Introduction and Course Overview

Instructor: Prof. Jason Pacheco
Outline

- Motivating examples of representation
- Efficient computation on graphical models
- Overview of course topics
- Course details (attendance, grading, etc.)
Why Graphical Models?

Structure simplifies both representation and computation

Representation
Complex global phenomena arise by simpler-to-specify local interactions

Computation
Inference / estimation depends only on subgraphs (e.g. dynamic programming, belief propagation, Gibbs sampling)
Why Graphical Models?

Structure simplifies both **representation** and **computation**

**Representation**
Complex global phenomena arise by simpler-to-specify local interactions

**Computation**
Inference / estimation depends only on subgraphs (e.g. dynamic programming, belief propagation, Gibbs sampling)

We will discuss inference later, but let’s focus on representation…
Outline

- Motivating examples of representation
- Efficient computation on graphical models
- Overview of course topics
- Course details (attendance, grading, etc.)
Problem: Given 3D protein backbone structure, estimate orientation of every side chain molecule.

Solution: Just physics of atomic interaction. Easy, right!? 
Protein Side Chain Prediction

Complex phenomena specified by simpler atomic interactions

Graphical Model

Nodes represent side chain orientations

Edges represent atomic interaction

Configuration Likelihoods

- Atomic Radius
- Side Chain Angle
Protein Side Chain Prediction

By exploiting graphical model structure we can scale computation to large macromolecules

[ Pacheco and Sudderth, ICML 2015 ]
Pose Estimation

**Graphical Model**

**Image (Data / Observation)**

**Problem:** Estimate orientation / shape / pose of human figure from an image

Model encodes likelihood of shape / pose / image consistency (e.g. skin color)

[ Pacheco, et al., NeurIPS 2014 ]
Pose Tracking

By composing single-frame model with temporal dynamics and motion prior we can do video tracking…
Kinematic Hand Tracking

Kinematic Prior

Structural Prior

Dynamic Prior

Sudderth et al., 2004
Hidden Markov Models

**Sequential models of discrete quantities of interest**

\[ \cdots \rightarrow x_{t-1} \rightarrow x_t \rightarrow x_{t+1} \rightarrow \cdots \]

**Example: Part-of-speech Tagging:**

\[ Y = \text{"I shot an elephant in my pajamas."} \]

\[ \mathcal{X} = \text{NP-V-Det-N-P-Det-N} \]

**Example: Speech Recognition**

\[ y = \text{b-ey-z-th-ih-er-em} \rightarrow \text{Bayes’ Theorem} \]

[ Source: Bishop, PRML ]

[ Source: nltk.org ]
Dynamical Models

**Sequential models of continuous quantities of interest**

Example: Nonlinear Time Series

Example: Multitarget Tracking
State-Space Models

Intracortical Brain-Computer Interface

Block 12: "Multiscale Semi-Markov Model"

[ Milstein, Pacheco, et al., NeurIPS 2017 ]
Outline

• Motivating examples of representation
• Efficient computation on graphical models
• Overview of course topics
• Course details (attendance, grading, etc.)
Why Graphical Models?

Structure simplifies both **representation** and **computation**

**Representation**
Complex global phenomena arise by simpler-to-specify local interactions

**Computation**
Inference / estimation depends only on subgraphs (e.g. dynamic programming, belief propagation, Gibbs sampling)
Computation in Graphical Models

This style of computation generalizes to all graphical models...

Example algorithms

- Belief propagation
- Gibbs sampling
- Particle filtering
- Viterbi decoder for HMMs
- Kalman filter (marginal inference)

Key Idea: Local computations only depend on the statistics of the current node and neighboring interactions.
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by passing messages forward and backward along chain.
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by passing messages forward and backward along chain.

\[ x^* = \arg \max_x p(x \mid y) \]
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by *passing messages* forward and backward along chain.

\[ x^* = \arg \max_x p(x \mid y) \]
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by passing messages forward and backward along chain.

$x^*_t+1 = \arg\max ...$

$x^* = \arg\max_{x} p(x \mid y)$
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by passing messages forward and backward along chain.

\[ x^*_t = \arg\max \ldots \]

\[ x^* = \arg\max_x \ p(x \mid y) \]
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by passing messages forward and backward along chain.

\[ x^* = \arg\max_x p(x \mid y) \]
Viterbi Decoder

Efficiently computes MAP estimate for state-space model by passing messages forward and backward along chain.

\[ x^* = \arg\max_x p(x \mid y) \]
Outline

- Motivating examples of representation
- Efficient computation on graphical models
- Overview of course topics
- Course details (attendance, grading, etc.)
We will cover **five** primary topics…

<table>
<thead>
<tr>
<th>Variational Inference</th>
<th>Advanced Markov chain Monte Carlo</th>
<th>Bayesian Nonparametrics</th>
<th>Bayesian Optimization</th>
<th>Bayesian Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficient methods for approximate posterior inference</td>
<td>Techniques for obtaining asymptotically exact inference while avoiding local optima</td>
<td>A class of probability models where model complexity is inferred from the data</td>
<td>Probabilistic methods for global optimization of smooth functions</td>
<td>Probabilistic uncertainty models for deep learning</td>
</tr>
</tbody>
</table>
Variational Inference

Uses Jensen’s inequality to bound quantities of inference

Jensen’s Inequality (for concave functions)

\[ f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)] \]

The logarithm is concave.

Variational Lower Bound

\[
\log p(y) \geq \mathbb{E}_q \left[ \log \frac{p(x, y)}{q(x)} \right]
\]

- Partition Function
- Marginal likelihood

Variational Approximation

\[ p(x) \]

\[ q(x) \]
Advanced Markov Chain Monte Carlo

Advanced MCMC techniques reduce sample complexity and avoid getting stuck in local energy minima.

Example: Parallel tempering exchange replicates across multiple MCMC chains running in (embarrassingly) parallel.
Bayesian Nonparametrics

Amount and nature of data drive model complexity

Example: Dirichlet process mixture models a distribution over an infinite number of mixture components.
Bayesian Optimization

Global optimization of random functions: \( \min_x f(x) \)

[Source: Ryan Adams]
Bayesian Optimization

Iteratively updates distribution over function value (regression)

[Source: Ryan Adams]
Bayesian Optimization

The function is well-approximated around the minimizer

[Source: Ryan Adams]
Bayesian Deep Learning

Neural networks are graphical models too…

…but they are *typically* not probabilistic
Bayesian Deep Learning

Combines deep learning with uncertainty models

[Source: Johnson et al., NIPS 2016]
Outline

• Motivating examples of representation
• Efficient computation on graphical models
• Overview of course topics
• Course details (attendance, grading, etc.)

Now for the bulleted lists of stuff...