Statistical and Computational Guarantees for the Baum-Welch Algorithm

Fanny Yang FANNY-YANG@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720-1776, USA

Sivaraman Balakrishnan SBALAKRI@BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Martin J. Wainwright WAINWRIG@BERKELEY.EDU
Department of Statistics
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720-1776, USA

Editor:

Abstract

The Hidden Markov Model (HMM) is one of the mainstays of statistical modeling of discrete time series, with applications including speech recognition, computational biology, computer vision and econometrics. Estimating an HMM from its observation process is often addressed via the Baum-Welch algorithm, which is known to be susceptible to local optima. In this paper, we first give a general characterization of the basin of attraction associated with any global optimum of the population likelihood. By exploiting this characterization, we provide non-asymptotic finite sample guarantees on the Baum-Welch updates, guaranteeing geometric convergence to a small ball of radius on the order of the minimax rate around a global optimum. As a concrete example, we prove a linear rate of convergence for a hidden Markov mixture of two isotropic Gaussians given a suitable mean separation and an initialization within a ball of large radius around (one of) the true parameters. To our knowledge, these are the first rigorous local convergence guarantees to global optima for the Baum-Welch algorithm in a setting where the likelihood function is nonconvex. We complement our theoretical results with thorough numerical simulations studying the convergence of the Baum-Welch algorithm and illustrating the accuracy of our predictions.

Keywords: Hidden Markov Models, Baum-Welch algorithm, EM algorithm, non-convex optimization, graphical models

1. 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 (Durbin, 1998), signal processing and speech recognition (Rabiner and Juang, 1993), control theory (Elliott et al., 1995), and econometrics (Kim and Nelson, 1999) among other disciplines. At a high level, a hidden Markov model is a Markov process split into an observable component and an unobserved or latent component. From a statistical standpoint, the use of latent states makes the HMM generic enough to model a variety of complex real-world time series, while the Markovian structure enables relatively simple computational procedures.

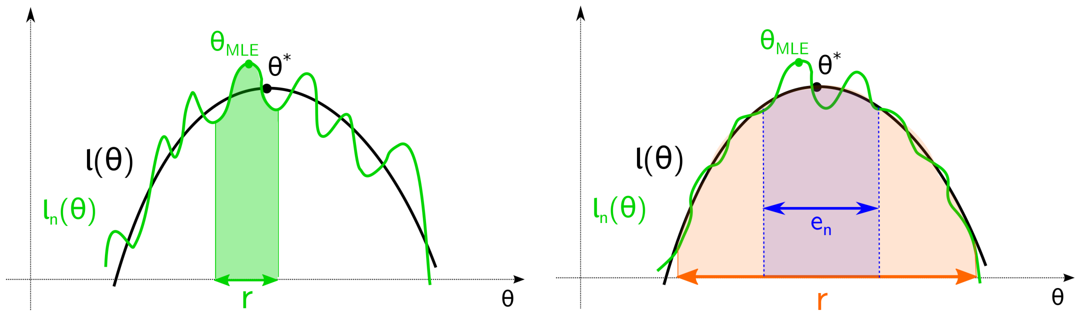
In applications of HMMs, an important problem is to estimate the state transition probabilities and the parameterized output densities based on samples of the observable component. From classical theory, it is known that under suitable regularity conditions, the maximum likelihood estimate (MLE) in an HMM has good statistical properties (Bickel et al., 1998). However, given the potentially nonconvex nature of the likelihood surface, computing the global maximum that defines the MLE is not a straightforward task. In fact, the HMM estimation problem in full generality is known to be computationally intractable under cryptographic assumptions (Terwijn, 2002). In practice, however, the Baum-Welch algorithm (Baum et al., 1970) is frequently applied and leads to good results. It can be understood as the specialization of the EM algorithm (Dempster et al., 1977) to the maximum likelihood estimation problem associated with the HMM. Despite its wide use in many applications, the Baum-Welch algorithm can get trapped in local optima of the likelihood function. Understanding when this undesirable behavior occurs—or does not occur—has remained an open question for several decades.

A more recent line of work (Mossel and Roch, 2006; Siddiqi et al., 2010; Hsu et al., 2012) has focused on developing tractable estimators for HMMs, via approaches that are distinct from the Baum-Welch algorithm. Nonetheless, it has been observed that the practical performance of such methods can be significantly improved by running the Baum-Welch algorithm using their estimators as the initial point; see, for instance, the detailed empirical study in Kontorovich et al. (Kontorovich et al., 2013). This curious phenomenon has been observed in other contexts (Chaganty and Liang, 2013), but has not been explained to date. Obtaining a theoretical characterization of when and why the Baum-Welch algorithm behaves well is the main objective of this paper.

subsectionRelated work and our contributions

Our work builds upon a framework for analysis of EM, as previously introduced by a subset of the current authors (Balakrishnan et al., 2014); see also the follow-up work to regularized EM algorithms (Yi and Caramanis, 2015; Wang et al., 2014). All of this past work applies to models based on i.i.d. samples, and as we show in this paper, there are a number of non-trivial steps required to derive analogous theory for the dependent variables that arise for HMMs. Before doing so, let us put the results of this paper in context relative to older and more classical work on Baum-Welch and related algorithms.

Under mild regularity conditions, it is well-known that the maximum likelihood estimate (MLE) for an HMM is a consistent and asymptotically normal estimator; for instance, see Bickel et al. (Bickel et al., 1998), as well as the expository works (Cappé et al., 2004; van Handel, 2008). On the algorithmic level, the original papers of Baum and co-authors (Baum et al., 1970; Baum and Petrie, 1966) showed that the Baum-Welch algorithm converges to a stationary point of the sample likelihood; these results are in the spirit of the classical convergence analysis of the EM algorithm (Wu, 1983; Dempster et al., 1977). These classical convergence results only provide a relatively weak guarantee—namely, that if the algorithm is initialized sufficiently close to the MLE, then it will converge to it. However, the classical analysis does not quantify the size of this neighborhood, and as a critical consequence, it



(a)

(b)

Figure 1: (a) A poorly behaved sample likelihood, for which there are many local optima at varying distances from the MLE. It would require an initialization extremely close to the MLE in order to ensure that the Baum-Welch algorithm would not be trapped at a sub-optimal fixed point. (b) A well-behaved sample likelihood, for which all local optima lie within an e_n-ball of the MLE, as well as the true parameter θ^*. In this case, the Baum-Welch algorithm, when initialized within a ball of large radius r, will converge to the ball of much smaller radius e_n. The goal of this paper is to give sufficient conditions for when the sample likelihood exhibits this favorable structure.

does not rule out the pathological type of behavior illustrated in panel (a) of Figure 1. Here the sample likelihood has multiple optima, both a global optimum corresponding to the MLE as well as many local optima *far away from the MLE* that are also fixed points of the Baum-Welch algorithm. In such a setting, the Baum-Welch algorithm will only converge to the MLE if it is initialized in an extremely small neighborhood.

In contrast, the goal of this paper is to give sufficient conditions under which the sample likelihood has the more favorable structure shown in panel (b) of Figure 1. Here, even though the MLE does not have a large basin of attraction, the sample likelihood has all of its optima (including the MLE) localized to a small region around the true parameter θ^*. Our strategy to reveal this structure, as in our past work (Balakrishnan et al., 2014), is to shift perspective: instead of studying convergence of Baum-Welch updates to the MLE, we study their convergence to an ϵ_n-ball of the true parameter θ^*, and moreover, instead of focusing exclusively on the sample likelihood, we first study the structure of the population likelihood, corresponding to the idealized limit of infinite data. Our first main result (Theorem 1) provides sufficient conditions under which there is a large ball of radius r, over which the population version of the Baum-Welch updates converge at a geometric rate to θ^*. Our second main result (Theorem 2) uses empirical process theory to analyze the finite-sample version of the Baum-Welch algorithm, corresponding to what is actually implemented in practice. In this finite sample setting, we guarantee that over the ball of radius r, the Baum-Welch updates will converge to an ϵ_n-ball with $\epsilon_n \ll r$, and most importantly, this ϵ_n-ball contains the true parameter θ^*. As a side-note, it also contains the MLE, but our theory does *not* guarantee convergence to the MLE, but rather to a point that is close to both the MLE and the true parameter θ^*.

These latter two results are abstract, applicable to a broad class of HMMs. We then specialize them to the case of a hidden Markov mixture consisting of two isotropic components, with means separated by a constant distance, and obtain concrete guarantees for this model. It is worth comparing these results to past work in the i.i.d. setting, for which the problem of Gaussian mixture estimation under various separation assumptions has been extensively studied (e.g., (Dasgupta, 1999; Vempala and Wang, 2004; Belkin and Sinha, 2010; Moitra and Valiant, 2010)). The constant distance separation required in our work is much weaker than the separation assumptions imposed in papers that focus on correctly labeling samples in a mixture model. Our separation condition is related to, but in general incomparable with the non-degeneracy requirements in other work (Hsu et al., 2012; Hsu and Kakade, 2013; Moitra and Valiant, 2010).

Finally, let us discuss the various challenges that arise in studying the dependent data setting of hidden Markov models, and highlight some important differences with the i.i.d. setting (Balakrishnan et al., 2014; Yi and Caramanis, 2015). In the non-i.i.d. setting, arguments passing from the population-based to sample-based updates are significantly more delicate. First of all, it is not even obvious that the population version of the Q-function—a central object in the Baum-Welch updates—even exists. 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. In particular, each term in the gradient of the likelihood is a function of all observations. 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. Nor do we pursue the approach of some past work on HMM estimation (e.g., (Hsu et al., 2012)), in which it is assumed that there are multiple independent samples of the HMM.[1] Instead, we directly analyze the Baum-Welch algorithm that practioners actually use—namely, one that applies to a single sample of an n-length HMM. In order to make the argument rigorous, we need to make use of more sophisticated techniques for proving concentration for dependent data (Yu, 1994; Nobel and Dembo, 1993).

The remainder of this paper is organized as follows. In Section 2, we introduce basic background on hidden Markov models and the Baum-Welch algorithm. Section 3 is devoted to the statement of our main results in the general setting, whereas Section 4 contains the more concrete consequences for the Gaussian output HMM. The main parts of our proofs are given in Section 5, with the more technical details deferred to the appendices.

2. Background and problem set-up

In this section, we introduce some standard background on hidden Markov models and the Baum-Welch algorithm.

2.1 Standard HMM notation and assumptions

We begin by defining a discrete-time hidden Markov model with hidden states taking values in a discrete space. Letting $\mathbb{Z}$ denote the integers, suppose that the observed random

[1] The rough argument here is that it is possible to reduce an i.i.d. sampling model by cutting the original sample into many pieces, but this is not an algorithm that one would implement in practice.

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 $[s] := \{1, \dots, s\}$. 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 $(z_1, \dots, z_n)$ is given by

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

where the vector β is 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 $(t-1) \to t$ are independent of the index t. 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 denote vectors belonging to the s-dimensional probability simplex.

We assume throughout that the Markov chain is aperiodic and recurrent, whence it has a unique stationary distribution $\bar{\pi}$, defined by the eigenvector equation $\bar{\pi}^T = \bar{\pi}^T A$. To be clear, both $\bar{\pi}$ and the matrix A depend on β, but we omit this dependence so as to simplify notation. We assume throughout that the Markov chain begins in its stationary state, so that $\pi_1 = \bar{\pi}$, and moreover, that it is reversible, meaning that

$$\bar{\pi}(j) A(j, k) = \bar{\pi}(k) A(k, j) \tag{2}$$

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

A key quantity in our analysis is the mixing rate of the Markov chain. In particular, we assume the existence of *mixing constant* $\epsilon_{\text{mix}} \in (0, 1]$ such that

$$\epsilon_{\text{mix}} \leq \frac{p(z_i | z_{i-1}; \beta)}{\bar{\pi}(z_i)} \leq \epsilon_{\text{mix}}^{-1} \tag{3}$$

for all $(z_i, z_{i-1}) \in [s] \times [s]$. This condition implies that the dependence on the initial distribution decays geometrically. More precisely, some simple algebra shows that

$$\sup_{\pi_1} \left\| \pi_1^T A^t - \pi_1^T \right\|_{\text{TV}} \leq c_0 \rho_{\text{mix}}^t \qquad \text{for all } t = 1, 2, \dots, \tag{4}$$

where $\rho_{\text{mix}} = 1 - \epsilon_{\text{mix}}$ denotes the *mixing rate* of the process, and c_0 is a universal constant. Note that as $\epsilon_{\text{mix}} \to 1^-$, the Markov chain has behavior approaching that of an i.i.d. sequence, whereas as $\epsilon_{\text{mix}} \to 0^+$, 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 we assume to be parameterized by a vector μ. Introducing the shorthand $x_1^n = (x_1, \dots, x_n)$ and $z_1^n = (z_1, \dots, z_n)$, the joint

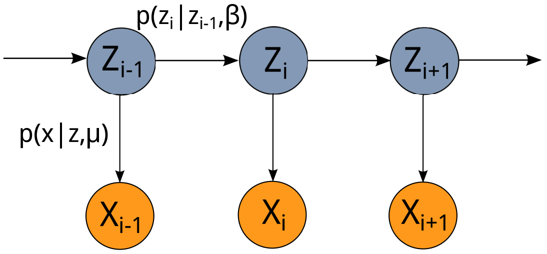


Figure 2: The hidden Markov model as a graphical model. The blue circles indicate observed variables Z_i, whereas the orange circles indicate latent variables X_i.

probability of the sequence (x_1^n, z_1^n) (also known as the complete likelihood) can be written in the form

$$p(z_1^n, x_1^n; \theta) = \pi_1(z_1) \prod_{i=2}^{n} p(z_i \mid z_{i-1}; \beta) \prod_{i=1}^{n} p(x_i | z_i; \mu), \tag{5}$$

where the pair $\theta := (\beta, \mu)$ parameterizes the transition and observation functions. Similarly we define the likelihood

$$p(x_1^n; \theta) = \sum_{z_1^n} p(z_1^n, x_1^n; \theta).$$

We define a form of complete likelihood [2]

$$p(z_0^n, x_1^n; \theta) = \pi_0(z_0) \prod_{i=1}^{n} p(z_i \mid z_{i-1}; \beta) \prod_{i=1}^{n} p(x_i | z_i; \mu), \tag{6}$$

where $\pi_0 = \bar{\pi}$.

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

$$A = \frac{1}{e^{\beta} + e^{-\beta}} \begin{bmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{bmatrix} = \begin{bmatrix} \zeta & 1-\zeta \\ 1-\zeta & \zeta \end{bmatrix}, \tag{7}$$

where $\zeta := \frac{e^{\beta}}{e^{\beta}+e^{-\beta}}$. By construction, this Markov chain is recurrent and aperiodic with the unique stationary distribution $\bar{\pi} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}^T$. Moreover, by calculating the eigenvalues of the transition matrix, we find that the mixing condition (4) holds with $\rho_{\text{mix}} := |2\zeta - 1| = |\tanh(\beta)|$.

Suppose moreover that the observed variables in $\mathbb{R}^d$ are conditionally Gaussian, say with

$$p(x_t | z_t; \mu) = \begin{cases} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\big\{ -\frac{1}{2\sigma^2} \|x - \mu\|_2^2 \big\} & \text{if } z_t = 1 \\ \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\big\{ -\frac{1}{2\sigma^2} \|x + \mu\|_2^2 \big\} & \text{if } z_t = 2. \end{cases} \tag{8}$$

[2] We have defined a complete likelihood that involves an additional hidden variable z_0 that does not have any associated observation x_0. This choice turns out to be convenient, but does preserve the usual relationship $\sum_{z_0^n} p(z_0^n, x_1^n; \theta) = p(x_1^n; \theta)$ between the ordinary and complete likelihoods in EM problems.

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 sequel.

2.2 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 guaranteed to ascend on the likelihood function of the hidden Markov model. 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 (5). The rescaled log likelihood of the sample path x_1^n is 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. For any choice of parameter θ' and positive integers $i \leq j$ and $a < b$, let $\mathbb{E}_{Z_i^j | x_a^b, \theta'}$ denote the expectation under the conditional distribution $p(Z_i^j \mid x_a^b; \theta')$. With this notation, the concavity of the logarithm and Jensen's inequality implies that for any choice of θ', we have the lower bound

$$\ell_n(\theta) \geq \underbrace{\frac{1}{n} \mathbb{E}_{Z_0^n | x_1^n, \theta'} \big[\log p(Z_0^n, x_1^n; \theta) \big]}_{Q_n(\theta \,|\theta')} + \underbrace{\frac{1}{n} \mathbb{E}_{Z_0^n | x_1^n, \theta'} \big[- \log p(Z_0^n \mid x_1^n; \theta') \big]}_{H_n(\theta')}. \tag{9}$$

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 EM operator $M_n : \Omega \mapsto \Omega$

$$M_n(\theta') = \arg\max_{\theta \in \Omega} Q_n(\theta \mid \theta'), \tag{10}$$

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

$$\begin{aligned} Q_n(\theta \mid \theta') = \frac{1}{n} \mathbb{E}_{Z_0 | x_1^n, \theta'} \big[\log \pi_0(Z_0; \beta) \big] + \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Z_{i-1}, Z_i | x_1^n, \theta'} \big[\log p(Z_i \mid Z_{i-1}; \beta) \big] \\ + \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{Z_i | x_1^n, \theta'} \big[\log p(x_i \mid Z_i; \mu) \big], \end{aligned} \tag{11}$$

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

$$Q_n(\theta \mid \theta') = Q_{1,n}(\mu \mid \theta') + Q_{2,n}(\beta \mid \theta') \tag{12}$$

where $Q_{1,n}(\mu \mid \theta') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_i|x_1^n,\theta'}\big[\log p(x_i \mid Z_i, \mu)\big]$, and $Q_{2,n}(\beta \mid \theta')$ collects the remaining terms. In order to compute the expectations defining this function (E-step), we need to determine the 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 obtained efficiently using a recursive message-passing algorithm, known either as the forward-backward or sum-product algorithm (Kschischang et al., 2001; Wainwright and Jordan, 2008).

In the M-step, the decomposition (12) suggests that the maximization over the two components (β, μ) can also be decoupled. Accordingly, with a slight abuse of notation, we often write

$$M_n^{\mu}(\theta') = \arg\max_{\mu\in\Omega_\mu} Q_{1,n}(\mu \mid \theta'), \quad \text{and} \quad M_n^{\beta}(\theta') = \arg\max_{\beta\in\Omega_\beta} Q_{2,n}(\beta \mid \theta')$$

for these two decoupled maximization steps, where Ω_β and Ω_μ denote the feasible set of transition and observation parameters respectively and $\Omega := \Omega_\beta \times \Omega_\mu$.

3. Main results

We now turn to a statement of our main results, along with a discussion of some of their consequences. The first step is to establish the existence of an appropriate population analog of the Q-function. Although the existence of such an object is a straightforward consequence of the law of large numbers in the case of i.i.d. data, it requires some technical effort to establish existence for the case of dependent data; in particular, we do so using k-truncated version of the full Q-function (see Proposition 1). This truncated 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 (see Theorem 1). We then use non-asymptotic forms of empirical process theory to show that under suitable conditions, the actual sample-based EM updates—i.e., the updates that are actually implemented in practice—are also well-behaved in this region with high probability (see Theorem 2). In subsequent analysis to follow in Section 4, we show that this initialization radius is suitably large for an HMM with Gaussian outputs.

3.1 Existence of population Q-function

In the analysis of Balakrishnan et al. Balakrishnan et al. (2014), 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. In their setting of i.i.d. data, the standard law of large numbers ensures that as the sample size n increases, the sample-based Q-function approaches its expectation, namely the function

$$\bar{Q}(\theta \mid \theta') = \mathbb{E}\big[Q_n(\theta \mid \theta')\big] \;=\; \mathbb{E}\big[\mathbb{E}_{Z_1|X_1,\theta'}\big[\log p(X_1, Z_1; \theta)\big]\big].$$

Here we use the shorthand $\mathbb{E}$ for the expectation over all samples X that are drawn from the joint distribution (in this case $\mathbb{E} := \mathbb{E}_{X_1^n|\theta^*}$).

When the samples are dependent, the quantity $\mathbb{E}\big[Q_n(\theta \mid \theta')\big]$ is no longer independent of n, and so an additional step is required. A reasonable candidate for a general definition of the population Q-function is given by

$$\bar{Q}(\theta \mid \theta') := \lim_{n \to +\infty} [\mathbb{E} Q_n(\theta \mid \theta')]. \tag{13}$$

Although it is clear that this definition is sensible in the i.i.d. case, it is necessary for dependent sampling schemes to prove that the limit given in definition (13) actually exists.

In this paper, we do so by considering a suitably truncated version of the sample-based Q-function. Similar arguments have been used in past work (e.g., Cappé et al. (2004); van Handel (2008)) to establish consistency of the MLE; here our focus is instead the behavior of the Baum-Welch algorithm. 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. Recall that $\mathbb{E}_{Z_i^j \mid x_a^b, \theta}$ denotes 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

$$\begin{aligned} Q_n^k(\theta \mid \theta') = \frac{1}{n} \Big[\mathbb{E}_{Z_0 \mid x_{-k}^k, \theta'} \log p(Z_1; \beta) + \sum_{i=1}^n \mathbb{E}_{Z_{i-1}^i \mid x_{i-k}^{i+k}, \theta'} \log p(Z_i \mid Z_{i-1}; \beta) \\ + \sum_{i=1}^n \mathbb{E}_{Z_i \mid x_{i-k}^{i+k}, \theta'} \log p(x_i \mid Z_i; \mu) \Big]. \end{aligned} \tag{14}$$

In an analogous fashion to the decomposition in equation (11), we can decompose Q_n^k in the form

$$Q_n^k(\theta \mid \theta') = Q_{1,n}^k(\mu \mid \theta') + Q_{2,n}^k(\beta \mid \theta').$$

We associate with this triplet of Q-functions the corresponding EM operators $M_n^k(\theta')$, $M_n^{\mu,k}(\theta')$ and $M_n^{\beta,k}(\theta')$ as in Equation (10). Note that as opposed to the function Q_n from equation (11), 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, we are guaranteed that the k-truncated population function and its decomposed analogs given by

$$\begin{aligned} \bar{Q}^k(\theta \mid \theta') &:= \lim_{n \to \infty} \mathbb{E} Q_n^k(\theta \mid \theta') = \mathbb{E} Q_{1,n}^k(\mu \mid \theta') + \lim_{n \to \infty} \mathbb{E} Q_{2,n}^k(\beta \mid \theta') \\ &:= \bar{Q}_1^k(\mu \mid \theta') + \bar{Q}_2^k(\beta \mid \theta') \end{aligned} \tag{15}$$

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

Our first result uses the existence of this truncated population object in order to show that the standard population Q-function from equation (13) is indeed well-defined. In doing so, we make use of the sup-norm

$$\|Q_1 - Q_2\|_\infty := \sup_{\theta, \theta' \in \Omega} \Big| Q_1(\theta \mid \theta') - Q_2(\theta \mid \theta') \Big|. \tag{16}$$

For a radius $r > 0$, define the ball $\mathbb{B}_2(r; \mu^*) = \{\mu \in \mathbb{R}^d \mid \|\mu - \mu^*\|_2 \leq r\}$. We require in the following that the observation densities satisfy the following boundedness condition

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{E}\Big[\max_{z_i \in [s]} \big|\log p(X_i | z_i; \mu)\big|\Big] < \infty. \tag{17}$$

Proposition 1 *Under the previously stated assumptions, the population function $\bar{Q}$ defined in equation* (13) *exists.*

The proof of this claim is given in Appendix A. It hinges on the following auxiliary claim, which bounds the difference between $\mathbb{E}Q_n$ and the k-truncated Q-function as

$$\|\mathbb{E}Q_n - \bar{Q}^k\|_\infty \leq \frac{c\, s^5}{\epsilon_{\text{mix}}^8 \bar{\pi}_{\min}^2} \big(1 - \epsilon_{\text{mix}} \bar{\pi}_{\min}\big)^k + \frac{1}{n} \log \bar{\pi}_{\min}^{-1}, \tag{18}$$

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 (3). Note that the dependencies on ϵ_{mix} and $\bar{\pi}_{\min}$ are not optimized here since it would not help to illustrate the main ideas more clearly. 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.

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

Our ultimate goal is to establish a bound on the difference between the sample-based Baum-Welch estimate and θ^*, in particular showing contraction of the Baum-Welch update towards the true parameter. Our strategy for doing so involves first analyzing the Baum-Welch iterates at the population level, which is the focus of this section.

The quantity $\bar{Q}$ is significant for the EM updates because the parameter θ^* satisfies the self-consistency property $\theta^* = \arg\max_\theta \bar{Q}(\theta \mid \theta^*)$. In the i.i.d. setting, the function $\bar{Q}$ can often be computed in closed form, and hence directly analyzed, as was done in past work Balakrishnan et al. (2014). In the HMM case, this function $\bar{Q}$ no longer has a closed form, so an alternative route is needed. Here we analyze the population version via the truncated function $\bar{Q}^k$ (15) instead, where k is a given truncation level (to be chosen in the sequel). Although θ^* is no longer a fixed point of $\bar{Q}^k$, the bound (18) combined with the assumption of strong concavity of $\bar{Q}^k$ imply an upper bound on the distance of the maximizers of $\bar{Q}^k$ and $\bar{Q}$.

With this setup, we consider an idealized population-level algorithm that, based on some initialization $\widetilde{\theta}^0 \in \Omega = \mathbb{B}_2\big(r; \mu^*\big) \times \Omega_\beta$, generates the sequence of iterates

$$\widetilde{\theta}^{t+1} = \overline{M}^k(\widetilde{\theta}^t) := \arg\max_{\theta \in \Omega} \bar{Q}^k(\theta \mid \widetilde{\theta}^t). \tag{19}$$

Since $\bar{Q}^k$ is an approximate version of $\bar{Q}$, the update operator $\overline{M}^k$ should be understood as an approximation to the idealized population EM operator $\overline{M}$ where the maximum is taken with respect to $\bar{Q}$. As part (a) of the following theorem shows, the approximation error is well-controlled under suitable conditions. We analyze the convergence of the sequence $\{\widetilde{\theta}^t\}_{t=0}^{\infty}$ in terms of the norm $\|\cdot\|_\star : \Omega_\mu \times \Omega_\beta \to \mathbb{R}^+$ given by

$$\|\theta - \theta^*\|_\star = \|(\mu, \beta) - (\mu^*, \beta^*)\|_\star := \|\mu - \mu^*\|_2 + \|\beta - \beta^*\|_2. \tag{20}$$

Contraction in this norm implies that both parameters μ, β converge linearly to the true parameter.

Conditions on $\bar{Q}^k$: Let us now introduce the conditions on the truncated function $\bar{Q}^k$ that underlie our analysis. For a radius $r > 0$, we concentrate on showing conditions in the Cartesian product

$$\Omega := \mathbb{B}_2(r; \mu^*) \times \Omega_\beta,$$

where Ω_β is the set of allowable HMM transition parameters. First, let us say that the function $\bar{Q}^k(\cdot \mid \theta')$ is $(\lambda_\mu, \lambda_\beta)$*-strongly concave* in Ω if

$$\bar{Q}_1^k(\mu_1 \mid \theta^*) - \bar{Q}_1^k(\mu_2 \mid \theta^*) - \langle \nabla_\mu \bar{Q}_1^k(\mu_2 \mid \theta^*), \mu_1 - \mu_2 \rangle \leq -\frac{\lambda_\mu}{2}\|\mu_1 - \mu_2\|_2^2 \tag{21a}$$

$$\text{and} \qquad \bar{Q}_2^k(\beta_1 \mid \theta^*) - \bar{Q}_2^k(\beta_2 \mid \theta^*) - \langle \nabla_\beta \bar{Q}_2^k(\beta_2 \mid \theta^*), \beta_1 - \beta_2 \rangle \leq -\frac{\lambda_\beta}{2}\|\beta_1 - \beta_2\|_2^2 \tag{21b}$$

for all $(\mu_1, \beta_1), (\mu_2, \beta_2) \in \Omega$.

Second, we impose *first-order stability* conditions on the gradients of each component of $\bar{Q}^k$:

- For each $\mu \in \Omega_\mu, \theta' \in \Omega$, we have

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu^*, \beta')\|_2 \leq L_{\mu,1}\|\mu' - \mu^*\|_2 \tag{22a}$$

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta^*)\|_2 \leq L_{\mu,2}\|\beta' - \beta^*\|_2, \tag{22b}$$

  We refer to this condition as L_μ-FOS for short.
- Secondly, for all $\beta \in \Omega_\beta, \theta' \in \Omega$, we require that

$$\|\nabla_\beta \bar{Q}_2^k(\beta \mid \mu', \beta') - \nabla_\beta \bar{Q}_2^k(\beta \mid \mu^*, \beta')\|_2 \leq L_{\beta,1}\|\mu' - \mu^*\|_2 \tag{23a}$$

$$\|\nabla_\beta \bar{Q}_2^k(\beta \mid \mu', \beta') - \nabla_\beta \bar{Q}_2^k(\beta \mid \mu', \beta^*)\|_2 \leq L_{\beta,2}\|\beta' - \beta^*\|_2. \tag{23b}$$

  We refer to this condition as L_β-FOS for short.

As we show in Section 4, these conditions hold for concrete models.

Convergence guarantee for $\bar{Q}^k$-updates: We are now equipped to state our main convergence guarantee for the updates. It involves the quantities

$$L := \max\{L_{\mu_1}, L_{\mu_2}\} + \max\{L_{\beta_1}, L_{\beta_2}\}, \quad \lambda := \min\{\lambda_\mu, \lambda_\beta\} \quad \text{and} \quad \kappa := \frac{L}{\lambda}, \tag{24}$$

with κ generally required to be smaller than one, as well as the additive norm $\|\cdot\|_\star$ from equation (20).

Part (a) of the theorem controls the *approximation error* induced by using the k-truncated function $\bar{Q}^k$ as opposed to the exact population function $\bar{Q}$, whereas part (b) guarantees a *geometric rate of convergence* in terms of κ defined above in equation (24).

Theorem 1 *(a)* Approximation guarantee: *Under the mixing condition* (4)*, density boundedness condition* (17)*, and* $(\lambda_\mu, \lambda_\beta)$*-strong concavity condition* (21)*, there is a universal constant* c_0 *such that*

$$\|\overline{M}^k(\theta) - \overline{M}(\theta)\|_\star^2 \leq \underbrace{c_0 \frac{s^5}{\lambda\, \epsilon_{\text{mix}}^8 \overline{\pi}_{\min}^2} \big(1 - \epsilon_{\text{mix}} \overline{\pi}_{\min}\big)^k}_{=:\varphi^2(k)} \qquad \text{for all } \theta \in \Omega, \tag{25}$$

where s *is the number of states, and* $\overline{\pi}_{\min} := \min_{\beta \in \Omega_\beta} \min_{j \in [s]} \overline{\pi}(j;\beta)$*.*

(b) Convergence guarantee: *Suppose in addition that the* (L_μ, L_β)*-FOS conditions* (22),(23) *holds with parameter* $\kappa \in (0,1)$ *as defined in* (24) *for* $\theta, \theta' \in \Omega = \mathbb{B}_2\big(r;\mu^*\big) \times \Omega_\beta$*, and that the truncation parameter* k *is sufficiently large to ensure that*

$$\varphi(k) \leq \big(1 - \kappa\big) r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2.$$

Then given an initialization $\widetilde{\theta}^0 \in \Omega$*, the iterates* $\{\widetilde{\theta}^t\}_{t=0}^\infty$ *generated by the* $\overline{M}^k$ *operator satisfy the bound*

$$\|\widetilde{\theta}^t - \theta^*\|_\star \leq \kappa^t \|\widetilde{\theta}^0 - \theta^*\|_\star + \frac{1}{1-\kappa}\varphi(k). \tag{26}$$

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 (25) and (26) show, this approximation error is controlled, and decays exponentially in k. The proof of the recursive bound (26) is based on first showing that

$$\|\overline{M}^k(\theta) - \overline{M}^k(\theta^*)\|_\star \leq \kappa \|\theta - \theta^*\|_\star \tag{27}$$

for any $\theta \in \Omega$. Inequality (27) 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.

Finally, when Theorem 1 is applied to a concrete model, the task is to find the biggest r and Ω_β such that the conditions in the theorem are satisfied, and we do so for the Gaussian output HMM in Section 4.

3.3 Sample-based results

We now turn to a result that applies to the sample-based form of the Baum-Welch algorithm—that is, corresponding to the updates that are actually applied in practice. For a tolerance parameter $\delta \in (0,1)$, we let $\varphi_n(\delta, k)$ be the smallest positive scalar such that

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{P}\Big[\|M_n(\theta) - M_n^k(\theta)\|_\star \geq \varphi_n(\delta,k)\Big] \leq \delta. \tag{28a}$$

This quantity bounds the approximation error induced by the k-truncation, and is the sample-based analogue of the quantity $\varphi(k)$ appearing in Theorem 1(a). For each $\delta \in (0,1)$, we let $\epsilon_n^\mu(\delta,k)$ and $\epsilon_n^\beta(\delta,k)$ denote the smallest positive scalars such that

$$\mathbb{P}\big[\|M_n^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta)\|_2 \geq \epsilon_n^\mu(\delta,k)\Big] \leq \delta, \quad \text{and} \quad \mathbb{P}\big[\|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)\|_2 \geq \epsilon_n^\beta(\delta,k)\Big] \leq \delta \tag{28b}$$

for all $\theta \in \mathbb{B}_2(\mu^*; r) \times \Omega_\beta$. Furthermore we define $\epsilon_n(\delta, k) := \epsilon_n^\mu(\delta, k) + \epsilon_n^\beta(\delta, k)$. For a given truncation level k, these values give an upper bound on the difference between the population and sample-based M-operators, as induced by having only a finite number n of samples.

Theorem 2 (Sample Baum-Welch) *Suppose that the truncated population EM operator $\overline{M}^k$ satisfies the local contraction bound (27) with parameter $\kappa \in (0,1)$ in Ω. For a given sample size n, suppose that (k, n) are sufficiently large to ensure that*

$$\varphi_n(\delta, k) + \epsilon_n^\mu\big(\delta, k\big) + \varphi(k) \leq (1-\kappa)\, r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2. \tag{29a}$$

Then given any initialization $\widehat{\theta}^0 \in \Omega$, with probability at least $1 - 2\delta$, the Baum-Welch sequence $\{\widehat{\theta}^t\}_{t=0}^\infty$ satisfies the bound

$$\|\widehat{\theta}^t - \theta^*\|_\star \leq \underbrace{\kappa^t \|\widehat{\theta}^0 - \theta^*\|_\star}_{\textit{Geometric decay}} + \underbrace{\frac{1}{1-\kappa}\Big\{2\varphi_n\big(\delta, k\big) + \epsilon_n\big(\delta, k\big) + \varphi(k)\Big\}}_{\textit{Residual error } e_n}. \tag{29b}$$

The bound (29b) 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. In particular, we can see that the residual error is of the same order as for the MLE and that the required initialization radius is optimal up to constants. Let us emphasize here that k as well as the truncated operators are purely theoretical objects which were introduced for the analysis.

4. Concrete results for the Gaussian output HMM

We now return to the concrete example of a Gaussian output HMM, as first introduced in Section 2.1, and specialize our general theory to it. Before doing so, let us make some preliminary comments about our notation and assumptions. Recall that our Gaussian output HMM is based on $s = 2$ hidden states, using the transition matrix from equation (7), and the Gaussian output densities from equation (8). For convenience of analysis, we let the hidden variables Z_i take values in $\{-1, 1\}$. In addition, we require that the mixing coefficient $\rho_{\text{mix}} = 1 - \epsilon_{\text{mix}}$ is bounded away from 1 in order to ensure that the mixing condition (3) is fulfilled. We denote the upper bound for ρ_{mix} as $b < 1$ so that $\rho_{\text{mix}} \leq b$ and $\epsilon_{\text{mix}} \geq 1 - b$. The feasible set of the probability parameter ζ and its log odds analog $\beta = \frac{1}{2}\log\left(\frac{\zeta}{1-\zeta}\right)$ are then given by

$$\Omega_\zeta = \Big\{\zeta \in \mathbb{R} \mid \frac{1-b}{2} \leq \zeta \leq \frac{1+b}{2}\Big\}, \quad \text{and} \quad \Omega_\beta = \Big\{\beta \in \mathbb{R} \ \mid\ |\beta| < \underbrace{\frac{1}{2}\log\big(\frac{1+b}{1-b}\big)}_{\beta_B}\Big\}. \tag{30}$$

4.1 Explicit form of Baum-Welch updates

We begin by deriving an explicit form of the Baum-Welch updates for this model. Using this notation, the Baum-Welch updates take the form

$$\widehat{\mu}^{t+1} = \frac{1}{n}\sum_{i=1}^{n}(2p(Z_i = 1 \mid x_1^n;\theta^t) - 1)x_i, \tag{31a}$$

$$\widehat{\zeta}^{t+1} = \Pi_{\Omega_\zeta}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{Z_i} p(Z_i = Z_{i+1} \mid x_1^n;\theta^t)\right), \text{ and} \tag{31b}$$

$$\widehat{\beta}^{t+1} = \frac{1}{2}\log\big(\frac{\widehat{\zeta}^{t+1}}{1-\widehat{\zeta}^{t+1}}\big), \tag{31c}$$

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 $Q_{1,n}(\cdot \mid \theta^t), Q_{2,n}(\cdot \mid \theta^t)$. 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.

4.2 Population and sample guarantees

We now use the results from Section 3 to show that the population and sample-based version of the Baum-Welch updates are linearly convergent in a ball around θ^* of fixed radius. In establishing the population-level guarantee, the key conditions which need to be fulfilled—and the one that are the most technically challenging to establish— are the (L_μ, L_β)-FOS conditions (22), (23). In particular, we want to show that these conditions hold with Lipschitz constants L_μ, L_β that decrease exponentially with the separation of the mixtures. As a consequence, we obtain that for large enough separation $\frac{L}{\lambda} < 1$, i.e. the EM operator is contractive towards the true parameter.

Throughout this section, we use c_0, c_1 to denote universal constants and C_0, C_1 for 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. 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

We begin by stating a result for 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 $\kappa = \kappa(\eta)$ that is given by

$$\kappa(\eta) := \frac{C_1\eta^2(\eta^2+1)\,\mathrm{e}^{-c_2\eta^2}}{1-b^2}. \tag{32}$$

Corollary 1 (Population Baum-Welch) *Consider a two-state Gaussian output HMM that is mixing (i.e. satisfies equation* (3)*), and with its SNR lower bounded as $\eta^2 \geq C$ 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 $\varphi(k) \le (1-\kappa)r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2$. Then for any initialization $\widetilde{\theta}^0 = (\widetilde{\mu}^0, \widetilde{\beta}^0) \in \mathbb{B}_2\big(r; \mu^*\big) \times \Omega_\beta$, the sequence $\{\widetilde{\theta}^t\}_{t=0}^\infty$ generated by $\overline{M}^k$ satisfies the bound*

$$\|\widetilde{\theta}^t - \theta^*\|_\star \le \kappa^t \|\widetilde{\theta}^0 - \theta^*\|_\star + \frac{1}{1-\kappa}\varphi(k) \tag{33}$$

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

From definition (32) it follows that as long as the signal-to-noise ratio η is larger than a universal constant, the convergence rate $\kappa(\eta) < 1$. The bound (33) then ensures a type of contraction and the pre-condition $\varphi(k) \le (1-\kappa)r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2$ can be satisfied by choosing the truncation parameter k large enough. If we use a finite truncation parameter k, then the 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 Baum-Welch iterates) *For a given tolerance $\delta \in (0,1)$, suppose that the sample size is lower bounded as $n \ge C_1(\sigma^2 + \|\mu^*\|_2^2) d \log^2(\frac{d}{\delta})$ for a sufficiently large C_1. Then under the conditions of Corollary 1, with probability at least $1-\delta$, we have*

$$\|\widetilde{\theta}^t - \theta^*\|_\star \le \kappa^t \|\widehat{\theta}^0 - \theta^*\|_\star + C \frac{\|\mu^*\|_2 (\frac{\|\mu^*\|_2^2}{\sigma^2} + 1) \log^2 n \sqrt{\frac{d \log^2(n/\delta)}{n}}}{1-\kappa}. \tag{34}$$

Remarks: As a consequence of the bound (34), 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 $\sqrt{\frac{d \log^6(n)}{n}}$.

In order to interpret this guarantee, note that in the case of symmetric Gaussian output HMMs as in Section 4, standard techniques can be used to show that the minimax rate of estimating μ^* in Euclidean norm scales as $\sqrt{\frac{d}{n}}$. If we could compute the MLE in polynomial time, then its error would also exhibit this scaling. The significance of Corollary 2 is that it shows that the Baum-Welch update achieves this minimax risk of estimation up to logarithmic factors.

Moreover, it should be noted that the initialization radius given here is essentially optimal up to constants. Because of the symmetric nature of the population log-likelihood, the all zeroes vector is a stationary point. Consequently, the maximum Euclidean radius of any basin of attraction for one of the observation parameters—that is, either μ^* or $-\mu^*$—can at most be $r = \|\mu^*\|_2$. Note that our initialization radius only differs from this maximal radius by only a small constant factor.

4.3 Simulations

In this section, we provide the results of simulations that confirm the accuracy of our theoretical predictions for two-state Gaussian output HMMs. In all cases, we update the estimates for the mean vector $\widehat{\mu}^{t+1}$ and transition probability $\hat{\zeta}^{t+1}$ according to equation (31); for convenience, we update ζ as opposed to β. The true parameters are denoted by μ^* and ζ^*.

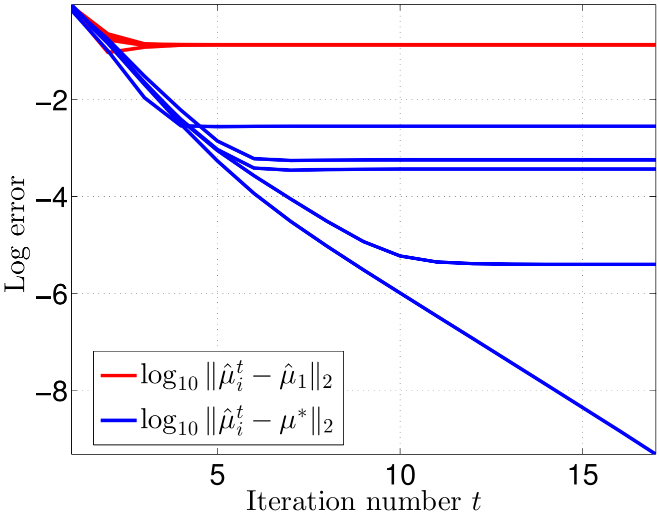


Figure 3: Plot of the convergence of the optimization error $\log \|\widehat{\mu}_i^t - \widehat{\mu}_1\|_2$, plotted in blue, and the statistical error $\log \|\widehat{\mu}_i^t - \mu^*\|_2$, plotted in red, for 5 different initializations. The parameter settings were $d = 10$, $n = 1000$, $\rho_{\text{mix}} = 0.6$ and SNR $\frac{\|\mu^*\|_2}{\sigma} = 1.5$. See the main text for further details.

In all simulations, we fix the mixing parameter to $\rho_{\text{mix}} = 0.6$, generate initial vectors $\widehat{\mu}^0$ randomly in a ball of radius $r := \frac{\|\mu^*\|_2}{4}$ around the true parameter μ^*, and set $\widehat{\zeta}^0 = \frac{1}{2}$. 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.

Figure 3 depicts the convergence behavior of the Baum-Welch updates, as assessed in terms 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 $i-$th trial by $\widehat{\mu}_i$. The curves in blue depict the *optimization error*—that is, the differences between the Baum-Welch iterates $\widehat{\mu}_i^t$ using the i-th initialization, and $\widehat{\mu}_1$. On the other hand, the red lines represent the *statistical error*—that is, 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. The convergence of the statistical error aligns with the theoretical prediction in upper bound (34) of Corollary 2. The (minimax-optimal) error floor in the curve corresponds to the residual error and the e_n–region in Figure 1. In addition, the blue optimization error curves show that for different initializations, the Baum-Welch algorithm converges to *different stationary points* $\widehat{\mu}_i$; however, all of these points have roughly the same distance from μ^*. This phenomenon highlights the importance of the change of perspective in our analysis—that is, focusing on the true parameter as opposed to the MLE. Given the presence of all these local optima in a small neighborhood of μ^*, the MLE basin of attraction must necessarily be much smaller than the initialization radius guaranteed by our theory.

Figure 4 shows how the convergence rate of the Baum-Welch algorithm depends on the underlying SNR parameter η^2; this behavior confirms the predictions given in Corollary 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.

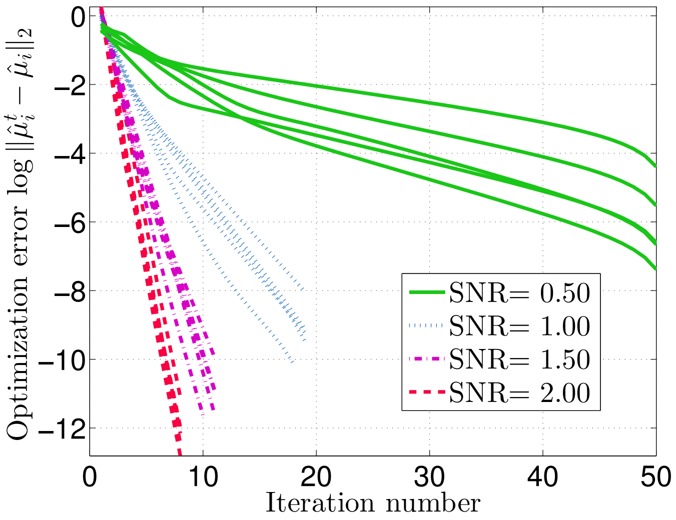


Figure 4: Plot of convergence behavior for different SNR, where for each curve, different parameters were chosen. The parameter settings are $d = 10$, $n = 1000$ and $\rho_{\text{mix}} = 0.6$.

5. Proofs

In this section, we collect the proofs of our main results. In all cases, we provide the main bodies of the proofs here, deferring the more technical details to the appendices.

5.1 Proof of Theorem 1

Throughout this proof, we make use of the shorthand $\widetilde{\rho}_{\text{mix}} = 1 - \epsilon_{\text{mix}}\overline{\pi}_{\text{min}}$. Also we denote the separate components of the population EM operators by $\overline{M}(\theta) =: (\overline{M}^{\mu}(\theta), \overline{M}^{\beta}(\theta))^T$ and their truncated equivalents by $\overline{M}^k(\theta) =: (\overline{M}^{\mu,k}(\theta), \overline{M}^{\beta,k}(\theta))^T$. We begin by proving the bound given in part (a). Since $\bar{Q} = \lim_{n\to\infty} \mathbb{E}[Q_n]$, we have

$$\|\bar{Q} - \bar{Q}^k\|_\infty = \| \lim_{n\to\infty} \mathbb{E}[Q_n] - \bar{Q}^k\|_\infty \;\le\; \frac{Cs^5}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k,$$

where we have exchanged the supremum and the limit before applying the bound (18). The same holds for the separate functions $\bar{Q}_1, \bar{Q}_2$.

Using this bound and the fact that for $\bar{Q}_1$ we have $\bar{Q}_1(\overline{M}^{\mu}(\theta) \mid \theta) \geq \bar{Q}_1(\overline{M}^{\mu,k}(\theta) \mid \theta)$, we find that

$$\bar{Q}_1(\overline{M}^{\mu}(\theta) \mid \theta) \geq \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta) - \frac{Cs^5}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k.$$

Since $\overline{M}^{\mu,k}(\theta)$ is optimal, the first-order conditions for optimality imply that

$$\langle \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta), \, \theta - \overline{M}^{\mu,k}(\theta) \rangle \leq 0 \qquad \text{for all } \theta \in \mathbb{B}_2\big(r; \theta^*\big).$$

By the strict concavity of $\bar{Q}^k(\cdot|\theta)$ for all θ, we obtain

$$\begin{aligned}
\frac{Cs^5}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k &\geq \bar{Q}_1(\overline{M}^{\mu}(\theta) \mid \theta) - \bar{Q}_1^k(\overline{M}^{\mu}(\theta) \mid \theta) \\
&\geq \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta) - \bar{Q}_1^k(\overline{M}^{\mu}(\theta) \mid \theta) - \frac{Cs^5}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k \\
&\geq \frac{\lambda_\mu}{2} \|\overline{M}^{\mu}(\theta) - \overline{M}^{\mu,k}(\theta)\|_2^2 - \frac{Cs^5}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k
\end{aligned}$$

and therefore $\|\overline{M}^{\mu}(\theta) - \overline{M}^{\mu,k}(\theta)\|_2^2 \leq 4 \frac{Cs^5}{\lambda \epsilon_{\text{mix}}^8 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k$. In particular, setting $\theta = \theta^*$ and identifiability, i.e. $\overline{M}^{\mu}(\theta^*) = \theta^*$, yields

$$\|\overline{M}^{\mu,k}(\theta^*) - \theta^*\|_2^2 \leq 4 \frac{Cs^5}{\lambda_\mu \epsilon_{\text{mix}}^6 \overline{\pi}_{\text{min}}^2} \widetilde{\rho}_{\text{mix}}^k,$$

and the equivalent bound can be obtained for $\overline{M}^{\beta,k}(\cdot)$ which yields the claim.

We now turn to the proof of part (b). Let us suppose that the recursive bound (27) holds, and use it to complete the proof of this claim. We first show that if $\widetilde{\mu}^t \in \mathbb{B}_2(r; \mu^*)$, then we must have $\widetilde{\mu}^{t+1} \in \mathbb{B}_2(r; \mu^*)$ as well. Indeed, if $\widetilde{\mu}^t \in \mathbb{B}_2(r; \mu^*)$, then we have by triangle inequality and the (L_μ, L_β)-FOS condition

$$\begin{aligned}
\|\overline{M}^{\mu,k}(\widetilde{\theta}^t) - \mu^*\|_2 &\leq \|\overline{M}^{\mu,k}(\widetilde{\theta}^t) - \overline{M}^{\mu,k}(\mu^*, \widetilde{\beta}^t)\|_2 + \|\overline{M}^{\mu,k}(\mu^*, \widetilde{\beta}^t) - \overline{M}^{\mu,k}(\mu^*, \beta^*)\|_2 + \|\overline{M}^{\mu,k}(\theta^*) - \mu^*\|_2 \\
&\leq \frac{L_{\mu,1}}{\lambda_\mu} \|\widetilde{\mu}^t - \mu^*\|_2 + \frac{L_{\mu,2}}{\lambda_\mu} \|\widetilde{\beta}^t - \beta^*\|_2 + \frac{\varphi(k)}{2} \\
&\leq \kappa(r + \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2) + \varphi(k) \leq r,
\end{aligned}$$

where the final step uses the assumed bound on φ. For the joint parameter update we in turn have

$$\begin{aligned}
\|\overline{M}^k(\widetilde{\theta}^t) - \theta^*\|_\star &\leq \|\overline{M}^k(\widetilde{\theta}^t) - \overline{M}^k(\theta^*)\|_\star + \|\overline{M}^k(\theta^*) - \theta^*\|_\star \\
&\leq \kappa \|\widetilde{\theta}^t - \theta^*\|_\star + \varphi(k).
\end{aligned} \tag{35}$$

By repeatedly applying inequality (35) and summing the geometric series, the claimed bound (26) follows.

It remains to prove the bound (27). Since the vector $\overline{M}^k(\theta^*)$ maximizes the function $\theta \mapsto \bar{Q}_1^k(\theta \mid \theta^*)$, we have the first-order optimality condition

$$\langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta^*) \mid \theta^*),\, \overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\rangle \leq 0, \qquad \text{valid for any } \theta.$$

Similarly, we have $\langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle \leq 0$, and adding together these two inequalities yields

$$0 \leq \langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta^*) \mid \theta^*) - \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\mu) \mid \theta),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle$$

On the other hand, by the λ-strong concavity condition, we have

$$\lambda_\mu \|\overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\|_2^2 \leq \langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta^*)) - \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta^*) \mid \theta^*),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle$$

Combining these two inequalities yields

$$\begin{aligned}\lambda_\mu \|\overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\|_2^2 &\leq \langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta^*) - \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle \\ &\leq \big[L_{\mu_1}\|\mu - \mu^*\|_2 + L_{\mu_2}\|\beta - \beta^*\|_2\big] \|\overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\|_2,\end{aligned}$$

and similarly we obtain $\lambda_\beta \|\overline{M}^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta^*)\|_2 \leq \big[L_{\beta_1}\|\mu - \mu^*\|_2 + L_{\beta_2}\|\beta - \beta^*\|_2\big]$. Adding both inequalities yields the claim (27).

5.2 Proof of Theorem 2

By the triangle inequality, for any step, with probability at least $1 - 2\delta$, we have

$$\begin{aligned}\|\widehat{\theta}^{t+1} - \theta^*\|_\star &\leq \|M_n(\widehat{\theta}^t) - M_n^k(\widehat{\theta}^t)\|_\star + \|M_n^k(\widehat{\theta}^t) - \overline{M}^k(\widehat{\theta}^t)\|_\star + \|\overline{M}^k(\widehat{\theta}^t) - \theta^*\|_\star \\ &\leq \varphi_n(\delta, k) + \epsilon_n(\delta, k) + \kappa\|\widehat{\theta}^t - \theta^*\|_\star + \varphi(k).\end{aligned}$$

In order to see that the iterates do not leave $\mathbb{B}_2\big(r; \mu^*\big)$, observe that

$$\begin{aligned}\|\widehat{\mu}^{t+1} - \mu^*\|_2 &\leq \|M_n^\mu(\widehat{\theta}^t) - M_n^{\mu,k}(\widehat{\theta}^t)\|_2 + \|M_n^{\mu,k}(\widehat{\theta}^t) - \overline{M}^{\mu,k}(\widehat{\theta}^t)\|_2 + \|\overline{M}^{\mu,k}(\widehat{\theta}^t) - \mu^*\|_2 \qquad (36) \\ &\leq \varphi_n(\delta, k) + \epsilon_n^\mu(\delta, k) + \kappa(\|\widehat{\mu}^t - \mu^*\|_2 + \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2) + \frac{\varphi(k)}{2}.\end{aligned}$$

Consequently, as long as $\|\widehat{\mu}^t - \mu^*\|_2 \leq r$, we also have $\|\widehat{\mu}^{t+1} - \mu^*\|_2 \leq r$ whenever

$$\varphi_n(\delta, k) + \varphi(k) + \epsilon_n^\mu(\delta, k) \leq (1 - \kappa)\, r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2.$$

Combining inequality (36) with the equivalent bound for β, we obtain

$$\|\widehat{\theta}^t - \theta^*\|_\star \leq \kappa\|\widehat{\theta}^{t-1} - \theta^*\|_\star + 2\varphi_n(\delta, k) + \epsilon_n(\delta, k) + \varphi(k)$$

Summing the geometric series yields the bound (29b).

5.3 Proof of Corollary 1

The boundedness condition (Assumption (17)) is easy to check since for $X \sim N(\mu^*, \sigma^2)$, the quantity $\sup_{\mu \in \mathbb{B}_2(r;\mu^*)} \mathbb{E}\big[\max\{\|X-\mu\|_2, \|X+\mu\|_2\}\big]$ is finite for any choice of radius $r < \infty$.

By Theorem 1, the k-truncated population EM iterates satisfy the bound

$$\|\widetilde{\theta}^t - \theta^*\|_\star \leq \kappa^t \|\widetilde{\theta}^0 - \theta^*\|_\star + \frac{1}{1-\kappa}\varphi(k), \tag{37}$$

if the strong concavity (21) and FOS conditions (22), (23) hold with suitable parameters.

In the remainder of proof—and the bulk of the technical work— we show that:

- strong concavity holds with $\lambda_\mu = 1$ and $\lambda_\beta = 1 - b^2$;
- the FOS conditions hold with

$$L_{\mu,1} = c\ (\eta^2+1)\varphi_2(\epsilon_{\text{mix}})\eta^2 \mathrm{e}^{-c\eta^2}, \quad \text{and} \quad L_{\mu,2} = c\sqrt{\|\mu^*\|_2^2 + \sigma^2}\varphi_2(\epsilon_{\text{mix}})\eta^2\mathrm{e}^{-c\eta^2}$$
$$L_{\beta,1} = c\frac{1-b}{1+b}\varphi_2(\epsilon_{\text{mix}})\eta^2\mathrm{e}^{-c\eta^2} \quad \text{and} \quad L_{\beta,2} = c\sqrt{\|\mu^*\|_2^2 + \sigma^2}\varphi_2(\epsilon_{\text{mix}})\eta^2\mathrm{e}^{-c\eta^2},$$

where $\varphi_2(\epsilon_{\text{mix}}) := \left(\frac{1}{\log(1/(1-\epsilon_{\text{mix}}))} + \frac{1}{\epsilon_{\text{mix}}}\right)$. Substuting these choices into the bound (37) and performing some algebra yields the claim.

5.3.1 Establishing strong concavity

We first show concavity of $\bar{Q}_1^k(\cdot \mid \theta')$ and $\bar{Q}_2^k(\cdot \mid \theta')$ separately. For strong concavity of $\bar{Q}_1^k(\cdot \mid \theta')$, observe that

$$\bar{Q}_1^k(\mu \mid \theta') = -\frac{1}{2}\mathbb{E}\left[p(z_0 = 1 \mid X_{-k}^k; \theta')\|X_0 - \mu\|_2^2 + (1 - p(z_0 = 1 | X_{-k}^k; \theta'))\|X_0 + \mu\|_2^2 + c\right],$$

where c is a quantity independent of μ. By inspection, this function is strongly concave in μ with parameter $\lambda_\mu = 1$.

On the other hand, we have

$$\bar{Q}_2^k(\beta \mid \theta') = \mathbb{E}_{X_{-k}^k | \theta^*} \sum_{z_0, z_1} p(z_0, z_1 \mid X_{-k}^k; \theta') \log\left(\frac{\mathrm{e}^{\beta z_0 z_1}}{\mathrm{e}^{\beta} + \mathrm{e}^{-\beta}}\right).$$

This function has second derivative $\frac{\partial^2}{\partial \beta^2}\bar{Q}_2^k(\beta \mid \theta') = -4\frac{\mathrm{e}^{-2\beta}}{(\mathrm{e}^{-2\beta}+1)^2}$. As a function of $\beta \in \Omega_\beta$, this second derivative is maximized at $\beta = \frac{1}{2}\log\left(\frac{1+b}{1-b}\right)$. Consequently, the function $\bar{Q}_2^k(\cdot \mid \theta')$ is strongly concave with parameter $\lambda_\beta = 1 - b^2$.

5.3.2 Separate FOS conditions

We now turn to proving that the FOS conditions in equations (22) and (23) hold. A key ingredient in our proof is the fact that the conditional density $p(z_{-k}^k \mid x_{-k}^k; \mu, \beta)$ belongs to the exponential family with parameters $\beta \in \mathbb{R}$, and $\gamma_i := \frac{\langle \mu, x_i \rangle}{\sigma^2} \in \mathbb{R}$ for $i = -k, \ldots, k$.

(See the book Wainwright and Jordan (2008) for more details on exponential families.) In particular, we have

$$\underbrace{p(z_{-k}^k \mid x_{-k}^k, \mu, \beta)}_{:=p(z_{-k}^k;\gamma,\beta)} = \exp\left\{\sum_{\ell=-k}^{k} \gamma_\ell z_\ell + \beta \sum_{\ell=-k}^{k-1} z_\ell z_{\ell+1} - \Phi(\gamma,\beta)\right\}, \tag{38}$$

where the function h absorbs various coupling terms. Note that this exponential family is a specific case of the following exponential family distribution

$$\underbrace{\tilde{p}(z_{-k}^k \mid x_{-k}^k, \mu, \beta)}_{:=\tilde{p}(z_{-k}^k;\gamma,\beta)} = \exp\left\{\sum_{\ell=-k}^{k} \gamma_\ell z_\ell + \sum_{\ell=-k}^{k-1} \beta_\ell z_\ell z_{\ell+1} - \Phi(\gamma,\beta)\right\}. \tag{39}$$

The distribution in (38) corresponds to (39) with $\beta_\ell = \beta$ for all ℓ and the so-called partition function Φ is given by

$$\Phi(\gamma,\beta) = \log \sum_z \exp\left\{\sum_{\ell=-k}^{k} \gamma_\ell z_\ell + \sum_{\ell=-k}^{k-1} \beta_\ell z_\ell z_{\ell+1}\right\}.$$

The reason to view our distribution as a special case of the more general one in (39) becomes clear when we consider the equivalence of expectations and the derivatives of the cumulant function

$$\left.\frac{\partial \Phi}{\partial \gamma_\ell}\right|_{\theta'} = \mathbb{E}_{Z_{-k}^k|x_{-k}^k,\theta'} Z_\ell \quad \text{and} \quad \left.\frac{\partial \Phi}{\partial \beta_0}\right|_{\theta'} = \mathbb{E}_{Z_{-k}^k|x_{-k}^k,\theta'} Z_0 Z_1, \tag{40}$$

where we recall that $\mathbb{E}_{Z_{-k}^k|x_{-k}^k,\theta'}$ is the expectation with respect to the distribution $\tilde{p}(Z_{-k}^k \mid x_{-k}^k; \mu', \beta')$ with $\beta_\ell = \beta'$. Note that in the following any value θ' for $\tilde{p}$ is taken to be on the manifold on which $\beta_\ell = \beta'$ for some β' since this is the manifold the algorithm works on. Also, as before, $\mathbb{E}$ denotes the expectation over the joint distribution of all samples X_ℓ drawn according to $p(\cdot;\theta^*)$, in this case X_{-k}^k.

Similarly to equations (40), the covariances of the sufficient statistics correspond to the second derivatives of the cumulant function

$$\left.\frac{\partial^2 \Phi}{\partial \beta_\ell \partial \beta_0}\right|_{\theta} = \operatorname{cov}(Z_0 Z_1, Z_\ell Z_{\ell+1} \mid X_{-k}^k, \theta) \tag{41a}$$

$$\left.\frac{\partial^2 \Phi}{\partial \gamma_\ell \partial \gamma_0}\right|_{\theta} = \operatorname{cov}(Z_0, Z_\ell \mid X_{-k}^k, \theta) \tag{41b}$$

$$\left.\frac{\partial^2 \Phi}{\partial \beta_\ell \partial \gamma_0}\right|_{\theta} = \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid X_{-k}^k, \theta). \tag{41c}$$

In the following, we adopt the shorthand

$$\begin{aligned}\operatorname{cov}(Z_\ell, Z_{\ell+1} \mid \gamma', \beta') &= \operatorname{cov}(Z_\ell, Z_{\ell+1} \mid X_{-k}^k, \theta') \\ &= \mathbb{E}_{Z_\ell^{\ell+1}|X_{-k}^k,\theta'}(Z_\ell - \mathbb{E}_{Z_\ell^{\ell+1}|X_{-k}^k,\theta'} Z_\ell)(Z_{\ell+1} - \mathbb{E}_{Z_\ell^{\ell+1}|X_{-k}^k,\theta'} Z_{\ell+1})\end{aligned}$$

where the dependence on β is occasionally omitted so as to simplify notation.

5.3.3 Proof of inequality (22a)

By an application of the mean value theorem, we have

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu^*, \beta')\| \le \underbrace{\left\| \mathbb{E} \sum_{\ell=-k}^{k} \frac{\partial^2 \Phi}{\partial \gamma_\ell \partial \gamma_0}\bigg|_{\theta=\widetilde{\theta}} (\gamma'_\ell - \gamma^*_\ell) X_0 \right\|}_{T_1}$$

where $\widetilde{\theta} = \theta' + t(\theta^* - \theta')$ for some $t \in (0,1)$. Since second derivatives yield covariances (see equation (41)), we can write

$$T_1 = \left\| \sum_{\ell=-k}^{k} \mathbb{E} X_0 \mathbb{E}\left[\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \frac{\langle \mu' - \mu^*, X_\ell \rangle}{\sigma^2} \,\middle|\, X_0 \right] \right\|_2,$$

so that it suffices to control the expected conditional covariance. By the Cauchy-Schwarz inequality and the fact that $\operatorname{cov}(XY) \le \sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}$ and $\operatorname{var}(Z_0 \mid X) \le 1$, we obtain the following bound on the expected conditional covariance by using Lemma 4 (see Appendix B.1)

$$\begin{aligned} \mathbb{E}\left[\left| \operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \mid X_0 \right| \right] &\le \sqrt{\mathbb{E}\left[\operatorname{var}(Z_0 \mid \widetilde{\gamma}) \mid X_0 \right]} \sqrt{\mathbb{E}\left[\operatorname{var}(Z_\ell \mid \widetilde{\gamma}) \mid X_0 \right]} \\ &\le \sqrt{\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)}. \end{aligned} \tag{42a}$$

Furthermore, by Lemma 5 and 6 (see Appendix B.1), we have

$$|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})| \le 2\rho_{\text{mix}}^{\ell}, \quad \text{and} \quad \left\| \mathbb{E}(\operatorname{var}(Z_0 | \widetilde{\gamma}_0))^{1/2} X_0 X_0^T \right\|_{op} \le C \mathrm{e}^{-c\eta^2}. \tag{42b}$$

From the definition of the operator norm, we have

$$\begin{aligned} \|\mathbb{E} \operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) X_0 X_\ell^T\|_{op} &= \sup_{\substack{\|u\|_2=1 \\ \|v\|_2=1}} \mathbb{E} \operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \langle X_0, v \rangle \langle X_\ell, u \rangle \\ &\le \sup_{\|v\|_2=1} \mathbb{E} |\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})| \langle X_0, v \rangle^2 + \sup_{\|u\|_2=1} \mathbb{E} |\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})| \langle X_\ell, u \rangle^2 \\ &= \|\mathbb{E} X_0 X_0^T \mathbb{E}\left[|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \mid X_0 \right]\|_{op} + \|\mathbb{E} X_\ell X_\ell^T \mathbb{E}\left[\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \mid X_\ell \right]\|_{op} \\ &\overset{(i)}{\le} 2 \min\{\rho_{\text{mix}}^{|\ell|} \|\mathbb{E} X_0 X_0^T\|_{op}, \|\mathbb{E} \operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)^{1/2} X_0 X_0^T\|_{op}\} \\ &\overset{(ii)}{\le} 2 \min\{(\|\mu^*\|_2^2 + \sigma^2)\rho_{\text{mix}}^{|\ell|}, C' \mathrm{e}^{-c\eta^2}\}, \end{aligned} \tag{43}$$

where inequality (i) makes use of inequalities (42a) and (42b), and step (ii) makes use of the second inequality in line (42b).

Combining inequalities (42a) and (43), we find that

$$\begin{aligned} T_1 &\le \frac{\|\mu' - \mu^*\|_2}{\sigma^2} \sum_{\ell=-k}^{k} \|\mathbb{E} \operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) X_0 X_\ell^T\|_{op} \\ &\le 2 \frac{\|\mu' - \mu^*\|_2}{\sigma^2} \sum_{\ell=-k}^{k} \min\{(\|\mu^*\|_2^2 + \sigma^2)\rho_{\text{mix}}^{|\ell|}, C \mathrm{e}^{-c\eta^2}\} \\ &\le 4(\eta^2 + 1)\big(mC\mathrm{e}^{-c\eta^2} + \frac{\rho_{\text{mix}}^m}{1 - \rho_{\text{mix}}}\big) \|\mu' - \mu^*\|_2. \end{aligned}$$

The last inequality follows from the proof of Corollary 1 in the paper Balakrishnan et al. (2014) if $\eta^2 > C$ for some universal constant C. By choosing $m = \frac{c\eta^2}{\log(1/\rho_{\text{mix}})}$ so as to optimize the tradeoff, we have shown that

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu^*, \beta')\| \leq L_{\mu,1}\|\mu' - \mu^*\|_2,$$

where $L_{\mu,1} = c\ \varphi_1(\eta)\varphi_2(\epsilon_{\text{mix}})\eta^2 \mathrm{e}^{-c\eta^2}$ as claimed.

5.3.4 Proof of inequality (22b)

The same argument via the mean value theorem guarantees that

$$\|\frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \mu', \beta') - \frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \mu', \beta^*)\| \leq \left\| \mathbb{E}\sum_{\ell=-k}^{k} \frac{\partial^2 \Phi}{\partial\beta_\ell \partial\gamma_0}\Big|_{\theta=\widetilde{\theta}} (\beta' - \beta^*) X_0 \right\|_2 .$$

In order to bound this quantity, we again use the equivalence (41) and bound the expected conditional covariance. Furthermore, Lemma 5 and 6 yield

$$\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}) \overset{(i)}{\leq} 2\rho_{\text{mix}}^{\ell} \quad \text{and} \quad \left\|\mathbb{E}\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0) X_0 X_0^T\right\|_{op} \overset{(ii)}{\leq} c\mathrm{e}^{-c\eta^2}. \tag{44}$$

Here inequality (ii) follows by combining inequality (54c) from Lemma 5 with the fact that $\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0) \leq 1$.

$$\begin{aligned}
\|\mathbb{E}X_0 \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma})\|_2 &= \sup_{\|u\|_2=1} \mathbb{E}\langle X_0, u\rangle \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}) \\
&\leq \sup_{\|u\|_2=1} \mathbb{E}|\langle X_0, u\rangle| \mathbb{E}\big[|\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma})| \mid X_0\big] \\
&\overset{(iii)}{\leq} \sup_{\|u\|_2=1} \mathbb{E}|\langle X_0, u\rangle| \min\{\rho_{\text{mix}}^{|\ell|}, (\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0))^{1/2}\} \\
&\overset{(iv)}{\leq} \min\{\sup_{\|u\|_2=1} \sqrt{\mathbb{E}\langle X_0, u\rangle^2}\rho_{\text{mix}}^{|\ell|}, \sup_{\|u\|_2=1} \sqrt{\mathbb{E}\langle X_0, u\rangle^2 \operatorname{var}(Z_0 \mid \widetilde{\gamma}_0))}\} \\
&\overset{(v)}{\leq} \min\{\rho_{\text{mix}}^{|\ell|}\sqrt{\|\mathbb{E}X_0X_0^T\|_{op}}, \sqrt{\|\mathbb{E}\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)X_0X_0^T\|_{op}}\} \\
&\overset{(vi)}{\leq} \min\{\rho_{\text{mix}}^{|\ell|}\sqrt{\|\mu^*\|_2^2 + \sigma^2}, C\mathrm{e}^{-c\eta^2}\}
\end{aligned}$$

where step (iii) uses inequality (44); step (iv) follows from the Cauchy-Schwarz inequality; step (v) follows from the definition of the operator norm; and step (vi) uses inequality (44) again.

Putting together the pieces, we find that

$$\begin{aligned}
\left\|\mathbb{E}\sum_{\ell=-k}^{k} \frac{\partial^2\Phi}{\partial\beta_\ell\partial\gamma_0} X_0\right\|_2 |\beta' - \beta^*| &\leq \sum_{\ell=-k}^{k} \|\mathbb{E}X_0\mathbb{E}[\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}) \mid X_0]\|_2\, |\beta' - \beta^*| \\
&\leq 4\sqrt{\|\mu^*\|_2^2 + \sigma^2}\left(c\, m\, \mathrm{e}^{-c\eta^2} + \frac{\rho_{\text{mix}}^m}{1 - \rho_{\text{mix}}}\right) |\beta' - \beta^*|.
\end{aligned}$$

Setting $m = \frac{c\eta^2}{\log(1/\rho_{\text{mix}})}$ to balance the tradeoff between the two terms, we find that inequality (22b) holds with $L_{\mu,2} = c\varphi_2(\epsilon_{\text{mix}})\sqrt{\|\mu^*\|_2 + \sigma^2}\eta^2 \mathrm{e}^{-c\eta^2}$, as claimed.

5.3.5 Proof of inequality (23a)

By the same argument via the mean value theorem, we find that

$$\Big\|\frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta', \mu') - \frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta', \mu^*)\Big\| \le \Big|\mathbb{E}\sum_{\ell=-k}^{k} \frac{\partial^2 \Phi}{\partial\gamma_\ell \partial\beta_0}\Big|_{\theta=\widetilde{\theta}} \frac{\langle \mu' - \mu^*, X_\ell\rangle}{\sigma^2}\Big|.$$

Equation (41) guarantees that $\frac{\partial^2\Phi}{\partial\gamma_\ell\partial\beta_0} = \operatorname{cov}(Z_0Z_1, Z_\ell \mid \gamma)$. Therefore, by similar arguments as in the proof of inequalities (22), we have

$$\begin{aligned}
T := &\Big|\sum_{\ell=-k}^{k} \mathbb{E}\langle \mu' - \mu^*, X_\ell\rangle \mathbb{E}[\operatorname{cov}(Z_0Z_1, Z_\ell \mid \widetilde{\gamma}_\ell, \beta')|X_\ell]\Big| \\
&\le \Big|\sum_{\ell=-k}^{k} \mathbb{E}|\langle \mu' - \mu^*, X_\ell\rangle| \min\{\rho_{\text{mix}}^{|\ell|}, (\operatorname{var}(Z_\ell \mid \widetilde{\gamma}_\ell, \beta'))^{1/2}\}\Big| \\
&\le \Big|\sum_{\ell=-k}^{k} \min\big\{\rho_{\text{mix}}^{|\ell|}, \sqrt{\mathbb{E}\operatorname{var}(Z_\ell \mid \widetilde{\gamma}_\ell, \beta')}\big\}\sqrt{\mathbb{E}\langle \mu' - \mu^*, X_\ell\rangle^2}\Big| \\
&\le \sqrt{\|\mu^*\|_2^2 + \sigma^2}\big(mc\,\mathrm{e}^{-c\eta^2} + 2\sum_{\ell=m+1}^{k} \rho_{\text{mix}}^{\ell}\big).
\end{aligned}$$

where we have used inequality (54b) from Lemma 6. Finally, setting $m = \frac{c\eta^2}{\log(1/\rho_{\text{mix}})}$ yields that the FOS condition holds with $L_{\beta,2} = c\sqrt{\|\mu^*\|_2 + \sigma^2}\varphi_2(\epsilon_{\text{mix}})\eta^2\mathrm{e}^{-c\eta^2}$, as claimed.

5.3.6 Proof of inequality (23b)

By the same mean value argument, we find that

$$\Big\|\frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta', \mu') - \frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta^*, \mu')\Big\| \le \Big|\mathbb{E}\sum_{\ell=-k}^{k} \frac{\partial^2 \Phi}{\partial\beta_\ell \partial\beta_0}\Big|_{\theta=\widetilde{\theta}} (\beta' - \beta^*)\Big|.$$

By the exponential family view in equality (41) it suffices to control the expected conditional covariance. Lemma 5 and 6 guarantee that

$$|\operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1} \mid X_{-k}^k, \widetilde{\gamma})| \le \rho_{\text{mix}}^{|\ell|}, \quad \text{and} \quad \mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta}) \le c\,\frac{1+b}{1-b}\,\mathrm{e}^{-c\eta^2}. \tag{45}$$

Furthermore, the Cauchy-Schwarz inequality combined with the bound (53a) from Lemma 4 yields

$$\begin{aligned}
\mathbb{E}\big|\operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma})\big| &\le \sqrt{\mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_{-k}^k, \widetilde{\beta})}\sqrt{\mathbb{E}\operatorname{var}(Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}_{-k}^k, \widetilde{\beta})} \\
&\le \sqrt{\mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta})}\sqrt{\mathbb{E}\operatorname{var}(Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}_\ell^{\ell+1}, \widetilde{\beta})} \\
&\le \mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta}).
\end{aligned} \tag{46}$$

Combining the bounds (45) and (46) yields

$$\begin{aligned}\left|\sum_{\ell=-k}^{k} \mathbb{E}\frac{\partial^2\Phi}{\partial\beta_\ell\partial\beta_0}(\beta'-\beta^*)\right| &\le \sum_{\ell=-k}^{k} \left|\mathbb{E}\operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}_{-k}^{k}, \widetilde{\beta})\right| |\beta'-\beta^*| \\ &\le \sum_{\ell=-k}^{k} \min\left\{\rho_{\text{mix}}^{|\ell|}, \mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta})\right\} |\beta'-\beta^*| \\ &\le 2\left(c\frac{1+b}{1-b}m\mathrm{e}^{-c\eta^2} + \sum_{l=m+1}^{k}\rho_{\text{mix}}^{\ell}\right)|\beta'-\beta^*| \\ &\le 2c\frac{1+b}{1-b}\varphi_2(\epsilon_{\text{mix}})\eta^2\mathrm{e}^{-c\eta^2}|\beta'-\beta^*|\end{aligned}$$

where the final inequality follows by setting $m = \frac{c\eta^2}{\log(1/\rho_{\text{mix}})}$. Therefore, the FOS condition holds with $L_{\beta,1} = c\frac{1-b}{1+b}\varphi_2(\epsilon_{\text{mix}})\eta^2\mathrm{e}^{-c\eta^2}$, as claimed.

5.4 Proof of Corollary 2

In order to prove this corollary, it is again convenient to separate the updates on the mean vectors μ from those applied to the transition parameter β. Recall the definitions of φ, φ_n and ϵ_n from equations (25) and (28a) respectively.

Using Theorem 2 we readily have that given any initialization $\widehat{\theta}^0 \in \Omega$, with probability at least $1-2\delta$, we are guaranteed that

$$\|\widehat{\theta}^T - \theta^*\|_\star \le \kappa^T\|\widehat{\theta}^0 - \theta^*\|_\star + \frac{2\varphi_n(\delta,k) + \epsilon_n(\delta,k) + \varphi(k)}{1-\kappa}. \tag{47}$$

In order to leverage the bound (47), we need to find appropriate upper bounds on the quantities $\varphi_n(\delta,k)$, $\epsilon_n(\delta,k)$.

Lemma 1 *Suppose that the truncation level satisifes the lower bound*

$$k \ge \log\left(\frac{C_\epsilon n}{\delta}\right)\left(\log\frac{1}{(1-\frac{\epsilon_{\text{mix}}}{2})}\right)^{-1} \qquad \text{where } C_\epsilon := \frac{36}{\epsilon_{\text{mix}}^3}. \tag{48a}$$

Then with the radius $r = \frac{\|\mu^*\|_2}{4}$, we have

$$\epsilon_n^\mu(\delta,k) \le C_0\|\mu^*\|_2\big(\frac{\|\mu^*\|_2^2}{\sigma^2}+1\big)\log(k^2/\delta)\sqrt{\frac{k^3 d\log n}{n}}, \quad \text{and} \tag{48b}$$

$$\epsilon_n^\beta(\delta,k) \le C_0\,\frac{\sqrt{\|\mu^*\|_2^2+\sigma^2}}{\sigma^2}\sqrt{\frac{k^3\log(k^2/\delta)}{n}}. \tag{48c}$$

Lemma 2 *Suppose that the SNR $\eta^2 = \frac{\|\mu^*\|_2^2}{\sigma^2}$ is lower bounded as $\eta^2 \ge C(\log\log d + \log\|\mu^*\|_2)$ for a sufficiently large C. Then with the radius $r = \frac{\|\mu^*\|_2}{4}$, we have*

$$\varphi_n^2(\delta,k) \le C_1\left\{\frac{1}{\sigma}\sqrt{\frac{d\log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2}{\sigma}\sqrt{\frac{\log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2^2}{\sigma^2}\right\}\varphi^2(k). \tag{49}$$

See Appendices C.1 and C.2, respectively, for the proofs of these two lemmas.

Using these two lemmas, we can now complete the proof of the corollary. From the definition (32) of κ, under the stated lower bound on η^2, we can ensure that $\kappa\{\frac{4\log\frac{1+b}{1-b}}{\|\mu^*\|_2}\} \le 1/2$. Under this condition, inequality (29a) with $r = \|\mu^*\|_2/4$ reduces to showing that

$$\varphi_n(\delta, k) + \epsilon_n^\mu(\delta, k) + \varphi(k) \le (1 - 2\kappa)\frac{\|\mu^*\|_2}{8}. \tag{50}$$

As long as $n \ge C_1(\sigma^2 + \|\mu^*\|_2^2)d\log^2(d/\delta)$ for a sufficiently large C_1, we are guaranteed that the bound (50) holds. Now the specific choice (48a) of k guarantees that

$$\varphi_n(\delta, k) + \epsilon_n^\mu(\delta, k) + \epsilon_n^\beta(\delta, k) + \varphi(k) \le C\|\mu^*\|_2(\frac{\|\mu^*\|_2^2}{\sigma^2} + 1)\log^2 n\sqrt{\frac{d\log^2(n/\delta)}{n}}. \tag{51}$$

Substituting the bound (51) into inequality (47) completes the proof of the corollary.

6. Discussion

In this paper, we provided general global convergence guarantees for the Baum-Welch algorithm as well as specific results for a hidden Markov mixture of two isotropic Gaussians. In contrast to the classical perspective of focusing on the MLE, we focused on bounding the distance between the Baum-Welch iterates and the true parameter. Under suitable regularity conditions, our theory guarantees that the iterates converge to an e_n-ball of the true parameter, where e_n represents a form of statistical error. It is important to note that our theory does not guarnatee convergence to the MLE itself, but rather to this ball that contains both the MLE and the true parameter. When applied to the Gaussian mixture HMM, we proved that the Baum-Welch algorithm achieves estimation error that is minimax optimal up to logarithmic factors. To the best of our knowledge, these are the first rigorous guarantees for the Baum-Welch algorithm that allow for a large initialization radius.

Acknowledgments

This work was partially supported by NSF grant CIF-31712-23800, ONR-MURI grant DOD-002888, and AFOSR grant FA9550-14-1-0016 to MJW.

Appendix A. Proof of Proposition 1

In order to show that the limit $\lim_{n\to\infty} \mathbb{E}Q_n(\theta \mid \theta')$ exists, it suffices to show that the sequence of functions $\{\mathbb{E}Q_1, \mathbb{E}Q_2, \dots, \mathbb{E}Q_n\}$ is Cauchy in the sup-norm (as defined previously in equation (16)). In particular, it suffices to show that for every $\epsilon > 0$ there is a positive integer $N(\epsilon)$ such that for $m, n \geq N(\epsilon)$,

$$\|\mathbb{E}Q_m - \mathbb{E}Q_n\|_\infty \leq \epsilon.$$

In order to do so, we make use of the previously stated bound (18) relating $\mathbb{E}Q_n$ to $\bar{Q}^k$. Taking this bound as given for the moment, an application of the triangle inequality yields

$$\|\mathbb{E}Q_m - \mathbb{E}Q_n\|_\infty \leq \|\mathbb{E}Q_m - \bar{Q}^k\|_\infty + \|\mathbb{E}Q_n - \bar{Q}^k\|_\infty \ \leq \ \epsilon,$$

the final inequality follows as long as we choose $N(\epsilon)$ and k large enough (roughly proportional to $\log(1/\epsilon)$).

It remains to prove the claim (18). In order to do so, we require an auxiliary lemma:

Lemma 3 (Approximation by truncation) *For a Markov chain satisfying the mixing condition* (3)*, we have*

$$\sup_{\theta' \in \Omega} \sup_x \sum_{z_i} |p(z_i \mid x_1^n; \theta') - p(z_i \mid x_{i-k}^{i+k}; \theta')| \leq \frac{Cs^4}{\epsilon_{\text{mix}}^8 \bar{\pi}_{\min}} \big(1 - \epsilon_{\text{mix}} \bar{\pi}_{\min}\big)^k \tag{52}$$

for all $i \in [0, n]$, where $\bar{\pi}_{\min} = \min_{j \in [s], \beta \in \Omega_\beta} \bar{\pi}(j; \beta)$.

See Appendix D.2 for the proof of this lemma.

Using Lemma 3, let us now prove the claim (18). Introducing the shorthand notation

$$h(X_i, z_i, \theta, \theta') := \log p(X_i \mid z_i; \theta) + \sum_{z_{i-1}} p(z_i \mid z_{i-1}; \theta') \log p(z_i | z_{i-1}, \theta),$$

we can verify by applying Lemma 3 that

$$\begin{aligned}
\|\mathbb{E}Q_n - \bar{Q}^k\|_\infty &= \Big| \sup_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n \sum_{z_i} \mathbb{E}(p(z_i \mid X_1^n, \theta') - p(z_i \mid X_{i-k}^{i+k}, \theta')) h(X_i, z_i, \theta, \theta') \Big| \\
&\quad + \Big| \frac{1}{n} \sup_{\theta, \theta'} \mathbb{E} \sum_{z_0} p(z_0 \mid X_1^n, \theta') \log p(z_0; \theta) \Big| \\
&\leq \sup_{\theta, \theta'} \sup_x \sum_{z_i} \big| p(z_i \mid x_1^n, \theta') - p(z_i \mid x_{i-k}^{i+k}, \theta') \big| \left[\mathbb{E} \frac{1}{n} \sum_{i=1}^n \max_{z_i \in [m]} |h(X_i, z_i, \theta, \theta')| \right] + \frac{1}{n} \log \bar{\pi}_{\min}^{-1} \\
&\leq \frac{cs^4}{\epsilon_{\text{mix}}^8 \bar{\pi}_{\min}} \big(1 - \epsilon_{\text{mix}} \bar{\pi}_{\min}\big)^k \left(\mathbb{E} \max_{z_i \in [m]} \big| \log p(X_i \mid z_i, \theta) \big| + s \log \bar{\pi}_{\min}^{-1} \right) + \frac{1}{n} \log \bar{\pi}_{\min}^{-1} \\
&\leq \frac{c'\, s^5}{\epsilon_{\text{mix}}^8 \bar{\pi}_{\min}^2} \big(1 - \epsilon_{\text{mix}} \bar{\pi}_{\min}\big)^k + \frac{1}{n} \log \bar{\pi}_{\min}^{-1},
\end{aligned}$$

where the last inequality follows from the boundedness condition (17) on the log output densities.

Appendix B. Technical details for Corollary 1

The proof of the corollary mainly involves proving the bound (37), and then showing that the conditions in the theorem hold for the special Gaussian output HMM.

B.1 Bounds on conditional variances

In this section, we collect some auxiliary bounds on conditional covariances in hidden Markov models. These results are used in the proof of Corollary 1.

Lemma 4 *For any HMM with observed-hidden states (X_i, Z_i), we have*

$$\mathbb{E}\Big[\operatorname{var}(Z_0 Z_1 \mid X_{-k}^{k})\Big] \leq \mathbb{E}\operatorname{var}(Z_0 Z_1 \mid X_0^1) \tag{53a}$$

$$\mathbb{E}\Big[\operatorname{var}(Z_0 \mid X_{-k}^{k}) \mid X_0\Big] \leq \operatorname{var}(Z_0 \mid X_0) \tag{53b}$$

where we have omitted the dependence on the parameters.

Proof We use the law of total variance, which guarantees that $\operatorname{var} Z = \mathbb{E}\big[\operatorname{var}(Z \mid X)\big] + \operatorname{var}\mathbb{E}[Z \mid X]$. Using this decomposition, we have

$$\mathbb{E}[\operatorname{var}(Z_0 \mid X_0^1) \mid X_0] \leq \operatorname{var}(Z_0 \mid X_0)$$
$$\mathbb{E}[\operatorname{var}(Z_0 Z_1 \mid X_0^2) \mid X_0^1] \leq \operatorname{var}(Z_0 Z_1 \mid X_0^1).$$

The result then follows by induction. ■

We now show that the expected conditional variance of the hidden state (or pairs thereof) conditioned on the corresponding observation (pairs of observations) decays exponentially with the SNR.

Lemma 5 *For a 2-state Markov chain with true parameter θ^*, we have for $\mu \in \mathbb{B}_2\big(\frac{\|\mu^*\|_2}{4}; \mu^*\big)$ and $\beta \in \Omega_\beta$*

$$\Big\|\mathbb{E} X_0 X_0^T (\operatorname{var}(Z_0 \mid \gamma_0, \beta))^{1/2}\Big\|_{op} \leq c_0\, \mathrm{e}^{-c\eta^2} \tag{54a}$$

$$\mathbb{E}\operatorname{var}(Z_\ell \mid \gamma_\ell, \beta) \leq c_0\, \mathrm{e}^{-c\eta^2} \tag{54b}$$

$$\mathbb{E}\operatorname{var}(Z_0 Z_1 \mid \gamma_0^1, \beta) \leq c_0 \frac{1+b}{1-b}\, \mathrm{e}^{-c\eta^2}. \tag{54c}$$

The last lemma provides rigorous confirmation of the intuition that the covariance between any pair of hidden states should decay exponentially in their separation ℓ:

Lemma 6 *For a 2-state Markov chain with mixing coefficient ϵ_{mix} and uniform stationary distribution, we have*

$$\max\Big\{\operatorname{cov}(Z_0, Z_\ell \mid \gamma),\ \operatorname{cov}(Z_0 Z_1, Z_\ell Z_{\ell+1} \mid \gamma),\ \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \gamma)\Big\} \leq 2\rho_{\text{mix}}^{\ell} \tag{55}$$

with $\rho_{\text{mix}} = 1 - \epsilon_{\text{mix}}$ for all $\theta \in \Omega$.

Lemma 5 is proven in Section B.2 whereas Lemma 6 is a mixing result and its proof is found in Section D.3.

B.2 Proof of Lemma 5

By definition of the Gaussian HMM example, we have $\mathrm{var}(Z_i \mid \widetilde{\gamma}_i) = \frac{4}{(\mathrm{e}^{\widetilde{\gamma}_i} + \mathrm{e}^{-\widetilde{\gamma}_i})^2}$. Moreover, following the proof of Corollary 1 in the paper Balakrishnan et al. (2014), we are guaranteed that $\mathbb{E}\,\mathrm{var}(Z_i \mid \widetilde{\gamma}_i) \leq 8\mathrm{e}^{-\frac{\eta^2}{32}}$ and $\|\mathbb{E}X_i X_i^T(\mathrm{var}(Z_i|\widetilde{\gamma}_i))^{1/2}\|_{op} \leq c_0 \mathrm{e}^{-\frac{\eta^2}{32}}$, from which inequalities (54a) and (54b) follow.

We now prove inequality (54c) for $\beta \in \Omega_\beta$ and $\mu \in \mathbb{B}_2\big(\frac{\|\mu^*\|_2}{4}; \mu^*\big)$. Note that

$$\begin{aligned}
\frac{1}{4}\,\mathrm{var}(Z_0 Z_1 \mid \gamma_0^1, \beta) &= \frac{\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1} + \mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}}{\left[\mathrm{e}^{\beta}(\mathrm{e}^{\gamma_0+\gamma_1} + \mathrm{e}^{-(\gamma_0+\gamma_1)}) + \mathrm{e}^{-\beta}(\mathrm{e}^{\gamma_0-\gamma_1} + \mathrm{e}^{-(\gamma_0-\gamma_1)})\right]^2} \\
&\leq \mathrm{e}^{2|\beta|} \frac{\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1} + \mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}}{(\mathrm{e}^{\gamma_0+\gamma_1} + \mathrm{e}^{-(\gamma_0+\gamma_1)} + \mathrm{e}^{\gamma_0-\gamma_1} + \mathrm{e}^{-(\gamma_0-\gamma_1)})^2} \\
&\leq \left(\frac{1+b}{1-b}\right)\left[\frac{\mathrm{e}^{|\gamma_0|}}{\mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}} + \frac{\mathrm{e}^{|\gamma_1|}}{\mathrm{e}^{2\gamma_1} + \mathrm{e}^{2\gamma_1}}\right]
\end{aligned}$$

where γ are now random variables and we used

$$\begin{aligned}
&(\mathrm{e}^{\gamma_0+\gamma_1} + \mathrm{e}^{-(\gamma_0+\gamma_1)} + \mathrm{e}^{\gamma_0-\gamma_1} + \mathrm{e}^{-(\gamma_0-\gamma_1)})^2 \\
&\geq \mathrm{e}^{-|\gamma_0|}(\mathrm{e}^{-\gamma_0} + \mathrm{e}^{\gamma_0})(\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1}) + \mathrm{e}^{-|\gamma_1|}(\mathrm{e}^{-\gamma_1} + \mathrm{e}^{\gamma_1})(\mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}) \\
&\geq (\mathrm{e}^{-|\gamma_0|} + \mathrm{e}^{-|\gamma_1|})(\mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0})(\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1})
\end{aligned}$$

It directly follows that

$$\begin{aligned}
\frac{1}{4}\mathbb{E}\,\mathrm{var}(Z_0 Z_1 \mid \gamma_0^1, \beta) &\leq 2\left(\frac{1+b}{1-b}\right)\mathbb{E}\left[\frac{1}{\mathrm{e}^{\gamma_0} + \mathrm{e}^{-3\gamma_0}}\mathbf{1}_{\gamma_0 \geq 0} + \frac{1}{\mathrm{e}^{3\gamma_0} + \mathrm{e}^{-\gamma_0}}\mathbf{1}_{\gamma_0 \leq 0}\right] \\
&\leq 2\left(\frac{1+b}{1-b}\right)(\mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbf{1}_{\gamma_0 \geq 0}] + \mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbf{1}_{\gamma_0 \leq 0}]) \\
&\leq 4\left(\frac{1+b}{1-b}\right)\mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbf{1}_{\gamma_0 \geq 0}]
\end{aligned}$$

where the last inequality follows from symmetry of the random variables X_i. It is easy to bound

$$\mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbf{1}_{\gamma_0 \geq 0}] = \mathbb{E}\,\mathrm{e}^{-\frac{\|\mu\|_2 V_1}{\sigma^2}}\mathbf{1}_{V_1 \geq 0} \leq 2\mathrm{e}^{-\frac{\eta^2}{32}}$$

by employing a similar procedure as in the proof of Corollary 1 in Balakrishnan et al. (2014). Inequality (54c) then follows.

Appendix C. Technical details for Corollary 2

In this section we prove Lemmas 1 and 2. In order to do so, we leverage the independent blocks approach used in the analysis of dependent data (see, for instance, the papers Yu (1994); Nobel and Dembo (1993)). For future reference, we state here an auxiliary lemma that plays an important role in both proofs.

Let $X_{-\infty}^{\infty}$ be a sequence sampled from a Markov chain with mixing rate $\rho_{\text{mix}} = 1 - \epsilon_{\text{mix}}$, and let $\overline{\pi}_{\min}$ be the minimum entry of the stationary distribution. Given some functions $f_1 : \mathbb{R}^{2k} \to \mathbb{R}^d$ and $f_2 : \mathbb{R} \to \mathbb{R}^d$ respectively, our goal is to control the difference between the functions

$$g_1(X) := \frac{1}{n} \sum_{i=1}^{n} f_1(X_{i-k}^{i+k}), \qquad g_2(X) := \frac{1}{n} \sum_{i=1}^{n} f_2(X_i) \tag{56a}$$

and their expectation. Defining $m_1 := \lfloor n/4k \rfloor$ and $m_2 := \lfloor n/k \rfloor$, we say that f_1 respectively f_2 is (δ, k)-concentrated if

$$\begin{aligned} \mathbb{P}\Big[\|\frac{1}{m_1} \sum_{i=1}^{m_1} f_1(\widetilde{X}_{i;2k}) - \mathbb{E} f_1(\widetilde{X}_{1;2k})\|_2 \geq \epsilon\Big] &\leq \frac{\delta}{8k}, \\ \mathbb{P}\Big[\|\frac{1}{m_2} \sum_{i=1}^{m_2} f_2(\widetilde{X}_i) - \mathbb{E} f_2(\widetilde{X}_1)\|_2 \geq \epsilon\Big] &\leq \frac{\delta}{2k} \end{aligned} \tag{56b}$$

where $\{\widetilde{X}_{i;2k}\}_{i\in\mathbb{N}}$ are a collection of i.i.d. sequences of length $2k$ drawn from the same Markov chain and $\{\widetilde{X}_i\}_{i\in\mathbb{N}}$ a collection of i.i.d. variables drawn from the same stationary distribution. In our notation, $\{\widetilde{X}_{i;2k}\}_{i\in\mathbb{N}}$ under $\mathbb{P}$ are distributed identically distributed as $\{X_{i;2k}\}_{i\in\mathbb{N}}$ under $\mathbb{P}_0$.

Lemma 7 *Consider functions f_1, f_2 that are (δ, k)-concentrated* (56b) *for a truncation parameter $k \geq \log\left(\frac{6n}{\overline{\pi}_{\min}^2 \epsilon_{\text{mix}}^3 \delta}\right) (\log \frac{1}{1-\epsilon_{\text{mix}}})^{-1}$. Then the averaged functions g_1, g_2 from equation* (56a) *satisfy the bounds*

$$\mathbb{P}\Big[\|g_1(X) - \mathbb{E} g_1(X)\|_2 \geq \epsilon\Big] \leq \delta \quad \textit{and} \quad \mathbb{P}\Big[\|g_2(X) - \mathbb{E} g_2(X)\|_2 \geq \epsilon\Big] \leq \delta. \tag{57}$$

Proof We prove the lemma for functions of the type (f_1, g_1); the proof for the case (f_2, g_2) is very similar. In order to simplify notation, we assume throughout the proof that the effective sample size n is a multiple of $4k$, so that the block size $m = \frac{n}{4k}$ is integral. By definition (56a), the function g is a function of the sequences $\{X_{1-k}^{1+k}, X_{2-k}^{2+k}, \ldots, X_{n-k}^{n+k}\}$. We begin by dividing these sequences into blocks. Let us define the subsets of indices

$$\begin{aligned} H_i^j &= \{4k(i-1) + k + j \mid 4k(i-1) + 3k + j\}, \quad \text{and} \\ R_i^j &= \{4k(i-1) - k + j \mid 4k(i-1) + k - 1 + j\}. \end{aligned}$$

With this notation, we have the decomposition

$$g(X) = \frac{1}{2}\left(\frac{1}{2k} \sum_{j=1}^{2k} \underbrace{\frac{1}{m} \sum_{i=1}^{m} f(X_{H_i^j})}_{g^{H^j}(X)} + \frac{1}{2k} \sum_{j=1}^{2k} \underbrace{\frac{1}{m} \sum_{i=1}^{m} f(X_{R_i^j})}_{g^{R^j}(X)} \right),$$

from which we find that

$$
\begin{aligned}
\mathbb{P}\big[\|g(X) - \mathbb{E}g(X)\|_2 \leq \epsilon\big] &\geq \mathbb{P}\Big(\bigcap_{j=1}^{2k}\{\|g^{H^j}(X) - \mathbb{E}g(X)\|_2 \leq \epsilon\} \cap \{\|g^{R^j}(X) - \mathbb{E}g(X)\|_2 \leq \epsilon\}\Big) \\
&\overset{\text{(i)}}{\geq} 1 - 4k\mathbb{P}(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon),
\end{aligned}
$$

where (i) follows using stationarity of the underlying sequence combined with the union bound.

In order to bound the probability $\mathbb{P}\big[\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon\big]$, it is convenient to relate it to the probability of the same event under the product measure $\mathbb{P}_0$ on the blocks $\{H_1^1, \ldots, H_m^1\}$. In particular, we have $\mathbb{P}(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon) \leq T_1 + T_2$, where

$$
\begin{aligned}
T_1 &:= \mathbb{P}_0(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon), \quad \text{and} \\
T_2 &:= |\mathbb{P}(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon) - \mathbb{P}_0(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon)|.
\end{aligned}
$$

By our assumed concentration (56b), we have $T_1 \leq \frac{\delta}{8k}$, and so it remains to show that $T_2 \leq \frac{\delta}{8k}$.

Now following standard arguments (e.g., see the papers Nobel and Dembo (1993); Yu (1994)), we first define

$$
\beta(k) = \sup_{A \in \sigma(\mathcal{S}_{-\infty:0}, \mathcal{S}_{k:\infty})} |\mathbb{P}(A) - \mathbb{P}_{-\infty}^0 \times \mathbb{P}_1^\infty(A)|, \tag{58}
$$

where $\mathcal{S}_{-\infty:0}$ and $\mathcal{S}_{k:\infty}$ are the σ-algebras generated by the random vector $X_{-\infty:0}$ and $X_{k:\infty}$ respectively, and $\mathbb{P}_{-\infty}^0 \times \mathbb{P}_1^\infty$ is the product measure under which the sequences $X_{-\infty:0}$ and $X_{1:\infty}$ are independent. Define $\mathcal{S}_i$ to be the σ-algebra generated by $X_{H_i^j}$ for $i = \{1, \ldots, m\}$; it then follows by induction that $\sup_{A \in \sigma(\mathcal{S}_1, \ldots, \mathcal{S}_m)} |\mathbb{P}(A) - \mathbb{P}_0(A)| \leq m\beta(k)$. An identical relation holds over the blocks R_i^j.

For our two-state HMM, Lemma 12 implies that

$$
\begin{aligned}
\beta(k) = |p(x) - p(x_{k:\infty})p(x_{-\infty}^0)| &\leq |p(x_{-\infty}^0 | x_k^n) - p(x_{-\infty}^0)| \\
&\leq |p(z_0 | x_k^n) - p(z_0)| \\
&\overset{\text{(i)}}{\leq} \frac{3}{\overline{\pi}_{\min}^2 \epsilon_{\mathrm{mix}}^3} \rho_{\mathrm{mix}}^k = \frac{3}{\overline{\pi}_{\min}^2 \epsilon_{\mathrm{mix}}^3} \mathrm{e}^{-k\log(1/\rho_{\mathrm{mix}})},
\end{aligned} \tag{59}
$$

where step (i) follows from inequality (72b). From our assumed lower bound on k, we conclude that $m\beta(k) \leq \frac{\delta}{8k}$, which completes the proof. ■

In the following sections we apply it in order to prove the bounds on the approximation and sample error of the M-operators.

C.1 Proof of Lemma 1

We prove each of the two inequalities in equations (48b) and (48c) in turn. With suitable choices of the function f in Lemma 7, we can prove the upper bounds stated in equations (48b) and (48c).

Proof of inequality (48b)**:** We use the notation from the proof of Lemma 7 and furthermore define the weights $w_\theta(X_{i-k}^{i+k-1}) = p(Z_0 = 1 \mid X_{i-k}^{i+k-1}, \theta)$, as well as the function $f_0(X_{i-k}^{i+k-1}) = (2w_\theta(X_{i-k}^{i+k-1}) - 1)X_0$. It is then possible to write the sample splitting EM operator explicitly as the average

$$M_n^{\mu,k}(\theta) = \arg\max_{\mu} \frac{1}{n}\Big[\sum_{i=1}^{n} \mathbb{E}_{Z_i \mid X_{i-k}^{i+k}, \theta'} \log p(x_i \mid Z_i, \mu)\Big] = \frac{1}{n}\sum_{i=1}^{n} f_0(X_{-k}^{i+k-1}).$$

We are now ready to apply Lemma 7 with the choices $f_1 = f_0$, $g_1(X) = M_n^{\mu,k}(\theta)$ and $\delta = \delta$, and $n = n$. According to Lemma 7, given that the lower bound on the truncation parameter k holds, we now need to show that f_0 is (δ, k)-concentrated, that means finding ϵ_n^μ such that

$$\mathbb{P}_0\left[\big\|\frac{1}{m}\sum_{i=1}^{m} f_0(\widetilde{X}_{i;2k}) - \mathbb{E} f_0(\widetilde{X}_{i;2k})\big\|_2 \geq \epsilon_n^\mu\right] \leq \frac{\delta}{8k},$$

where $\mathbb{P}_0$ denotes the product measure over the independent blocks and $m := m_1 = \lfloor n/4k \rfloor$.

Let X_i be the middle element of the block $\widetilde{X}_{i;2k}$ and Z_i, V_i the corresponding latent and noise variable. We can then write $X_i = Z_i + V_i$ where V_i are zero-mean Gaussian random variables with covariance matrix $\sigma^2 I$.

With a minor abuse of notation, let us use $X_{i,\ell}$ to denote ℓ^{th} element in the block $\widetilde{X}_{i;2k} = (X_{i,1}, \ldots, X_{i,2k})^T$, and write $\widetilde{X} = \{\widetilde{X}_{i;2k}\}_{i=1}^{n}$. In view of Lemma 7, our objective is to find the smallest scalar ϵ_n^μ such that

$$\mathbb{P}\big[\sup_{\theta\in\Omega} \|\frac{1}{m}\sum_{i=1}^{m} \underbrace{(2w_\theta(\widetilde{X}_{i;2k}) - 1)X_{i,k} - \mathbb{E}(2w_\theta(\widetilde{X}_{i;2k}) - 1)X_{i,k}}_{f_\theta(\widetilde{X}_{i;2k})}\|_2 \geq \epsilon_n^\mu\big] \leq \frac{\delta}{8k} \tag{60}$$

For each unit norm vector $u \in \mathbb{R}^d$, define the random variable

$$\tilde{V}_m(\widetilde{X}; u) = \sup_{\theta\in\Omega} \frac{1}{m}\sum_{i=1}^{m} (2w_\theta(\widetilde{X}_{i;2k}) - 1)\langle X_{i,k}, u\rangle - \mathbb{E}(2w_\theta(\widetilde{X}_{i;2k}) - 1)\langle X_{i,k}, u\rangle.$$

Let $\{u^{(1)}, \ldots, u^{(T)}\}$ denote a 1/2-cover of the unit sphere in $\mathbb{R}^d$; by standard arguments, we can find such a set with cardinality $\log T \leq d \log 5$. Using this covering, we have

$$\sup_{\theta\in\Omega} \|\frac{1}{m}\sum_{i=1}^{m} f_\theta(\widetilde{X}_{i;2k})\|_2 = \sup_{\|u\|_2 \leq 1} \tilde{V}_m(\widetilde{X}; u) \leq 2 \max_{j\in[T]} \tilde{V}_m(\widetilde{X}; u^{(j)}),$$

where the inequality follows by a discretization argument. Consequently, we have

$$\begin{aligned}\mathbb{P}\big[\sup_{\theta\in\Omega} \|\frac{1}{m}\sum_{i=1}^{m} f_\theta(\widetilde{X}_{i;2k})\|_2 \geq \epsilon_n^\mu\big] &\leq \mathbb{P}\big[\max_{j\in[T]} \tilde{V}_m(\widetilde{X}; u^{(j)}) \geq \frac{\epsilon_n^\mu}{2}\big] \\ &\leq T \max_{j\in[T]} \mathbb{P}\big[\tilde{V}_m(\widetilde{X}; u^{(j)}) \geq \frac{\epsilon_n^\mu}{2}\big].\end{aligned}$$

The remainder of our analysis focuses on bounding the tail probability for a fixed unit vector u, in particular ensuring an exponent small enough to cancel the $T \leq e^{d \log 5}$ pre-factor. By Lemma 2.3.7 of van der Vaart and Wellner Van Der Vaart and Wellner (1996), for any $t > 0$, we have

$$\mathbb{P}_X\big[\tilde{V}_m(\widetilde{X}; u) \geq t\big] \leq c\mathbb{P}_{X,\epsilon}\big[V_m(\widetilde{X}; u) \geq \frac{t}{4}\big],$$

where $V_m(\widetilde{X}; u) = \sup_{\theta \in \Omega} \big|\frac{1}{m}\sum_{i=1}^m \epsilon_i(2w_\theta(\widetilde{X}_{i;2k}) - 1)\langle X_{i,k}, u\rangle\big|$, and $\{\epsilon_i\}_{i=1}^m$ is a sequence of i.i.d. Rademacher variables.

We now require a sequence of technical lemmas; see Section C.3 for their proofs. Our first lemma shows that the variable $V_m(\widetilde{X}; u)$, viewed as a function of the Rademacher sequence, is concentrated:

Lemma 8 *For each fixed $(\widetilde{X}, u)$, we have*

$$\mathbb{P}_\epsilon\big[V_m(\widetilde{X}; u) \geq \mathbb{E}_\epsilon V_m(\widetilde{X}; u) + t\big] \leq 2\mathrm{e}^{-\frac{t^2}{16L_m^2(\widetilde{X};u)}}, \tag{61}$$

where $L_m(\widetilde{X}; u) = \frac{1}{m}\sqrt{\sum_{i=1}^m \langle X_{i,k}, u\rangle^2}$.

Our next lemma bounds the expectation with respect to the Rademacher random vector:

Lemma 9 *There exists a universal constant c such that for each fixed $(\widetilde{X}; u)$, we have*

$$\mathbb{E}_\epsilon V_m(\widetilde{X}; u) \leq \underbrace{c\frac{\|\mu^*\|_2}{\sigma^2}\sqrt{\log m}\,\Big[\sum_{\ell=1}^{2k}\mathbb{E}_{\tilde{\epsilon}}\|\frac{1}{m}\sum_{i=1}^m \tilde{\epsilon}_{i,\ell}X_{i,\ell}\langle X_{i,k}, u\rangle\|_2\Big]}_{M_m(\widetilde{X};u)} + \underbrace{\mathbb{E}_g\big|\frac{1}{m}\sum_{i=1}^m g_{i,2k+1}\langle X_{i,k}, u\rangle\big|}_{N_m(\widetilde{X};u)} \tag{62}$$

where $\epsilon, \tilde{\epsilon} \in \mathbb{R}^m$ are random vectors with i.i.d. Rademacher components, and g is a random vector with i.i.d. $N(0,1)$ components.

We now bound the three quantities $L_m(\widetilde{X}; u)$, $M_m(\widetilde{X}; u)$, and $N_m(\widetilde{X}; u)$ appearing in the previous two lemmas. In particular, let us introduce the quantities $L' = cL\|\mu^*\|_2\big(\frac{\|\mu^*\|_2^2}{\sigma^2} + 1\big)$, $L'' = L\sqrt{\|\mu^*\|_2^2 + \sigma^2}$ and $L = \frac{\sqrt{8}}{1-\rho_{\text{mix}}}$.

Lemma 10 *Define the event*

$$\mathcal{E} = \Bigg\{L_m(\widetilde{X}; u) \leq \sqrt{\frac{2(\|\mu^*\|_2^2 + \sigma^2)\log\frac{1}{\delta}}{m}}, \quad M_m(\widetilde{X}; u) \leq L'k\sqrt{\frac{d\log m \log\frac{k}{\delta}}{m}}$$
$$\text{and } N_m(\widetilde{X}; u) \leq cL''\sqrt{\frac{d\log\frac{1}{\delta}}{m}}\Bigg\}.$$

Then we have $\mathbb{P}\big[\mathcal{E}\big] \geq 1 - \delta\mathrm{e}^{-c'\,d}$ for a universal constant $c' > 0$.

In conjunction, Lemmas 8 and 9 imply that conditionally on the event $\mathcal{E}$, we have

$$\mathbb{E}_\epsilon\big[V_m(\widetilde{X};u)\big] \le c\|\mu^*\|_2(\frac{\|\mu^*\|_2^2}{\sigma^2}+1)k\sqrt{\frac{d\log m\log\frac{k}{\delta}}{m}}.$$

Combining this bound with Lemma 10 yields

$$\begin{aligned} T\,\mathbb{P}_{\widetilde{X},\epsilon}\big[\tilde{V}_m(\widetilde{X};u)\ge t\big] &\le T\,\mathbb{P}_{X,\epsilon}\big[V_m(\widetilde{X};u)\ge \frac{t}{4}\mid\mathcal{E}\big] + T\,\mathbb{P}\big[\mathcal{E}^c\big] \\ &\le 2\mathrm{e}^{4d+\log k - ck^2 d\log m\log\frac{k}{\delta}} + \mathrm{e}^{4d-\tilde{c}d\log\frac{k}{\delta}} \\ &\le \delta, \end{aligned}$$

where the final inequality follows by setting $t/4 = c\|\mu^*\|_2(\frac{\|\mu^*\|_2^2}{\sigma^2}+1)k\log(\frac{k}{\delta})\sqrt{\frac{d\log m}{m}}$. After rescaling δ by $8k$ and setting $m = \frac{n}{4k}$, the result follows after an application of Lemma 7.

Proof of inequality (48c)**:** In order to bound $|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)|$, we need a few extra steps. First, let us define new weights

$$v_\theta(X_{i-k}^{i+k-1}) = p(Z_0 = Z_1 = 1 \mid X_{i-k}^{i+k-1},\theta) + p(Z_0 = Z_1 = -1 \mid X_{i-k}^{i+k-1},\theta),$$

and also write the update in the form

$$\begin{aligned} M_{\zeta,n}^k(\theta) &= \arg\max_{\zeta\in\Omega_\zeta}\Big\{\mathbb{E}_{Z_1|X_{i-k}^{i+k},\theta}\log p(Z_1\mid\zeta) + \sum_{i=2}^n \mathbb{E}_{Z_{i-1}^i|X_{i-k}^{i+k},\theta}\log p(Z_i\mid Z_{i-1},\zeta)\Big\} \\ &= \arg\max_{\zeta\in\Omega_\zeta}\Big\{\frac{1}{2} + \sum_{i=2}^n \mathbb{E}_{Z_{i-1}^i|X_{i-k}^{i+k},\theta}\log p(Z_i\mid Z_{i-1},\zeta)\Big\} \\ &= \Pi_{\Omega_\zeta}\Big(\frac{1}{n}\sum_{i=2}^n v_\theta(X_{i-k}^{i+k-1})\Big), \end{aligned}$$

where we have reparameterized the transition probabilities with ζ via the equivalences $\beta = h(\zeta) := \frac{1}{2}\log\left(\frac{\zeta}{1-\zeta}\right)$. Note that the original EM operator is obtained via the transformation $M_n^{\beta,k}(\theta') = h(M_{\zeta,n}^k(\theta'))$ and we have $\overline{M}_{\zeta,n}^k(\theta) = \Pi_{\Omega_\zeta}\mathbb{E}v_\theta(X_{i-k}^{i+k-1})$ by definition.

Given this set-up, we can now pursue an argument similar to that of inequality (48b). The new weights remain Lipschitz with the same constant—that is, we have the bound $|v_\theta(\widetilde{X}_{i;2k}) - v_{\theta'}(\widetilde{X}_{i;2k})| \le L\|\tilde{\theta}_i - \tilde{\theta}_i'\|_2$. As a consequence, we can write

$$\mathbb{P}\big[\sup_{\theta\in\Omega}|\frac{1}{m}\sum_{i=1}^m v_\theta(\widetilde{X}_{i;2k}) - \mathbb{E}v_\theta(\widetilde{X}_{i;2k})| \ge \epsilon_n^\beta\big] \le \frac{\delta}{8k},$$

with ϵ_n^β defined as in the lemma statement. In this case, it is not necessary to perform the covering step, nor to introduce extra Rademacher variables after the Gaussian comparison step; consequently, the two constants ϵ_n^β and ϵ_n^μ differ by a factor of $\sqrt{d\log n}$ modulo constants.

Applying Lemma 7 then yields a tail bound for the quantity $\big|\frac{1}{n}\sum_{i=1}^n v_\theta(\widetilde{X}_{i;2k}) - \mathbb{E}v_\theta(\widetilde{X}_{i;2k})\big|$ with probability δ. Since projection onto a convex set only decreases the distance, we find that

$$\mathbb{P}\left[|M_{\zeta,n}^k(\theta) - \overline{M}_{\zeta,n}^k(\theta)| \geq C\frac{\sqrt{\|\mu^*\|_2^2+\sigma^2}}{\sigma^2}\sqrt{\frac{k^3\log(k^2/\delta)}{n}}\right] \leq \delta.$$

In order to prove the result, the last step needed is the fact that

$$\frac{1}{2}\Big|\log\frac{x}{1-x} - \log\frac{y}{1-y}\Big| \leq \frac{1}{\tilde{x}(1-\tilde{x})}|x-y| \leq \frac{2}{1-b^2}|x-y| =: L|x-y|$$

for $x, y, \tilde{x} \in \Omega_\zeta$. Since $M_{\zeta,n}^k(\theta) \in \Omega_\zeta$ we finally arrive at

$$\mathbb{P}\Big[|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)| \geq C(1-b^2)\frac{\sqrt{\|\mu^*\|_2^2+\sigma^2}}{\sigma^2}\sqrt{\frac{k^3\log\left(\frac{k^2}{\delta}\right)}{n}}\Big] \leq \delta$$

and the proof is complete.

C.2 Proof of Lemma 2

Since $\|\theta\|_2 \leq \|\theta\|_\star \leq \sqrt{2}\|\theta\|_2$, it is sufficient to show that

$$\sup_{\theta\in\mathbb{B}_2(r;\theta^*)} \mathbb{P}\Big[\|M_n(\theta) - M_n^k(\theta)\|_2^2 \geq c_1\xi(n,\delta)\varphi^2(k)\Big] \leq \delta$$

with $\varphi^2(k) := \frac{c_0 s^5}{\lambda\epsilon_{\text{mix}}^8\overline{\pi}_{\min}^2}(1-\epsilon_{\text{mix}}\overline{\pi}_{\min})^k$ We first claim that

$$\|M_n(\theta) - M_n^k(\theta)\|_2^2 \leq \frac{2\|Q_n - Q_n^k\|_\infty}{\lambda}, \quad \text{where} \quad \lambda := 4\big(\exp(\beta_B) + \exp(-\beta_B)\big)^{-2}. \tag{63}$$

In Section 5.3.1. we showed that population operators are strongly concave with parameter at least λ. We make the added observation that using our parameterization, the sample Q functions $Q_n^k(\cdot \mid \theta'), Q_n(\cdot \mid \theta')$ are also strongly concave. This is because the concavity results for the population operators did not use any property of the covariates in the HMM, in particular not the expectation operator, and the single term $\frac{1}{n}\mathbb{E}\sum_{z_0} p(z_0 \mid X_1^n, \beta')\log p(z_0;\beta) = \frac{1}{n}\log\frac{1}{2}$ is constant for all $\beta \in \Omega_\beta$. From this λ-strong concavity, the bound (63) follows immediately.

Given the bound (63), the remainder of the proof focuses on bounding the difference $\|Q_n - Q_n^k\|_\infty$. Recalling the shorthand notation

$$h(X_i, z_i, \theta, \theta') = \log p(X_i|z_i,\theta) + \sum_{z_{i-1}} p(z_i|z_{i-1},\theta')\log p(z_i|z_{i-1},\theta),$$

we have

$$\begin{aligned}
\|Q_n - Q_n^k\|_\infty &= \Big|\sup_{\theta,\theta'\in\Omega}\frac{1}{n}\sum_{i=1}^n\sum_{z_i}(p(z_i|X_1^n,\theta') - p(z_i|X_{i-k}^{i+k},\theta'))h(X_i,z_i,\theta,\theta')\Big| \qquad (64)\\
&\leq \sup_{\theta,\theta'\in\Omega}\sup_X\sum_{z_i}\big|p(z_i \mid X_1^n,\theta') - p(z_i \mid X_{i-k}^{i+k},\theta')\big| \ \left[\frac{1}{n}\sum_{i=1}^n\max_{z_i\in[m]}\big|h(X_i,z_i,\theta,\theta')\big|\right].
\end{aligned}$$

It is easy to verify that $\max_{z_i\in[m]} |h(X_i, z_i, \theta, \theta')| \le \max_{z_i} \mid \log p(X_i|z_i, \theta) \mid + 2\log(\bar{\pi}_{\min}^{-1})$, and moreover, Lemma 3 implies that

$$\sup_{\theta,\theta'\in\Omega} \sup_{X} \sum_{z_i} \mid p(z_i \mid X_1^n, \theta') - p(z_i \mid X_{i-k}^{i+k}, \theta')| \le \frac{32C\big(1-\epsilon_{\text{mix}}\bar{\pi}_{\min}\big)^k}{\epsilon_{\text{mix}}^8}.$$

Combining these inequalities yields the upper bound

$$\|Q_n - Q_n^k\|_\infty \le \frac{32C\big(1-\epsilon_{\text{mix}}\bar{\pi}_{\min}\big)^k}{\epsilon_{\text{mix}}^8}\Big[\mathbb{E}\max_{z_i\in[m]} \mid \log p(X_i|z_i,\theta)| + e_n(X) + 2\log(\bar{\pi}_{\min}^{-1})\Big]$$

where

$$e_n(X) := \Big|\frac{1}{n}\sum_{i=1}^n \max_{z_i\in[m]}\Big|\log p(X_i \mid z_i, \theta)\Big| - \mathbb{E}\max_{z_i\in[m]}\Big|\log p(X_i \mid z_i,\theta)\Big|\Big|.$$

By assumption, we have that $\mathbb{E}\max_{z_i\in[m]}|\log p(X_i \mid z_i, \theta)|$ is bounded by an appropriately large universal constant. Putting these together, we find that

$$\|M_n(\theta) - M_n^k(\theta)\|_2^2 \le \frac{64c\big(1-\epsilon_{\text{mix}}\bar{\pi}_{\min}\big)^k}{\lambda\epsilon_{\text{mix}}^8}(1+e_n(X)).$$

In order to complete the proof, it suffices to show that

$$\mathbb{P}\Big(e_n(X) \ge c_0\Big(\frac{1}{\sigma}\sqrt{\frac{d\log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2}{\sigma}\sqrt{\frac{\log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2^2}{\sigma^2}\Big)\Big) \le \delta,$$

where c_0 is a universal constant and $C_\epsilon = \frac{36}{\epsilon_{\text{mix}}^3}$.

Observe that we have

$$\begin{aligned} e_n(X) &= \frac{1}{2n\sigma^2}\sum_{i=1}^n \big[\max\{\|X_i+\mu\|_2^2, \|X_i-\mu\|_2^2\} - \mathbb{E}\max\{\|X_i+\mu\|_2^2, \|X_i-\mu\|_2^2\}\big] \\ &= \frac{1}{2n\sigma^2}\sum_{i=1}^n \big(\|X_i\|_2^2 - \mathbb{E}\|X_i\|_2^2\big) + \frac{1}{n\sigma^2}\sum_{i=1}^n \big(|X_i^T\mu| - \mathbb{E}|X_i^T\mu|\big)\,. \end{aligned}$$

Note that we are again dealing with a dependent sequence so that we cannot use usual Hoeffding type bounds. For some $\tilde{k}$ to be chosen later on, and $m = n/\tilde{k}$ using the proof idea of Lemma 7 with $f_2(X_i) = |X_i^T\mu|$ and $f_2(X_i) = \|X_i\|_2^2$, we can write

$$\mathbb{P}(e_n(X) \ge \frac{t}{2\sigma^2}) \le \tilde{k}\Big(\underbrace{\mathbb{P}_0\Big(|\frac{1}{m}\sum_{i=1}^m \|X_i\|_2^2 - \mathbb{E}\|X_i\|_2^2| \ge \frac{t}{2}\Big)}_{T_1} + \underbrace{\mathbb{P}_0\Big(|\frac{1}{m}\sum_{i=1}^m |X_i^T\mu| - \mathbb{E}|X_i^T\mu|| \ge \frac{t}{4}\Big)}_{T_2} + m\beta(\tilde{k})\Big),$$

where $\beta(\tilde{k})$ was previously defined in equation (58). We claim that the choices

$$t := c_1\Big(\sigma\sqrt{\frac{d\log(\tilde{k}/\delta)}{m}} + \sigma\|\mu^*\|_2\sqrt{\frac{\log(\tilde{k}/\delta)}{m}} + \|\mu^*\|_2^2\Big), \quad \text{and} \quad \tilde{k} := \frac{C_2\log(\frac{36n}{\epsilon_{\text{mix}}^3\delta})}{\log 1/(1-\epsilon_{\text{mix}})},$$

suffice to ensure that $\mathbb{P}(e_n(X) \geq t/(2\sigma^2)) \leq \delta$. Given this claim, the lemma follows immediately since $\frac{t}{2\sigma^2} \geq \xi(n, \delta)$ for big enough constant c_0.

Accordingly, it remains to verify the sufficiency of the chosen t and $\tilde{k}$. The bound (59) implies that

$$m\beta(\tilde{k}) \leq \frac{12m\rho_{\text{mix}}^{\tilde{k}}}{\epsilon_{\text{mix}}^3} \leq \frac{\delta}{3\tilde{k}}.$$

In the sequel we develop bounds on T_1 and T_2. For T_1, observe that since $X_i \sim Z_i\mu^* + \epsilon_i$ where ϵ_i is a Gaussian vector with covariance $\sigma^2 I$ and Z_i independent under $\mathbb{P}_0$, standard χ^2 tail bounds imply that

$$\mathbb{P}_0\Big[|\frac{1}{m}\sum_{i=1}^m \|X_i\|_2^2 - \mathbb{E}\|X_i\|_2^2| \geq \frac{t}{2}\Big] \leq \frac{\delta}{3\tilde{k}}.$$

Finally, we turn our attention to the term T_2. Observe that,

$$X_i^T\mu \sim \frac{1}{2}\mathcal{N}(\mu^T\mu^*, \sigma^2\|\mu\|_2^2) + \frac{1}{2}\mathcal{N}(-\mu^T\mu^*, \sigma^2\|\mu\|_2^2),$$

so that $\mid X_i^T\mu \sim |\mathcal{N}(\mu^T\mu^*, \sigma^2\|\mu\|_2^2)|$. Denote $U_i = \mid X_i^T\mu \mid$. Letting ϵ denote a Rademacher random variable, observe that

$$\mathbb{E}\exp(tU_i) \overset{\text{(i)}}{\leq} \mathbb{E}\exp(2t\epsilon U_i) \overset{\text{(ii)}}{\leq} \exp\big(2t^2\sigma^2\|\mu\|_2^2 + 2t\mu^T\mu^*\big),$$

where (i) follows using symmetrization, and (ii) follows since the random variable ϵU_i is a Gaussian mixture. Observe that

$$\mathbb{E}U_i \overset{\text{(iii)}}{\leq} |\mu^T\mu^*| + \sigma\|\mu\|_2 \overset{\text{(iv)}}{\leq} \underbrace{2(\sigma + \|\mu^*\|_2)\|\mu^*\|_2}_{M},$$

where we have used for (iii) that U_i is a folded normal, and for (iv) that $\|\mu - \mu^*\|_2 \leq \frac{\|\mu^*\|_2}{4}$. Setting $D := 4\sigma\|\mu^*\|_2\sqrt{\frac{\log(6\tilde{k}/\delta)}{m}}$ observe that $\frac{t}{4} \geq 2M + D$ for big enough c_1. Thus, applying the Chernoff bound yields

$$\begin{aligned} T_2 \leq \mathbb{P}_0\Big[|\frac{1}{m}\sum_{i=1}^m U_i - \mathbb{E}U_i| \geq 2M + D\Big] &\leq \mathbb{P}_0\Big(\mid \frac{1}{m}\sum_{i=1}^m U_i \mid \geq M + D\Big) \\ &\leq 2\inf_{t\geq 0}\Big\{\mathbb{E}\exp\Big(\frac{t}{m}\sum_{i=1}^m U_i - Mt - Dt\Big)\Big\}, \\ &\leq 2\exp\Big(-\frac{mD^2}{8\sigma^2\|\mu\|_2^2}\Big) \leq \frac{\delta}{3\tilde{k}}. \end{aligned}$$

By combining the bounds on T_1 and T_2, some algebra shows that our choices of $t, \tilde{k}$ yield the claimed bound—namely, that $\mathbb{P}\big[e_n(X) \geq t/(2\sigma^2)\big] \leq \delta$.

C.3 Proofs of technical lemmas

In this section, we collect the proofs of various technical lemmas cited in the previous sections.

C.3.1 Proof of Lemma 8

We use the following concentration theorem (e.g., Ledoux (1997)): suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is coordinate-wise convex and L-Lipschitz with respect to the Euclidean norm. Then for any i.i.d. sequence of variables $\{X_i\}_{i=1}^n$ taking values in the interval $[a, b]$, we have

$$\mathbb{P}\big[f(X) \geq \mathbb{E}f(X) + \delta\big] \leq \mathrm{e}^{-\frac{\delta^2}{4L^2(b-a)^2}} \tag{65}$$

We consider the process without absolute values (which introduces the factor of two in the lemma) and see that $\epsilon := (\epsilon_1, \dots, \epsilon_n)$ is a random vector with bounded entries and that the supremum over affine functions is convex.
It remains to show that the function $\epsilon \mapsto V_m(\widetilde{X}, u)$ is Lipschitz with $L_m(\widetilde{X}; u)$ as follows

$$\begin{aligned}\big|\sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i f_\theta(\widetilde{X}_{i;2k}) - \sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i' f_\theta(\widetilde{X}_{i;2k})\big| &\leq \frac{1}{m}\big|\sum_{i=1}^m(\epsilon_i - \epsilon_i')f_\theta(\widetilde{X}_{i;2k})\big| \\ &\leq \frac{1}{m}\sqrt{\sum_{i=1}^m (2w_\theta(\widetilde{X}_{i;2k}) - 1)^2\langle X_{i,k}, u\rangle^2}\|\epsilon - \epsilon'\|_2 \\ &\leq L_m(\widetilde{X}; u))\|\epsilon - \epsilon'\|_2\end{aligned}$$

where $\theta = \arg\max_{\theta\in\Omega}\sum_i \epsilon_i f_\theta(\widetilde{X}_{i;2k})$ in the last line and we use that $|2w_\theta(\widetilde{X}_{i;2k}) - 1| \leq 1$.

C.3.2 Proof of Lemma 9

The proof consists of three steps. First, we observe that the Rademacher complexity is upper bounded by the Gaussian complexity. Then we use Gaussian comparison inequalities to reduce the process to a simpler one, followed by a final step to convert it back to a Rademacher process.

Relating the Gaussian and Rademacher complexity: Let $g_i \sim \mathcal{N}(0, 1)$. It is easy to see that using Jensen's inequality and the fact that $\epsilon_i|g_i| \overset{d}{=} g_i$

$$\begin{aligned}\mathbb{E}_\epsilon \sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i f_\theta(\widetilde{X}_{i;2k}) &= \sqrt{\frac{2}{\pi}}\mathbb{E}_\epsilon \sup_\theta \frac{1}{m}\sum_{i=1}^m \mathbb{E}_g[|g_i|]f_\theta(\widetilde{X}_{i;2k}) \\ &\leq \sqrt{\frac{2}{\pi}}\mathbb{E}_g \sup_\theta \frac{1}{m}\sum_{i=1}^m g_i f_\theta(\widetilde{X}_{i;2k}).\end{aligned}$$

Lipschitz continuity: Define

$$\tilde{\theta}_i := (\gamma_i, \beta) = (\frac{\langle\mu, X_{i,1}\rangle}{\sigma^2}, \dots, \frac{\langle\mu, X_{i,2k}\rangle}{\sigma^2}, \beta), \tag{66}$$

Now we can use results in the proof of Corollary 1 to see that $\tilde{\theta} \mapsto F(\theta; \widetilde{X}_{i;2k}) := f_\theta(\widetilde{X}_{i;2k})$ is Lipschitz in the Euclidean norm, i.e. there exists an L, only dependent on ρ_{mix} such that

$$|F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) - F(\tilde{\theta}'_i; \widetilde{X}_{i;2k})| \leq L \|\tilde{\theta}_i - \tilde{\theta}'_i\|_2 |\langle X_{i,k}, u\rangle| \tag{67}$$

For this we directly use results (exponential family representation) that were used to show Corollary 1. We overload notation and write $X_\ell := X_{1,\ell}$ and analyze Lipschitz continuity for the first block. First note that $F(\theta, X_{0;2k}) = \mathbb{E}_{Z_k | X_1^{2k}, \theta} Z_k$. By Taylor's theorem, we then have

$$\begin{aligned}
|F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) - F(\tilde{\theta}'_i; \widetilde{X}_{i;2k})| &= |\langle X_{i,k}, u\rangle| |\mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, \theta} Z_k - \mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, \theta'} Z_k| \\
&\leq |\langle X_{i,k}, u\rangle| |\mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu, \beta)} Z_k - \mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu', \beta)} Z_k| \\
&+ |\langle X_{i,k}, u\rangle| |\mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu', \beta)} Z_k - \mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu', \beta')} Z_k|
\end{aligned}$$

Let us examine each of the summands separately. By the Cauchy-Schwartz inequality and Lemma 6, we have

$$\begin{aligned}
|\mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu, \beta)} Z_k - \mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu', \beta)} Z_k| &= \frac{1}{\sigma} | \sum_{\ell=1}^{2k} \frac{\partial^2 \Phi}{\partial \gamma_\ell \partial \gamma_0} \Big|_{\theta = \tilde{\theta}} (\gamma_\ell - \gamma'_\ell)| \\
&= \Big| \sum_{\ell=1}^{2k} \operatorname{cov}(Z_0, Z_\ell \mid \widetilde{X}_{i;2k}, \tilde{\theta}) (\langle \mu, X_\ell\rangle - \langle \mu', X_\ell\rangle) \Big| \\
&\leq \sqrt{(\sum_{\ell=1}^{2k} 4\rho_{\text{mix}}^{2\ell}) \sum_{\ell=1}^{2k} (\gamma_\ell - \gamma'_\ell)^2},
\end{aligned}$$

as well as

$$\begin{aligned}
|\mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu', \beta)} Z_k - \mathbb{E}_{Z_k | \widetilde{X}_{i;2k}, (\mu', \beta')} Z_k| &= | \sum_{\ell=1}^{2k} \frac{\partial^2 \Phi}{\partial \beta_\ell \partial \gamma_0} \Big|_{\theta = \tilde{\theta}} (\beta - \beta')| \\
&= \Big| \sum_{\ell=1}^{2k} \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{X}_{i;2k}, \tilde{\theta}) (\beta - \beta') \Big| \\
&\leq \frac{2}{1 - \rho_{\text{mix}}} |\beta - \beta'|.
\end{aligned}$$

Combining these two bounds yields

$$|F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) - F(\tilde{\theta}'_i; \widetilde{X}_{i;2k})|^2 \leq \langle X_{i,k}, u\rangle^2 L \big(\sum_{\ell=1}^{2k} (\gamma_\ell - \gamma'_\ell)^2 + (\beta - \beta')^2 \big) = \langle X_{i,k}, u\rangle^2 L^2 \|\tilde{\theta}_i - \tilde{\theta}'_i\|_2^2$$

with $L^2 = \frac{8}{(1 - \rho_{\text{mix}})^2}$.

Applying the Sudakov-Fernique Gaussian comparison: Let us introduce the shorthands $X_\theta = \frac{1}{m}\sum_i g_i f_\theta(\widetilde{X}_{i;2k})$, and

$$Y_\theta = \frac{1}{m} L \sum_i \big(\sum_{\ell=1}^{2k} g_{i\ell}\frac{\langle \mu, X_{i,\ell}\rangle}{\sigma^2} + g_{i,2k+1}\beta\big)\langle X_{i,k}, u\rangle.$$

By construction, the random variable $X_\theta - X_\theta'$ is a zero-mean Gaussian variable with variance

$$\begin{aligned} \mathbb{E}_g(X_\theta - X_{\theta'})^2 &= \sum_i (F(\tilde\theta;\widetilde{X}_{i;2k}) - F(\tilde\theta';\widetilde{X}_{i;2k}))^2 \\ &\le L^2\sum_i\langle X_{i,k},u\rangle^2\big(\sum_{\ell=1}^{2k}(\gamma_{i,\ell}-\gamma'_{i,\ell})^2 + (\beta-\beta')^2\big)\\ &= \mathbb{E}_g(Y_\theta - Y_{\theta'})^2 \end{aligned} \tag{68}$$

By the Sudakov-Fernique comparison Ledoux and Talagrand (2013),we are then guaranteed that $\mathbb{E}\sup_\theta X_\theta \le \mathbb{E}\sup_\theta Y_\theta$. Therefore, it is sufficient to bound

$$\mathbb{E}_g \sup_{\theta\in\Omega} Y_\theta = \underbrace{\mathbb{E}_g \sup_\theta \frac{L}{\sigma^2 n}\sum_{i=1}^m\sum_{\ell=1}^{2k} g_{i\ell}\langle \mu, X_{i,\ell}\rangle\langle X_{i,k},u\rangle}_{T_1} + \underbrace{\mathbb{E}_g \sup_\theta \frac{L}{n}\sum_{i=1}^m g_{i,2k+1}\beta\langle X_{i,k},u\rangle}_{T_2}.$$

Converting back to a Rademacher process: We now convert the term T_1 back to a Rademacher process, which allows us to use sub-exponential tail bounds. We do so by re-introducing additional Rademacher variables, and then removing the term $\max_i |g_i|$ via the Ledoux-Talagrand contraction theorem Ledoux and Talagrand (2013). Given a Rademacher variable ϵ_{il} independent of g, note the distributional equivalence $\epsilon_{il} g_{il} \overset{d}{=} g_{i\ell}$. Then consider the function $\phi_{i\ell}(g_{i\ell}) := g_{i\ell}h_{i\ell}$ with $h_{i\ell} := \langle \mu, X_{i\ell}\rangle\langle X_{i,k},u\rangle$ for which it is easy to see that

$$|\phi_{i\ell}(g_{i\ell},h_{i\ell}) - \phi_{i\ell}(g_{i\ell},h'_{i\ell})| \le |g_{i\ell}||h_{i\ell}-h'_{i\ell}| \tag{69}$$

Applying Theorem 4.12. in Ledoux and Talgrand Ledoux and Talagrand (2013) yields

$$\mathbb{E}\sup_\theta \frac{1}{m}\sum_{i=1}^m\sum_{\ell=1}^{2k}\epsilon_{i\ell}g_{i\ell}\langle \mu, X_{i\ell}\rangle\langle X_{i,k},u\rangle \le \mathbb{E}_g\|g\|_\infty \mathbb{E}_\epsilon \sup_\theta \frac{1}{m}\sum_{i=1}^m\sum_{\ell=1}^{2k}\epsilon_{i\ell}\langle \mu, X_{i,\ell}\rangle\langle X_{i,k},u\rangle.$$

Putting together the pieces yields the claim (62).

C.3.3 Proof of Lemma 10

We prove that the probability of each of the events corresponding to the inequalities is smaller than $\frac{\delta}{3}\mathrm{e}^{-\tilde{c}d}$.

Bounding L_m: We start by bounding $L_m(\widetilde{X};u)$, for which we note that

$$\sum_{i=1}^n \langle X_{i,k},u\rangle^2 \le \sum_{i=1}^m \|\mu^*\|_2^2 + \sum_{i=1}^m\langle n_{i,k},u\rangle^2 + \sum_{i=1}^m \langle \mu^*,u\rangle\langle n_{i,k},u\rangle$$

implies that $L_m(\widetilde{X};u) \le \sqrt{\frac{2(\|\mu^*\|_2^2+\sigma^2)\log\frac{1}{\delta}}{m}}$ with probability at least $1-\frac{\delta}{3}\mathrm{e}^{-\tilde{c}d}$.

Bounding N_m: In order to bound $N_m(\widetilde{X};u)$, we first introduce an extra Rademacher random variable into its definition; doing so does not change its value (now defined by an expectation over both g and the Rademacher variables). We now require a result for a product of the form ϵgh where g, h are independent Gaussian random variables.

Lemma 11 *Let (ϵ, g, h) be independent random variables, with ϵ Rademacher, $g \sim \mathcal{N}(0, \sigma_g^2)$, and $h \sim \mathcal{N}(0, \sigma_h^2)$. Then the random variable $Z = \epsilon gh$ is a zero-mean sub-exponential random variable with parameters $(\frac{\sigma_g^2\sigma_h^2}{2}, \frac{1}{4})$.*

Proof Note that $g' = \alpha h$ with $\alpha = \frac{\sigma_g}{\sigma_h}$ is identically distributed as g. Therefore, we have

$$gh = \frac{1}{\alpha} g g' = \frac{1}{4\alpha}[(g-g')^2 + (g+g')^2]$$

The random variables $g - g'$ and $g + g'$ are independent and therefore $(g-g')^2, (g+g')^2$ are sub-exponential with parameters $\nu^2 = 4\sigma_g^4$, $b = \frac{1}{4}$. This directly yields

$$\mathbb{E}e^{\lambda\epsilon[(g+g')^2-(g-g')^2]} \leq e^{4\lambda^2\sigma_g^4}$$

for $|\lambda| \leq \frac{1}{b}$. Therefore $\mathbb{E}e^{\lambda\epsilon gh} \leq e^{\frac{\lambda^2\sigma_g^2\sigma_h^2}{4}}$, which shows that ϵgh is sub-exponential with parameters $(\frac{\sigma_g^2\sigma_h^2}{2}, \frac{1}{4})$. ■

Returning to the random variable $N_m(\widetilde{X};u)$, each term $\epsilon_i g_{i,2k+1}\langle X_{i,k}, u\rangle$ is a sub-exponential random variable with mean zero and parameter $\nu^2 = \|\mu^*\|_2^2 + \frac{\sigma^2}{2}$. Consequently, there are universal constants such that $N_m(\widetilde{X};u) \leq cL\nu\sqrt{\frac{d\log\frac{1}{\delta}}{m}}$ with probability at least $1 - \frac{\delta}{3}e^{-\tilde{c}d}$.

Bounding M_m: Our next claim is that with probability at least $\frac{\delta}{3}$, we have

$$\mathbb{E}_\epsilon\|\frac{1}{m}\sum_{i=1}^{m}\epsilon_{i\ell}X_{i,\ell}\langle X_{i,k}, u\rangle\|_2 \leq (\|\mu^*\|_2^2 + \sigma^2)\sqrt{\frac{d\log\frac{k}{\delta}}{m}}, \tag{70}$$

which then implies that $M_m(\widetilde{X};u) \leq c\|\mu^*\|_2\big(\frac{\|\mu^*\|_2}{\sigma^2} + 1\big)k\sqrt{\frac{d\log m\log\frac{k}{\delta}}{n}}$. In order to establish this claim, we first observe that by Lemma 11, the random variable $\epsilon_\ell\langle X_{i,\ell}, u\rangle\langle X_{i,k}, u\rangle$ is zero mean, sub-exponential with parameter at most $\nu^2 = (\|\mu^*\|_2^2 + \sigma^2)^2$. The bound then follows by the same argument used to bound the quantity N_m.

Appendix D. Mixing related results

In the following we will use the shorthand notations $w_\theta(x_{-k}^k) := p(z_0 = -1 \mid x_{-k}^k, \theta)$ and $\pi_k^\theta := p(z_0 \mid x_{-k}^0, \theta)$ which we refer to as weights and the filtering distribution respectively.

Introducing the shorthand notation $p_\mu(x_k) := \sum_{z_k}\sum_{z_{k-1}} p(x_k \mid z_k)p(z_k \mid z_{k-1})\mu(z_{k-1})$, we define the filter operator

$$F_i\nu(z_i) := \frac{\sum_{z_{i-1}} p(x_i \mid z_i)p(z_i \mid z_{i-1})\nu(z_{i-1})}{\sum_{z_i}\sum_{z_{i-1}} p(x_i \mid z_i)p(z_i \mid z_{i-1})\nu(z_{i-1})} = \sum_{z_{i-1}} \frac{p(x_i \mid z_i)p(z_i \mid z_{i-1})}{p_\nu(x_i)}\nu(z_{i-1}).$$

where the observations x are fix. Using this notation, the filtering distribution can then be rewritten in the more compact form $\pi_k^\theta = p(z_0 \mid x_{-k}^0, \theta) = F_k \dots F_1 \mu$. Similarly, we define

$$K_{j|i}(z_j \mid z_{j-1}) := \frac{p(z_j \mid z_{j-1}) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}{\sum_{z_j} p(z_j \mid z_{j-1}) p(x_j \mid z_j) p(x_{j+1}^i | z_j)}, \quad \text{and} \quad \nu_{\ell|i} := \frac{p(x_{\ell+1}^i \mid z_\ell) \nu(z_\ell)}{\sum_{z_\ell} p(x_{\ell+1}^i \mid z_\ell) \nu(z_\ell)}$$

Note that $\epsilon_{\text{mix}} C_0 \leq p(x_{\ell+1}^i \mid z_\ell) \leq \epsilon_{\text{mix}}^{-1} C_0$ where

$$C_0 = \sum_{z_i \dots z_{\ell+1}} p(x_{\ell+1} \mid z_{\ell+1}) \dots p(x_i \mid z_i) p(z_i \mid z_{i-1}) \dots p(z_{\ell+2} \mid z_{\ell+1}) \pi(z_{\ell+1})$$

and therefore

$$\sup_x \frac{\sup_z p(x_{\ell+1}^i \mid z_\ell)}{\inf_z p(x_{\ell+1}^i \mid z_\ell)} \leq \epsilon_{\text{mix}}^{-2}. \tag{71}$$

With these definitions, it can be verified (e.g., see Chapter 5 of van Handel van Handel (2008)) that $F_i \dots F_{\ell+1} \nu = \nu_{\ell+1|i}^T K_{\ell+1|i} \dots K_{i|i}$, where $\nu^T K := \int \nu(x') K(x|x') \mathrm{d}x'$. (In the discrete setting, this relation can be written as the row vector ν being right multiplied by the kernel matrix K.)

We also observe that

$$\begin{aligned} \pi_k^\theta - \pi_k^{\theta'} &= F_k^\theta \cdots F_1^\theta \mu^\theta - F_k^{\theta'} \cdots F_1^{\theta'} \mu^{\theta'} \\ &= \sum_{i=0}^{k-1} F_k^\theta \dots F_{i+1}^\theta (F_i^\theta \pi_{i-1}^{\theta'} - F_i^{\theta'} \pi_{i-1}^{\theta'}) + F_k^\theta \pi_{k-1}^{\theta'} - F_k^{\theta'} \pi_{k-1}^{\theta'}, \end{aligned}$$

where $F_0^\theta \pi_0^{\theta'} := \mu^\theta$ and $F_0^{\theta'} \pi_0^{\theta'} := \mu^{\theta'}$. For simplicity, we write $F_k := F_k^\theta$ and $F_k' := F_k^{\theta'}$.

D.1 Consequences of mixing

In this technical appendix we derive several useful consequences of the geometric mixing condition on the stochastic process Z_i.

Lemma 12 *For any geometrically ρ_{mix}-mixing and time reversible Markov chain with s states, there is a universal constant c such that*

$$\sup_x |p(z_i \mid x_{i+k}) - p(z_i)| \leq c\, \rho_{\text{mix}}^k, \quad \textit{and} \tag{72a}$$

$$\sup_{z_0} \left| p(z_0 \mid x_k^n) - p(z_0) \right| < \frac{c(s+1)}{\pi_{\min}^3 \epsilon_{\text{mix}}^3} \rho_{\text{mix}}^k. \tag{72b}$$

Proof We first prove the bound (72a). Using time reversibility and the defining of mixing we obtain

$$
\begin{aligned}
\max_x (p(z_0 \mid x_k) - \pi(z_0)) &= \sum_{z_k} (p(z_0 \mid z_k) - \overline{\pi}(z_0)) p(z_k \mid x_k) \\
&\leq \sum_{z_k} p(z_k \mid x_k) \max_{z_k} |(p(z_0 \mid z_k) - \overline{\pi}(z_0))| \\
&\leq \max_{z_k} \left| \frac{p(z_k \mid z_0)\overline{\pi}(z_0)}{\overline{\pi}(z_k)} - \frac{\overline{\pi}(z_0)\overline{\pi}(z_k)}{\overline{\pi}(z_k)} \right| \\
&\leq \frac{\overline{\pi}(z_0)}{\overline{\pi}(z_k)} \max_{z_k} |p(z_k \mid z_0) - \overline{\pi}(z_0)| \leq \frac{1}{\overline{\pi}_{\text{min}}} \rho_{\text{mix}}^k
\end{aligned}
$$

where $p(z_k \mid z_0) = P(Z_k = z_k \mid Z_0 = z_0)$ and $p(z_0 \mid z_k) = P(Z_0 = z_0 \mid Z_k = z_k)$.

Using this result we can now prove inequality (72b). By definition, we have

$$
p(z_0) = \frac{p(x_{k+1}^n \mid x_k) p(x_k) p(z_0)}{p(x_{k+1}^n, x_k)}, \quad \text{and} \quad p(z_0 \mid x_k, x_{k+1}^n) = \frac{p(x_{k+1}^n \mid x_k, z_0) p(x_k \mid z_0) p(z_0)}{p(x_{k+1}^n, x_k)}
$$

and therefore

$$
\begin{aligned}
|p(z_0) - p(z_0 \mid x_k^n)| \leq \frac{p(x_k)p(z_0)}{p(x_k^n)} &|p(x_{k+1}^n \mid x_k) - p(x_{k+1}^n \mid x_k, z_0)| \\
&+ \frac{p(x_{k+1}^n \mid x_k, z_0) p(z_0)}{p(x_{k+1}^n \mid x_k)} |p(x_k) - p(x_k \mid z_0)| \quad (73)
\end{aligned}
$$

In the following we bound each of the two differences. Note that

$$
\begin{aligned}
|p(x_{k+1}^n \mid x_k, z_0) - p(x_{k+1}^n \mid x_k)| &= \sum_{z_k} \sum_{z_{k+1}} p(x_{k+1}^n \mid z_{k+1}) p(z_{k+1} \mid z_k) |p(z_k \mid x_k, z_0) - p(z_k \mid x_k)| \\
&\leq \sup_{z_k, x_k} |p(z_k \mid x_k, z_0) - p(z_k \mid x_k)| \sum_{z_k} p(x_{k+1}^n \mid z_k) \qquad (74)
\end{aligned}
$$

The last term $\sum_{z_k} p(x_{k+1}^n \mid z_k)$ is bounded by s for s-state models. Using the bound (72a), we obtain

$$
|p(x_k \mid z_0) - p(x_k)| = \frac{|p(z_0 \mid x_k) - \overline{\pi}(z_0)| p(x_k)}{\overline{\pi}(z_0)} \leq \frac{p(x_k)}{\overline{\pi}_{\text{min}}^2} \rho_{\text{mix}}^k \qquad (75)
$$

which yields

$$
\begin{aligned}
|p(z_k \mid x_k, z_0) - p(z_k \mid x_k)| &= p(x_k \mid z_k) \left| \frac{p(z_k \mid z_0)}{p(x_k \mid z_0)} - \frac{\overline{\pi}(z_k)}{p(x_k)} \right| \\
&\leq \frac{p(x_k \mid z_k)}{p(x_k \mid z_0)} |p(z_k \mid z_0) - \overline{\pi}(z_k)| + \frac{\overline{\pi}(z_k) p(x_k \mid z_k)}{p(x_k) p(x_k \mid z_0)} |p(x_k \mid z_0) - p(x_k)| \\
&\leq \frac{p(x_k \mid z_k)}{p(x_k \mid z_0)} \left(\rho_{\text{mix}}^k + \frac{1}{\overline{\pi}_{\text{min}}^2} \rho_{\text{mix}}^k \right) \\
&\leq \frac{1}{p(z_k \mid z_0)} \frac{s+1}{\overline{\pi}_{\text{min}}} \rho_{\text{mix}}^k \leq \frac{2}{\overline{\pi}_{\text{min}}^3 \epsilon_{\text{mix}}} \rho_{\text{mix}}^k. \qquad (76)
\end{aligned}
$$

The last statement is true because one can check that for all $t \in \mathbb{N}$ we have

$$\min_{z_k, z_0} p(z_k \mid z_0) = \min_{ij} (A^t)_{ij} \geq \min_{ij} (A)_{ij} \geq \epsilon_{\text{mix}} \overline{\pi}_{\min}$$

for any stochastic matrix A which satisfies the mixing condition (3).

Substituting (74) with (76) and (75) into (73), we obtain

$$\begin{aligned}
|p(z_0) - p(z_0 \mid x_k^n)| &\leq \frac{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_0)}{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_k \mid x_k)} \frac{2}{\overline{\pi}_{\min}^2 \epsilon_{\text{mix}}} \rho_{\text{mix}}^k + \frac{p(x_{k+1}^n \mid x_k, z_0) p(z_0)}{p(x_{k+1}^n \mid x_k)} \frac{\rho_{\text{mix}}^k}{\overline{\pi}_{\min}} \\
&\leq \left(\frac{2s}{\overline{\pi}_{\min}^3 \epsilon_{\text{mix}}^3} + \frac{s}{\epsilon_{\text{mix}}^2 \overline{\pi}_{\min}} \right) \rho_{\text{mix}}^k \leq \frac{2s+1}{\overline{\pi}_{\min}^3 \epsilon_{\text{mix}}^3} \rho_{\text{mix}}^k
\end{aligned}$$

where we use (71) to see that

$$\frac{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_k \mid x_k, z_0)}{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_k \mid x_k)} \leq \frac{\max_{z_k} p(x_{k+1}^n \mid z_k)}{\min_{z_k} p(x_{k+1}^n \mid z_k)} \leq \epsilon_{\text{mix}}^{-2}$$

and similarly for the first term. ■

Lemma 13 (Filter stability) *For any mixing Markov chain which fulfills condition* (3), *the following holds*

$$\|F_i \dots F_1 (\nu - \nu')\|_\infty \leq \epsilon_{\text{mix}}^{-2} \widetilde{\rho}_{\text{mix}}^i \|\nu - \nu'\|_1$$

where $\widetilde{\rho}_{\text{mix}} = 1 - \epsilon_{\text{mix}} \overline{\pi}_{\min}$. *In particular we have*

$$\sup_{z_i} |p(z_i \mid x_i^1) - p(z_i \mid x_{-n}^i)| \leq 2 \epsilon_{\text{mix}}^{-2} \widetilde{\rho}_{\text{mix}}^i. \tag{77}$$

Proof Given the mixing assumption (3) we can show that $K_{j|i}(x|y) \geq \epsilon p_{j|i}(x)$ with $\epsilon = \epsilon_{\text{mix}} \overline{\pi}_{\min}$ for some probability distribution $p_{j|i}(\cdot)$. This is because we can lower bound

$$\begin{aligned}
K_{j|i}(z_j \mid z_{j-1}) &= \frac{p(z_j \mid z_{j-1}) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}{\sum_{z_j} p(z_j \mid z_{j-1}) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)} \\
&\geq \underbrace{\frac{\epsilon_{\text{mix}} \overline{\pi}(z_j) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}{\sum_{z_j} \frac{\overline{\pi}(z_j)}{\overline{\pi}_{\min}} p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}}_{=: \epsilon p_j(z_j)}
\end{aligned}$$

with $\epsilon = \epsilon_{\text{mix}} \overline{\pi}_{\min}$. This allows us to define the stochastic matrix

$$Q_{j|i} = \frac{1}{1-\epsilon} (K_{j|i} - \epsilon P_{j|i}) \text{ or } K_{j|i} = \epsilon P_{j|i} + (1-\epsilon) Q_{j|i}.$$

where $(P_{j|i})_{k\ell} = p_{j|i}(\ell)$. Using $\widetilde{\rho}_{\mathrm{mix}} = 1 - \epsilon$ we then obtain by induction and using inequality (71)

$$\begin{aligned}
\|(\nu_{1|i} - \nu'_{1|i})K_{1|i}\dots K_{i|i}\|_\infty &\leq \prod_{j=1}^{i}(1-\epsilon)\|(\nu_{1|i} - \nu'_{1|i}) \otimes_{j=1}^{i} Q_{j|i}\|_2 \\
&\leq \widetilde{\rho}_{\mathrm{mix}}^{i} \|\nu_{1|i} - \nu'_{1|i}\|_2 \prod_{j=1}^{i} \|Q_{j|i}^{T}\|_{op} \leq \widetilde{\rho}_{\mathrm{mix}}^{i} \|\nu_{1|i} - \nu'_{1|i}\|_2 \\
&\leq \widetilde{\rho}_{\mathrm{mix}}^{i} \left\| \frac{p(x_2^i \mid \cdot)\nu(\cdot)}{\sum_{z_1} p(x_2^i \mid z_1)\nu(z_1)} - \frac{p(x_2^i \mid z_1)\nu'(z_1)}{\sum_{z_\ell} p(x_2^i \mid \cdot)\nu'(\cdot)} \right\|_2 \\
&\leq \widetilde{\rho}_{\mathrm{mix}}^{i} \Big[\Big\| \frac{p(x_2^i \mid \cdot)}{\sum_{z_1} p(x_2^i \mid z_1)\nu(z_1)} (\nu(\cdot) - \nu'(\cdot)) \Big\|_2 \\
&+ \Big| \Big(\frac{\sup_z p(x_2^i \mid z_1)}{\sum_{z_1} p(x_2^i \mid z_1)\nu(z_1)} - \frac{\sup_z p(x_2^i \mid z_1)}{\sum_{z_1} p(x_2^i \mid z_1)\nu'(z_1)} \Big) \Big| \|\nu'(\cdot)\|_2 \Big] \\
&\leq \widetilde{\rho}_{\mathrm{mix}}^{i} \Big(\frac{\sup_z p(x_2^i \mid z_1)}{\inf_z p(x_2^i \mid z_1)} \Big)^2 \|\nu - \nu'\|_1 \leq \epsilon_{\mathrm{mix}}^{-2} \widetilde{\rho}_{\mathrm{mix}}^{i} \|\nu - \nu'\|_1,
\end{aligned}$$

since $Q_{j|i}$ are stochastic matrices and $\|\nu\|_2 \leq \|\nu\|_1$ for probability vectors. The second statement is readily derived by substituting $\nu(z_1) = p(z_1)$ and $\nu'(z_1) = p(z_1 \mid x_1^n)$. ■

D.2 Proof of Lemma 3

Recall the shorthand $\widetilde{\rho}_{\mathrm{mix}} = 1 - \epsilon_{\mathrm{mix}}\overline{\pi}_{\min}$. For the general case observe that

$$\begin{aligned}
&\sup_{z_i} |p(z_i \mid x_1^n) - p(z_i \mid x_{i-k:i+k})| \\
&\leq |p(z_i|x_{i+1}^n)p(z_i|x_1^i) - p(z_i|x_{i+1}^{i+k})p(z_i|x_{i-k}^i)| \frac{p(x_{i+1}^n)}{p(x_{i+1}^n|x_1^n)p(z_i)} \\
&+ \left| \frac{p(x_{i+k+1}^n|x_{i+1}^{i+k})p(x_1^{i-k-1}|x_{i-k}^i)}{p(x_{i+k+1}^n, x_1^{i-k-1}|x_{i-k}^{i+k})} - 1 \right| \frac{p(x_{i+1}^{i+k})}{p(x_{i-k}^{i+k}|x_{i-k}^i)} \frac{1}{p(z_i)}.
\end{aligned}$$

From Lemma 13 we directly obtain the following upper bounds

$$\begin{aligned}
\sup_{z,x} |p(z_i \mid x_1^i) - p(z_i \mid x_{i-k}^i)| &\leq \epsilon_{\mathrm{mix}}^{-2} \widetilde{\rho}_{\mathrm{mix}}^{k} \\
\sup_{z,x} |p(z_i \mid x_{i+1}^n) - p(z_i \mid x_{i+1}^{i+k})| &\leq \epsilon_{\mathrm{mix}}^{-2} \widetilde{\rho}_{\mathrm{mix}}^{k}
\end{aligned}$$

where the latter follows because of reversibility assumption (2) of the Markov chain. Inequality (71) can also be used to show that $\frac{p(x_{i+1}^n)}{p(x_{i+1}^n|x_1^i)} \leq s\epsilon_{\mathrm{mix}}^{-2}$. The first term of the sum is therefore bounded above by $2\frac{s\widetilde{\rho}_{\mathrm{mix}}^{k}}{\overline{\pi}_{\min}\epsilon_{\mathrm{mix}}^{4}}$.

For the second term, we again use inequality (71) and Lemma 13 to observe that

$$\begin{aligned}
&\sup_x \frac{|p(x_{i+k+1}^n|x_{i+1}^{i+k}) - p(x_{i+k+1}^n|x_1^{i+k})|}{p(x_{i+k+1}^n|x_1^{i+k})} \\
&\leq \sup_x \frac{\sum_{z_i} p(x_{i+k+1}^n \mid z_{i+k})|p(z_{i+k}|x_{i+1}^{i+k}) - p(z_{i+k}|x_1^{i+k})|}{\sum_{z_i} p(x_{i+k+1}^n|z_{i+k})p(z_{i+k}|x_1^{i+k})} \\
&\leq \sup_x \frac{\sup_z p(x_{i+k+1}^n|z_{i+k})}{\inf_z p(x_{i+k+1}^n|z_{i+k})} \sum_{z_{i+k}} |p(z_{i+k}|x_{i+1}^{i+k}) - p(z_{i+k}|x_1^{i+k})| \leq C\epsilon_{\text{mix}}^{-4}\widetilde{\rho}_{\text{mix}}^k s, \\
&\sup_x \frac{|p(x_1^{i-k-1} \mid x_{i-k}^i) - p(x_1^{i-k-1} \mid x_{i-k}^{i+k})|}{p(x_1^{i-k-1} \mid x_{i-k}^{i+k})} \leq C\epsilon_{\text{mix}}^{-4}\widetilde{\rho}_{\text{mix}}^k s
\end{aligned}$$

and $\frac{p(x_{i+1}^{i+k})}{p(x_{i-k}^{i+k}|x_{i-k}^i)} \leq s\epsilon_{\text{mix}}^2$ as well as $\frac{p(x_1^{i-k-1}|x_{i-k}^i)}{p(x_1^{i-k-1}|x_{i-k}^{i+k})} \leq s\epsilon_{\text{mix}}^{-2}$. The second term is therefore bounded by

$$\begin{aligned}
&\left|\frac{p(x_{i+k+1}^n|x_{i+1}^{i+k})p(x_1^{i-k-1}|x_{i-k}^i)}{p(x_{i+k+1}^n, x_1^{i-k-1}|x_{i-k}^{i+k})} - 1\right| \frac{p(x_{i+1}^{i+k})}{p(x_{i-k}^{i+k}|x_{i-k}^i)} \frac{1}{p(z_i)} \\
\leq &\frac{|p(x_{i+k+1}^n|x_{i+1}^{i+k}) - p(x_{i+k+1}^n|x_1^{i+k})|}{p(x_{i+k+1}^n|x_1^{i+k})} \frac{p(x_1^{i-k-1} \mid x_{i-k}^i)}{p(x_1^{i-k-1} \mid x_{i-k}^{i+k})} + \frac{|p(x_1^{i-k-1} \mid x_{i-k}^i) - p(x_1^{i-k-1} \mid x_{i-k}^{i+k})|}{p(x_1^{i-k-1} \mid x_{i-k}^{i+k})} \\
\leq &\frac{Cs^2\widetilde{\rho}_{\text{mix}}^k}{\epsilon_{\text{mix}}^6}
\end{aligned}$$

D.3 Proof of Lemma 6

The latter inequality is valid in our particular case because

$$\begin{aligned}
|\operatorname{cov}(z_0, z_\ell \mid x_0, \ldots, x_k)| &= |\sum_{z_0, z_\ell} z_0 z_\ell p(z_\ell \mid z_0, x)p(z_0|x) - \sum_{z_0} z_0 p(z_0|x) \sum_{z_\ell} z_\ell p(z_\ell|x)| \\
&= |\sum_{z_0, z_\ell} z_0 z_\ell p(z_0|x)(p(z_\ell|z_0, x) - p(z_\ell \mid x))| \\
&\leq \sup_{z_\ell, z_0} |p(z_\ell|z_0, x) - p(z_\ell \mid x)| \sum_{z_0} \sum_{z_\ell} |z_0 z_\ell| p(z_0 \mid x)
\end{aligned}$$

Let us now show that $\sup_{z_\ell, z_0} \left|p(z_\ell \mid z_0, x) - p(z_\ell \mid x)\right| \leq \rho_{\text{mix}}^\ell$. Introducing the shorthand $\Delta(\ell) = p(z_\ell = 1 \mid z_0 = 1, x) - p(z_\ell = 1 \mid z_0 = -1, x)$, we first claim that

$$|\Delta(1)| \leq \rho_{\text{mix}} \tag{78}$$

To establish this fact, note that

$$\begin{aligned}
\Delta(1) &= \left|\frac{p(x \mid z_\ell = 1)}{p(x \mid z_{\ell-1} = 1)} p(z_\ell = 1 \mid z_{\ell-1} = 1) - \frac{p(x \mid z_\ell = 1)}{p(x \mid z_{\ell-1} = -1)} p(z_\ell = 1 \mid z_{\ell-1} = -1)\right| \\
&= \frac{ap}{ap + b(1-p)} - \frac{a(1-p)}{a(1-p) + bp} \\
&= \frac{ab}{(ap + b(1-p))(a(1-p) + bp)}(2p - 1)
\end{aligned}$$

where we write $a = p(x \mid z_\ell = 1)$ and $b = p(x \mid z_\ell = -1)$. The denominator is minimized at $p = 1$ so that inequality (78) is shown. The same argument shows that $|\Delta(-1)| \leq \rho_{\text{mix}}$.

Induction step: Assume that $\Delta(\ell - 1) \leq \rho_{\text{mix}}^{\ell-1}$. It then follows that

$$\begin{aligned}
&|p(z_\ell = 1 \mid z_0 = 1, x) - p(z_\ell = 1 \mid z_0 = -1, x)| \\
&= |\sum_{z_{\ell-1}} p(z_\ell = 1 \mid z_{\ell-1}, x)p(z_{\ell-1} \mid z_0 = 1, x) - p(z_\ell = 1 \mid z_{\ell-1}, x)p(z_{\ell-1}|z_0 = -1, x)| \\
&= \Delta(1)\Delta(\ell - 1) \leq \rho_{\text{mix}}^{\ell}
\end{aligned}$$

Since

$$p(z_\ell = 1 \mid z_0 = -1, x) - p(z_\ell = 1 \mid z_0 = 1, x) = -p(z_\ell = -1 \mid z_0 = -1, x) + p(z_\ell = -1 \mid z_0 = 1, x)$$

we use the shorthand $s = p(z_0 = 1 \mid x)$ to obtain

$$\begin{aligned}
&\sup_{z_\ell, z_0} \mid p(z_\ell \mid z_0, x) - p(z_\ell \mid x)| \\
&= \sup_{b_\ell, b_0} p(z_\ell = b_\ell \mid z_0 = b_0, x) - [(p(z_\ell = b_\ell \mid z_0 = 1, x)s + p(z_\ell = b_\ell \mid z_0 = -1, x)(1 - s)] \\
&\leq (1 - s)|\Delta(\ell)| \leq \rho_{\text{mix}}
\end{aligned}$$

which proves the bound for $\text{cov}(Z_0, Z_1 \mid \gamma)$.

For the two state mixing we define $\widetilde{\Delta}(\ell) = p(z_\ell = 1 \mid z_1 z_0 = 1, x) - p(z_\ell = 1 \mid z_1 z_0 = -1, x)$ and can readily see that $|\widetilde{\Delta}(1)| \leq \rho_{\text{mix}}$ and

$$\begin{aligned}
&|p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = 1, x) - p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = -1, x)| \\
&= [p(z_{\ell+2} = 1 \mid z_{\ell+1} = 1, x) - p(z_{\ell+2} = -1 \mid z_{\ell+1} = -1, x)]\tilde{\Delta}(2)
\end{aligned}$$

Using equation (78), we obtain

$$|\widetilde{\Delta}(2)| = |p(z_1 = 1 \mid z_0 = 1, x) - p(z_1 = -1 \mid z_0 = -1, x)|\widetilde{\Delta}(1) \leq \rho_{\text{mix}} \tag{79}$$

from which it directly follows that

$$|p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = 1, x) - p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = -1, x)| \leq \rho_{\text{mix}}$$

The rest follows the same arguments as above and the bound for $\text{cov}(Z_0 Z_1, Z_\ell Z_{\ell+1} \mid \gamma)$ in inequality (55) is shown.

Finally, the bound for $\text{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \gamma)$ in inequality (55) follows in a straightforward way using the relation (79) and induction with equation (78), as above.

References

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.

L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, pages 1554–1563, 1966.

L.E. 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*, pages 164–171, 1970.

M. Belkin and K. Sinha. Toward learning Gaussian mixtures with arbitrary separation. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 407–419, 2010.

P.J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general Hidden Markov Models. *The Annals of Statistics*, 26(4):1614–1635, 08 1998. doi: 10.1214/aos/1024691255.

O. Cappé, E. Moulines, and T. Rydén. Hidden Markov Models, 2004.

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

S. Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 634–644, 1999.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, pages 1–38, 1977.

R. Durbin. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998. ISBN 9780521620413. URL `https://books.google.co.in/books?id=R5P2GlJvigQC`.

R.J. Elliott, L. Aggoun, and J.B. Moore. *Hidden Markov Models: Estimation and Control*. Applications of Mathematics. Springer, 1995. ISBN 9780387943640. URL `https://books.google.co.in/books?id=aR6-ASc_efQC`.

D. Hsu and S.M. Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 11–20, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1859-4. doi: 10.1145/2422436.2422439. URL `http://doi.acm.org/10.1145/2422436.2422439`.

D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning Hidden Markov Models. *J. Comput. Syst. Sci.*, 78(5):1460–1480, 2012.

C.J. Kim and C.R. Nelson. *State-space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Applications*. MIT Press, 1999. ISBN 9780262112383. URL `https://books.google.co.in/books?id=eQFsQgAACAAJ`.

L. A. Kontorovich, B. Nadler, and R. Weiss. On learning parametric-output HMMs. In *Proc. 30th International Conference Machine Learning*, pages 702–710, June 2013.

F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Info. Theory*, 47(2):498–519, February 2001.

Michel Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, 1:63–87, 1997.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*, volume 23. Springer Science & Business Media, 2013.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 93–102, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4244-7. doi: 10.1109/FOCS.2010.15. URL `http://dx.doi.org/10.1109/FOCS.2010.15`.

E. Mossel and S. Roch. Learning nonsingular phylogenies and Hidden Markov Models. *The Annals of Applied Probability*, 16(2):583–614, 05 2006.

A. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters*, 17(3):169–172, 1993.

L.R. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Pearson Education Signal Processing Series. Pearson Education, 1993. ISBN 9788129701381. URL `https://books.google.co.in/books?id=hoVLAAAACAAJ`.

S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank Hidden Markov Models. In *Proc. 13th International Conference on Artificial Intelligence and Statistics*, 2010.

S. Terwijn. On the learnability of Hidden Markov Models. In *Proceedings of the 6th International Colloquium on Grammatical Inference: Algorithms and Applications*, ICGI '02, pages 261–268, London, UK, UK, 2002. Springer-Verlag.

Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

R. van Handel. Hidden Markov Models. *Unpublished lecture notes*, 2008.

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, June 2004. ISSN 0022-0000. doi: 10.1016/j.jcss.2003.11.008. URL `http://dx.doi.org/10.1016/j.jcss.2003.11.008`.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1—305, December 2008.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

J. C. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

Xinyang Yi and Constantine Caramanis. Regularized EM algorithms: A unified framework and provable statistical guarantees. *arXiv preprint arXiv:1511.08551*, 2015.

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.