Probabilistic Graphical Models

Raquel Urtasun and Tamir Hazan

TTI Chicago

April 25, 2011

Summary

Past two weeks

- Exact inference via VE
- Exact inference via message-passing

This week

- Exact inference via optimization
- Approximate inference via optimization

Exact inference

- The computational complexity and memory requirements of exact inference are exponential with the tree-width.
- This is prohibitive for a large set of applications.
- In this week we will see approximations that construct an approx. of P_Φ that is simple to do inference over.
- The general principle exploited is locality.
- The target class (i.e., approximation) is called $\mathcal{Q}$.
- We seek a Q that best approximates P_Φ.
- Queries will be done over Q.

Inference as Optimization

There are three types of approx. methods:

1. Methods that use clique tree message passing on structures other than cliques, e.g., loopy BP. They optimize approximate versions of the energy functional.
2. Methods that use message passing on clique trees with approximate messages, e.g., expectation propagation (EP).
3. Generalizations of mean field methods. They use the exact energy functional, but restrict attention to a class $\mathcal{Q}$ that have a particular simple factorization.

Exact inference as optimization

- Assume we have a factorized distribution

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

  with $\mathbf{U}_{\Phi} = Scope(\phi) \subseteq \mathcal{X}$.

- The result of Sum-Product BP is a calibrated tree, with calibrated set of beliefs.
- In exact inference we find beliefs that match the distribution defined by an initial set of factors.
- We can interpret exact inference as searching over the set of distributions $\mathcal{Q}$ that are representable by the cluster tree to find a distribution Q^* that matches P_{Φ}.
- Thus we search for a calibrated distribution that is "as close as possible" to P_{Φ}.
- Many possible ways: L_2, L_1, relative entropy, etc.

Relative Entropy

- The **relative entropy** or **KL divergence** between P_1 and P_2 is

$$\mathbf{D}(P_1||P_2) = \mathbf{E}_{\mathcal{X}\sim P_1}\left[\ln \frac{P_1(\mathcal{X})}{P_2(\mathcal{X})}\right]$$

- $\mathbf{D}(P_1||P_2) \geq 0$ and is 0 iff $P_1(\mathcal{X}) = P_2(\mathcal{X})$.
- The relative entropy is not symmetric (remember lecture on M-projection $\mathbf{D}(P_\Phi||Q)$ and I-projection $\mathbf{D}(Q||P_\Phi)$).
- M-projection is more adequate, as is the number of bits lost when coding P_Φ using Q.
- However, the M-projection requires marginals over P_Φ to compute

$$Q = \underset{Q}{\text{argmin}}\, \mathbf{D}(P_\Phi||Q)$$

and the I-projection does not to compute

$$Q = \underset{Q}{\text{argmin}}\, \mathbf{D}(Q||P_\Phi)$$

Representation I

- We want to search over Q that minimizes $\mathbf{D}(Q||P_\Phi)$.
- Suppose we are given a cluster tree $\mathcal{T}$ for P_Φ: $\mathcal{T}$ satisfies running intersection and family preserving properties.
- Suppose we are given a set of beliefs

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

  with β_i the beliefs over $\mathbf{C}_i$ and $\mu_{i,j}$ the beliefs over $\mathbf{S}_{i,j}$.
- The set of beliefs satisfy the clique tree invariant

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

- The set of beliefs $\mathbf{Q}$ satisfy the **marginal consistency constraints** if

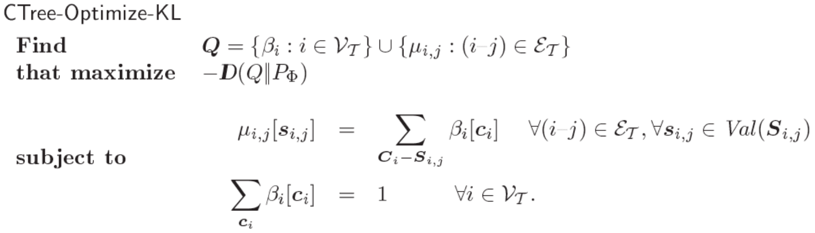
$$\forall i \in \mathcal{V}_\mathcal{T}, \;\; \beta_i(\mathbf{c}_i) = Q(\mathbf{c}_i), \qquad \forall (i-j) \in \mathcal{E}_\mathcal{T}, \;\; \mu_{i,j}(\mathbf{s}_{i,j}) = Q(\mathbf{s}_{i,j})$$

- The beliefs correspond to the marginals of Q.

Representation II

- We are searching over a set of distributions Q that are representable by a set of beliefs $\mathbf{Q}$ over the cliques and sepsets in a particular clique tree structure.
- We have make two decisions on Q:
  1. Space of distributions we are considering, i.e., all distributions such as $\mathcal{T}$ is an I-map.
  2. Representation of the distributions, i.e., a set of calibrated clique beliefs.
- We can now do exact inference by maximizing $-\mathbf{D}(Q||P_\Phi)$

Optimization Program



CTree-Optimize-KL

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

that maximize $-\boldsymbol{D}(Q \| P_\Phi)$

subject to

$$\begin{aligned} \mu_{i,j}[\boldsymbol{s}_{i,j}] &= \sum_{\boldsymbol{C}_i - \boldsymbol{S}_{i,j}} \beta_i[\boldsymbol{c}_i] \quad \forall (i – j) \in \mathcal{E}_\mathcal{T}, \forall \boldsymbol{s}_{i,j} \in Val(\boldsymbol{S}_{i,j}) \\ \sum_{\boldsymbol{c}_i} \beta_i[\boldsymbol{c}_i] &= 1 \qquad \forall i \in \mathcal{V}_\mathcal{T}. \end{aligned}$$

- When solving this we look at different configurations that satisfies the marginal consistency constraints, and select the configuration that is closer to P_Φ.
- If $\mathcal{T}$ is an I-map of P_Φ then there is a unique solution of this optimization.
- It can be found by the exact inference algorithms we have already seen.
- We can search for Q that minimizes $\mathbf{D}(Q||P_\Phi)$.
- However we have to sum over all possible instantiations of $\mathcal{X}$.

Energy Functional

Theorem: $\mathbf{D}(Q||P_\Phi) = \ln Z - F(\hat{P}_\Phi, Q)$, where $F(\hat{P}_\Phi, Q)$ is the energy functional

$$F(\hat{P}_\Phi, Q) = \mathbf{E}_{\mathcal{X}\sim Q}\left[\ln \hat{P}(\mathcal{X})\right] + \mathbf{H}_Q(\mathcal{X}) = \sum_{\phi\in\Phi} \mathbf{E}_{\mathcal{X}\sim Q}\left[\ln \phi\right] + \mathbf{H}_Q(\mathcal{X})$$

Proof: Let's write

$$\mathbf{D}(Q||P_\Phi) = \mathbf{E}_{\mathcal{X}\sim Q}\left[\ln Q(\mathcal{X})\right] - \mathbf{E}_{\mathcal{X}\sim Q}\left[\ln P_\Phi(\mathcal{X})\right]$$

using product form of P_Φ

$$\ln P_\Phi(\mathcal{X}) = \sum_{\phi\in\Phi} \ln \phi(\mathbf{U}_\phi) - \ln Z$$

Since $\mathbf{H}_Q(\mathcal{X}) = -\mathbf{E}_{\mathcal{X}\sim Q}\left[\ln Q(\mathcal{X})\right]$ then

$$\mathbf{D}(Q||P_\Phi) = -\mathbf{H}_Q(\mathcal{X}) - \mathbf{E}_{\mathcal{X}\sim Q}\left[\sum_{\phi\in\Phi} \ln \phi(\mathbf{U}_\phi)\right] + \mathbf{E}_{\mathcal{X}\sim Q}\left[\ln Z\right]$$

Z does not depend on Q.

Helmholtz Free Energy

$$\mathbf{D}(Q||P_\Phi) = -\mathbf{H}_Q(\mathcal{X}) - \mathbf{E}_{\mathcal{X}\sim Q}\left[\sum_{\phi\in\Phi} \ln \phi(\mathbf{U}_\phi)\right] + \ln Z$$

- As Z does not depend on Q, minimizing the relative entropy is equivalent to maximizing the energy functional $F(\hat{P}_\Phi, Q)$.
- This is called the **(Helmholtz) Free Energy**.

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

- It contains two terms, the **energy** term and the **entropy** term.
- Choice of Q important so that we can evaluate both terms.

Optimizing the Energy Functional

- We pose the problem of finding a good approx. Q as the one of maximizing the energy functional (minimizing the relative entropy).
- By choosing appropriate Q we can evaluate the energy functional and also maximize it.
- As $\mathbf{D}(Q||P_\Phi) \geq 0$, then $\ln Z \geq F(\hat{P}_\Phi, Q)$.
- The energy functional is a lower bound on the logarithm of the partition function.
- Computing the partition function is one of the hardest queries of inference. This gives us a lower bound.
- We now look into **variational methods**, which are inference methods that optimize this energy functional.
- We introduce additional degrees of freedom over which we optimize to get the best approximation.

Exact inference as optimization

- Reformulate the optimization problem in terms of the energy functional.
- For the case of calibrated trees, we can simplify the objective function.

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

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

where ψ_i is the set of initial potentials

$$\psi_i = \prod_{\phi, \alpha(\phi) = i} \phi$$

- Let's examine these expectations.
- Importantly all the terms are **local**.

Equivalence of energy functionals

Prop: If $\mathbf{Q}$ is a set of calibrated beliefs for $\mathcal{T}$ and Q is defined as

$$Q(\mathcal{X}) = \frac{\prod_{i \in \mathcal{V}_\mathcal{T}} \beta_i}{\prod_{(i-j) \in \mathcal{E}_\mathcal{T}} \mu_{i,j}}$$

then $\hat{F}(\hat{P}_\Phi, \mathbf{Q}) = F(\hat{P}_\Phi, Q)$.

Proof: Since $\ln \psi_i = \sum_{\phi, \alpha(\phi)=i} \ln \phi$ and $\beta_i(\mathbf{c}_i) = Q(\mathbf{c}_i)$ we have

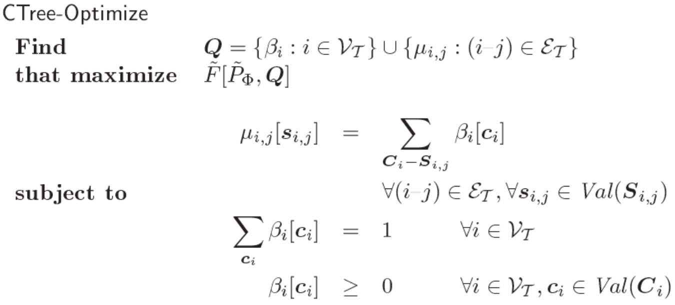
$$\sum_i \mathbf{E}_{\mathbf{C}_i \sim \beta_i}[\ln \psi_i] = \sum_\phi \mathbf{E}_{\mathbf{C}_i \sim Q}[\ln \phi]$$

Moreover

$$\mathbf{H}_Q(\mathcal{X}) = \sum_{i \in \mathcal{V}_\mathcal{T}} \mathbf{H}_{\beta_i}(\mathbf{C}_i) - \sum_{(i-j) \in \mathcal{E}_\mathcal{T}} H_{\mu_{i,j}}(\mathbf{S}_{i,j})$$

Exact inference as optimization

- If Q factorizes according to $\mathcal{T}$, we can represent it with a set of calibrated beliefs.
- We impose marginal consistency constraint so that neighboring beliefs agree on the marginal distribution, i.e., the beliefs are calibrated.
- We can now derive a new optimization



CTree-Optimize

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

that maximize $\tilde{F}[\tilde{P}_\Phi, \boldsymbol{Q}]$

subject to

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

$$\sum_{\boldsymbol{c}_i} \beta_i[\boldsymbol{c}_i] = 1 \qquad \forall i \in \mathcal{V}_{\mathcal{T}}$$

$$\beta_i[\boldsymbol{c}_i] \geq 0 \qquad \forall i \in \mathcal{V}_{\mathcal{T}}, \boldsymbol{c}_i \in Val(\boldsymbol{C}_i)$$

Lagrange multipliers

- The method of Lagrange multipliers provides a strategy for finding the maxima and minima of a function subject to constraints

$$\begin{aligned} \max_{x,y} \quad & f(x,y) \\ \text{subject to} \quad & g(x,y) = c \end{aligned}$$

- We introduce a new variable λ called the Lagrange multiplier and write the Lagrange function

$$L(x,y,\lambda) = f(x,y) + \lambda(g(x,y) - c)$$

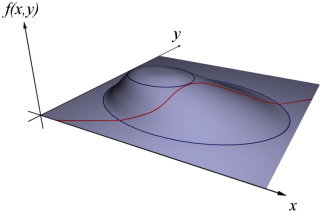
  λ can be added or subtracted.

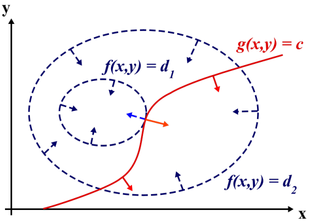
- If $f(x,y)$ is maximum for the original constrained problem, then there exists a λ such that (x,y,λ) is a stationary point for the Lagrange function.
- Stationary points are those points where the partial derivatives of L are zero.
- Not all stationary points yield a solution of the original problem.
- Thus, the method of Lagrange multipliers yields a necessary condition for optimality in constrained problems

Contours and conditions I

Consider a 2D example

$$\begin{aligned} \max_{x,y} \quad & f(x,y) \\ \text{subject to} \quad & g(x,y) = c \end{aligned}$$





- We can visualize the contours $f(x,y) = d$ for values of d and the contour of g given by $g(x,y) = c$.
- While moving along the contour line for $g = c$ the value of f can vary.
- Only when the contour line for $g = c$ meets contour lines of f tangentially, we do not increase or decrease the value of f.

Contours and conditions II

- The contour lines of f and g touch when the tangent vectors of the contour lines are parallel.
- This is the same as saying that the gradients of f and g are parallel.
- Thus we want points (x, y) where $g(x, y) = c$ and

  $$\nabla_{x,y} f = -\lambda \nabla_{x,y} g$$

  where

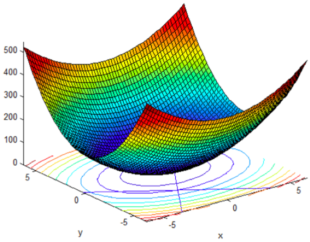
  $$\nabla_{x,y} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right), \qquad \nabla_{x,y} g = \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y} \right)$$

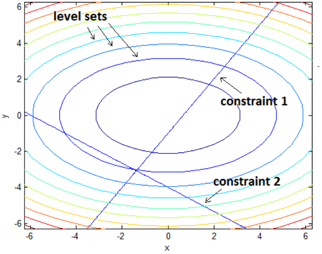
- λ is required as the two gradients might not have the same magnitude.
- To incorporate these conditions into one equation, we introduce an auxiliary function

  $$L(x, y, \lambda) = f(x, y) + \lambda(g(x, y) - c)$$

  and solve $\nabla_{x,y,\lambda} L(x, y, \lambda) = 0$.
- This is the method of Lagrange multipliers.
- Note that $\nabla_{x,y,\lambda} L(x, y, \lambda) = 0$ implies $g(x, y) = c$.

Handling multiple constraints I





- If we consider only the points that satisfy the constraints then a point $(p, f(p))$ is a stationary point of f iff the constraints at that point do not allow movement in a direction where f changes value.
- Once we have located the stationary points, we need to test if its a minimum, a maximum or just a stationary point that is neither.

Handling multiple constraints II

- Consider the level set of f at $(p, f(p))$.
- Let $\{v_L\}$ be the set of vectors containing the directions in which we can move and still remain in the same level set.
- Thus, for every vector v in $\{v_L\}$ we have

$$\Delta f = \frac{df}{dx_1} v_{x_1} + \cdots + \frac{df}{dx_N} v_{x_N}$$

  with v_{x_k} the x_k-th component of v.
- Thus we can write $\nabla f \cdot v = 0$, with $\nabla f = [\frac{df}{dx_1}, \cdots, \frac{df}{dx_N}]^T$.
- All directions from this point that do not change the value of f must be perpendicular to $\nabla f(p)$.
- We can also write $\nabla g \cdot v = 0$.

Single constraint revisited

- At stationary points the direction that changes f is in the same direction that violates the constraint so

$$\nabla f(p) = \lambda \nabla g(p) \qquad \Rightarrow \qquad \nabla f(p) - \lambda \nabla g(p) \; = \; 0$$

- We only do this test when the point $g(p) = 0$, we have 2 eq. that when solved, identify all constrained stationary points:

$$\begin{cases} g(p) = 0 & \text{means point satisfies constraint} \\ \nabla f(p) - \lambda \nabla g(p) = 0 & \text{means point is a stationary point} \end{cases}$$

- Fully expanded, there are $N + 1$ simultaneous equations that need to be solved for the $N + 1$ variables which are λ and $x_1, \; x_2, \ldots, x_N$:

$$\begin{aligned} g(x_1, x_2, \ldots, x_N) &= 0 \\ \frac{df}{dx_1}(x_1, x_2, \ldots, x_N) - \lambda \frac{dg}{dx_1}(x_1, x_2, \ldots, x_N) &= 0 \\ &\vdots \\ \frac{df}{dx_N}(x_1, x_2, \ldots x_N) - \lambda \frac{dg}{dx_N}(x_1, x_2, \ldots, x_N) &= 0 \end{aligned}$$

Multiple constraints

- If there is more than one constraint active together, each constraint contributes a direction that will violate it.
- Together, these violation directions form a violation space.
- The direction that changes f at p is in the violation space defined by the constraints $g_1, g_2, \ldots, g_M$ if and only if:

$$\sum_{k=1}^{M} \lambda_k \nabla g_k(p) = \nabla f(p) \quad \Rightarrow \quad \nabla f(p) - \sum_{k=1}^{M} \lambda_k \nabla g_k(p) = 0$$

- Add equations to guarantee that we only perform this test when we are at a point that satisfies every constraint:

$$\begin{aligned} g_1(p) &= 0 \\ &\vdots \\ g_M(p) &= 0 \\ \nabla f(p) - \sum_{k=1}^{M} \lambda_k \nabla g_k(p) &= 0 \end{aligned}$$

Lagrangian

- Every equation equal to zero is exactly what one would have to do to solve for the unconstrained stationary points of the Lagrangian

$$L\left(x_1, \ldots, x_N, \lambda_1, \ldots, \lambda_M\right) = f\left(x_1, \ldots, x_N\right) - \sum_{k=1}^{M} \lambda_k g_k\left(x_1, \ldots, x_N\right)$$

- Solving the equation above for its unconstrained stationary points generates exactly the same stationary points as solving for the constrained stationary points of f under the constraints $g_1, g_2, \ldots, g_M$.
- The function above is called a **Lagrangian**.
- The scalars $\lambda_1, \lambda_2, \ldots, \lambda_M$ are called **Lagrange Multipliers**.
- This optimization method itself is called **The Method of Lagrange Multipliers**.
- This method is generalized by the **Karush-Kuhn-Tucker conditions**, which can also take into account inequality constraints of the form $h(x) \leq c$.

KKT Conditions

- Let's consider the following optimization problem

$$\begin{array}{ll} \min_{x} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad h_j(x) = 0 \end{array}$$

- Suppose that the objective function, i.e., the function to be minimized, is $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and the constraint functions are $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ and $h_j : \mathbb{R}^n \rightarrow \mathbb{R}$.
- Suppose they are continuously differentiable at a point x^* .
- If x^* is a local minimum that satisfies some regularity conditions, then there exist constants μ_i $(i = 1, \ldots, m)$ and λ_j $(j = 1, \ldots, l)$, called KKT multipliers, such that the following properties are satisfied.

KKT Conditions

Stationarity

$$\nabla f(x^*) + \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*) = 0,$$

Primal feasibility

$$g_i(x^*) \leq 0, \text{ for all } i = 1, \ldots, m$$
$$h_j(x^*) = 0, \text{ for all } j = 1, \ldots, l$$

Dual feasibility

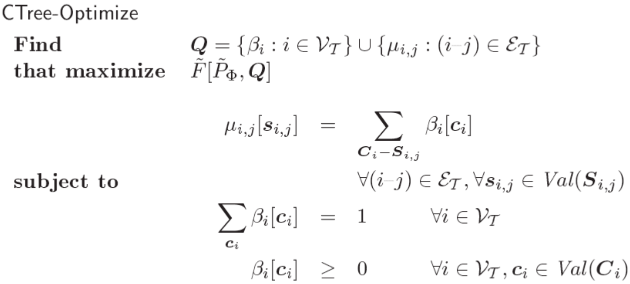
$$\mu_i \geq 0, \text{ for all } i = 1, \ldots, m$$

Complementary slackness

$$\mu_i g_i(x^*) = 0, \text{for all } i = 1, \ldots, m.$$

Our optimization problem

- Assume that the potentials are strictly positive.
- We can look for stationary points of the optimization problem



CTree-Optimize

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

that maximize $\tilde{F}[\tilde{P}_\Phi, \boldsymbol{Q}]$

subject to

$$
\begin{array}{rcll}
\mu_{i,j}[\boldsymbol{s}_{i,j}] & = & \displaystyle\sum_{\boldsymbol{C}_i - \boldsymbol{S}_{i,j}} \beta_i[\boldsymbol{c}_i] & \\
 & & \forall (i–j) \in \mathcal{E}_{\mathcal{T}}, \forall \boldsymbol{s}_{i,j} \in Val(\boldsymbol{S}_{i,j}) & \\
\displaystyle\sum_{\boldsymbol{c}_i} \beta_i[\boldsymbol{c}_i] & = & 1 & \forall i \in \mathcal{V}_{\mathcal{T}} \\
\beta_i[\boldsymbol{c}_i] & \geq & 0 & \forall i \in \mathcal{V}_{\mathcal{T}}, \boldsymbol{c}_i \in Val(\boldsymbol{C}_i)
\end{array}
$$

- In this case there is a single maximum.
- We use the method of Lagrange multipliers to characterize the stationary points.

Formal statement

Theorem: A set of beliefs **Q** is a stationary point of the C-Tree-Optimize algorithm iff there exist a set of factors $\{\delta_{i \to j}(\mathbf{S}_{i,j}) : (i - j) \in \mathcal{E}_{\mathcal{T}}\}$ such that

$$\delta_{i \to j} \propto \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \psi_i \left(\prod_{k \in Nb_i - \{j\}} \delta_{k \to i} \right)$$

and moreover we have

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

Proof: In the next set of slides by means of the method of Lagrange multipliers.

Lagrangian

- We don't need to impose the constraint that the beliefs are positive when the factors are positive, as this will already be satisfied.
- We write the Lagrangian as

$$L = \hat{F}(\hat{P}_\Phi, Q) - \sum_{i \in \mathcal{V}_\mathcal{T}} \lambda_i \left(\sum_{\mathbf{c}_i} \beta_i(\mathbf{c}_i) - 1 \right) - \sum_i \sum_{j \in Nb_i} \sum_{\mathbf{s}_{i,j}} \lambda_{j \to i}(\mathbf{s}_{i,j}) \left(\sum_{\mathbf{c}_i - \mathbf{s}_{i,j}} \beta_i(\mathbf{c}_i) - \mu_{i,j}(\mathbf{s}_{i,j}) \right)$$

where Nb_i is the number of neighbors of $\mathbf{C}_i$ in the clique tree.

- Two types of Lagrange multipliers: marginalization constrains and for sum to one.
- The Lagrangian L is a function of $\{\beta_i\}$, $\{\mu_{i,j}\}$ and the Lagrange multipliers $\{\lambda_i\}$, $\{\lambda_{i \to j}\}$.
- To find the maximum of the Lagrangian, we take its partial derivatives with respect to $\beta_i(\mathbf{c}_i)$, $\mu_{i,j}(\mathbf{s}_{i,j})$ and the Lagrange multipliers.

Stationary points

- The derivatives are

$$\begin{aligned}\frac{\partial L}{\partial \beta_i(\mathbf{c}_i)} &= \ln \psi(\mathbf{c}_i) - \ln \beta_i(\mathbf{c}_i) - 1 - \lambda_i - \sum_{j \in Nb_i} \lambda_{j \to i}(\mathbf{s}_{i,j}) \\ \frac{\partial L}{\partial \mu_{i,j}(\mathbf{s}_{i,j})} &= \ln \mu_{i,j}(\mathbf{s}_{i,j}) + 1 + \lambda_{i \to j}(\mathbf{s}_{i,j}) + \lambda_{j \to i}(\mathbf{s}_{i,j})\end{aligned}$$

- At the stationary point these derivatives are zero, so we get

$$\begin{aligned}\beta_i(\mathbf{c}_i) &= \exp\{-1 - \lambda_i\}\psi_i(\mathbf{c}_i) \prod_{j \in Nb_i} \exp(-\lambda_{j \to i}(\mathbf{s}_{i,j}) \\ \mu_{i,j}(\mathbf{s}_{i,j}) &= \exp\{-1\} \exp\{-\lambda_{i \to j}(\mathbf{s}_{i,j}) \exp\{-\lambda_{j \to i}(\mathbf{s}_{i,j})\}\end{aligned}$$

- The beliefs are functions of the form $\exp\{\lambda_{i \to j}(\mathbf{s}_{i,j})\}$, and $\mu_{i,j}(\mathbf{s}_{i,j})$ is the product of two such terms.
- These play the role of messages, we define

$$\delta_{i \to j}(\mathbf{s}_{i,j}) \triangleq \exp\{-\lambda_{i \to j}(\mathbf{s}_{i,j}) - \frac{1}{2}\}$$

Deriving message passing

- We can now write

$$\begin{aligned}\beta_i(\mathbf{c}_i) &= \exp\{-\lambda_i - 1 + \frac{1}{2}|Nb_i|\}\psi_i(\mathbf{c}_i) \prod_{j \in Nb_i} \delta_{j \to i}(\mathbf{s}_{i,j}) \\ \mu_{i,j}(\mathbf{s}_{i,j}) &= \delta_{i \to j}(\mathbf{s}_{i,j})\delta_{j \to i}(\mathbf{s}_{i,j})\end{aligned}$$

- Combining this with the marginalization over the sepset we have

$$\begin{aligned}\delta_{i \to j}(\mathbf{s}_{i,j}) &= \frac{\mu_{i,j}(\mathbf{s}_{i,j})}{\delta_{j \to i}(\mathbf{s}_{i,j})} = \frac{\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{C}_i, \mathbf{s}_{i,j})}{\delta_{j \to i}(\mathbf{s}_{i,j})} \\ &= \exp\{-\lambda_i - 1 + \frac{1}{2}|Nb_i|\} \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \psi(\mathbf{c}_i) \prod_{k \in Nb_i - \{j\}} \delta_{k \to i}(\mathbf{s}_{i,k})\end{aligned}$$

- The messages $\delta_{i \to j}$ depend on other messages, and $\exp\{-\lambda_i - 1 + \frac{1}{2}|Nb_i|\}$ is a constant.
- Combining this with $\sum_{\mathbf{c}_i} \beta_i(\mathbf{c}_i) = 1$, we can solve for the λ_i to ensure that this constant normalizes the β_i.

Formal statement and algorithm

Theorem: A set of beliefs $\mathbf{Q}$ is a stationary point of the C-Tree-Optimize algorithm iff there exist a set of factors $\{\delta_{i\to j}(\mathbf{S}_{i,j}) : (i-j) \in \mathcal{E}_\mathcal{T}\}$ such that

$$\delta_{i\to j} \propto \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \psi_i \left(\prod_{k \in Nb_i - \{j\}} \delta_{k \to i} \right)$$

and moreover we have

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

- The fix point equations define the relationship that must hold when we find the optimal Q.
- We can apply the equation as assignments and define an algorithm (init messages to 1).
- We can guarantee that this converges to a solution satisfying all equations.
- A particular order reconstructs the sum-product algorithm.