On the Convergence of the Mean Shift Algorithm with the Gaussian Kernel

Youness Aliyari Ghassabeh*

Department of Mathematics and Statistics, Queen's University, Kingston, ON, K7L 3N6

Abstract

The mean shift (MS) algorithm is a non-parametric, iterative technique that has been used to find modes of an estimated probability density function (pdf). Although the MS algorithm has been widely used in many applications, such as clustering, image segmentation, and object tracking, a rigorous proof for its convergence is still missing. This paper tries to fill some of the gaps between theory and practice by presenting specific theoretical results about the convergence of the MS algorithm. To achieve this goal, first we show that all the stationary points of an estimated pdf using a certain class of kernel functions are inside the convex hull of the data set. Then the convergence of the sequence generated by the MS algorithm for an estimated pdf with isolated stationary points will be proved. Finally, we present a sufficient condition for the estimated pdf using the Gaussian kernel to have isolated stationary points.

Keywords: Mean Shift Algorithm, Mode Estimate Sequence, Convex Hull, Isolated Stationary Points, Kernel Function, Gaussian KDE, Convergence.

1. Introduction

The modes of a probability density function (pdf) play an essential role in many applications, including classification [1], clustering [2], multi-valued regression [3], image segmentation [4], and object tracking [5]. Due to the lack of knowledge about the pdf, a nonparametric technique is proposed to find an estimate for the gradient of a pdf [6]. The gradient of a pdf at a continuity point is estimated using the sample observations that fall in the vicinity of that point. By equating the gradient estimate to zero, we can find an equation for the modes of a pdf. The mean shift (MS) algorithm is a simple, non-parametric, and iterative method introduced by Fukunaga and Hostetler [6] for finding modes of an estimated pdf. The algorithm was generalized by Cheng [7] in order to show that the MS algorithm is a mode-seeking process on a surface constructed with a shadow kernel. Later, the algorithm became popular in the machine learning society when its potential usage

*Corresponding author. Phone +1613314363.
Email address: aliyari@cs.toronto.edu (Youness Aliyari Ghassabeh)

for feature space analysis was studied [4].

The MS algorithm shifts each data point to the weighted average of the data set in each iteration. It starts from one of the data points and iteratively improves the mode estimate. The algorithm can be used as a clustering tool, where each mode represents a cluster. In contrast to the k-mean clustering approach, the mean shift algorithm does not require any prior knowledge of the number of clusters and there is no assumption of the shape of the clusters. The algorithm has been successfully used for applications such as image segmentation [8, 9], edge detection [10, 11], object tracking [5, 12], information fusion [13], and noisy source vector quantization [14][15].

In spite of using the MS algorithm in different applications, a rigorous proof for the convergence of the algorithm is still missing in the literature. The authors in [4] claimed that the MS algorithm generates a convergent sequence. But a crucial step for the convergence proof of the sequence in [4] is not correct. In another work, it was shown that the MS algorithm with the Gaussian kernel is an instance of the expectation maximization (EM) algorithm and hence the generated sequence converges to a mode of the estimated pdf [17]. However, without additional conditions, the EM algorithm may not converge.

In this paper, we first show that the gradient of the estimated pdf cannot be zero outside the convex hull of the data set. The previous statement implies that all the stationary points of the estimated pdf must be inside the convex hull. Then, we consider the MS algorithm in D-dimensional space ($D \geq 1$) and prove that if the estimated pdf has isolated stationary points then the MS algorithm converges to a mode inside the convex hull of the data set. Furthermore, we provide a sufficient condition for the pdf estimate using the Gaussian kernel to have isolated stationary points.

The organization of the paper is as follows. In Section 2, a brief review of the MS algorithm is given. The incompleteness of the previously given proofs for the convergence of the MS algorithm is discussed in Section 3. The convergence proof of the MS algorithm with the isolated stationary points is given in Section 4. Furthermore, a sufficient condition to have isolated stationary points for an estimated pdf using the Gaussian kernel is given in Section 4. The concluding remarks are given in Section 5.

2. Mean shift algorithm

A D-variate kernel $K : \mathbb{R}^D \rightarrow \mathbb{R}$ is a non-negative real-valued function that satisfies the following conditions [18]

$$\int_{\mathbb{R}^D} K(\boldsymbol{x})d\boldsymbol{x} = 1, \;\; \lim_{\|\boldsymbol{x}\|\to\infty} \|\boldsymbol{x}\|^D K(\boldsymbol{x}) = 0, \int_{\mathbb{R}^D} \boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} = 0, \;\; \int_{\mathbb{R}^D} \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})d\boldsymbol{x} = c_K\boldsymbol{I},$$

where c_K is a constant and $\boldsymbol{I}$ is the identity matrix. Let $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, n$ be a sequence of n independent and identically distributed (iid) random variables. The kernel density estimate $\hat{f}$ at an arbitrary point $\boldsymbol{x}$ using a kernel $K(\boldsymbol{x})$ is given by

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{x}_i), \tag{1}$$

where $K_{\mathbf{H}}(\boldsymbol{x}) = |\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\boldsymbol{x})$, $\mathbf{H}$ is a symmetric positive definite $D \times D$ matrix called the bandwidth matrix, and $|\mathbf{H}|$ denotes the determinant of $\mathbf{H}$. A special class of kernels, called radially symmetric kernels, has been widely used for pdf estimation. Radially symmetric kernels are defined by $K(\boldsymbol{x}) = c_{k,D}k(\|\boldsymbol{x}\|^2)$, where $c_{k,D}$ is a normalization factor that causes $K(\boldsymbol{x})$ to integrate to one and $k : [0, \infty) \to [0, \infty)$ is called the profile of the kernel. The profile of a kernel is assumed to be a non-negative, non-increasing, and piecewise continuous function that satisfies $\int_0^\infty k(x)dx < \infty$. Two widely used kernel functions are the Epanechnikov kernel and the Gaussian kernel, both of which are defined by [19],

1. Epanechnikov kernel

$$K_E(\boldsymbol{x}) = \begin{cases} \frac{1}{2}c_D^{-1}(D+2)(1 - \|\boldsymbol{x}\|^2) & \text{if } \|\boldsymbol{x}\| \leq 1 \\ 0 & \text{if } \|\boldsymbol{x}\| > 1, \end{cases}$$

where c_D is the volume of the unit D-dimensional sphere.

2. Gaussian kernel

$$K_N(\boldsymbol{x}) = (2\pi)^{-D/2} \exp\Big(-\frac{\|\boldsymbol{x}\|^2}{2}\Big).$$

The probability density estimation that results from this technique is asymptotically unbiased and consistent in the mean square sense [20]. For the sake of simplicity, the bandwidth matrix $\mathbf{H}$ is chosen to be proportional to the identity matrix, i.e., $\mathbf{H} = h^2\boldsymbol{I}$. Then, by using the profile k and the bandwidth h, the estimated pdf changes to the following well-known form [19]

$$\hat{f}_{h,k}(\boldsymbol{x}) = \frac{c_{k,D}}{nh^D}\sum_{i=1}^{n} k(\|\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\|^2). \tag{2}$$

Assuming that k is differentiable with derivative k', taking the gradient of (2) yields [4]

$$\nabla\hat{f}_{h,k}(\boldsymbol{x}) = \frac{2c_{k,D}}{nh^{D+2}}\Big[\sum_{i=1}^{n} g\big(\|\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\|^2\big)\Big]\Big[\frac{\sum_{i=1}^{n}\boldsymbol{x}_i g(\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\|^2)}{\sum_{i=1}^{n} g(\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\|^2)} - \boldsymbol{x}\Big], \tag{3}$$

where $g(x) = -k'(x)$. The first term in the above equation is proportional to the density estimate at $\boldsymbol{x}$ using kernel $G(\boldsymbol{x}) = c_{g,D}g(\|\boldsymbol{x}\|^2)$. The second term is called the mean shift (MS) vector, $\boldsymbol{m}_{h,g}(\boldsymbol{x})$, and (3) can be rewritten in the following form [4]

$$\nabla \hat{f}_{h,k}(\boldsymbol{x}) = \hat{f}_{h,g}(\boldsymbol{x}) \frac{2c_{k,D}}{h^2 c_{g,D}} \boldsymbol{m}_{h,g}(\boldsymbol{x}). \tag{4}$$

The above expression indicates that the MS vector computed with bandwidth h and profile g is proportional to the normalized gradient density estimate obtained with the profile k (normalization is done by density estimate with profile g). Therefore, the MS vector always points toward the direction of the maximum increase in the density function. In fact, the MS algorithm is an instance of the gradient ascent algorithm with an adaptive step size [21].

The modes of the estimated density function are located at the zeros of the gradient function, i.e., $\nabla \hat{f}(\boldsymbol{x}) = 0$. Equating (3) to zero reveals that the modes of the estimated pdf are fixed points of the following function

$$\mathbf{m}_{h,g}(\boldsymbol{x}) + \boldsymbol{x} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\big(\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\|^2\big)}{\sum_{i=1}^{n} g\big(\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\|^2\big)}. \tag{5}$$

The MS algorithm initializes the mode estimate sequence to be one of the observed data. The mode estimate $\boldsymbol{y}_j$ in the jth iteration is updated as

$$\boldsymbol{y}_{j+1} = \boldsymbol{y}_j + \boldsymbol{m}(\boldsymbol{y}_j) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\big(\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\|^2\big)}{\sum_{i=1}^{n} g\big(\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\|^2\big)}. \tag{6}$$

The MS algorithm iterates this step until the norm of the difference between two consecutive mode estimates becomes less than some predefined threshold. Typically n instances of the MS algorithm are run in parallel, with the ith instance initialized to the ith data point.

3. Incompleteness of the previous proofs

Although the MS algorithm has been used widely in different applications, a rigorous proof for the convergence of the algorithm has not been given. The following statement is claimed to be true about the MS algorithm [4]: if the kernel K has a convex, monotonically decreasing, and bounded profile, the mode estimate sequence $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ and the sequence $\{\hat{f}_{h,k}(\boldsymbol{y}_j)\}$ converge. The authors in [4] successfully showed that the sequence $\{\hat{f}_{h,k}(\boldsymbol{y}_j)\}_{j=1,2,\ldots}$ is an increasing and convergent sequence. However, the second part of the statement of Theorem 1 in [4], which claims that the sequence $\{\boldsymbol{y}_i; i = 1, 2, \ldots\}$ converges, is not correct. The authors in [4] claimed that the mode estimate sequence $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ is a Cauchy sequence and therefore it converges to a point in the convex hull of the data set. To show that the mode estimate sequence is a Cauchy sequence, they used the following inequality (See Eq. $(A.7)$ in [4])

$$\begin{aligned} &\|\boldsymbol{y}_{j+m} - \boldsymbol{y}_{j+m-1}\|^2 + \|\boldsymbol{y}_{j+m-1} - \boldsymbol{y}_{j+m-2}\|^2 + \ldots + \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \\ &\geq \|\boldsymbol{y}_{j+m} - \boldsymbol{y}_j\|^2, \text{ for } j, m > 0. \end{aligned}$$

It is clear that the above inequality is not correct in general, hence the mode estimate sequence is not necessarily a Cauchy sequence. It is also clear that the convergence of $\hat{f}_{h,k}(\boldsymbol{y}_j)$ does not imply the convergence of $\{\boldsymbol{y}_j\}$ (the implication in the reverse direction is true since $\hat{f}$ is a continuous function). Through further manipulation of the proof in [4], it can be shown that the norm of difference between two consecutive mode estimate converges to zero, i.e., $\lim_{j\to\infty} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| = 0$, which does not imply convergence of $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$. The following inequality is proved in [4]

$$\hat{f}_{h,k}(\boldsymbol{y}_{j+1}) - \hat{f}_{h,k}(\boldsymbol{y}_j) \geq \frac{c_{k,D}}{nh^{D+2}} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \sum_{i=1}^{n} g\Big(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\Big),$$

where $g(x) = -k'(x)$, and the mode estimate $\boldsymbol{y}_j$ is defined in (6). If $k(x)$ is a convex and strictly decreasing function such that $0 < |k'(x)| < \infty$ for all $x \geq 0$, then $g(x) = -k'(x)$ is always positive. Let $M(j) = \min\{g\Big(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\Big), i = 1, \ldots, n\}$. Since $\boldsymbol{y}_j$ lies in the convex hull $\mathcal{C}$ of the data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, we have $M(j) \geq g(\frac{a^2}{h^2})$, where $a = \text{diam}\,\mathcal{C} < \infty$ is the diameter of $\mathcal{C}$. Let $\varphi = g(\frac{a^2}{h^2})$. Hence, the above inequality can be simplified as follows

$$\begin{aligned} \hat{f}_{h,k}(\boldsymbol{y}_{j+1}) - \hat{f}_{h,k}(\boldsymbol{y}_j) &\geq \frac{c_{k,D}}{nh^{D+2}} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \sum_{i=1}^{n} g\Big(\|\frac{\boldsymbol{y}_j - \boldsymbol{x}_i}{h}\|^2\Big) \\ &\geq \frac{c_{k,D}}{nh^{D+2}} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 n M(j) \\ &\geq \frac{c_{k,D}}{h^{D+2}} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \varphi. \end{aligned}$$

Therefore, we have

$$\Big(\hat{f}_{h,k}(\boldsymbol{y}_{j+1}) - \hat{f}_{h,k}(\boldsymbol{y}_j)\Big) \frac{h^{D+2}}{\varphi c_{k,D}} \geq \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\|^2 \geq 0.$$

Since $\hat{f}_{h,k}(\boldsymbol{y}_{j+1})$ is a convergent sequence [4], the limit of the left side of the above inequality as $j \to \infty$ is zero. Therefore, the following limit relation holds

$$\lim_{j\to\infty} \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| = 0. \tag{7}$$

This implies that the norm of the difference between two consecutive mode estimates converges to zero. According to the definition of the mean shift vectors, it is obvious that the mode estimate sequence $\{\boldsymbol{y}_j\}$ is in the convex hull of the data set, i.e., $\boldsymbol{y}_j \in \mathcal{C}, j = 1, 2, \ldots$. Therefore $\{\boldsymbol{y}_j\}_{j=1,2\ldots,}$ is a bounded sequence, satisfying the above limit. Despite the claim in [4], the last two properties are not enough to prove the convergence of $\{\boldsymbol{y}_j\}_{j=1,2\ldots,}$. For example, consider the sequence $\{\boldsymbol{z}_j\}_{j=1,2,\ldots} \in \mathbb{R}^2$ defined as follows

$$\boldsymbol{z}_j = \Big(\sin(2\pi \sum_{k=1}^{j} \frac{1}{k}), \cos(2\pi \sum_{k=1}^{j} \frac{1}{k})\Big),\ j = 1, 2, \ldots$$

The above sequence is bounded and satisfies the inequality

$$\|\boldsymbol{z}_j - \boldsymbol{z}_{j+1}\| \leq 2\pi \frac{1}{j+1}.$$

The left side is the length of the chord connecting two consecutive members of the sequence, and the right side is the geodesic distance along the unit circle between those two members. It can be observed that the right side of the above inequality goes to zero as $j \to \infty$, but $\{\boldsymbol{z}_j\}$ is not a convergent sequence.

In another work, Carreira-Perpiñán [17] showed that the MS algorithm with the Gaussian kernel $K(\boldsymbol{x}) = c\, e^{-\|\boldsymbol{x}\|^2}$ is an instance of the EM algorithm and claimed that this fact implies the convergence of $\{\boldsymbol{y}_j\}$. Authors in [22] incorrectly claimed the EM algorithm converges (see Theorem 2 in [22]). A counterexample in [23] shows that a sequence may satisfy all the hypotheses of Theorem 2 in [22] but converges to a unit circle instead of converging to a single point. In other words, without additional conditions, the EM algorithm may not converge (the stringent regularity conditions for the convergence of the EM algorithm has been discussed in [24]). Incidentally, the error in the original proof for the convergence of the EM algorithm in [22] and the error in the proof of the convergence of the MS algorithm in [4] are both due to the same incorrect use of the triangle inequality. Also, in a footnote in [17], Carreira-Perpiñán claimed that according to Morse theory [34], the modes of a Gaussian mixture are always isolated. But, to my knowledge, the Morse theory does not imply the isolatedness of modes of a Gaussian mixture. In fact, a general and useful condition to have a set of isolated stationary points for the estimated pdf using the Gaussian kernel still seems to be missing in the literature. Finding the number of modes of a pdf estimate using the Gaussian kernel is still an open problem and needs to be investigated.

On the positive side, the authors in [25][16] claimed that an estimated pdf has a finite number of modes and using this assumption they showed that the MS algorithm generates a convergent sequence. Unfortunately, the authors in [25][16] did not provide a proof to support their claim about the finiteness of the number of the stationary points of an estimated pdf. In another work, Carreira-Perpiñán made a claim, without proof, that the estimated pdf using the Gaussian kernel always has a finite number of stationary points [26]. However, to my knowledge, there has not been a rigorous proof in the literature to show the finiteness of the set of stationary points of the estimated pdf for commonly used kernels such as the Gaussian kernel. In two recent works, the convergence of sequence generated by the MS algorithm in the one-dimensional space ($D = 1$) was investigated [27, 28]. The authors in [27] showed that the MS algorithm with an analytic kernel (e.g., the Gaussian kernel) generates a convergent sequence in the one-dimensional space. The author in [28] proved that for the MS algorithm in the one-dimensional space with a certain class of kernel functions, there exists some $N > 0$ such that for all $j > N$ the mode estimate sequence $\{\boldsymbol{y}_j\}_{j=1,2,...}$ is a monotone and convergent sequence. The special one-dimensional case has limited use in applications and the authors in [27, 28] could not generalize the convergence results for the MS algorithm for the general D-dimensional ($D > 1$) case.

In this way, it appears that the convergence of the sequence generated by the MS algorithm in the D-dimensional ($D > 1$) case has not yet been proved.

4. Theoretical results for the convergence of the MS algorithm

In this section, we first show that all the stationary points of the estimated pdf are inside the convex hull of the data set. Then, we consider the MS algorithm with the Gaussian kernel and find a sufficient condition to have isolated stationary points. The Gaussian kernel has been widely used in various applications, and its properties have been extensively studied in the literature. Later in this section, we prove that if the stationary points of the estimated pdf are isolated then the mode estimate sequence generated by the MS algorithm converges.

4.1. Isolated stationary points using the Gaussian kernel

Let $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, n$ be the input data. From (2), the estimated pdf using the Gaussian kernel is given by $\hat{f}(\boldsymbol{x}) = c\sum_{i=1}^{n} k(\|(\boldsymbol{x} - \boldsymbol{x}_i)/h\|^2)$, where $k(x) = \exp(-x/2)$ and $c = (2\pi)^{-D/2}/(nh^D)$. Let $\mathcal{C}$

denote the convex hull of the data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. The authors in [30] showed that all the stationary points of the estimated pdf using the Gaussian kernel are inside the convex hull of the data set. In the following lemma, we prove the same result for a wide class of kernels K with a strictly decreasing and differentiable profile k.

Lemma 1. *If a kernel function K has a strictly decreasing differentiable profile k, such that $|k'(x)| > 0$ for all $x > 0$, then the gradient of the estimated pdf using the kernel K and bandwidth h is nonzero outside the convex hull of the data set.*

Lemma 1 guarantees that for a certain class of kernel functions, e.g., the Gaussian kernel, all the stationary points of the estimated pdf lie inside the convex hull $\mathcal{C}$.

Now, we are in a position to introduce a sufficient condition for the stationary points of the estimated pdf using the Gaussian kernel to be isolated. The probability density estimate using the Gaussian kernel with the covariance matrix $\boldsymbol{\Sigma}$ is given by $\hat{f}(\boldsymbol{x}) = c_N \sum_{i=1}^{n} \exp(-\frac{(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)}{2})$, where $c_N > 0$ is a normalization factor to ensure that $\hat{f}(\boldsymbol{x})$ integrates to one. The gradient and Hessian matrix of the estimated pdf are given by

$$\nabla \hat{f}(\boldsymbol{x}) = c_N \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}) \exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)/2),$$
$$\boldsymbol{H}(\boldsymbol{x}) = c_N \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(-\boldsymbol{I} + (\boldsymbol{x}-\boldsymbol{x}_i)(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1}) \exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)/2).$$

Let

$$C(\boldsymbol{x}) = \sum_{i=1}^{n} \exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)/2),$$
$$\boldsymbol{A}(\boldsymbol{x}) = \sum_{i=1}^{n} (\boldsymbol{x}-\boldsymbol{x}_i)(\boldsymbol{x}-\boldsymbol{x}_i)^T \exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)/2).$$

Let S denote the set of stationary points of the estimated pdf, i.e., $S = \{\boldsymbol{x}^* : \nabla \hat{f}(\boldsymbol{x}^*) = \boldsymbol{0}\}$. Since $\hat{f}(\boldsymbol{x})$ has partial derivatives of arbitrarily high order, we have the following lemma

Lemma 2. *If the Hessian matrix at the stationary points is of full rank, the stationary points are isolated.*

We provide a sufficient condition for $\boldsymbol{\Sigma}$ such that the Hessian matrix at the stationary points has full rank. If the Hessian matrix $\boldsymbol{H}$ is not full rank, then there exists a vector $\boldsymbol{v} \neq \boldsymbol{0}$ such that $\boldsymbol{H}\boldsymbol{v} = \boldsymbol{0}$. This is equivalent to $\boldsymbol{A}(\boldsymbol{x})\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = C(\boldsymbol{x})\boldsymbol{v}$. By expanding the last equality, we obtain

$$\Big(\boldsymbol{x}\boldsymbol{x}^T C(\boldsymbol{x}) - 2\boldsymbol{x}\sum_{i=1}^{n} \boldsymbol{x}_i^T \exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)/2) \\ + \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \exp(-(\boldsymbol{x}-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)/2)\Big)\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = C(\boldsymbol{x})\boldsymbol{v}. \quad (8)$$

By definition, at a stationary point $\boldsymbol{x}^*$, we have

$$\boldsymbol{x}^* = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i \exp(-\frac{(\boldsymbol{x}^*-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{x}_i)}{2})}{C(\boldsymbol{x}^*)}. \quad (9)$$

Then, equation (8) at a stationary point $\boldsymbol{x}^*$ can be simplified to

$$\overbrace{\Big(-\boldsymbol{x}^*\boldsymbol{x}^{*T} C(\boldsymbol{x}^*) + \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \exp(-(\boldsymbol{x}^*-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{x}_i)/2)\Big)}^{\boldsymbol{B}(\boldsymbol{x}^*)}\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = C(\boldsymbol{x}^*)\boldsymbol{v}. \quad (10)$$

The above equality implies that if the Hessian matrix is not of full rank at a stationary point $\boldsymbol{x}^*$, then $C(\boldsymbol{x}^*)$ is an eigenvalue of $\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}$.

Let $\boldsymbol{\Sigma}$ be a symmetric, positive definite matrix. We show that if $\boldsymbol{\Sigma}$ satisfies a certain condition, then $C(\boldsymbol{x}^*)$ can never be an eigenvalue of $\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}$. We need the following lemmas

Lemma 3. *Let $\boldsymbol{\Sigma}$ be a nonsingular $D \times D$ matrix and $\boldsymbol{x} \in \mathbb{R}^D$. Then, for any $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}$ has rank one and its only nonzero eigenvalue $\hat{\lambda}$ is $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$.*

Lemma 4. **[31]** *Let $\|.\|$ be any matrix norm on $\mathbb{C}^{D\times D}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_D$ be the (real or complex) eigenvalues of $\boldsymbol{A} \in \mathbb{C}^{D\times D}$. Then, we have*

$$\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|,$$

where $\rho(\boldsymbol{A})$ is the spectral radius of $\boldsymbol{A}$ and is defined as $\rho(\boldsymbol{A}) = \max_i |\lambda_i|$.

Lemma 5. **[31]** *Let $\boldsymbol{A}$ be a $D \times D$ matrix. Let $\boldsymbol{A}^*$ denote the conjugate transpose of $\boldsymbol{A}$. Then $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$ have the same eigenvalues.*

Lemma 6. **[31]** *Let $\boldsymbol{A}$ be a $D \times D$ Hermitian matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$. Then*

$$\max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{x}\|^2} = \lambda_1.$$

Lemma 7. **[32]** *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $D \times D$ Hermitian matrices. Then we have the following inequality*

$$\lambda_{max}(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_{max}(\boldsymbol{A}) + \lambda_{max}(\boldsymbol{B}),$$

where λ_{max} denotes the largest eigenvalue.

Lemma 8. **[31]** *Let $\boldsymbol{A}$ be an arbitrary $D \times D$ matrix. Then the induced matrix norm by L_2 vector norm is given by*

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{max}(\boldsymbol{A}^*\boldsymbol{A})},$$

where $\|\boldsymbol{A}\|_2$ is also called the spectral norm.

From Lemma 8, the spectral norm of a $D \times D$ matrix $\boldsymbol{A}$ induced by L_2 vector norm is given by

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{max}(\boldsymbol{A}^*\boldsymbol{A})},$$

where λ_{max} denotes the largest eigenvalue of $\boldsymbol{A}^*\boldsymbol{A}$. Note that $\boldsymbol{A}^*\boldsymbol{A}$ is a positive semi-definite matrix, therefore $\lambda_{max} \geq 0$. Using the triangle inequality for the norm of any two $D \times D$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have $\|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ [31]. Using Lemma 4 and triangle inequality for spectral norm, we have

$$\begin{aligned}
\rho(\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}) &\leq \|\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}\|_2 \\
&= \| -\boldsymbol{x}^*{\boldsymbol{x}^*}^T C(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1} + \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T \exp(-(\boldsymbol{x}^*-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{x}_i)/2)\boldsymbol{\Sigma}^{-1}\|_2 \\
&\leq \|\boldsymbol{x}^*{\boldsymbol{x}^*}^T C(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}\|_2 + \sum_{i=1}^{n} \|\boldsymbol{x}_i\boldsymbol{x}_i^T \exp(-(\boldsymbol{x}^*-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{x}_i)/2)\boldsymbol{\Sigma}^{-1}\|_2 \\
&= C(\boldsymbol{x}^*)\|\boldsymbol{x}^*{\boldsymbol{x}^*}^T\boldsymbol{\Sigma}^{-1}\|_2 + \sum_{i=1}^{n} \exp(-(\boldsymbol{x}^*-\boldsymbol{x}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^*-\boldsymbol{x}_i)/2)\|\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\|_2. \quad (11)
\end{aligned}$$

Using Lemma 5 for $\|\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\|_2, i = 1, 2, \ldots, n$, we have

$$\begin{aligned}
\|\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\|_2 &= \sqrt{\lambda_{max}(\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1})} = \sqrt{\lambda_{max}(\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T)} \\
&= \sqrt{\lambda_{max}(\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-2}\boldsymbol{x}_i\boldsymbol{x}_i^T)} = a_i\sqrt{\lambda_{max}(\boldsymbol{x}_i\boldsymbol{x}_i^T)} = a_i\sqrt{\|\boldsymbol{x}_i\|^2} = a_i\|\boldsymbol{x}_i\|, \quad (12)
\end{aligned}$$

where $a_i = \sqrt{\boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-2} \boldsymbol{x}_i}$ and $\|\boldsymbol{x}_i\|^2$ is the largest eigenvalue of $\boldsymbol{x}_i \boldsymbol{x}_i^T$. Combining (11) and (12), we obtain

$$\begin{aligned}
\rho(\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}) &\le C(\boldsymbol{x}^*) a^* \|\boldsymbol{x}^*\| + \sum_{i=1}^{n} \exp(-(\boldsymbol{x}^* - \boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^* - \boldsymbol{x}_i)/2) a_i \|\boldsymbol{x}_i\| \\
&\le C(\boldsymbol{x}^*) a^* \|\boldsymbol{x}^*\| + a_{max} \|\boldsymbol{x}_{max}\| \sum_{i=1}^{n} \exp(-(\boldsymbol{x}^* - \boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}^* - \boldsymbol{x}_i)/2) \\
&= C(\boldsymbol{x}^*) a^* \|\boldsymbol{x}^*\| + a_{max} \|\boldsymbol{x}_{max}\| C(\boldsymbol{x}^*) \\
&\le C(\boldsymbol{x}^*) a^* \|\boldsymbol{x}_{max}\| + a_{max} \|\boldsymbol{x}_{max}\| C(\boldsymbol{x}^*), \qquad (13)
\end{aligned}$$

where $a^* = \sqrt{\boldsymbol{x}^{*T} \boldsymbol{\Sigma}^{-2} \boldsymbol{x}^*}$, $a_{max} = \max_i a_i$, and $\|\boldsymbol{x}_{max}\| = \max_i \|\boldsymbol{x}_i\|$. Let $\|\boldsymbol{x}^*\|^2 = b$ (b is unknown but less than $\|\boldsymbol{x}_{max}\|^2$), then from Lemma 6, $a^* \le \sqrt{b \lambda_{max}(\boldsymbol{\Sigma}^{-2})} \le \|\boldsymbol{x}_{max}\| \lambda_{max}(\boldsymbol{\Sigma}^{-1})$.

If $\|\boldsymbol{x}_{max}\|^2 \lambda_{max}(\boldsymbol{\Sigma}^{-1}) + a_{max} \|\boldsymbol{x}_{max}\| < 1$, then we observe that $\rho(\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}) < C(\boldsymbol{x}^*)$. This means $C(\boldsymbol{x}^*)$ cannot be an eigenvalue of $\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}$, which contradicts (10). Therefore, we have the following result.

Lemma 9. *Let $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, n$. Let $\|\boldsymbol{x}_{max}\|^2$ denote the largest norm among all $\boldsymbol{x}_i, i = 1, \ldots, n$. Let $a_{max} = \max_i \sqrt{\boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-2} \boldsymbol{x}_i}$. Let $\hat{f}(\boldsymbol{x})$ denote the estimated pdf using the Gaussian kernel with the covariance matrix $\boldsymbol{\Sigma}$. If $\|\boldsymbol{x}_{max}\|^2 \lambda_{max}(\boldsymbol{\Sigma}^{-1}) + a_{max} \|\boldsymbol{x}_{max}\| < 1$, then the Hessian matrix of the estimated pdf at the stationary points is of full rank and the stationary points are isolated.*

Remark. Note that for the special case that $\boldsymbol{\Sigma} = h^2 \boldsymbol{I}$, using Lemma 3 we know the only nonzero eigenvalue of $\boldsymbol{x}_i \boldsymbol{x}_i^T / h^2, i = 1, \ldots, n$ is equal to $\boldsymbol{x}_i^T \boldsymbol{x}_i / h^2$. Then using Lemma 7, we obtain

$$\begin{aligned}
\lambda_{max}(\boldsymbol{B}(\boldsymbol{x}^*)/h^2) &\le \lambda_{max}(-\boldsymbol{x}^* \boldsymbol{x}^{*T} C(\boldsymbol{x}^*)/h^2) \\
&\quad + \sum_{i=1}^{n} \lambda_{max}(\boldsymbol{x}_i \boldsymbol{x}_i^T \exp(-(\boldsymbol{x}^* - \boldsymbol{x}_i)^T (\boldsymbol{x}^* - \boldsymbol{x}_i)/(2h^2))/h^2) \\
&\le \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i / h^2 \exp(-(\boldsymbol{x}^* - \boldsymbol{x}_i)^T (\boldsymbol{x}^* - \boldsymbol{x}_i)/(2h^2)) \\
&\le \|\boldsymbol{x}_{max}\|^2 C(\boldsymbol{x}^*)/h^2, \qquad (14)
\end{aligned}$$

where $\lambda_{max}(\boldsymbol{A})$ denotes the largest eigenvalue of $\boldsymbol{A}$ and $\|\boldsymbol{x}_{max}\|^2 = \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\|^2$.

If $\|\boldsymbol{x}_{max}\|^2/h^2 < 1$, then we observe from (14) that $\lambda_{max}(\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}) < C(\boldsymbol{x}^*)$. This means $C(\boldsymbol{x}^*)$ cannot be an eigenvalue of $\boldsymbol{B}(\boldsymbol{x}^*)\boldsymbol{\Sigma}^{-1}$, which contradicts equation (10). Therefore, we have the following result

Lemma 10. *Let $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, n$. Let $\hat{f}(\boldsymbol{x})$ denote the estimated pdf using the Gaussian kernel with the covariance matrix $\boldsymbol{\Sigma} = h^2\boldsymbol{I}$. Let $\|\boldsymbol{x}_{max}\|^2 = \max_{i=1,\ldots,n} \|\boldsymbol{x}_i\|^2$. If $\|\boldsymbol{x}_{max}\| < h$, then the Hessian matrix of the estimated pdf at the stationary points is of full rank and the stationary points are isolated.*

Note that the number of the stationary points of a Gaussian kernel density estimate would be invariant under centering or location shift of data points. Thus, one can can expect that if $\max_{1\leq i\leq n} \|\boldsymbol{x}_i - \boldsymbol{x}_0\| < h$ for some proper centering point $\boldsymbol{x}_0$, then Lemma 10 would hold and the sufficient condition will be more practical.

4.2. Convergence proof when the set of stationary points is finite

Assuming that the stationary points are isolated, then the total number of stationary points of the estimated pdf inside the convex hull $\mathcal{C}$ cannot be infinite. Since the stationary points are inside the closed and bounded set $\mathcal{C}$, an infinite number of stationary points would have a convergent subsequence whose limit would not be isolated. By continuity, the limit point is also a stationary point and it is not isolated, which contradicts the fact that each stationary point is isolated. Hence, the number of stationary points is finite. Next, we show that when the stationary points of the estimated pdf are isolated, then the mode estimate sequence $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ is a convergent sequence. We prove the following theorem

Theorem 1. *Let $\boldsymbol{x}_i \in \mathbb{R}^D, i = 1, \ldots, n$. Assume that the stationary points of the estimated pdf are isolated. Then the mode estimate sequence $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ converges.*

Proof. Let $\mathcal{C}$ denote the convex hull of the data set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Let S denote the set of stationary points of the estimated pdf $\hat{f}_{h,k}$, i.e., $S = \{\boldsymbol{x}_i^* : \|\nabla \hat{f}_{h,k}(\boldsymbol{x}_i^*)\| = 0\}$. Let ζ be the smallest distance between the points in S, i.e., $\zeta = \min\{\|\boldsymbol{x}_i^* - \boldsymbol{x}_j^*\| : \boldsymbol{x}_i^*, \boldsymbol{x}_j^* \in S, i \neq j\}$. Since S is finite, we have $\zeta > 0$. Let $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ be the mode estimate sequence generated by the MS algorithm. From the definition, it is clear that the mode estimate sequence $\{\boldsymbol{y}_j\}_{j=1,2,\ldots}$ is always inside the convex hull $\mathcal{C}$. Equation (7) implies that there exists $N_1 > 0$ such that $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \frac{\zeta}{3}$ for all $j \geq N_1$. Combining (4), (6), and (7), we get

$$\lim_{j\to\infty} \nabla \hat{f}_{h,k}(\boldsymbol{y}_j) = \boldsymbol{0}. \tag{15}$$

Assume $S = \{\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_M^*\}$ and define $B(\boldsymbol{x}_i^*, \zeta/3)$ as the open ball of radius $\zeta/3$ centered at $\boldsymbol{x}_i^*$. Then the gradient of the estimated pdf outside of these balls is nonzero, i.e., $\nabla \hat{f}_{h,k}(\boldsymbol{y}_j) \neq 0,\ \boldsymbol{y}_j \notin B(\boldsymbol{x}_i^*, \zeta/3),\ i = 1, 2, \ldots, M$. If these open balls are removed from the convex hull of the data set, then the remaining set is compact. The norm of the gradient is a continuous function and attains its minimum value, say c, over this compact set. From (15), we can find N_2 such that $\|\nabla \hat{f}_{h,k}(\boldsymbol{y}_j)\| < c$ for all $j \geq N_2$. Thus for $j \geq N_2$, $\boldsymbol{y}_j$ cannot be outside $\bigcup_{i=1}^{M} B(\boldsymbol{x}_i^*, \zeta/3)$. Letting $N = \max\{N_1, N_2\}$, we will prove that for all $j > N$, if $\boldsymbol{y}_j \in B(\boldsymbol{x}_i, \zeta/3)$ then $\boldsymbol{y}_{j+1} \in B(\boldsymbol{x}_i, \zeta/3)$. We know that for $j \geq N$, $\boldsymbol{y}_{j+1} \in \bigcup_{i=1}^{M} B(\boldsymbol{x}_i^*, \zeta/3)$. Assume $\boldsymbol{y}_{j+1} \in B(\boldsymbol{x}_k^*, \zeta/3),\ k \neq i$. Then by the triangle inequality

$$\|\boldsymbol{x}_k^* - \boldsymbol{x}_i^*\| = \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_{j+1} + \boldsymbol{x}_k^* - \boldsymbol{x}_i^* + \boldsymbol{y}_j - \boldsymbol{y}_j\| \leq \|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| + \|\boldsymbol{y}_j - \boldsymbol{x}_i^*\| + \|\boldsymbol{x}_k^* - \boldsymbol{y}_{j+1}\|.$$

Since by definition of ζ, $\|\boldsymbol{x}_k^* - \boldsymbol{x}_i^*\| \geq \zeta$ and by assumption $\|\boldsymbol{y}_j - \boldsymbol{x}_i^*\| \leq \zeta/3$ and $\|\boldsymbol{y}_{j+1} - \boldsymbol{x}_k^*\| \leq \zeta/3$, we have

$$\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| \geq \|\boldsymbol{x}_k^* - \boldsymbol{x}_i^*\| - \|\boldsymbol{y}_j - \boldsymbol{x}_i^*\| - \|\boldsymbol{x}_k^* - \boldsymbol{y}_{j+1}\| \geq \zeta - \frac{\zeta}{3} - \frac{\zeta}{3} = \frac{\zeta}{3}.$$

This contradicts that $\|\boldsymbol{y}_{j+1} - \boldsymbol{y}_j\| < \frac{\zeta}{3}$. Therefore if $\boldsymbol{y}_j \in B(\boldsymbol{x}_i^*, \zeta/3)$, then $\boldsymbol{y}_{j+1} \in B(\boldsymbol{x}_i^*, \zeta/3)$ for all $j \geq N$. Since $\boldsymbol{y}_j \in \bigcup_{i=1}^{M} B(\boldsymbol{x}_i^*, \zeta/3)$ for all $j \geq N$, we obtain that there is an index i such that $\boldsymbol{y}_j \in B(\boldsymbol{x}_i^*, \zeta/3)$ for all $j \geq N$. Since $\|\nabla \hat{f}_{h,K}(\boldsymbol{y}_j)\| \to 0$ and $\boldsymbol{x}_i^*$ is the unique zero of $\|\nabla \hat{f}_{h,K}\|$ in $B(\boldsymbol{x}_i^*, \zeta/3)$, by the continuity of $\|\nabla \hat{f}_{h,K}\|$ we have $\lim_{j\to\infty} \boldsymbol{y}_j = \boldsymbol{x}_i^*$. □

Theorem 1 guarantees the convergence of the mode estimate sequence when the modes of the estimated pdf are isolated. Lemma 10 also provides sufficient conditions to have isolated stationary points. Using a fully parameterized $\boldsymbol{\Sigma}$ increases the computational complexity of the Gaussian pdf estimate. Furthermore, finding a covariance matrix $\boldsymbol{\Sigma}$ that satisfies the sufficient condition in Lemma 9 is a challenging task, especially when the size of the input data set is large. Therefore, in practice in order to reduce the computational cost, the covariance matrix $\boldsymbol{\Sigma}$ is chosen either as a diagonal matrix $\boldsymbol{\Sigma} = diag(h_1^2, h_2^2, \ldots, h_D^2)$ or is proportional to the identity matrix $\boldsymbol{\Sigma} = h^2\boldsymbol{I}$. The main advantage of the latter case is that only one parameter, h (the bandwidth), needs to be set in advance. When the covariance matrix is chosen proportional to the identity matrix, Lemma 10 states that the modes of the Gaussian pdf estimate are isolated if $h^2 \geq \|\boldsymbol{x}_{max}\|^2$. Choosing a large value of the bandwidth h generates a smooth pdf estimate with low estimation variance, at the expense of introducing a large bias into the estimation [19]. The latter is not practically desirable, since a large bias will lead to a poor estimation of the pdf that results in an inaccurate mode estimate. Furthermore, it has been shown that conditions for the consistency[1] of such a Gaussian pdf estimate are $h_n \to 0$ and $nh_n \to \infty$,

[1] A consistent pdf estimate $\hat{f}(\boldsymbol{x})$ is an estimator having the property that as the number of data points increases indefinitely, the

as $n \to \infty$ [19]. It is clear that the first consistency condition contradicts the sufficient condition given in Lemma 10.

Therefore, the theoretical conditions provided by Lemma 9 and Lemma 10 for a Gaussian pdf estimate to have isolated stationary points are of limited use in practice.

5. Conclusion

In this paper, we first reviewed the given proofs in the literature for the convergence of the MS algorithm and discussed their incompleteness. Then we studied some theoretical properties of the MS algorithm. In particular, we showed that for a certain class of kernel functions, the gradient of the estimated pdf is always nonzero outside the convex hull of the data set. Then we proved that the MS algorithm with isolated stationary points generates a convergent sequence. We also provided a sufficient condition for the MS algorithm with the Gaussian kernel to have isolated stationary points. Although the given conditions in Lemma 9 and Lemma 10 guarantee for a Gaussian pdf estimate to have isolated stationary points, they have limited use in practice. Specifically, satisfying the sufficient condition in Lemma 10 will generate a biased estimation of the pdf that leads to inaccurate mode estimates. Unfortunately, a general and useful condition that leads to a set of isolated stationary points of the estimated pdf for commonly used kernels (e.g., the Gaussian kernel) still seems to be missing.

Acknowledgment

The author would like to thank the anonymous first reviewer for his/her valuable comments and suggestions that improved the quality of the paper.

Appendix

Proof of Lemma 1.

Let $\boldsymbol{t} \notin \mathcal{C}$ be an arbitrary point outside the convex hull $\mathcal{C}$. Since the input data is a finite set, $\mathcal{C}$ is a bounded closed set. Therefore, there exists $\boldsymbol{x}_0 \in \mathcal{C}$ such that $\boldsymbol{x}_0$ has the smallest distance to $\boldsymbol{t}$

$$d(\boldsymbol{x}_0, \boldsymbol{t}) = \inf_{\boldsymbol{x} \in \mathcal{C}} d(\boldsymbol{x}, \boldsymbol{t}) > 0,$$

resulting sequence of estimates converges in probability to $f(\boldsymbol{x})$.

where $d(\boldsymbol{x}, \boldsymbol{t}) = \|\boldsymbol{x} - \boldsymbol{t}\|$. Since the profile function k is strictly decreasing and $|k'(x)| > 0$, we have $k'(x) < 0, x \in (0, \infty)$. The estimated pdf and the gradient of the estimated pdf at point $\boldsymbol{t} \notin \mathcal{C}$ are computed as follows

$$\begin{aligned} \hat{f}(\boldsymbol{t}) &= c \sum_{i=1}^{n} k(\|(\boldsymbol{t} - \boldsymbol{x}_i)/h\|^2) \\ \nabla \hat{f}(\boldsymbol{t}) &= \frac{c}{h^2} \sum_{i=1}^{n} 2(\boldsymbol{t} - \boldsymbol{x}_i) k'(\|(\boldsymbol{t} - \boldsymbol{x}_i)/h\|^2). \end{aligned} \tag{16}$$

The directional derivative $D_{\boldsymbol{u}}$ in the direction of the unit vector $\boldsymbol{u} = \frac{\boldsymbol{x}_0 - \boldsymbol{t}}{\|\boldsymbol{x}_0 - \boldsymbol{t}\|}$ at point $\boldsymbol{t}$ is given by

$$D_{\boldsymbol{u}}(\boldsymbol{t}) = \nabla \hat{f}(\boldsymbol{t}) \cdot \boldsymbol{u}, \tag{17}$$

where $\boldsymbol{x} \cdot \boldsymbol{y}$ denotes the inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^D$. We will show that $D_{\boldsymbol{u}}(\boldsymbol{t})$ is positive. Because the profile k is a strictly decreasing function, we have

$$k'(\|(\boldsymbol{t} - \boldsymbol{x}_i)/h\|^2) < 0.$$

It follows from (16) that it suffices to show that $(\boldsymbol{t} - \boldsymbol{x}_i) \cdot \boldsymbol{u} < 0, i = 1, \ldots, n$. According to the separating hyperplane theorem [33], there exists a hyperplane P with normal vector $\boldsymbol{u} = \frac{\boldsymbol{x}_0 - \boldsymbol{t}}{\|\boldsymbol{x}_0 - \boldsymbol{t}\|}$ that contains $\boldsymbol{x}_0$ and separates $\boldsymbol{t}$ and $\mathcal{C}$. The hyperplane P is defined by

$$\begin{aligned} P &= \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{x}_0) \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) = 0\} \\ &= \{\boldsymbol{x} : \boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) = c\}, \end{aligned}$$

where $c = \boldsymbol{x}_0 \cdot (\boldsymbol{x}_0 - \boldsymbol{t})$. Let P_- and P_+ be the half spaces separated by the hyperplane P such that $\mathcal{C} \subset P_+$ and $\boldsymbol{t} \in P_-$, i.e., $P_+ = \{\boldsymbol{x} : \boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) \geq c\}$ and $P_- = \{\boldsymbol{x} : \boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) \leq c\}$. Consider a new hyperplane $\hat{P}$ with the same normal vector $\boldsymbol{u}$ that contains $\boldsymbol{t}$. The new hyperplane $\hat{P}$ is parallel to P and is defined by

$$\begin{aligned} \hat{P} &= \{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{t}) \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) = 0\} \\ &= \{\boldsymbol{x} : \boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) = \hat{c}\}, \end{aligned}$$

where $\hat{c} = \boldsymbol{t} \cdot (\boldsymbol{x}_0 - \boldsymbol{t})$. The half spaces $\hat{P}_-$ and $\hat{P}_+$ corresponding to $\hat{P}$ are $\hat{P}_+ = \{\boldsymbol{x} : \boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) \geq \hat{c}\}$ and $\hat{P}_- = \{\boldsymbol{x} : \boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) \leq \hat{c}\}$. Since $\mathcal{C} \subset P_+$, we have $\boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) \geq c$ for $\boldsymbol{x} \in \mathcal{C}$. Since $\hat{c} + \|\boldsymbol{x}_0 - \boldsymbol{t}\|^2 = c$, we obtain $\boldsymbol{x} \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) > \hat{c}$ for all $\boldsymbol{x} \in \mathcal{C}$. The last inequality naturally holds for $\boldsymbol{x} = \boldsymbol{x}_i, i = 1, \ldots, n$, so that

$$\boldsymbol{x}_i \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) > c - (\boldsymbol{x}_0 - \boldsymbol{t}) \cdot (\boldsymbol{x}_0 - \boldsymbol{t}),$$

which is easily seen to be equivalent to

$$(\boldsymbol{x}_i - \boldsymbol{t}) \cdot (\boldsymbol{x}_0 - \boldsymbol{t}) > 0, \ i = 1, \ldots, n.$$

From the above inequality and equations (16) and (17), we conclude that $D_{\mathbf{u}}(\boldsymbol{t}) > 0$ for all $\boldsymbol{t} \notin \mathcal{C}$. Therefore, the gradient of the estimated pdf cannot be zero outside of the convex hull, so all stationary points of $\hat{f}(\boldsymbol{x})$ must lie in $\mathcal{C}$[2]. □

Proof of Lemma 2.

This result can be deduced from the inverse function theorem [35]. The inverse function theorem states that if $f : \mathbb{R}^n \to \mathbb{R}^n$ is a continuously differentiable function on some open set containing $\mathbf{a} \in \mathbb{R}$, such that $|Jf(\mathbf{a}) \neq \mathbf{0}|$, where J denotes the Jacobian of f, then there is some open set V containing $\mathbf{a}$ and an open W containing $f(\mathbf{a})$ such that $f : V \to W$ has a continuous inverse $f^{-1} : W \to V$, which is differentiable for all $\boldsymbol{y} \in W$. Therefore, if f denotes the pdf estimate, then the Hessian matrix is the Jacobian of the gradient of f. If the Hessian matrix is of full rank at some stationary point $\boldsymbol{x}^*$, then its determinant is nonzero and, based on the inverse function theorem, the stationary point $\boldsymbol{x}^*$ is isolated. □

Proof of Lemma 3.

First, we show that $\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}$ has rank one. Since Rank$(\boldsymbol{\Sigma}) = D$ and Rank$(\boldsymbol{x}\boldsymbol{x}^T) = 1$, we have [36]

$$\text{Rank}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}) \leq \min\{\text{Rank}(\boldsymbol{x}\boldsymbol{x}^T), \text{Rank}(\boldsymbol{\Sigma}^{-1})\} = 1. \tag{18}$$

Also, according to the Sylvester's rank inequality [37], we have

$$\text{Rank}(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}) \geq \text{Rank}(\boldsymbol{\Sigma}^{-1}) + \text{Rank}(\boldsymbol{x}\boldsymbol{x}^T) - D = 1. \tag{19}$$

Using (18) and (19), Rank$(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}) = 1$. Assume $\boldsymbol{y}$ is an eigenvector of $\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}$ so that $\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \lambda\boldsymbol{y}$. If $\lambda \neq 0$, then $\lambda\boldsymbol{y} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y})\boldsymbol{x}$, so $\boldsymbol{y}$ is a constant multiple of $\boldsymbol{x}$. Setting $\boldsymbol{y} = \boldsymbol{x}$, we obtain that

[2] An alternative proof for the special case of the Gaussian kernel can be found at [29].

$\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ is the only nonzero eigenvalue of $\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}$. □

References

[1] S. Avidan, “Ensemble tracking,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 261-271, Feb. 2007.

[2] J. Li, S. Ray, B. G. Lindsay, “A nonparametric statistical approach to clustering via mode identification,” *Journal of Machine Learning Research,* vol. 8, pp. 1687-1723, Aug. 2007.

[3] J. Einbeck, G. Tutz, “The fitting of multifunctions: an approach to nonparametric multimodal regression,” *COMPSTAT 2006, Proceedings in Computational Statistics,* Rome, Italy, pp. 1243-1250, Aug. 2006.

[4] D. Comanicio, P. Meer, “Mean shift: a robust approach toward feature space analysis,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 603-619, May 2002.

[5] D. Comaniciu, V. Ramesh, P. Meer, “Kernel-based object tracking,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 25, pp. 564-575, May 2003.

[6] K. Fukunaga, L. D. Hostetler, “Estimation of the gradient of a density function, with applications in pattern recognition,” *IEEE Trans. on Inform. Theory,* vol. 21, pp. 32-40, Jan. 1975.

[7] Y. Cheng, “Mean shift, mode seeking and clustering,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 17, no. 8, pp. 790-799, Aug. 1995.

[8] J. Wang, B. Thiesson, Y. Xu, M. Cohen,“Image and video segmentation by anisotropic kernel mean shift,” *In Proc. European Conference on Computer Vision,* Prague, Czech Republic, vol. 2, pp. 238-250, 2004.

[9] H. Zhou, G. Schaefer, M. E. Celebi, F. Minrui, “Bayesian image segmentation with mean shift,” *In Proc. 16th IEEE International Conference on Image Processing (ICIP),* Cairo, Egypt, pp. 2405-2408, Nov. 2009.

[10] Y. Zhu, R. He, N. Xiong, P. Shi, Z. Zhang, “Edge detection based on fast adaptive mean shift algorithm,” *In Proc. International Con. on Computational Science and Engineering,* Vancouver, Canada, pp. 1034-1039, Aug. 2009.

[11] H. Guo, P. Guo, Q. Liu, “Mean shift-based edge detection for color image,” *In Proc. International Conference on Neural Networks and Brain (ICNNB),* Beijing, China, pp. 1118-1122, Oct. 2005.

[12] A. Yilmaz, “Object tracking by asymmetric kernel mean shift with automated scale and orientation selection,” *In Proc. IEEE Conference on Computer Vision and Pattern Recognition,* Minnesota, USA, pp. 18-23, Jun. 2007.

[13] H. Chen, P. Meer, “Robust fusion of uncertain information,” *IEEE Trans. Systems, Man, ans Cybernetics-Part B: Cybernetics,* vol. 35, no. 5, pp. 578-586, 2005.

[14] Y. Aliyari Ghassabeh, T. Linder, G. Takahara, “On noisy source vector quantization via a subspace constrained mean shift algorith,” *Proc. 26th Biennial Symp. on Communications,* Kingston, On., Canada, pp. 107-110, 2012.

[15] Y. Aliyari Ghassabeh, T. Linder, G. Takahara, “On the convergence and applications of mean shift type algorithm,” *25th IEEE Canadian Conference on Electrical and Computer Engineering, CCECE 2012,* Montreal, QC., Canada, 2012.

[16] Y. Aliyari Ghassabeh, “Asymptotic stability of equilibrium points of mean shift algorithm,” *Machine Learning*, pp.1-11, DOI 10.1007/s10994-014-5435-2, Mar. 2014.

[17] M. A. Carreira-Perpiñán, “Gaussian mean shift is an EM algorithm,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 767-776, May 2007.

[18] M. P. Wand, M. Jones, *Kernel Smoothing*, Chapman and Hall, 1995.

[19] B. W. Silverman, *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, 1986.

[20] E. Parzan, “On estimation of probability density function and mod,” *Annual of Mathematical Statistics,* vol. 33, pp. 1065-1967, 1962.

[21] M. Fashing, C. Tomasi, “Mean shift is a bound optimization,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 27, no. 3, pp. 471-474, Mar. 2005.

[22] A. P. Dempster, N. M. Laird, D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B,* vol. 39, pp. 1-38, 1977.

[23] R. A. Boyles, “On the convergence of the EM algorithm,” *Journal of the Royal Statistical Society: Series B,* vol. 45, pp. 47-50, Jan. 1983.

[24] C. F. J. Wu, “On the convergence properties of the EM algorithm,” *The Annals of Statistics,* vol. 11, pp. 95-103, 1983.

[25] X. Li, Z. Hu, F. Wu, “A note on the convergence of the mean shift,” *Pattern Recognition,* vol. 40, pp. 1756-1762, Jun. 2007.

[26] M. A. Carreira-Perpiñán, C. K. I. Williams, “On the number of modes of a gaussian mixture,” *Scale Space Method in Computer Vision*, L. Griffin and M. Lillholm, Eds, pp. 625-649, 2003.

[27] Y. Aliyari Ghassabeh, T. Linder, G. Takahara, “On some convergence properties of the subspace constrained mean shift,” *Pattern Recognition,* vol. 46, no. 11, pp. 3140-3147, 2013.

[28] Y. Aliyari Ghassabeh, “On the convergence of the mean shift algorithm in the one-dimensional space,” *Pattern Recognition Letters,* vol. 34, no. 12, pp. 1423-1427, 2013.

[29] B. Wallace, “On the critical points of Gaussian mixtures,” Department of Mathematics and Statistics, Queen’s University, July 2013.

[30] S. Ray, B. G. Lindsay, “The topography of multivariate normal mixture,” *The Annal of Statistics,* vol. 33, no. 5, pp. 2042-2065, 2005.

[31] R. A. Horn, C R. Johnson, *Matrix Analysis,* Cambridge University Press, 1990.

[32] A. Knutson, T. Tao, “Honey combs and sums of Hermitian matrices,” *Notices Amer. Math. Soc.,* vol. 48, no. 2, pp. 175186, 2001.

[33] A. Ostaszewski, *Advanced Mathematical Method,* Cambridge University Press, 1990.

[34] J. Milnor, *Morse Theory*, Princton University Press, 1963.

[35] S. G. Krantz, H. R. Parks, *A Primer of Real Analysis and Functions,* Springer, 2002.

[36] G. Matsaglia, G. P. H. Styan, “Equalities and inequalities for ranks of matrices” *Linear and Multilinear Algebra,* vol. 2, no. 2, pp. 269-292, 1974.

[37] K. B. Datta, *Matrix and Linear Algebra,* Prentice-Hall of India, 2004.