Learning HMMs with Nonparametric Emissions via Spectral Decompositions of Continuous Matrices

Maruan Al-Shedivat* Kirthevasan Kandasamy* Eric P. Xing

Summary

- **NP-HMM-SPEC:** Fast spectral learning algorithm with provable guarantees for HMMs with *nonparametric emissions.*
- Perturbation theory results are extended to *continuous matrices.*
- Sample complexity bounds for nonparametric HMMs are established.
- NP-HMM-SPEC algorithm uses Chebyshev polynomial expansion which enables fast continuous matrix operations.

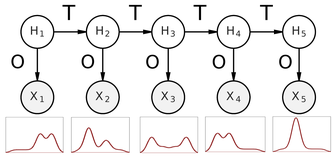
Code: https://github.com/alshedivat/nphmm

1. Background & Motivation

Spectral learning of discrete HMMs [Hsu et al., 2009]

- **Simple to implement.**
  Based on estimating $P_1 = P(x_1), P_{21} = P(x_1, x_2), P_{3x1} = P(x_1, x_2 = x, x_3)$.
  Key step: compute the SVD of the estimated correlation matrix, $\hat{P}_{21}$.
- **Theoretically appealing.**
- **An order of magnitude faster than EM.**

What if the observations are continuous variables?



- Transition: $T \in \mathbb{R}^{m \times m}$.
- Observation densities, $O_{1:m}$, are β-smooth.

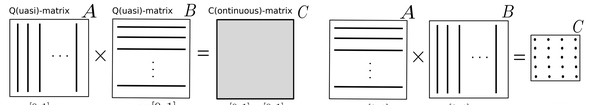
- **Discretize the observation space**, e.g., using bucketing.
  Drawbacks: *simple, but does not work well.*
- **Use a parametric model for emissions**, e.g., mixture of gaussians.
  Drawbacks: *introduces potentially irrelevant biases.*
- **Embed HMM into an RKHS** [Song et al., 2010].
  Drawbacks: *does not recover predictive probabilities and is slower.*

Our proposal

- Use continuous linear algebra with fast Chebyshev approximations.
- Virtually re-use the same algorithm with similar guarantees & speed.

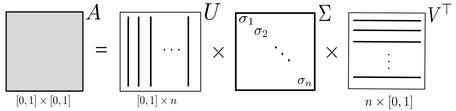
2. Continuous Linear Algebra

Continuous matrix operations



$$C(x,y) = \sum_{i=1}^{n} A(x,i)B(i,y) \qquad C_{ij} = \int_0^1 A(i,x)B(x,j)dx$$

SVD:



Algorithm: NP-HMM-SPEC

Input: Data $\{X^{(j)} = (X_1^{(j)}, X_2^{(j)}, X_3^{(j)})\}_{j=1}^N$, number of states m.

1. Obtain estimates $\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}$ for P_1, P_{21}, P_{321} via KDE:
$$\widehat{P}_{\mathcal{I}}(t) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{h_{\mathcal{I}}} K_h\left(t - X_{\mathcal{I}}^{(j)}\right), \text{ where } \mathcal{I} \in \{1, 21, 321\}.$$
2. Compute Chebyshev polynomial representations for $\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}$.
3. Compute the (m-truncated) C-matrix SVD, $\widehat{P}_{21} = \widehat{U}\widehat{S}\widehat{V}^\top$.
   Let $\widehat{U} \in \mathbb{R}^{[0,1] \times m}$ be the first m left singular vectors of $\widehat{P}_{21}$.
4. Compute the observable representation:
$$\widehat{b}_1 = \widehat{U}^\top \widehat{P}_1, \quad \widehat{b}_\infty = (P_{21}^\top \widehat{U})^\dagger \widehat{P}_1, \quad \widehat{B}(x) = (\widehat{U}^\top \widehat{P}_{3x1})(\widehat{U}^\top \widehat{P}_{21})^\dagger$$

Output: Estimates $\widehat{b}_1, \widehat{b}_\infty, \widehat{B}(x)$.

Theorem (sample complexity for the bound on L_1-error)
With probability $1 - \delta$, the error $|P(x_1, \ldots, x_t) - \widehat{P}(x_1, \ldots, x_t)|_{L_1} \leq \epsilon$ when

Nonparameteric HMM [this work]

$$\frac{N}{\log N} \gtrsim \left(\frac{t}{\epsilon}\right)^{2+3/\beta} \frac{\kappa(O)^{2+3/\beta} m^{1+3/\beta}}{\sigma_m(P_{21})^{4+4/\beta}} \left(\log \frac{1}{\delta}\right)^{1+3/2\beta}$$

for $t \geq 1$ and observation **Q-matrix** $\boldsymbol{O}$, $\kappa(O) = \sigma_1(O)/\sigma_m(O)$.

Discrete HMM [Hsu et al., 2009]

$$N \gtrsim \left(\frac{t}{\epsilon}\right)^{2} \frac{\kappa(O)^2 m}{\sigma_m(P_{21})^4} \left(\log \frac{1}{\delta}\right)$$

for $t \geq 1$ and observation matrix O, $\kappa(O) = \sigma_1(O)/\sigma_m(O)$.

3. Empirical Results

On synthetic data

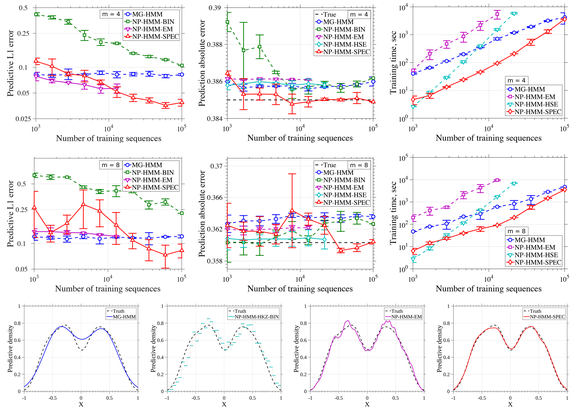


Figure 1: Performance (upper panels) and predictive densities (lower panel).

On real data

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 $\pm$ standard error on the 3 real datasets.

Key Takeaways

NP-HMM-SPEC:

- Provides guarantees in terms of $\boldsymbol{L_1}$ **error**.
- **Orders of magnitude faster** than competitors.
- Returns **full predictive densities** (in contrast to RKHS embeddings).

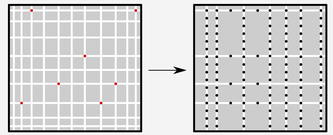
Ongoing research & open questions:

- **Our computational methodology is only suitable for 1D.** Efficient continuous linear algebra in multiple dimensions is an open question.
- Joint model learning & Chebyshev density approximation.

4. Additional Details & Proof Sketches

Chebyshev technology Townsend [2014]

The low-rank decomposition of a cmatrix is done in two stages:



Stage 1: Chebyshev mesh & matrix GE
Complexity: $O(k^3)$

Stage 2: refine mesh & function GE
Complexity: $O(k^2(m+n))$

$$f(x) \approx \sum_{j=1}^{k} d_j c_j(y) r_j(x)$$

$c_j(y)$ and $r_j(x)$ are Chebyshev polynomials that match $f(x)$ at the mesh points

Theorem 4.6 (Convergence of the functional GE) [Townsend, 2014].
If $A \in \mathbb{R}^{[a,b] \times [c,d]}$ is a continuous analytic C-matrix with bounded columns, under certain conditions

$\implies$ GE series converge to A **absolutely, uniformly, and geometrically**.

Proof sketches

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

Assumption 2. All emission densities belong to the *Hölder class*, $\mathcal{H}_1(\beta, L)$, i.e., are β-smooth functions.

Theorem (Concentration bound for KDE).

- Let $f \in \mathcal{H}_d(\beta, L)$ be a density on $[0, 1]^d$.
- The number of samples, N, satisfies $N/\log N \gtrsim \epsilon^{-(2+d/\beta)}$.

$$\implies \mathbb{P}\left(\|\hat{f} - f\|_{L^2} > \epsilon\right) \lesssim \exp\left(-N^{\frac{2\beta}{2\beta+d}}\epsilon^2\right),$$

Perturbation theory for Q/C-matrices

Lemma (Wedin's Sine Theorem for C-matrices).

- $A, \tilde{A}, E \in \mathbb{R}^{[0,1] \times [0,1]}$ where $\tilde{A} = A + E$ and $\mathrm{rank}(A) = m$.
- $U, \tilde{U} \in \mathbb{R}^{[a,b] \times m}$ – the first m left singular vectors of A and $\tilde{A}$.

$$\implies \forall x \in \mathbb{R}^m \ \|\tilde{U}^\top U x\|_2 \geq \|x\|_2 \sqrt{1 - 2\|E\|_{L^2}^2/\sigma_m(\tilde{A})^2}$$

Lemma (Pseudo-inverse Theorem for Q-matrices).

- $A, \tilde{A}, E \in \mathbb{R}^{[0,1] \times [0,1]}$ where $\tilde{A} = A + E$

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

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain

* equal contribution

Correspondence: {alshedivat, kandasamy, epxing}@cs.cmu.edu