Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 9: Parameter Learning in Fully-Observed UGMs

February 25, 2025

Reading: See course homepage

Today

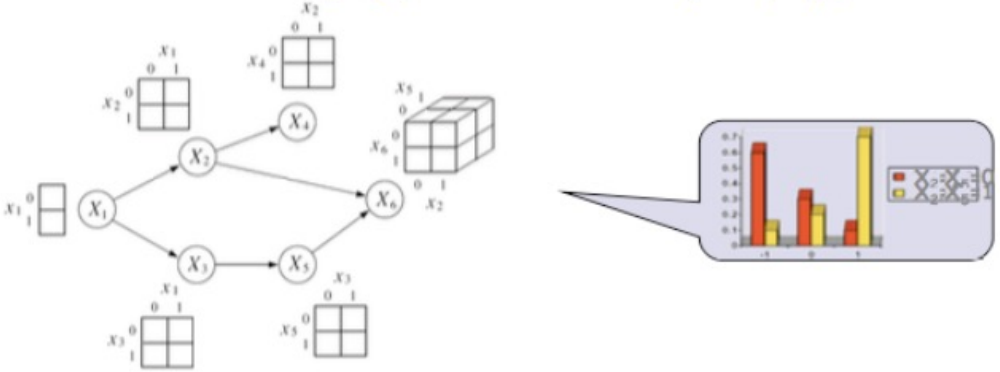
- Parameter Learning in Undirected Graphical Models
  - Iterative Proportional Fitting
  - Generalized Iterative Scaling

Parameter Learning in Fully-Observed Undirected Graphical Models

Recall: MLE for BNs

- If we assume the parameters for each CPD are globally independent, and all nodes are fully observed, then the log-likelihood function decomposes into a sum of local terms, one per node

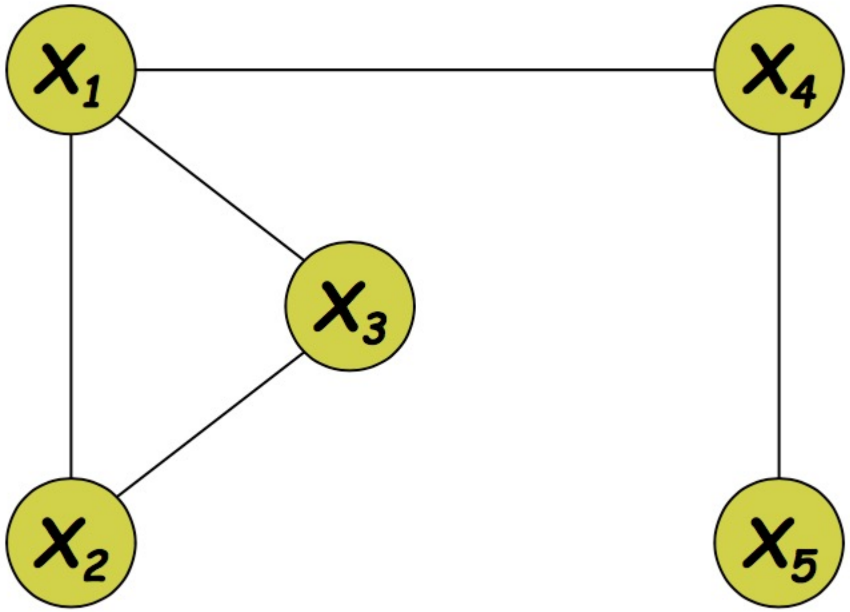
$$\ell(\theta; D) = \log p(D \mid \theta) = \log \prod_n \left(\prod_i p(x_{n,i} \mid \mathbf{x}_{n,\pi_i}, \theta_i) \right) = \sum_i \left(\sum_n \log p(x_{n,i} \mid \mathbf{x}_{n,\pi_i}, \theta_i) \right)$$



$$\theta_{ijk}^{ML} = \frac{n_{ijk}}{\sum_{i,j',k} n_{ij'k}}$$

- MLE-based parameter estimation of GM reduces to local est. of each GLIM.

What about for Undirected GMs?



Main challenge: Clique potentials are not probabilities, so MLE may not decompose into estimates for individual parameters.

MLE for Undirected GMs

- For **directed** models, the log-likelihood decomposes into a sum of terms, one per family (node plus parents).
- For undirected models, the log-likelihood does not decompose, because the normalization constant Z is a function of **all** parameters.

$$P(x_1,\ldots,x_n) = \frac{1}{Z}\prod_{c\in C}\psi_c(\mathbf{x}_c) \qquad Z = \sum_{x_1,\ldots,x_n}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

- In general, we need to do inference to learn parameters for undirected models, even in the fully observed case.

Likelihood for UGMs with tabular clique potentials

- Sufficient statistics: Summarize the number of times that a configuration x is observed in a dataset D as:

$$m(\mathbf{x}) \stackrel{\text{def}}{=} \sum_n \delta(\mathbf{x}, \mathbf{x}_n) \quad \text{(total count)}, \qquad \text{and} \qquad m(\mathbf{x}_c) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_{V \setminus c}} m(\mathbf{x}) \quad \text{(clique count)}$$

Number of times configuration x is seen in dataset

Number of times clique configuration x_c is seen in dataset

Likelihood for UGMs with tabular clique potentials

- Sufficient statistics: Summarize the number of times that a configuration x is observed in a dataset D as:

$$m(\mathbf{x}) \stackrel{\text{def}}{=} \sum_n \delta(\mathbf{x}, \mathbf{x}_n) \quad \text{(total count)}, \qquad \text{and} \qquad m(\mathbf{x}_c) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_{V \setminus c}} m(\mathbf{x}) \quad \text{(clique count)}$$

Number of times configuration x is seen in dataset

Number of times clique configuration x_c is seen in dataset

- The log-likelihood is then:

$$\log p(D|\theta) = \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z$$

$$p(D|\theta) = \prod_n \prod_{\mathbf{x}} p(\mathbf{x} \mid \theta)^{\delta(\mathbf{x}, \mathbf{x}_n)}$$

$$\log p(D|\theta) = \sum_n \sum_{\mathbf{x}} \delta(\mathbf{x}, \mathbf{x}_n) \log p(\mathbf{x} \mid \theta) = \sum_{\mathbf{x}} \sum_n \delta(\mathbf{x}, \mathbf{x}_n) \log p(\mathbf{x} \mid \theta)$$

$$\ell = \sum_{\mathbf{x}} m(\mathbf{x}) \log\left(\frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c)\right)$$

$$= \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z$$

Derivative of Log-likelihood

- Log-likelihood $\log p(D|\theta) = \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z$

- First term: $\frac{\partial \ell_1}{\partial \psi_c(\mathbf{x}_c)} = m(\mathbf{x}_c) / \psi_c(\mathbf{x}_c)$

- Second term:

$$\frac{\partial \log Z}{\partial \psi_c(\mathbf{x}_c)} = \frac{1}{Z} \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\sum_{\tilde{\mathbf{x}}} \prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$$

$$= \frac{1}{Z} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$$

$$= \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{1}{\psi_c(\tilde{\mathbf{x}}_c)} \frac{1}{Z} \prod_d \psi_d(\tilde{\mathbf{x}}_d)$$

$$= \frac{1}{\psi_c(\mathbf{x}_c)} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) p(\tilde{\mathbf{x}}) = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

Derivative of Log-likelihood

- Putting it together

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N\frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

- Set equal to zero:

$$p^*_{MLE}(\mathbf{x}_c) = \frac{m(\mathbf{x}_c)}{N} \stackrel{\text{def}}{=} \tilde{p}(\mathbf{x}_c)$$

- But the UGM is parameterized by ψ_c not p.

Case 1: The model is decomposable

- If the model is **decomposable** and **all the clique potentials are defined on maximal cliques**, then:
  - The MLE of clique potentials are equal to the empirical marginals (or conditionals) of the corresponding clique.

- Example: Chain $X_1 - X_2 - X_3$

$$p_{MLE}(X1, X2, X3) = \frac{\tilde{p}(X1, X2)\tilde{p}(X2, X3)}{\tilde{p}(X2)}$$

$$p_{MLE}(X1, X2) = \sum_{X3} \tilde{p}(X1, X2, X3) = \tilde{p}(X1|X2) \sum_{X3} \tilde{p}(X2, X3) = \tilde{p}(X1, X2)$$

$$p_{MLE}(X2, X3) = \tilde{p}(X2, X3)$$

$$\tilde{\psi}_{12}^{MLE}(x_1, x_2) = \tilde{p}(x_1, x_2) \qquad \tilde{\psi}_{23}^{MLE}(x_2, x_3) = \frac{\tilde{p}(x_2, x_3)}{\tilde{p}(x_2)} = \tilde{p}(x_3|x_2)$$

Case 2: The model is NON-decomposable

- If the model is **non-decomposable** (clique potentials are defined on non-maximal cliques), then we cannot equate MLE of clique potentials to empirical marginals (or conditionals).

- Two iterative algorithms:
  - Iterative Potential Fitting
  - Generalized Iterative Scaling

Iterative Proportional Fitting (IPF)

- From the log-likelihood: $\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N\frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$

- Let's rewrite in a different way: $\frac{m(\mathbf{x}_c)}{N\psi_c(\mathbf{x}_c)} = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$ or $\frac{\tilde{p}(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$
  - The clique potentials implicitly appear in the model marginal $p(\mathbf{x}_c) = f(\psi_c(\mathbf{x}_c))$

- Let's forget a closed form solution and focus on a **fixed-point iteration** method

$$\frac{\tilde{p}(\mathbf{x}_c)}{\psi_c^{(t+1)}(\mathbf{x}_c)} = \frac{p(\mathbf{x}_c)}{\psi_c^{(t)}(\mathbf{x}_c)} \quad \Rightarrow \quad \psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c)\frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$$

- Need to run inference for $p^{(t)}(\mathbf{x}_c)$

Properties of IPF Updates

- Set of fixed-point equations:

$$\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c)\frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$$

- We can show that it is also a coordinate ascent algorithm (coordinates=parameters of clique potentials)

- At each step, it will increase the log-likelihood, and it will converge to a global maximum.

$$\max \ell \Leftrightarrow \min KL\big(\tilde{p}(x) \,\|\, p(x \mid \theta)\big) = \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x \mid \theta)}$$

- Maximizing the log likelihood is equivalent to minimizing the KL divergence (cross entropy)
- The max-entropy principle to parameterization offers a dual perspective to the MLE.

$$\begin{aligned} \min_p \quad & \mathrm{KL}\big(p(x) \,\|\, h(x)\big) \\ & \stackrel{\text{def}}{=} \sum_x p(x) \log \frac{p(x)}{h(x)} = -\mathrm{H}(p) - \sum_x p(x) \log h(x) \\ \text{s.t.} \quad & \sum_x p(x) f_i(x) = \alpha_i \\ & \sum_x p(x) = 1 \end{aligned}$$

So far...

- **Decomposable graphs:** MLE for clique potentials correspond to empirical marginals or conditionals
- **Non-decomposable graphs:**
  - If clique potentials are parameterized as full tables:
    - Iterative Proportional Fitting

$$\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c)\frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$$

  - Cost of clique potentials as full tables is **exponential** in the number of variables in the clique.

  Can we represent UGMs more compactly and still estimate parameters?

Feature-parameterized clique potentials

Recall Parameter Sharing in BNs…

Parameters:

- $P(Z_i \mid Z_{i-1})$
- $P(X_i \mid Z_i)$
- $P(Z_1)$

→ $d(|Z|^2 + |X||Z|) + |Z|$

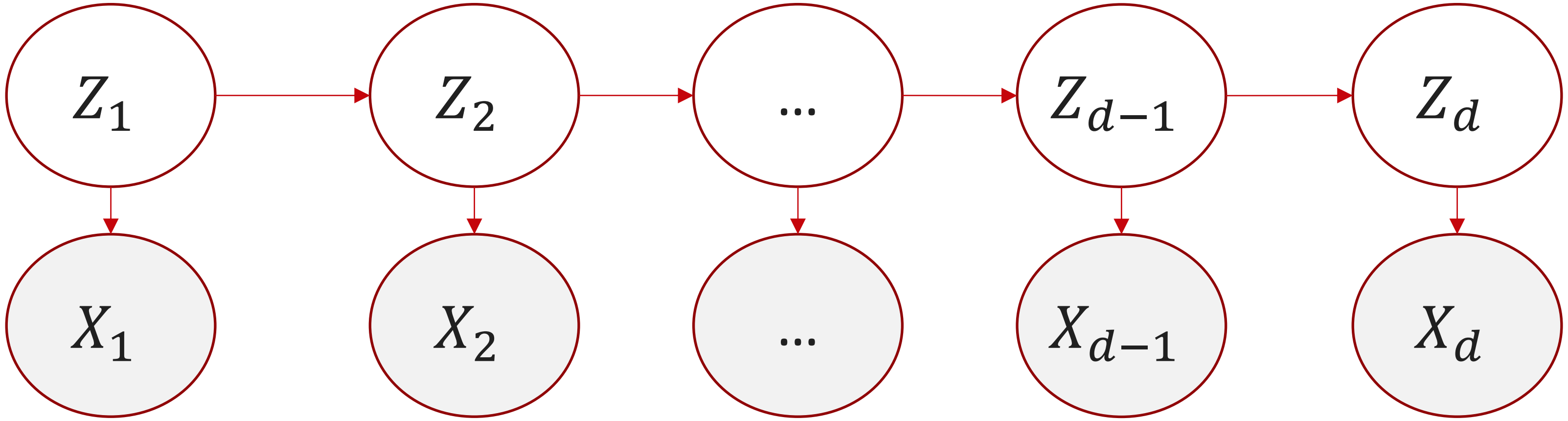
HMM Definition:

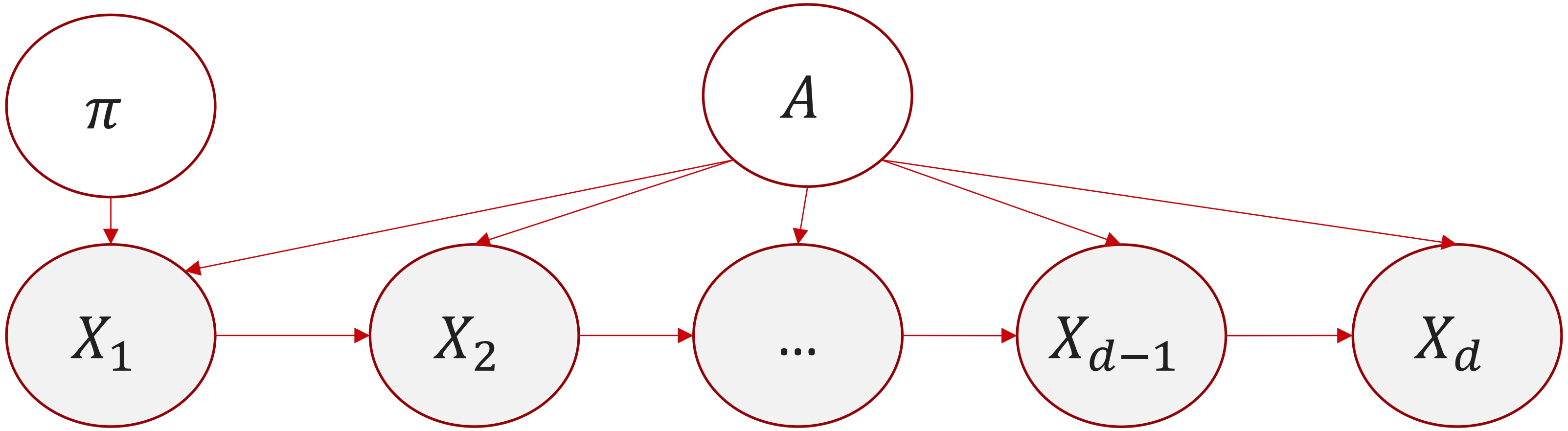
- $P(Z_i = k \mid Z_{i-1} = j) = T_{jk}$
- $P(X_i = k \mid Z_i = j) = E_{jk}$
- $P(Z_1)$

→ $|Z|^2 + |X||Z| + |Z|$

Alternate Definition:

→ $|X| + |X|^2$





Features

- A "feature" is a function that is non-zero for a few particular inputs and zero otherwise.
- Key idea: Instead of modeling all possible feature values in a big table, model specific groupings of feature values together.
- Example:
  - Let a clique correspond to three consecutive characters.
  - How would we define p(c1, c2, c3)?
    - All possible character combinations we need $26^3 - 1$ parameters.
    - But there are sequences that are unlikely: kfd
    - Define a feature like "ing": 1 if c1=i,c2=n,c3=g. 0 otherwise.

Features as Potentials

- Each feature function can be converted to a potential by exponentiating it. We can multiply these together to get a clique potential.
- Example:

$$\psi_c(c_1, c_2, c_3) = e^{\theta_{\text{ing}} f_{\text{ing}}} \times e^{\theta_{\text{?ed}} f_{\text{?ed}}} \times \ldots$$

$$= \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(c_1, c_2, c_3) \right\}$$

- There is still an exponential number of settings, but only K parameters (θ_k)
- A nice benefit of undirected graphical models: we don't have to normalize each feature.

Combining Features

- Each feature function has a weight θ_k which represents the numerical strength of the feature and whether it increases or decreases the probability of a clique.
- The marginal over the clique is a generalized exponential family distribution (a GLM):

$$p(c_1, c_2, c_3) \propto \exp\left\{ \begin{matrix} \theta_{\text{ing}} f_{\text{ing}}(c_1, c_2, c_3) + \theta_{\text{?ed}} f_{\text{?ed}}(c_1, c_2, c_3) + \\ \theta_{\text{qu?}} f_{\text{qu?}}(c_1, c_2, c_3) + \theta_{\text{zzz}} f_{\text{zzz}}(c_1, c_2, c_3) + \cdots \end{matrix} \right\}$$

- The features may be overlapping across cliques

$$\psi_c(\mathbf{x}_c) \stackrel{\text{def}}{=} \exp\left\{ \sum_{i \in I_c} \theta_k f_k(\mathbf{x}_{c_i}) \right\}$$

Feature-based model

- Joint distribution: $p(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_c \psi_c(\mathbf{x}_c) = \frac{1}{Z(\theta)} \exp\left\{ \sum_c \sum_{i \in I_c} \theta_k f_k(\mathbf{x}_{c_i}) \right\}$

- We can drop sum over c: $p(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}_{c_i}) \right\}$

- What are the sufficient statistics for this model?
- The features

- We need to learn weighting parameters θ_k

MLE of Feature-based model

$$\ell(\theta; D) \propto \frac{1}{N} \sum_n \log p(x_n \mid \theta)$$

$$= \sum_x \tilde{p}(x) \log p(x \mid \theta)$$

$$= \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \log Z(\theta)$$

- Problem: Z is a function of the parameters.
- Solution: Let's maximize a lower-bound of the log-likelihood.

$$\ell(\theta; D) \geq \tilde{\ell}(\theta; D) = \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^t)} - \log Z(\theta^t) + 1$$

MLE of Feature-based model: GIS Update

- A bit more math gets us to the **Generalized Iterative Scaling (GIS)** update rule:

$$\theta_i^{t+1} = \theta_i^t + \log \frac{E_{data}[f_i(x)]}{E_{p(x;\theta^t)}[f_{i(x)}]}$$

Empirical expectation

Expectation under current model distribution

Summary

- **Iterative Proportional Fitting (IPF)**:

$$\psi_c^{t+1}(x_c) = \psi_c^t(x_c)\frac{\tilde{p}(x_c)}{p^t(x_c)}$$

Empirical distribution

Current model distribution

- **Generalized Iterative Scaling (GIS)**:

$$\theta_i^{t+1} = \theta_i^t + \log\frac{E_{data}[f_i(x)]}{E_{p(x;\theta^t)}[f_{i(x)}]}$$

Empirical expectation

Expectation under current model distribution

Summary

Why don't we just do gradient descent for UGMs?

The partition function!

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

$$Z = \sum_{x_1,\ldots,x_n} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON