Mach Learn (2015) 98:359–368
DOI 10.1007/s10994-014-5435-2

Asymptotic stability of equilibrium points of mean shift algorithm

Youness Aliyari Ghassabeh

Received: 20 April 2013 / Accepted: 1 February 2014 / Published online: 12 March 2014

Abstract The mean shift (MS) algorithm is a popular non-parametric technique that has been widely used in statistical pattern recognition and machine learning. The algorithm iteratively tries to find modes of an estimated probability density function. These modes play an important role in many applications, such as clustering, image segmentation, feature extraction, and object tracking. The modes are fixed points of a discrete, nonlinear dynamical system. Although the algorithm has been successfully used in many applications, a theoretical study of its convergence is still missing in the literature. In this paper, we first consider the iteration index as a continuous variable and, by introducing a Lyapunov function, show that the equilibrium points are asymptotically stable. We also show that the proposed function can be considered as a Lyapunov function for the discrete case with the isolated stationary points. The availability of a Lyapunov function for continuous and discrete cases shows that if the MS iterations start in a neighborhood of an equilibrium point, the generated sequence remains close to that equilibrium point and finally converges to it.

Keywords Mean shift algorithm · Lyapunov function · Mode estimate sequence · Asymptotically stable · Convex function · Equilibrium point · Fixed point

1 Introduction

The mean shift (MS) algorithm is a non-parametric mode seeking technique that was introduced by Fukunaga and Hostetler (Jan. 1975) and later developed by Cheng (1995) and Comanicio and Meer (2002). The algorithm starts from one of the data points and iteratively shifts each data point to the weighted average of the data set in order to find the stationary points of an estimated probability density function (pdf). Modes of an estimated pdf have

Editor: James Cussens.

Y. Aliyari Ghassabeh (✉)
Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada
e-mail: aliyari@mast.queensu.ca

been used in a wide range of applications, including image segmentation (Comanicio and Meer 2002; Wang et al. 2004), object tracking (Comaniciu et al. 2000, 2003), noisy source vector quantization (Aliyari Ghassabeh et al. 2012b), and nonlinear dimensionality reduction (Aliyari Ghassabeh et al. 2012a). The main advantage of the MS algorithm is that it does not require any prior knowledge of the number of clusters and there is no assumption for the shape of the clusters. The MS algorithm generates a sequence, called the mode estimate sequence, in order to estimate modes of an estimated pdf. In the original paper, the authors claimed that the mode estimate sequence is a convergent sequence (Comanicio and Meer 2002), but the given proof was not correct. Later in another work, Carreira-Perpinán (2007) showed that the MS algorithm with a Gaussian kernel is an expectation maximization (EM) algorithm and therefore the generated sequence converges to a mode of the estimated pdf. However, there are situations when the EM algorithm may not converge Boyles (1983), as a result of which the convergence of the MS algorithm does not follow. The author in Carreira-Perpinán (2007) also assumed the iteration index to be a continuous variable; in addition, for a special case when all the terms in the Gaussian mixture model have the same diagonal bandwidth matrix, this author introduced a strict Lyapunov function in order to show that an equilibrium point of the system is an asymptotically stable point

In two recent works, the convergence of the MS algorithm in the one-dimensional space ($d = 1$) is investigated (Aliyari Ghassabeh et al. 2013; Aliyari Ghassabeh 2013). The authors in Aliyari Ghassabeh et al. (2013) showed that the MS algorithm with an analytic kernel (e.g., Gaussian kernel) generates a convergent sequence in the one-dimensional space. The author in Aliyari Ghassabeh (2013) proved that for the MS algorithm in the one-dimensional space with certain class of kernels, the mode estimate sequence is a monotone and convergent sequence. However, the authors in Aliyari Ghassabeh et al. (2013) and Aliyari Ghassabeh (2013) could not generalize the convergence result to a high dimensional space ($d > 1$).

In this paper, we first generalize the results given in Carreira-Perpinán (2007) for the iteration index as a continuous variable. In particular, we assume that each term in the pdf estimate using the Gaussian kernel has a unique covariance matrix instead of assuming a constant diagonal bandwidth matrix for all the terms. Then, we introduce a strict Lyapunov function and show that it satisfies the required condition for an equilibrium point to be asymptotically stable. We also investigate the discrete case with isolated stationary points and show that the proposed Lyapunov function for the continuous case can be used for the discrete case as well. The availability of a Lyapunov function guarantees the asymptotic stability of the system (i.e., the mode estimate sequence remains close to an equilibrium point and finally converges to it). In Sect. 2, I give a short introduction to the MS algorithm. I also provide a brief review of the Lyapunov stability theory in Sect. 3. The main theoretical results are given in Sect. 4. The concluding remarks are given in Sect. 5.

2 Mean shift algorithm

Let $\mathbf{x}_i \in \mathbb{R}^d,\ i = 1, \ldots, n$ be a set of n independent and identically distributed (iid) random variables. The multivariate kernel density estimation using kernel K and bandwidth matrix $\mathbf{H}$ is given by Silverman (1986)

$$\hat{f}_{K,\mathbf{H}}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K\left(\mathbf{H}^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right),$$

where the kernel K is a non-negative, real-valued, and integrable function with a compact support satisfying the following conditions (Wand 1995)

$$\int_{\mathbb{R}^d} K(\mathbf{x})d\mathbf{x} = 1, \quad \lim_{\|\mathbf{x}\|\to\infty} \|\mathbf{x}\|^d K(\mathbf{x}) = 0, \quad \int_{\mathbb{R}^d} \mathbf{x}K(\mathbf{x})d\mathbf{x} = 0.$$

For simplicity, we assume a specific class of kernel functions called radially symmetric kernels that are defined in terms of a profile k.

Definition 1 A profile $k : [0, \infty) \to [0, \infty)$ is a non-negative, non-increasing, and piecewise continuous function that satisfies $\int_0^\infty k(x)dx < \infty$ and $K(\boldsymbol{x}) = c_{k,d}k(\|\boldsymbol{x}\|^2)$, where $c_{k,d}$ is a normalization factor that causes $K(\boldsymbol{x})$ to integrate to one.

Furthermore, the shadow of a profile k is defined by Cheng (1995)

Definition 2 A profile h is called the shadow of a profile k if and only if

$$h(x) = a + b \int_x^\infty k(t)dt,$$

where $b > 0$ and $a \in \mathbb{R}$ is a constant.

To reduce the computational cost, in practice the bandwidth matrix $\mathbf{H}$ is chosen to be proportional to the identity matrix, i.e., $\mathbf{H} = h\mathbf{I}$. The estimated pdf using the profile function and with only one bandwidth parameter simplifies to the following form

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

The modes of an estimated pdf are zeros of the gradient function. Taking the gradient of (1) and equating it to zero reveals that modes of the estimated pdf are zeros of the following function (fixed points of $\mathbf{m}_{h,g}(\mathbf{x}) + \mathbf{x}$)

$$\mathbf{m}_{h,g}(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}, \tag{2}$$

where $g(x) = -k'(x)$. The vector $\mathbf{m}_{h,g}$ is called the MS vector (Comanicio and Meer 2002).

Note that the fixed point of a function $f : \mathbb{R}^d \to \mathbb{R}^d$ is any value $\boldsymbol{x} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}) = \boldsymbol{x}$, whereas a stationary point of f is any value $\boldsymbol{y} \in \mathbb{R}^d$ such that $\nabla f(\boldsymbol{y}) = \mathbf{0}$.

The MS vector can alternatively be represented by Comanicio and Meer (2002)

$$\mathbf{m}_{h,g}(\mathbf{x}) = c\frac{\nabla \hat{f}_k(\mathbf{x})}{\hat{f}_g(\mathbf{x})}, \tag{3}$$

where c is a scalar depending on the bandwidth h, and $\hat{f}_k$ represents the pdf estimate using the profile k. The above expression shows that at an arbitrary point $\mathbf{x}$, a MS vector is proportional to the normalized density gradient estimate at $\mathbf{x}$. The MS algorithm starts from one of the data points and updates the mode estimate iteratively. The mode estimate in the kth iteration is updated by

$$\begin{aligned} \mathbf{y}_{k+1} &= \mathbf{m}_{h,g}(\mathbf{y}_k) + \mathbf{y}_k \\ &= \frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{y}_k-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{y}_k-\mathbf{x}_i}{h}\right\|^2\right)}. \end{aligned} \tag{4}$$

It can be shown that the norm of the difference between two consecutive mode estimates converges to zero (Aliyari Ghassabeh et al. 2012a), i.e., $\lim_{k\to\infty} \|\mathbf{y}_{k+1} - \mathbf{y}_k\| = 0$. Therefore the MS algorithm terminates the iterations until the norm of the difference between two consecutive mode estimates becomes less than some predefined threshold. The convergence of the algorithm for the special one dimensional case ($d = 1$) is proved (Aliyari Ghassabeh 2013), but unfortunately the convergence result has not been generalized for higher dimensions, i.e. $d > 1$.

3 Lyapunov stability theory

Consider a general nonlinear dynamical system (Luenberger 1979)

$$\dot{\mathbf{x}} = f(\mathbf{x}(t), t), \quad \text{Continuous case,}$$
$$\mathbf{x}(k+1) = f(\mathbf{x}(k), k), \quad \mathbf{x}(0) = \mathbf{x}_0, \qquad \text{Discrete case,}$$

where $\mathbf{x} \in \mathbb{U} \subset \mathbb{R}^d$, $\mathbb{U}$ is a neighborhood of the origin and $f : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous and differentiable function. An equilibrium point $\mathbf{x}^*$ for continuous and discrete cases is defined as follows (Antsaklis 2006).

Definition 3 A vector $\mathbf{x}^*$ is called an equilibrium point from time t_0 for the continuous case if $f(\mathbf{x}^*, t) = 0, \forall t \geq t_0$ and is called an equilibrium (or fixed) point from time k_0 for the discrete case if $f(\mathbf{x}^*, k) = \mathbf{x}^*, \forall k > k_0$.

An equilibrium point $\mathbf{x}^*$ is called Lyapunov stable if solutions starting close enough to the equilibrium point remain close enough forever. Formally speaking, we have (Antsaklis 2006):

Definition 4 An equilibrium point $\mathbf{x}^*$ is called Lyapunov stable if for every $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that if $\|\mathbf{x}(0) - \mathbf{x}^*\| < \delta(\epsilon)$ then $\|\mathbf{x}(t) - \mathbf{x}^*\| < \epsilon$ for all $t \geq 0$ (the Lyapunov stability is defined similarly for the discrete case).

The equilibrium point $\mathbf{x}^*$ is said to be asymptotically stable if it is Lyapunov stable and if there exists $\delta > 0$ such that if $\|\mathbf{x}(0) - \mathbf{x}^*\| < \delta$ then $\lim_{t\to\infty} \|\mathbf{x}(t) - \mathbf{x}^*\| = 0$ (Antsaklis 2006).

Let $\mathbf{x}^*$ denote an equilibrium point for a continuous dynamic system. Lyapunov's second method states that if there exists a continuous, differentiable function $V(\mathbf{x}) : E \to \mathbb{R}$, where $E \subset \mathbb{R}^d$ is a neighborhood of $\mathbf{x}^*$, such that $V(\mathbf{x}^*) = 0$ and $V(\mathbf{x}) > 0$ if $\mathbf{x} \neq \mathbf{x}^*$, then $\mathbf{x}^*$ is asymptotically stable if $\dot{V}(\mathbf{x}) < 0$ for all $\mathbf{x} \in E \backslash \{\mathbf{x}^*\}$ (Tripathi 2008). For a discrete time system, the theorem is slightly different: if there exist a continuous, differentiable function $V(\mathbf{x}) : E \to \mathbb{R}$, where E is defined as before, such that $V(\mathbf{x}^*) = 0$ and $V(\mathbf{x}) > 0$ if $\mathbf{x} \neq \mathbf{x}^*$, then $\mathbf{x}^*$ is asymptotically stable if $\Delta V(\mathbf{x}) = V(\mathbf{x}_{k+1}) - V(\mathbf{x}_k) < 0$ for all $\mathbf{x} \in E \backslash \{\mathbf{x}^*\}$ (Haddad and Chellaboina 2008).

4 Theoretical results

In this section, we first consider the iteration index for the MS algorithm to be continuous and generalize the results in Carreira-Perpinán (2007). Then we show that the proposed function can also be used as a Lyapunov function for the discrete case with isolated stationary points, which shows that the fixed points of the MS algorithm are asymptotically stable.

4.1 Continuous case

Carreira-Perpinán (2007) investigated the MS algorithm with a Gaussian kernel and considered the iteration index to be a continuous variable. The Gaussian MS algorithm with a continuous iteration index can be written as follows (Carreira-Perpinán 2007)

$$\dot{\mathbf{x}} = \frac{\sum_{i=1}^{n} \Sigma_i^{-1}(\mathbf{x}_i - \mathbf{x}) \exp\left(-(\mathbf{x}_i - \mathbf{x})^t \Sigma_i^{-1}(\mathbf{x}_i - \mathbf{x})/2\right)}{\sum_{i=1}^{n} \exp\left(-(\mathbf{x}_i - \mathbf{x})^t \Sigma_i^{-1}(\mathbf{x}_i - \mathbf{x})/2\right)} = \frac{\nabla \hat{f}(\mathbf{x})}{\hat{f}(\mathbf{x})}, \tag{5}$$

where Σ_i is the covariance matrix for ith component in the Gaussian mixture model. For simplicity the author in Carreira-Perpinán (2007) assumed that $\Sigma_i = h^2\mathbf{I}$. Then the above continuous dynamical system reduces to

$$\dot{\mathbf{x}} = \nabla\left(h^2 \log(\hat{f}(\mathbf{x}))\right), \tag{6}$$

where $f(\mathbf{x})$ is defined in (1) and has a Gaussian kernel with the bandwidth matrix $h^2\mathbf{I}$. The Lyapunov function in neighborhood E of any equilibrium point $\mathbf{x}^*$ is defined by Carreira-Perpinán (2007)

$$V(\mathbf{x}) = h^2 \log \frac{\hat{f}(\mathbf{x}^*)}{\hat{f}(\mathbf{x})}. \tag{7}$$

It is not difficult to show that $V(\mathbf{x}^*) = 0$ and $V(\mathbf{x}) > 0$ for all $\mathbf{x} \in E\backslash\{\mathbf{x}^*\}$, i.e. V is positive definite in $E\backslash\{\mathbf{x}^*\}$. The author in Carreira-Perpinán (2007) also showed that $\dot{V}(\mathbf{x}) < 0$ for all $\mathbf{x} \in E\backslash\{\mathbf{x}^*\}$. Therefore, the equilibrium point $\mathbf{x}^*$ is asymptotically stable point for the dynamical system. The author in Carreira-Perpinán (2007) mentioned that finding a Lyapunov function for the general case (5) is more difficult.

In recent work, the authors provided a sufficient condition for the MS algorithm with the Gaussian kernel to have a unique mode in the convex hull of the data set (Theorem 2 in Liu et al. 2013). They showed that if the MS algorithm has a unique mode in the convex hull of the data set, then the mode is globally stable and the mode estimated sequence is an exponentially convergent sequence (Theorem 3 in Liu et al. 2013). The provided sufficient condition in Liu et al. (2013) depends on the data set and the covariance matrix of each Gaussian term in the pdf estimate. In general, it may be a difficult task to choose the covariance matrices to satisfy the provided sufficient condition. Furthermore, the MS algorithm with a unique mode has limited use in practice. The MS algorithm has been widely used in applications such as image segmentation and clustering, which require the algorithm to have multiple modes.

We propose a Lyapunov function for the general case (5) in order to guarantee the asymptotic stability of the algorithm for a general Gaussian mixture model. Let $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \ldots, n$ denote our samples. The density estimate using the Gaussian kernel is given by

$$\hat{f}(\mathbf{x}) = c \sum_{i=1}^{n} \exp\left(-(\mathbf{x} - \mathbf{x}_i)^t \Sigma_i^{-1}(\mathbf{x} - \mathbf{x}_i)/2\right), \tag{8}$$

where c is the normalization factor and $\Sigma_i, i = 1, \ldots, n$ is the covariance matrix for ith sample. Let $N(\mathbf{x}_i, \Sigma_i) = \exp(-(\mathbf{x} - \mathbf{x}_i)^t \Sigma_i^{-1}(\mathbf{x} - \mathbf{x}_i)/2)$ denote the Gaussian function at $\mathbf{x}$, then the gradient estimate at $\mathbf{x}$ using $\hat{f}(\mathbf{x})$ is computed by

$$\begin{aligned}\nabla \hat{f}(\mathbf{x}) &= c\sum_{i=1}^{n} \Sigma_i^{-1}(\mathbf{x}_i - \mathbf{x})N(\mathbf{x}_i, \Sigma_i)\\ &= c\sum_{i=1}^{n} \Sigma_i^{-1}\mathbf{x}_i N(\mathbf{x}_i, \Sigma_i) - c\sum_{i=1}^{n} \Sigma_i^{-1}N(\mathbf{x}_i, \Sigma_i)\mathbf{x}. \qquad (9)\end{aligned}$$

Multiplying both sides of (9) by $[\sum_{i=1}^{n} \Sigma_i^{-1}N(\mathbf{x}_i, \Sigma_i)]^{-1}/c$, we obtain

$$\begin{aligned}\left[\sum_{i=1}^{n} \Sigma_i^{-1}N(\mathbf{x}_i, \Sigma_i)\right]^{-1} \frac{\nabla \hat{f}(\mathbf{x})}{c} &= \left[\sum_{i=1}^{n} \Sigma_i^{-1}N(\mathbf{x}_i, \Sigma_i)\right]^{-1} \sum_{i=1}^{n} \Sigma_i^{-1}\mathbf{x}_i N(\mathbf{x}_i, \Sigma_i) - \mathbf{x}\\ &= \mathbf{m}(\mathbf{x}). \qquad (10)\end{aligned}$$

Consider a nonlinear continuous system $\dot{\mathbf{x}}(t) = \mathbf{m}(\mathbf{x}(t))$, where $\mathbf{m} : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous, differentiable function defined in (10). Let $\mathbf{x}^*$ be an equilibrium point of the system, i.e., $m(\mathbf{x}^*) = 0$. Let $V(\mathbf{x}) = \hat{f}(\mathbf{x}^*) - \hat{f}(\mathbf{x})$, where $V : E \to \mathbb{R}$ is a continuous, differentiable function and $E \subset \mathbb{R}^d$ is an open neighborhood around $\mathbf{x}^*$ such that $\hat{f}(\mathbf{x}^*) > \hat{f}(\mathbf{x})$ for all $\mathbf{x} \in E$.[1] Since $\mathbf{x}^*$ is a mode of the estimated pdf in local neighborhood E, then $V(\mathbf{x}) = f(\mathbf{x}^*) - f(\mathbf{x}) > 0$ for all $\mathbf{x} \in E\backslash\{\mathbf{x}^*\}$ and $V(\mathbf{x}^*) = 0$, i.e., $V(\mathbf{x})$ is an strictly positive definite in local neighborhood E. Now it is time to show that the function $\dot{V}(\mathbf{x})$ is negative definite, i.e., $\dot{V}(\mathbf{x}) < 0$ for all $\mathbf{x} \in E\backslash\{\mathbf{x}^*\}$ and $\dot{V}(\mathbf{x}^*) = 0$. By taking the derivative of V using the chain rule, we have

$$\begin{aligned}\dot{V}(\mathbf{x}) &= \dot{\mathbf{x}}^t \nabla V\\ &= -\dot{\mathbf{x}}^t \nabla \hat{f}(\mathbf{x}) = -\mathbf{m}(\mathbf{x}(t))^t \nabla \hat{f}(\mathbf{x})\\ &= -\left(\left[\sum_{i=1}^{n} \Sigma_i^{-1}N(\mathbf{x}_i, \Sigma_i)\right]^{-1} \frac{\nabla \hat{f}(\mathbf{x})}{c}\right)^t \nabla \hat{f}(\mathbf{x})\\ &= \frac{-1}{c}\nabla \hat{f}(\mathbf{x})^t \left[\sum_{i=1}^{n} \Sigma_i^{-1}N(\mathbf{x}_i, \Sigma_i)\right]^{-1} \nabla \hat{f}(\mathbf{x}) < 0.\end{aligned}$$

The last inequality is true since the weighted sum of the inverse of the covariance matrices is a positive definite matrix. It is also obvious that $\dot{V}(\mathbf{x}^*) = 0$. Therefore, $V(\mathbf{x}) = \hat{f}(\mathbf{x}^*) - \hat{f}(\mathbf{x})$ is a strict Lyapunov function for the continuous dynamical system in (5) and $\mathbf{x}^*$ is locally asymptotically stable, i.e., if we start from any point $\mathbf{x}_0 \in E$ then the mode estimate sequence remains close to $\mathbf{x}^*$ and finally will converge to $\mathbf{x}^*$.

4.2 Discrete case

For the discrete case, Fashing and Tomasi proved the following theorem (Theorem 2 in Fashing and Tomasi 2005).

Theorem 1 *The MS procedure with a piecewise constant profile k is equivalent to Newton's method applied to a density estimate using the shadow of k.*

Theorem 1 implies that for a very special class of profile functions, piecewise constant profiles, the MS algorithm tends to be equivalent to Newton's method. A piecewise constant

[1] We assume that the density estimate $\hat{f}$ has isolated stationary points, therefore such neighborhood E exists.

profile (e.g., uniform profile) defines a piecewise constant kernel. Piecewise constant kernels (e.g., uniform kernels) have limited use in kernel density estimation, since the pdf estimate using a piecewise constant kernel is a non-smooth function that is not desirable. Theorem 1 is not correct for widely used kernels (e.g., Gaussian kernel) and therefore the MS algorithm in general is not equivalent to Newton's method. Furthermore, even for a piecewise constant profile k, Theorem 1 does not necessarily imply the convergence of the sequence. There are situations where Newton's method diverges. For example, consider function $f(x) = x^{1/3}$: starting at point $x_1 = a (a \in \mathbb{R})$, Newton's method generates the following sequence

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = -2x_n.$$

It is clear that for $a \neq 0$ the sequence $\{x_n\}_{n=1,2,...}$ grows instead of converging, hence Newton's method fails to find the root $x = 0$. The authors in Fashing and Tomasi (2005) showed that the MS procedure at step $k, k \geq 1$ maximizes a quadratic function $\rho_k(\boldsymbol{x})$ (Theorem 3 in Fashing and Tomasi 2005). Furthermore, they proved that $\rho_k(\boldsymbol{x})$ can be considered as a (lower) bounding function for the density estimate $\hat{f}(\boldsymbol{x})$, where the bounding function $\rho_k(\boldsymbol{x})$ for $\hat{f}(\boldsymbol{x})$ is defined by Salakhutdinov et al. (2003)

Definition 5 Let $\hat{f}(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ denote our objective function, where $\mathcal{X} \subset \mathbb{R}^D$, $D \geq 1$. The bounding function $\rho_k(\boldsymbol{x})$ for $\hat{f}(\boldsymbol{x})$ is a function such that $\rho_k(\boldsymbol{x}^*) = \hat{f}(\boldsymbol{x}^*)$ at some point $\boldsymbol{x}^* \in \mathcal{X}$ and $\rho_k(\boldsymbol{x}) \leq \hat{f}(\boldsymbol{x})$ for every other $\boldsymbol{x} \in \mathcal{X}$.

The authors in Fashing and Tomasi (2005) showed that the MS algorithm with profile k is a quadratic bound maximization over a density estimate $\hat{f}$ using the shadow of k (Theorem 4 in Fashing and Tomasi 2005). This result implies that the pdf estimate along the sequence generated by the MS algorithm is an increasing sequence, i.e., $\hat{f}(\boldsymbol{y}_{k+1}) \geq \rho_k(\boldsymbol{y}_{k+1}) > \rho_k(\boldsymbol{y}_k) = \hat{f}(\boldsymbol{y}_k)$ (Fashing and Tomasi 2005).

Assume we are interested in maximizing a scalar valued function $L(\theta)$ of a free parameter vector Θ. The bound maximizer algorithms (e.g., EM algorithm for maximum likelihood learning in latent variable models) never worsen the objective function. In other words, the bound maximizer algorithms generate a sequence $\{\Theta_k\}_{k=1,2,...}$ such that $L(\theta_{k+1}) > L(\theta_k), k \geq 1$ (Salakhutdinov et al. 2003). However, a bound maximizer algorithm (e.g., EM algorithm) without additional conditions may not converge (Wu 1983). For example, Boyles presented a counterexample that satisfies all the hypotheses of Theorem 2 in Dempster et al. (1977) but converges to a unit circle instead of converging to a single point (Boyles 1983). Thus, showing that the MS algorithm is a bound optimization is not enough to prove the convergence of mode estimate sequence.

From (4) and (10), the discrete dynamical system for the MS algorithm is

$$\mathbf{y}(k+1) = \mathbf{m}(\mathbf{y}(k)) + \mathbf{y}(k), \tag{11}$$

where $\mathbf{y}(k)$ is the mode estimate at kth iteration. Let $\mathbf{y}^*$ denote the equilibrium point of (11), then $\mathbf{y}^*$ is a fixed point of (11), which implies $\mathbf{m}(\mathbf{y}^*) = 0$. Consider the proposed Lyapunov function $V(\mathbf{y}) = \hat{f}(\mathbf{y}^*) - \hat{f}(\mathbf{y})$. For any isolated mode $\mathbf{y}^*$ of the estimated pdf there is an open neighborhood E around $\mathbf{y}^*$ such that the estimated pdf attains its maximum at $\mathbf{y}^*$, i.e., $V(\mathbf{y}) = \hat{f}(\mathbf{y}^*) - \hat{f}(\mathbf{y}) > 0$ for all points $\mathbf{y} \in E \backslash \{\mathbf{y}^*\}$. It is clear that $V(\mathbf{y}^*) = 0$, therefore $V(\mathbf{x})$ is a strict Lyapunov function in E. To show that $\Delta V(\mathbf{y}) < 0$, we need the following lemma.[2]

[2] The authors in Comanicio and Meer (2002) assumed that $\boldsymbol{y}_j = \mathbf{0}$ and based on this assumption, they showed that the pdf estimate is an increasing sequence along the mode estimate sequence. Here, we relax the assumption $\boldsymbol{y}_j = \mathbf{0}$ and prove the monotonicity of the $\hat{f}(\boldsymbol{y}_j)$. The proof is a reproduction of the proof in Comanicio and Meer (2002), except $\boldsymbol{y}_j \neq \mathbf{0}$.

Lemma 1 *If the profile* k *is a convex and strictly decreasing function, then the density estimate values* $\hat{f}$ *are increasing along the mode estimate sequence.*

Proof Let $\mathbf{y}_j \neq \mathbf{y}_{j+1}$, we show that $\hat{f}(\mathbf{y}_{j+1}) > \hat{f}(\mathbf{y}_j)$. From Equation (1), we have

$$\begin{aligned}\hat{f}(\mathbf{y}_{j+1}) - \hat{f}(\mathbf{y}_j) &= \frac{c_{k,d}}{nh^d}\left[\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{y}_{j+1}-\mathbf{x}_i}{h}\right\|^2\right) - \sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\right]\\ &= \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n}\left[k\left(\left\|\frac{\mathbf{y}_{j+1}-\mathbf{x}_i}{h}\right\|^2\right) - \left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\right]\\ &=> \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(\left\|\frac{\mathbf{y}_{j+1}-\mathbf{x}_i}{h}\right\|^2 - \left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right),\end{aligned}$$

where the last inequality is true since the convexity of the profile function k implies that $k(x_2) - k(x_1) => k'(x_1)(x_2 - x_{x1})$. By expanding the terms in the right side of the above inequality and using Eq. (4), we have

$$\begin{aligned}\hat{f}(\mathbf{y}_{j+1}) - \hat{f}(\mathbf{y}_j) &=> \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(\left\|\frac{\mathbf{y}_{j+1}-\mathbf{x}_i}{h}\right\|^2 - \left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\\ &= \frac{c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(\left\|\mathbf{y}_{j+1}\right\|^2 + \|\mathbf{x}_i\|^2 - 2\mathbf{y}_{j+1}\cdot\mathbf{x}_i - \left\|\mathbf{y}_j\right\|^2\right.\\ &\quad \left. - \|\mathbf{x}_i\|^2 + 2\mathbf{y}_j\cdot\mathbf{x}_i\right)\\ &= \frac{c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(\left\|\mathbf{y}_{j+1}\right\|^2 - \left\|\mathbf{y}_j\right\|^2 - 2\left(\mathbf{y}_{j+1}-\mathbf{y}_j\right)\cdot\mathbf{x}_i\right)\\ &= \frac{c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(\left\|\mathbf{y}_{j+1}\right\|^2 - \left\|\mathbf{y}_j\right\|^2 - 2\left(\mathbf{y}_{j+1}-\mathbf{y}_j\right)\cdot\mathbf{y}_{j+1}\right)\\ &= \frac{c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(-\left\|\mathbf{y}_{j+1}\right\|^2 - \left\|\mathbf{y}_j\right\|^2 + 2\mathbf{y}_j\cdot\mathbf{y}_{j+1}\right)\\ &= -\frac{c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} k'\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)\left(\left\|\mathbf{y}_{j+1}-\mathbf{y}_j\right\|\right)^2 > 0,\end{aligned}$$

where $\cdot$ denotes the inner product. The last inequality comes from the fact that the profile function k is strictly decreasing, therefore its derivative is strictly less than zero, i.e., $k'(x) < 0$. Therefore, the sequence $\{\hat{f}(\mathbf{y}_j)\}_{j=1,2,\ldots}$ is strictly increasing and for an arbitrary j, we have $\hat{f}(\mathbf{y}_{j+1}) - \hat{f}(\mathbf{y}_j) > 0$. □

Using Lemma (1), we have

$$\begin{aligned}\Delta V(\mathbf{y}) &= V(\mathbf{y}(k+1)) - V(\mathbf{y}(k))\\ &= \hat{f}(\mathbf{y}^*) - \hat{f}(\mathbf{y}(k+1)) - \hat{f}(\mathbf{y}^*) + \hat{f}(\mathbf{y}(k))\\ &= \hat{f}(\mathbf{y}(k)) - \hat{f}(\mathbf{y}(k+1)) < 0. \qquad (12)\end{aligned}$$

The last inequality holds since from Lemma 1 the sequence $\{\hat{f}(\mathbf{y}(k))\}_{k=1,2,\ldots}$ is an increasing sequence. Therefore, for the discrete dynamical system in (11), function V is a Lyapunov function and equilibrium point $\mathbf{y}^*$ is asymptotically stable.

Remarks

1. For the Lyapunov function in Carreira-Perpinán (2007), it is required that all the covariance matrices be the same and proportional to the identity matrix, i.e., $\Sigma_i = h^2\mathbf{I}, i = 1, 2, \ldots, n$. But for the proposed Lyapunov function, there is no constraint on the covariance matrices except being positive definite.
2. The systems in (5) or (11) can have many equilibrium points and, as long as the equilibrium points are isolated, the above argument works. For each equilibrium point $\mathbf{x}_i^*, i = 1, 2, \ldots$ ($\mathbf{y}^*$ for the discrete case), there is an open neighborhood E_i such that the estimated pdf $\hat{f}(\mathbf{x})$ attains its maximum at $\mathbf{x}_i^*$ on E_i.
3. By proving the asymptotic stability of the isolated equilibrium points of the MS algorithm, we showed that if we start from a point close to an specific equilibrium point, then the MS algorithm remains close to the equilibrium point and finally converges to it.
4. In real world applications, where digital computers store numbers in floating point representation, the MS algorithm may not converge exactly to a fixed point due to the rounding error. As mentioned before, the MS algorithm stops when the distance between two mode estimates becomes less that some predefined threshold, i.e., $\|\mathbf{y}_{k+1} - \mathbf{y}_k\| < \epsilon$. By choosing a small threshold we can guarantee that the stopping point is close enough to the fixed point.

5 Conclusion

The MS algorithm is a widely used technique for estimating modes of an estimated pdf. Although the algorithm has been used in many applications, it seems that the study of the theoretical properties of the algorithm has been missing in the literature. In this paper, we generalized the asymptotic stability results in Carreira-Perpinán (2007) by introducing a Lyapunov function for the MS algorithm with a continuous iteration index. The author in Carreira-Perpinán (2007) proposed a Lyapunov function for the MS algorithm with the Gaussian kernel when all terms in the pdf estimate have equal covariance matrices that are proportional to the identity matrix. In our case, there is no constraint on the covariance matrices and they just need to be positive definite matrices. We also showed that the proposed function satisfies the required condition for an equilibrium (fixed) point of the discrete MS algorithm with isolated stationary points to be asymptotically stable. In other words, we proved that for the MS algorithm with isolated stationary points, if we start the iterations close enough to an equilibrium point, then the mode estimate sequence remains close to that point and finally converges to it.

References

Aliyari Ghassabeh, Y. (2013). On the convergence of the mean shift algorithm in the one-dimensional space. *Pattern Recognition Letters, 34*(12), 1423–1427.

Aliyari Ghassabeh, Y., Linder, T., & Takahara, G. (2012a, Apil). On the convergence and applications of mean shift type algorithms. In *Proceedings of 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Montreal, Canada*.

Aliyari Ghassabeh, Y., Linder, T., & Takahara, G. (2012b, May) On noisy source vector quantization via a subspace constrained mean shift algorithm. In *Proceesing of 26th Biennial Symposium on Communications (QBSC), Kingston, Canada*.

Aliyari Ghassabeh, Y., Linder, T., & Takahara, G. (2013). On some convergence properties of the subspace constrained mean shift. *Pattern Recognition*, *46*(11), 3140–3147.

Antsaklis, P. J., & Michel, A. N. (2006). *Linear systems*. Berlin: Springer.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *45*, 47–50.

Carreira-Perpinán, M. A. (2007, May). Gaussian mean shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*, 767–776.

Cheng, Y. (1995, August). Mean shift, mode seeking and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*, 790–799.

Comanicio, D., & Meer, P. (2002, May). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 603–619.

Comaniciu, D., Ramesh, V., & Meer, P. (2000) Real-time tracking of non-rigid objects using mean shift. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition, Princeton, USA* (pp. 142–149).

Comaniciu, D., Ramesh, V., & Meer, P. (2003, May) Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 564–575.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*, 1–38.

Fashing, M., & Tomasi, C. (2005). Mean shift is a bound optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3), 471–474.

Fukunaga, K., & Hostetler, L. D. (1975, January). Estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21*, 32–40.

Haddad, W. M., & Chellaboina, V. (2008). *Nonlinear dynamical systems and control: A Lyapunov-based approach*. Princeton: Princeton University Press.

Liu, Y., Li, S. Z., Wuc, W., & Huanga, R. (2013). Dynamics of a mean-shift-like algorithm and its applications on clustering. *Information Processing Letters*, *13*, 8–16.

Luenberger, D. G. (1979). *Introduction to dynamic systems: Theory, models, and applications*. New York: Wiley.

Salakhutdinov, R., Roweis, S., & Ghahramani, Z. (2003, August) On the convergence of bound optimization algorithms. In *Proceedings 19th Conference in Uncertainty in Artificial Intelligence (UAI 03), Acapulco, Mexico*.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Tripathi, S. M. (2008). *Modern control systems: An introduction*. Sudbury: Jones and Bartlett Publishers.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.

Wang, J., Thiesson, B., Xu, Y., & Cohen, M. (2004). Image and video segmentation by anisotropic kernel mean shift. In *Proceedings of European Conference on Computer Vision, Prague, Czech Republic* (Vol. 2, pp. 238–250).

Wu, C. F. J. (1983). On theconvergence properties of the EM algorithm. *The Annals of Statistics*, *11*, 95–103.