance of our research by demonstrating how the research outputs apply to practice.

[Gans et al., 2007] investigates the posit of some bandit consumer choice models:

4.2 Revenue Management

Revenue management is a data-driven system to price perishable assets tactically at the micro-market level to maximize expected revenue or profit. Some critical reviews before 2009 can be found in the book [Gallego and Topaloglu, 2019]. There are some extra summary papers like [Strauss et al., 2018], [Klein et al., 2020].

4.2.1 Traditional RM

Keywords 4.1

- Protection Level, Booking Limit, Littlewood's Rule

Assumption 4.2.1.1: What does **"traditional"** means in RM?

1. The traditional RM system doesn't consider the choice model, in particular, it assumes the demands are independent random variables;
2. Further assumption: consumer will leave without purchasing if preferred fare class is unavailable (holds when gaps in fares are large enough);
3. The capacity is fixed, the capacity's marginal profit is zero(can be relaxed);
4. All booked consumers would arrive (another circumstance see 4.2.2).

Assumption 4.2.1.2: Single Resource RM

1. The units of capacity is c, pricing at multiple different level $p_n < \cdots < p_1$;
2. Low-before-high fare class arrival order: D_2 before D_1 for example (this is the worst case for revenue);
3. **Protection level** for customer j: leave $y \in \{0, 1, \cdots, c\}$ for $D_{j-1},, D_1$; $c - y$ is the **booking limit** which serves D_j;

😎 So the problem is to solve the optimal protection level given the current consumer level j.

Let $V_j(x)$ be the optimal revenue given D_j coming in, x units remained. $V_0(x) = 0$ by design. Let y be the protection level for $D_{j-1}, \cdots, D_1$: sales at $p_j = \min\{x - y, D_j\}$. The remaining capacity for $D_{j-1}, \cdots, D_1$ is $x - \min\{x - y, D_j\} = max\{y, x - D_j\}$. Now let $W_j(y, x)$ be the optimal solution. We have:

$$W_j(y,x) = p_j \mathbb{E}\{\min\{x-y, D_j\}\} + \mathbb{E}\{V_{j-1}(\max\{y, x-D_j\})\} \tag{4.2.1.1}$$

$$V_j(x) = \max_{y\in\{0,\ldots,x\}} W_j(y,x) = \max_{y\in\{0,\ldots,x\}} \left\{p_j \mathbb{E}\{\min\{x-y, D_j\}\} + \mathbb{E}\{V_{j-1}(\max\{y, x-D_j\})\}\right\} \tag{4.2.1.2}$$

Proposition 4.2.1.1: Structure of the Optimal Policy

$$y^*_{j-1} = \max\{y \in \mathbb{N}_+ : \Delta V_{j-1}(y) > p_j\}. \tag{4.2.1.3}$$

The maximizer of $W_j(y,x)$ is given by $y^*_j, , y^*_1$

Remark 4.2.1

The optimal solution for $y_j$is independent of the distribution of D_j;

Corollary 4.2.1.1: When $j = 2$:

Theorem 4.2.1.1: Littlewood's rule

$$y^*_1 = \max\{y \in \mathbb{N}_+ : \mathbb{P}\{D_1 \geq y\} > r\} \tag{4.2.1.4}$$

;

Remark 4.2.2

The Littlewoood's Rule:

1. The solution depends on the **fare ratio**: $r := p_2/p_1$;
2. When the distribution of D_2 is continuous: $F_1(y) = \mathbb{P}\{D_1 \leq y\}$. The optimal protection level is $y^*_1 = F_1^{-1}(1-r) = \mu_1 + \sigma_1 \notin^{-1} (1-r)$:
   (a) if $r > \frac{1}{2}$, $y^*_1 < \mu_1$ and y^*_1 decreases with σ_1;
   (b) if $r < \frac{1}{2}$, $y^*_1 < \mu_1$ and y^*_1 increases with σ_1;
   (c) if $r = \frac{1}{2}$, $y^*_1 = \mu_1$;
3. Using Littlewood's rule would result in some D_1 served by competitors (high spill rates). Solution: add penalty to save more seats for the high fare consumers:

$$y^*_1 = \max\left\{y \in \mathbb{N}_+ : \mathbb{R}\{D_1 \geq y\} > \frac{p_2}{p_1 + \rho}\right\} \tag{4.2.1.5}$$

4.2.2 Overbooking

4.2.3 Traditional Consumer Choice Model

4.2.4 Current Consumer Choice Model

4.3 Dynamic Pricing

4.3.1 Basic Pricing Theory

This subsection summarizes the basic pricing theory for **multi-product momnpoly** firms.

Perspective from the Firm

The firm's profit function is given by:

$$R(p,z) := (p-z)'d(p) = \sum_{i=1}^{n}(p_i - z_i)d_i(p_1,\dots,p_n), \tag{4.3.1.1}$$

Where $z = (z_1, \cdots, z_n)$ is the variable cost vector, $p = (p_1, \cdots, p_n)$ is the price vector, $d(p)$ is the demands. Currently it's popular to set $p_i \in [0, \infty]$, where by setting $p_i = \infty$ is equivalent to not offering product i.

Consider the revenue as a function only of z, given by:

$$\mathcal{R}(z) := \max_{p \in X} R(p,z), \tag{4.3.1.2}$$

Where X is the set of allowable prices.

Theorem 4.3.1.1: The Decreasing Covex Property

$\mathcal{R}(z)$ is **decreasing convex** in z.

Proof 4.3.1

The decreasing property is obvious since $R(p,z)$ is decreasing in z, here is the proof for convexity.
For any $z, \tilde{z}$:

$$\begin{aligned}
\mathcal{R}(\alpha z + (1-\alpha)\tilde{z}) &= \max_{p \in X} R(p, \alpha z + (1-\alpha)\tilde{z}) \\
&= \max_{p \in X} R(\alpha p + (1-\alpha)p, \alpha z + (1-\alpha)\tilde{z})) \\
&= \max_{p \in X} [\alpha(p-z)' + (1-\alpha)(p-\tilde{z})']\, d(p) \\
&= \max_{p \in X} [\alpha R(p,z) + (1-\alpha) R(p,\tilde{z})] \\
&\leq \alpha \max_{p \in X} R(p,z) + (1-\alpha) \max_{p \in X} R(p,\tilde{z}) \\
&= \alpha \mathcal{R}(z) + (1-\alpha)\mathcal{R}(\tilde{z}).
\end{aligned} \tag{4.3.1.3}$$

☺ This proof is very similar to proving the convexity of the hyperplane.

Q.E.D.

By transforming the revenue function as a function only to z is useful. Because by **Jensen's Inequality**, $\mathbb{E}[\mathcal{R}(\mathbb{Z})] \geq \mathcal{R}(\mathbb{E}[\mathbb{Z}])$. Their differences can be interpreted as the differnce between a dynamic pricing policy $p(Z)$ that responds to changes in Z and a static policy $\mathcal{R}(\mathbb{E}[\mathbb{Z}])$, the larger the variance of Z, the larger the gap between them.

Perspective form the Consumers

To answer whether consumers are better off with pricing policy $p(Z)$ or $p(\mathbb{E}(Z))$, we can frame this using the **utility theory** ([Chen and Gallego, 2019]). Assume that consumers purchase a non-negative vector $q = (q_1, q_2, \cdots, q_n)$ of products, their *net utility* is given by:

$$S(q,p) := U(q) - q'p, \tag{4.3.1.4}$$

where $U(q)$ us an increasing concave function. Then the *optimal surplus* can be conputed by:

$$\mathcal{S}(p) := \max_{q \geq 0} S(q,p) \tag{4.3.1.5}$$

where the solution $q^\star = d(p) = \nabla^{-1} U(p)$. under the first-order condition. Similar to the analysis from the firm's perspective, we can get $\mathbb{E}[\mathcal{S}(P)] \geq \mathcal{S}(\mathbb{E}[P])$. And at last we can achieve:

$$\mathbb{E}[\mathcal{S}(p(\mathbb{Z}))] \geq \mathcal{S}(\mathbb{E}[p(\mathbb{Z})]) \geq \mathcal{S}(p(\mathbb{E}[Z])). \tag{4.3.1.6}$$

4.3.2 Dynamic Pricing and Reinforcement Learning

[Misra et al., 2019] is probably the first paper that incoperates reinforcement learning (MAB) algorithm into dynamic pricing.

Assumption 4.3.2.1: [Misra et al., 2019]

For consumers:

1. She has stable preferences v_i doesn't change over time;
2. Her choice satisfies weak axim of revealed preference: if $v_i > p$, then she would buy it;
3. Heterogeneity among consumers can be separated as observable and unobservable heterogeneity: $v_i = f(\mathbb{Z}_i) + \nu_i$;
4. There are bound for $\nu_i : \nu_i \in [-\delta, \delta]$

Model Setting:

1. There are K prices that a firm can choose: $p \in \{p_1, \dots, p_K\}$;
2. sample mean $\bar{\pi}_{kt} = 1/n_{kt} \sum_{\tau=1}^{n_{kt}} \pi_{k\tau}$, a policy is defined as $p_t = \Psi(\{p_\tau, \pi_\tau | \tau = 1, \dots, t-1\})$;
3. The regret is given by: $\text{Regret}(\Psi, \{\pi(\mathfrak{p}_{\mathfrak{t}})\}, t) = \pi^* t - \sum_{k=1}^{K} \pi(p_k) \mathbb{E}[n_{kt}]$

[Calvano et al., 2020]

4.4 Platform Operations Management

[Rietveld and Schilling, 2021] reviews literature in platform competitions.

[Wang et al., 2023] uses game theory framework to analyze the cross-licensing policy initiated by Qualcomm, which provides some insights for the up-stream manufacturing company:

1. The supplier shouldn't adopt the cross-licensing policy if the inferior manufacturer's cost of innovation is high;
2. Cross-licensing may achieve a higher level of total innovation if the superior manufacturer's cost of innovation is low;
3. The superior manufacturer can benefit from cross-licensing, if innovation is costly but the manufacturers' costs of innovation are similar.

4.4.1 Hotel Platform

An overview of Airbnb

[Dolnicar, 2021] provides a comprehensive illustration of all aspects of Airbnb, including the business model, competitive landscape, and the regulations of Airbnb.

[Guttentag, 2019] reviewed some tier c papers about the progress on Airbnb, their main focus is on the loyalty and motivation of guests and hosts, Airbnb's regulation and culture, as well as Airbnb's impact on the tourism sector.

Airbnb makes money by renting out property that it doesn't own, the hosts can be an individual or a company (But Airbnb doesn't own property, it's just an intermediary). ([Folger, 2023]) In 2017, Airbnb invested in Niido, a hotel-like apartment program managed by Airbnb. In 2020, Airbnb ended its partnership with Niido apartments. During the whole process, only 2 apartments were in service.

[Zhang et al., 2022] studies how Airbnb property demand changed after the acquisition of *verified* images. Variables description:

- Treatment variable: 212 properties had verified photos by the end of April 2017, the remaining 7,211 did not;
- Property demand: purchased date, the number of days in a month in which the property was open , blocked, and booked. $\frac{booked}{open} \times 100$
- Property Price: the price is endogenous because of the random demand shocks, characteristics of competing properties were used as IVs (The logic is that the characteristics of competing products are unlikely to be correlated with unobserved shocks in the demand for the focal property. However, the proximity of the characteristics of a property and its

competitors influences the competition and as a result, the property markup and price). Cost-related variables are collected including residential utility fees.

- Property Photos: CNN architecture was used to predict the quality of a photo (dummy variable).

Other Identification Techniques:

- DiD model: $DEMAND_{itcym} = INTERCEPT + \alpha TREATIND_{it} + \lambda CONTROLS_{it} + PROPERTY_i + SEASONALITY_{cym} + \varepsilon_{it}$;
- PSW method: calculate the prosensity score $\widehat{ps_i(X_i)}$ and use IPTW ($\omega_i(T, X_i) = \frac{T}{\widehat{ps_i(X_i)}} + \frac{1-T}{1-\widehat{ps_i(X_i)}}$) to weight the sample.
- Relative time model was used to test the common trends assumption, Rosenbaum bounds test was used to test the validation of PSW methods.

Besides the work above, the authors investigate what makes a picture good. They listed 3 components (composition, color, figure-ground relationship) including 12 attributes and used the following regression to give some human-interpretable suggestions:

$$\begin{aligned}
DEMAND_{itcym} =& INTERCEPT + \alpha TREATIND_{it} \\
&+ \mu IMAGE\_COUNT_{it} \\
&+ \rho_1 \mathrm{R\_\_IO}_{it} \\
&+ \rho_2 BEDROOM\_PHOTO\_RATIO_{it} \\
&+ \rho_3 \mathrm{I}N\_\_\mathrm{IO}_{it} \\
&+ \rho_4 LIVING\_PHOTO\_RATIO_{it} \\
&+ \eta IMAGE\_ATTRIBUTES_{it} \\
&+ \lambda CONTROLS_{it} + SEASONALITY_{cym} \quad + PROPERTY_i + \epsilon_{it}
\end{aligned} \tag{4.4.1.1}$$

[Chen et al., 2023] investigates the professional players' effects on the non-professional host. They define host who has more than one properties on the platform simultaneously as professional players, and use a quasi-experiment (OHOH policy) and DiD model to analysis whether competition effects or differentiation effects is dominant.

They predict 2 propositions:

1. If differentiation effects dominates, then OHOH would not affect the supply and price of non-professional hosts;

2. If competition effects dominate, then OHOH would increase the supply and price of non-professional hosts.

$$Y_{it} = \mu_i + \nu_t + \beta\cdot 1(\text{Policy})_{it} + \gamma' \mathbf{X}_{it} + \varepsilon_{it} \tag{4.4.1.2}$$

Their results show that the competition effects dominate the role of professional hosts. [Farronato and Fradkin investigate the peer's entry on consumer's welfare in the accommodation industry using a structural model.

Intuition: Peer hosts are responsive to market conditions, expand supply as hotels fill up, and keep hotel prices down as a result.

4.4.2 Platform Owner's Entry

Keywords 4.2

Complementory Markets, Spillover Effects

[Chen and Tsai, 2023] is the first paper considering the asymmetric information between the platform owner and the third-party sellers, and the effects of the information disadvantage on consumer's welfare. They assemble data from Amazon: 122000 products, each with two sellers offering the same product. *The information asymmetry*: when Amazon's competitors make sales, Amazon adjusts its prices accordingly; conversely, third-party sellers do not react to their competitors' sales.

After observing the empirical evidence of the information asymmetry, they design a theoretical model and itentify the parameters. Using the structural regression, they find that by giving information to third-party sellers. Both the Amazon, third-party sellers and the consumers' welfare would be increased.

[Zhu and Liu, 2018] surveys empirical studies that examine the direct entry of platform owners into complementary product spaces.

[Zhu and Liu, 2018] studies the entry of Amazon platform. Logit regression is adopted to verify the following **Hypothesis**:

- Platform owners are more likely to compete with a complementor when its products are successful;

- Platform owners are less likely to compete with a complementor when its products require significant platform-specific investments to grow.

Identification Techniques:

- The sales ranking is used as proxy variable for the sales of the products;

- To overcome the impact of the referral rates by category-level fixed effects;

- To measure the seller's platform-specific investment, they calculate the seller's average answers;

[He et al., 2020] investigates the effects and the mechanisms of platform owner's entry on third-party's online and offline demands using a B2B shopping platform's data. They use DiD to identify that the platform owner's entry does harm the demands (more in online channels than offline channels).
What's more, they propose three mechanisms to explain the effects:

1. Competition Effects: The owner can appropriate value from the third-party sellers (only significant for online channels);

2. Spillover Effects: increasing the exposure or awareness of the products (not the same as the mobile app market);

3. Disintermediation effect: sellers would use defensive strategy to transact outside the platform (mostly in offline channel) see [Gu and Zhu, 2021] and [Ha et al., 2022].

Finally, DDD and PSM were adopted to identify the heterogeneity of the effects between large and small third-party stores.
[Deng et al., 2023] use data from JD.com and provide an unexpected result, thet [Wen and Zhu, 2019] and [Foerderer et al., 2018] focus on complementors' reactions, especially on innovation strategy for the platform owner entry.
[Shi et al., 2023] investigates the timing of the platform owner's entry on the value creation.

Definition 4.4.2.1: Timing of Owner's Entry

- **Platform Complementors**: actors that offer an application that brings additional value to platform users when used in combination with the platform;

- **Early-entry**: the entry occurs when the ratio between the current and the eventual complementary market's size is low;

- **Late-entry**: entry to a relatively mature complementary market;

- **Value creation**: the activities geared toward increasing the perceived attractiveness of the platform ecosystem among customers and measure it as changes to complement popularity among customers. (proxy variable)

Identification Techniques:

- Use the *the number of reviews* as the proxy for the popularity of complements (dependent variables);

- *functional specificity* measures the heterogeneity of a complement based on the complexity of services offered by the compliment;

- Follows [Zhu and Liu, 2018] to account for platform-specific investments by *interfacce coupling*: whether the complement connects with the platform core;
- To verify the exogeneity of Amazon's entry decision, a logit regression is conducted to test the number of reviews (popularity) does not influence Amazon's decision;
- PSM method is adopted.

$$Reviews_{it} = \alpha + \beta Treated_i \times After_t + \delta Controls_{it} + C_i + T_i + \epsilon_{it} \tag{4.4.2.1}$$

Many papers analyze the platform owner's entry from a game-theory perspective.
[Hagiu et al., 2020] investigate the owner's entry decision in the complementary markets.

Assumption 4.4.2.1: Creating platforms by hosting rivals

- There are two companies M and S: M sells product A and B_M, S can only sell B_S;
- The costs to produce are all set to zero;
- The number of customers is normalized to one, λ_A customers are only interested in A (A-type), $1 - \lambda_A$ are interested in both (B-type);
- $u_A > 0$ for both type of customers, $u_B > 0$ and $u_S = u_B + \Delta$ for B-type;
- There are costs for consumers to go to each store with a cost σ, and $0 < \sigma < \min\{u_A, u_B\}, \Delta < \sigma$.

The authors analyze two conditions: without hosting and hosting. By solving the equilibrium given two conditions, they find the following conclusion:

1. In the without hosting condition, M would dominant the whole market, the profit for S is zero;
2. Even without transfer for S to sell products in the platform M, M can still benefit if $\lambda_A \leq \sigma / u_A$ and $\Delta > \lambda_A (u_A - \sigma)^+ F / (1 - \lambda_A)$.

[Cheng et al., 2022] analyze the owner's entry ob the incumbent sellers in the E-commerce platform.

Assumption 4.4.2.2: Sell-on Contract

Without the entry, the demands for the two incumbent sellers are:

$$\begin{aligned} D_1^s &= \frac{1}{2}[1 - p_1^s + \theta(p_2^s - p_1^s)] \\ D_2^s &= \frac{1}{2}[1 - p_2^s + \theta(p_1^s - p_2^s)] \end{aligned} \tag{4.4.2.2}$$

After the entry, it is modified to:

$$\begin{aligned} D_r^o &= \frac{1}{2+a}\left[a - p_r^o + \frac{1}{2}[\delta(p_1^o - p_r^o) + \delta(p_2^o - p_r^o)]\right] \\ D_1^o &= \frac{1}{2+a}\left[1 - p_1^o + \frac{1}{2}[\theta(p_2^o - p_1^o) + \delta(p_r^o - p_1^o)]\right] \\ D_2^o &== \frac{1}{2+a}\left[1 - p_2^o + \frac{1}{2}[\theta(p_1^o - p_2^o) + \delta(p_r^o - p_2^o)]\right] \end{aligned} \tag{4.4.2.3}$$

Where a captures the strength of pltform's own brand. θ, δ are the cross sensitivity.

They find that in both the sell-to and sell-on conditions, the entry of the platform owner may not harm the incumbent sellers. What's more, comparing to the entry of a new third-party sellers, the incumbent sellers profit more when facing the entry of the owner. [Lai et al., 2022] investigates the introduction of Amazon's fullfillment program (FBA) on the third-party sellers in the E-commerce platform.

Assumption 4.4.2.3: FBA on the third-party sellers

- Both Amazon and the third-party sell substitute products, at price of p_A and p_S;
- Amazon procure the products from OEM supplier at price w with a cost c_0, the third-party obtain its products with a cost g;
- The third party pays r share of its revenue to Amazon as commission;
- The delivery cost for different means are all c, the Amazon provides a better service $S_A > S_s$, the third party can pay $T > c$ to use FBA;

Without FBA, the demands are given by:

$$\begin{aligned} q_A &= Q_A - p_A + \beta p_S + \alpha s_A - \eta s_S, \\ q_S &= Q_S - p_S + \beta p_A + \alpha s_S - \eta s_A, \end{aligned} \tag{4.4.2.4}$$

With FBA, the demands are given by:

$$\begin{aligned} q_A &= Q_A - p_A + \beta p_S + \alpha s_A - \eta s_A, \\ q_S &= Q_S - p_S + \beta p_A + \alpha s_A - \eta s_A. \end{aligned} \tag{4.4.2.5}$$

They assume are informations are common knowledge.

There are three decision makers: Amazon: p_A, T, Third-party: p_S, FBA, OEM: w:

$$\begin{aligned} \Pi_{A,N} &= (p_A - w - c)q_A(p_A, p_S) + rp_S q_S(p_A, p_S) \\ \Pi_{S,N} &= [(1-r)p_S - g - c]q_S(p_A, p_S) \\ \Pi_{M,N} &= (w - c_0)q_A(p_A, p_S) \end{aligned} \tag{4.4.2.6}$$

To given insights which policy would benefit different parties, they took the following steps:

1. In the without-FBA condition, given w, solve the equilibrium for $p_A^\star(w), p_S^\star(w)$;
2. Substituting them into the decision model of OEM, solve the optimal $w^\star$ in the SOSC condition.
3. In the FBA condition, repete above steps, solve the equilibrium for $p_A^\star(w), p_S^\star(w), w^\star$;
4. Substituting them into the decision model of OEM, solve the optimal $T^\star$.

They find some interesting insights:

1. The third party benifits from FBA when $T < \bar{T}$;
2. Amazon can benefit from FBA if η is either small or large;
3. The FBA program can achieve a *win-win-win* outcome for he third-party seller, Amazon and its OEM supplier.

4.4.3 Consumer Polarization

Consumer polarization is a topic in consumer research. **Group-Polarization Hypothesis** suggests that group discussion generally produces attitudes that are more extreme in the direction of the average of prediscussion attitudes in a variety of situations. Works like [Rao and Steckel, 1991] provides a mathematical presentation for this phenomenon (in the domain of preference):

$$U_{\mathrm{s}} = \sum_{i=1}^{m} \lambda_i u_i + \phi(\bar{u} - K)0 \leq \lambda_i \leq 1, \sum_{i=1}^{m} \lambda_i = 1, \phi \geq 0 \tag{4.4.3.1}$$

In this model, the $\bar{u}$ is the algebraic mean of all consumers' utility, and K is the **Pivot Point**. Rewrite this formula:

$$\begin{aligned} U_{\mathrm{g}} &= \sum_{i=1}^{m} \left(\lambda_i + \frac{\phi}{m} \right) u_i - \phi K & & w_0 \leq 0; \\ & & & \sum_{i=1}^{m} w_i \geq 1; \\ &= w_0 + w_1 u_1 + w_2 u_2 + \cdots + w_m u_m & & 0 \leq \tfrac{w_0}{1 - \sum w_i} \leq 1. \end{aligned} \tag{4.4.3.2}$$

Remark 4.4.1

- [Zhao et al., 2023] use experiment results to suggest that eWOM (electronic word of mouth) polarization (the degree of eWOM to which positive and negative sentiments are simultaneously strong) would decrease the consumers' intention to purchase, mediating by the enhancement of attitude ambivalence. (Ambivalence is a psychological state where a person endorses both positive and negative attitudinal positions)
- [Iyer and Yoganarasimhan, 2021] use game theory framework, to get the conclusion that sequential decision-making could reduce the polarization.

4.4.4 Network Effect

Network Effect and **Network Externality**:

[Narayan et al., 2011] verifies that peer influence affects attribute preferences via a Bayesian updating mechanism. In their model, the utility is given as follows:

$$U_{ijp}^{R} = X_{jp} \beta_i^{R} + \lambda_i \varepsilon_{ijp}^{R} \tag{4.4.4.1}$$

Where U_{ijp} is the utility of consumer i for product j given choice set p, X_{jp} is the attribute of product j in the choice set p, β_i^R is the customers' weights. The Bayesian updating process is given below:

$$\beta_{ik}^{R} = \rho_{ik} \beta_{ik}^{l} + (1 - \rho_{ik}) \frac{\sum_{i=1, i \neq i}^{N} w_{ii} \beta_{ik}^{l}}{\max \left[\left(\sum_{i=1, i \neq i}^{N} w_{ii} \right), 1 \right]}, \qquad \text{where } 0 \leq \rho_{ik} \leq 1. \tag{4.4.4.2}$$

Other research on peer influence:

Remark 4.4.2

The consumers' interaction and social connections have a proposition proposed in [Zhang et al., 2017] for their goal attainment and spending: a positive linear term plus a negative squared term;

4.4.5 Online Gaming

Many industrial news about online gaming can be found in [Chen et al., 2017]. In [Lei, 2022], the dissertation fully discussed loot box pricing, matchmaking, and price discrimination with fairness constraints.

Play-Duration and Spending

[Zhang et al., 2017]'s work shows a nonlinear effect of social connections and interactions on consumers' goal attainment and spending: A positive linear terms and a negative squared term. Mechanism: functional in providing useful information or tips that can facilitate goal attainment, but would raise information overload problems.

Player engagement can be embodied by many specific metrics, such as time or money spent in the game, the number of matches played within a time window, or churn risk. [Chen et al., 2017] define churn risk as the proportion of total players stopping playing the game over a period of time.

Matchmaking

Matchmaking connects multiple players to participate in online PvP games. (PvP(Player-versus-Player) games, which cover many popular genres, such as multiplayer online battle arena (MOBA), first-person shooting (FPS), and e-sports, have increased worldwide popularity in recent years.)

The past matchmaking strategy matches similar skilled players in the same round (SBMM), the current MM system focuses on improving the players' engagement and decreasing the churn rate. For example, in [Chen et al., 2017] EOMM (Engagement Optimization MatchMaking) is proposed to minimize the churn rate.

[Chen et al., 2021] propose an algorithm to maximize the cumulative active players.

Assumption 4.4.5.1: Chen 2021 MatchMaking

1. players can have heterogeneous skill levels: level 1 to level K;
2. the outcome of each match is a Bernoulli random variable: $p_{kj} = 1 - p_{jk}, p_{kk} = 0.5$, $p_{kj} > 0.5$ if $k > j$;
3. player's skill level is fixed: *relative* level;
4. and their state depends on the win-loss outcomes of the past m matches: $g \in \mathcal{G}$ ($2^m + 1$ possible cardinality);
5. A geometric losing churn model: players churn with a fixed probability, starting from the second loss in a row;
6. $P^k_{win}, P^k_{lose} \in [0,1]^{|\mathcal{G}|\times|\mathcal{G}|}$ is the transition matrix of level k player's engagement state;
7. $M_{kj} = p_{kj}P^k_{win} + (1 - p_{kj})P^k_{lose}$ is the aggregate transition matrix. ($\bar{\mathrm{G}}$ is the reduced aggregate transition matrix);
8. using the fluid matching model and assume players are infinitely divisible;

The **Dynamic Programming** formulation: $f_{kg,jg'}$ is the amount of kg players matched with jg' players, s^t_{kg} is the number of kg players at time t.
FB *flow balance* constraints:

$$
\begin{aligned}
\sum_{j=1}^{K}\sum_{g'\in\bar{\mathcal{G}}} f^t_{kg,jg'} &= s^t_{kg}, k = 1,\dots,K, \forall g \in \bar{\mathcal{G}},\\
\sum_{j=1}^{K}\sum_{g'\in\bar{\mathcal{G}}} f^t_{jg',kg} &= s^t_{kg}, k = 1,\dots,K, \forall g \in \bar{\mathcal{G}},\\
f^t_{kg,jg'} &= f^t_{jg',kg}, j = 1,\dots,K, k = 1,\dots,K, \forall g \in \bar{\mathcal{G}}, g' \in \bar{\mathcal{G}}\\
f^t_{kg,jg'} &\geq 0, j = 1,\dots,K, k = 1,\dots,K, \forall g \in \bar{\mathcal{G}}, g' \in \bar{\mathcal{G}}
\end{aligned}
\tag{4.4.5.1}
$$

ED *evolution of demographics*:

$$
\mathbf{s}^{t+1}_k = \sum_{j=1,\dots,K}\left(\mathbf{f}^t_{kj}\mathbf{1}\right)^\top\left(\bar{M}_{kj} + N_k\right) k = 1,\dots,K \tag{4.4.5.2}
$$

The value-to-go function is:

$$V^{\pi}(\mathbf{s}^{t}) = \sum_{k=1}^{K} \sum_{g \in \bar{\mathcal{G}}} s_{kg}^{t+1} + \gamma V^{\pi}(\mathbf{s}^{t+1}) \quad \text{subject to (FB), (ED).} \tag{4.4.5.3}$$

The above model can be formulated in a linear programming style:

Theorem 4.4.5.1: Chen 2021 MM LP Formulation

$$\begin{aligned} V^{*}(\mathbf{s}^{0}) &= \max \sum_{t=1}^{\infty} \gamma^{t-1} \sum_{k} \sum_{g \in \bar{\mathcal{G}}} s_{kg}^{t} \\ \text{s.t.} & \sum_{j=1}^{K} \sum_{g' \in \bar{\mathcal{G}}} f_{kg,jg'}^{t} = s_{kg}^{t}, \forall k, \forall g \in \bar{\mathcal{G}}, t = 0, 1, \dots \\ & \sum_{j=1}^{K} \sum_{g' \in \bar{\mathcal{G}}} f_{jg',kg}^{t} = s_{kg}^{t}, \forall k, \forall g \in \bar{\mathcal{G}}, t = 0, 1, \dots \\ & f_{kg,jg'}^{t} = f_{jg',kg}^{t}, j = 1, \dots, K, k = 1, \dots, K, \forall g \in \bar{\mathcal{G}}, g' \in \bar{\mathcal{G}}, t = 0, 1, \dots \\ & f_{kg,jg'}^{t} \geq 0, j = 1, \dots, K, k = 1, \dots, K, \forall g \in \bar{\mathcal{G}}, g' \in \bar{\mathcal{G}}, t = 0, 1, \dots \\ & \mathbf{s}_{k}^{t+1} = \sum_{j=1,\dots,K} \left(\mathbf{f}_{kj}^{t} \mathbf{1}\right)^{\top} \left(\bar{M}_{kj} + N_{k}\right), \forall k, t = 0, 1, \dots \end{aligned} \tag{4.4.5.4}$$

Remark 4.4.3

- Using an optimal matchmaking policy instead of SBMM may reduce the required bot ratio significantly while maintaining the same level of engagement.

4.5 Behavioral Operations Management

4.6 Data-Driven Operations Management

Chapter 5

Miscellaneous

5.1 Notes on Tools

5.1.1 LaTeX Shortcuts

There are 6x6 colors in the preset preamble:

aa	ab	ac	ad	ae	af
ba	bb	bc	bd	be	bf
ca	cb	cc	cd	ce	cf
da	db	dc	dd	de	df
ea	eb	ec	ed	ee	ef
fa	fb	fc	fd	fe	ff

Using `\href{URL}{text}` to refer a website.
Using `\eq` to write equation, `\tab` to get an unordered list, `\lis` to get an ordered list.

$$E = mc^2 \tag{5.1.1.1}$$

- item 1;
- item 2;
- item 3.

1. item 1;
2. item 2;

3. item 3.

Format of Words

```
\hl{highlighted}, \ul{underlined},
↪  \st{strikethrough}\\
\rt{red}, \yt{yellow}, \bt{blue},
↪  \gt{green}
```

highlighted, underlined, ~~strikethrough~~
red, yellow, blue, green

Shortcuts

```
\RR, \NN, \ZZ, \QQ\\
\bA, \bB, \bC, \bD
```

$\mathbb{R}$, $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$
$\mathbb{A}, \mathbb{B}, \mathbb{C}, \mathbb{D}$

Shortcuts

```
\tbf{Text}, \tit{Text}\\
\cA, \cB, \cC, \cD
```

Text, *Text*
A, B, C, D

Emoji

```
\emogood, \emobad, \emocool,
↪  \emoheart, \emotree
```

😊, ☹, 😎, ♥, 🌳

Use `\ass`, `\ax`, `\thm`, `\co`, `\pro`, `\defi`, `\re`, `\key`, `\ex`, `\proo` to use preset tcolorboxes template.

Assumption 5.1.1.1: Example

1. item 1;

2. item 2;

3. item 3.

Axiom 5.1.1.1: Exmaple

Test

Theorem 5.1.1.1: Exmaple

Test

Corollary 5.1.1.1: Exmaple

Test

Proposition 5.1.1.1: Exmaple

Test

Definition 5.1.1.1: Exmaple

Test

Remark 5.1.1

Expamle

Keywords 5.1

Example

Example 5.1.1.1

Problem example

Solution:

The solution is omitted.

Proof 5.1.1: proposition 5.1.1

The Formal Proof

Q.E.D.

Using `\label, \ref` to refer to chapters 5, sections 5.1, equations 5.1.1, and boxes 5.1.1.

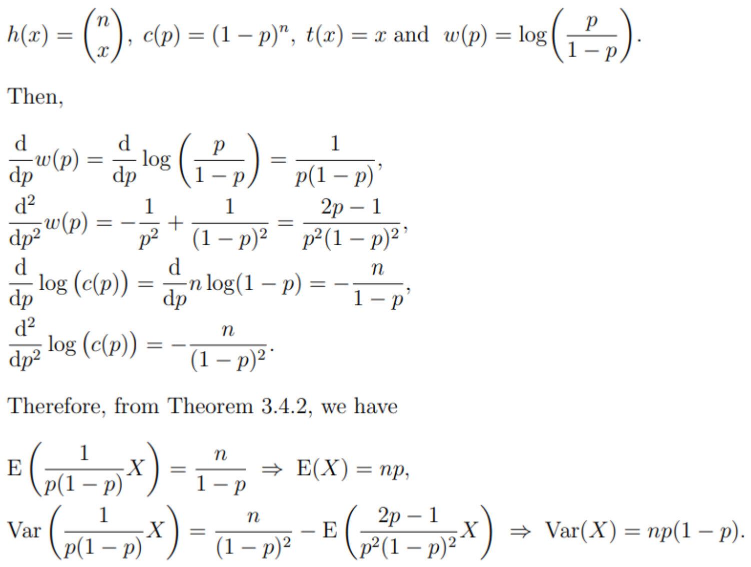
Using `\cite` to cite the literature in apa style. For example: [Klein et al., 2020]

Using `\sep` to insert a horizontal line with words in the middle:

Compilation

Cleaning all the auxiliary files: LaTeXmk → BibTeX → LaTeXmk → LaTeXmk. Or, zip the main files and upload them to Overleaf.

Put photos in the *pic* file and use `\fig` to show it.



$$h(x) = \binom{n}{x}, \ c(p) = (1-p)^n, \ t(x) = x \text{ and } \ w(p) = \log\left(\frac{p}{1-p}\right).$$

Then,

$$\frac{\mathrm{d}}{\mathrm{d}p}w(p) = \frac{\mathrm{d}}{\mathrm{d}p}\log\left(\frac{p}{1-p}\right) = \frac{1}{p(1-p)},$$
$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}w(p) = -\frac{1}{p^2} + \frac{1}{(1-p)^2} = \frac{2p-1}{p^2(1-p)^2},$$
$$\frac{\mathrm{d}}{\mathrm{d}p}\log\left(c(p)\right) = \frac{\mathrm{d}}{\mathrm{d}p}n\log(1-p) = -\frac{n}{1-p},$$
$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}\log\left(c(p)\right) = -\frac{n}{(1-p)^2}.$$

Therefore, from Theorem 3.4.2, we have

$$\mathrm{E}\left(\frac{1}{p(1-p)}X\right) = \frac{n}{1-p} \ \Rightarrow \ \mathrm{E}(X) = np,$$
$$\mathrm{Var}\left(\frac{1}{p(1-p)}X\right) = \frac{n}{(1-p)^2} - \mathrm{E}\left(\frac{2p-1}{p^2(1-p)^2}X\right) \ \Rightarrow \ \mathrm{Var}(X) = np(1-p).$$

Figure 5.1: Example.png

Using \alg to write the pseudo code:

Algorithm 1 Example Code

```
Require: $n \geq 0$
Ensure: $y = x^n$
  $y \leftarrow 1$
  $X \leftarrow x$
  $N \leftarrow n$
  while $N \neq 0$ do
      if $N$ is even then
          $X \leftarrow X \times X$
          $N \leftarrow \frac{N}{2}$                                  ▷ This is a comment
      else if $N$ is odd then
          $y \leftarrow y \times X$
          $N \leftarrow N - 1$
      end if
  end while
```

List of Figures

List of Algorithms

Bibliography

[Agarwal et al., 2019] Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32.

[Angrist and Pischke, 2009] Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press.

[Angrist and Pischke, 2014] Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The Path from Cause to Effect*. Princeton university press.

[Bertrand and Schoar, 2006] Bertrand, M. and Schoar, A. (2006). The role of family in family firms. *Journal of economic perspectives*, 20(2):73–96.

[Bertsimas and Tsitsiklis, 1997] Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to Linear Optimization*, volume 6. Athena scientific Belmont, MA.

[Brusco et al., 2017] Brusco, M. J., Singh, R., Cradit, J. D., and Steinley, D. (2017). Cluster analysis in empirical OM research: Survey and recommendations. *International Journal of Operations & Production Management*, 37(3):300–320.

[Calvano et al., 2020] Calvano, E., Calzolari, G., Denicolo, V., and Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297.

[Cao et al., 2022] Cao, J., Xu, Y., and Zhang, C. (2022). Clans and calamity: How social capital saved lives during China's Great Famine. *Journal of Development Economics*, 157:102865.

[Casella and Berger, 2021] Casella, G. and Berger, R. L. (2021). *Statistical Inference*. Cengage Learning.

[Chen et al., 2021] Chen, M., Elmachtoub, A. N., and Lei, X. (2021). Matchmaking strategies for maximizing player engagement in video games. *Available at SSRN 3928966*.

[Chen and Gallego, 2019] Chen, N. and Gallego, G. (2019). Welfare analysis of dynamic pricing. *Management Science*, 65(1):139–151.

[Chen and Tsai, 2023] Chen, N. and Tsai, H.-T. (2023). Price competition under information (dis) advantage. *Available at SSRN 4420175*.

[Chen et al., 2023] Chen, W., Wei, Z., and Xie, K. (2023). Regulating professional players in peer-to-peer markets: Evidence from Airbnb. *Management Science*, 69(5):2893–2918.

[Chen et al., 2017] Chen, Z., Xue, S., Kolen, J., Aghdaie, N., Zaman, K. A., Sun, Y., and Seif El-Nasr, M. (2017). EOMM: An Engagement Optimized Matchmaking Framework. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1143–1150, Perth Australia. International World Wide Web Conferences Steering Committee.

[Cheng et al., 2022] Cheng, H. K., Jung, K. S., Kwark, Y., and Pu, J. (2022). Impact of own brand product introduction on optimal pricing models for platform and incumbent sellers. *Information Systems Research*.

[Cheng et al., 2021] Cheng, J., Dai, Y., Lin, S., and Ye, H. (2021). Clan culture and family ownership concentration: Evidence from China. *China Economic Review*, 70:101692.

[Choi et al., 2016] Choi, T.-M., Cheng, TCE., and Zhao, X. (2016). Multi-methodological research in operations management. *Production and Operations Management*, 25(3):379–389.

[Deng et al., 2023] Deng, Y., Tang, C. S., Wang, W., and Yoo, O. S. (2023). Can third-party sellers benefit from a platform's entry to the market? *Service Science*.

[Dolnicar, 2021] Dolnicar, S. (2021). *Airbnb Before, During and After COVID-19*. The University of Queensland.

[Farronato and Fradkin, 2022] Farronato, C. and Fradkin, A. (2022). The welfare effects of peer entry: The case of Airbnb and the accommodation industry. *American Economic Review*, 112(6):1782–1817.

[Fisher, 2007] Fisher, M. (2007). Strengthening the empirical base of operations management. *Manufacturing & Service Operations Management*, 9(4):368–382.

[Fisher and Raman, 2022] Fisher, M. and Raman, A. (2022). Innovations in retail operations: Thirty years of lessons from Production and Operations Management. *Production and Operations Management*, 31(12):4452–4461.

[Foerderer et al., 2018] Foerderer, J., Kude, T., Mithas, S., and Heinzl, A. (2018). Does platform owner's entry crowd out innovation? Evidence from Google photos. *Information Systems Research*, 29(2):444–460.

[Folger, 2023] Folger, J. (2023). How Airbnb Works—for Hosts, Guests, and the Company Itself.

[Gallego and Topaloglu, 2019] Gallego, G. and Topaloglu, H. (2019). *Revenue Management and Pricing Analytics*, volume 279 of *International Series in Operations Research & Management Science*. Springer New York, New York, NY.

[Gans et al., 2007] Gans, N., Knox, G., and Croson, R. (2007). Simple models of discrete choice and their performance in bandit experiments. *Manufacturing & Service Operations Management*, 9(4):383–408.

[Gu and Zhu, 2021] Gu, G. and Zhu, F. (2021). Trust and disintermediation: Evidence from an online freelance marketplace. *Management Science*, 67(2):794–807.

[Guttentag, 2019] Guttentag, D. (2019). Progress on airbnb: A literature review. *Journal of Hospitality and Tourism Technology*, 10(4):814–844.

[Ha et al., 2022] Ha, A. Y., Tong, S., and Wang, Y. (2022). Channel structures of online retail platforms. *Manufacturing & service operations management*, 24(3):1547–1561.

[Hagiu et al., 2020] Hagiu, A., Jullien, B., and Wright, J. (2020). Creating platforms by hosting rivals. *Management Science*, 66(7):3234–3248.

[He et al., 2020] He, S., Peng, J., Li, J., and Xu, L. (2020). Impact of platform owner's entry on third-party stores. *Information Systems Research*, 31(4):1467–1484.

[Hwang, 2017] Hwang, K. (2017). *Cloud Computing for Machine Learning and Cognitive Applications*. Mit Press.

[Iyer and Yoganarasimhan, 2021] Iyer, G. and Yoganarasimhan, H. (2021). Strategic Polarization in Group Interactions. *Journal of Marketing Research*, 58(4):782–800.

[Klein et al., 2020] Klein, R., Koch, S., Steinhardt, C., and Strauss, A. K. (2020). A review of revenue management: Recent generalizations and advances in industry applications. *European Journal of Operational Research*, 284(2):397–412.

[Kumar and Tang, 2022] Kumar, S. and Tang, C. S. (2022). Expanding the boundaries of the discipline: The 30th-anniversary issue of Production and Operations Management. *Production and Operations Management*, 31(12):4257–4261.

[Lai et al., 2022] Lai, G., Liu, H., Xiao, W., and Zhao, X. (2022). "Fulfilled by amazon": A strategic perspective of competition at the e-commerce platform. *Manufacturing & Service Operations Management*, 24(3):1406–1420.

[Lei, 2022] Lei, X. (2022). *Revenue Management in Video Games and with Fairness*. PhD thesis, Columbia University.

[Luenberger et al., 1984] Luenberger, D. G., Ye, Y., et al. (1984). *Linear and Nonlinear Programming*, volume 2. Springer.

[Misra et al., 2019] Misra, K., Schwartz, E. M., and Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252.

[Mithas et al., 2022] Mithas, S., Chen, Y., Lin, Y., and De Oliveira Silveira, A. (2022). On the causality and plausibility of treatment effects in operations management research. *Production and Operations Management*, 31(12):4558–4571.

[Narayan et al., 2011] Narayan, V., Rao, V. R., and Saunders, C. (2011). How Peer Influence Affects Attribute Preferences: A Bayesian Updating Mechanism. *Marketing Science*, 30(2):368–384.

[Peters et al., 2017] Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

[Rao and Steckel, 1991] Rao, V. R. and Steckel, J. H. (1991). A polarization model for describing group preferences. *Journal of Consumer Research*, 18(1):108–118.

[Rietveld and Schilling, 2021] Rietveld, J. and Schilling, M. A. (2021). Platform competition: A systematic and interdisciplinary review of the literature. *Journal of Management*, 47(6):1528–1563.

[Roth and Singhal, 2022] Roth, A. M. and Singhal, V. R. (2022). Pioneering role of the Production and Operations Management in promoting empirical research in operations management. *Production and Operations Management*, 31(12):4529–4543.

[Roth, 2007] Roth, A. V. (2007). Applications of empirical science in manufacturing and service operations. *Manufacturing & Service Operations Management*, 9(4):353–367.

[Shi et al., 2023] Shi, R., Aaltonen, A., Henfridsson, O., and Gopal, R. D. (2023). Comparing platform owners' early and late entry into complementary markets. *MIS Quarterly*.

[Spirtes, 2010] Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, 11(5).

[Strauss et al., 2018] Strauss, A. K., Klein, R., and Steinhardt, C. (2018). A review of choice-based revenue management: Theory and methods. *European Journal of Operational Research*, 271(2):375–387.

[Thrun and Littman, 2000] Thrun, S. and Littman, M. L. (2000). Reinforcement learning: An introduction. *AI Magazine*, 21(1):103–103.

[Wang et al., 2023] Wang, J., Huang, T., and Lee, J. (2023). Cross-licensing in a Supply Chain with Asymmetric Manufacturers.

[Wen and Zhu, 2019] Wen, W. and Zhu, F. (2019). Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal*, 40(9):1336–1367.

[Yang, 2019] Yang, H. (2019). Family clans and public goods: Evidence from the New Village beautification project in South Korea. *Journal of Development Economics*, 136:34–50.

[Zhang, 2019] Zhang, C. (2019). Family support or social support? The role of clan culture. *Journal of Population Economics*, 32:529–549.

[Zhang, 2020] Zhang, C. (2020). Clans, entrepreneurship, and development of the private sector in China. *Journal of Comparative Economics*, 48(1):100–123.

[Zhang et al., 2017] Zhang, C., Phang, C. W., Wu, Q., and Luo, X. (2017). Nonlinear Effects of Social Connections and Interactions on Individual Goal Attainment and Spending: Evidences from Online Gaming Markets. *Journal of Marketing*, 81(6):132–155.

[Zhang et al., 2022] Zhang, S., Lee, D., Singh, P. V., and Srinivasan, K. (2022). What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Management Science*, 68(8):5644–5666.

[Zhao et al., 2023] Zhao, P., Ma, Z., Gill, T., and Ranaweera, C. (2023). Social media sentiment polarization and its impact on product adoption. *Marketing Letters*.

[Zhu and Liu, 2018] Zhu, F. and Liu, Q. (2018). Competing with complementors: An empirical look at Amazon. com. *Strategic management journal*, 39(10):2618–2642.