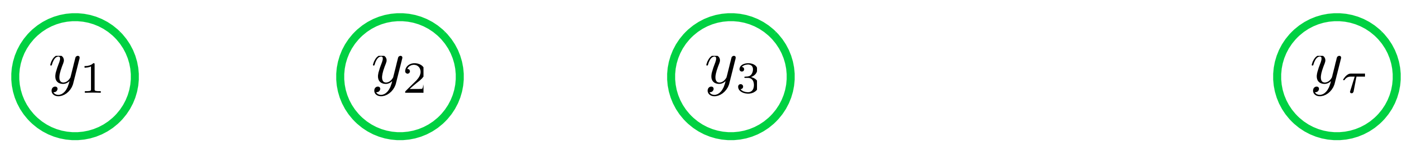
Reduced-Rank Hidden Markov Models

Sajid M. Siddiqi
Byron Boots
Geoffrey J. Gordon

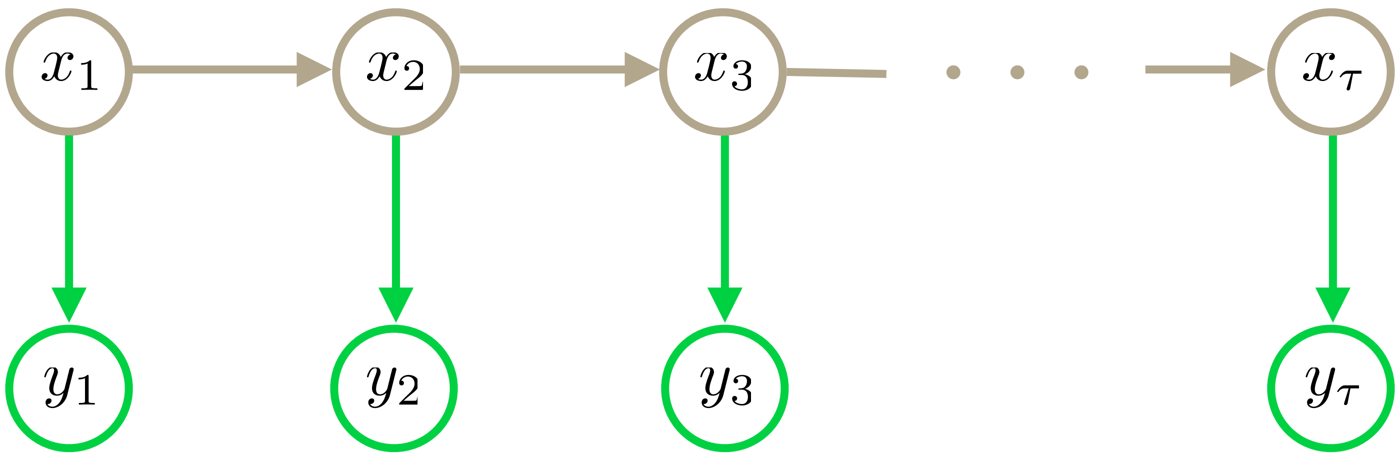
Select Lab

Carnegie Mellon University



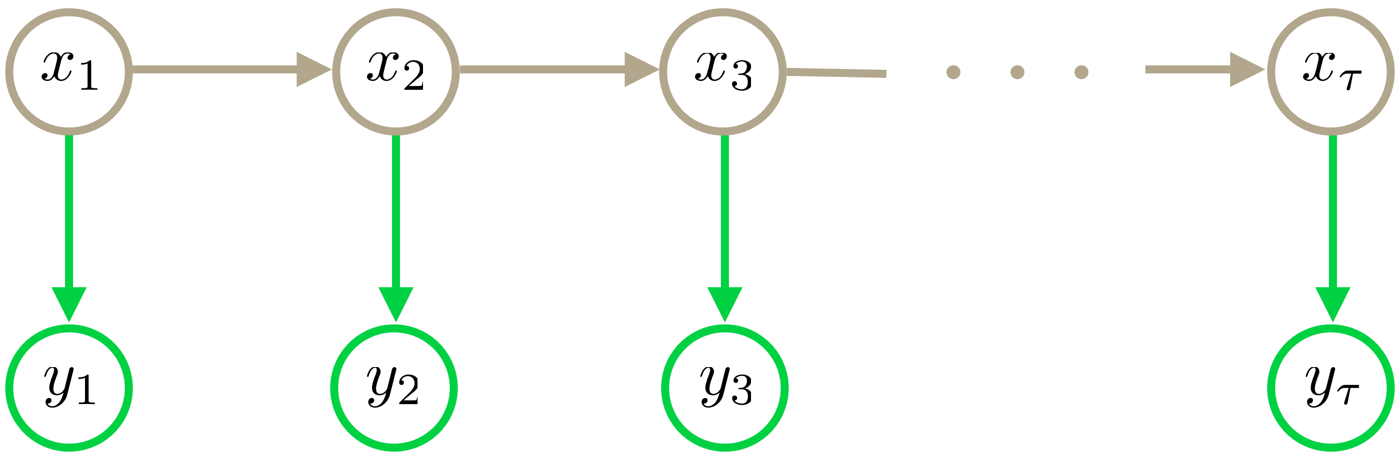
Sequence of observations: $Y = [y_1 \; y_2 \; y_3 \; \ldots \; y_\tau]$

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



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

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

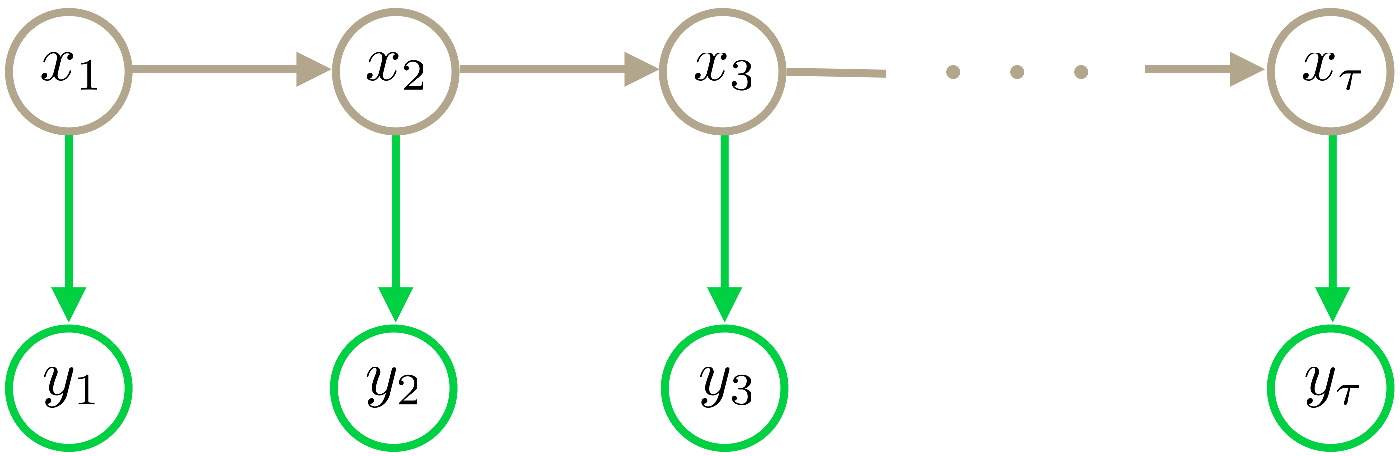


Sequence of observations: $Y = [y_1 \; y_2 \; y_3 \; \ldots \; y_\tau]$

Hidden variable is discrete and Markovian

Hidden Markov Models (HMMs)

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

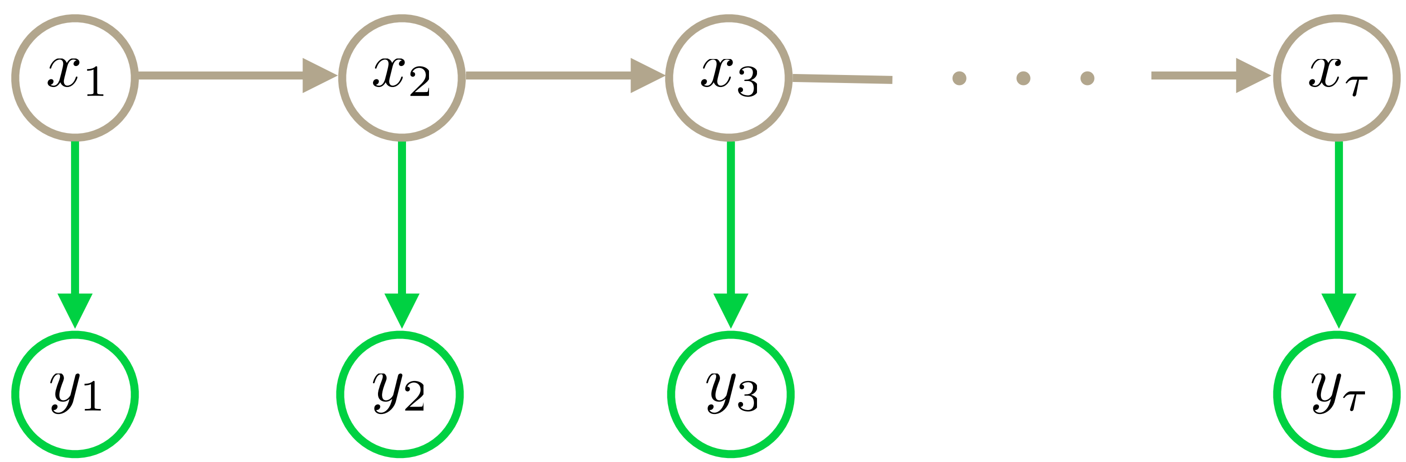


Sequence of observations: $Y = [y_1 \; y_2 \; y_3 \; \ldots \; y_\tau]$

Hidden variable is discrete and Markovian

Hidden Markov Models (HMMs)

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



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

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

Today

This work:

- Generalize spectral learning algorithm to larger class of models
- Supply tighter finite sample bounds
- Apply algorithm to high dimensional data

Overview

In particular we introduce a new model:

Overview

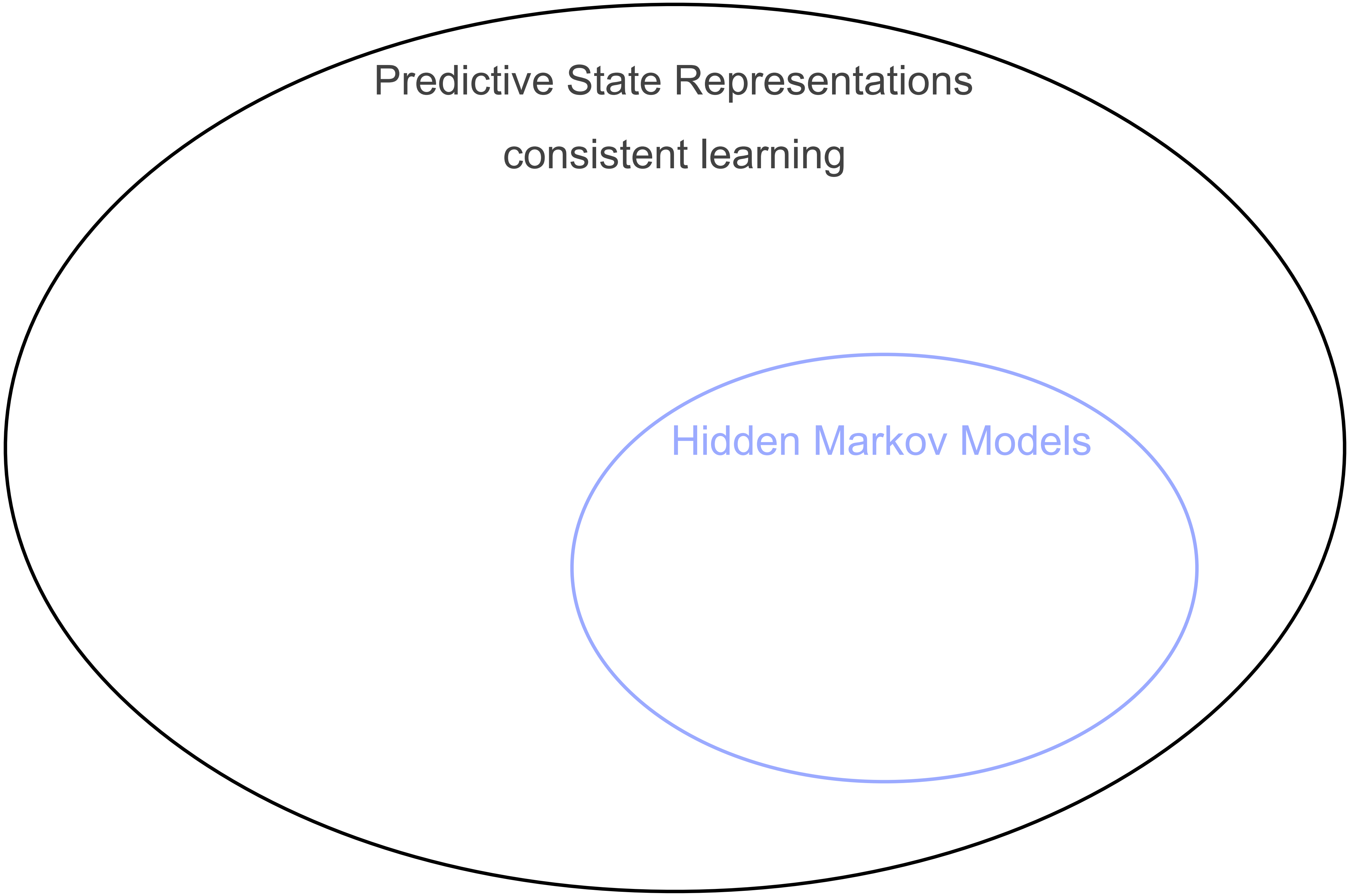
In particular we introduce a new model:

Hidden Markov Models

consistent learning with
finite sample bounds

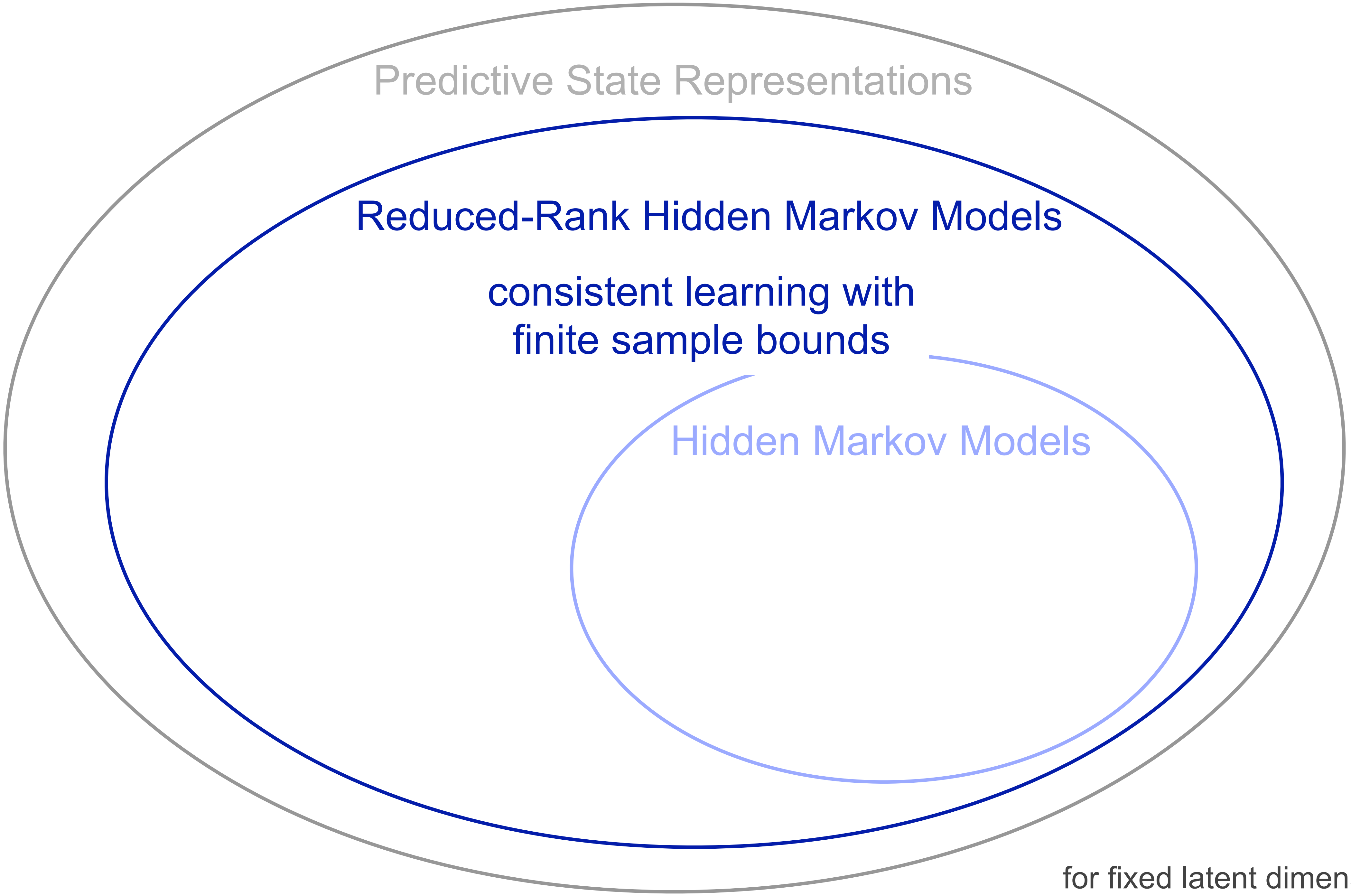
Overview

In particular we introduce a new model:



Overview

In particular we introduce a new model:



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

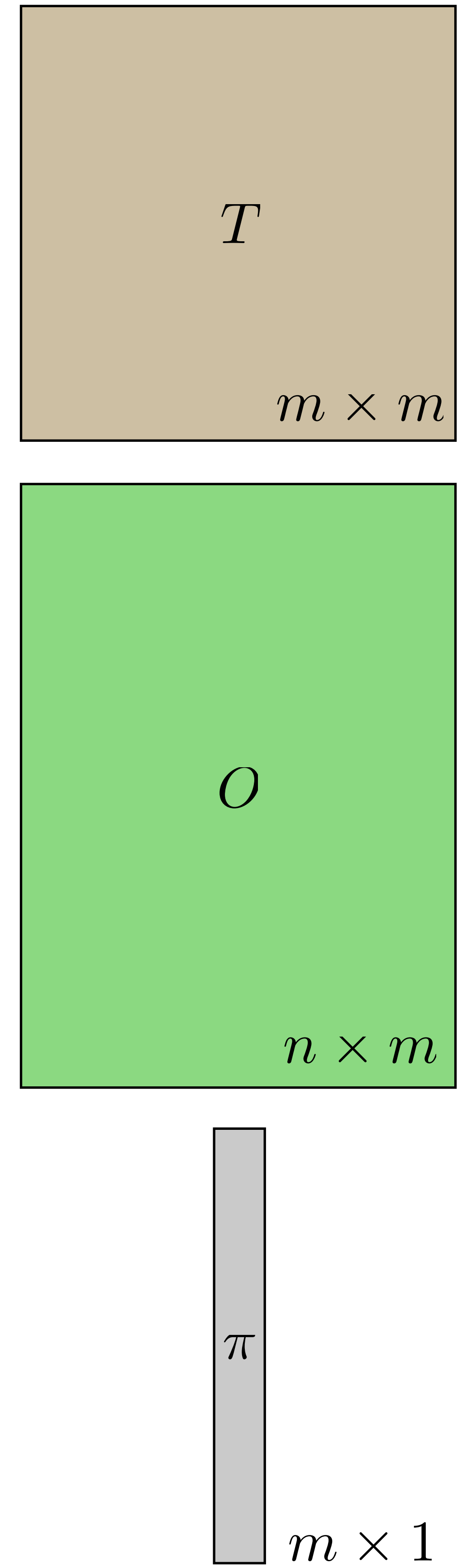
$$T_{i,j} = \Pr\left[x_{t+1} = i \mid x_t = j\right]$$

$O : n \times m$ 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

$$\pi_i = \Pr[x_1 = i]$$

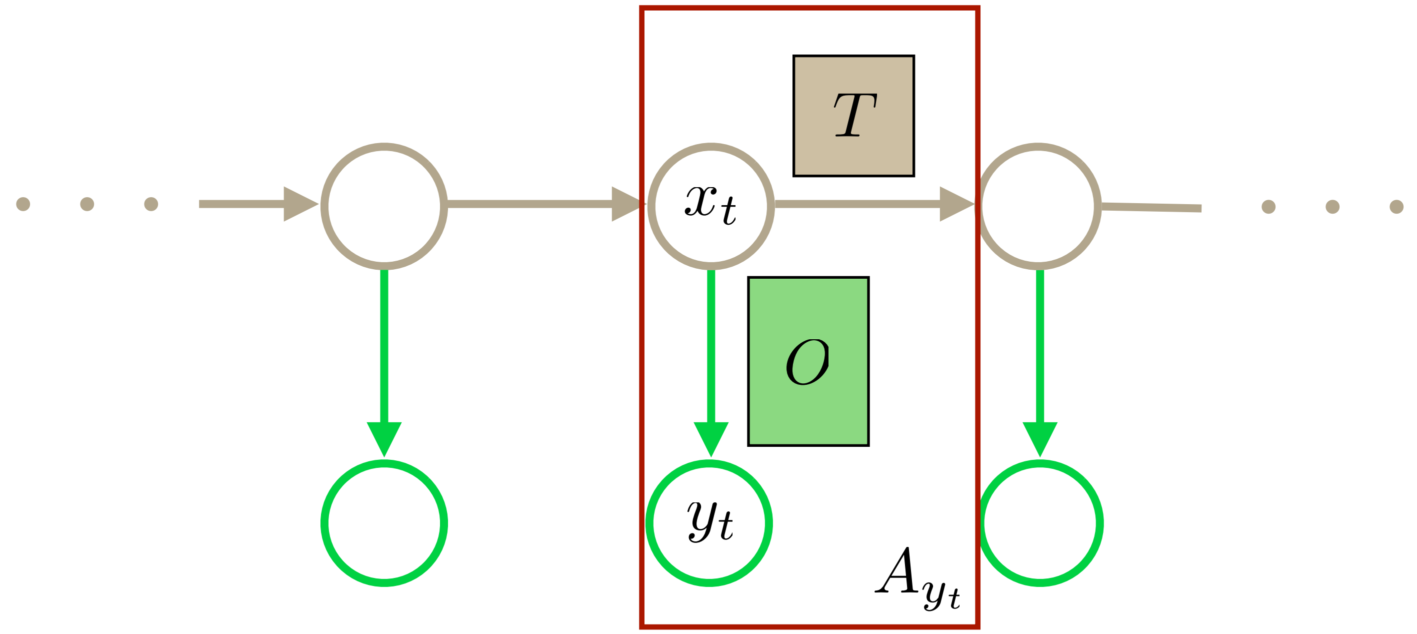


Observable Operators

[Schützenberger, 1961; Jaeger, 2000]

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

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



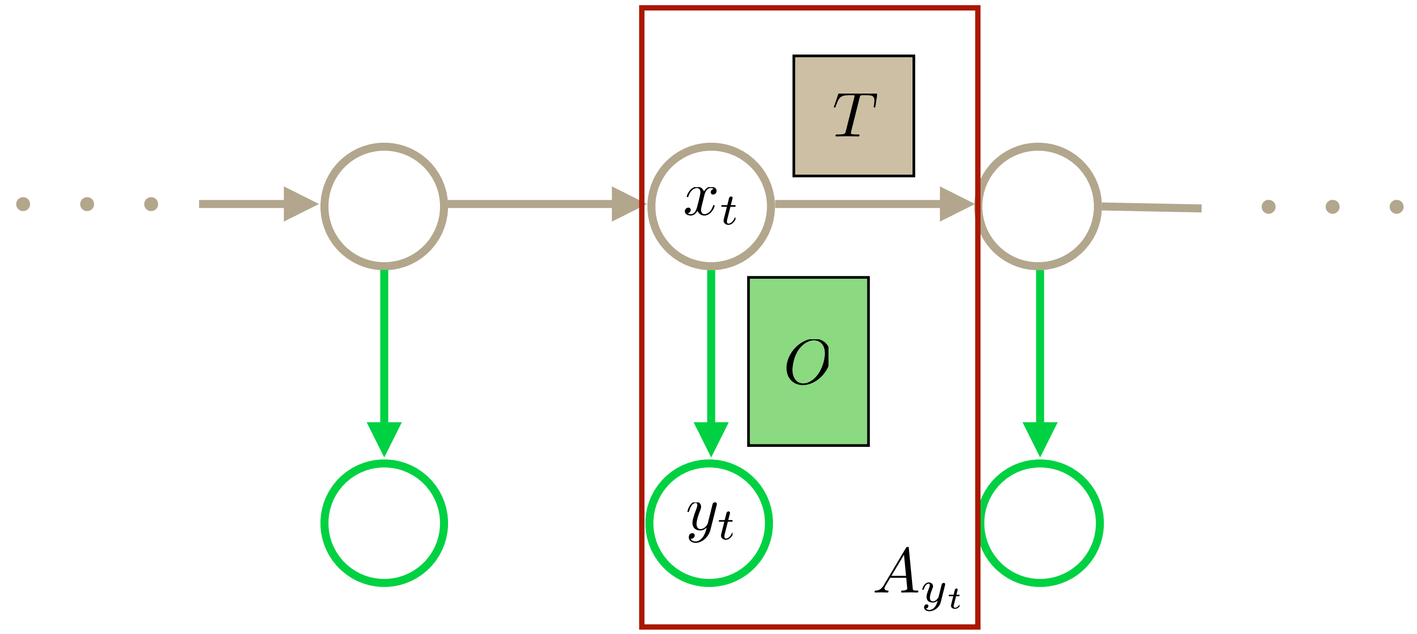
Observable Operators

[Schützenberger, 1961; Jaeger, 2000]

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

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

$$A_y = T\mathrm{diag}(O_{y,\cdot})$$



Observable Operators

[Schützenberger, 1961; Jaeger, 2000]

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

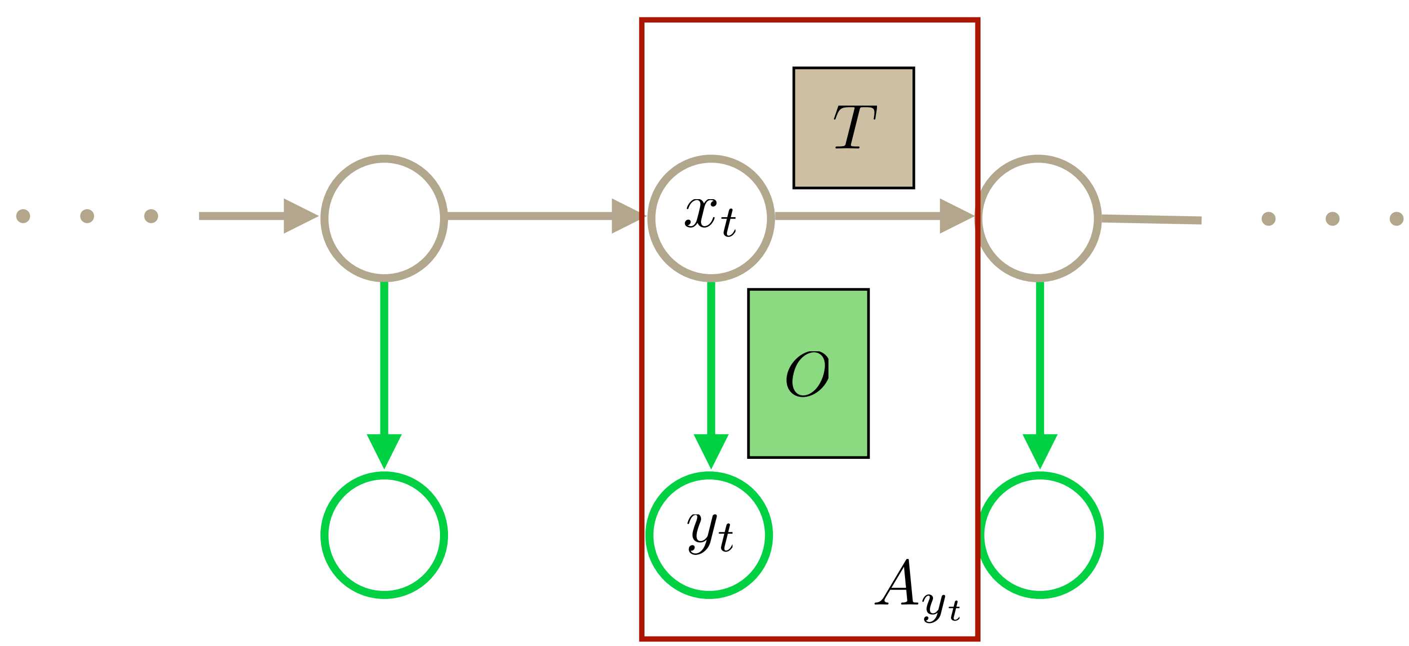
$$[A_y]_{i,j} \equiv \Pr[x_{t+1} = i \wedge y_t = y \mid x_t]$$

$$A_y = T \underbrace{\mathrm{diag}(O_{y,\cdot})}$$

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

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

$$= \sum_{x_{\tau+1}} \Pr[x_{\tau+1} \mid x_\tau] \Pr[y_\tau \mid x_\tau] \ldots \sum_{x_3} \Pr[x_3 \mid x_2] \Pr[y_2 \mid x_2] \sum_{x_2} \Pr[x_2 \mid x_1] \Pr[y_1 \mid x_1] \Pr[x_1]$$

Inference in HMMs

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

$$= \sum_{x_{\tau+1}} \Pr[x_{\tau+1} \mid x_\tau] \Pr[y_\tau \mid x_\tau] \ldots \sum_{x_3} \Pr[x_3 \mid x_2] \Pr[y_2 \mid x_2] \sum_{x_2} \Pr[x_2 \mid x_1] \Pr[y_1 \mid x_1] \Pr[x_1]$$

$$1_m^\mathsf{T} T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

Inference in HMMs

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

$$= \sum_{x_{\tau+1}} \Pr[x_{\tau+1} \mid x_\tau] \Pr[y_\tau \mid x_\tau] \ldots \sum_{x_3} \Pr[x_3 \mid x_2] \Pr[y_2 \mid x_2] \sum_{x_2} \Pr[x_2 \mid x_1] \Pr[y_1 \mid x_1] \Pr[x_1]$$

$$1_m^\top T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

Inference in HMMs

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

$$= \sum_{x_{\tau+1}} \Pr[x_{\tau+1} \mid x_\tau] \Pr[y_\tau \mid x_\tau] \ldots \sum_{x_3} \Pr[x_3 \mid x_2] \Pr[y_2 \mid x_2] \sum_{x_2} \Pr[x_2 \mid x_1] \Pr[y_1 \mid x_1] \Pr[x_1]$$

$$1_m^\top T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

$$1_m^\top A_{y_\tau} \ldots A_{y_2} A_{y_1} \pi$$

Inference in HMMs

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

$$= \sum_{x_{\tau+1}} \Pr[x_{\tau+1} \mid x_\tau] \Pr[y_\tau \mid x_\tau] \ldots \sum_{x_3} \Pr[x_3 \mid x_2] \Pr[y_2 \mid x_2] \sum_{x_2} \Pr[x_2 \mid x_1] \Pr[y_1 \mid x_1] \Pr[x_1]$$

$$1_m^\top T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

$$1_m^\top A_{y_\tau} \ldots A_{y_2} A_{y_1} \pi$$

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

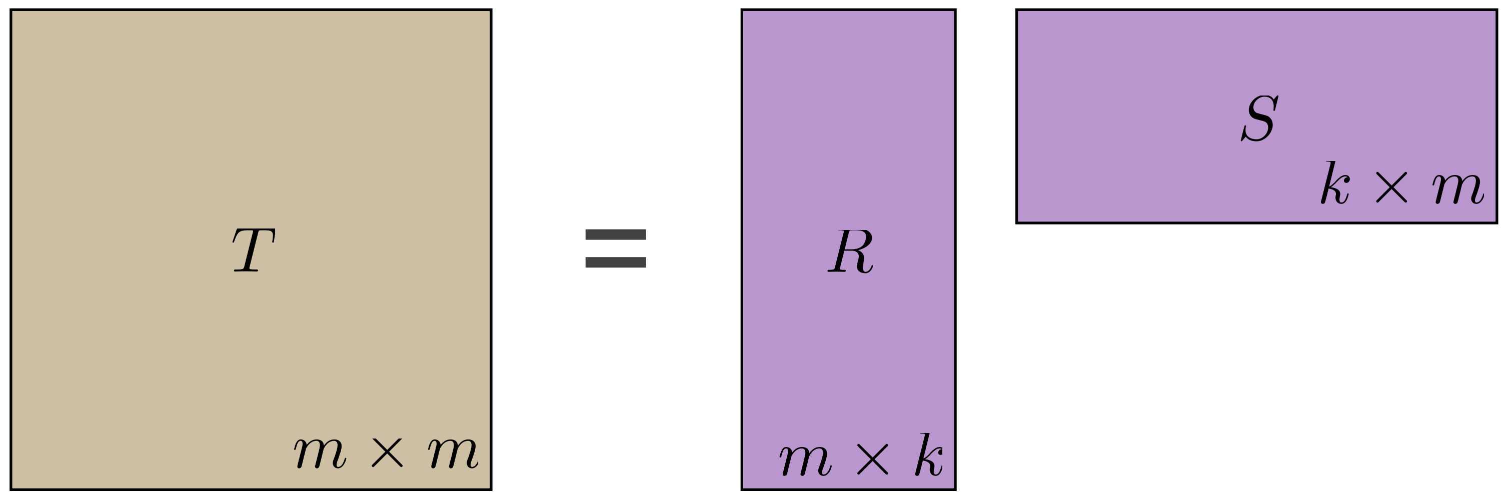
We can take advantage of this fact to perform efficient inference and learning

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

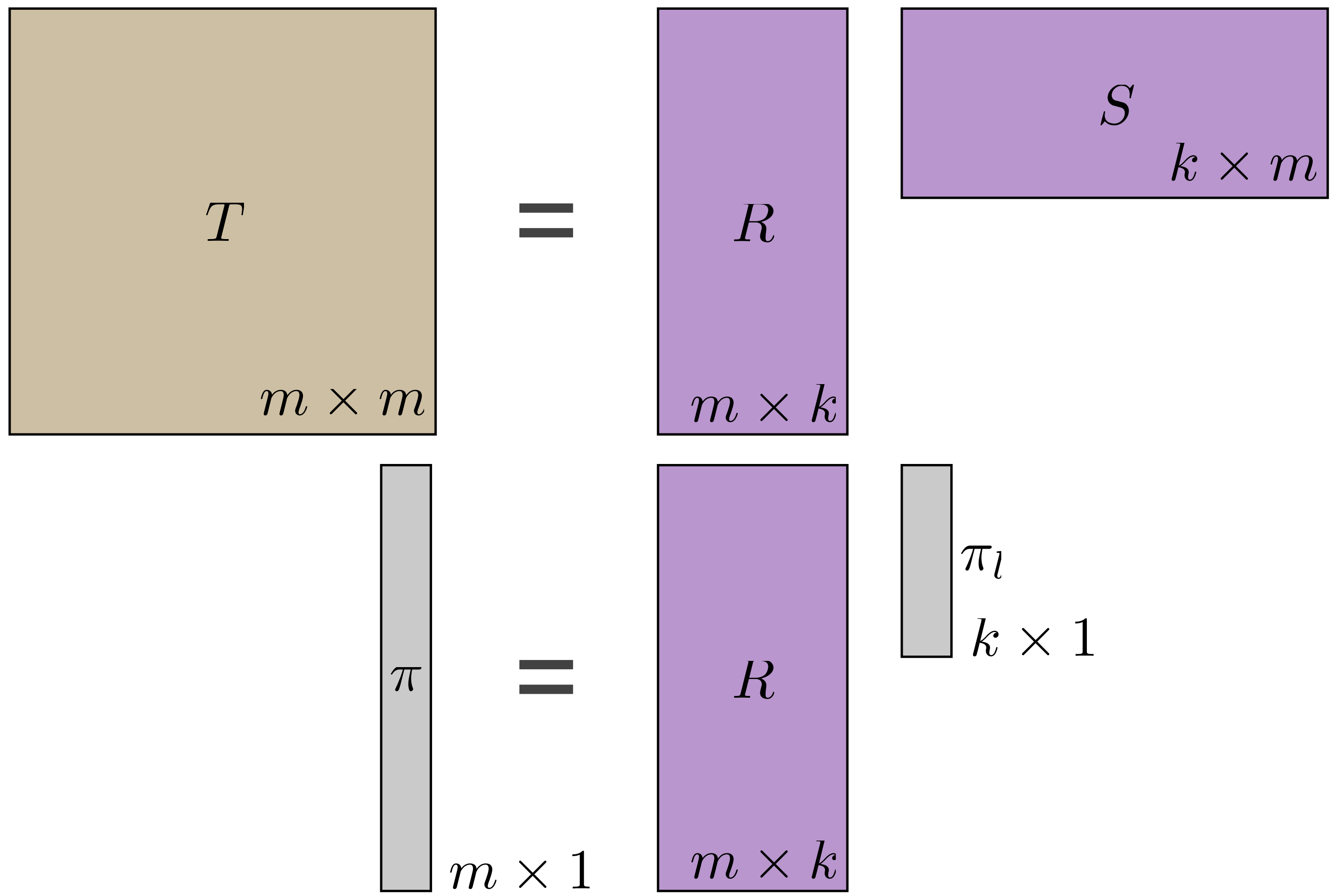


Parameters:

T : column-stochastic with factors R and S

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

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

π : prior distribution over states with factors R and π_l

Inference in RR-HMMs

$$\Pr\left[y_1,\ y_2,\ y_3,\ \ldots, y_\tau\right]$$

can be expressed as

$$1_m^{\mathsf{T}} T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_3,\cdot}) T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

Inference in RR-HMMs

$$\Pr\left[y_1,\ y_2,\ y_3,\ \ldots, y_\tau\right]$$

can be expressed as

$$1_m^\mathsf{T} T \operatorname{diag}(O_{y_\tau,\cdot}) \ldots T \operatorname{diag}(O_{y_3,\cdot}) T \operatorname{diag}(O_{y_2,\cdot}) T \operatorname{diag}(O_{y_1,\cdot}) \pi$$

$$1_m^\mathsf{T} RS \operatorname{diag}(O_{y_\tau,\cdot}) \ldots RS \operatorname{diag}(O_{y_3,\cdot}) RS \operatorname{diag}(O_{y_2,\cdot}) RS \operatorname{diag}(O_{y_1,\cdot}) R\pi_l$$

Inference in RR-HMMs

$$\Pr\left[y_1,\ y_2,\ y_3,\ \ldots, y_\tau\right]$$

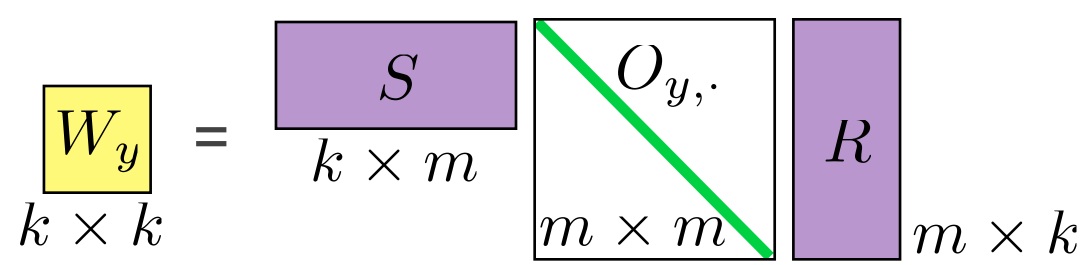
can be expressed as

$$1_m^{\mathsf{T}} T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_3,\cdot}) T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

$$\underbrace{1_m^{\mathsf{T}} R}\underbrace{S \mathrm{diag}(O_{y_\tau,\cdot})} \ldots \underbrace{R S \mathrm{diag}(O_{y_3,\cdot})}\underbrace{R S \mathrm{diag}(O_{y_2,\cdot})}\underbrace{R S \mathrm{diag}(O_{y_1,\cdot}) R}\pi_l$$

Can group terms into $k \times k$ observable operators W_y

$$W_y \equiv S \mathrm{diag}(O_{y,\cdot}) R$$



Inference in RR-HMMs

$$\Pr\left[y_1,\ y_2,\ y_3,\ \ldots, y_\tau\right]$$

can be expressed as

$$1_m^\top T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T\mathrm{diag}(O_{y_3,\cdot}) T\mathrm{diag}(O_{y_2,\cdot}) T\mathrm{diag}(O_{y_1,\cdot})\pi$$

$$1_m^\top R S\mathrm{diag}(O_{y_\tau,\cdot}) \ldots RS\mathrm{diag}(O_{y_3,\cdot}) RS\mathrm{diag}(O_{y_2,\cdot}) RS\mathrm{diag}(O_{y_1,\cdot}) R\pi_l$$

$$\rho^\top W_{y_\tau} \ldots W_{y_3} W_{y_2} W_{y_1} \pi_l$$

where

$$W_y \equiv S\mathrm{diag}(O_{y,\cdot})R$$

Inference in RR-HMMs

$$\Pr\left[y_1,\ y_2,\ y_3,\ \ldots, y_\tau\right]$$

can be expressed as

$$1_m^\mathsf{T} T \mathrm{diag}(O_{y_\tau,\cdot}) \ldots T \mathrm{diag}(O_{y_3,\cdot}) T \mathrm{diag}(O_{y_2,\cdot}) T \mathrm{diag}(O_{y_1,\cdot}) \pi$$

$$1_m^\mathsf{T} R S \mathrm{diag}(O_{y_\tau,\cdot}) \ldots R S \mathrm{diag}(O_{y_3,\cdot}) R S \mathrm{diag}(O_{y_2,\cdot}) R S \mathrm{diag}(O_{y_1,\cdot}) R \pi_l$$

$$\rho^\mathsf{T} W_{y_\tau} \ldots W_{y_3} W_{y_2} W_{y_1} \pi_l$$

where

$$W_y \equiv S \mathrm{diag}(O_{y,\cdot}) R$$

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

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

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

$$P_{3,y,1} = OA_yT\mathrm{diag}(\pi)O^{\mathsf{T}}$$

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

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

$$P_{3,y,1} = OA_yT\mathrm{diag}(\pi)O^{\top}$$

$$A_y \equiv T\mathrm{diag}(O_{y,\cdot})$$

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

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

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

$$A_y \equiv T\text{diag}(O_{y,\cdot})$$

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

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

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

$$P_{3,y,1} = OA_yT\mathrm{diag}(\pi)O^\top$$

$$A_y \equiv T\mathrm{diag}(O_y, \cdot)$$

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

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

$$P_{3,y,1} = OA_yT\mathrm{diag}(\pi)O^\top$$

$$A_y \equiv T\mathrm{diag}(O_y, \cdot)$$

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

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

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

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

$$P_{3,y,1} = OA_yT\mathrm{diag}(\pi)O^\top$$

$$A_y \equiv T\mathrm{diag}(O_{y,\cdot})$$

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

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

similarity transform of the true HMM parameter A_y

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: $\widehat{P}_{2,1}$ and $\widehat{P}_{3,y,1}$

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

3. Find observable operators $\widehat{B}_y = (\widehat{U}^\top \widehat{P}_{3,y,1})(\widehat{U}^\top \widehat{P}_{2,1})^\dagger$

Spectral Learning for HMM Parameters

Pros

Transformed parameters allow HMM inference!
(other terms cancel)

sense
learn
act

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

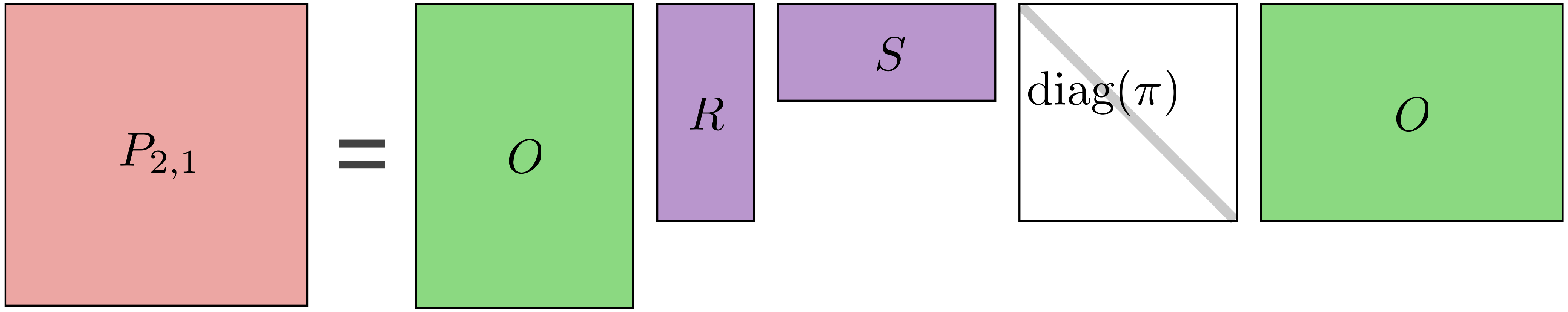
The rank of $P_{2,1}$ and $P_{3,y,1}$ depends on R and S

$$\begin{aligned} P_{2,1} &= OT\mathrm{diag}(\pi)O^{\mathsf{T}} \\ &= ORS\mathrm{diag}(\pi)O^{\mathsf{T}} \end{aligned}$$

Spectral Learning for RR-HMMs

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

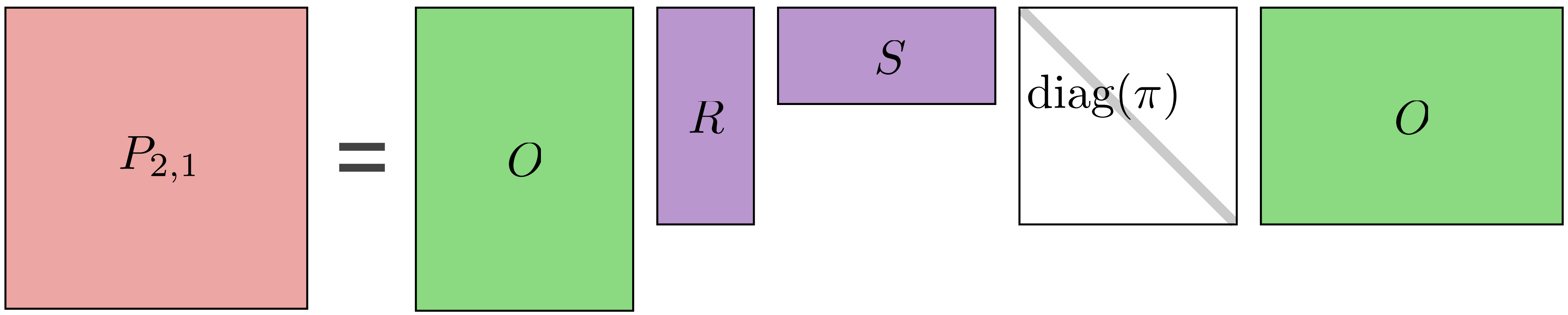
$$\begin{aligned} P_{2,1} &= OT\mathrm{diag}(\pi)O^{\mathsf{T}} \\ &= ORS\mathrm{diag}(\pi)O^{\mathsf{T}} \end{aligned}$$



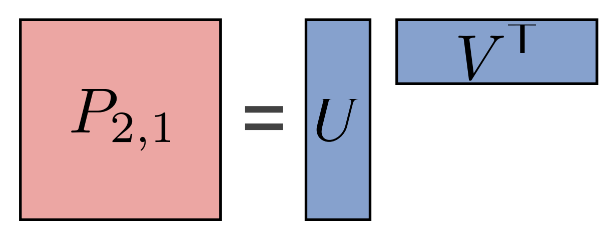
Spectral Learning for RR-HMMs

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

$$\begin{aligned} P_{2,1} &= OT\mathrm{diag}(\pi)O^\top \\ &= ORS\mathrm{diag}(\pi)O^\top \end{aligned}$$



Thin SVD $UV^\top$ splits $P_{2,1}$ "inside" RS



Spectral Learning for RR-HMMs

We can show that:

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

Spectral Learning for RR-HMMs

We can show that:

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

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:

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

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:

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

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:

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

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, π

Bound on Error in Probability Estimates

N training sequences of length 3 each

Mild assumptions on RR-HMM parameters R, S, O, π

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

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

Bound on Error in Probability Estimates

N training sequences of length 3 each

Mild assumptions on RR-HMM parameters R, S, O, π

To bound error on joint probability estimates by ϵ 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

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

Bound on Error in Probability Estimates

N training sequences of length 3 each

Mild assumptions on RR-HMM parameters R, S, O, π

To bound error on joint probability estimates by ϵ 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 $\propto \dfrac{1}{\epsilon^2}, \dfrac{1}{\sigma_k(OR)^2}, \dfrac{1}{\sigma_k(P_{2,1})^4}, \log\left(\dfrac{1}{\eta}\right)$

Bound on Error in Probability Estimates

N training sequences of length 3 each

Mild assumptions on RR-HMM parameters R, S, O, π

To bound error on joint probability estimates by ϵ 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 $\propto \dfrac{1}{\epsilon^2}, \dfrac{1}{\sigma_k(OR)^2}, \dfrac{1}{\sigma_k(P_{2,1})^4}, \log\left(\dfrac{1}{\eta}\right)$

Bound on Error in Probability Estimates

N training sequences of length 3 each

Mild assumptions on RR-HMM parameters R, S, O, π

To bound error on joint probability estimates by ϵ 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 $\propto \dfrac{1}{\epsilon^2}, \dfrac{1}{\sigma_k(OR)^2}, \dfrac{1}{\sigma_k(P_{2,1})^4}, \log\left(\dfrac{1}{\eta}\right)$

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, π

To bound error on joint probability estimates by ϵ 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 $\propto \frac{1}{\epsilon^2}, \frac{1}{\sigma_k(OR)^2}, \frac{1}{\sigma_k(P_{2,1})^4}, \log\left(\frac{1}{\eta}\right)$

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, π

To bound error on joint probability estimates by ϵ 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 $\propto \frac{1}{\epsilon^2}, \frac{1}{\sigma_k(OR)^2}, \frac{1}{\sigma_k(P_{2,1})^4}, \log\left(\frac{1}{\eta}\right)$

large if observations are uninformative

large if transitions are highly stochastic

Proof Intuition

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

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

3. Propagate bound to 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

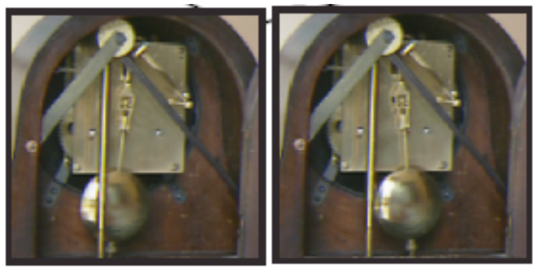
Experimental Results

Statistical Consistency:

See paper for an assessment of consistency on a toy problem

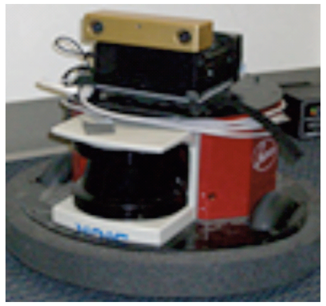
Clock Pendulum Video Texture:

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

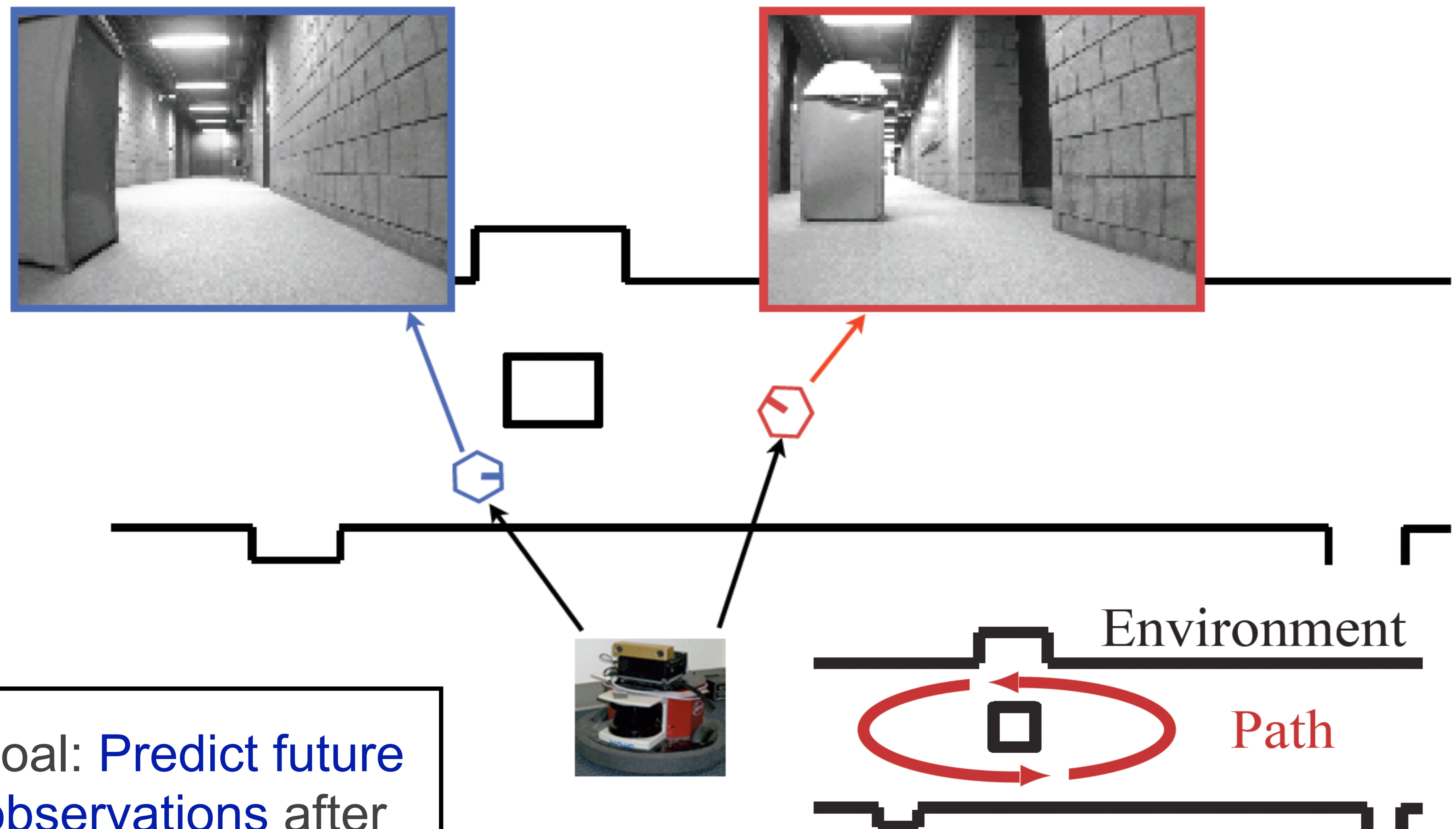
HMM

LDS

RR-HMM

Experimental Results

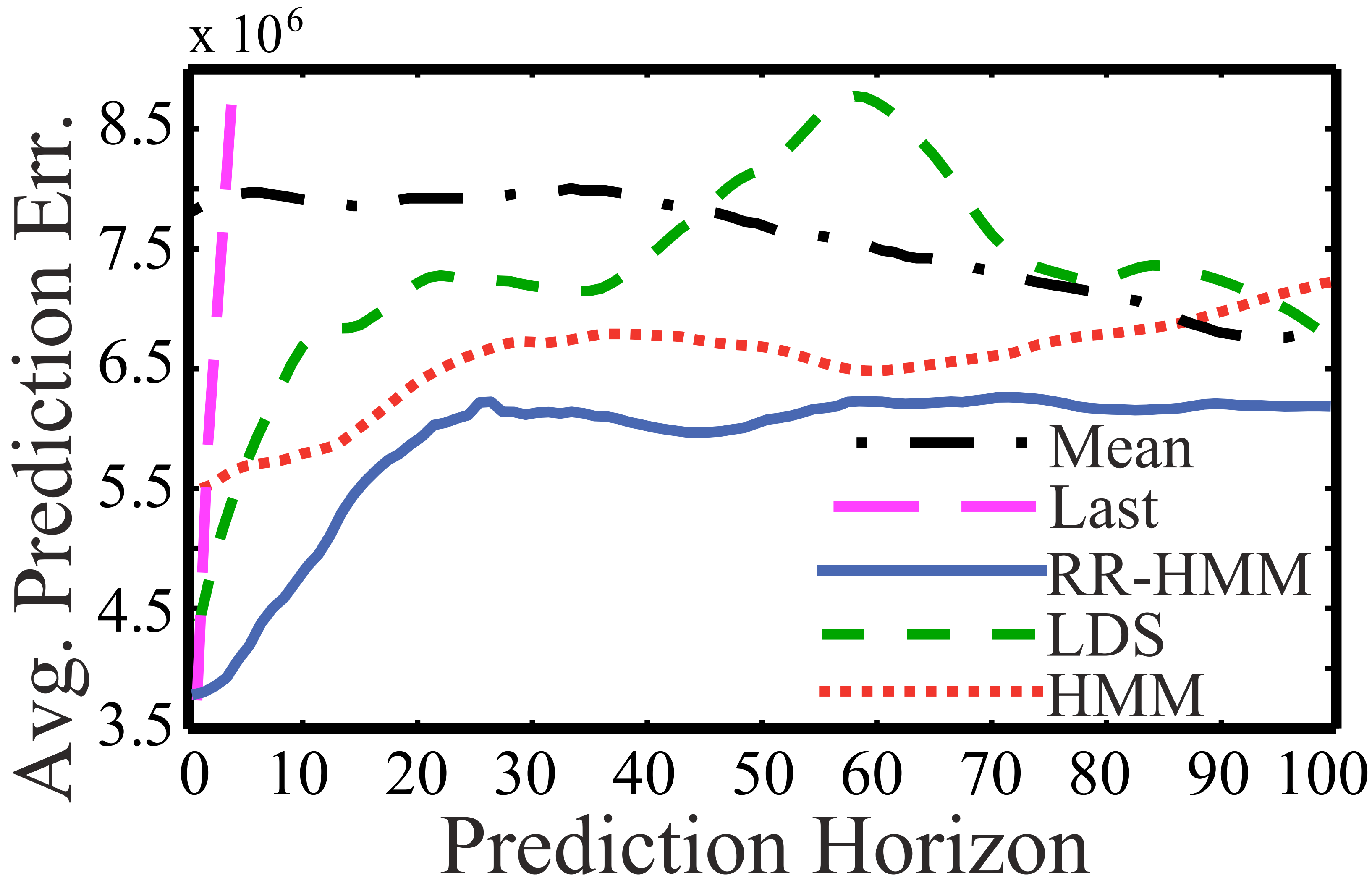
Mobile Robot Vision



Goal: Predict future observations after initial tracking.

Experimental Results

Mobile Robot Vision



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!

sense
learn
act