Estimating the Accuracy of Spectral Learning for HMMs

Farhana Ferdousi Liza(✉) and Marek Grześ(✉)

School of Computing, University of Kent, Canterbury CT2 7NZ, UK
{fl207,mgrzes}@kent.ac.uk

Abstract. Hidden Markov models (HMMs) are usually learned using the expectation maximisation algorithm which is, unfortunately, subject to local optima. Spectral learning for HMMs provides a unique, optimal solution subject to availability of a sufficient amount of data. However, with access to limited data, there is no means of estimating the accuracy of the solution of a given model. In this paper, a new spectral evaluation method has been proposed which can be used to assess whether the algorithm is converging to a stable solution on a given dataset. The proposed method is designed for real-life datasets where the true model is not available. A number of empirical experiments on synthetic as well as real datasets indicate that our criterion is an accurate proxy to measure quality of models learned using spectral learning.

Keywords: Spectral learning · HMM · SVD · Evaluation technique

1 Introduction

Learning parameters of dynamical systems and latent variable models using spectral learning algorithms is fascinating because of their capability of globally optimal parameter estimation. The inherent problem of local optima in many existing local search methods, such as Expectation Maximisation (EM), Gradient Descent, Gibbs Sampling, or Metropolis Hastings, led to the development of the spectral learning algorithms which are based on the method of moments (MoM). The goal of the MoM is to estimate the parameters, θ, of a probabilistic model, $p(x|\theta)$, from training data, $X = \{x_n\}$ where $n = 1 \dots N$. The basic idea is to compute several sample moments (empirical moments) of the model on the training data, $\phi_i(X) = \frac{1}{N}\sum_{n=1}^{N} f_i(x_n)$, and then to alter the parameters so that the expected moments under the model, $\langle f_i(x)\rangle_{p(x|\theta)} = \int f_i(x)p(x|\theta)dx$, are identical with the empirical moments, that is $\phi_i(X) = \langle f_i(x)\rangle_{p(x|\theta)}$. In short, the method of moments involves equating empirical moments with theoretical moments.

Hsu et al. [9] proposed an efficient and accurate MoM-based algorithm for discrete Hidden Markov Models (HMMs) that provides a theoretical guarantee for a unique and globally optimal parameter estimation. However, the algorithms that are based on MoM require large amounts of data to equate empirical moments with the theoretical moments [7,8].

C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 46–56, 2016.
DOI: 10.1007/978-3-319-44748-3_5

In real-life experiments, the practitioners would like to know what are the minimal data requirements that can guarantee that a particular model is learned near-optimally. When spectral learning does not have access to a sufficient amount of data, the estimates may be far from the global optima, and in some cases, the parameter estimates may lie outside the domain of the parameter space [6,15]. The practitioners are not able to judge how far their parameters can be from the optimal solution of a given model when the true model is not available, which is a normal scenario in practice. As a consequence, when the empirical model learned using spectral learning does not perform well, the practitioner does not know whether the solution is sub-optimal and the model is still correct, or whether the model is simply wrong. In this paper, we design a new method that can approximate the convergence of spectral learning.

The contribution of this work is that we provide a way to verify whether a particular dataset is sufficient to train a HMM using spectral learning [9]. In the current big data era, the proposed criterion can also be deployed in a system where there is possibility of more incoming data over time. In particular, the proposed measure can comfortably be incorporated into the online spectral learning algorithms, such as [2]. A number of authors have been dealing with the other features of spectral leaning, such as low rank, scalability and insufficient statistics; however, to the best of our knowledge, there is no work where the basis vector rotation-based measure would be used as a proxy method to estimate the convergence of spectral learning algorithms.

2 Background

A hidden Markov model (HMM) is a probabilistic system that models Markov processes with unobserved (hidden) states. A discrete HMM can be defined as follows: let $(x_1, x_2, \dots)$ be a sequence of discrete observations from a HMM where $x_t \in \{1, \dots, n\}$ is the observation, and $(h_1, h_2, \dots)$ be a sequence of hidden states where $h_t \in \{1, \dots, m\}$ is the hidden state at time step t. The parameters of an HMM are $\langle \pi, T, O \rangle$ where $\pi \in \mathbb{R}^m$ is the initial state distribution, $T \in \mathbb{R}^{m \times m}$ is the transition matrix, and $O \in \mathbb{R}^{n \times m}$ is the observation matrix. $T(i, j) = P(i|j)$ is the probability of going to state i if the current state is j, $\pi(j)$ is the probability of starting in state j, and $O(q, j) = P(q|j)$ is the probability of emitting symbol q if the current state is j.

In general, the term 'spectral' refers to the use of eigenvalues, eigenvectors, singular values and singular vectors. Singular values and singular vectors can be obtained from the singular value decomposition (SVD) [1]. The SVD of a matrix A is a factorisation of A into a product of three matrices $A = UDV^\top$ where the columns of U and V are orthonormal and the matrix D is diagonal with positive real entries known as singular values. U and V are called left and right singular vectors. The k dimensional subspace that best fits the data in A can be specified by the top k left singular vectors in matrix U. The spectral algorithm for HMMs uses the SVD to retrieve a tractable subspace where the hidden state dynamics are preserved [9].

A convenient way to calculate the probability of a HMM sequence $(x_1, x_2, \ldots, x_t)$ using the matrix operator [5,10], $A_x = T\text{diag}(O_{x,1}, \ldots, O_{x,m})$ for $x = 1, \ldots, n$, is as follows:

$$Pr(x_1, \ldots, x_t) = 1_m^\top A_{x_t} \ldots A_{x_1} \pi. \tag{2.1}$$

The spectral learning algorithm for HMMs learns a representation that is based on this observable operator view of HMMs; this is an observable view because every A_x represents state transitions for a given observation x. However, in this case, the set of 'characteristic events' revealing a relationship between the hidden states and observations have to be known or estimated from data that requires the knowledge of the T and O matrices. For a real dataset, we don't have exact T and O matrices. To relax this requirement, Hsu et al. [9] has used a transformed set of operators (in a tractable subspace) based on low order empirical moments of the data. In practice, the following empirical moment matrices have to be estimated from the data:

$$P_1 \in \mathbb{R}^n \quad [P_1]_i = Pr(x_1 = i) \tag{2.2}$$

$$P_{2,1} \in \mathbb{R}^{n \times n} \quad [P_{2,1}]_{ij} = Pr(x_2 = i, x_1 = j) \tag{2.3}$$

$$P_{3,x,1} \in \mathbb{R}^{n \times n} \quad [P_{3,x,1}]_{ij} = Pr(x_3 = i, x_2 = x, x_1 = j). \tag{2.4}$$

The resulting transformed operators for the HMM are then computed as follows:

$$\hat{b}_1 = \hat{U}^\top \hat{P}_1;\ \hat{b}_\infty = (\hat{P}_{2,1}^\top \hat{U})^+ \hat{P}_1;\ \hat{B}_x = \hat{U}^\top \hat{P}_{3,x,1} (\hat{U}^\top \hat{P}_{2,1})^+ \ \forall_x \in [n] \tag{2.5}$$

where $\hat{U}$ is computed by performing SVD on the correlation matrix $P_{2,1}$. If T and O in the underlying HMM are of rank m and $\hat{\pi} > 0$, then it can be shown that

$$\hat{Pr}(x_1, \ldots, x_t) = b_\infty^\top B_{x_t} \ldots B_{x_1} b_1 \tag{2.6}$$

which means that using the parameters learned by the spectral learning algorithm, a full joint probability of a sequence can be computed without knowing the exact T and O matrices. In a real situation, one does not have the optimal empirical moments and has to approximate the moment matrices, $\hat{P}_1$, $\hat{P}_{2,1}$ and $\hat{P}_{3,x,1}$, from a finite amount of data and consequently to approximate the operators $\hat{b}_1, \hat{b}_\infty$ and $\hat{B}_x$. Hsu et al. [9] has proved that the joint probability estimates of a HMM sequence are consistent (i.e. they converge to the correct model) when the number (N) of sampled observations tends to infinity:

$$\lim_{N \to \infty} \sum_{x_1, \ldots, x_t} |Pr(x_1, \ldots, x_t) - \hat{Pr}(x_1, \ldots, x_t)| = 0.$$

3 Main Method

Based on our empirical observations, in this paper, we have proposed a convergence criterion for a spectral learning algorithm for HMMs [9] with finite data.

Since the true convergence cannot be determined with certainty when the true model is unavailable, our method is a proxy measure that approximates the difference from the true model. Technically speaking, here convergence means that the minimal training data requirement is satisfied to yield empirical moments that are sufficiently close to the real moments. Note that when the essential amount of data is not available, the parameter estimates based on empirical moments may not even be in the domain of the parameter space (e.g. probabilities can be negative [6,15]). Our observations lead to a straightforward methodology that can approximate whether the algorithm has access to a sufficient amount of data. Specifically, we apply the spectral learning algorithm on a number of sub-sets of the training data where the size of the sub-sets is increased in subsequent iterations and each subset contains data of the previous sub-set to observe the effect of the increasing dataset. The spectral learning solution uses one of the orthonormal matrices, $\hat{U}$, as described in the previous section, to define a solution to the overall learning problem. Our main method is based on an observation that the bases contained in $\hat{U}$ rotate when the algorithm is executed on different sub-sets of the training data. By rotation, we mean the angle change between any two subsequent basis vectors contained in corresponding $\hat{U}$. Note that every sub-set defines one basis. The key point is that the magnitude of those rotations diminishes when the size of the sub-set grows. In Sect. 5, experimental results in (Fig. 2) confirm this claim, where the angle change differences become smaller as the size of the training sub-sets becomes larger.

Therefore, our hypothesis is that the magnitude of those rotations (measured as an angle change difference between two successive basis) can be a good proxy to determine the convergence (or equivalently data sufficiency) of the spectral learning algorithm. In order to justify our hypothesis, we show empirically on synthetic data that the learned model is an accurate approximation of the true model, when the angle that quantifies the magnitude of the rotations of the successive bases is sufficiently small. In short, we show empirically that when the rotation is small, the error is small as well. So, we can treat the rotation as a proxy to quantify the error.

In our approach, the original one-shot spectral learning algorithm for HMMs has been converted to a multi-step procedure for multiple training sub-sets described above. When the basis rotation between two successive corresponding basis vectors is less than a required value (e.g. 10^{-5}), our empirical experiments show that the spectral learning solution converges to a stable solution in the parameter space. However, if the required rotation (angle change difference) between two corresponding bases is not achievable with the training data at hand, then the spectral learning solution cannot be considered as reliable, and in that case another suitable parameter estimation technique should be used for the task. On convergence, the original spectral algorithm should be applied to the whole dataset to compute the final parameters. The rotation (angle change difference) is calculated using the maximum value of the dot product between each successive corresponding basis vectors in $\hat{U}$.

The next sections will show empirical evidence that smaller basis rotations are correlated with the real error on data when the true model is known, and therefore the magnitude of the basis rotation can be considered as a proxy that can assess the quality of the learned parameters for a particular model.

4 Experimental Methodology

4.1 Evaluation

In order to seek empirical evidence to support our hypothesis, we need a notion of an error function that can measure the quality of the HMM model learned from data, and we want to show empirically that our proxy measure is correlated with that error function. Then, we will conclude that our proxy method is a good indicator of the quality of the learned model on real data where the true error cannot be computed because the ground truth is not known.

Zhao and Poupart [15] used a normalized L_1 error that uses the sum of t^{th} roots of absolute errors, where t is the length of test sequences and τ is the set of all test sequences. This approach relies on the probability of seeing a certain sequence of outputs, $Pr(x_1, \ldots, x_t)$.

$$L_1 = \sum_{(x_1,\ldots,x_t)\in\tau} |Pr(x_1, \ldots, x_t) - \hat{Pr}(x_1, \ldots, x_t)|^{\frac{1}{t}}. \tag{4.1}$$

The error bounds for spectral learning in HMMs were derived in Hsu et al. [9, Sect. 4.2.3] for a similar measure. Moreover, unlike other approaches, such as Kullback-Leibler divergence [14], this method does not use a logarithm in its computation, and is robust against negative probabilities.

Spectral learning for HMMs [9] uses the transformed operators and Eq. 2.6 to calculate joint probabilities. Certainly, when one knows the exact model (which is true in the case of synthetic HMMs), the exact probability $Pr(x_1, ..., x_t)$ can be calculated using Eq. 2.1, which either involves multiplication of the exact transition, T, and emission matrices, O, or combined matrices, A_x. Calculating the error by comparing such exact measures reveals how close the estimated model is to the exact model. The normalized L_1 error serves this purpose without using the exact T and O because, as shown in [15], it uses probabilities of sequences and handles negative values. As a result, this leads to a measure that can indicate whether a model is well-fitted or not. We use this error to show that the angle change difference can indicate that the model is well-fitted. For a model to be well-fitted, in the spectral algorithms, the empirical moments have to be sufficient. In that sense, the angle change difference can also be a valid indication of the sufficiency of empirical moments.

4.2 Experimental Settings

The performance of HMM learning algorithms, in general, depends on the linear independence of the rows of (T and O) for a maximum discrimination of the

state and observation symbols. However, as T and O are inextricably linked to the model execution, [3] defined Inverse Condition Number (ICN) for indicating the linear independence of T and O. ICN was calculated as a ratio between the smallest and largest singular value of a row augmented matrix of T and O. A row augmented matrix is a matrix obtained by appending the columns of two given matrices. Such ICN was also demonstrated in [4] with Local Search method (LM) based parameter estimation techniques for HMMs. If ICN is close to 1 then the HMM is well-conditioned; if the ICN is close to 0 then HMM is ill-conditioned. While experimenting with the synthetic HMM systems, ICN was used to verify how our proposed convergence measure works with ill-conditioned HMMs.

Table 1. Description of Benchmark and some random HMMs (here Ex. or ex. are abbreviation of the word Example. Ex. are used in our analysis and plots)

HMM	Reference	ICN	HMM	Reference	ICN
Ex.1 (m = 2, n = 3)	[11, p. 26 ex. 1]	0.5731	Ex.2 (m = 2, n = 6)	[11, p. 40 ex. 2]	0.7338
Ex. 3 (m = 2, n = 2)	[12, p. 79 ex. 1]	0.5881	Ex. 4 (m = 2, n = 10)	Random	0.6756
Ex. 5 (m = 3, n = 8)	[11, p. 26 ex. 3]	0.6158	Ex. 6 (m = 3, n = 3)	Random	0.6101
Ex. 7 (m = 3, n = 10)	[11, p. 26 ex. 4]	0.6219	Ex. 8 (m = 3, n = 3)	[12, p. 80 ex. 4]	0.2070
Ex. 9 (m = 3, n = 3)	[12, p. 80 ex. 5]	0.3305	Ex. 10 (m = 3, n = 3)	[12, p. 81 ex. 6]	0.4612
Ex. 11 (m = 3, n = 3)	[12, p. 81 ex. 7]	0.3678	Ex. 12 (m = 3, n = 3)	[12, p. 79 ex. 2]	0.3715
Ex. 13 (m = 3, n = 10)	Random	0.6355	Ex. 14 (m = 3, n = 12)	Random	0.6528
Ex. 15 (m = 3, n = 20)	Random	0.5475	Ex. 16 (m = 2, n = 2)	Random	0.9800
Ex. 17 (m = 2, n = 2)	Random	0.1800	Ex. 18 (m = 2, n = 2)	Random	0.0200

The proposed convergence measure was tested on both synthetic datasets and on one real dataset. The normalised L_1 error can be computed on synthetic data only since the true model is required. The synthetic datasets are the HMM benchmarks used in [11–13]. All benchmark HMMs are summarised in the Table 1 where HMM column has the number of states (m) and the number of possible observations (n) used in the experiment, the Reference column has the source of O and T matrices, and the ICN column has the value of the ICN for the corresponding HMM. In our analysis, we have also used additional random HMM systems generated by rejection sampling method to obtain different ICN values.

The observation triples for training [9] were generated by sampling from the corresponding HMMs. The real dataset was arranged into triples using a sliding window approach. This dataset is based on web-navigation data from msnbc.com[1] and consists of 989818 time-series sequences and 17 observable symbols. The testing data to compute the L_1 error for every synthetic HMM consists of 20000 observation sequences of length 50.

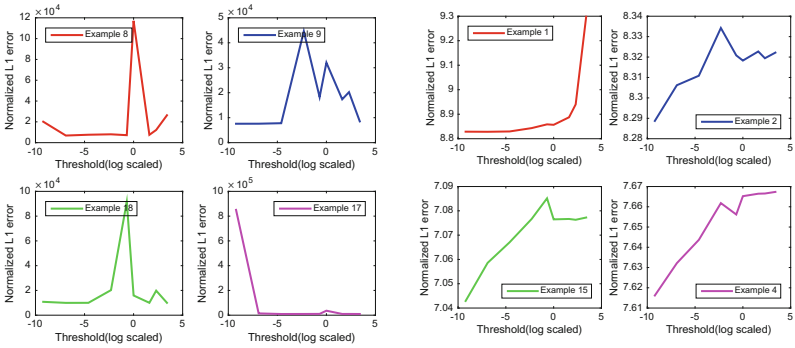
To show that the basis vector angle change difference can be a good indication of the sufficiency of the empirical moments, each HMM was trained incrementally (with subsets of training data) for different required maximum angle change difference (θ), and for each case the normalised L_1 error was calculated.

[1] https://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data.

In all our experiments, we added 20 additional training examples in each subsequent sub-set. Our goal is to show that, when maximum requirement for angle change difference is smaller, L_1 is also smaller on well-conditioned HMMs. The chosen θ were 30, 10, 5, 2, 1, 0.5, 0.1, 0.01, 0.001, 0.0001, and 0.00001 degrees. To visualise and validate the convergence on smaller θs, the T and O matrices were recovered from spectral parameters using Appendix C of [9]. The next section will show the quality of those matrices as a function the ICN and the required maximum angle change difference.

5 Experimental Results

The L_1 error was calculated for different θ values for each benchmark HMM system as described in the previous section. When θ is large (30, 10, 5, 2, 1, 0.5, 0.1, 0.01), the spectral learning solution of joint probability generates many negative probabilities for test sequences and the error pattern is thus inconclusive. As a negative probability for a sequence makes L_1 larger, for large θs, small spikes can be seen in Fig. 1b. However, for well-conditioned HMMs, the error reduces when the θ becomes smaller (0.001, 0.0001, and 0.00001 in Fig. 1b). This indicates that a smaller θ implies a well-fitted model. In a practical application where the exact model is not known, our angle change measure can inform a practitioner about the quality of the current parameters for her model. For the same experiment, the basis vector angle change difference for different θ were plotted in (Fig. 2). To achieve a larger θ, a small training dataset is sufficient. However, to achieve a small θ, comparatively larger training data is required. Thus, the angle change difference is correlated with training data requirements as well as the L_1 measure of correctness. A similar result of the angle change difference for different θ was found on all other synthetic HMMs systems and on a real dataset.



(a) Ill-condtioned ($ICN < 0.4$) (b) Well-conditioned ($ICN \geq 0.4$)

Fig. 1. L_1 error with chosen θs for different example HMMs

Table 2. Recovered T and O matrices for a well-conditioned HMM (ICN ≥ 0.4) with different θs. (See that smaller θ retrieves T and O closer to the true parameters)

$\theta = 5$	$\theta = 2$	$\theta = 0.01$
$T = \begin{bmatrix} 1.6486 & -0.5497 \\ 0.3941 & 0.5716 \end{bmatrix}$	$T = \begin{bmatrix} 0.9695 + 0.4851i & 1.5889 - 0.0733i \\ 0.2789 - 0.8469i & -0.7950 - 0.5653i \end{bmatrix}$	$T = \begin{bmatrix} 0.7074 & 0.8103 \\ 0.1287 & -0.0615 \end{bmatrix}$
$O = \begin{bmatrix} 0.5590 & 0.1565 \\ 0.1537 & 0.3937 \\ 0.2541 & 0.4830 \end{bmatrix}$	$O = \begin{bmatrix} 0.3430 + 0.1563i & 0.3430 - 0.1563i \\ 0.5470 + 0.0000i & 0.2667 + 0.0000i \\ 0.2502 + 0.0860i & 0.2502 - 0.0860i \end{bmatrix}$	$O = \begin{bmatrix} 0.5332 & 0.1293 \\ 0.5322 & 0.1239 \\ 0.2178 & 0.4636 \end{bmatrix}$
$\theta = 0.0001$	$\theta = 0.00001$	**True parameter**
$T = \begin{bmatrix} 0.9336 & 0.3400 \\ 0.0628 & 0.6582 \end{bmatrix}$	$T = \begin{bmatrix} 0.8874 & 0.3503 \\ 0.1106 & 0.6330 \end{bmatrix}$	$T = \begin{bmatrix} 0.9000 & 0.3000 \\ 0.1000 & 0.7000 \end{bmatrix}$
$O = \begin{bmatrix} 0.2873 & 0.9300 \\ 0.4739 & 0.0198 \\ 0.2391 & 0.0499 \end{bmatrix}$	$O = \begin{bmatrix} 0.2435 & 0.8298 \\ 0.5080 & 0.1066 \\ 0.2476 & 0.0646 \end{bmatrix}$	$O = \begin{bmatrix} 0.2500 & 0.8000 \\ 0.5000 & 0.1000 \\ 0.2500 & 0.1000 \end{bmatrix}$

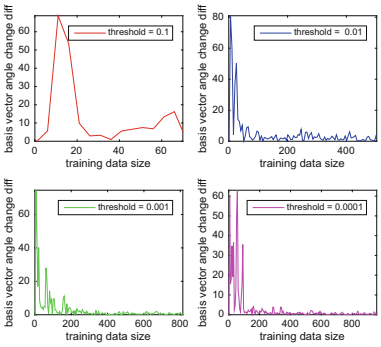


Fig. 2. Synthetic data (Ex. 1)

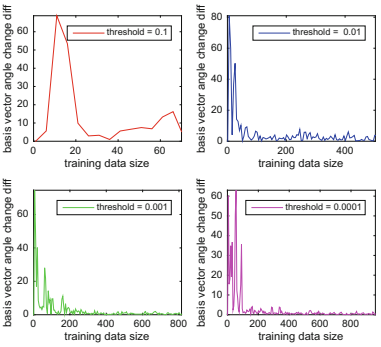


Fig. 3. Real data ($\theta = 0.001$)

The actual number of hidden states is usually not known for a real dataset. The angle change difference for different numbers of hidden states (m) (Fig. 3) shows that more hidden states leads to a higher model complexity (i.e. higher m) and as a result more training data is required to achieve the same θ, 0.001. For instance, to achieve $\theta = 0.001$ with the number of hidden states $m = 8$, $20 \times 4 \times 10^4$ training examples are required, whereas to achieve the same θ with $m = 4$, 20×10^4 training examples are sufficient. Here, 20 is the number of training examples (observations) added to each subsequent training sub-set. Thus, the angle change difference is also correlated with model complexity.

L_1 is not possible to calculate for the real dataset because of the absence of the ground truth as the exact model for real dataset is not known. This is true for all real datasets in general. However, the angle change based criterion can be used with ease to determine the sufficiency of the training data (consequently, empirical moments), and therefore the fitness of the model for real data. Therefore, by using angle change difference as a measure of convergence, it is possible to determine the required training data for sufficient empirical moments based

on model complexity, m. From the empirical evidence, we observed that the θ value of 10^{-5} gives satisfactory result in most cases. However, by taking smaller θ, we would be more confident about the solution.

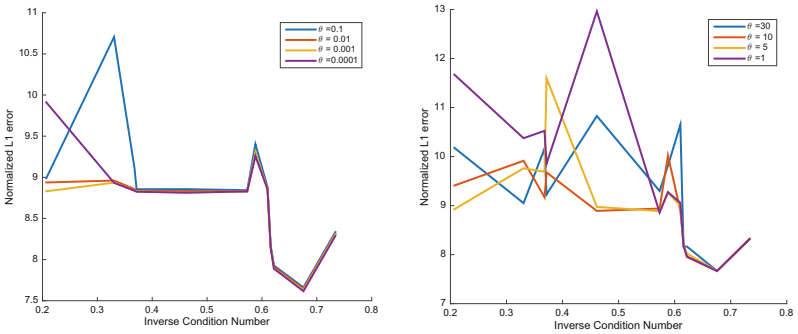
In our experiments, the T and O matrices were extracted to visualize the convergence. It was also observed that when the angle change is small, the recovered T and O matrices are in the parameter space and they are close to the exact model (Table 2). This is another confirmation that leads to a conclusion that the model fitness or empirical moment sufficiency for a well-conditioned HMM can be determined using a certain small angle change difference as a convergence criterion.

On well-conditioned HMMs, when the empirical moments are sufficient, the L_1 error is reduced monotonously (e.g. for a particular HMM, L_1 gets smaller when the θ becomes smaller (Fig. 1b). This is not the case on ill-conditioned HMMs (Fig. 1a) because of the ICN uncertainty existing in the observable space. If a HMM is ill-conditioned (i.e. if ICN is close to 0 or $ICN < 0.4$), L_1 is not correlated in the same way with angle change difference, and it does not lead to conclusive results. For example, in Fig. 4a, for a HMM with $ICN = 0.2$, the L_1 error is lower for $\theta = 0.1$ whereas the error is higher for $\theta = 0.0001$.

This is a feature of a particular HMM, and not a problem with the parameter estimation techniques, because ill-conditioned HMMs are almost impossible to learn even with substantial amount of data at hand (Table 3). Therefore, for ill-conditioned HMMs, the angle change difference is not a proxy for model fitness. This is due to the ambiguity and the uncertainty feature of HMMs [3]. However, on well-conditioned HMMs with $ICN \geq 0.4$, the L_1 is monotonously consistent (Fig. 4a), and shows that the model is fitted well. In Fig. 4b, the non-monotonous L_1 corresponds to a model that was not trained enough, and the angle change difference is large. That means a smaller θ will result in convergence when the error will follow the monotonous trend. If the error is not monotonous even with small θ then its means the HMM is ill-conditioned. In that case, HMM will not be able to model the data. Our empirical experiments confirm that the angle change difference is a useful proxy for sufficient empirical moments and consequently model fitness for well-conditioned HMMs.

Table 3. Recovered T and O matrices for an ill-conditioned HMM (ICN ≤ 0.2) with the required maximum angle change difference $\theta = 0.00001$

True system	Estimated system	True system	Estimated system
$T = \begin{bmatrix} 0.51 & 0.49 \\ 0.49 & 0.51 \end{bmatrix}$	$T = \begin{bmatrix} 1.0372 & 1.3084 \\ -0.0372 & -0.3084 \end{bmatrix}$	$T = \begin{bmatrix} 0.50 & 0.10 & 0.20 \\ 0.20 & 0.60 & 0.40 \\ 0.30 & 0.30 & 0.40 \end{bmatrix}$	$T = \begin{bmatrix} 0.9793 & -0.5696 & 3.5463 \\ -0.3589 & 2.2337 & -4.6428 \\ 0.2551 & -0.4065 & 1.3400 \end{bmatrix}$
$O = \begin{bmatrix} 0.49 & 0.51 \\ 0.51 & 0.49 \end{bmatrix}$	$O = \begin{bmatrix} 0.4899 & 0.0979 \\ 0.5101 & 0.9121 \end{bmatrix}$	$O = \begin{bmatrix} 0.20 & 0.40 & 0.70 \\ 0.70 & 0.40 & 0.10 \\ 0.10 & 0.20 & 0.20 \end{bmatrix}$	$O = \begin{bmatrix} 0.2686 & 0.5274 & 1.0800 \\ 0.5980 & 0.2478 & 0.0552 \\ 0.2275 & 0.1330 & -0.1376 \end{bmatrix}$



(a) θ sufficiently small (b) Under-trained model due to large θ

Fig. 4. Angle change difference and model error

6 Conclusion

In this paper, we have proposed a basis vector angle change based convergence criterion for spectral learning that is known to require large amounts of data. These algorithms usually work in one-shot and on real data there is no indication for the practitioner about the convergence. As a result, it is likely that a practitioner will end up using an under-trained model. We showed how a one-shot algorithm can be trained several times for several sub-sets of the available training data of a growing size. The advantage is that in this way one can approximate the convergence of the algorithm to check whether the dataset provides sufficient empirical moments that allow the algorithm to produce parameters that lead to a small error. We demonstrated our method on spectral learning for HMMs and showed empirically that our claims are justified in that case. We have tested our proposed method on synthetic and real data and showed empirically that our method can indicate sufficiency of the empirical moments. As a result, the practitioners can check whether they need more data or whether they need a different model if the predictive power of their solution is not satisfactory.

For the sake of computational efficiency, one could apply our method in conjunction with an online algorithm [2] which relies on incremental SVD. However, in this study, our goal was to investigate the convergence of spectral learning and, thus, using standard, non-incremental SVD had better methodological justification.

We know that, in theory [9], spectral learning for HMMs will converge to an optimal solution given a sufficient amount of data. In the face of our results, it would be interesting to compare spectral learning with local search methods, such as EM, for different magnitudes of the angle change.

References

1. Baker, K.: Singular Value Decomposition Tutorial. Ohio State University (2005)
2. Boots, B., Gordon, G.: An online spectral learning algorithm for partially observable nonlinear dynamical systems. In: Proceedings of AAAI (2011)
3. Caelli, T., McCane, B.: Components analysis of hidden Markov models in computer vision. In: Proceedings of 12th International Conference on Image Analysis and Processing, pp. 510–515, September 2003
4. Davis, R.I.A., Lovell, B.C.: Comparing and evaluating HMM ensemble training algorithms using train and test and condition number criteria. Pattern Anal. Appl. **6**(4), 327–336 (2003)
5. Even-Dar, E., Kakade, S.M., Mansour, Y.: The value of observation for monitoring dynamic systems. In: IJCAI, pp. 2474–2479 (2007)
6. Glaude, H., Enderli, C., Pietquin, O.: Spectral learning with proper probabilities for finite state automaton. In: Proceedings of ASRU. IEEE (2015)
7. Hall, A.R., et al.: Generalized Method of Moments. Oxford University Press, Oxford (2005)
8. Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica: J. Econometric Soc. **50**, 1029–1054 (1982)
9. Hsu, D., Kakade, S.M., Zhang, T.: A spectral algorithm for learning hidden Markov models. J. Comput. Syst. Sci. **78**(5), 1460–1480 (2012)
10. Jaeger, H.: Observable operator models for discrete stochastic time series. Neural Comput. **12**(6), 1371–1398 (2000)
11. Mattfeld, C.: Implementing spectral methods for hidden Markov models with real-valued emissions. CoRR abs/1404.7472 (2014). http://arxiv.org/abs/1404.7472
12. Mattila, R.: On identification of hidden Markov models using spectral and non-negative matrix factorization methods. Master's thesis, KTH Royal Institute of Technology (2015)
13. Mattila, R., Rojas, C.R., Wahlberg, B.: Evaluation of Spectral Learning for the Identification of Hidden Markov Models, July 2015. http://arxiv.org/abs/1507.06346
14. Vanluyten, B., Willems, J.C., Moor, B.D.: Structured nonnegative matrix factorization with applications to hidden Markov realization and clustering. Linear Algebra Appl. **429**(7), 1409–1424 (2008)
15. Zhao, H., Poupart, P.: A sober look at spectral learning. CoRR abs/1406.4631 (2014). http://arxiv.org/abs/1406.4631