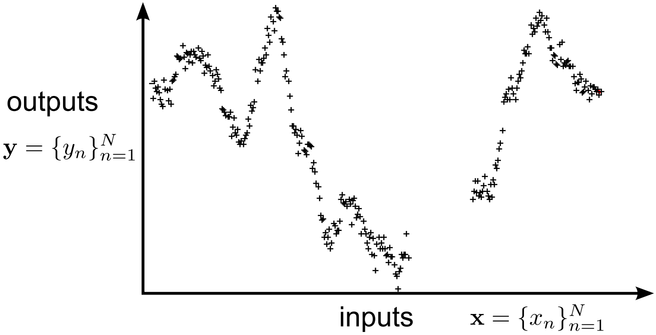
A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation

Dr. Richard E. Turner (ret26@cam.ac.uk)
Computational and Biological Learning Lab, Department of Engineering, University of Cambridge

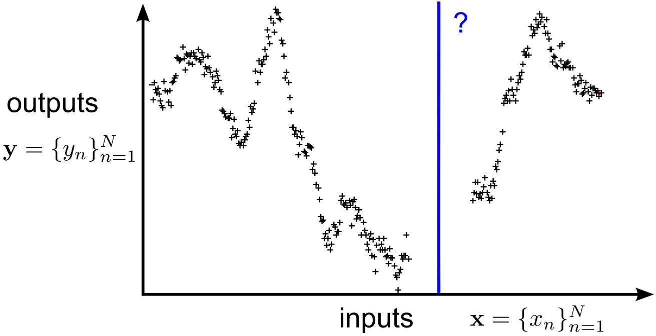
...joint work with Thang Bui, Cuong Nguyen and Josiah Yan

Manfred Opper is a God

Motivation: Gaussian Process Regression



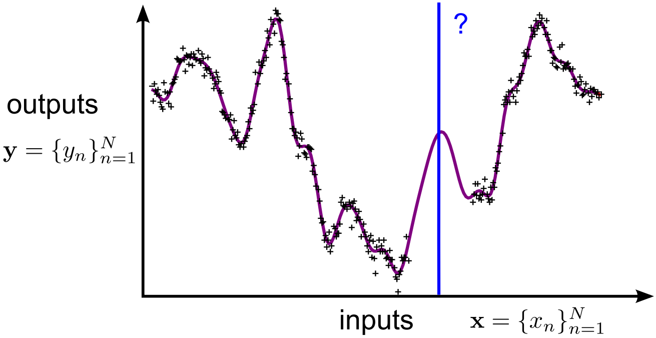
Motivation: Gaussian Process Regression



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta)$$

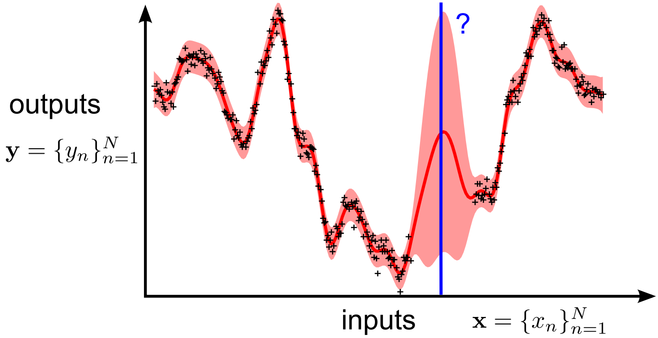
$$p(y_n|f, x_n, \theta)$$



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta) \xrightarrow{\text{inference \& learning}} p(f|\mathbf{y}, \mathbf{x}, \theta)$$

$$p(y_n|f, x_n, \theta) \qquad\qquad p(\mathbf{y}|\mathbf{x}, \theta)$$



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta)$$

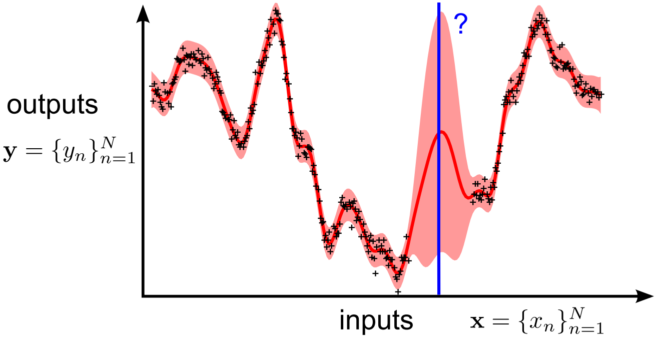
$$p(y_n|f, x_n, \theta)$$

$\xrightarrow{\text{inference \& learning}}$

intractabilities
computational $\mathcal{O}(N^3)$
analytic

$$p(f|\mathbf{y}, \mathbf{x}, \theta)$$

$$p(\mathbf{y}|\mathbf{x}, \theta)$$



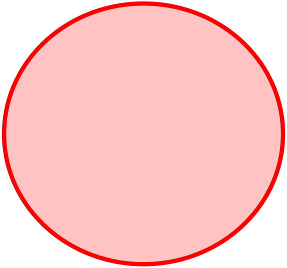
A Brief History of Gaussian Process Approximations

FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

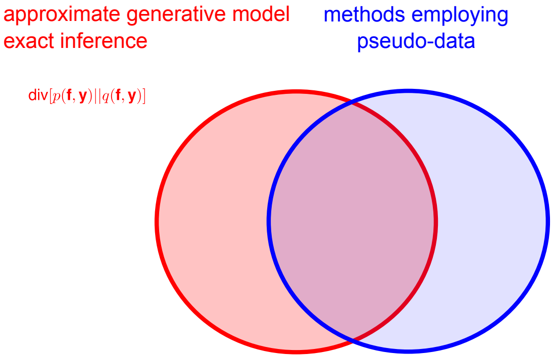
approximate generative model
exact inference

div$[p(\mathbf{f}, \mathbf{y})||q(\mathbf{f}, \mathbf{y})]$



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations



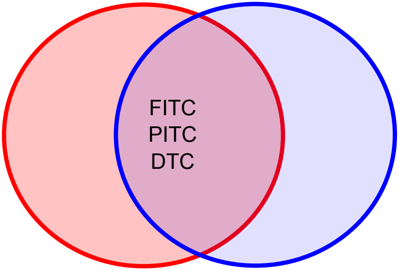
FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

div$[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

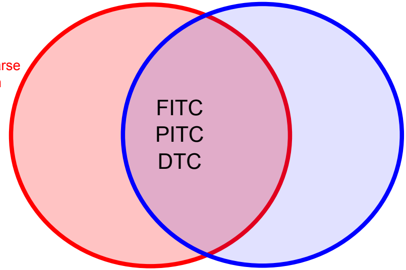
A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

div$[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

A Unifying View of Sparse Approximate Gaussian Process Regression Quinonero-Candela & Rasmussen, 2005 (FITC, PITC, DTC)



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

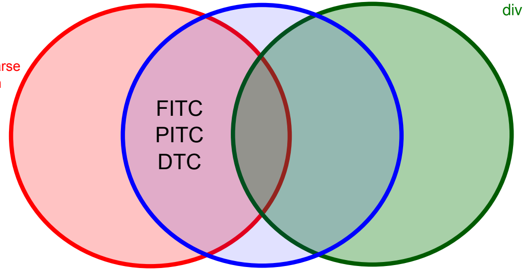
methods employing
pseudo-data

exact generative model
approximate inference

div$[p(\mathbf{f}, \mathbf{y}) || q(\mathbf{f}, \mathbf{y})]$

div$[p(\mathbf{f}|\mathbf{y}) || q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression Quinonero-Candela & Rasmussen, 2005 (FITC, PITC, DTC)



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

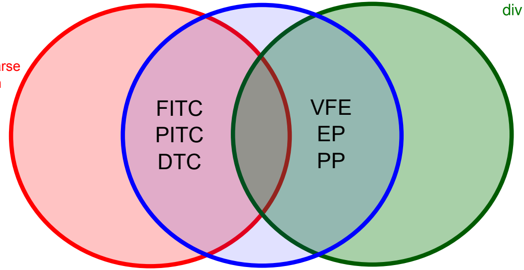
methods employing
pseudo-data

exact generative model
approximate inference

div$[p(\mathbf{f}, \mathbf{y})||q(\mathbf{f}, \mathbf{y})]$

div$[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse
Approximate Gaussian
Process Regression
Quinonero-Candela &
Rasmussen, 2005
(FITC, PITC, DTC)



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

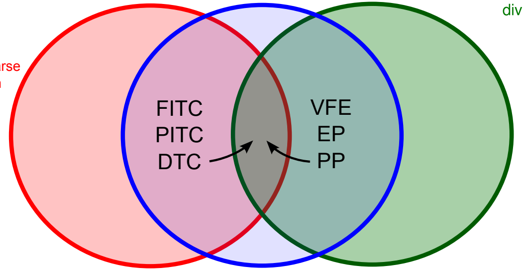
methods employing
pseudo-data

exact generative model
approximate inference

div$[p(\mathbf{f}, \mathbf{y})||q(\mathbf{f}, \mathbf{y})]$

div$[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression Quinonero-Candela & Rasmussen, 2005 (FITC, PITC, DTC)



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

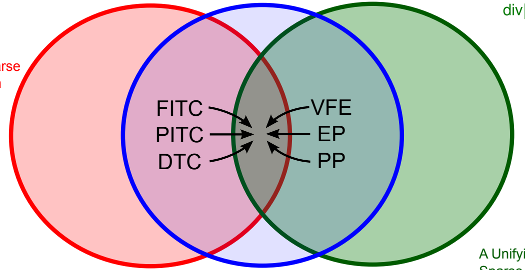
methods employing
pseudo-data

exact generative model
approximate inference

div$[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

div$[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression
Quinonero-Candela & Rasmussen, 2005
(FITC, PITC, DTC)

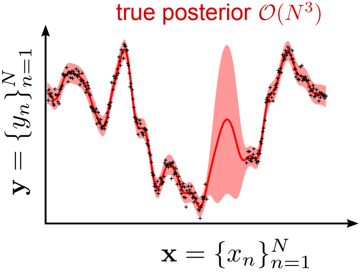


A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation
Bui, Yan and Turner, 2016
(VFE, EP, FITC, PITC ...)

FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods."
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

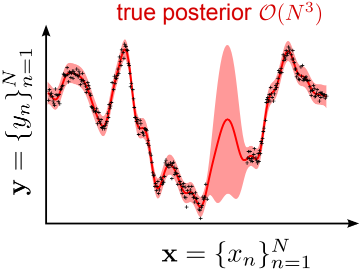
EP pseudo-point approximation

$$p^*(f) = \; p(f, \mathbf{y}|\mathbf{x}, \theta)$$



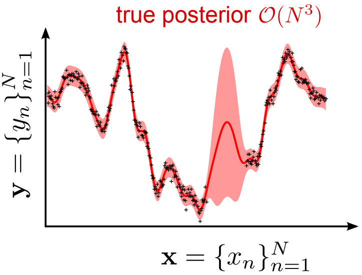
EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta)$$
$$= p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)}$$



EP pseudo-point approximation

$$
\begin{aligned}
p^*(f) &= p(f, \mathbf{y}|\mathbf{x}, \theta) \\
&= p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)} \\
&= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \; \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}
\end{aligned}
$$

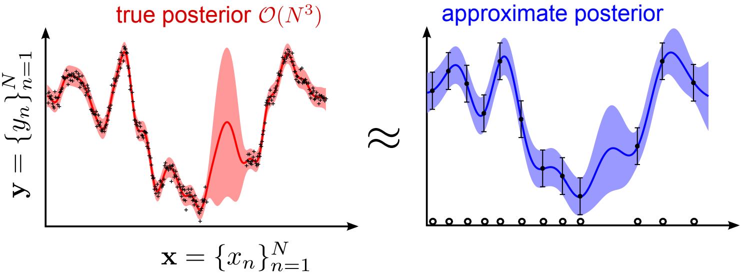


EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta)$$

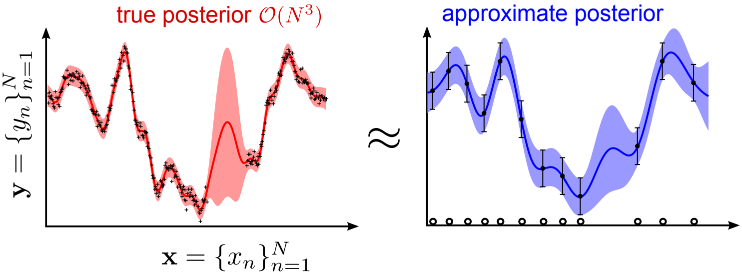
$$= p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)} \qquad q^*(f) = p(f|\theta) \prod_{n=1}^{N} \underline{t_n(f)}$$

$$= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \; \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}$$



EP pseudo-point approximation

$$
\begin{aligned}
p^*(f) &= p(f, \mathbf{y}|\mathbf{x}, \theta) \\
&= p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)} \\
&= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \; \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}
\end{aligned}
\qquad
\begin{aligned}
q^*(f) &= p(f|\theta) \prod_{n=1}^{N} \underline{t_n(f)} \\
&= \underline{Z_{\mathrm{EP}}} \; \underline{q(f)}
\end{aligned}
$$



EP pseudo-point approximation

$$
\begin{aligned}
p^*(f) &= p(f, \mathbf{y} | \mathbf{x}, \theta) \\
&= p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n | f, x_n, \theta)} \\
&= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \; \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}
\end{aligned}
$$

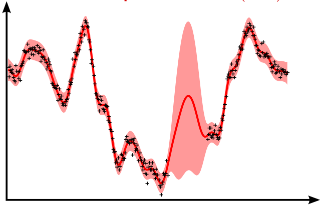
$$
\begin{aligned}
q^*(f) &= p(f|\theta) \prod_{n=1}^{N} \underline{t_n(f)} \\
&= \underline{Z_{\mathrm{EP}}} \; \underline{q(f)}
\end{aligned}
$$

$$t_n(f) = \mathcal{N}(\mathbf{u}; \mu_n, \Sigma_n)$$

$$\dim(\mathbf{u}) = M \quad f = \{\mathbf{u}, f_{\neq \mathbf{u}}\}$$

true posterior $\mathcal{O}(N^3)$

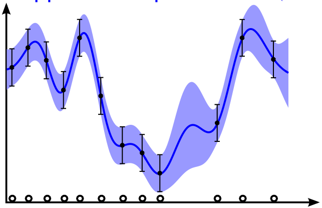
$\mathbf{y} = \{y_n\}_{n=1}^{N}$



$\mathbf{x} = \{x_n\}_{n=1}^{N}$

$\approx$

approximate posterior $\mathcal{O}(NM^2)$



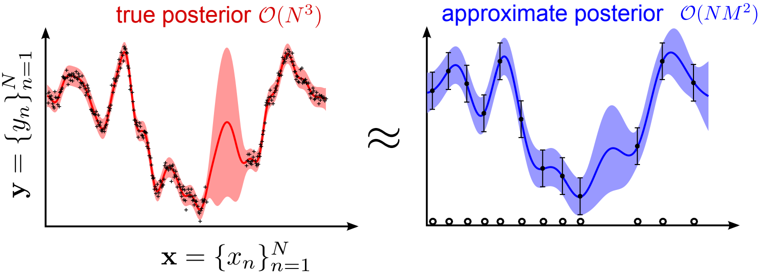
EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)} = \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}$$

$$q^*(f) = p(f|\theta) p(\tilde{\mathbf{y}}|\mathbf{u}, \tilde{\Sigma}) = p(f|\theta) \prod_{n=1}^{N} \underline{t_n(f)} = \underline{Z_{\mathrm{EP}}}\ \underline{q(f)}$$

$$t_n(f) = \mathcal{N}(\mathbf{u}; \mu_n, \Sigma_n)$$

$$\dim(\mathbf{u}) = M \quad f = \{\mathbf{u}, f_{\neq \mathbf{u}}\}$$



EP pseudo-point approximation

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta)$$
$$= p(f|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)}$$
$$= \underbrace{p(\mathbf{y}|\mathbf{x}, \theta)}_{\text{marginal likelihood}} \; \underbrace{p(f|\mathbf{y}, \mathbf{x}, \theta)}_{\text{posterior}}$$

$$q^*(f) = p(f|\theta) p(\tilde{\mathbf{y}}|\mathbf{u}, \tilde{\Sigma})$$

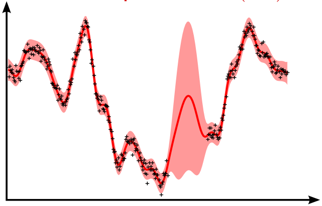
exact joint of new GP regression model

$$= p(f|\theta) \prod_{n=1}^{N} \underline{t_n(f)}$$
$$= \underline{Z_{\mathrm{EP}}} \; \underline{q(f)}$$

$$t_n(f) = \mathcal{N}(\mathbf{u}; \mu_n, \Sigma_n)$$
$$\dim(\mathbf{u}) = M \quad f = \{\mathbf{u}, f_{\neq \mathbf{u}}\}$$

true posterior $\mathcal{O}(N^3)$

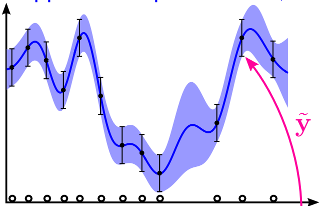
$\mathbf{y} = \{y_n\}_{n=1}^N$



$\mathbf{x} = \{x_n\}_{n=1}^N$

$\approx$

approximate posterior $\mathcal{O}(NM^2)$



$\tilde{\mathbf{y}}$ $\tilde{\Sigma}$

input locations of 'pseudo' data $\tilde{\mathbf{x}}$

outputs and covariance 'pseudo' data

EP algorithm

EP algorithm

1. remove

cavity

$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

take out one pseudo-observation likelihood

EP algorithm

1. remove

$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one
pseudo-observation
likelihood

2. include

$$p_n^{\mathrm{tilt}}(f) = q^{\backslash n}(f)p(y_n|f, x_n, \theta)$$

tilted

add in one
true observation
likelihood

EP algorithm

1. remove

$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one pseudo-observation likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\backslash n}(f) p(y_n | f, x_n, \theta)$$

tilted

add in one true observation likelihood

KL between unnormalised stochastic processes

3. project

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \, \mathrm{KL}\left[p_n^{\text{tilt}}(f) || q^*(f)\right]$$

project onto approximating family

EP algorithm

1. remove $$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$ take out one pseudo-observation likelihood

cavity

2. include $$p_n^{\text{tilt}}(f) = q^{\backslash n}(f) p(y_n | f, x_n, \theta)$$ add in one true observation likelihood

tilted

KL between unnormalised stochastic processes

3. project $$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \operatorname{KL}\left[p_n^{\text{tilt}}(f) || q^*(f)\right]$$ project onto approximating family

4. update $$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\backslash n}(f)}$$ update pseudo-observation likelihood

EP algorithm

1. remove $$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$ take out one pseudo-observation likelihood

cavity

2. include $$p_n^{\text{tilt}}(f) = q^{\backslash n}(f)p(y_n|f, x_n, \theta)$$ add in one true observation likelihood

tilted

KL between unnormalised stochastic processes

3. project $$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \operatorname{KL}\left[p_n^{\text{tilt}}(f)||q^*(f)\right]$$ project onto approximating family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$
2. Gaussian regression: matches moments everywhere

4. update $$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\backslash n}(f)}$$ update pseudo-observation likelihood

EP algorithm

1. remove

$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one pseudo-observation likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\backslash n}(f)p(y_n|f, x_n, \theta)$$

tilted

add in one true observation likelihood

KL between unnormalised stochastic processes

3. project

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \, \mathrm{KL}\left[p_n^{\text{tilt}}(f)||q^*(f)\right]$$

project onto approximating family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$
2. Gaussian regression: matches moments everywhere

4. update

$$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\backslash n}(f)}$$
$$= z_n \mathcal{N}(\mathrm{K}_{f_n \mathbf{u}} \mathrm{K}_{\mathbf{uu}}^{-1} \mathbf{u}; g_n, v_n)$$

update pseudo-observation likelihood

rank 1

Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u})$$

suppressed θ & x_n

Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

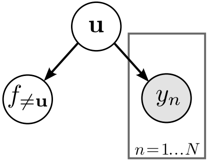
Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_n f_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

$q^*(f)$



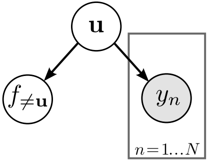
Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

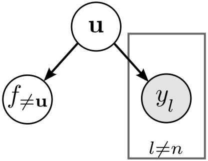
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

$$q^*(f)$$



$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$



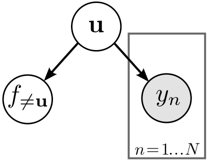
Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

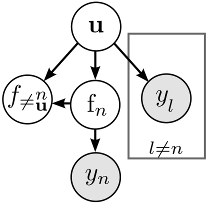
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

$q^*(f)$



$p_n^{\text{tilt}}(f) = q^{\backslash n}(f)p(y_n|f, x_n, \theta)$



Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

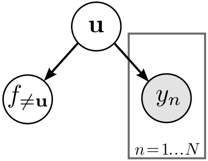
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

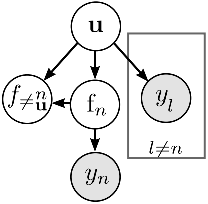
suppressed θ & x_n

$$p_n^{\mathrm{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$q^*(f)$

$p_n^{\mathrm{tilt}}(f) = q^{\backslash n}(f)p(y_n|f, x_n, \theta)$





Fixed points of EP = FITC approximation

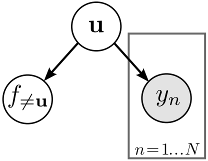
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

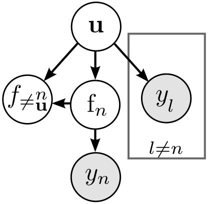
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$ suppressed θ & x_n

$$p_n^{\mathrm{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f)$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\; p_n^{\mathrm{tilt}}(f)$$





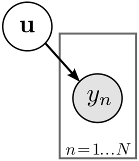
Fixed points of EP = FITC approximation

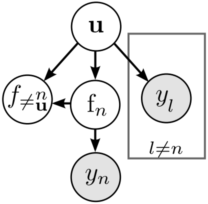
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_n f_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$ suppressed θ & x_n

$$p_n^{\text{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad\qquad \int \mathrm{d}f_{\neq\mathbf{u}}\; p_n^{\text{tilt}}(f)$$





Fixed points of EP = FITC approximation

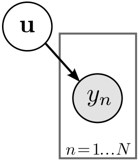
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

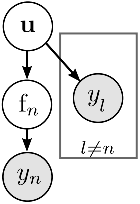
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

$$p_n^{\mathrm{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad\qquad \int \mathrm{d}f_{\neq\mathbf{u}}\; p_n^{\mathrm{tilt}}(f)$$





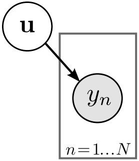
Fixed points of EP = FITC approximation

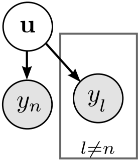
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$ suppressed θ & x_n

$$p_n^{\mathrm{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad\qquad \int \mathrm{d}f_{\neq\mathbf{u}}\; p_n^{\mathrm{tilt}}(f)$$





Fixed points of EP = FITC approximation

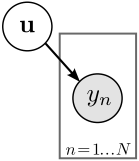
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

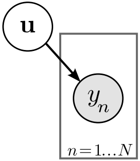
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

$$p_n^{\mathrm{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad\qquad \int \mathrm{d}f_{\neq\mathbf{u}}\; p_n^{\mathrm{tilt}}(f)$$





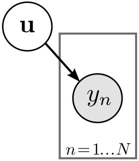
Fixed points of EP = FITC approximation

$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

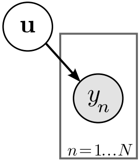
$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$ suppressed θ & x_n

$$p_n^{\mathrm{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad\qquad \int \mathrm{d}f_{\neq\mathbf{u}}\, p_n^{\mathrm{tilt}}(f)$$



=
equivalent



Fixed points of EP = FITC approximation

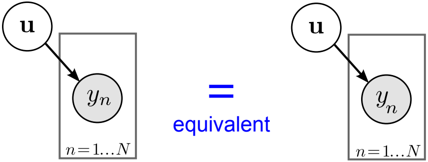
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

$$p_n^{\text{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad \int \mathrm{d}f_{\neq\mathbf{u}}\, p_n^{\text{tilt}}(f)$$



Csato & Opper (2002)

Qi, Abdel-Gawad & Minka (2010)

Fixed points of EP = FITC approximation

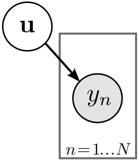
$$t_n(\mathbf{u}) = p(y_n|\mathbf{u}, x_n, \theta) = \mathcal{N}(y_n; \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathbf{u}; \mathrm{K}_{f_nf_n} - \mathrm{K}_{f_n\mathbf{u}}\mathrm{K}_{\mathbf{uu}}^{-1}\mathrm{K}_{\mathbf{u}f_n} + \sigma_y^2)$$

$$q^*(f) = p(f)\prod_{n=1}^{N} t_n(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})\prod_{n=1}^{N} p(y_n|\mathbf{u})$$

suppressed θ & x_n

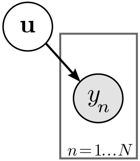
$$p_n^{\text{tilt}}(f) = p(f)p(y_n|f)\prod_{l\neq n} t_l(\mathbf{u}) = p(f_{\neq\mathbf{u}}|\mathbf{u})p(\mathbf{u})p(y_n|f)\prod_{l\neq n} p(y_l|\mathbf{u})$$

$$\int \mathrm{d}f_{\neq\mathbf{u}}\, q^*(f) \qquad \int \mathrm{d}f_{\neq\mathbf{u}}\; p_n^{\text{tilt}}(f)$$



=

equivalent



Csato & Opper (2002)

Qi, Abdel-Gawad & Minka (2010)

Interpretation resolves philosophical issues with FITC (increase M with N)

FITC likelihood > GP likelihood => EP over-estimates (marginal) likelihood

EP algorithm

1. remove

$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})}$$

cavity

take out one pseudo-observation likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\backslash n}(f) p(y_n | f, x_n, \theta)$$

tilted

add in one true observation likelihood

3. project

KL between unnormalised stochastic processes

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \operatorname{KL}\left[p_n^{\text{tilt}}(f) || q^*(f)\right]$$

project onto approximating family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$
2. Gaussian regression: matches moments everywhere

4. update

$$t_n(\mathbf{u}) = \frac{q^*(f)}{q^{\backslash n}(f)}$$

$$= z_n \mathcal{N}(\mathrm{K}_{f_n \mathbf{u}} \mathrm{K}_{\mathbf{u}\mathbf{u}}^{-1} \mathbf{u}; g_n, v_n)$$

update pseudo-observation likelihood

rank 1

Power EP algorithm (as tractable as EP)

1. remove

$$q^{\backslash n}(f) = \frac{q^*(f)}{t_n(\mathbf{u})^{\alpha}}$$

cavity

take out fraction of pseudo-observation likelihood

2. include

$$p_n^{\text{tilt}}(f) = q^{\backslash n}(f) p(y_n|f, x_n, \theta)^{\alpha}$$

tilted

add in fraction of true observation likelihood

3. project

KL between unnormalised stochastic processes

$$q^*(f) = \underset{q^*(f)}{\operatorname{argmin}} \mathrm{KL}\left[p_n^{\text{tilt}}(f) || q^*(f)\right]$$

project onto approximating family

1. minimum: moments matched at pseudo-inputs $\mathcal{O}(NM^2)$
2. Gaussian regression: matches moments everywhere

4. update

$$t_n(\mathbf{u})^{\alpha} = \frac{q^*(f)}{q^{\backslash n}(f)}$$

$$t_n(\mathbf{u}) = z_n \mathcal{N}(\mathrm{K}_{f_n \mathbf{u}} \mathrm{K}_{\mathbf{uu}}^{-1} \mathbf{u}; g_n, v_n)$$

update pseudo-observation likelihood

rank 1

Power EP: a unifying framework

$\alpha \to 0$ — α — $\alpha = 1$

VFE
Titsias, 2009

FITC
Csato and Opper, 2002
Snelson and Ghahramani, 2005

$$t_n(\mathbf{u}) = \mathcal{N}(\mathbf{K}_{\mathrm{f}_n\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}; y_n, \alpha D_{\mathrm{f}_n\mathrm{f}_n} + \sigma_y^2)$$

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{K}_{\mathbf{u}\mathrm{f}}\overline{\mathbf{K}}_{\mathbf{ff}}^{-1}\mathbf{y}, \mathbf{K}_{\mathbf{uu}} - \mathbf{K}_{\mathbf{uf}}\overline{\mathbf{K}}_{\mathbf{ff}}^{-1}\mathbf{K}_{\mathbf{fu}})$$

$$\log \mathcal{Z}_{\mathrm{PEP}} = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\overline{\mathbf{K}}_{\mathbf{ff}}| - \frac{1}{2}\mathbf{y}^\intercal\overline{\mathbf{K}}_{\mathbf{ff}}^{-1}\mathbf{y} + \frac{1-\alpha}{2\alpha}\sum_n \log\left(1 + \alpha D_{\mathrm{f}_n\mathrm{f}_n}/\sigma_y^2\right)$$

$$\overline{\mathbf{K}}_{\mathbf{ff}} = \mathbf{Q}_{\mathbf{ff}} + \alpha \mathrm{diag}(\mathbf{D}_{\mathbf{ff}}) + \sigma_y^2\mathrm{I} \qquad \mathbf{D}_{\mathbf{ff}} = \mathbf{K}_{\mathbf{ff}} - \mathbf{Q}_{\mathbf{ff}}$$

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)$$

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)$$

$$q^*(f) = p(f|\theta) \prod_{k=1}^{K} t_k(\mathbf{u})$$

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \underline{\prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)}$$

$$q^*(f) = p(f|\theta) \prod_{k=1}^{K} \underline{t_k(\mathbf{u})}$$

$\alpha = 1$

PITC / BCM
Schwaighofer & Tresp, 2002,
Snelson 2006,

$\alpha \to 0$

VFE
Titsias, 2009

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \underbrace{\prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)}$$

$$q^*(f) = p(f|\theta) \prod_{k=1}^{K} \underline{t_k(\mathbf{u})}$$

$\alpha = 1$

PITC / BCM
Schwaighofer & Tresp, 2002,
Snelson 2006,

$\alpha \to 0$

VFE
Titsias, 2009

Place pseudo-data in different space: interdomain transformations

$g(z) = \int w(z, z') f(z') \mathrm{d}z'$ (linear transform)

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)$$

$$q^*(f) = p(f|\theta) \prod_{k=1}^{K} t_k(\mathbf{u})$$

$\alpha = 1$

PITC / BCM
Schwaighofer & Tresp, 2002,
Snelson 2006,

$\alpha \to 0$

VFE
Titsias, 2009

Place pseudo-data in different space: interdomain transformations

$g(z) = \int w(z, z') f(z') \mathrm{d}z'$ (linear transform)

$$p^*(f, g) = p(f, g|\theta) \prod_{n=1}^{N} p(y_n|f, x_n, \theta)$$

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \underline{\prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)}$$

$$q^*(f) = p(f|\theta) \prod_{k=1}^{K} \underline{t_k(\mathbf{u})}$$

$\alpha = 1$

PITC / BCM
Schwaighofer & Tresp, 2002,
Snelson 2006,

$\alpha \to 0$

VFE
Titsias, 2009

Place pseudo-data in different space: interdomain transformations

$g(z) = \int w(z, z') f(z') \mathrm{d}z'$ (linear transform)

$$p^*(f, g) = p(f, g|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)}$$

$$q^*(f, g) = p(f, g|\theta) \prod_{n=1}^{N} \underline{t_n(\mathbf{u})}$$

pseudo-data
in new space

$g = \{\mathbf{u}, g_{\neq \mathbf{u}}\}$

Power EP: a unifying framework

Approximate blocks of data: structured approximations

$$p^*(f) = p(f, \mathbf{y}|\mathbf{x}, \theta) = p(f|\theta) \prod_{k=1}^{K} \underline{\prod_{n \in \mathcal{K}_n} p(y_n|f, x_n, \theta)}$$

$$q^*(f) = p(f|\theta) \prod_{k=1}^{K} \underline{t_k(\mathbf{u})}$$

$\alpha = 1$

PITC / BCM
Schwaighofer & Tresp, 2002,
Snelson 2006,

$\alpha \to 0$

VFE
Titsias, 2009

Place pseudo-data in different space: interdomain transformations

$g(z) = \int w(z, z') f(z') \mathrm{d}z'$ (linear transform)

$$p^*(f, g) = p(f, g|\theta) \prod_{n=1}^{N} \underline{p(y_n|f, x_n, \theta)}$$

$$q^*(f, g) = p(f, g|\theta) \prod_{n=1}^{N} \underline{t_n(\mathbf{u})}$$

pseudo-data
in new space

$g = \{\mathbf{u}, g_{\neq \mathbf{u}}\}$

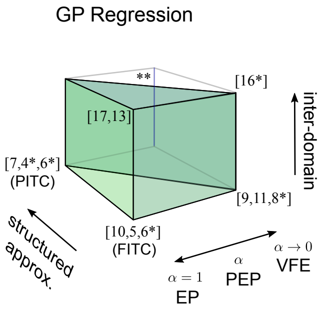
$\alpha = 1$

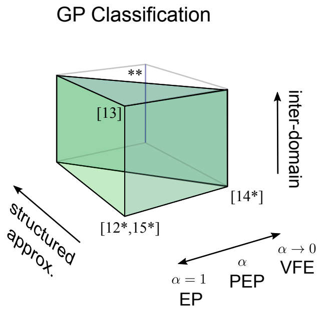
Figueiras-Vidal & Lázaro-Gredilla 2009

$\alpha \to 0$

Tobar et al. 2015
Matthews et al, 2016

Power EP: a unifying framework





[4] Quiñonero-Candela et al. 2005
[5] Snelson et al., 2005
[6] Snelson, 2006
[7] Schwaighofer, 2002

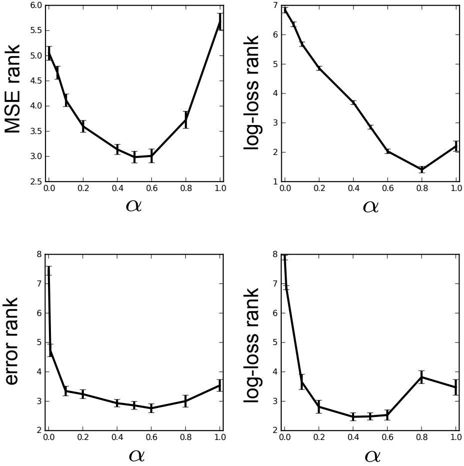
* = optimised pseudo-inputs
** = structured versions of VFE recover VFE

[8] Titsias, 2009
[9] Csató, 2002
[10] Csató et al., 2002
[11] Seeger et al., 2003
[12] Naish-Guzman et al, 2007
[13] Qi et al., 2010
[14] Hensman et al., 2015
[15] Hernández-Lobato et al., 2016
[16] Matthews et al., 2016
[17] Figueiras-Vidal et al., 2009

How should I set the power parameter α?

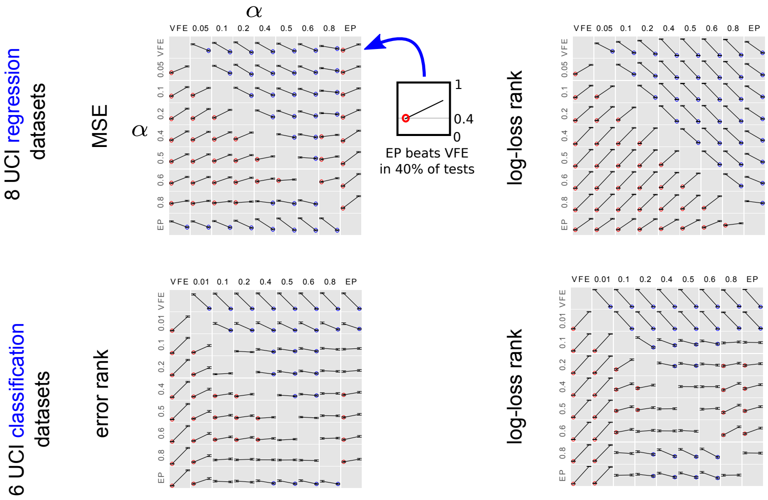
8 UCI regression datasets
20 random splits
M = 0 - 200
hypers and inducing
inputs optimised

6 UCI classification datasets
20 random splits
M = 10, 50, 100
hypers and inducing
inputs optimised



α = 0.5 does well on average

How should I set the power parameter α?



α = 0.5 does well on average

Streaming / Online Sparse Approximations

Goal: Online posterior update (using old posterior and new data batch).

Two new innovations for **online learning and inducing input optimisation**

1. naïve approach: use previous approximate posterior as prior

$$\overbrace{q^{(\mathsf{new})}(f)}^{\text{new posterior}} \approx \overbrace{p(\mathbf{y}^{(\mathsf{new})}|f)}^{\text{new likelihood}} \; \overbrace{q^{(\mathsf{old})}(f)}^{\text{old posterior}}$$

Goal: Online posterior update (using old posterior and new data batch).

Two new innovations for **online learning and inducing input optimisation**

1. better approach: only take likelihood terms from old posterior

$$\overbrace{q^{(\text{new})}(f)}^{\text{new posterior}} \approx \overbrace{p(\mathbf{y}^{(\text{new})}|f)}^{\text{new likelihood}} \overbrace{\frac{q^{(\text{old})}(f)}{p(f|\theta^{(\text{old})})}}^{\text{old likelihoods}} \overbrace{p(f|\theta^{(\text{new})})}^{\text{original prior}}$$

Streaming / Online Sparse Approximations

Goal: Online posterior update (using old posterior and new data batch).

Two new innovations for **online learning and inducing input optimisation**

1. better approach: only take likelihood terms from old posterior

$$\overbrace{q^{(\text{new})}(f)}^{\text{new posterior}} \approx \overbrace{p(\mathbf{y}^{(\text{new})}|f)}^{\text{new likelihood}} \overbrace{\frac{q^{(\text{old})}(f)}{p(f|\theta^{(\text{old})})}}^{\text{old likelihoods}} \overbrace{p(f|\theta^{(\text{new})})}^{\text{original prior}}$$

2. naïve approach: use same pseudo-points throughout

$$q^{(\text{old})}(f) = p(f_{\neq \mathbf{u}}|\mathbf{u}, \theta^{(\text{old})})q(\mathbf{u})$$
$$q^{(\text{new})}(f) = p(f_{\neq \mathbf{u}}|\mathbf{u}, \theta^{(\text{new})})q(\mathbf{u})$$

Goal: Online posterior update (using old posterior and new data batch).

Two new innovations for **online learning and inducing input optimisation**

1. better approach: only take likelihood terms from old posterior

$$\underbrace{q^{(\text{new})}(f)}_{\text{new posterior}} \approx \underbrace{p(\mathbf{y}^{(\text{new})}|f)}_{\text{new likelihood}} \; \overbrace{\frac{q^{(\text{old})}(f)}{p(f|\theta^{(\text{old})})}}^{\text{old likelihoods}} \; \underbrace{p(f|\theta^{(\text{new})})}_{\text{original prior}}$$

2. better approach: decouple sets of pseudo-points

$$q^{(\text{old})}(f) = p(f_{\neq \mathbf{u}^{(\text{old})}}|\mathbf{u}^{(\text{old})}, \theta^{(\text{old})})\, q(\mathbf{u}^{(\text{old})})$$
$$q^{(\text{new})}(f) = p(f_{\neq \mathbf{u}^{(\text{new})}}|\mathbf{u}^{(\text{new})}, \theta^{(\text{new})})\, q(\mathbf{u}^{(\text{new})})$$

Streaming / Online Sparse Approximations

Goal: Online posterior update (using old posterior and new data batch).

Two new innovations for **online learning and inducing input optimisation**

1. better approach: only take likelihood terms from old posterior

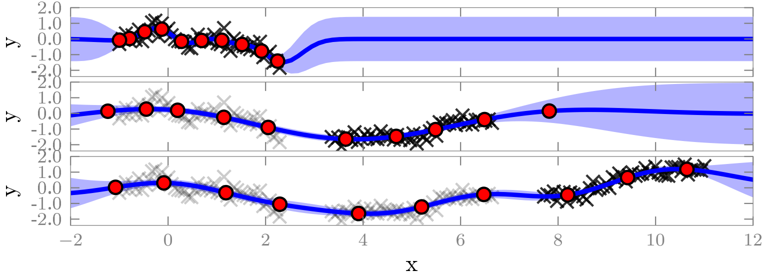
$$\overbrace{q^{(\text{new})}(f)}^{\text{new posterior}} \approx \overbrace{p(\mathbf{y}^{(\text{new})}|f)}^{\text{new likelihood}} \overbrace{\frac{q^{(\text{old})}(f)}{p(f|\theta^{(\text{old})})}}^{\text{old likelihoods}} \overbrace{p(f|\theta^{(\text{new})})}^{\text{original prior}}$$

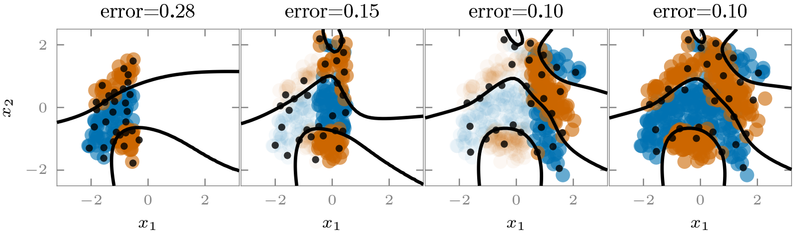
2. better approach: decouple sets of pseudo-points

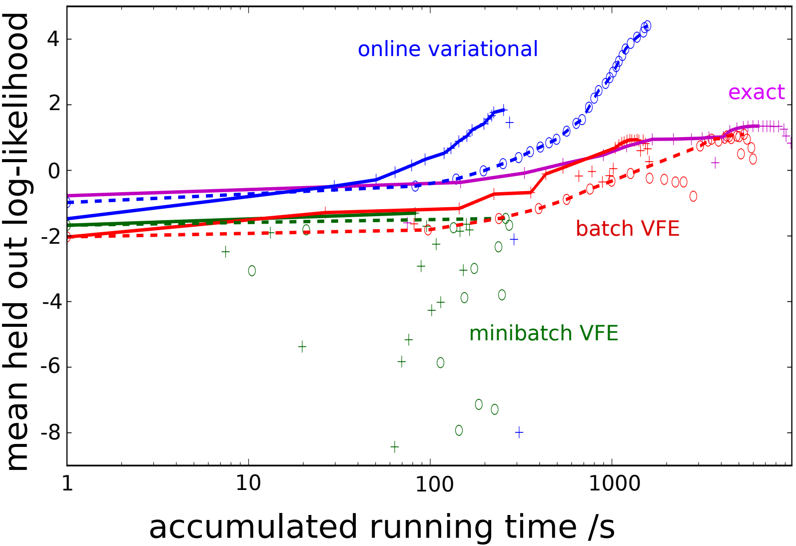
$$q^{(\text{old})}(f) = p(f_{\neq \mathbf{u}^{(\text{old})}}|\mathbf{u}^{(\text{old})}, \theta^{(\text{old})})\, q(\mathbf{u}^{(\text{old})})$$
$$q^{(\text{new})}(f) = p(f_{\neq \mathbf{u}^{(\text{new})}}|\mathbf{u}^{(\text{new})}, \theta^{(\text{new})})\, q(\mathbf{u}^{(\text{new})})$$

VFE is now the best Power EP method (inducing point clumping)

Online Sparse Approximations: Regression and Classification






online variational
exact
batch VFE
minibatch VFE
mean held out log-likelihood
accumulated running time /s
4
2
0
-2
-4
-6
-8
1
10
100
1000

Summary

- Provided a unifying framework for Gaussian Process Approximation methods using pseudo-points via PEP
- FITC and PITC are EP in disguise and they use the same approximating distribution as VFE
- Intermediate powers in PEP perform best on average in batch setting (more theory and empirical work needed)
- VFE methods perform best in the online setting

Core material:

- A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation, arXiv preprint 2016
- Streaming Sparse Gaussian Process Approximations, arXiv preprint 2017

VFE is best for online inference and learning