2011

Convergence Of The Mean Shift Algorithm And Its Generalizations

Ting Hu
University of Central Florida

CONVERGENCE OF THE MEAN SHIFT ALGORITHM AND ITS GENERALIZATIONS

by

TING HU
B.S. Science and Technology University of Hunan, 2005
M.S. Zhejiang University, 2007

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematics
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Summer Term
2011

Major Professor: Xin Li

To Zhijun Qian

and our child Emily Qian

ABSTRACT

Mean shift is an effective iterative algorithm widely used in image analysis tasks like tracking, image segmentation, smoothing, filtering, edge detection and etc. It iteratively estimates the modes of the probability function of a set of sample data points based in a region.

Mean shift was invented in 1975, but it was not widely used until the work by Cheng in 1995. After that, it becomes popular in computer vision. However the convergence, a key character of any iterative algorithm, has been rigorously proved only very recently, but with strong assumptions.

In this thesis, the method of mean shift is introduced systematically first and then the convergence is established under more relaxed assumptions. Finally, generalization of the mean shift method is also given for the estimation of probability density function using generalized multivariate smoothing functions to meet the need for more real life applications.

ACKNOWLEDGMENTS

It is hard to fully express my grateful gratitude and appreciation for those who guided and supported me through these last two years in only few words here.

I would like to express my sincere gratitude to my advisor, Professor Li Xin, for his efforts to plant the seeds of this work and his constant assistance and support. I wish to thank him deeply from my heart for giving me this precious opportunity to pursue my study in the fascinating mathematical kingdom. He made my master experience in United States one of the most memorable events in my life. His personality, leadership experience, and critical thinking provided an exemplary example for me to follow in my career.

I am very grateful to my thesis approval committee members: Professor Han Deguang and Professor Yong Jiongmin. I thank Professor Yong for his patient and kindness while guiding me throughout many difficult problems. His advanced calculus and optimization courses are enlightening and helpful for my research. I would like to thank Professor Han for his encouragement while I was working as a teaching assistant for his linear algebra course.

A special mention has to be made of all the professors at Department of Mathematics for creating an active and illuminating academic atmosphere for doing research. Special thanks are extended to Professor Zuhair Nashed, Professor Kuppalapalle Vajravelu, Professor Jianjian Ren, Professor Cynthia Y.Young and Professor Marianna Y.Pensky for their expert guidance during those wonderful courses and especially for their instructive guidance to me.

I am also grateful to the University of Central Florida faculty and staff for their support and would like to thank Ms. Norma for her work on preparation and submission of my files and documents. I would also take this opportunity to thanks Ms. Tammy Muhs, Ms. Michelle Taylor,

Mr. Keith Carlson and Mr. David Schweitzer for their consideration and careful arrangements while I was working in the Math Lab.

Great thanks to my friends Ge Lei, Shao Haimei, Qiu Hong, Keri Ann Hagerman, Saliha Pehlivan, Li Pan, Don Porchia, George Nguyen, Rida Benhaddou, Yang Ye, Cheng Teng, Yonggi Park, Roman Krylov, and all of my colleagues. I could not be the one today without such profound discussions with them during my study at UCF.

I would express my deep love and thanks to my husband Zhijun Qian. He gives me encouragement and freedom for anything I would like to do. He has always been supportive and understanding, and he never complained that I could not be a responsible wife while I was busy studying. I cannot thank him enough for being such a perfect husband.

I would like to thank my mom and my dad for being supportive and encouragement for me in my life. Thank my little brother for being with my parents and bringing them so much happiness and laughs. I would also like to thank my parents-in-law – because of their sacrifice I could continue taking my courses when my baby was only four months old. Taking care of my baby made them tired and sleepless everyday but they never complain about it.

And last, but certainly not the least, I would like to thank the most important person in my life, my daughter Emily Yunhan Qian, for being healthy and happy with her grandparents and being patient while waiting her mommy coming home from school.

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND 1

1.1. Digital image representation[15,16] 1

1.2. Statistical background[17,18] 3

1.2.1. Random variables, distribution and density functions 3

1.2.2. Gaussian random variable 5

1.2.3. Conditional expected values 5

1.2.4. Density estimation 6

1.3. Outline of Thesis 7

CHAPTER 2: INTRODUCTION TO MEAN SHIFT METHOD 8

2.1. Original definition of mean shift 8

2.1.1. Probability density estimates 9

2.1.2. Gradient estimates and their properties 10

2.1.3. Mean shift gradient estimates 12

2.1.4. Mean shift normalized gradient estimates 16

2.1.5. Earliest applications 16

2.2. Improved mean shift algorithm 17

2.2.1. Introduction of Cheng's generalization 18

2.2.2. Introduction of D.Comaniciu's applications and proofs of mean shift 19

2.3. Basic ideas and steps of mean shift algorithm 20

CHAPTER 3: CONVERGENCY OF MEAN SHIFT ALGORITHM 23

3.1. Background 23

3.1.1. Sequence and its convergence 23

3.1.2. Cauchy sequence 24

3.2. Previous proofs of convergence and their mistakes 25

3.3. Corrected convergence results and their proofs 29

3.3.1. Assumptions and preliminaries 29

3.3.2. Proofs 32

CHAPTER 4: NEW RESULTS 42

4.1. Need for improvement on the convergence of the sequence of mean shifts 42

4.2. A new result on the convergence 42

4.3. Generalizations of mean shifts and their convergence 47

4.3.1. Multi-kernels 47

4.3.2. Multivariate kernel functions 53

CHAPTER 5: CONCLUSIONS AND FUTURE WORK 59

5.1. Major Contributions 59

5.2. Future Work 59

LIST OF FIGURES

Fig. 1. 1: Cameraman image ..... **Error! Bookmark not defined.**

Fig. 3. 1: First 1000 members of sequence $\{y_j\}$ 28

Fig. 3. 2: $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_1) \right\}$ 36

Fig.4. 1: S_i, $S_{\eta,0}$, $X_i^{'}$ and $X_0^{'}$ 44

LIST OF TABLES

Table 1.1 Top left corner of cameraman image represented by matrix 3

CHAPTER 1: BACKGROUND

In this thesis, we will focus on the properties of a method, referred to as the mean shift method, in computer vision. In particular, we will study its convergence property. Scientists in computer vision area did numerous researches and have made significant progress in mimicking the physiological process of visual system of human on computer. Due to the wide applications of computer vision systems in machine intelligence and homeland security, intensive researches and strong interests have emerged and are swarming into this promising field. Even a small step of improvement or refinement is possible to induce significant effect in the application. Today, with a huge supply of powerful algorithms, computers could achieve many abilities such as clustering, tracking, segmentation, smoothing and filtering, etc.

1.1. *Digital image representation*[15,16]

Digital photo is a well known example of two-dimensional digital image.

The pixels are the smallest element of a digital image and they are often represented by squares. They are normally arranged in a two-dimensional grid and the whole of them represent the image. Since digital image are often rectangular, the pixels in it located by array and could be treated as a matrix. For gray image, it could be represented by a two dimensional matrix and the value of each elements in the matrix could be represented by a function $f(x, y)$. The function $f(x, y)$ is a two dimensional function, x and y are spatial coordinates and the value of $f(x, y)$ is proportional to the brightness of the image at that point. Thus, a digital image looks like this:

$$f(x,y)=\begin{bmatrix} f(0,0) & f(0,1) & \text{L} & f(0,m-1) \\ f(1,0) & f(1,1) & \text{L} & f(1,m-1) \\ \text{M} & \text{M} & \text{O} & \text{M} \\ f(n-1,0) & f(n-1,1) & \text{L} & f(n-1,m-1) \end{bmatrix}$$

Take a simple gray image for example, *Fig.1.1* is a gray image with 512×512 pixels and could be represented by a two-dimensional 512×512 matrix. The exact value of each element in this matrix could be obtained by transferring the image to data. Since this matrix contains 512×512 entries, only the top left corner of this matrix is shown as *table 1.1*. We could see a digital image is composed by a certain number of data and it is easy to handle by playing with those data under some mathematic operations.

Fig. 1. 1: Cameraman image

Table 1.1 Top left corner of cameraman image represented by matrix

73	72	73	76	81	80	73	67	28	32
77	75	74	77	81	80	76	71	31	35
197	195	193	194	197	197	196	193	194	198
192	190	188	188	189	190	190	191	194	198
164	164	164	162	161	162	164	165	154	157
150	152	154	153	151	150	151	153	152	153
155	160	164	164	161	158	158	160	159	158
149	155	161	162	158	154	154	155	160	158
157	156	155	154	154	155	156	157	157	157
157	156	155	154	154	155	156	157	157	157

More information on the basics of digital images can be found in, for example, [15,16].

1.2. *Statistical background*[17,18]

From section 1.1 above, we could see that a digital image is actually a matrix. For such a small gray image, it contains 512×512 elements and it is not easy to find out the explicit form of the function which could represent the value of the elements in it. In order to obtain the underlying information from those large set of sample data of a digital image, an estimation of *f(x,y)* is given using the density estimation form statistics.

1.2.1. Random variables, distribution and density functions

Definition 1.1: A *random variable* is a function, which maps each event or outcome $\xi\in S$ to real number $X(\xi)$. If the mapping $X(\xi)$ is such that the random variable $X(\xi)$ takes on a finite or countable infinite number of values, then $X(\xi)$ is a *discrete random variable*;

whereas, if the range of $X(\xi)$ is an uncountable infinite number of points, $X(\xi)$ is a *continuous random variable*.

Definition 1.2: The *cumulative distribution function* or *cdf* of a random variable X, denoted by

$$F_X(x) = P(X \le x), \text{ for all } x.$$

Definition 1.3: The *probability density function* or *pdf* of a continuous random variable X is (assuming the limit exists)

$$f_X(x) = \lim_{\varepsilon \to 0} \frac{P(x \le X < x + \varepsilon)}{\varepsilon}, \text{ for all } x.$$

From the properties of cumulative distribution, for continuous random variables,

$$P(x \le X < x + \varepsilon) = F_X(x + \varepsilon) - F_X(x)$$

So

$$f_X(x) = \lim_{\varepsilon \to 0} \frac{F_X(x + \varepsilon) - F_X(x)}{\varepsilon} = \frac{dF_X(x)}{dx}$$

From the properties of the *cdf* s, we can infer several important properties of probability density function as follows:

$(1)\ f_X(x) \ge 0$

$(2)\ f_X(x) = \dfrac{dF_X(x)}{dx}$

$(3)\ F_X(x) = \int_{-\infty}^{x} f_X(t)dt$

$(4)\ \int_{-\infty}^{\infty} f_X(t)dt = 1$

$$(5)\int_a^b f_X(t)dt = P(a < X \le b)$$

1.2.2. Gaussian random variable

In the study of image processing, the Gaussian random variable is the most important random variable and most commonly used in the computer vision research area.

Definition 1.4: A *Gaussian random variable* is one whose probability density function can be written in the general form

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2} \right)$$

where m is the mean and σ is the standard deviation. In general, the Gaussian *pdf* is centered about the point $x = m$ and has a width that is proportional to σ. This random variable is referred to as a *normal random variable*. Furthermore, for the special case when $m = 0$ and $\sigma = 1$, it is called a *standard normal*.

1.2.3. Conditional expected values

As specified in **Definition 1.5**, the conditional expected value of a random variable is a weighted average of the values the random variable can take on, weighted by the conditional *pdf* of the random variable.

Definition 1.5: (*conditional expected value*) The expected value of a random variable X, conditioned on some event A is

$$E(X \mid A) = \begin{cases} \int_{-\infty}^{\infty} x f_{X|A}(x) dx & X \text{ is continuous} \\ \sum_k x_k f_{X|A}(x_k) & X \text{ is discrete} \end{cases}$$

Similarly, the expected value of a function $g(X)$ of random variable X , conditioned on some event A is

$$E(g(X) \mid A) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx & X \text{ is continuous} \\ \sum_k g(x_k) f_{X|A}(x_k) & X \text{ is discrete} \end{cases}$$

1.2.4. Density estimation

The goal of density estimation is to obtain an estimation of distribution from the observed data. Probability density function represents the data in the whole sample space from that population. The density estimation methods are classified into parametric estimation and nonparametric estimation. If the distribution is known (up to the unknown parameters) in advance or is assumed to be an exact form, then the problem becomes the estimation of the parameters in the distribution. This is defined to be parametric estimation. Another approach is to estimate the distribution directly from the measured data which is known as nonparametric estimation. This method requires a large amount of observed data and requires intensive numerical computation.

1.3. *Outline of Thesis*

The outline of this thesis will be as follows.

The first chapter summarizes the background information and the basic definitions.

The second chapter recalls the previous work in research and major achievements in the field of mean shift algorithm.

The third chapter focused on the previous attempts and proofs of convergence, which is the key property of mean shift algorithm.

The fourth chapter presents some new results and their proofs on the convergence of mean shift algorithm. It also contains further improvement and generalization of the mean shift proposed by this thesis.

The fifth chapter gives the conclusion and the scope of future work.

CHAPTER 2: INTRODUCTION TO MEAN SHIFT METHOD

The mean shift algorithm is an effective iterative statistical method to find the modes of the probability density function. It is a procedure for finding the local maximum of the probability density function of an unknown distribution by the given set of samples obeying that distribution. In 1975, Fukunaga and Hostetler[1] developed an algorithm, which they referred to as the mean shift algorithm, to estimate the zeros of the gradient of the probability density function by using the kernel based density estimation method. They also applied it to clustering and data filtering. Cheng[4] gave a more systematic study and generalized the mean shift algorithm in 1995. This is studied further by D. Comaniciu[5-9]. The following is an introduction to some important aspects of this algorithm.

2.1. *Original definition of mean shift*

In the 1975 paper by Fukunaga and Hostetler[1],they studied about the application of estimation of the gradient of density functions in pattern recognition. A key idea was introduced by estimating the gradient of density functions of a point was introduced by using the sample observations within a small region around it. In this process, a new term “mean shift” was introduced and a new algorithm named “mean shift algorithm” was given.

2.1.1. Probability density estimates

In most image processing and pattern recognition problems, very little is known about the true probability density function or even the form of it. So the exact gradient of a density function could not be obtained directly from the real probability density function (*pdf*) of the sample data in most cases since the lack of knowledge about the explicit form of *pdf* . A straight forward approach to estimate the density gradient would be to first approximate the probability density function and then take its gradient.

Based on the idea above, a form of differentiable nonparametric multivariate estimators of the probability density functions introduced by Cacoullos'[3] (*eq. 2.1*) is adopted, which is an extension of Parzen's[2] univariate kernel estimates.

$$\hat{f}(X) \equiv \left(Nh^n\right)^{-1} \sum_{j=1}^{N} k(h^{-1}(X - X_j)) \qquad \textit{Eq. 2.1}$$

where X_1 , X_2 ,..., X_N is a set of N independent and identically distributed n -dimensional random vectors defined as

$$X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \mathrm{M} \\ x_{in} \end{pmatrix} \qquad \textit{Eq. 2.2}$$

Function $k(Y)$ satisfies

$$\sup_{Y \in R^n} \left|k(Y)\right| < \infty \qquad \textit{Eq. 2.3}$$

$$\int_{R^n} \left|k(Y)\right| dY < \infty \qquad \textit{Eq. 2.4}$$

$$\lim_{\|Y\|\to\infty} \|Y\|^n \left|k(Y)\right| = 0 \qquad \textit{Eq. 2.5}$$

$$\int_{R^n} k(Y)\,dY = 1 \qquad \textit{Eq. 2.6}$$

and $h(N)$ is a function of the sample size N.

2.1.2. Gradient estimates and their properties

By taking the gradient of the proposed density estimates (*eq. 2.1*), the gradient estimates are

$$\nabla_x \hat{f}(X) \equiv \left(Nh^n\right)^{-1} \sum_{j=1}^{N} \nabla_x k(h^{-1}(X - X_j)) \qquad \textit{Eq. 2.7}$$

Let $Y = h^{-1}(X - X_j)$, then $\nabla_x Y = h^{-1}$. Based on the chain rule, eq. 2.*7* can be changed to

$$\nabla_x \hat{f}(X) \equiv \left(Nh^{n+1}\right)^{-1} \sum_{j=1}^{N} \nabla k(h^{-1}(X - X_j)) \qquad \textit{Eq. 2.8}$$

where

$$\nabla k(Y) \equiv \left(\frac{\partial k(Y)}{\partial y_1}, \frac{\partial k(Y)}{\partial y_2}, \ldots, \frac{\partial k(Y)}{\partial y_n}\right)^{\mathrm{T}} \qquad \textit{Eq. 2.9}$$

Some conditions on $k(X)$ and $h(N)$ are made in order to guarantee asymptotic unbiasedness, consistency, and uniform consistency of the gradient estimate.

1. Asymptotically unbiased.

   An estimate is said unbiased if the distance between the average of the collection of estimates and the single parameter being estimated is zero. If the gradient estimate is

asymptotically unbiased, it means that the mean value of the estimate converges to the true value of the gradient as $N \to \infty$.This property is true when[1]

$$\lim_{N\to\infty} h(N) = 0 \qquad \textit{Eq. 2.10}$$

2. Consistent in quadratic mean.

   This property means that for large N, the variance in the estimate is close to the true value. In order to content this property, in addition to *eq.2.5*, *h* must satisfy

$$\lim_{N\to\infty} Nh^n(N) = \infty \qquad \textit{Eq. 2.11}$$

   In addition to *eq. 2.3* to *2.5*, the kernel function is such that

$$\sup_{Y\in R^n} \left| k_i'(Y) \right| < \infty \qquad \textit{Eq. 2.12}$$

$$\int_{R^n} \left| k_i'(Y) \right| dY < \infty \qquad \textit{Eq. 2.13}$$

$$\lim_{\|Y\|\to\infty} \|Y\|^n k_i'(Y) = 0 \qquad \textit{Eq. 2.14}$$

   where

$$k_i'(Y) \equiv \frac{\partial k(Y)}{\partial y_i} \qquad \textit{Eq. 2.15}$$

3. Uniform consistency.

   Meeting this property means that the estimate is of high probability close to the true value for large values of *N*. To satisfy this property, all the conditions for *h* mentioned above must be satisfied.

In conclusion, with those properties, Fukunaga and Hostetler proved that their estimate of gradient of density is quite accurately close to the true value of gradient. The proofs of properties of gradient estimate are given by Fukunaga and Hostetler[1] in the appendix of their paper. We will recall some of these properties in the next two subsections.

2.1.3. Mean shift gradient estimates

Since the function *k* can be any kernel function satisfying *eq.2.3* to *eq.2.6*, Fukunaga and Hostetler[1] choose the Guassian kernel function which is a well known differentiable multivariate kernel function and satisfying those conditions to be *k*. The Guassian kernel function is

$$k(X) \equiv (2\pi)^{-n/2} \exp(-\frac{1}{2} X^{\mathrm{T}} X) \qquad \textit{Eq. 2.16}$$

We know

$$X^{\mathrm{T}} X = (x_1, x_2, ..., x_n) \begin{pmatrix} x_1 \\ x_2 \\ \mathrm{M} \\ x_n \end{pmatrix} = x_1^{\,2} + x_2^{\,2} + \mathrm{L} + x_n^{\,2} \qquad \textit{Eq. 2.17}$$

Let $g(X) = X^{\mathrm{T}} X$, then

$$\nabla_x X^{\mathrm{T}} X = \nabla_x g = \begin{pmatrix} \frac{\partial g}{\partial x_1} \\ \frac{\partial g}{\partial x_2} \\ \mathrm{M} \\ \frac{\partial g}{\partial x_n} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \\ \mathrm{M} \\ 2x_n \end{pmatrix} = 2X \qquad \textit{Eq. 2.18}$$

so

$$\begin{aligned} \nabla_x k(X) &\equiv (2\pi)^{-n/2} \nabla_x \exp(-\frac{1}{2} X^{\mathrm{T}} X) \\ &= (2\pi)^{-n/2} \cdot (-\frac{1}{2}) \cdot 2X \exp(-\frac{1}{2} X^{\mathrm{T}} X) \\ &= -(2\pi)^{-n/2} \cdot X \exp(-\frac{1}{2} X^{\mathrm{T}} X) \end{aligned} \qquad \textit{Eq. 2.19}$$

Substitute it into *eq.2.7* for $\nabla_x k(X)$ to get

$$\begin{aligned} \nabla_x \hat{f}(X) &\equiv (Nh^n)^{-1} \sum_{j=1}^{N} \nabla_x k(h^{-1}(X - X_j)) \\ &= (Nh^n)^{-1} \sum_{j=1}^{N} -(2\pi)^{-n/2} \frac{1}{h^2} (X - X_j) \exp\left[-\frac{1}{2h^2} (X - X_j)^{\mathrm{T}} (X - X_j) \right] \\ &= (N)^{-1} \sum_{j=1}^{N} (X_j - X) h^{-(n+2)} (2\pi)^{-n/2} \exp\left[-\frac{1}{2h^2} (X - X_j)^{\mathrm{T}} (X - X_j) \right] \end{aligned} \qquad \textit{Eq. 2.20}$$

For convenience, change *j* to *i* in the above result and get the general form of estimate of the density gradient

$$\nabla_x \hat{f}(X) \equiv (N)^{-1} \sum_{i=1}^{N} (X_i - X) h^{-(n+2)} (2\pi)^{-n/2} \exp\left[-\frac{1}{2h^2} (X - X_i)^{\mathrm{T}} (X - X_i) \right] \qquad \textit{Eq. 2.21}$$

This result is very familiar since it is similar to $\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(X_i - X\right)$, which is just the mean of the $X_i - X$. In this function, $X_i - X$ is defined as "shift" of each point from *X*. Therefore the term "shift" is a noun at the beginning. Fukunaga and Hostetler also defined $h^{-(n+2)}\left(2\pi\right)^{-n/2}\exp\left[-\frac{1}{2h^2}\left(X - X_i\right)^{\mathrm{T}}\left(X - X_i\right)\right]$ as "weighting factor".

The same general form will result if the kernel function is of the form

$$k\left(X\right) = g(X^{\mathrm{T}}X) \qquad \textit{Eq. 2.22}$$

A simple kernel with this form is

$$k\left(X\right) = \begin{cases} c\left(1 - X^{\mathrm{T}}X\right), & X^{\mathrm{T}}X \le 1 \\ 0, & X^{\mathrm{T}}X > 1 \end{cases} \qquad \textit{Eq. 2.23}$$

where

$$c = \pi^{-n/2}\left(\frac{n+2}{2}\right)\Gamma\left(\frac{n+2}{2}\right) \qquad \textit{Eq. 2.24}$$

Taking the gradient of *eq.2.23* and substituting to *eq.2.8,* we obtain the gradient estimate as

$$\begin{aligned} \nabla_x \hat{f}\left(X\right) &\equiv \left(Nh^{n+1}\right)^{-1}\sum_{X_i \in S_h(X)} \nabla k\left(h^{-1}\left(X - X_i\right)\right) \\ &= \left(Nh^{n+1}\right)^{-1} 2ch^{-1}\sum_{X_i \in S_h(X)}\left(X_i - X\right) \qquad \textit{Eq. 2.25} \\ &= \left(Nh^{n+2}\right)^{-1} 2c\sum_{X_i \in S_h(X)}\left(X_i - X\right) \end{aligned}$$

where

$$S_h\left(X\right) \equiv \left\{Y : \left(Y - X\right)^{\mathrm{T}}\left(Y - X\right) \le h^2\right\} \qquad \textit{Eq. 2.26}$$

Note that $S_h(X)$ is a small neighborhood around, that contains all points Y such that the Euclidean distance $d(Y,X)$ between Y and X is less than or equal to h. Then we could know the volume of this neighborhood is

$$\nu_h(X) \equiv \int_{S_h(X)} dY = \frac{h^n \pi^{n/2}}{\Gamma\left(\frac{n+2}{2}\right)} \qquad \textit{Eq. 2.27}$$

Substituting *eq.2.24* and *eq.2.26* into *eq.2.25*, we obtain as the gradient estimate,

$$\begin{aligned}
\nabla_x \hat{f}(X) &= \left(Nh^{n+2}\right)^{-1} 2c \sum_{X_i \in S_h(X)} (X_i - X) \\
&= \left(Nh^{n+2}\right)^{-1} 2\pi^{-n/2} \left(\frac{n+2}{2}\right) \Gamma\left(\frac{n+2}{2}\right) \sum_{X_i \in S_h(X)} (X_i - X) \\
&= \left(\frac{\Gamma\left((n+2)/2\right)}{h^n \pi^{n/2}}\right) \frac{n+2}{Nh^2} \sum_{X_i \in S_h(X)} (X_i - X) \qquad \textit{Eq. 2.28} \\
&= \frac{1}{\nu_h(X)} \cdot \frac{n+2}{Nh^2} \sum_{X_i \in S_h(X)} (X_i - X) \\
&= \frac{k}{N\nu_h(X)} \cdot \frac{n+2}{h^2} \sum_{X_i \in S_h(X)} \frac{1}{k} (X_i - X)
\end{aligned}$$

In *eq.2.28,* $M_h(X) \equiv \sum_{X_i \in S_h(X)} \frac{1}{k}(X_i - X)$ is the sample mean shift of the observations in the small region $S_h(X)$ around X.

2.1.4. Mean shift normalized gradient estimates

Comparing *eq.2.28* and *eq.2.1*, we could see that the term in *eq.2.28*, which is $\frac{k}{Nv_h(X)}$, is identical to *eq.2.1* if *k* is uniform over the regions $S_h(X)$. So we let $f(X) = \frac{k}{Nv_h(X)}$ and take it to the left side of *eq.2.28*, we get

$$\frac{\nabla_x f(X)}{f(X)} = \frac{n+2}{h^2} M_h(X) \qquad \textit{Eq. 2.29}$$

We also know

$$\frac{\nabla_x f(X)}{f(X)} = \nabla_x \ln f(X) \qquad \textit{Eq. 2.30}$$

Hence from *eq.2.29* and *eq.2.30*, we get

$$\nabla_x \ln f(X) = \frac{n+2}{h^2} M_h(X) \qquad \textit{Eq. 2.31}$$

This shows that the estimation of normalized gradient is simple and easy to calculate based on the mean-shift method.

2.1.5. Earliest applications

After mean shift between data points in a small region and *X* is used to estimate gradient of probability density, the earliest applications of it began. Those applications focused on mode clustering and data filtering.

Assume that a point in a pattern or a signal could be represented by a vector X_i,

$$X_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \mathrm{M} \\ x_{in} \end{pmatrix} \qquad Eq.\ 2.32$$

First, the gradient at the observation point could be estimated base on a selected region of data points as described in the previous subsection. Then, the observation point is moved, or called shifted (here shift is a verb), in the direction of gradient of this sample to the mode. This process is continued iteratively until no substantial shift of points occurs. This is based on the precondition that the estimation is convergent. The proof of its convergence as a crucial problem as it is was left unsolved and it is one of our research tasks in this thesis.

In conclusion, Fukunaga and Hostetler[1] in their illustrious paper invented the term "mean-shift" and defined the original meaning of it, and they also introduced weighting factor which is an important topic that is investigated later by many researchers. But this algorithm did not cause sufficient attention for a long time until twenty years later when Cheng[4] in another classical paper gave a systematic study of the mean shift algorithm, which will be introduced in next section.

2.2. *Improved mean shift algorithm*

Fukunaga and Hostetler[1] proposed the mean shift algorithm as a cluster analysis method, and their intuition that mean shift is a gradient ascent need to be verified, the convergence of mean shift need to be proved, and its relationship with other similar algorithms needs to be illuminated.

2.2.1. Introduction of Cheng's generalization

In 1995, Cheng[4] testified the intuition and established that mean shift is a gradient ascent in a formal and rigorous proof. He also studied the convergence of the algorithm based on the Gaussian kernel as well as a specified general form of the probability density function. He improved and generalized this algorithm in the following three aspects.

First, he introduced several different forms of kernels, which include non-flat kernels such as Gaussian kernel, Epanechnikov kernel, and biweight kernel.

Second, points in data can be weighted. So the contribution of each sample data is allowed to be different, which extends the real applications of mean shift algorithm in computer vision.

Third, the (generalized) algorithm could be performed on any subset of X. The original data points themselves were kept constant. Another set, a copy of the original data set, was allowed to move around in the Euclidean space.

Cheng also introduced the concept "shadow" of a kernel in order to clarify the relationships between different kernels. He proved that mean shift on any kernel is equivalent to the gradient ascent on the density estimated with its shadow[4].

In addition to the above contribution, he showed some particular behaviors of mean shift in clustering and studied how the probability strategy can be applied in weight assignment. He also studied the global optimization application of mean shift and gave specific examples. Furthermore, he suggested that some computational obstacles should be resolved to enable the widely application of mean shift.

2.2.2. Introduction of D.Comaniciu's applications and proofs of mean shift

D.Comaniciu and his co-workers[5-9] made mean shift algorithm popular by successfully applying the algorithm to feature space analysis and achieving many good application results in image smoothing, segmentation, real-time object tracking, etc. They realized the application of this useful algorithm and enabled its benefit to us through its wide application in computer vision area.

Moreover, they indicated that under some specifically assumptions, mean shift algorithm will convergent to the nearest density gradient of feature space. This is concluded in a theorem as follows which will be very useful for later proofs.

Theorem: If the kernel K has a convex and monotonically decreasing profile, the sequences $\{\hat{f}(y_j), j=1,2,...\}$ and $\{y_j, j=1,2,...\}$ converge, and $\{\hat{f}(y_j), j=1,2,...\}$ is monotonically increasing.

The above theorem itself is correct and the proof method of this theorem provided in their papers is ingenious. In addition, this theorem guarantees the application of mean shift application under the above specified conditions.

However, in their discussion of the convergence of $\{y_j, j=1,2,...\}$, they made several mathematical mistakes which are not easy to be detected.

First, in Refs. [6,7], their proofs were essentially based on the inaccurate conclusion which says $\{y_j, j=1,2,...\}$ converges when $\|y_{j+1}-y_j\|$ converges to zero as j goes to infinite .

As is well known and also pointed out by Li[10], this is incorrect. This will be discussed later in next chapter and the counterexamples will be provided.

Another mistake appeared in Ref. [8] where a key step of the proof in Ref. [8] is

$$\left\|y_{j+m}-y_{j+m-1}\right\|^2+\mathrm{L}\ +\left\|y_{j+1}-y_j\right\|^2\geq\left\|y_{j+m}-y_j\right\|^2,$$

which does not hold in general. This is also proved and the counterexample will be shown in Chapter 3.

Those mistakes were pointed out by Li et al[10], they also provided corrections. However, there are still some gaps exist. The detail of their proofs, defections and our corrections for their proofs will be shown in next chapter.

2.3. *Basic ideas and steps of mean shift algorithm*

First, feature space is considered as a probability density function $f(X), X\in\mathbf{R}^m$. Let $\{X_i, 1\leq i\leq N\}$ be an independently and identically distributed sample data set. If $f(X)$ is estimated by

$$\hat{f}(X)=\sum_{i=1}^{N}w_iK_i(X) \qquad \textit{Eq. 2.33}$$

where $\sum_{i=1}^{N}w_i=1$, K is kernel function . If a kernel function is specified. It will be easy to figure out its gradient

$$\nabla\hat{f}(X)=\sum_{i=1}^{N}w_i\nabla K_i \qquad \textit{Eq. 2.34}$$

If the kernel function is specified as

$$K_i(X) = ck\left(\left\|X - X_i\right\|^2\right) \qquad \textit{Eq. 2.35}$$

Then

$$\nabla \hat{f}(X) = \sum_{i=1}^{N} w_i \nabla K_i = \sum_{i=1}^{N} 2w_i ck'\left(\left\|X - X_i\right\|^2\right)(X - X_i) \qquad \textit{Eq. 2.36}$$

Let

$$L_i(X) = -2w_i ck'\left(\left\|X - X_i\right\|^2\right) \qquad \textit{Eq. 2.37}$$

We have

$$\nabla \hat{f}(X) = \sum_{i=1}^{N} L_i(X)(X_i - X) = \sum_{i=1}^{N} L_i(X)\left[\left(\sum_{i=1}^{N} L_i(X)\right)^{-1} \sum_{i=1}^{N} L_i(X) X_i - X\right] \qquad \textit{Eq. 2.38}$$

In the right side of the above equation, the part $\left(\sum_{i=1}^{N} L_i(X)\right)^{-1} \sum_{i=1}^{N} L_i(X) X_i - X$ is defined as mean shift vector. Substitute X by y_i . $y_j (j = 1,2,3...)$ represents the iterative point that used calculate the gradient of probability density and the mean shift vector in every iterative step. y_1 is the initial iterative point. Then the mean shift vector would be written as

$$ms_k(y_j) = \left(\sum_{i=1}^{N} L_i(y_j)\right)^{-1} \sum_{i=1}^{N} L_i(y_j) X_i - y_j \qquad \textit{Eq. 2.39}$$

Then the mean shift procedure could be simplified as the following four steps:

1. Compute mean shift vector by *eq.2.39,*
2. Translate the Kernel window

$$y_{j+1} = y_j + ms_k\left(y_j\right) = \left(\sum_{i=1}^{N} L_i\left(y_j\right)\right)^{-1} \sum_{i=1}^{N} L_i\left(y_j\right) X_i \quad , \qquad \textit{Eq. 2.40}$$

3. Recalculate the new mean shift vector,

4. Repeat till the result is convergent.

The result get from the above process will reach the local maximum of the probability density function of the feature space.

CHAPTER 3: CONVERGENCY OF MEAN SHIFT ALGORITHM

3.1. *Background*

We briefly review the definitions of the concepts of sequence, convergence, and Cauchy sequences that will be important to our discussion.

3.1.1. Sequence and its convergence

Sequence

A sequence is an ordered list of objects. Like a set, a sequence contains members (also called elements or terms of the sequence). The number of terms (possibly infinite) in a sequence is called the length of the sequence. Unlike a set, order matters, and exactly the same elements can appear multiple times at different positions in the sequence. A sequence can be reviewed as a discrete function. For example, (C, R, Y) is a sequence of letters that differs from (Y, C, R), as the ordering matters. Sequences can be finite, as in this example, or infinite, such as the sequence of all even positive integers (2, 4, 6,...). In this work, we will consider sequences of real numbers or points from the Euclidean space R^N. For convenience, we will write $\{x_n\}$ for a sequence x_1, x_2, …

Definition of convergence for a sequence

The limit of a sequence $\{x_n\}$ is, intuitively, the unique number or point l (if it exists) such that the terms of the sequence become arbitrarily close to l for "large" values of n. If the limit exists, then we say that the sequence is convergent and that it converges to l. This can be

described more precisely as: the sequence $\{x_n\}$ has limit l if for every positive real number ε, there is a positive integer N such that for all natural numbers $m > N$, $|x_m - l| < \varepsilon$.

3.1.2. Cauchy sequence

A Cauchy sequence is a sequence whose elements become arbitrarily close to each other as the sequence progresses. To be more precise, given any positive number, we can always drop some terms from the start of the sequence, so that the maximum of the distances between any two of the remaining elements is smaller than that number.

For Real numbers R

A sequence x_1, x_2, x_3 ,…of real numbers is called Cauchy, if for every positive real number ε, there is a positive integer N such that for all natural numbers $m, n > N$, $|x_m - x_n| < \varepsilon$. In a similar way one can define Cauchy sequences of complex numbers. Cauchy formulated such a condition by requiring $|x_m - x_n|$ be infinite small for every pair of infinite m, n.

In a metric space M

To define Cauchy sequences in any metric space, the absolute value $|x_m - x_n|$ is replaced by $d(x_m, x_n)$, which is the distance function between x_m and x_n. If for every positive real number ε > 0, there is a positive integer N such that for all natural numbers m, n, the distance $d(x_m, x_n) < \varepsilon$. Roughly speaking, the terms of the sequence are getting closer and closer together in a way that suggests that the sequence ought to have a limit in M. Nonetheless, such a limit

does not always exist within M. A special metric space that we will mainly consider is the Euclidean space R^N where the distance function is given by the Euclidean distance:

$$d(x,y) = \| x - y \| = \sqrt{\sum_{k=1}^{N} (x_k - y_k)^2} \qquad \textit{Eq. 3.1}$$

Not every sequence is convergent. The following important theorem tells us when a sequence in R and R^N is convergent.

Theorem: A sequence of points in R or R^N has the Cauchy property if and only if it is a convergent sequence.

3.2. *Previous proofs of convergence and their mistakes*

As we know, convergence is a prerequisite for any iterative algorithm. Otherwise, the computing will not end. The proof of convergence of mean shift sequence $\{y_j, j = 1,2,...\}$ deduced by Cheng[4] is base on the following assumptions:

(1) $k(x) = e^{-\|x\|^2}$.

(2) The density of random variable x is $f(x) = e^{-\gamma^2 \|x\|^2}$, $\gamma < \beta$.

However, it is difficult to guarantee the second assumption in real applications since the true value of γ is unknown. Hence, its applicability is confined to some area.

Comanniciu, Ramesh and Meer[6-8]attempted to prove the convergence of mean shift sequence $\{y_j, j = 1,2,...\}$ under the more general assumption that $k(x)$ is convex and

monotonically decreasing, and $w_i = 1/N$. However, the proofs in Refs.[6-8] contain mathematical mistakes as we mentioned before in Chapter 2. First, in Refs. [6,7], their proofs were essentially based on the inaccurate conclusion which says $\{y_j, j = 1,2,...\}$ converges when $\|y_{j+1} - y_j\|$ converges to zero as j goes to infinite . As is well known and also pointed out by Li[10] in the following example, this is incorrect.

Example 1

Let $y_j = \sum_{i=1}^{j} 1/i$,then $\|y_{j+1} - y_j\| = \frac{1}{j+1} \to 0 (j \to \infty)$. However, it is well known that $\{y_j, j = 1,2,...\}$ does not converge.

This well-known example is not very appropriate for a counterexample for the proofs in Refs.[6,7] since the sequence $\{y_j, j = 1,2,...\}$ is bounded there while the sequence in above example is not. We provide the following counterexample:

Example 2

We give an intuitive example of a bounded sequence $\{y_j\}$ that is not convergent, while the sequence satisfies $\| y_{j+1} - y_j \| \le \frac{1}{j+1} \to 0$ as $j \to \infty$.

We could generate an infinite sequence $\{y_j\}$ based on the idea by "folding" a sequence that goes to infinity through letting the y_j's move back and forward between a bounded area

while the distance between y_j and y_{j+1} becomes closer as j gets bigger. But y_j get accumulated near the lower bound and the super bound, hence $\{y_j\}$ is not convergent.

For example, in the case of a sequence of real numbers, the first 1000 members of this sequence could be obtained by the following algorithm,

```
x(1)=1, y(1)=0,
  for j=1:1000
      if  (((y(j)+x(j)/j)>=0)&&((y(j)+x(j)/j)<=1))
          y(j+1)=y(j)+x(j)/j
          x(j+1)=x(j)
      else
          y(j+1)=y(j)-x(j)/j
          x(j+1)=-x(j)
      end
  end
```

The first few elements of the sequence $\{y_j\}$ are:

0 1.0000 0.5000 0.1667 0.4167 0.6167 0.7833 0.9262 0.8012 0.6901 0.5901 0.4992 0.4158 0.3389 …

This sequence could be easily shown by the algorithm above that the distance between y_j and y_{j+1} is no more than $\frac{1}{j}$ which tends to 0 as $j \to \infty$, whereas the sequence itself is not convergent. These properties are shown on *Figure 3.1*.

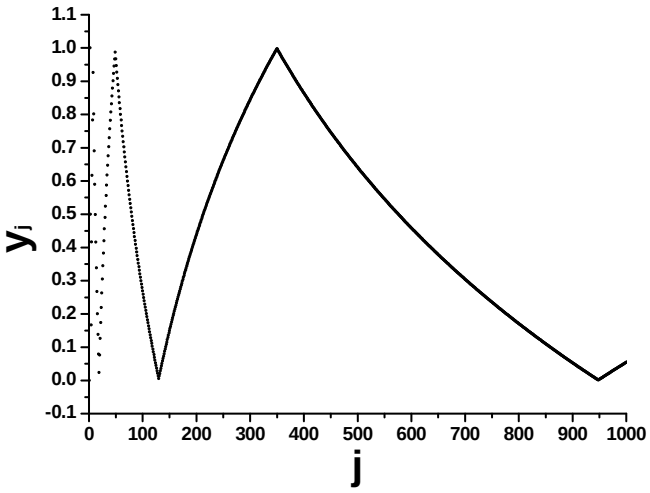

Fig. 3. 1: First 1000 members of sequence $\{y_j\}$

Another mistake appeared in Ref. [8] where a key step of the proof in Ref. [8] is

$$\left\|y_{j+m}-y_{j+m-1}\right\|^2+\mathrm{L}\ +\left\|y_{j+1}-y_j\right\|^2\geq\left\|y_{j+m}-y_j\right\|^2$$

which does not hold in general. The following is an example provided by Ref. [10].

Example 3

Let $m=2$, then

$$\begin{aligned}\left\|y_{j+2}-y_j\right\|^2&=\left\|y_{j+2}-y_{j+1}+y_{j+1}-y_j\right\|^2\\&=\left\|y_{j+2}-y_{j+1}\right\|^2+\left\|y_{j+1}-y_j\right\|^2+2\left(y_{j+2}-y_{j+1}\right)^{\mathrm{T}}\left(y_{j+1}-y_j\right)\end{aligned}$$

We also know $\left(y_{j+2}-y_{j+1}\right)^{\mathrm{T}}\left(y_{j+1}-y_j\right)\geq 0$ from Theorem 2 in Ref.[8], hence

$$\left\|y_{j+2}-y_j\right\|^2\geq\left\|y_{j+2}-y_{j+1}\right\|^2+\left\|y_{j+1}-y_j\right\|^2$$

It is conflict with $\left\|y_{j+m}-y_{j+m-1}\right\|^2+\mathrm{L}\ +\left\|y_{j+1}-y_j\right\|^2\geq\left\|y_{j+m}-y_j\right\|^2$ as claimed in Ref. [8].

Therefore, due to the mistakes mentioned above, the convergence of the sequence of mean shifts was not verified for general kernels before Ref. [10] provided a correct proof under some restrictive assumptions. The next section will discuss the statements and proofs given in Ref. [10].

3.3. *Corrected convergence results and their proofs*

3.3.1. Assumptions and preliminaries

We follow the treatment given in Ref. [10], there are some assumptions and preliminaries we should know before stating the results and proofs of convergence.

Definition 1. Function $k(x)$ is called a bounded kernel if , on $[0,+\infty)$, it satisfies:

(1) (positivity) $k(x) \geq 0$.

(2) (decreasing) $k(x_1) \geq k(x_2), 0 \leq x_1 \leq x_2 < +\infty$.

(3) (integrability) $\int_0^{\infty} k(x)dx < \infty$.

(4) (boundedness) $0 < k(0) < +\infty$.

Given a bounded kernel function $k(x)$, the density estimation of random variable X is defined as

$$\hat{f}(X) = \sum_{i=1}^{N} w_i \frac{1}{\int k\left(h^{-2}(X - X_i)^{\mathrm{T}} \sum_i^{-1} (X - X_i)\right) dX} \cdot k\left(h^{-2}(X - X_i)^{\mathrm{T}} \sum_i^{-1} (X - X_i)\right) \qquad \textit{Eq. 3.2}$$

where $h > 0$ is a fixed constant to control the size of the window (region of sample), the matrix $\sum_i^{-1}$ is a positive definite matrix to represent the local structure around X_i, w_i is the prior

probability for X_i, which is also called the weight of sample X_i. Hence $\sum_{i=1}^{N} w_i = 1$, $w_i > 0$, *i=1,2,...,N.*

Write

$$\hat{f}(X) = \sum_{i=1}^{N} w_i K_i(X) \qquad \textit{Eq. 3.3}$$

where

$$K_i(X) = ck\left(\left\|X - X_i\right\|_{H_i}^2\right)$$

$$\left\|X - X_i\right\|_{H_i}^2 = (X - X_i)^{\mathrm{T}} H_i (X - X_i)$$

$$H_i = \sum\nolimits_i^{-1} / h^2, \quad H = \{H_i, 1 \le i \le N\}$$

Then, we could find that $c = \dfrac{1}{\int k\left(h^{-2}(X - X_i)^T \sum_i^{-1}(X - X_i)\right) dX} > 0$ is a constant to ensure that $K_i(X)$ is a probability density function.

Assume that *k* is differentiable, from *eq.3.3*, we could get the gradient of probability density

$$\begin{aligned} \nabla \hat{f}(X) &= \sum_{i=1}^{N} w_i \nabla K_i(X) \\ &= \sum_{i=1}^{N} 2 w_i c k'\left(\left\|X - X_i\right\|_{H_i}^2\right) H_i (X - X_i) \end{aligned} \qquad \textit{Eq. 3.4}$$

Eq.3.4 could be simplified as

$$\begin{aligned}\nabla \hat{f}(X) &= \sum_{i=1}^{N} L_i(X)(X_i - X) \\ &= \sum_{i=1}^{n} L_i(X) X_i - \sum_{i=1}^{n} L_i(X) X\end{aligned} \qquad \textit{Eq. 3.5}$$

where

$$L_i(X) = -2w_i c k'\left(\left\|X - X_i\right\|_{H_i}^2\right) H_i \qquad \textit{Eq. 3.6}$$

From the decreasing property of $k(x)$, we know that $k'(x) < 0$. So $L_i(X)$ should also be positive definite and invertible. Then the *eq.3.5* could be rewritten as

$$\begin{aligned}\nabla \hat{f}(X) &= \sum_{i=1}^{N} L_i(X) X_i - \sum_{i=1}^{N} L_i(X) X \\ &= \sum_{i=1}^{N} L_i(X) \left[\left(\sum_{i=1}^{N} L_i(X)\right)^{-1} \sum_{i=1}^{N} L_i(X) X_i - X\right]\end{aligned} \qquad \textit{Eq. 3.7}$$

Then we get a term $ms_k(X) = \left(\sum_{i=1}^{N} L_i(X)\right)^{-1} \sum_{i=1}^{N} L_i(X) X_i - X$, which represents the mean shift vector. Let X be y_j above:

$$ms_k(y_j) = \left(\sum_{i=1}^{N} L_i(y_j)\right)^{-1} \sum_{i=1}^{N} L_i(y_j) X_i - y_j \qquad \textit{Eq. 3.8}$$

Then we get the iterative procedure of mean shift algorithm

$$y_{j+1} = y_j + ms_k(y_j) = \left(\sum_{i=1}^{N} L_i(y_j)\right)^{-1} \sum_{i=1}^{N} L_i(y_j) X_i \qquad \textit{Eq. 3.9}$$

3.3.2. Proofs

We now discuss the results and proofs of Ref. [10] on the convergence of the mean shift algorithms. We will present the proofs in such a manner that best suits for our later extension to be given in next chapter. Much more detail is added in the presentation and some necessary clarifications are also provided. In particular, the end of the proof of Theorem 2 is augmented to fill a small technical gap in the original proof given in Ref. [10].

Indeed, we will prove that both the estimation of the probability density sequence $\{\hat{f}(y_j), j=1,2,...\}$ and the iterative sequence $\{y_j, j=1,2,...\}$ are convergent. To obtain these results for general kernels, Li, Hu, and Wu in Ref. [10] assume some more properties on the kernel functions.

Definition 2: A function $k:[0,+\infty]\to R$ is smoothly convex if a bounded and continuous k' exists and satisfies

$$k(x_2)-k(x_1)>k'(x_1)(x_2-x_1), \forall x_1\geq 0, x_2\geq 0, x_1\neq x_2 \qquad \textit{Eq. 3.10}$$

Remark. The above definition is taken directly from Ref. [10]. The phrase "a bounded and continuous k' exists" could be rephrased as "k has bounded and continuous derivative". Indeed, Definition 2 basically says (equivalently): a function $k:[0,+\infty]\to R$ is smoothly convex if and only if k is convex and has bounded and continuous derivative.

Based on the definition above, two theorems are given in Ref.[10], one for the convergence of $\{\hat{f}(y_j), j=1,2,...\}$, the other $\{y_j\}$.

Theorem 1: If the kernel $k(x)$ is smoothly convex, then the sequence $\{\hat{f}(y_j), j=1,2,...\}$ converges and monotonically increases to its limit.

***Proof*:**

From **Definition 1**, kernel $k(x)$ is bounded. Therefore, from *eq.3.3*, we also know $\{\hat{f}(y_j), j=1,2,...\}$ are bounded. To prove the theorem, we need only to verify that it is non-decreasing. Let $j=1,2,...$

(1). If $y_{j+1}=y_j$, then it is evident that $\hat{f}(y_{j+1}) \geq \hat{f}(y_j)$.

(2). If $y_{j+1} \neq y_j$, then from *eq.3.3*, we have

$$\begin{aligned}\hat{f}(y_{j+1})-\hat{f}(y_j) &= \sum_{i=1}^{N} w_i\left[K_i(y_{j+1})-K_i(y_j)\right] \\ &= \sum_{i=1}^{N} w_i\left[ck\left(\left\|y_{j+1}-X_i\right\|_{H_i}^2\right)-ck\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\right] \\ &= \sum_{i=1}^{N} w_i c\left[k\left(\left\|y_{j+1}-X_i\right\|_{H_i}^2\right)-k\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\right] \\ &\overset{(eq.3.10)}{\geq} \sum_{i=1}^{N} w_i ck'\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\left(\left\|y_{j+1}-X_i\right\|_{H_i}^2-\left\|y_j-X_i\right\|_{H_i}^2\right)\end{aligned} \qquad \textit{Eq. 3.11}$$

Let $g(x)=-k'(x)$, the above inequality equation becomes

$$\hat{f}(y_{j+1})-\hat{f}(y_j) \geq \sum_{i=1}^{N} w_i cg\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\left(\left\|y_j-X_i\right\|_{H_i}^2-\left\|y_{j+1}-X_i\right\|_{H_i}^2\right) \qquad \textit{Eq. 3.12}$$

Since

$$\begin{aligned}\left\|y_{j+1}-X_i\right\|_{H_i}^2 &= \left\|y_{j+1}-y_j+y_j-X_i\right\|_{H_i}^2 \\ &= \left\|y_{j+1}-y_j\right\|_{H_i}^2+\left\|y_j-X_i\right\|_{H_i}^2+2\left(y_{j+1}-y_j\right)^{\mathrm{T}} H_i\left(y_j-X_i\right)\end{aligned}$$

Then we could get

$$\left\|y_i - X_i\right\|_{H_i}^2 - \left\|y_{j+1} - X_i\right\|_{H_i}^2 = -\left\|y_{j+1} - y_j\right\|_{H_i}^2 - 2\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i \left(y_j - X_i\right) \qquad \textit{Eq. 3.13}$$

Hence from *eq.3.12* and *eq.3.13*, we have

$$\hat{f}\left(y_{j+1}\right) - \hat{f}\left(y_j\right) \geq \sum_{i=1}^{N} w_i cg\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left[-\left\|y_{j+1} - y_j\right\|_{H_i}^2 - 2\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i \left(y_j - X_i\right)\right] \qquad \textit{Eq. 3.14}$$

From *eq.3.9*

$$y_{j+1} = \left(\sum_{i=1}^{N} L_i\left(y_j\right)\right)^{-1} \sum_{i=1}^{N} L_i\left(y_j\right) X_i \qquad \textit{Eq. 3.15}$$

So we have

$$\left(\sum_{i=1}^{N} L_i\left(y_j\right)\right) y_{j+1} = \sum_{i=1}^{N} L_i\left(y_j\right) X_i \qquad \textit{Eq. 3.16}$$

Then by *eq.3.6*,

$$\left[\sum_{i=1}^{N} 2w_i ck'\left(\left\|y_j - X_i\right\|_{H_i}^2\right) H_i\right] y_{j+1} = \sum_{i=1}^{N} 2w_i ck'\left(\left\|y_j - X_i\right\|_{H_i}^2\right) H_i X_i \qquad \textit{Eq. 3.17}$$

Multiply both sides by $\left(y_{j+1} - y_j\right)^{\mathrm{T}} / 2$ from left,

$$\left[\sum_{i=1}^{N} w_i ck'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\right] y_{j+1} = \sum_{i=1}^{N} w_i ck'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i X_i \quad \textit{Eq. 3.18}$$

So we get

$$
\begin{aligned}
&\sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - X_i\right) \\
&= \sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i y_i - \sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i X_i \\
&= \sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i y_i - \sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i y_{j+1} \\
&= \sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - y_{j+1}\right) \\
&= -\sum_{i=1}^{N} w_i c k'\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 \\
&= \sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2
\end{aligned}
\qquad \textit{Eq. 3.19}
$$

Hence, the right side in *eq.3.14* is

$$
\begin{aligned}
&\sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left[-\left\|y_{j+1} - y_j\right\|_{H_i}^2 - 2\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - X_i\right)\right] \\
&= -\sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 - \sum_{i=1}^{N} 2 w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - X_i\right) \\
&= -\sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 - \left[-2\sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2\right] \\
&= \sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2
\end{aligned}
\qquad \textit{Eq. 3.20}
$$

Thus, inequality equation *eq.3.14* becomes

$$
\hat{f}\left(y_{j+1}\right) - \hat{f}\left(y_j\right) \geq \sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 \qquad \textit{Eq. 3.21}
$$

From the assumption on $k(x)$, we know $k'(x) < 0$, which means $g(x) > 0$. We also know $c > 0$, $w_i > 0$, $\sum_i^{-1}$ is a positive definite matrix and $H_i = \sum_i^{-1} / h^2$, so

$\sum_{i=1}^{N} w_i cg\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2$ should be positive, thus the following inequality holds when $y_{j+1} \neq y_j$:

$$\hat{f}\left(y_{j+1}\right) - \hat{f}\left(y_j\right) > 0 \qquad \textit{Eq. 3.22}$$

Therefore, the theorem is proved.

Theorem 2: If the $k(x)$ is smoothly convex, and the number of critical points of $\hat{f}\left(y_j\right)$ is finite on $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_1) \right\}$, then the iterative sequence $\left\{ y_j, j = 1, 2, ... \right\}$ converges.

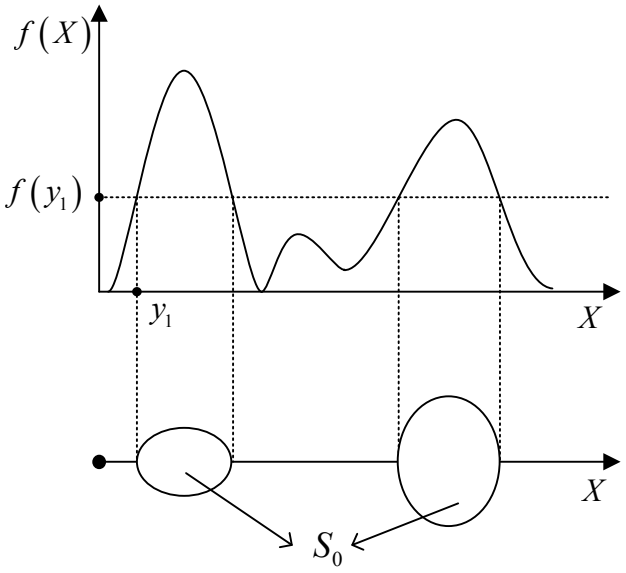

Fig. 3. 2: $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_1) \right\}$

Remark: Li, Wu and Hu[10] stated that "the number of critical points of $\hat{f}(y_j)$ is finite on $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_1) \right\}$" can always be satisfied in practice as the critical points usually represent the modes or classes in real applications. One of the goals of this thesis is to remove or reduce this technical assumption. Our results are presented and proved in the next chapter.

***Proof*:**

As for the iterative sequence $\{y_j, j = 1,2,...\}$, if there exists $j_0 > 0$ such that $y_{j_0} = y_{j_0+1}$, it can be easily seen that $y_{j_0} = y_{j_0+1} = y_{j_0+2} = \mathrm{L}$ from *eq.3.9*. Therefore, $\{y_j, j = 1,2,...\}$ converges in this case.

Assume $y_j \neq y_{j+1}$ for any $j > 0$, and let a be the minimal eigenvalue of the positive definite matrices $\{H_j\}$, then $a > 0$ and $\|y_{j+1} - X_i\|_{H_i}^2 \geq a \|y_{j+1} - X_i\|^2$. Because $k(x)$ is smoothly convex and decreasing from **Definition 1** and **2**, we have $k'(x) < 0$. We also know that $c > 0$, $w_i > 0$, $\sum_i^{-1}$ is a positive definite matrix and $H_i = \sum_i^{-1} / h^2$. Therefore, there exists $b > 0$ such that

$$\sum_{i=1}^{N} w_i c g\left(\|y - X_i\|_{H_i}^2 \right) = -\sum_{i=1}^{N} w_i c k'\left(\|y - X_i\|_{H_i}^2 \right) \geq b, \text{ for all } y \in S_0 \qquad \textit{Eq. 3.23}$$

Hence, from inequality *eq.3.21*

$$\begin{aligned}\hat{f}\left(y_{j+1}\right)-\hat{f}\left(y_j\right) &\geq \sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1}-y_j\right\|_{H_i}^2 \\ &\geq a\sum_{i=1}^{N} w_i c g\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1}-y_j\right\|^2 \\ &\geq ab\left\|y_{j+1}-y_j\right\|^2\end{aligned}$$

Therefore, from **Theorem 1**, we have

$$\left\|y_{j+1}-y_j\right\|^2 \to 0 (j\to\infty) \qquad \textit{Eq. 3.24}$$

From **Definition 2**, we know $k'(x)$ is bounded and from *eq.3.6* $L_i(X) = -2w_i c k'\left(\left\|X - X_i\right\|_{H_i}^2\right)H_i$, $\sum_{i=1}^{N} L_i\left(y_j\right)$ is also bounded. Therefore, from *eq.3.7* to *eq.3.9,*

$$\begin{aligned}\nabla\hat{f}\left(y_j\right) &= \sum_{i=1}^{N} L_i\left(y_j\right)\left[\left(\sum_{i=1}^{N} L_i\left(y_j\right)\right)^{-1}\sum_{i=1}^{N} X_i L_i\left(y_j\right) - y_j\right] \\ &= \sum_{i=1}^{N} L_i\left(y_j\right)\left(y_{j+1}-y_j\right) \qquad \textit{Eq. 3.25} \\ &\to 0 \qquad (j\to\infty)\end{aligned}$$

Because the number of critical points of $\hat{f}(X)$ is finite on $S_0 = \left\{y \middle| \hat{f}(y) \geq \hat{f}\left(y_1\right)\right\}$, without loss of generality, assume that there are m_0 critical points in S_0: $\left\{X_k', 1\leq k\leq m_0\right\}$. Of course, we have $\nabla\hat{f}\left(X_k'\right) = 0, 1\leq k\leq m_0$ and $\nabla\hat{f}(X)\neq 0, X\in S_0$ but $X\notin\left\{X_k', 1\leq k\leq m_0\right\}$.

Let

$$d_0 \overset{\Delta}{=} \min\left\{\left\|X_j' - X_k'\right\|, 1\leq j\neq k\leq m_0\right\} \qquad \textit{Eq. 3.26}$$

$$S_{\varepsilon,i} = \left\{X \middle| \left\|X - X_i'\right\| < \varepsilon, X\in S_0\right\},\ 1\leq i\leq m_0, \text{ where } 0\leq\varepsilon\leq d_0/3. \qquad \textit{Eq. 3.27}$$

From **Definition 2** and *eq.3.25*, we know $\nabla \hat{f}(X)$ is continuous and $\nabla \hat{f}(X) \neq 0$ on the bounded closed set $V_0 = S_0 - \bigcup_{i=1}^{m_0} S_{\varepsilon,i}$. Therefore, $\min_{X \in V_0} \left\| \nabla \hat{f}(X) \right\| \neq 0$, and there exists $c_\varepsilon > 0$ satisfying

$$\left\| \nabla \hat{f}(X) \right\| > c_\varepsilon \ , X \in V_0 \qquad \textit{Eq. 3.28}$$

From *eq.3.24* and *eq.3.25*, there exists $N_\varepsilon > 0$ satisfying the following two inequalities simultaneously

$$\left\| y_{j+1} - y_j \right\| < \varepsilon, j \geq N_\varepsilon \qquad \textit{Eq. 3.29}$$

$$\left\| \nabla \hat{f}(y_j) \right\| < c_\varepsilon, j \geq N_\varepsilon \qquad \textit{Eq. 3.30}$$

From *eq.3.28* and *eq.3.30*, $\left\{ y_j, j \geq N_\varepsilon \right\} \notin V_0$, we also have $V_0 = S_0 - \bigcup_{i=1}^{m_0} S_{\varepsilon,i}$, then we could get

$$\left\{ y_j, j \geq N_\varepsilon \right\} \in \bigcup_{i=1}^{m_0} S_{\varepsilon,i} . \qquad \textit{Eq. 3.31}$$

We now use mathematical induction to show that $\left\{ y_j, j \geq N_\varepsilon \right\}$ all lie in the same $S_{\varepsilon,i}$ for some *i*. Assume that we have two consecutive points y_j and y_{j+1} in two different $S_{\varepsilon,i}$: $y_j \in S_{\varepsilon,i_1}$ and $y_{j+1} \in S_{\varepsilon,i_2}$ for some $j \geq N_\varepsilon$, $1 \leq i_1 \neq i_2 \leq m_0$.

（1）If $j = N_\varepsilon$, then from *eq.3.26* and *eq.3.27*

$$
\begin{aligned}
\left\| y_{j+1} - y_j \right\| &= \left\| y_{N_\varepsilon+1} - y_{N_\varepsilon} \right\| \\
&= \left\| y_{N_\varepsilon+1} - X'_{i_2} + X'_{i_2} - X'_{i_1} + X'_{i_1} - y_{N_\varepsilon} \right\| \\
&\geq \left\| X'_{i_2} - X'_{i_1} \right\| - \left\| y_{N_\varepsilon+1} - X'_{i_2} \right\| - \left\| X'_{i_1} - y_{N_\varepsilon} \right\| \\
&\geq \min \left\| X'_{i_1} - X'_{i_2} \right\| - \max \left\| y_{N_\varepsilon+1} - X'_{i_2} \right\| - \max \left\| X'_{i_1} - y_{N_\varepsilon} \right\| \qquad \textit{Eq. 3.32} \\
&> d_0 - \varepsilon - \varepsilon \\
&> 3\varepsilon - \varepsilon - \varepsilon \\
&= \varepsilon
\end{aligned}
$$

which contradicts *eq.3.29*. Thus, y_{N_ε} and $y_{N_\varepsilon+1}$ can only lie inside the same neighborhoods S_{ε,i_0} defined around a critical point X'_{i_0}.

（2）Assume that for $m > N_\varepsilon$, y_{N_ε} ,…, y_m all lie in the same S_{ε,i_0} but $y_{m+1} \notin S_{\varepsilon,i_0}$. Assume that $y_{m+1} \in S_{\varepsilon,i_{m+1}}$. Then we have

$$
\begin{aligned}
\left\| y_{m+1} - y_m \right\| &= \left\| y_{m+1} - X'_{i_{m+1}} + X'_{i_{m+1}} - X'_{i_0} + X'_{i_0} - y_m \right\| \\
&\geq \left\| X'_{i_{m+1}} - X'_{i_0} \right\| - \left\| y_{m+1} - X'_{i_{m+1}} \right\| - \left\| X'_{i_0} - y_m \right\| \\
&\geq \min \left\| X'_{i_{m+1}} - X'_{i_0} \right\| - \max \left\| y_{m+1} - X'_{i_{m+1}} \right\| - \max \left\| X'_{i_0} - y_m \right\| \qquad \textit{Eq. 3.33} \\
&> d_0 - \varepsilon - \varepsilon \\
&> \varepsilon
\end{aligned}
$$

This is a contradiction again since we know $\left\| y_{m+1} - y_m \right\| < \varepsilon$ for $m > N_\varepsilon$ according to *eq.3.29*. Hence y_{m+1} should be also inside S_{ε,i_0}.

Thus, by induction, all $\left\{ y_j, j \geq N_\varepsilon \right\}$, $y_j = X_j^*$ can only lie inside one of the neighborhoods S_{ε,i_0} defined around a critical point X'_{i_0} and $\left\| y_j - X'_{i_0} \right\| < \varepsilon$ for $j \geq N_\varepsilon$. That means $\left\{ y_j, j = 1, 2, ... \right\}$ is convergent. Theorem 2 is proved.

Remark. Between " $\left\| y_j - X'_{i_0} \right\| < \varepsilon$ for $j \geq N_\varepsilon$ " and the sentence " That means $\left\{ y_j, j = 1, 2, ... \right\}$ is convergent. " at the end of the previous proof, there is a gap: Is i_0 independent of ε? The proof itself only shows that for every $\varepsilon > 0$, there is an i_0 and an N_ε such that $\left\| y_j - X'_{i_0} \right\| < \varepsilon$ for $j \geq N_\varepsilon$. There is no information on how i_0 depends on $\varepsilon > 0$. The following simple argument shows that once one such i_0 is found, for an $\varepsilon < d_0 / 3$, the same i_0 works for all ε' with $0 < \varepsilon' < \varepsilon$:

Claim: Let i_0 be the one index found in the previous proof and let $0 < \varepsilon' < \varepsilon$, then there is an $N_{\varepsilon'} \geq N_\varepsilon$ such that $\left\| y_j - X'_{i_0} \right\| < \varepsilon'$ for $j \geq N_{\varepsilon'}$.

Proof of the claim: As the previous proof shows, for the given $\varepsilon' > 0$, there is an i_0^* and an $N_{\varepsilon'} \geq 1$ such that $\left\| y_j - X'_{i_0^*} \right\| < \varepsilon'$ for $j \geq N_{\varepsilon'}$. We need to show that $i_0 = i_0^*$. Assume, to the contrary, that $i_0 \neq i_0^*$. Then, let $j = N_\varepsilon + N_{\varepsilon'}$. We have

$$\left\| y_j - X'_{i_0} \right\| < \varepsilon \text{ and } \left\| y_j - X'_{i_0^*} \right\| < \varepsilon' \qquad \textit{Eq. 3.34}$$

Therefore,

$$\left\| X'_{i_0} - X'_{i_0^*} \right\| \leq \left\| X'_{i_0} - y_j \right\| + \left\| y_j - X'_{i_0^*} \right\| < \varepsilon + \varepsilon' < 2\varepsilon < d_0 \qquad \textit{Eq. 3.35}$$

contradicting the definition of d_0.

This proves the claim.

CHAPTER 4: NEW RESULTS

4.1. *Need for improvement on the convergence of the sequence of mean shifts*

In **Theorem 2**, a result stated and proved in Ref. [10] and discussed in Chapter 3, there is an additional condition that "the critical points of $\hat{f}(y_j)$ is finite on $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_1) \right\}$", however, it is really difficult to make sure whether the critical point of $\hat{f}(y_j)$ is finite or not, and it is not easy for us to count the number of it anyway. Even though Li, Hu, and Wu argued in Ref. [10] that this assumption could be satisfied in many applications, we do have examples in applications where the critical points could be a whole curve. Hence we need to prove that the iterative sequence $\{y_j, j = 1, 2, ...\}$ converges without this restrictive condition. In this section, we will give a convergence result that allows S_0 contain infinitely many critical points $\{X_i', i \geq 1\}$ of $\hat{f}(X)$.

4.2. *A new result on the convergence*

Theorem 3: If the kernel function $k(x)$ is convex, and if $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_1) \right\}$ contains possibly infinitely many critical points $\{X_i', i \geq 1\}$ of $\hat{f}(X)$ with at most one accumulation point X_0', then the iterative sequence $\{y_j, j = 1, 2, ...\}$ converges.

Proof: Similarly to the proof of **Theorem 2**, if $y_j = y_{j+1}$ for any $j > 0$, then the iterative sequence $\{y_j, j = 1,2,...\}$ will be a constant sequence, therefore, convergent.

Assume $y_j \neq y_{j+1}$ for any $j > 0$. If there are only finitely many critical points in S_0, then the same proof given for Theorem 2 establishes the convergence. Now, assume that there are infinitely many critical points $\{X_i', i \geq 1\}$ of $\hat{f}(X)$ in S_0 and there is one accumulation point. For convenience, we will assume the whole sequence is convergent and $\lim_{i\to\infty} X_i' = X_0'$. Then we have $\nabla \hat{f}(X_i') = 0$ and $\nabla \hat{f}(X) \neq 0$ for $X \in S_0$, $X \notin \{X_i', i \geq 1\}$. Since $\lim_{i\to\infty} X_i' = X_0'$, for any $\eta > 0$, there exists an integer $N_\eta > 0$ such that for any $i \geq N_\eta$, $\|X_i' - X_0'\| \leq \eta$. So, there are only finitely many *i* such that $\|X_i' - X_0'\| > \eta$. For convenience, assume $\|X_i' - X_0'\| > \eta$ for all $1 \leq i < N_\eta$. Let

$$d_0 \stackrel{\Delta}{=} \min\{\|X_j' - X_k'\|, 1 \leq j \neq k < N_\eta\} > 0$$

$$d_1 \stackrel{\Delta}{=} \min\{\|X_i' - X_0'\|, 1 \leq i < N_\eta\} - \eta > 0$$

$$0 < \varepsilon(\eta) < \min\{d_0/3, d_1/2\}$$

$$S_i = \{X \,|\, \|X - X_i'\| < \varepsilon, X \in S_0\}, 1 \leq i < N_\eta$$

$$S_{\eta,0} = \{X \,|\, \|X - X_0'\| < \eta, X \in S_0\}$$

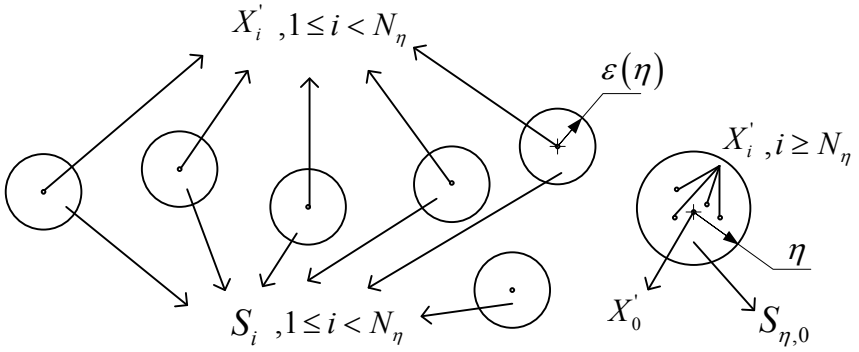

Fig.4. 1: $S_i, S_{\eta,0}, X_i^{'}$ *and* $X_0^{'}$.

Let $N = N_\eta - 1$, we know $\nabla \hat{f}(X)$ is continuous and $\nabla \hat{f}(X) \neq 0$ on the bounded closed set $V_0 = S_0 - \bigcup_{i=1}^{N} S_i - S_{\eta,0}$. Therefore, $\min_{X \in V_0} \left\| \nabla \hat{f}(X) \right\| \neq 0$, and there exists $c_\varepsilon > 0$ satisfying $\left\| \nabla \hat{f}(X) \right\| > c_\varepsilon$ for $X \in V_0$. From **Theorem 2**, we know $y_{j+1} - y_j \to 0 (j \to \infty)$ and $\nabla \hat{f}(y_j) \to 0 (j \to \infty)$. So there exists an integer $M_\varepsilon > 0$ such that $\left\| y_{j+1} - y_j \right\| < \varepsilon, j \ge M_\varepsilon$ and $\left\| \nabla \hat{f}(y_j) \right\| < c_\varepsilon, j \ge M_\varepsilon$. Hence $\{ y_j, j \ge M_\varepsilon \} \notin V_0$, then we could get $y_j \in \bigcup_{i=1}^{N} S_i + S_{\eta,0}$ for $j \ge M_\varepsilon$. Now, we show that $\{ y_j, j \ge M_\varepsilon \}$ is contained in a single S_i for some *i* with $0 \le i < N_\eta$.

If we have two consecutive points y_j and y_{j+1} in two different S_i, $1 \le i < N_\eta$. We will get a contradiction as in the proof of Theorem 2. Next, we show that y_j and y_{j+1} could not be located separately as one in S_i while another in $S_{\eta,0}$. We give the proof for such case only.

(1) For $j = M_\varepsilon$, and if $y_{j+1} \in S_i, 1 \le i < N_\eta$ but $y_j \in S_{\eta,0}$

$$\begin{aligned}\left\|y_{j+1}-y_j\right\| &= \left\|y_{M_\varepsilon+1}-y_{M_\varepsilon}\right\| \\ &= \left\|y_{M_\varepsilon+1}-X_i^{'}+X_i^{'}-X_0^{'}+X_0^{'}-y_{M_\varepsilon}\right\| \\ &\geq \left\|X_i^{'}-X_0^{'}\right\|-\left\|y_{M_\varepsilon+1}-X_i^{'}\right\|-\left\|X_0^{'}-y_{M_\varepsilon}\right\| \\ &\geq \min\left\|X_i^{'}-X_0^{'}\right\|-\left\|y_{M_\varepsilon+1}-X_i^{'}\right\|-\left\|X_0^{'}-y_{M_\varepsilon}\right\| \\ &> d_1+\eta-\varepsilon-\eta \\ &> 2\varepsilon-\varepsilon \\ &= \varepsilon\end{aligned}$$

contradicts $\left\|y_{j+1}-y_j\right\|<\varepsilon$, $j\geq M_\varepsilon$. So if y_{M_ε} is in $S_{\eta,0}$, then $y_{M_\varepsilon+1}$ must be also in $S_{\eta,0}$.

(2) Assume that for $m>M_\varepsilon$, $y_{N_\varepsilon},\ldots,y_m$ all lie in $S_{\eta,0}$, we assume $y_{m+1}\notin S_{\eta,0}$ but $y_{m+1}\in S_i$ for $1\leq i<N_\eta$, then

$$\begin{aligned}\left\|y_{m+1}-y_m\right\| &= \left\|y_{m+1}-X_i^{'}+X_i^{'}-X_0^{'}+X_0^{'}-y_m\right\| \\ &\geq \left\|X_i^{'}-X_0^{'}\right\|-\left\|y_{m+1}-X_i^{'}\right\|-\left\|X_0^{'}-y_m\right\| \\ &\geq \min\left\|X_i^{'}-X_0^{'}\right\|-\left\|y_{m+1}-X_i^{'}\right\|-\left\|X_0^{'}-y_m\right\| \\ &> d_1+\eta-\varepsilon-\eta \\ &> 2\varepsilon-\varepsilon \\ &= \varepsilon\end{aligned}$$

This result is contradicted with $\left\|y_{j+1}-y_j\right\|<\varepsilon, j\geq M_\varepsilon$, hence $y_{m+1}\in S_{\eta,0}$, by induction, $y_j(j\geq M_\varepsilon)$ will all stay in $S_{\eta,0}$.

The above process is not enough to show the iterative sequence $\left\{y_j, j=1,2,\ldots\right\}$ is convergent to $X_0^{'}$ or $X_i^{'}$ since there are infinite many critical points $X_i^{'}$ near the accumulation point $X_0^{'}$ in $S_{\eta,0}$. The iterative sequence may still converge to any one $X_i^{'}\left(i\geq N_\eta\right)$ inside the

neighborhood $S_{\eta,0}$ or may not converge at all. Hence, a further step is needed to make sure the sequence $\{y_j, j=1,2,...\}$ is either convergent to $X_i^{'}$ or $X_0^{'}$. We only need to consider two cases:

Case 1: If y_j ($j \geq M_{\varepsilon(\eta)}$, ε is indeed dependent on η) all stay in $S_i (i \geq 1)$, then the argument given after the proof of Theorem 2 shows that the sequence $\{y_j, j=1,2,...\}$ must converge to $X_i^{'}$.

Case 2: If $y_j (j \geq M_{\varepsilon(\eta)})$ all stay in $S_{\eta,0}$, then we have to repeat the proof of the first half of this proof to the subsequence $\{y_j, j \geq M_{\varepsilon(\eta)}\}$ with η replaced by $\eta/2$ to obtain that $\{y_j, j \geq M_{\varepsilon(\eta/2)}\}$ all lie in one of S_i (with the corresponding (smaller) $\varepsilon(\eta/2)$) or $S_{\eta/2,0}$, which is the neighborhood of $X_0^{'}$. Again, we need to consider two cases: (i) if it is S_i for some $i > 0$, then the sequence converges to $X_i^{'}$; (ii) if it is $S_{\eta/2,0}$, we have to repeat the argument for a sub-subsequence $\{y_j, j \geq M_{\varepsilon(\eta/4)}\}$ with $\eta/2$ replaced by $\eta/4$.

By continuing in this fashion, we will arrive at the conclusion that either the subsequence is contained in S_i for some $i > 0$ and therefore must converge to $X_i^{'}$ (please note that ε must be decreased every time as well) or it is contained in $S_{\eta/2^n,0}$. If we have to continue forever, then the sequence must satisfy the property that $\{y_j, j \geq M_{\varepsilon(\eta/2^n)}\} \in S_{\eta/2^n,0}$ for every *n=1,2,…*, which implies that the sequence $\{y_j\}$ converges to $X_0^{'}$. This completes our proof of Theorem 3.

4.3. *Generalizations of mean shifts and their convergence*

In this section, we give some generalizations of the mean shifts and establish their convergence. Our main idea is to make the choice of kernel functions more adaptive and more multidimensional.

(1) Note that the same kernel function is used for different sample points in the definition of the kernel estimation of the density. Hence the difference and the anisotropy of different samples was not taken into account during the prove process.

(2) No sufficient attention has been paid to the difference of sample contributions. As we know, the peripheral samples are less reliable since they are often more influenced by noise. Hence, different samples in different location should be ideally treated differently.

4.3.1. Multi-kernels

Assumptions and preliminaries

We assume all kernel functions $k_i(x)\,(i \geq 1)$ satisfy the requirements in **Definition 1** in Chapter 3. Recall that we have

(1) (positivity) $k_i(x) \geq 0$.

(2) (decreasing) $k_i(x_1) \geq k_i(x_2), 0 \leq x_1 \leq x_2 < +\infty$.

(3) (integrability) $\int_0^{\infty} k_i(x)dx < \infty$.

(4) (boundedness) $0 < k_i(0) < +\infty$.

Given many bounded kernel function $k_i(x)$, the density estimation of random variable x is defined as

$$\hat{f}(X) = \sum_{i=1}^{N} w_i k_i\left(\|X - X_i\|_{H_i}^2\right) \qquad \textit{Eq. 4.1}$$

Where

$$\|X - X_i\|_{H_i}^2 = (X - X_i)^{\mathrm{T}} H_i (X - X_i),$$

$$H_i = \sum\nolimits_i^{-1} / h^2, \quad H = \{H_i, 1 \le i \le n\},$$

$h > 0$ is a fixed value to ensure the size of the window (region of sample), w_i is the prior probability for x_i, which is also called the weight of sample X_i. Hence $\sum_{i=1}^{n} w_i = 1$, $w_i > 0$. The matrix $\sum_i^{-1}$ is a positive definite matrix to represent the local structure around X_i. The only difference is that a single kernel is replaced by *n* kernels. This allows us to mix different types of kernels in the density estimation. We found that all arguments used in Chapter 3 can be modified to this more general model and we now describe all the important steps in our verification.

Assume that $k_i(x)$ is differentiable, from *eq.4.1*, we could get the gradient of probability density

$$\nabla \hat{f}(X) = \sum_{i=1}^{N} 2 w_i k_i'\left(\|X - X_i\|_{H_i}^2\right) H_i (X - X_i) \qquad \textit{Eq. 4.2}$$

Eq.4.2 could be simplified as

$$\begin{aligned}\nabla \hat{f}(X) &= \sum_{i=1}^{N} L_i(X)(X_i - X) \\ &= \sum_{i=1}^{N} L_i(X) X_i - \sum_{i=1}^{n} L_i(X) X\end{aligned} \qquad \textit{Eq. 4.3}$$

where

$$L_i(X) = -2 w_i k_i'\left(\left\| X - X_i \right\|_{H_i}^2\right) H_i \qquad \textit{Eq. 4.4}$$

From the definition of $k(x)$, we could know $k_i'(x) < 0$ for $i \geq 1$. So $L_i(X)$ should also be positive definite and invertible. Then the *eq.4.3* could be rewritten as

$$\begin{aligned}\nabla \hat{f}(X) &= \sum_{i=1}^{N} L_i(X) X_i - \sum_{i=1}^{N} L_i(X) X \\ &= \sum_{i=1}^{N} L_i(X) \left[\left(\sum_{i=1}^{N} L_i(X) \right)^{-1} \sum_{i=1}^{N} L_i(X) X_i - X \right]\end{aligned} \qquad \textit{Eq. 4.5}$$

Then we get a term $ms_k(X) := \left(\sum_{i=1}^{N} L_i(X) \right)^{-1} \sum_{i=1}^{N} L_i(X) X_i - X$, which represents the mean shift vector. Let X be y_j above:

$$ms_k(y_j) = \left(\sum_{i=1}^{N} L_i(y_j) \right)^{-1} \sum_{i=1}^{N} L_i(y_j) X_i - y_j \qquad \textit{Eq. 4.6}$$

Then we get the iterative procedure of mean shift algorithm

$$y_{j+1} = y_j + ms_k(y_j) = \left(\sum_{i=1}^{N} L_i(y_j) \right)^{-1} \sum_{i=1}^{N} L_i(y_j) X_i \qquad \textit{Eq. 4.7}$$

Note that we have the same appearance (except that a single kernel is replaced by multiple kernels).

Proofs

We will prove that both the estimation of the probability density sequence $\{\hat{f}(y_j), j=1,2,...\}$ and the iterative sequence $\{y_j, j=1,2,...\}$ are convergent.

By **Definition 2**, if all $k_i(x)$ are smoothly convex, there exists $k_i^{'}(x)$ that is bounded and continuous and satisfies

$$k_i(x_2)-k_i(x_1)>k_i^{'}(x_1)(x_2-x_1), \forall x_1 \geq 0, x_2 \geq 0, x_1 \neq x_2 \qquad \textit{Eq. 4.8}$$

Theorem 4: If all $k_i(x)$ are smoothly convex, then the sequence $\{\hat{f}(y_j), j=1,2,...\}$ converges and monotonically increases to its limit.

***Proof*:**

From **Definition 1**, kernel $k_i(x)$ is bounded. Therefore, from *eq.4.1*, we also know $\{\hat{f}(y_j), j=1,2,...\}$ are bounded.

For any $j=1,2,...$

(1). If $y_{j+1}=y_j$, then it is clear that $\hat{f}(y_{j+1}) \geq \hat{f}(y_j)$.

(2). If $y_{j+1} \neq y_j$, then from *eq.4.1* and *eq.4.8*, we have

$$\begin{aligned}\hat{f}(y_{j+1})-\hat{f}(y_j) &= \sum_{i=1}^{N} w_i\left[k_i\left(\|y_{j+1}-X_i\|_{H_i}^2\right)-k_i\left(\|y_j-X_i\|_{H_i}^2\right)\right] \\ &\geq \sum_{i=1}^{N} w_i k_i^{'}\left(\|y_j-X_i\|_{H_i}^2\right)\left(\|y_{j+1}-X_i\|_{H_i}^2-\|y_j-X_i\|_{H_i}^2\right)\end{aligned}$$

Let $g_i(x)=-k_i^{'}(x)$, the above inequality equation becomes

$$\hat{f}\left(y_{j+1}\right)-\hat{f}\left(y_j\right)\geq\sum_{i=1}^{N}w_i g_i\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\left(\left\|y_j-X_i\right\|_{H_i}^2-\left\|y_{j+1}-X_i\right\|_{H_i}^2\right) \qquad \textit{Eq. 4.9}$$

We know that

$$\begin{aligned}\left\|y_{j+1}-X_i\right\|_{H_i}^2&=\left\|y_{j+1}-y_j+y_j-X_i\right\|_{H_i}^2\\&=\left\|y_{j+1}-y_j\right\|_{H_i}^2+\left\|y_j-X_i\right\|_{H_i}^2+2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)\end{aligned} \qquad \textit{Eq. 4.10}$$

Then we could get

$$\left\|y_i-X_i\right\|_{H_i}^2-\left\|y_{j+1}-X_i\right\|_{H_i}^2=-\left\|y_{j+1}-y_j\right\|_{H_i}^2-2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right) \qquad \textit{Eq. 4.11}$$

Hence from *eq.4.9* and *eq.4.11*, we have

$$\hat{f}\left(y_{j+1}\right)-\hat{f}\left(y_j\right)\geq\sum_{i=1}^{N}w_i g_i\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\left[-\left\|y_{j+1}-y_j\right\|_{H_i}^2-2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)\right] \qquad \textit{Eq. 4.12}$$

From *eq.4.7*

$$y_{j+1}=\left(\sum_{i=1}^{N}L_i\left(y_j\right)\right)^{-1}\sum_{i=1}^{N}L_i\left(y_j\right)X_i \qquad \textit{Eq. 4.13}$$

So we have

$$\left(\sum_{i=1}^{N}L_i\left(y_j\right)\right)y_{j+1}=\sum_{i=1}^{N}L_i\left(y_j\right)X_i \qquad \textit{Eq. 4.14}$$

Then by *eq.4.4*,

$$\left[\sum_{i=1}^{N}2w_i k_i'\left(\left\|y_j-X_i\right\|_{H_i}^2\right)H_i\right]y_{j+1}=\sum_{i=1}^{N}2w_i k_i'\left(\left\|y_j-X_i\right\|_{H_i}^2\right)H_i X_i \qquad \textit{Eq. 4.15}$$

Multiply both sides by $\left(y_{j+1}-y_j\right)^{\mathrm{T}}/2$ from left,

$$\left[\sum_{i=1}^{N}w_i k_i'\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\right]y_{j+1}=\sum_{i=1}^{N}w_i k_i'\left(\left\|y_j-X_i\right\|_{H_i}^2\right)\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i X_i \qquad \textit{Eq. 4.16}$$

So we get

$$
\begin{aligned}
&\sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - X_i\right) \\
&= \sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i y_i - \sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i X_i \\
&= \sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i y_i - \sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i y_{j+1} \\
&= \sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - y_{j+1}\right) \\
&= -\sum_{i=1}^{N} w_i k_i^{'}\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 \\
&= \sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2
\end{aligned}
\qquad \textit{Eq. 4.17}
$$

Hence, the right side in *eq.4.12* is

$$
\begin{aligned}
&\sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left[-\left\|y_{j+1} - y_j\right\|_{H_i}^2 - 2\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - X_i\right)\right] \\
&= -\sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 - \sum_{i=1}^{N} 2 w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left(y_{j+1} - y_j\right)^{\mathrm{T}} H_i\left(y_j - X_i\right) \\
&= -\sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 - \left[-2\sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2\right] \\
&= \sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2
\end{aligned}
\qquad \textit{Eq. 4.18}
$$

Thus, inequality equation *eq.4.12* becomes

$$
\hat{f}\left(y_{j+1}\right) - \hat{f}\left(y_j\right) \geq \sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2 \qquad \textit{Eq. 4.19}
$$

From the assumption on $k_i(x)$, we know $k_i^{'}(x) < 0$, which means $g_i(x) > 0$. We also know $w_i > 0$, $\sum_i^{-1}$ is a positive definite matrix and $H_i = \sum_i^{-1} / h^2$, so

$\sum_{i=1}^{N} w_i g_i\left(\left\|y_j - X_i\right\|_{H_i}^2\right)\left\|y_{j+1} - y_j\right\|_{H_i}^2$ should be positive, thus the following inequality holds when $y_{j+1} \neq y_j$:

$$\hat{f}\left(y_{j+1}\right) - \hat{f}\left(y_j\right) > 0 \qquad \textit{Eq. 4.20}$$

Therefore, the theorem is proved.

Theorem 5: If all $k_i(x)$ is smoothly convex, and the number of critical points of $\hat{f}\left(y_j\right)$ is finite on $S_0 = \left\{ y \middle| \hat{f}(y) \geq \hat{f}\left(y_1\right) \right\}$ or infinite but with one accumulation point, then the iterative sequence $\left\{y_j, j = 1, 2, ...\right\}$ converges.

The proof method is similar to the proof of **Theorem 2** and **Theorem 3**. Hence, to save space, we will not present the detailed proofs again.

4.3.2. Multivariate kernel functions

In the previous proofs, the probability density function is restricted as a specialized form. This is not reasonable and could not be satisfied in real application. Hence we generalize a function that could represents most forms of probability density function, then the property of convergence of this iterative method could be easily shown and allow the estimations of probability density function to have different compositions.

Assumptions and preliminaries

There are some assumptions and preliminaries we should know before the proof of convergence.

Definition 3: A continuously differentiable function of multivariable $f(x): S \subseteq \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$ is convex if

$$f\left(x_2\right) - f\left(x_1\right) \geq \left(\nabla f\left(x_1\right)\right)^{\mathrm{T}} \left(x_2 - x_1\right)$$

Definition 4. Function $K\left(t_1, t_2, \mathrm{L}\ , t_N\right)$ is bounded on $[0, +\infty)$, satisfies:

(1) (positivity) $K\left(t_1, t_2, \mathrm{L}\ , t_N\right) \geq 0$.

(2) (decreasing for every variable)

$K\left(t_1, \mathrm{L}\ , t_i, \mathrm{L}\ , t_N\right) \geq K\left(t_1, \mathrm{L}\ , t_i^*, \mathrm{L}\ , t_N\right), 1 \leq i \leq N, 0 \leq t_i \leq t_i^* < +\infty$.

(3) (integrability) $\int_0^{\infty} K\left(t_1, t_2, \mathrm{L}\ , t_N\right) dt_i < \infty$, $1 \leq i \leq N$.

(4) (boundedness) $0 < K\left(t_1, t_2, \mathrm{L}\ , t_N\right) < +\infty$.

Assume that we have a sample of size N. We estimate the density of random variable X as

$$\hat{f}\left(X\right) = K\left(\left\|X - X_1\right\|_{H_1}^2, \left\|X - X_2\right\|_{H_2}^2, ..., \left\|X - X_N\right\|_{H_N}^2\right) \qquad \textit{Eq. 4.21}$$

where

$$\left\|X - X_i\right\|_{H_i}^2 = \left(X - X_i\right)^{\mathrm{T}} H_i \left(X - X_i\right)$$

$$H_i = \sum\nolimits_i^{-1} / h^2$$

Assume that $K:\mathbb{R}^n \to \mathbb{R}$ is differentiable, from *eq.4.21*, we could get the gradient of probability density

$$\begin{aligned}\nabla \hat{f}(X) &= \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i (X - X_i) \\ &= \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X - \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X_i\end{aligned} \qquad \textit{Eq. 4.22}$$

where $\mathrm{D}_i K = \dfrac{\partial K}{\partial t_i}$. From the definition of $K(t_1, t_2, \mathrm{L}\ , t_N)$, we know that $\mathrm{D}_i K < 0$ since for every variable $K(t_1, t_2, \mathrm{L}\ , t_N)$ is decreasing. *Eq.4.22* could be written as

$$\begin{aligned}\nabla \hat{f}(x) &= \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X - \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X_i \\ &= -\sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i \left[\left(\sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i \right)^{-1} \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X_i - X \right]\end{aligned} \qquad \textit{Eq. 4.23}$$

Then we get a term $ms_k(X) = \left(\sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i \right)^{-1} \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X_i - X$, which represents the mean shift vector. Let X be y_j in $ms_k(X)$ above:

$$ms_k(y_j) = \left(\sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i \right)^{-1} \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X_i - y_j \qquad \textit{Eq. 4.24}$$

Then we get the iterative procedure of mean shift algorithm

$$y_{j+1} = y_j + ms_k(y_j) = \left(\sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i \right)^{-1} \sum_{i=1}^{N} \mathrm{D}_i K \cdot 2H_i X_i \qquad \textit{Eq. 4.25}$$

Proofs

We now prove that both the estimation of the probability density sequence $\{\hat{f}(y_j), j=1,2,...\}$ and the iterative sequence $\{y_j, j=1,2,...\}$ are convergent.

Theorem 6: If $K(t_1,t_2,\mathrm{L}\ ,t_N)$ is smoothly convex, the sequence $\{\hat{f}(y_j), j=1,2,...\}$ converges and monotonically increases to its limit.

Proof:

From **Definition 4**, kernel K is bounded. Therefore, from *eq.4.21*, we also know $\{\hat{f}(y_j), j=1,2,...\}$ is bounded.

For any $j=1,2,...$

(1) If $y_{j+1}=y_j$, then it is evident that $\hat{f}(y_{j+1})\geq\hat{f}(y_j)$.

(2) If $y_{j+1}\neq y_j$, then from *eq.4.21*, and by **Definition 3**, we have

$$\hat{f}(y_{j+1})-\hat{f}(y_j)\geq\sum_{i=1}^{N}\mathrm{D}_i K\cdot\left[\left\|y_{j+1}-X_i\right\|_{H_i}^2-\left\|y_j-X_i\right\|_{H_i}^2\right] \qquad \textit{Eq. 4.26}$$

We know that

$$\begin{aligned}\left\|y_{j+1}-X_i\right\|_{H_i}^2&=\left\|y_{j+1}-y_j+y_j-X_i\right\|_{H_i}^2\\&=\left\|y_{j+1}-y_j\right\|_{H_i}^2+\left\|y_j-X_i\right\|_{H_i}^2+2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)\end{aligned} \qquad \textit{Eq. 4.27}$$

Then we could get

$$\left\|y_{j+1}-X_i\right\|_{H_i}^2-\left\|y_i-X_i\right\|_{H_i}^2=\left\|y_{j+1}-y_j\right\|_{H_i}^2+2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right) \qquad \textit{Eq. 4.28}$$

Hence from *eq.4.26* and *eq.4.28*, we have

$$\hat{f}\left(y_{j+1}\right)-\hat{f}\left(y_j\right)\geq\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left[\left\|y_{j+1}-y_j\right\|_{H_i}^2+2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)\right] \qquad \textit{Eq. 4.29}$$

From *eq.4.25*

$$y_{j+1}=\left(\sum_{i=1}^{N}\mathrm{D}_iK\cdot 2H_i\right)^{-1}\sum_{i=1}^{N}\mathrm{D}_iK\cdot 2H_iX_i \qquad \textit{Eq. 4.30}$$

So we have

$$\left(\sum_{i=1}^{N}\mathrm{D}_iK\cdot 2H_i\right)y_{j+1}=\sum_{i=1}^{N}\mathrm{D}_iK\cdot 2H_iX_i \qquad \textit{Eq. 4.31}$$

Multiply both sides by $\left(y_{j+1}-y_j\right)^{\mathrm{T}}/2$ from left

$$\left[\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\right]y_{j+1}=\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_iX_i \qquad \textit{Eq. 4.32}$$

So we get

$$\begin{aligned}\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)&=\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-y_{j+1}\right)\\&=-\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2\end{aligned} \qquad \textit{Eq. 4.33}$$

Hence, the right side in *eq.4.29* is

$$\begin{aligned}&\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left[\left\|y_{j+1}-y_j\right\|_{H_i}^2+2\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)\right]\\&=\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2+2\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left(y_{j+1}-y_j\right)^{\mathrm{T}}H_i\left(y_j-X_i\right)\\&=\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2-2\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2\\&=-\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2\end{aligned} \qquad \textit{Eq. 4.34}$$

Thus, inequality equation *eq.4.29* becomes

$$\hat{f}\left(y_{j+1}\right)-\hat{f}\left(y_j\right)\geq-\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2 \qquad \textit{Eq. 4.35}$$

Since $\mathrm{D}_iK<0$, we also know $\sum_i^{-1}$ is a positive definite matrix and $H_i=\sum_i^{-1}/h^2$, so $-\sum_{i=1}^{N}\mathrm{D}_iK\cdot\left\|y_{j+1}-y_j\right\|_{H_i}^2$ should be positive, thus the following inequality holds when $y_{j+1}\neq y_j$:

$$\hat{f}\left(y_{j+1}\right)-\hat{f}\left(y_j\right)>0 \qquad \textit{Eq. 4.36}$$

Therefore, the theorem is proved.

Theorem 7: If $K\left(t_1,t_2,\mathrm{L}\ ,t_N\right)$ is smoothly convex, and the number of critical points of $\hat{f}\left(y_j\right)$ is finite on $S_0=\left\{y\middle|\hat{f}\left(y\right)\geq\hat{f}\left(y_1\right)\right\}$ or infinite but with one accumulation point, then the iterative sequence $\left\{y_j,j=1,2,...\right\}$ converges.

The proof method is similar to the proof of **Theorem 2** and **Theorem 3**. Hence we will not present the detailed proofs again.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

5.1. *Major Contributions*

The major contributions in this thesis are summarized as follows.

1. Fill a small gap in the argument of Ref. [10] on the convergence of the sequence of mean shifts (end of proof of Theorem 2)
2. Give a new result when the assumption of "finitely many" critical points is replaced by "infinitely many" but with only one accumulation point in the convergence of the mean shifts (Theorem 3 and its proof)
3. Give two generalization of the mean shifts to multi-kernels and to multivariate kernels and verified the convergence results (Theorems 4 and 5)

Other minor contributions include providing detailed proofs and giving a counterexample of a divergent bounded sequence whose difference of consecutive terms tends to zero.

5.2. *Future Work*

The promising results presented here warrant future investigation. Suggested future work is as follows:

1. The assumption of the "at most one accumulation points" in Theorem 3 on the convergence of the mean shifts is a good starting point when considering infinitely many critical points in S_0. One interesting problem is to see if we can remove this assumption altogether.

2. It is of interest to study the limit of the sequence of sets $S_{j-1} = \left\{ y \middle| \hat{f}(y) \geq \hat{f}(y_j) \right\}, j = 1, 2....$ From Theorem 2, this sequence is monotone (getting smaller) as j increases. It is also known that each set is closed and bounded. How is the limit of this sequence related to the set of all modes of the density function?
3. Numerical experiments need to be performed. In particular, with the multi-kernel and multivariate kernel generalizations, applications that require these extensions should be treated and compared with the single kernel case.

REFERENCES

[1] K. Fukunaga, L.D. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Trans. Inf. Theory 21 (1975) 32 – 40.

[2] E. Parzen, "On estimation of a probability density function and mode," Ann. Math. Statistics.,vol.33, pp.1065-1067, 1962.

[3] T. Cacoullos, "Estimation of a multivariate density," Ann. Inst. Statistics.Math.,vol.18, pp.179-189, 1966.

[4] Y.Z. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 790 – 799.

[5] D. Comaniciu, P. Meer. Mean Shift Analysis and Applications, IEEE Int. Conf. Computer Vision (ICCV'99), Kerkyra, Greece, 1197-1203, 1999.

[6] D. Comaniciu, V. Ramesh, P. Meer, Real-time tracking of non-rigid objects using mean shift, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2000, pp.142 – 149.

[7] D. Comaniciu, V. Ramesh, P. Meer, The variable bandwidth mean shift and data-driven scale selection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 1, 2001, pp. 438 – 445.

[8] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002).

[9] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Trans. Pattern Anal. Mach. Intell. 25 (5) (2003) 564 – 575.

[10] Xiangru Li, Zhangyi Hu, Fuchao Wu, A note on the convergence of the mean shift, Pattern Recognition , Vol. 40, pp. 1756-1762, 2007.

[11] M. Fashing, C. Tomasi, Mean shift is a bound optimization, IEEE Trans. Pattern Anal. Mach. Intell. 27 (3) (2005) 471 – 474.

[12] M. Singh, H. Arora, N. Ahuja, A robust probabilistic estimation framework for parametric image models, European Conference on Computer Vision, vol. 1, 2004, pp. 508 – 522.

[13] Shih-Hung Liao, Shih-Yu Chiu, Leu-Shing Lan, A dual-mode mean-shift algorithm, 51st Midwest Symposium on Circuits and Systems, 2008, pp.334 – 337.

[14] B.W.Silverman, Density estimation for statistics and data analysis, Chapman and Hall, ISBN:0-412-24620-1,1986

[15] Rafael C. Gonzalez, Richard E. Woods, Digital Image Processing, Addison-Wesley Publishing Company, ISBN:0-201-50803-6, TA1632.G66, 1992

[16] Maria Petrou, Panagiota Bosdogianni, Image Processing: The Fundamentals, John Wiley & Sons Ltd, ISBN:0-471-99883-4, TA1637.P48, 1999

[17] Scott L. Miller, Donald G. Childers, Probability and Random Processes With Applications to Signal Processing and Communications, Elsevier Academic Press, ISBN:0-12-172651-7(hard cover: alk. paper), TK5102.9.M556, 2004

[18] George Casella, Roger L. Berger, Statistical Inference, Brooks/Cole Publishing Company, ISBN:0-534-11958-1, QA276.C37, 1990

[19] Stephen Boyed, Lieven Vandenberghe, Convex Oprimization, Cambridge University Press, ISBN:0-521-83378-7, QA402.5.B69, 2004