Fifty-third Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 2, 2015

Statistical and Computational Guarantees for the Baum-Welch Algorithm

Fanny Yang*, Sivaraman Balakrishnan† and Martin J. Wainwright†,*
Department of Statistics†, and
Department of Electrical Engineering and Computer Sciences*
UC Berkeley, Berkeley, CA 94720

Abstract— **The Hidden Markov Model (HMM) is one of the main-stays of statistical modeling of discrete time series and is widely used in many applications. Estimating an HMM from its observation process is often addressed via the Baum-Welch algorithm, which performs well empirically when initialized reasonably close to the truth. This behavior could not be explained by existing theory which predicts susceptibility to bad local optima. In this paper we aim at closing the gap and provide a framework to characterize a sufficient basin of attraction for any *global* optimum in which Baum-Welch is guaranteed to converge linearly to an "optimally" small ball around the global optimum. The framework is then used to determine the linear rate of convergence and a sufficient initialization region for Baum-Welch applied on a two component isotropic hidden Markov mixture of Gaussians.**

I. Introduction

Hidden Markov models (HMMs) are one of the most widely applied statistical models of the last 50 years, with major success stories in computational biology, signal processing and speech recognition, control theory, and econometrics among other disciplines. An important problem to be solved is estimating the state transition probabilities and the parameterized output densities based on samples of the observable component.

The dominant approach to this parameter estimation problem is via the Baum-Welch algorithm [1], a specialization of the EM algorithm [2] to the maximum likelihood problem associated with the HMM. Despite its use in practice, the Baum-Welch algorithm can get trapped in local optima of the likelihood function, and rigorously characterizing when the algorithm performs well has remained an open question for several decades.

With this context, the main contribution of this paper is a theoretical analysis of the Baum-Welch procedure, in particular providing sufficient conditions and finite-sample error bounds on the difference between its iterates and the true parameter. We first analyze the Baum-Welch algorithm at the population level and provide sufficient conditions which can be used to characterize both the basin of attraction of the population global optimum of the likelihood function, as well as the rate of convergence of the Baum-Welch algorithm.

In the finite-sample case, Baum-Welch only converges to a local maximum of the sample log likelihood, while the same rate of convergence and basin of attraction as in the population analysis apply. We can guarantee that given an initialization in the basin, the final Baum-Welch estimate lies within a ball around the true parameter with a radius in the order of the minimax rate of the problem. This result provides an explanation why EM combined with spectral methods estimators initialization has empirically performed well. To our knowledge, these are the only rigorous guarantees establishing a form of global convergence for this widely used algorithm.

As an application of this theory, we consider the problem of estimating a hidden Markov mixture of two isotropic Gaussians. For this example, we show that when the mixture components are separated by $\mathcal{O}(\log \log d)$, the globally optimal solution of the population has a basin of attraction of size on the order of the mean separation. Note that this is in fact also the largest possible initialization radius one can hope for.

A. Related work

A systematic study of the statistical properties of the maximum likelihood estimator (MLE) was undertaken by Bickel et al. [5], who established that it is consistent and asymptotically normal. On the other hand, the original papers of Baum and co-authors [3], [1] showed that the Baum-Welch algorithm converges to a stationary point of the sample likelihood, which is in the spirit of the classical convergence analysis of the EM algorithm [4], [2]. Existing analysis therefore only guarantees convergence towards the MLE, which is by [5] close to the true value, when the initialization is close enough.

In practice, however, a two-step procedure with an efficient initialization for Baum-Welch often converges close to a very good estimate of the true parameter

978-1-5090-1824-6/15/$31.00 ©2015 IEEE

(see [6], [7], [8], [9]). The goal of this paper is thus to obtain a result that can characterize the necessary basin of attraction, in which linear convergence to a minimax-optimal ball around the true parameter is guaranteed. Note that as opposed to classical analysis, we do not make a statement about convergence to the MLE specifically.

Our analysis builds upon a framework for studying the EM algorithm on i.i.d. samples, previously introduced by Balakrishnan et al. [10]. In the non-i.i.d. setting, arguments passing from the population-based to sample-based updates are significantly more delicate. From a technical standpoint, various gradient smoothness conditions are much more difficult to establish since the gradient of the likelihood no longer decomposes over the samples as in the i.i.d. setting. Finally, in order to establish the finite-sample behavior of the Baum-Welch algorithm, we can no longer appeal to standard i.i.d. concentration and empirical process techniques; instead, we need to make use of more sophisticated techniques for dependent data, including the independent block technique [11], [12]. Full proofs of the results can be found in the long version of the paper on arxiv.

II. Background and problem set-up

A. Standard HMM notation and assumptions

We begin by defining the notion of a hidden Markov model. Letting $\mathbb{Z}$ denote the integers, suppose that the observed random variables $\{X_i\}_{i\in\mathbb{Z}}$ take values in $\mathbb{R}^d$, and the latent random variables $\{Z_i\}_{i\in\mathbb{Z}}$ take values in the discrete space $[s] := \{1,\ldots,s\}$. The Markov structure is imposed on the sequence of latent variables. In particular, if the variable Z_1 has some initial distribution π_1, then the joint probability of a particular sequence $(z_1,\ldots,z_n)$ is given by

$$p(z_1,\ldots,z_n;\beta) = \pi_1(z_1;\beta)\prod_{i=1}^{n} p(z_i \mid z_{i-1};\beta),$$

where the vector β denotes a particular parameterization of the initial distribution and Markov chain transition probabilities. We restrict our attention to the homogeneous case, meaning that the transition probabilities for step $(s-1) \mapsto s$ are independent of the index s. Consequently, if we define the transition matrix $A \in \mathbb{R}^{s\times s}$ with entries

$$A(j,k;\beta) := p(z_2 = k \mid z_1 = j;\beta),$$

then the marginal distribution π_i of Z_i can be described by the matrix vector equation

$$\pi_i^T = \pi_1^T A^{i-1},$$

where π_i and π_1 should be understood as probability vectors in the s-dimensional simplex.

Throughout, we assume that the Markov chain is aperiodic and recurrent, which ensures that it has a unique stationary distribution $\bar{\pi}$ defined by the eigenvector equation $\bar{\pi}^T = \bar{\pi}^T A$. Note that $\bar{\pi}$ is in fact a function $\bar{\pi}(\cdot \mid \beta)$ which depends on β as does A. However we will omit the dependence whenever it produces unnecessary notational clutter. We consider Markov chains which are already in their stationary state so that $\pi_1 = \bar{\pi}$. In addition, we also assume that the chain is reversible, meaning that

$$\bar{\pi}(j)A(j,k) = \bar{\pi}(k)A(k,j) \tag{1}$$

for all pairs $j,k \in [s]$.

Another key quantity in our analysis is the mixing rate of the Markov chain. We assume throughout that the stationary distribution $\bar{\pi}$ is strictly positive and introduce the *mixing constant* $\epsilon_{\mathrm{mix}} \in (0,1]$ which satisfies

$$\epsilon_{\mathrm{mix}} \leq \frac{p(z_i|z_{i-1},\beta)}{\bar{\pi}(z_i)} \leq \epsilon_{\mathrm{mix}}^{-1} \tag{2}$$

for all $(z_i, z_{i-1}) \in [s]\times[s]$. This condition implies that the dependence on the initial distribution decays geometrically—more specifically

$$\sup_{\pi_1} \left\|\pi_1^T A^t - \pi_1^T\right\|_{\mathrm{TV}} \leq c_0 \rho_{\mathrm{mix}}^t \tag{3}$$

for all $t = 1,2,\ldots$, where $\rho_{\mathrm{mix}} = 1-\epsilon_{\mathrm{mix}}$ denotes the *mixing rate* of the process, and c_0 is a universal constant. Note that as $\epsilon_{\mathrm{mix}} \to 1^-$, the Markov chain has behavior approaching that of an i.i.d. sequence, whereas as $\epsilon_{\mathrm{mix}} \to 0^+$, its behavior becomes increasingly "sticky".

Associated with each latent variable Z_i is an observation $X_i \in \mathbb{R}^d$. We use $p(x_i|z_i,\mu)$ to denote the density of X_i given that $Z_i = z_i$, an object that is parameterized by a vector μ. Introducing the shorthand $x_1^n = (x_1,\ldots,x_n)$ and $z_0^n = (z_0,\ldots,z_n)$, the joint probability of the sequence (x_1^n, z_0^n) (also known as the complete likelihood) can be written in the form

$$p(z_0^n, x_1^n;\theta) = \pi_0(z_0;\beta)\prod_{i=1}^{n} p(z_i \mid z_{i-1},\beta)\prod_{i=1}^{n} p(x_i|z_i,\mu) \tag{4}$$

where the pair $\theta := (\beta,\mu)$ parameterizes the transition and observation functions. The unobserved state at time step $t=0$ is introduced for notational convenience later on. The likelihood then reads

$$p(x_1^n;\theta) = \sum_{z_0^n} p(z_0^n, x_1^n;\theta).$$

A simple example: A special case helps to illustrate these definitions. In particular, suppose that we have a Markov chain with $s = 2$ states. Consider a matrix of transition probabilities $A \in \mathbb{R}^{2\times 2}$ of the form

$$A = \frac{1}{e^{\beta} + e^{-\beta}} \begin{bmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{bmatrix} = \begin{bmatrix} \zeta & 1-\zeta \\ 1-\zeta & \zeta \end{bmatrix}, \tag{5}$$

where $\zeta := \frac{e^{\beta}}{e^{\beta}+e^{-\beta}}$. By construction, this Markov chain is recurrent and aperiodic with the unique stationary distribution $\overline{\pi} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}^T$. Moreover, by calculating the eigenvalues of the transition matrix, we find that the mixing condition (3) holds with $\rho_{\mathrm{mix}} := |2\zeta - 1| = |\tanh(\beta)|$.

Suppose moreover that the observed variables are conditionally Gaussian, say with

$$p(x_t|z_t, \mu) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\|x-\mu\|_2^2} & \text{if } z_t = 1 \\ \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\|x+\mu\|_2^2} & \text{if } z_t = 2 \end{cases} \tag{6}$$

With this choice, the marginal distribution of each X_t is a two-state Gaussian mixture with mean vectors μ and $-\mu$, and covariance matrices $\sigma^2 I_d$. We provide specific consequences of our general theory for this special case in the Section IV.

B. Baum-Welch updates for HMMs

We now describe the Baum-Welch updates for a general discrete-state hidden Markov model. As a special case of the EM algorithm, the Baum-Welch algorithm is known to ascend the likelihood function. It does so indirectly, by first computing a lower bound on the likelihood (E-step) and then maximizing this lower bound (M-step).

For a given integer $n \geq 1$, suppose that we observe a sequence $x_1^n = (x_1, \ldots, x_n)$ drawn from the marginal distribution over X_1^n defined by the model (4). The rescaled log likelihood of the sample path x_1^n is then given by

$$\ell_n(\theta) := \frac{1}{n} \log \Big(\sum_{z_0^n} p(z_0^n, x_1^n; \theta) \Big).$$

The EM likelihood is based on lower bounding the likelihood via Jensen's inequality. Using $\mathbb{E}_{Z_1^n|x_1^n,\theta'}$ to denote the expectation under the conditional distribution $p(Z_1^n|x_1^n;\theta')$, the concavity of the logarithm and Jensen's inequality imply that for any choice of θ', we have

$$\ell_n(\theta) \geq \underbrace{\frac{1}{n}\mathbb{E}_{Z_0^n|x_1^n,\theta'}\big[\log p(x_1^n, Z_0^n, \theta)\big]}_{Q_n(\theta\,|\,\theta')} + H_n(\theta')$$

where H_n is a function only of θ'. For a given choice of θ', the E-step corresponds to the computation of the function $\theta \mapsto Q_n(\theta \mid \theta')$. The M-step is defined by the arg max operator $M_n(\theta') = \arg\max_{\theta\in\Omega} Q_n(\theta \mid \theta')$, where Ω is the set of feasible parameter vectors. Overall, given an initial vector $\theta^0 = (\beta^0, \mu^0)$, the EM algorithm generates a sequence $\{\theta^t\}_{t=0}^{\infty}$ according to the recursion $\theta^{t+1} = M_n(\theta^t)$.

This description can be made more concrete for an HMM, in which case the Q-function takes the form

$$\begin{aligned} Q_n(\theta \mid \theta') = \frac{1}{n}\mathbb{E}_{Z_0|x_1^n,\theta'}\big[\log \pi_0(Z_0 \mid \beta)\big] \\ + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_{i-1}^i|x_1^n,\theta'}\big[\log p(Z_i \mid Z_{i-1}, \beta)\big] \\ + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_i|x_1^n,\theta'}\big[\log p(x_i \mid Z_i, \mu)\big], \end{aligned} \tag{7}$$

where the dependence of π_0 on β comes from the assumption that $\pi_0 = \overline{\pi}$. Note that the Q-function can be decomposed as a sum of a term which is solely dependent on μ, and another one which only depends on β—that is

$$Q_n(\theta \mid \theta') = Q_{1,n}(\mu|\theta') + Q_{2,n}(\beta|\theta') \tag{8}$$

where $Q_{1,n}(\mu|\theta') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_i|x_1^n,\theta'}\big[\log p(x_i|Z_i,\mu)\big]$. In order to compute the expectations defining this function, we need to compute marginal distributions over the singletons Z_i and pairs (Z_i, Z_{i+1}) under the joint distribution $p(Z_0^n \mid x_1^n, \theta')$. These marginals can be computed efficiently using a recursive message-passing algorithm, known either as the forward-backward or sum-product algorithm [13].

In the M-step, decomposition (8) suggests that the maximization over the two components (β, μ) can also be decoupled. Accordingly, with a slight abuse of notation, we often write $M_n^{\beta}(\theta') = \arg\max_{\beta\in\Omega_\beta} Q_{2,n}(\beta|\theta')$ and $M_n^{\mu}(\theta') = \arg\max_{\mu\in\Omega_\mu} Q_{1,n}(\mu|\theta')$ for these two decoupled maximization steps, where $\Omega := \Omega_\beta \times \Omega_\mu$ is the feasible set of parameters.

III. Main results

We now turn to a statement of our main results, along with a discussion of some of their consequences. We begin by establishing the existence of an appropriate population analog of the Q-function. Although the existence of such an object is obvious in the case of i.i.d. data, it requires some technical effort to establish existence for the case of dependent data. In our proof, we make use of a k-truncated version of the Q-function based on the mixing property and stationarity of the Markov chain. This object plays a central role in the remainder of our analysis. In particular, we first analyze

a version of the Baum-Welch updates on the expected k-truncated Q-function for an extended sequence of observations x_{1-k}^{n+k}, and provide sufficient conditions for these population-level updates to be contractive. We then use non-asymptotic forms of empirical process theory to show that under suitable conditions, the actual sample-based EM updates—i.e., what is actually implemented in practice—are well-behaved with high probability.

A. Existence of population Q-function

In the analysis of Balakrishnan et al. [10] for EM applied to i.i.d. data, the central object is the notion of a population Q-function—namely, the function that underlies the EM algorithm in the idealized limit of infinite data. When the samples are dependent, the quantity $\mathbb{E}\big[Q_n(\theta \mid \theta')\big]$ is no longer independent of n as opposed to the i.i.d. setting. A reasonable candidate for a general definition of the population Q-function is therefore given by

$$\bar{Q}(\theta \mid \theta') := \lim_{n \to +\infty} [\mathbb{E} Q_n(\theta \mid \theta')]. \tag{9}$$

It is clear that this definition is sensible in the i.i.d. case, but for dependent sampling schemes, it is necessary to prove that the limit given in definition (9) actually exists.

In this paper, we do so by considering a suitably truncated version of the sample-based Q-function. Let us consider a sequence $\{(X_i, Z_i)\}_{i=1-k}^{n+k}$, assumed to be drawn from the stationary distribution of the overall chain. For positive integers $i < j$ and $a < b$, we let $\mathbb{E}_{Z_i^j | x_a^b, \theta}$ denote expectations taken over the distribution $p(Z_i^j \mid x_a^b, \theta)$. Then, for a positive integer k to be chosen, we define

$$\begin{aligned} Q_n^k(\theta|\theta') = \frac{1}{n}\Big[& \mathbb{E}_{Z_0 | x_{-k}^k, \theta'} \log p(Z_1; \beta) \\ & + \sum_{i=1}^{n} \mathbb{E}_{Z_{i-1}^i | x_{i-k}^{i+k}, \theta'} \log p(Z_i \mid Z_{i-1}, \beta) \\ & + \sum_{i=1}^{n} \mathbb{E}_{Z_i | x_{i-k}^{i+k}, \theta'} \log p(x_i \mid Z_i, \mu) \Big]. \end{aligned}$$

Similarly as in equation (7), we can then decompose Q_n^k into $Q_n^k(\theta|\theta') = Q_{1,n}^k(\mu|\theta') + Q_{2,n}^k(\beta|\theta')$, yielding the corresponding arg max operators $M_n^k(\theta'), M_{\tilde{n}}^{\mu,k}(\theta'), M_{\tilde{n}}^{\beta,k}(\theta')$. Note that, as opposed to the function Q_n from equation (7), the definition of Q_n^k involves variables Z_i, Z_{i-1} that are not conditioned on the full observation sequence x_1^n, but instead only on a $2k$ window centered around the index i. By construction, the k-truncated population function (and the decomposed equivalents) given by

$$\begin{aligned} \bar{Q}^k(\theta|\theta') &:= \lim_{n \to \infty} \mathbb{E} Q_n^k(\theta|\theta') && (10) \\ &= \mathbb{E} Q_{1,n}^k(\mu|\theta') + \lim_{n \to \infty} \mathbb{E} Q_{2,n}^k(\beta|\theta') \\ &=: \bar{Q}_1^k(\mu \mid \theta') + \bar{Q}_2^k(\beta \mid \theta') && (11) \end{aligned}$$

is then well-defined: in particular, due to stationarity of the random sequences $\{p(z_i \mid X_{i-k}^{i+k})\}_{i=1}^n, \{p(z_{i-1}^i \mid X_{i-k}^{i+k})\}_{i=1}^n$, the expectation is independent of the sample size n.

Our first result uses the existence of this truncated population object to establish the existence of the standard population Q-function from equation (9). In doing so, we make use of the sup-norm

$$\|Q_1 - Q_2\|_\infty := \sup_{\theta, \theta' \in \Omega} \Big| Q_1(\theta|\theta') - Q_2(\theta|\theta') \Big|. \tag{12}$$

Furthermore, we require in the following that the observation densities satisfy the following boundedness condition

$$\sup_{\theta \in \mathbb{B}_2(r; \theta^*)} \mathbb{E}\Big[\max_{z_i \in [s]} \big| \log p(X_i | z_i, \mu) \big| \Big] < \infty. \tag{13}$$

Proposition 1 *Under the previously stated assumptions, the population function $\bar{Q}$ defined in equation* (9) *exists.*

The proof of this statement hinges on the following auxiliary claim, which bounds the difference between $\mathbb{E} Q_n$ and the k-truncated Q-function as

$$\begin{aligned} \|\mathbb{E} Q_n - \bar{Q}^k\|_\infty \leq \frac{c\, s^5}{\epsilon_{\mathrm{mix}}^8 \bar{\pi}_{min}^2} \big(1 - \epsilon_{\mathrm{mix}} \bar{\pi}_{min}\big)^k \\ + \frac{1}{n} \log \bar{\pi}_{min}^{-1}, \end{aligned} \tag{14}$$

where $\bar{\pi}_{min} := \min_{\beta \in \Omega_\beta, j \in [s]} \bar{\pi}(j \mid \beta)$ is the minimum probability in the stationary distribution, and ϵ_{mix} is the mixing constant from equation (2) in the range of feasible transition parameters $\beta \in \Omega_\beta$. Since this bound holds for all n, it shows that the population function $\bar{Q}$ can be uniformly approximated by $\bar{Q}^k$, with the approximation error decreasing geometrically as the truncation level k grows. This fact plays an important role in the analysis to follow.

B. Analysis of updates based on $\bar{Q}^k$

Let θ^* be some fixed global maximum of the population likelihood function and assume throughout the paper that θ^* is equal to the true parameter for the sake of identifiability. A key step for establishing linear convergence is to prove contraction of the EM update towards this local maximum. Our strategy to show this

first involves analyzing some properties of the EM iterates at the population level via the truncated function $\overline{Q}^k$ instead of $\overline{Q}$, where k is a given truncation level (to be chosen in the sequel). The bound (14) provides us with uniform control on the difference between $\overline{Q}^k$ and $\overline{Q}$.

Viewing the maximum θ^* as fixed, consider the function $q^k(\theta) := \overline{Q}^k(\theta \mid \theta^*)$. Suppose that there is a radius $r > 0$ such that q^k is λ-*strongly concave* over the Euclidean ball $\mathbb{B}_2(r; \theta^*)$ of radius r centered at θ^*—that is

$$q^k(\theta_1) - q^k(\theta_2) - \langle \nabla q^k(\theta_2), \theta_1 - \theta_2 \rangle \leq -\frac{\lambda}{2} \|\theta_1 - \theta_2\|_2^2 \quad (15)$$

for all $\theta_1, \theta_2 \in \mathbb{B}_2(r; \theta^*)$. In addition, suppose that the following *first-order stability* condition (or FOS(L) for short) holds:

$$\|\nabla_\theta \overline{Q}^k(\theta|\theta') - \nabla_\theta \overline{Q}^k(\theta|\theta^*)\|_2 \leq L\|\theta' - \theta^*\|_2 \quad (16)$$

for all $\theta, \theta' \in \mathbb{B}_2\big(r; \theta^*\big)$. Our analysis to follow shows that these conditions hold for concrete models.

With this setup, we consider an idealized algorithm that, based on some initialization $\widetilde{\theta}^0 \in \mathbb{B}_2(r; \theta^*)$, generates the sequence of iterates $\widetilde{\theta}^{t+1} = \overline{M}^k(\widetilde{\theta}^t)$. Since $\overline{Q}^k$ is an approximate version of $\overline{Q}$, the update operator $\overline{M}^k$ should be understood as an approximation to the idealized population EM operator $\overline{M}$. The following theorem shows that the approximation error is well-controlled under suitable conditions and that the sequence $\{\widetilde{\theta}^t\}_{t-0}^\infty$ converges to a particular neighborhood of θ^*. Recall hereby that $\overline{\pi}(\cdot \mid \beta)$ denotes the stationary distribution of the Markov chain on a discrete space with s elements, a transition parameter β and $\overline{\pi}_{min} = \min_{\beta \in \Omega_\beta} \min_{j \in [s]} \overline{\pi}(j \mid \beta)$.

Theorem 1 *(a)* Approximation guarantee: *Under the mixing condition* (3)*, density boundedness condition* (13)*, and strong concavity condition* (15)*, there is a universal constant c_0 such that for all $\theta \in \Omega$*

$$\|\overline{M}^k(\theta) - \overline{M}(\theta)\|_2^2 \leq \underbrace{c_0 \frac{s^5}{\lambda\, \epsilon_{\mathrm{mix}}^8 \overline{\pi}_{min}^2} \big(1 - \epsilon_{\mathrm{mix}} \overline{\pi}_{min}\big)^k}_{=:\varphi^2(k)} \quad (17)$$

(b) Convergence guarantee: *Suppose in addition that the FOS(L) condition* (16) *holds with parameter $L \in (0, \lambda)$ over the ball $\mathbb{B}_2\big(r; \theta^*\big)$ and that the truncation parameter k is sufficiently large to ensure that $\varphi(k) \leq r\big(1 - \frac{L}{\lambda}\big)$. Defining $\kappa := \frac{L}{\lambda}$, the iterates $\{\widetilde{\theta}^t\}_{t=0}^\infty$ generated by the $\overline{M}^k$ operator satisfy the bound*

$$\|\widetilde{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\widetilde{\theta}^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\varphi(k). \quad (18)$$

Note that the subtlety here is that θ^* is no longer a fixed point of the operator $\overline{M}^k$, due to the error induced by the k^{th}-order truncation. Nonetheless, under a mixing condition, as the bounds (17) and (18) show, this approximation error is controlled, and decays exponentially in k. The proof of the recursive bound (18) is based on showing that

$$\|\overline{M}^k(\theta) - \overline{M}^k(\theta^*)\|_2 \leq \frac{L}{\lambda}\|\theta - \theta^*\|_2 \quad (19)$$

holds for any $\theta \in \mathbb{B}_2(r; \theta^*)$. Inequality (19) is equivalent to stating that the operator $\overline{M}^k$ is contractive, i.e. that applying $\overline{M}^k$ to the pair θ and θ^* always decreases the distance.

C. Sample-based results

Theorem 1 is a population-level result applying to the (truncated) EM operators based on an extended sequence of samples. In this section we leverage the good behavior of the truncated population-based analogs to analyze a sample-based version of the EM algorithm in which each update uses a "fresh" and finite set of samples. More concretely, given a batch of n observations and a fixed iteration number T, suppose that we split the full collection of samples into a set of T subsets, each[1] of size $\tilde{n} := n/T$. The *sample-splitting EM updates* are given by

$$\widehat{\theta}^{t+1} := M_{\tilde{n}}(\widehat{\theta}^t) \qquad \text{for } t = 0, 1, \ldots, T-1,$$

where we recall that $M_n(\theta') = \arg\max_{\theta \in \Omega_\theta} Q_n^k(\theta|\theta')$ and where the update at time t uses the t^{th} subset of samples which we denote by S^t. For this reason, strictly speaking, the operator $M_{\tilde{n}}$ should be further indexed by t, but we drop this dependence to simplify notation.

For a given sample size $n \geq 1$ and number of iterations T, we adopt the shorthand $\tilde{n} = n/T$ for the number of samples used in each iteration. For a tolerance parameter $\delta \in (0, 1)$, we let $\varphi_{\tilde{n}}(\delta, k)$ be the smallest positive scalar such that

$$\sup_{\theta \in \mathbb{B}_2(r; \theta^*)} \mathbb{P}\Big[\|M_{\tilde{n}}(\theta) - M_{\tilde{n}}^k(\theta)\|_2 \geq \varphi_{\tilde{n}}(\delta, k)\Big] \leq \delta. \quad (20a)$$

This quantity bounds the approximation error for a fixed number of samples, and thus is conceptually related to the population-level quantity $\varphi(k)$ appearing

[1]To simplify notation, we assume that n is divisible by T.

in Theorem 1(a). Similarly, for each $\delta \in (0,1)$, we let $\epsilon_{\tilde{n}}(\delta; k)$ denote the smallest scalar such that

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{P}\Big[\|M^k_{\tilde{n}}(\theta) - \overline{M}^k(\theta)\|_2 \geq \epsilon_{\tilde{n}}(\delta, k)\Big] \leq \delta. \tag{20b}$$

For a given truncation level k, it gives an upper bound on the difference between the population and sample-based M-operators.

Theorem 2 (Sample splitting) *Suppose that the truncated population EM operator $\overline{M}^k$ satisfies the local contraction bound* (19) *with parameter $\kappa \in (0,1)$ on the ball $\mathbb{B}_2(r;\theta^*)$. For a given sample size n and iteration number T, suppose that (k,n) are sufficiently large to ensure that*

$$\varphi_{\tilde{n}}(\delta/T, k) + \epsilon_{\tilde{n}}\big(\delta/T, k\big) \leq (1-\kappa)\, r. \tag{21}$$

Then given any initialization $\widehat{\theta}^0 \in \mathbb{B}_2(r;\theta^*)$, with probability at least $1-2\delta$, the sample-splitting EM sequence $\{\widehat{\theta}^t\}_{t=0}^{\infty}$ satisfies the bound

$$\|\widehat{\theta}^t - \theta^*\|_2 \leq \underbrace{\kappa^t \|\widehat{\theta}^0 - \theta^*\|_2}_{\text{Geometric decay}} + \underbrace{\frac{1}{1-\kappa}\Big\{\varphi_{\tilde{n}}\big(\delta/T, k\big) + \epsilon_{\tilde{n}}\big(\delta/T, k\big)\Big\}}_{\text{Residual error}} \tag{22}$$

for all $t = 1, \ldots, T-1$.

The bound (22) shows that the distance between $\widehat{\theta}^t$ and θ^* is bounded by two terms: the first decays geometrically as t increases, and the second term corresponds to a residual error term that remains independent of t. Thus, by choosing the iteration number T larger than $\frac{\log(2r/\epsilon)}{\log \kappa}$, we can ensure that the first term is at most ϵ. The residual error term can be controlled by requiring that the sample size n is sufficiently large, and then choosing the truncation level k appropriately. We provide a concrete illustration of this procedure in the following section, where we analyze the case of Gaussian output HMMs.

IV. Guarantees for Gaussian output HMMs

Let us now return to the concrete example of a Gaussian output HMM, as first introduced in Section II-A. Recall that our Gaussian output HMM is based on $s = 2$ hidden states, using the transition matrix from equation (5), and the Gaussian output densities from equation (6). For convenience of analysis, we let the hidden variables z_i take values in $\{-1, 1\}$. In addition, we require that the mixing coefficient $\rho_{\mathrm{mix}} = 1 - \epsilon_{\mathrm{mix}}$ is bounded away from 1 in order to ensure that the mixing condition (2) is fulfilled. We denote the upper bound for ρ_{mix} as $b < 1$ so that $\rho_{\mathrm{mix}} \leq b$ and $\epsilon_{\mathrm{mix}} \geq 1 - b$. The feasible set of the probability parameter ζ and its log odds analog $\beta = \frac{1}{2}\log\big(\frac{\zeta}{1-\zeta}\big)$ are then given by

$$\Omega_\zeta = \left\{\zeta \in \mathbb{R} \mid \frac{1-b}{2} \leq \zeta \leq \frac{1+b}{2}\right\}, \quad \text{and}$$

$$\Omega_\beta = \Big\{\beta \in \mathbb{R} \;\mid\; |\beta| < \underbrace{\frac{1}{2}\log\big(\frac{1+b}{1-b}\big)}_{\beta_B}\Big\}.$$

A. Explicit form of Baum-Welch updates

We begin by deriving an explicit form of the Baum-Welch updates for this model. Given n observations and T iterations of the algorithm, the sample-splitting EM update at iteration t operates on a set of subsamples S^t of size n/T, which we denote by $x(S^t)$. Using this notation, the Baum-Welch updates take the form

$$\widehat{\mu}^{t+1} = \frac{1}{|S^t|}\sum_{i \in S^t}(2p(Z_i = 1 \mid x(S^t), \theta^t) - 1)x_i, \tag{23}$$

$$\widehat{\zeta}^{t+1} = \Pi_{\Omega_\zeta}\left(\frac{\sum_{i \in S^t}\sum_{Z_i} p(Z_i = Z_{i+1} \mid x(S^t), \theta^t)}{|S^t|}\right),$$

$$\widehat{\beta}^{t+1} = \frac{1}{2}\log\big(\frac{\widehat{\zeta}^{t+1}}{1-\widehat{\zeta}^{t+1}}\big),$$

where Π_{Ω_ζ} denotes the Euclidean projection onto the set Ω_ζ. Note that the maximization steps are carried out on the decomposed Q-functions $Q_{1,n}(\cdot|\theta^t), Q_{2,n}(\cdot|\theta^t)$. In addition, since we are dealing with a one-dimensional quantity β, the projection of the unconstrained maximizer onto the interval Ω_ζ is equivalent to the constrained maximizer over the feasible set Ω_ζ. This step is in general not valid for higher dimensional transition parameters.

B. Population and sample guarantees

We begin by using the results from Section III to show that the population and sample-based version of the Baum-Welch updates are linearly convergent. In establishing the population-level guarantee, the key condition which needs to be fulfilled is the $FOS(L)$ condition (16) with explicit dependence of the Lipschitz constant L on model parameters such as the separation of the mixtures.

Throughout this section, we use the notation c_0, c_1 to denote universal constants, with no dependence on any problem parameters. In order to ease notation, our explicit tracking of parameter dependence is limited to the standard deviation σ and Euclidean norm $\|\mu^*\|_2$, which together determine the signal-to-noise ratio $\eta^2 := \frac{\|\mu^*\|_2^2}{\sigma^2}$ of the mixture model. We use the notation C_0, C_1 to denote quantities that do not depend on $(\|\mu^*\|_2, \sigma)$, but

may depend on other parameters such as $\bar{\pi}_{min}$, ρ_{mix}, b, and so on.

We begin by stating a result that applies to the sequence $\{\widetilde{\theta}^t\}_{t=0}^{\infty}$ obtained by repeatedly applying the k-truncated population-level Baum-Welch update operator $\overline{M}^k$. Our first corollary establishes that this sequence is linearly convergent, with a convergence rate that is given by

$$\kappa(\eta) := C_1\eta^2(\eta + \|\mu^*\|_2)\, \mathrm{e}^{-c_2\eta^2}\, \log d. \tag{24}$$

Corollary 1 (Population Baum-Welch) *Consider a two-state Gaussian output HMM that is mixing* (2)*, and suppose that the SNR lower bounded as $\eta^2 \geq C \log\log d$ for a sufficiently large constant C. Given the radius $r = \frac{\|\mu^*\|_2}{4}$, suppose that the truncation parameter k is sufficiently large to ensure that $\varphi(k) \leq (1 - \frac{\kappa}{1-b^2})r$. Then for any initialization $\widetilde{\theta}^0 = (\widetilde{\mu}^0, \widetilde{\beta}^0) \in \mathbb{B}_2\big(r; \mu^*\big) \times \Omega_\beta$, then the sequence $\{\widetilde{\theta}^t\}_{t=0}^{\infty}$ generated by $\overline{M}^k$ satisfies the bound*

$$\max\Big\{\|\widetilde{\mu}^t - \mu^*\|_2, |\widetilde{\beta}^t - \beta^*|\Big\} \leq \kappa^t \|\widetilde{\theta}^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\varphi(k) \tag{25}$$

for all iterations $t = 1, 2, \ldots$

Definition (24) yields that as long as the signal-to-noise ratio η is larger than a constant multiple of $\log\log d$, we can ensure that $\kappa(\eta) \leq 1 - b^2 < 1$. The bound (25) then guarantees a type of contraction and also that the pre-condition $\varphi(k) \leq (1 - \frac{\kappa}{1-b^2})r$ can be satisfied by choosing the truncation parameter k large enough. If we use a finite truncation parameter k, then contraction occurs up to the error floor given by $\varphi(k)$, which reflects the bias introduced by truncating the likelihood to a window of size k. At the population level (in which the effective sample size is infinite), we could take the limit $k \to \infty$ so as to eliminate this bias. However, this is no longer possible in the finite sample setting, in which we must necessarily have $k \ll n$.

Corollary 2 (Sample-splitting Baum-Welch iterates) *For a given tolerance $\delta \in (0,1)$ and iteration number T, suppose that the sample size is lower bounded as $n \geq C_1(\sigma^2 + \|\mu^*\|_2^2)Td\log^2(\frac{d}{\delta})$ for a sufficiently large constant C_1. Then under the conditions of Corollary 1, with probability at least $1-\delta$, we have*

$$\max\Big\{\|\widehat{\mu}^T - \mu^*\|_2, |\widehat{\beta}^T - \beta^*|\Big\} \leq \widetilde{\kappa}^T\|\widehat{\theta}^0 - \theta^*\|_2 + C\frac{\sigma\sqrt{\frac{Td\log^2(n/\delta)}{n}} + (\|\mu^*\|_2 + 1)\sqrt{\frac{T\log^2(n/\delta)}{n}}}{1-\widetilde{\kappa}}, \tag{26}$$

where $\widetilde{\kappa} = \sqrt{2}\kappa$.

As a consequence of the bound (26), if we are given a sample size $n \gtrsim d\log^2 d$, then taking $T \approx \log n$ iterations is guaranteed to return an estimate $(\widehat{\mu}^T, \widehat{\beta}^T)$ with error of the order $\sqrt{\frac{d\log^3(n)}{n}}$.

C. Simulations

In this section, we discuss how simulations confirm our theoretical predictions for the two-state Gaussian output HMMs described in Section IV. The EM updates for the mean vector $\widehat{\mu}^{t+1}$ and transition probability $\hat{\zeta}^{t+1}$ are performed according to equation (23), where ζ is chosen instead of β for simplicity of computation. In the sequel, we denote the final parameter estimates by $\widehat{\mu}$ and $\widehat{\zeta}$ whereas the true parameters are μ^* and ζ^*.

In all simulations, the mixing parameter ρ_{mix} is fixed to be 0.6, since a discussion of the dependence on the parameters in this limited case is not expected to transfer to the general asymmetric case. The initializations $\widehat{\mu}^0$ are generated randomly in a ball of radius $r := \frac{\|\mu^*\|_2}{4}$ around the true parameter μ^* whereas $\widehat{\zeta}^0 = \frac{1}{2}$. Finally, the estimation error of the mean vector μ is computed as $\log_{10}\|\hat{\mu} - \mu^*\|_2$. Since the transition parameter estimation errors behave similarly to the observation parameter in simulations, we omit the corresponding figures here.

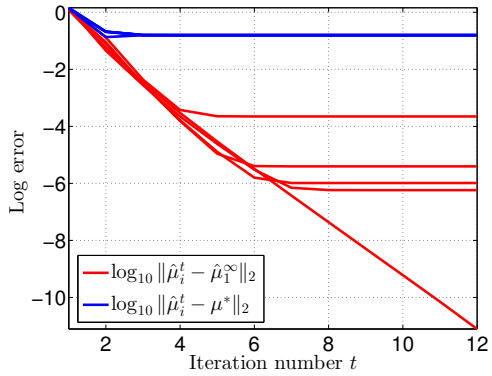

Fig. 1: Plot of the convergence of the optimization error $\log\|\widehat{\mu}_i^t - \widehat{\mu}_1^\infty\|_2$ and statistical error $\log\|\widehat{\mu}_i^t - \mu^*\|_2$ for 5 different initializations. The parameter settings were $d = 10$, $n = 1000$, $\rho_{\mathrm{mix}} = 0.6$ and SNR $\frac{\|\mu^*\|_2}{\sigma} = 1.5$.

Figure 1 depicts the convergence behavior of both the optimization and the statistical error. Here we run the Baum-Welch algorithm for a fixed sample sequence X_1^n drawn from a model with SNR $\eta^2 = 1.5$ and $\zeta = 0.2$, using different random initializations in the ball around μ^* with radius $\frac{\|\mu^*\|_2}{4}$. We denote the final estimate of the first trial by $\widehat{\mu}_1^\infty$. The curves in red depict the

differences between the Baum-Welch iterates $\widehat{\mu}_i^t$ using the i-th initialization, and $\widehat{\mu}_1^\infty$, whereas the blue lines represent the statistical error, i.e. the distance of the iterates from the true parameter μ^*.

For both family of curves we observe linear convergence in the first few iterations until an error floor is reached. This aligns with the upper bound (26) in Corollary 2. In addition, the red curves show that for different initializations, the Baum-Welch algorithm converges to different stationary points $\widehat{\mu}_i$ which however all have the same distance from μ^*.

The dependence of the convergence rate on the SNR η^2 is confirmed by simulation results shown in Figure 2. Lines of the same color represent different random draws of parameters given a fix SNR. Clearly, the convergence is linear for high SNR, and the rate decreases with decreasing SNR. It is also noticeable that for low SNR, the convergence ceases to be linear and even seems to elicit sublinear behavior. This observation suggests a phase transition at a particular value of the SNR and aligns with our theoretical predictions in inequality (26) only hold for sufficiently high SNR $\eta^2 \geq c \log \log d$.

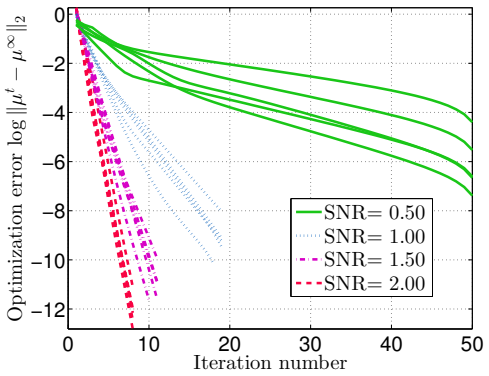

Fig. 2: Plot of convergence behavior for different SNR, where for each curve, different parameters were chosen. The parameter settings are $d = 10$, $n = 1000$ and $\rho_{\text{mix}} = 0.6$.

V. Discussion

In this paper, we developed tools to provide general global convergence guarantees for the Baum-Welch algorithm as well as specific results for a hidden Markov mixture of two isotropic Gaussians. When studying the Baum-Welch algorithm in cases when the underlying i.i.d. mixture is identifiable, we ideally hope to directly leverage results for the EM algorithm in the i.i.d. case together with appropriate mixing conditions on the underlying Markov chain in order to establish rigorous convergence guarantees. Our results confirm that (essentially) such a result is true for the mixture of two isotropic Gaussians and we conjecture that a fully general version of this result is possible. This conjecture if true would give a straightforward recipe for proving rigorous global convergence guarantees for the Baum-Welch algorithm, via analogous results in the i.i.d. setting.

References

[1] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The Annals of Mathematical Statistics*, pp. 164–171, 1970.

[2] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[3] L. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The Annals of Mathematical Statistics*, pp. 1554–1563, 1966.

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

[5] P. Bickel, Y. Ritov, and T. Rydén, "Asymptotic normality of the maximum-likelihood estimator for general Hidden Markov Models," *Ann. Statist.*, vol. 26, no. 4, pp. 1614–1635, 08 1998.

[6] E. Mossel and S. Roch, "Learning nonsingular phylogenies and Hidden Markov Models," *Ann. Appl. Probab.*, vol. 16, no. 2, pp. 583–614, 05 2006.

[7] D. Hsu, S. Kakade, and T. Zhang, "A spectral algorithm for learning Hidden Markov Models," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012.

[8] L. A. Kontorovich, B. Nadler, and R. Weiss, "On learning parametric-output HMMs," in *Proc. 30th International Conference Machine Learning*, June 2013, pp. 702–710.

[9] A. Chaganty and P. Liang, "Spectral experts for estimating mixtures of linear regressions," *arXiv preprint arXiv:1306.3729*, 2013.

[10] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *arXiv preprint arXiv:1408.2156*, 2014.

[11] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *The Annals of Probability*, pp. 94–116, 1994.

[12] A. Nobel and A. Dembo, "A note on uniform laws of averages for dependent processes," *Statistics & Probability Letters*, vol. 17, no. 3, pp. 169–172, 1993.

[13] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1—305, December 2008.