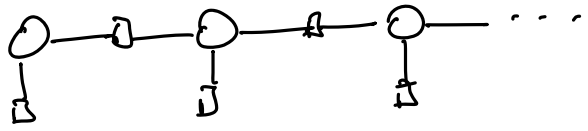
LECTURE 6 - INFERENCE (CONT.)

1. Belief Propagation

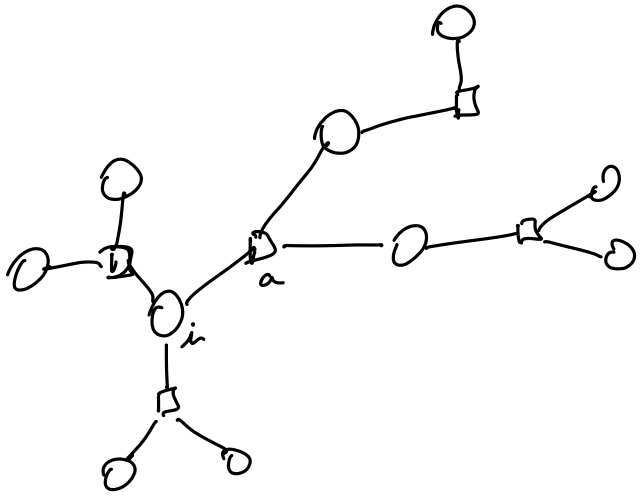
last time : BP on Ising chains



→ Forward ($\nu_{\to j}$) and backward ($\nu_{j\leftarrow}$) messages computed using dynamic programming.

- BP on trees

$G = (V, A, E)$
$E \subseteq V \times A$



$$p(x) = \frac{1}{Z} \prod_{a \in A} \phi_a(x_{N(a)})$$

→ At iteration t, two messages for each edge (i,a)

- $\nu_{i \to a}^{(t)}(x_i)$
- $\hat{\nu}_{a \to i}^{(t)}(x_i)$

→ BP / sum-product equations:

(*) $$\nu_{j\to a}^{(t+1)}(x_j) \propto \prod_{b\in N(j)\setminus\{a\}} \hat{\nu}_{b\to j}^{(t)}(x_j)$$

(**) $$\hat{\nu}_{a\to j}^{(t)}(x_j) \propto \sum_{x_{N(a)\setminus\{j\}}} \phi_a(x_{N(a)}) \prod_{k\in N(a)\setminus\{j\}} \nu_{k\to a}^{(t)}(x_k)$$

→ Estimate marginals:

$$\nu_i^{(t)}(x_i) \propto \prod_{a\in N(i)} \hat{\nu}_{a\to i}^{(t-1)}(x_i)$$

→ Algorithm:

Initialize $\nu_{i\to a}^{(0)}(\cdot)$ as i.i.d. with some distr. p_0.

For $t=0,\dots,t_{max}$

- For each $(j,a)\in E$, compute $\hat{\nu}_{a\to j}^{(t)}$ using (**)
- For each $(j,a)\in E$ compute $\nu_{a\to j}^{(t+1)}$ using *
- If no message changed: return current messages.

return "not converged".

Theorem (BP is exact on trees)

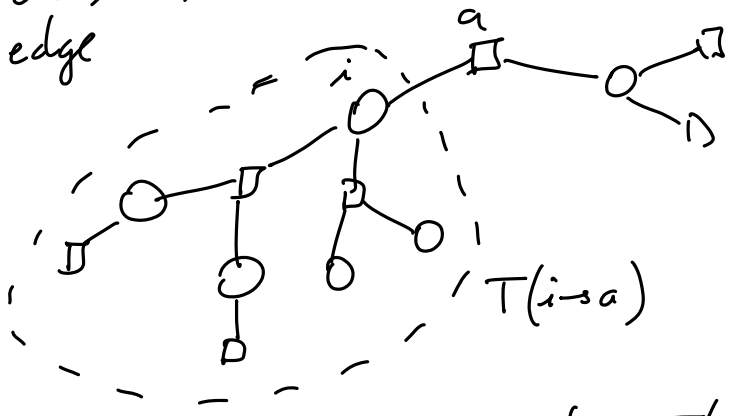
Let t^* = maximum distance between any two variable nodes (diameter)

1. BP updates converge after at most t^* iterations, irrespective of initial conditions.
   For any $(i,a) \in E$, $t > t^*$, $\nu_{i\to a}^{(t)} = \nu_{i\to a}^{*}$ (fixed-point messages)
   $\hat{\nu}_{a\to i}^{(t)} = \hat{\nu}_{a\to i}^{*}$

2. The fixed-point messages provide exact marginals.
   i.e. $p(x_i) = \nu_i^*(x_i)$ for all $i \in V$.

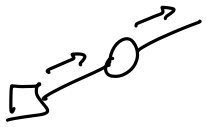
proof (sketch):

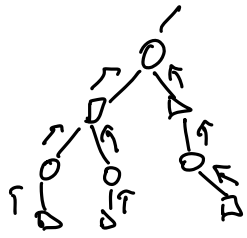
For an edge $(i,a) \in E$, let $T(i\to a)$ be the subtree rooted at this edge



Denote by $t^*(i\to a)$ the depth of the tree $T(i\to a)$.
We can show by induction on $t^*(i\to a)$ that the message $\nu_{i\to a}^{(t)}$ converges to the marginal of x_i in the subgraph $T(i\to a)$ after $t^*(i\to a)$ steps.

(exercise = similar to DP recursion in the chain)





The marginals are then given by.

$$p(x_i) \propto \sum_{x_{V\setminus\{i\}}} \prod_a \phi_a(x_{N(a)})$$

$$\propto \prod_{b \in N(i)} \sum_{x_{N(b)\setminus\{i\}}} \phi_b(x_{N(b)}) \prod_{j \in N(b)\setminus\{i\}} m^*_{j\to b}(x_j)$$

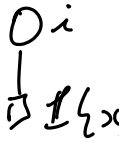
$$= m^*_i(x_i)$$

□

Remarks:

- conditioning can be done using the same BP equations by adding factors $\mathbb{1}\{x_i = x_i^\circ\}$

$$p(x \mid x_B = x_B^\circ) \propto \prod_a \phi_a(x_{N(a)}) \cdot \mathbb{1}\{x_B = x_B^\circ\}$$



- complexity depends – linearly on depth of tree
  – exponentially on size of factors

BP on general graphs ("Loopy BP")

→ Note that one can run the BP algorithm on any graph

→ In general, BP does not converge to the correct marginals.

→ Yet, it often works well in practice ("Loopy BP")

→ It can be shown that the fixed points of the BP equations correspond to the stationary points of a quantity known as the "Bethe Free Energy", which is often a good approximation of the Free Energy $\log Z$

Max-product and Min-sum

- Sometimes we care about optimization instead of marginals

  Ex: MAP inference

  $$\max_{z} \; p(z \mid x = x_0)$$

  (image reconstruction, sequence decoding)

- Similar algorithms to BP can be derived for such optimization problems.

→ **Max-product** : $(\Sigma, \Pi) \mapsto (\max, \Pi)$

$$\nu_{j\to a}^{(t+1)}(x_j) \propto \prod_{b \in N(j)\setminus\{a\}} \hat{\nu}_{b\to j}^{(t)}(x_j)$$

$$\hat{\nu}_{a\to j}^{(t)}(x_j) \propto \max_{x_{N(a)\setminus\{j\}}} \phi_a(x_{N(a)}) \prod_{h \in N(a)\setminus\{j\}} \nu_{h\to a}^{(t)}(x_h)$$

max-marginals : $\nu_i^{(t)}(x_i) \propto \prod_{a \in A(i)} \hat{\nu}_{a\to i}^{(t-1)}(x_i)$

→ **Min-sum** : $(\Sigma, \Pi) \mapsto (\min, \Sigma)$

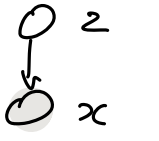
$$\max_x \prod_a e^{-E_a(x_{N(a)})} \iff \min_x \sum_a E_a(x_{N(a)})$$

Similar updates, defined up to an additive constant (replaces normalization)

---

2. Learning in latent-variable models

Latent-variable models : ex:



(unobs.)
(observed)

Goal: estimate parameters θ $p(x, z; \theta)$ when z is not observed

Maximum likelihood estimator:

$$\max_{\theta} \sum_i \log p(x_i;\theta)$$

$$= \max_{\theta} \sum_i \log \int p(x_i, z; \theta)\, dz$$

↖ depending on discrete or continuous z.

→ this problem can be highly non-convex!

Q: How can we solve this?

- **K-means**

→ clustering with discrete assignments

- $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$: data
- $\mu_k \in \mathbb{R}^p$, $k = 1, \dots, K$: cluster centers (parameters)
- $z_i \in [K]$: cluster assignments

→ Objective:

$$\min_{\mu, z} \left\{ L(\mu, z) := \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2 \right\}$$

Algorithm

Repeat until convergence

→ $z_i = \arg\min_k \|x_i - \mu_k\|^2$

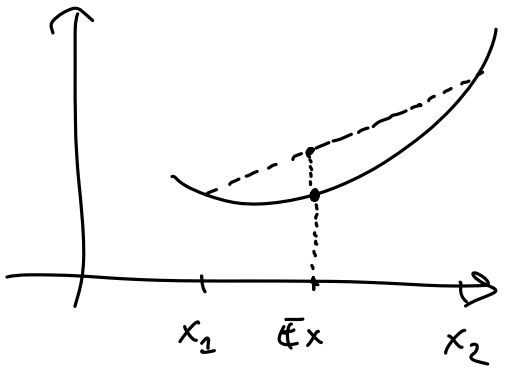
→ $\mu_k = \frac{1}{|\{i : z_i = k\}|} \sum_{i : z_i = k} x_i$

→ alternate between minimization of $L(\mu, z)$ w.r.t. z and μ.

EM algorithm

- Recall Jensen's inequality:

→ if φ is convex, then $\varphi(\mathbb{E}X) \leq \mathbb{E}\,\varphi(X)$



$$X = \begin{cases} x_1 & \text{w.p. } \theta \\ x_2 & \text{— } 1-\theta \end{cases}$$

→ we have equality ($\varphi(\mathbb{E}x) = \mathbb{E}\,\varphi(x)$) if and only if $X = x_0$ almost surely (constant)

- "Variational" lower bound:

$$\log p(x;\theta) = \log \sum_z p(x,z;\theta)$$

$$= \log \sum_z q(z)\,\frac{p(x,z;\theta)}{q(z)} \quad \left(\text{for some } q(z) \geq 0,\ \sum_z q(z) = 1\right)$$

$$= \log \mathbb{E}_q\left[\frac{p(x,z;\theta)}{q(z)}\right]$$

$$\overset{(*)}{\geq} \mathbb{E}_q\left[\log \frac{p(x,z;\theta)}{q(z)}\right] \quad \text{(Jensen's, since log is concave)}$$

We thus have:

$$\log p(x;\theta) \geq \underbrace{\mathbb{E}_q[\log p(x,z;\theta)] - \mathbb{E}_q[\log q(z)]}_{=: L(\theta, q)}$$

→ Note that we have equality in (*) when $q(z) \propto p(x,z;\theta)$

i.e. $q(z) = p(z|x;\theta)$

→ <u>E-M algorithm</u> (Expectation - Maximization)

For $t \geq 0$:

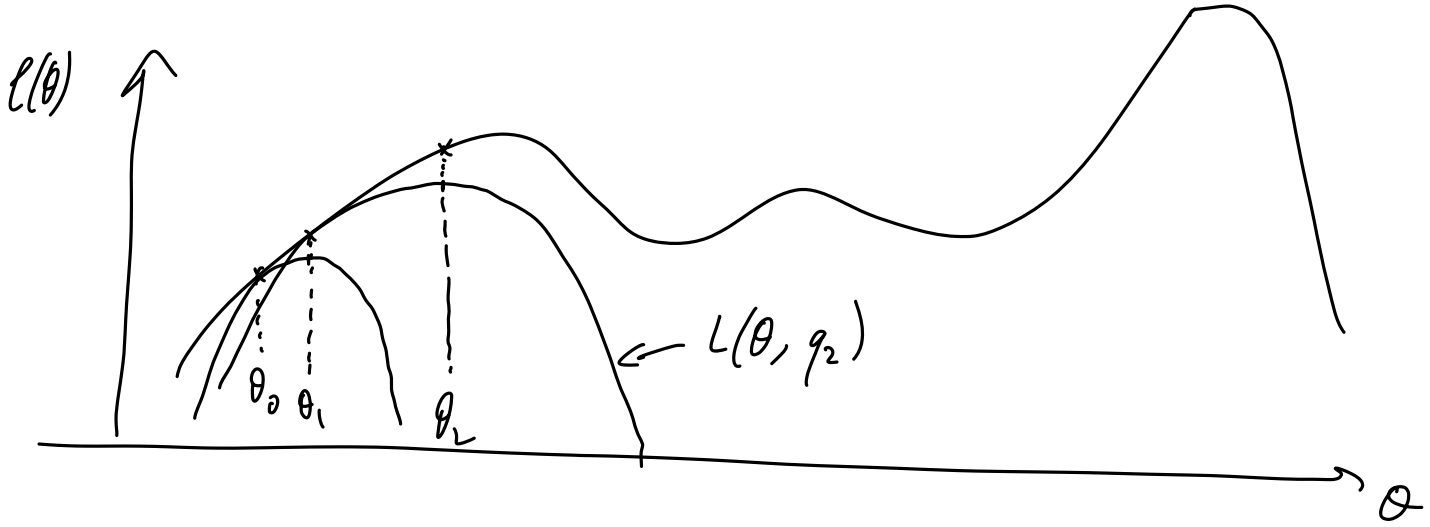
- $q_{t+1} \in \arg\max_q L(\theta_t, q) \quad \left[= p(z|x;\theta_t) \right]$

  (E-step)

- $\theta_{t+1} \in \arg\max_\theta L(\theta, q_{t+1}) \quad \left[= \arg\max_\theta \mathbb{E}_{q_{t+1}}[\log p(x,z;\theta)] \right]$

  (M-step)

"complete" log-likelihood



Remarks:

- corresponds to alternating maximizations on $L(\theta, q)$
- $\ell(\theta) = \max_q L(\theta, q) = L(\theta, p(\cdot|x,\theta))$

$\ell(\theta_t)$ is non-decreasing $\Rightarrow$ θ_t converges to stationary point (not necessarily global maximum)

- First example of a variational method (more next week)

- The gap between $\log p(x;\theta)$ and $L(\theta, q)$ corresponds to a Kullback-Leibler divergence (relative entropy)

$$D_{KL}(p, q) = E_p\left[\log \frac{p(x)}{q(x)}\right]$$

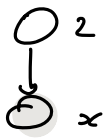
$$\ell(\theta) - L(\theta, q) = \log p(x;\theta) - E_q\left[\log \frac{p(z,x;\theta)}{q(z)}\right]$$

$$= E_q\left[\log \frac{q(z)}{p(z|x;\theta)}\right]$$

$$= D_{KL}(q, p(\cdot|x;\theta))$$

Example: Gaussian Mixtures

- $z \sim \pi$
- $x|z=k \sim \mathcal{N}(\mu_k, \Sigma_k)$



$\theta = (\pi, \{\mu_k\}, \{\Sigma_k\})$ $\quad \sum_k \pi_k = 1$

→ n points x_i, corresponding latents: z_i

→ E-step:

$$\underbrace{p(z_i = k \mid x_i; \theta)}_{=: \tau_{ik}} \propto \mathcal{N}(x_i \mid \mu_k, \Sigma_k) \cdot \pi_k$$

→ M-step:

- complete log-likelihood:

$$\sum_i \log p(x_i, z_i, \theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{z_i = k\} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{z_i = k\} \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

$$\mathbb{E}\left[\sum_i \log p(x_i, z_i, \theta) \mid x; \theta_t \right] = \sum_{i=1}^{n} \sum_{k=1}^{K} p(z_i = k \mid x_i; \theta_t) \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} p(z_i = k \mid x_i; \theta_t) \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

($p(z_i = k \mid x_i; \theta_t) = \tau_{ik}^t$)

$$\Rightarrow \begin{cases} \pi_k^{t+1} \propto \sum_{i=1}^{n} \tau_{ik}^t \\ \mu_k^{t+1} = \dfrac{\sum_i \tau_{ik}^t x_i}{\sum_i \tau_{ik}^t} \\ \Sigma_k^{t+1} = \dfrac{\sum_i \tau_{ik}^t (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T}{\sum_i \tau_{ik}^t} \end{cases}$$