An Online Sample Based Method for Mode Estimation using ODE Analysis of Stochastic Approximation Algorithms

Chandramouli Kamanchi\textsuperscript{1} Raghuram Bharadwaj Diddigi\textsuperscript{2} Prabuchandran K.J.\textsuperscript{3} Shalabh Bhatnagar\textsuperscript{4}

Abstract—One of the popular measures of central tendency that provides better representation and interesting insights of the data compared to the other measures like mean and median is the metric mode. If the analytical form of the density function is known, mode is an argument of the maximum value of the density function and one can apply the optimization techniques to find mode. In many of the practical applications, the analytical form of the density is not known and only the samples from the distribution are available. Most of the techniques proposed in the literature for estimating the mode from the samples assume that all the samples are available beforehand. Moreover, some of the techniques employ computationally expensive operations like sorting. In this work we provide a computationally effective, online iterative algorithm that estimates the mode of a unimodal smooth density given only the samples generated from the density. Asymptotic convergence of the proposed algorithm using an ordinary differential equation (ODE) based analysis is provided. We also prove the stability of estimates by utilizing the concept of regularization. Experimental results further demonstrate the effectiveness of the proposed algorithm.

I. INTRODUCTION

There are many metrics that are used to represent the central tendency of the data. Among them, the popular ones are mean, median and mode. Mean is extensively studied due to its simplicity, linearity and ease of estimation via sample averages. However mean is susceptible to outliers. For example, when we are estimating the mean from a finite number of samples, one bad outlier can shift the estimate far away from the original mean. Also, in some of the applications, mean may not be the desired quantity to analyze. For example, it is often interesting to know the income that majority of the population in a country earns rather than the average income of the country which is a skewed quantity.

In this work, we focus on finding the mode of a distribution when the analytical form of it is not known. That is, we are given only the samples of the distribution and we need to estimate the mode from these samples. We first discuss some of the works that have been reported in the literature for estimating the mode. This problem of estimation of the mode of a unimodal density has been first considered in [1]. In [1] a sequence of density functions is iteratively constructed from the samples and the respective modes are calculated as maximum likelihood estimates. It is shown that these estimates of mode converge in probability to the actual mode.

We can broadly classify the solution techniques of mode estimation problem into two groups. The first group comprises a non-parametric way of estimating the mode where the mode is estimated directly from the sample data without constructing the density function. The second group of methods comprises the parametric way of estimation in which the density function is approximately constructed and the mode is computed using optimization techniques.

In [2], a non-parametric estimator (popularly known as Grenander’s estimate) for estimating the mode directly from the samples is proposed. The later developments of mode estimation methods are based on the idea that the mode is situated at the center of an interval of selected length that contains majority of observed points. A sequence of nested intervals and intermediate mode estimates is constructed based on the above described idea and mode is taken to be the point that these intermediate estimates of mode converge to. Different ways of selecting the interval lengths are studied in [3], [4] along with their convergence properties. A variant of this idea involves selecting the interval of shortest length that contains some desired number of points instead of deciding the lengths of interval. The estimation methods in [5]–[7] are based on this idea. Detailed survey of the above discussed techniques along with their robustness is extensively studied in [8], [9].

In [10], a parametric method of estimating the mode is proposed. The idea here is to fit the samples to a normal distribution. Then the mode is estimated by calculating the mean of this fitted normal distribution. This idea is recently extended to find the multivariate mode in [11]. In [12], multivariate mean shift algorithm is proposed for estimating the multivariate mode from the samples. The idea here is to iteratively shift the estimated mode towards the actual mode using Gaussian kernels. In [13], minimum volume peeling method is proposed to estimate the multivariate mode from the sample data. The idea here is to iteratively construct subsets of the set of samples with minimum volume and discard the remaining points. The mode is then calculated by averaging the points in the constructed subset. This is based on the observation that mode is generally situated in the minimum volume set of a given fixed number of samples. An effective way of selecting the subset of points is discussed in [13].

Most of the algorithms considered in the literature so
far make the assumption that all the samples are available upfront. These techniques cannot be extended to the case of streaming data where the samples arrive online one at a time. Also, the non-parametric techniques (refer [9]) require the samples to be in a sorted order.

Our proposed algorithm is fundamentally different from the above algorithms in the sense that ours is an online algorithm that works with the data as it becomes available. This enables us to work with online samples without storing them in the memory. Also, we do not resort to any computationally expensive operations like sorting. In addition, our algorithm works for both univariate and multivariate distributions. We provide a convergence analysis of our proposed technique and show the robustness of our technique using simulation results.

Our work is closest to [14]. In [14] a gradient-like recursive procedure for estimating the mode is proposed and convergence is provided utilizing the theory of martingales. In our work to mitigate the lack of analytical form of density, we construct a kernel based representation (refer section II) of the density function and use stochastic gradient ascent to calculate mode. Our work is different from [14] in the following ways.

- Our proposed algorithm is based on assumptions different from those of [14]. Moreover, we prove the stability of the mode estimates by introducing the concept of regularization [15].
- We demonstrate the effectiveness of our algorithm by providing empirical evaluation on well-known distributions.
- Our convergence proof utilizes the well-known ODE based analysis of stochastic approximation algorithms [16]. To the best of our knowledge, ours is the first work that makes use of ODE based analysis in the context of mode estimation.

The rest of the paper is organized as follows. Section II explains the necessary background and preliminaries. We provide the motivation for our work as well as present our algorithm in Section III. Section IV describes the convergence analysis. Section V presents the results of our numerical experiments. Finally Section VI presents conclusions and future work.

II. BACKGROUND AND PRELIMINARIES

To begin, suppose we have a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and a random vector \(X : \Omega \rightarrow \mathbb{R}^p\) with a smooth density function \(f(x)\). A mode of the random vector \(X\) is defined as an argument of the maximum of \(f\). Suppose we have a unique mode i.e. \(f\) has a unique maximizer then the mode is a measure of central tendency of the distribution of \(X\). A natural problem that arises is the estimation of the mode given a sequence of independent and identically distributed samples of \(X\) denoted by \(X_1, X_2, \ldots, X_n\). In what follows we provide necessary formal definitions and prove some properties that are utilized in the motivation and the convergence of our proposed algorithm to solve this problem.

A. Approximation of Identity

Definition 1: The convolution of two functions \(u\) and \(v\) on \(\mathbb{R}^p\) is defined as
\[
(u * v)(x) := \int_{\mathbb{R}^p} u(x-t)v(t)dt \quad (x \in \mathbb{R}^p).
\]

Definition 2: Given a function \(K : \mathbb{R}^p \rightarrow \mathbb{R}\) and \(\epsilon > 0\) we define
\[
K_\epsilon(x) := \epsilon^{-p}K(\frac{x}{\epsilon}).
\]

For a given function \(K\) the family of functions \(\{K_\epsilon\}_{\epsilon > 0}\) is called approximation of the identity.

Lemma 1: Suppose \(\int_{\mathbb{R}^p} |K(t)|dt < \infty\). Then, given any \(\epsilon > 0\)

1) \(\int_{\mathbb{R}^p} K_\epsilon = \int_{\mathbb{R}^p} K\).

2) \(\int_{||x||>\delta} |K_\epsilon| \to 0\) as \(\epsilon \to 0\), for any fixed \(\delta > 0\).

Proof: For statement 1 choose \(y = \frac{x}{\epsilon}\). So \(dx = \epsilon^p dy\).

Then we have
\[
\int_{\mathbb{R}^p} K_\epsilon = \int_{\mathbb{R}^p} K(\frac{x}{\epsilon})dx = \int_{\mathbb{R}^p} K(y)dy = \int_{\mathbb{R}^p} K.
\]

Again for statement 2, let \(y = x/\epsilon\) and choose \(\delta > 0\). Now
\[
\int_{||x||>\delta} |K_\epsilon(x)|dx = \frac{1}{\epsilon^p} \int_{||x||>\delta} |K(\frac{x}{\epsilon})|dx = \int_{||y||>\delta/\epsilon} |K(y)|dy
\]

Since \(\int_{\mathbb{R}^p} |K(t)|dt < \infty\) and \(\delta/\epsilon \to \infty\) as \(\epsilon \to 0\), the proof is complete. \(\square\)

Theorem 1: Let \(u_\epsilon := u * K_\epsilon\), where \(\int_{\mathbb{R}^p} K = 1\), and \(K(x) = o(||x||^{-p})\) as \(||x|| \to \infty\). Then \(u_\epsilon \to u\) as \(\epsilon \to 0\) at each point of continuity of \(u\).

Proof: Let \(u\) be continuous at \(x\). By the definition of continuity, given \(\eta > 0\), there exists \(\delta > 0\) such that \(|u(x-t) - u(x)| < \eta\) if \(||t|| < \delta\). Since \(\int_{\mathbb{R}^p} K_\epsilon = \int_{\mathbb{R}^p} K = 1\) from Lemma 1 and hypothesis we have,

\[
|u_\epsilon(x) - u(x)| = \left| \int u(x-t)K_\epsilon(t)dt - u(x) \int K_\epsilon(t)dt \right|
\]

\[
= \left| \int_{||t|| < \delta} (u(x-t) - u(x))K_\epsilon(t)dt \right|
\]

\[
+ \int_{||t|| \geq \delta} (u(x-t) - u(x))K_\epsilon(t)dt.
\]

\[
\leq \eta \int_{||t|| < \delta} |K_\epsilon(t)|dt + \int_{||t|| \geq \delta} |u(x-t) - u(x)||K_\epsilon(t)|dt
\]

\[
\leq \eta ||K||_1 + \int_{||t|| \geq \delta} |u(x-t)||K_\epsilon(t)|dt
\]

The third term approaches zero with \(\epsilon\) by Lemma 1. It is enough to show that the second term approaches zero. From
the hypothesis $|K(x)| = \mu(x)||x||^{-p}$ for some non-negative $\mu(x)$ where $\mu(x) \to 0$ as $||x|| \to \infty$. We have

$$\int |u(x-t)||K_\epsilon(t)dt|$$

$$= \int |u(x-t)||\mu\left(\frac{t}{\epsilon}\right)||t||^{-p}dt$$

$$\leq \delta^{-p}\left\{ \sup_{||t||\geq \delta}\mu\left(\frac{t}{\epsilon}\right) \right\} \int |u(x-t)|dt$$

$$\leq \delta^{-p}\left\{ \sup_{||t||\geq \delta}\mu\left(\frac{t}{\epsilon}\right) \right\}||u||_1$$

Again from the hypothesis $\mu\left(\frac{t}{\epsilon}\right) \to 0$ as $|\frac{t}{\epsilon}| \to \infty$. So $\sup_{||t||\geq \delta}\mu\left(\frac{t}{\epsilon}\right) \to 0$ as $\epsilon \to 0$. This concludes the proof.

**Corollary 1:** Let $\nabla u_\epsilon = \nabla u * K_\epsilon$, where $\int_{\mathbb{R}^p} K_\epsilon = 1$, and $K(x) = o(||x||^{-p})$ as $||x|| \to \infty$. Then $\nabla u_\epsilon \to \nabla u$ as $\epsilon \to 0$ at each point of continuity of $\nabla u$. Here the convolution, $\nabla u * K_\epsilon$, is performed component wise.

**Proof:** The result is obtained by applying Theorem 1 to each component of $\nabla u$.

**B. Stochastic Optimization**

Stochastic optimization [17] deals with the study of iterative algorithms of the type

$$x_{k+1} = x_k + a_k[\nabla h(x_k) + N_{k+1}]$$

(1)

Here $x_k \in \mathbb{R}^p$, $k \geq 0$ are the parameters that are updated according to (1). The function $h : \mathbb{R}^p \to \mathbb{R}$ is an underlying cost function whose minimum we are interested in finding. Also, $a_k, k \geq 0$ is a prescribed step-size sequence. Further, $N_{k+1}, k \geq 0$ constitute the noise terms. We state here a theorem that is utilized in the convergence analysis of our algorithm. Consider the following assumptions [18], [19].

**A1.** The step-sizes $a_k, k \geq 0$ satisfy the requirements:

$$a_k > 0 \quad \forall k, \sum_k a_k = \infty, \sum_k a_k^2 < \infty.$$ 

**A2.** The sequence $N_k, k \geq 0$ is a martingale difference sequence with respect to the following increasing sequence of sigma fields:

$$\mathcal{F}_k := \sigma\{x_0, N_1, \ldots, N_k\}, k \geq 0.$$ 

Thus, in particular, $\forall k \geq 0$,

$$\mathbb{E}[N_{k+1}|\mathcal{F}_k] = 0 \text{ a.s.}$$

Further $N_k, k \geq 0$ are square integrable and

$$\mathbb{E}[||N_{k+1}||^2|\mathcal{F}_k] \leq C(1 + ||x_k||^2) \text{ a.s.}$$

for a given constant $C > 0$.

**A3.** The function $\nabla h : \mathbb{R}^p \to \mathbb{R}^p$ is Lipschitz continuous.

**A4.** The functions $\nabla h_c(x) := \frac{\nabla h(x)}{c}$, $c \geq 1, x \in \mathbb{R}^p$, satisfy $\nabla h_c(x) \to \nabla h_\infty(x)$ as $c \to \infty$, uniformly on compacts. Furthermore, the o.d.e

$$\dot{x}(t) = \nabla h_\infty(x(t))$$

(2)

has the origin as the unique globally stable equilibrium [20]. Consider the ordinary differential equation

$$\dot{x}(t) = \nabla h(x(t)).$$

(3)

Let $H$ denote the compact set of asymptotically stable equilibrium points of the ODE (3).

**Theorem 2:** Under (A1)-(A4), $\sup_n ||x_n|| < \infty$ (stability) a.s. Further $x_k \to H$ almost surely as $k \to \infty$.

**Proof:** Follows as a consequence of Theorem 2 and 7 of [18].

**III. MOTIVATION AND ALGORITHM**

In this section we motivate and present our iterative algorithm for estimating the mode of a unimodal density. The idea of computing the mode is described below. Let $f$ denote the unimodal density function. As mode is the maximizer of the density function, we can estimate the mode by gradient ascent as follows:

$$m_{n+1} = m_n + a_n \nabla f(m_n),$$

(4)

where $a_n$ and $m_n$ are the step-size and current mode estimate, respectively at time $n$.

We introduce a function $g$ defined as follows:

$$g(m) = f(m) - \frac{1}{2}\lambda||m||^2,$$

(5)

where $\lambda > 0$ is the regularization coefficient [15]. The idea is to find a $m$ that maximizes the function $g(m)$. This is done to maintain the stability of the estimates in our algorithm (refer proposition 3 in Section IV). Therefore the gradient ascent update is performed as follows:

$$m_{n+1} = m_n + a_n \nabla g(m_n)$$

(6)

$$= m_n + a_n(\nabla f(m_n) - \lambda m_n)$$

(7)

It remains to be shown that solution obtained using this update equation (7) converges to the mode obtained using the update equation (4) as $\lambda \to 0$. Let

$$\dot{m}(\lambda) := \arg \max g(m)$$

and

$$m^* := \arg \max f(m).$$

It is easy to see that

$$\dot{m}(0) = m^*.$$ 

From the continuity of $\arg \max$ function given by the Maximum Theorem [21] we have as $\lambda \to 0$,

$$\dot{m}(\lambda) \to \dot{m}(0) = m^*.$$ 

The update equation (7), however needs the information of $\nabla f(m_n)$, which is not known. We therefore make use of the ideas in the previous section II to estimate $\nabla f(m)$ as follows. To make the notations easy, we replace $m_n$ with $m$ and derive
Applying Corollary 1 to \(\nabla f\) with the kernel \(K\) we get for small \(\epsilon > 0\)
\[
\nabla f(m) \approx \nabla f_{\epsilon}(m) = \int_{\mathbb{R}^p} \nabla f(m - t) K_{\epsilon}(t) dt.
\]
(8)

By the properties of convolution
\[
\int_{\mathbb{R}^p} \nabla f(m - t) K_{\epsilon}(t) dt = \int_{\mathbb{R}^p} \nabla K_{\epsilon}(m - t) f(t) dt = E_{X \sim \nu}[\nabla K_{\epsilon}(m - X)].
\]
(9)

Note that there are several valid choices of functions \(K\) (also called kernels) to obtain approximation of identity.

Now, (7) can be re-written using stochastic gradient ascent as follows:
\[
m_{n+1} = m_n + a_n (\nabla K_{\epsilon}(m_n - X_{n+1}) - \lambda m_n),
\]
(10)

where \(X_{n+1}\) is the sample obtained at time \(n + 1\).

In the following table, we indicate some of the popular kernels [22].

| Name of the Kernel | Analytical Form |
|-------------------|-----------------|
| Gaussian          | \( \frac{1}{\sqrt{2\pi}} e^{-x^2} \) |
| Cauchy            | \( \frac{1}{\pi (1 + x^2)} \) |
| Fejer             | \( \frac{\sin(x)}{\pi x} \) |
| Multivariate Gaussian | \( \frac{1}{(2\pi)^{n/2}} \exp(-\frac{1}{2} A X^T A^{-1} X) \) |

**TABLE I**

**EXAMPLES OF KERNELS**

It is easily verified that the kernels defined in Table I satisfy the hypotheses of Corollary 1. The full algorithm for estimating the mode from online streaming data is described in Algorithm 1.

**Algorithm 1 Calculation of the mode**

**Input:** \(\lambda\): A small regularization coefficient.

\(X_{n+1}\) : Sample input at time \(n + 1\).

\(m_n\) : Current estimate of mode

**Output:** mode \(m_{n+1}\) estimated from samples \(X_1, X_2, \ldots, X_{n+1}\)

1. **procedure** MODE(\(X_{n+1}, m_n, a_n\))
2. \(d_{n+1} = \nabla K_{\epsilon}(m_n - X_{n+1}) - \lambda m_n\)
3. \(m_{n+1} = m_n + a_n d_{n+1}\)
4. **return** \(m_{n+1}\)

Let \(m_0\) denote an initial mode estimate and \(\epsilon\), a small constant. The algorithm works as follows. At time \(n + 1\), the algorithm takes as input the current mode estimate \(m_n\) and the sample \(X_{n+1}\). It then computes the direction \(d_{n+1}\) and updates current approximation of the mode \(m_n\) along \(d_{n+1}\) as shown in step 2. The output of the algorithm is the updated mode estimate computed from samples obtained so far, i.e., \(X_1, \ldots, X_{n+1}\). We prove the convergence of the algorithm in the next section.

**IV. CONVERGENCE ANALYSIS**

Let \(\mathcal{F}_k = \sigma(m_j, 0 \leq j \leq k; X_j, 0 < j \leq k), k \geq 0\) be a sequence of sigma fields. Observe that \(\mathcal{F}_k\) forms a filtration. Let \(d_{k+1} = \nabla K_{\epsilon}(m_k - X_{k+1}) - \lambda m_k\). Note that \(d_k\) is \(\mathcal{F}_k\)-measurable. Moreover \(d_k\) is integrable i.e. \(E[|d_k|] < \infty\) under the assumption that \(\nabla K_{\epsilon}\) is Lipschitz continuous. Now the basic algorithm can be written as
\[
m_{k+1} = m_k + a_k d_{k+1} = m_k + a_k (E[d_{k+1}|\mathcal{F}_k] + N_{k+1})
\]
(11)

where \(N_{k+1} = d_{k+1} - E[d_{k+1}|\mathcal{F}_k]\) is a mean zero term. Also, \(N_{k+1}\) constitutes a martingale difference sequence (see proof of Proposition 1). Here \(E[.]\) is the expectation with respect to the density \(f\).

Our convergence analysis rests on Theorem 2. Our algorithm is in the form of the general iterative scheme (11) with \(\nabla h(m_k) = E[d_{k+1}|\mathcal{F}_k] = \nabla f_{\epsilon}(m_k) - \lambda m_k\) and \(N_{k+1} = d_{k+1} - E[d_{k+1}|\mathcal{F}_k]\). We choose Gaussian kernel for our analysis of the algorithm. Similar analysis can be carried out for other choice of kernels. We validate the assumptions of Theorem 2 and apply them below.

The choice \(a_k = 1/k, k \geq 1\) assures assumption A1. The following proposition validates assumption A2.

**Proposition 1:** \((N_k, \mathcal{F}_k), k \geq 0\) is a martingale difference sequence with
\[
E[|N_{k+1}|^2|\mathcal{F}_k] \leq C(1 + |m_k|^2),
\]
for all \(n \geq 0\) and for some \(C > 0\).

**Proof:** It is easy to see that \(E[N_{k+1}|\mathcal{F}_k] = E[d_{k+1} - E[d_{k+1}|\mathcal{F}_k]|\mathcal{F}_k] = 0\). So clearly \((N_k, \mathcal{F}_k)\) is a martingale difference sequence. Now
\[
E[|N_{k+1}|^2|\mathcal{F}_k] \leq 2 \left( E[|d_{k+1}|^2|\mathcal{F}_k] + E[|E[d_{k+1}|\mathcal{F}_k]|^2|\mathcal{F}_k] \right) \\
\leq 4 E[|d_{k+1}|^2|\mathcal{F}_k].
\]

The first inequality follows from the simple identity \((a - b)^2 \leq 2(a^2 + b^2)\), while the second inequality follows from a simple application of Jensen’s inequality. Since the higher derivatives and in particular Hessian of \(K_{\epsilon}\) is bounded, it follows that \(\nabla K_{\epsilon}\) is Lipschitz continuous. We have
\[
|\nabla K_{\epsilon}(m)| - |\nabla K_{\epsilon}(0)| \leq |\nabla K_{\epsilon}(m) - \nabla K_{\epsilon}(0)| \leq L|m|,
\]
where \(L > 0\) is the Lipschitz constant. Hence for all \(m,\)
\[
|\nabla K_{\epsilon}(m)| \leq C_0(1 + |m|),
\]
(12)
where $C_0 = \max\{\|\nabla K_\epsilon(0)\|, L\}$. Therefore,

$$
E[\|N_{k+1}\|^2|F_k]
\leq 4E[\|d_{k+1}\|^2|F_k]
\leq 8E[\|\nabla K_\epsilon(m_k - X_{k+1})\|^2 + \|\lambda m_k\|^2|F_k]
\leq 8E[2C_0^2(1 + \|m_k - X_{k+1}\|^2)^2 + \lambda^2\|m_k\|^2|F_k]
\leq 8(2C_0^2 + (4C_0^2 + \lambda^2)\|m_k\|^2 + 4C_0^2\|X_{k+1}\|^2)|F_k|
= (C + \|m_k\|^2)
$$

where $C_1 = E[\|X_{k+1}\|^2]$ and $C = 8\max\{2C_0^2 + 4C_0^2\lambda + \lambda^2\}$. This completes the proof.

The following lemma is useful in proving assumption A3 (see Proposition 2).

**Lemma 2:** $E[d_{k+1}|F_k] = \nabla f(x_k) - \lambda m_k$.

**Proof:** Now $E[d_{k+1}|F_k] = E[\nabla K_\epsilon(m_k - X_{k+1})|F_k] - \lambda m_k$. Also $\nabla K_\epsilon$ is analytic and has a power series expansion around $m_k$. Using power series of $\nabla K_\epsilon$, linearity of the expectation and independence of $X_{k+1}$ from $F_k$ we obtain $E[\nabla K_\epsilon(m_k - X_{k+1})|F_k] = E[\nabla K_\epsilon(m_k - X)] = \nabla f(x_k)$.

Owing to Lemma 2, our iterative update (11) transforms into $m_{k+1} = m_k + a_k(\nabla f(x_k) - \lambda m_k + N_{k+1})$ and we validate assumption A3 below.

**Proposition 2:** $\nabla f(x) - \lambda m$ is Lipschitz continuous.

**Proof:** Now for any $x$ and $y$

$$
\|\nabla f(x) - \nabla f(y) - \lambda(x - y)\|
\leq \|E[\nabla K_\epsilon(x - T) - \nabla K_\epsilon(y - T)]\| + \lambda\|x - y\|
\leq \|E[\nabla K_\epsilon(x - T) - \nabla K_\epsilon(y - T)]\| + \lambda\|x - y\|
\leq (L + \lambda)\|x - y\|
$$

where $L$ is the Lipschitz constant of $\nabla K_\epsilon$.

The following proposition proves assumption A4.

**Proposition 3:** The ODE $\dot{m} = h_\infty(m)$ has the origin as its unique globally asymptotically stable equilibrium point.

**Proof:** From the definition of $h_\infty(m)$, see assumption A4, we have

$$
h_\infty(m) = \lim_{c \to \infty} \frac{\nabla f(x(c m)) - \lambda c m}{c}
= \lim_{c \to \infty} \frac{E[\nabla K_\epsilon(cm - X)]}{c} - \lambda m
= \lim_{c \to \infty} \frac{1}{c} \int_{\mathbb{R}^p} \frac{2(x - cm)}{e^{3/\sqrt{\pi}} c^3} e^{-\frac{\|x - cm\|^2}{c^2}} f(x)dx - \lambda m
= \lim_{c \to \infty} \frac{1}{c} \int_{\mathbb{R}^p} \frac{2x}{e^{3/\sqrt{\pi}} c^3} e^{-\frac{\|x - cm\|^2}{c^2}} f(x)dx
- \lambda m
= -\lambda m
$$

where the facts that $e^{-\frac{\|x - cm\|^2}{c^2}} \leq 1$ and $\int_{\mathbb{R}^p} \|f(x)\|dx < \infty$ are utilized. By the application of Dominated Convergence Theorem [22] we have

$$
\frac{2m}{e^{3/\sqrt{\pi}}} \lim_{c \to \infty} \frac{1}{\int_{\mathbb{R}^p} e^{-\frac{\|x - cm\|^2}{c^2}} f(x)dx}
= \frac{2m}{e^{3/\sqrt{\pi}}} \lim_{c \to \infty} \int_{\mathbb{R}^p} e^{-\frac{\|x - cm\|^2}{c^2}} f(x)dx = 0.
$$

So we have that $h_\infty(m) = -\lambda m$.

Now for the system $\dot{m} = h_\infty(m) = -\lambda m$, clearly origin is an equilibrium point. Also for any initial point $m_0$, $m(t) = m_0 \exp(-\lambda t)$ is the solution of the system and $m(t) \to 0$ as $t \to \infty$. Therefore origin is the unique globally asymptotically stable equilibrium point. This concludes the proof.

**Remark 1:** Note that the regularization coefficient $\lambda$ plays a key role in establishing the stability of mode estimates. To see the effect of regularization term consider the iterates

$$
m_{n+1} = m_n + a_n \nabla f(x(m_n)).
$$

The $h_\infty(.)$ corresponding to this update equation is identically 0 there by violating assumption A4.

Consider now the following ODE:

$$
\dot{x} = \nabla f(x(t)) - \lambda x
$$

Let $\dot{H}$ be the set of asymptotically stable equilibrium points of (13).

**Remark 2:** From our assumptions as $\lambda \to 0$, $\dot{H} \to \{m^*\}$

We have the following as our main result.

**Theorem 3:** $m_n$ obtained from Algorithm 1 satisfies $m_n \to \dot{H}$ a.s.

**Proof:** The result follows from the foregoing and Theorem 2.

**V. EXPERIMENTS**

In this section, we discuss the numerical performance of our algorithm. We implement our algorithm on known popular distributions. We collect $10^6$ samples from a known distribution and apply our algorithm for estimating the mode. The initial mode is selected as the average of initial 1000 points. We consider Gaussian kernel for univariate distributions and multivariate Gaussian kernel for bivariate normal and Dirichlet distributions (see Table [1]) for our experiments. The regularization coefficient $\lambda$ is chosen to be $10^{-5}$ and $c$ is set to 1. We perform 100 runs of the experiment and estimated mode is calculated as the mean of modes obtained.
over the 100 runs. The following Table II illustrates the performance of our algorithm (estimated mode) on standard distributions. We have also indicated the actual mode of the distribution in the Table II. The code for our experiments can be found at [1].

| Distribution | Actual Mode | Estimated Mode ± Std.Dev |
|--------------|-------------|--------------------------|
| Normal       | 10          | 9.91783 ± 0.333930       |
| Gamma        | 5           | 5.182626 ± 0.036357      |
| Exponential  | 0           | 0.697886 ± 0.007722      |
| Weibull      | 0           | 0.697192 ± 0.007544      |
| Beta         | 1           | 0.900645 ± 0.001356      |
| Bivariate Normal | [20; 15] | 20.030644 ± 15.015614 ± 0.00 |
| Dirichlet    | [0.5; 0.5] | [0.498404; 0.501579] ± 0.00 |

TABLE II
PERFORMANCE OF OUR PROPOSED ALGORITHM ON STANDARD DISTRIBUTIONS

It is interesting to note that, though Exponential and Weibull densities are not smooth and do not satisfy our assumptions, the estimated mode obtained by our algorithm is closer to the actual mode.

In the Figure 1, we indicate the performance of our algorithm with different initial points. For this purpose, we select Normal distribution with mean 10. We implement our algorithm with initial points 5, 10 and 15 and plot the estimated mode over initial 50,000 iterations. We observe that the estimates of the mode in all the three cases converge towards the actual mode having value 10 as the number of iterations increase. This shows that the proposed algorithm is not very sensitive with respect to the initial mode point. These results confirm the practical utility of our algorithm.

Newton’s method is known to converge faster than the gradient descent method. This estimation of mode may possibly lead to new ideas in Bayesian Statistics. For example, the core idea in Maximum Aposteriori Estimate (MAP) is to find the mode of the posterior distribution from the given samples. Another interesting future direction is to obtain finite sample error bounds and rate of convergence results for our algorithm.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an online computationally efficient algorithm for computing the mode from the samples of an unknown density. We have provided the proofs for the stability of the iterates and convergence of our algorithm. Next, we did numerical experiments on standard distributions and showed the effectiveness of our algorithm in practice.

In future, we wish to propose second order updates using say the Newton’s method in the place of gradient descent.

REFERENCES

[1] E. Parzen, “On estimation of a probability density function and mode,” The annals of mathematical statistics, vol. 33, no. 3, pp. 1065–1076, 1962.
[2] U. Grenander, “Some direct estimates of the mode,” The Annals of Mathematical Statistics, pp. 131–138, 1965.
[3] H. Chernoff, “Estimation of the mode,” Annals of the Institute of Statistical Mathematics, vol. 16, no. 1, pp. 31–41, 1964.
[4] E. J. Wegman, “A note on the estimation of the mode,” The Annals of Mathematical Statistics, pp. 1909–1915, 1971.
[5] T. Dalenius, “The mode—a neglected statistical parameter,” Journal of the Royal Statistical Society. Series A (General), pp. 110–117, 1965.
[6] J. Venter, “On estimation of the mode,” The Annals of Mathematical Statistics, pp. 1446–1455, 1967.
[7] T. Robertson and J. D. Cryer, “An iterative procedure for estimating the mode,” Journal of the American Statistical Association, vol. 69, no. 348, pp. 1012–1016, 1974.
[8] D. R. Bickel, “Robust estimators of the mode and skewness of continuous data,” Computational statistics & data analysis, vol. 39, no. 2, pp. 153–163, 2002.
[9] D. R. Bickel and R. Frühwirth, “On a fast, robust estimator of the mode: comparisons to other robust estimators with applications,” Computational Statistics & Data Analysis, vol. 50, no. 12, pp. 3500–3530, 2006.
[10] D. R. Bickel, “Robust and efficient estimation of the mode of continuous data: the mode as a viable measure of central tendency,” Journal of statistical computation and simulation, vol. 73, no. 12, pp. 899–912, 2003.
[11] C.-Y. Hsu and T.-J. Wu, “Efficient estimation of the mode of continuous multivariate data,” Computational Statistics & Data Analysis, vol. 63, pp. 148–159, 2013.
[12] L. D. Griffin and M. Lillholm, “A multiscale mean shift algorithm for mode estimation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.
[13] T. Kirschstein, S. Liebscher, G. C. Porzio, and G. Ragozini, “Minimum volume peeling: A robust nonparametric estimator of the multimode,” Computational Statistics & Data Analysis, vol. 93, pp. 456–468, 2016.
[14] A. B. Tsybakov, “Recursive estimation of the mode of a multivariate distribution,” Problemy Peredachi Informatsii, vol. 26, no. 1, pp. 38–45, 1990.
[15] A. N. Tikhonov, A. Goncharsky, V. Stepanov, and A. G. Yagola, Numerical methods for the solution of ill-posed problems. Springer Science & Business Media, 2013, vol. 308.
[16] L. Ljung, “Analysis of recursive stochastic algorithms,” IEEE transactions on automatic control, vol. 22, no. 4, pp. 551–575, 1977.
[17] H. Kushner and G. G. Yin, Stochastic approximation and recursive algorithms and applications. Springer Science & Business Media, 2003, vol. 35.
[18] V. S. Borkar, Stochastic approximation: a dynamical systems viewpoint. Springer, 2009, vol. 48.
[19] S. Bhatnagar, H. Prasad, and L. Prashanth, Stochastic recursive algorithms for optimization: simultaneous perturbation methods. Springer, 2012, vol. 434.
[20] M. W. Hirsch, S. Smale, and R. L. Devaney, Differential equations, dynamical systems, and an introduction to chaos. Academic press, 2012.
[21] C. Berge, Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity. Courier Corporation, 1997.
[22] R. L. Wheeden, Measure and integral: an introduction to real analysis. CRC press, 2015, vol. 308.

Fig. 1. Performance of proposed algorithm with different initial points

[https://github.com/raghudiddigi/Mode-Estimation]