Journal of Multivariate Analysis 135 (2015) 1–10

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel

Youness Aliyari Ghassabeh

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

HIGHLIGHTS

- Incompleteness of the previous proofs for the convergence of MS algorithm is reviewed.
- I showed the gradient function is always nonzero outside the convex hull of the data.
- The convergence of the MS algorithm with isolated stationary points is proved.
- A sufficient condition for Gaussian KDE to have isolated stationary points is given.

ARTICLE INFO

Article history:
Received 11 April 2014
Available online 11 December 2014

AMS subject classifications:
26B25
26B35
40A05
40A30
62G05

Keywords:
Mean shift algorithm
Mode estimate sequence
Convex hull
Isolated stationary points
Kernel function
Gaussian KDE
Convergence

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.

1. Introduction

The modes of a probability density function (pdf) play an essential role in many applications, including classification [6], clustering [24], multi-valued regression [16], image segmentation [12], and object tracking [13]. Due to the lack of knowledge about the pdf, a nonparametric technique is proposed to find an estimate for the gradient of a pdf [18]. 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 [18] for finding modes of an estimated pdf. The algorithm was generalized by Cheng [11] in order to show that the MS algorithm is a mode-seeking process on a surface constructed with

E-mail address: aliyari@cs.toronto.edu.

http://dx.doi.org/10.1016/j.jmva.2014.11.009
0047-259X/

a shadow kernel. Later, the algorithm became popular in the machine learning society when its potential usage for feature space analysis was studied [12].

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 [33,36], edge detection [37,19], object tracking [13,35], information fusion [10], and noisy source vector quantization [3,4].

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 [12] claimed that the MS algorithm generates a convergent sequence. But a crucial step for the convergence proof of the sequence in [12] 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 [8]. 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 $K : \mathbb{R}^D \to \mathbb{R}$ is a non-negative real-valued function that satisfies the following conditions [32]

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

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 (i.i.d.) 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 [30],

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 [28]. 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 [30]

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