Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices

Kirthevasan Kandasamy[*]
Carnegie Mellon University
Pittsburgh, PA 15213
kandasamy@cs.cmu.edu

Maruan Al-Shedivat[*]
Carnegie Mellon University
Pittsburgh, PA 15213
alshedivat@cs.cmu.edu

Eric P. Xing
Carnegie Mellon University
Pittsburgh, PA 15213
epxing@cs.cmu.edu

Abstract

Recently, there has been a surge of interest in using spectral methods for estimating latent variable models. However, it is usually assumed that the distribution of the observations conditioned on the latent variables is either discrete or belongs to a parametric family. In this paper, we study the estimation of an m-state hidden Markov model (HMM) with only smoothness assumptions, such as Hölderian conditions, on the emission densities. By leveraging some recent advances in continuous linear algebra and numerical analysis, we develop a computationally efficient spectral algorithm for learning nonparametric HMMs. Our technique is based on computing an SVD on nonparametric estimates of density functions by viewing them as *continuous matrices*. We derive sample complexity bounds via concentration results for nonparametric density estimation and novel perturbation theory results for continuous matrices. We implement our method using Chebyshev polynomial approximations. Our method is competitive with other baselines on synthetic and real problems and is also very computationally efficient.

1 Introduction

Hidden Markov models (HMMs) [1] are one of the most popular statistical models for analyzing time series data in various application domains such as speech recognition, medicine, and meteorology. In an HMM, a discrete hidden state undergoes Markovian transitions from one of m possible states to another at each time step. If the hidden state at time t is h_t, we observe a random variable $x_t \in \mathcal{X}$ drawn from an *emission* distribution, $O_j = \mathbb{P}(x_t | h_t = j)$. In its most basic form $\mathcal{X}$ is a discrete set and O_j are discrete distributions. When dealing with continuous observations, it is conventional to assume that the emissions O_j belong to a parametric class of distributions, such as Gaussian.

Recently, spectral methods for estimating parametric latent variable models have gained immense popularity as a viable alternative to the Expectation Maximisation (EM) procedure [2–4]. At a high level, these methods estimate higher order moments from the data and recover the parameters via a series of matrix operations such as singular value decompositions, matrix multiplications and pseudo-inverses of the moments. In the case of discrete HMMs [2], these moments correspond exactly to the joint probabilities of the observations in the sequence.

Assuming parametric forms for the emission densities is often too restrictive since real world distributions can be arbitrary. Parametric models may introduce incongruous biases that cannot be reduced even with large datasets. To address this problem, we study *nonparametric* HMMs only assuming some mild smoothness conditions on the emission densities. We design a spectral algorithm for this setting. Our methods leverage some recent advances in continuous linear algebra [5, 6] which views two-dimensional functions as continuous analogues of matrices. Chebyshev polynomial approximations enable efficient computation of algebraic operations on these continuous objects [7, 8]. Using these ideas, we extend existing spectral methods for discrete HMMs to the continuous nonparametric setting. Our main contributions are:

[*] Joint lead authors.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

1. We derive a spectral learning algorithm for HMMs with nonparametric emission densities. While the algorithm is similar to previous spectral methods for estimating models with a finite number of parameters, many of the ideas used to generalise it to the nonparametric setting are novel, and, to the best of our knowledge, have not been used before in the machine learning literature.
2. We establish sample complexity bounds for our method. For this, we derive concentration results for nonparametric density estimation and novel perturbation theory results for the aforementioned continuous matrices. The perturbation results are new and might be of independent interest.
3. We implement our algorithm by approximating the density estimates via Chebyshev polynomials which enables efficient computation of many of the continuous matrix operations. Our method outperforms natural competitors in this setting on synthetic and real data and is computationally more efficient than most of them. Our Matlab code is available at `github.com/alshedivat/nphmm`.

While we focus on HMMs in this exposition, we believe that the ideas presented in this paper can be easily generalised to estimating other latent variable models and predictive state representations [9] with nonparametric observations using approaches developed by Anandkumar et al. [3].

Related Work: Parametric HMMs are usually estimated using maximum likelihood principle via EM techniques [10] such as the Baum-Welch procedure [11]. However, EM is a local search technique, and optimization of the likelihood may be difficult. Hence, recent work on spectral methods has gained appeal. Our work builds on Hsu et al. [2] who showed that discrete HMMs can be learned efficiently, under certain conditions. The key idea is that any HMM can be completely characterised in terms of quantities that depend entirely on the observations, called the *observable representation*, which can be estimated from data. Siddiqi et al. [4] show that the same algorithm works under slightly more general assumptions. Anandkumar et al. [3] proposed a spectral algorithm for estimating more general latent variable models with parametric observations via a moment matching technique.

That said, we are aware of little work on estimating latent variable models, including HMMs, when the observations are *nonparametric*. A commonly used heuristic is the nonparametric EM [12], which lacks theoretical underpinnings. This should not be surprising because EM is degenerate for most nonparametric problems as a maximum likelihood procedure [13]. Only recently, De Castro et al. [14] have provided a minimax-type of result for the nonparametric setting. In their work, Siddiqi et al. [4] proposed a heuristic based on kernel smoothing, to modify the discrete algorithm for continuous observations. Further, their procedure cannot be used to recover the joint or conditional probabilities of a sequence, which would be needed to compute probabilities of events and other inference tasks.

Song et al. [15, 16] developed an RKHS-based procedure for estimating the Hilbert space embedding of an HMM. While they provide theoretical guarantees, their bounds are in terms of the RKHS distance of the true and estimated embeddings. This metric depends on the choice of the kernel and it is not clear how it translates to a suitable distance measure on the observation space such as an L^1 or L^2 distance. While their method can be used for prediction and pairwise testing, it cannot recover the joint and conditional densities. On the contrary, our model provides guarantees in terms of the more interpretable total variation distance and is able to recover the joint and conditional probabilities.

2 A Pint-sized Review of Continuous Linear Algebra

We begin with a pint-sized review on continuous linear algebra which treats functions as continuous analogues of matrices. Appendix A contains a quart-sized review. Both sections are based on [5, 6]. While these objects can be viewed as operators on Hilbert spaces which have been studied extensively in the years, the above line of work simplified and specialised the ideas to functions.

A *matrix* $F \in \mathbb{R}^{m\times n}$ is an $m \times n$ array of numbers where $F(i,j)$ denotes the entry in row i, column j. m or n could be (countably) infinite. A *column qmatrix* (quasi-matrix) $Q \in \mathbb{R}^{[a,b]\times m}$ is a collection of m functions defined on $[a,b]$ where the row index is continuous and column index is discrete. Writing $Q = [q_1, \dots, q_m]$ where $q_j : [a,b] \to \mathbb{R}$ is the j^{th} function, $Q(y,j) = q_j(y)$ denotes the value of the j^{th} function at $y \in [a,b]$. $Q^\top \in \mathbb{R}^{m\times[a,b]}$ denotes a row qmatrix with $Q^\top(j,y) = Q(y,j)$. A *cmatrix* (continuous-matrix) $C \in \mathbb{R}^{[a,b]\times[c,d]}$ is a two dimensional function where both row and column indices are continuous and $C(y,x)$ is the value of the function at $(y,x) \in [a,b] \times [c,d]$. $C^\top \in \mathbb{R}^{[c,d]\times[a,b]}$ denotes its transpose with $C^\top(x,y) = C(y,x)$. Qmatrices and cmatrices permit all matrix multiplications with suitably defined inner products. For example, if $R \in \mathbb{R}^{[c,d]\times m}$ and $C \in \mathbb{R}^{[a,b]\times[c,d]}$, then $CR = T \in \mathbb{R}^{[a,b]\times m}$ where $T(y,j) = \int_c^d C(y,s)R(s,j)\mathrm{d}s$.

A cmatrix has a singular value decomposition (SVD). If $C \in \mathbb{R}^{[a,b]\times[c,d]}$, it decomposes as an infinite sum, $C(y,x) = \sum_{j=1}^{\infty} \sigma_j u_j(y) v_j(x)$, that converges in L^2. Here $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ are the singular values of C. $\{u_j\}_{j\geq 1}$ and $\{v_j\}_{j\geq 1}$ are functions that form orthonormal bases for $L^2([a,b])$ and $L^2([c,d])$, respectively. We can write the SVD as $C = U\Sigma V^\top$ by writing the singular vectors as infinite qmatrices $U = [u_1, u_2 \dots]$, $V = [v_1, v_2 \dots]$, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2 \dots)$. If only $m < \infty$ first singular values are nonzero, we say that C is of rank m. The SVD of a qmatrix $Q \in \mathbb{R}^{[a,b]\times m}$ is, $Q = U\Sigma V^\top$ where $U \in \mathbb{R}^{[a,b]\times m}$ and $V \in \mathbb{R}^{m\times m}$ have orthonormal columns and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_m)$ with $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$. The rank of a column qmatrix is the number of linearly independent columns (i.e. functions) and is equal to the number of nonzero singular values. Finally, as for the finite matrices, the pseudo inverse of the cmatrix C is $C^\dagger = V\Sigma^{-1}U^\top$ with $\Sigma^{-1} = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, \dots)$. The pseudo inverse of a qmatrix is defined similarly.

3 Nonparametric HMMs and the Observable Representation

Notation: Throughout this manuscript, we will use $\mathbb{P}$ to denote probabilities of events while p will denote probability density functions (pdf). An HMM characterises a probability distribution over a sequence of hidden states $\{h_t\}_{t\geq 0}$ and observations $\{x_t\}_{t\geq 0}$. At a given time step, the HMM can be in one of m hidden states, i.e. $h_t \in [m] = \{1, \dots, m\}$, and the observation is in some bounded continuous domain $\mathcal{X}$. Without loss of generality, we take[2] $\mathcal{X} = [0,1]$. The nonparametric HMM will be completely characterised by the initial state distribution $\pi \in \mathbb{R}^m$, the state transition matrix $T \in \mathbb{R}^{m\times m}$ and the emission densities $O_j : \mathcal{X} \to \mathbb{R}, j \in [m]$. $\pi_i = \mathbb{P}(h_1 = i)$ is the probability that the HMM would be in state i at the first time step. The element $T(i,j) = \mathbb{P}(h_{t+1} = i | h_t = j)$ of T gives the probability that a hidden state transitions from state j to state i. The emission function, $O_j : \mathcal{X} \to \mathbb{R}_+$, describes the pdf of the observation conditioned on the hidden state j, i.e. $O_j(s) = p(x_t = s | h_t = j)$. Note that we have $O_j(x) > 0$, $\forall x$ and $\int O_j(\cdot) = 1$ for all $j \in [m]$. In this exposition, we denote the emission densities by the qmatrix, $O = [O_1, \dots, O_m] \in \mathbb{R}_+^{[0,1]\times m}$.

In addition, let $\widetilde{O}(x) = \mathrm{diag}(O_1(x), \dots, O_m(x))$, and $A(x) = T\widetilde{O}(x)$. Let $x_{1:t} = \{x_1, \dots, x_t\}$ be an ordered sequence and $x_{t:1} = \{x_t, \dots, x_1\}$ denote its reverse. For brevity, we will overload notation for A for sequences and write $A(x_{t:1}) = A(x_t)A(x_{t-1})\dots A(x_1)$. It is well known [2, 17] that the joint probability density of the sequence $x_{1:t}$ can be computed via $p(x_{1:t}) = \mathbf{1}_m^\top A(x_{t:1})\pi$.

Key structural assumption: Previous work on estimating HMMs with continuous observations typically assumed that the emissions, O_j, take a parametric form, e.g. Gaussian. Unlike them, we only make mild nonparametric smoothness assumptions on O_j. As we will see, to estimate the HMM well in this problem we will need to estimate entire pdfs well. For this reason, the nonparametric setting is significantly more difficult than its parametric counterpart as the latter requires estimating only a finite number of parameters. When compared to the previous literature, this is the crucial distinction and the main challenge in this work.

Observable Representation: The observable representation is a description of an HMM in terms of quantities that depend on the observations [17]. This representation is useful for two reasons: (i) it depends only on the observations and can be directly estimated from the data; (ii) it can be used to compute joint and conditional probabilities of sequences even without the knowledge of T and O and therefore can be used for inference and prediction. First, we define the joint densities, P_1, P_{21}, P_{321}:

$$P_1(t) = p(x_1 = t), \quad P_{21}(s,t) = p(x_2 = s, x_1 = t), \quad P_{321}(r,s,t) = p(x_3 = r, x_2 = s, x_1 = t),$$

where $x_i, i = 1,2,3$ denotes the observation at time i. Denote $P_{3x1}(r,t) = P_{321}(r,x,t)$ for all x. We will find it useful to view both $P_{21}, P_{3x1} \in \mathbb{R}^{[0,1]\times[0,1]}$ as cmatrices. We will also need an additional qmatrix $U \in \mathbb{R}^{[0,1]\times m}$ such that $U^\top O \in \mathbb{R}^{m\times m}$ is invertible. Given one such U, the observable representation of an HMM is described by the parameters $b_1, b_\infty \in \mathbb{R}^m$ and $B : [0,1] \to \mathbb{R}^{m\times m}$,

$$b_1 = U^\top P_1, \qquad b_\infty = (P_{21}^\top U)^\dagger P_1, \qquad B(x) = (U^\top P_{3x1})(U^\top P_{21})^\dagger \tag{1}$$

As before, for a sequence, $x_{t:1} = \{x_t, \dots, x_1\}$, we define $B(x_{t:1}) = B(x_t)B(x_{t-1})\dots B(x_1)$. The following lemma shows that the first m left singular vectors of P_{21} are a natural choice for U.

Lemma 1. *Let $\pi > 0$, T and O be of rank m and U be the qmatrix composed of the first m left singular vectors of P_{21}. Then $U^\top O$ is invertible.*

[2] We discuss the case of higher dimensions in Section 7.

To compute the joint and conditional probabilities using the observable representation, we maintain an *internal state*, b_t, which is updated as we see more observations. The internal state at time t is

$$b_t = \frac{B(x_{t-1:1})b_1}{b_\infty^\top B(x_{t-1:1})b_1}. \tag{2}$$

This definition of b_t is consistent with b_1. The following lemma establishes the relationship between the observable representation and the internal states to the HMM parameters and probabilities.

Lemma 2 (Properties of the Observable Representation)**.** *Let* $\mathrm{rank}(T) = \mathrm{rank}(O) = m$ *and* $U^\top O$ *be invertible. Let* $p(x_{1:t})$ *denote the joint density of a sequence* $x_{1:t}$ *and* $p(x_{t+1:t+t'}|x_{1:t})$ *denote the conditional density of* $x_{t+1:t+t'}$ *given* $x_{1:t}$ *in a sequence* $x_{1:t+t'}$*. Then the following are true.*

1. $b_1 = U^\top O\pi$
2. $b_\infty = \mathbf{1}_m^\top (U^\top O)^{-1}$
3. $B(x) = (U^\top O)A(x)(U^\top O)^{-1} \ \ \forall x \in [0,1]$.
4. $b_{t+1} = B(x_t)b_t/(b_\infty^\top B(x_t)b_t)$.
5. $p(x_{1:t}) = b_\infty^\top B(x_{t:1})b_1$.
6. $p(x_{t+t':t+1}|x_{1:t}) = b_\infty^\top B(x_{t+t':t+1})b_t$.

The last two claims of the Lemma 2 show that we can use the observable representation for computing the joint and conditional densities. The proofs of Lemmas 1 and 2 are similar to the discrete case and mimic Lemmas 2, 3 & 4 of Hsu et al. [2].

4 Spectral Learning of HMMs with Nonparametric Emissions

The high level idea of our algorithm, NP-HMM-SPEC, is as follows. First we will obtain density estimates for P_1, P_{21}, P_{321} which will then be used to recover the observable representation b_1, b_∞, B by plugging in the expressions in (1). Lemma 2 then gives us a way to estimate the joint and conditional probability densities. For now, we will assume that we have N i.i.d sequences of triples $\{X^{(j)}\}_{j=1}^N$ where $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, X_3^{(j)})$ are the observations at the first three time steps. We describe learning from longer sequences in Section 4.3.

4.1 Kernel Density Estimation

The first step is the estimation of the joint probabilities which requires a nonparametric density estimate. While there are several techniques [18], we use kernel density estimation (KDE) since it is easy to analyse and works well in practice. The KDE for P_1, P_{21}, and P_{321} take the form:

$$\widehat{P}_1(t) = \frac{1}{N}\sum_{j=1}^N \frac{1}{h_1} K\left(\frac{t - X_1^{(j)}}{h_1}\right), \qquad \widehat{P}_{21}(s,t) = \frac{1}{N}\sum_{j=1}^N \frac{1}{h_{21}^2} K\left(\frac{s - X_2^{(j)}}{h_{21}}\right) K\left(\frac{t - X_1^{(j)}}{h_{21}}\right),$$

$$\widehat{P}_{321}(r,s,t) = \frac{1}{N}\sum_{j=1}^N \frac{1}{h_{321}^3} K\left(\frac{r - X_3^{(j)}}{h_{321}}\right) K\left(\frac{s - X_2^{(j)}}{h_{321}}\right) K\left(\frac{t - X_1^{(j)}}{h_{321}}\right). \tag{3}$$

Here $K : [0,1] \to \mathbb{R}$ is a symmetric function called a smoothing kernel and satisfies (at the very least) $\int_0^1 K(s)\mathrm{d}s = 1$, $\int_0^1 sK(s)\mathrm{d}s = 0$. The parameters h_1, h_{21}, h_{321} are the bandwidths, and are typically decreasing with N. In practice they are usually chosen via cross-validation.

4.2 The Spectral Algorithm

```
Algorithm 1  NP-HMM-SPEC
Input:  Data $\{X^{(j)} = (X_1^{(j)}, X_2^{(j)}, X_3^{(j)})\}_{j=1}^N$, number of states $m$.
  • Obtain estimates $\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}$ for $P_1, P_{21}, P_{321}$ via kernel density estimation (3).
  • Compute the cmatrix SVD of $\widehat{P}_{21}$. Let $\widehat{U} \in \mathbb{R}^{[0,1]\times m}$ be the first $m$ left singular vectors of $\widehat{P}_{21}$.
  • Compute the parameters observable representation. Note that $\widehat{B}$ is a $\mathbb{R}^{m\times m}$ valued function.
        $\widehat{b}_1 = \widehat{U}^\top \widehat{P}_1$,    $\widehat{b}_\infty = (P_{21}^\top \widehat{U})^\dagger \widehat{P}_1$,    $\widehat{B}(x) = (\widehat{U}^\top \widehat{P}_{3x1})(\widehat{U}^\top \widehat{P}_{21})^\dagger$
```

The algorithm, given above in Algorithm 1, follows the roadmap set out at the beginning of this section. While the last two steps are similar to the discrete HMM algorithm of Hsu et al. [2], the SVD, pseudoinverses and multiplications are with q/c-matrices. Once we have the estimates $\widehat{b}_1$, $\widehat{b}_\infty$, and $\widehat{B}(x)$ the joint and predictive (conditional) densities can be estimated via (see Lemma 2):

$$\widehat{p}(x_{1:t}) = \widehat{b}_\infty^\top \widehat{B}(x_{t:1})\widehat{b}_1, \qquad \widehat{p}(x_{t+t':t+1}|x_{1:t}) = \widehat{b}_\infty^\top \widehat{B}(x_{t+t':t+1})\widehat{b}_t. \tag{4}$$

Here $\widehat{b}_t$ is the estimated internal state obtained by plugging in $\widehat{b}_1, \widehat{b}_\infty, \widehat{B}$ in (2). Theoretically, these estimates can be negative in which case they can be truncated to 0 without affecting the theoretical results in Section 5. However, in our experiments these estimates were never negative.

4.3 Implementation Details

C/Q-Matrix operations using Chebyshev polynomials: While our algorithm and analysis are conceptually well founded, the important practical challenge lies in the efficient computation of the many aforementioned operations on c/q-matrices. Fortunately, some very recent advances in the numerical analysis literature, specifically on computing with Chebyshev polynomials, have rendered the above algorithm practical [6, Ch.3-4]. Due to the space constraints, we provide only a summary. Chebyshev polynomials is a family of orthogonal polynomials on compact intervals, known to be an excellent approximator of one-dimensional functions [19, 20]. A recent line of work [5, 8] has extended the Chebyshev technology to two dimensional functions enabling the mentioned operations and factorisations such as QR, LU and SVD [6, Sections 4.6-4.8] of continuous matrices to be carried efficiently. The density estimates $\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}$ are approximated by Chebyshev polynomials to within machine precision. Our implementation makes use of the Chebfun library [7] which provides an efficient implementation for the operations on continuous and quasi matrices.

Computation time: Representing the KDE estimates $\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}$ using Chebfun was roughly linear in N and is the brunt of the computational effort. The bandwidths for the three KDE estimates are chosen via cross validation which takes $\mathcal{O}(N^2)$ effort. However, in practice the cost was dominated by the Chebyshev polynomial approximation. In our experiments we found that NP-HMM-SPEC runs in linear time in practice and was more efficient than most alternatives.

Training with longer sequences: When training with longer sequences we can use a sliding window of length 3 across the sequence to create the triples of observations needed for the algorithm. That is, given N samples each of length $\ell^{(j)}, j = 1, \dots, N$, we create an augmented dataset of triples $\{\, \{(X_t^{(j)}, X_{t+1}^{(j)}, X_{t+2}^{(j)})\}_{t=1}^{\ell^{(j)}-2} \,\}_{j=1}^{N}$ and run NP-HMM-SPEC with the augmented data. As is with conventional EM procedures, this requires the additional assumption that the initial state is the stationary distribution of the transition matrix T.

5 Analysis

We now state our assumptions and main theoretical results. Following [2, 4, 15] we assume i.i.d sequences of triples are used for training. With longer sequences, the analysis should only be modified to account for the mixing of the latent state Markov chain, which is inessential for the main intuitions. We begin with the following regularity condition on the HMM.

Assumption 3. *$\pi > 0$ element-wise. $T \in \mathbb{R}^{m\times m}$ and $O \in \mathbb{R}^{[0,1]\times m}$ are of rank m.*

The rank condition on O means that emission pdfs are linearly independent. If either T or O are rank deficient, then the learner may confuse state outputs, which makes learning difficult[3]. Next, while we make no parametric assumptions on the emissions, some smoothness conditions are used to make density estimation tractable. We use the Hölder class, $\mathcal{H}_1(\beta, L)$, which is standard in the nonparametrics literature. For $\beta = 1$, this assumption reduces to L-Lipschitz continuity.

Assumption 4. *All emission densities belong to the Hölder class, $\mathcal{H}_1(\beta, L)$. That is, they satisfy,*

$$\textit{for all } \alpha \le \lfloor \beta \rfloor,\ j \in [m],\ s,t \in [0,1] \qquad \left| \frac{\mathrm{d}^\alpha O_j(s)}{\mathrm{d}s^\alpha} - \frac{\mathrm{d}^\alpha O_j(t)}{\mathrm{d}t^\alpha} \right| \le L|s-t|^{\beta - |\alpha|}.$$

Here $\lfloor \beta \rfloor$ is the largest integer ***strictly*** *less than β.*

[3] Siddiqi et al. [4] show that the discrete spectral algorithm works under a slightly more general setting. Similar results hold for the nonparametric case too but will restrict ourselves to the full rank setting for simplicity.

Under the above assumptions we bound the total variation distance between the true and the estimated densities of a sequence, $x_{1:t}$. Let $\kappa(O) = \sigma_1(O)/\sigma_m(O)$ denote the condition number of the observation qmatrix. The following theorem states our main result.

Theorem 5. *Pick any sufficiently small $\epsilon > 0$ and a failure probability $\delta \in (0, 1)$. Let $t \geq 1$. Assume that the HMM satisfies Assumptions 3 and 4 and the number of samples N satisfies,*

$$\frac{N}{\log(N)} \;\geq\; C\, m^{1+\frac{3}{2\beta}} \frac{\kappa(O)^{2+\frac{3}{\beta}}}{\sigma_m(P_{21})^{4+\frac{4}{\beta}}} \left(\frac{t}{\epsilon}\right)^{2+\frac{3}{\beta}} \log\left(\frac{1}{\delta}\right)^{1+\frac{3}{2\beta}}.$$

Then, with probability at least $1 - \delta$, the estimated joint density for a t-length sequence satisfies $\int |p(x_{1:t}) - \widehat{p}(x_{1:t})| \mathrm{d}x_{1:t} \leq \epsilon$. Here, C is a constant depending on β and L and $\widehat{p}$ is from (4).

Synopsis: Observe that the sample complexity depends critically on the conditioning of O and P_{21}. The closer they are to being singular, the more samples is needed to distinguish different states and learn the HMM. It is instructive to compare the results above with the discrete case result of Hsu et al. [2], whose sample complexity bound[4] is $N \gtrsim m \frac{\kappa(O)^2}{\sigma_m(P_{21})^4} \frac{t^2}{\epsilon^2} \log \frac{1}{\delta}$. Our bound is different in two regards. First, the exponents are worsened by additional $\sim \frac{1}{\beta}$ terms. This characterizes the difficulty of the problem in the nonparametric setting. While we do not have any lower bounds, given the current understanding of the difficulty of various nonparametric tasks [21–23], we think our bound might be unimprovable. As the smoothness of the densities increases $\beta \to \infty$, we approach the parametric sample complexity. The second difference is the additional $\log(N)$ term on the left hand side. This is due to the fact that we want the KDE to concentrate around its expectation in L^2 over $[0, 1]$, instead of just point-wise. It is not clear to us whether the $\log$ can be avoided.

To prove Theorem 5, first we will derive some perturbation theory results for c/q-matrices; we will need them to bound the deviation of the singular values and vectors when we use $\widehat{P}_{21}$ instead of P_{21}. Some of these perturbation theory results for continuous linear algebra are new and might be of independent interest. Next, we establish a concentration result for the kernel density estimator.

5.1 Some Perturbation Theory Results for C/Q-matrices

The first result is an analog of Weyl's theorem which bounds the difference in the singular values in terms of the operator norm of the perturbation. Weyl's theorem has been studied for general operators [24] and cmatrices [6]. We have given one version in Lemma 21 of Appendix B. In addition to this, we will also need to bound the difference in the singular vectors and the pseudo-inverses of the truth and the estimate. To our knowledge, these results are not yet known. To that end, we establish the following results. Here $\sigma_k(A)$ denotes the k^{th} singular value of a c/q-matrix A.

Lemma 6 (Simplified Wedin's Sine Theorem for Cmatrices)**.** *Let $A, \tilde{A}, E \in \mathbb{R}^{[0,1]\times[0,1]}$ where $\tilde{A} = A + E$ and $\mathrm{rank}(A) = m$. Let $U, \tilde{U} \in \mathbb{R}^{[a,b]\times m}$ be the first m left singular vectors of A and $\tilde{A}$ respectively. Then, for all $x \in \mathbb{R}^m$, $\|\tilde{U}^\top U x\|_2 \geq \|x\|_2 \sqrt{1 - 2\|E\|_{L^2}^2/\sigma_m(\tilde{A})^2}$.*

Lemma 7 (Pseudo-inverse Theorem for Qmatrices)**.** *Let $A, \tilde{A}, E \in \mathbb{R}^{[a,b]\times m}$ and $\tilde{A} = A + E$. Then,*

$$\sigma_1(A^\dagger - \tilde{A}^\dagger) \;\leq\; 3 \max\{\sigma_1(A^\dagger)^2, \sigma_1(A^\dagger)^2\}\, \sigma_1(E).$$

5.2 Concentration Bound for the Kernel Density Estimator

Next, we bound the error for kernel density estimation. To obtain the best rates under Hölderian assumptions on O, the kernels used in KDE need to be of *order* β. A β order kernel satisfies,

$$\int_0^1 K(s)\mathrm{d}s = 1, \qquad \int_0^1 s^\alpha K(s)\mathrm{d}s = 0, \text{for all } \alpha \leq \lfloor\beta\rfloor, \qquad \int_0^1 s^\beta K(s)\mathrm{d}s \leq \infty. \tag{5}$$

Such kernels can be constructed using Legendre polynomials [18]. Given N i.i.d samples from a d dimensional density f, where $d \in \{1, 2, 3\}$ and $f \in \{P_1, P_{21}, P_{321}\}$, for appropriate choices of the bandwidths h_1, h_{21}, h_{321}, the KDE $\hat{f} \in \{\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}\}$ concentrates around f. Informally, we show

$$\mathbb{P}\left(\|\hat{f} - f\|_{L^2} > \varepsilon\right) \;\lesssim\; \exp\left(-\log(N)^{\frac{d}{2\beta+d}} N^{\frac{2\beta}{2\beta+d}} \varepsilon^2\right). \tag{6}$$

[4] Hsu et al. [2] provide a more refined bound but we use this form to simplify the comparison.

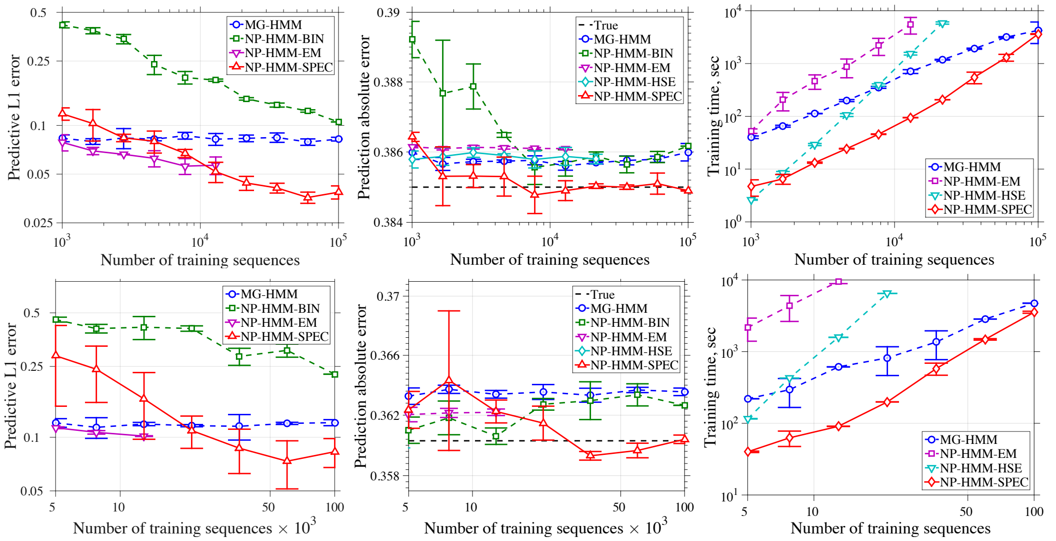

Figure 1: The upper and lower panels correspond to $m = 4$ $m = 8$ respectively. All figures are in log-log scale and the x-axis is the number of triples used for training. **Left:** L_1 error between true conditional density $p(x_6|x_{1:5})$, and the estimate for each method. **Middle:** The absolute error between the true observation and a one-step-ahead prediction. The error of the true model is denoted by a black dashed line. **Right:** Training time.

for all sufficiently small ε and $N/\log N \gtrsim \varepsilon^{-2+\frac{d}{\beta}}$. Here $\lesssim, \gtrsim$ denote inequalities ignoring constants. See Appendix C for a formal statement. Note that when the observations are either discrete or parametric, it is possible to estimate the distribution using $O(1/\varepsilon^2)$ samples to achieve ε error in a suitable metric, say, using the maximum likelihood estimate. However, the nonparametric setting is inherently more difficult and therefore the rate of convergence is slower. This slow convergence is also observed in similar concentration bounds for the KDE [25, 26].

A note on the Proofs: For Lemmas 6, 7 we follow the matrix proof in Stewart and Sun [27] and derive several intermediate results for c/q-matrices in the process. The main challenge is that several properties for matrices, e.g. the CS and Schur decompositions, are not known for c/q-matrices. In addition, dealing with various notions of convergences with these infinite objects can be finicky. The main challenge with the KDE concentration result is that we want an L^2 bound – so usual techniques (such as McDiarmid's [13, 18]) do not apply. We use a technical lemma from Giné and Guillou [26] which allows us to bound the L^2 error in terms of the VC characteristics of the class of functions induced by an i.i.d sum of the kernel. The proof of theorem 5 just mimics the discrete case analysis of Hsu et al. [2]. While, some care is needed (e.g. $\|x\|_{L^2} \leq \|x\|_{L^1}$ does not hold for functional norms) the key ideas carry through once we apply Lemmas 21, 6, 7 and (6). A more refined bound on N that is tighter in $\operatorname{polylog}(N)$ terms is possible – see Corollary 25 and equation 13 in the appendix.

6 Experiments

We compare NP-HMM-SPEC to the following. MG-HMM: An HMM trained using EM with the emissions modeled as a mixture of Gaussians. We tried $2, 4$ and 8 mixtures and report the best result. NP-HMM-BIN: A naive baseline where we bin the space into n intervals and use the discrete spectral algorithm [2] with n states. We tried several values for n and report the best. NP-HMM-EM: The Nonparametric EM heuristic of [12]. NP-HMM-HSE: The Hilbert space embedding method of [15].

Synthetic Datasets: We first performed a series of experiments on synthetic data where the true distribution is known. The goal is to evaluate the estimated models against the *true* model. We generated triples from two HMMs with $m = 4$ and $m = 8$ states and nonparametric emissions. The details of the set up are given in Appendix E. Figure 1 presents the results.

First we compare the methods on estimating the one step ahead conditional density $p(x_6|x_{1:5})$. We report the L^1 error between the true and estimated models. In Figure 2 we visualise the estimated one step ahead conditional densities. NP-HMM-SPEC outperforms all methods on this metric. Next, we compare the methods on the prediction performance. That is, we sample sequences of length 6 and test how well a learned model can predict x_6 conditioned on $x_{1:5}$. When comparing on squared error, the best predictor is the mean of the distribution. For all methods we use the mean of $\widehat{p}(x_6|x_{1:5})$ except

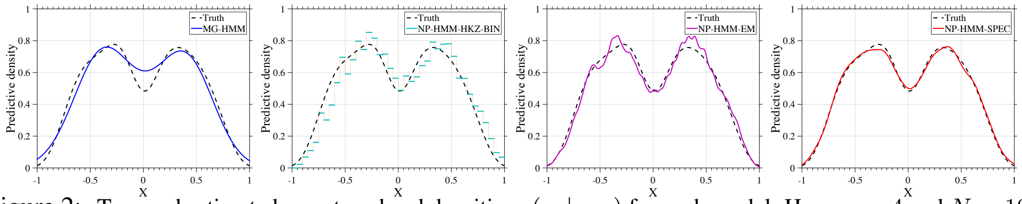

Figure 2: True and estimated one step ahead densities $p(x_4|x_{1:3})$ for each model. Here $m = 4$ and $N = 10^4$.

Dataset	MG-HMM	NP-HMM-BIN	NP-HMM-HSE	NP-HMM-SPEC
Internet Traffic	0.143 ± 0.001	0.188 ± 0.004	0.0282 ± 0.0003	$\mathbf{0.016 \pm 0.0002}$
Laser Gen	0.33 ± 0.018	0.31 ± 0.017	0.19 ± 0.012	$\mathbf{0.15 \pm 0.018}$
Patient Sleep	0.330 ± 0.002	0.38 ± 0.011	$\mathbf{0.197 \pm 0.001}$	0.225 ± 0.001

Table 1: The mean prediction error and the standard error on the 3 real datasets.

for NP-HMM-HSE for which we used the mode since the mean cannot be computed. No method can do better than the true model (shown via the dotted line) in expectation. NP-HMM-SPEC achieves the performance of the true model with large datasets. Finally, we compare the training times of all methods. NP-HMM-SPEC is orders of magnitude faster than NP-HMM-HSE and NP-HMM-EM.

Note that the error of MG-HMM—a parametric model—stops decreasing even with large data. This is due to the bias introduced by the parametric assumption. We do not train NP-HMM-EM for longer sequences because it is too slow. A limitation of the NP-HMM-HSE method is that it cannot recover conditional probabilities, so we exclude it from that experiment. We could not include the method of [4] in our comparisons since their code was not available and their method is not straightforward to implement. Further, their method cannot compute joint/predictive probabilities.

Real Datasets: We compare all the above methods (except NP-HMM-EM which was too slow) on prediction error on 3 real datasets: internet traffic [28], laser generation [29] and sleep data [30]. The details on these datasets are in Appendix E. For all methods we used the mode of the conditional distribution $p(x_{t+1}|x_{1:t})$ as the prediction as it performed better. For NP-HMM-SPEC, NP-HMM-HSE,NP-HMM-BIN we follow the procedure outlined in Section 4.3 to create triples and train with the triples. In Table 1 we report the mean prediction error and the standard error. NP-HMM-HSE and NP-HMM-SPEC perform better than the other two methods. However, NP-HMM-SPEC was faster to train (and has other attractive properties) when compared to NP-HMM-HSE.

7 Conclusion

We proposed and studied a method for estimating the observable representation of a Hidden Markov Model whose emission probabilities are smooth nonparametric densities. We derive a bound on the sample complexity for our method. While our algorithm is similar to existing methods for discrete models, many of the ideas that generalise it to the nonparametric setting are new. In comparison to other methods, the proposed approach has some desirable characteristics: we can recover the joint/conditional densities, our theoretical results are in terms of more interpretable metrics, the method outperforms baselines and is orders of magnitude faster to train.

In this exposition only focused on one dimensional observations. The multidimensional case is handled by extending the above ideas and technology to multivariate functions. Our algorithm and the analysis carry through to the d-dimensional setting, *mutatis mutandis*. The concern however, is more practical. While we have the technology to perform various c/q-matrix operations for $d = 1$ using Chebyshev polynomials, this is not *yet* the case for $d > 1$. Developing efficient procedures for these operations in the high dimensional settings is a challenge for the numerical analysis community and is beyond the scope of this paper. That said, some recent advances in this direction are promising [8, 31].

While our method has focused on HMMs, the ideas in this paper apply for a much broader class of problems. Recent advances in spectral methods for estimating parametric predictive state representations [32], mixture models [3] and other latent variable models [33] can be generalised to the nonparamatric setting using our ideas. Going forward, we wish to focus on such models.

Acknowledgements

The authors would like to thank Alex Townsend, Arthur Gretton, Ahmed Hefny, Yaoliang Yu, and Renato Negrinho for the helpful discussions. This work was supported by NIH R01GM114311, AFRL/DARPA FA87501220324.

References

[1] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*, 1989.

[2] Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. A Spectral Algorithm for Learning Hidden Markov Models. In *COLT*, 2009.

[3] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. *arXiv preprint arXiv:1203.0683*, 2012.

[4] Sajid M. Siddiqi, Byron Boots, and Geoffrey J. Gordon. Reduced-Rank Hidden Markov Models. In *AISTATS*, 2010.

[5] Alex Townsend and Lloyd N Trefethen. Continuous analogues of matrix factorizations. In *Proc. R. Soc. A*, 2015.

[6] Alex Townsend. *Computing with Functions in Two Dimensions*. PhD thesis, University of Oxford, 2014.

[7] Tobin A Driscoll, Nicholas Hale, and Lloyd N Trefethen. Chebfun guide. *Pafnuty Publ*, 2014.

[8] Townsend, Alex and Trefethen, Lloyd N. An extension of chebfun to two dimensions. *SIAM J. Scientific Computing*, 2013.

[9] Michael L Littman, Richard S Sutton, and Satinder P Singh. Predictive representations of state. In *NIPS*, volume 14, pages 1555–1561, 2001.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 1977.

[11] Lloyd R Welch. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, 2003.

[12] Tatiana Benaglia, Didier Chauveau, and David R Hunter. An em-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526, 2009.

[13] Larry Wasserman. *All of Nonparametric Statistics*. Springer-Verlag NY, 2006.

[14] Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. Minimax adaptive estimation of nonparametric hidden markov models. *arXiv preprint arXiv:1501.04787*, 2015.

[15] Le Song, Byron Boots, Sajid M Siddiqi, Geoffrey J Gordon, and Alex Smola. Hilbert space embeddings of hidden markov models. In *ICML*, 2010.

[16] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric Estimation of Multi-View Latent Variable Models. pages 640–648, 2014.

[17] Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 2000.

[18] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

[19] L. Fox and I. B. Parker. *Chebyshev polynomials in numerical analysis*. Oxford U.P. cop., 1968.

[20] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. Society for Industrial and Applied Mathematics, 2012.

[21] Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *Ann. of Stat.*, 1995.

[22] James Robins, Lingling Li, Eric Tchetgen, and Aad W van der Vaart. Quadratic semiparametric Von Mises Calculus. *Metrika*, 69(2-3):227–247, 2009.

[23] Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabás Póczos, Larry Wasserman, and James Robins. Nonparametric Von Mises Estimators for Entropies, Divergences and Mutual Informations. In *NIPS*, 2015.

[24] Woo Young Lee. Weyl's theorem for operator matrices. *Integral Equations and Operator Theory*, 1998.

[25] Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John D. Lafferty, and Larry A. Wasserman. Forest Density Estimation. *Journal of Machine Learning Research*, 12:907–951, 2011.

[26] Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'IHP Probabilités et statistiques*, 2002.

[27] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[28] Vern Paxson and Sally Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 1995.

[29] U Hübner, NB Abraham, and CO Weiss. Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared NH 3 laser. *Physical Review A*, 1989.

[30] Santa Fe Time Series Competition. `http://www-psych.stanford.edu/ andreas/Time-Series/SantaFe.html`.

[31] Hashemi, B. and Trefethen, L. N. Chebfun to three dimensions. *In preparation*, 2016.

[32] Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive State Representations: A New Theory for Modeling Dynamical Systems. In *UAI*, 2004.

[33] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models. *JMLR*, 2014.