Convergence of the Expectation-Maximization Algorithm Through Discrete-Time Lyapunov Stability Theory

Orlando Romero[†]   Sarthak Chatterjee[‡]   Sérgio Pequito[†]

Abstract— In this paper, we propose a dynamical systems perspective of the Expectation-Maximization (EM) algorithm. More precisely, we can analyze the EM algorithm as a nonlinear state-space dynamical system. The EM algorithm is widely adopted for data clustering and density estimation in statistics, control systems, and machine learning. This algorithm belongs to a large class of iterative algorithms known as *proximal point methods*. In particular, we re-interpret limit points of the EM algorithm and other local maximizers of the likelihood function it seeks to optimize as equilibria in its dynamical system representation. Furthermore, we propose to assess its convergence as asymptotic stability in the sense of Lyapunov. As a consequence, we proceed by leveraging recent results regarding discrete-time Lyapunov stability theory in order to establish asymptotic stability (and thus, convergence) in the dynamical system representation of the EM algorithm.

I. Introduction

With the ever-expanding size and complexity of datasets used in the field of statistics, control systems, and machine learning, there has been a growing interest in developing algorithms that efficiently find the solution to the optimization problems that arise in these settings. For example, a fundamental problem in exploratory data mining is the problem of *cluster analysis*, where the central task is to group objects into subgroups (*i.e.*, *clusters*) such that the objects in a particular cluster share several characteristics (or, features) with those that are, in some sense, sufficiently different from objects in different clusters [1], [2].

The Expectation-Maximization (EM) algorithm [3] is one of the most popular methods used in distribution-based clustering analysis and density estimation [4], [5]. Given a dataset, we can assume that the data is distributed according to a finite mixture of Gaussian distributions whose parameters are randomly initialized and iteratively improved using the EM algorithm that seeks to maximize the likelihood that the data is justified by the distributions. This leads to finding the finite Gaussian mixture that hopefully best fits the dataset in question.

A current trend in optimization, machine learning, and control systems, is that of leveraging on a dynamical systems interpretation of iterative optimization algorithms [6]–[8]. The key idea is to view the estimates themselves in the iterations of the algorithm as a state vector at different discrete instances of time (in particular, the initial approximation is viewed as the initial state), while the mechanism itself

[†]Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy NY, 12180, USA.

[‡]Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy NY, 12180, USA.

used to construct each subsequent estimate is modeled as a state-space dynamical system. Then, local optimizers and convergence in the optimization algorithm roughly translate to equilibria and asymptotic stability (in the sense of Lyapunov) in its dynamical system interpretation.

The convergence of the EM algorithm has been studied from the point-of-view of general point-to-set notions of convergence of optimization algorithms such as Zangwill's convergence theorem [9]. Works such as [10] and [11] provide proofs of the convergence of the sequence of estimates generated by the EM algorithm.

The main contribution of this paper is to present a dynamical systems perspective of the convergence of the EM algorithm. The convergence of the EM algorithm is well known. However, our nonlinear stability analysis approach is intended to help open the field to new iteration schemes by possible addition of an artificial external input in the dynamical system representation of the EM algorithm. Then, leveraging tools from feedback systems theory, we could design a control law that translates to an accelerated convergence of the algorithm for specific subclassess of distributions.

The rest of the paper is organized as follows. In Section II, we briefly review the problem of maximum likelihood estimation and the EM algorithm. In Section III, we propose a dynamical systems perspective of the EM algorithm and propose a particular generalized EM (GEM) algorithm. In Section IV we establish our main convergence results by leveraging discrete-time Lyapunov stability theory. Finally, Section V concludes the paper.

II. Expectation-Maximization Algorithm

In this section, we recall the *Expectation-Maximization* (EM) algorithm. Let $\theta \in \Theta \subseteq \mathbb{R}^p$ be some vector of unknown (but deterministic) parameters characterizing a distribution of interest, which we seek to infer from a collected dataset $y \in \mathbb{R}^m$ (from now on assumed fixed). To estimate θ from the dataset y, we first need a statistical model (*i.e.*, an indexed class of probability distributions $\{p_\theta(y) : \theta \in \Theta\}$). The function $\mathcal{L} : \Theta \to \mathbb{R}$ given by

$$\mathcal{L}(\theta) = p_\theta(y) \qquad (1)$$

denotes the *likelihood function*. The objective is to compute a *maximum likelihood estimate* (MLE)

$$\hat{\theta}_{\text{MLE}} \in \operatorname*{argmax}_{\theta \in \Theta} \mathcal{L}(\theta). \qquad (2)$$

We will treat

$$\underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta) = \left\{ \theta \in \Theta : \mathcal{L}(\theta) = \max_{\theta' \in \Theta} \mathcal{L}(\theta') \right\} \quad (3)$$

as a set, unless it consists of a single point $\theta^\star$, in which case we may use $\theta^\star = \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta)$.

Next, we will introduce some assumptions that will ensure well-definedness throughout this paper.

Assumption 1. $\mathcal{L}(\theta) > 0$ for every $\theta \in \Theta$. ○

This assumption is simply a mild technical condition intended to avoid pathological behaviors, and is satisfied by most mixtures of distributions used in practice (*e.g.*, Gaussian, Poisson, Beta). Furthermore, we surely have $\mathcal{L}(\theta) > 0$ for at least *some* $\theta \in \Theta$ (since, otherwise, the dataset y is entirely useless regarding maximum likelihood estimation), and thus it suffices that we disregard from Θ any θ such that $\mathcal{L}(\theta) = 0$.

The underlying assumption for the EM algorithm is that there exists some latent (non-observable) random vector $x \in \mathcal{X} \subseteq \mathbb{R}^n$ for which we possess a "complete" statistical model $\{p_\theta(x, y) : x \in \mathcal{X}, \theta \in \Theta\}$ (as opposed to the "incomplete" model $\{p_\theta(y) : \theta \in \Theta\}$), and for which maximizing the expected value of the *complete log-likelihood function* is easier than the *incomplete likelihood function*. However, since x is latent, the idea behind the EM algorithm is to iteratively maximize the expected complete log-likelihood.

Assumption 2. $\mathcal{X} = \{x \in \mathbb{R}^n : p_\theta(x, y) > 0\}$ does not depend on $\theta \in \Theta$. ○

Together with Assumption 1, this assumption will further allow us to avoid certain pathological cases. Specifically, we can properly define the expected log-likelihood function (hereafter, also referred to as the Q-function) $Q : \Theta \times \Theta \to \mathbb{R}$, defined as

$$Q(\theta, \theta') \stackrel{\text{def}}{=} \mathbb{E}_{p_{\theta'}(\cdot|y)}[\log p_\theta(\cdot, y)] \quad (4a)$$

$$= \int_{\mathcal{X}} p_{\theta'}(x|y) \log p_\theta(x, y) \, dx, \quad (4b)$$

where $p_{\theta'}(x|y) = p_{\theta'}(x, y)/p_{\theta'}(y)$. With all these ingredients and assumptions, we summarize the EM algorithm in Algorithm 1. Notice that, the term $Q(\cdot, \theta_k)$ in Algorithm 1 denotes the expected complete log-likelihood function for any given iteration k.

Remark 1. In practice, the iterations of Algorithm 1 are computed until some stopping criterion is achieved, such that it approximates θ_∞. ○

III. DYNAMICAL SYSTEM INTERPRETATION OF THE EM ALGORITHM AND CONVERGENCE

Formally, the convergence of the EM algorithm is concerned with the existence and characteristics of the limit of the sequence $\{\theta_k\}_{k \in \mathbb{Z}_+}$ as $k \to \infty$. In particular, *local convergence* refers to the property of the sequence $\{\theta_k\}_{k \in \mathbb{Z}_+}$ converging to the *same* point $\theta^\star$ for every initial approximation θ_0 that is sufficiently close to $\theta^\star$. On the other

---

Algorithm 1 Expectation-Maximization (EM)

Input: Observed data $y \in \mathbb{R}^m$, complete statistical model $\{p_\theta(x, y) : x \in \mathcal{X}, \theta \in \Theta\}$, and initial approximation θ_0 of $\hat{\theta}_{\text{MLE}} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(\theta)$.

Output: θ_∞ such that hopefully $p_{\theta_\infty}(y) \approx \max_{\theta \in \Theta} \mathcal{L}(\theta)$.

1: **for** $k = 0, 1, 2, \ldots$ **do**
2:    **E-step:** Compute $Q(\theta, \theta_k)$
3:    **M-step:** Determine $\theta_{k+1} \in \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta, \theta_k)$
4: **end for**
5: **return** $\theta_\infty = \lim_{k \to \infty} \theta_k$, if it exists.

---

hand, *global convergence* refers to convergence to the same point for *any* initial approximation. Ideally, $\theta^\star$ is a global maximizer (or at least a local one) of the likelihood function. In practice, Algorithm 1 may converge to other stationary points of the likelihood function [3], [10].

We will now see how the EM algorithm (such as many other iterative optimization algorithms) can be interpreted as a dynamical system in state-space, for which *convergence* translates to *(asymptotic) stability* in the sense of Lyapunov.

To start, recall that any discrete-time time-invariant nonlinear dynamical system in state-space can be described by its dynamics, which are of the form

$$\begin{cases} \theta[k + 1] = F(\theta[k]), & k \in \mathbb{Z}_+ = \{0, 1, 2, \ldots\}, \\ \theta[0] = \theta_0, \end{cases} \quad (\mathcal{S})$$

and where $\theta[k]$ denotes the *state* of the system and $F : \Theta \to \Theta$ is some known function. In particular, any F that satisfies

$$F(\theta') \in \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta, \theta') \quad (5)$$

for every $\theta' \in \Theta$ represents a particular realization of the different iterations of Algorithm 1. For the sake of simplicity, let us make the following assumption.

Assumption 3. $Q(\cdot, \theta')$ has a unique global maximizer in Θ for each $\theta' \in \Theta$. ○

Remark 2. Assumption 3 does not imply that the likelihood function has a unique global maximizer, and subsequently, the MLE may still be non-unique. Furthermore, under Assumption 3, the sequence $\{\theta_k\}_{k \in \mathbb{Z}_+}$ generated by Algorithm 1 is unique for each $\theta_0 \in \Theta$, the function $F^{\text{EM}} : \Theta \to \Theta$ given by

$$F^{\text{EM}}(\theta') = \underset{\theta \in \Theta}{\operatorname{argmax}} Q(\theta, \theta') \quad (\text{EM})$$

for $\theta' \in \Theta$ is uniquely defined, and $(\mathcal{S})$ with $F = F^{\text{EM}}$ captures the dynamical evolution emulated by Algorithm 1, *i.e.* $\theta[k] = \theta_k$ for every $k \in \mathbb{Z}_+$. ○

Recall that, for a dynamical system of the form $(\mathcal{S})$, we say that $\theta^\star$ is an *equilibrium* of the system if $\theta[0] = \theta^\star$ implies that $\theta[k] = \theta^\star$ for every $k \in \mathbb{Z}_+$. In other words, if $\theta^\star$ is a fixed point of $F(\theta)$, *i.e.*, $F(\theta^\star) = \theta^\star$. For self-consistency, we now formally define Lyapunov stability.

Definition 1 (Lyapunov stability). Let $\theta^\star$ be an equilibrium of the dynamical system $(\mathcal{S})$. We say that $\theta^\star$ is *stable* if the trajectory $\theta[k]$ is arbitrarily close to $\theta^\star$ provided that it starts sufficiently close to $\theta^\star$. In other words, if, for any $\varepsilon > 0$, there exists some $\delta > 0$ such that $\theta_0 \in B_\delta(\theta^\star)$ implies that $\theta[k] \in B_\varepsilon(\theta^\star)$ for every $k \in \mathbb{Z}_+$.

Further, we say that $\theta^\star$ is *(locally) asymptotically stable* if, apart from being stable, the trajectory $\theta[k]$ converges to $\theta^\star$ provided that it starts sufficiently close to $\theta^\star$. In other words, if $\theta^\star$ is stable and there exists some $\delta > 0$ such that $\theta_0 \in B_\delta(\theta^\star)$, implies that $\theta[k] \to \theta^\star$ as $k \to \infty$. $\circ$

From the previous definition, it readily follows that asymptotic stability of the trajectory $\theta[k]$ generated by $(\mathcal{S})$ with $F = F^{\text{EM}}$ is nearly equivalent to local convergence of the EM algorithm, since $\theta_k = \theta[k] \to \theta^\star$ for any θ_0 in some sufficiently small compact ball centered around $\theta^\star$. Therefore, to establish local convergence of the EM algorithm from the point of view of the asymptotic stability of the corresponding dynamical system, we first need to establish that the points of interest (*i.e.*, the local maxima of the likelihood function) are equilibria of the system (*i.e.*, fixed points of F^{EM}).

Let $\theta^\star \in \Theta$ be a local maximizer of $\mathcal{L}(\theta)$. More precisely, let us start by considering $\theta^\star = \hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$. Notice that the Q-function can be re-written as follows:

$$Q(\theta, \theta') = \mathbb{E}_{p_{\theta'}(\cdot|y)}[\log p_\theta(\cdot, y)] \tag{6a}$$
$$= \log p_\theta(y) + \mathbb{E}_{p_{\theta'}(\cdot|y)}[\log p_\theta(\cdot|y)] \tag{6b}$$
$$= \log \mathcal{L}(\theta) - \mathcal{D}_{\text{KL}}(\theta'\|\theta) - \mathcal{H}(\theta') \tag{6c}$$

where $\mathcal{D}_{\text{KL}}(\theta'\|\theta)$ denotes the Kullback-Leibler divergence from $p_\theta(\cdot|y)$ to $p_{\theta'}(\cdot|y)$, and $\mathcal{H}(\theta')$ denotes the differential Shannon entropy of $p_{\theta'}(\cdot|y)$ [12]. Since the entropy term in (6c) does not depend on θ, then

$$F^{\text{EM}}(\theta') = \underset{\theta \in \Theta}{\arg\max} \{\log \mathcal{L}(\theta) - \mathcal{D}_{\text{KL}}(\theta'\|\theta)\} \tag{7}$$

for $\theta' \in \Theta$.

Remark 3. In general, the Kullback-Leibler divergence from an arbitrary distribution $q(x)$ to another $p(x)$, denoted as $\mathcal{D}_{\text{KL}}(p\|q)$, satisfies the following two properties: *(i)* $\mathcal{D}_{\text{KL}}(p\|q) \geq 0$ for every p and q, a result known as *Gibbs' inequality*; and *(ii)* $\mathcal{D}_{\text{KL}}(p\|q) = 0$ if and only if $p = q$ almost everywhere. $\circ$

Therefore, we can now state the following.

Proposition 1. *The MLE is an equilibrium of the dynamical system $(\mathcal{S})$ with $F = F^{\text{EM}}$.* $\diamond$

Proof. Inspecting (7) at $\theta' = \hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$, it readily follows that $\log \mathcal{L}(\theta)$ and $-\mathcal{D}_{\text{KL}}(\hat{\theta}_{\text{MLE}}\|\theta)$ are both maximized at $\theta = \hat{\theta}_{\text{MLE}}$, which in turn implies that $F^{\text{EM}}(\hat{\theta}_{\text{MLE}}) = \hat{\theta}_{\text{MLE}}$. ∎

However, it should be clear that the previous argument does not hold for non-global maximizers of the likelihood function. One approach to get around this issue is to consider a specific variant of the *generalized* EM algorithm (GEM)[1]. More precisely, we propose a GEM algorithm that searches for a global maximizer of $Q(\theta, \theta_k)$ in a restricted parameter space: the closure of the δ-ball $B_\delta(\theta_k) = \{\theta \in \Theta : \|\theta - \theta_k\| < \delta\}$ centered around θ_k, denoted as $\overline{B}_\delta(\theta_k)$. We call this algorithm the δ-EM algorithm (not to be confused with the α-EM algorithm [13]), which is summarized in Algorithm 2.

Algorithm 2 δ-Expectation-Maximization (δ-EM)

Input: Restricted parameter space radius $\delta > 0$, observed data $y \in \mathbb{R}^m$, complete statistical model $\{p_\theta(x, y) : x \in \mathcal{X}, \theta \in \Theta\}$, and initial approximation θ_0 of $\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Theta} \mathcal{L}(\theta)$.

Output: θ_∞ such that hopefully $p_{\theta_\infty}(y) \approx \max_{\theta \in \Theta} p_\theta(y)$.

1: **for** $k = 1, 2, \ldots$ **do**
2:     **E-step:** Compute $Q(\theta, \theta_k)$
3:     **M-step:** Determine $\theta_{k+1} \in \underset{\theta \in \overline{B}_\delta(\theta_k) \cap \Theta}{\arg\max} Q(\theta, \theta_k)$
4: **end for**
5: **return** $\theta_\infty = \lim_{k \to \infty} \theta_k$, if it exists.

Next, we make the following simplifying assumption.

Assumption 4. $Q(\cdot, \theta')$ has a unique global maximizer in $\overline{B}_\delta(\theta') \cap \Theta$, for every $\theta' \in \Theta$ and $\delta > 0$. $\circ$

Naturally, we can interpret Algorithm 2 as the dynamical system $(\mathcal{S})$ with $F = F^{\delta-\text{EM}} : \Theta \to \Theta$ given by

$$F^{\delta-\text{EM}}(\theta') \overset{\text{def}}{=} \underset{\theta \in \overline{B}_\delta(\theta_k) \cap \Theta}{\arg\max} Q(\theta, \theta') \tag{δ-EM}$$
$$= \underset{\theta \in \overline{B}_\delta(\theta') \cap \Theta}{\arg\max} \{\log \mathcal{L}(\theta) - \mathcal{D}_{\text{KL}}(\theta'\|\theta)\}, \tag{8}$$

where (8) was derived following a similar argument that led to (7). Therefore, similar to Proposition 1, we have the following result.

Proposition 2. *Any local maximizer of the likelihood function is an equilibrium of the dynamical system $(\mathcal{S})$ with $F = F^{\delta-\text{EM}}$, provided that $\delta > 0$ is small enough.* $\diamond$

Proof. Let $\theta^\star \in \Theta$ be a local maximizer of $\mathcal{L}(\theta)$. Inspecting (8) at $\theta' = \theta^\star$, it readily follows that $-\mathcal{D}_{\text{KL}}(\theta^\star\|\theta)$ is maximized at $\theta = \theta^\star$. Furthermore, if $\delta > 0$ is small enough, then $\log \mathcal{L}(\theta)$ is also maximized (in $\overline{B}_\delta(\theta^\star) \cap \Theta$) at $\theta = \theta^\star$, which implies that $F^{\delta-\text{EM}}(\theta^\star) = \theta^\star$. ∎

IV. LOCAL CONVERGENCE THROUGH DISCRETE-TIME LYAPUNOV STABILITY THEORY

In this section, we will discuss how we can establish the local convergence of the EM algorithm by exploiting classical results from Lyapunov stability theory in the dynamical system interpretation of the EM algorithm.

---

[1]A GEM algorithm is any variant of the EM algorithm where the M-step is replaced by a search of some $\theta_{k+1} \in \Theta$ such that $Q(\theta_{k+1}, \theta_k) > Q(\theta_k, \theta_k)$, if one exists (otherwise $\theta_{k+1} = \theta_k$), not necessarily in $\arg\max_{\theta \in \Theta} Q(\theta, \theta_k)$.

To start, we state the discrete-time version of the Lyapunov theorem (Theorem 1.2 in [14]). First, recall that a function $\mathcal{V} : \Theta \to \mathbb{R}$ is said to be *positive semidefinite*, if $\mathcal{V}(\theta) \geq 0$ for every $\theta \in \Theta$. Furthermore, we say that $\mathcal{V}$ is *positive definite* with respect to $\theta^\star \in \Theta$, if $\mathcal{V}(\theta^\star) = 0$ and $\mathcal{V}(\theta) > 0$ for $\theta \in \Theta \setminus \{\theta^\star\}$.

Theorem 1 (Lyapunov Stability [14]). *Let $\theta^\star \in \Theta$ be an equilibrium of the dynamical system $(\mathcal{S})$ in the interior of Θ and let $\delta > 0$ be such that $\overline{B}_\delta(\theta^\star) \subseteq \Theta$. Suppose that F is continuous and there exists some continuous function $\mathcal{V} : B_\delta(\theta^\star) \to \mathbb{R}$ (called a Lyapunov function) such that $\mathcal{V}$ and $-\Delta\mathcal{V}$ are, respectively, positive definite with respect to $\theta^\star$ and positive semidefinite, where $\Delta\mathcal{V}(\theta) \stackrel{\text{def}}{=} \mathcal{V}(F(\theta)) - \mathcal{V}(\theta)$. Then, $\theta^\star$ is stable. If $-\Delta\mathcal{V}$ is also positive definite with respect to $\theta^\star$, then $\theta^\star$ is asymptotically stable.* ◇

Remark 4. The statement of the theorem, as stated in Theorem 1.2 of [14], assumes local Lipschitz continuity of F. This is likely a residual from the classical assumption of local Lipschitz continuity of F in continuous systems with dynamics of the form $\dot{\theta}(t) = F(\theta(t))$, which is required by the Picard-Lindelöf theorem to ensure unique existence of a solution to the differential equation $\dot{\theta}(t) = F(\theta(t))$ for each initial state $\theta(0) = \theta_0$. For discrete-time systems, on the other hand, the unique existence of the trajectory is immediate. However, a careful analysis of the argument used in the proof found in [14] reveals that the continuity of F is nevertheless implicitly needed to ensure the continuity of $\Delta\mathcal{V}$, since the extreme value theorem is invoked for $\Delta\mathcal{V}$. ○

In order to leverage the previous theorem to establish local convergence of the EM algorithm to local maxima of the likelihood function, we need to propose a *candidate* Lyapunov function. However, before doing so, we need to ensure that $F = F^{\text{EM}}$ is continuous, which is attained by imposing some regularity on the likelihood function.

Assumption 5. $\mathcal{L}(\theta)$ is twice continuously differentiable. ○

Subsequently, we obtain the following result.

Lemma 1. F^{EM} and $F^{\delta-\text{EM}}$ are both continuous. ◇

Proof. Under Assumption 5, it readily follows from (4b) that $Q(\cdot, \cdot)$ is continuous in both of its arguments. Let $\{\theta'_k\}_{k \in \mathbb{Z}_+} \subseteq \Theta$ be a sequence converging to $\theta' \in \Theta$. Note that, for each $k \in \mathbb{Z}_+$, we have $Q(F^{\text{EM}}(\theta'_k), \theta'_k) \geq Q(\theta, \theta'_k)$ for every $\theta \in \Theta$. Taking the limit when $k \to \infty$, and leveraging the continuity of Q, we have $Q\left(\lim_{k \to \infty} F^{\text{EM}}(\theta'_k), \theta'\right) \geq Q(\theta, \theta')$ for every $\theta \in \Theta$. Consequently,

$$Q\left(\lim_{k \to \infty} F^{\text{EM}}(\theta'_k), \theta'\right) = \max_{\theta \in \Theta} Q(\theta, \theta'), \tag{9}$$

and therefore,

$$\lim_{k \to \infty} F^{\text{EM}}(\theta'_k) = \operatorname*{argmax}_{\theta \in \Theta} Q(\theta, \theta') = F^{\text{EM}}(\theta'). \tag{10}$$

This same argument can be readily adapted for $F^{\delta-\text{EM}}$. ∎

Let $\theta^\star \in \Theta$ be a local maximizer of $\mathcal{L}(\theta)$. Once again, let us start by considering $\theta^\star = \hat{\theta}_{\text{MLE}} \in \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta)$.

From Proposition 1, we know that $\theta^\star$ is an equilibrium of $(\mathcal{S})$ for $F = F^{\text{EM}}$. Next, a naive guess of a candidate Lyapunov function would be to consider $\mathcal{V}(\theta) = \mathcal{L}(\theta)$, since this would satisfy $\mathcal{V}(\theta) \geq 0$ for every $\theta \in \Theta$, but $\mathcal{V}(\theta^\star) > 0$; hence, it is not a Lyapunov function. Notwithstanding, if we subtract $\mathcal{L}(\theta^\star)$ from the previous candidate, *i.e.*, $\mathcal{V}(\theta) = \mathcal{L}(\theta) - \mathcal{L}(\theta^\star)$, then $\mathcal{V}(\theta^\star) = 0$ and $\mathcal{V}(\theta) < 0$ for $\theta \in \Theta \setminus \operatorname{argmax}_{\theta' \in \Theta} \mathcal{L}(\theta')$. As a consequence, it should be clear that

$$\mathcal{V}(\theta) = \mathcal{L}(\theta^\star) - \mathcal{L}(\theta) \tag{11}$$

appears to be the ideal candidate, since $\mathcal{V}(\theta) \geq 0$ for every $\theta \in \Theta$ and $\mathcal{V}(\theta^\star) = 0$. Yet, $\mathcal{V}$ may be only positive semidefinite instead of positive definite (with respect to $\theta^\star$), since $\mathcal{V}(\theta) = 0$ if and only if $\theta \in \operatorname{argmax}_{\theta' \in \Theta} \mathcal{L}(\theta')$. To circumvent this issue, we will need to assume that $\theta^\star$ is an isolated maximizer[2] of $\mathcal{L}(\theta)$.

Lemma 2. *Suppose that $\theta^\star \in \Theta$ is an isolated maximizer of $\mathcal{L}(\theta)$. Then, $\mathcal{V} : B_r(\theta^\star) \to \mathbb{R}$ given by (11) with $F = F^{\text{EM}}$ or $F = F^{\delta-\text{EM}}$ is positive definite with respect to $\theta^\star$ and $-\Delta\mathcal{V}$ is positive semidefinite, provided that $r > 0$ is small enough.* ◇

Proof. Let us focus on the case $F = F^{\text{EM}}$. The positive definiteness follows from (11) and the definition of isolated maximizer. On the other hand, since $\mathcal{D}_{\text{KL}}(\theta \| \theta) = 0$, it follows from (7) that

$$\log \mathcal{L}(F^{\text{EM}}(\theta)) \geq \log \mathcal{L}(\theta) + \underbrace{\mathcal{D}_{\text{KL}}(\theta \,\|\, F^{\text{EM}}(\theta))}_{\geq 0} \tag{12}$$

for every $\theta \in \Theta$. Thus, from the strict monotonicity of the logarithm function, it follows that $-\Delta\mathcal{V}(\theta) = \mathcal{L}(F^{\text{EM}}(\theta)) - \mathcal{L}(\theta)$ is indeed positive semidefinite. The case $F = F^{\delta-\text{EM}}$ follows by essentially the same argument. ∎

Remark 5. We are now in conditions to establish the stability of isolated MLEs in the dynamical system that represents the EM algorithm. Further, through a similar argument, the stability of arbitrary isolated local maximizers of the likelihood for the dynamical system that represents the δ-EM algorithm with small enough $\delta > 0$. However, non-asymptotic stability does not seem to translate into any interesting aspect of the convergence of the EM or δ-EM algorithms. ○

In order to establish the positive definiteness of $-\Delta\mathcal{V}$, we first need to characterize the equilibria of the dynamical systems that represent Algorithms 1 and 2.

Lemma 3. *Every fixed point F^{EM} and $F^{\delta-\text{EM}}$ in the interior of Θ is a stationary point of the likelihood function.* ◇

Proof. Let $\theta^\star \in \Theta$ be a fixed point of F^{EM} or $F^{\delta-\text{EM}}$ in the interior of Θ. From Assumption 5, and from (7) and (8), it

---

[2]We say that $\theta^\star \in \Theta$ is an *isolated* maximizer (stationary point) of $\mathcal{L}(\theta)$ if it is the only local maximizer (stationary point) of $\mathcal{L}(\theta)$ in $B_r(\theta^\star) \cap \Theta$ for some small enough $r > 0$.

follows that

$$\frac{\partial}{\partial \theta} \Big\{ \log \mathcal{L}(\theta) - \mathcal{D}_{\mathrm{KL}}(\theta^\star \| \theta) \Big\} \Big|_{\theta = \theta^\star} = 0, \qquad (13)$$

where the divergence term actually vanishes, since $\theta = \theta^\star$ is a (global) minimizer of $\mathcal{D}_{\mathrm{KL}}(\theta^\star \| \theta)$. Thus, it follows that $\theta^\star$ is a stationary point of $\mathcal{L}(\theta)$. ∎

Next, we will need to assume additional regularity on the likelihood function. A very common assumption for a parameterized statistical model is for the parameterization to be injective, meaning that each distribution in the model is indexed by exactly one instance of the parameter space.

Assumption 6. $\theta \mapsto p_\theta(\cdot | y)$ is injective. ○

Equipped with the last two assumptions (Assumption 5 and 6), we are ready to establish the positive definiteness of $-\Delta \mathcal{V}$, and subsequently, the local convergence of Algorithm 1 to ML estimates.

Theorem 2 (Local Convergence of EM to MLE). *If* $\nabla^2 \mathcal{L}(\hat{\theta}_{\mathrm{MLE}}) \prec 0$, *then the sequence* $\{\theta_k\}_{k \in \mathbb{Z}_+}$ *generated by Algorithm 1 converges to* $\hat{\theta}_{\mathrm{MLE}} = \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta)$ *for every initial approximation* $\theta_0 \in \Theta$ *in a small enough open ball centered around* $\hat{\theta}_{\mathrm{MLE}}$. ◇

Proof. First, recall from Proposition 1 that $\hat{\theta}_{\mathrm{MLE}}$ is an equilibrium of $(\mathcal{S})$ with $F = F^{\mathrm{EM}}$ (*i.e.*, a fixed point of F^{EM}), and from Lemma 1 that F^{EM} is continuous. Furthermore, $\hat{\theta}_{\mathrm{MLE}}$ is in the interior of Θ since $\nabla^2 \mathcal{L}(\hat{\theta}_{\mathrm{MLE}}) \prec 0$.

Let $r > 0$ be small enough such that $\hat{\theta}_{\mathrm{MLE}}$ is the only stationary point of $\mathcal{L}(\theta)$ in $B_r(\hat{\theta}_{\mathrm{MLE}}) \cap \Theta$. Such $r > 0$ can be chosen since $\hat{\theta}_{\mathrm{MLE}}$ is itself a stationary point of $\mathcal{L}(\theta)$ and $\nabla^2 \mathcal{L}(\hat{\theta}_{\mathrm{MLE}}) \prec 0$ (which, together, they ensure isolated stationarity). In particular, $\hat{\theta}_{\mathrm{MLE}}$ is an isolated maximizer. Then, from Lemma 2, the function $\mathcal{V} : B_r(\hat{\theta}_{\mathrm{MLE}}) \to \mathbb{R}$ given by (11) with $\theta^\star = \hat{\theta}_{\mathrm{MLE}}$ is positive definite with respect to $\hat{\theta}_{\mathrm{MLE}}$, and $-\Delta \mathcal{V}$ is positive semidefinite.

Note that the inequality of the divergence term in (12) is strict for $\theta \in B_r(\hat{\theta}_{\mathrm{MLE}}) \setminus \{\hat{\theta}_{\mathrm{MLE}}\}$. This is because, from Assumption 6, $\mathcal{D}_{\mathrm{KL}}(\theta \| F^{\mathrm{EM}}(\theta)) = 0$ if and only if θ is a fixed point of F^{EM}. But such a point needs to be a stationary point (see Lemma 3), which would lead to the contradiction $\theta = \hat{\theta}_{\mathrm{MLE}}$, since $\hat{\theta}_{\mathrm{MLE}}$ is the only stationary point of $\mathcal{L}(\theta)$ in $B_r(\hat{\theta}_{\mathrm{MLE}}) \setminus \{\hat{\theta}_{\mathrm{MLE}}\}$. Therefore, $\log \mathcal{L}(F^{\mathrm{EM}}(\theta)) > \log \mathcal{L}(\theta)$, since $-\Delta \mathcal{V}(\theta) = \mathcal{L}(F^{\mathrm{EM}}(\hat{\theta}_{\mathrm{MLE}})) - \mathcal{L}(\theta) > 0$ for every $\theta \in B_r(\hat{\theta}_{\mathrm{MLE}}) \setminus \{\hat{\theta}_{\mathrm{MLE}}\}$. Finally, since $\hat{\theta}_{\mathrm{MLE}}$ is a fixed point of F^{EM}, then $\Delta \mathcal{V}(\hat{\theta}_{\mathrm{MLE}}) = 0$, which concludes that $-\Delta \mathcal{V}$ is positive definite. The conclusion follows by invoking Theorem 1, since $\hat{\theta}_{\mathrm{MLE}}$ was just proved to be asymptotically stable. ∎

Theorem 3 (Local Convergence of δ-EM to Local Maxima). *If* $\theta^\star \in \Theta$ *is such that* $\nabla \mathcal{L}(\hat{\theta}^\star) = 0$ *and* $\nabla^2 \mathcal{L}(\theta^\star) \prec 0$, *and* $\delta > 0$ *is small enough, then the sequence* $\{\theta_k\}_{k \in \mathbb{Z}_+}$ *generated by Algorithm 2 converges to* $\theta^\star$ *for every initial approximation* $\theta_0 \in \Theta$ *in a small enough open ball centered around* $\theta^\star$. ◇

Proof. The proof follows similar steps to those in the proof of Theorem 2 with the following adaptations: first, replace F^{EM} by $F^{\delta-\mathrm{EM}}$, and secondly, $\hat{\theta}_{\mathrm{MLE}}$ by $\theta^\star$. ∎

Notice that, the reason why Theorem 3 cannot be readily adapted for the EM algorithm (as opposed to the δ-EM algorithm) is that local maximizers may fail to be fixed points of F^{EM} and therefore, equilibria of $(\mathcal{S})$ with $F = F^{\mathrm{EM}}$. To circumvent this limitation, we will focus on the *limit points* of the EM algorithm. First, recall that $\theta^\star \in \Theta$ is a fixed point of Algorithm 1, if there exists some $\theta_0 \in \Theta$ such the $\theta_k \to \theta^\star$ as $k \to \infty$ for the sequence $\{\theta_k\}_{k \in \mathbb{Z}_+}$ generated by Algorithm 1, which is captured by the following result.

Lemma 4. *If* $\theta^\star \in \Theta$ *is a limit point of Algorithm 1, then it is also a fixed point of* F^{EM}. ◇

Proof. Let $\theta_0 \in \Theta$ be such that $\theta_k \to \theta^\star$ as $k \to \infty$, where $\{\theta_k\}_{k \in \mathbb{Z}_+}$ was generated by Algorithm 1. Then, by the continuity of F^{EM}, it follows that $F^{\mathrm{EM}}(\theta^\star) = \lim_{k \to \infty} F^{\mathrm{EM}}(\theta_k) = \lim_{k \to \infty} \theta_{k+1} = \theta^\star$. ∎

Remark 6. Notice that, while not every local maximizer of the likelihood function is a limit point of the EM algorithm, the same is not true for the δ-EM algorithm, provided that $\delta > 0$ is sufficiently small and the local maximizer is sufficiently regular (*i.e.*, an isolated stationary point of the likelihood function).

Upon Remark 6, and the convergence results established before, we can now establish the following claim.

Theorem 4 (Local Convergence of EM to its Limit Points). *If* $\theta^\star \in \Theta$ *is a limit point of Algorithm 1 such that* $\nabla^2 \mathcal{L}(\theta^\star) \prec 0$, *then the sequence* $\{\theta_k\}_{k \in \mathbb{Z}_+}$ *generated by Algorithm 1 converges to* $\theta^\star$ *for every initial approximation* $\theta_0 \in \Theta$ *in a small enough open ball centered around* $\theta^\star$. ◇

Proof. The proof follows similar steps to those in the proof of Theorem 2, where $\hat{\theta}_{\mathrm{MLE}}$ is replaced by $\theta^\star$, and followed by invoking Lemma 4 instead of Proposition 1. ∎

We conclude this section by exploring how the notion of *exponential stability* can be leveraged to bound the *convergence rate* of the EM algorithm. First, recall that the (linear) rate of convergence for a sequence $\{\theta_k\}_{k \in \mathbb{Z}_+}$ is the number $0 \le \mu \le 1$ given by

$$\mu = \lim_{k \to \infty} \frac{\|\theta_{k+1} - \theta^\star\|}{\|\theta_k - \theta^\star\|}, \qquad (14)$$

provided that the limit exists. Additionally, recall that an equilibrium $\theta^\star \in \Theta$ of $(\mathcal{S})$ is said to be *exponentially stable* if there exist constants $c, \gamma > 0$ such that, for every $\theta_0 \in \Theta$ in a sufficiently small open ball centered around $\theta^\star$, we have $\|\theta[k] - \theta^\star\| \le c \cdot e^{-\gamma k} \|\theta_0 - \theta^\star\|$ for every $k \in \mathbb{Z}_+$.

Remark 7. If $\theta^\star$ is an exponentially stable equilibrium of $(\mathcal{S})$ with $F = F^{\mathrm{EM}}$, then $\{\theta_k\}_{k \in \mathbb{Z}}$ generated by Algorithm 1

converges to $\theta^\star$ with linear convergence rate

$$\mu = \lim_{k \to \infty} \frac{\|\theta_{k+1} - \theta^\star\|}{\|\theta_k - \theta^\star\|} \tag{15a}$$

$$\leq \lim_{k \to \infty} \frac{c \cdot e^{-\gamma \times 0}\|\theta_k - \theta^\star\|}{\|\theta_k - \theta^\star\|} \tag{15b}$$

$$= c, \tag{15c}$$

for every initial approximation $\theta_0 \in \Theta$ in a sufficiently small open ball centered around $\theta^\star$. ∘

The following theorem (adapted from Theorem 5.7 in [14]) allows us to ensure exponential stability, and subsequently to bound the linear convergence rate, provided our Lyapunov function satisfies some additional (mild) regularity conditions.

Theorem 5 (Exponential Stability). *Let $\theta^\star \in \Theta$ be an equilibrium of $(\mathcal{S})$ in the interior of Θ, with $B_\delta(\theta^\star) \subseteq \Theta$ for some small enough $\delta > 0$, and $F : \Theta \to \Theta$ continuous. Let $\mathcal{V} : B_\delta(\theta^\star) \to \mathbb{R}$ be a continuous and positive definite function (with respect to $\theta^\star$) such that*

$$\mathcal{V}(\theta) \leq a\|\theta - \theta^\star\|^2, \tag{16a}$$

$$-\Delta\mathcal{V}(\theta) \geq b\|\theta - \theta^\star\|^2, \tag{16b}$$

for every $\theta \in \Theta$, for some constants $a, b > 0$. Then, $\theta^\star$ is exponentially stable. More precisely, we have $\|\theta[k] - \theta^\star\| \leq c \cdot e^{-\gamma k}$ for every $k \in \mathbb{Z}$ and $\theta_0 \in B_r(\theta^\star)$ with $r > 0$ small enough, where $c = d/a$ with

$$d = \lim_{\delta \to 0} \max_{\theta \in \bar{B}_r(\theta^\star) \backslash B_\delta(\theta^\star)} \frac{\mathcal{V}(\theta)}{\|\theta - \theta^\star\|}, \tag{17}$$

and $\gamma = \log a - \log(a - b)$. ◇

Equipped with this result, we are now ready to establish the following sufficient conditions.

Proposition 3. *Let $\theta^\star \in \Theta$ be a limit point of Algorithm 1 such that $\nabla^2 \mathcal{L}(\theta^\star) \prec 0$. Suppose that there exist constants $a, b > 0$ such that*

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^\star) \geq -a\|\theta - \theta^\star\|^2, \tag{18a}$$

$$\mathcal{L}(F^{\text{EM}}(\theta)) - \mathcal{L}(\theta) \geq b\|\theta - \theta^\star\|^2, \tag{18b}$$

or

$$\mathcal{L}(\theta)/\mathcal{L}(\theta^\star) \geq \exp\{-a\|\theta - \theta^\star\|^2\}, \tag{19a}$$

$$\mathcal{D}_{\text{KL}}(\theta\|F^{\text{EM}}(\theta)) \geq b\|\theta - \theta^\star\|^2, \tag{19b}$$

for every $\theta \in B_\delta(\theta^\star)$ in some small enough $\delta > 0$. Then, $\theta^\star$ is exponentially stable. ◇

Proof. From Lemma 4, it follows that $\theta^\star$ is an equilibrium of $(\mathcal{S})$ with $F = F^{\text{EM}}$. The result follows from invoking Theorem 5. First, to see that condition (16a) is verified, we notice that this is equivalent to (18a) for $\mathcal{V}(\theta) = \mathcal{L}(\theta^\star) - \mathcal{L}(\theta)$ (which is continuous and positive definite with respect to $\theta^\star$ in $B_\delta(\theta^\star)$). On the other hand, condition (16b) readily follows from (18b). Similarly, (16) follows from (19) for $\mathcal{V}(\theta) = \log \mathcal{L}(\theta^\star) - \log \mathcal{L}(\theta)$ (which is also continuous and positive definite with respect to $\theta^\star$ in $B_\delta(\theta^\star)$). ∎

Remark 8. (18a) holds for $= -\frac{1}{2} \min_{\theta \in B_\delta(\theta^\star)} \lambda_{\min}[\nabla^2 \mathcal{L}(\theta)]$. ∘

Lastly, as consequence of Theorems 4 and 5, we have the following result.

Theorem 6 (Explicit Bound for Convergence Rate of EM). *Under the conditions of Proposition 3, the linear convergence rate of Algorithm 1 can be bounded as $\mu \leq d/a$, where (17) defined through $\mathcal{V}(\theta) = \mathcal{L}(\theta^\star) - \mathcal{L}(\theta)$ for the case (18), and $\mathcal{V}(\theta) = \log \mathcal{L}(\theta^\star) - \log \mathcal{L}(\theta)$ for the case (19).* ◇

V. CONCLUSION AND FUTURE WORK

We proposed a dynamical systems perspective of the Expectation-Maximization (EM) algorithm, by analyzing the evolution of its estimates as a nonlinear state-space dynamical system. In particular, we drew parallels between limit points and equilibria, convergence and asymptotic stability, and we leveraged results on discrete-time Lyapunov stability theory to establish local convergence results for the EM algorithm. Finally, we derived conditions that allow us to construct explicit bounds on the linear convergence rate of the EM algorithm by establishing exponential stability in the dynamical system that represents it. Future work will be dedicated to leveraging tools from integral quadratic constraints (IQCs) to study the rate of convergence of EM-like algorithms by including artificial control inputs optimally designed through feedback.

REFERENCES

[1] C. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[4] R. D. Nowak, "Distributed em algorithms for density estimation and clustering in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2245–2253, Aug 2003.

[5] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

[6] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 57–95, 2016.

[7] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization," in *Proceedings of the 2011 IEEE Conference on Decision and Control and European Control Conference*, Dec 2011, pp. 3800–3805.

[8] Z. E. Nelson and E. Mallada, "An integral quadratic constraint framework for real-time steady-state optimization of linear time-invariant systems," in *Proceedings of the 2018 American Control Conference*, June 2018, pp. 597–603.

[9] W. I. Zangwill, *Nonlinear programming: a unified approach*. Prentice-Hall Englewood Cliffs, NJ, 1969, vol. 196, no. 9.

[10] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, pp. 95–103, 1983.

[11] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM review*, vol. 26, no. 2, pp. 195–239, 1984.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.

[13] Y. Matsuyama, "The -em algorithm: Surrogate likelihood maximization using -logarithmic information measures," *Information Theory, IEEE Transactions on*, vol. 49, pp. 692 – 706, 04 2003.

[14] N. Bof, R. Carli, and L. Schenato, "Lyapunov theory for discrete time systems," *arXiv:1809.05289*, 2018.