Clustering Time Series and the Surprising Robustness of HMMs

Mark Kozdoba
markk@tx.technion.ac.il

Shie Mannor
shie@ee.technion.ac.il

Abstract

Suppose that you are given a time series where consecutive samples are believed to come from a probabilistic source, and that the source changes from time to time. Your objective is to learn the distribution of each source and to cluster the samples according to the source that generated them. A standard approach to this problem is to model the data as a hidden Markov model (HMM). However, due to the Markov property and stationarity of HMMs, simple examples can be given where this approach yields poor results for the clustering. We propose a more general, non-stationary model of the data, where the only restriction is that the sources can not change too often. Even though the model governing the sources may not be Markovian, we show that that a maximum likelihood HMM estimator can still be used. Specifically, we show that a maximum-likelihood HMM estimator produces the correct second moment of the data, and the results can be extended to higher moments. In contrast to the existing consistency and misspecification results involving maximum likelihood for HMMs, our approach yields bounds for finite sample sizes.

1 INTRODUCTION

Let \( x = (x_1, x_2, \ldots, x_N) \) be a sequence of symbols over some finite alphabet \( X \), where each symbol is sampled from one of \( k \) sources, with distributions \( \mu_1, \ldots, \mu_k \). Given the sequence \( x \), we are interested in inferring the distribution of the sources, and in the classification of the samples, in the sense of determining for each sample \( x_i \), which source produced it. A well known toy instance of this problem is the Unfair Casino, where the sources are biased coins. A classical real world application, with Gaussian sources, is in Speech Recognition, see [Gales and Young, 2007] for an overview of the modern approach. In general, applications appear in virtually any field involving time series or sequential data. For instance, in Financial times series [Mamon and Elliott, 2007], Biological sequence analysis [Yoon, 2009], Computer Vision [Horst, 2002] and Activity Recognition [Trabelsi et al., 2013], [Kim et al., 2010], to name a few. In order to be able to distinguish between the sources, one clearly needs some conditions on how the sources change as time progresses. Indeed, if the source is chosen
independently at each time \( i \), it is easy to see that the sources are indistinguishable and one effectively sees a single source with distribution equal to the empirical distribution of the data. A natural assumption on the underlying sequence of sources (also referred to as states) \( s = (s_1, \ldots, s_N) \) is that it forms a Markov chain, and the resulting model is a Hidden Markov Model (HMM). Given a sequence of data, \( x \), one can find a maximum likelihood HMM, and use the Viterbi sequence (the most likely state sequence \( s \) given the data) for classification. This approach is a standard, and often effective tool in a wide variety of applications, including all of the applications mentioned above. However, the Markov chain assumption on the state sequence \( s \) has certain theoretical and practical limitations. Perhaps the most important limitation is the stationarity. For instance, it is easy to see that if one learns a model on a path on which the data tends to stay in a single state for longer periods of time and then tries to classify a sequence with shorter stay durations, then the chain will stay at states for longer time, resulting in low classification accuracy (details and additional examples are given in Section 4). However, observe that if one knows the distributions \( \mu_1, \ldots, \mu_k \) of the states, then one may often perform classification differently. Consider the sliding window approach, which in its simplest form consists of taking a fixed window of the data, \( w_i = (x_i, \ldots, x_{i+l}) \), and for each \( i \) choosing \( \mu_j \) which maximizes the joint likelihood of the samples in \( w_i \). The standard tradeoff of this approach is that for larger \( l \) it is easier to distinguish between the sources \( \mu_j \), but if \( l \) is too large, then the window \( w_i \) may contain data generated by multiple sources, for a large fraction of indices \( i \). Assume for the sake of discussion that we found the right \( l \), namely an \( l \) such that on one hand, if \( w_i \) contains samples from a single source then maximum likelihood reliably determines which \( \mu_j \) generated \( w_i \), and on the other hand the process usually stays at single state for longer than \( l \). In such a situation, the window approach will work well. Moreover, note that as long as the requirement that a process stays at a single state at least \( l \) units of times is met, no additional regularity conditions on the state changes are required for the classification to work. In particular the state changes process does not need to be a Markov chain, and may even be non-stationary. In fact, one does not need to assume any probabilistic model on the state changes. This represents a significant relaxation of the assumptions about the data.

Motivated by this observation, we define an Interval Model \( I \) of the data to be a finite or infinite sequence of consecutive intervals in \( \mathbb{N} \), \( I_1, I_2, \ldots \), and a mapping \( \tau : \mathbb{N} \to \{1, \ldots, k\} \) such that for any \( i \in I_l \), \( x_i \) has distribution \( \mu_{\tau(l)} \) and all \( x_i \) are independent. In addition, we assume that for every \( l \), \( |I_l| > m \) for some \( m > 0 \). The parameter \( m \) is assumed to be large enough for the sliding window approach to work, but we do not assume that \( m \) is known. As discussed above, classification for this model is easy (see Section 4 for details). Thus the main question is: given a sample \( x \) from the model \( I \), how and when can the distributions \( \mu_i \) be learned?

We observe empirically in Section 4 that while classification using HMM and Viterbi sequences can give poor results on samples generated from \( I \), and despite the fact that the data is not generated by an HMM, the distributions
that maximum-likelihood HMM estimator learns on a sample $x$ are correct. Thus, given the data $x$, one can first infer the distributions using an HMM, and then classify the samples using the window approach (the required length of a window, $l$ can be easily estimated once the measures $\mu_i$ are known). This two-stage approach significantly improves the classification accuracy, as discussed in Section 4.

The main objective of this paper is a result that explains why the maximum-likelihood HMM produces the correct measures on samples from an Interval Model $I$. We concentrate on the case where $\mu_i$ are distributions over some finite set $X$, but the approach can be extended to other classes, including, for instance, the Gaussians.

First, we show that for a typical sample $x = (x_1, \ldots, x_N)$ of length $N$ from $I$, there is an HMM $H_0$ that assigns a log-likelihood of at least

$$\frac{1}{N} \log P_{H_0}(x) \geq -\frac{\log m}{m} - \sum_{i \leq k} w_i H(\mu_i) \tag{1}$$

to the sample $x$, where $w_i$ are proportions of samples from $\mu_i$ in $x$, and $H(\mu_i)$ are the entropies of the distributions. The precise statement of this result is in Lemma 3.2. As detailed in the proofs, the term $-\sum_{i \leq k} w_i H(\mu_i)$ is the normalized log-likelihood that $I$ itself assigns to a typical sample $x$, and it represents the true likelihood of the data. Note that the error term $-\frac{\log m}{m}$ decreases with increasing $m$. The log-likelihood (1) is achieved on an HMM that has emission distributions $\mu_i$ identical to those of $I$, and the probability of a state change in this HMM is of order $\frac{1}{m}$.

As the next step, we would like to show that if an HMM $H$ has the wrong emission probabilities $\mu_i$, then with high probability over samples $x$, the log-likelihood that $H$ assigns to $x$, $\frac{1}{N} \log P_{H}(x)$, is smaller than (1). By standard arguments this will imply that $H$ can not be the maximum-likelihood estimator, and hence the maximum-likelihood estimator has the correct emission distributions.

We note that $\frac{1}{N} \log P_{H}(x)$ is a somewhat complicated quantity. There are no known methods of maximizing it analytically as a function of $H$, which is the reason why sequential optimization methods, such as the Baum-Welch algorithm, are used to approximate maximum-likelihood estimators in practice. Moreover, for finite $N$, there currently exist no upper bounds on $\frac{1}{N} \log P_{H}(x)$ in the literature that depend explicitly on the parameters of $H$. See Section 2 for a more detailed discussion.

In order to obtain upper bounds on $\frac{1}{N} \log P_{H}(x)$, we use second moments as a proxy for $H$ having emissions $\mu_i$ close to those of $I$. Recall that our sources take values in a finite set $\mathcal{X}$. A second moment of a sample $x$ is an $|\mathcal{X}| \times |\mathcal{X}|$ matrix given by

$$M(x)(a, b) = \frac{1}{N - 1} \left| \{i \leq N - 1 \ | \ x_i = a \land x_{i+1} = b \} \right|, \tag{2}$$

for every $a, b \in \mathcal{X}$. Thus $M(x)(a, b)$ indicates how often the pair $a, b$ appears in $x$. Note that, by definition, $M(x)$ is a probability distribution on the set $\mathcal{X} \times \mathcal{X}$.
The second moment of a random sequence $X = (X_1, \ldots, X_N)$ sampled from $I$ is the expected value, $M_X = \mathbb{E}_{x \sim X} M(x)$ and represents the typical occurrence frequencies for the pair $a, b$. For an HMM with a stationary distribution $\pi$ on the states, transition matrix $p_{ij}$ and emission distributions $\nu_j$, one can define the expected second moment as

$$M_{H, \text{mean}}(a, b) = \sum_{i,j \leq k} \pi_i p_{ij} \nu_i(a) \nu_j(b). \tag{3}$$

The frequencies $M_{H, \text{mean}}(a, b)$ are what one expects to see on a typical sample from $H$, for large $N$. However, $H$ also assigns some non-negligible (on the normalized log scale, $\frac{1}{N} \log \mathbb{P}_H (y)$) likelihood to sequences $y$ with $M(y)$ different from $M_{H, \text{mean}}$ and for our purposes such $M(y)$ have to be considered too. Here and in what follows, we denote by $M_H$ the moment achievable by $H$ which is closest to $M_X$ in a particular sense. The precise definition we require is somewhat technical and will be given in Section 3.

The main result of this paper states the following: Let $M_X$ be the expected second moment of Interval Model $I$ and let $M_H$ be the second moment of an HMM $H$ as discussed above. If

$$\|M_X - M_H\|_{TV} \geq D, \tag{4}$$

then with (exponentially-) high probability over $x$,

$$\frac{1}{N} \log \mathbb{P}_H (x) \leq -D^2 - \sum_{i \leq k} w_i H(\mu_i), \tag{5}$$

where in (4), $M_X$ and $M_H$ are regarded as measures on the set $\mathbb{X} \times \mathbb{X}$, and $\|\cdot\|_{TV}$ is the total variation distance. The detailed statement of this result is given in Theorem 3.3. Combining (4) and (5), we may conclude using standard arguments that the second moment of the maximum likelihood estimator satisfies up to log terms

$$\|M_X - M_H\|_{TV} \leq \frac{1}{\sqrt{m}}. \tag{6}$$

Thus the maximum likelihood estimator approximates the second moment of the data. It is now natural to ask how much information the second moment $M_H$ contains about the emission distributions $\nu_i$ of $H$? In particular, is it true that if $M_X = M_H$ then the model $I$ and $H$ have the same set of emission distributions? In general, the answer to this question is negative. Elegant counterexamples can be found in [Chang, 1996] (see also [Anandkumar et al., 2012]). However, it is also well known and easy to see that the column space of the second moment matrices is spanned by the emission distributions. We will see that a similar statement holds for our, somewhat non-classical matrices $M_X$ and $M_H$. Therefore, if $M_X$ and $M_H$ are known, we can reconstruct the $k$-dimensional

---

1In these examples both $H$ and $X$ are HMMs, and the moments are expected moments. However, the examples can easily be modified to fit our slightly different setting.
subspaces \( \text{span} \{ \mu_j \} \subset \mathbb{R}^{|X|} \) and \( \text{span} \{ \nu_j \} \subset \mathbb{R}^{|X|} \) spanned by emissions of \( I \) and \( H \) respectively. Note that in order to specify a measure on \(|X|\) points one needs \(|X| - 1\) parameters, but if one knows that the measure belongs to a given \(k\)-dimensional subspace, then only \(k - 1\) parameters are required. Since \(k\) is typically much smaller than \(|X|\), this means that the second moment contains most of the information about the emissions.

Finally, we note that our approach can be easily extended to moments higher than two. Indeed, the main combinatorial tool used in this paper is the type theory for second moments of Markov chains as developed in [Csiszár et al., 1987]. This theory has an analog for higher moments (also [Csiszár et al., 1987]). Consequently, while all the essential ideas of our argument are present already in the second moment case, an extension to higher moments can be carried out with only simple technical changes, using the higher moment type theory.

There are two main contributions in this paper. First, we propose a new two stage scheme for clustering time series, which is based on a new model of sequential data. The new model makes significantly lighter assumptions on the data than the standard HMM model, and the classification procedure suggested by the model is significantly more accurate in certain natural cases. Our second and main contribution is a theorem about the behaviour of a maximum likelihood HMM estimator on samples from the Interval Model, as discussed above. In addition, we note that our theorem can be viewed as a statistical consistency result and our method of proof differs significantly from the existing methods for HMM related consistency results in the literature. See Section 2 for the details.

The rest of this paper is organized as follows: In Section 2 we review the literature. Section 3 contains definitions and statements of the results. In Section 4 we discuss several numeric examples and provide the details of the proposed new classification method. The Appendix contains the proofs and in Appendix A.4 we compare standard Viterbi sequence and our proposed classification approach on the S&P 500 stock market index data.

2 LITERATURE

As discussed in the introduction, HMMs are a central tool in sequential data analysis. It is useful to distinguish between two types of applications of HMMs. The first type, to which we refer as a “structural HMM”, contains applications in which the transition matrix of the states plays a crucial role in the modelling of the dynamics of the system. In such systems there may exist, for instance, states that are much more likely to occur if a certain other state occurred one unit of time before them, and are unlikely to occur otherwise. On the other hand, in the second type of applications, which we call the “smoothing HMMs”, there is no reason for a particular pattern of transitions to occur. The transition matrix in these applications is close to the identity matrix, and the main purpose of the Markov process on the states is to keep the system in a fixed state until a strong evidence of state change ap-
pears in the data. In this type of applications, the Markov chain acts more as a tool for change point detection ([Basseville and Nikiforov, 1993]), with a fixed amount of clusters, rather than as a dynamics model. For instance, some applications in Speech Recognition [Gales and Young, 2007], and Biology [Yoon, 2009] belong to the structural type, while many applications in Finance [Mamon and Elliott, 2007], Computer Vision [Horst, 2002], and Activity Recognition (for instance [Trabelsi et al., 2013], [Kim et al., 2010]) belong to the smoothing type. While the approach of our paper may apply to systems of the first type, it is mainly intended for systems of the second type, where a system stays in a given state for long enough intervals and when state change occurs it has no particular statistical regularity. A common theme for both types of the applications, however, is that while the underlying Markov assumption is convenient, it does not necessarily hold for a given data, and is not usually verified. More importantly, as discussed in the introduction, this assumption can result in poor classification performance. Hence, the objective of this paper is to remove this assumption.

Our main result, Theorem 3.3, can be viewed as a statistical consistency result. The classical consistency result for HMMs due to [Baum and Petrie, 1966] (see also [Petrie, 1969] and [Le Gland and Mevel, 2000] for other approaches) states that if \( (X_i)_{i=1}^{\infty} \) is an infinite sample from an HMM \( H \), and \( H_n \) is a sequence of maximum likelihood estimators for the growing sequences \( (X_i)_{i=1}^{n} \), then \( H_n \) converges to \( H \) almost surely. On the other hand, our result shows how HMMs behave on a sample \( (X_i)_{i=1}^{N} \) which is not generated by HMM, but rather by an Interval Model. In addition, the results and methods in the above mentioned literature are essentially of asymptotic character. These methods can not be applied to a finite vector \( (X_i)_{i=1}^{N} \). In contrast, using the approach of this paper one can obtain estimates for a finite, fixed length sample. This is a crucial difference, since even when the model \( I \) is considered with infinite sample sizes, the sequence \( H_n \) of finite sample maximum likelihood HMM approximations does not necessarily converge. All estimates in this paper are essentially inequalities for finite \( N \), and all \( \varepsilon \) that appear in the statements can be written explicitly as functions of \( N \) and other parameters of the problem.

Moments of the data play an important role in our approach. In recent years, moments of the data have been used for parameter estimation in various mixture models. For instance, in [Arora et al., 2012], [Arora et al., 2013], it was shown that for several types of mixture models, the underlying distributions \( \mu_j \) can be inferred from the second moment of the data under a certain “anchor words” assumption on \( \mu_j \)s. In [Anandkumar et al., 2012] it was shown that for a sufficiently large number of samples and under lighter assumptions on \( \mu_j \), the third moment of the data can be used to reconstruct \( \mu_j \) for a variety of mixtures, including the HMM. Note that the use of moments in this paper is different. Our estimator is the classical maximum likelihood estimator rather than an estimator based on moments. We use moments only as a tool to show that properties of the estimator approximate the properties of the true model.

Finally, we make essential use of Type Theory for Markov chains. The results we use are taken from [Csiszár et al., 1987], where second order and
higher order type theory is developed, based on path counting estimates due to Whittle, 1955, Billingsley, 1961.

3 DEFINITIONS AND RESULTS

In Sections 3.1 and 3.2 we give the notions necessary to state the results. Section 3.3 contains the statements and an outline of the proof of the main theorem.

3.1 MODELS

An Interval Model is a tuple \( I = (\{I_l\}_{l \in \mathbb{N}}, \{\mu_i\}_{i=1}^k, \tau, m) \), where \( I_l \) is a sequence of consecutive intervals, \( I_l = [b_l, e_l] \subset \mathbb{N} \), such that \( b_1 = 1 \), and \( b_{l+1} = e_l + 1 \) for all \( l \), \( \mu_i \) are probability measures on a fixed ground set \( \mathcal{X} \), \( \tau : \mathbb{N} \to \{1, \ldots, k\} \) is an assignment of distributions to intervals, and \( m > 0 \) is such that \( |I_l| \geq m \) for all \( l \in \mathbb{N} \). We say that a sequence of random variables with values in \( \mathcal{X} \), \( X = X_1, X_2, \ldots \), is distributed according to interval model \( I \), denoted \( X \sim I \), if \( X_i \) are independent and for every \( l \in \mathbb{N} \) and \( i \in I_l \), \( X_i \) has distribution \( \mu_{\tau(l)} \).

For any finite \( N \), the weights \( \{w_j\} \) are the proportions of each of the states \( \mu_j \) in the data,

\[
    w_j = w_j(N) = \frac{1}{N} |\{i \leq N \mid i \in I_l \land \tau(l) = j\}|. \tag{7}
\]

Although not strictly a necessary requirement, in what follows we assume for convenience that the model \( I \) is such that \( w_j(N) \) converges with \( N \to \infty \) to some limiting values, which will also be denoted \( w_j \).

For each time \( i \in \mathbb{N} \) we define \( \kappa(i) = \tau(l) \) where \( l \) is such that \( i \in I_l \), to be the index of the distribution of \( X_i \).

A Hidden Markov Model, HMM, is a tuple \( H = (S, \{\nu_i\}_{i=1}^k, \{p_{ij}\}_{i,j=1}^k) \) where \( S = \{S_1, \ldots, S_k\} \) is a state space, \( \nu_i \) are corresponding emission probabilities, and \( p_{ij} = P(X_{t+1} = S_j \mid X_t = S_i) \) is the transition matrix.

For a sequence \( x = (x_1, \ldots, x_{N+1}) \), the log-likelihood of \( x \) under the HMM \( H \) with initial distribution \( \pi \) is defined by

\[
    L(x, H, \pi) = \frac{1}{N+1} \log \left( \sum_{s=s_1, \ldots, s_{N+1}} \pi(s_1) \prod_{i=1}^N p_{s_i, s_{i+1}} \prod_{i=1}^{N+1} \nu_{s_i}(x_i) \right), \tag{8}
\]

where the sum is over all possible paths in the underlying Markov chain.

For any finite set \( S \), denote by \( \Delta_S \) the set of all probability measures on a set \( S \).

Finally, for any two probability distributions \( \mu, \nu \in \Delta_{\mathcal{X}} \), define the Kullback-Leibler divergence (or relative entropy) by

\[
    D(\nu \mid \mu) = \sum_{a \in \mathcal{X}} \nu(a) \log \frac{\nu(a)}{\mu(a)}, \tag{9}
\]
and the entropy as
\[ H(\mu) = - \sum_{a \in \mathcal{X}} \mu(a) \log \mu(a). \] (10)

### 3.2 MOMENTS

For a sequence \( x = (x_1, \ldots, x_{N+1}) \), the second moment is a probability distribution \( M(x) \in \Delta_{\mathcal{X} \times \mathcal{X}} \), defined by
\[ M(x)(a, b) = \frac{1}{N} | \{ i \leq N \mid x_i = a \land x_{i+1} = b \} | \] (11)
for all \( a, b \in \mathcal{X} \). The second moment is the frequency of observing consecutive pairs of symbols. For a random vector \( X = (X_1, \ldots, X_{N+1}) \), the second moment is the expected second moment,
\[ M_X = \mathbb{E}_X M(X). \] (12)

If \( X_i \) are independent and have the same distribution \( \mu \), then \( M_X(a, b) = \mu(a) \cdot \mu(b) \). Recall the tensor notation for operators – for any two vectors \( v, w \in \mathbb{R}^\mathcal{X} \), \( v \otimes w \) is a rank 1 linear operator \( \mathbb{R}^\mathcal{X} \to \mathbb{R}^\mathcal{X} \), which acts by \( (v \otimes w)(u) = \langle u, v \rangle \cdot w \) for all \( u \in \mathbb{R}^\mathcal{X} \). In particular, for \( a, b \in \mathcal{X} \) we have \( \langle (v \otimes w)\delta_a, \delta_b \rangle = v(a) \cdot w(b) \).

Hence, the moment \( M_X \) is commonly written as
\[ M_X = \mu \otimes \mu. \] (13)

Next, we compute the expected second moment of an interval model \( I = I(\{I_l\}_{l \in \mathbb{N}}, \{\mu_l\}_{l=1}^k, \tau, m) \). Fix \( N \) and set
\[ c_{rl} = | \{ i < N + 1 \mid \kappa(i) = r \land \kappa(i + 1) = l \} | \] (14)
for the counts of state transitions in the model \( I \) up to time \( N + 1 \). Then, if \( X \sim I \),
\[ M_X = \frac{1}{N} \sum_{r,l \leq k} c_{rl} \mu_r \otimes \mu_l. \] (15)

An important observation is that since every interval in \( I \) is of length at least \( m \), up to time \( N \) there are at most \( N/m \) intervals, and hence at most \( N/m \) transitions between different states. Equivalently,
\[ \frac{1}{N} \sum_{r \neq l} c_{rl} \leq \frac{1}{m}. \] (16)

It is thus useful to think of \( M_X \) roughly as composed from the pure second moments, and an \( 1/m \) distortion term representing mixed moments,
\[ M_X \sim \sum_{r \leq k} w_r \cdot \mu_r \otimes \mu_r + \frac{1}{m} M'. \] (17)
For an HMM $H$ and some initial distribution $\pi \in \Delta_S$, define the expected second moment as the expectation of observing $a, b$ in two steps starting from $\pi$,

$$M_{H,\pi}(a, b) = \sum_{i,j=1}^{k} \pi(i)p_{ij} \cdot \nu_i(a)\nu_j(b). \quad (18)$$

Let us write (18) in a slightly different form.

$$M_{H,\pi}(a, b) = (\pi_1 \nu_1(a), \ldots, \pi_k \nu_k(a)) \left( \begin{array}{ccc} \cdots & \cdots & \cdots \\ p_{ij} & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{array} \right) \left( \begin{array}{c} \nu_1(b) \\ \cdots \\ \nu_k(b) \end{array} \right). \quad (19)$$

For every $a, b \in \mathcal{X}$, we denote the $1 \times k$ vector on the left of (19) as $\phi_a = (\bar{w}_1 \mu_1(a), \ldots, \bar{w}_k \mu_k(a))$, and the $k \times 1$ vector on the right as $\chi_b = (\nu_1(b), \ldots, \nu_k(b))^T$. Equation (19) can then be equivalently rewritten as

$$M_{H,\pi}(a, b) = \phi_a \cdot p \cdot \chi_b, \quad (20)$$

for all $a, b \in \mathcal{X}$. We refer to (20) as a decoupled representation of the moment. Observe that the moment of the data also admits a similar decoupled representation. Indeed, consider the $k \times k$ matrix $u$ with $u_{ij} = \frac{c_{ij}}{\bar{w}_i}$, with $c_{ij}$ defined as in (14). Define also the marginalization

$$\bar{w}_i = \sum_{j \leq k} w_{ij}, \quad (21)$$

and a transition matrix $U$ by $U_{ij} = \frac{u_{ij}}{\bar{w}_i}$. Then we have

$$M_X(a, b) = \phi_a \cdot U \cdot \chi_b \quad (22)$$

for all $a, b \in \mathcal{X}$, where $\phi_a = (\bar{w}_1 \mu_1(a), \ldots, \bar{w}_k \mu_k(a))$, and $\chi_b = (\mu_1(b), \ldots, \mu_k(b))$ as in (20). In other words, for the purposes of computation of the moments, $\bar{w}$ can be considered as a starting probability, and $U$ as the transition matrix of the model $I$.

Next, given an HMM $H = H(S, \{\nu_i\}_{i=1}^{k}, \{p_{ij}\}_{i,j=1}^{k})$, and a set of arbitrary vectors $\phi = \{\phi_a\}_{a \in \mathcal{X}} \in \mathbb{R}^k$, define the generalized second moment of $H$ as a matrix $M_{\phi,H} \in \Delta_{\mathcal{X} \times \mathcal{X}}$ given by

$$M_{\phi,H}(a, b) = \phi_a \cdot p \cdot \chi_b. \quad (23)$$

We call a set of vectors $\phi = \{\phi_a\}_{a \in \mathcal{X}}$ as above proper if all the entries of all $\phi_a$ are non-negative, and

$$\sum_{a \in \mathcal{X}} \sum_{j \leq k} \phi_a(j) = 1. \quad (24)$$
If \( \phi \) is a proper system, define a probability measure \( m_{\phi} \) on \( X \) by

\[
m_{\phi}(a) = \sum_{j \leq k} \phi_a(j).
\]

(25)

We conclude this section by stating the connection between column spaces of \( M_X \) and \( M_{\phi,H} \), and spaces spanned by \( \{\mu_j\}_{j \leq k} \) and \( \{\nu_j\}_{j \leq k} \) respectively.

Note that for any matrix \( M \), the column space of \( M \) coincides with the image of \( M \), \( \text{Im}(M) \), as an operator \( \mathbb{R}^X \to \mathbb{R}^X \).

Lemma 3.1.

1. If \( X \sim I \) for an interval model \( I \). Then \( \text{Im}(M_X) \subset \text{span}\{\mu_j\}_{j \leq k} \).

2. For an HMM \( H \) and an arbitrary set \( \{\phi_a\}_{a \in X} \), \( \text{Im}(M_{\phi,H}) \subset \text{span}\{\nu_j\}_{j \leq k} \).

The proof is given in Appendix A.1.

\[ \text{3.3 RESULTS} \]

As discussed in the Introduction, the following Lemma asserts that there exists an HMM attaining a high likelihood for most samples \( X \) from \( I \).

Lemma 3.2. Given an interval model \( I \), there is an HMM \( H \) and an initial distribution \( \pi \), such that if \( X = (X_1, \ldots, X_N) \sim I \) then with probability at least \( 1 - \varepsilon_N \),

\[
L(X, H, \pi) \geq -\frac{\log km}{m} - \sum_j w_j H(\mu_j) - \varepsilon_N,
\]

(27)

where \( \varepsilon_N \to 0 \) with \( N \to \infty \).

The proof is given in Appendix A.2.

We are now ready to state the main result of this paper. Let \( X = (X_1, \ldots, X_N) \) be generated by model \( I \), and let \( H \) be an HMM. Define

\[
D_\delta = \inf_{\phi \in P_\delta} \|M_X - M_{\phi,H}\|_{TV},
\]

(28)

where

\[
P_\delta = \{ \phi \mid \phi \text{ is proper and } \|m_{\phi} - \bar{M}_X\|_{TV} \leq \delta \}.
\]

(29)

In other words, \( D_\delta \) measures how well \( M_X \) can be approximated by a generalized moment \( M_{\phi,H} \) where \( \phi \) can be any proper system with \( m_{\phi} \) close to the marginal \( \bar{M}_X \). To gain some intuition into this quantity, consider the case when \( D_\delta \) is small, and \( M_X \) has the maximal rank, \( k \). Then, standard matrix perturbation theory results (for instance [Welin, 1972]) imply that \( \text{Im}(M_X) \) is close to \( \text{Im}(M_{\phi,H}) \) and hence \( \text{span}\{\mu_j\}_{j \leq k} \) is close to \( \text{span}\{\nu_j\}_{j \leq k} \) by Lemma 3.1.
Theorem 3.3. For an Interval Model \( I = I(\{I_i\}_{i \in \mathbb{N}}, \{\mu_i\}_{i=1}^k, \tau, m) \), an HMM \( H = H(S, \{\nu_i\}_{i=1}^k, \{p_{ij}\}_{i,j=1}^k) \), and every \( \delta > \frac{1}{m} \) the following holds: If \( D_\delta \geq 2^\delta \), then there is \( r = r_{I,H,\delta} \) such that with probability at least \( 1 - 2^{-rN} \) over \( X \), for every initial distribution \( \pi \),

\[
L(x, H, \pi) \leq -D_\delta^2 - \sum_j w_j H(\mu_j) + \varepsilon_N, \tag{30}
\]

where \( \varepsilon_N \to 0 \) with \( N \to \infty \).

The algorithmic implications of this result were discussed in the introduction and the proof is given in Appendix A.3. Here we briefly describe the main idea of the proof. Define a neighbourhood \( U_\delta \subset \Delta_{X \times X} \) of \( M_X \) by

\[
U_\delta = \{ M \in \Delta_{X \times X} \mid \| M - M_X \|_{TV} \leq \delta \}, \tag{31}
\]

and denote by \( O_\delta \) the set of all sequences \( x = (x_1, \ldots, x_N) \) such that \( M(x) \in U_\delta \). Roughly speaking, the proof of (30) can be seen as a combination of two different uses of type theory. First, the type theory of Markov chains can be used to show that if an HMM \( H \) satisfies (28), then the likelihood given to the full set \( O_\delta \) by \( H \) is at most \( 2^{-ND_\delta^2} \).

\[
P_H (O_\delta) = \sum_{x \in O_\delta} 2^{NL(x, H, \pi)} \leq 2^{-ND_\delta^2}. \tag{32}
\]

On the other hand, the standard type theory for independent sequences together with additional concentration results can be used to show that \( O_\delta \) contains a subset \( X^l \subset O_\delta \) of size at least \( 2^{N \cdot (\sum_j w_j H(\mu_j))} \) such that all \( x \in X^l \) are equiprobable with respect to \( X \) and \( X^l \) is of nearly full measure, \( P_X (X^l) \geq 1 - \varepsilon \). Combining these two statements, one obtains

\[
\frac{1}{|X^l|} \sum_{x \in X^l} P_H (x) \leq \frac{1}{|X^l|} P_H (O_\delta) \leq 2^{-N(D_\delta^2 + \sum_j w_j H(\mu_j))}. \tag{33}
\]

Note that (33) is in fact an averaged version of (30). The corresponding high probability formulation can be easily obtained via Markov’s inequality.

4 EXAMPLES

In Section 4.1 we present two examples which demonstrate two different problems occurring when classification is done using the standard Viterbi sequence. In Section 4.2 we describe an alternative method of classification, the Window HMM, and show how the performance is improved on the above examples. In Appendix A.4 we compare the clustering of the S&P 500 stock market index using the two methods. All the experiments in this section were repeated at least 10 times, and the Baum-Welch algorithm was run with 10 restarts each time.
4.1 MISCLASSIFICATION EXAMPLES

Consider two measures on a set $X = \{1, \ldots, 100\}$, defined as follows: Let $\hat{\mu}_1$ be a uniform measure on $\{1, \ldots, 80\}$, $\hat{\mu}_2$ be uniform on $\{20, \ldots, 100\}$, and $u$ be uniform on the whole $X$. Define the source measures $\mu_1, \mu_2$ by

$$
\mu_1 = \frac{1}{2}(\hat{\mu}_1 + u), \quad \mu_2 = \frac{1}{2}(\hat{\mu}_2 + u),
$$

as shown in Figure 1.

![Figure 1: Measures $\mu_1, \mu_2$ and their empirical counterparts.](image)

To get an idea about the difficulty of distinguishing between these measures from samples, consider the following elementary classifier $E$: Let $x = (x_1, \ldots, x_m)$ be a vector such that either all $x_i$ are sampled independently from $\mu_1$ or all $x_i$ are sample independently form $\mu_2$. Given such $x$, the elementary classifier $E$ computes the likelihoods $L_1(x) = \prod_{i \leq m} \mu_1(x_i)$ and $L_2(x) = \prod_{i \leq m} \mu_2(x_i)$. Then $E$ outputs $E(x) = 1$ if $L_1(x) \geq L_2(x)$ and $E(x) = 2$ otherwise. Given that $x$ is sampled from $\mu_1$, what is the probability that $E(x) = 1$? This probability is pictured in Figure 2 as a function of the sample length $m$, obtained by simulation.

![Figure 2: Classification Accuracy of $E$ as a function of $m$.](image)

In particular, the value at $m = 50$ is $P_{x \sim \mu_1} (E(x) = 1) \approx 0.95$, and the value at $m = 100$ is near 0.98. In other words, to distinguish between $\mu_1$ and $\mu_2$ using $E$ with probability at least 0.95, one needs samples of length $m > 50$.

As our first example, consider an interval model $I$ where all intervals $I_j$ are of length $m = 50$, and the emissions alternate between $\mu_1$ and $\mu_2$. We generate a sample $x = (x_1, \ldots, x_N)$ of length $N = 1000m$ from $I$ and fit an HMM $H$.
with two states to the data $x$, using the standard Baum-Welch algorithm. The emissions found by the algorithm are shown in Figure 1 and approximate the true emissions $\mu_1, \mu_2$ well. The transitions found by the algorithm satisfy $p_{11} \approx p_{22} \approx 0.98$, and represent the correct average waiting time until a state change occurs. Next, we compute the Viterbi sequence given the data, $s = s_1, \ldots, s_N$, where $s_i = 0$ or $s_i = 1$ for each $i$. Let $t = t_1, \ldots, t_N$ be the true sequence of states, that is $t_i = 0$ for $i \in [1, 50]$, $t_i = 1$ for $i \in [51, 100]$ etc. Note that here for convenience we index the states by 0, 1 rather than 1, 2 as in main text. Define the classification error of $s$ to be the fraction of times where $s$ and $t$ differ, possibly after a permutation in assignment of labels to states:

$$e(s) = \min \frac{1}{N} \left\{ \sum_{i \leq N} |t_i - s_i|, \sum_{i \leq N} |t_i - (1 - s_i)| \right\}.$$  

(35)

The error on the Viterbi sequence obtained as described above is $e(s) \approx 0.3$. A fragment of the true sequence and the Viterbi sequence is shown in Figure 3. Note that the maximal possible error is 0.5 and therefore a value of error of 0.3 is quite high.

Figure 3: Example 1. Left: Standard Viterbi Sequence, Right: Window HMM Viterbi. Ground truth in blue, classification results in green.

We mention that if we change $m$ to $m = 100$ in Example 1, the error is $e(s) \approx 0.19$, and we have $e(s) \approx 0.08$ for $m = 200$.

In the second example we consider the error due to changing length of the intervals. Fix a constant $L = 1000$. Let $I$ be an interval model consisting of $2L$ alternating intervals of length $m = 100$, $I_1, \ldots, I_{2L}$, followed by two long intervals, $I_{2L+1}, I_{2L+2}$, such that $|I_{2L+1}| = |I_{2L+2}| = 9 \cdot L \cdot m$. We refer to the intervals $I_1, \ldots, I_{2L}$ as the fast changing fragment, and to $I_{2L+1}, I_{2L+2}$ as the slow changing fragment. The total length of the data is $N = 20 \cdot L \cdot m$, 10% of which is the fast fragment. Fitting an HMM to a sample from this model we again obtain emissions that approximate well $\mu_1$ and $\mu_2$. The Viterbi sequence $s$ has a classification error $e(s) \approx 0.043$. Although this may seem as a small error, a closer look reveals that on the intervals $I_{2L+1}, I_{2L+2}$ the sequence $s$ achieves exact classification, while on the faster changing intervals $I_1, \ldots, I_{2L}$ the error
is $e(s) \approx 0.43$, close to the error of a random guess. The situation is shown in Figure 4. Observe that as mentioned earlier, on an HMM trained on $I_1, \ldots, I_{2L}$ alone, without the slow changing intervals, we have $e(s) \approx 0.19$. Therefore we conclude that the classification performance of the standard Viterbi sequence decreases when multiple time scales are present in the data.

![Figure 4: Viterbi classification, Example 2. Left: full data, Right: zoom on the fast changes fragment.](image)

### 4.2 WINDOW CLASSIFICATION

In this section we propose a classification method motivated by the structure of the Interval Model. We assume that the data $x = (x_1, \ldots, x_N)$ to be clustered comes from some Interval Model $I = I(\{I_l\}_{l \in \mathbb{N}}, \{\mu_l\}_{l=1}^k, \tau, m_0)$, and that the measures $\mu_1, \ldots, \mu_k$ are known. In addition, we assume that the parameter $m_0$ of the model is large enough so that classifier $E$ on samples of length $m_0$ distinguishes between the sources with probability at least $0.95$. Specifically, let $m$ be the minimal integer such that for every $i \leq k$,

$$
\mathbb{P}_{y \sim \mu_i} \left( \{ y \mid \mu_i(y) \geq \mu_j(y) \text{ for all } j \neq i \} \right) \geq 0.95,
$$

where $y = (y_1, \ldots, y_m)$ is an i.i.d sample of length $m$ from $\mu_i$. Note that given $\mu_1, \ldots, \mu_k$ such $m$ can be easily found by simulation. We find such $m$ as the first step of the classification, and we assume that $m_0 \geq m$.

Next, given a sample $x = (x_1, \ldots, x_N)$, for every $i \leq N - m$ construct a window sample $w_i = (x_{i+m}, x_{i+m-1})$. We call a window $w_i$ pure if it is contained in a single interval, that is $[i, i + m) \subset I_l$ for some $l \in \mathbb{N}$. Observe that if a window $w_i$ is pure, then (36) implies that we can correctly classify the samples $x_i, \ldots, x_{i+m-1}$ with probability at least 0.95. Note that if $m_0$ is significantly larger than $m$, this observation is sufficient to classify most of the samples correctly (indeed, for any $m_0$, at least $\frac{m_0-m}{m_0} N$ of the windows $w_i$ are pure). In what follows we refine this observation by modelling also non-pure

---

2 The value 0.95 is a convenient practical choice, but one can use other values depending on the data and application requirements.
windows. This refinement allows us to do classification in cases such as Example 1 in Section 4.1, where \( m_0 \) is not necessarily large compared to \( m \).

Consider a non-pure window \( w_i = (x_i, \ldots, x_{i+m-1}) \), such that \( [i, i + t] \subset I_l \) and \( [i + t, i + m) \subset I_{l+1} \) for some \( t > 0 \). The expected empirical distribution of this window is a convex combination of distributions corresponding to \( I_l \) and \( I_{l+1} \),

\[
E \frac{1}{m} \sum_{j \in [i, i+m)} \delta_{x_j} = \frac{t}{m} \mu_\tau(l) + \frac{m-t}{m} \mu_\tau(l+1).
\]

Thus, as the index \( i \) moves between \( I_l \) and \( I_{l+1} \), the distribution of the window interpolates linearly between \( \mu_\tau(l) \) and \( \mu_\tau(l+1) \). We model this dynamics by an HMM which we call a Window HMM, and which we now define. See the concluding remarks in this section for a discussion of the difference between classification using Window HMM and HMM.

A Window HMM \( W \) for measures \( \mu_1, \ldots, \mu_k \) and window length \( m \) is an HMM with \( k + k \cdot (k-1) \cdot (m-1) \) states, \( S = \{a_i\}_{i=1}^k \cup \{T_{ij}^l\} \), where \( a_i \) are pure states, and \( T_{ij}^l \) for \( i, j \leq k, i \neq j, \) and \( 1 \leq l \leq m-1 \) are transition states. The alphabet for the emissions of \( W \) consists of \( m \)-tuples over \( \mathcal{X} \), and the emission distributions are products of \( \mu_j \)'s or of their linear combinations. Specifically, for an \( m \)-tuple \( y = (y_1, \ldots, y_m) \),

\[
P_W (y \mid a_i) = \prod_{s=1}^m \mu_i(y_s) \tag{38}
\]

for every pure state \( a_i \), and

\[
P_W (y \mid T_{ij}^l) = \prod_{s=1}^m \left( \frac{m-l}{m} \mu_i + \frac{l}{m} \mu_j \right)(y_s) \tag{39}
\]

for every transition state \( T_{ij}^l \). Given an input sequence \( x = (x_1, \ldots, x_N) \), we construct the sequence of windows \( \{w_i\}_{i=1}^{N-m} \) and feed it as input to the HMM \( W \). Thus, the states \( a_i \) of \( W \) model the pure windows, and states \( T_{ij}^l \) model intermediate windows on the intersection between intervals of \( \mu_i \) and \( \mu_j \). The transition probabilities for states of \( W \) model the transition of the windows between intervals. Specifically, choose some \( p_0 \in [0, 1] \) and set

\[
P_W (a_i \mid a_i) = p_0, \quad P_W \left(T_{ij}^l \mid a_i\right) = \frac{1-p_0}{k-1} \tag{40}
\]

for all \( i, j \leq k \),

\[
P_W \left(T_{i+1}^l \mid T_{ij}^l\right) = 1 \tag{41}
\]

for all \( i, j \leq k \) and \( l \leq m-2 \), and

\[
P_W \left(a_j \mid T_{ij}^{m-1}\right) = 1 \tag{42}
\]

for all \( i, j \leq k \). Figure 5 illustrates the transitions for the case \( k = 2 \) and \( m = 2 \). Thus the paths of \( W \) either stay at a single pure node or move between pure
nodes $a_i$ and $a_j$ via a sequence of states $T_{ij}^1, \ldots, T_{ij}^{m-1}$, simulating the transition of a window between intervals. As we discuss later in this section, the precise value of $p_0$ has practically no influence on the performance of the final algorithm. We choose $p_0 = 0.9$ in all experiments.

The full classification algorithm is as follows: Given a sample $x = (x_1, \ldots, x_N)$, and measures $\{\mu_i\}_{i \leq k}$, find $m$ that satisfies (36), construct the Window HMM $W$ and window sequence $w_i$. Find the Viterbi sequence of states $s = s_1, \ldots, s_{N-m}$ for the HMM $W$ and the input data $w_1, \ldots, w_{N_m}$. For each $i \leq N - m$, classify index $i$ as a source $j$ if $s_i = a_j$ or $s_i = T_{jl}^h$ for some $h \leq k$ and $l \leq m - 1$.

We now describe the performance of the Window HMM classification on Examples 1 and 2 of Section 4.1. For Example 1, we have used window length $m = 50$, which satisfies the condition (36). The classification using $W$ achieves an error $e \approx 0.06$. This is a significant improvement over $e \approx 0.3$ for the standard Viterbi sequence. On Example 1, Window HMM switches between the states more often than standard HMM, as shown in Figure 3.

For Example 2 we also used $W$ with window $m = 50$. With this $m$, $W$ has an error $e \approx 0.04$. Note that this is practically the same error as the error achieved by the standard HMM in Section 4.1 However, $W$ errs differently. It is almost exact on the initial fast changing sequence, but has some errors on the long intervals. Figure 6 shows this and should be compared with Figure 4. Note that the error of the order 0.04 for $W$ is not surprising, since one can not perfectly distinguish between samples from $\mu_1$ and $\mu_2$ using window of length $m = 100$. Indeed, as we have seen in Section 4.1, even when the data is generated from by a single distribution, the classifier $E$ with $m = 100$ has error 0.05.

We conclude that the Window HMM is significantly more sensitive than the standard Viterbi sequence to the changes between the data sources, and does not depend on the change of time scale, but can introduce a small constant error rate due to possible missclassifications for a fixed window size.

Since both standard HMM and Window HMM eventually use the Viterbi sequence, one can ask why the approaches differ? We conclude this section by discussing a crucial difference in a way Window HMM and a standard Viterbi sequence perform classification. Consider first an HMM $H$ with two states and coin toss emission probabilities. Specifically, for any $\varepsilon, \delta < 1$, let $\mathcal{X} = \{0, 1\}$, set

![Figure 5: Transition Structure of Window HMM](image)

- $a_1$
- $a_2$
- $a_3$
- $a_4$
- $a_5$
- $a_6$
- $a_7$
- $a_8$
- $a_9$
- $a_{10}$
- $a_{11}$
- $a_{12}$
- $a_{13}$
- $a_{14}$
- $a_{15}$
- $a_{16}$

- $T_{12}^1$
- $T_{23}^2$
- $T_{34}^3$
- $T_{45}^4$
- $T_{56}^5$
- $T_{67}^6$
- $T_{78}^7$
- $T_{89}^8$
- $T_{910}^9$
- $T_{1011}^{10}$
- $T_{1112}^{11}$
- $T_{1213}^{12}$
- $T_{1314}^{13}$
- $T_{1415}^{14}$
- $T_{1516}^{15}$
- $T_{161}^{16}$
Figure 6: Window classification, Example 2. Left: full data, Right: zoom on the fast changes fragment.

\[ \mu_1(0) = 1 - \delta, \mu_2(0) = \delta \]

and let

\[ p = \left( \begin{array}{cc} 1 - \varepsilon, \varepsilon \\ \varepsilon, 1 - \varepsilon \end{array} \right) \]

be the transition matrix. Let \( \pi \) such that \( \pi(0) = 1 \) be the initial distribution, and let \( x = (0, 0, 0, 0, 0, 1, 1, 1, 1) \) be the data sequence. If the two coins are similar, say \( \delta = 0.4 \), and \( \varepsilon \) is somewhat small, then the Viterbi sequence will be \( s_0 = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1) \) despite the change of value in the data. For concreteness, it is easy to check that this holds for \( \delta = 0.4 \) and every \( \varepsilon < 0.1 \). We can say that in this case, 1 appearing in the data for 5 time units is not a sufficient evidence for the HMM to change state. This is precisely what happens when one uses the standard Viterbi sequence in Examples 1 and 2, and on some fragments of the SP500 data discussed in Appendix A.4. Let \( s_1 \) be the true sequence \( s_1 = (1, 1, 1, 2, 2, 2, 2, 2, 2, 2) \). By writing explicitly the likelihoods for paths \( s_0, s_1 \),

\[ P_H(s_0) = (1 - \varepsilon)^9 \cdot \mu_1^5(0) \mu_1^4(1), \]

\[ P_H(s_1) = (1 - \varepsilon)^8 \cdot \varepsilon \cdot \mu_2^2(0) \mu_2^3(1), \]

we see that path \( s_1 \) is preferred when \( (\mu_2(1)/\mu_1(1))^5 \geq \varepsilon \). We call the ratio \( r_x = \mu_2(1)/\mu_1(1) \) the evidence of sample 1 about the state. Note that as coins become more different, \( \delta \to 0 \), \( r_1 \) grows to infinity. We can formulate this observation as follows: The more evidence an individual sample contains about the state, the less the Viterbi path depends on the transition probabilities between the states (\( \varepsilon \) in this case).

Returning to the Window HMM classifier \( W \), note that a single sample for \( W \) is a window \( w_i = (x_i, \ldots, x_{i+m-1}) \) of length \( m \). By standard information theoretic arguments, if the window is pure and generated from some \( \mu_j \), then the ratio

\[ \prod_{s \leq m-1} \mu_j(x_{i+s})/ \prod_{s \leq m-1} \mu_h(x_{i+s}) \]

17
grows to infinity exponentially with $m$, for any $h \neq i$, with high probability over $x$. Hence each sample for $W$ contains strong evidence about the state. In particular, for pure windows the state chosen by the Viterbi path of $W$ is chosen essentially by the maximum likelihood criterion, independently of the transition parameters of $W$. This is in contrast to the classification using the standard Viterbi sequence, where dependence on the transitions is stronger and can lead to missclassifications as in Examples 1 and 2.

References

[Anandkumar et al., 2012] Anandkumar, A., Hsu, D., and Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. In COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland, pages 33.1–33.34.

[Arora et al., 2013] Arora, S., Ge, R., Halpern, Y., Mimno, D. M., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013.

[Arora et al., 2012] Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models – going beyond svd. In Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, FOCS '12, pages 1–10, Washington, DC, USA. IEEE Computer Society.

[Basseville and Nikiforov, 1993] Basseville, M. and Nikiforov, I. V. (1993). Detection of Abrupt Changes: Theory and Application.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. Ann. Math. Statist., 37(6).

[Billingsley, 1961] Billingsley, P. (1961). Statistical methods in markov chains. Ann. Math. Statist., 32(1).

[Chang, 1996] Chang, J. T. (1996). Full reconstruction of markov models on evolutionary trees: Identifiability and consistency. Mathematical Biosciences, 137.

[Cover and Thomas, 2006] Cover, T. M. and Thomas, J. A. (2006). Elements of Information Theory.

[Csiszár et al., 1987] Csiszár, I., Cover, T. M., and Choi, B.-S. (1987). Conditional limit theorems under markov conditioning. IEEE Transactions on Information Theory, 33(6).
[Dvoretzky et al., 1956] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27(3).

[Gales and Young, 2007] Gales, M. and Young, S. (2007). The application of hidden markov models in speech recognition. *Found. Trends Signal Process.*, 1(3).

[Horst, 2002] Horst, B. (2002). *Hidden Markov Models: Applications in Computer Vision*.

[Kim et al., 2010] Kim, E., Helal, S., and Cook, D. (2010). Human activity recognition and pattern discovery. *Pervasive Computing, IEEE*, 9(1).

[Le Gland and Mevel, 2000] Le Gland, F. and Mevel, L. (2000). Exponential forgetting and geometric ergodicity in hidden markov models. *Mathematics of Control, Signals and Systems*, 13(1).

[Mamon and Elliott, 2007] Mamon, R. and Elliott, R. (2007). *Hidden Markov Models in Finance*.

[Petrie, 1969] Petrie, T. (1969). Probabilistic functions of finite state markov chains. *Ann. Math. Statist.*, 40(1).

[Sanov, 1957] Sanov, I. (1957). On the probability of large deviations of random variables. *Mut. Sb., vol. 42, pp. 11-44, 1957 (in Russian). English translation in Sel. Transl. Math. Statist. Probab., vol. 1, 1961*.

[Trabelsi et al., 2013] Trabelsi, D., Mohammed, S., Chamroukhi, F., Oukhellou, L., and Amirat, Y. (2013). An unsupervised approach for automatic activity recognition based on hidden markov model regression. *Automation Science and Engineering, IEEE Transactions on*, 10(3).

[Wedin, 1972] Wedin, P.-. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

[Whittle, 1955] Whittle, P. (1955). Some distributions and moment formulae for the markov chain. *J. Roy. Stat. Sot., Ser. B*, 17.

[Yoon, 2009] Yoon, B.-J. (2009). Hidden markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6).
A APPENDIX

A.1 PROOF OF LEMMA 3.1

Proof. The statement for $M_X$ follows directly from (15). To show the statement for $M_{\phi,H}$, for any $\phi_a \in \mathbb{R}^k$ and $i \leq k$ let $\phi_i(a)$ be the $i$-th coordinate of $\phi_a$. For every $i \leq k$, define $\hat{\phi}_i \in \mathbb{R}^X$ by $\hat{\phi}_i(a) = \phi_a(i)$. Then by the definition, (23),

\[ M_{H,\pi}(a, b) = \sum_{i \leq k} \sum_{j \leq k} p_{ij} \phi_i(a) \nu_j(b), \tag{47} \]

and hence

\[ M_{H,\pi} = \sum_{i,j \leq k} p_{ij} \hat{\phi}_i \otimes \nu_j, \tag{48} \]

therefore concluding the proof. $\square$

A.2 PROOF OF LEMMA 3.2

Proof. Consider an HMM $H$ with $k$ states, $S = \{1, \ldots, k\}$, with emission probabilities equal to those of the model $I$, $\mu_i$, and some transition matrix, $p_{ij}$. In order to show the lower bound on $L$, it suffices to consider a single path of the HMM. Let $s = s_1, \ldots, s_N$ be a sequence of states of $H$ that follows precisely the sequence of states in $I$, so that $s_i = \kappa(i)$. Let the initial distribution $\pi$ be a delta measure concentrated on the first state of $I$, $\kappa(1)$. Recall that the likelihood $L(x, H, \pi) = \frac{1}{N} \log P_{H,\pi}(x)$ is given by a sum (8). The contribution of a single path $s$ is this sum is

\[ \frac{1}{N} \sum_{i,j \leq k} c_{ij} \log p_{ij} + \frac{1}{N} \sum_{i=1}^k \sum_{j} \log \mu_j(x_i^j), \tag{49} \]

where $c_{ij}$ are the transition counts of the model $I$, as in (14), and $x_i^j$ are the entries of $X$, rearranged so that for all $i$, $x_i^j$ are entries sampled from $\mu_j$. Consider the second term first,

\[ \frac{1}{N} \sum_{j=1}^k \sum_{i} \log \mu_j(x_i^j) = \frac{1}{N} \sum_{j=1}^k \frac{c_{jj}}{c_{jj}} \sum_{i} \log \mu_j(x_i^j). \tag{50} \]

Clearly, by the law of large numbers, $\frac{1}{c_{jj}} \sum_i \log \mu_j(x_i^j) \to -H(\mu_j)$ with $N \to \infty$, and we assume here that $\frac{c_{jj}}{c_{jj}}$ has a limit value $w_j$.

Next, the first term in (49),

\[ \frac{1}{N} \sum_{i,k \leq k} c_{ij} \log p_{ij} \tag{51} \]

controls the underlying Markov chain probability of the path $s$. Since, the total number of transitions between different states in the model is small, see (16).
this probability is large when $p_{ii}$ are close to 1. In particular, by choosing $p_{ii} = 1 - \frac{1}{m}$ and $p_{ij} = \frac{1}{(k-1)m}$ for all $i$ and $j \neq i$, we obtain
\[
\frac{1}{N} \sum_{i,k \leq k} c_{ij} \log p_{ij} \leq -\frac{\log km}{m}.
\]

(A.3) PROOF OF THEOREM 3.3

Let $x = (x_1, \ldots, x_N)$ be distributed according to an interval model $I$. First we show that the empirical second moment of a sample $x$ is close to its expected second moment, $M_X$. Specifically, for an appropriate total variation neighbourhood $U$ of $M_X$ in $\Delta_{X \times X}$, we show that $M(x) \in U$. Since $X$ is composed mostly from chunks of i.i.d samples, this can be obtained by a variety of standard results. We give an elementary argument in Lemma A.1. Note however, that by taking $N$ large enough, one can ensure that the pure components, $\mu_r \otimes \mu_r$, as defined in (15), are well approximated. On the other hand, as noted above, (16), $X$ contains relatively few mixed transitions (that is, $X_i \sim \mu_r$ and $X_{i+1} \sim \mu_l$ for $r \neq l$). We therefore do not expect $\mu_r \otimes \mu_l$ to be sampled correctly, and bound them by their total mass, which gives rise to the term $1/m$ in (53).

**Lemma A.1.** Let $X = (X_1, \ldots, X_N)$ be distributed according to the interval model $I = I(\{I_l\}_{l \in N}, \{\mu_i\}_{k i=1}^r, \tau, m)$. Then
\[
\Pr_X \left( \|M(x) - M_X\|_{TV} \geq \varepsilon + \frac{1}{m} \right) \leq 2^{-c \frac{\varepsilon^2}{\log \|X\|}}
\]
where $c > 0$ is an absolute constant.

**Proof.** We consider samples from each state separately first. Let
\[
A_j = \{i \leq N + 1 \mid \kappa(i) = j\}
\]
be the set of indices $i$ such that $X_i \sim \mu_j$. Divide $A_j$ into a set of odd pairs and even pairs as follows:
\[
A_j^1 = \{(i, i+1) \mid i \in A_j, \text{ i is odd}\},
\]
\[
A_j^2 = \{(i, i+1) \mid i \in A_j, \text{ i is even}\}.
\]
For instance, if $(1, 2, 3, 4, 5, 6) \subseteq A_j$, then $(1, 2), (3, 4), (5, 6)$ are odd pairs, and $(2, 3), (4, 5)$ are even. Then the pairs in each $A_j^t$ are mutually independent. Hence they can be considered i.i.d samples from the measure $\mu_j \times \mu_j$. To estimate how well independent empirical samples of $\mu_j \times \mu_j$ approximate $\mu_j \times \mu_j$, we use the Dvoretzky Kiefer Wolfowitz inequality, [Dvoretzky et al., 1956], which bounds the sup distance between the empirical and true distribution. For $a, b \in X$, let
\[
S(a, b) = \mu_j(a)\mu_j(b) - \frac{1}{|A_j^t|} \sum_{(i, i+1) \in A_j^t} \delta(X_i = a) \cdot \delta(X_{i+1} = b)
\]

(57)
be the difference between the empirical and the true measures. Then
\[
P \left( \max_{a,b \in \mathcal{X}} |S(a, b)| \geq \frac{\varepsilon}{|\mathcal{X}|^2} \right) \leq 2^{-c|A| |\mathcal{X}|^2}. \tag{58}
\]
Combining the estimates (58) for all \( A_j^1, A_j^2 \) via the union bound, summing over all \( a, b \in \mathcal{X} \) obtain the TV distance, and recalling that the proportion of transitions where \( \kappa(i) \neq \kappa(i + 1) \) is at most \( 1/m \) completes the proof. \( \square \)

Define
\[
U_{\varepsilon} = \left\{ M \in \Delta_{\mathcal{X} \times \mathcal{X}} \mid \|M - M_{\mathcal{X}}\|_{TV} \leq \varepsilon + \frac{1}{m} \right\}. \tag{59}
\]

Lemma A.1 states that \( M(x) \in U_{\varepsilon} \) with high probability over \( I \).

Next, fix an HMM \( H \) and \( \varepsilon > 0 \), and let \( M_H = M_{\phi,H} \) be the generalized second moment attaining the infimum in (28) with \( \delta = \varepsilon + \frac{1}{m} \). Thus
\[
D = D_{\delta} = \|M_{\mathcal{X}} - M_H\|_{TV} = \inf_{\phi \in P_\delta} \|M_{\mathcal{X}} - M_{\phi,H}\|_{TV}. \tag{60}
\]

Let
\[
O_{\varepsilon} = \{ y = y_1, \ldots, y_{N+1} \mid M(y) \in U_{\varepsilon} \} \tag{61}
\]
be the set of all data sequences \( y \) with \( M(y) \in U_{\varepsilon} \). Using the type theory for second moments of Markov chains, we will show that (60) implies
\[
P_{H,\pi}(O_{\varepsilon}) \leq 2^{-ND^2}, \tag{62}
\]
for every initial distribution \( \pi \). Equivalently, under \( H \), probability of observing a sequence \( y \) with \( M(y) \in U_{\varepsilon} \) is at most \( 2^{-ND^2} \). We first prove Theorem 3.3 assuming (62), and then prove (62).

The inequality (62) bounds the likelihood under \( H \) of all \( y \) such that \( M(y) \in U_{\varepsilon} \). To prove Theorem 3.3 we need to bound the likelihood \( L(x, H, \pi) \) of individual sequence \( x \) produced by \( I \). Let
\[
H(X) = H(X_1, \ldots, X_N) = N \sum_{i \leq k} w_i H(\mu_i) \tag{63}
\]
be the entropy of a sample \( X = (X_1, \ldots, X_N) \) from \( I \).

**Lemma A.2.** Let \( X = (X_1, \ldots, X_N) \sim I \). Then there exist at least at least \( 2^{N(H(X) - \varepsilon_N)} \) different sequences \( x^l = (x_1, \ldots, x_N) \), such that for every such \( x \),
\[
\left| -\frac{1}{N} \log P_X(x^l) - H(X) \right| \leq \varepsilon_N \tag{64}
\]
and
\[
\sum_l P_X(x^l) \geq 1 - \varepsilon_N, \tag{65}
\]
where \( \varepsilon_N \to 0 \) with \( N \to \infty \).

22
This Lemma is the standard Asymptotic Equipartition Property (AEP) for independent variables (see [Cover and Thomas, 2006]), the proof is omitted. Note that on one hand sequences $x^l$ are a set of almost full probability by (65), and on the other hand, by Lemma A.1, $\mathbb{P}_X(M(x) \in U_\varepsilon)$ is almost full. Therefore, possibly by slightly changing $\varepsilon_N$, we may assume that sequences $x^l$ also satisfy $M(x^l) \in U_\varepsilon$, or equivalently, $x_1 \in O_\varepsilon$. Denote by $X^l$ the set of sequences $x^l$.

Since the sequences in $X^l$ are distinct and $x^l \in O_\varepsilon$, using (62) we can write
\[
\sum_l \mathbb{P}_H(x^l) \leq \mathbb{P}_H(O_\varepsilon) \leq 2^{-N D^2}.
\]
Or equivalently,
\[
\frac{1}{|X^l|} \sum_l \mathbb{P}_H(x^l) \leq 2^{-N(D^2 + \frac{1}{2} \log |X^l|)}.
\]
Since by Lemma A.2 $|X^l| \geq 2^{N(H(X) - \varepsilon_N)}$, we obtain
\[
\frac{1}{|X^l|} \sum_l \mathbb{P}_H(x^l) \leq 2^{-N(D^2 + \sum_{i \leq k} w_i H(\mu_i) - \varepsilon_N)}.
\]
Note that (68) is essentially the statement of Theorem 3.3 on average over $x^l$. It remains to apply Markov inequality to this average to obtain
\[
\mathbb{P}_X \left( L(x, H) \geq -D^2/2 - \sum_j w_j H(\mu_j) + \varepsilon_N \right) \leq 2^{-N \frac{D^2}{2}},
\]
therefore concluding the proof of Theorem 3.3, up to the the proof of (62). Note the uniform average over $|X^l|$ in (68) was replaced with probability over $X$ in (69). This is possible since in Lemma A.2 the size of $|X^l|$ depends on the size of the maximal type class, and $x^l$ inside the same type class are equiprobable.

To prove (62), we use a standard construction to transform an HMMs into a Markov chain in a special form. This converts the problem of bounding the likelihood of data under an HMM to a problem of bounding a likelihood of a certain set of paths in the chain. Indeed, given an HMM $H = (S, \{\nu_i\}_{i=1}^k, \{p_{ij}\}_{i,j=1}^k)$, construct a Markov chain $H' = (S', p')$ with state space $S' = S \times \mathbb{X}$, and transition probabilities
\[
p'_{(i,a),(j,b)} = p_{ij} \nu_j(b).
\]
For a state $(i, a) \in S'$, we refer to $a$ the data component of the state. Clearly, by observing a random walk of $H'$ and looking only at the data component, we get a distribution over the data that is identical to that of the HMM. Note that for a single data vector $x = (x_1, \ldots, x_{N+1})$, there are exactly $k^{N+1}$ paths of the chain $H'$ yielding the data $x$.

Next, we use type theory for Markov chains to obtain deviation bounds on the empirical second moment of a random walk. Similarly to second moment
of the data, for a Markov chain $H' = (S', p')$, and a path $s = s_1, s_2, \ldots, s_{N+1}$, where $s_i \in S'$, define the second moment $M(s) \in \Delta_{S' \times S'}$ by

$$M(s)(u, v) = \frac{1}{N} \left| \{i \leq N \mid s_i = u \land s_{i+1} = v \} \right|,$$

(71)

for all $u, v \in S'$. For a subset $\Pi \subset \Delta_{S' \times S'}$, the second order type theory provides bounds of the form

$$P_{H'} (M(s) \in \Pi) \leq 2^{-N \cdot D},$$

(73)

where $D$ is a suitably defined distance between the set $\Pi$ and the transition matrix $p'$. Statement (73) is a Markov chain analog of Sanov’s theorem for i.i.d sequences ([Sanov, 1957], [Cover and Thomas, 2006]). We use a second moment deviation inequality due to [Csiszár et al., 1987], stated as Lemma A.3. Note that type theory provides estimates on moments of paths of the chain $H'$, which take values in $\Delta_{S' \times S'}$, while our assumptions are about moments of the data, $M(x) \in \Delta_{X \times X}$. The connection between the two moments is given by a linear transformation. Indeed, consider a Markov chain $H' = (S', p')$ corresponding to an HMM $H$. Define a linear map $T : \Delta_{S' \times S'} \to \Delta_{X \times X}$ by

$$T(M')(a, b) = \sum_{i,j \leq k} M'((i, a), (j, b)).$$

(74)

If $M'$ is the second moment of a path of the chain, then $T(M')$ is the second moment of the data. The following inequality follows from the chain rule for relative entropy (see [Cover and Thomas, 2006]): For any $M_1, M_2 \in \Delta_{S' \times S'}$,

$$D(T(M_1)|T(M_2)) \leq D(M_1, M_2).$$

(75)

To state the deviation result, Lemma A.3 we require some additional notation. For any measure $M \in \Delta_{S' \times S'}$, define the marginalizations $\bar{M}, \bar{\bar{M}} \in \Delta_{S'}$ by

$$\bar{M}(u) = \sum_{v \in S'} M(u, v), \quad \bar{\bar{M}}(u) = \sum_{v \in S'} M(v, u).$$

(66)

Moreover, given $M \in \Delta_{S' \times S'}$, define the related transition matrix to be

$$M(v|u) = \frac{M(u, v)}{\bar{M}(u)}.$$

(67)

A measure $M \in \Delta_{S' \times S'}$ is called stationary, if $\bar{M} = \bar{\bar{M}}$. Such measure is a stationary measure of a random walk given by the transition matrix $M(v|u)$. We denote by $\Delta_{S' \times S'}^0$ the set of all stationary measures.

Finally, we introduce a quantity that will control the deviations of moments. Given a transition matrix $p' = p_{uv}$ and a measure $M \in \Delta_{S' \times S'}$, define

$$D(M|p') = \sum_{u,v \in S'} M(u, v) \log \frac{M(u, v)}{p_{uv}} = \frac{\sum_{u,v \in S'} M(u, v) \log M(u, v)}{\sum_{u,v \in S'} M(u, v) p_{uv}}.$$

(78)

(79)
The quantity $D(M|p')$ differs from the standard Kullback-Leibler divergence since $p'$ is not a measure. However, as follows from (74), we can write $D(M|p') = D(M|z)$, where $D(t|z)$ is the standard KL divergence and $z \in \Delta_{S' \times S'}$ is defined by $z(u, v) = \bar{M}(u) \cdot p'_uv$.

For any closed set $\Pi \subset \Delta_{S' \times S'}$, denote $\Pi_0 = \Delta^0_{S' \times S'} \cap \Pi$.

**Lemma A.3** ([Csiszar et al., 1987]). Let $C = (S', p')$ be a Markov chain and let $X = X_1, \ldots, X_{N+1}$ be a random walk generated by the chain. For a closed convex $\Pi \subset \Delta_{S' \times S'}$, set

$$
D' = \min_{M \in \Pi_0} D(M|p').
$$

Then for any $\varepsilon \geq 0$, there is $N_\varepsilon$, such that for all $N \geq N_\varepsilon$,

$$
P_C(M(X) \in \Pi) \leq 2^{-N(D' - \varepsilon)}.
$$

We remark that although not stated explicitly in [Csiszar et al., 1987], the dependence of $N_\varepsilon$ on $\varepsilon$ can be made explicit.

Lemma A.3 provides us with likelihood estimates that depend on the parameters of the unfolded Markov chain $H'$. To obtain the bound (62) for an HMM $H$ we apply Lemma A.3 to the Markov chain $H'$ with the set $\Pi \subset \Delta_{S' \times S'}$ given by

$$
\Pi = T^{-1}(U_\varepsilon) = \{M' \mid T(M') \in U_\varepsilon\},
$$

where $U_\varepsilon$ was defined in (59).

Choose some $M' \in \Pi$. In what follows we show that if $D$ is given by (60), then

$$
D(T(M')|T(M)p')) \geq 2D^2.
$$

Note that by (75) this implies

$$
D(M'|\bar{M}'p') \geq 2D^2,
$$

and hence $D' \geq 2D^2$ in Lemma A.3 therefore proving (62).

Next, to obtain (83), observe that by Pinsker’s Inequality (see [Cover and Thomas, 2006]), it is sufficient to show that

$$
\|T(M') - T(\bar{M}'p')\|_{TV} \geq D.
$$

Recall that by definition $T(M') \in U_\varepsilon$, and hence

$$
\|T(M') - M_X\|_{TV} \leq \varepsilon + \frac{1}{m}.
$$

Thus to obtain (85) it is sufficient to show that

$$
\|M_X - T(\bar{M}'p')\|_{TV} \geq D - \varepsilon - \frac{1}{m}.
$$

\[\text{Pinsker’s Inequality: } 2 \|\mu - \nu\|_{TV}^2 \leq D(\mu|\nu) \text{ for all measures } \mu, \nu.\]

25
Let us now write the explicit expression for $T(\bar{M}'p')$.

\[ T(\bar{M}'p')(a, b) = \sum_{i,j \leq k} \bar{M}'((i,a))p_j(i,a, i, j, b) = \sum_{i,j \leq k} \bar{M}'((i,a))p_{ij} \mu_j(b). \]  

(88)

(89)

In addition, observe that by definition,

\[ \sum_{i \leq k} \bar{M}''((i,a)) = \sum_{i \leq k} \sum_{j \leq k} \sum_{b \in \mathcal{X}} M''((i,a), (j,b)) = \sum_{b \in \mathcal{X}} M(a, b) = \bar{M}(a). \]  

(90)

(91)

(92)

For every $a, b \in \mathcal{X}$, denote $\phi_a = (M'(1,a) , \ldots , M'(k,a)) \in \mathbb{R}^k$, and $\chi_b = (\nu_1(b), \ldots , \nu_k(b))$. Then we can rewrite (88) in the decoupled representation form (see (20)) as

\[ T(\bar{M}'p')(a, b) = \phi_a \cdot p \cdot \chi_b. \]  

(93)

Moreover, since $T(M') \in U_\varepsilon$, the marginals satisfy

\[ \|\bar{M}_X - \bar{M}\|_{TV} \leq \varepsilon + \frac{1}{m}. \]  

(94)

Therefore, by (92) and (94), the condition (28) for $\phi_a$ in the main Theorem holds.

### A.4 S&P 500 DATA

In this section we consider the logarithmic daily returns of the S&P 500 stock market index for the years 1950 to 2014. This data can be obtained, for instance, from Yahoo Finance website. We fit an HMM with 3 states to the data, to obtain the emission distributions $\mu_1, \mu_2, \mu_3$ (blue, red and green respectively in Figures 7, 8). The distribution functions of the resulting distributions are shown in Figure 7.

All three distributions are essentially mean zero, and have different standard deviations, as shown in Table 1. The transition probabilities distributions of the HMM satisfy $p_{ii} \geq 0.98$ for $i = 1, 2, 3$. We have clustered the data using the standard Viterbi sequence, and using the Window HMM classification with $m = 20$. This value of $m$ was chosen to satisfy the condition (36). A fragment of the clustering is shown in Figure 8.

\footnote{The values of the returns were modelled as a discrete real valued variable with 200 possible values. The discretization was performed by partitioning the real line into 200 bins corresponding to the 200 quantiles of the full distribution of the data.}
Figure 7: Distribution functions for the emissions for SP500.

Table 1: Standard Deviations of S&P 500 Emissions

| Distribution | Standard Deviation |
|--------------|-------------------|
| $\mu_1$ (blue) | 0.004             |
| $\mu_2$ (red)  | 0.014             |
| $\mu_3$ (green) | 0.05              |

The HMM and Window HMM classification differ in total on 15% of the data. A typical difference in classification on this data looks similar to the one presented in Figure 8 – a long sequence of the same color for the HMM classification is broken into a number of subsequences of different colors by the Window HMM classification. To assess the quality of the partition found by the Window HMM we performed the following experiment: Consider all the periods in the data where the HMM assigns color red and Window HMM assigns color blue, and such that the length of the period is at least 10 days. There are 25 such periods in the data, ranging in length from 10 to 60 days, and comprising 706 days in total. One can ask then what is the standard deviation of the data computed only on those 706 days. Since the HMM clusters these days as red, one expects the standard deviation to be close to 0.014. On the other hand, since Window HMM clusters these days as blue, one expects the standard deviation to be near 0.004. The true standard deviation on these days is 0.005. Moreover, individually on 20 out of the 25 periods comprising the 706 days the standard deviation did not exceed 0.006 and it was smaller than 0.01 for all the periods. This computation shows that the periods should indeed belong to the blue cluster rather than the red one, and they can be detected using the Window HMM but not the standard HMM classification.
Figure 8: A Fragment of S&P500 Clustering. Top raw: Viterbi sequence, Middle: Window HMM Viterbi, Bottom: Difference indicator.