Lecture5. Note Title 1/18/2010

Belief Propagation.

Limitations of MFT → doesn't deal well with pairwise responses

→ change free energy to

$$\mathcal{F}_{bethe}(b) = \sum_{ij} \sum_{x_i,x_j} b_{ij}(x_i,x_j)\,\psi_{ij}(x_i,x_j) + \sum_i \sum_{x_i} b_i(x_i)\,\phi_i(x_i)$$

+ Entropy Term!

Here we explicitly have pairwise terms $b_{ij}(x_i,x_j)$.

The pseudomarginals are $\{ b_{ij}(x_i,x_j), b_i(x_i)\}$.

Problems: (i) how do we combine the pseudomarginals together to form a consistent distribution?

(ii) how do we define the entropy?

There are two (related) solutions. Both require defining an entropy term of form:

$$\sum_{i,j} \rho_{ij} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log b_{ij}(x_i,x_j) + \sum_i \rho_i \sum_{x_i} b_i(x_i) \log b_i(x_i)$$

Solution (1): $\rho_{ij}=1,\ \forall ij$ $\rho_i = -(n_i-1)$, n_i no. of inputs to node i

this gives the Bethe Free Energy. → non-convex

⇒ Belief Propagation (BP) algorithm.

Solution (2): $\rho_i = 1, \forall i$, & $\rho_{ij} \leq 1$.

chosen so that the Free Energy is convex

⇒ TRW Tree-Reweighted Algorithm & variants

Note: in both cases the "entropy" is an approximation.

BP is more famous than TRW (+variants).

But TRW-variants are arguably better – can give guaranteed convergence to global optima

(2) Bethe Free Energy and Belief Propagation.

BP works by "message passing"

messages $m_{ij}(x_j)$ from node i to node j & state x_j

$$\begin{cases} b_i(x_i) = \frac{1}{Z_i} e^{-\phi_i(x_i)} \prod_k m_{ki}(x_i) & Z_i \text{ normalization} \\ b_{ij}(x_i,x_j) = \frac{1}{Z_{ij}} e^{-\psi_{ij}(x_i,x_j)} \prod_{k\neq j} m_{ki}(x_i) \prod_{l\neq i} m_{lj}(x_j) \end{cases}$$

Update equation:

$$m_{ij}^{t+1}(x_j) = \sum_{x_i} e^{-\psi_{ij}(x_i,x_j) - \phi_j(x_j)} \prod_{n\neq j} m_{ni}^t(x_i)$$

Properties: (i) Algorithm will always converge to correct solution if the graph has no closed loops (tree) ~ like dynamic programming.

(ii) The algorithm will often converge to a good approximate solution if the no. of closed loops is small.

(iii) The algorithm can converge to bad results, or even fail to converge in some cases.

The algorithm looks mysterious — why the messages? — but here is some insight into it.

First: the algorithm re-parametrizes the distribution. At any time step

$$\frac{\prod_{ij} b_{ij}(x_i,x_j)}{\left(\prod_i b_i(x_i)\right)^{n_i-1}} \propto e^{-\{\sum_{ij}\psi_{ij}(x_i,x_j) + \sum_i \phi_i(x_i)\}}$$

the e-constraint admissibility

Second: at a fixed point of the algorithm it can be shown that $\sum_{x_j} b_{ij}(x_i,x_j) = b_i(x_i) \quad \forall\, i, x_i$

the m-constraint consistency

The Bethe Free Energy must be supplemented by consistency constraints

Lagrange multiplier

$$\sum_{i,j,x_j} \lambda_{ij}(x_j) \left\{ \sum_{x_i} b_{ij}(x_i,x_j) - b_j(x_j) \right\}$$

(3) It can be shown that the extrema of the Bethe free energy obey the consistency and the admissibility constraints. (and vice versa).

BP obeys the admissibility constraint at all times and converges (if it does) to a solution which satisfies the consistency constraint.

By contrast, applying CCCP to Bethe gives an algorithm that maintains the consistency constraint and converges to a state that satisfies the admissibility constraint
→ But for Bethe, the update rule for CCCP cannot be computed analytically and requires minimizing a convex energy. This is a double loop algorithm which is slower than BP (and needs the inner loop to converge).

<u>Where</u> do the message come from:

$$\mathcal{F}[\underset{\text{pseudo marginals}}{b};\underset{\text{Lagrange multipliers}}{\lambda}] = \underset{\text{Bethe}}{\mathcal{F}[b]} + \underset{\text{constraints.}}{\sum_{ij;x_j} \lambda_{ij}(x_j)\left(\sum_{x_j} b_{ij}(x_i,x_j) - b_j(x_j)\right)}$$

<u>Go</u> to the dual free energy $\hat{\mathcal{F}}[\lambda]$ by solve $\frac{\partial \mathcal{F}[b;\lambda]}{\partial b} = 0$ to obtain $\mathcal{F}(b)$.
(can be done)

<u>perform</u> dynamics in the dual space variables λ
→ observe that $\lambda_{ij}(x_j)$ look similar to messages $m_{ij}(x_j)$
can equate $\lambda_{ji}(x_i) = -\sum_{k \neq j} \log m_{ki}(x_i)$

<u>Big Problem:</u>
if $\mathcal{F}[b:\lambda]$ is convex, then $\hat{\mathcal{F}}[\lambda]$ is concave and maxima of $\hat{\mathcal{F}}[\lambda]$ correspond to minima of $\mathcal{F}[b:\lambda]$
$b^* = b(\lambda)$
Not true if $\mathcal{F}[b:\lambda]$ is non-convex. Explain why BP can fail to converge — if graph is a tree, Bethe is convex

(4) The convex free energies are better behaved.

$$F[b,\lambda] = \underset{\substack{\uparrow \\ \text{convex}}}{F[b]} + \sum_{ij, x_j} \lambda_{ij}(x_j) \left\{ \sum_{x_i} b_{ij}(x_i,x_j) - b_i(x_j) \right\}$$

Again, we can solve for $b(\lambda)$ by setting $\frac{\partial F}{\partial b} = 0$.

this gives a concave dual

$$\hat{F}[\lambda] = F[b(\lambda),\lambda].$$

Algorithm that increase the dual free energy.

- coordinate descent, cccp will give the maximum of the dual will will correspond to the minimum of the free energy.

A growing number of algorithms – similar to BP.

But this is only a bound, the solution may not be a good estimate of the marginals (except for a tree.

TCBO. Express distribution as $p(x) = \frac{1}{Z} e^{\sum_\alpha \theta_\alpha(x_\alpha)} = \prod_\alpha \psi_\alpha$ ← unary & pairwise

TRW : iterate over edges $i \to j$ (constraints on order)

(1) $$m_{i\to j}(x_j) \propto \max_{x_i} \psi_i(x_i) \psi_{ij}^{1/\rho_{ij}}(x_i,x_j) \frac{\prod_{k\in N(i)\setminus j} m_{k\to i}^{\rho_{ik}}(x_i)}{m_{j\to i}^{1-\rho_{ij}}(x_i)}$$

(2) $$b_i(x_i) \propto \psi_i(x_i) \prod_{j\in N(i)} m_{j\to i}^{\rho_{ij}}(x_i)$$

$$b_{ij}(x_i,x_j) \propto \psi_i(x_i)\psi_j(x_j)\Psi_{ij}^{1/\rho_{ij}}(x_i,x_j) \frac{\prod_{k\in N(i)\setminus j} m_{k\to i}^{\rho_{ik}}(x_i)}{m_{j\to i}^{1-\rho_{ij}}(x_i)} \frac{\prod_{k\in N(j)\setminus i} m_{k\to j}^{\rho_{jk}}(x_j)}{m_{i\to j}^{1-\rho_{ij}}(x_j)}$$

(5) Where does the convex free energy come from?

Wainwright et al.

Express: $P(\underline{x}) = e^{\underline{\theta}\cdot\underline{\phi}(\underline{x}) - \underline{\Phi}(\underline{\theta})}$

Define a distribution $\rho(T)$ over the spanning trees of the graph.

On each tree define potentials $\underline{\theta}_T$ s.t. $\sum_T \underline{\theta}_T \rho(T) = \underline{\theta}$.

Log partition $\underline{\Phi}(\underline{\theta}) = \underline{\Phi}(\sum_T \rho(T)\underline{\theta}_T) \leq \sum_T \rho(T)\underline{\Phi}(\underline{\theta}_T)$

$\underline{\Phi}(.)$ convex, Jensen's inequality.

Obtain Free energy by minimizing $\sum_T \rho(T)\underline{\Phi}(\underline{\theta}_T)$ subject to constraint $\sum_T \underline{\theta}_T \rho(T) = \underline{\theta}$.

After some manipulation this yields the convex free energy with $\underline{\theta} = \{\theta_i(x_i), \psi_{ij}(x_i,x_j)\}$. Convex since each term $\rho(T)$ in the summation is convex (since it is a tree)

Comment: exploits the idea that inference is easy over trees - so break down the graph into the set of all spaning trees.

(6) Low-Tempature → MAP

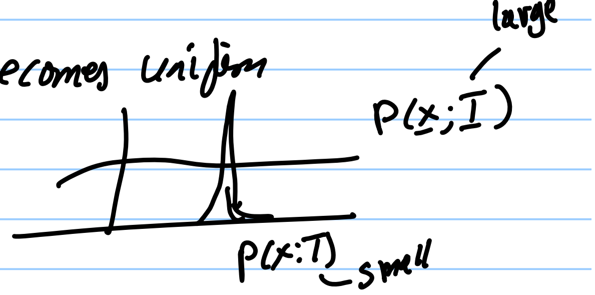
$$P(\underline{x}) \rightarrow \frac{1}{Z(T)}\{P(\underline{x})\}^{1/T}$$

$$P(\underline{x}) = \frac{1}{Z}e^{-E(\underline{x})}$$

$$P(\underline{x};T) = \frac{1}{Z[T]}e^{-E(\underline{x})/T}$$

As $T \rightarrow 0$ $P(\underline{x}:T)$ gets peaked about the lowest energy.

As $T \rightarrow 0$ $P(\underline{x};T)$ becomes uniform



Calculate free energies as a function of T

$$\mathcal{F}[B;T] = \sum_{\underline{x}} E(\underline{x})B(\underline{x}) + T\, E_{ent}(B(\underline{x}))$$

entropy term.

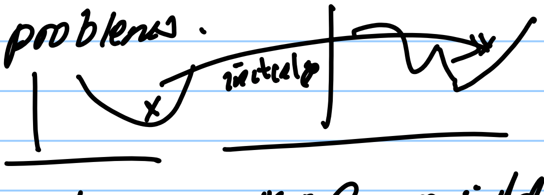
If $E_{ent}(B(x))$ is convex,
↳ then $\mathcal{F}[B:T]$ becomes convex for large T

For Bethe/Convex Free Energy. $\mathcal{F}[B:T]$ relates to a linear programming problem as $T \rightarrow 0$ (current research)

Points to be made:

- For some free energies there is a critical temperature T_c → above T_c solution is trivial.
- Continuation method — deterministic annealing

— minimize $\mathcal{F}[B,T]$ at large T, use this to initialize algorithm at smaller T. This heuristic works well for many problems. (but no guarantees)



— solution as $T \rightarrow 0$ can become MAP provided certain conditions apply.