INFERENCE AS OPTIMIZATION

Wenbao Li   Songling Liu

DM Lab

May 25, 2016

Outline

11.1 Introduction

In the previous chapters, we have learn some exact inference methods such as elimination, message passing and the junction-tree algorithm.

The Limitation

However,the computational and space complexity of the exact inference is exponential in the tree-width.

- Computational complexity:$e^{tree-width}$
- Space complexity:$e^{tree-width}$

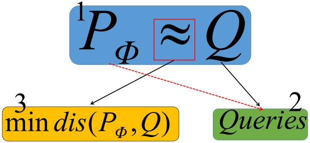
11.1 Introduction

Approximate Inference Technologies

In this chapter we will introduce a class of approximate inference technologies which solve the inference problems which can be understood as an optimization problem.

Some Common Principles

For each method, there are some common principles.



- Define a target class $\mathbf{Q}$ of "easy" distributions Q and then find the "best" approximation to P_{Φ}.
- Inference on Q rather than on P_{Φ}.
- Same target function.

11.1 Introduction

Approximate Inference Process

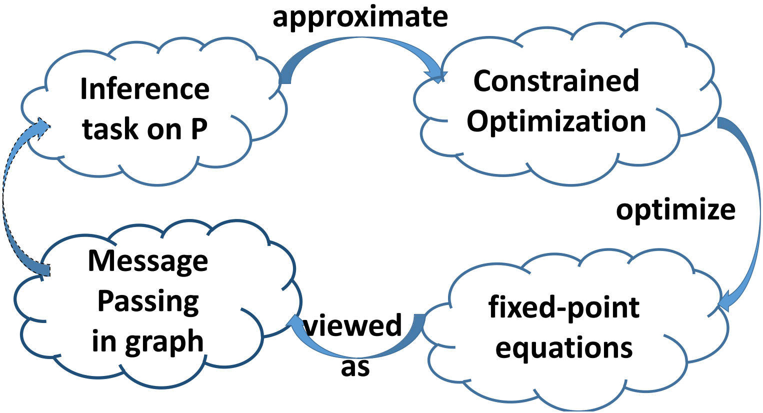


Figure: Process of Approximate Inference Methods

11.1 Introduction

Three categories of Approximate Inference Methods

In the following section, the approximate inference methods mainly fall into three categories.

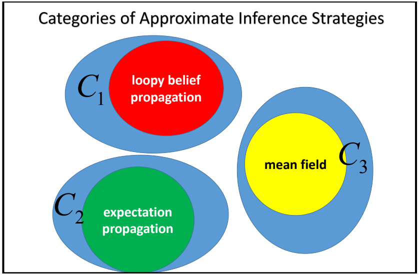


Figure: C_1:use clique-tree message passing schemes on structures other than trees(maybe graph).C_2:use message propagation on clique trees with approximate messages.C_3:generalize the mean field method originating in statistical physics.

11.1 Introduction

Two perspectives

Each of these algorithms can be described from two perspectives:

- as a message passing algorithm
- as an optimization problem consisting of an objective and a constraint space.

Common process

- First:a simple variant of the algorithm.
- Then:optimization perspective on the algorithm.
- Finally:generalizations of the simple algorithm.

11.1.1 Exact inference revisited

Casting exact inference as an optimization problem

Assume we have a factorized distribution of the form

$$P_\Phi(\mathcal{X}) = \frac{1}{Z} \prod_{\phi \in \Phi} \phi(\mathbf{U}_\phi) \tag{1}$$

- Factors ϕ in Φ(因子)
- $\mathbf{U}_\phi = Scope[\phi] \subseteq \mathcal{X}$(辖域)

Queries on the P_Φ which include:

- queries about marginal probabilities of variables
- queries about the partition function Z

11.1.1 Exact inference revisited

Casting exact inference as an optimization problem

Beliefs for cluster tree can represent a distribution.

$$\tilde{P}_{\Phi}(\mathcal{X}) = \frac{\prod_{i \in \mathcal{V}_{\mathcal{T}}} \beta_i(\mathbf{C}_i)}{\prod_{(i-j) \in \mathcal{E}_{\mathcal{T}}} \mu_{i,j}(\mathbf{S}_{i,j})} \tag{2}$$

Thus,

exact inference $\rightarrow$ search Q^* that matches P_{Φ} over $\mathbf{Q}$

Another description: searching for a calibrated distribution that is as close as possible to P_{Φ}.

11.1.1 Exact inference revisited

Casting exact inference as an optimization problem

Definition

The relative entropy which measure the distance of P_1 and P_2 is defined as follows:

$$\mathbb{D}(P_1||P_2) = \mathbb{E}_{P_1}\left[\ln \frac{P_1(\mathcal{X})}{P_2(\mathcal{X})}\right]. \tag{3}$$

- Non-negative
- 0 if and only if $P_1 = P_2$.
- Not symmetric. $\mathbb{D}(P_1||P_2) != \mathbb{D}(P_2||P_1)$.

Two ways of projections(which to choose?)(*See chapter 8.5*):

- *M-projection:* $\min \mathbb{D}(P_\Phi||Q)$
- *I-projection:* $\min \mathbb{D}(Q||P_\Phi)$ WHY?

11.1.1 Exact inference revisited

Casting exact inference as an optimization problem

$\mathcal{T}$ is clique tree of P_Φ, given a set of beliefs

$$\mathbf{Q} = \{\beta_i : i \in \mathcal{V}_\mathcal{T}\} \cup \{\mu_{i,j} : (i - j) \in \mathcal{E}_\mathcal{T}\} \tag{4}$$

where $\mathbf{C}_i$ denotes clusters in $\mathcal{T}$, β_i denotes beliefs over $\mathbf{C}_i$, and $\mu_{i,j}$ denotes beliefs over $\mathbf{S}_{i,j}$ of edges in $\mathcal{T}$.

As in definition 10.6, the set of beliefs in $\mathcal{T}$ defines a distribution Q by the formula

$$Q(\mathcal{X}) = \frac{\prod_{i \in \mathcal{V}_\mathcal{T}} \beta_i(\mathbf{C}_i)}{\prod_{(i-j) \in \mathcal{E}_\mathcal{T}} \mu_{i,j}(\mathbf{S}_{i,j})} \tag{5}$$

And the beliefs correspond to marginals of the distribution Q defined by eqution 5.

11.1.1 Exact inference revisited

Casting exact inference as an optimization problem

Consider two decisions when deciding on the representation of $\mathbf{Q}$:

- space of distribution(所有以$\mathcal{T}$为I-map的分布)
- representation of these distributions(作为校准的团置信的一个集合)

11.1.1 Exact inference revisited

The optimization problem

CTree-Optimize-KL:

Find:

$$\mathbf{Q} = \{\beta_i : i \in \mathcal{V}_\mathcal{T}\} \cup \{\mu_{i,j} : (i - j) \in \mathcal{E}_\mathcal{T}\}$$

maximizing:

$$-\mathbb{D}(Q||P_\Phi)$$

subject to:

$$\mu_{i,j}[\mathbf{s}_{i,j}] = \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{c}_i), \forall (i - j) \in \mathcal{E}_\mathcal{T}, \forall \mathbf{s}_{i,j} \in Val(S_{i,j})$$

$$\sum_{\mathbf{c}_i} \beta_i(\mathbf{c}_i) = 1, \forall i \in \mathcal{V}_\mathcal{T}$$

11.1.1 Exact inference revisited

Theorem

If $\mathcal{T}$ is an I-map of P_Φ, then there is a unique solution to CTree-Optimize-KL.

This optimum can be found using the exact inference algorithms we developed in chapter 10.

11.1.2 The Energy Functional(能量泛函)

- Instead of ~~searching over the space of all calibrated cluster trees~~, we can search over a space of "simple" distributions.
- Find an approximate one instead of ~~equivalent one~~.Moreover, we can design the set of distributions where we can perform inference efficiently.

11.1.2 The Energy Functional

Theorem

$\mathbb{D}(Q||P_\Phi) = \ln Z - F[\tilde{P}_\Phi, Q]$.

where $F[\tilde{P}_\Phi, Q]$ *is the energy functional*

$$F[\tilde{P}_\Phi, Q] = \mathbb{E}_Q[\ln \tilde{P}(\mathcal{X})] + \mathbb{H}_Q(\mathcal{X}) = \sum_{\phi \in \Phi} \mathbb{E}_Q[\ln \phi] + \mathbb{H}_Q(\mathcal{X}). \tag{6}$$

Proof see next page.

energy functional $=$ *energy term* $+$ *entropy term*

- Energy term:the expectations of the logarithms of factors in Φ.
- Entropy term:the entropy of Q.

11.1.3 Optimizing the Energy Functional

Proof

Proof.

$$\mathbb{D}(Q||P_\Phi) = \mathbb{E}_Q\left[\ln \frac{Q(\mathcal{X})}{P_\Phi(\mathcal{X})}\right] \text{ (relative entropy definition)}$$
$$= \mathbb{E}_Q\left[\ln Q(\mathcal{X}) - \ln P_\Phi(\mathcal{X})\right] \text{ (expansion)}$$
$$= \mathbb{E}_Q\left[\ln Q(\mathcal{X})\right] - \mathbb{E}_Q\left[\ln P_\Phi(\mathcal{X})\right] \text{ (expansion)}$$
$$= -\mathbb{H}_Q(\mathcal{X}) - \mathbb{E}_Q\left[\ln\left(\frac{\prod_{\phi\in\Phi}\phi(U_\phi)}{Z}\right)\right] \text{ (factor form of distribution)}$$
$$= -\mathbb{H}_Q(\mathcal{X}) - \mathbb{E}_Q\left[\sum_{\phi\in\Phi}\ln\phi(U_\phi) - \ln Z\right] \text{ (expansion)}$$
$$= -\mathbb{H}_Q(\mathcal{X}) - \sum_{\phi\in\Phi}\left[\mathbb{E}_Q \ln\phi(U_\phi)\right] + \ln Z \text{ (expansion)}$$
$$= \ln Z - F[\tilde{P}_\Phi, Q]$$

□

11.1.3 Optimizing the Energy Functional

Problem transformation

- Find good approximation Q
- $\min$ Relative entropy
- $\max$ Energy functional

Energy functional:*lower bound* on the logarithm of the partition function Z, for any choice of Q.
So, inference methods $\Leftrightarrow$ strategies for optimizing the energy functional.
Variational Methods:(这个名字指的是一种通过引入新的变分参数来增加自由度，然后优化这些参数，从而解决问题的通用策略。)

11.2 Exact Inference as Optimization

Variational approach and exact inference

Factored Energy Functional:

Definition

Given a cluster tree $\mathcal{T}$ with a set of beliefs $\mathbf{Q}$ and an assignment α that maps factors in P_Φ to clusters in $\mathcal{T}$, we define the factored energy functional:

$$\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}] = \sum_{i \in \mathcal{V}_\mathcal{T}} \mathbb{E}_{\mathbf{C}_i \sim \beta_i}[\ln \psi] + \sum_{i \in \mathcal{V}_\mathcal{T}} \mathbb{H}_{\beta_i}(\mathbf{C}_i) - \sum_{(i-j) \in \mathcal{E}_\mathcal{T}} \mathbb{H}_{\mu_{i,j}}(\mathbf{S}_{i,j}), \tag{7}$$

where ψ_i is the initial potential assigned to $\mathbf{C}_i$: $\psi_i = \prod_{\phi, \alpha(\phi)=i} \phi$, and $\mathbb{E}_{\mathbf{C}_i \sim \beta_i}[\cdot]$ denotes expectation on the value $\mathbf{C}_i$ given the beliefs β_i

All the terms are local.

11.2 Exact Inference as Optimization

Variational approach and exact inference

Proposition

If $\mathbf{Q}$ is a set of calibrated beliefs for $\mathcal{T}$, and Q is defined by equation 5, then

$$\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}] = \tilde{F}[\tilde{P}_\Phi, Q]$$

Proof.

Note that $\ln \psi_i = \sum_{\phi, \alpha(\phi)=i} \ln \phi$. Moreover, since $\beta_i(\mathbf{c}_i) = Q(\mathbf{c}_i)$, we conclude that $\sum_i \mathbb{E}_{\mathbf{C}_i - \beta_i}[\ln \psi_i] = \sum_\phi \mathbb{E}_{\mathbf{C}_i - Q}[\ln \phi]$.
It remains to show that
$\mathbb{H}_Q(\mathcal{X}) = \sum_{i \in \mathcal{V}_\mathcal{T}} \mathbb{H}_{\beta_i}(\mathbf{C}_i) - \sum_{(i-j) \in \mathcal{E}_\mathcal{T}} \mathbb{H}_{\mu_{i,j}} \mathbf{S}_{i,j}$.
This equality follows directly from equation 5 and theorem 10.4. □

11.2 Exact Inference as Optimization

Variational approach and exact inference

Reformulating CTree-Optimize-KL(Energy functional form).

Optimization Problem

CTree-Optimize:

Find: $\mathbf{Q} = \{\beta_i : i \in \mathcal{V}_\mathcal{T}\} \cup \{\mu_{i,j} : (i-j) \in \mathcal{E}_\mathcal{T}\}$.

Maximizing: $\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}]$.

Subject to:

$$\mu_{i,j}[\mathbf{s}_{i,j}] = \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{c}_i), \forall (i-j) \in \mathcal{E}_\mathcal{T}, \forall \mathbf{s}_{i,j} \in Val(S_{i,j}) \quad (8)$$

$$\sum_{\mathbf{c}_i} \beta_i(\mathbf{c}_i) = 1, \forall i \in \mathcal{V}_\mathcal{T} \quad (9)$$

$$\beta_i(\mathbf{c}_i) \geq 0, \forall i \in \mathcal{V}_\mathcal{T}, \mathbf{c}_i \in \mathbf{Val}(\mathbf{C}_i) \quad (10)$$

11.2.1 Fix-point Characterization

Lagrange optimizing

$$
\begin{aligned}
\mathcal{J} &= \tilde{F}\left[\tilde{P}_{\Phi}, \mathbf{Q}\right] \\
&- \sum_{i \in \mathcal{V}_{\mathcal{T}}} \lambda_i \left(\sum_{\mathbf{c}_i} \beta_i\left(\mathbf{c}_i\right) - 1 \right) \\
&- \sum_{i} \sum_{j \in Nb_i} \sum_{\mathbf{s}_{i,j}} \lambda_{j \to i}\left[\mathbf{s}_{i,j}\right] \left(\sum_{\mathbf{c}_i \sim \mathbf{s}_{i,j}} \beta_i\left(\mathbf{c}_i\right) - \mu_{i,j}\left[\mathbf{s}_{i,j}\right] \right),
\end{aligned} \tag{11}
$$

11.2.1 Fix-point Characterization

Lagrange optimizing

Derivation:

$$\frac{\partial}{\partial \beta_i(\mathbf{c}_i)} \mathcal{J} = \ln \psi_i[\mathbf{c}_i] - \ln \beta_i(\mathbf{c}_i) - 1 - \lambda_i - \sum_{j \in Nb_i} \lambda_{j \to i}[\mathbf{s}_{i,j}]. \quad (12)$$

$$\frac{\partial}{\partial \mu_{i,j}[\mathbf{s}_{i,j}]} \mathcal{J} = \ln \mu_{i,j}[\mathbf{s}_{i,j}] + 1 + \lambda_{i \to j}[\mathbf{s}_{i,j}] + \lambda_{j \to i}[\mathbf{s}_{i,j}]. \quad (13)$$

11.2.1 Fix-point Characterization

Lagrange optimizing

Equating each derivative to 0, rearranging terms, and exponentiating, we get:

$$\beta_i(\mathbf{c}_i) = exp\{-1 - \lambda_i\}\,\psi_i[\mathbf{c}_i]\prod_{j \in Nb_i} exp\{-\lambda_{j\to i}[\mathbf{s}_{i,j}]\} \tag{14}$$

$$\mu_{i,j}[\mathbf{s}_{i,j}] = exp\{-1\}\,exp\{-\lambda_{i\to j}[\mathbf{s}_{i,j}]\}\,exp\{-\lambda_{j\to i}[\mathbf{s}_{i,j}]\} \tag{15}$$

11.2.1 Fix-point Characterization

Lagrange optimizing

We define

$$\delta_{i\to j}\left[\mathbf{s}_{i,j}\right] = exp\left\{-\lambda_{i\to j}\left[\mathbf{s}_{i,j}\right] - \frac{1}{2}\right\} \tag{16}$$

Rewrite the equations as

$$\beta_i\left(\mathbf{c}_i\right) = exp\left\{-\lambda_i - 1 + \frac{1}{2}\left|Nb_i\right|\right\}\psi_i\left(\mathbf{c}_i\right)\prod_{j\in Nb_i}\delta_{j\to i}\left[\mathbf{s}_{i,j}\right] \tag{17}$$

$$\mu_{i,j}\left[\mathbf{s}_{i,j}\right] = \delta_{i\to j}\left[\mathbf{s}_{i,j}\right]\delta_{j\to i}\left[\mathbf{s}_{i,j}\right] \tag{18}$$

11.2.1 Fix-point Characterization

Lagrange optimizing

Combining these equations with the first constraint equation:rewrite

$$\begin{aligned}
\delta_{i\to j}\left[\mathbf{s}_{i,j}\right] &= \frac{\mu_{i,j}\left[\mathbf{s}_{i,j}\right]}{\delta_{j\to i}\left[\mathbf{s}_{i,j}\right]} \\
&= \frac{\sum_{\mathbf{c}_i\sim\mathbf{s}_{i,j}}\beta_i\left(\mathbf{c}_i\right)}{\delta_{j\to i}\left[\mathbf{s}_{i,j}\right]} \\
&= \exp\left\{-\lambda_i - 1 + \frac{1}{2}\left|Nb_i\right|\right\} \\
&\times \sum_{\mathbf{c}_i\sim\mathbf{s}_{i,j}}\psi_i\left(\mathbf{c}_i\right)\prod_{k\in Nb_i-\{j\}}\delta_{k\to i}\left[\mathbf{s}_{i,k}\right] \\
&= constant \times \sum_{\mathbf{c}_i\sim\mathbf{s}_{i,j}}\psi_i\left(\mathbf{c}_i\right)\prod_{k\in Nb_i-\{j\}}\delta_{k\to i}\left[\mathbf{s}_{i,k}\right]
\end{aligned} \tag{19}$$

11.2.1 Fix-point Characterization

Lagrange optimizing

Theorem

A set of beliefs $\mathbf{Q}$ is a stationary point of CTree-Optimize if and only if there exists a set of factors $\{\delta_{i\to j}[\mathbf{S}_{i,j}] : (i-j) \in \mathcal{E}_{\mathcal{T}}\}$ such that

$$\delta_{i\to j} \propto \sum_{\mathbf{c}_i \sim \mathbf{s}_{i,j}} \psi_i\left(\mathbf{c}_i\right) \prod_{k \in Nb_i - \{j\}} \delta_{k\to i}\left[\mathbf{s}_{i,k}\right] \tag{20}$$

and moreover, we have that

$$\begin{aligned} \beta_i &\propto \psi_i \left(\prod_{j \in Nb_i} \delta_{j\to i} \right) \\ \mu_{i,j} &= \delta_{j\to i} \cdot \delta_{i\to j} \end{aligned}$$

11.2.1 Fix-point Characterization

Lagrange optimizing

- The solution of the optimization problem
- fixed-point equations
- solving the fixed point equations by an easy iterative approach

11.3 Propagation-Based Approximation

Message propagation in cluster graph

- General message passing algorithm in a cluster graph.
- Derived from a set of fixed-point equations induced by the stationary points of an approximate energy functional.

11.3.1 A simple example

Consider a Markov Network

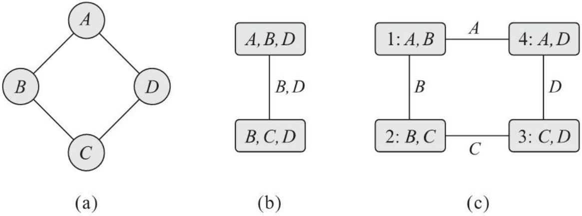
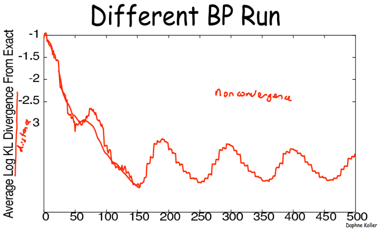


Figure 11.1 An example of a cluster graph. (a) A simple network. (b) A clique tree for the network in (a). (c) A cluster graph for the same network.

exact inference on (b). Inference on (c).

11.3.1 Propagation-Based Approximation

Existing problem



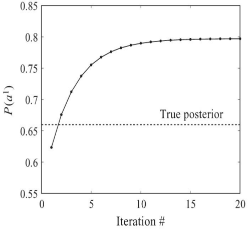


Figure 11.2 An example run of loopy belief propagation in the simple network of figure 11.1a. In this run, all potentials prefer consensus assignments over nonconsensus ones. In each iteration, we perform message passing for all the edges in the cluster graph of figure 11.1b.

Two problems:

- Convergence
- Calibrated cluster graph : True probability distribution

11.3.2 Cluster-graph Belief Propagation

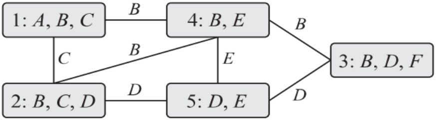
Generalized running intersection property

Definition

We say that $\mathcal{U}$ satisfies the **running intersection property** if , whenever there is a variable X such that $X \in \mathbf{C}_i$ and $X \in \mathbf{C}_j$,then there is a single path between $\mathbf{C}_i$ and $\mathbf{C}_j$ for which $X \in \mathbf{S}_c$ for all edges e in the path.

- Must exist $\rightarrow$ Message about X flows across the cluster containing it.
- At most one. $\rightarrow$ Stop the cycles.

11.3.2 Cluster-graph Belief Propagation(cont.)



In cluster tree, RIP$\Rightarrow$ $\mathbf{S}_{i,j} = \mathbf{C}_i \cap \mathbf{C}_j$.
In cluster graph, No.

11.3.2 Cluster-graph Belief Propagation(cont.)

Calibrated cluster graph

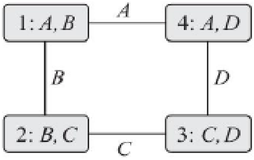
Calibrated Cluster graph : if for every $(i - j)$ connecting cluster $\mathbf{C}_i$ and $\mathbf{C}_j$, we have that $\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j$.

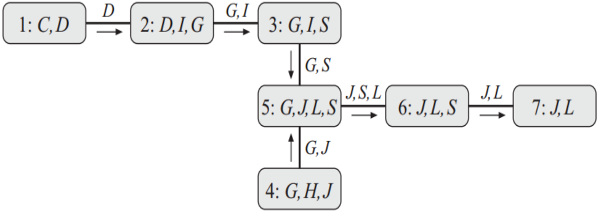
- weak than calibrate cluster tree.
- agree only on those variables in the sepset.

How to calibrate cluster graph ?

CGraph-SP-Calibrate

Just like **CTree-SP-Calibrate**





- Cluster graph contains loops.Not like the cluster tree, there is no cluster ready in cluster graph.

So : initialize $\delta_{i \to j} = 1$ for every edge $(i - j) \in \mathcal{E}_{\mathcal{U}}$.

CGraph-SP-Calibrate

Alg:Calibration using sum-product belief propagation in a cluster graph

Procedure CGraph-SP-Calibrate (
  Φ, // Set of factors
  $\mathcal{U}$ // Generalized cluster graph Φ
)
  Initialize-CGraph
  while graph is not calibrated
    Select $(i\text{–}j) \in \mathcal{E}_{\mathcal{U}}$
    $\delta_{i\to j}(\boldsymbol{S}_{i,j}) \leftarrow$ SP-Message(i, j)
  for each clique i
    $\beta_i \leftarrow \psi_i \cdot \prod_{k\in \mathrm{Nb}_i} \delta_{k\to i}$
  return $\{\beta_i\}$

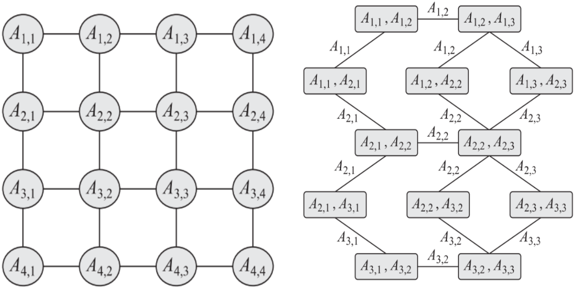
Procedure Initialize-CGraph (
  $\mathcal{U}$
)
  for each cluster $\boldsymbol{C}_i$
    $\beta_i \leftarrow \prod_{\phi\,:\,\alpha(\phi)=i} \phi$
  for each edge $(i\text{–}j) \in \mathcal{E}_{\mathcal{U}}$
    $\delta_{i\to j} \leftarrow 1$
    $\delta_{j\to i} \leftarrow 1$

Procedure SP-Message (
  i, // sending clique
  j // receiving clique
)
  $\psi(\boldsymbol{C}_i) \leftarrow \psi_i \cdot \prod_{k\in(\mathrm{Nb}_i-\{j\})} \delta_{k\to i}$
  $\tau(\boldsymbol{S}_{i,j}) \leftarrow \sum_{\boldsymbol{C}_i-\boldsymbol{S}_{i,j}} \psi(\boldsymbol{C}_i)$
  return $\tau(\boldsymbol{S}_{i,j})$

Same as **CGraph-BU-Calibrate**.Initialize $\mu_{i,j} = 1$.
They are instances of a general class of algorithms called *cluster-graph belief propagation*, which passes messages over cluster graphs.

Lower costs of cluster-graph belief propagation than running exact inference

Another example



- Exact inference in $n * n$ grid network(exponential in n)
- A round of propagations in the generalized cluster graph (linear in the size of the grid: n^2)

11.3.3 Properties of Cluster-Graph Belief Propagation

Cluster graph invariant(不变量)

Theorem

Let $\mathcal{U}$ be a generalized cluster graph over a set of factors Φ. Consider the set of beliefs $\{\beta_i\}$ and sepset $\{\mu_{i,j}\}$ at any iteration of CGraph-BU-Calibrate; then

$$\tilde{P}_\Phi(\mathcal{X}) = \frac{\prod_{i \in \mathcal{V}_\mathcal{U}} \beta_i[\mathbf{C}_i]}{\prod_{(i-j) \in \mathcal{E}_\mathcal{U}} \mu_{i,j}[\mathbf{S}_{i,j}]} \tag{21}$$

where $\tilde{P}_\Phi(\mathcal{X}) = \prod_{\phi \in \Phi} \phi$ is the unnormalized distribution defined by Φ.

11.3.3 Properties of Cluster-Graph Belief Propagation

Tree Consistency

第10章中，在校准的聚类树中，聚类上的置信是联合分布的边缘概率。可以从中读出所关心变量的边缘的概率。

这是否在校准的聚类图中也成立呢?

计算的概率是一个近似，近似质量如何?

11.3.3 Properties of Cluster-Graph Belief Propagation

Tree Consistency

Theorem

Assume that $\mathcal{T}$ is a sub-tree of calibrated cluster graph $\mathcal{U}$, we can think of it as defining a distribution

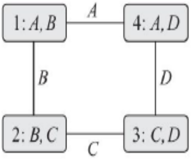
$$P_{\mathcal{T}}(\mathcal{X}) = \frac{\prod_{i \in \mathcal{V}_{\mathcal{T}}} \beta_i[\mathbf{C}_i]}{\prod_{(i-j) \in \mathcal{E}_{\mathcal{T}}} \mu_{i,j}[\mathbf{S}_{i,j}]} \tag{22}$$

If the cluster graph is calibrated, then by definition so is $\mathcal{T}$. And so, because $\mathcal{T}$ is a tree that satisfies the running intersection property, we can apply theorem 10.4 , and we conclude that

$$\beta_i(\mathbf{C}_i) = P_{\mathcal{T}}(\mathbf{C}_i) \tag{23}$$

11.3.3 Properties of Cluster-Graph Belief Propagation

Tree Consistency:example



- Delete cluster $\mathbf{C}_4 = \{A, D\} \Rightarrow$ A suitable cluster tree.
- $\beta_1(A, B) = P_{\mathcal{T}}(A, B) \Rightarrow \beta_1(A, B) \neq P_{\Phi}(A, B)$

11.3.4 Analyzing Convergence*

Not mention

- Cluster tree $\Rightarrow$ Converge
- Many network $\Rightarrow$ Don't converge

11.3.5 How to Construct Cluster Graphs

Compromise between cost and accuracy

聚类图的结构确定算法所执行的传播步骤，并且因此确定了什么类型的信息可以在传播的过程中传递。这些选择直接对结果的质量产生影响。

11.3.5 How to Construct Cluster Graphs

Compromise between cost and accuracy

For example:

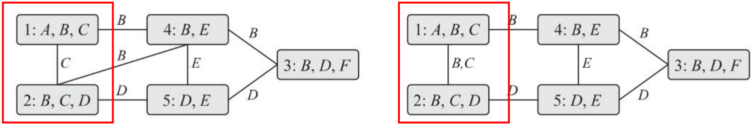


Figure 11.3 **Two examples of generalized cluster graph for an MRF** with potentials over $\{A, B, C\}$, $\{B, C, D\}$, $\{B, D, F\}$, $\{B, E\}$ and $\{D, E\}$.

Cluster graph $\boldsymbol{U_2}$ capture the strong dependencies between B and C.

On the other hand, we have to make sure a **valid** cluster graph.

11.3.5 Construct Cluster Graphs

11.3.5.1 Pairwise Markov Networks

Definition

A Pairwise Markov Networks is an undirected graph whose nodes are $X_1, ..., X_n$ and each edge $X_i \leftrightarrow X_j$ is associated with a factor(potential) $\phi(X_i \leftrightarrow X_j)$.(From Chapter 4.1)

For each potential, we introduce a corresponding cluster, and put edges between the clusters that have overlapping scope. In other words, there is an edge between the cluster $\mathbf{C}_{(i,j)}$ that corresponds to the edge $X_i \leftrightarrow X_j$ and the clusters $\mathbf{C}_i$ and $\mathbf{C}_j$ that correspond to the univariate factors over X_i and X_j.

Pairwise Markov Networks

Example:PMN $\Rightarrow$cluster graph

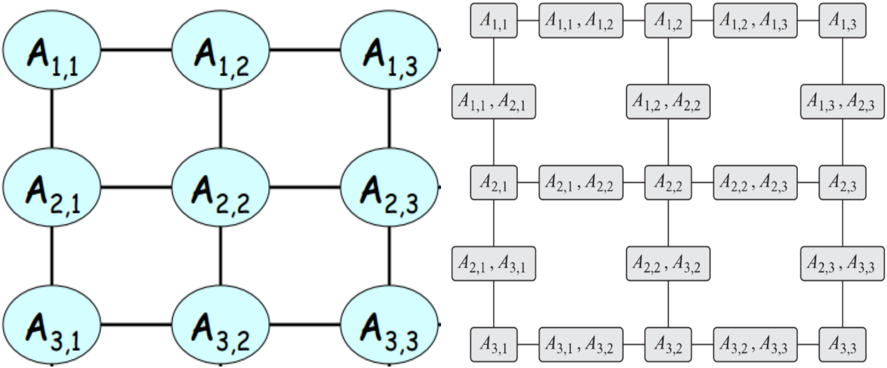


Figure: (a)A 3*3 grid network (b)A generalized cluster graph for 3*3 grid when viewed as pairwise MRF

11.3.5.2 Bethe cluster graph

What is Bethe cluster graph

Big clusters(Scope of factor for each $\phi \in \Phi$) + univariate clusters + edges between them

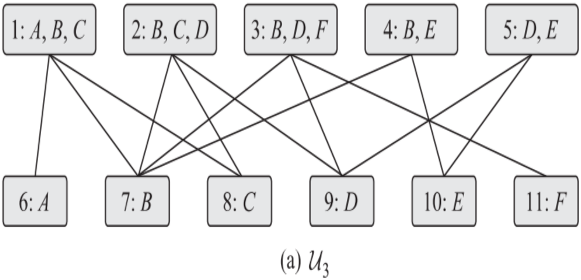


Figure: (a) Bethe factorization.

11.3.5.3 Beyond Marginal Probabilities

Some improvement

- Limitation of BetheCG:Lost the interaction between variables.
- Solution one:Merge some of the large clusters. $\Rightarrow$ Brings costs.
- Solution two:Add a mediate distribution over B and C.

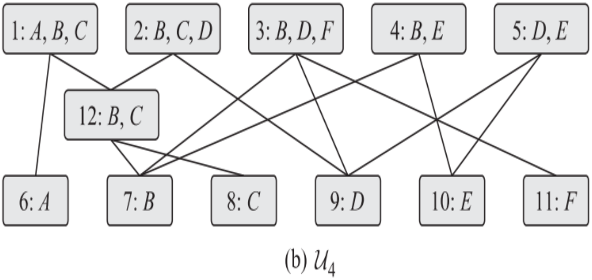


Figure: (b) Capturing interactions between A, B, C and $\{B, C, D\}$

Some change

Approximate Bethe CG

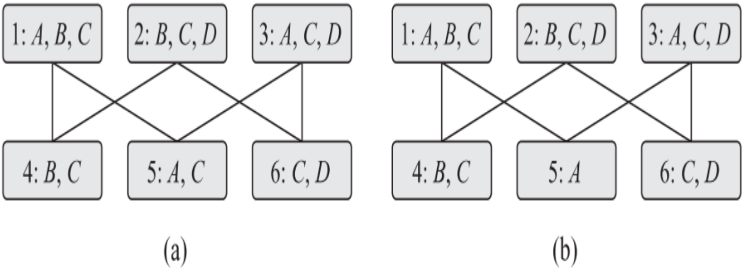


Figure: (a) Invalid (b)A way to be valid

.

11.3.6 Variational Analysis

Energy Functional Review(More see 11.1)

- Energy functional:

$$F[\tilde{P}_\Phi, Q] = \mathbb{E}_Q[\ln \tilde{P}(\mathcal{X})] + \mathbb{H}_Q(\mathcal{X}) = \sum_{\phi \in \Phi} \mathbb{E}_Q[\ln \phi] + \mathbb{H}_Q(\mathcal{X}).$$

- Factored energy functional(An approximation for cluster graph):

$$\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}] = \sum_{i \in \mathcal{V}_\mathcal{T}} \mathbb{E}_{\mathbf{C}_i - \beta_i}[\ln \psi] + \sum_{i \in \mathcal{V}_\mathcal{T}} \mathbb{H}_{\beta_i}(\mathbf{C}_i) - \sum_{(i-j) \in \mathcal{E}_\mathcal{T}} \mathbb{H}_{\mu_{i,j}}(\mathbf{S}_{i,j})$$

11.3.6 Variational Analysis

CTree optimization problem Review

Optimization Problem

CTree-Optimize:

Find: $\mathbf{Q} = \{\beta_i : i \in \mathcal{V}_\mathcal{T}\} \cup \{\mu_{i,j} : (i-j) \in \mathcal{E}_\mathcal{T}\}$.

Maximizing: $\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}]$.

Subject to:

$$\mu_{i,j}[\mathbf{s}_{i,j}] = \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{c}_i), \forall (i-j) \in \mathcal{E}_\mathcal{T}, \forall \mathbf{s}_{i,j} \in Val(S_{i,j}) \quad (24)$$

$$\sum_{\mathbf{c}_i} \beta_i(\mathbf{c}_i) = 1, \forall i \in \mathcal{V}_\mathcal{T} \quad (25)$$

$$\beta_i(\mathbf{c}_i) \geq 0, \forall i \in \mathcal{V}_\mathcal{T}, \mathbf{c}_i \in \mathbf{Val}(\mathbf{C}_i) \quad (26)$$

11.3.6 Variational Analysis

The fixed-point equations Review

The fixed-point equations:

Theorem

A set of beliefs $\mathbf{Q}$ is a stationary point of CTree-Optimize if and only if there exists a set of factors $\{\delta_{i\to j}[\mathbf{S}_{i,j}] : (i-j) \in \mathcal{E}_{\mathcal{T}}\}$ such that

$$\delta_{i\to j} \propto \sum_{\mathbf{c}_i \sim \mathbf{s}_{i,j}} \psi_i(\mathbf{c}_i) \prod_{k \in Nb_i - \{j\}} \delta_{k\to i}[\mathbf{s}_{i,k}] \tag{27}$$

and moreover, we have that

$$\begin{aligned} \beta_i &\propto \phi_i \left(\prod_{j \in Nb_i} \delta_{j\to i} \right) \\ \mu_{i,j} &= \delta_{j\to i} \cdot \delta_{i\to j} \end{aligned}$$

11.3.6 Variational Analysis

Why variational analysis

Cluster graph belief propagation , approximate ? !

$$\Longrightarrow$$

Variational analysis provides the relative proof.

Message $\Leftarrow$ Fixed-point equations

11.3.6 Variational Analysis

How to get the formalism:Step 1

- First, exact energy functional is hard to optimize.
- Factored energy functional is defined by entropy of cluster and sepset (Local information).

Approximate energy functional.

11.3.6 Variational Analysis

How to get the formalism:Step 2

The whole space of optimized $\mathbf{Q}$ is hard to search the optimal solution.

Definition

So more precisely, consider some cluster graph $\mathcal{U}$,for a distribution P , we define $\mathbf{Q}_P = \{P(\mathbf{C}_i)\}_{i \in \mathcal{V}_\mathcal{U}} \cup \{P(\mathbf{S}_{i,j})\}_{i-j \in \mathcal{E}_\mathcal{U}}$. We now define the marigal polytope(边缘可剖分空间) of $\mathcal{U}$ to be

$$Marg[\mathcal{U}] = \{\mathbf{Q}_P : P(\mathcal{X})\} \quad (28)$$

But obtaining this kind of space is very hard as exact inference.

Approximate constraint space

11.3.6 Variational Analysis

How to get the formalism:Step 2

To avoid these problems, we perform our optimization over the local consistency polytope(局部一致的可剖分空间):

$$Local[\mathcal{U}] = \left\{ \begin{array}{c} \{\beta_i : i \in \mathcal{V}_{\mathcal{U}}\} \cup \\ \{\mu_{i,j} : (i\text{–}j) \in \mathcal{E}_{\mathcal{U}}\} \end{array} \middle| \begin{array}{lcll} \mu_{i,j}[\boldsymbol{s}_{i,j}] & = & \sum_{\boldsymbol{C}_i - \boldsymbol{S}_{i,j}} \beta_i(\boldsymbol{c}_i) & \forall (i\text{–}j) \in \mathcal{E}_{\mathcal{U}}, \forall \boldsymbol{s}_{i,j} \in Val(\boldsymbol{S}_{i,j}) \\ 1 & = & \sum_{\boldsymbol{c}_i} \beta_i(\boldsymbol{c}_i) & \forall i \in \mathcal{V}_{\mathcal{U}} \\ \beta_i(\boldsymbol{c}_i) & \geq & 0 & \forall i \in \mathcal{V}_{\mathcal{U}}, \boldsymbol{c}_i \in Val(\boldsymbol{C}_i). \end{array} \right\} \quad (11.16)$$

Some keywords : *pseudo-marginal distributions(伪边缘分布), calibrated*

Approximate constraint space

11.3.6 Variational Analysis

Optimization problem description

Optimization Problem

CGraph-Optimize:

Find:$\mathbf{Q} = \{\beta_i : i \in \mathcal{V}_{\mathcal{U}}\} \cup \{\mu_{i,j} : (i-j) \in \mathcal{E}_{\mathcal{U}}\}$.

Maximizing:$\tilde{F}[\tilde{P}_{\Phi}, \mathbf{Q}]$.

Subject to:

$$\mathbf{Q} \in Local[\mathcal{U}] \tag{29}$$

Thus, our optimization problem contains two approximations:

- Approximate energy functional;
- Approximate optimized variable's space(the space of pseudo-marginals)

11.3.6 Variational Analysis

Fix-point equations

Theorem

A set of beliefs $\mathbf{Q}$ *is a stationary point of CGree-Optimize if and only if for every edge* $(i - j) \in \mathcal{E}_{\mathcal{U}}$ *there are auxiliary factors* $\{\delta_{i \to j}[\mathbf{S}_{i,j}] : (i - j) \in \mathcal{E}_{\mathcal{U}}\}$ *such that*

$$\delta_{i \to j} \propto \sum_{\mathbf{c}_i \sim \mathbf{s}_{i,j}} \psi_i(\mathbf{c}_i) \prod_{k \in Nb_i - \{j\}} \delta_{k \to i}[\mathbf{s}_{i,k}] \tag{30}$$

and moreover, we have that

$$\begin{aligned} \beta_i &\propto \phi_i \left(\prod_{j \in Nb_i} \delta_{j \to i} \right) \\ \mu_{i,j} &= \delta_{j \to i} \cdot \delta_{i \to j} \end{aligned}$$

11.3.6 Variational Analysis

Convergence point and stationary point

Proposition

$\mathbf{Q}$ *is the convergence point of applying CGraph-SP-Calibrate*$(\phi,\mathcal{U})$ *if and only if* $\mathbf{Q}$ *is a stationary point of* $\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}]$.

Proposition

At convergence of CGraph-BU-Calibrate, the set of beliefs is a stationary point of $\tilde{F}[\tilde{P}_\Phi, \mathbf{Q}]$.

Conclusion

Take home message

- Optimization format of inference.(Its optimal point:The fixed-point equation)
- CGraph-SP-Calibrate/CGraph-BU-Calibrate(cluster-graph belief propagation)
- Equivalence property of above two process.

Next...section 4 - 6

Next 3 parts will be presented by Songling Liu.