Probabilistic & Unsupervised Learning
Approximate Inference

Expectation Propagation

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

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

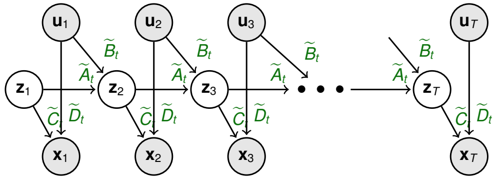
Term 1, Autumn 2023

Intractabilities and approximations

- Inference – computational intractability
  - Gibbs sampling, other MCMC
  - Factored variational approx
  - Loopy BP/EP/Power EP
  - Recognition models
- Inference – analytic intractability
  - Laplace approximation (global)
  - (Sequential) Monte-Carlo
  - Message approximations (linearised, sigma-point, Laplace)
  - Assumed-density methods and Expectation-Propagation
  - Parametric variational approx
  - Recognition models
- Learning – intractable partition function
  - Sampling parameters
  - Constrastive divergence
  - Score-matching
- Posterior estimation and model selection
  - Laplace approximation / BIC
  - Monte-Carlo
  - (Annealed) importance sampling
  - Reversible jump MCMC
  - Variational Bayes

Not a complete list!

Nonlinear state-space model (NLSSM)



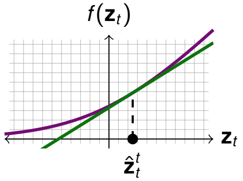
$$\mathbf{z}_{t+1} = f(\mathbf{z}_t, \mathbf{u}_t) + \mathbf{w}_t$$
$$\mathbf{x}_t = g(\mathbf{z}_t, \mathbf{u}_t) + \mathbf{v}_t$$

$\mathbf{w}_t$, $\mathbf{v}_t$ usually still Gaussian.

Extended Kalman Filter (EKF): linearise nonlinear functions about current estimate, $\hat{\mathbf{z}}_t^t$:

$$\mathbf{z}_{t+1} \approx \underbrace{f(\hat{\mathbf{z}}_t^t, \mathbf{u}_t)}_{\widetilde{B}_t\mathbf{u}_t} + \underbrace{\frac{\partial f}{\partial \mathbf{z}_t}\bigg|_{\hat{\mathbf{z}}_t^t}}_{\widetilde{A}_t} (\mathbf{z}_t - \hat{\mathbf{z}}_t^t) + \mathbf{w}_t$$

$$\mathbf{x}_t \approx \underbrace{g(\hat{\mathbf{z}}_t^{t-1}, \mathbf{u}_t)}_{\widetilde{D}_t\mathbf{u}_t} + \underbrace{\frac{\partial g}{\partial \mathbf{z}_t}\bigg|_{\hat{\mathbf{z}}_t^{t-1}}}_{\widetilde{C}_t} (\mathbf{z}_t - \hat{\mathbf{z}}_t^{t-1}) + \mathbf{v}_t$$



Run the Kalman filter (smoother) on non-stationary linearised system ($\widetilde{A}_t, \widetilde{B}_t, \widetilde{C}_t, \widetilde{D}_t$):

- Adaptively approximates non-Gaussian messages by Gaussians.
- Local linearisation depends on central point of distribution ⇒ approximation degrades with increased state uncertainty. May work acceptably for close-to-linear systems.

Can base EM-like algorithm on EKF/EKS (or alternatives).

Other message approximations

Consider the forward messages on a latent chain:

$$P(\mathbf{z}_t|\mathbf{x}_{1:t}) = \frac{1}{Z} P(\mathbf{x}_t|\mathbf{z}_t) \int d\mathbf{z}_{t-1}\, P(\mathbf{z}_t|\mathbf{z}_{t-1}) P(\mathbf{z}_{t-1}|\mathbf{x}_{1:t-1})$$

We want to approximate the messages to retain a tractable form (e.g. Gaussian).

$$\tilde{P}(\mathbf{z}_t|\mathbf{x}_{1:t}) \approx \frac{1}{Z} P(\mathbf{x}_t|\mathbf{z}_t) \int d\mathbf{z}_{t-1}\, \underbrace{P(\mathbf{z}_t|\mathbf{z}_{t-1})}_{\mathcal{N}(f(\mathbf{z}_{t-1}),\, Q)}\; \underbrace{\tilde{P}(\mathbf{z}_{t-1}|\mathbf{x}_{1:t-1})}_{\mathcal{N}(\hat{\mathbf{z}}_{t-1},\, V_{t-1})}$$

- Linearisation at the peak (EKF) is only one approach.
- Laplace filter: use mode and curvature of integrand.
- Sigma-point ("unscented") filter: next slide.
- Parametric variational:

  $$\operatorname{argmin} \mathbf{KL}\left[\mathcal{N}(\hat{\mathbf{z}}_t, \hat{V}_t) \middle\| \int d\mathbf{z}_{t-1} \ldots \right].$$

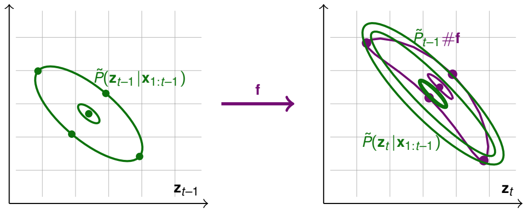
  Needs Gaussian expectations of log $\int$ ⇒ Monte-Carlo integration (later lecture).
- The other KL:

  $$\operatorname{argmin} \mathbf{KL}\left[\int d\mathbf{z}_{t-1} \ldots \middle\| \mathcal{N}(\hat{\mathbf{z}}_t, \hat{V}_t)\right]$$

  needs only first and second moments of nonlinear message ⇒ EP.

The Sigma-point filter

- Historical interest, but also a useful intuition for what comes next.


$\tilde{P}(\mathbf{z}_{t-1}|\mathbf{x}_{1:t-1})$
$\mathbf{z}_{t-1}$
$\mathbf{f}$
$\tilde{P}_{t-1}\#\mathbf{f}$
$\tilde{P}(\mathbf{z}_t|\mathbf{x}_{1:t-1})$
$\mathbf{z}_t$

- Approximates pushed-forward belief from time t–1.
- Evaluate $\mathbf{f}(\hat{\mathbf{z}}_{t-1}), \mathbf{f}(\hat{\mathbf{z}}_{t-1} \pm \sqrt{\lambda}\mathbf{v})$ for eigenvalues, eigenvectors $\hat{V}_{t-1}\mathbf{v} = \lambda\mathbf{v}$.
- "Fit" Gaussian to these $2K + 1$ $\boldsymbol{\sigma}-$points:

$$\mathcal{N}\Big(\underbrace{\tfrac{1}{2K+1} \sum_{i=1}^{2K+1} \mathbf{f}(\boldsymbol{\sigma}_i)}_{\boldsymbol{\mu}}, \tfrac{1}{2K+1} \sum_{i=1}^{2K+1} \mathbf{f}(\boldsymbol{\sigma}_i)\mathbf{f}(\boldsymbol{\sigma}_i)^{\mathsf{T}} - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathsf{T}} + \mathrm{Q}\Big)$$

- Incorporate noise process.
- Equivalent to evaluation of mean and covariance of $\tilde{P}_{t-1}\#\mathbf{f}$ by Gaussian quadrature.
- One form of "Assumed Density Filtering" (and of calculations for EP).

Variational learning

Free energy:

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

E-steps:

- Exact EM: $q(\mathcal{Z}) = \operatorname*{argmax}_q \mathcal{F} = P(\mathcal{Z}|\mathcal{X},\theta)$
  - Saturates bound: converges to local maximum of likelihood.
- (Factored) variational approximation:

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

  - Increases bound: converges, but not necessarily to ML.
- Other approximations: $q(\mathcal{Z}) \approx P(\mathcal{Z}|\mathcal{X},\theta)$
  - Usually no guarantees, but if learning converges it may be more accurate than the factored approximation

Approximating the posterior

Linearisation (or local Laplace, sigma-point and other such approaches) seem *ad hoc*. A more principled approach might look for an approximate q that is closest to P in some sense.

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

Open choices:

- form of the metric D
- nature of the constraint space $\mathcal{Q}$

- Variational methods: $D = \mathbf{KL}[q\|P]$.
  - Choosing $\mathcal{Q}$ = {tree-factored distributions} leads to efficient message passing.
- Can we use other divergences?

The other KL

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

For a factored approximation the (clique) marginals obtained by minimising this KL are correct:

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

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

Perversely, this means finding the best q for this KL is intractable!

But it raises the hope that approximate minimisation might still yield useful results.

Approximate optimisation

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

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

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

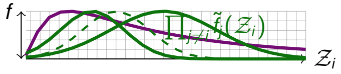
Consider q with the same factorisation, but potentially approximated sites: $q(\mathcal{Z}) \propto \prod_{i=1}^{N} \tilde{f}_i(\mathcal{Z}_i)$.

We would like to minimise (at least in some sense) $\mathbf{KL}[P\|q]$.
Possible optimisations:

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

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

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



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.

Note: we will ignore normalisation for now, but return to this later.

Local updates

Each EP update involves a KL minimisation:

$$\tilde{f}_i^{\text{new}}(\mathcal{Z}) \leftarrow \operatorname*{argmin}_{f\in\{\tilde{f}\}} \mathbf{KL}[f_i(\mathcal{Z}_i)q_{\neg i}(\mathcal{Z})\|f(\mathcal{Z}_i)q_{\neg i}(\mathcal{Z})] \qquad \left[q_{\neg i}(\mathcal{Z}) \overset{\text{def}}{=} \prod_{j\neq i}\tilde{f}_j(\mathcal{Z}_j)\right]$$

Separate the contextual factor: $q_{\neg i}(\mathcal{Z}) = q_{\neg i}(\mathcal{Z}_i)q_{\neg i}(\mathcal{Z}_{\neg i}|\mathcal{Z}_i) \qquad [\mathcal{Z}_{\neg i} \overset{\text{def}}{=} \mathcal{Z}\backslash\mathcal{Z}_i]$

Then:

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

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

Expectation Propagation (EP)

```
Input $f_1(\mathcal{Z}_1)\dots f_N(\mathcal{Z}_N)$
Initialize $\tilde{f}_1(\mathcal{Z}_1) = \operatorname{argmin}_{f\in\{\tilde{f}\}} \mathbf{KL}[f_1(\mathcal{Z}_1)\|f(\mathcal{Z}_1)]$, $\tilde{f}_i(\mathcal{Z}_i) = 1$ for $i > 1$, $q(\mathcal{Z}) \propto \prod_i \tilde{f}_i(\mathcal{Z}_i)$
repeat
  for $i = 1 \dots N$ do
    Delete: $q_{\neg i}(\mathcal{Z}) \leftarrow \frac{q(\mathcal{Z})}{\tilde{f}_i(\mathcal{Z}_i)} = \prod_{j\neq i}\tilde{f}_j(\mathcal{Z}_j)$
    Project: $\tilde{f}_i^{\text{new}}(\mathcal{Z}) \leftarrow \operatorname{argmin}_{f\in\{\tilde{f}\}} \mathbf{KL}[f_i(\mathcal{Z}_i)q_{\neg i}(\mathcal{Z}_i)\|f(\mathcal{Z}_i)q_{\neg i}(\mathcal{Z}_i)]$
    Include: $q(\mathcal{Z}) \leftarrow \tilde{f}_i^{\text{new}}(\mathcal{Z}_i)\, q_{\neg i}(\mathcal{Z})$
  end for
until convergence
```

Message Passing

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

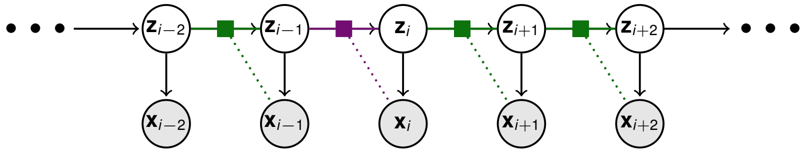
$$q_{\neg i}(\mathcal{Z}_i) = \prod_{j \in \mathrm{ne}(i)} M_{j \to i}(\mathcal{Z}_j \cap \mathcal{Z}_i)$$

- 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). $\Rightarrow$ belief propagation.
- In loopy graphs, we can use loopy belief propagation. In that case

$$q_{\neg i}(\mathcal{Z}_i) = \prod_{j \in \mathrm{ne}(i)} M_{j \to i}(\mathcal{Z}_j \cap \mathcal{Z}_i)$$

  becomes an approximation to the **true** cavity distribution (or we can recast the approximation directly in terms of messages $\Rightarrow$ later lecture).
- 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.
- In either case, message updates can be scheduled in any order.
- No guarantee of convergence (but see "power-EP" methods).

EP for a NLSSM



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

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

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

$$q_{\neg i}(\mathbf{z}_i, \mathbf{z}_{i-1}) = \int\limits_{\substack{\mathbf{z}_1 \ldots \mathbf{z}_{i-2} \\ \mathbf{z}_{i+1} \ldots \mathbf{z}_i}} \prod_{i' \neq i} \tilde{f}_{i'}(\mathbf{z}_{i'}, \mathbf{z}_{i'-1}) = \underbrace{\int\limits_{\mathbf{z}_1 \ldots \mathbf{z}_{i-2}} \prod_{i' < i} \tilde{f}_{i'}(\mathbf{z}_{i'}, \mathbf{z}_{i'-1})}_{\alpha_{i-1}(\mathbf{z}_{i-1})} \underbrace{\int\limits_{\mathbf{z}_{i+1} \ldots \mathbf{z}_n} \prod_{i' > i} \tilde{f}_{i'}(\mathbf{z}_{i'}, \mathbf{z}_{i'-1})}_{\beta_i(\mathbf{z}_i)}$$

with both α and β Gaussian.

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

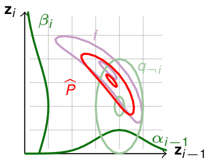
NLSSM EP message updates

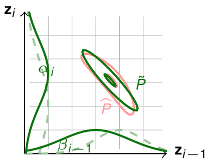
$$\tilde{f}_i(\mathbf{z}_i, \mathbf{z}_{i-1}) = \underset{f \in \mathcal{N}}{\operatorname{argmin}}\, \mathbf{KL}\left[f(\mathbf{z}_i, \mathbf{z}_{i-1})q_{\neg i}(\mathbf{z}_i, \mathbf{z}_{i-1}) \middle\| f(\mathbf{z}_i, \mathbf{z}_{i-1})q_{\neg i}(\mathbf{z}_i, \mathbf{z}_{i-1})\right] = \underset{f \in \mathcal{N}}{\operatorname{argmin}}\, \mathbf{KL}\Big[\underbrace{\phi_i(\mathbf{z}_i, \mathbf{z}_{i-1})\psi_i(\mathbf{z}_i)\alpha_{i-1}}_{\widehat{P}(\mathbf{z}_{i-1}, \mathbf{z}_i)}$$

$$\tilde{P}(\mathbf{z}_{i-1}, \mathbf{z}_i) = \underset{P \in \mathcal{N}}{\operatorname{argmin}}\, \mathbf{KL}\left[\widehat{P}(\mathbf{z}_{i-1}, \mathbf{z}_i) \middle\| P(\mathbf{z}_{i-1}, \mathbf{z}_i)\right] \qquad \tilde{f}_i(\mathbf{z}_i, \mathbf{z}_{i-1}) = \frac{\tilde{P}(\mathbf{z}_{i-1}, \mathbf{z}_i)}{\alpha_{i-1}(\mathbf{z}_{i-1})\beta_i(\mathbf{z}_i)}$$

$$\alpha_i(\mathbf{z}_i) = \int\limits_{\mathbf{z}_1 \ldots \mathbf{z}_{i-1}} \prod_{i' < i+1} \tilde{f}_{i'}(\mathbf{z}_{i'}, \mathbf{z}_{i'-1}) = \int\limits_{\mathbf{z}_{i-1}} \alpha_{i-1}(\mathbf{z}_{i-1})\tilde{f}_i(\mathbf{z}_i, \mathbf{z}_{i-1}) = \frac{1}{\beta_i(\mathbf{z}_i)} \int\limits_{\mathbf{z}_{i-1}} \tilde{P}(\mathbf{z}_{i-1}, \mathbf{z}_i)$$

$$\beta_{i-1}(\mathbf{z}_{i-1}) = \int\limits_{\mathbf{z}_{i+1} \ldots \mathbf{z}_i} \prod_{i' > i} \tilde{f}_{i'}(\mathbf{z}_{i'}, \mathbf{z}_{i'-1}) = \int\limits_{\mathbf{z}_i} \beta_i(\mathbf{z}_i)\tilde{f}_i(\mathbf{z}_i, \mathbf{z}_{i-1}) = \frac{1}{\alpha_{i-1}(\mathbf{z}_{i-1})} \int\limits_{\mathbf{z}_i} \tilde{P}(\mathbf{z}_{i-1}, \mathbf{z}_i)$$





Moment Matching

Each EP update involves a KL minimisation:

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

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

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

So minimum is found by matching sufficient stats or moment matching.

Numerical issues

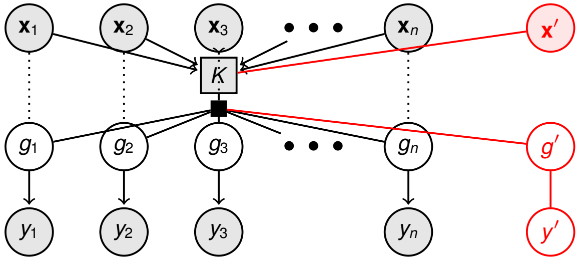
How do we calculate $\langle \mathrm{T}(x) \rangle_{\widehat{P}}$?

Often analytically tractable, but even if not requires a (relatively) low-dimensional integral:

- Quadrature methods.
  - Classical Gaussian quadrature (same Gauss, but nothing to do with the distribution) gives an iterative version of Sigma-point methods.
  - Positive definite joints, but not guaranteed to give positive definite messages.
  - Heuristics include skipping non-positive-definite steps, or damping messages by interpolation or exponentiating to power < 1.
  - Other quadrature approaches (*e.g.* GP quadrature) may be more accurate, and may allow formal constraint to pos-def cone.
- Laplace approximation.
  - Equivalent to Laplace propagation.
  - As long as messages remain positive definite will converge to global Laplace approximation.

EP for Gaussian process classification

EP provides a succesful framework for Gaussian-process modelling of non-Gaussian observations (*e.g.* for classification).

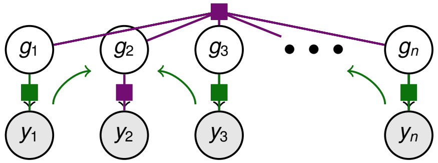


Recall:

- A GP defines a multivariate Gaussian distribution on any finite subset of random vars $\{g_1 \dots g_n\}$ drawn from a (usually uncountable) potential set indexed by "inputs" $\mathbf{x}_i$.
- The Gaussian parameters depend on the inputs: $(\boldsymbol{\mu} = [\mu(\mathbf{x}_i)],\ \Sigma = [K(\mathbf{x}_i, \mathbf{x}_j)])$.
- If we think of the gs as function values, a GP provides a prior over functions.
- In a GP regression model, noisy observations y_i are conditionally independent given g_i.
- No parameters to learn (though often hyperparameters); instead, we make predictions on test data directly: [assuming $\mu = 0$, and matrix Σ incorporates diagonal noise]

$$P(y'|\mathbf{x}', \mathcal{D}) = \mathcal{N}\left(\Sigma_{x',X}\Sigma_{X,X}^{-1}\mathbf{z},\ \Sigma_{x',x'} - \Sigma_{x',X}\Sigma_{X,X}^{-1}\Sigma_{X,x'}\right)$$

GP EP updates



- We can write the GP joint on g_i and y_i as a factor graph:

$$P(g_1 \dots g_n, y_1, \dots y_n) = \underbrace{\mathcal{N}(g_1 \dots g_n | \mathbf{0}, K)}_{f_0(\mathcal{G})} \prod_i \underbrace{\mathcal{N}\left(y_i | g_i, \sigma_i^2\right)}_{f_i(g_i)}$$

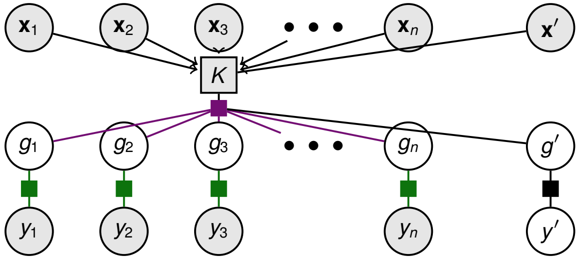
- The same factorisation applies to non-Gaussian $P(y_i|g_i)$ (*e.g.* $P(y_i{=}1) = 1/(1 + e^{-g_i})$).
- EP: approximate non-Gaussian $f_i(g_i)$ by Gaussian $\tilde{f}_i(g_i) = \mathcal{N}\left(\tilde{\mu}_i, \tilde{\psi}_i^2\right)$.
- $q_{\neg i}(g_i)$ can be constructed by the usual GP marginalisation. If $\Sigma = K + \text{diag}\left[\tilde{\psi}_1^2 \dots \tilde{\psi}_n^2\right]$

$$q_{\neg i}(g_i) = \mathcal{N}\left(\Sigma_{i,\neg i}\Sigma_{\neg i,\neg i}^{-1}\tilde{\boldsymbol{\mu}}_{\neg i},\ K_{i,i} - \Sigma_{i,\neg i}\Sigma_{\neg i,\neg i}^{-1}\Sigma_{\neg i,i}\right)$$

- The EP updates thus require calculating Gaussian expectations of $f_i(g)g^{\{1,2\}}$:

$$\tilde{f}_i^{\text{new}}(g_i) = \mathcal{N}\left(\int dg\, q_{\neg i}(g) f_i(g) g,\ \int dg\, q_{\neg i}(g) f_i(g) g^2 - (\tilde{\mu}_i^{\text{new}})^2\right) / q_{\neg i}(g_i)$$

EP GP prediction



- Once appoximate site potentials have stabilised, they can be used to make predictions.
- Introducing a test point changes K, but does not affect the *marginal* $P(g_1 \dots g_n)$ (by consistency of the GP).
- The unobserved output factor provides no information about g' ($\Rightarrow$ constant factor on g')
- Thus no change is needed to the approximating potentials $\tilde{f}_i$.
- Predictions are obtained by marginalising the approximation: [let $\tilde{\Psi} = \text{diag}\left[\tilde{\psi}_1^2 \dots \tilde{\psi}_n^2\right]$]

$$P(y'|\mathbf{x}', \mathcal{D}) = \int dg'\, P(y'|g') \mathcal{N}\Big(g' \mid K_{x',X}(K_{X,X} + \tilde{\Psi})^{-1}\tilde{\boldsymbol{\mu}},$$
$$K_{x',x'} - K_{x',X}(K_{X,X} + \tilde{\Psi})^{-1}K_{X,x'}\Big)$$

Normalisers

- As long as our approximating class is a tractable exponential family, normalisers can be computed as needed.
- Consider an approximating class written

  $$\tilde{f}_i(\mathcal{Z}_i) \propto e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_i - \Phi(\boldsymbol{\theta}_i)}$$

  i.e., define a single sufficient statistic vector on all latents, setting entries in $\boldsymbol{\theta}_i$ to 0 for suff stat functions that take cliques other than $\mathcal{Z}_i$.
- Then

  $$q(\mathcal{Z}) \propto \prod_i \tilde{f}_i \propto e^{T(\mathcal{Z})\cdot\sum\theta_i - \sum\Phi(\theta_i)}$$

  and so we can simply renormalise at the end as usual:

  $$q(\mathcal{Z}) = e^{T(\mathcal{Z})\cdot\sum\theta_i - \Phi(\sum\theta_i)}\,.$$
- However, to compute an approximation to the likelihood $\int d\mathcal{Z}\ \prod_i f_i(\mathcal{Z}_i)$ we need to keep track of the site integrals.

Computing likelihoods – keeping track of normalisers

- Define unnormalised ExpFam approximating sites $\tilde{f}_i = \tilde{C}_i e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_i}$.

  Write $\boldsymbol{\theta} = \sum\boldsymbol{\theta}_j$ for the natural parameters of $q(\mathcal{Z})$ and $\boldsymbol{\theta}_{\neg i} = \sum_{j\neq i}\boldsymbol{\theta}_j$ for the natural parameters of $q_{\neg i}(\mathcal{Z})$.

  Let $\Phi(\boldsymbol{\theta}) = \log\int e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}}$ be the (tractable) ExpFam log normaliser.
- Now, at each EP step minimise the "unnormalised KL":

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

  This matches the zeroth moment of $f_i(\mathcal{Z}_i)q_{\neg i}(\mathcal{Z})$ as well as the expected sufficient statistics as before. That is:

  $$\int \tilde{C}_i e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_i}\prod_{\neg i}\tilde{C}_j e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_j} = \int f_i(\mathcal{Z}_i)\prod_{\neg i}\tilde{C}_j e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_j} \quad\Rightarrow\quad \tilde{C}_i = e^{\Phi_i(\boldsymbol{\theta}_{\neg i}) - \Phi(\boldsymbol{\theta})}$$

  where Φ_i is the log-normaliser of the "tilted" ExpFam $\widehat{P}_i(\mathcal{Z}) \propto f(\mathcal{Z}_i)e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}}$.
- The likelihood approximation is then:

  $$\log\int\prod_{i=1}^{N} f_i(\mathcal{Z}_i) \approx \log\int\prod_{i=1}^{N}\tilde{f}_i(\mathcal{Z}_i) = \Phi(\boldsymbol{\theta}) + \sum\log\tilde{C}_i \stackrel{\text{def}}{=} \tilde{\ell}$$

Learning

EP yields approximate *inferential* posteriors. To learn (hyper)parameters we can use:

- Approximate Bayesian inference (analagous to VB)
  - may be difficult to construct a coherent normalisable exponential family approximation on both latents and parameters.
- Approximate EM – maximize $\langle\log P(\mathcal{X},\mathcal{Z})\rangle_{q_{EP}(\mathcal{Z})}$.
  - Practical, but no coherent cost function (unlike variational inference), so no guarantee of convergence even if EP itself converges.
- Direct maximisation of EP log-likelihood estimate.
  - Consistent, although convergence guarantees still difficult.
  - Seems challenging as we need to differentiate through (iteration-based) dependence of approximate $q(\mathcal{Z})$ and $\tilde{C}_i$s.
  - However, proves to be simpler than it sounds.

EP log-likelihood optimisation for learning

Let true potentials f_i depend on model (hyper)parameters η.
We have

$$\nabla_\eta\tilde{\ell} = \nabla_\eta\Phi(\boldsymbol{\theta}) + \sum_{i=1}^{N}\nabla_\eta\log\tilde{C}_i = \boldsymbol{\mu}\cdot\nabla_\eta\boldsymbol{\theta} + \sum_{i=1}^{N}\nabla_\eta\log\tilde{C}_i \qquad (*)$$

using the standard ExpFam moment-generating result with mean parameters $\boldsymbol{\mu} = \langle T(\mathcal{Z})\rangle_{q(\mathcal{Z})}$.
Now, zeroth-moment matching implies that at EP convergence:

$$\log\tilde{C}_i = \Phi_i(\boldsymbol{\theta}_{\neg i}) - \Phi(\boldsymbol{\theta}) \Rightarrow \nabla_\eta\log\tilde{C}_i = \nabla_\eta\Phi_i(\boldsymbol{\theta}_{\neg i}) - \boldsymbol{\mu}\cdot\nabla_\eta\boldsymbol{\theta} \qquad (**)$$

but $\Phi_i(\boldsymbol{\theta}_{\neg i}) = \log\int f_i(\mathcal{Z}_i)e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_{\neg i}}$ depends on η in two ways: *directly* through f_i and *indirectly* through the converged $\boldsymbol{\theta}_{\neg i}$.

$$\begin{aligned}\nabla_\eta\Phi_i(\boldsymbol{\theta}_{\neg i}) &= \partial_{\boldsymbol{\theta}_{\neg i}}\Phi_i(\boldsymbol{\theta}_{\neg i})\cdot\nabla_\eta\boldsymbol{\theta}_{\neg i} + e^{-\Phi_i(\boldsymbol{\theta}_{\neg i})}\int\nabla_\eta f_i(\mathcal{Z}_i)\ e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_{\neg i}}\,d\mathcal{Z}\\ &= \langle T(\mathcal{Z})\rangle_{\widehat{P}_i}\cdot\nabla_\eta\boldsymbol{\theta}_{\neg i} + \int\nabla_\eta\log f_i(\mathcal{Z}_i)\ f_i(\mathcal{Z}_i)e^{\mathbf{T}(\mathcal{Z})\cdot\boldsymbol{\theta}_{\neg i} - \Phi_i(\boldsymbol{\theta}_{\neg i})}\,d\mathcal{Z}\\ &= \boldsymbol{\mu}\cdot\nabla_\eta\boldsymbol{\theta}_{\neg i} + \langle\nabla_\eta\log f_i(\mathcal{Z}_i)\rangle_{\widehat{P}_i} \qquad (***)\end{aligned}$$

by EP moment matching at convergence!

EP log-likelihood optimisation for learning

So putting it all together:

$$
\begin{aligned}
\nabla_\eta \tilde{\ell} &= \boldsymbol{\mu} \cdot \nabla_\eta \boldsymbol{\theta} + \sum_{i=1}^{N} \nabla_\eta \log \tilde{C}_i && (*) \\
&= \boldsymbol{\mu} \cdot \nabla_\eta \boldsymbol{\theta} + \sum_{i=1}^{N} \Big(\nabla_\eta \Phi_i(\boldsymbol{\theta}_{\neg i}) - \boldsymbol{\mu} \cdot \nabla_\eta \boldsymbol{\theta} \Big) && (**) \\
&= \boldsymbol{\mu} \cdot \nabla_\eta \boldsymbol{\theta} + \sum_{i=1}^{N} \Big(\boldsymbol{\mu} \cdot \nabla_\eta \boldsymbol{\theta}_{\neg i} - \boldsymbol{\mu} \cdot \nabla_\eta \boldsymbol{\theta} + \langle \nabla_\eta \log f_i(\mathcal{Z}_i) \rangle_{\widehat{P}_i} \Big) && (***) \\
&= \boldsymbol{\mu} \cdot \nabla_\eta \Big(\boldsymbol{\theta} + \sum_{i=1}^{N} (\boldsymbol{\theta}_{\neg i} - \boldsymbol{\theta}) \Big) + \sum_{i=1}^{N} \langle \nabla_\eta \log f_i(\mathcal{Z}_i) \rangle_{\widehat{P}_i} \\
&= \boldsymbol{\mu} \cdot \nabla_\eta \Big(\sum_{i=1}^{N} \boldsymbol{\theta}_i + \sum_{i=1}^{N} (\boldsymbol{\theta}_{\neg i} - \boldsymbol{\theta}) \Big) + \sum_{i=1}^{N} \langle \nabla_\eta \log f_i(\mathcal{Z}_i) \rangle_{\widehat{P}_i} \\
&= \boldsymbol{\mu} \cdot \nabla_\eta \sum_{i=1}^{N} (\boldsymbol{\theta} - \boldsymbol{\theta}) + \sum_{i=1}^{N} \langle \nabla_\eta \log f_i(\mathcal{Z}_i) \rangle_{\widehat{P}_i} \\
&= \sum_{i=1}^{N} \langle \nabla_\eta \log f_i(\mathcal{Z}_i) \rangle_{\widehat{P}_i}
\end{aligned}
$$

and the gradient can be computed provided EP converges.

A final generalisation: alpha divergences and Power EP

- Alpha divergences

$$D_\alpha[p\|q] = \frac{1}{\alpha(1-\alpha)} \int dx \left(\alpha p(x) + (1-\alpha) q(x) - p(x)^\alpha q(x)^{1-\alpha} \right)$$

$$D_{-1}[p\|q] = \frac{1}{2} \int dx \, \frac{(p(x) - q(x))^2}{p(x)}$$

$$\lim_{\alpha \to 0} D_\alpha[p\|q] = \mathbf{KL}[q\|p] \qquad \text{Note: } \lim_{\alpha \to 0} \frac{(p(x)/q(x))^\alpha}{\alpha} = \log \frac{p(x)}{q(x)}$$

$$D_{\frac{1}{2}}[p\|q] = 2 \int dx \, (p(x)^{\frac{1}{2}} - q(x)^{\frac{1}{2}})^2$$

$$\lim_{\alpha \to 1} D_\alpha[p\|q] = \mathbf{KL}[p\|q]$$

$$D_2[p\|q] = \frac{1}{2} \int dx \, \frac{(p(x) - q(x))^2}{q(x)}$$

- Local (EP) minimisation gives fixed-point updates that blend messages (to power α) with previous site approximations.

$$\tilde{f}_i^{\text{new}} = \operatorname*{argmin}_{f \in \{\tilde{f}\}} \mathbf{KL}\Big[f_i(\mathcal{Z}_i)^\alpha \tilde{f}_i(\mathcal{Z}_i)^{1-\alpha} q_{\neg i}(\mathcal{Z}) \Big\| f(\mathcal{Z}_i) q_{\neg i}(\mathcal{Z}) \Big]$$

- Small changes (for $\alpha < 1$) lead to more stable updates, and more reliable convergence.