
A Guide to Data Science



From mathematics to algorithms.

作者: Jinxiong & Zhang
时间: December 21, 2018
邮箱: jinxiongzhang@qq.com

Version: 3.00

目 录



1	Abstract	1
2	Statistical Foundation	2
2.1	Basic Probability	2
2.1.1	Random Variable	3
2.1.2	Discrete Distribution	3
2.1.3	Continuous Distribution	4
2.1.4	Bivariate Distribution	5
2.2	Representation of Random Variable	6
2.2.1	Mixture Representation	7
2.3	Bayes' Theorem	7
2.3.1	Conditional Probability	7
2.3.2	Bayes's Formula and Inverse Bayes' Formula	9
2.4	What determines a distribution?	11
2.4.1	Expectation, Variance and Entropy	11
2.4.2	Moment and Moment Generating Function	14
2.5	Sampling Methods	15
2.5.1	Sampling from Discrete Distribution	15
2.5.2	Sampling from Continuous Distribution	16
3	Numerical Optimization	19
3.1	Gradient Descent and More	19
3.2	Mirror Gradient Method	20
3.2.1	Projected Gradient Descent	20
3.2.2	Mirror Descent	21
3.3	Variable Metric Methods	22
3.3.1	Newton's Method	22
3.3.2	The Fisher Scoring Algorithm	22
3.3.3	Quasi-Newton Methods	23
3.3.4	Natural Gradient Descent	23
3.4	Expectation Maximization Algorithm	24
3.4.1	Generalized EM Algorithm	25

3.5	Alternating Direction Method of Multipliers	25
3.6	Stochastic Gradient Descent	26
4	ElegantBook 设置说明	30
4.1	编译方式	30
4.2	选项设置	30
4.3	数学环境简介	31
4.4	可编辑的字段	31
5	ElegantBook 写作示例	32
5.1	Economics and Differentiable Function	32
5.2	Bibliography	36
	参考文献	36





第 1 章 Abstract



This is a brief introduction to data science. **Data science** can be seen as an extension of statistics.

- All sciences are, in the abstract, mathematics.
- All judgments are, in their rationale, statistics. by C.P. Rao

On Chomsky and the Two Cultures of Statistical Learning or Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author) provides more discussion of statistical model or data science.



"The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work—that is, correctly to describe phenomena from a reasonably wide area."

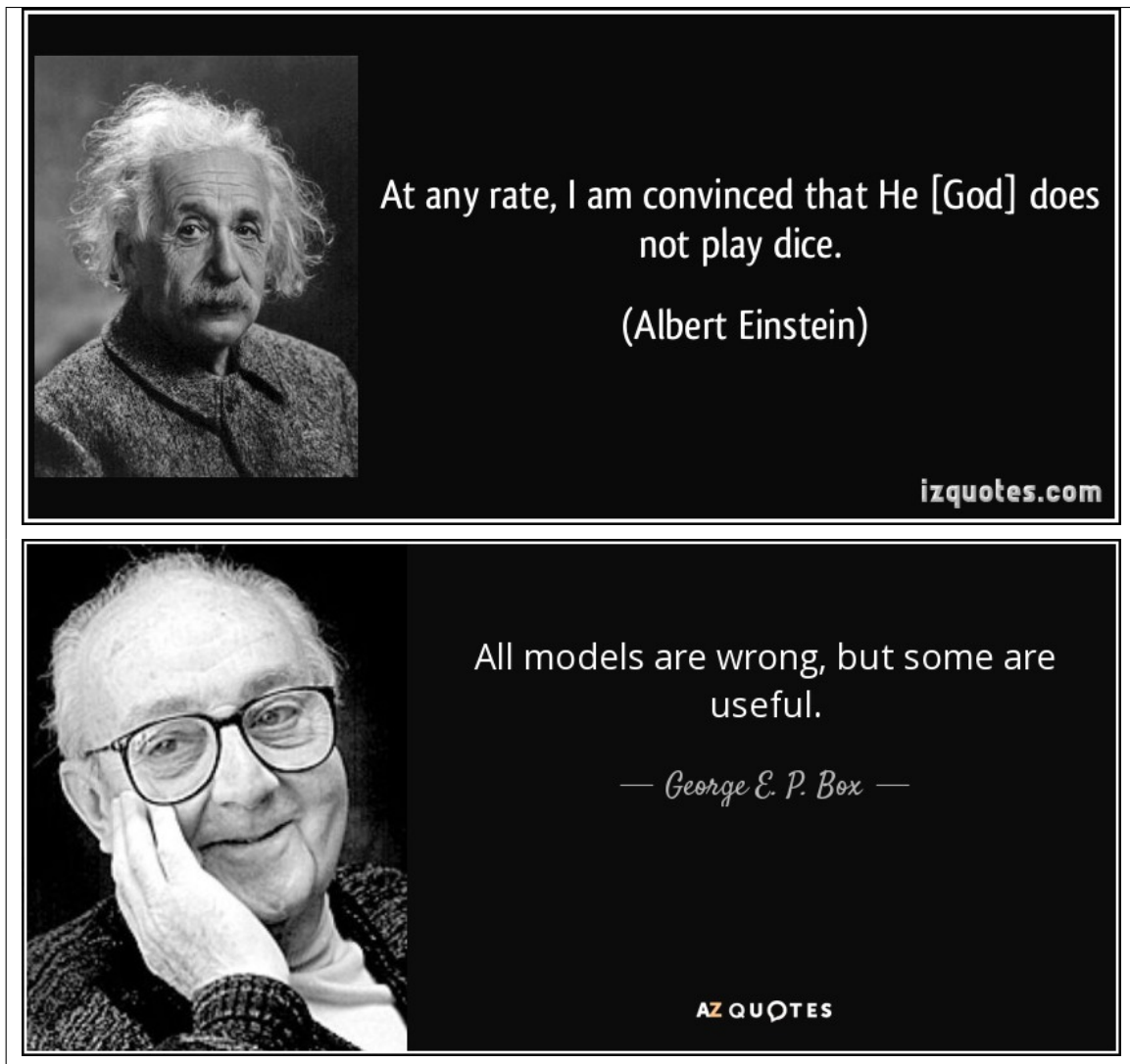
John von Neumann

第 2 章 Statistical Foundation



"Statistics, in a nutshell, is a discipline that studies the best ways of dealing with randomness, or more precisely and broadly, variation", Professor Xiao-Li Meng said.

表 2.1: On mathematical model



2.1 Basic Probability

Probability theory is regarded as the theoretical foundation of statistics, which provides the ideal models to measure the stochastic phenomenon, randomness, chance or luck. It is the distribution theory bridging probability and statistics.

The project [Seeing theory](#) is a brief introduction to statistics and probability theory and provides a [.pdf file](#).

2.1.1 Random Variable

Definition 2.1: Random Variable

A random variable is a mapping

$$X : \Omega \rightarrow \mathbb{R}$$

that assigns a real number $X(\omega)$ to each outcome ω .

- Ω is the sample space.
- ω in Ω are called sample outcomes or realization.
- Subsets of Ω are called Event.



Definition 2.2: Probability

A function P is called probability function with respect to the known sample space Ω and all the event $\forall \omega \in \Omega$ if

Properties	Conditions
Nonnegative	$P(\omega) \geq 0 \forall \omega \in \Omega$
Normalized	$P(\Omega) = 1$
Additive	If $\omega_1, \omega_2, \dots \in \Omega$ are disjoint, then $P(\bigcup_{i=1}^{\infty} \omega_i) = \sum_{i=1}^{\infty} P(\omega_i)$



These properties are called [Kolmogorov axioms](#).

2.1.2 Discrete Distribution

Random variable or event is countable or listed, then we can list the probability of every event. If so, we can compute the probability of a random variable less than the given value. For example, $P(X \leq 10) = \sum_{i=1}^{10} P(X = i)$ if the random variable X only takes positive integers.

Definition 2.3: Probability mass function

The probability mass function of a discrete random variable X is defined as

$$f_X(x) = P_X(X = x)$$

for all x .



The left subindex X is to point out that the probability is with respect to the random variable X and it can omit if no confusion.



2.1.3 Continuous Distribution

If random variables may take on a continuous range of values, it is hard or valueless to compute the probability of a given value. It is usually to find the cumulative distribution function.

Definition 2.4: Cumulative Distribution Function

A function $F(x)$ is called cumulative distribution function (CDF in short) if

- $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$;
- The function $F(x)$ is monotone increasing function with x ;
- $\lim_{x \rightarrow x_0^+} F(x) = F(x_0)$ for all x_0 .



In probability, $\forall x, F(x) = P(X \leq x)$ for some random variable X .

For example, if the random variable X has the cumulative distribution function (CDF)

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

we say the random variable X is uniformly distributed in $[0, 1]$. We can see [this link](#).

The probability density function can be seen as the counterpart of pmf. For the continuous random variable X , its CDF $F_X(x)$ is not required to be differentiable. But if so, we can compute the derivative of the CDF $F_X(x)$: $\frac{dF_X(x)}{dx} = f_X(x)$.

Definition 2.5: Probability Density Function

We call the function $f_X(x)$ is the probability density function with respect to the continuous random variable X if

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

for all x , where $F_X(x)$ is the probability cumulative function of the random variable X . ♣

For example, the random variable X uniformly distributed in $[0, 1]$ has the probability density function

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

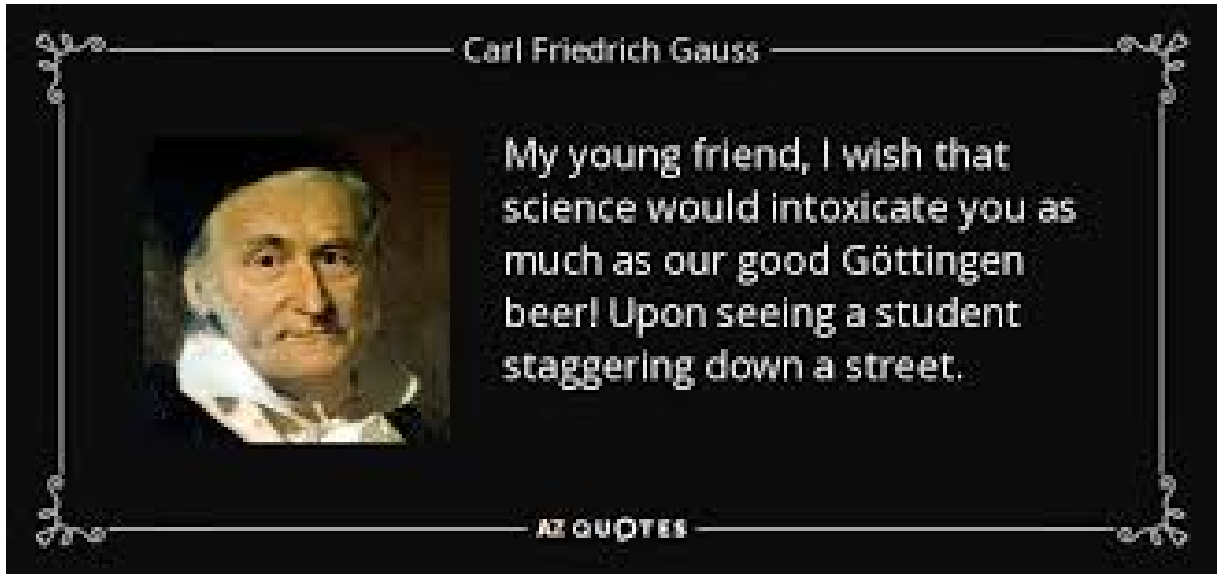
If the pdf f_X is positive in the set S , we call the set S is the support set or support of the distribution.

One common probability - **normal or Gaussian distribution** - is often given in the pdf:

$$f(x|\sigma^2, \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$



where σ^2 is given positive and finite, μ is given and finite. And its support is $(-\infty, \infty)$.



The density function of a power law distribution is in the form of

$$f_X(x) = x^{-\alpha}, x \in [1, \infty)$$

where α is real. It is also called **Pareto-type distribution**.

The power law can be used to describe a phenomenon where a small number of items is clustered at the top of a distribution (or at the bottom), taking up 95% of the resources. In other words, it implies a small amount of occurrences is common, while larger occurrences are rare.

Every probability distribution is an ideal mathematical model that describes some real event. Here is [a field guide to continuous distribution](<http://threeplusone.com/FieldGuide.pdf>).

2.1.4 Bivariate Distribution

The cumulative probability function and probability density function are really functions with some required properties. We wonder the bivariate functions in probability theory.

Definition 2.6: Joint Distribution Function

The joint distribution function $F : \mathbb{R}^2 \rightarrow [0, 1]$, where X and Y are discrete variables, is given by

$$F_{(X,Y)}(x, y) = P(X \leq x, Y \leq y).$$



Their joint mass function $f : \mathbb{R}^2 \rightarrow [0, 1]$ is given by

$$f_{(X,Y)}(x, y) = P(X = x, Y = y).$$



Definition 2.7: Joint Density Function

The joint distribution function $F_{(X,Y)} : \mathbb{R}^2 \rightarrow [0, 1]$, where X and Y are continuous variables, is given by $F_{(X,Y)}(x, y) = P(X \leq x, Y \leq y)$. And their joint density function if $f_{X,Y}(x, y)$ satisfies that

$$F_{(X,Y)}(x, y) = \int_{v=-\infty}^y \int_{u=-\infty}^x f_{(X,Y)}(u, v) du dv$$

for each $x, y \in \mathbb{R}$.

**Definition 2.8: Marginal Distribution**

The marginal distributed of X and Y are

$$F_X(x) = P(X \leq x) = F_{(X,Y)}(x, \infty) \quad \text{and} \quad F_Y(y) = P(Y \leq y) = F_{(X,Y)}(\infty, y)$$

for discrete random variables;

$$F_X(x) = \int_{-\infty}^x \left(\int_{\mathbb{R}} f_{(X,Y)}(u, y) dy \right) du \quad \text{and} \quad F_Y(y) = \int_{-\infty}^y \left(\int_{\mathbb{R}} f_{(X,Y)}(x, v) dx \right) dv$$

for continuous random variables and the marginal probability density function is

$$f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x, y) dy, \quad f_Y(y) = \int_{\mathbb{R}} f_{(X,Y)}(x, y) dx.$$



Two random variables X and Y are called as identically distributed if $P(X \in A) = P(Y \in A)$ for any $A \subset \mathbb{R}$.

2.2 Representation of Random Variable

The random variable is nearly no sense without its probability distribution. We can choose the CDF or pdf to depict the probability, which depends on the case.

We know that a function of random variable is also a random variable. For example, supposing that X is a random variable, the function $g(X) = e^X$ is a random variable as well as $g(X) = \ln(X)$.

The function of random variable must have its distribution. We can take it to generate more distributions.

Theorem 2.1: S

Suppose that X is with the cumulative distribution function CDF $F_X(x)$ and $Y = F_X(X)$, then Y is *uniformly* distributed in the interval $[0, 1]$.



Let X and Y be two different random variables, $Z = X + Y$ or $Z = X \cdot Y$ are typical



functions of random variables, especially X is discrete while Y is continuous.

2.2.1 Mixture Representation

Let P_1, P_2, \dots, P_n be probability distribution and $\sum_{i=1}^n \lambda_i = 1, \lambda_i > 0 \quad \forall i \in \{1, 2, \dots, n\}$, we can get

$$P = \sum_{i=1}^n \lambda_i P_i$$

is also a probability distribution. If X is distributed in P , i.e. $X \sim P$, its probability is computed as

$$P(X = x) = \sum_{i=1}^n \lambda_i P_i(X = x)$$

for discrete random variable or

$$P(X \leq x) = \sum_{i=1}^n \lambda_i P_i(X \leq x)$$

for continuous random variable.

A random variable X is said to have a mixture distribution if the distribution of X depends on a quantity that also has a distribution.

Sometimes, it is quite difficult if we directly generate a random vector $X \sim f_X(x)$, but the augmented vector $(X, Z) \sim f_{(X,Z)}(x, z)$ is relatively easy to generate such as [Box-Muller algorithm]. And we can represent $f_X(x)$ as the marginal probability distribution of $f_{(X,Z)}(x, y)$ in integral form

$$\int_{\mathbb{R}} f_{(X,Z)}(x, z) dz.$$

Thanks to Professor Tian Guoliang(Gary) in SUSTech, who brought me to computational statistics. More information on probability can be founded in [The Probability web](#).

2.3 Bayes' Theorem

Bayes theorem is the foundation of Bayesian statistics in honor of the statistician [Thomas Bayes](#).

2.3.1 Conditional Probability

If the set A is the subset of B , i.e. $A \subset B$, we know that if $x \in A$, $x \in B$. We can say if A happened, the event B must have happened. However, if B happened, what is probability when A happened? It should be larger than the probability of A when B did not happened and it is also larger than the probability of A when the events including B happened. We know that $A \cap B = A$ when A is the subset of B .



Definition 2.9: Conditional Probability

Supposing $A, B \in \Omega$ and $P(B) > 0$, the conditional probability of A given B (denoted as $P(A|B)$) is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$



The vertical bar $|$ means that the right one is given or known and is parameter rather than variable.

Here is a [visual explanation](#).

Definition 2.10: Statistical Independence

If $P(A \cap B) = P(A)P(B)$, we call the event A and B are statistically independent.



From the definition of conditional probability, it is obvious that:

- $P(B) = \frac{P(A \cap B)}{P(A|B)}$;
- $P(A \cap B) = P(A|B)P(B)$;
- $P(A \cap B) = P(B|A)P(A)$.

Thus

- $P(A|B)P(B) = P(B|A)P(A)$;
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(B|A) \frac{P(A)}{P(B)}$;
- $P(B) = \frac{P(B|A)P(A)}{P(A|B)} = \frac{P(B|A)}{P(A|B)} P(A)$.

Definition 2.11: Conditional Distribution Function

The conditional distribution function of Y given $X = x$ is the function $F_{Y|X}(y|x)$ given by

$$F_{Y|X}(y|x) = \int_{-\infty}^x \frac{f_{(X,Y)}(x, v)}{f_X(x)} dv = \frac{\int_{-\infty}^x f_{(X,Y)}(x, v) dv}{f_X(x)}$$

for the support of $f_X(x)$ and the conditional probability density function of $F_{Y|X}$, written $f_{Y|X}(y|x)$, is given by

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

for any x such that $f_X(x) > 0$.

**Definition 2.12: Conditional Probability for Continuous Random Variable**

If X and Y are non-generate and jointly continuous random variables with density



$f_{X,Y}(x, y)$ then, if B has positive measure,

$$P(X \in A | Y \in B) = \frac{\int_{y \in B} \int_{x \in A} f_{X,Y}(x, y) xy}{\int_{y \in B} \int_{x \in \mathbb{R}} f_{X,Y}(x, y) xy}.$$



Definition 2.13: Chain Rule of Probability

$$P(x^{(1)}, \dots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^n P(x^{(i)} | x^{(1)}, \dots, x^{(i-1)})$$



Definition 2.14: Conditional independence

The event A and B are **conditionally independent** given C if and only if

$$P(A \cap B | C) = P(A | C)P(B | C).$$



Theorem 2.2: Total Probability Theorem

Supposing the events A_1, A_2, \dots are disjoint in the sample space Ω and $\bigcup_{i=1}^{\infty} A_i = \Omega$, B is any subset of Ω , we have

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B | A_i)P(A_i).$$



Note:

- For the discrete random variable, the total probability theorem tells that any event can be decomposed to the basic event in the event space.
- For the continuous random variable, this theorems means that

$$\int_{\mathbb{B}} f_X(x) dx = \sum_{i=1}^{\infty} \int_{\mathbb{B} \cap A_i} f_X(x) x.$$

2.3.2 Bayes's Formula and Inverse Bayes' Formula

Theorem 2.3: Bayes's Formula

Supposing the events A_1, A_2, \dots are disjoint in the sample space Ω and $\bigcup_{i=1}^{\infty} A_i = \Omega$, B is any subset of Ω , we have

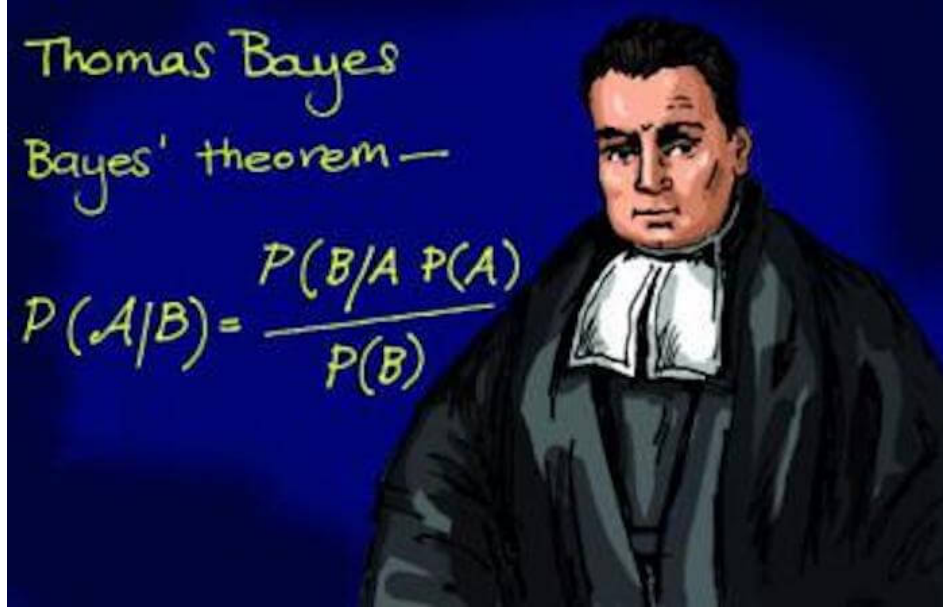
$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{i=1}^{\infty} P(B | A_i)P(A_i)} = \frac{P(B | A_i)P(A_i)}{P(B)}.$$





Note:

- The probability $P(A_i)$ is called prior probability of event A_i and the conditional probability $P(A_i|B)$ is called posterior probability of the event A_i in Bayesian statistics.
- For any A_i , $P(A_i|B) \propto P(B|A_i)P(A_i)$ and $P(B)$ is the normalization constant as the sum of $P(B|A_i)P(A_i)$.



The joint pdf of the random variables X and Y can be expressed as

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

in some proper conditions. Thus we get by division

$$f_Y(y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_{X|Y}(x|y)}. \quad (1)$$

Integrating this identity with respect to y on support of $f_Y(y)$, we immediately have the **point-wise formula** as shown below

$$f_X(x) = \left\{ \int_{f_{X|Y}(x|y) \neq 0} \frac{f_{Y|X}(y|x)}{f_{X|Y}(x|y)} dy \right\}^{-1}. \quad (2)$$

Now substitute (2) into (1), we obtain the dual form of IBF for $f_Y(y)$ and hence by symmetry we obtain the **function-wise formula** of $f_Y(y)$ at y_0 as shown in (f1), or the sampling formula in (f2) when the normalizing constant is omitted.

$$f_X(x) = \left\{ \int_{f_{Y|X}(y_0|x) \neq 0} \frac{f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)} dx \right\}^{-1} \frac{f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)} \quad (f1)$$

$$\propto \frac{f_{X|Y}(x|y_0)}{f_{Y|X}(y_0|x)} \quad (f2)$$

There are more information on **inverse Bayes' formula**:



表 2.2: The Inventor of IBF



- <http://web.hku.hk/~kaing/Background.pdf>
- http://101.96.10.64/web.hku.hk/~kaing/Section1_3.pdf
- <http://web.hku.hk/~kaing/HKSSinterview.pdf>
- <http://web.hku.hk/~kaing/HKSSstalk.pdf>

2.4 What determines a distribution?

The cumulative density function (CDF) or probability density function(pdf) is roughly equal to random variable, i.e. one random variable is always attached with CDF or pdf. However, what we observed is not the variable itself but its realization or sample.


- In theory, what properties do some CDFs or pdfs hold? For example, are they all integrable?
- In practice, the basic question is how to determine the CDF or pdf if we only observed some samples?

2.4.1 Expectation, Variance and Entropy

Expectation and variance are two important factors that characterize the random variable. The expectation of a discrete random variable is defined as the weighted sum.




Definition 2.15: Mean

For a discrete random variable X , of which the range is x_1, \dots, x_∞ , its mean $\mathbb{E}(X)$ is defined as $\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i P(X = x_i)$. 

The mean of a discrete random variable is a weighted sum. Average is one special expectation with equiprob, i.e. $P(X = x_1) = P(X = x_2) = \dots = P(X = x_i) = \dots \forall i$.

The expectation of a continuous random variable is defined as one special integration.

Definition 2.16: Expectation

For a continuous random variable X with the pdf $f(x)$, if the integration $\int_{-\infty}^{\infty} |x|f(x)dx$ exists, the value $\int_{-\infty}^{\infty} xf(x)dx$ is called the (mathematical) **expectation** of X , which is often denoted as $\mathbb{E}(X)$. 

Note that **NOT** all random variables have expectation. For example, the expectation of standard **Cauchy distribution**

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

is undefined.

The expectation is the center of the probability density function in many cases. What is more,


$$\arg \max_b \mathbb{E}(X - b)^2 = \mathbb{E}(X)$$

which implies that $\mathbb{E}(X)$ is the most likeliest to appear in expectation.


Definition 2.17: Expectation of Random vector

Now suppose that X is a random vector $X = (X_1, \dots, X_d)$, where $X_j, j = 1, \dots, d$, is real-valued. Then $\mathbb{E}(X)$ is simply the vector

$$(\mathbb{E}(X_1), \dots, \mathbb{E}(X_d))$$

where $\mathbb{E}(X_i) = \int_{\mathbb{R}} x_i f_{X_i} dx_i \forall i \in \{1, 2, \dots, d\}$ and $f_{X_i}(x_i)$ is the marginal probability density function of X_i . 

Definition 2.18: Variance

If the random variable has finite expectation $\mathbb{E}(X)$, then the expectation of $(X - \mathbb{E}(X))^2$ is called the variance of X (denoted as $Var(X)$), i.e. $\sum_{i=1}^{\infty} (x_i - \mathbb{E}(X))^2 P(X = x_i)$ for discrete random variable and $\int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f_X(x) dx$ for continuous random variable. 



It is obvious that

$$\text{Var}(X) = \int_{\mathbb{R}} (x - \mathbb{E}(X))^2 f_X(x) dx = \int_{\mathbb{X}} (x - \mathbb{E}(X))^2 f_X(x) dx$$

where the set \mathbb{X} is the support of X .

In general, let X be a random variable with pdf $f_X(x)$, the expectation of the random variable $g(X)$ is equal to the integration $\int_{\mathbb{R}} g(x) f_X(x) dx$ if it exists.

In analysis, the expectation of a random variable is the integration of some functions.

Definition 2.19: Conditional Expectation

The conditional expectation of Y given X can be defined as

$$\Psi(X) = \mathbb{E}(Y|X) = \int_{\mathbb{R}} y f_{Y|X}(y|x) dy.$$



It is a **parameter-dependent integral**, where the parameter x can be considered as fixed constant. See [the content in Calculus II](#).

Definition 2.20: Tower Rule

Tower Rule Let X be random variable on a sample space Ω , let the events A_1, A_2, \dots are disjoint in the sample space Ω and $\bigcup_{i=1}^{\infty} A_i = \Omega$, B is any subset of Ω , $\mathbb{E}(X) = \sum_{i=1}^{\infty} \mathbb{E}(X|A_i)P(A_i)$. In general, the conditional expectation $\Psi(X) = \mathbb{E}(Y|X)$ satisfies $\mathbb{E}(\Psi(X)) = \mathbb{E}(Y)$.



The probability of an event can be expressed as the expectation, i.e.

表 2.3: Probability as Integration

$$P(x \in A) = \int_{\mathbb{R}} \mathbb{I}_A f_X(x) dx$$

where

$$\mathbb{I}_A = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}. \quad (2.1)$$

It is called as the indicator function of A .

Definition 2.21: Entropy

The Shannon entropy is defined as

$$H = - \sum_{i=1}^{\infty} P(X = x_i) \ln P(X = x_i)$$

for the discrete random variable X with the range $\{x_1, \dots, x_n\}$ with the convention $P(X = x) = 0$ that $P(X = x) \ln \frac{1}{P(X=x)} = 0$ and

$$H = \int_{\mathbb{X}} \ln\left(\frac{1}{f_X(x)}\right) f_X(x) dx$$




for the continuous random variable X with the support \mathbb{X} .



The Shannon entropy is often related to *finite* state discrete random variable, i.e. $n < \infty$ and the value of random variable is not involved in its entropy.

表 2.4: Claude Elwood Shannon, 1916-2001



Information is the #resolution of
uncertainty.

— Claude Shannon —

AZ QUOTES

There is a **Visual information**
at <http://colah.github.io/posts/2015-09-Visual-Information/>.

2.4.2 Moment and Moment Generating Function

In calculus, we know that some proper functions can be extended as a Taylor series in a interval.

Moment is a specific quantitative measure of the shape of a function.

Definition 2.22: Moments

The n -th moment of a real-valued continuous function $f_X(x)$ of a real variable about a value c is

$$\mu_n = \int_{x \in \mathbb{R}} (x - c)^n f_X(x) dx.$$



And the constant c always take the expectation $\mathbb{E}(X)$ or 0.

Definition 2.23: Moment Generating Function

Moment generating function of a random variable X is the expectation of the random variable of e^{tX} , i.e.

$$M_X(t) = \mathbb{E}(e^{tX})$$



表 2.5: Moment generating function

$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$ for continuous random variable
$M_X(t) = \sum_{i=1}^{\infty} e^{tx_i} P(x = x_i)$ for discrete random variable

It is **Laplace transformation** applied to probability density function. And it is also related with cumulants <http://scholarpedia.org/article/Cumulants>.

The moments are not unique, i.e. the different distribution may have the same moments such as

$$f_1(x) = \frac{1}{\sqrt{2\pi x}} \exp(-\log(x)^2/2)$$

$$f_2(x) = f_1(x)(1 + \sin(2\pi \log x)).$$

2.5 Sampling Methods

2.5.1 Sampling from Discrete Distribution

Definition 2.24: Inverse transform technique

Let F be a probability cumulative function of a random variable taking non-negative integer values, and let U be uniformly distributed on interval $[0, 1]$. The random variable given by $X = k$ if and only if $F(k-1) < U < F(k)$ has the distribution function F . ♣

Sampling from Categorical Distribution

The *categorical distribution* (also called a *generalized Bernoulli distribution*, *multinoulli distribution*) is a discrete probability distribution that describes the possible results of a random variable that can take on one of K possible categories, with the probability of each category separately specified.

The category can be represented as **one-hot vector**, i.e. the vector $(1, 0, \dots, 0)$ is referred as the first category and the others are as similar as it. This representation is of some advantages:

1. Each category is identified by its position and equal to each other in norm;
2. The one-hot vectors can not be compared as the real number in value;
3. The probability mass function of the K categories distribution can be written in a compact form - $P(X) = [p_1, p_2, \dots, p_K] \cdot X^T$, where
 - The probability $p_i \geq 0, \forall i \in [1, \dots, K]$ and $\sum_{i=1}^K p_i = 1$;
 - The random variable X is one-hot vector and X^T is the transpose of X .

Sampling it by **inverse transform sampling**:

1. Pick a **uniformly distributed** number between 0 and 1.



2. Locate the greatest number in the CDF whose value is less than or equal to the number just chosen., i.e. $F^{-1}(u) = \inf\{x : F(x) \leq u\}$.
3. Return the category corresponding to this CDF value.

See more on [Categorical distribution](#).

The categorical variable cannot be ordered, how to compute the CDF?

2.5.2 Sampling from Continuous Distribution

Direct Methods

Inverse transform technique

Theorem 2.4: Inverse transform technique

Let F be a probability cumulative function of a continuous random variable, and let U be uniformly distributed on interval $[0, 1]$. The random variable $X = F^{-1}(U)$ has the distribution function F .



Khinchine's (1938) theorem: Suppose the random variable X with density given by

$$f_X(x) = \int_x^\infty z^{-1} f_Z(z) dz$$

or This mixture can be represented equivalently by

$$Z \sim f_Z(z), z > 0 \quad \text{and} \quad (1)$$

$$X|(Z = z) \sim \text{Unif}(0, z). \quad (2)$$

Hence $\frac{X}{Z}|(Z = z) = \frac{X}{z}|(Z = z) \sim \text{Unif}(0, 1)$ not depending on z , so that

$$\frac{X}{Z}|(Z = z) = \frac{X}{z}|(Z = z) \sim \text{Unif}(0, 1) \quad (2.2)$$

$$\frac{X}{Z} \stackrel{d}{=} U \sim \text{Unif}(0, 1) \quad (2.3)$$

$$X = ZU \quad (3)$$

and U and Z are mutually independent.

If $\nabla f_X(x)$ exists and $f_X(\infty) = 0$, we can obtain the following equation by Newton-Leibiniz's theorem

$$f_X(x) = - \int_x^\infty \nabla f_X(z) dz$$

and comparing (4) and (0), it is obvious: $\nabla f_X(z) = -z^{-1} f_Z(z)$.

See more information on [On Khinchine's Theorem and Its Place in Random Variate Generation](#) and [Reciprocal symmetry, unimodality and Khinchine's theorem](#).



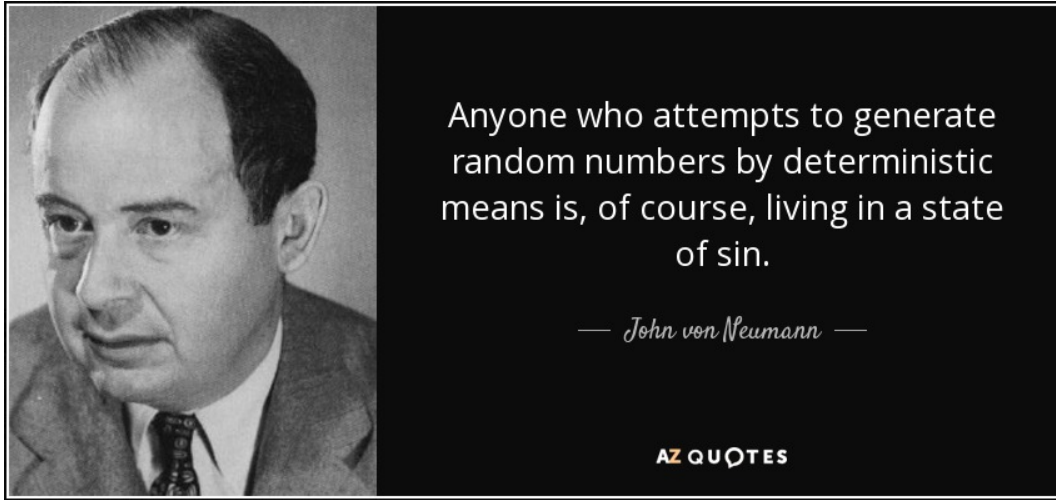
Rejection sampling

The **rejection sampling** is to draw samples from a distribution with the help of a proposal distribution.

1. Obtain a sample y from distribution Y and a sample u from $\text{Unif}(0, 1)$ (the uniform distribution over the unit interval). 2. Check whether or not $u < f(y)/Mg(y)$.

- If this holds, accept y as a sample drawn from f ;
- if not, reject the value of y and return to the sampling step.

The algorithm will take an average of M iterations to obtain a sample. **Ziggural algorithm** is an application of rejection sampling to draw samples from Gaussian distribution.



Sampling-importance-resampling

It is also called SIR method in short. As name shown, it consists of two steps: sampling and importance sampling. The SIR method generates an approximate i.i.d. sample of size m from the target density $f(x)$, $x \in \mathbb{X} \subset \mathbb{R}^n$.

The SIR without replacement is as following:

1. Draw $X^{(1)}, X^{(2)}, \dots, X^{(J)}$ independently from the proposal density $g(\cdot)$.
2. Select a subset $\{X^{(k_i)}\}_{i=1}^m$ from $\{X^{(i)}\}_{i=1}^J$ via resampling without replacement from the discrete distribution $\{\omega_i\}_{i=1}^J$ where $w_i = \frac{f(X^{(i)})}{g(X^{(i)})}$ and $\omega_i = \frac{w_i}{\sum_{i=1}^J w_i}$.

The Conditional Sampling Method

The conditional sampling method due to the prominent Rosenblatt transformation is particularly available when the joint distribution of a d vector is very difficult to generate but one marginal distribution and $d - 1$ univariate conditional distributions are easy to simulate.



It is based on the chain rule of probability:

$$f(x) = f(x_d) \prod_{i=1}^{d-1} f_k(x_k | x_{k+1}, x_{k+2}, \dots, x_d)$$

where $x = (x_1, x_2, \dots, x_d)^T$ and $f(x)$ is the probability density function.

To generate X from $f(x)$, we only need to generate x_d from the marginal density $f_d(x_d)$, then to generate x_k sequentially from the conditional density $f_k(x_k | x_{k+1}, x_{k+2}, \dots, x_d)$.



第 3 章 Numerical Optimization



IN *A Few Useful Things to Know about Machine Learning*, Pedro Domingos put up a relation:

LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION.

- Representation as the core of the note is the general (mathematical) **model** that computer can handle.
- Evaluation is **criteria**. An evaluation function (also called objective function, cost function or scoring function) is needed to distinguish good classifiers from bad ones.
- Optimization is to aimed to find the parameters that optimizes the evaluation function, i.e.

$$\arg \min_{\theta} f(\theta) = \{\theta^* | f(\theta^*) = \min f(\theta)\} \text{ or } \arg \max_{\theta} f(\theta) = \{\theta^* | f(\theta^*) = \max f(\theta)\}.$$

The objective function to be minimized is also called cost function.

Evaluation is always attached with optimization; the evaluation which cannot be optimized is not a good evaluation in machine learning.

3.1 Gradient Descent and More

Each iteration of a line search method computes a search direction p^k and then decides how far to move along that direction. The iteration is given by

$$x^{k+1} = x^k + \alpha_k p^k$$

where the positive scalar α^k is called the step length. The success of a line search method depends on effective choices of both the direction p^k and the step length α_k .

Note: we use the notation x^k and α_k to represent the k th iteration of the vector variables x and k th step length, respectively. Most line search algorithms require p^k to be a descent direction — one for which $\langle p^k, \nabla f_k \rangle < 0$ — because this property guarantees that the function f can be reduced along this direction, where ∇f_k is the gradient of objective function f at the k th iteration point x_k i.e. $\nabla f_k = \nabla f(x^k)$.

Gradient descent and its variants are to find the local solution of the unconstrained optimization problem:

$$\min f(x)$$

where $x \in \mathbb{R}^n$.

Its iterative procedure is:

$$x^{k+1} = x^k - \alpha_k \nabla_x f(x^k)$$

where x^k is the k th iterative result, $\alpha_k \in \{\alpha | f(x^{k+1}) < f(x^k)\}$ and particularly $\alpha_k = \arg \min_{\alpha} \{f(x^k - \alpha \nabla_x f(x^k))\}$.

Some variants of gradient descent methods are not line search method. For example, the **heavy ball method**:

$$x^{k+1} = x^k - \alpha_k \nabla_x f(x^k) + \rho_k (x^k - x^{k-1})$$

where the momentum coefficient $\rho_k \in [0, 1]$ generally and the step length α_k cannot be determined by line search.

Nesterov accelerated gradient method is defined by

$$x^k = y^k - \alpha^{k+1} \nabla_x f(y^k) \quad \text{Descent} \quad (3.1)$$

$$y^{k+1} = x^k + \rho^k (x^k - x^{k-1}) \quad \text{Momentum} \quad (3.2)$$

where the momentum coefficient $\rho_k \in [0, 1]$ generally.

3.2 Mirror Gradient Method

It is often called **mirror descent**. It can be regarded as non-Euclidean generalization of **projected gradient descent** to solve some constrained optimization problems.

3.2.1 Projected Gradient Descent

Projected gradient descent is aimed to solve convex optimization problem with explicit constraints, i.e.

$$\arg \min_{x \in \mathbb{S}} f(x)$$

where $\mathbb{S} \subset \mathbb{R}^n$. It has two steps:

$$z^{k+1} = x^k - \alpha_k \nabla_x f(x^k) \quad \text{Gradient descent} \quad (3.3)$$

$$x^{k+1} = Proj_{\mathbb{S}}(z^{k+1}) = \arg \min_{x \in \mathbb{S}} \|x - z^{k+1}\|^2 \quad \text{Projection} \quad (3.4)$$



表 3.1: Inventor of Nesterov accelerated Gradient



3.2.2 Mirror Descent

Mirror descent can be regarded as the non-Euclidean generalization via replacing the ℓ_2 norm or Euclidean distance in projected gradient descent by **Bregman divergence**.

Bregman divergence is induced by convex smooth function f :

$$B(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

where $\langle \cdot, \cdot \rangle$ is inner product. Especially, when f is quadratic function, the Bregman divergence induced by f is

$$B(x, y) = x^2 - y^2 - \langle 2y, x - y \rangle = x^2 + y^2 - 2xy = (x - y)^2$$

i.e. the Euclidean distance. A wonderful introduction to Bregman divergence is **Meet the Bregman Divergences** by **Mark Reid** at <http://mark.reid.name/blog/meet-the-bregman-divergences.html>. It is given by:

$$z^{k+1} = x^k - \alpha_k \nabla_x f(x^k) \quad \text{Gradient descent} \quad (3.5)$$

$$x^{k+1} = \arg \min_{x \in \mathbb{S}} B(x, z^{k+1}) \quad \text{Bregman projection} \quad (3.6)$$

One special method is called **entropic mirror descent** when $f = e^x$ and \mathbb{S} is simplex. See more on the following link list.



3.3 Variable Metric Methods

3.3.1 Newton's Method

NEWTON'S METHOD and QUASI-NEWTON METHODS are classified to variable metric methods.

It is also to find the solution of unconstrained optimization problems, i.e.

$$\min f(x)$$

where $x \in \mathbb{R}^n$.

Newton's method is given by

$$x^{k+1} = x^k - \alpha^{k+1} H^{-1}(x^k) \nabla_x f(x^k)$$

where $H^{-1}(x^k)$ is inverse of the Hessian matrix of the function $f(x)$ at the point x^k . It is called **Newton–Raphson algorithm** in statistics. Especially when the log-likelihood function $\ell(\theta)$ is well-behaved, a natural candidate for finding the MLE is the Newton–Raphson algorithm with quadratic convergence rate.

3.3.2 The Fisher Scoring Algorithm

In maximum likelihood estimation, the objective function is the log-likelihood function, i.e.

$$\ell(\theta) = \sum_{i=1}^n \log P(x_i|\theta)$$

where $P(x_i|\theta)$ is the probability of realization $X_i = x_i$ with the unknown parameter θ . However, when the sample random variable $\{X_i\}_{i=1}^n$ are not observed or realized, it is best to replace negative Hessian matrix (i.e. $-\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}$) of the likelihood function with the **observed information matrix**:

$$J(\theta) = \mathbb{E}\left(-\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}\right) = - \int \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} f(x_1, \dots, x_n|\theta) dx_1 \cdots dx_n$$

where $f(x_1, \dots, x_n|\theta)$ is the joint probability density function of X_1, \dots, X_n with unknown parameter θ .

And the **Fisher scoring algorithm** is given by

$$\theta^{k+1} = \theta^k + \alpha_k J^{-1}(\theta^k) \nabla_{\theta} \ell(\theta^k)$$

where $J^{-1}(\theta^k)$ is the inverse of observed information matrix at the point θ^k .

See <http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/scoring.pdf> or <https://wiseodd.github.io/techblog/2018/03/11/fisher-information/> for more information.

Fisher scoring algorithm is regarded as an example of **Natural Gradient Descent** in information geometry such as <https://wiseodd.github.io/techblog/2018/03/14/natural-gradient/>.



3.3.3 Quasi-Newton Methods

Quasi-Newton methods, like steepest descent, require only the gradient of the objective function to be supplied at each iterate. By measuring the changes in gradients, they construct a model of the objective function that is good enough to produce superlinear convergence. The improvement over steepest descent is dramatic, especially on difficult problems. Moreover, since second derivatives are not required, quasi-Newton methods are sometimes more efficient than Newton's method.

In optimization, quasi-Newton methods (a special case of *variable-metric methods*) are algorithms for finding local maxima and minima of functions. Quasi-Newton methods are based on Newton's method to find the stationary point of a function, where the gradient is 0. In quasi-Newton methods the Hessian matrix does not need to be computed. The Hessian is updated by analyzing successive gradient vectors instead. Quasi-Newton methods are a generalization of the secant method to find the root of the first derivative for multidimensional problems. In multiple dimensions the secant equation is under-determined, and quasi-Newton methods differ in how they constrain the solution, typically by adding a simple low-rank update to the current estimate of the Hessian. One of the chief advantages of quasi-Newton methods over Newton's method is that the Hessian matrix (or, in the case of quasi-Newton methods, its approximation) B does not need to be inverted. The Hessian approximation B is chosen to satisfy

$$\nabla f(x^{k+1}) = \nabla f(x^k) + B(x^{k+1} - x^k),$$

which is called the **secant equation** (the Taylor series of the gradient itself). In more than one dimension B is underdetermined. In one dimension, solving for B and applying the Newton's step with the updated value is equivalent to the [secant method]https://www.wikiwand.com/en/Secant_method. The various quasi-Newton methods differ in their choice of the solution to the *secant equation* (in one dimension, all the variants are equivalent).

3.3.4 Natural Gradient Descent

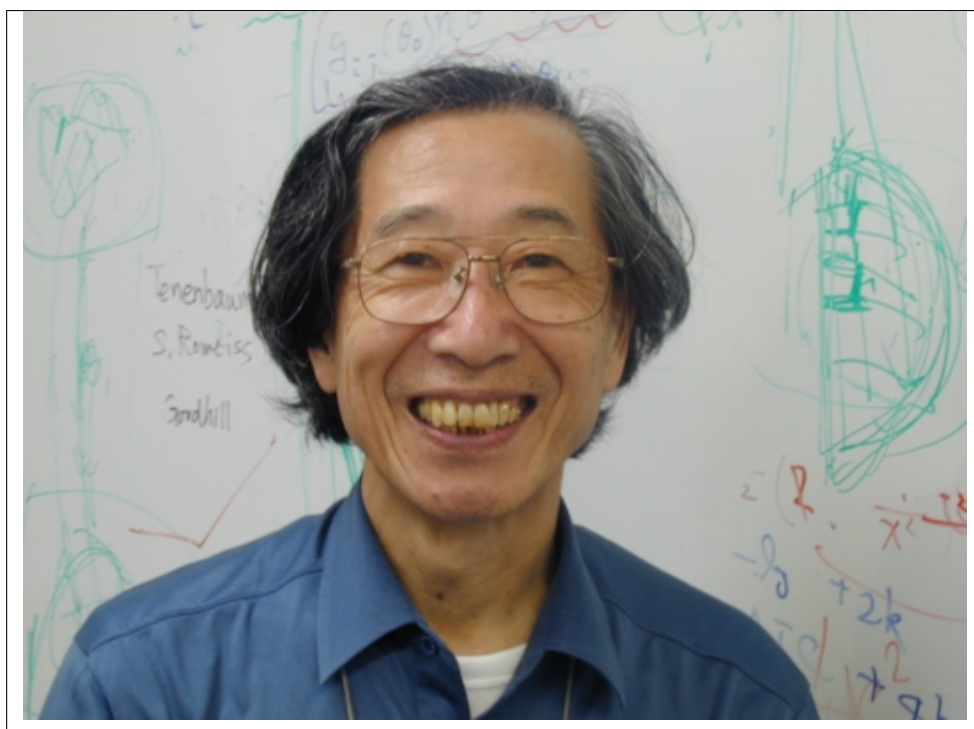
Natural gradient descent is to solve the optimization problem $\min_{\theta} L(\theta)$ by

$$\theta^{(t+1)} = \theta^{(t)} - \alpha_{(t)} F^{-1}(\theta^{(t)}) \nabla_{\theta} L(\theta^{(t)})$$

where $F^{-1}(\theta^{(t)})$ is the inverse of Riemann metric at the point $\theta^{(t)}$. And Fisher scoring algorithm is a typical application of Natural Gradient Descent to statistics. *Natural gradient descent* for manifolds corresponding to exponential families can be implemented as a first-order method through mirror descent <https://www.stat.wisc.edu/~raskutti/publication/MirrorDescent.pdf>.



表 3.2: Originator of Information Geometry



3.4 Expectation Maximization Algorithm

Expectation-Maximization algorithm, popularly known as the EM algorithm has become a standard piece in the statistician's repertoire. It is used in incomplete-data problems or latent-variable problems such as Gaussian mixture model in maximum likelihood estimation. The basic principle behind the EM is that instead of performing a complicated optimization, one augments the observed data with latent data to perform a series of simple optimizations.

Let $\ell(\theta|Y_{obs}) \triangleq \log L(\theta|Y_{obs})$ denote the log-likelihood function of observed datum Y_{obs} . We augment the observed data Y_{obs} with latent variables Z so that both the complete-data log-likelihood $\ell(\theta|Y_{obs}, Z)$ and the conditional predictive distribution $f(z|Y_{obs}, \theta)$ are available. Each iteration of the EM algorithm consists of an expectation step (E-step) and a maximization step (M-step). Specifically, let $\theta^{(t)}$ be the current best guess at the MLE $\hat{\theta}$. The E-step is to compute the **Q** function defined by

$$Q(\theta|\theta^{(t)}) = \mathbb{E}(\ell(\theta|Y_{obs}, Z)|Y_{obs}, \theta^{(t)}) \quad (3.7)$$

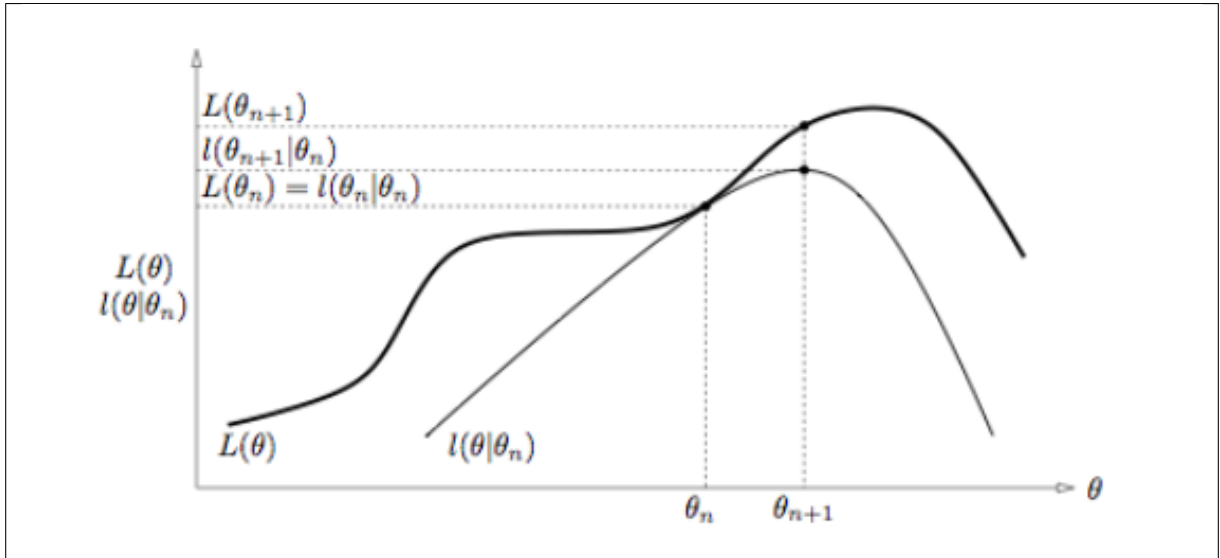
$$= \int_Z \ell(\theta|Y_{obs}, Z) \times f(z|Y_{obs}, \theta^{(t)}) dz, \quad (3.8)$$

and the M-step is to maximize **Q** with respect to θ to obtain

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta|\theta^{(t)}).$$



表 3.3: Diagram of EM algorithm



3.4.1 Generalized EM Algorithm

Each iteration of the **generalized EM** algorithm consists of an expectation step (E-step) and an ascent step instead of maximization step (M-step). Specifically, let $\theta^{(t)}$ be the current best guess at the MLE $\hat{\theta}$. The E-step is to compute the **Q** function defined by

$$Q(\theta|\theta^{(t)}) = \mathbb{E}(\ell(\theta|Y_{obs}, Z)|Y_{obs}, \theta^{(t)}) \quad (3.9)$$

$$= \int_Z \ell(\theta|Y_{obs}, Z) \times f(z|Y_{obs}, \theta^{(t)}) dz, \quad (3.10)$$

and the another step is to find θ that satisfies $Q(\theta^{t+1}|\theta^t) > Q(\theta^t|\theta^t)$, i.e.

$$\theta^{(t+1)} \in \{\hat{\theta} | Q(\hat{\theta}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})\}.$$

It is not to maximize the conditional expectation.

See more on the book **The EM Algorithm and Extensions, 2nd Edition by Geoffrey McLachlan, Thriyambakam Krishna** at <https://www.wiley.com/en-cn/The+EM+Algorithm+and+Extensions,+2nd+Edition-p-9780471201700>.

See more on <https://www.stat.berkeley.edu/~aldous/Colloq/lange-talk.pdf>

3.5 Alternating Direction Method of Multipliers

Alternating direction method of multipliers is called **ADMM** shortly. It is aimed to solve the following convex optimization problem:

$$\min F(x, y) \{ = f(x) + g(y) \} \quad (\text{cost function})$$

$$Ax + By = b \quad (\text{constraint})$$



where $f(x)$ and $g(y)$ are convex; A and B are matrices.

Define the augmented Lagrangian:

$$L_\mu(x, y) = f(x) + g(y) + \lambda^T(Ax + By - b) + \frac{\mu}{2}\|Ax + By - b\|_2^2.$$

It is iterative procedure at k th step:

1. $x^{k+1} = \arg \min_x L_\mu(x, y^k, \lambda^k);$
2. $y^{k+1} = \arg \min_y L_\mu(x^{k+1}, y, \lambda^k);$
3. $\lambda^{k+1} = \lambda^k + \mu(Ax^{k+1} + By^{k+1} - b).$

Algorithm 1 ADMM

Require: the initial points: x_0, y_0, λ_0

Ensure: $k = 0$

while $k \neq N$ **do**

$$x^{k+1} = \arg \min_x L_\mu(x, y^k, \lambda^k)$$

$$y^{k+1} = \arg \min_y L_\mu(x^{k+1}, y, \lambda^k)$$

$$\lambda^{k+1} = \lambda^k + \mu(Ax^{k+1} + By^{k+1} - b)$$

$$k \leftarrow k + 1$$

end while



Note: Thanks to *Professor He Bingsheng* who taught me this.

3.6 Stochastic Gradient Descent

Stochastic gradient descent takes advantages of stochastic or estimated gradient to replace the true gradient in gradient descent. It is *stochastic gradient* but may not be *descent*. The name **stochastic gradient methods** may be more appropriate to call the methods with stochastic gradient. It can date back upto **stochastic approximation**.

It is aimed to solve the problem with finite sum optimization problem, i.e.

$$\arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n f(\theta|x_i)$$

where $n < \infty$ and $\{f(\theta|x_i)\}_{i=1}^n$ are in the same function family and $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ are constants while $\theta \in \mathbb{R}^p$ is the variable vector.

The difficulty is that p , that the dimension of θ , is tremendous. In other words, the model is **overparameterized**. And the number n is far larger than p generally, i.e. $n \gg p \gg d$. What is worse, the functions $\{f(\theta|x_i)\}_{i=1}^n$ are not convex in most case.

The stochastic gradient method is defined as

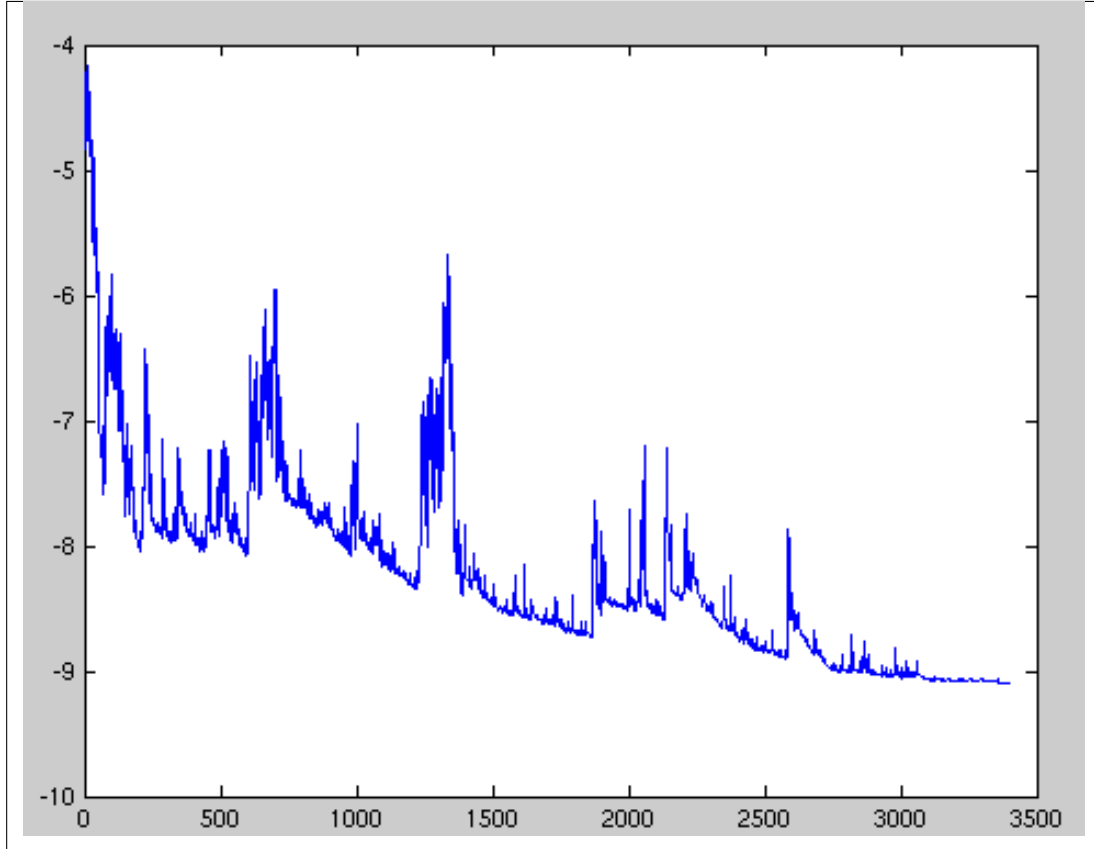
$$\theta^{k+1} = \theta^k - \alpha_k \frac{1}{m} \sum_{j=1}^m \nabla f(\theta^k|x'_j)$$



where x'_j is stochastically draw from $\{x_i\}_{i=1}^n$ and $m \ll n$.

It is the fact $m \ll n$ that makes it possible to compute the gradient of finite sum objective function and its side effect is that the objective function is not always descent. There is fluctuations in the total objective function as gradient steps with respect to mini-batches are taken.

表 3.4: The fluctuations in the objective function as gradient with respect to mini-batches are taken



An heuristic proposal for avoiding the choice and for modifying the learning rate while the learning task runs is the bold driver (BD) method. The learning rate increases exponentially if successive steps reduce the objective function f , and decreases rapidly if an “accident” is encountered (if objective function f increases), until a suitable value is found. After starting with a small learning rate, its modifications are described by the following equation:

$$\alpha_{k+1} = \begin{cases} \rho \alpha_k, & f(\theta^{k+1}) < f(\theta^k); \\ \eta^n \alpha_k, & f(\theta^{k+1}) > f(\theta^k) \text{ using } \alpha_k, \end{cases}$$

where ρ is close to 1 such as $\rho = 1.1$ in order to avoid frequent “accidents” because the objective function computation is wasted in these cases, η is chosen to provide a rapid reduction ($\eta = 0.5$), and n is the minimum integer such that the reduced rate η^n succeeds in diminishing the objective function.



The fact that the sample size is far larger than the dimension of parameter, $n \gg p$, that makes it expensive to compute total objective function $f(\theta) = \sum_{i=1}^n f(\theta|x_i)$. Thus it is not clever to determine the learning rate α_k by line search. And most stochastic gradient methods are to find proper step length α_k to make it converge at least in convex optimization. The variants of gradient descent such as momentum methods or mirror gradient methods have their stochastic counterparts.

- It is simplest to set the step length a constant, such as $\alpha_k = 3 \times 10^{-3} \forall k$.
- There are decay schemes, i.e. the step length α_k diminishes such as $\alpha_k = \frac{\alpha}{k}$, where α is constant.
- And another strategy is to tune the step length adaptively such as **AdaGrad**, **ADAM**.

PS: the step length α_k is called ****learning rate**** in machine learning and stochastic gradient descent is also named as *incremental gradient descent method* such as the survey at http://www.mit.edu/~dimitrib/Incremental_Survey_LIDS.pdf in some case.

See the following links for more information on *stochastic gradient descent*.

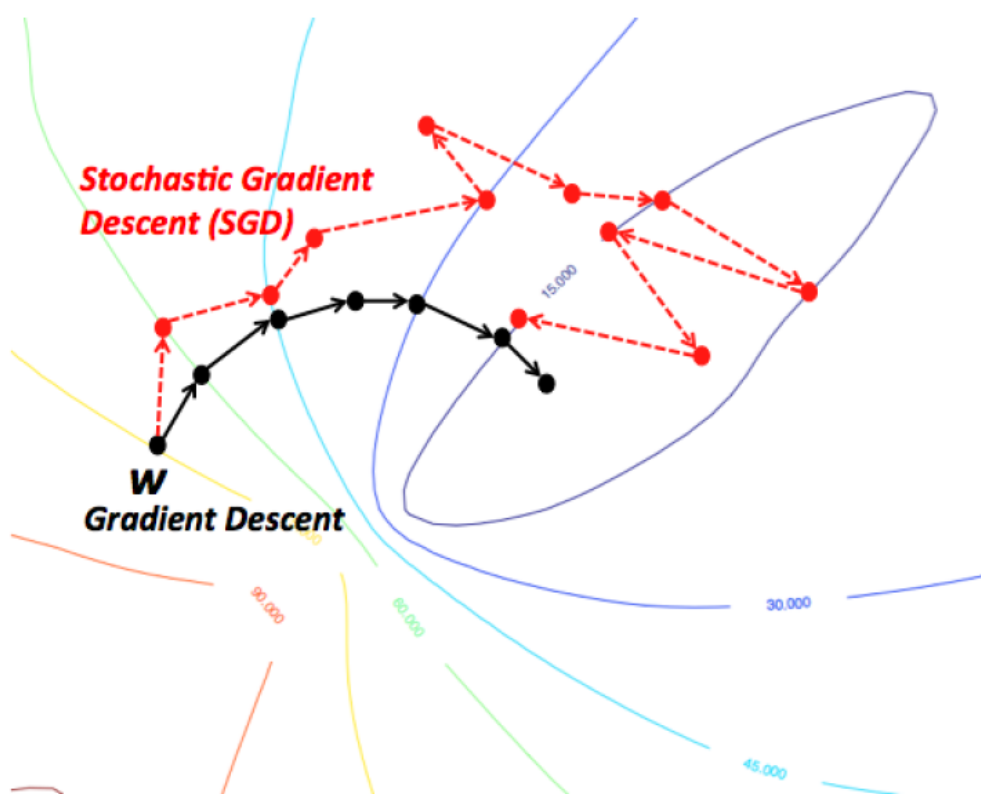


图 3.1: The Differences of Gradient Descent and Stochastic Gradient Descent



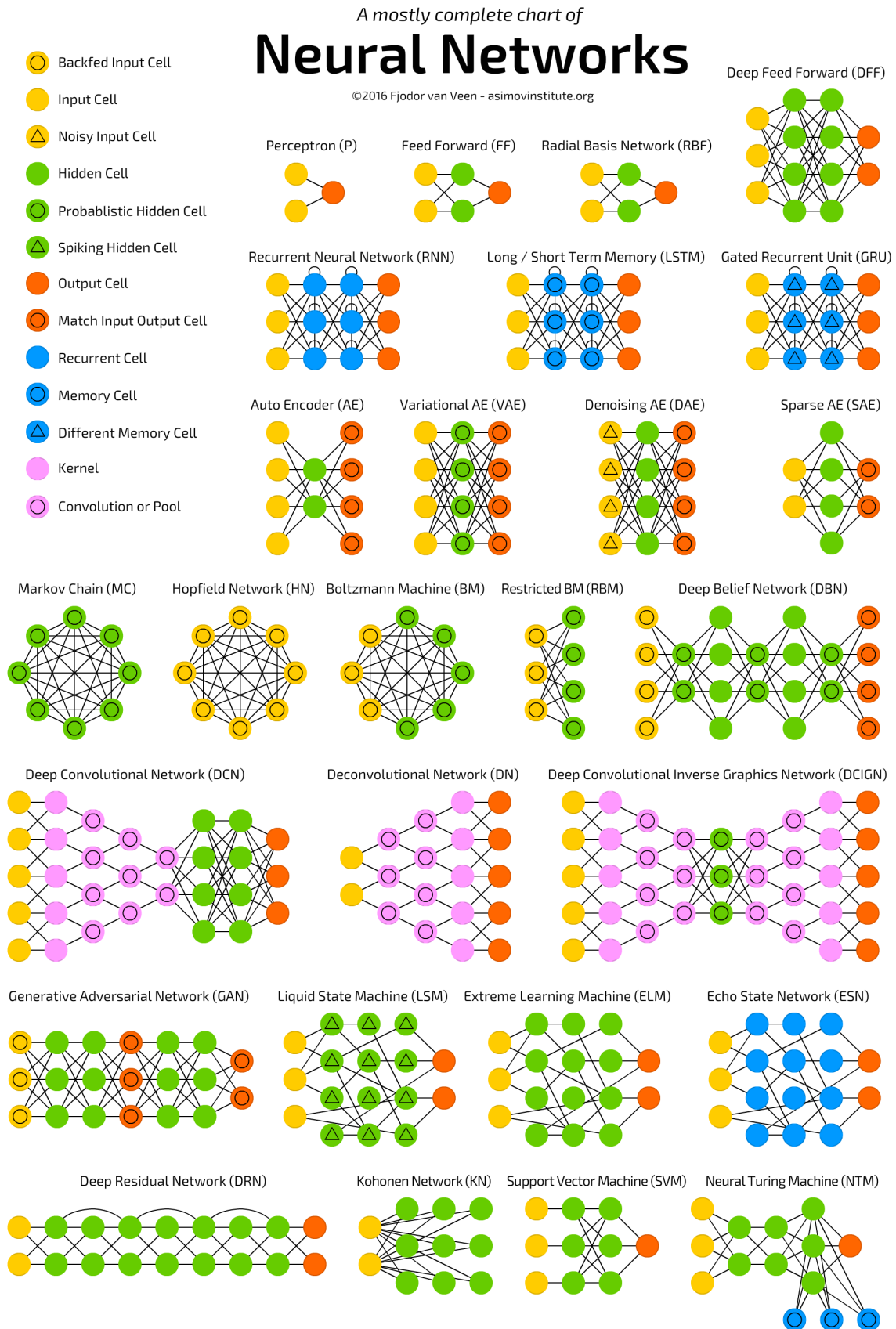


图 3.2: Neural Networks



第 4 章 ElegantBook 设置说明



4.1 编译方式

本模板基于基础的 book 文类，所以 book 的选项对于本模板也是有效的。默认编码为 UTF-8，推荐使用 T_EX Live 编译。作者编写环境为 Win10(64bit) + T_EX Live 2018。由于使用的是 ctex 宏包，所以支持 pdf_latex 以及 xel_atex 编译。

4.2 选项设置

本文特殊选项设置共有 2 类，分为 主题颜色设置 以及 章标题显示风格设置。

第 1 类为 主题颜色设置，内置 3 组颜色主题，分别为 green（默认），cyan，blue，另外还有一个自定义的选项 nocolor，用户 必须 在使用模板的时候选择某个颜色主题或选择 nocolor 选项。调用颜色主题 green 的方法为 `\documentclass[green]{elegantbook}` 或者使用 `\documentclass[color=green]{elegantbook}`。需要改变颜色的话请选择 nocolor 选项或者使用 `color=none`，然后在导言区定义 main、second、third 颜色，具体的方法如下：

```
\definecolor{main}{RGB}{70,70,70}    %定义 main 颜色值
\definecolor{second}{RGB}{115,45,2}   %定义 second 颜色值
\definecolor{third}{RGB}{0,80,80}     %定义 third 颜色值
```










	green	cyan	blue	主要使用的环境
main				definition
second				theorem lemma corollary
third				proposition

表 4.1: ElegantBook 模板中的三套颜色主题

第 2 类为 章标题显示风格，包含 hang（默认）与 display 两种风格，区别在于章标题单行显示（hang）与双行显示（display），本说明使用了 hang。调用方式为

`\documentclass[hang]{elegantbook}` 或者 `\documentclass[titlestyle=hang]{elegantbook}`。

综合起来，同时调用三个选项使用 `\documentclass[color=X,titlestyle=Y]{elegantbook}`。
其中 `X` 可以选择 `green,cyan,blue,none`；`Z` 可以选择 `hang` 或者 `display`。

4.3 数学环境简介

在我们这个模板中，定义了三大类环境

1. 定理类环境，包含标题和内容两部分。根据格式的不同分为3种
 - **definition** 环境，含有一个可选项，编号以章节为单位，颜色为 `main`；
 - **theorem**、**lemma**、**corollary** 环境，颜色为主颜色 `second`，编号均以章节为单位；
 - **proposition** 环境，含有一个可选项，编号以章节为单位，颜色为 `third`。
2. 示例类环境，有 **example**、**exercise** 环境，自动编号，编号以章节为单位。
3. 证明类环境，有 **proof**、**note** 环境，特点是，有引导符或者结尾符，**note** 环境有引导符号，**proof** 环境有证明完毕标志。
4. 结论类环境，有 **conclusion**、**assumption**、**property**，**remark**、**solution** 环境，三者均以粗体的引导词为开头，和普通段落格式一致。

4.4 可编辑的字段

在模板中，可以编辑的字段分别为作者 `\author`、邮箱 `\email`、中文标题 `\zhtitle`、中文标题结尾 `\zhend`、英文标题 `\entitle`、英文标题结尾 `\enend`、名言 `\myquote`、版本号 `\version`。并且，可以根据自己的喜好把封面水印效果的 `cover.pdf` 替换掉，以及封面中用到的 `logo.pdf`。



第 5 章 ElegantBook 写作示例



5.1 Economics and Differentiable Function

Economics focuses on the behaviour and interactions of economic agents and how economies work. Microeconomics analyzes basic elements in the economy, including individual agents and markets, their interactions, and the outcomes of interactions. Individual agents may include, for example, households, firms, buyers, and sellers. Macroeconomics analyzes the entire economy (meaning aggregated production, consumption, savings, and investment) and issues affecting it, including unemployment of resources (labour, capital, and land), inflation, economic growth, and the public policies that address these issues (monetary, fiscal, and other policies). See glossary of economics.

$$\begin{aligned} \max(\min) \quad & \mathbb{E} \int_{t_0}^{t_1} f(t, x, u) dt \\ \text{s.t.} \quad & dx = g(t, x, u)dt + \sigma(t, x, u)dz \\ & k(0) = k_0 \text{ given} \end{aligned}$$

where z is stochastic process or white noise or wiener process.

Other broad distinctions within economics include those between positive economics, describing "what is", and normative economics, advocating "what ought to be"; between economic theory and applied economics; between rational and behavioural economics; and between mainstream economics and heterodox economics.

Economic analysis can be applied throughout society, in business, finance, health care, and government. Economic analysis is sometimes also applied to such diverse subjects as crime, education, the family, law, politics, religion, social institutions, war, science, and the environment.

Definition 5.1: Differenzierbarkeit

Eine Funktion $f : I \rightarrow \mathbb{R}$ auf einem Intervall I heit in $x_0 \in I$ differenzierbar oder linear approximierbar, wenn der Grenzwert

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

existiert. Bei Existenz heit dieser Grenzwert Ableitung oder Differentialquotient von

f in x_0 und man schreibt für ihn

$$f'(x_0) \quad \text{oder} \quad \frac{df}{dx}(x_0).$$



The discipline was renamed in the late 19th century, primarily due to Alfred Marshall, from "political economy" to "economics" as a shorter term for "economic science". At that time, it became more open to rigorous thinking and made increased use of mathematics, which helped support efforts to have it accepted as a science and as a separate discipline outside of political science and other social sciences.

Example 5.1: E and F be two events such that $P(E) = P(F) = 1/2$, and $P(E \cap F) = 1/3$, let $\mathcal{F} = \sigma(Y)$, X and Y be the indicate function of E and F respectively. How to compute $\mathbb{E}[X \mid \mathcal{F}]$?

Some subsequent comments criticized the definition as overly broad in failing to limit its subject matter to analysis of markets. From the 1960s, however, such comments abated as the economic theory of maximizing behaviour and rational-choice modelling expanded the domain of the subject to areas previously treated in other fields. There are other criticisms as well, such as in scarcity not accounting for the macroeconomics of high unemployment.

Exercise 5.1: let $S = l^\infty = \{(x_n) \mid \exists M \text{ such that } \forall n, |x_n| \leq M, x_n \in \mathbb{R}\}$, $\rho_\infty(x, y) = \sup_{n \geq 1} |x_n - y_n|$, show that (l^∞, ρ_∞) is complete.

Theorem 5.1: Mittelwertsatz für n Variable

Es sei $n \in \mathbb{N}$, $D \subseteq \mathbb{R}^n$ eine offene Menge und $f \in C^1(D, \mathbb{R})$. Dann gibt es auf jeder Strecke $[x_0, x] \subset D$ einen Punkt $\xi \in [x_0, x]$, so dass gilt

$$f(x) - f(x_0) = \text{grad } f(\xi)^\top (x - x_0)$$



Note: 在本模板中，引理 (lemma)，推论 (corollary) 的样式和定理的样式一致，包括颜色，仅仅只有计数器的设置不一样。在这个例稿中，我们将不给出引理推论的例子。

Gary Becker, a contributor to the expansion of economics into new areas, describes the approach he favours as "combin[ing the] assumptions of maximizing behaviour, stable preferences, and market equilibrium, used relentlessly and unflinchingly." One commentary characterizes the remark as making economics an approach rather than a subject matter but with great specificity as to the "choice process and the type of social interaction that [such] analysis involves."

Economic efficiency measures how well a system generates desired output with a given set of inputs and available technology. Efficiency is improved if more output is generated without changing inputs, or in other words, the amount of "waste" is reduced. A widely



accepted general standard is Pareto efficiency, which is reached when no further change can make someone better off without making someone else worse off.

Proposition 5.1: 最优性原理

如果 u^* 在 $[s, T]$ 上为最优解, 则 u^* 在 $[s, T]$ 任意子区间都是最优解, 假设区间为 $[t_0, t_1]$ 的最优解为 u^* , 则 $u(t_0) = u^*(t_0)$, 即初始条件必须还是在 u^* 上。

Microeconomics examines how entities, forming a market structure, interact within a market to create a market system. These entities include private and public players with various classifications, typically operating under scarcity of tradable units and light government regulation. The item traded may be a tangible product such as apples or a service such as repair services, legal counsel, or entertainment.

Corollary 5.1

假设 $V(\cdot, \cdot)$ 为值函数, 则根据最大值原理 5.1, 有如下推论

$$V(k, z) = \max \left\{ u(zf(k) - y) + \beta \mathbb{E}V(y, z') \right\}$$

Proof: 因为 $y^* = \alpha\beta zk^\alpha$, $V(k, z) = \alpha/1 - \alpha\beta \ln k_0 + 1/1 - \alpha\beta \ln z_0 + \Delta$ 。

$$\begin{aligned} \text{右边} &= \left\{ u(zf(k) - y) + \beta \mathbb{E}V(y, z') \right\} \\ &= \ln(zk^\alpha - \alpha\beta zk^\alpha) + \beta \mathbb{E} \left[\frac{\alpha}{1 - \alpha\beta} \ln y + \frac{1}{1 - \alpha\beta} \ln z' + \Delta \right] \\ &= \ln(1 - \alpha\beta)zk^\alpha + \beta \left\{ \mathbb{E} \left[\frac{\alpha}{1 - \alpha\beta} \ln \alpha\beta zk^\alpha \right] + \frac{1}{1 - \alpha\beta} \mathbb{E}[\ln z'] + \Delta \right\} \end{aligned}$$

利用 $\mathbb{E}[\ln z'] = 0$, 并将对数展开得

$$\begin{aligned} \text{右边} &= \ln(1 - \alpha\beta) + \ln z + \alpha \ln k + \frac{\alpha\beta}{1 - \alpha\beta} [\ln \alpha\beta + \ln z + \alpha \ln k] + \frac{\beta}{1 - \alpha\beta} \mu + \beta\Delta \\ &= \frac{\alpha}{1 - \alpha\beta} \ln k + \frac{1}{1 - \alpha\beta} \ln z + \Delta \end{aligned}$$

所以 左边 = 右边, 证毕。

□

Properties: Properties of Cauchy Sequence

1. $\{x_k\}$ is cauchy sequence then $\{x_k^i\}$ is cauchy sequence.
2. $x_k \in \mathbb{R}^n$, $\rho(x, y)$ is Euclidean, then cauchy is equivalent to convergent, (\mathbb{R}^n, ρ) metric space is complete.



Note: conclusion、assumption、property、remark、solution 的环境效果是一样的。

Various market structures exist. In perfectly competitive markets, no participants are large enough to have the market power to set the price of a homogeneous product. In other words, every participant is a "price taker" as no participant influences the price of a product. In the real world, markets often experience imperfect competition.



Scarcity is represented in the figure by people being willing but unable in the aggregate to consume beyond the PPF (such as at X) and by the negative slope of the curve.[32] If production of one good increases along the curve, production of the other good decreases, an inverse relationship. This is because increasing output of one good requires transferring inputs to it from production of the other good, decreasing the latter.

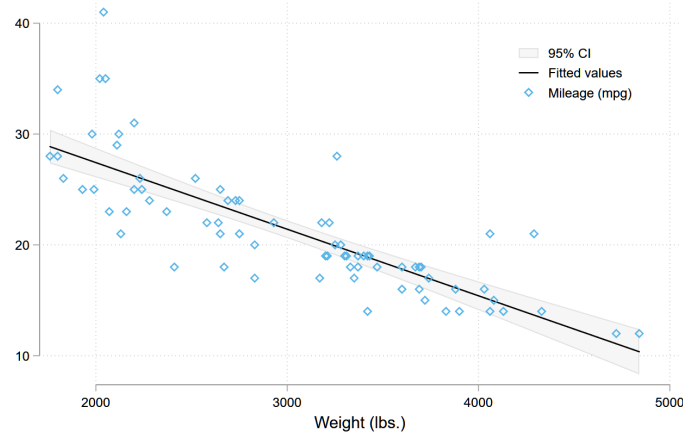


图 5.1: The Relationship between MPG and Weight

Definition 5.2: Contraction mapping

(S, ρ) is the metric space, $T : S \rightarrow S$, If there exists $\alpha \in (0, 1)$ such that for any x and $y \in S$, the distance

$$\rho(Tx, Ty) \leq \alpha \rho(x, y) \quad (5.1)$$

Then T is a **contraction mapping**. ♣

Remarks:

1. $T : S \rightarrow S$, where S is a metric space, if for any $x, y \in S$, $\rho(Tx, Ty) < \rho(x, y)$ is not contraction mapping.
2. Contraction mapping is continuous map.

Conclusions: In theory, in a free market the aggregates (sum of) of quantity demanded by buyers and quantity supplied by sellers may reach economic equilibrium over time in reaction to price changes; in practice, various issues may prevent equilibrium, and any equilibrium reached may not necessarily be morally equitable. For example, if the supply of healthcare services is limited by external factors, the equilibrium price may be unaffordable for many who desire it but cannot pay for it.



5.2 Bibliography

This template uses Bib_TE_X to generate the bibliography, the default bibliography style is aer. ? use data from a major peer-to-peer lending marketplace in China to study whether female and male investors evaluate loan performance differently. You can add bib items (from Google Scholar, Mendeley, EndNote, and etc.) to reference.bib file, and cite the bibkey in the tex file.

