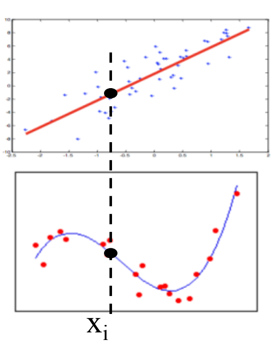
CS 3750 Advanced Machine Learning

Gaussian Processes: classification

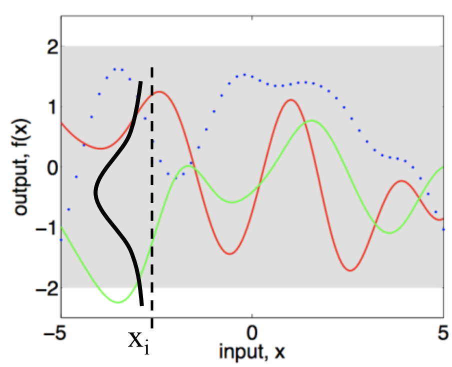
Jinpeng Zhou
jiz150@pitt.edu

Gaussian Processes (GP)

$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \boldsymbol{\varepsilon}$



$p(\boldsymbol{y} \mid \mathbf{x})$



GP is a collection of $\mathbf{f}(\mathbf{x})$ such that:

any set of $(\mathbf{f}(\mathbf{x_1}), \ldots, \mathbf{f}(\mathbf{x}_n))$ has a joint Gaussian distribution.

Weight-Space View

$y = f(\mathbf{x}) + \varepsilon = \mathbf{x}^{\mathrm{T}}\mathbf{w} + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$

$\Rightarrow \mathbf{y} \sim N(\mathbf{X}^{\mathrm{T}}\mathbf{w}, \sigma^2 \boldsymbol{I})$

Assume $\mathbf{w} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{p}})$, then:

$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{y} \mid \mathbf{X})} \sim N\left(\frac{1}{\sigma^2} A^{-1}\mathbf{X}\mathbf{y}, A^{-1}\right)$$

where $A = \frac{1}{\sigma^2} X X^T + \Sigma_p^{-1}$

Posterior on weights is Gaussian.

Weight-Space View

$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) \sim N\left(\frac{1}{\sigma^2} A^{-1}\mathbf{X}\mathbf{y}, A^{-1}\right), \qquad A = \frac{1}{\sigma^2} X X^T + \Sigma_p^{-1}$

Thus (Similar result when use basis function: $\mathbf{f}(\mathbf{x}) = \boldsymbol{\Phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}$):

$$p(f_* \mid \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(f_* \mid \mathbf{x}_*, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) d\mathbf{w}$$

$$= N\left(\frac{1}{\sigma^2} \mathbf{x}_*^T A^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*^T A^{-1}\mathbf{x}_*\right) = N(\mu, \Sigma)$$

- f_* has a Gaussian distribution and its Σ doesn't depend on $\mathbf{y}$
- $\mathbf{x}$ always appears as $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ ("scalar product" → kernel trick)

Kernel → Distribution of f_* → Prediction

Function-space View

Given:

mean function $m(\mathbf{x})$

covariance function $k(\mathbf{x}, \mathbf{x}')$

Define GP as:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

Such that for each $\mathbf{x_i}$:

$$f(\mathbf{x_i}) \sim N\left(m(\mathbf{x_i}), k(\mathbf{x_i}, \mathbf{x_i})\right)$$

$$\text{cov}\left(\mathbf{f}(\mathbf{x_i}), \mathbf{f}(\mathbf{x_j})\right) = k\left(\mathbf{x_i}, \mathbf{x_j}\right)$$

Function-space View

Given joint Gaussian distribution:

$$\begin{bmatrix} x_A \\ x_B \end{bmatrix} \sim N(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix})$$

The conditional densities are also Gaussian:

$$x_A \mid x_B \sim N(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_A - x_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$$

$$x_B \mid x_A \sim N(\ldots, \ldots)$$

Function-space View

Consider training and test points.

$$\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_*\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{m}(\boldsymbol{X})\\ \boldsymbol{m}(\boldsymbol{X}_*)\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)\\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right)$$

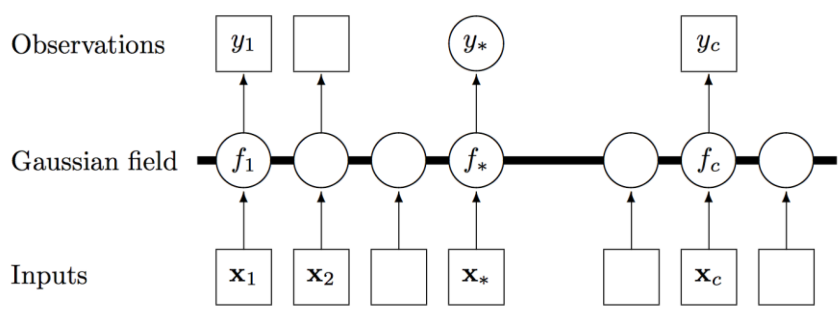
$\boldsymbol{K}$ is the covariance matrix.

$y = f(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$. Thus:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_*\end{bmatrix} \mid \mathbf{X}, \mathbf{X}_* \sim N\left(\begin{bmatrix}\boldsymbol{m}(\boldsymbol{X})\\ \boldsymbol{m}(\boldsymbol{X}_*)\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \boldsymbol{\sigma}^2\boldsymbol{I} & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)\\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) + \boldsymbol{\sigma}^2\boldsymbol{I}\end{bmatrix}\right)$$

$$\mathbf{y}_* \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{X}_* \sim N(\ldots, \ldots)$$

GP for Regression (GPR)

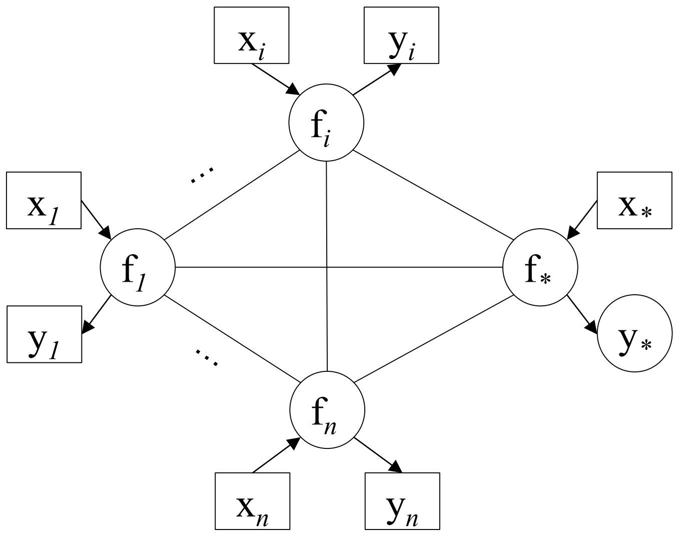


Squares are observed, circles are latent

The thick bar is like a “bus” connected all nodes

Each y is conditionally independent of all other nodes given f

GPR



GPR

$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}$, where $\mathbf{f} \sim \mathrm{GP}(\mathrm{m} = 0, \mathrm{k})$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2)$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim N(0, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \boldsymbol{\sigma}^2\boldsymbol{I} & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*) \\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) + \boldsymbol{\sigma}^2\boldsymbol{I} \end{bmatrix})$$

Prediction: $\mathbf{y}_* \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{X}_* \sim N(\ldots,\ldots) = N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$

Where:

$$\boldsymbol{\mu}^* = \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\left(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \boldsymbol{\sigma}^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) + \boldsymbol{\sigma}^2\boldsymbol{I} - \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\left(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \boldsymbol{\sigma}^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)$$

GP Classification (GPC)

Map output into [0, 1] by using response function:

- Sigmoid function (logistic):

$$\sigma(x) = (1 + e^{-x})^{-1}$$

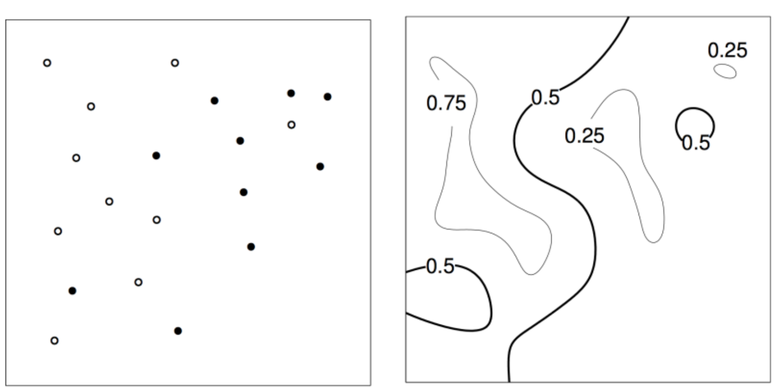
- Cumulative normal (probit):

$$\Phi(x) = \int_{-\infty}^{x} N(t|0,1)dt$$

GPC examples

Squared exponential kernel with hyperparameter length-scale = 0.1

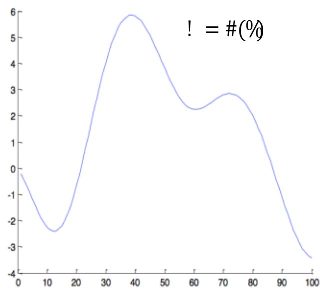
Logistic response function.



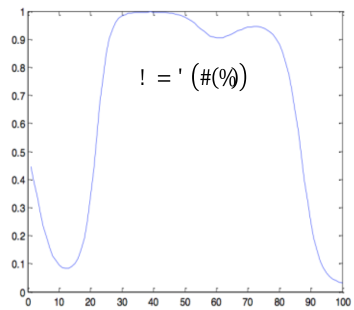
(a) (b)

GPC

GP over f



GP over f
Non-G over y



GPC

Target $\mathbf{y} = \boldsymbol{\sigma}(\mathbf{f}(\mathbf{x}))$ is a non-Gaussian.

E.g., it's a Bernoulli for class 0, 1:

$$\boldsymbol{p}(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} \sigma(\mathrm{f}(\mathrm{x_i}))^{y_i}(1 - \sigma(\mathrm{f}(\mathrm{x_i}))^{1-y_i}$$

How to predict $\mathbf{p}(\mathbf{y}_* \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{X}_*)$?

GPC

Assume single test case $\mathbf{x}_*$, predict its class 0 or 1.

Assume sigmoid function and a GP prior over $\mathbf{f}$.

Thus, try to involve $\mathbf{f}$:

$$p(\mathrm{y}_* = 1 \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathrm{y}_* = 1, \mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$$

$$= \int p(\mathrm{y}_* = 1 \mid \mathrm{f}_*, \cancel{\mathbf{y}, \boldsymbol{X}, \mathbf{x}_*})\, p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$$

$$= \int \sigma(\mathrm{f}_*)\, p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$$

Approximation

$$p(\mathrm{y}_* = 1 \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int \sigma(\mathrm{f}_*)\, p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$$

Note that $\sigma(\mathrm{f}_*)$ is a sigmoid function.

If $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian, convolution of a sigmoid and a Gaussian can be computed as:

$$\int \sigma(t) N(\mu, s^2) dt \approx \sigma\left(\sqrt{1 + \frac{\pi s^2}{8}}\, \mu\right)$$

Otherwise, analytically intractable!

If $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian…

$$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathrm{f}_*, \mathbf{f} \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathbf{f}$$

$$= \int \frac{p(\mathbf{y} \mid \mathrm{f}_*, \mathbf{f}, \boldsymbol{X}, \mathbf{x}_*)\ p(\mathrm{f}_*, \mathbf{f}, \boldsymbol{X}, \mathbf{x}_*)}{p(\mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)} d\mathbf{f}$$

$$= \int \frac{p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{X})\ p(\mathbf{f}, \boldsymbol{X}, \mathbf{x}_*)\ p(\mathrm{f}_* \mid \mathbf{f}, \boldsymbol{X}, \mathbf{x}_*)}{p(\mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)} d\mathbf{f}$$

$$= \int \frac{p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{X})\ p(\mathbf{f}, \boldsymbol{X})}{p(\mathbf{y}, \boldsymbol{X})}\ p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$$

$$= \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})\ p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$$

If $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian…

$$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})\ p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$$

Recall from GP regression: $\mathbf{y}_* \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{X}_* \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$

$$\boldsymbol{\mu}^* = \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})\left(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\sigma}^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}$$

$$\boldsymbol{\Sigma}^* = \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*) + \boldsymbol{\sigma}^2 \boldsymbol{I} - \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})\left(\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\sigma}^2 \boldsymbol{I}\right)^{-1} \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*)$$

Replace $\boldsymbol{X}_*$ with $\mathbf{x}_*$:

$$p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) \sim N(\boldsymbol{k}_*^T \boldsymbol{C}_n^{-1} \boldsymbol{y},\ c - \boldsymbol{k}_*^T \boldsymbol{C}_n^{-1} \boldsymbol{k}_*)$$

Where: $\boldsymbol{k}_* = \boldsymbol{K}(\boldsymbol{X}, \mathbf{x}_*)$ $\quad \boldsymbol{C}_n = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\sigma}^2 \boldsymbol{I}$

$c = k(\mathbf{x}_*, \mathbf{x}_*) + \boldsymbol{\sigma}^2$

If $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian…

$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})\, p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f})\, d\mathbf{f}$

Now $p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f})$ is a Gaussian!

IF:

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ is a Gaussian

THEN:

$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian!

Is $p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ a Gaussian?

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \frac{\boldsymbol{p}(\boldsymbol{y} \mid \mathbf{f}, \boldsymbol{X})\, \boldsymbol{p}(\boldsymbol{f} \mid \boldsymbol{X})}{\boldsymbol{p}(\boldsymbol{y} \mid \boldsymbol{X})}$

$\boldsymbol{p}(\mathbf{f} \mid \boldsymbol{X}) \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}), \quad \boldsymbol{p}(\boldsymbol{y} \mid \mathbf{f}, \boldsymbol{X}) = \prod_{i=1}^{n} p(y_i \mid f_i)$

$\Rightarrow p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \frac{\boldsymbol{N}(\mathbf{0}, \boldsymbol{K})}{\boldsymbol{p}(\boldsymbol{y} \mid \boldsymbol{X})} \prod_{i=1}^{n} p(y_i \mid f_i)$

Note: $p(y_i \mid f_i)$ is Non-Gaussian (e.g., Bernoulli)

$\Rightarrow p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ is Non-Gaussian.

Approximation

$p(\mathrm{y}_* = 1 \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int \sigma(\mathrm{f}_*)\, p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$

~~If $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian,~~ *Try to approximate* $p(\mathrm{f}_* \mid \mathbf{y}, \mathbf{X}, \mathbf{x}_*)$ *as Gaussian*, … be computed as …

- Laplace Approximation
- Expectation Propagation (EP)
- *Variational Approximation, Monte Carlo Sampling, etc.*

If $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian…

$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})\, p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$

Now $p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f})$ is a Gaussian!

IF:

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ is ~~a~~ <u>*approximated as*</u> Gaussian

THEN:

$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is ~~a~~ <u>*approximated as*</u> Gaussian!

Laplace Approximation

Goal: use a Gaussian to approximate $p(\boldsymbol{x})$

Gaussian: $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Idea:

- μ should match the mode: $argmax_x(p(\boldsymbol{x}))$
- Relation between $\exp(k(x-\mu)^2)$ and $p(\boldsymbol{x})$

Laplace Approximation

- μ should match the mode: $argmax_x(p(\boldsymbol{x}))$

  $p'(\mu) = 0,$ μ is placed at the mode

- Relation between $\exp(k(x-\mu)^2)$ and $p(\boldsymbol{x})$

  Taylor expansion of $\boldsymbol{ln\, p}(\boldsymbol{x})$ at $\boldsymbol{\mu}$

Let $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{ln\, p}(\boldsymbol{x})$, Taylor expansion of $t(x)$ at $\boldsymbol{\mu}$:

$t(x) = t(\mu) + \frac{t'(\mu)}{1!}(x-\mu) + \frac{t''(\mu)}{2!}(x-\mu)^2 + \cdots$

Laplace Approximation

Taylor quadratic approximation:

$$t(x) = t(\mu) + \frac{t'(\mu)}{1!}(x-\mu) + \frac{t''(\mu)}{2!}(x-\mu)^2 + \cdots$$

$$\approx t(\mu) + \frac{t'(\mu)}{1!}(x-\mu) + \frac{t''(\mu)}{2!}(x-\mu)^2$$

Note that $t'(\mu) = \frac{p'(\mu)=0}{p(\mu)} = 0$, thus:

$$t(x) \approx t(\mu) + \frac{t''(\mu)}{2!}(x-\mu)^2$$

Laplace Approximation

$$t(x) \approx t(\mu) + \frac{t''(\mu)}{2!}(x-\mu)^2 \qquad // \; t(x) = \; ln\, p(x)$$

$$\Rightarrow \exp\big(t(x)\big) \approx \exp\left(t(\mu) + \frac{t''(\mu)}{2!}(x-\mu)^2\right)$$

$$\Rightarrow p(x) \approx exp\big(t(\mu)\big)\; exp\left(\frac{t''(\mu)(x-\mu)^2}{2!}\right)$$

$$\Rightarrow p(x) \approx p(\mu)\; exp\left(-\frac{A(x-\mu)^2}{2}\right)$$

Where: $A = -\,t''(\mu) = -\frac{d^2}{dx^2} ln\, p(x)\,|_{x=\mu}$

Laplace Approximation

$$p(x) \approx p(\mu)\ exp\left(-\frac{A(x-\mu)^2}{2}\right)$$

$$p(\mathbf{x}) \approx p(\boldsymbol{\mu})\ exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})^T A(\mathbf{x}-\boldsymbol{\mu})}{2}\right)$$

Corresponding Gaussian approximation:

$$p(\mathbf{x}) \approx N(\boldsymbol{\mu},\ \boldsymbol{A}^{-1})$$

Where:

$$\frac{d}{d\mathbf{x}} ln\, p(\mathbf{x})\,|_{\mathbf{x}=\boldsymbol{\mu}} = \mathbf{0}$$

$$-\frac{d^2}{d\mathbf{x}^2} ln\, p(\mathbf{x})\,|_{\mathbf{x}=\boldsymbol{\mu}} = \mathbf{A}$$

Laplace Approximation

Back to $p(\mathbf{f} \mid \mathbf{y}, X)$:

$$\ln p(\mathbf{f} \mid \mathbf{y}, X) = \ln \frac{p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid \mathbf{X})}{p(\mathbf{y} \mid \mathbf{X})}$$

$$= \ln p(\mathbf{y} \mid \mathbf{f}) + \ln p(\mathbf{f} \mid \mathbf{X}) - \ln p(\mathbf{y} \mid \mathbf{X})$$

$\ln p(\mathbf{y} \mid \mathbf{X})$ is independent of **f**, and we only need to do 1st & 2nd derivatives on **f**

Thus, define $\Psi(\mathbf{f})$ for derivative calculation:

$$\Psi(\mathbf{f}) = \ln p(\mathbf{y} \mid \mathbf{f}) + \ln p(\mathbf{f} \mid \mathbf{X})$$

Laplace Approximation

$\Psi(\mathbf{f}) = \ln p(\mathbf{y} \mid \mathbf{f}) + \ln p(\mathbf{f} \mid \mathbf{X})$

Note that:

- $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} \sigma(\mathrm{f}(\mathrm{x_i}))^{y_i}(1 - \sigma(\mathrm{f}(\mathrm{x_i})))^{1-y_i}$

$$= \prod_{i=1}^{n} (1 - \sigma(\mathrm{f}(\mathrm{x_i}))) \left(\frac{\sigma(\mathrm{f}(\mathrm{x_i}))}{1-\sigma(\mathrm{f}(\mathrm{x_i}))}\right)^{y_i}$$

$$= \prod_{i=1}^{n} \frac{1}{1+e^{\mathrm{f}(\mathrm{x_i})}} \left(e^{\mathrm{f}(\mathrm{x_i})}\right)^{y_i}$$

- $\mathbf{f} \mid \mathbf{X} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{C_n})$

Laplace Approximation

$\Psi(\mathbf{f}) = \ln p(\mathbf{y} \mid \mathbf{f}) + \ln p(\mathbf{f} \mid \mathbf{X})$

- $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{n} \frac{1}{1+e^{\mathrm{f}(\mathrm{x_i})}} \left(e^{\mathrm{f}(\mathrm{x_i})}\right)^{y_i}$
- $\mathbf{f} \mid \mathbf{X} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{C_n})$

Thus:

$$\Psi(\mathbf{f}) = \mathbf{y}^{\mathbf{T}}\mathbf{f} - \sum_{i=1}^{n} \ln\left(1 + e^{\mathrm{f}(\mathrm{x_i})}\right) - \frac{\mathbf{f}^{\mathbf{T}}\mathbf{C}_{\mathbf{n}}^{-1}\mathbf{f}}{\mathbf{2}} - \frac{\mathbf{n}\ln 2\pi}{\mathbf{2}} - \frac{\ln|C_n|}{\mathbf{2}}$$

Laplace Approximation

$\Psi(\mathbf{f}) = \ln p(\mathbf{y} \mid \mathbf{f}) + \ln p(\mathbf{f} \mid \mathbf{X})$

$= \mathbf{y}^{\mathbf{T}}\mathbf{f} - \sum_{i=1}^{n} \ln\left(1 + e^{\mathrm{f}(\mathrm{x_i})}\right) - \frac{\mathbf{f}^{\mathbf{T}} C_n^{-1} \mathbf{f}}{2} - \frac{\mathbf{n} \ln 2\pi}{2} - \frac{\ln|C_n|}{2}$

$\nabla\Psi(\mathbf{f}) = \nabla \ln p(\mathbf{y} \mid \mathbf{f}) - \mathbf{C}_{\mathbf{n}}^{-1}\mathbf{f} = \mathbf{y} - \boldsymbol{\sigma}(\mathbf{f}) - \mathbf{C}_{\mathbf{n}}^{-1}\mathbf{f}$

where $\boldsymbol{\sigma}(\mathbf{f}) = [\boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_1)), \ldots, \boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_n))]^T$

$\nabla\nabla\Psi(\mathbf{f}) = \nabla\nabla \ln p(\mathbf{y} \mid \mathbf{f}) - \mathbf{C}_{\mathbf{n}}^{-1} = -\boldsymbol{W}_{\boldsymbol{n}} - \boldsymbol{C}_{\boldsymbol{n}}^{-1}$

$$W_n = \begin{bmatrix} \boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_1))(1-\boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_1))) & & \\ & \ddots & \\ & & \boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_n))(1-\boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_n))) \end{bmatrix}$$

Laplace Approximation

Corresponding Gaussian approximation:

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) \approx N(\boldsymbol{\mu}, \boldsymbol{A}^{-1})$

Where:

$\nabla\Psi(\boldsymbol{\mu}) = \mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}) - \mathbf{C}_{\mathbf{n}}^{-1}\boldsymbol{\mu} = 0$

$\Rightarrow \mathbf{y} - \boldsymbol{\sigma}(\boldsymbol{\mu}) = \mathbf{C}_{\mathbf{n}}^{-1}\boldsymbol{\mu}$

$\mathbf{?}\Rightarrow \boldsymbol{\mu}$

$-\nabla\nabla\Psi(\boldsymbol{\mu}) = \boldsymbol{W}_{\boldsymbol{n}} + \mathbf{C}_{\mathbf{n}}^{-1}$

$\Rightarrow \mathbf{A} = \boldsymbol{W}_{\boldsymbol{n}} + \mathbf{C}_{\mathbf{n}}^{-1}$

Laplace Approximation

$\boldsymbol{\sigma}(\mathbf{f}) = [\boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_1)), \dots, \boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}_n))]^T$

Cannot directly solve **f** from: $\mathbf{y} - \boldsymbol{\sigma}(\mathbf{f}) = \mathbf{C}_\mathbf{n}^{-1}\mathbf{f}$.

Recall its motivation:

place **f** at the mode: maximum $p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$

Thus, instead of maximum, try to find $\mathbf{f}^*$ for the minimum of $-\ln p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$, with iteration:

$$\mathbf{f}^{new} = \mathbf{f}^{old} - (\nabla\nabla\Psi)^{-1}\nabla\Psi$$

Finally, get the Gaussian approximation:

$$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) \approx N(\mathbf{f}^*, \ (\boldsymbol{W}_\boldsymbol{n} + \mathbf{C}_\mathbf{n}^{-1})^{-1})$$

Laplace Approximation

$$p(\mathrm{y}_* = 1 \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int \sigma(\mathrm{f}_*)\, p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$$

$$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})\, p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$$

Now we have:

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ is approximated as Gaussian.

$p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f})$ is a Gaussian.

Thus, $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian $\sim N(\mu, s^2)$

Thus, $p(\mathrm{y}_* = 1 \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) \approx \sigma\left(\sqrt{1 + \frac{\pi s^2}{8}}\mu\right)$

Expectation Propagation

$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) \, p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \frac{p(\mathbf{y} \mid \mathbf{f}) \, p(\mathbf{f} \mid \mathbf{X})}{p(\mathbf{y} \mid \mathbf{X})}$

$= \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}) \prod_{i=1}^{n} p(y_i \mid f_i)$

Normalization term:

$Z = p(\mathbf{y} \mid \mathbf{X}) = \int p(\mathbf{f} \mid \mathbf{X}) \prod_{i=1}^{n} p(y_i \mid f_i) \, d\mathbf{f}$

Expectation Propagation

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}) \prod_{i=1}^{n} p(y_i \mid f_i)$

$p(y_i \mid f_i)$ is Non-Gaussian due to response function.

Try to find a Gaussian t_i to approximate it:

$p(y_i \mid f_i) \approx t_i(f_i \mid \tilde{Z}_i, \ \tilde{\mu}_i, \ \tilde{\sigma}_i{}^2) = \tilde{Z}_i \, N(f_i \mid \tilde{\mu}_i, \tilde{\sigma}_i{}^2)$

$\Rightarrow \ \prod_{i=1}^{n} p(y_i \mid f_i) \approx \prod_{i=1}^{n} t_i(f_i \mid \tilde{Z}_i, \ \tilde{\mu}_i, \ \tilde{\sigma}_i{}^2)$

$= \mathrm{N}(\tilde{\boldsymbol{\mu}}, \ \tilde{\Sigma}) \prod_{i=1}^{n} \tilde{Z}_i,$ where $\tilde{\Sigma}$ is $diag(\tilde{\sigma}_i{}^2)$

Expectation Propagation

$$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}) \prod_{i=1}^{n} p(y_i \mid f_i)$$

$$\approx \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}) \prod_{i=1}^{n} t_i(f_i \mid \tilde{Z}_i, \ \tilde{\mu}_i, \ \tilde{\sigma}_i{}^2)$$

$$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) \approx \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}) \mathrm{N}(\tilde{\boldsymbol{\mu}}, \ \tilde{\boldsymbol{\Sigma}}) \prod_{i=1}^{n} \tilde{Z}_i$$

Thus:

$$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) \approx q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \mathrm{N}(\boldsymbol{\mu}, \Sigma)$$

Where: $\boldsymbol{\mu} = \Sigma \tilde{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}$ $\quad \Sigma = \left(\boldsymbol{K}^{-1} + \tilde{\Sigma}^{-1}\right)^{-1}$

Expectation Propagation

How to choose $\tilde{Z}_i, \ \tilde{\mu}_i, \ \tilde{\sigma}_i{}^2$ that defines t_i?

$\tilde{Z}_i, \ \tilde{\mu}_i, \ \tilde{\sigma}_i{}^2$ that minimize the difference between

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ and $q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$:

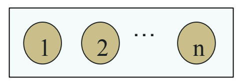
$\mathrm{A} \prod_{i=1}^{n} p(y_i \mid f_i) \approx \mathrm{A} \prod_{i=1}^{n} t_i(f_i \mid \tilde{Z}_i, \ \tilde{\mu}_i, \ \tilde{\sigma}_i{}^2)$ where $\mathrm{A} = \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K})$

Using Kullback-Leibler (KL) divergence: KL (p(x) ‖ q(x))

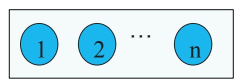
- $\min_{\{t_i\}} KL(\mathrm{A} \prod_{i=1}^{n} p(y_i \mid f_i) \,||\, A \prod_{i=1}^{n} t_i)$ intractable ✗
- $\min_{t_i} KL\left(Ap(y_i \mid f_i) \prod_{j \neq i} t_j \,||\, At_i \prod_{j \neq i} t_j\right)$ iterative on t_i ✓

Expectation Propagation

Default

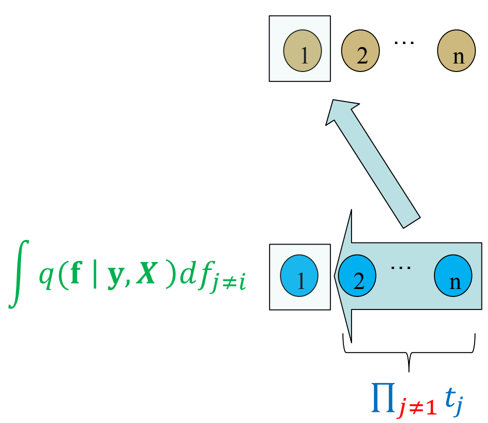


intractable ✗



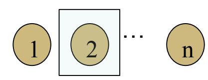
Expectation Propagation

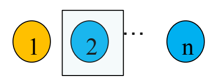
1^{st} Iteration, based on marginal for t_1



Expectation Propagation

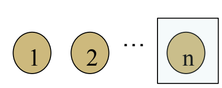
2st Iteration, based on marginal for t_2

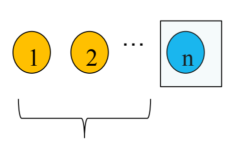




Expectation Propagation

nth Iteration, based on marginal for t_n

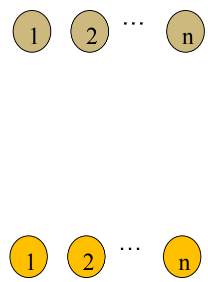




$\prod_{j \neq n} t_j$

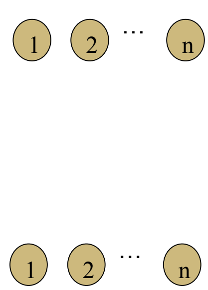
Expectation Propagation

Repeat until convergence



Expectation Propagation

Repeat until convergence



Expectation Propagation

$q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = A \prod_{i=1}^{n} t_i = \mathrm{N}(\boldsymbol{\mu}, \Sigma)$

$t_i(f_i \mid \tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \tilde{Z}_i N(f_i \mid \tilde{\mu}_i, \tilde{\sigma}_i^2)$

Iteratively update t_i

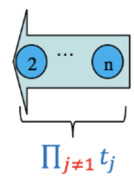
1. Start from current approximate posterior, leave out the current t_i to get the *cavity distribution* $q_{-i}(f_i)$:

$$q_{-i}(f_i) = \int A \prod_{j\neq i} t_j \, df_{j\neq i} = \frac{t_i \int A \prod_{j\neq i} t_j df_{j\neq i}}{t_i} = \frac{\int A \prod_{j=1}^{n} t_j df_{j\neq i}}{t_i}$$

$$= \frac{\int q(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) df_{j\neq i}}{t_i} = \frac{q(f_i \mid \mathbf{y}, \boldsymbol{X})}{t_i} = \frac{\mathrm{N}(f_i \mid \mu_i, \sigma_i^2)}{\tilde{Z}_i N(f_i \mid \tilde{\mu}_i, \tilde{\sigma}_i^2)} = \mathrm{N}(f_i \mid \mu_{-i}, \sigma_{-i}^2)$$

where: $\sigma_i^2 = \Sigma_{ii}$ $\quad \sigma_{-i}^2 = \left(\sigma_i^{-2} - \tilde{\sigma}_i^{-2}\right)^{-1}$

$$\mu_{-i} = \sigma_{-i}^2\left(\sigma_i^{-2}\mu_i - \tilde{\sigma}_i^{-2}\tilde{\mu}_i\right)$$



Expectation Propagation

$p(y_i \mid f_i) \approx t_i$

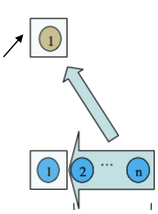
$p(y_i \mid f_i)$ is from response function

Iteratively update t_i

2. Find the new Gaussian marginal $\hat{q}(f_i)$: $q(f_i \mid \mathbf{y}, \boldsymbol{X})$ which best approximates the product of $q_{-i}(f_i)$ and $p(y_i \mid f_i)$:

Non-Gaussian Marginal $p(f_i \mid \mathbf{y}, \boldsymbol{X})$

$$q_{-i}(f_i)\, p(y_i \mid f_i) \approx \hat{q}(f_i) = \hat{Z}_i N\left(\hat{\mu}_i, \hat{\sigma}_i^2\right)$$



By:

$$\min_{t_i} KL\left(q_{-i}(f_i)\, p(y_i \mid f_i) \,||\, \hat{q}(f_i)\right) \quad \text{// tractable}$$

Expectation Propagation

Iteratively update t_i

3. Update parameters $\tilde{Z}_i,\ \tilde{\mu}_i,\ \tilde{\sigma}_i{}^2$ of t_i based on $\widehat{Z}_i$, $\widehat{\mu}_i$, $\widehat{\sigma}_i{}^2$ from found $\hat{q}(f_i)$. E.g.:

Recall: $q_{-i}(f_i) = \frac{\overset{\text{old}}{\hat{q}(f_i)}}{t_i} = \mathrm{N}(f_i \mid \mu_{-i},\ \sigma_{-i}^2)$

where $\sigma_{-i}^2 = \left(\overset{\text{old}}{\sigma_i^{-2}} - \tilde{\sigma}_i{}^{-2}\right)^{-1}$

Now we have new $\hat{q}(f_i)$,

$$\sigma_{-i}^2 = \left(\overset{\text{new}}{\hat{\sigma}_i{}^{-2}} - \tilde{\sigma}_i{}^{-2}\right)^{-1} \Rightarrow \tilde{\sigma}_i{}^2 = \left(\hat{\sigma}_i{}^{-2} - \sigma_{-i}^{-2}\right)^{-1}$$

Expectation Propagation

$p(\mathrm{y}_* = 1 \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int \sigma(\mathrm{f}_*)\, p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) d\mathrm{f}_*$

$p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*) = \int p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})\, p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f}) d\mathbf{f}$

$p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X}) = \frac{1}{Z} \boldsymbol{N}(\mathbf{0}, \boldsymbol{K}) \prod_{i=1}^{n} p(y_i \mid f_i)$

Now we have:

$p(y_i \mid f_i)$ is approximated as Gaussian.

Subsequently $p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$ is a Gaussian.

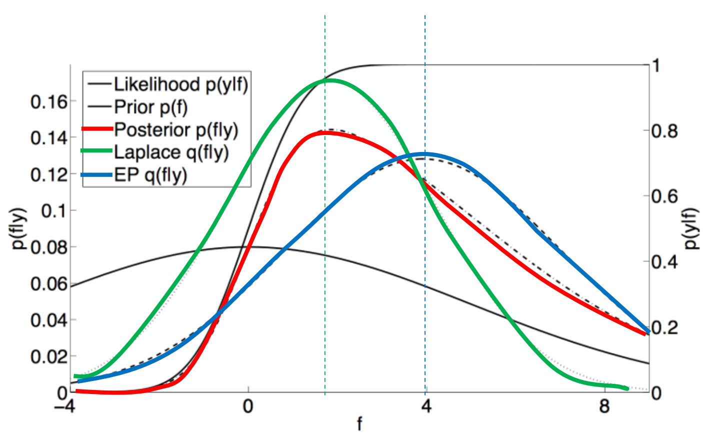
$p(\mathrm{f}_* \mid \boldsymbol{X}, \mathbf{x}_*, \mathbf{f})$ is a Gaussian.

Thus, $p(\mathrm{f}_* \mid \mathbf{y}, \boldsymbol{X}, \mathbf{x}_*)$ is a Gaussian $\sim N(\mu, s^2)$

Approximation

- Laplace Approximation
  - $p(\mathbf{f} \mid \mathbf{y}, \boldsymbol{X})$
  - 2nd order Taylor approximation
  - Mean $\boldsymbol{\mu}$ is placed at the mode
  - Covariance $\boldsymbol{A}^{-1}$ is also related to the mode
- EP
  - $p(y_i \mid f_i)$
  - Iteratively update t_i by minimizing KL divergence

Approximation



Laplace peaks at posterior mode
EP has a more accurate placement of probability mass

GPC

- Nonparametric classification based on Bayesian methodology
- Classification decision is directly made from observed data
- The computational cost is $O(N^3)$ in general, due to the covariance matrix inversion. (*DeepMind's newly proposed "Neural Processes" achieved O(N) by GP+NN*)

References

1. Ebden, Mark. "Gaussian processes for regression: A quick introduction." *The Website of Robotics Research Group in Department on Engineering Science, University of Oxford* 91 (2008): 424-436.
2. Williams, Christopher K., and Carl Edward Rasmussen. "Gaussian processes for machine learning." *the MIT Press* 2, no. 3 (2006): 4.
3. Kuss, Malte, and Carl E. Rasmussen. "Assessing approximations for Gaussian process classification." In *Advances in Neural Information Processing Systems*, pp. 699-706. 2006.
4. Milos Hauskrecht. "CS2750 Machine Learning: Linear regression". http://people.cs.pitt.edu/~milos/courses/cs2750/Lectures/Class8.pdf
5. Nicholas, Zabaras. "Kernel Introduction to Gaussian Methods and Processes." https://www.dropbox.com/s/2yonhmzn57lvv1d/Lec27-KernelMethods.pdf?dl=0
6. Yee Whye The. "Expectation and Belief Propagation". https://www.stats.ox.ac.uk/~teh/teaching/probmodels/lecture4bp.pdf