Acta Appl Math (2007) 96: 309–326
DOI 10.1007/s10440-007-9102-5

On Adjusted Viterbi Training

Alexey Koloydenko · Meelis Käärik · Jüri Lember

Published online: 13 April 2007

Abstract The *EM algorithm* is a principal tool for parameter estimation in the hidden Markov models, where its efficient implementation is known as the *Baum–Welch* algorithm. This paper is however motivated by applications where EM is replaced by *Viterbi training*, or *extraction* (VT), also known as the *Baum–Viterbi* algorithm. VT is computationally less intensive and more stable, and has more of an intuitive appeal. However, VT estimators are also biased and inconsistent. Recently, we have proposed elsewhere the *adjusted Viterbi training* (VA), a new method to alleviate the above imprecision of the VT estimators while preserving the computational advantages of the baseline VT algorithm. The key difference between VA and VT is that asymptotically, the true parameter values are *a fixed point* of VA (and EM), but not of VT. We have previously studied VA for a special case of Gaussian mixtures, including simulations to illustrate its improved performance. The present work proves the asymptotic fixed point property of VA for general hidden Markov models.

Keywords Baum · Computational efficiency · Consistent · Consistency · EM · Hidden Markov models · Maximum likelihood · Mixture models · Parameter estimation · Viterbi extraction · Viterbi training

J. Lember is supported by Estonian Science Foundation Grant 5694.

A. Koloydenko (✉)
Division of Statistics, University of Nottingham, Nottingham NG7 2RD, UK
e-mail: alexey.koloydenko@maths.nottingham.ac.uk

M. Käärik
University of Tartu, Liivi 2-503, Tartu 50409, Estonia
e-mail: meelis.kaarik@ut.ee

J. Lember
University of Tartu, Liivi 2-507, Tartu 50409, Estonia
e-mail: jyril@ut.ee

1 Introduction

We consider procedures to estimate parameters of a finite state hidden Markov model (HMM) given observations $x_1, \ldots, x_n$. Let Y be a Markov chain with state space $S = \{1, 2, \ldots, K\}$, transition matrix $\mathbb{P} = (P_{ij})$, and initial distribution π. To every state $l \in S$ there corresponds an emission distribution P_l with density f_l that is known up to the parametrization $f_l(x; \theta_l)$. When Y_k, $k \geq 1$, is in state l, an observation x_k on X_k is emitted according to P_l and independent of everything else. Y can also be called a *regime*.

A standard method to compute the (locally) maximal likelihood estimates of $(\pi, \mathbb{P}, \theta_1, \theta_2, \ldots, \theta_K)$, the HMM parameters, is the EM algorithm. The computationally efficient implementation of EM in the present context is also known as the *Baum–Welch* or simply *Baum*, or *forward–backward algorithm* [1, 3, 7, 12, 16, 18, 37, 38]. Since EM can in practice be slow and computationally expensive, it is commonly replaced by *Viterbi extraction*, or *training*, (VT), also known as the *Baum–Viterbi* algorithm. VT appears to have been introduced in [17] by F. Jelinek and his colleagues at IBM in the context of speech recognition where it has been used extensively ever since [12, 16, 31, 34, 38, 39, 44–46]. Its computational stability and intuitive appeal [12] have also made VT popular in natural language modeling [35], image analysis [19, 28], and bioinformatics [4, 10, 11, 23, 30, 36]. VT is also related to constrained vector quantization [9, 15]. The main idea of the method is to replace the computationally costly expectation (E-step) of the EM algorithm with an appropriate maximization step that generally requires less intensive computations. In speech recognition, essentially the same training procedure was also described by Rabiner et al. in [20, 39] (see also [37, 38]) as a variation of the *Lloyd algorithm* used in vector quantization. In that context, VT has gained the name of *segmental K-means* [12, 20]. The analogy with vector quantization is especially pronounced when the underlying chain is trivialized to i.i.d. variables, thus producing an i.i.d. sample from a mixture distribution. For such mixture models, VT was also described by Gray et al. in [9], where the training algorithm was considered in the vector quantization context under the name of *entropy constrained vector quantization (ECVQ)*. A better known name for VT in the mixture case is *Classification EM (CEM)*, [8, 13], stressing that instead of the mixture likelihood, CEM maximizes the *Classification Likelihood* [4, 8, 13, 32]. VT-CEM was particularly suitable for the early efforts in image segmentation [42, 43]. Also, for the uniform mixture of Gaussians with a common covariance matrix of the form $\sigma^2 I$ and unknown σ, VT, or CEM, is equivalent to the *k-means clustering* [8, 9, 13, 41].

We presently focus on the case when π and $\mathbb{P}$ (the mixing weights, in the mixture case), the regime parameters, are known. This might seem overly restrictive in general, but does not appear to be entirely unrealistic in such applications as speech recognition. The overall flexibility of such applications is usually attained via flexible models for the emission distributions (e.g. mixtures of many high-dimensional Gaussians), whereas the regime is modeled as simply as possible, if not simply fixed. Moreover, all training procedures considered in this paper and including our adjusted Viterbi training (VA), extend relatively easily to the general case as illustrated in [25]; assuming the regime to be known, however, greatly simplifies the exposition.

The VT algorithm for estimation of the emission parameters can be described as follows. Fix an initial value of the parameters $\theta^{(0)}$ and find a realization of Y to maximize the likelihood of the given observations. Such an n-tuple of states is called a *Viterbi*, or *forced, alignment*. Every alignment partitions the original sample into subsamples corresponding to distinct states. If regarded as an i.i.d. sample from P_l, the subsample corresponding to state l gives rise to $\hat{\mu}_l^n$, the maximum likelihood estimate (MLE) of θ_l. These estimates replace the

current parameter values and are subsequently used to obtain an alignment in the next step of the training, and so on. It can be shown that in general this procedure converges in finitely many steps [20]; also, it is usually much faster than the Baum algorithm. (Although the two algorithms scale essentially as K^2n, the E-part of the Baum algorithm additionally requires expensive evaluations of the densities f_l at every data point and for all $l = 1, 2, \ldots, K$.)

Despite its attractiveness, VT has a significant theoretical disadvantage that might also affect its performance in applications. The VT estimators are generally biased and not consistent. This has been noted, at least in the case of mixtures, since [4], with a specific caveat issued in [47]. (In Sect. 4.1, we illustrate numerically an appreciable bias in VT estimation of an HMM that is more general than an i.i.d. mixture.) The fact that the VT estimators are biased and inconsistent is not particularly surprising. Indeed, in contrast to EM, VT's objective is different from increasing the likelihood of the parameters given the observed data x. Instead, VT increases the joint likelihood of the (hidden) state sequence and the parameters, given the observed data x. It is true that under certain reasonable conditions [21, 33], the difference between the two objective functions vanishes as D, the dimension of the emission X_i, grows relative to $\log(K)$, which can be realistic in *isolated word recognition* [33]. Even though, this does not imply closeness of the parameter estimates obtained by EM and VT since both perform local rather than global optimization [12].

Certainly, unbiasedness and consistency are neither necessary nor sufficient for a procedure to perform well in applications [43]. However, there are a number of indications that some applications, such as *segment-based speech recognition* [44], do benefit from staying faithful to the standard, i.e. EM-type, likelihood maximization. Perhaps, such indications should be interpreted with caution since a real application integrates its HMM, or, more often a hierarchy of interacting HMMs, into a complex system, making it difficult to isolate its particular factor, such as parameter estimation, for evaluation. Nonetheless, it is acknowledged, for example in [44], that conventional speech recognizers would prefer the "smoother convergence" of EM, presumably over the more abrupt, greedy one of VT. Furthermore, it appears that concessions to using VT in segment-based speech recognizers are more due to domain specific complications associated with a direct implementation of the Baum algorithm, and less due to the (ten-fold) speed advantage of VT over Baum [44]. It also appears consistent with these observations that other applications [35] propose compromises using VT with more than one best alignment, or several perturbations of the best alignment. There are other considerations (e.g. related to initialization) in favor of the Baum–Welch algorithm for use in segment-based speech recognition [44].

Motivated by the above considerations, we have attempted to investigate the following question: Is it possible to adjust VT in an analytic way so that adjusted training still enjoys the good properties of VT (fast convergence and overall computational feasibility) while the adjusted estimators become less biased or more consistent? In particular, we focus on a special property of the EM algorithm that VT lacks. This property ensures that the true parameters are asymptotically a fixed point of the algorithm. In other words, for a sufficiently large sample, the EM algorithm "recognizes" the true parameters and does not change them significantly. In contrast to this, an iteration of VT would in general disturb the correct values noticeably (Sect. 4.1). We have thus proposed in [25] to modify VT in order to make the true parameters an asymptotically fixed point of VA, the resulting algorithm. The idea of reducing the bias of VT also appeared in [21], where a sequentially (in time) adjusted VT was proposed based on random delays and suitable for *on-line* processing of virtually infinite processes. Although VA is also based on the asymptotic properties of the process, it is substantially different from the *sequential segmental K-means* of [21] as, for one instance,

it uses the entire batch of observations $x_1, \ldots, x_n$. (Although our default setting has been *off-line*, on-line implementations of VA might also be considered in the future.)

In order to understand VA it is crucial to understand the asymptotic behavior of $\hat{\mu}_l^n$, the maximum likelihood estimators based on the subsamples obtained from the alignment. Since the alignment depends on $\theta^{(0)}$, the initial values of the parameters, so does $\hat{\mu}_l^n(\theta^{(0)}, x_1, \ldots, x_n)$. *Thus, for θ_l^* to be asymptotically fixed for every $l \in S$ means the following: Assuming θ_l^* are the true parameters and the alignments are based on θ_l^*,*

$$\hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n) \underset{n\to\infty}{\to} \theta_l^*, \quad \text{a.s.} \tag{1}$$

The reason why VT does not enjoy the desired fixed point property is that (1) need not hold in general [4, 47]. Hence, in order to improve VT in the above sense, one needs to study the asymptotics of $\hat{\mu}_l^n$. In particular, the following questions should be answered: Does the sequence $\hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n)$ converge (a.s.) at all? If yes, then what is $\mu_l(\theta^*)$, its limit? These questions have been essentially answered in [24]. Namely, it has been shown (under certain mild conditions) that the empirical measures $P_l^n(\cdot; \theta^*, X_1, X_2, \ldots, X_n)$ obtained via the Viterbi alignment (with true parameters) do converge weakly to a limiting probability measure $Q_l(\cdot; \theta^*)$ and that in general $Q_l(\theta^*) \neq P_l(\theta^*)$. Formally, for every $l \in S$, there exists a probability measure Q_l such that for any Q_l-integrable g

$$\int g(x) P_l^n(dx; \theta^*, X_1, \ldots, X_n) \underset{n\to\infty}{\to} \int g(x) Q_l(dx; \theta^*), \quad \text{a.s.}, \tag{2}$$

which implies $P_l^n(\theta^*, X_1, \ldots, X_n) \Rightarrow Q_l(\theta^*)$, a.s. [2] ("$\Rightarrow$" denotes the weak convergence of probability measures). In order to obtain the above results, Viterbi alignments, or paths, have to be extended at infinitum. Earlier attempts to consider convergence of Viterbi paths appear in [5, 6] with a more general and more complete treatment of the problem to be found in [24, 26].

In this paper, we show that under general conditions on the densities $f_l(x; \theta_l)$, convergence (2) implies convergence of $\hat{\mu}_l^n$, i.e.

$$\hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n) \underset{n\to\infty}{\to} \mu_l(\theta^*), \quad \text{a.s.},$$

$$\text{where } \mu_l(\theta^*) \overset{\text{def}}{=} \arg\max_{\theta_l' \in \Theta_l} \int \ln f_l(x; \theta_l') Q_l(dx; \theta^*). \tag{3}$$

Since in general $Q_l \neq P_l$, it is most likely that $\mu_l(\theta^*) \neq \theta_l^*$. *Reduction of the bias $\mu_l(\theta) - \theta_l$ is the main feature of the adjusted Viterbi training.*

The rest of the paper is organized as follows. In Sect. 2, we present the HMM framework and adjusted Viterbi training formally. In Sect. 3, we prove convergence (3), *which is the theoretical result of the paper*. Simulations in Sect. 4.1 illustrate the discrepancy between the measures P_l and Q_l, as well as the performance of VA. A concluding discussion follows in Sect. 4.2.

2 Adjusted Viterbi Training

2.1 The Model

Assume Y to be irreducible and aperiodic with transition matrix $\mathbb{P} = (p_{ij})$ and assume the initial distribution π to be also the stationary distribution of Y. We consider the hidden

Markov model (HMM), in which to every state $l \in S$ there corresponds an *emission distribution* P_l on $(\mathcal{X}, \mathcal{B})$. We assume $\mathcal{X}$ and $\mathcal{B}$ to be a separable metric space and the corresponding Borel σ-algebra, respectively. Let f_l be the density of P_l with respect to a reference measure λ on $(\mathcal{X}, \mathcal{B})$, where λ can be, for example, the Lebesgue measure.

In this model, to any realization $y_1, y_2, \ldots$ of Y there corresponds a sequence of independent random variables, $X_1, X_2, \ldots$, where X_n has the distribution P_{y_n}. We do not know the realizations y_n (the Markov chain Y is hidden), as we only observe the process $X = X_1, X_2, \ldots$, or, more formally:

Definition 2.1 We say that the stochastic process X is a hidden Markov model if there is a (measurable) function h such that for each n,

$$X_n = h(Y_n, e_n), \quad \text{where } e_1, e_2, \ldots \text{ are i.i.d. and independent of } Y. \tag{4}$$

Hence, the emission distribution P_l is the distribution of $h(l, e_n)$. The distribution of X is completely determined by the regime parameters $(\pi, \mathbb{P})$ and the emission distributions P_l, $l \in S$. The process X is also α-mixing and, therefore, ergodic [12, 14, 27].

2.2 Viterbi Alignment and Training

Suppose we observe $x_1, \ldots, x_n$, the first n elements of X. Throughout the paper we assume that the sample $x_1, \ldots, x_n$ is generated by an HMM with regime parameters $(\pi, \mathbb{P})$ and with emission densities $f_l(x; \theta_l^*)$, where $\theta^* = (\theta_1^*, \ldots, \theta_K^*)$ are the unknown true parameters. We assume that the regime parameters $\mathbb{P}$ and π are known, but the emission densities are known only up to the parametrization $f_l(\cdot; \theta_l)$, $\theta_l \in \Theta_l$.

A key concept of the paper is the *Viterbi alignment*, which is any sequence of states $q_1, \ldots, q_n \in S$ that maximizes the likelihood of observing $x_1, \ldots, x_n$. In other words, the Viterbi alignment is a maximum-likelihood estimator of the realization of $Y_1, \ldots, Y_n$, treated as a set of unknown parameters, for given $x_1, \ldots, x_n$. In the following, the Viterbi alignment will be referred to as the *alignment*. We start with the formal definition of the alignment. Let $q_1, \ldots, q_n$, $q_i \in S$ denote a sequence of states and define $\Lambda(q_1, \ldots, q_n; x_1, \ldots, x_n; \theta)$ to be the likelihood function $\mathbf{P}(Y_i = q_i,\ i = 1, \ldots, n) \prod_{i=1}^n f_{q_i}(x_i)$.

Definition 2.2 For each $n \geq 1$, let the set of alignments be defined as follows:

$$\mathcal{V}_\theta(x_1, \ldots, x_n) = \left\{ v \in S^n \colon \forall w \in S^n \ \Lambda(v; x_1, \ldots, x_n; \theta) \geq \Lambda(w; x_1, \ldots, x_n; \theta) \right\}. \tag{5}$$

Overloading the term, we also refer to any map $v_\theta : \mathcal{X}^n \mapsto \mathcal{V}_\theta(x_1, \ldots, x_n)$ as an alignment.

The non-uniqueness of alignment might cause problems when dealing with asymptotics [24]. Specifying a unique $v_\theta \in \mathcal{V}_\theta(x_1, \ldots, x_n)$ for every n and $x_1, \ldots, x_n$ (and every θ) in a consistent manner is discussed in [6, 24]. From now on, $v_\theta \in \mathcal{V}_\theta$ is assumed to be chosen uniquely in accordance with [24]. For many cases in practice, however, $\mathcal{V}_\theta$ consists of a single alignment to begin with.

Recall that *Viterbi training* provides a common shortcut to computing MLE of θ^*, especially in situations where D, the dimension of X is high, n is large, and f_ls are complex. VT replaces the computationally expensive expectation (E-)step by an appropriate maximization step that is based on the alignment. We now describe the Viterbi training in the HMM case.

Viterbi Training

1. Choose an initial value $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_K^{(j)})$, $j = 0$.
2. Given $\theta^{(j)}$, obtain the alignment

$$v_{\theta^{(j)}}(x_1, \ldots, x_n) = (v_1, \ldots, v_n)$$

and partition the sample $x_1, \ldots, x_n$ into K sub-samples, where the observation x_k belongs to the l^{th} subsample if and only if $v_k = l$. Equivalently, define (at most) K empirical measures

$$\hat{P}_l^n(A; \theta^{(j)}, x_1, \ldots, x_n) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n I_{A\times l}(x_i, v_i)}{\sum_{i=1}^n I_l(v_i)}, \quad A \in \mathcal{B},\ l \in S. \tag{6}$$

3. For every sub-sample find MLE given by:

$$\hat{\mu}_l^n(\theta^{(j)}, x_1, \ldots, x_n) = \arg\max_{\theta_l \in \Theta_l} \int \ln f_l(x; \theta_l)\, \hat{P}_l^n(dx; \theta^{(j)}, x_1, \ldots, x_n), \tag{7}$$

and take

$$\theta_l^{(j+1)} = \hat{\mu}_l^n(\theta^{(j)}, x_1, \ldots, x_n), \quad l \in S.$$

If for some $l \in S$ $v_i \neq l$ for any $i = 1, \ldots, n$ (lth subsample is empty), then the empirical measure $\hat{P}_l^n$ is formally undefined, in which case we take $\theta_l^{(j+1)} = \theta_l^{(j)}$. We will be ignoring this special case from now on.

The Viterbi training can be interpreted as follows. Suppose that at some step j, $\theta^{(j)} = \theta^*$ and hence $v_{\theta^{(j)}}$ is obtained using the true parameters. Let $y_1, \ldots, y_n$ be the actual hidden realization of Y. The training is then based on the assumption that the alignment $v_{\theta^{(j)}}(x_1, \ldots, x_n) = (v_1, \ldots, v_n)$ is perfect, i.e., $v_i = y_i$, $i = 1, \ldots, n$, or nearly perfect. If the alignment were indeed perfect, the empirical measures $\hat{P}_l^n$, $l \in S$, would be obtained from the i.i.d. sample generated from $P_l(\theta^*)$, and the MLE $\hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n)$ would be a natural estimator to use. Clearly, under these assumptions $\hat{P}_l^n(\theta^*, X_1, \ldots, X_n) \Rightarrow P_l(\theta^*)$ a.s. and, provided that $\{f_l(\cdot; \theta) : \theta \in \Theta_l\}$ is a P_l-Glivenko–Cantelli class and Θ_l is equipped with some suitable metric, $\lim_{n\to\infty} \hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n) = \theta_l^*$ a.s. Hence, if n is sufficiently large, then $\hat{P}_l^n \approx P_l$ and $\theta_l^{(j+1)} = \hat{\mu}_l^n(\theta^*, x_1, \ldots, x_n) \approx \theta_l^* = \theta_l^{(j)}$, $\forall l \in S$, i.e. $\theta^{(j)} = \theta^*$ would be (approximately) a fixed point of the training algorithm.

A weak point of the above argument is, of course, that the alignment in general is not perfect (even when the parameters used to find it, are the true ones). That is, generally $v_i \neq y_i$. In particular, this implies that the empirical measures $\hat{P}_l^n(\theta^*, X_1, \ldots, X_n)$ are not obtained from an i.i.d. sample taken from $P_l(\theta^*)$, and can be rather far from that. Hence, we have no reason to believe that $\hat{P}_l^n(\theta^*, X_1, \ldots, X_n) \Rightarrow P_l(\theta^*)$ a.s. and $\lim_{n\to\infty} \hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n) = \theta_l^*$ a.s. Moreover, we do not even know whether the sequences of empirical measures $\{\hat{P}_l^n(\theta^*, X_1, \ldots, X_n)\}$ and MLE estimators $\{\hat{\mu}_l^n(\theta^*, X_1, \ldots, X_n)\}$ converge (a.s.) at all. We thus present Theorem 3.2, the theoretical result of this paper, which states that if Θ_l is a closed subset of $\mathbb{R}^d$, then, under certain assumptions on classes $\{f_l(\cdot; \theta_l) : \theta_l \in \Theta_l\}$, convergence (2) (proven in [24] and implying $\hat{P}_l^n(\theta^*, X_1, \ldots, X_n) \Rightarrow Q_l(\theta^*)$, a.s.) yields convergence (3). In an attempt to reduce the bias $\theta_l^* - \mu_l(\theta^*)$ (in general $\mu_l(\theta^*) \neq \theta_l^*$), we have proposed *adjusted Viterbi training* as follows: Assuming (3), consider the mapping

$$\theta \mapsto \mu_l(\theta), \quad l = 1, \ldots, K. \tag{8}$$

The function (8) is independent of the sample, hence the following correction is well-defined:

$$\Delta_l(\theta) = \theta_l - \mu_l(\theta), \quad l = 1, \ldots, K. \tag{9}$$

Adjusted Viterbi Training

1. Choose an initial value $\theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_K^{(j)})$, $j = 0$.
2. Given $\theta^{(j)}$, obtain the alignment and define the empirical measures $\hat{P}_l^n(\theta^{(j)}, x_1, \ldots, x_n)$ as in (6).
3. For every $l \in S$, find $\hat{\mu}_l^n(\theta^{(j)}, x_1, \ldots, x_n)$ as in (7).
4. For each $l \in S$, define

$$\theta_l^{(j+1)} = \hat{\mu}_l^n(\theta^{(j)}, x_1, \ldots, x_n) + \Delta_l(\theta^{(j)}),$$

where Δ_l as in (9).

Note that, as desired, for a sufficiently large n, the adjusted training algorithm has θ^* as its (approximately) fixed point. Indeed, suppose $\theta^{(j)} = \theta^*$. From (3), $\hat{\mu}_l^n(\theta^{(j)}, x_1, \ldots, x_n) = \hat{\mu}_l^n(\theta^*, x_1, \ldots, x_n) \approx \mu_l(\theta^*) = \mu_l(\theta^{(j)})$, for all $l \in S$. Hence,

$$\theta_l^{(j+1)} = \hat{\mu}_l(\theta^*, x_1, \ldots, x_n) + \Delta_l(\theta^*) \approx \mu_l(\theta^*) + \Delta_l(\theta^*) = \theta_l^* = \theta^{(j)}, \quad l \in S. \tag{10}$$

In [25], we have considered i.i.d. sequences $X_1, X_2, \ldots$ distributed according to mixture densities $\sum_{l=1}^K \pi_l f_l(x; \theta^*)$ with mixing weights π_l. In this particular case of HMM, the alignment is trivial and convergences (2) and (3) follow directly from SLLN and consistency of MLE, respectively; functions (9) are relatively easy to find. Thus, in this special case, the adjusted Viterbi training algorithm is easy to implement. The simulations in [22, 25] have also shown encouraging results to illustrate the main features of the proposed idea.

3 Convergence $\hat{\mu}_l^n \underset{n\to\infty}{\to} \mu_l$

We now study convergence (3), the theoretical underpinning of adjusted Viterbi training. Since (3) is intimately related to consistency of MLE (in a non-i.i.d. setting), a variety of relevant results have been proved in the literature. Many of those are based on the Glivenko–Cantelli property of classes $\{\ln f_l(\cdot; \theta_l) : \theta_l \in \Theta_l\}$, which is in turn proved by several large deviation bounds such as Azuma–Hoeffding's inequality. Since our empirical measures are not based on any of the standard processes, such as i.i.d., martingales, or Markov chains, the aforementioned results do not apply directly. In finite-dimensional models, one can use the so-called Wald's consistency proof (see, e.g. [48]), which has also been applied in the HMM context [40]. In the following, we consider the case where for every $l \in S$, $\Theta_l \subset \mathbb{R}^d$, equipped with the Euclidean norm $\|\|$. In this case, Wald's technique is easy to adapt, making the following assumptions about the classes $\{f_l(\cdot; \theta) : \theta \in \Theta_l\}$.

Assumptions For every $l \in S$,

(0) Θ_l is closed;
(1) there exists $\theta_l \in \Theta_l$ such that $\int |\ln f_l(x; \theta_l)| Q_l(dx) < \infty$;
(2) there exists a Q_l-integrable function G_l such that $\ln f_l(x; \theta_l) \le G_l(x)$, $\forall \theta_l \in \Theta_l$, $x \in \mathcal{X}$;
(3) $\theta_l \mapsto \ln f_l(x; \theta_l)$ is continuous for every $x \in \mathcal{X}$;
(4) for every $x \in \mathcal{X}$, $\lim_{\|\theta_l\| \to \infty} f_l(x; \theta_l) = 0$.

Let us fix an arbitrary $l \in S$ and let us also refer to $\theta_l^* \in \Theta_l^*$, Q_l, $\hat{P}_l^n(\theta^*, x_1, \ldots, x_n)$, and $\hat{\mu}_l(\theta^*, x_1, \ldots, x_n)$, simply as $\theta^* \in \Theta^*$, Q, P_n, and μ_n, respectively. Let $\phi(\theta, x)$ stand for

$G_l(x) - \ln f_l(x;\theta)$. By (2) we have $\phi(\theta, x) \geq 0$, $\forall\, \theta \in \Theta$, $\forall x \in \mathcal{X}$. With this notation, we have

$$\mu_n = \arg\inf_{\theta\in\Theta} \int \phi(\theta, x) P_n(dx),$$

and

$$\Phi \stackrel{\text{def}}{=} \inf_{\theta\in\Theta} \int \phi(\theta, x) Q(dx) \stackrel{\text{by(1)}}{<} \infty. \tag{11}$$

Let

$$\mathcal{M} \stackrel{\text{def}}{=} \left\{\theta\colon \int \phi(\theta, x) Q(dx) = \Phi\right\}.$$

By (3), $\theta \mapsto \int \phi(\theta, x) Q(dx)$ is lower-semicontinuous. Hence $\mathcal{M}$ is closed. It also follows from (4) that $\mathcal{M}$ is non-empty. We are going to prove that $\mu_n \to \mathcal{M}$, a.s. Often the limit is unique, $\mu_n \to \mu$, a.s. (where $\mu = \mu_l(\theta^*)$). The regenerativity argument exploited in [24] yields

$$\int g dP_n \to \int g dQ, \quad \text{a.s.}, \tag{12}$$

where g is an arbitrary Q-integrable function. From now on, we assume (12). The first step is to prove that the sequence μ_n is a.s. bounded.

Proposition 3.1 *There exists $R_0 < \infty$ such that*

$$\boldsymbol{P}(\|\mu_n\| < R_0, \ \textit{eventually}) = 1.$$

Proof Let $\mu \in \mathcal{M}$. By (12),

$$\int \phi(\mu, x) P_n(dx) \to \int \phi(\mu, x) Q(dx) = \Phi, \quad \text{a.s.} \tag{13}$$

On the other hand, by the definition of μ_n, for each n

$$\int \phi(\mu_n, x) P_n(dx) \leq \int \phi(\mu, x) P_n(dx),$$

which together with (13) implies

$$\limsup_n \int \phi(\mu_n, x) P_n(dx) \leq \Phi, \quad \text{a.s.} \tag{14}$$

By (4) and monotone convergence,

$$\lim_{R\nearrow\infty} \int \inf_{\|\theta\|>R} \phi(\theta, x) Q(dx) = \infty. \tag{15}$$

By (3), the function $x \mapsto \inf_{\|\theta\|>R} \phi(\theta, x)$ is measurable. Choose R_0 sufficiently large for

$$\int \inf_{\|\theta\|>R_0} \phi(\theta, x) Q(dx) \geq \Phi + 1 \tag{16}$$

to hold. Suppose $\|\mu_n\| > R_0$ i.o. (infinitely often), that is, $\|\mu_{n_k}\| > R_0$ for some subsequence $\{\mu_{n_k}\}$. Then,

$$\int \phi(\mu_{n_k}, x) P_{n_k}(dx) \geq \int \inf_{\|\theta\| > R_0} \phi(\theta, x) P_{n_k}(dx). \tag{17}$$

By (12) and (16)

$$\int \inf_{\|\theta\| > R_0} \phi(\theta, x) P_n(dx) \to \int \inf_{\|\theta\| > R_0} \phi(\theta, x) Q(dx) \geq \Phi + 1, \quad \text{a.s.}$$

Thus, given (17),

$$\limsup_n \int \phi(\mu_n, x) P_n(dx) \geq \limsup_k \int \phi(\mu_{n_k}, x) P_{n_k}(dx) \geq \Phi + 1, \quad \text{a.s.},$$

contradicting (14). □

Theorem 3.2 *Given* (12), *any sequence* $\{\mu_n\}$ *with*

$$\mu_n = \arg\max_{\theta \in \Theta} \int \ln f_l(x; \theta) P_n(dx), \tag{18}$$

satisfies

$$\mu_n \to \mathcal{M}, \quad a.s. \tag{19}$$

Proof It follows from (4) and (11) that there exists $R^0 < \infty$ such that $\|\mu\| \leq R^0$ for all $\mu \in \mathcal{M}$. Clearly, redefining $R_0 \stackrel{\text{def}}{=} \max\{R_0, R^0\}$ does not affect Proposition 3.1. Without loss of generality we restrict the parameter space $\Theta^R \stackrel{\text{def}}{=} \{\theta \in \Theta\colon \|\theta\| \leq R_0\}$. Let $\epsilon > 0$ be arbitrary and consider the set

$$\Theta_\epsilon \stackrel{\text{def}}{=} \{\theta \in \Theta^R : d(\theta, \mathcal{M}) \geq \epsilon\}, \quad \text{where } d(\theta, \mathcal{M}) = \min_{\mu \in \mathcal{M}} \|\theta - \mu\|.$$

By (0), the set Θ_ϵ is closed and, therefore, compact. For every $\theta \in \Theta_\epsilon$, $\int \phi(\theta, x) Q(dx) > \Phi$, and moreover, there exists a $\gamma > 0$ (possibly depending on ϵ) such that

$$\inf_{\theta \in \Theta_\epsilon} \int \phi(\theta, x) Q(dx) \geq \gamma + \Phi. \tag{20}$$

Using the technique of Wald, we show

$$\liminf_n \inf_{\theta \in \Theta_\epsilon} \int \phi(\theta, x) P_n(dx) \geq \Phi + \frac{\gamma}{2}, \quad \text{a.s.} \tag{21}$$

Hence, if $d(\mu_n, \mathcal{M}) > \epsilon$ i.o., then

$$\limsup_n \int \phi(\mu_n, x) P_n(dx) \geq \Phi + \frac{\gamma}{2},$$

which contradicts (14). This finishes the proof of the theorem as soon as we prove (21). To prove (21), note that for every $\theta' \in \Theta_\epsilon$,

$$\sup_{\delta > 0} \int \inf_{\theta : \|\theta - \theta'\| < \delta} \phi(\theta, x) Q(dx) = \lim_{\delta \searrow 0} \int \inf_{\theta : \|\theta - \theta'\| < \delta} \phi(\theta, x) Q(dx),$$

hence by monotone convergence, we have:

$$\sup_{\delta>0}\int \inf_{\theta:\|\theta-\theta'\|<\delta}\phi(\theta,x)Q(dx) = \int \lim_{\delta\searrow 0}\inf_{\theta:\|\theta-\theta'\|<\delta}\phi(\theta,x)Q(dx)$$
$$= \int \phi(\theta',x)Q(dx) \geq \Phi+\gamma.$$

Therefore, around every $\theta' \in \Theta_\epsilon$, there exists an open ball $B(\theta')$ such that

$$\int \inf_{\theta\in B(\theta')}\phi(\theta,x)Q(dx) \geq \frac{\gamma}{2}+\Phi. \tag{22}$$

The balls $B(\theta')$ form an open cover of Θ_ϵ. Since the set Θ_ϵ is compact, there is a finite subcover $\{B(\theta_i)\}$. Now

$$\inf_{\theta\in\Theta_\epsilon}\int \phi(\theta,x)P_n(dx) = \min_i \inf_{\theta\in B(\theta_i)}\int \phi(\theta,x)P_n(dx) \geq \min_i \int \inf_{\theta\in B(\theta_i)}\phi(\theta,x)P_n(dx),$$

and since $\min_i \int \inf_{\theta\in B(\theta_i)}\phi(\theta,x)P_n(dx) \underset{n\to\infty}{\to} \min_i \int \inf_{\theta\in B(\theta_i)}\phi(\theta,x)Q(dx) \geq \Phi+\frac{\gamma}{2}$ a.s., we finally obtain

$$\liminf_n \inf_{\theta\in\Theta_\epsilon}\int \phi(\theta,x)P_n(dx) \geq \Phi+\frac{\gamma}{2} \quad \text{a.s.},$$

as required. □

Let us briefly discuss validity of assumptions (0)–(4). Assumption (0) guarantees the compactness of Θ_ϵ, and can be relaxed provided Θ contains the closed ball (centered at the origin) of radius R_0. Assumption (1) ensures that $\Phi < \infty$, which is usual and very natural. At first, this condition might appear difficult to verify given that the measures Q_l are, in general, analytically not known. However, using ergodic theory, one can show the existence of $V = (V_1, V_2, \ldots)$, a stationary process taking values in S, such that

$$Q_l(A) = \mathbf{P}(X_1 \in A | V_1 = l) \leq \frac{\mathbf{P}(X_1\in A)}{\mathbf{P}(V_1=l)} = \sum_{i=1}^K a_i P_i(A), \tag{23}$$

where $0 \leq a_i = \pi_i \mathbf{P}(V_1 = l)^{-1} < \infty$ [26]. Therefore, $Q_l \ll \lambda$ with the corresponding relation on the derivatives:

$$q_l := \frac{dQ_l}{d\lambda} \leq \sum_{i=1}^K a_i f_i, \quad \lambda-\text{a.s.}, \quad \text{where } f_i = f_i(\cdot;\theta_i^*).$$

Hence a function h is Q_l-integrable if it is P_i-integrable for each $i \in S$.

Assumption (2) is both most important and most restrictive but can be replaced by the following weaker conditions ([40]):

$$\forall\theta'\in\Theta \quad \exists\delta>0\colon \int \sup_{\theta:\|\theta-\theta'\|<\delta}\big(\ln f(x,\theta)\big)^+ Q(dx) < \infty, \tag{24}$$

$$\exists R>0\colon \int \sup_{\theta\|\theta\|>R}\big(\ln f(x,\theta)\big)^+ Q(dx) < \infty, \tag{25}$$

where $a^+ = \max\{0, a\}$. However, since Q_l is a probability measure, the condition (2) holds, for example, if the family $\{f_l(\cdot;\theta_l) : \theta_l \in \Theta_l\}$ is uniformly bounded. This holds for many models. Assumptions (3) and (4) are essentially determined by the parametrization of the model. Assumption (3) is valid for most of the models in practice and guarantees that the (uncountable) infima in the proof of Proposition 3.1 are measurable. Often, (3) is replaced by the weaker upper-semicontinuity assumption. Assumption (4) guarantees boundedness of μ_n and $\mathcal{M}$. Note that for bounded Θ, the assumption (4) can be dropped, since then the set Θ is already compact. If Θ is unbounded, then (4) is needed only to ensure the existence of R_0 such that (16) holds. Hence, one can replace (4) with (16), which is much more general. The latter depends on usually unknown Φ, but it is implied, for example, by the following condition

$$\lim_{R\to\infty} \int \sup_{\theta:\|\theta\|>R} \big(\ln f(x,\theta)\big)^+ Q(dx) = 0. \tag{26}$$

Clearly (26) is weaker than (4) and also implies (25). Hence, for unbounded parameter domains, the conditions (2) and (4) can be replaced by the more general (24) and (26).

Example 3.3 Consider a shift (location) parameter family. Let $\Theta_l = \mathcal{X} = \mathbb{R}^d$, $l = 1, \ldots, K$, and let λ be the Lebesgue measure. Suppose g_l are continuous bounded strictly positive densities on $\mathcal{X}$, and consider the families $f_l(x;\theta_l) = g_l(x-\theta_l)$, where θ_l is the location parameter. Assumption (0) holds trivially; (1) holds, if there exists θ_l such that

$$\int \big|\ln g(x-\theta_l)\big| g(x-\theta_i^*)\lambda(dx) < \infty$$

for every $i = 1, \ldots, K$. By boundedness of g_l, assumption (2) holds. Assumption (3) holds since g_l is continuous and (4) holds, because θ is location parameter.

In particular, assumptions (1)–(4) are fulfilled by the families of Laplacian distributions and the (multivariate) normal distributions with known covariance-matrices. These classes are used in Philips speech recognition models [34, 45]. For these classes, $\mathcal{M}$ consists of one element, only. Finally, Theorem 3.2 implies (3).

4 Simulations and Discussions

4.1 Simulations

We carry out simulations to demonstrate the discrepancy between P and Q-measures as well as the improvement in performance of the Viterbi training algorithm due to the adjustment. We consider a simple HHM, where the underlying MC has the following transition matrix

$$\begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

$\epsilon \in (0, 0.5]$, and the emission distributions are univariate normal with unit variance and unknown means, $P_1 = \mathcal{N}(\theta_1, 1)$, $P_2 = \mathcal{N}(\theta_2, 1)$. In this model, there are two emission parameters, θ_1 and θ_2 and one regime parameter, ϵ. Without loss of generality, assume $\theta_1 < \theta_2$ and let $a = 0.5(\theta_2 - \theta_1)$. With $\epsilon = 0.5$, this model reduces to the i.i.d. (mixture) model studied in [25].

First, we study the measures Q_l. In our model, the shape of (the density of) Q_l depends on the emission parameters through their difference a only. We therefore estimate the densities of Q_l for several values of a and ϵ. Figures 1, 2, 3, 4 provide several such examples with dashed and solid curves representing the Q_l and P_l densities, respectively. The true means θ_l^* as well as the (estimated) fixed points $\mu_l(\theta^*)$ are also marked, highlighting the correction $\Delta_l(\theta^*) = \theta_l^* - \mu_l(\theta^*)$.

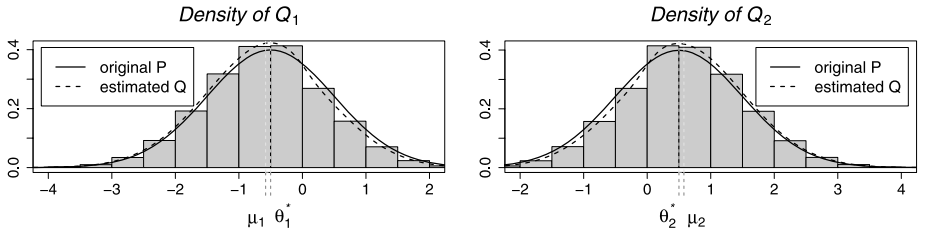

Fig. 1 Densities of Q_l, $\epsilon = 0.2$, $a = 0.5$

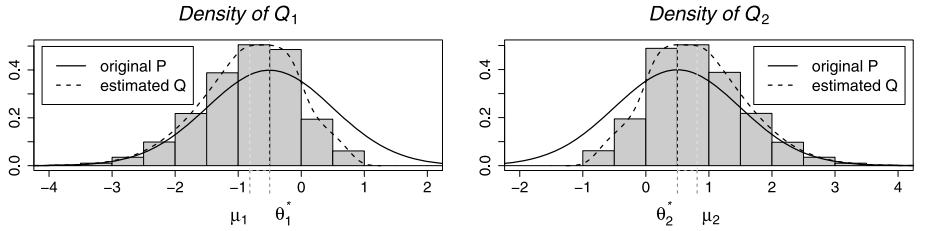

Fig. 2 Densities of Q_l, $\epsilon = 0.4$, $a = 0.5$

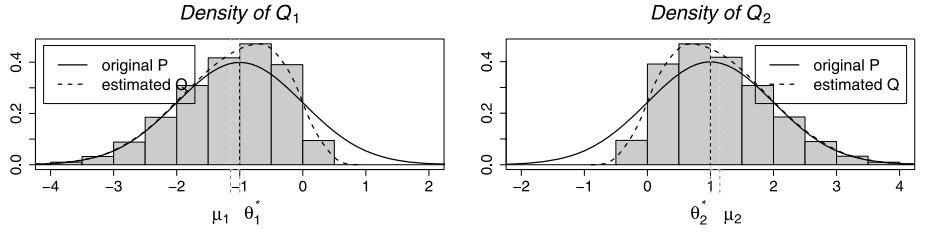

Fig. 3 Densities of Q_l, $\epsilon = 0.4$, $a = 1.0$

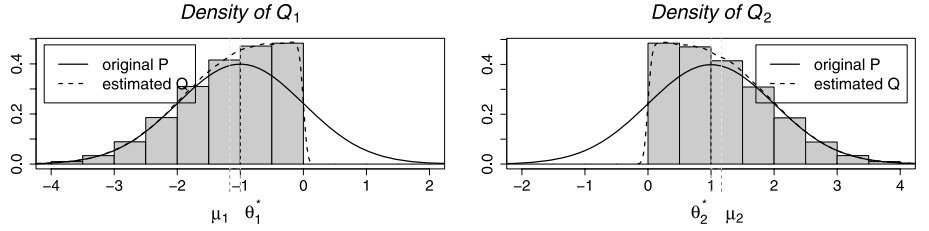

Fig. 4 Densities of Q_l, $\epsilon = 0.5$, $a = 1.0$

Since the shape of the density of Q_l depends on $\theta = (\theta_1, \theta_2)$ through $\theta_2 - \theta_1$, then $\mu_l(\theta_1 + c, \theta_2 + c) = \mu_l(\theta_1, \theta_2) + c$ for all $c \in \mathbb{R}$ (8), and therefore Δ also depends on θ only through $\theta_2 - \theta_1$, or a. Thus, to implement the VA algorithm, the correction function $\Delta(a; \epsilon)$ ($a \in (0, \infty)$, $\epsilon \in (0, 0.5]$) is needed. Except for $\epsilon = 0.5$, however, this function is not known analytically, hence needs to be approximated. To this effect, we use an (a, ϵ) mesh with $a = 0.1, 0.2, \ldots, 3.0$ and $\epsilon = 0.08, 0.09, \ldots, 0.5$ to simulate Q_l via HMM samples of size 10^6 for each (a, ϵ) node of the mesh. Thus, we approximate (8), and ultimately $\Delta(a)$, stochastically at every node of the mesh. Finally, for all other (a, ϵ) values, the correction function is obtained by linear interpolation. The results are presented in Fig. 5.

Note that generally Δ decreases as a increases (vanishing as the two density curves move infinitely far apart). Except for the case of independent mixtures ($\epsilon = 0.5$), however, there appears to be a range of near-zero a values (i.e. almost identical parameter values) with the opposite behavior, which might be interesting to investigate in more detail.

To assess precision of our approximation, at least for $\epsilon = 0.5$, in Fig. 6 we compare the approximation with the analytic result which is immediately available for this case:

$$\Delta(a; 0.5) = 2\big(\phi(a) - a\Phi(-a)\big),$$

where ϕ and Φ are the density and cumulative distribution function of the standard normal distribution, respectively. Clearly, the difference between analytic and approximate Δ is insignificant in this case.

Based on the Δ function obtained above, we apply the adjusted Viterbi training and compare it with the VT and EM algorithms. Tables 1, 2, 3 present simulation results obtained

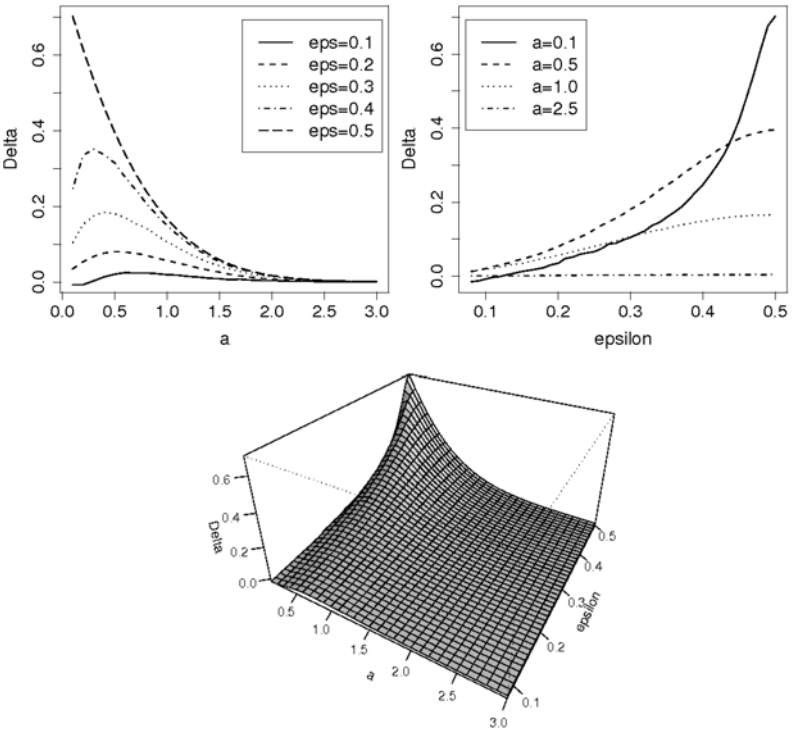

Fig. 5 Estimated correction function $\Delta(a, \epsilon)$

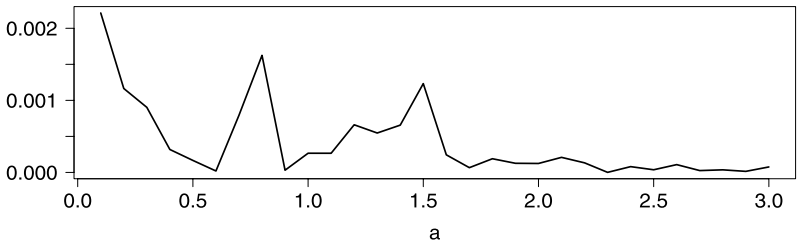

Fig. 6 The absolute difference between estimated and analytic functions $\Delta(a; 0.5)$

Table 1 Performance analysis. $\epsilon = 0.2$, $a = 0.2$, starting from 1st and 3rd quartiles

	EM	VT	VA
Step 0	$(-0.689, 0.687)$	$(-0.689, 0.687)$	$(-0.689, 0.687)$
Step 1	$(-0.477, 0.475)$	$(-0.537, 0.536)$	$(-0.460, 0.459)$
Step 2	$(-0.385, 0.384)$	$(-0.474, 0.474)$	$(-0.359, 0.358)$
Step 3	$(-0.335, 0.333)$	$(-0.445, 0.445)$	$(-0.305, 0.307)$
Step 4	$(-0.303, 0.301)$	$(-0.429, 0.430)$	$(-0.273, 0.274)$
Step 5	$(-0.281, 0.279)$	$(-0.420, 0.422)$	$(-0.252, 0.254)$
Step 6	$(-0.265, 0.264)$		$(-0.239, 0.241)$
Step 7	$(-0.253, 0.252)$		$(-0.229, 0.232)$
Step 8	$(-0.244, 0.243)$		
L_1 error	0.087	0.442	0.061
L_2 error	0.061	0.312	0.043
L_∞ error	0.044	0.222	0.032

Table 2 Performance analysis. $\epsilon = 0.4$, $a = 0.5$, starting from 1st and 3rd quartiles

	EM	VT	VA
Step 0	$(-0.763, 0.764)$	$(-0.763, 0.764)$	$(-0.763, 0.764)$
Step 1	$(-0.632, 0.633)$	$(-0.854, 0.856)$	$(-0.632, 0.634)$
Step 2	$(-0.575, 0.575)$	$(-0.860, 0.864)$	$(-0.572, 0.575)$
Step 3	$(-0.545, 0.545)$		$(-0.541, 0.543)$
Step 4	$(-0.528, 0.528)$		$(-0.521, 0.525)$
Step 5	$(-0.517, 0.518)$		$(-0.511, 0.515)$
Step 6	$(-0.511, 0.511)$		
L_1 error	0.022	0.724	0.026
L_2 error	0.016	0.512	0.019
L_∞ error	0.011	0.364	0.015

from samples of size 10^6. The parameters are initialized to the first and third quartiles and the stopping rule is for the L_∞-distance between successive updates to fall below 0.01. Viterbi training is seen to be quickest to terminate, but its estimates are evidently biased. On the other hand, accuracy of adjusted Viterbi training is comparable to that of the EM algorithm, while VA terminates somewhat more rapidly than EM. Given the fact that each step of EM

Table 3 Performance analysis. $\epsilon = 0.5$, $a = 1.0$, starting from 1st and 3rd quartiles

	EM	VT	VA
Step 0	$(-1.050, 1.053)$	$(-1.050, 1.053)$	$(-1.050, 1.053)$
Step 1	$(-1.013, 1.015)$	$(-1.166, 1.169)$	$(-1.014, 1.016)$
Step 2	$(-1.003, 1.005)$	$(-1.165, 1.169)$	$(-1.004, 1.006)$
L_1 error	0.008	0.334	0.010
L_2 error	0.006	0.236	0.007
L_∞ error	0.005	0.169	0.006

Table 4 Performance analysis. $\epsilon = 0.2$, $a = 0.2$, true initial parameters

	EM	VT	VA
Step 0	$(-0.200, 0.200)$	$(-0.200, 0.200)$	$(-0.200, 0.200)$
Step 1	$(-0.198, 0.202)$	$(-0.252, 0.254)$	$(-0.198, 0.200)$
Step 2		$(-0.298, 0.302)$	
Step 3		$(-0.333, 0.339)$	
Step 4		$(-0.357, 0.367)$	
Step 5		$(-0.373, 0.386)$	
Step 6		$(-0.383, 0.399)$	
Step 7		$(-0.387, 0.408)$	
L_1 error	0.003	0.396	0.002
L_2 error	0.002	0.280	0.002
L_∞ error	0.002	0.208	0.002

Table 5 Performance analysis. $\epsilon = 0.4$, $a = 0.5$, true initial parameters

	EM	VT	VA
Step 0	$(-0.500, 0.500)$	$(-0.500, 0.500)$	$(-0.500, 0.500)$
Step 1	$(-0.501, 0.500)$	$(-0.812, 0.814)$	$(-0.497, 0.499)$
Step 2		$(-0.857, 0.861)$	
Step 3		$(-0.860, 0.865)$	
L_1 error	0.001	0.725	0.004
L_2 error	0.001	0.513	0.003
L_∞ error	0.001	0.365	0.003

requires significantly more intensive computations, one should expect the overall run time of VA to be appreciably less than that of EM.

We also test the three algorithms for the fixed point property (using the same stopping rule as before). It is evident from Tables 4, 5, 6 that both EM and VA do approximately satisfy this property, whereas VT moves the true parameters to a notably different location.

4.2 Discussion

The simulations above, as well as those in [22, 25], show that the proposed adjustment typically improves precision of the Viterbi training estimators. Moreover, accuracy of VA is already comparable with that of EM. Since the introduced correction does not depend on the data, the adjustment does not increase the amount of computations per data point. As also

Table 6 Performance analysis. $\epsilon = 0.5$, $a = 1.0$, true initial parameters

	EM	VT	VA
Step 0	$(-1.000, 1.000)$	$(-1.000, 1.000)$	$(-1.000, 1.000)$
Step 1	$(-0.998, 1.000)$	$(-1.165, 1.167)$	$(-0.998, 1.000)$
Step 2		$(-1.165, 1.167)$	
L_1 error	0.002	0.332	0.002
L_2 error	0.002	0.235	0.002
L_∞ error	0.002	0.167	0.002

shown by the simulations, in most cases of replacing VT by VA, the number of iterations does not increase drastically either. However, in implementing VA, the central issue is the availability of the correction function (8). In the special case of the i.i.d. mixture model, function (8) is essentially available analytically. Even in the high-dimensional setting with many components, when the required expressions might become unattractive, reasonable work-arounds can still be found [25].

Apart from the i.i.d. case, exact theoretical calculations of the correction function are generally impossible since the measures Q_l are not known analytically. Hence, the correction function should be computed approximately, perhaps in a stochastic manner. Here (Sect. 4.1), we estimate this function on the regular rectangular grid using linear interpolation. Since our point estimates at the grid nodes are precise, and the grid is sufficiently dense, the obtained approximation is rather accurate. Although such a procedure requires a significant effort, we point out that all the computations are done *off-line* and can be reused with the same model (regime and emission parameter values).

Another, computationally less demanding approach, is the so called *stochastically adjusted Viterbi training* (SVA) proposed in [22]. Instead of estimating the correction at every point as in the previous approach, SVA estimates the correction at every iteration (by simulations) and, therefore, only at the points visited by the algorithm. Clearly, if the number of iterations is relatively small, this method should overall require less computation. On the other hand, if a model is to be used repeatedly, estimating the correction function off-line as above, might still be preferable.

Thus far, we have primarily discussed estimation of the emission parameters. Under a complicated regime model with unknown regime parameters, one promising approach to estimating the emission parameters can be called *independent training*. In this approach, the data are treated as if they were generated by an i.i.d. mixture. The justification of this approach is as follows. If the regime is a stationary process with marginal probabilities π_i (as in the present paper), then the data $x_1, \ldots, x_n$ are a sample from the mixture distribution $\sum_{l=1}^{K} \pi_l f_l(x; \theta^*)$. Pretending to be dealing with an i.i.d. sample, one looses all the information about the dependence structure (regime) but not about the emission distribution. Hence, the corresponding estimators of the emission parameters need not deteriorate, and, for some applications, might actually be sufficiently accurate. At the same time, training in the i.i.d. case is usually significantly easier. Note, in particular, that the transition matrix under independent training is fully determined by the stationary distribution π. Even if π is not (or partly) known, one can still train the model assuming i.i.d. mixtures with unknown weights. Let us reiterate that VA is also easily applied for mixtures with unknown weights. The simulations in [25] clearly show that in this case VA terminates more rapidly than EM. Hence, independent training is applicable even with little knowledge about the transition structure, and can also be extended to settings that are more general than HMM.

All in all, however, estimation of the mixture parameters (in the i.i.d. case) remains to be an important issue, with MLE being a natural choice. In this case MLE is also known as the maximum likelihood independent estimator, or MLIE. The properties of MLIE are studied by Lindgren [29], who shows that it is consistent and asymptotically normal. Lindgren also compares accuracy of MLIE with that of MLE based on the full Markov model (both computed via EM). He concludes that, unless dependence is very strong, MLIE performs as well as MLE. His results are generalized by Ryden [40], who introduces a more general version of independent training. Again, although the EM algorithm is a natural procedure for computing MLIE, cheap alternatives such as Viterbi training, are also appreciated. To this end, note the following observation supported by the presented simulations. The adjustment of the mixture VT toward MLIE is more significant than the adjustment of the full VT (for the actual HMM) toward the true MLE. Hence, the adjusted Viterbi training is worth to consider for independent training as well.

References

1. Baum, L., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Stat. **37**, 1554–1563 (1966)
2. Billingsley, P.: Probability and Measure. Wiley, New York (1995)
3. Bilmes, J.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report 97-021. International Computer Science Institute, Berkeley (1998)
4. Bryant, P., Williamson, J.: Asymptotic behaviour of classification maximum likelihood estimates. Biometrika **65**(2), 273–281 (1978)
5. Caliebe, A.: Properties of the maximum a posteriori path estimator in hidden Markov models. IEEE Trans. Inf. Theory **48**(7), 41–51 (2006)
6. Caliebe, A., Rösler, U.: Convergence of the Maximum a posteriori path estimator in hidden Markov models. IEEE Trans. Inf. Theory **48**(7), 1750–1758 (2002)
7. Cappé, O., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer, Berlin (2005)
8. Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. **14**(3), 315–332 (1992)
9. Chou, P., Lookbaugh, T., Gray, R.: Entropy-constrained vector quantization. IEEE Trans. Acoust. Speech Signal Process. **37**(1), 31–42 (1989)
10. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
11. Ehret, G., Reichenbach, P., Schindler, U., Horvath, C., Fritz, S., Nabholz, M., Bucher, P.: DNA binding specificity of different STAT proteins. J. Biol. Chem. **276**(9), 6675–6688 (2001)
12. Ephraim, Y., Merhav, N.: Hidden Markov processes. IEEE Trans. Inf. Theory **48**(6), 1518–1569 (2002)
13. Fraley, C., Raftery, A.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **97**(458), 611–631 (2002)
14. Genon-Catalot, V., Jeantheau, T., Laredo, C.: Stochastic volatility models as hidden Markov models and statistical applications. Bernoulli **6**, 1051–1079 (2000)
15. Gray, R., Linder, T., Li, J.: A Lagrangian formulation of Zador's entropy-constrained quantization theorem. IEEE Trans. Inf. Theory **48**(3), 695–707 (2000)
16. Huang, X., Ariki, Y., Jack, M.: Hidden Markov Models for Speech Recognition. Edinburgh University Press, Edinburgh (1990)
17. Jelinek, F.: Continuous speech recognition by statistical methods. Proc. IEEE **64**, 532–556 (1976)
18. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (2001)
19. Joshi, D., Li, J., Wang, J.: A computationally efficient approach to the estimation of two- and three-dimensional hidden Markov models. IEEE Trans. Image Process. **37**(1), 31–42 (2006)
20. Juang, B.-H., Rabiner, L.: The segmental K-means algorithm for estimating parameters of hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **38**(9), 1639–1641 (1990)
21. Kogan, J.: Hidden Markov models estimation via the most informative stopping times for the Viterbi algorithm. In: Levinson, S., Shepp, L. (eds.) Image Models (and their Speech Model Cousins). The IMA Volumes in Mathematics and Its Applications, vol. 80, pp. 115–130

22. Kolde, R.: Estimating of mixture density parameters with adjusted Viterbi training (in Estonian). Bachelor Theses, Tartu University (2005)
23. Krogh, A.: An introduction to hidden Markov models for biological sequences. In: Computational Methods in Molecular Biology. Elsevier, Amsterdam (1998)
24. Lember, J., Koloydenko, A.: Adjusted Viterbi training for hidden Markov models. Technical Report 07-01, School of Mathematical Sciences, Nottingham University, http://www.maths.nottingham.ac.uk/personal/pmzaak/VA/AVT2.pdf (2007)
25. Lember, J., Koloydenko, A.: Adjusted Viterbi training: A proof of concept. Probab. Eng. Inf. Sci. **21**(3) (2007, to appear)
26. Lember, J., Koloydenko, A.: Adjusted Viterbi training for hidden Markov models. Bernoulli (2007, in revision)
27. Leroux, B.: Maximum-likelihood estimation for hidden Markov models. Stoch. Process. Appl. **40**, 127–143 (1992)
28. Li, J., Gray, R., Olshen, R.: Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models. IEEE Trans. Inf. Theory **46**(5), 1826–1841 (2000)
29. Lindgren, G.: Markov regime models for mixed distributions and switching regression. Scand. J. Stat. **5**(5), 81–91 (1978)
30. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, V., Borodovsky, M.: Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33**(20), 6494–6506 (2005)
31. McDermott, E., Hazen, T.: Minimum classification error training of landmark models for real-time continuous speech recognition. In: Proc. ICASSP. Montreal, Quebec, http://groups.csail.mit.edu/sls/publications/2004/McDermott_Hazen_ICASSP04.pdf (May 2004)
32. McLachlan, G., Peel, D.: Finite Mixture Models. Probability and Statistics. Wiley, New York (2000)
33. Merhav, N., Ephraim, Y.: Hidden Markov modelling using a dominant state sequence with application to speech recognition. Comput. Speech Lang. **5**(6), 327–339 (1991)
34. Ney, H., Steinbiss, V., Haeb-Umbach, R., Tran, B., Essen, U.: An overview of the Philips research system for large vocabulary continuous speech recognition. Int. J. Pattern Recognit. Artif. Intell. **8**(1), 33–70 (1994)
35. Och, F., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. A digital archive of research papers in computational linguistics: http://acl.ldc.upenn.edu/P/P00/P00-1056.pdf (2000)
36. Ohler, U., Niemann, H., Liao, G., Rubin, G.: Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. Bioinformatics **17**(Suppl. 1), S199–S206 (2001)
37. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
38. Rabiner, L., Juang, B.: Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River (1993)
39. Rabiner, L., Wilpon, J., Juang, B.: A segmental K-means training procedure for connected word recognition. AT&T Tech. J. **64**(3), 21–40 (1986)
40. Ryden, T.: Consistent and asymptotically normal parameter estimates for hidden Markov models. Ann. Stat. **22**(4), 1884–1895 (1993)
41. Sabine, M., Gray, R.: Global convergence and empirical consistency of the generalized Lloyd algorithm. IEEE Trans. Inf. Theory **32**(2), 148–155 (1986)
42. Sclove, S.: Application of the conditional population-mixture model to image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **5**, 428–433 (1983)
43. Sclove, S.: Author's reply. IEEE Trans. Pattern Anal. Mach. Intell. **5**, 657–658 (1984)
44. Shu, I., Hetherington, L., Glass, J.: Baum–Welch training for segment-based speech recognition. In: Proc. ASRU, St. Thomas, US Virgin Islands, http://groups.csail.mit.edu/sls/publications/2003/ASRU_Shu.pdf (December 2003)
45. Steinbiss, V., Ney, H., Aubert, X., Besling, S., Dugast, C., Essen, U., Geller, D., Haeb-Umbach, R., Kneser, R., Meyer, H., Oerder, M., Tran, B.: The Philips research system for continuous-speech recognition. Philips J. Res. **49**, 317–352 (1995)
46. Ström, N., Hetherington, L., Hazen, T., Sandness, E., Glass, J.: Acoustic modeling improvements in a segment-based speech recognizer. In: Proc. IEEE ASRU Workshop Keystone, CO, USA, MIT Comp. Sci. and AI Lab., Spoken Language Systems, http://www.sls.lcs.mit.edu/sls/publications/1999/asru99-strom.pdf (1999)
47. Titterington, D.M.: Comments on "Application of the conditional population-mixture model to image segmentation". IEEE Trans. Pattern Anal. Mach. Intell. **6**, 656–657 (1984)
48. van der Vaart, A.: Asymptotic Statistics. Cambridge University Press, Cambridge (2000)