Probabilistic & Unsupervised Learning

Expectation Propagation

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London

Term 1, Autumn 2010

Variational Recapitulation

Free energy:

$$\mathcal{F}(q,\theta) = \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q(\mathcal{Y}|\mathcal{X})} + \mathbf{H}[q] = \log P(\mathcal{X}|\theta) - \mathbf{KL}[q(\mathcal{Y})\|P(\mathcal{Y}|\mathcal{X},\theta)] \le \ell(\theta)$$

E-steps:

- Exact EM:

$$q(\mathcal{Y}) = \operatorname*{argmax}_{q} \mathcal{F} = P(\mathcal{Y}|\mathcal{X},\theta)$$

  – Saturates bound: converges to max likelihood.

- (Factored) variational approximation:

$$q(\mathcal{Y}) = \operatorname*{argmax}_{q_1(\mathcal{Y}_1)q_2(\mathcal{Y}_2)} \mathcal{F} = \operatorname*{argmin}_{q_1(\mathcal{Y}_1)q_2(\mathcal{Y}_2)} \mathbf{KL}[q_1(\mathcal{Y}_1)q_2(\mathcal{Y}_2)\|P(\mathcal{Y}|\mathcal{X},\theta)]$$

  – Increases bound: provably converges, but not necc. to ML.

- Other approximations:

$$q(\mathcal{Y}) \approx P(\mathcal{Y}|\mathcal{X},\theta)$$

  – Usually no guarantee, but if converges may be more accurate than factored var. approx.

Approximation

Makes sense to consider q closest to P in some sense.

$$q = \operatorname*{argmin}_{q \in \mathcal{Q}} D(P \| q)$$

– metric for closeness?

– constraint space $\mathcal{Q}$?

Variational methods use $D =$ **KL**$[q \| P]$. Factored constraints lead to efficient message passing approaches. What about other divergences?

The Other KL

What about the 'other' KL ($q = \operatorname{argmin} \textbf{KL}[P\|q]$)?

Crucially, for a factored approximation the (clique) marginals are correct:

$$\begin{aligned}
\operatorname*{argmin}_{q_i} \textbf{KL}\Big[P(\mathcal{Y}|\mathcal{X})\Big\|\prod q_j(\mathcal{Y}_j|\mathcal{X})\Big] &= \operatorname*{argmin}_{q_i} - \int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X}) \log \prod_j q_j(\mathcal{Y}_j|\mathcal{X}) \\
&= \operatorname*{argmin}_{q_i} - \sum_j \int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X}) \log q_j(\mathcal{Y}_j|\mathcal{X}) \\
&= \operatorname*{argmin}_{q_i} - \int d\mathcal{Y}_i\; P(\mathcal{Y}_i|\mathcal{X}) \log q_i(\mathcal{Y}_i|\mathcal{X}) \\
&= P(\mathcal{Y}_i|\mathcal{X})
\end{aligned}$$

and the marginals are what we need for learning (although if factored over disjoint sets as in var. approx. some cliques will be missing).

Perversely, this means finding the best q for this KL is intractrable! But if we can minimise it approximately we might still get decent results.

Approximate Optimisation

The posterior distribution in a graphical model is a (normalised) product of factors:

$$P(\mathcal{Y}|\mathcal{X}) = \frac{P(\mathcal{Y},\mathcal{X})}{P(\mathcal{X})} = \frac{1}{Z}\prod_i P(y_i|\mathsf{pa}(y_i)) \propto \prod_{i=1}^{N} f_i(\mathcal{Y}_i)$$

where the $\mathcal{Y}_i$ are not necessarily disjoint. In the language of EP the f_i are called sites.

Consider q with the same factorisation, but potentially approximated sites: $q(\mathcal{Y}) \stackrel{\text{def}}{=} \prod_{i=1}^{N} \tilde{f}_i(\mathcal{Y}_i)$

Possible optimisations:

$$\min_{q(\mathcal{Y}_i)} \mathbf{KL}\left[\prod_{i=1}^{N} f_i(\mathcal{Y}_i)\middle\|\prod_{i=1}^{N} \tilde{f}_i(\mathcal{Y}_i)\right]$$ (global: intractable)

$$\min_{\tilde{f}_i(\mathcal{Y}_i)} \mathbf{KL}\left[f_i(\mathcal{Y}_i)\middle\|\tilde{f}_i(\mathcal{Y}_i)\right]$$ (local, fixed: simple, inaccurate)

$$\min_{\tilde{f}_i(\mathcal{Y}_i)} \mathbf{KL}\left[f_i(\mathcal{Y}_i)\prod_{j\neq i} \tilde{f}_j(\mathcal{Y}_i)\middle\|\tilde{f}_i(\mathcal{Y}_i)\prod_{j\neq i} \tilde{f}_j(\mathcal{Y}_i)\right]$$ (local, contextual: iterative, accurate) $\longleftarrow$ EP

Expectation Propagation (EP)

Input $f_1(\mathcal{Y}_1) \ldots f_N(\mathcal{Y}_N)$
Initialize $\tilde{f}_1(\mathcal{Y}_1) = \operatorname{argmin}_{f \in \{\tilde{f}\}} \mathbf{KL}[f_1(\mathcal{Y}_1) \| f(\mathcal{Y}_1)]$, $\tilde{f}_i(\mathcal{Y}_i) = 1$ for $i > 1$,
$q(\mathcal{Y}) \propto \prod_i \tilde{f}_i(\mathcal{Y}_i)$
repeat
  for $i = 1 \ldots N$ **do**
    Deletion: $q_{\neg i}(\mathcal{Y}) \leftarrow \dfrac{q(\mathcal{Y})}{\tilde{f}_i(\mathcal{Y}_i)} = \prod_{j \neq i} \tilde{f}_j(\mathcal{Y}_j)$
    Projection: $\tilde{f}_i^{\mathrm{new}}(\mathcal{Y}) \leftarrow \operatorname*{argmin}_{f \in \{\tilde{f}\}} \mathbf{KL}[f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \| f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y})]$
    Inclusion: $q(\mathcal{Y}) \leftarrow \tilde{f}_i^{\mathrm{new}}(\mathcal{Y}_i)\, q_{\neg i}(\mathcal{Y})$
  end for
until convergence

Expectation? Propagation?

EP is really two ideas:

- Approximation of factors, usually by “projection” to exponential families. This involves finding expected sufficient statistics, hence expectation.
- Local divergence minimization in the context of other factors. This leads to a message passing approach, hence propagation.

Local updates

Each EP update involves a KL minimisation:

$$\tilde{f}_i^{\text{new}}(\mathcal{Y}) \leftarrow \underset{f \in \{\tilde{f}\}}{\operatorname{argmin}} \, \mathbf{KL}[f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \| f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y})]$$

Write $q_{\neg i}(\mathcal{Y}) = q_{\neg i}(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_{\neg i} | \mathcal{Y}_i)$. Then:

$$\begin{aligned}
\min_f \mathbf{KL}[f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \| f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y})] \\
&= \max_f \int d\mathcal{Y}_i d\mathcal{Y}_{\neg i} \; f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \log f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \\
&= \max_f \int d\mathcal{Y}_i d\mathcal{Y}_{\neg i} \; f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_{\neg i} | \mathcal{Y}_i) \big(\log f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_i) + \log q_{\neg i}(\mathcal{Y}_{\neg i} | \mathcal{Y}_i) \big) \\
&= \max_f \int d\mathcal{Y}_i \; f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_i) \big(\log f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_i) \big) \int d\mathcal{Y}_{\neg i} \; q_{\neg i}(\mathcal{Y}_{\neg i} | \mathcal{Y}_i) \\
&= \min_f \mathbf{KL}[f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_i) \| f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}_i)]
\end{aligned}$$

$q_{\neg i}(\mathcal{Y}_i)$ is sometimes called the cavity distribution.

Message Passing

The cavity distribution (in a tree) can be further broken down into a product of terms from each neighboring clique:

$$q_{\neg i}(\mathcal{Y}_i) = \prod_{j \in \mathrm{ne}(i)} m(\mathcal{Y}_j \cap \mathcal{Y}_i)$$

Once the ith site has been approximated, the messages can be passed on to neighbouring cliques by marginalising to the shared variables (SSM example follows).

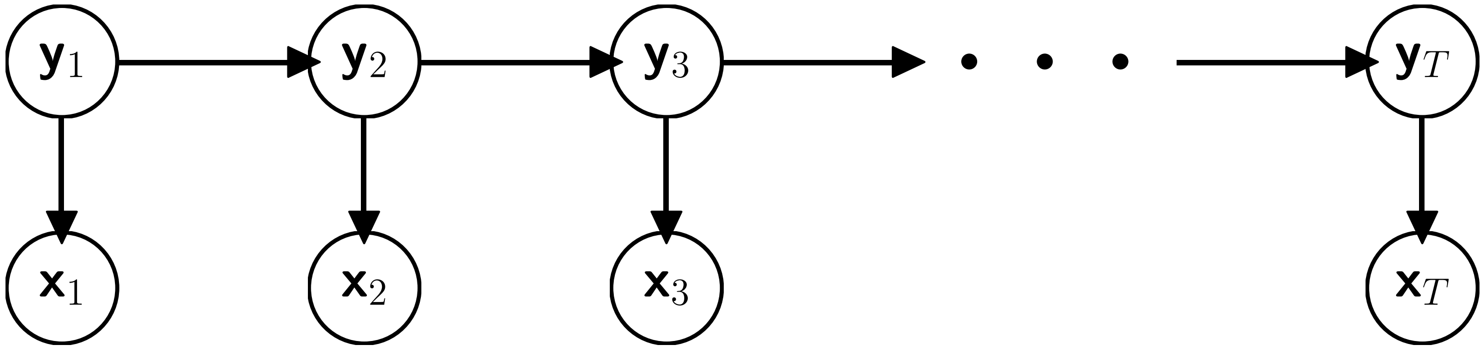
This is exactly the same as belief propagation.

In loopy graphs, we can use loopy belief propagation. In that case

$$q_{\neg i}(\mathcal{Y}_i) = \prod_{j \in \mathrm{ne}(i)} m(\mathcal{Y}_j \cap \mathcal{Y}_i)$$

becomes an approximation to the **true** cavity distribution.
For some approximations (e.g. Gaussian) may be able to compute true loopy cavity using approximate sites, even if computing exact message would have been intractable.

EP for a NLSSM



$$p(\mathbf{y}_i|\mathbf{y}_{i-1}) = \phi_i(\mathbf{y}_i, \mathbf{y}_{i-1}) \qquad \textit{e.g.}\ \exp(-\|\mathbf{y}_i - h_s(\mathbf{y}_{i-1})\|^2/2\sigma^2)$$

$$p(\mathbf{x}_i|\mathbf{y}_i) = \psi_i(\mathbf{y}_i) \qquad \textit{e.g.}\ \exp(-\|\mathbf{x}_i - h_o(\mathbf{y}_i)\|^2/2\sigma^2)$$

Then $f_i(\mathbf{y}_i, \mathbf{y}_{i-1}) = \phi_i(\mathbf{y}_i, \mathbf{y}_{i-1})\psi_i(\mathbf{y}_i)$. As ϕ_i and ψ_i are non-linear, inference is not generally tractable. Assume $\tilde{f}_i(\mathbf{y}_i, \mathbf{y}_{i-1})$ is Gaussian. Then,

$$q_{\neg t}(\mathbf{y}_i, \mathbf{y}_{i-1}) = \sum_{\substack{\mathbf{y}_1 \cdots \mathbf{y}_{t-2} \\ \mathbf{y}_{t+1} \cdots \mathbf{y}_i}} \prod_{i' \neq i} \tilde{f}_{i'}(\mathbf{y}_{i'}, \mathbf{y}_{i'-1}) = \underbrace{\sum_{\mathbf{y}_1 \cdots \mathbf{y}_{i-2}} \prod_{i' < i} \tilde{f}_{i'}(\mathbf{y}_{i'}, \mathbf{y}_{i'-1})}_{\alpha_{i-1}(\mathbf{y}_{i-1})} \underbrace{\sum_{\mathbf{y}_{i+1} \cdots \mathbf{y}_i} \prod_{i' > i} \tilde{f}_{i'}(\mathbf{y}_{i'}, \mathbf{y}_{i'-1})}_{\beta_i(\mathbf{y}_i)}$$

with both α and β Gaussian.

$$\tilde{f}_i(\mathbf{y}_i, \mathbf{y}_{i-1}) = \underset{f \in \mathcal{N}}{\operatorname{argmin}}\, \mathbf{KL}\left[\phi_i(\mathbf{y}_i, \mathbf{y}_{i-1})\psi_i(\mathbf{y}_i)\alpha_{i-1}(\mathbf{y}_{i-1})\beta_i(\mathbf{y}_i) \middle\| f(\mathbf{y}_i, \mathbf{y}_{i-1})\alpha_{i-1}(\mathbf{y}_{i-1})\beta_i(\mathbf{y}_i)\right]$$

Moment Matching

Each EP update involves an KL minimisation:

$$\tilde{f}_i^{\mathrm{new}}(\mathcal{Y}) \leftarrow \operatorname*{argmin}_{f\in\{\tilde{f}\}} \mathbf{KL}[f_i(\mathcal{Y}_i)q_{\neg i}(\mathcal{Y})\|f(\mathcal{Y}_i)q_{\neg i}(\mathcal{Y})]$$

Usually, both $q_{\neg i}(\mathcal{Y}_i)$ and $\tilde{f}$ are in the same exponential family. Let $q(x) = \frac{1}{Z(\boldsymbol{\theta})}e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}}$. Then

$$\begin{aligned}
\operatorname*{argmin}_q \mathbf{KL}\big[p(x)\big\|q(x)\big] &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbf{KL}\left[p(x)\middle\|\frac{1}{Z(\boldsymbol{\theta})}e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}}\right] \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} -\int dx\; p(x)\log\frac{1}{Z(\boldsymbol{\theta})}e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}} \\
&= \operatorname*{argmin}_{\boldsymbol{\theta}} -\int dx\; p(x)\mathbf{S}(x)\cdot\boldsymbol{\theta} + \log Z(\boldsymbol{\theta}) \\
\frac{\partial}{\partial\boldsymbol{\theta}} &= -\int dx\; p(x)\mathbf{S}(x) + \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial}{\partial\boldsymbol{\theta}}\int dx\; e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}} \\
&= -\langle\mathbf{S}(x)\rangle_p + \frac{1}{Z(\boldsymbol{\theta})}\int dx\; e^{\mathbf{S}(x)\cdot\boldsymbol{\theta}}\mathbf{S}(x) \\
&= -\langle\mathbf{S}(x)\rangle_p + \langle\mathbf{S}(x)\rangle_q
\end{aligned}$$

So minimum is found by matching sufficient stats. This is usually moment matching.

How do we calculate $\langle\mathbf{S}(x)\rangle_p$? Low dimensional integral $\longrightarrow$ Quadrature, Laplace approx …

EP Summary

Input $f_1(\mathcal{Y}_1) \dots f_N(\mathcal{Y}_N)$
Initialize $\tilde{f}_1(\mathcal{Y}_1) = \operatorname{argmin}_{f \in \{\tilde{f}\}} \textbf{KL}[f_1(\mathcal{Y}_1) \| f(\mathcal{Y}_1)]$, $\tilde{f}_i(\mathcal{Y}_i) = 1$ for $i > 1$,
$q(\mathcal{Y}) \propto \prod_i \tilde{f}_i(\mathcal{Y}_i)$
repeat
  for $i = 1 \dots N$ **do**
    Deletion: $q_{\neg i}(\mathcal{Y}) \leftarrow \dfrac{q(\mathcal{Y})}{\tilde{f}_i(\mathcal{Y}_i)} = \prod_{j \neq i} \tilde{f}_j(\mathcal{Y}_j)$
    Projection: $\tilde{f}_i^{\text{new}}(\mathcal{Y}) \leftarrow \operatorname*{argmin}_{f \in \{\tilde{f}\}} \textbf{KL}[f_i(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y}) \| f(\mathcal{Y}_i) q_{\neg i}(\mathcal{Y})]$
    Inclusion: $q(\mathcal{Y}) \leftarrow \tilde{f}_i^{\text{new}}(\mathcal{Y}_i)\, q_{\neg i}(\mathcal{Y})$
  end for
until convergence

- Minimizes the opposite KL to variational methods.
- KL minimisation (projection) only depends on $q_{\neg i}(\mathcal{Y})$ marginalised to $\mathcal{Y}_i$.
- $\tilde{f}_i(\mathcal{Y})$ in exponential family $\rightarrow$ projection step is moment matching.
- Update order need not be sequential.
- Loopy belief propagation and assumed density filtering are special cases.
- No convergence guarantee (although convergent forms can be developed).
- The names (deletion, projection, inclusion) are not the same as in (Minka, 2001).

More...

- EP for GP classification.
- Computing moments:
  - Often exact computational possible
  - Numerical quadrature $\Rightarrow$ "unscented" methods
- Other projection methods:
  - Laplace $\Rightarrow$ Laplace propagation
- Computing normalisers.
  - "Unnormalised KL":

$$\mathbf{KL}[p\|q] = \int dx\, p(x) \log \frac{p(x)}{q(x)} + \int dx\, \big(q(x) - p(x)\big)$$

    equivalent to (separately) keeping track of site integrals.

More...

- Inconsistent updates:
  - skipping
  - partial steps
  - power EP
- Alpha divergences

$$
\begin{aligned}
D_\alpha[p\|q] &= \frac{1}{\alpha(1-\alpha)} \int dx\, \alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha} \\
D_{-1}[p\|q] &= \frac{1}{2} \int dx\, \frac{(p(x)-q(x))^2}{p(x)} \\
\lim_{\alpha\to 0} D_\alpha[p\|q] &= \textbf{KL}[q\|p] \\
D_{\frac{1}{2}}[p\|q] &= 2 \int dx\, (p(x)^{\frac{1}{2}} - q(x)^{\frac{1}{2}})^2 \\
\lim_{\alpha\to 1} D_\alpha[p\|q] &= \textbf{KL}[p\|q] \\
D_2[p\|q] &= \frac{1}{2} \int dx\, \frac{(p(x)-q(x))^2}{q(x)}
\end{aligned}
$$

More. . .