An Introduction to Graphical Models and Variational Inference
Machine Learning Reading Group

Niall Twomey & Rafael Poyiadzi

University of Bristol

24th of November
The sinus example

Economy of Representation
Instead of $2^5 - 1$ parameters, we only need 10.
Undirected Graphical Models

- Global Markov Independence $A \perp B \mid C$
  - Independence based on separation

- Local Markov Independence $X \perp \text{TheRest} \mid ABCD$
  - ABCD Markov blanket

A distribution $P$ factorizes over $G$ if there exist
- subset of variables $D_1 \subseteq \mathcal{X}$, ..., $D_m \subseteq \mathcal{X}$ ($D_i$ are maximal in $G$)
- non-negative potentials (factors/functions) $\Psi_1(D_1), \ldots$
- such that

$$P(X_1, X_2, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \Psi_i(D_i)$$

where

$$Z = \sum_{x_1, x_2, \ldots, x_n} \prod_{i=1}^{m} \Psi_i(D_i)$$

Also known as Gibbs distributions, Markov random fields, undirected graphical models
All factors over single variables or pairs of variables

- Node potentials $\Psi_i(x_i)$
- Edge potentials $\Psi_{ij}(x_i, x_j)$

Factorization

$$P(X) = \prod_{i \in V} \Psi_i(x_i) \prod_{(i,j) \in E} \Psi_{ij}(x_i, x_j)$$
Approaches to Inference: Diagram

- Inference
  - Exact
    - VE
    - JT
    - BP
  - Approximate
    - Stochastic
      - Gibbs, M-H
      - (MC)
    - Deterministic
      - Cluster
      - ~MP
        - LBP
        - EP
    - Variational
Variable Elimination on Chains

\[ P(E) \]

\[ = \sum_{d} \sum_{c} \sum_{b} P(c|b)P(d|c)P(E|d) \left( \sum_{a} P(a)P(b|a) \right) \]

\[ = \sum_{d} \sum_{c} \sum_{b} P(c|b)P(d|c)P(E|d)P(b) \]

Equivalent to matrix-vector multiplication, \(|\text{Val}(A)| \times |\text{Val}(B)|\)

Computational Complexity for a chain of length \(k\)

- Each step costs \(O(|\text{Val}(X_i)| \times |\text{Val}(X_{i+1})|)\) operations: \(O(kn^2)\)
- Compare to naïve summation: \(O(n^k)\)
Another Example

- Query: $P(A|h)$, need to renormalize over $A$

- Initial factors
  - $P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)P(h|e,f)$
  - $\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)P(g|e)m_h(e,f)$
  - $\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)P(f|a)m_h(e,f)$
  - $\Rightarrow P(a)P(b)P(c|b)P(d|a)P(e|c,d)m_f(a,e)$
  - $\Rightarrow P(a)P(b)P(c|b)P(d|a)m_e(a,c,d)$
  - $\Rightarrow P(a)P(b)P(c|b)m_d(a,c)$
  - $\Rightarrow P(a)P(b)m_c(a,b)$
  - $\Rightarrow P(a)m_b(a)$

- Step 3: renormalize
  - $P(a,h) = P(a)m_b(a)$, compute $P(h) = \sum_a P(a)m_b(a)$
  - $\Rightarrow P(a|h) = \frac{P(a)m_b(a)}{\sum_a P(a)m_b(A)}$
Junction Tree Algorithm

- Moralize the graph
- Triangulate the graph
- Obtain clique tree
- Obtain junction tree
- Run local message passing on clique level instead
Variational Bounds

\[ \log(x) = \min_{\lambda} \{ \lambda x - \log \lambda - 1 \} \]
\[ g(x) = \frac{1}{1 + e^{-x}} = \min_{\lambda} \{ e^{\lambda x - H(\lambda)} \} \]
\[ \leq \lambda x - \log \lambda - 1 \]
\[ \leq e^{\lambda x - H(\lambda)} \]
Convex Duality

• for concave $f(x)$:

$$f(x) = \min_\lambda \{ \lambda^T x - f^*(\lambda) \}$$

$$f^*(\lambda) = \min_x \{ \lambda^T x - f(x) \}$$

• yields bounds:

$$f(x) \leq \lambda^T x - f^*(\lambda)$$

$$f^*(\lambda) \leq \lambda^T x - f(x)$$
Variational calculus: Brachristochrone
Variational Inference

- Loops are causing the problem for exact inference
- Junction tree can still results in large cliques (eg. Grid)
- Variational inference: approximate distribution with loopy graphs using distribution with tree or simpler topologies

Kullback-Leibler (KL) divergence is often used

\[ KL(P||Q) = \sum_x P(X) \ln \frac{P(X)}{Q(X)} \]

\[ KL(P||Q) = \int_{-\infty}^{\infty} P(X) \ln \frac{P(X)}{Q(X)} dX \]
How complex is $p$ for GMMs?

Posterior is intractable for large $n$, and we might want to add priors

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}$$
Main idea

- We create a **variational distribution** over the latent variables $q(z_1:m | \nu)$
- Find the settings of $\nu$ so that $q$ is close to the posterior
Evidence Lower Bound (ELBO)

- Apply Jensen’s inequality on log probability of data

\[
\log p(x) = \log \left[ \int_z p(x, z) \right] \\
= \log \left[ \int_z p(x, z) \frac{q(z)}{q(z)} \right] \\
= \log \left[ \mathbb{E}_q \left[ \frac{p(x, z)}{q(z)} \right] \right] \\
\geq \mathbb{E}_q \left[ \log p(x, z) \right] - \mathbb{E}_q \left[ \log q(z) \right]
\]

- Fun side effect: Entropy
- Maximizing the ELBO gives as tight a bound on log probability
Jensen’s inequality

$$\log P(V) = \log \Sigma_H P(H, V)$$

$$= \log \Sigma_H Q(H|V) \cdot \frac{P(H, V)}{Q(H|V)}$$

$$\geq \Sigma_H Q(H|V) \log \left[ \frac{P(H, V)}{Q(H|V)} \right]$$

$$KL(Q||P) = \Sigma_H Q(H|V) \log \left[ \frac{Q(H|V)}{P(H|V)} \right]$$
Kullback-Leibler Divergence

- $\text{argmin}_Q \ KL(P||Q)$
  - True distribution $P$ defines support of difference
  - The correct direction, but intractable to compute

- $\text{argmin}_P \ KL(Q||P)$
  - Approximate distribution defines support
  - Tends to give overconfident results, but tractable

```
argmin_Q D(P||Q)  argmin_Q D(Q||P)  argmin_Q D(Q||P)
```
Minimising $\text{KL}(q || p) \leftrightarrow$ Maximising ELBO

- Plug into KL divergence

$$\text{KL}(q(z)||p(z|x)) = \mathbb{E}_q \left[ \log \frac{q(z)}{p(z|x)} \right]$$

$$= \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z|x)]$$

$$= \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z,x)] + \log p(x)$$

$$= - (\mathbb{E}_q [\log p(z,x)] - \mathbb{E}_q [\log q(z)]) + \log p(x)$$

- Negative of ELBO (plus constant); minimizing KL divergence is the same as maximizing ELBO
The equations

\[ \max_Q F(P, Q) = \max_{Q_i} \sum_{D_i} E_Q[\ln \Psi(D_i)] + H(Q) \]

subject to \( \forall i, \sum_{x_i} Q_i(x_i) = 1, \text{ and } Q_i(x_i) \geq 0 \)

Constrained optimization
- Add \( \lambda \), form Lagrangian multipliers
- Take derivative, set to zeros

\( Q \) is a stationary point of mean field approximation iff \( \forall i \):

\[
Q_i(x_i) = \frac{1}{Z_i} \exp \left( \sum_{D_j : X_i \in D_j} E_Q[\ln \Psi(D_j)] \right)
\]
The Algorithm

- $Q_i$ only needs to consider factors that intersect $X_i$
  
  \[ Q_i(X_i) = \frac{1}{Z_i} \exp \left( \sum_{D_j: X_i \in D_j} E_Q[\ln \Psi(D_j)] \right) \]

- It is like fixing the variables in the Markov blanket to their average (mean)
- Then renormalize

- Eg.
  
  \[ Q_1(X_1) = \frac{1}{Z_1} \exp\left( E_{Q(X_2)}[\ln \Psi(X_1, X_2)] + E_{Q(X_3)}[\ln \Psi(X_1, X_3)] + E_{Q(X_4)}[\ln \Psi(X_1, X_4)] + E_{Q(X_5)}[\ln \Psi(X_1, X_5)] \right) \]
Graphical Model for Bayesian Linear Regression

- The joint distribution of all the variables is given by
  \[ p(y, X, w, \alpha) = p(y|X, w)p(w|\alpha)p(\alpha). \]

- The likelihood for \( w \) is given by
  \[ p(y|X, w) = \prod_{t=1}^{N} \mathcal{N}(y_t|w^T x_t, \beta^{-1}). \]

- The prior over \( w \) is given by
  \[ p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1} I). \]

- The prior over \( \alpha \) is given by
  \[ p(\alpha) = \text{Gam}(\alpha|a_0, b_0). \]
Find a variational posterior distribution \( q(w, \alpha) = q(w)q(\alpha) \) which is an approximation to the posterior distribution \( p(w, \alpha|X, y) \),

\[
\begin{align*}
q(w, \alpha) &= q(w)q(\alpha) \\
&\approx p(w, \alpha|X, y).
\end{align*}
\]

Compute the optimal variational distribution \( q^*(w) \) and \( q^*(\alpha) \) using VBEM:

\[
\begin{align*}
\log q^*(w) &= \mathbb{E}_w \log p(X, y, w, \alpha) + \text{const}, \\
\log q^*(\alpha) &= \mathbb{E}_\alpha \log p(X, y, w, \alpha) + \text{const}.
\end{align*}
\]
Optimized $q^*(w)$

\[
\log q^*(w) = \mathbb{E}_\alpha \log [p(y | X, w)p(w | \alpha)p(\alpha)] + \text{const}
= \log p(y | X, w) + \mathbb{E}_\alpha p(w | \alpha) + \text{const}
= -\frac{\beta}{2} \sum_{t=1}^{N} (y_t - w^T x_t)^2 - \frac{1}{2} \mathbb{E}_\alpha \{\alpha\} w^T w + \text{const}
= -\frac{\beta}{2} (y - Xw)^T (y - Xw) - \frac{1}{2} \mathbb{E}_\alpha \{\alpha\} w^T w + \text{const}
= -\frac{1}{2} w^T \left( \mathbb{E}_\alpha \{\alpha\} I + \beta X^T X \right) w + \beta w^T X^T y + \text{const},
\]

which is the of a Gaussian distribution, leading to

\[
q^*(w) = \mathcal{N}(w | m_N, S_N),
\]

\[
m_N = \beta S_N X^T y,
\]

\[
S_N = \left( \mathbb{E}_\alpha \{\alpha\} I + \beta X^T X \right)^{-1}.
\]
\[
\log q^*(\alpha) = \mathbb{E}_w \log p(X, y, w, \alpha) + \text{const}
\]
\[
= \mathbb{E}_w \log [p(y|X, w)p(w|\alpha)p(\alpha)] + \text{const}
\]
\[
= \log p(\alpha) + \mathbb{E}_w \log p(w|\alpha) + \text{const}
\]
\[
= (\alpha_0 - 1) \log \alpha - b_0 \alpha + \frac{D}{2} \log \alpha - \frac{\alpha}{2} \mathbb{E}_w \{w^T w\} + \text{const}
\]
\[
= \left( a_0 + \frac{D}{2} - 1 \right) \log \alpha - \left( b_0 + \frac{\alpha}{2} \mathbb{E}_w \{w^T w\} \right) \alpha,
\]
which is the log of a Gamma distribution, leading to
\[
q^*(\alpha) = \text{Gam}(\alpha | a_N, b_N),
\]
\[
a_N = a_0 + \frac{D}{2},
\]
\[
b_N = b_0 + \frac{1}{2} \mathbb{E}_w \{w^T w\}.
\]
Algorithm Outline

- Re-estimate $q^*(\alpha)$:
  \[ q^*(\alpha) = \text{Gam}(\alpha | a_N, b_N), \]
  \[ a_N = a_0 + \frac{D}{2}, \]
  \[ b_N = b_0 + \frac{1}{2} \mathbb{E}_w \{ w^T w \}, \]
  \[ \mathbb{E}_w \{ w^T w \} = m_N^T m_N + \text{tr} \{ S_N \}. \]

- Re-estimate $q^*(w)$:
  \[ q^*(w) = \mathcal{N}(w | m_N, S_N), \]
  \[ m_N = \beta S_N X^T y, \]
  \[ S_N = \left( \mathbb{E}_\alpha \{ \alpha \} I + \beta X^T X \right)^{-1}, \]
  \[ \mathbb{E}_\alpha \{ \alpha \} = \frac{a_N}{b_N}. \]

Predictive Distribution

The predictive distribution over $y_\ast$, given a new input $x_\ast$, is evaluated using the Gaussian variational posterior:

\[ p(y_\ast | x_\ast, X, y) = \int p(y_\ast | w, x_\ast) p(w | X, y) dw \]
\[ = \int p(y_\ast | w, x_\ast) p(w | X, y) dw \]
\[ \approx \int p(y_\ast | w, x_\ast) q(w) dw \]
\[ = \int \mathcal{N}(y_\ast | w^T x_\ast, \beta^{-1}) \mathcal{N}(w | m_N, S_N) dw \]
\[ = \mathcal{N}(y_\ast | m_N^T x_\ast, \beta^{-1} + x_\ast^T S_N x_\ast). \]
Forward Direction

- $KL(P || Q)$
  
  \[
  = \sum_x P(X) \ln \frac{P(X)}{Q(X)} = \sum_x P(X) \ln P(X) - \sum_x P(X) \ln Q(X)
  \]

  - no Q in it, ignore
  - cross entropy

Given $Q \propto \prod_i \Psi(X_i)$, it is equal to
\[
\sum_i \sum_x P(X) \ln \Psi(X_i) + \text{const.}
\]

We need solve inference problems
\[
\sum_{x_1, \ldots, x_n} P(X_1, \ldots, X_n) \ln \Psi(X_i)
\]
which are hard in the first place!
Gibbs Sampling for Ising

- Binary MRF $G = (V,E)$ with pairwise clique potentials

$$p(x; \theta) \propto \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \right\}$$

1. pick a node $s$ at random
2. sample $u \sim \text{Uniform}(0,1)$
3. update node $s$:
   $$x_s^{(m+1)} = \begin{cases} 
   1 & \text{if } u \leq \left\{ 1 + \exp\left[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} x_t^{(m)})\right]\right\}^{-1} \\
   0 & \text{otherwise}
   \end{cases}$$
4. goto step 1

Naive MF for Ising

- use a variational mean parameter at each site $\nu_s \in (0, 1)$

1. pick a node $s$ at random
2. update its parameter:
   $$\nu_s \leftarrow \left\{ 1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \nu_t)]\right\}^{-1}$$
3. goto step 1
Exponential Representations via Maximum Entropy

\[ p^* := \arg \max_{p \in \mathcal{P}} H(p) \text{ subject to } \mathbb{E}_p[\phi_\alpha(X)] = \mu_\alpha \text{ for all } \alpha \in \mathcal{I}, \]
\[ \mu_\alpha := \frac{1}{n} \sum_{i=1}^{n} \phi_\alpha(X_i), \text{ for all } \alpha \in \mathcal{I}, \]

where each \( \alpha \) in some set \( \mathcal{I} \) indexes a function \( \phi_\alpha : \mathcal{X} \rightarrow \mathbb{R} \). For example, if we set \( \phi_1(x) = x \) and \( \phi_2(x) = x^2 \), then the observations correspond to empirical versions of the first and second moments of the random variable \( X \). Based on the \( |\mathcal{I}| \)-dimensional vector of empirical expectations \( \hat{\mu} = (\hat{\mu}_\alpha, \alpha \in \mathcal{I}) \), our goal is to infer a full probability distribution over the random variable \( X \). In particular, we represent it can be shown — by calculus of variations in the general continuous case, and by ordinary calculus in the discrete case — that the optimal solution \( p^* \) takes the form

\[ p_\theta(x) \propto \exp \left\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(x) \right\}, \]

where \( \theta \in \mathbb{R}^d \) represents a parameterization of the distribution in exponential family form.

Basics of Exponential Families

\[ p_\theta(x_1, x_2, \ldots, x_m) = \exp \left\{ \langle \theta, \phi(x) \rangle - A(\theta) \right\}, \]
\[ A(\theta) = \log \int_{\mathcal{X}^m} \exp(\theta, \phi(x)) \nu(dx). \]

let \( \phi = (\phi_\alpha, \alpha \in \mathcal{I}) \) be a collection of functions \( \phi_\alpha : \mathcal{X}^m \rightarrow \mathbb{R} \), known either as potential functions or sufficient statistics.

For a given vector of sufficient statistics \( \phi \), let \( \theta = (\theta_\alpha, \alpha \in \mathcal{I}) \) be an associated vector of canonical or exponential parameters.

Examples of Graphical Models in Exponential Form

Ising Model

\[ p_\theta(x) = \exp \left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s, t) \in E} \theta_{st} x_s x_t - A(\theta) \right\}, \]

Gaussian MRF

\[ p_\theta(x) = \exp \left\{ \langle \theta, x \rangle + \frac{1}{2} \langle x, \Theta x^T \rangle - A(\theta, \Theta) \right\}, \]
\[ \log Z \geq \sup_Q \{ E_Q[\psi(x)] + H[Q(x)] \} \]
\[ \log Z \geq \sup_Q \{ E_Q[\theta^T \phi(x)] + H[Q(x)] \} \]
\[ \log Z \geq \sup_Q \{ \theta^T E_Q[\phi(x)] + H[Q(x)] \} \]

Most common specialized family:
- **"log-linear models"**
  \[ \psi(x) = \sum_c \theta_c \phi_c(x_c) = \theta^T \phi(x) \]
  (natural parameters of EFs)
- **linear in parameters** \( \theta \)
- **clique potentials** \( \phi(x) \)
  (sufficient statistics of EFs)

\[ A(\theta) \geq \sup_{\mu \in M} \{ \theta^T \mu - A^*(\mu) \} \]

\( M = \text{set of all moment parameters realizable under subclass } Q \)