CHAPTER 2

Preliminaries

In this chapter, we provided some necessary notations, definitions, and the main theory that was used throughout this thesis.

2.1 Basic Probability Theory

A process that has random outcomes is called a *random experiment*. The set of all possible outcomes of a random experiment is called the *sample space* and is denoted Ω. A combination of outcomes, a subset of Ω, is called an *event*. The set of events is denoted $\mathcal{F}$. It is assumed that the set $\mathcal{F}$ is a *σ-algebra*.

Definition 2.1.1. ([3]) Let Ω be a set, then $\mathcal{F}$ is a *σ-algebra* of subsets of Ω if $\mathcal{F}$ has the following properties:

1. $\Omega \in \mathcal{F}$,
2. $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$,
3. If $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The experiment selects an outcome in Ω according to a *probability measure* P, which is a set function that maps $\mathcal{F}$ into $\mathbb{R}$.

Definition 2.1.2. ([3]) Let Ω be a sample space and $\mathcal{F}$ be a σ-algebra of subsets of Ω. A set function $P : \mathcal{F} \to \mathbb{R}$ is a *probability measure* if the following conditions hold

1. $P(A) \geq 0$ for every $A \in \mathcal{F}$,
2. $P(\Omega) = 1$,
3. If $\{A_i\}_{i=1}^{\infty}$ is a sequence of mutually exclusive events in $\mathcal{F}$, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

A triplet $(\Omega, \mathcal{F}, P)$ consisting of the sample space Ω, the σ-algebra of subsets of Ω, and a probability measure P defined on $\mathcal{F}$ is called a *probability space*.

Definition 2.1.3. ([3]) Let $(\Omega, \mathcal{F}, P)$ be a probability space, A *random variable* X is a real-valued function that assigns the value $X(\omega) \in \mathbb{R}$ to each outcome $\omega \in \Omega$. That is, $X : \Omega \to \mathbb{R}$.

Definition 2.1.4. ([3]) Let $X : \Omega \to \mathbb{R}$ be a random variable. The *distribution function* of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

Definition 2.1.5. ([3]) Let $X : \Omega \to \mathbb{R}$ be a random variable and F_X its distribution function. If there exists a nonnegative integrable function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds,$$

then f_X is called the *probability density function* of X.

Definition 2.1.6. ([3]) The *mean value* of a random variable X is defined as:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx.$$

Definition 2.1.7. ([3]) The *variance* of a random variable X is defined as:

$$\text{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

Moreover, the *covariance* between two random variables X and Y is defined as:

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

Proposition 2.1.8. *([12]) Let X, Y, Z be random variables and $a, b, c, d \in \mathbb{R}$. Then*

1. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$,
2. $\text{var}(aX + b) = a^2\text{var}(X)$,
3. $\text{cov}(aX + b, cY + d) = ac\text{cov}(X, Y)$,
4. $\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$.

Definition 2.1.9. ([11]) If the random variable X has the probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x-\mu)^2}{2\sigma^2}),\ x \in \mathbb{R},$$

we say that X has a normal distribution with mean μ and variance σ^2 and it is denoted that $X \sim \mathcal{N}(\mu, \sigma^2)$.

Definition 2.1.10. ([11]) If the k-dimensional random vector $\mathbf{X}$, vector of the random variables $X_1, X_2, ..., X_k$, has the probability density function

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\big),\ \mathbf{x} \in \mathbb{R}^k,$$

we say that $\mathbf{X}$ has a multivariate normal distribution with k-dimensional mean vector $\boldsymbol{\mu}$ and $k \times k$ covariance matrix $\boldsymbol{\Sigma}$ and it is denoted that $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Proposition 2.1.11. *([12]) Let X be a normal random variable with mean μ and variance σ^2. Then the random variable Y defined by $Y = aX + b$, for $a, b \in \mathbb{R}$ with $a \neq 0$, has a normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$.*

Definition 2.1.12. ([11]) For any events $A, B \in \mathcal{F}$ such that $P(B) \neq 0$, the *conditional probability* of A given B is defined by

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

Theorem 2.1.13. *([11]) (**Bayes' theorem**) For any events $A, B \in \mathcal{F}$ and $P(B) \neq 0$,*

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

2.2 Stochastic Process and Stochastic Differential Equations

The following definitions are extensions for sequence of random variables, the idea being that of a family of random variables depending on time.

Definition 2.2.1. ([21]) A collection of random variables which indexed by time t, denoted by $\{X_t\}_{t\geq 0}$ is called a *stochastic process.*

Definition 2.2.2. ([21]) We say that a stochastic process $\{X_t\}_{t\geq 0}$ has *independent increments* if the increments

$$X_{t_1} - X_{t_0}, X_{t_2} - X_{t_1}, ..., X_{t_{n-1}} - X_{t_{n-2}}, X_{t_n} - X_{t_{n-1}}$$

are independent random variables for any $t_0 < t_1 < \cdots < t_n$.

Definition 2.2.3. ([21]) A stochastic process $\{B_t\}_{t\geq 0}$ is called a *Wiener process* with start in $x \in \mathbb{R}$ if the following holds:

- $B_0 = x$,
- for all $0 \leq s \leq t$, we have $B_t - B_s \sim \mathcal{N}(0, t - s)$,
- the process has independent increments.

We say that $\{B_t\}_{t\geq 0}$ is a *standard Wiener process* if $x = 0$.

Definition 2.2.4. ([3]) The *Itô stochastic differential equation* of a stochastic process X_t on the interval $[0, T]$ has the form

$$dX_t = f(t, X_t)dt + g(t, X_t)dB_t, \tag{2.2.1}$$

for $0 \leq t \leq T$, where $\{B_t\}_{t\geq 0}$ is a standard Wiener process and f, g satisfy the following conditions:

- $\mathbb{E}|f(t)|^2 \leq k_1$ and $\mathbb{E}|g(t)|^2 \leq k_2$ for any $t \in [0,T]$ and $k_1, k_2 \in \mathbb{R}^+$,
- $\mathbb{E}|f(t_2) - f(t_1)|^2 \leq k_3|t_2 - t_1|$ and $\mathbb{E}|g(t_2) - g(t_1)|^2 \leq k_4|t_2 - t_1|$ for any $t_1, t_2 \in [0,T]$ and $k_3, k_4 \in \mathbb{R}^+$,
- f and g are measurable on $[0,T] \times \Omega$.

Theorem 2.2.5. *([3]) (**Itô's formula**) Let X_t be a stochastic process that satisfies (2.2.1) and assume that $F := F(t, X_t)$ is a twice continuously differentiable function. Let*

$$\widetilde{f}(t, X_t) = \frac{\partial F}{\partial t} + f(t, X_t)\frac{\partial F}{\partial X_t} + \frac{1}{2}g^2(t, X_t)\frac{\partial^2 F}{\partial X_t^2},$$

and

$$\widetilde{g}(t, X_t) = g(t, X_t)\frac{\partial F}{\partial X_t}.$$

Then F satisfies the stochastic differential equation,

$$dF = \widetilde{f}(t, X_t)dt + \widetilde{g}(t, X_t)dB_t.$$

In Chapter 3, as we simulated the stochastic processes in order to obtain results by the particle filter method, here we provided some background on the *Euler-Maruyama* method.

Next, we define $\widetilde{X}_t$ as numerical approximation of X_t.

Proposition 2.2.6. *([3]) Let X_t be a stochastic process that satisfies (2.2.1). Then, the Euler-Maruyama approximation of X_t on the interval $[0,T]$ has the form*

$$\widetilde{X}_{t_n} = \widetilde{X}_{t_{n-1}} + f(t_{n-1}, \widetilde{X}_{t_{n-1}})\Delta t_n + g(t_{n-1}, \widetilde{X}_{t_{n-1}})\Delta B_{t_n},\ \widetilde{X}_{t_0} = x_0,$$

for $n = 1, 2, ..., N$, where $\Delta B_{t_n} = B_{t_n} - B_{t_{n-1}} \sim \mathcal{N}(0, \Delta t_n)$, $\Delta t_n = T/N$ and $t_n = n\Delta t_n$.

Example 2.2.7. Assume that X_t satisfies the stochastic differential equation

$$dX_t = aX_t dt + bX_t dB_t, \tag{2.2.2}$$

for $0 \leq t \leq T$. Let $F(t, X_t) = \log X_t$. By using Itô's formula with (2.2.2), we obtain that F satisfies the stochastic differential equation

$$dF = \big(\frac{\partial F}{\partial t} + aX_t\frac{\partial F}{\partial X_t} + \frac{1}{2}b^2X_t^2\frac{\partial^2 F}{\partial X_t^2}\big)dt + bX_t\frac{\partial F}{\partial X_t}dB_t$$

$$= \big(0 + aX_t\frac{1}{X_t} - \frac{1}{2}b^2X_t^2\frac{1}{X_t^2}\big)dt + bX_t\frac{1}{X_t}dB_t$$
$$= (a - \frac{1}{2}b^2)dt + bdB_t.$$

By using the Euler-Maruyama method with (2.2.2), the discritized equation is given as

$$\widetilde{X}_{t_n} = \widetilde{X}_{t_{n-1}} + a\widetilde{X}_{t_{n-1}}\Delta t_n + b\widetilde{X}_{t_{n-1}}\Delta B_{t_n},$$

where $\Delta t_n = t_n - t_{n-1}$ and $\Delta B_{t_n} = B_{t_n} - B_{t_{n-1}}$.

2.3 Parameter Estimation Methods

In this section, we consider a methodology for obtaining estimators that is applicable to many types of problems: particle filter and maximum likelihood estimation method.

2.3.1 Particle Filter Method

In this section, we start with introducing the state estimation problem. The filtering problem is investigated. We consider the SIS particle filter method to solve this problem. The algorithm is construced.

State Estimation Problem

In order to define the state estimation problem, we consider the model containing two sequences of random vectors [23]:

1. The hidden state sequence, $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$, is given by

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{p}_{k-1}), \tag{2.3.1}$$

where $\mathbf{f}_k$ is a possibly nonlinear and time-dependent function of the state $\mathbf{x}_{k-1}$ and $\{\mathbf{p}_k\}_{k\in\mathbb{N}}$ is an independent and identically distributed state noise.

The sequence $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ is assumed to be a Markov process, that is

$$p(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{x}_{k-2}, ..., \mathbf{x}_1, \mathbf{x}_0) = p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{2.3.2}$$

2. The observation state sequence, $\{\mathbf{z}_k\}_{k\in\mathbb{N}}$, is given by

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, \mathbf{q}_k), \tag{2.3.3}$$

where $\mathbf{h}_k$ is a possibly nonlinear and time-dependent function of the state $\mathbf{x}_k$ and $\{\mathbf{q}_k\}_{k\in\mathbb{N}}$ is an independent and identically distributed observation noise.

The sequence $\{\mathbf{z}_k\}_{k\in\mathbb{N}}$ is assumed to be a Markov process with respect to the history of $\mathbf{x}_k$, that is

$$p(\mathbf{z}_k|\mathbf{x}_k, \mathbf{x}_{k-1}, ..., \mathbf{x}_1, \mathbf{x}_0) = p(\mathbf{z}_k|\mathbf{x}_k). \tag{2.3.4}$$

The state estimation problem aims at obtaining information about $\mathbf{x}_k$ based on the hidden model (2.3.1) and $\mathbf{z}_{1:k} := \{\mathbf{z}_1, ..., \mathbf{z}_k\}$ given by the observation model (2.3.3). The hidden-observation model can be illustrated by Figure 2.1.

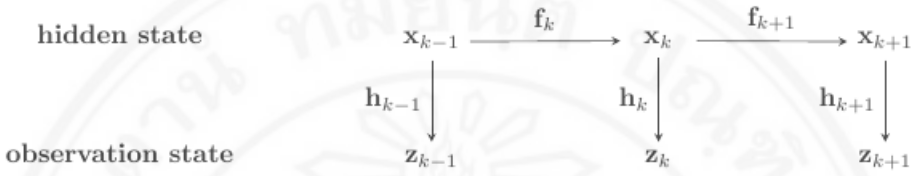

Figure 2.1: A graphical model of hidden-observation model

Different problems can be considered with this model, namely [18]:

1. The prediction problem, concerned with the determination of $p(\mathbf{x}_k|\mathbf{z}_{1:k-1})$.
2. The filtering problem, concerned with the determination of $p(\mathbf{x}_k|\mathbf{z}_{1:k})$.
3. The fixed-lag smoothing problem, concerned with the determination of $p(\mathbf{x}_k|\mathbf{z}_{1:k+p})$, where p is the fixed-lag.

In this work, we focused on solving the filtering problem using the filtering method. The aim of this method is to estimate $p(\mathbf{x}_k|\mathbf{z}_{1:k})$, the posterior distribution based on observation of $\{\mathbf{z}_1, ..., \mathbf{z}_k\}$. By assuming that the initial distribution $p(\mathbf{x}_0|\mathbf{z}_0) = p(\mathbf{x}_0)$ is available and from Bayes' theorem, we have

$$\begin{aligned}
p(\mathbf{x}_k|\mathbf{z}_{1:k}) &= p(\mathbf{x}_k|\mathbf{z}_k|\mathbf{z}_{1:k-1}) \\
&= \frac{p(\mathbf{z}_k|\mathbf{x}_k|\mathbf{z}_{1:k-1})p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \\
&= \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{p(\mathbf{z}_k|\mathbf{z}_{1:k-1})} \\
&= \frac{p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})}{\int p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1})d\mathbf{x}_k} \\
&\propto p(\mathbf{z}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{z}_{1:k-1}) \\
&\propto p(\mathbf{z}_k|\mathbf{x}_k) \int p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})d\mathbf{x}_{k-1}.
\end{aligned}$$

Note that, the posterior distribution is described by three terms, the transition density of hidden state $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ is defined by the equation (2.3.2), the likelihood density $p(\mathbf{z}_k|\mathbf{x}_k)$ defined by the observation model (2.3.4), and suppose that the required posterior distribution $p(\mathbf{x}_{k-1}|\mathbf{z}_{1:k-1})$ at time $k-1$ is avaliable.

The most widely known filtering method called the Kalman filter, can be applied to solve a linear model with Gaussian noise. If the linearity or the Gaussian condition is not valid, the extended Kalman filter (EKF) and the unscented Kalman filter (UKF) can be used. Similarly, Monte Carlo method have been developed in order to represent the posterior distribution in terms of random samples and associated weights, denoted as particle filter, do not to require the hypothesis of the Kalman filter. Hence, the particle filter technique offer an alternative method to solve general filtering problem.

In this work, we treat $\{\mathbf{z}_k\}_{k\in\mathbb{N}}$ as the stock price process and $\{\mathbf{x}_k\}_{k\in\mathbb{N}}$ as the volatility process. For solving filtering problem, we use the particle filter method.

Particle Filter Method

The particle filter method [4] is a Monte Carlo technique for the solution of the state estimation problem. The particle filter is also known as the bootstrap filter, condensation algorithm, interacting particle approximations and survival of the fittest. The key idea of particle filter is to represent the required posterior distribution by a set of random samples with associated weights and to compute estimates based on these samples and weights. As the number of sample becomes large, the result become an equivalent representation to the usual functional description of the posterior distribution.

Let $\{\mathbf{x}_k^{(i)}, w_k^{(i)}\}_{i=1}^N$ denote N *particles* that characterizes the posterior distribution where $\{\mathbf{x}_k^{(i)}\}_{i=1}^N$ is a set of support points with associated weights $\{w_k^{(i)}\}_{i=1}^N$. The weights are normalized such that $\sum_i w_k^{(i)} = 1$.

Then, the posterior distribution can be approximated as [4]:

$$p(\mathbf{x}_k|\mathbf{z}_{1:k}) \approx \sum_{i=1}^{N} w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)}), \tag{2.3.5}$$

where δ is a dirac-delta function.

There are various kinds of particle filters, such as sequential importance sampling (SIS) filter, sampling importance resampling (SIR) filter, auxiliary sampling importance resampling (ASIR) filter and regularized particle filter (RPF). In this work, we focus on the SIS particle filter method.

The idea to generate SIS particle filter simulation are as followed: Firstly, the initials

$\{\mathbf{x}_0^{(i)}\}_{i=1}^N$ need to be drawn from a prior distribution $p(\mathbf{x}_0)$ and the initial weight of each particle is set to equal $1/N$. After the initial draw is completed, there will be two main calculation steps for each time step, *update* and *analysis*. For time step k, the update step is for update the values of $\{\mathbf{x}_k^{(i)}\}_{i=1}^N$ from $\{\mathbf{x}_{k-1}^{(i)}\}_{i=1}^N$ based on *importance density function*, denoted by $\pi(\cdot)$. Note that the importance density function, which normally be different from (2.3.1), is chosen such that it is easier to simulate $\mathbf{x}_k$ compare to (2.3.1). After the update step is completed, the values of $\{w_k^{(i)}\}_{i=1}^N$ will be updated in the analysis step. The weight are updated using the principle of *importance weight sampling* [9]. There are many choices of updated importance weight sampling. In this research, we choose

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\mathbf{z}_k|\mathbf{x}_k^{(i)})p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})}{\pi(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)}, \tag{2.3.6}$$

and we select the optimal importance density function as defined in [9],

$$\pi(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k) = p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k). \tag{2.3.7}$$

Degeneracy Problem and Resampling

A common problem with the SIS particle filter is the degeneracy phenomenon. After some iteration, only few particles would have non-zero importance weights [9]. A suitable measure of degeneracy of the algorithm is the *effective sample size*, N_{eff} introduced in [20], which is defined by

$$N_{eff} = \frac{N}{1 + var(w_k^{(i)})},$$

where $w_k^{(i)}$ is the normalized weight obtained using (2.3.6). Note that the effective sample size can not be evaluated exactly, but we can use an approximate, $\widetilde{N}_{eff}$, which is calculated by

$$\widetilde{N}_{eff} = \frac{1}{\sum_{i=1}^N (w_k^{(i)})^2}. \tag{2.3.8}$$

When $\widetilde{N}_{eff}$ is below a predefined threshold N_T (say $\frac{N}{3}$,$\frac{N}{2}$ or $\frac{2N}{3}$ [1]), this suggest a degeneracy problem. To solve this issue, an intuitive solution is to multiply the particles with high normalized importance weights and discard the particles with low normalized importance weights, which is done in the *resampling* step.

The basic idea of resampling is to eliminate particles that have small weights and to concentrate on particles with large weights. The resampling step involves generating a new set of $\{\mathbf{x}_k^{(i)*}\}_{i=1}^N$ and a new uniform weight $w_k^{(i)*} = \frac{1}{N}$ for each $i = 1, \ldots, N$.

There are many types of resampling [15], such as multinomial resampling, stratified resampling, systematic resampling and residual resampling. In this research, we use the *systematic resampling* [19].

Algorithm

In general, the algorithm for the SIS particle filter is as followed:

Step 1: For time step $k = 0$ and for $i = 1, 2, ..., N$, draw the samples $\mathbf{x}_0^{(i)}$ from the prior distribution $p(x_0)$ and set $w_0^{(i)} = \frac{1}{N}$.

Step 2: For time step $k = 1, 2, 3, ..., M$.

Step 2.1: For $i = 1, 2, ..., N$, draw the samples $\mathbf{x}_k^{(i)}$ from the important density function $\pi(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, \mathbf{z}_k)$.

Step 2.2: For $i = 1, 2, ..., N$, calculate the important weights $w_k^{(i)}$ according to (2.3.6) and normalize.

Step 2.3: Compute $\widetilde{N}_{eff}$ according to (2.3.8). If $\widetilde{N}_{eff} < N_T$, generate the new particle set by systematic resampling.

Step 2.4: Calculate the posterior distribution $p(\mathbf{x}_k|\mathbf{z}_{1:k})$ according to (2.3.5).

2.3.2 Maximum Likelihood Estimation Method

In this section, we present the maximum likelihood estimation method, which is a statistical method that finds the most likely value for the parameter based on the observed data.

Definition 2.3.1. ([22]) Suppose that we observe $X_1, X_2, ..., X_n$, a set of random variables from a distribution that depends on the vector of parameter $\boldsymbol{\alpha}$ with probability density function $f(x_i|\boldsymbol{\alpha})$. Then, the joint probability density function of the observed sample given by

$$f(x_1, ..., x_n|\boldsymbol{\alpha}).$$

Moreover, The *log-likelihood function* considers the density function $f(x_1, ..., x_n|\boldsymbol{\alpha})$ as a function of $\boldsymbol{\alpha}$ where observe sample is taken to be fixed. That is,

$$L(\boldsymbol{\alpha}|x_1, ..., x_n) = \ln f(x_1, ..., x_n|\boldsymbol{\alpha}).$$

Theorem 2.3.2. *([22]) The maximum likelihood estimators of* $\boldsymbol{\alpha}$ *are taken to be the points that maximize the function* $L(\boldsymbol{\alpha}|x_1, ..., x_n)$ *with respect to* $\boldsymbol{\alpha}$*. That is,* $\widehat{\boldsymbol{\alpha}}$ *is the maximum likelihood estimator of* $\boldsymbol{\alpha}$ *if*

$$L(\widehat{\boldsymbol{\alpha}}|x_1, ..., x_n) = \sup_{\boldsymbol{\alpha} \in A} L(\boldsymbol{\alpha}|x_1, ..., x_n),$$

where $\boldsymbol{A}$ *is the set of all vector parameters. Assume that* $L(\boldsymbol{\alpha}|x_1, ..., x_n)$ *is a twice differentiable function, candidates for the maximum likelihood estimator of* $\boldsymbol{\alpha}$ *have the properties*

$$\frac{\partial}{\partial \boldsymbol{\alpha}} L(\boldsymbol{\alpha}|x_1, ..., x_n)|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} = 0,$$

and

$$\frac{\partial^2}{\partial \boldsymbol{\alpha}^2} L(\boldsymbol{\alpha}|x_1, ..., x_n)|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} < 0.$$

Example 2.3.3. Suppose that $X_1, X_2, ..., X_n$ is a set of independent and identically distributed random variables from $\mathcal{N}(\mu, \sigma^2)$ distribution. We will denote the corresponding observed sample as $x_1, x_2, ..., x_n$. The likelihood function in this case is given by

$$\begin{aligned}
L(\mu, \sigma^2|x_1, ..., x_n) &= f(x_1, ..., x_n|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x_i - \mu)^2}{2\sigma^2}) \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\big(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\big).
\end{aligned}$$

Therefore, the log-likelihood function is given by

$$\ln L(\mu, \sigma^2|x_1, ..., x_n) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

The first partial derivatives are

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2|x_1, ..., x_n) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu),$$

and

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2|x_1, ..., x_n) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 \\
&= -\frac{n}{2(\sigma^2)^2} \left(\sigma^2 - \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \right).
\end{aligned}$$

Setting these partial derivatives equal to zero and solving for μ and σ give

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Obviously, the second partial derivative with respect to μ and σ are negative, we obtain that $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators.