Probabilistic Graphical Models

and Marginal Estimation Algorithms

Olivier Peltre

IMJ-PRG

01.10.2020

CRIL Lens

Outline

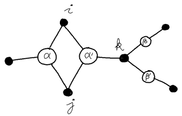
A graphical model is defined by:

- a collection of *variables* $x_i \in E_i$ for $i \in \Omega$
- a collection of *factors* $f_{i_1 \dots i_n}(x_{i_1}, \dots, x_{i_n})$ for $i_1, \dots, i_n \in \Omega$

Notations: for any $\alpha = \{i_1, \dots, i_n\} \subseteq \Omega$:

- $E_\alpha = \prod_{i \in \alpha} E_i$ local configuration space
- $f_\alpha(x_\alpha) = f_{i_1 \dots i_n}(x_{i_1}, \dots, x_{i_n})$ local observable

Factor graph:



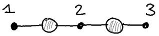
Definition (Graphical Model)

$$p(x) = \frac{1}{Z} \prod_{\alpha \subseteq \Omega} f_\alpha(x_\alpha)$$

Applications: Statistical physics, decoding (telecoms), Bayesian inference, Boltzmann machines...

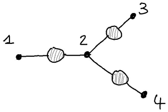
I - Graphical Models

Examples:



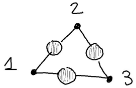
$$p(x) = f(x_1, x_2) f(x_2, x_3)$$

$$\Rightarrow p(x) = \frac{p(x_1, x_2) p(x_2, x_3)}{p(x_2)}$$



$$p(x) = f(x_1, x_2) f(x_2, x_3) f(x_2, x_4)$$

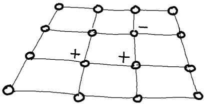
$$\Rightarrow p(x) = \frac{p(x_1, x_2) p(x_2, x_3) p(x_2, x_3)}{p(x_2)^2}$$



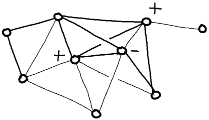
$$p(x) = f(x_1, x_2) f(x_2, x_3) f(x_3, x_1)$$

II - Thermodynamics

- Collection of atoms $i \in \Omega$
- Each carrying a spin $x_i \in \{\pm 1\}$
- Graph structure $i \sim j \;\Leftrightarrow\; i$ spatially close to j



(a)



(b)

Crystal

Spin Glass

1. Energy and Potentials

$H(x) \in \mathbb{R}^{E_\Omega}$ hamiltonian (total energy) $\rightarrow$ *sum of local interaction potentials*

Definition (Hamiltonian with pairwise interactions)

$$H(x) = \sum_i h_i(x_i) + \sum_{i \sim j} h_{ij}(x_i, x_j)$$

Magnetism: $x_i \in \{\pm 1\}$

- $h_i(x_i) = -b_i \, x_i$ — $b_i \in \mathbb{R}$ local field (bias)
- $h_{ij}(x_i, x_j) = -w_{ij} \, x_i x_j$ — $w_{ij} \in \mathbb{R}$ local coupling (weight)

Example: Ising model of ferromagnetism

- $\Omega = \mathbb{Z}^d$ regular lattice
- $B \in \mathbb{R}$ macroscopic field
- $W = \pm 1$ (anti)ferromagnetic crystal

$$H(x) = -B \sum_i x_i - W \sum_{i \sim j} x_i x_j$$

Ferromagnetic case: $W = +1$

- $h_{ij}(x_i, x_j) = -1$ for $++$ and $--$
- $h_{ij}(x_i, x_j) = +1$ for $+-$ and $-+$

2. Probabilistic Model

Definition (Gibbs state)

At equilibrium temperature $T = \frac{1}{\theta}$ *with a thermostat*

$$p(x|\theta) = \frac{\mathrm{e}^{-\theta H(x)}}{Z(\theta)}$$

where $Z(\theta) = \sum_{x \in \{\pm 1\}^{\Omega}} \mathrm{e}^{-\theta H(x)}$ *is the partition function*

N.B. $p(x|\theta) = \frac{1}{Z(\theta)} \prod_{\alpha} \mathrm{e}^{-\theta h_{\alpha}(x_{\alpha})}$ is a graphical model.

Problem: computing $Z(\theta)$ is intractable...

- High temperature: $p(x|\theta) \underset{\theta \to 0}{\longrightarrow} \frac{1}{|E_{\Omega}|}$ uniform distribution
- Low temperature: $p(x|\theta) \underset{\theta \to \infty}{\longrightarrow} \delta_{x_{min}}$ energy minimum

Example: Ising model

$$H(x) = -B\sum_i x_i - \sum_{i\sim j} x_i x_j$$

High temperature:

$x_i \sim \mathcal{U}(-1,+1)$

$\mathbb{E}[x_i] = 0$

Low temperature:

$x_i = \mathrm{sign}(B)$

$\mathbb{E}[x_i] = \mathrm{sign(B)}$ $\rightarrow$ magnetization

N.B. $B \rightarrow 0^{\pm}$ *singularity* $\rightarrow$ spontaneous $\pm$ magnetization

Phase transition: $\mathbb{E}[x_i]$ is not an analytic function of temperature

Theorem (Free Energy Principle)

$p(x|\theta) = \frac{1}{Z(\theta)}\, \mathrm{e}^{-\theta H(x)}$ *minimises the variational free energy functional:*

$$\mathcal{F}_\theta(q) = \mathbb{E}_q[H] - \frac{1}{\theta}\mathbb{E}_q[-\ln(q)]$$

whose minimum is the free energy:

$$F(\theta) = -\frac{1}{\theta} \ln \sum_x \mathrm{e}^{-\theta H(x)} = -\frac{1}{\theta} \ln Z(\theta)$$

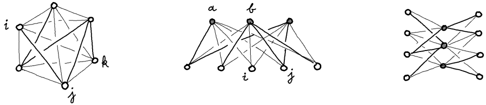
Introducing dependencies in the hamiltonian itself, temperature acts by scalings

- $S(p) = -\sum_x p(x) \ln p(x)$ Shannon entropy
- $\mathbb{F}(H) = -\ln \sum_x \mathrm{e}^{-H(x)}$ free energy

Legendre transform $S(p) \to \mathbb{F}(H)$:

- $\mathcal{F}(p, H) = \langle\, p \,|\, H \,\rangle - S(p)$ variational free energy

III - Boltzmann Machines



(a) (b) (c)

Restricted Boltzmann Machine: bipartite graph (b)

- x observed variables (image)
- y latent variables (abstract features)

$$p(x,y) = \frac{1}{Z}\,\mathrm{e}^{-H(x,y)} \qquad \text{where} \qquad H(x,y) = \sum_i h_i(x_i) + \sum_j h_j(y_j) + \sum_{i\sim j} h_{ij}(x_i,y_j)$$

Train the network tho generate similar images

Maximise the log-likelihood of N training samples $\bar{x}^1, \ldots, \bar{x}^N$

Definition

$$\begin{aligned}\mathcal{L} &= -\frac{1}{N}\sum_{s=1}^{N} \ln p(\bar{x}^s) \\ &= -\frac{1}{N}\sum_{s=1}^{N} \ln \sum_{y} p(\bar{x}^s, y)\end{aligned}$$

$\mathcal{L}$ may be rewritten in terms of free and effective energies

$$\mathcal{L} = -\frac{1}{N}\sum_{s=1}^{N} \ln \frac{\sum_y \mathrm{e}^{-H(\bar{x}^s, y)}}{\sum_x \sum_y \mathrm{e}^{-H(x,y)}}$$

Perform gradient descent on $\mathcal{L}$ w.r.t. model parameters $h_i, h_j, h_{ij} \dots$

$\rightarrow$ Estimate the directional derivatives $\frac{\partial \mathcal{L}}{\partial h_{ij}}$?

Proposition

$$\frac{\partial \mathcal{L}}{\partial h_{ij}} = -\frac{1}{N} \sum_{s=1}^{N} \mathbb{E}\left[h_{ij}(\bar{x}_i^s, y_j) \,\middle|\, \bar{x}^s\right] + \mathbb{E}\left[h_{ij}(x_i, y_j)\right]$$

Markov Properties: $p(y|\bar{x}^s) = \prod_j p(y_j|\bar{x}^s)$ ($\leftarrow$ RBM assumption)

Yet the second term requires **marginal distributions** $p_{ij}(x_i, y_j)$ to be estimated.

Contrastive Divergence (Hinton): Estimate the gradient $\frac{\partial \mathcal{L}}{\partial h_{ij}}$ via Gibbs sampling

Assume $x_i, y_j \in \{\pm 1\}$ and $H(x, y) = \sum_i a_i x_i + \sum_j b_j y_j + \sum_{i \sim j} w_{ij} x_i y_j$

CD relies on the bipartite structure of the interaction graph (**R**BM):

$$p(y|x) = \prod_j p(y_j|x) \quad \text{with} \quad p(y_j = 1|x) = \sigma\Big(\sum_{i \sim j} w_{ij} x_i + b_j\Big)$$

$$p(x|y) = \prod_i p(x_i|y) \quad \text{with} \quad p(x_i = 1|y) = \sigma\Big(\sum_{j \sim i} w_{ij} y_j + a_i\Big)$$

Starting from some configuration $x = \bar{x}^s$

- Draw $y \sim p(y|x)$
- Draw $x \sim p(x|y)$

Eventually loop k times (CD-k) and average over samples to estimate $\mathbb{E}[w_{ij} x_i y_j]$

IV - Marginal Estimation Algorithms

Gibbs sampling (CD) works well for training RBMs

However performs poorly on other tasks (e.g. conditional BMs)

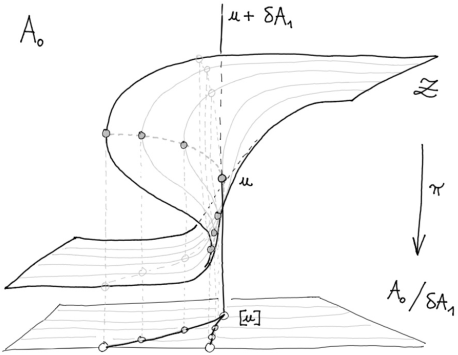
Limitations:

- high correlation between samples
- iterate long enough to approximate the stationary measure

Belief Propagation (BP) and message-passing algorithms take a different approach to approximate the local marginals $p_\alpha(x_\alpha)$.

Local beliefs $q_\alpha(x_\alpha)$ are iterated upon until marginal consistency is reached.

Total energy $H(x) = \sum_\alpha u_\alpha(x_\alpha)$ is conserved at each step.



Analogous to a diffusion equation $\dot{u} = \delta\varphi$, where $u \sim$ *potential*, $\varphi \sim$ *heat flux*.