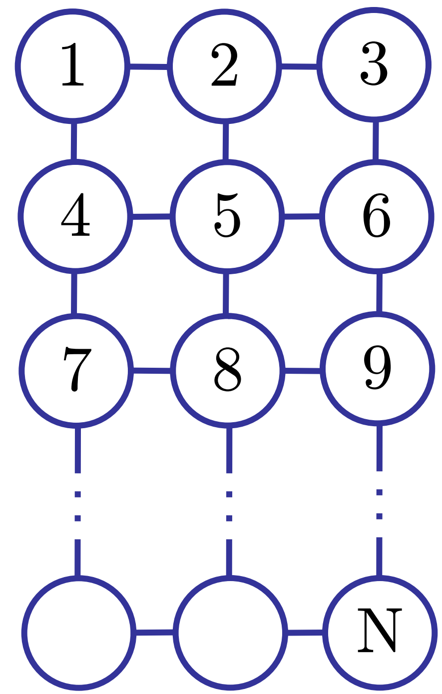
Constrained Approximate Maximum Entropy Learning (CAMEL)

Varun Ganapathi, David Vickrey, John Duchi, Daphne Koller

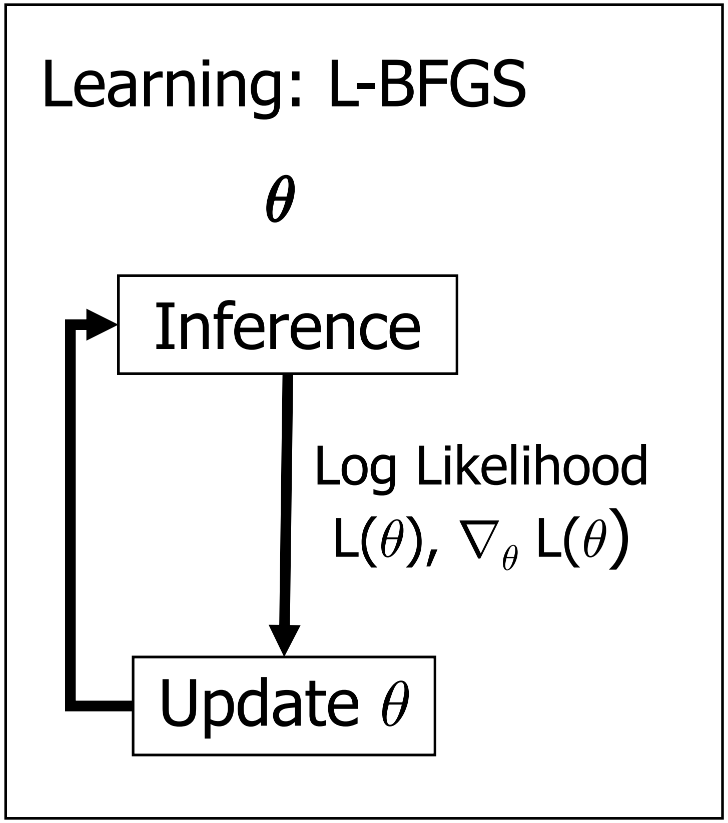
Stanford University

Undirected Graphical Models

- Undirected graphical model:
  - Random vector: $(X_1, X_2, \ldots, X_N)$
  - Graph G = (V,E) with N vertices
  - θ: Model parameters
- Inference
  - Intractable when densely connected
  - Approximate Inference (e.g., BP) can work well
- How to learn θ given data?



Maximizing Likelihood with BP

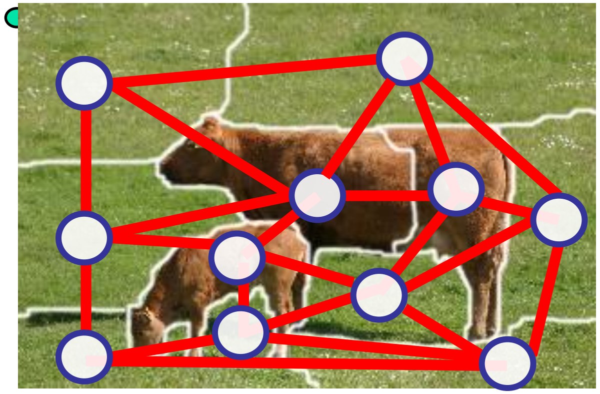


- MRF Likelihood is convex
- CG/LBFGS
- Estimate gradient with BP*
  - BP is finding fixed point of non-convex problem
    - Multiple local minima
    - Convergence
- Unstable double-loop learning algorithm

* Shental et al., 2003; Taskar et al., 2002; Sutton & McCallum, 2005

Multiclass Image Segmentation

- Goal: Image segmentation & labeling
- Model: Conditional Random Field
  - Nodes: Superpixel class labels
  - Edges: Dependency relations
- Dense network with tight loops
- Potentials => BP converges anyway
- However, BP in inner loop of learning almost never converges



Simplified Example

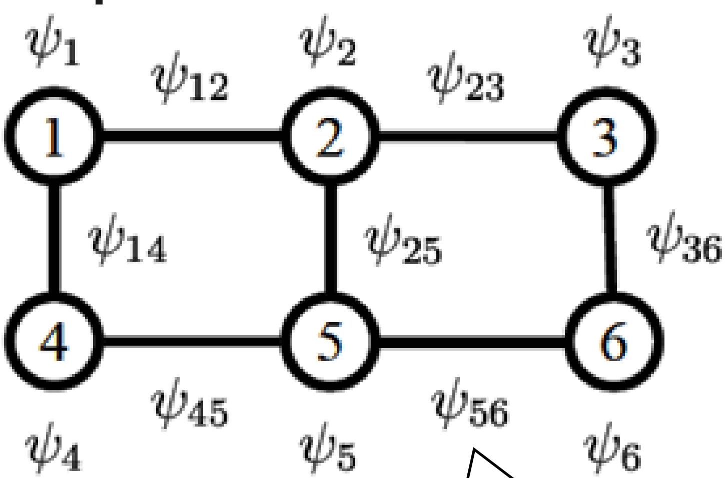
(Gould et al., Multi-Class Segmentation with Relative Location Prior, IJCV 2008)

Our Solution

Unified variational objective for parameter learning

- Can be applied to any entropy approximation
- Convergent algorithm for non-convex entropies
- Accomodates parameter sharing, regularization, conditional training
- Extends several existing objectives/methods
  - Piecewise training (Sutton and McCallum, 2005)
  - Unified propagation and scaling (Teh and Welling, 2002)
  - Pseudo-moment matching (Wainwright et al, 2003)
  - Estimating the wrong graphical model (Wainwright, 2006)

Log Linear Pairwise MRFs



Edge Potentials

Node Potentials

$$\psi_c = \exp(\theta^T F(x_c))$$

$$p(x) \propto \prod_{c \in \mathcal{C}} \psi_c(x_c)$$

$$\boldsymbol{\pi} = \{\pi_c(x_c) \ : \ c \in \mathcal{C}\}$$

Cliques

(pseudo) marginals

All results apply to general MRFs

Maximum Entropy

$$\begin{aligned} \text{maximize}_Q \quad & \mathrm{H}_Q(\boldsymbol{X}) && \text{Entropy} \\ \text{subject to} \quad & \mathrm{E}_Q[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] && \text{Moment Matching} \\ & \textstyle\sum_{\boldsymbol{x}} Q(\boldsymbol{x}) = 1 && \text{Normalization} \\ & Q(\boldsymbol{x}) \geq 0 && \text{Non-negativity} \end{aligned}$$

- Equivalent to maximum likelihood
- Intuition
- Regularization and conditional training can be handled easily (see paper)
- Q is exponential in number of variables

Maximum Entropy

$$
\begin{aligned}
\text{maximize}_Q \quad & \mathrm{H}_Q(\boldsymbol{X}) \\
\text{subject to} \quad & \mathrm{E}_Q[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] \\
& \textstyle\sum_{\boldsymbol{x}} Q(\boldsymbol{x}) = 1 \\
& Q(\boldsymbol{x}) \geq 0
\end{aligned}
$$

Entropy

Moment Matching

Normalization

Non-negativity

Marginals

$$
\begin{aligned}
\text{maximize}_\pi \quad & \textstyle\sum_c n_c \mathrm{H}(\pi_c) \\
\text{subject to} \quad & \mathrm{E}_{\boldsymbol{\pi}}[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] \\
& \textstyle\sum_{x_c} \pi_c[x_c] = 1 \\
& \pi \geq 0 \\
& \textstyle\sum_{x_i} \pi_{ij}[x_i, x_j] = \pi_j[x_j]
\end{aligned}
$$

Approximate Entropy

Moment Matching

Normalization

Non-negativity

Local Consistency

CAMEL

$$\begin{array}{ll} \text{maximize}_\pi & \sum_c n_c \mathrm{H}(\pi_c) \\ \text{subject to} & \mathrm{E}_{\boldsymbol{\pi}}[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] \\ & \sum_{x_c} \pi_c[x_c] = 1 \\ & \pi \geq 0 \\ & \sum_{x_i} \pi_{ij}[x_i, x_j] = \pi_j[x_j] \end{array}$$

Approximate Entropy
Moment Matching
Normalization
Non-negativity
Local Consistency

- Concavity depends on counting numbers n_c
- Bethe (non-concave):
  - Singletons: n_c = 1 - deg(x_i)
  - Edge Cliques: n_c = 1

Simple CAMEL

$$
\begin{array}{lll}
\text{maximize}_\pi & \sum_c \mathrm{H}(\pi_c) & \text{Approximate Entropy} \\
\text{subject to} & \mathrm{E}_{\boldsymbol{\pi}}[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] & \text{Moment Matching} \\
& \sum_{x_c} \pi_c[x_c] = 1 & \text{Normalization} \\
& \pi \geq 0 & \text{Non-negativity} \\
& \sum_{x_i} \pi_{ij}[x_i, x_j] = \pi_j[x_j] & \text{Local Consistency}
\end{array}
$$

- Simple concave objective:
  - for all c, $n_c = 1$

Piecewise Training*

maximize_{π}	$\sum_c \mathrm{H}(\pi_c)$	Approximate Entropy
subject to	$\mathrm{E}_{\boldsymbol{\pi}}[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}]$	Moment Matching
	$\sum_{x_c} \pi_c[x_c] = 1$	Normalization
	$\pi \geq 0$	Non-negativity
	$\cancel{\sum_{x_i} \pi_{ij}[x_i, x_j] = \pi_j[x_j]}$	~~Local Consistency~~

- Simply drop the marginal consistency constraints
- Dual objective is the sum of local likelihood terms of cliques

* Sutton & McCallum, 2005

Convex-Concave Procedure

- Objective:
  Convex(x) + Concave(x)
- Used by Yuille, 2003
- Approximate Objective:
  $g^T x$ + Concave(x)
- Repeat:
  - Maximize approximate objective
  - Choose new approximation
- Guaranteed to converge to fixed point

Algorithm

- Repeat
  - Choose g to linearize about current point

$$
\begin{array}{ll}
\text{maximize}_\pi & \sum_c \mathrm{H}(\pi_c) + g^T \boldsymbol{\pi} \\
\text{subject to} & \mathrm{E}_{\boldsymbol{\pi}}[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] \\
& \sum_{x_c} \pi_c[x_c] = 1 \\
& \pi \geq 0 \\
& \sum_{x_i} \pi_{ij}[x_i, x_j] = \pi_j[x_j]
\end{array}
$$

Approximate Entropy

Moment Matching

Normalization

Non-negativity

Local Consistency

  - Solve unconstrained dual problem

Dual Problem

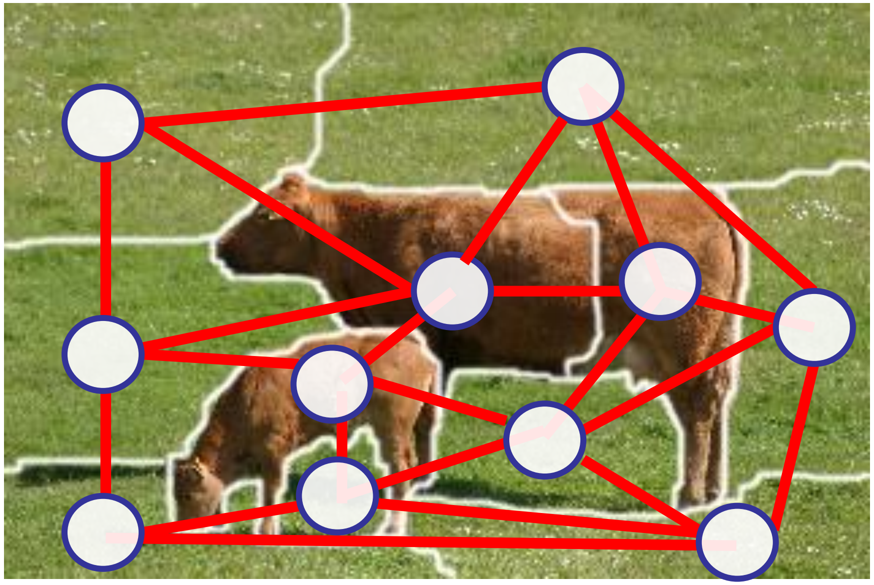
- Sum of local likelihood terms
  - Similar to multiclass logistic regression
  - g is a bias term for each cluster
  - Local consistency constraints reduce to another feature
  - Lagrange multipliers that correspond to weights and messages
- Simultaneous inference and learning
  - Avoids problem of setting convergence threshold

Experiments

- Algorithms Compared:
  - Double loop with BP in inner loop
    - Residual Belief Propagation (Elidan et al., 2006)
    - Save messages between calls
    - Reset messages during line search
    - 10 restarts with random messages
  - Camel + Bethe
  - Simple Camel
  - Piecewise (Simple Camel w/o local consistency)
- All used L-BFGS (Zhu et al, 1997)
- BP at test time

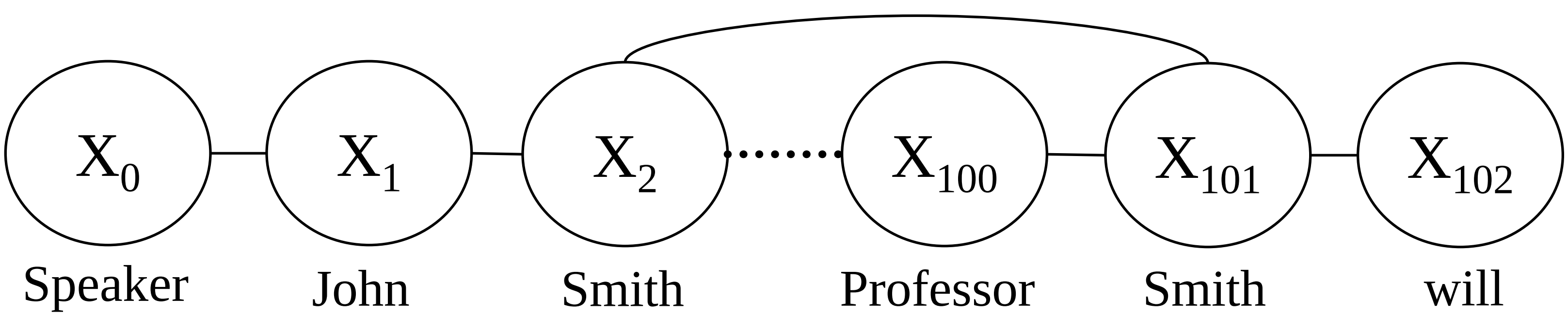
Segmentation

- Variable for each superpixel
  - 7 Classes: Rhino,Polar Bear, Water, Snow, Vegetation, Sky, Ground
- 84 parameters
- Lots of loops
- Densely connected

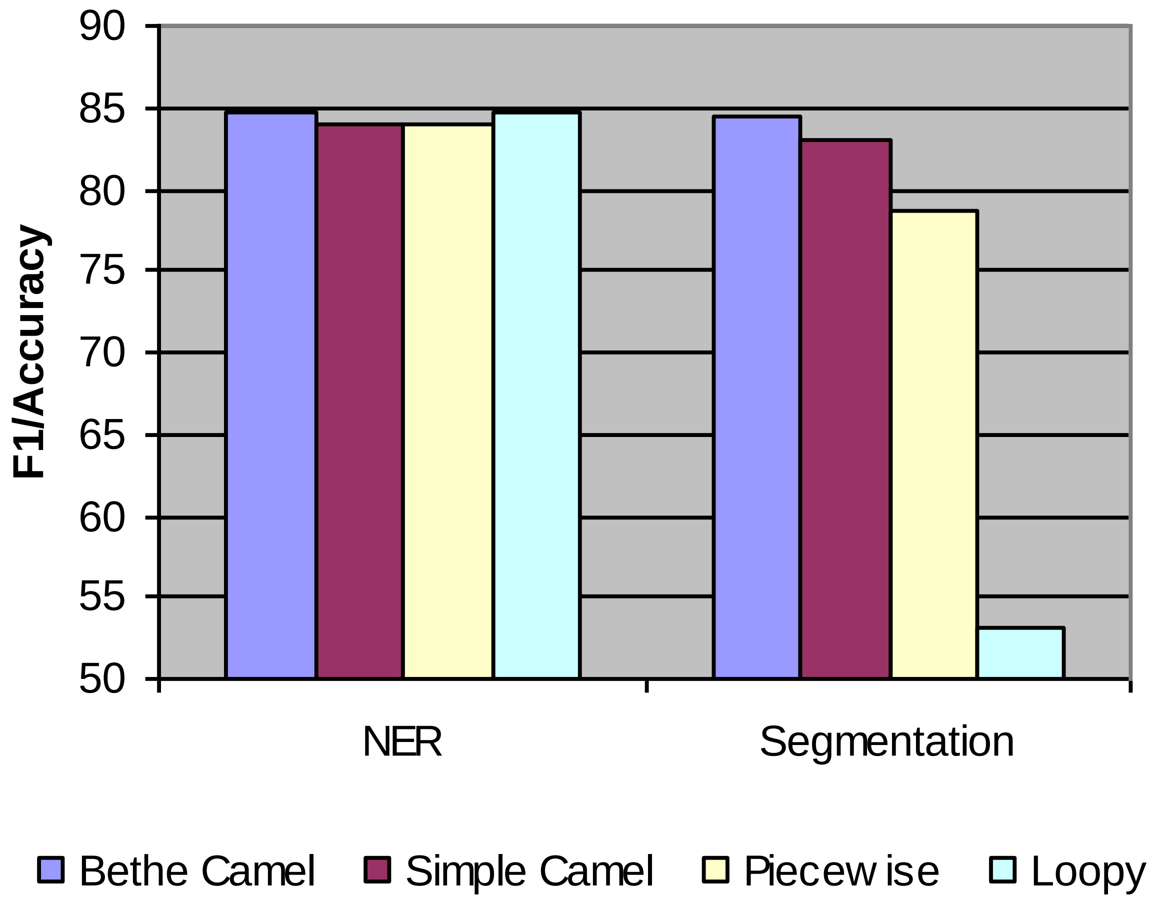


Named Entity Recognition

- Variable for each word
  - 4 Classes: Person, Location, Organization, Misc.
- Skip Chain CRF (Sutton and McCallum, 2004)
  - Words connected in a chain
  - Long-range dependencies for repeated words
- ~400k features, ~3 million weights

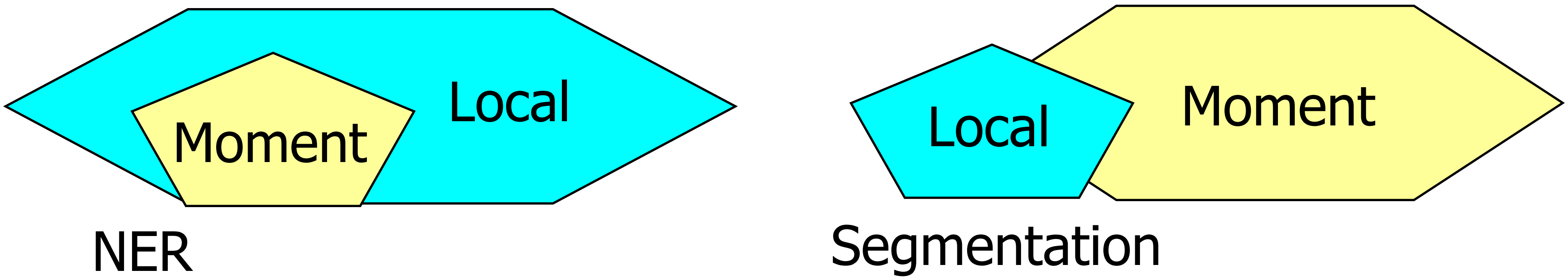


Results



- Small number of relinearizations (<10)

Discussion



- Local consistency constraints add **good** bias
- NER has millions of moment-matching constraints
  - Moment matching $\Rightarrow$ learned distribution $\approx$ empirical $\Rightarrow$ local consistency naturally satisfied
- Segmentation has only 84 parameters
  - $\Rightarrow$ Local consistency rarely satisified

Conclusions

- CAMEL algorithm unifies learning and inference
  - Optimizes **Bethe** approximation to entropy
  - Repeated convex optimization with simple form
    - Only few iterations required (can stop early too!)
  - Convergent
  - Stable
- Our results suggest that constraints on the probability distribution are more important to learning than the entropy approximations

Future Work

- For inference, evaluate relative benefit of approximations to entropy and constraints
- Learn with tighter outer bounds on marginal polytope
- New optimization methods to exploit structure of constraints

Related Work

- Unified Propagation and Scaling-Teh & Welling, 2002
  - Similar idea in using Bethe entropy and local constraints for learning
  - No parameter sharing, conditional training and regularization
  - Optimization (updates one coordinate at a time) procedure does not work well when there is large amount of parameter sharing
- Pseudo-moment matching-Wainwright et al, 2003
  - No parameter sharing, conditional training, and regularization
  - Falls out of our formulation because it corresponds to case where there is only one feasible point in the moment-matching constraints

Running Time

- NER dataset
  - piecewise is about twice as fast
- Segmentation dataset
  - Pay large cost because you have many more dual parameters (several per edge)
  - But you get an improvement

LBP as Optimization

- Bethe Free Energy

$$E_Q \sum_l \log \psi_l - \sum_i \mathrm{H}(\pi_i) - \sum_{ij} H\Big(\sum_{\boldsymbol{c}_i \setminus \boldsymbol{s}_{ij}} \pi_i(\boldsymbol{c}_i) \Big)$$

- Constraints on pseudo-marginals
  - Pairwise Consistency: $\sum_x \pi_{ij} = \pi_j$
  - Local Normalization: $\sum \pi_i = 1$
  - Non-negativity: $\pi_i \geq 0$

Optimizing Bethe CAMEL

Solve

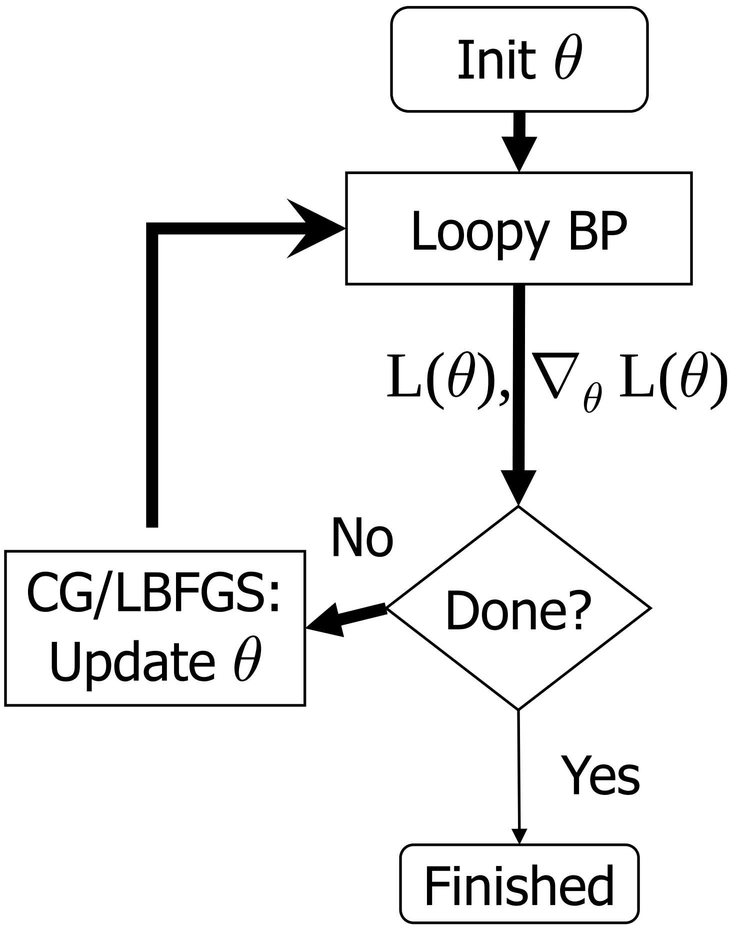
$$\begin{array}{ll} \text{maximize}_\pi & \sum_{i,j \in E} c_{ij} \mathrm{H}(\pi_{ij}) + \sum_{i \in V} c_i \mathrm{H}(\pi_i) + \mathbf{g}^T \pi \\ \text{subject to} & \mathrm{E}_{\boldsymbol{\pi}}[\boldsymbol{f}] = \mathrm{E}_{\hat{\mathrm{P}}}[\boldsymbol{f}] \\ & \sum_{x_c} \pi_c[x_c] = 1 \\ & \pi \geq 0 \\ & \sum_{x_i} \pi_{ij}[x_i, x_j] = \pi_j[x_j] \end{array}$$

Relinearize

$$\mathrm{g} \leftarrow \nabla_\pi \left(\sum_i \deg(i)\, \mathrm{H}(\pi_i)\right) [\pi^*]$$

Similar concept used in CCCP algorithm (Yuille et al, 2002)

Maximizing Likelihood with BP



- Goal:
  - Maximize likelihood of data
- Optimization difficult:
  - Inference doesn't converge
  - Inference has multiple local minima
  - CG/LBFGS fail!

Loopy BP searches for a fixed point of a non-convex problem (Yedidia et. al, Generalized Belief Propagation, 2002)