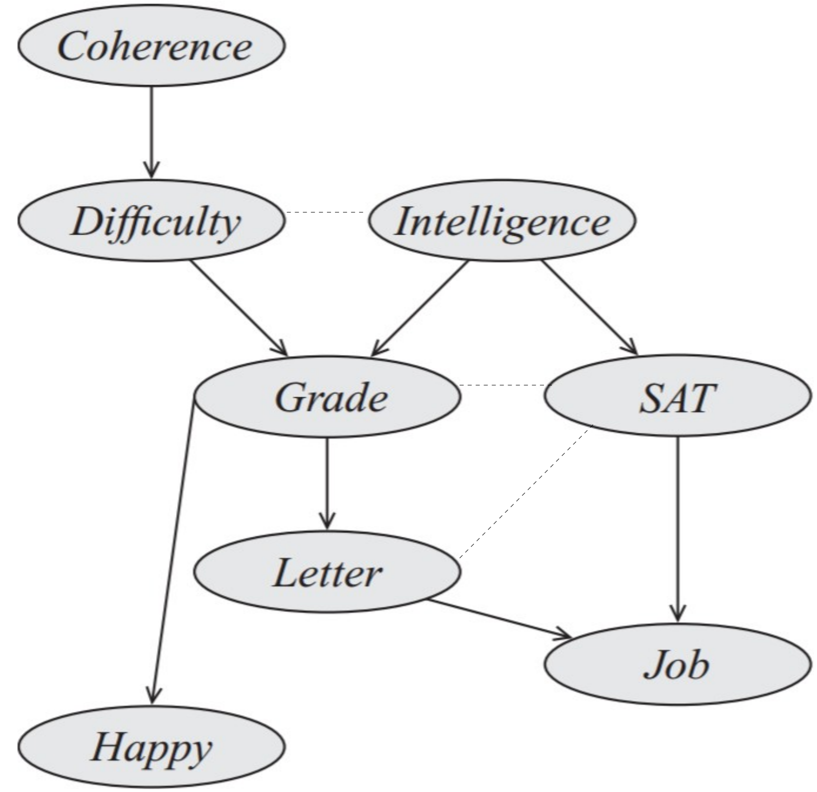
Lecture 7: Variational Inference

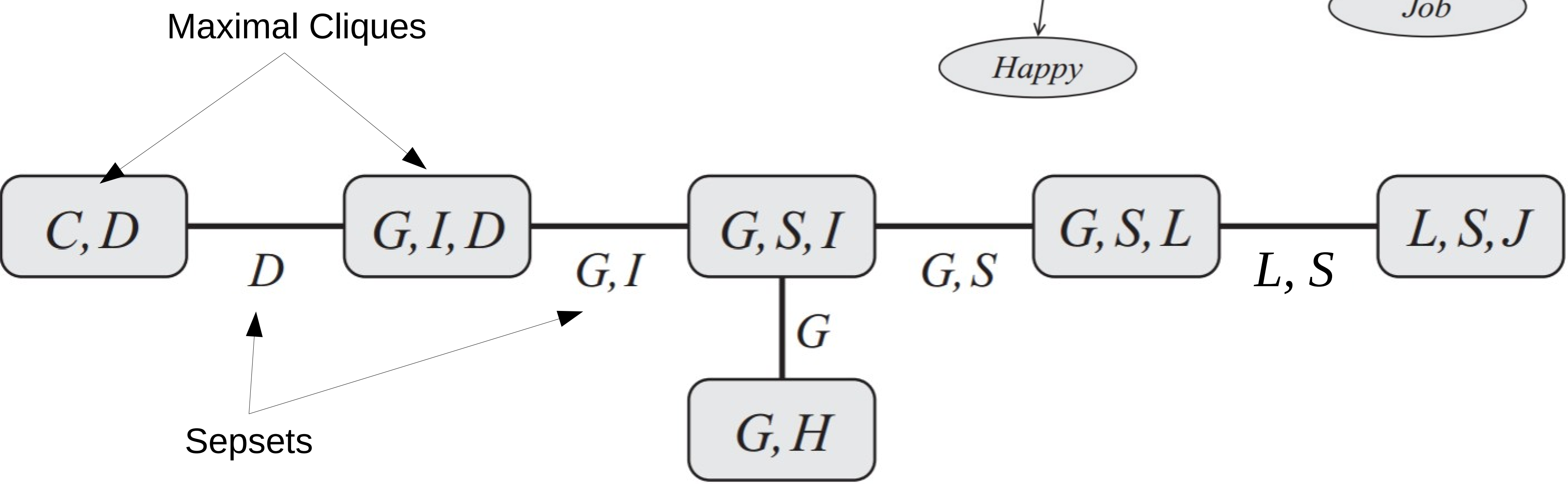
Probabilistic Graphical Models, Koller and Friedman:

- Chap 11

- Variational Methods, Junction Tree Algorithm, Loopy Belief Propagation, Lower Bounds

Clique Graphs

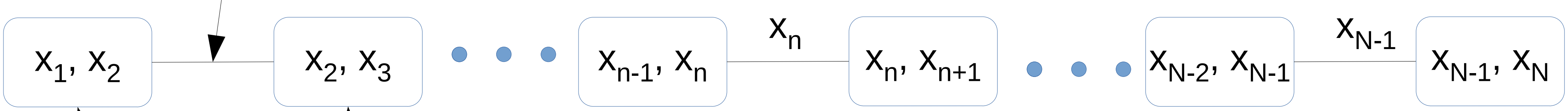




Koller and Freidman – PGM Principles and Techniques

Sum-Product Algorithm

$\delta_{1\to 2}(x_2) = \Sigma_{x1}\, \psi_1(x_1,x_2)$



$\psi_1 = \phi_1(x_1,x_2)$

$\psi_2 = \phi_2(x_2,x_3)\ \Pi\ \delta_{1\to 2}(x_2)$

Notice Product is over incoming messages and uses the original factor, ϕ_2 , even when the return message is computed. (Different when we do the belief update version)

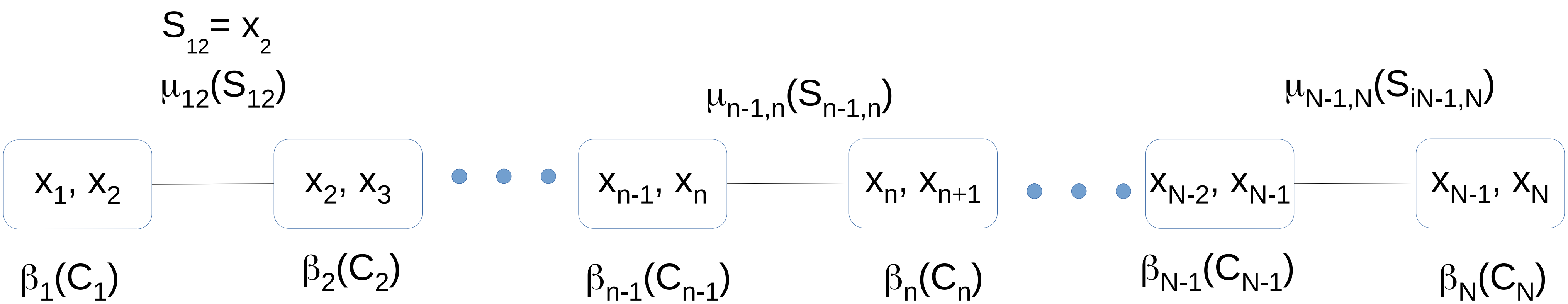
$\beta_i(C_i) = \phi_i(C_i)\ \Pi\ \delta_{j\to i}(S_{ij}) \quad = \Sigma_{X\text{-}C_i}\ \Pi_j\ \phi_j(C_j) \quad = Z\ P(C_i)$ = 'Cluster belief'

Over all Neighbors j

Product over all Clusters j

Sum over all variables not in C_i

Belief Update Message Passing



$$\beta_i(C_i) = \phi_i(C_i) \prod_j \delta_{j \rightarrow i}(S_{ij})$$

$$\mu_{ij}(S_{ij}) = \Sigma_{C_i - S_{ij}} \; \beta_i(C_i) \; = \; \Sigma_{C_j - S_{ij}} \; \beta_j(C_j)$$

$$\tilde{P}(\chi) = \frac{\prod_i \beta_i(C_i)}{\prod_{i-j} \; \mu_{ij}(S_{ij})}$$

Projections - Entropy

- Entropy:

  $H_P(\chi) = E_P[- \ln P(\chi)]$

- Relative Entropy:

  $D(P \parallel Q) = E_P[\ln (P(\chi) / Q(\chi))$

  $= -H_p(\chi) + E_P[- \ln Q(\chi)] >= 0$

- A (Kullback-Leibler) distance but not symmetric.

Projections - I and M

- I -Projection of P to Q

  $Q^{I} = \arg\min_{Q} D(Q \parallel P)$

  – Focus more on peaks in P.

- M -Projection of P to Q

  $Q^{M} = \arg\min_{Q} D(P \parallel Q)$

  – Focus more on spread of P.

  Notice Q is also being restricted to a given family

M - Projections – Q is Exponential

- $Q_\theta\,(\chi) = A(\chi)\ \exp\{ < t(\theta), \tau(\chi) > \} / Z(\theta)$
- If we find parameters, θ, such that:

  $E_{Q\theta}[\,\tau(\chi)\,] = E_P[\,\tau(\chi)\,]$

  then Q_θ is the M – Projection of P.

- This is what leads to the moment matching method and the name M projection.

I - Projections

- $Q^I = \arg\min_Q D(Q \,\|\, P)$
- This is often much easier to work with since the expectations use the simpler Q.
- Here is a simple Q:

$$Q = \frac{\prod_i \beta_i(C_i)}{\prod_{i-j} \mu_{ij}(S_{ij})}$$

- Subject to: $\mu_{ij}(S_{ij}) = \Sigma_{Ci-Sij} \; \beta_i(C_i) = \Sigma_{Cj-Sij} \; \beta_j(C_j)$

$$1 = \Sigma_{Ci} \; \beta_i(C_i)$$

Ctree-Optimize-KL

- $Q^I = arg\ min_Q\ D(Q \parallel P)$ $\qquad Q = \dfrac{\Pi_i\ \beta_i(C_i)}{\Pi_{i\text{-}j}\ \ \mu_{ij}(S_{ij})}$
- We know that we can do this optimization with message passing if the clusters form a tree. It gives 0 as the KL divergence (Relative Entropy)

The Energy Functional

$$D(Q \parallel P) = \ln Z - F(\tilde{P}(\chi), Q)$$

$\tilde{P}$ is the unnormalize P

$$F(\tilde{P}(\chi), Q) = E_Q[\ln \tilde{P}(\chi)] + H_Q(\chi)$$

$$= \sum_{\phi} E_Q[\ln \phi] - E_Q[\ln Q(\chi)]$$

The ϕ are factors of a factor graph.

- Z is the partition function (normalization) of $\tilde{P}$.
- The sum in the ‘Free Energy’ is called the energy term
- The other term is the entropy term
- If the entropy is tractable (which we can choose) and the factors only involve small subsets of the variables then we have something here we can deal with.
- We shall either minimize the Relative entropy or maximize the Energy Functional

Message Passing

$$D(Q \parallel P) = \ln Z - F(\tilde{P}(\chi), Q)$$

$$F(\tilde{P}(\chi), Q) = \sum_{\phi} E_Q[\ln \phi] - E_Q[\ln Q(\chi)]$$

The book shows how if Q is:

$$Q = \frac{\prod_i \beta_i(C_i)}{\prod_{i-j} \mu_{ij}(S_{ij})}$$

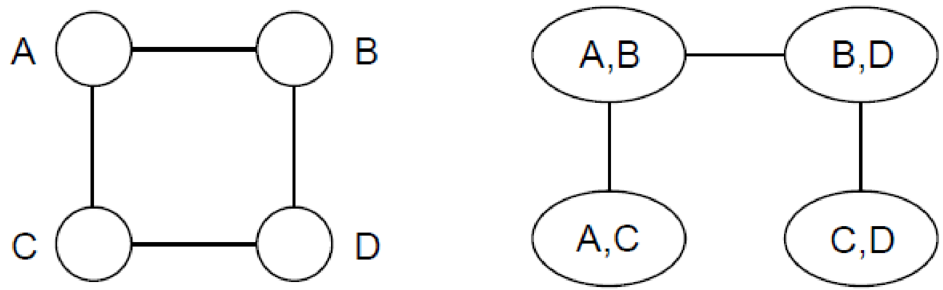
- Subject to:

$$\mu_{ij}(S_{ij}) = \Sigma_{C_i - S_{ij}} \beta_i(C_i) = \Sigma_{C_j - S_{ij}} \beta_j(C_j)$$

$$1 = \Sigma_{C_i} \beta_i(C_i)$$

Then one can derive message passing by using lagrange multipliers to solve this.

Message Passing in Clique Trees



- Note that C appears in two non-neighboring cliques.
- Question: What guarantee do we have that the probability associated with C in these two cliques will be the same?
- Answer: Nothing. In fact this is a problem with the algorithm as described so far. It is not true that in general local consistency implies global consistency.

From Jordan, PGM Intro

Loopy Belief Propagation

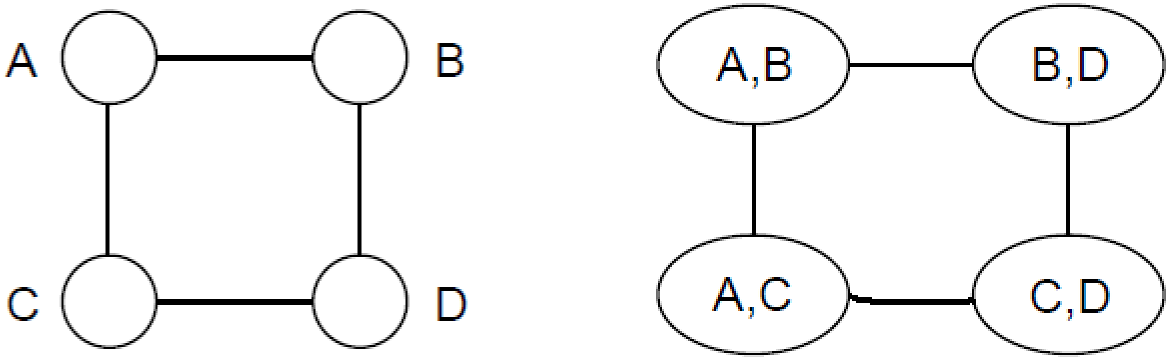
Belief update on general graphs.

Messages are passed around until convergence (not guaranteed!).

Approximate but tractable for large graphs.

Sometime works well, sometimes not at all.

Loopy Message Passing

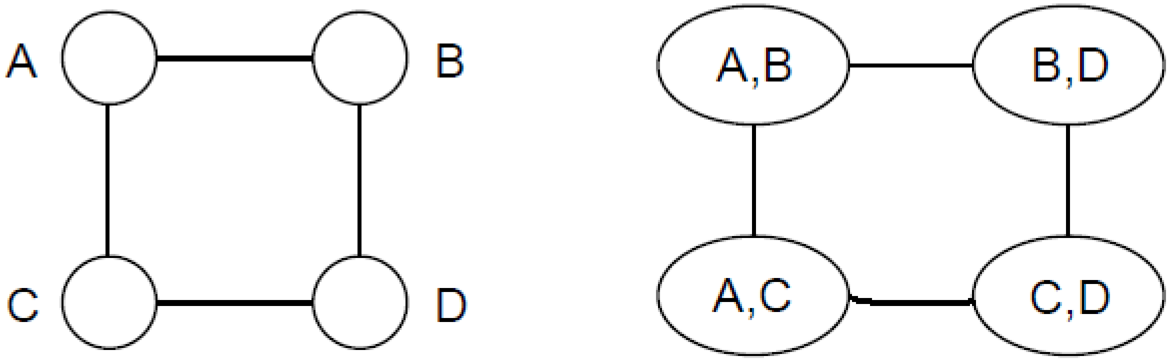


- Running intersection property is generalized to: for any two cliques with a common variable X there is **one** path on which the X is is in all the sepsets.
- In order to get this to hold we may have to define some sepsets to not be the intersection of the clusters on either side.

  Thus $\mu_{ij}(S_{ij}) = \Sigma_{Ci-Sij}\ \beta_i(C_i) = \Sigma_{Cj-Sij}\ \beta_j(C_j)$ is now a weaker constraint as there may be stuff they do not agree on.
- They will agree on the marginal of a variable but not the joint marginal of all common variables if we have running intersection

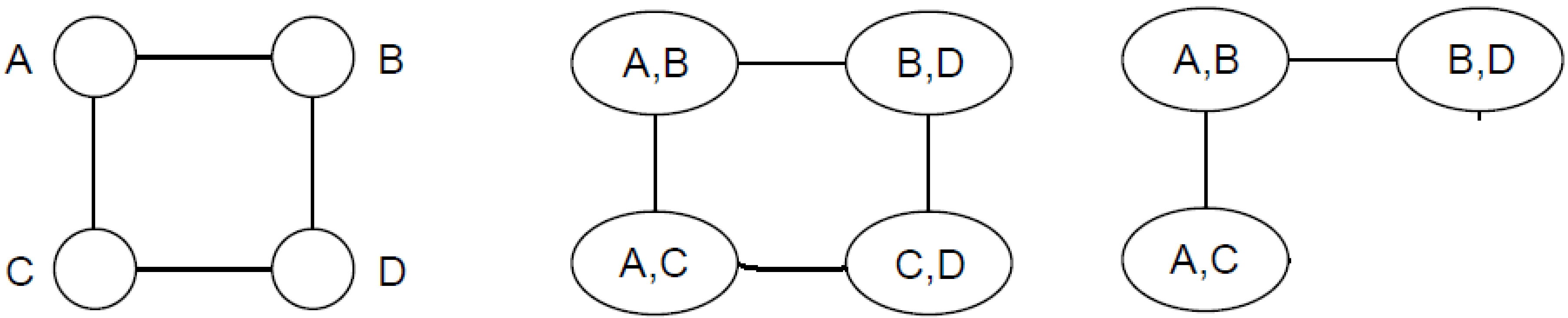
Koller and Freidman – PGM Principles and Techniques

Loopy Message Passing



- There is an issue as to where to start sending messages as there are no leaves. This is solved by essentially removing the rule of waiting until a node has full info before sending.
- So it sets all incoming messages to 1 at the start.
- This kind of inference can be much faster than exact inference if the proper tree would have huge cliques.

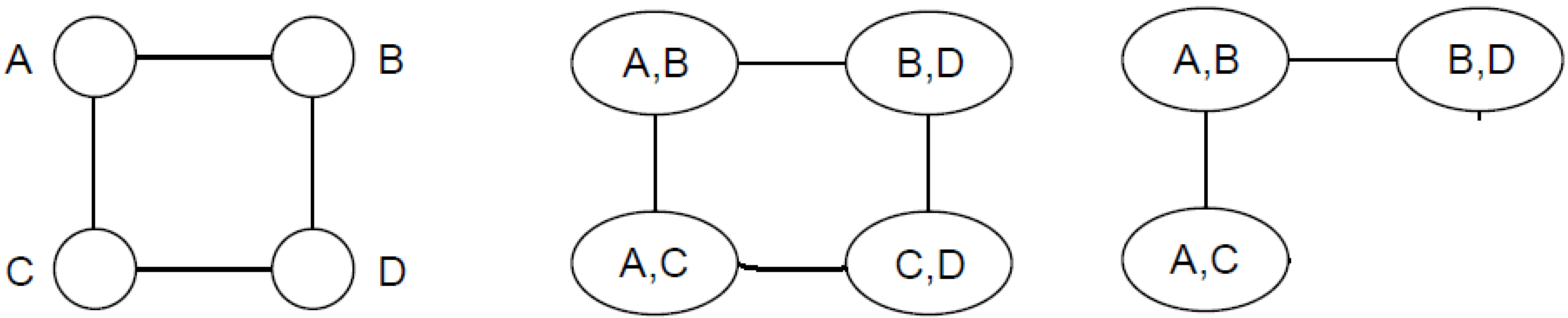
Loopy Message Passing



- $\tilde{P}(\chi) = \dfrac{\prod_i \beta_i(C_i)}{\prod_{i-j} \mu_{ij}(S_{ij})}$ is still invariant under message passing
- Imagine that we do manage to calibrate the graph (not garrenteed).
- If we focus on a subtree of the graph that happens to have the running intersection property.
- Since its a tree it defines a marginal as above for its cliques in terms of summing the invariant and furthermore the clique beliefs will be equal to marginalizing that invariant.

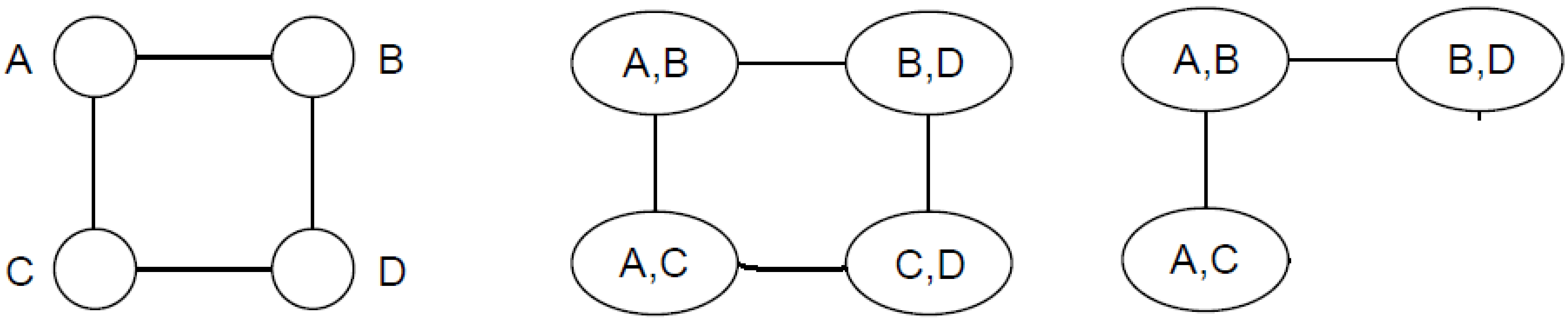
Koller and Freidman – PGM Principles and Techniques

Loopy Message Passing



- $P(A, B, C, D)=P_{tree}(A,B,C,D) \dfrac{\beta_4(C,D)}{\mu_4(C)\mu_3(D)}$
- This implies that in general
- $P(A, B) \neq P_{tree}(A,B) = \beta_2(A,B)$
- Even when the loopy belief is calibrated.
- (see 11.3.3.2 if this makes no sense)

Loopy Message Passing



- $P(A, B, C, D) = P_{tree}(A,B,C,D) \dfrac{\beta_4(C,D)}{\mu_4(C)\mu_3(D)}$
- Compare to the non-loopy case if (CD) clique had instead been (CE)
- $P(A, B, C, D, E) = P_{tree}(A,B,C,D) \dfrac{\beta_4(C,E)}{\mu_4(C)}$
- Now marginalizing E will cause the extra term to be 1.

Variational Inference

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$.

We can lower bound the marginal likelihood using Jensen's inequality:

$$\begin{aligned}
\ln p(\mathcal{D}) &= \ln \int p(\mathcal{D},\theta) d\theta = \ln \int q(\theta) \frac{P(\mathcal{D},\theta)}{q(\theta)} d\theta \\
&\geq \int q(\theta) \ln \frac{p(\mathcal{D},\theta)}{q(\theta)} d\theta = \underbrace{\int q(\theta) \ln p(\mathcal{D},\theta) d\theta + \underbrace{\int q(\theta) \ln \frac{1}{q(\theta)} d\theta}_{\text{Entropy functional}}}_{\text{Variational Lower-Bound}} \\
&= \ln p(\mathcal{D}) - \mathrm{KL}(q(\theta)||p(\theta|D)) = \mathcal{L}(q)
\end{aligned}$$

Using: $p(D,\theta) = p(\theta \mid D)\, p(D)$

where $\mathrm{KL}(q||p)$ is a Kullback–Leibler divergence. It is a non-symmetric measure of the difference between two probability distributions q and p.

The goal of variational inference is to maximize the variational lower-bound w.r.t. approximate q distribution, or minimize $\mathrm{KL}(q||p)$.

Figure: Ruslan Salakhutdinov, BCS and CSAIL, MIT

Variational Inference

Key Idea: Approximate intractable distribution $p(\theta|D)$ with simpler, tractable distribution $q(\theta)$ by minimizing $\mathrm{KL}(q(\theta)||p(\theta|D))$.

We can choose a fully factorized distribution: $q(\theta) = \prod_{i=1}^{D} q_i(\theta_i)$, also known as a mean-field approximation.

The variational lower-bound takes form:

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\theta) \ln p(\mathcal{D}, \theta) d\theta + \int q(\theta) \ln \frac{1}{q(\theta)} d\theta \\
&= \int q_j(\theta_j) \underbrace{\left[\ln p(\mathcal{D}, \theta) \prod_{i \neq j} q_i(\theta_i) d\theta_i \right]}_{\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]} d\theta_j + \sum_i \int q_i(\theta_i) \ln \frac{1}{q(\theta_i)} d\theta_i
\end{aligned}
$$

Suppose we keep $\{q_{i \neq j}\}$ fixed and maximize $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$.

Figure: Ruslan Salakhutdinov, BCS and CSAIL, MIT

$$\int q_j(\theta_j) \underbrace{\left[\ln p(\mathcal{D}, \theta) \prod_{i \neq j} q_i(\theta_i) d\theta_i \right]}_{\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]} d\theta_j + \sum_i \int q_i(\theta_i) \ln \frac{1}{q(\theta_i)} d\theta_i$$

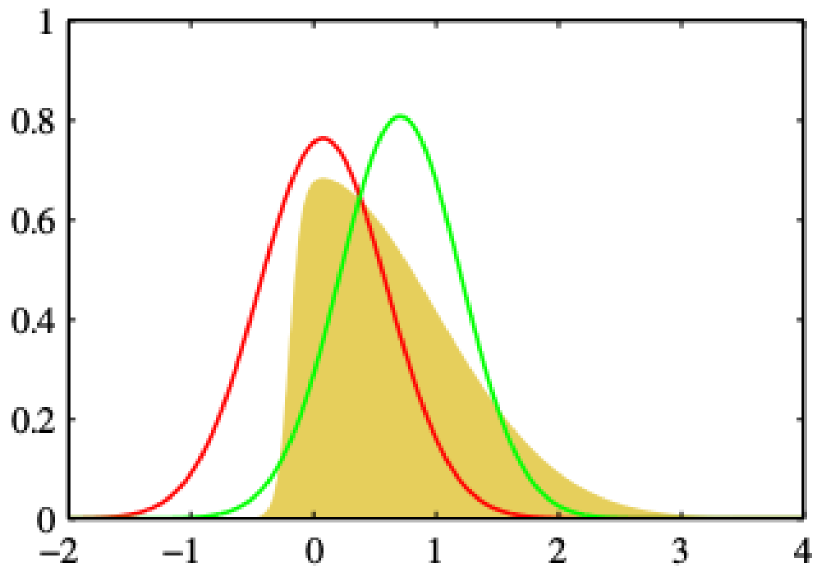
So take the functional derivative wrt q_j and set everything left under the integral to 0.

0 = $\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]$ - $ln\ q_j$ – 1

$ln\ q_j$ = $\mathbb{E}_{i \neq j}[\ln p(\mathcal{D}, \theta)]$ – 1

Then exponentiate both sides (and normalize)

Variational Approximation



The plot shows the original distribution (yellow), along with the Laplace (red) and variational (green) approximations.

By maximizing $\mathcal{L}(q)$ w.r.t. all possible forms for the distribution $q_j(\theta_j)$ we obtain a general expression:

$$q_j^*(\theta_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\ln p(\mathcal{D}, \theta)])}{\int \exp(\mathbb{E}_{i\neq j}[\ln p(\mathcal{D}, \theta)])d\theta_j}$$

Iterative Procedure: Initialize all q_j and then iterate through the factors replacing each in turn with a revised estimate.

Convergence is guaranteed as the bound is convex w.r.t. each of the factors q_j (see Bishop, chapter 10).

Figure: Ruslan Salakhutdinov, BCS and CSAIL, MIT

Tutorial 8 Variational Inference
Olga Mikheeva

Excellent concise description of theory and Example of using mean field coordinate ascent to solve a GMM.

- We want to find: $p(\boldsymbol{z}|\boldsymbol{x}) = \dfrac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})}$

(z are the data and x the latent variables)

- But: $$p(\boldsymbol{x}) = \int p(\boldsymbol{z}, \boldsymbol{x}) d\boldsymbol{z}.$$
- is intractable.

Lets find an I-Projection

$$q^*(z) = argmin_{q(z)\in\mathcal{Q}} KL(q(z)||p(z|\boldsymbol{x}))$$

$$\begin{aligned} KL(q(z)||p(z|\boldsymbol{x})) &= E_{q(z)}[\log q(z)] - E_{q(z)}[\log p(z|\boldsymbol{x})] \\ &= E_{q(z)}[\log q(z)] - E_{q(z)}[\log p(z, \boldsymbol{x})] + \log p(\boldsymbol{x}) \end{aligned}$$

- Hard term is there again but now without any q.
- Evidence Lower Bound:

$$\begin{aligned} ELBO(q) &= E_{q(z)}[\log p(z, \boldsymbol{x})] - E_{q(z)}[\log q(z)] \\ &= E_{q(z)}[\log p(z)] + E_{q(z)}[\log p(\boldsymbol{x}|z)] - E_{q(z)}[\log q(z)] \end{aligned}$$

- Match prior + match data + reduce spread

Lets find an I-Projection

$$q^*(\boldsymbol{z}) = argmin_{q(\boldsymbol{z})\in\mathcal{Q}} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

$$\begin{aligned} KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}|\boldsymbol{x})] \\ &= E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z},\boldsymbol{x})] + \log p(\boldsymbol{x}) \end{aligned}$$

$$\begin{aligned} ELBO(q) &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z},\boldsymbol{x})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\ &= E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z})] + E_{q(\boldsymbol{z})}[\log p(\boldsymbol{x}|\boldsymbol{z})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \end{aligned}$$

$$\log p(\boldsymbol{x}) = ELBO(q) + KL(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$$

Mean Field Approximation

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_i(z_i)$$

$$ELBO(q) = E_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - E_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})]$$

- Coordinate ascent says take one term, j, to maximize by plugging in and setting the variational derivative to 0:

$$q_j^*(z_j) \propto \exp\{E_{-j}[\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})]\}$$

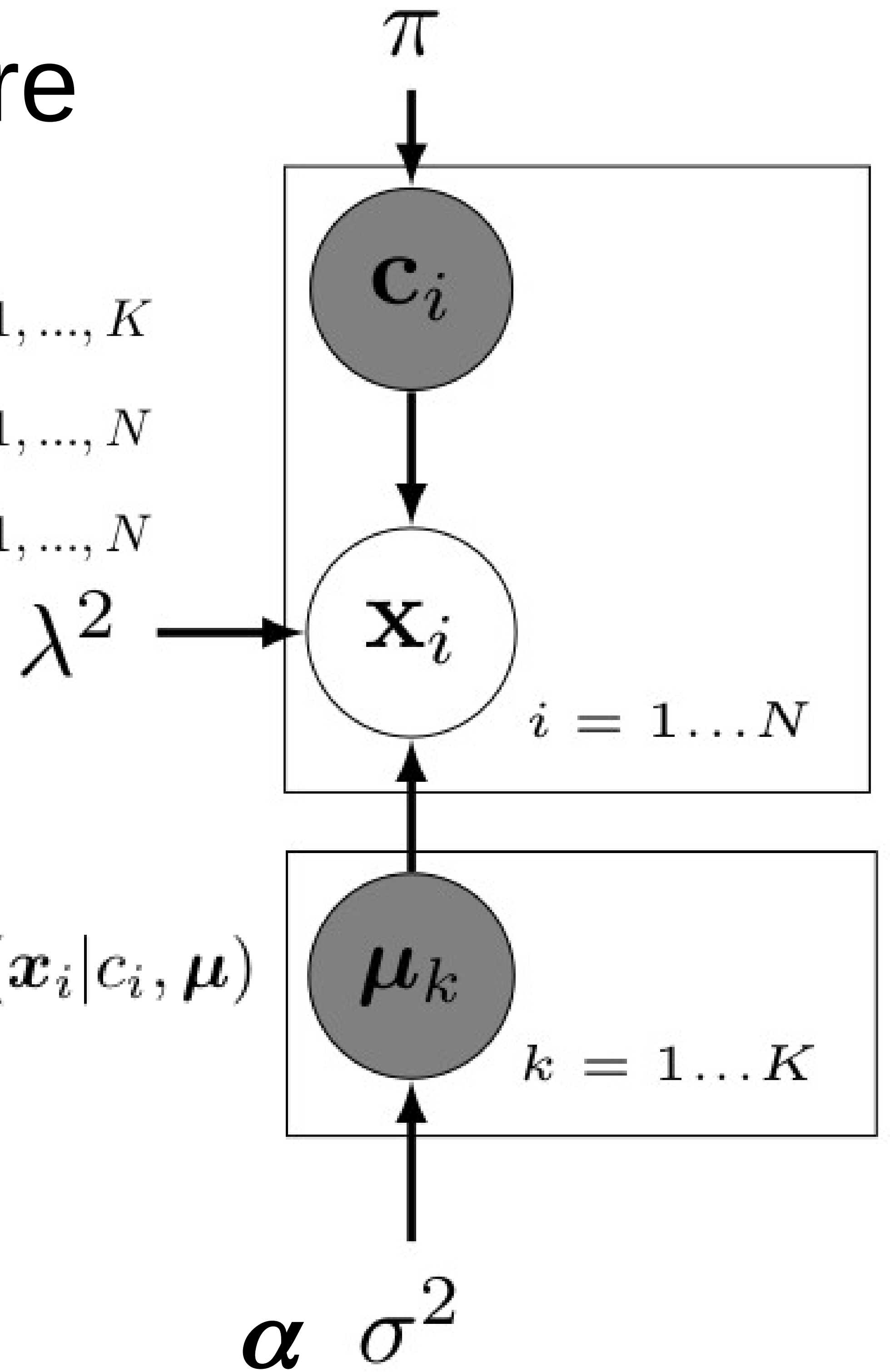
Gaussian Mixture Model

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\alpha}, \sigma^2 \boldsymbol{I}) \qquad k = 1, ..., K$$

$$c_i \sim Categorical\Big(\frac{1}{K}, ..., \frac{1}{K}\Big) \qquad i = 1, ..., N$$

$$\boldsymbol{x}_i | c_i, \boldsymbol{\mu} \sim \mathcal{N}(c_i^T \boldsymbol{\mu}, \lambda^2 \boldsymbol{I}) \qquad i = 1, ..., N$$

The dim of x is p.



$$p(\boldsymbol{\mu}, \mathbf{c}, \boldsymbol{x}) = \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) \prod_{i=1}^{N} p(c_i) p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})$$

Categorical dist. = multinomial dist.
π = (1/K,...1/K)

Mean Field

$$p(\boldsymbol{\mu}, \boldsymbol{c}) \approx q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{i=1}^{N} q(c_i)$$

$$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \boldsymbol{m}_k, s_k^2 \boldsymbol{I})$$

$$q(c_i) \sim Categorical(\boldsymbol{\phi}_i)$$

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

Lets take the first term: $\sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)]$

ELBO

Alert: remember this slide exists if you are dong the tutorial as you are asked to do this.

The dim of x is p.

$$\sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)]$$

$$= \sum_{k=1}^{K} -\tfrac{1}{2}\left[p \log(2\pi\sigma^2) + \int \boldsymbol{d\mu}\; N(\mu_k - m_k, s_k^2 I)\,(\mu_k - \alpha)^2 / \sigma^2\right]$$

$$= -\tfrac{1}{2}\left[Kp \log(2\pi\sigma^2) + \sum_{k=1}^{K} \left(m_k^2 + p\boldsymbol{S}_k^2 + \alpha^2 - 2\boldsymbol{m}_k \bullet \boldsymbol{\alpha}\right) / \sigma^2\right]$$

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

$$\sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] = -N \log K \quad : \text{all terms are } 1/K$$

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i | c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

- Third term is sort of like first, but now we have $c_i{}^T\boldsymbol{\mu}$ together and λ replaces σ and more...
- That leads to a sum over k and a ϕ_{ik} factor.
- Also the sum of i is to N.

ELBO

$$\mathcal{L}(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\phi}) =$$

$$= \sum_{k=1}^{K} E_q[\log p(\boldsymbol{\mu}_k)] + \sum_{i=1}^{N} E_q\Big[\log p(c_i)\Big] + \sum_{i=1}^{N} E_q\Big[\log p(\boldsymbol{x}_i|c_i, \boldsymbol{\mu})\Big]$$

$$- \sum_{k=1}^{K} E_q\Big[\log q(\boldsymbol{\mu}_k)\Big] - \sum_{i=1}^{N} E_q\Big[\log q(c_i)\Big]$$

- Forth term is easy as it ends up as moments of a normal distribution. End up with an expression with p, K and s.

- Last term is easy too and will only involve the ϕ_{ik}

Structured Variational Approximations

- In the LDA tutorial we will see we can simplify the graph by removing edges to get an approximation.
- An example might be to define Q over a 2D grid as only having edges long the rows but not the columns.
- The book shows how you can get something like the mean field approximation but for entire clique potentials rather than one variable at a time.
- So called cluster mean field.