The Infinite Hidden Markov Model

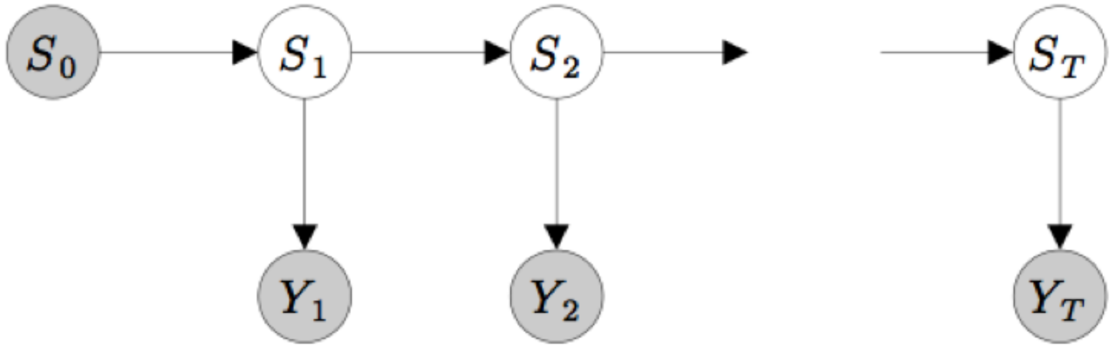
Matthew J. Beal, Zoubin Ghahramani, Carl Edward Rasmussen
University College London
NIPS'02

Presented by **Phuong Nguyen**

Motivation: Modeling time series

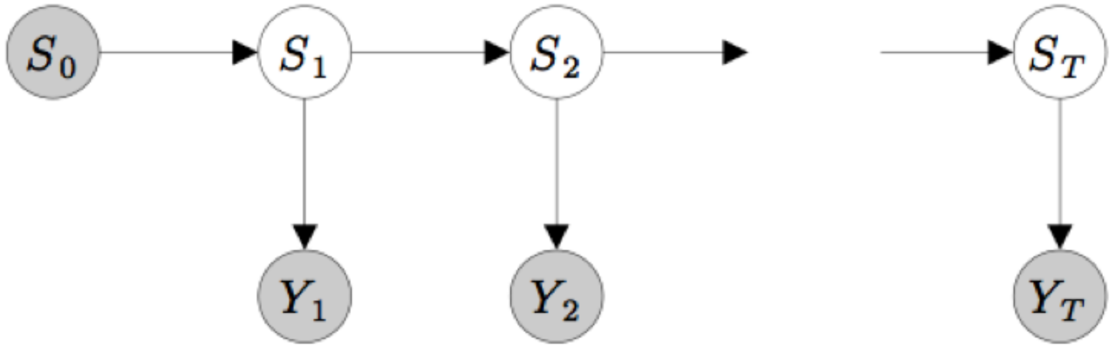
- Given a sequence of observations $\{y_1, y_2, \ldots, y_n\}$, for example:
  - Sequence of images, or words
  - Speech signals
  - Stock prices
  - etc.
- Goal: To build a probabilistic model of the data - something that can predict $P(y_t|y_{t-1}, y_{t-2}, y_{t-3}, \ldots)$

Hidden Markov Model: Causal structure and "hidden variables"



- NLP (e.g., POS tagging):
  - S: part of speech of word
  - Y: word
- Vision:
  - S: object identities, poses, illumination
  - Y: image pixel values

Hidden Markov Model



- Core: hidden K-state Markov chain $s_t \in \{1, 2, \ldots, K\}$
  - Sequence of hidden states has Markov dynamics
  - Observations are independent of all other states
- Parameters
  - Transition matrix: $P(s_t \mid s_{t-1})$
  - Emission matrix: $P(y_t \mid s_t)$

Choosing the number of hidden states

- How do we choose K, the number of hidden states, in an HMM?
- Can we define a model with an *unbounded* number of hidden states, and a suitable inference algorithm?

Idea: Using Dirichlet Process to model transition and emission mechanisms

Dirichlet Process: k finite states

- Drawing n samples $\{c_1, c_2, \ldots, c_n\}$ that take on values $\{1, 2, \ldots, k\}$ with proportion given by π

$$P(c_1, \ldots, c_n | \boldsymbol{\pi}) = \prod_{j=1}^{k} \pi_j^{n_j} \ , \qquad \text{with} \quad n_j = \sum_{n'=1}^{n} \delta(c_{n'}, j)$$

- Put π under a conjugate prior

$$P(\boldsymbol{\pi}|\beta) \sim \text{Dirichlet}(\beta/k, \ldots, \beta/k) \ = \ \frac{\Gamma(\beta)}{\Gamma(\beta/k)^k} \prod_{j=1}^{k} \pi_j^{\beta/k-1}$$

Dirichlet Process: k finite states

$$P(c_1, \ldots, c_n|\boldsymbol{\pi}) = \prod_{j=1}^{k} \pi_j^{n_j} \;, \qquad \text{with} \quad n_j = \sum_{n'=1}^{n} \delta(c_{n'}, j)$$

$$P(\boldsymbol{\pi}|\beta) \sim \mathrm{Dirichlet}(\beta/k, \ldots, \beta/k) \;=\; \frac{\Gamma(\beta)}{\Gamma(\beta/k)^k} \prod_{j=1}^{k} \pi_j^{\beta/k-1}$$

- Joint & conditional probability:

$$P(c_1, \ldots, c_n|\beta) = \int d\boldsymbol{\pi}\; P(c_1, \ldots, c_n|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\beta) = \frac{\Gamma(\beta)}{\Gamma(n+\beta)} \prod_{j=1}^{k} \frac{\Gamma(n_j + \beta/k)}{\Gamma(\beta/k)}$$

$$P(c_d = j|\mathbf{c}_{-d}, \beta) = \frac{n_{-d,j} + \beta/k}{n - 1 + \beta}$$

Dirichlet Process: Infinite states

- What if the number of states is infinite?
- Conditional probability when taking the limit:

$$P(c_d = j \mid \mathbf{c}_{-d}, \beta) = \begin{cases} \frac{n_{-d,j}}{n-1+\beta} & j \in \{1, \ldots, K\} \text{ i.e. represented} \\ \frac{\beta}{n-1+\beta} & \text{for all unrepresented } j\text{, combined} \end{cases}$$

+ where K is the number of presented states,

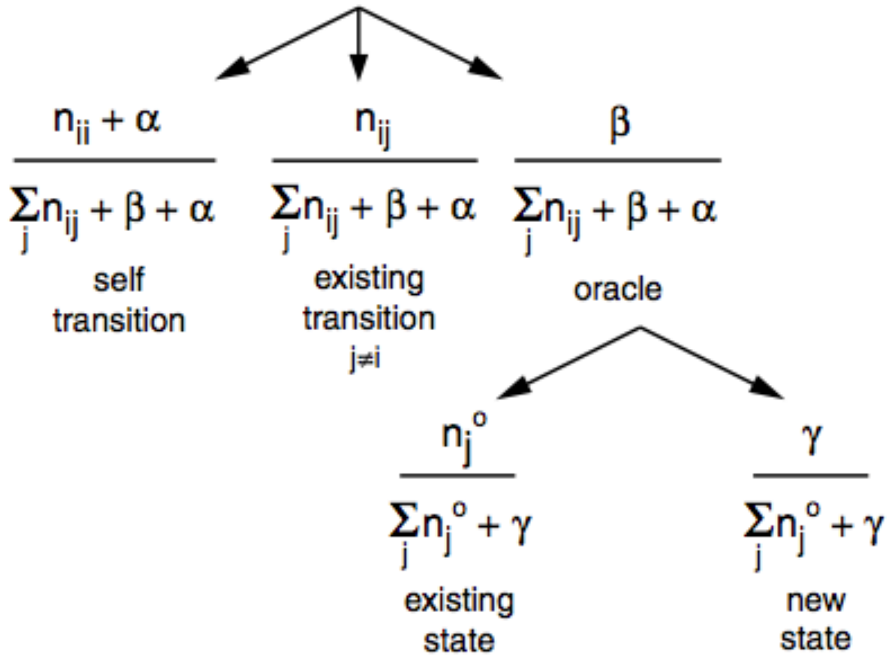
+ β control the tendency to populate a previously unrepresented state

Dirichlet Process: Infinite states

- Take-away results
  - We can integrate out the infinite number of transitions parameters
  - Under DP, there is a natural tendency to use existing transitions => typical trajectories
- Problem:
  - State trajectories under the prior would never visit the same state twice

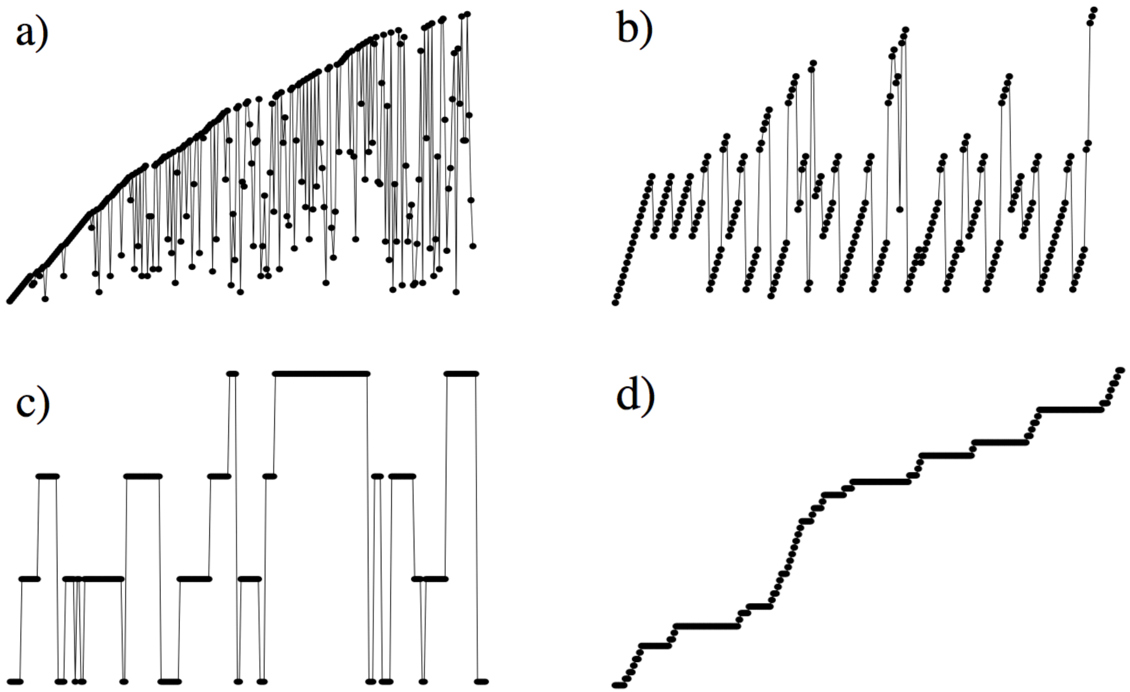
Solution: Hierarchical DP a model for transition and emission for an infinite HMM

HDP: Hidden state transition mechanism



- n_{ij} is the number of previous transitions from i to j
- α, β, and γ are hyperparameters
- Probability of transition from i to j proportional to n_{ij}
- With prob. proportional to β γ jump to a new state

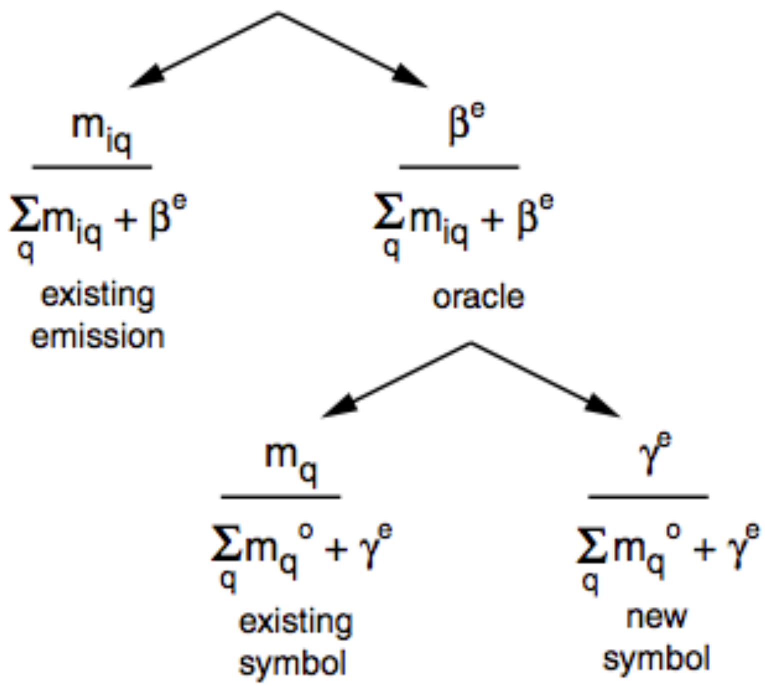
HDP's hidden state transition mechanism: Effects of parameters



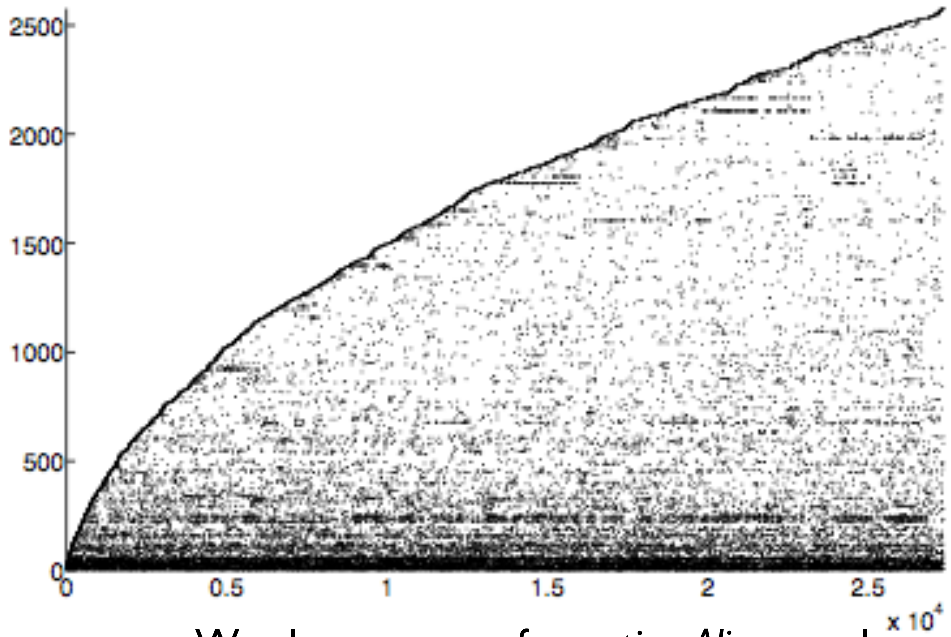
	α	β	γ
(a)	0.1	1000	100
(b)	0	0.1	100
(c)	8	2	2
(d)	1	1	10000

HDP: Emission mechanism

- Identical to transition mechanism, except that: there is no self-transition



State emission generative mechanism



Word occurrence for entire *Alice* novel

Inference in Infinite HMM

- What are the unknowns in iHMM?
  - Hidden state sequence $s = \{s_1, s_2, \ldots, s_T\}$
  - Five hyperparameters $\{\alpha, \beta, \gamma, \beta^e, \gamma^e\}$
- Inference procedure:
  1. Instantiate a random hidden state sequence $\{s_1, s_2, \ldots, s_T\}$
  2. For $t = 1, \ldots, T$
     1. - Gibbs sample s_t given hyperparameter settings, count matrices, and observations.
     2. - Update count matrices to reflect new s_t; this may change K, the number of represented hidden states.
  3. End
  4. Update hyperparameters $\{\alpha, \beta, \gamma, \beta^e, \gamma^e\}$
  5. Go to step 2.

Hyperparameter Optimization

- Hyperparameter approximation:

$$P(\alpha, \beta|\mathbf{s}) \propto \mathcal{G}(a_\alpha, b_\alpha)\mathcal{G}(a_\beta, b_\beta) \prod_{i=1}^{K} \frac{\beta^{K^{(i)}-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(n_{ii}+\alpha)}{\Gamma(\sum_j n_{ij}+\alpha+\beta)} \ ,$$

$$P(\beta^e|\mathbf{s}, \mathbf{y}) \propto \mathcal{G}(a_{\beta^e}, b_{\beta^e}) \prod_{i=1}^{K} \frac{{\beta^e}^{K^{e(i)}}\Gamma(\beta^e)}{\Gamma(\sum_q m_{iq}+\beta^e)} \ ,$$

$$P(\gamma|\mathbf{s}) \propto \mathcal{G}(a_\gamma, b_\gamma)\frac{\gamma^K\Gamma(\gamma)}{\Gamma(T^o+\gamma)} \ , \qquad P(\gamma^e|\mathbf{s}, \mathbf{y}) \propto \mathcal{G}(a_{\gamma^e}, b_{\gamma^e})\frac{\gamma^{K^e}\Gamma(\gamma^e)}{\Gamma(T^{oe}+\gamma^e)}$$

- Optimize hyperparameters using maximum a posteriori (MAP)

Estimating Likelihood of Observable Sequence

- Issues:
  - Estimating likelihood involves intractable sums over state trajectories
  - The number of distinct states grows with the sequence length
- How to estimate the likelihood effectively?

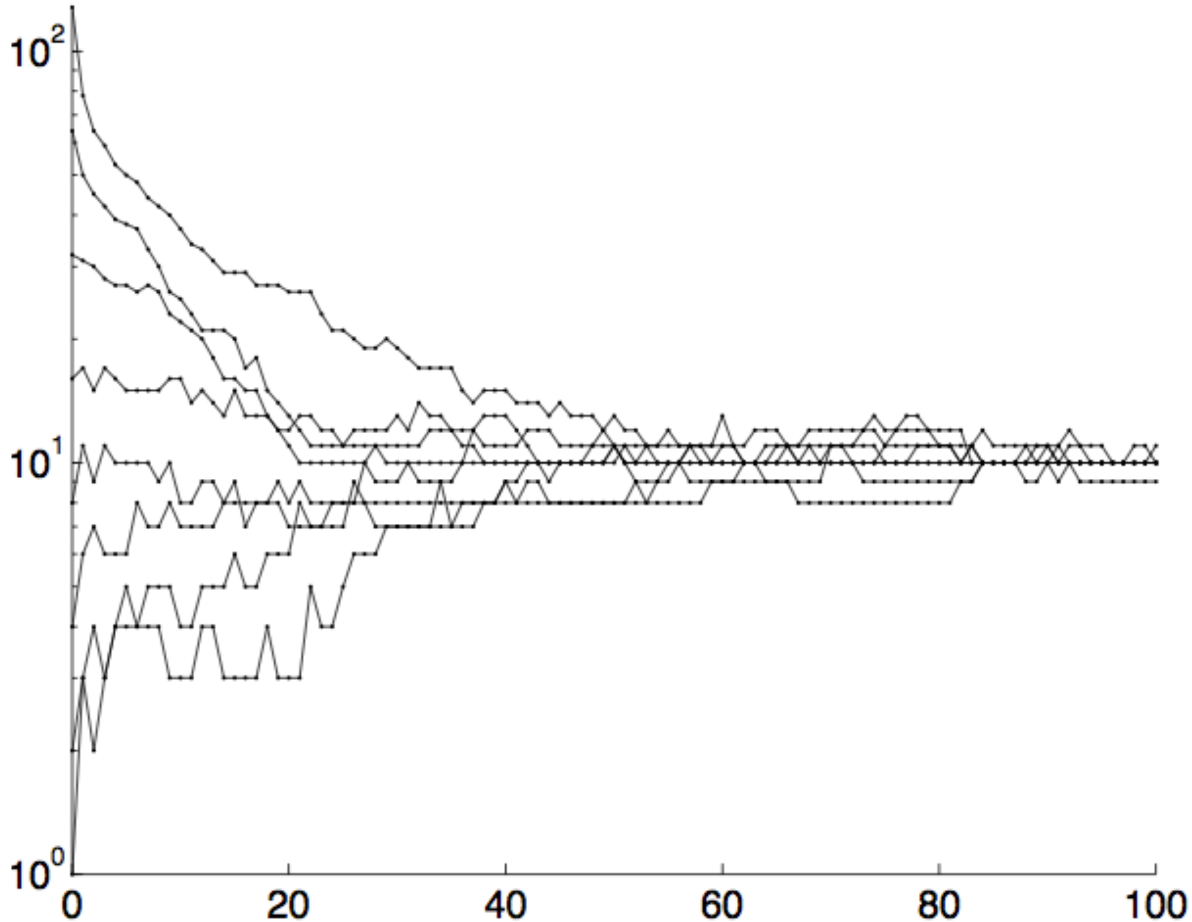
Estimating Likelihood of Observable Sequence

- Solution: Particle Filtering method

1. Compute $l_t^r = P(y_t|s_t = s_t^r)$ for each particle r.
2. Calculate $l_t = (1/R)\sum_r l_t^r \approx P(y_t|y_1, \ldots, y_{t-1})$.
3. Resample R particles $s_t^r \sim (1/\sum_{r'} l_t^{r'}) \sum_{r'} l_t^{r'} \delta(s_t, s_t^{r'})$.
4. Update transition and emission tables n^r, m^r for each particle.
5. For each r sample forward dynamics: $s_{t+1}^r \sim P(s_{t+1}|s_t = s_t^r, n^r, m^r)$; this may cause particles to land on novel states. Update n^r and m^r.
6. If $t < T$, Goto 1 with $t = t + 1$.

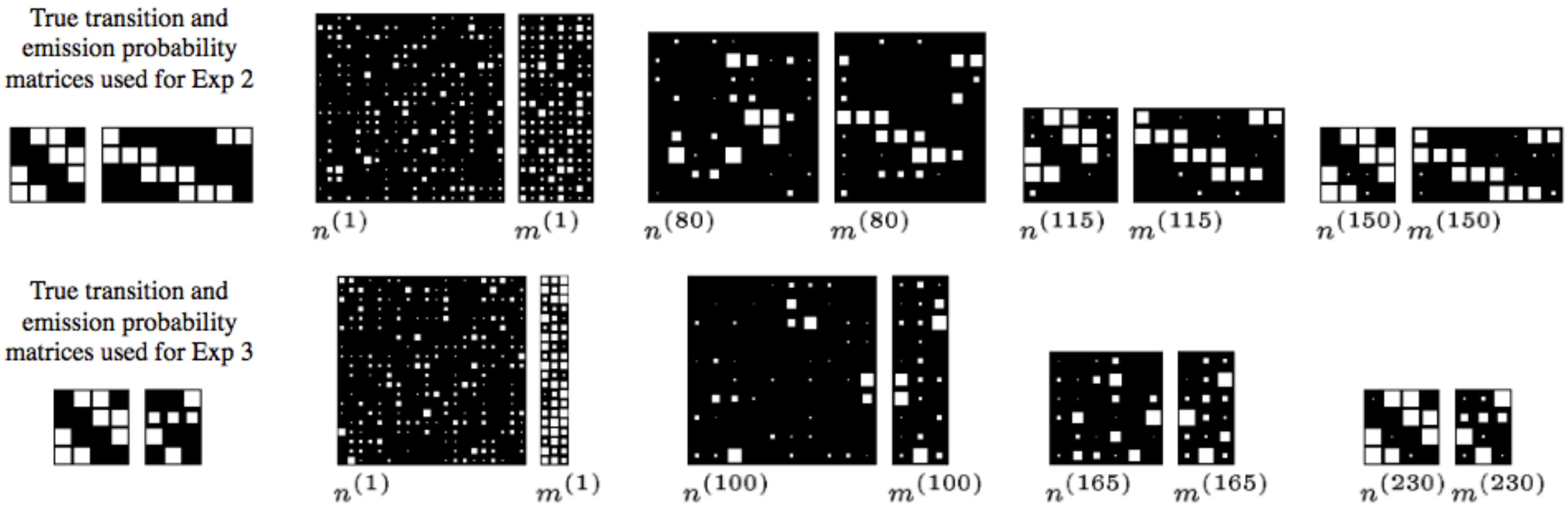
EXPERIMENTS

Synthetic experiments: Number of hidden states



Discovering the number of hidden states

Synthetic experiments: Expansive and Compressive



Expansive (top row – 4 states, 8 symbol) and **Compressive** (bottom row – 4 states, 3 symbols)

Further Reading

- Teh, Jordan, Beal and Blei (2005) (HDP paper)
  – Showed that iHMMs can be derived from hierarchical Dirichlet; processes, and provided a more efficient Gibbs sampler
- Van Gael, Saatci, Teh, and Ghahramani, 2008 (Beam Sampling paper)
  – Derived a much more efficient sampler based on Dynamic Programming

QUESTIONS?