Graphical Models, Exponential Families and Variational Inference

4.2 - 4.3
Bethe Kikuchi and Expectation Propagation

Vincent Adam
Alessandro Davide Ialongo

February 23, 2015

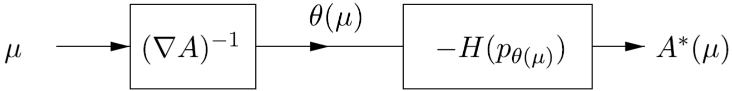
Reminder chap.3

- Set of realisable mean parameters
$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p \text{ s.t. } \mathbb{E}_p[\phi(X)] = \mu\}$$

- conjugate dual of the log partition function
$A(\theta) = \int dx\, exp(\langle \theta, \phi(x) \rangle)$

$$A(\theta) = \sup_{\mu} \{\langle \theta, \mu \rangle - A^*(\mu)\}$$

- Bijection



Reminder chap.4

- Bethe Approximation to the Entropy (for a graph $G = (V, E)$)

  $-A^*(\tau) \approx H_{\text{Bethe}}(\tau) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st})$

- Bethe Variational problem

$$\max_{\tau \in \mathbb{L}(G)} \{\langle \theta, \tau \rangle + H_{\text{Bethe}}(\tau)\}$$

- Deriving the **sum-product/belief propagation algorithm** for (pairwise) graphs in general:
  - Exact on trees
  - Using the Bethe Variational Approximation on loopy graphs

Section 4.2 (p98-109)

Outline

- Computing marginals for a more general class of distributions
- Represented by **Hypergraphs** and **Hypertrees** (e.g. junction trees)
- Same kind of approximations as in the standard case:
  - Approximation of the hypergraph's actual entropy $H_{app}(\tau) \approx H(p_\mu)$
  - Constructing an outer bound $\mathbb{L}_t(G)$ to the hypergraph's marginal polytope $\mathbb{M}_t(G)$
- Leading to **Generalized Belief Propagation**

4.2.1 (p99)

Hypergraphs and Hypertrees

- Generalization of pairwise MRF: edges can be between an arbitrary number of vertices
- **Hypergraphs:** $G = (V, E)$:
  - V vertex set as before: $\{1, ..., m\}$
  - E hyperedge set: $E \subseteq P(V)$ = power set of V
    - e.g. $V = \{1, 2, 3, 4\}$ - $E = \{\{1\}, \{2, 3\}, \{1, 2, 3\}, \{1, 3, 4\}\}$
- **Maximal hyperedge**: one not included into any other (i.e. $\{1, 2, 3\}, \{1, 3, 4\}$)
- **Hypertrees** or acyclic hypergraphs: hypergraphs whose maximal hyperedges and their intersection specify a junction tree
  - Hypertree width = size of the largest hyperedge - 1
- Hypergraph with maximal hyperedges of size two: generalization of pairwise MRF
- Hypertree: generalization of the Junction tree

4.2.1(p100)

Poset - Partially Ordered Set

- Set inclusion induces a **partial ordering** on the set of hyperedges E:
  - Only partial since not $\forall g, \forall h \in E : g \subseteq h \vee h \subseteq g$, i.e. we can have disjoint and partially disjoint hyperedges
- Visual representation, **Poset diagram** (displaying inclusion relations):

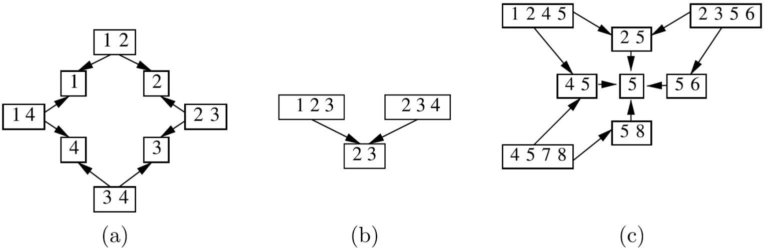


Figure : 4.4 (p100)

Appendix E.1 (p286-287)

Moebius Function

- Associated with any poset there is a **Moebius function**: $\omega : E \times E \to \mathbb{R}$ (Appendix E.1, p286):
  - Base cases: $\omega(g, g) = 1$, $\omega(g, h) = 0$ if $g \not\subseteq h$
  - Recursively: $\omega(g, h) = -\sum_{\{f | g \subseteq f \subset h\}} \omega(g, f)$
  - Also defined as the multiplicative inverse of the zeta function $\zeta(g, h) = \begin{cases} 1 & \text{if } g \subseteq h \\ 0 & \text{otherwise} \end{cases}$:
    - $\sum_{f \in E} \omega(g, f)\zeta(f, h) = \sum_{\{f | g \subseteq f \subseteq h\}} \omega(g, f) = \delta(g, h)$
    - So values of $\omega(g, h)$ can be found by inverting the matrix of zeta values $Z(i, j) = \zeta(g_i, g_j)$ for some indexing of the hyperedge set E

Appendix E.1 (p286-287)

Moebius Function Examples

- If $E = P(\{1, ..., m\})$ then $\omega(g, h) = (-1)^{|h \setminus g|} \mathbb{I}(g \subseteq h)$
- Example (4.4(b)): $E = \{\{23\}, \{123\}, \{234\}\}$:

$$Z^{-1} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \Omega$$

4.2.2 (p100-101)

Hypertree-based Factorization

- By the Moebius inversion formula (Lemma E.1 p287), for real-valued functions Υ and Ω on a poset E:

$$\Omega(h) = \sum_{g \subseteq h} \Upsilon(g) \qquad \text{and} \qquad \Upsilon(h) = \sum_{g \subseteq h} \Omega(g)\omega(g,h) \qquad \forall h \in E$$

- Applying it to the set of marginals $\mu = \{\mu_h | h \in E\}$ we gain a new set of functions $\phi = \{\phi_h | h \in E\}$:

$$\log \mu_h(x_h) = \sum_{g \subseteq h} \log \phi_g(x_g) \qquad \text{and, conversely}$$

$$\log \phi_h(x_h) = \sum_{g \subseteq h} \omega(g,h) \log \mu_g(x_g)$$

- This gives us an alternative factorization for all hypertrees containing all (but not only) the intersections between maximal hyperedges (which includes junction trees):

$$p_\mu(x) = \prod_{h \in E} \phi_h(x_h; \mu) \qquad (4.42)$$

4.2.2 - Example 4.4 (p101-102) I

Hypertree Factorization

- If the hypergraph $G = (V, E)$ is actually a tree, E contains the tree's vertices and its pairwise edges
- then, by (4.42): $p_\mu(x) = \prod_{s \in V} \phi_s(x_s) \prod_{(s,t) \in E} \phi_{st}(x_s, x_t)$
- and since $\forall g: \ \omega(g, g) = 1$ and $\omega(\{s\}, \{s, t\}) = -1$,
- and $\log \phi_h(x_h) = \sum_{g \subseteq h} \omega(g, h) \log \mu_g(x_g)$:

$$p_\mu(x) = \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$$

Recovering tree factorization (4.8)

4.2.2 - Example 4.4 (p101-102) II

Hypertree Factorization

- Practical example (Figure 4.4(c)):

$$E=\{\{5\},\{2,5\},\{4,5\},\{5,6\},\{5,8\},\{1,2,4,5\},\{2,3,5,6\},\{4,5,7,8\}\}$$

- For vertices: $\phi_s = \mu_s$, e.g. $\log \mu_5 = \sum_{g \subseteq \{5\}} \log \phi_g = \log \phi_5$
- Pairwise functions: e.g.
  $\log \mu_{25} = \sum_{g \subseteq \{2,5\}} \log \phi_g = \log \mu_5 + \log \phi_{25} \Rightarrow \phi_{25} = \frac{\mu_{25}}{\mu_5}$
- Recurring over hyperedge size: $\phi_{1245} = \frac{\mu_{1245}\mu_5}{\mu_{25}\mu_{45}}$
- Overall:

$$p_\mu(x)=\prod_{h\in E} \phi_h(x_h) = \frac{\mu_{1245}\mu_{2356}\mu_{4578}}{\mu_{25}\mu_{45}} = \frac{\prod_{c\in C} \mu_c(x_c)}{\prod_{s\in S}[\mu_s(x_s)]^{d(S)-1}}=(2.12)$$

4.2.2 (p102-103)

Entropy Decomposition

- From the hyperedge factorization of the joint $p_\mu(x)$ (4.42) follows a **local decomposition of the entropy**. To see this we define:
- **Hyperedge Entropy:** $H_h(\mu_h) = -\sum_{x_h} \mu_h(x_h) \log \mu_h(x_h)$
- **Multi-information:** $I_h(\mu_h) = \sum_{x_h} \mu_h(x_h) \log \phi_h(x_h)$
- So, again by (4.42), on hypertrees:

  $$H_{hypertree}(\mu) = -\sum_{h \in E} I_h(\mu_h) \qquad (4.45)$$
- Alternatively: $H_{hypertree}(\mu) = \sum_{h \in E} c(h) H_h(\mu_h) \qquad (4.47)$

where: $\quad c(h) = \sum_{\{e|h \subseteq e\}} \omega(h, e) \quad$ the "overcounting numbers"

For trees: $c(\{s\}) = d(s) - 1$ and $c(\{s, t\}) = 1$, giving us the reformulation of the Bethe entropy (4.15)

4.2.3 (p104-105) I

Kikuchi and Related Approximations

- In section 4.1 we formed tree-based approximations (both on entropy and marginal polytope)
- Now we form *hypertree*-based approximations
- Let $\tau = \{\tau_{h \in E}\}$ be a collection of hyperedge local marginals:

$$H_{app}(\tau) = \sum_{h \in E} c(h) H_h(\tau_h) \Longrightarrow \text{like Bethe, exact for (hyper)-trees}$$

$\mathbb{L}_t(G) = \text{set of pseudomarginals:}$

$$\left\{ \tau \geq 0 \mid \sum_{x'_h} \tau_h(x'_h) = 1, \text{ and } \sum_{\{x'_h \mid x'_g = x_g\}} \tau_h(x'_h) = \tau_g(x_g);\ \forall h, g \subset h \right\}$$

4.2.3 (p104-105) II

Kikuchi and Related Approximations

- Again, the set of pseudomarginals $\mathbb{L}_t(G)$ outer bounds the corresponding set of globally valid marginals $\mathbb{M}_t(G)$
  - the subscript t is the treewidth (i.e. the minimum width across all possible tree decompositions of G)
- The above gives us the **Hypertree Approximation of the Variational Principle**:

$$\max_{\tau \in \mathbb{L}_t(G)} \{\langle \theta, \tau \rangle + H_{app}(\tau)\} \qquad (4.53)$$

- If G is a pairwise MRF, $H_{app}(\tau) = H_{bethe}(\tau)$ (by the overcounting numbers) and $\mathbb{L}_t(G) = \mathbb{L}(G)$ since they enforce the same constraints over the same set of (hyper-)edges
- Then the above approximation becomes the Bethe Variational Problem (BVP) (4.16)

4.2.3 - Example 4.6 (p105-106) I

Kikuchi Approximation

- In figure 4.5 we use a Kikuchi clustering (a) to approximate the joint distribution of a 3×3 lattice. This produces a hypergraph (b):

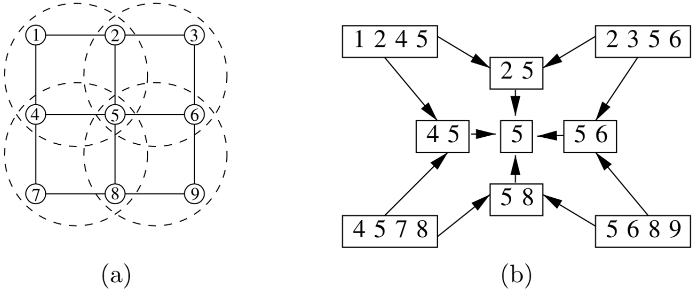


Figure : 4.5 (p106)

4.2.3 - Example 4.6 (p105-106) II

Kikuchi Approximation

- As an example we try to find the structure of the approximate entropy $H_{app} = \sum_{h \in E} c(h) H_h$ for the graph above
- We have: $c(h) = \sum_{\{e | h \subseteq e\}} \omega(h, e)$, so:
  - $c(h) = 1$ for the maximal edges $(\{1245\}, \{2356\}, \{4578\}, \{5689\})$ since they have no supersets and $\omega(g, g) = 1$
  - $c(\{25\}) = \sum_{e \in \{\{2,5\}, \{1245\}, \{2356\}\}} \omega(\{25\}, e) = 1 - 1 - 1 = -1$ (likewise for other pairwise hyperedges)
  - $c(\{5\}) = 1$ by a similar argument
- Therefore:

$$H_{app} = [H_{1245} + H_{2356} + H_{4578} + H_{5689}] - [H_{25} + H_{45} + H_{56} + H_{58}] + H_5$$

4.2.4 (p106-107)

Generalized Belief Propagation

- Different methods to solve the hypertree variational problem (4.53). As for sum-product W&J choose a Lagrangian approach (Yedidia [269]). In particular messages are passed from "parents" to "children"
- Define:
  - Ancestors: $\mathcal{A}(h) = \{g \in E \mid h \subset g\}$, $\mathcal{A}^+(h) = \mathcal{A}(h) \cup h$
  - Descendants: $\mathcal{D}(h) = \{g \in E \mid g \subset h\}$, $\mathcal{D}^+(h) = \mathcal{D}(h) \cup h$
- A message $M_{f \to g}(x_g)$ from hyperedge f to g is a functions over the state space of x_g

$$\tau_h(x_h) \propto \left[\prod_{g \in \mathcal{D}^+(h)} \exp(\theta(x_g)) \right] \left[\prod_{g \in \mathcal{D}^+(h)} \prod_{f \in Par(g) \setminus \mathcal{D}^+(h)} M_{f \to g}(x_g) \right]$$

4.1 and 4.2

Comparing Salient Formulae

- Entropies:

$$H_{app}(\tau) = \sum_{h \in E} c(h) H_h(\tau_h)$$

$$H_{Bethe}(\tau) = \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) =$$

$$-\sum_{s \in V} (d_s - 1) H_s(\tau_s) + \sum_{(s,t) \in E} H_{st}(\tau_{st})$$

- Messages:

$$\tau_h(x_h) \propto \left[\prod_{g \in \mathcal{D}^+(h)} \exp(\theta(x_g)) \right] \left[\prod_{g \in \mathcal{D}^+(h)} \prod_{f \in Par(g) \setminus \mathcal{D}^+(h)} M_{f \to g}(x_g) \right]$$

$$M_{ts}(x_s) \propto \sum_{x_t} \left[\exp(\theta_{st}(x_s, x_t) + \theta_t(x_t)) \prod_{u \in N(t) \setminus s} M_{ut}(x_t) \right]$$

4.2.4 - Example 4.7 (p108)

Parent to Child for Kikuchi

- Consider Kikuchi clustering of 3×3 lattice:

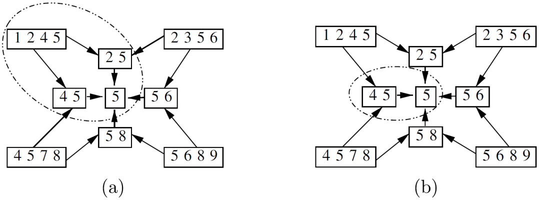


Figure : 4.6 (p109)

(a) $\tau_{1245} \propto$
$\psi_{1245}\psi_{25}\psi_{45}\psi_{5} \times M_{(2356)\to(25)} M_{(4578)\to(45)} M_{(56)\to(5)} M_{(58)\to(5)}$

(b) $\tau_{45} \propto$
$\psi_{45}\psi_{5} \times M_{(1245)\to(45)} M_{(4578)\to(45)} M_{(25)\to(5)} M_{(56)\to(5)} M_{(58)\to(5)}$

Appendix D (p280-285)

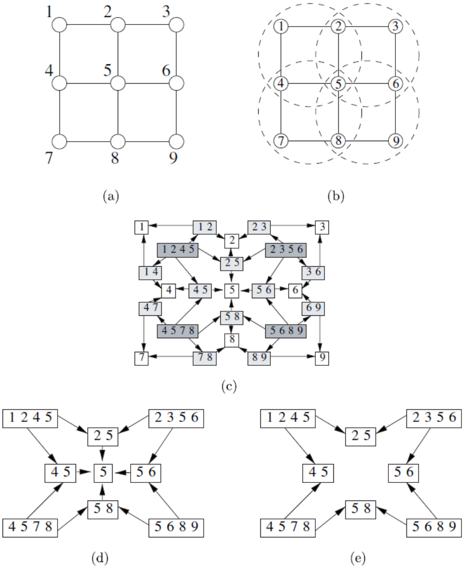


Figure : D.1 (p282)

Expectation Propagation Algorithms

Entropy Approximations Based on Term Decoupling (p111)

- $(X_1, ..., X_m) \in \mathbb{R}^m$
- $\underbrace{\phi = (\phi_1, ..., \phi_{d_T})}_{\textit{Tractable}}$ and $\underbrace{\Phi = (\Phi^1, ..., \Phi^{d_I})}_{\textit{Intractable}}$ sufficient statistics

Entropy Approximations Based on Term Decoupling (p111)

- $(X_1, \ldots, X_m) \in \mathbb{R}^m$
- $\underbrace{\phi = (\phi_1, \ldots, \phi_{d_T})}_{\textit{Tractable}}$ and $\underbrace{\Phi = (\Phi^1, \ldots, \Phi^{d_I})}_{\textit{Intractable}}$ sufficient statistics

The $(\phi, \Phi)-$Exponential Family

- parameters $\theta, \tilde{\theta} \leftrightarrow \phi, \Phi$
- $p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp\left(\langle \theta, \phi(x) \rangle\right) \exp\left(\langle \tilde{\theta}, \Phi(x) \rangle\right)$
- base model $p(x; \theta, \overrightarrow{0}) \propto f_0(x) \exp\left(\langle \theta, \phi(x) \rangle\right)$
  (no intractable component)

Entropy Approximations Based on Term Decoupling (p111)

- $(X_1, \ldots, X_m) \in \mathbb{R}^m$
- $\underbrace{\phi = (\phi_1, \ldots, \phi_{d_T})}_{\textit{Tractable}}$ and $\underbrace{\Phi = (\Phi^1, \ldots, \Phi^{d_I})}_{\textit{Intractable}}$ sufficient statistics

The $(\phi, \Phi)-$Exponential Family

- parameters $\theta, \tilde{\theta} \leftrightarrow \phi, \Phi$
- $p(x; \theta, \tilde{\theta}) \propto f_0(x) \exp\left(\langle \theta, \phi(x) \rangle\right) \exp\left(\langle \tilde{\theta}, \Phi(x) \rangle\right)$
- base model $p(x; \theta, \overrightarrow{0}) \propto f_0(x) \exp\left(\langle \theta, \phi(x) \rangle\right)$
  (no intractable component)

The $(\phi, \Phi^i)-$Exponential Family : "Φ^i-Augmented"

- $p(x; \theta, \tilde{\theta}^i) \propto f_0(x) \exp\left(\langle \theta, \phi(x) \rangle\right) \exp\left(\langle \tilde{\theta}^i, \Phi^i(x) \rangle\right)$

Example Tractable/Intractable Partitioning (p112)

Mixture Model

- Likelihood $p(y|X = x) = (1 - \alpha)\mathcal{N}(y; 0, \sigma_0^2\mathbb{I}) + \alpha\mathcal{N}(y; x, \sigma_1^2\mathbb{I})$
- Prior $X \sim \mathcal{N}(0, \Sigma)$

Example Tractable/Intractable Partitioning (p112)

Mixture Model

- Likelihood $p(y|X=x) = (1-\alpha)\mathcal{N}(y;0,\sigma_0^2\mathbb{I}) + \alpha\mathcal{N}(y;x,\sigma_1^2\mathbb{I})$
- Prior $X \sim \mathcal{N}(0,\Sigma)$
- Posterior

$$
\begin{aligned}
p(x|y^1\ldots,y^n) &\propto \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)\prod_i p(y^i|X=x) \\
&\propto \underbrace{\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)}_{Tractable=base}\underbrace{\exp\left\{\sum_i \log\ p(y^i|X=x)\right\}}_{Intractable,\ d_I=|\mathcal{Y}|}
\end{aligned}
$$

Example Tractable/Intractable Partitioning (p112)

Mixture Model

- Likelihood $p(y|X=x) = (1-\alpha)\mathcal{N}(y;0,\sigma_0^2\mathbb{I}) + \alpha\mathcal{N}(y;x,\sigma_1^2\mathbb{I})$
- Prior $X \sim \mathcal{N}(0,\Sigma)$
- Posterior

$$
\begin{aligned}
p(x|y^1\ldots,y^n) &\propto \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)\prod_i p(y^i|X=x) \\
&\propto \underbrace{\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)}_{Tractable=base}\underbrace{\exp\left\{\sum_i \log\ p(y^i|X=x)\right\}}_{Intractable,\, d_I=|\mathcal{Y}|}
\end{aligned}
$$

"Φ^i-Augmented" corresponds to having a single observation and is a tractable case (2 components, otherwise $2^{|\mathcal{Y}|}$)

"Φ^i-Augmented", tractable

In the $(\phi, \Phi^i)-$Exponential Family

- Likelihood tractable
- Entropy tractable

“Φ^i-Augmented”, tractable

In the $(\phi, \Phi^i)-$Exponential Family

- Likelihood tractable
- Entropy tractable

In what follows, use these 1-augmented families to

- approximate $\mathbb{M}(G)$
- approximate the entropy

Acceptable means for the $(\phi, \Phi)-$ExpFam (pp. 113-114)

Notation

- $\mu = \mathbb{E}[\phi(x)]$, $\tilde{\mu} = \mathbb{E}[\Phi(x)]$
- $\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu}) \,|\, (\mu, \tilde{\mu}) = \mathbb{E}\left[(\phi(x), \Phi(x))\right] \text{ for some } p\}$
- Same for base (Φ empty) or "Φ^i-Augmented"

Acceptable means for the $(\phi, \Phi)-$ExpFam (pp. 113-114)

Notation

- $\mu = \mathbb{E}[\phi(x)]$, $\tilde{\mu} = \mathbb{E}[\Phi(x)]$
- $\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu}) \,|\, (\mu, \tilde{\mu}) = \mathbb{E}\left[(\phi(x), \Phi(x))\right] \text{ for some } p\}$
- Same for base (Φ empty) or "Φ^i-Augmented"

Projection operator ('cropping') on acceptable means

- acceptable mean: $(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi)$
- projection $(\tau, \tilde{\tau}) \stackrel{\Pi^i}{\rightarrow} (\tau, \tilde{\tau}^i)$

Acceptable means for the $(\phi, \Phi)-$ExpFam (pp. 113-114)

Notation

- $\mu = \mathbb{E}[\phi(x)]$, $\tilde{\mu} = \mathbb{E}[\Phi(x)]$
- $\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu}) \,|\, (\mu, \tilde{\mu}) = \mathbb{E}\left[(\phi(x), \Phi(x))\right] \text{ for some } p\}$
- Same for base (Φ empty) or "Φ^i-Augmented"

Projection operator ('cropping') on acceptable means

- acceptable mean: $(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi)$
- projection $(\tau, \tilde{\tau}) \xrightarrow{\Pi^i} (\tau, \tilde{\tau}^i)$

Approximating $\mathcal{M}(\phi, \Phi)$

$$\begin{aligned}\mathcal{L}(\phi, \Phi) &= \left\{(\tau, \tilde{\tau}) \,|\tau \in \mathcal{M}(\phi), \quad \Pi^i\,(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i) \quad \forall i = 1, ..., d_I\right\} \\ &= \cap_i \left\{(\tau, \tilde{\tau}) \,|\tau \in \mathcal{M}(\phi), \quad \Pi^i\,(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i)\right\}\end{aligned}$$

Acceptable means for the $(\phi, \Phi)-$ExpFam (pp. 113-114)

Notation

- $\mu = \mathbb{E}[\phi(x)]$, $\tilde{\mu} = \mathbb{E}[\Phi(x)]$
- $\mathcal{M}(\phi, \Phi) = \{(\mu, \tilde{\mu}) \,|\, (\mu, \tilde{\mu}) = \mathbb{E}\left[(\phi(x), \Phi(x))\right] \text{ for some } p\}$
- Same for base (Φ empty) or "Φ^i-Augmented"

Projection operator ('cropping') on acceptable means

- acceptable mean: $(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi)$
- projection $(\tau, \tilde{\tau}) \stackrel{\Pi^i}{\rightarrow} (\tau, \tilde{\tau}^i)$

Approximating $\mathcal{M}(\phi, \Phi)$

$$\begin{aligned}
\mathcal{L}(\phi, \Phi) &= \left\{(\tau, \tilde{\tau}) \,|\tau \in \mathcal{M}(\phi), \quad \Pi^i(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i) \quad \forall i = 1, ..., d_I\right\} \\
&= \cap_i \left\{(\tau, \tilde{\tau}) \,|\tau \in \mathcal{M}(\phi), \quad \Pi^i(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i)\right\}
\end{aligned}$$

Remark:

- intersection of convex sets
- $\mathcal{M}(\phi, \Phi) \subseteq \mathcal{L}(\phi, \Phi)$

Approximating $\mathcal{M}$ and $H(\tau, \tilde{\tau})$ (pp. 114-115)

Approximating $\mathcal{M}$

$$\mathcal{L}(\phi, \Phi) = \left\{(\tau, \tilde{\tau}) \,|\, \tau \in \mathcal{M}(\phi), \quad \Pi^{i}(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^{i}) \quad \forall i = 1, ..., d_I\right\}$$

Approximating $\mathcal{M}$ and $H(\tau, \tilde{\tau})$ (pp. 114-115)

Approximating $\mathcal{M}$

$$\mathcal{L}(\phi, \Phi) \quad = \quad \left\{(\tau, \tilde{\tau})\,|\tau \in \mathcal{M}(\phi), \quad \Pi^{i}\,(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^{i}) \quad \forall i = 1, ..., d_I\right\}$$

Approximating $H(\tau, \tilde{\tau})$

- $H(\tau, \tilde{\tau})$ is not tractable , but $H(\tau, \tilde{\tau}^{I})$ tractable

$$\begin{aligned} H_{ep}(\tau, \tilde{\tau}) &= H(\tau) + \sum_{I} \left[H(\tau, \tilde{\tau}^{I}) - H(\tau)\right] \\ &= \sum_{I=1}^{d_I} H(\tau, \tilde{\tau}^{I}) - (d_I - 1)\,H(\tau) \end{aligned}$$

Approximating $\mathcal{M}$ and $H(\tau, \tilde{\tau})$ (pp. 114-115)

Approximating $\mathcal{M}$

$$\mathcal{L}(\phi, \Phi) = \left\{(\tau, \tilde{\tau}) \,|\, \tau \in \mathcal{M}(\phi), \quad \Pi^{i}(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^{i}) \quad \forall i = 1, \ldots, d_I\right\}$$

Approximating $H(\tau, \tilde{\tau})$

- $H(\tau, \tilde{\tau})$ is not tractable , but $H(\tau, \tilde{\tau}^{I})$ tractable

$$\begin{aligned} H_{ep}(\tau, \tilde{\tau}) &= H(\tau) + \sum_{I} \left[H(\tau, \tilde{\tau}^{I}) - H(\tau)\right] \\ &= \sum_{I=1}^{d_I} H(\tau, \tilde{\tau}^{I}) - (d_I - 1)\, H(\tau) \end{aligned}$$

Final optimization problem

$$\max_{(\tau,\tau') \in \mathcal{L}(\phi,\Phi)} \left\{\langle \tau, \theta \rangle + \langle \tilde{\tau}, \tilde{\theta} \rangle + H_{ep}(\tau, \tau')\right\}, \text{ eq. } (4.69)$$

Example 4.9 - Sum-Product and Bethe Approximation

Pairwise Markov random field on Graph $G = (V, E)$

- base: $p(x; \theta, \overrightarrow{0}) \propto \prod_{s \in V} \exp\left(\theta_s(x_s)\right)$
- $\Phi^{uv}-$augmented (one edge!):
  $p(x; \theta, \tilde{\theta}^{uv}) \propto \left[\prod_{s \in V} \exp\left(\theta_s(x_s)\right)\right] \exp\left(\tilde{\theta}^{uv}(x_u, x_v)\right)$

Example 4.9 - Sum-Product and Bethe Approximation

Pairwise Markov random field on Graph $G = (V, E)$

- base: $p(x; \theta, \overrightarrow{0}) \propto \prod_{s \in V} \exp\left(\theta_s(x_s)\right)$
- Φ^{uv}−augmented (one edge!):
  $p(x; \theta, \tilde{\theta}^{uv}) \propto \left[\prod_{s \in V} \exp\left(\theta_s(x_s)\right)\right] \exp\left(\tilde{\theta}^{uv}(x_u, x_v)\right)$

Calulating entropies (for a parameterization through means)

- $H(\tau_1 ... \tau_m) = \sum_{s \in V} H(\tau_s)$
- $H(\tau_1 ... \tau_m, \tau_{uv}) = \sum_{s \in V} H(\tau_s) + \underbrace{[H(\tau_{uv}) - H(\tau_u) - H(\tau_v)]}_{-I(\tau_{uv})} =$

  $H_{ep}(\tau_1 ... \tau_m, \tau_{uv})$

Example 4.9 - Sum-Product and Bethe Approximation

Pairwise Markov random field on Graph $G = (V, E)$

$$\begin{aligned}\mathcal{L}(\phi, \Phi) &= \left\{ (\tau, \tilde{\tau}) \,|\, \underbrace{\tau \in \mathcal{M}(\phi)}_{\text{normalization}}, \underbrace{(\tau, \tau_{uv}) \in \mathcal{M}(\phi, \Phi^{uv})}_{\text{marginalization}}, \forall (u, v) \in E \right\} \\ &= \mathbb{L}(G)\end{aligned}$$

4.3.2 Optimality in terms of Moment-Matching

Recall

$$\mathcal{L}(\phi, \Phi) \quad = \quad \cap_i \left\{ (\tau, \tilde{\tau}) \, | \tau \in \mathcal{M}(\phi), \quad \Pi^i \, (\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i) \right\}$$

4.3.2 Optimality in terms of Moment-Matching

Recall

$$\mathcal{L}(\phi, \Phi) \quad = \quad \cap_i \left\{ (\tau, \tilde{\tau}) \,|\tau \in \mathcal{M}(\phi), \quad \Pi^i\,(\tau, \tilde{\tau}) \in \mathcal{M}(\phi, \Phi^i) \right\}$$

Another construction

- 1-Expand (and decouple)

$$\{\tau \in \mathcal{M}(\phi)\} \otimes_i \left\{ \left(\eta^i, \tilde{\tau}^i\right) |\Pi^i \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i) \right\}$$

4.3.2 Optimality in terms of Moment-Matching

Recall

$$\mathcal{L}(\phi,\Phi) \quad = \quad \cap_i \left\{(\tau,\tilde{\tau})\,|\tau \in \mathcal{M}(\phi), \quad \Pi^i\,(\tau,\tilde{\tau}) \in \mathcal{M}(\phi,\Phi^i)\right\}$$

Another construction

- 1-Expand (and decouple)
  $$\{\tau \in \mathcal{M}(\phi)\} \otimes_i \left\{\left(\eta^i,\tilde{\tau}^i\right)|\Pi^i\left(\eta^i,\tilde{\tau}^i\right) \in \mathcal{M}(\phi,\Phi^i)\right\}$$

Expansion from $(\tau,\tilde{\tau}) \to \left\{\tau,(\eta^i,\tilde{\tau}^i), i = 1..d_l\right\}$

4.3.2 Optimality in terms of Moment-Matching

Recall

$$\mathcal{L}(\phi,\Phi) \quad = \quad \cap_i \left\{ (\tau,\tilde{\tau}) \,|\tau \in \mathcal{M}(\phi), \quad \Pi^i\,(\tau,\tilde{\tau}) \in \mathcal{M}(\phi,\Phi^i) \right\}$$

Another construction

- 1-Expand (and decouple)

$$\{\tau \in \mathcal{M}(\phi)\} \otimes_i \left\{ \left(\eta^i,\tilde{\tau}^i\right) |\Pi^i \left(\eta^i,\tilde{\tau}^i\right) \in \mathcal{M}(\phi,\Phi^i) \right\}$$

Expansion from $(\tau,\tilde{\tau}) \rightarrow \left\{ \tau, (\eta^i,\tilde{\tau}^i), i = 1..d_I \right\}$

- 2-Couple back

$$\{\tau \in \mathcal{M}(\phi)\} \otimes_i \left\{ \left(\eta^i,\tilde{\tau}^i\right) |\Pi^i \left(\eta^i,\tilde{\tau}^i\right) \in \mathcal{M}(\phi,\Phi^i) \right\} \text{ and}$$
$$\forall i,j \quad (\tau_i,\tilde{\tau}_i) = (\tau_j,\tilde{\tau}_j)$$

No secret here, just more variables, coupled together.

4.3.2 Optimality in terms of Moment-Matching

Constrained optimization problem

$$\max_{\{\tau,(\eta^i,\tilde{\tau}^i)\}} \left\{ \langle \tau, \theta \rangle + \textstyle\sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + \underbrace{H(\tau) + \sum_i \left[H(\eta^i, \tilde{\tau}^i) - H(\eta^i) \right]}_{F(\tau,(\eta^i,\tilde{\tau}^i))} \right\}$$

$$\text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i)$$
$$\text{and } \tau = \eta^i$$

4.3.2 Optimality in terms of Moment-Matching

Constrained optimization problem

$$\max_{\{\tau,(\eta^i,\tilde{\tau}^i)\}} \left\{ \langle \tau, \theta \rangle + \sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) \right\}$$

$$\begin{gathered} \text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i) \\ \text{and } \tau \in \mathcal{M}(\phi) \\ \text{and } \tau = \eta^i \end{gathered}$$

Associated Partial Lagrangian

$$L(\tau; \lambda) = \langle \tau, \theta \rangle + \sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\begin{gathered} \text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i) \\ \text{and } \tau \in \mathcal{M}(\phi) \end{gathered}$$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \textstyle\sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i)$$
$$\text{and } \tau \in \mathcal{M}(\phi)$$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } (\eta^i, \tilde{\tau}^i) \in \mathcal{M}(\phi, \Phi^i)$$
$$\text{and } \tau \in \mathcal{M}(\phi)$$

For an optimal solution $\{\tau, (\eta^i, \tilde{\tau}^i), i = 1..d_l\}$

$$\begin{aligned}
\nabla_\tau L(\tau, \lambda) &= 0 \\
\nabla_{(\eta^i, \tilde{\tau}^i)} L(\tau, \lambda) &= 0, \quad \text{for } i = 1...d_l \\
\nabla_\lambda L(\tau, \lambda) &= 0 \quad \text{(constraint)}
\end{aligned}$$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i) \text{ and } \tau \in \mathcal{M}(\phi)$$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \textstyle\sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i) \text{ and } \tau \in \mathcal{M}(\phi)$$

For an optimal solution $\left\{\tau, (\eta^i, \tilde{\tau}^i), i = 1..d_I\right\}$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } \left(\eta^i, \tilde{\tau}^i\right) \in \mathcal{M}(\phi, \Phi^i) \text{ and } \tau \in \mathcal{M}(\phi)$$

For an optimal solution $\left\{\tau, (\eta^i, \tilde{\tau}^i), i = 1..d_I\right\}$

$$\nabla_\tau L(\tau, \lambda) = 0$$
$$\Rightarrow q(x;\theta,\lambda) \propto f_0(x) \exp\left\{\langle \theta + \sum_i \lambda_i, \phi(x) \rangle\right\} \in \mathcal{M}(\phi)$$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \textstyle\sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \textstyle\sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } (\eta^i, \tilde{\tau}^i) \in \mathcal{M}(\phi, \Phi^i) \text{ and } \tau \in \mathcal{M}(\phi)$$

For an optimal solution $\left\{\tau, (\eta^i, \tilde{\tau}^i), i = 1..d_I\right\}$

$\nabla_\tau L(\tau, \lambda) = 0$
$\Rightarrow q(x;\theta,\lambda) \propto f_0(x) \exp\{\langle \theta + \sum_i \lambda_i, \phi(x) \rangle\} \in \mathcal{M}(\phi)$

$\nabla_{(\eta^i, \tilde{\tau}^i)} L(\tau, \lambda) = 0$
$\Rightarrow q^i(x;\theta,\tilde{\theta}^i,\lambda) \propto f_0(x) \exp\left\{\langle \theta + \sum_{l \neq i} \lambda_i, \phi(x) \rangle + \langle \tilde{\theta}^i, \Phi^i(x) \rangle\right\} \in \mathcal{M}(\phi, \Phi^i)$

Solving the optimization problem

$$L(\tau;\lambda) = \langle \tau, \theta \rangle + \sum_i \langle \tilde{\tau}^i, \tilde{\theta}^i \rangle + F(\tau, (\eta^i, \tilde{\tau}^i)) + \sum_i \langle \lambda^i, \tau - \eta^i \rangle$$

$$\text{subject to } (\eta^i, \tilde{\tau}^i) \in \mathcal{M}(\phi, \Phi^i) \text{ and } \tau \in \mathcal{M}(\phi)$$

For an optimal solution $\{\tau, (\eta^i, \tilde{\tau}^i), i = 1..d_I\}$

$$\nabla_\tau L(\tau, \lambda) = 0$$
$$\Rightarrow q(x;\theta,\lambda) \propto f_0(x) \exp\{\langle \theta + \sum_i \lambda_i, \phi(x) \rangle\} \in \mathcal{M}(\phi)$$

$$\nabla_{(\eta^i, \tilde{\tau}^i)} L(\tau, \lambda) = 0$$
$$\Rightarrow q^i(x;\theta,\tilde{\theta}^i,\lambda) \propto f_0(x) \exp\left\{\langle \theta + \sum_{l \neq i} \lambda_i, \phi(x) \rangle + \langle \tilde{\theta}^i, \Phi^i(x) \rangle\right\} \in \mathcal{M}(\phi, \Phi^i)$$

$$\nabla_\lambda L(\tau, \lambda) = 0 \Rightarrow \tau = \mathbb{E}_q[\phi(x)] \equiv \mathbb{E}_{q^i}[\phi(x)] = \eta^i$$

EP Summary

Expectation-propagation (EP) updates:

(1) At iteration $n = 0$, initialize the Lagrange multiplier vectors $(\lambda^1, \ldots, \lambda^{d_I})$.

(2) At each iteration, $n = 1, 2, \ldots,$ choose some index $i(n) \in \{1, \ldots, d_I\}$, and

(a) Using Equation (4.78), form the augmented distribution $q^{i(n)}$ and compute the mean parameter

$$\eta^{i(n)} := \int q^{i(n)}(x)\phi(x)\nu(dx) = \mathbb{E}_{q^{i(n)}}[\phi(X)]. \quad (4.80)$$

(b) Using Equation (4.77), form the base distribution q and adjust $\lambda^{i(n)}$ to satisfy the moment-matching condition

$$\mathbb{E}_q[\phi(X)] = \eta^{i(n)}. \quad (4.81)$$

EP : Examples

Example 1:

- simple graph: (1)-(2)
- $p(x_1, x_2) \propto \exp\left(\theta_1(x_1) + \theta_2(x_2) + \underbrace{\theta(x_1, x_2)}_{intractable}\right)$

EP : Examples

Example 1:

- simple graph: (1)-(2)
- $p(x_1, x_2) \propto \exp\left(\theta_1(x_1) + \theta_2(x_2) + \underbrace{\theta(x_1, x_2)}_{intractable}\right)$

EP updates

- $q(x_1, x_2; \theta, \lambda) \propto \exp\left(\theta_1(x_1) + \lambda_{12}(x_1)\right) \exp\left(\theta_2(x_2) + \lambda_{12}(x_2)\right)$

EP : Examples

Example 1:

- simple graph: (1)-(2)
- $p(x_1, x_2) \propto \exp\left(\theta_1(x_1) + \theta_2(x_2) + \underbrace{\theta(x_1, x_2)}_{intractable}\right)$

EP updates

- $q(x_1, x_2; \theta, \lambda) \propto \exp\left(\theta_1(x_1) + \lambda_{12}(x_1)\right) \exp\left(\theta_2(x_2) + \lambda_{12}(x_2)\right)$
- $q^{12}(x_1, x_2; \theta, \lambda) \propto \exp\left(\theta_1(x_1) + \theta_2(x_2) + \theta(x_1, x_2)\right)$

EP : Examples

Example 1:

- simple graph: (1)-(2)
- $p(x_1, x_2) \propto \exp\left(\theta_1(x_1) + \theta_2(x_2) + \underbrace{\theta(x_1, x_2)}_{intractable}\right)$

EP updates

- $q(x_1, x_2; \theta, \lambda) \propto \exp(\theta_1(x_1) + \lambda_{12}(x_1)) \exp(\theta_2(x_2) + \lambda_{12}(x_2))$
- $q^{12}(x_1, x_2; \theta, \lambda) \propto \exp(\theta_1(x_1) + \theta_2(x_2) + \theta(x_1, x_2))$
- $\mathbb{E}_{q^{12}(x_1)}(\phi(x_1)) = \mathbb{E}_{q(x_1)}(\phi(x_1))$ (message passing , board)

EP : Examples

Example 2: Mixture of Gaussians

- $\mathcal{M}(\phi, \Phi) = \left\{ \mathbb{E}[X], \mathbb{E}[XX^T], \mathbb{E}\left[\log p(y^i|X)\right], \ i = 1..n \right\}$
- Lagrange multipliers $(\lambda^i, \Lambda^i) \in R^m \times R^{m \times m}$

EP : Examples

Example 2: Mixture of Gaussians

- $\mathcal{M}(\phi, \Phi) = \left\{ \mathbb{E}[X], \mathbb{E}[XX^T], \mathbb{E}\left[\log p(y^i|X)\right], \; i = 1..n \right\}$
- Lagrange multipliers $(\lambda^i, \Lambda^i) \in R^m \times R^{m \times m}$

EP updates

- $q(x, \Sigma; (\lambda^i, \Lambda^i)) \propto \exp\left\{ \langle \sum_i \lambda^i, x \rangle + \langle -\frac{1}{2}\Sigma^{-1} + \sum_i \Lambda^i, xx^T \rangle \right\}$

EP : Examples

Example 2: Mixture of Gaussians

- $\mathcal{M}(\phi, \Phi) = \left\{\mathbb{E}[X], \mathbb{E}[XX^T], \mathbb{E}\left[\log\ p(y^i|X)\right]\ ,\ i = 1..n\right\}$
- Lagrange multipliers $(\lambda^i, \Lambda^i) \in R^m \times R^{m\times m}$

EP updates

- $q(x, \Sigma; (\lambda^i, \Lambda^i)) \propto \exp\left\{\langle \sum_i \lambda^i, x\rangle + \langle -\frac{1}{2}\Sigma^{-1} + \sum_i \Lambda^i, xx^T\rangle\right\}$
- $q^i(x, \Sigma; (\lambda^i, \Lambda^i)) \propto$ $\exp\left\{\langle \sum_{l\neq i} \lambda^l, x\rangle + \langle -\frac{1}{2}\Sigma^{-1} + \sum_{l\neq i} \Lambda^l, xx^T\rangle + \langle \tilde{\theta}^i, \log\ p(y^i|x)\rangle\right\}$

That's it for today