Outline

- Graphical Models and Recursive Inference (Pat)
  - Markovianity & Factorization
  - Exponential Families: Ising & Gaussian Models
  - Illustrative Example: $3 \times 3$ Ising Model
  - “Belief” Propagation: Sum/Max-Product & Gaussian Elim.
  - Exact inference gets hard! Many approximate methods...

- Model Identification in Exponential Families (Jason)
  - Convexity & Duality in Exponential Families
  - Variational Principles for Inference & Learning
  - Information Geometry & Iterative Projection Methods
Graphical Models: Markovianity

- Graph \( G = (V, E) \) defines family of probability distributions
  - Node set \( V \) identifies random vector \( x = (x_1, \ldots, x_{|V|}) \)
  - Edge set \( E \) indicates Markov properties with "separation"

\[ p(x_A, x_B|x_S) = p(x_A|x_S)p(x_B|x_S) \]

- **Definition:** Random vector \( x \) is Markov on \( G \) if and only if, for every triplet \( A, S, B \subset V \) such that \( S \) separates \( A \) and \( B \),
Graphical Models: Factorization

- Let edge set $E$ define $p(x)$ as product of “local” functions
- But is there a notion of “local” applicable for general $G$?
  - Choose domains $C \subset V$ over maximal cliques of $G$

- For each $C$, choose potential function $\psi_C : \mathcal{X}_C \rightarrow (0, \infty)$

**Definition:** $p(x)$ factors over $G$ if, for at least one collection $\{\psi_C\}$ of (maximal) clique potentials,

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C) \quad (Z \in \mathbb{R} \text{ for normalization})$$
Graphical Models: Punchline & Asides

- **Theorem (Hammersley-Clifford):** “x Markov on G” and “p(x) factors over G” define equivalent families of distributions
  \[ \Rightarrow \text{Graph structure tied to complexity of inference/learning} \]

- Connection to *Boltzmann distribution* in statistical physics
  \[ p(x) = \frac{1}{Z} \exp (-H(x)) \quad \text{(energy} \ H(x) = -\sum_C \log [\psi_C(x_C)]) \]

- *Factor graphs* characterize more specific “local” structure
Exponential Family Models

- All distributions on $\mathcal{X}$ that can be expressed in the form

$$p(x) = \exp \left[ \theta' \phi(x) - \Psi(\theta) \right] \quad (\Psi : \mathbb{R}^d \to \mathbb{R} \text{ for normalization})$$

with parameters $\theta \in \mathbb{R}^d$ and features $\phi : \mathcal{X} \to \mathbb{R}^d$

- Ising Models: if $x_i \in \{+1, -1\}$, then $d = |V| + |E|$ and

$$p(x) \propto \exp \left[ \sum_{(i,j) \in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \right]$$

- Gaussian Models: if $x \sim N(J^{-1}h, J^{-1})$, let $\theta = (h, J)$ so

$$p(x) \propto \exp \left[ -\frac{1}{2} x'Jx + h'x \right]$$

with matrix $J$ sparse in correspondence with edge set $E$
Illustrative Example: $3 \times 3$ Ising Model

Graphical Model

$x_i \in \{-1, +1\}$

Samples

$p(x) = 3.016 \times 10^{-1}$

$p(x) = 1.885 \times 10^{-2}$

$p(x) = 4.713 \times 10^{-3}$

$p(x) = 1.178 \times 10^{-3}$

$p(x) = 2.946 \times 10^{-4}$

$p(x) = 7.364 \times 10^{-5}$

$p(x) = 4.603 \times 10^{-6}$

$p(x) = 1.798 \times 10^{-8}$

$\theta_i = 0, \; i \in V$

“uniform”

$\theta_{ij} = 0.7, \; (i, j) \in E$

“attractive”
Illustrative Example: 3 × 3 Ising Model

Graphical Model

\[ x_i \in \{-1, +1\} \]

```
\begin{array}{ccc}
  x_1 & x_2 & x_3 \\
  x_4 & x_5 & x_6 \\
  x_7 & x_8 & x_9 \\
\end{array}
```

\[ \theta_i = 0, \quad i \in V \]

"uniform"

\[ \theta_{ij} = 0.7, \quad (i, j) \in E \]

"attractive"

Inference

\[ p(x_i | y) = 0.603 \]

\[ p(x_i | y) = 0.181 \]

\[ p(x_i | y) = 0.0855 \]

\[ p(x_i | y) = 0.262 \]

\[ p(x_i | y) = 0.0113 \]
Inference Problems & Variable Elimination

- **Marginalization**: compute \( p(x_A) = \sum_{x_{V \setminus A}} p(x) \)
  - Elimination of nodes \( V \setminus A \) by summation/integration
  - Basic operation to compute conditionals and likelihoods

- **Example**: let \( p(x) \propto \psi_{12}\psi_{13}\psi_{24}\psi_{35}\psi_{256} \) with \( |\mathcal{X}_i| = r \) for \( i \in V \)
  - Direct computation of \( p(x_1) = \sum_{x_2,\ldots,x_6} p(x) \) scales as \( r^6 \)
  - Exploiting factorization in computation of \( p(x_1) \) scales as \( r^3 \)

\[
p(x_1) = \frac{1}{Z} \sum_{x_2} \psi_{12} \sum_{x_3} \psi_{13} \sum_{x_4} \psi_{24} \sum_{x_5} \psi_{35} \sum_{x_6} \psi_{256}
\]

- **Max-Marginalization**: compute \( \nu(x_A) = \max_{x_{V \setminus A}} p(x) \)
  - Elimination of nodes \( V \setminus A \) by maximization
  - Basic operation to compute a mode of \( p(x) \) (with caveat!)
Recursive Inference: “Message-Passing”

- Discrete-variable chain with $|V| = 4$ \( \Rightarrow \) $p(x) \propto \psi_{12}\psi_{23}\psi_{34}$

\[
p(x_1) = \frac{1}{Z} \sum_{x_2} \psi_{12}(x_1, x_2) \sum_{x_3} \psi_{23}(x_2, x_3) \sum_{x_4} \psi_{34}(x_3, x_4)
\]

- Key idea: apply most efficient elimination ordering
  - Marginalization at all nodes share intermediate terms $m_i$
  - “Message” interpretation useful for distributed settings
“Belief” Propagation on Trees

- Markov tree: \[ p(x) \propto \prod_{i \in V} \psi(x_i) \prod_{(i,j) \in E} \psi(x_i, x_j) \]

- Sum-Product algorithm efficiently finds all marginals \( p(x_i) \)

\[
m_{j \rightarrow i}(x_i) = \sum_{x_j} \psi(x_i, x_j) \left( \psi(x_j) \prod_{k \in N(j) \setminus i} m_{k \rightarrow j}(x_j) \right)
\]

\[
p(x_v) \propto \psi(x_v) \prod_{i \in N(v)} m_{i \rightarrow v}(x_v)
\]

- Max-Product algorithm efficiently finds all max-marginals \( \nu(x_i) \)
Gaussian Elimination (GE) is a form of BP!

- Consider solution of $Jx = h$ by *Gaussian elimination*. Partition $V = A \cup B$ and eliminate variables $B$ from equations $A$ we obtain $\hat{J}_A x_A = \hat{h}_A$ where:

\[
\hat{J}_A = J_A - J_{A,B} J_B^{-1} J_{B,A} \\
\hat{h}_A = h_A - J_{A,B} J_B^{-1} h_B
\]

This is the *Schur complement* form of Gaussian elimination.

- Let $K(x; h, J) = \exp\{-\frac{1}{2} x' J x + h' x\}$. Then,

  1. *Integration:* $\int_{x_B} K(x_A, x_B; h, J) dx_B \propto K(x_A; \hat{h}_A, \hat{J}_A)$
  2. *Maximization:* $\max_{x_B} K(x_A, x_B; h, J) = K(x_A; \hat{h}_A, \hat{J}_A)$

Consequently, Gaussian BP involves identical steps as in GE.

- The *Kalman filter* is also a form of BP on a Gauss-Markov chain but is based on a directed (causal) factorization.
Inference on Graphs with Cycles

- Still variable elimination...but complicated by “entanglement”

- Junction Tree algorithm performs exact computation
  - Key idea: aggregate nodes to equivalent tree
  - Tractable if aggregates are low-order (i.e., low “treewidth”)
“Loopy Belief” Propagation

- Iterate BP equations at each node, ignorant of cycles

\[
\begin{align*}
    m_{j\rightarrow i}^{(t+1)}(x_i) &= \sum_{x_j} \psi(x_i, x_j) \left( \psi(x_j) \prod_{k \in N(j) \setminus i} m_{k\rightarrow j}^{(t)}(x_j) \right) \\
    N(j) \setminus i
\end{align*}
\]

- Need not converge: approximation if it does converge
  - Connection to coding: LDPC codes and “turbo codes”
  - Connection to physics: minimizing Bethe free energy
More about Exponential Families\textsuperscript{a}...

- The \textit{cumulant-generating function} plays a central role:

\[ \Psi(\theta) = \log \int \exp\{\theta \cdot \phi(x)\} dx \]

e.g., \( \Psi(\theta) = -\frac{1}{2} \log \det J(\theta) + \text{const} \) (Gaussian).

- Moment-generating property:

\[ \nabla \Psi(\theta) = \mathbb{E}_\theta \{\phi(x)\} \equiv \eta(\theta) \]

where \( \eta \) are the \textit{moments} \( \equiv \) marginal probabilities (discrete), means, variances and edge-covariances (Gaussian).

- The curvature of \( \Psi(\theta) \) is the \textit{Fisher information matrix}:

\[ \nabla^2 \Psi(\theta) = \mathbb{E}_\theta \{(\phi(x) - \eta(\theta))'(\phi(x) - \eta(\theta))\} \]

This is a spd covariance matrix, hence \( \Psi(\theta) \) is convex.

\textsuperscript{a}Barndorff-Nielsen '78.
Variational Principles

**Fenchel duality** [Fenchel ’49; Rockafellar ’74] The *convex conjugate* of Ψ equals the *negative entropy* as a function of the moments.

\[ \Psi^*(\eta) \equiv \sup_{\theta} \{ \eta \cdot \theta - \Psi(\theta) \} = -h(\eta) \]

Due to convexity of Ψ it holds that \((\Psi^*)^* = \Psi\).

**Learning** Given a desired set of moments \(\eta^*\) the corresponding parameters \(\theta^*\) minimize the convex function:

\[ f(\theta) = \Psi(\theta) - \eta^* \cdot \theta \]

In ML parameter estimation, \(\eta^*\) are the empirical moments.

**Inference** Given \(\theta^*\) the corresponding moments \(\eta^*\) minimize the convex function:

\[ g(\eta) = \Psi^*(\eta) - \theta^* \cdot \eta \]

Leads to approximate inference [Wainwright & Jordan ’03].
Information Geometry\textsuperscript{a}

- The \textit{Bregman distance}\textsuperscript{b} induced by $\Psi(\theta)$ equals the \textit{Kullback-Leibler divergence}.

$$D(\theta^*||\theta) = \Psi(\theta) - \{\Psi(\theta^*) + \nabla\Psi(\theta^*) \cdot (\theta - \theta^*)\}$$

Similar relation holds between $\Psi^*(\eta)$ and $D(\eta||\eta^*)$.

- \textit{Information Projection}: let $p \in \mathcal{F}$ and let $\mathcal{E} \subset \mathcal{F}$ is affine in $\theta$.

$$p_{\mathcal{E}} \equiv \arg\min_{q \in \mathcal{E}} D(p\|q)$$

Optimality condition: $(\eta(q) - \eta(p)) \perp (\theta(\mathcal{E}) - \theta(p_{\mathcal{E}}))$.

- \textit{Pythagorean Relation}: $p_{\mathcal{E}}$ is unique member of $\mathcal{E}$ satisfying

$$D(p\|q) = D(p\|p_{\mathcal{E}}) + D(p_{\mathcal{E}}\|q)$$

for all $q \in \mathcal{E}$.

\textsuperscript{a}Chentsov ’72; Efron ’78; Amari ’01.

\textsuperscript{b}Bregman ’67; Bauschke & Bowein ’97.
\[ D(p\|q) = D(p\|p\varepsilon) + D(p\varepsilon\|q) \]

\[ D(p\|p\varepsilon) = h(p\varepsilon) - h(p) \]
IPF as Projection onto Convex Sets

Iterate over cliques \( \{C_k\} \) of graph \( G \), update potentials to enforce marginal constraints...

- **Iterative Proportional Fitting:**\(^a\) marginal pmfs \( p(x_{C_k}) \)

\[
q^{(k+1)}(x) = q^{(k)}(x) \times \frac{p(x_{C_k})}{q^{(k)}(x_{C_k})}
\]

- **Covariance Selection:**\(^b\) marginal covariances \( P_{C_k} \)

\[
J^{(k+1)}_{C_k} = J^{(k)}_{C_k} + (P^{-1}_{C_k} - (P^{(k)}_{C_k})^{-1})
\]

- **Projection Interpretation:**\(^c\) \( M_k \subset \mathcal{F} \) affine in \( \eta \) imposes marginal moment constraints on clique \( C_k \).

\[
q^{(k+1)} = \arg\min_{p \in M_k} D(p\|q^{(k)})
\]

\(^a\)Kullback ’68.
\(^b\)Dempster ’77; Speed & Kiiveri ’86.
\(^c\)Csiszar ’75.
Expectation-Maximization as Alternating Projections

- Let $\mathcal{F} = \{p_\theta(x, y)\}$ be an exponential family, given observations $y_1, \ldots, y_n$, select $\theta$ to maximize the (marginal) log-likelihood:

$$f(\theta) \equiv \sum_i \log \int p_\theta(x, y_i) dx$$

Typically non-convex, possibly many local minima!

- Expectation-Maximization\(^a\) (Alternating Projections): Let $q^{(0)} \in \mathcal{F}$ and $\mathcal{D} = \{p(x, y) | \int p(x, y) dy = p^*(y)\}$

1. (E-step) $p^{(k+1)} = \arg \min_{p \in \mathcal{D}} \mathcal{D}(p \| q^{(k)})$ (inference)
2. (M-step) $q^{(k+1)} = \arg \min_{q \in \mathcal{F}} \mathcal{D}(p^{(k+1)} \| q)$ (IPF)

$\Rightarrow$ local minima of $f(\theta)$.

\(^a\)Dempster, Laird & Rubin ’77.
$$\mathcal{F}$$

$$D$$

$$q^{(0)}$$

$$q^{(1)}$$

$$q^*$$

$$p^{(0)}$$

$$p^{(1)}$$

$$p^*$$

$$f(\theta)$$
Summary: Exponential Family Graphical Models

- Graphical models combine graph theory and probability theory
- Exponential family representation links to convex analysis
- Lead to principled approximations for large-scale problems
  - Inference: compute marginals/modes of a given $p(x)$
  - Learning: design parameterized $p(x)$ given sample data
- Active research topics
  - Approximate Inference
  - Model Selection