Reduced-Rank Hidden Markov Models

Sajid M. Siddiqi
Byron Boots
Geoffrey J. Gordon

Select Lab
Carnegie Mellon University
Sequence of observations: \( Y = [y_1 \ y_2 \ y_3 \ \ldots \ y_\tau] \)
Assume a hidden variable that explains the observations: $X = [x_1 \ x_2 \ x_3 \ \ldots \ x_\tau]$
Assume a hidden variable that explains the observations: \( X = [x_1 \ x_2 \ x_3 \ldots \ x_\tau] \)

Sequence of observations: \( Y = [y_1 \ y_2 \ y_3 \ldots \ y_\tau] \)

Hidden variable is discrete and Markovian
Hidden Markov Models (HMMs)

Assume a hidden variable that explains the observations: $X = [x_1 \ x_2 \ x_3 \ \ldots \ x_\tau]$

Sequence of observations: $Y = [y_1 \ y_2 \ y_3 \ \ldots \ y_\tau]$

Hidden variable is discrete and Markovian
Hidden Markov Models (HMMs)

Assume a **hidden variable** that explains the observations: \( X = [x_1 \ x_2 \ x_3 \ldots x_\tau] \)

Sequence of **observations**: \( Y = [y_1 \ y_2 \ y_3 \ldots y_\tau] \)

Hidden variable is **discrete** and **Markovian**

Popular for modeling: **biological sequences**, **speech**, etc.
Previous Work

Would like to **learn** a HMM from sequences of observations
Previous Work

Would like to learn a HMM from sequences of observations

A popular approach is **Expectation-Maximization** (Baum-Welch)

- Tries to find a maximum-likelihood solution
- Suffers from local maxima
- Impractical (data & computation) for large hidden state spaces
Previous Work

Would like to learn a HMM from sequences of observations

A popular approach is **Expectation-Maximization** (Baum-Welch)

- Tries to find a maximum-likelihood solution
- Suffers from local maxima
- Impractical (data & computation) for large hidden state spaces

Many attempts to reduce local maxima, e.g.

**STACS** - [Siddiqi,Gordon,Moore 2008]

Best-first Model Merging - [Stolcke & Omohundro 1994]

These techniques have not eliminated the problem
Previous Work

An interesting alternative approach:

[Hsu, Kakade, Zhang, 2008]

• A closed-form spectral algorithm for identifying HMMs

• Consistent, finite sample bounds

• No local optima, but small loss in statistical efficiency
This work:

• Generalize spectral learning algorithm to larger class of models
• Supply tighter finite sample bounds
• Apply algorithm to high dimensional data
In particular we introduce a **new model**: 
In particular we introduce a new model:

Hidden Markov Models
consistent learning with finite sample bounds
In particular we introduce a new model:

Predictive State Representations
consistent learning

Hidden Markov Models
In particular we introduce a **new model**: Predictive State Representations

**Reduced-Rank Hidden Markov Models**

consistent learning with finite sample bounds

**Hidden Markov Models**

for fixed latent dimension $k$
Outline

1. Preliminaries

2. Hidden Markov Models

3. Reduced-Rank Hidden Markov Models

4. Learning RR-HMMs & Bounds

5. Empirical Results
HMM Definition

$m$: number of discrete states
$n$: number of discrete observations

$T: m \times m$ column-stochastic transition matrix

$$T_{i,j} = \Pr \left[ x_{t+1} = i \mid x_t = j \right]$$

$O: n \times m$ column stochastic observation matrix

$$O_{i,j} = \Pr \left[ y_t = i \mid x_t = j \right]$$

$\pi: m \times 1$ prior distribution over states

$$\pi_i = \Pr \left[ x_1 = i \right]$$
Observable Operators

[Schützenberger, 1961; Jaeger, 2000]

For each $y \in \{1, \ldots, n\}$, define an $m \times m$ matrix

$$[A_y]_{i,j} \equiv \Pr[x_{t+1} = i \land y_t = y \mid x_t]$$
Observable Operators

[Schützenberger, 1961; Jaeger, 2000]

For each $y \in \{1, \ldots, n\}$, define an $m \times m$ matrix

$$[A_y]_{i,j} \equiv \Pr[x_{t+1} = i \land y_t = y \mid x_t]$$

$$A_y = T \text{diag}(O_y, \cdot)$$
Observable Operators

[Schützenberger, 1961; Jaeger, 2000]

For each \( y \in \{1, \ldots, n\} \), define an \( m \times m \) matrix

\[
[A_y]_{i,j} \equiv \Pr[x_{t+1} = i \wedge y_t = y \mid x_t]
\]

\[
A_y = T \text{diag}(O_y, \cdot)
\]

transition probability

observation likelihood

\[
A_y = \Pr[x_{t+1} \mid x_t] \Pr[y \mid x_t]
\]
Inference in HMMs

\[ \Pr[y_1, y_2, \ldots, y_\tau] \]
Inference in HMMs

\[
\text{Pr}[y_1, y_2, \ldots, y_\tau] = \sum_{x_{\tau+1}} \text{Pr}[x_{\tau+1} | x_\tau] \text{Pr}[y_\tau | x_\tau] \ldots \sum_{x_3} \text{Pr}[x_3 | x_2] \text{Pr}[y_2 | x_2] \sum_{x_2} \text{Pr}[x_2 | x_1] \text{Pr}[y_1 | x_1] \text{Pr}[x_1]
\]
Inference in HMMs

\[
Pr[y_1, y_2, \ldots, y_\tau] = \sum_{x_{\tau+1}} \Pr[x_{\tau+1} | x_\tau] \Pr[y_\tau | x_\tau] \ldots \sum_{x_3} \Pr[x_3 | x_2] \Pr[y_2 | x_2] \sum_{x_2} \Pr[x_2 | x_1] \Pr[y_1 | x_1] \Pr[x_1] \times 1_1^\top T \text{diag}(O_{y_\tau,}) \ldots T \text{diag}(O_{y_2,}) T \text{diag}(O_{y_1,}) \pi
\]
Inference in HMMs

\[
Pr[y_1, y_2, \ldots, y_\tau] = \sum_{x_{\tau+1}} \Pr[x_{\tau+1} | x_\tau] \Pr[y_\tau | x_\tau] \ldots \sum_{x_3} \Pr[x_3 | x_2] \Pr[y_2 | x_2] \sum_{x_2} \Pr[x_2 | x_1] \Pr[y_1 | x_1] \Pr[x_1] \\
\rightarrow 1^T_m T \text{diag}(O_{y_\tau, \tau}) \ldots T \text{diag}(O_{y_2, 2}) T \text{diag}(O_{y_1, 1}) \pi
\]
Inference in HMMs

\[
\sum_{x_{\tau+1}} \Pr[x_{\tau+1} | x_{\tau}] \Pr[y_{\tau} | x_{\tau}] \ldots \sum_{x_3} \Pr[x_3 | x_2] \Pr[y_2 | x_2] \sum_{x_2} \Pr[x_2 | x_1] \Pr[y_1 | x_1] \Pr[x_1] \\
1^T_m T \text{diag}(O_{y_{\tau}, \ldots}) \ldots T \text{diag}(O_{y_2, \ldots}) T \text{diag}(O_{y_1, \ldots}) \pi \\
1^T_m A_{y_{\tau}} \ldots A_{y_2} A_{y_1} \pi
\]
Inference in HMMs

Pr[y_1, y_2, ..., y_\tau]

\[ = \sum_{x_{\tau+1}} \Pr[x_{\tau+1} | x_\tau] \Pr[y_\tau | x_\tau] \cdots \sum_{x_3} \Pr[x_3 | x_2] \Pr[y_2 | x_2] \sum_{x_2} \Pr[x_2 | x_1] \Pr[y_1 | x_1] \Pr[x_1] \]

Inference in an HMM is: \(O(\tau m^2)\)
Problems with HMMs

- HMMs that model smoothly evolving systems require a very large number of discrete states
- Inference and learning for such models is hard
Outline

1. Preliminaries

2. Hidden Markov Models

3. Reduced-Rank Hidden Markov Models

4. Learning RR-HMMs & Bounds

5. Empirical Results
Reduced-Rank Hidden Markov Models

**Idea**: Even if we have a very large number of discrete states, sometimes distribution lies in a real-valued subspace.

We can take advantage of this fact to perform efficient inference and learning.
Reduced-Rank Hidden Markov Models

We formulate a Reduced-Rank Hidden Markov Model (RR-HMM)
Reduced-Rank Hidden Markov Models

We formulate a Reduced-Rank Hidden Markov Model (RR-HMM) with a low-rank transition matrix.

Parameters:
- $T$: column-stochastic with factors $R$ and $S$
We formulate a Reduced-Rank Hidden Markov Model (RR-HMM) with a low-rank transition matrix

\[ T = R S \]

Parameters:
- \( T \): column-stochastic with factors \( R \) and \( S \)
- \( O \): column-stochastic \( n \times m \) observation matrix
- \( \pi \): prior distribution over states with factors \( R \) and \( \pi_l \)
Inference in RR-HMMs

\[ \Pr [y_1, y_2, y_3, \ldots, y_\tau] \]

can be expressed as

\[ 1_m^\top T \text{diag}(O_{y_\tau,}) \ldots T \text{diag}(O_{y_3,}) T \text{diag}(O_{y_2,}) T \text{diag}(O_{y_1,}) \pi \]
Inference in RR-HMMs

\[
\Pr[y_1, y_2, y_3, \ldots, y_\tau] \\
\text{can be expressed as}
\]

\[
1_m^T T \text{diag}(O_{y_\tau,}) \ldots T \text{diag}(O_{y_3,}) T \text{diag}(O_{y_2,}) T \text{diag}(O_{y_1,}) \pi
\]

\[
1_m^T R S \text{diag}(O_{y_\tau,}) \ldots R S \text{diag}(O_{y_3,}) R S \text{diag}(O_{y_2,}) R S \text{diag}(O_{y_1,}) R \pi_l
\]
Inference in RR-HMMs

\[
\Pr[y_1, y_2, y_3, \ldots, y_\tau]
\]
can be expressed as

\[
1^T_m T \text{diag}(O_{y_\tau, \cdot}) \ldots T \text{diag}(O_{y_3, \cdot}) T \text{diag}(O_{y_2, \cdot}) T \text{diag}(O_{y_1, \cdot}) \pi
\]

\[
1^T_m RS \text{diag}(O_{y_\tau, \cdot}) \ldots RS \text{diag}(O_{y_3, \cdot}) RS \text{diag}(O_{y_2, \cdot}) RS \text{diag}(O_{y_1, \cdot}) R \pi_l
\]

Can group terms into \( k \times k \) observable operators \( W_y \)

\[
W_y \equiv S \text{diag}(O_{y, \cdot}) R
\]

\[
W_y = \begin{bmatrix} S & O_{y, \cdot} & R \end{bmatrix}_{k \times k}^{k \times m} \begin{bmatrix} \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \end{bmatrix}_{m \times m}^{m \times k}
\]
Inference in RR-HMMs

\[ \Pr[ y_1, y_2, y_3, \ldots, y_\tau ] \]

can be expressed as

\[ 1^T_m T \text{diag}(O_{y_\tau},) \ldots T \text{diag}(O_{y_3},) T \text{diag}(O_{y_2},) T \text{diag}(O_{y_1},) \pi \]

\[ 1^T_m R S \text{diag}(O_{y_\tau},) \ldots R S \text{diag}(O_{y_3},) R S \text{diag}(O_{y_2},) R S \text{diag}(O_{y_1},) R \pi_l \]

\[ \rho^T W_{y_\tau} \ldots W_{y_3} W_{y_2} W_{y_1} \pi_l \]

where

\[ W_y \equiv S \text{diag}(O_{y,}) R \]
Inference in RR-HMMs

\[
\Pr[y_1, y_2, y_3, \ldots, y_\tau] \\
\text{can be expressed as} \\
1^T_m \text{diag}(O_{y_\tau, \cdot}) \ldots \text{diag}(O_{y_3, \cdot}) \text{diag}(O_{y_2, \cdot}) \text{diag}(O_{y_1, \cdot}) \pi \\
1^T_m R \text{diag}(O_{y_\tau, \cdot}) \ldots R \text{diag}(O_{y_3, \cdot}) R \text{diag}(O_{y_2, \cdot}) R \text{diag}(O_{y_1, \cdot}) R \pi_l \\
\rho^T W y_\tau \ldots W y_3 W y_2 W y_1 \pi_l \\
\text{where} \\
W_y \equiv S \text{diag}(O_{y, \cdot}) R
\]

Inference in a RR-HMM is only: \(O(\tau k^2)\)
Outline

1. Preliminaries

2. Hidden Markov Models

3. Reduced-Rank Hidden Markov Models

4. Learning RR-HMMs & Bounds

5. Empirical Results
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

**Idea:** Recover observable HMM parameters from probabilities of doubles and triples of observations

\[
P_{2,1}^{i,j} \equiv \Pr[y_2 = i, y_1 = j]
\]

\[
P_{3,y,1}^{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j]
\]

**Idea:** Recover observable HMM parameters from probabilities of doubles and triples of observations
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

\[
P_{2,1}^{i,j} \equiv \Pr[y_2 = i, y_1 = j]
\]

\[
P_{3,y,1}^{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j]
\]
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

\[
[P_{2,1}]_{i,j} \equiv \Pr[y_2 = i, y_1 = j]
\]

\[
[P_{3,y,1}]_{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j]
\]

2. Matrices factor into HMM parameters

\[
P_{2,1} = OT\text{diag}(\pi)O^T
\]

\[
P_{3,y,1} = OA_yT\text{diag}(\pi)O^T
\]
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

\[
[P_{2,1}]_{i,j} \equiv \Pr[y_2 = i, y_1 = j]
\]

\[
[P_{3,y,1}]_{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j]
\]

2. Matrices factor into HMM parameters

\[
P_{2,1} = OT \text{diag}(\pi)O^\top
\]

\[
P_{3,y,1} = OA_y T \text{diag}(\pi)O^\top
\]

\[
A_y \equiv T \text{diag}(O_y, \cdot)
\]
Spectral Learning for HMM Parameters


Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

$$[P_{2,1}]_{i,j} \equiv \Pr[y_2 = i, y_1 = j]$$

$$[P_{3,y,1}]_{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j]$$

2. Matrices factor into HMM parameters

$$P_{2,1} = OT \, \text{diag}(\pi) O^\top$$

$$P_{3,y,1} = OA_y T \, \text{diag}(\pi) O^\top$$

3. Pick a $U$ s.t. $(U^\top O)$ is invertible
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

\[ [P_{2,1}]_{i,j} \equiv \Pr[y_2 = i, y_1 = j] \]

\[ [P_{3,y,1}]_{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j] \]

2. Matrices factor into HMM parameters

\[ P_{2,1} = O T \text{diag}(\pi) O^\top \]

\[ P_{3,y,1} = O A_y T \text{diag}(\pi) O^\top \]

\[ A_y \equiv T \text{diag}(O_y, \cdot) \]

3. Pick a \( U \) s.t. \((U^\top O)\) is invertible

Then: \( B_y \equiv (U^\top P_{3,y,1})(U^\top P_{2,1})^\dagger = (U^\top O)A_y(U^\top O)^{-1} \)
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

\[ [P_{2,1}]_{i,j} \equiv \Pr[y_2 = i, y_1 = j] \]

\[ [P_{3,y,1}]_{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j] \]

2. Matrices factor into HMM parameters

\[ P_{2,1} = OT \text{diag}(\pi)O^T \]

\[ P_{3,y,1} = OA_y T \text{diag}(\pi)O^T \]

\[ A_y \equiv T \text{diag}(O_y, \cdot) \]

3. Pick a \( U \) s.t. \((U^T O)\) is invertible

\[ B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T O)A_y(U^T O)^{-1} \]

similarity transform of the true HMM parameter \( A_y \)
Spectral Learning for HMM Parameters

[HSU, KAKADE, ZHANG, 2008]

Idea: Recover observable HMM parameters from probabilities of doubles and triples of observations

1. Define

\[ P_{2,1}^{i,j} \equiv \Pr[y_2 = i, y_1 = j] \]

\[ P_{3,y,1}^{i,j} \equiv \Pr[y_3 = i, y_2 = y, y_1 = j] \]

2. Matrices factor into HMM parameters

\[ P_{2,1} = OT \text{diag}(\pi)O^T \]

\[ P_{3,y,1} = O A_y T \text{diag}(\pi)O^T \]

\[ A_y \equiv T \text{diag}(O_y, \cdot) \]

3. Pick a \( U \) s.t. \((U^T O)\) is invertible

Then: \( B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T O)A_y(U^T O)^{-1} \)

other parameters can be recovered up to a linear transform as well
Spectral Learning for HMM Parameters

[Hsu, Kakade, Zhang, 2008]

The algorithm:

1. Look at triples of observations $\langle y_1, y_2, y_3 \rangle$ in the data
   estimate frequencies: $\hat{P}_{2,1}$ and $\hat{P}_{3,y,1}$

2. Compute SVD of $\hat{P}_{2,1}$ to find a matrix of the top $m$
   singular vectors $\hat{U}$

3. Find observable operators $\hat{B}_y = (\hat{U}^T \hat{P}_{3,y,1})(\hat{U}^T \hat{P}_{2,1})^\dagger$
Spectral Learning for HMM Parameters

Pros

Transformed parameters allow HMM inference!
(other terms cancel)
Spectral Learning for HMM Parameters

Pros

Transformed parameters allow HMM inference! (other terms cancel)

Can prove finite sample error bounds
Spectral Learning for HMM Parameters

Pros and Cons

Transformed parameters allow HMM inference!
(Other terms cancel)

Can prove finite sample error bounds

However:
Spectral Learning for HMM Parameters

Pros and Cons

Transformed parameters allow HMM inference!
(other terms cancel)

Can prove finite sample error bounds

However:
Inference in large HMMs is still expensive
(data and computation)
Spectral Learning for HMM Parameters

**Pros and Cons**

Transformed parameters allow HMM inference!
(Other terms cancel)

Can prove finite sample error bounds

**However:**

Inference in large HMMs is still expensive
(Data and computation)

Error bounds vacuous if $T$ is low rank.
Spectral Learning for RR-HMMs

The rank of $P_{2,1}$ and $P_{3,y,1}$ depends on $R$ and $S$

$$P_{2,1} = OT \text{diag}(\pi)O^T$$

$$= ORS \text{diag}(\pi)O^T$$
Spectral Learning for RR-HMMs

The rank of $P_{2,1}$ and $P_{3,y,1}$ depends on $R$ and $S$

\[ P_{2,1} = OT \text{diag}(\pi)O^T \]
\[ = ORS \text{diag}(\pi)O^T \]
Spectral Learning for RR-HMMs

The rank of $P_{2,1}$ and $P_{3,y,1}$ depends on $R$ and $S$

$$P_{2,1} = OT \text{diag}(\pi) O^T$$
$$= ORS \text{diag}(\pi) O^T$$

Thin SVD $UV^T$ splits $P_{2,1}$ “inside” $RS$
Spectral Learning for RR-HMMs

We can show that:

\[ B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T OR) W_y (U^T OR)^{-1} \]
Spectral Learning for RR-HMMs

We can show that:

\[ B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T OR)W_y(U^T OR)^{-1} \]

This is a similarity transform of the RR-HMM parameter \( W_y \)
Can estimate other parameters up to a linear transform as well
Spectral Learning for RR-HMMs

We can show that:

\[ B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T OR)W_y (U^T OR)^{-1} \]

This is a similarity transform of the RR-HMM parameter \( W_y \). Can estimate other parameters up to a linear transform as well.

Parameters allow accurate RR-HMM inference (other terms cancel)
Spectral Learning for RR-HMMs

We can show that:

\[ B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T OR) W_y (U^T OR)^{-1} \]

This is a similarity transform of the RR-HMM parameter \( W_y \)
Can estimate other parameters up to a linear transform as well

Parameters allow accurate RR-HMM inference
(other terms cancel)

Learning and inference are independent of \( m \)
Spectral Learning for RR-HMMs

We can show that:

\[ B_y \equiv (U^T P_{3,y,1})(U^T P_{2,1})^\dagger = (U^T OR)W_y(U^T OR)^{-1} \]

This is a similarity transform of the RR-HMM parameter \( W_y \)
Can estimate other parameters up to a linear transform as well

Parameters allow accurate RR-HMM inference
(other terms cancel)

Learning and inference are independent of \( m \)

A \( k \)-dimensional RR-HMM is considerably more expressive than a \( k \)-state HMM (example in paper, and see experiments below)
Bound on Error in Probability Estimates

$N$ training sequences of length $3$ each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$
Bound on Error in Probability Estimates

\( N \) training sequences of length 3 each

Mild assumptions on RR-HMM parameters \( R, S, O, \pi \)

To bound error on joint probability estimates by \( \epsilon \) with probability \( 1 - \eta \)

\[
\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \hat{\Pr}[y_1, \ldots y_t] \right| \leq \epsilon \quad \text{w.p.} \quad 1 - \eta
\]
Bound on Error in Probability Estimates

$N$ training sequences of length 3 each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$

To bound error on joint probability estimates by $\epsilon$ with probability $1 - \eta$

$$\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \widehat{\Pr}[y_1, \ldots y_t] \right| \leq \epsilon \quad w.p. \quad 1 - \eta$$

$N$ must be larger than a term that is

$$\propto \left(\#\text{timesteps}\right)^2, \text{ rank } k, \#\text{observations}$$
Bound on Error in Probability Estimates

$N$ training sequences of length 3 each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$

To bound error on joint probability estimates by $\epsilon$ with probability $1 - \eta$

$$\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \widehat{\Pr}[y_1, \ldots, y_t] \right| \leq \epsilon \quad w.p. \quad 1 - \eta$$

$N$ must be larger than a term that is

$$\propto (\#\text{timesteps})^2, \text{ rank } k, \#\text{observations}$$

as well as

$$\propto \frac{1}{\epsilon^2}, \frac{1}{\sigma_k(OR)^2}, \frac{1}{\sigma_k(P_{2,1})^4}, \log \left( \frac{1}{\eta} \right)$$
Bound on Error in Probability Estimates

$N$ training sequences of length 3 each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$

To bound error on joint probability estimates by $\epsilon$ with probability $1 - \eta$

$$\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \widehat{\Pr}[y_1, \ldots y_t] \right| \leq \epsilon \quad w.p. \quad 1 - \eta$$

$N$ must be larger than a term that is

$\propto (\#\text{timesteps})^2, \text{ rank } k, \#\text{observations}$

as well as

$\propto \frac{1}{\epsilon^2} \frac{1}{\sigma_k(OR)^2}, \frac{1}{\sigma_k(P_{2,1})^4}, \log \left( \frac{1}{\eta} \right)$
Bound on Error in Probability Estimates

$N$ training sequences of length 3 each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$

To bound error on joint probability estimates by $\epsilon$ with probability $1 - \eta$

$$
\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \widehat{\Pr}[y_1, \ldots, y_t] \right| \leq \epsilon \quad \text{w.p.} \quad 1 - \eta
$$

$N$ must be larger than a term that is

$\propto (\#\text{timesteps})^2, \text{rank} \ k, \ #\text{observations}$

as well as $\propto \frac{1}{\epsilon^2}, \frac{1}{\sigma_k(OR)^2}, \frac{1}{\sigma_k(P_{2,1})^4}, \log \left( \frac{1}{\eta} \right)$

large if observations are uninformative
Bound on Error in Probability Estimates

$N$ training sequences of length $3$ each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$

To bound error on joint probability estimates by $\epsilon$ with probability $1 - \eta$

$$\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \hat{\Pr}[y_1, \ldots, y_t] \right| \leq \epsilon \quad \text{w.p.} \quad 1 - \eta$$

$N$ must be larger than a term that is

$$\propto (\text{#timesteps})^2, \text{ rank } k, \text{ #observations}$$

as well as

$$\propto \frac{1}{\epsilon^2}, \frac{1}{\sigma_k(OR)^2}, \frac{1}{\sigma_k(P_{2,1})^4}, \log \left( \frac{1}{\eta} \right)$$

large if observations are uninformative
large if transitions are highly stochastic
Bound on Error in Probability Estimates

$N$ training sequences of length 3 each

Mild assumptions on RR-HMM parameters $R, S, O, \pi$

To bound error on joint probability estimates by $\epsilon$ with probability $1 - \eta$

$$\sum_{y_1, \ldots, y_t} \left| \Pr[y_1, \ldots, y_t] - \widehat{\Pr}[y_1, \ldots y_t] \right| \leq \epsilon \quad \text{w.p.} \quad 1 - \eta$$

$N$ must be larger than a term that is

$\propto (\#\text{timesteps})^2, \text{ rank } k, \#\text{observations}$

as well as

$$\propto \frac{1}{\epsilon^2} \frac{1}{\sigma_k(OR)^2} \frac{1}{\sigma_k(P_{2,1})^4} \log \left( \frac{1}{\eta} \right)$$

large if observations are uninformative

large if transitions are highly stochastic
Proof Intuition

1. Bound \textit{# samples} needed to estimate $P_{2,1}$ and $P_{3,y,1}$ using standard tail inequality bounds

2. Bound \textit{resulting parameter estimation error} by analyzing how errors in $P_{2,1}$ affect its SVD

3. Propagate bound to \textit{error in joint probabilities} computed using estimated parameters
Additional Extensions

See paper for how to:

1. Model systems that require sequences of observations to disambiguate state

2. Use Kernel Density Estimation for continuous observations

3. Use features computed from observations
Outline

1. Preliminaries

2. Hidden Markov Models

3. Reduced-Rank Hidden Markov Models

4. Learning RR-HMMs & Bounds

5. Empirical Results
**Experimental Results**

**Statistical Consistency:**
See paper for an assessment of consistency on a toy problem

**Clock Pendulum Video Texture:**
Learning a smoothly evolving system

**Mobile Robot Vision:**
Assess long range prediction accuracy
Experimental Results

Video Textures

given a short video

Learn 3 models: HMM, LDS, RR-HMM

constrain dimensionality (10) to test expressivity
Experimental Results

**Video Textures**

given a short video

Learn 3 models: HMM, LDS, RR-HMM

constrain dimensionality (10) to test expressivity
Experimental Results

Video Textures

Simulations from models trained on clock data

HMM     LDS     RR-HMM
Experimental Results

Video Textures

Simulations from models trained on clock data

HMM

LDS

RR-HMM
Experimental Results

Video Textures

Simulations from models trained on clock data

HMM
LDS
RR-HMM
Experimental Results

Video Textures

Simulations from models trained on clock data

HMM

LDS

RR-HMM
Experimental Results

Mobile Robot Vision

Goal: Predict future observations after initial tracking.
Experimental Results

Mobile Robot Vision

![Graph showing the comparison of various prediction error metrics across different prediction horizons. The graph plots the average prediction error against the prediction horizon for four methods: Mean, Last, RR-HMM, and LDS. Each method is represented by a different line color, with Mean in black, Last in purple, RR-HMM in blue, and LDS in green. The horizontal axis represents the prediction horizon in steps, ranging from 0 to 100, while the vertical axis shows the average prediction error in units of $10^6$. The graph illustrates trends in error reduction as the prediction horizon increases.]
Conclusion

Summary:

• Introduced the RR-HMM: a model with many of the benefits of a large-state-space HMM, but without the associated inefficiency during inference and learning

• Supplied a spectral learning algorithm and finite sample bounds for the RR-HMM

• Successfully applied the RR-HMM to high dimensional data
Conclusion

Summary:
• Introduced the RR-HMM: a model with many of the benefits of a large-state-space HMM, but without the associated inefficiency during inference and learning.

• Supplied a spectral learning algorithm and finite sample bounds for the RR-HMM

• Successfully applied the RR-HMM to high dimensional data

Related Work:
• Hilbert Space Embeddings of Hidden Markov Models (ICML-2010) [L. Song, B. Boots, S. M. Siddiqi, G. Gordon, A. Smola]

• Closing the Learning-Planning Loop with Predictive State Representations (RSS-2010) [B. Boots, S. M. Siddiqi, G. Gordon]
Thank you!