Gaussian Processes
Covariance Functions and Classification

Carl Edward Rasmussen

Max Planck Institute for Biological Cybernetics
Tübingen, Germany

Gaussian Processes in Practice, Bletchley Park, July 12th, 2006

Outline

Covariance functions encode structure. You can learn about them by

- sampling,
- optimizing the marginal likelihood.

GP's with various covariance functions are equivalent to many well known models, large neural networks, splines, relevance vector machines...

- infinitely many Gaussian bumps regression
- Rational Quadratic and Matérn

Quick two-page recap of GP regression

Approximate inference for Gaussian process classification: Replace the non-Gaussian intractable posterior by a Gaussian. Expectation Propagation.

From random functions to covariance functions

Consider the class of functions (sums of squared exponentials):

$$\begin{aligned} f(x) &= \lim_{n\rightarrow\infty} \frac{1}{n}\sum_i \gamma_i \exp(-(x-i/n)^2), \quad \text{where} \;\; \gamma_i \sim \mathcal{N}(0,1),\ \forall i \\ &= \int_{-\infty}^{\infty} \gamma(u)\exp(-(x-u)^2)du, \quad \text{where} \;\; \gamma(u) \sim \mathcal{N}(0,1),\ \forall u. \end{aligned}$$

The mean function is:

$$\mu(x) \;=\; E[f(x)] \;=\; \int_{-\infty}^{\infty} \exp(-(x-u)^2) \int_{-\infty}^{\infty} \gamma p(\gamma) d\gamma du \;=\; 0,$$

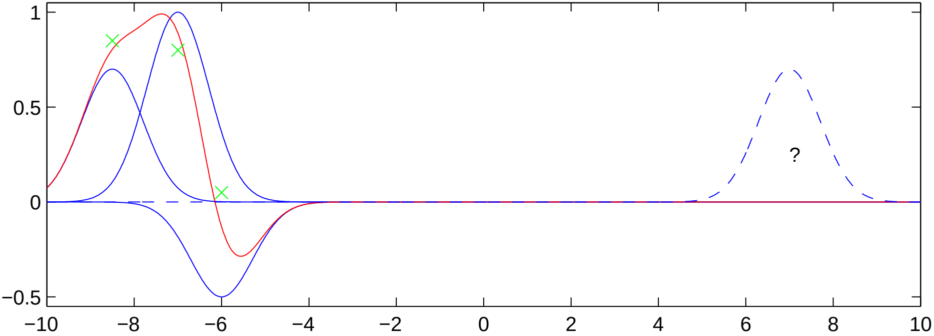
and the covariance function:

$$E[f(x)f(x')] \;=\; \int \exp\big(-(x-u)^2 - (x'-u)^2\big)du$$

$$= \int \exp\big(-2(u-\frac{x+x'}{2})^2 + \frac{(x+x')^2}{2} - x^2 - x'^2\big)\big)du \;\propto\; \exp\big(-\frac{(x-x')^2}{2}\big).$$

Thus, the squared exponential covariance function is equivalent to regression using infinitely many Gaussian shaped basis functions placed everywhere, not just at your training points!

Why it is dangerous to use only finitely many basis functions?



Rational quadratic covariance function

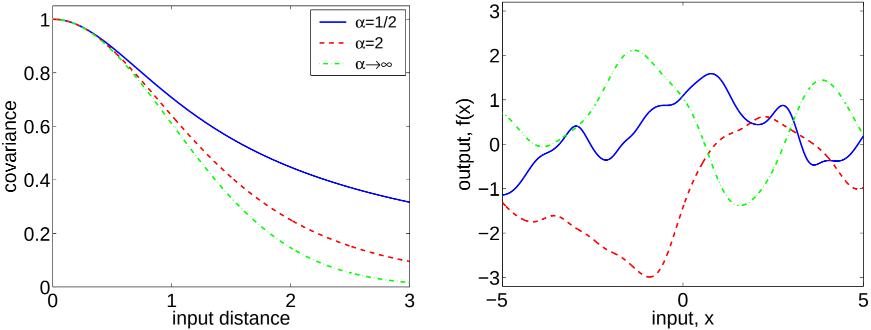
The *rational quadratic* (RQ) covariance function:

$$k_{\mathrm{RQ}}(r) \;=\; \Big(1 + \frac{r^2}{2\alpha\ell^2}\Big)^{-\alpha}$$

with $\alpha,\ \ell > 0$ can be seen as a *scale mixture* (an infinite sum) of squared exponential (SE) covariance functions with different characteristic length-scales. Using $\tau = \ell^{-2}$ and $p(\tau|\alpha,\beta) \propto \tau^{\alpha-1}\exp(-\alpha\tau/\beta)$:

$$\begin{aligned} k_{\mathrm{RQ}}(r) \;&=\; \int p(\tau|\alpha,\beta) k_{\mathrm{SE}}(r|\tau) d\tau \\ &\propto\; \int \tau^{\alpha-1}\exp\Big(-\frac{\alpha\tau}{\beta}\Big)\exp\Big(-\frac{\tau r^2}{2}\Big)d\tau \;\propto\; \Big(1 + \frac{r^2}{2\alpha\ell^2}\Big)^{-\alpha}. \end{aligned}$$

Rational quadratic covariance function II



The limit $\alpha \to \infty$ of the RQ covariance function is the SE.

Matérn covariance functions

Stationary covariance functions can be based on the Matérn form:

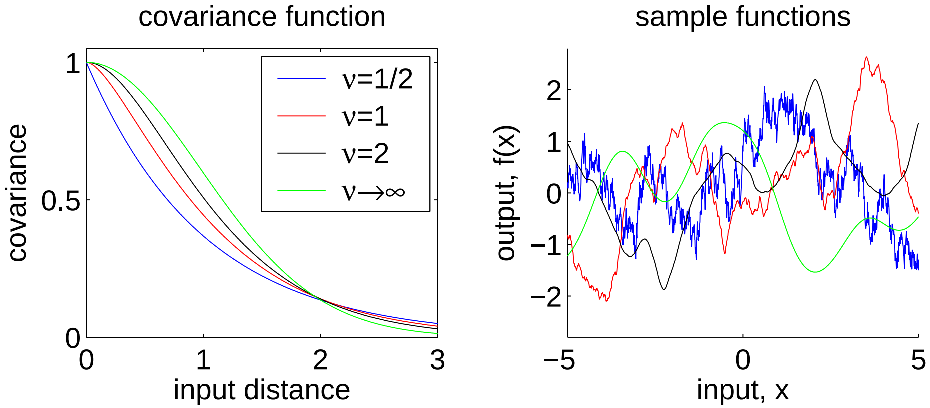
$$k(\mathbf{x},\mathbf{x}') = \frac{1}{\Gamma(\nu)2^{\nu-1}}\Big[\frac{\sqrt{2\nu}}{\kappa}|\mathbf{x}-\mathbf{x}'|\Big]^{\nu} K_{\nu}\Big(\frac{\sqrt{2\nu}}{\kappa}|\mathbf{x}-\mathbf{x}'|\Big),$$

where K_ν is the modified Bessel function of second kind of order ν, and κ is the characteristic length scale.

Sample functions from Matérn forms are $\lfloor \nu - 1 \rfloor$ times differentiable. Thus, the hyperparameter ν can control the degree of smoothness

Matérn covariance functions II

Univariate Matérn covariance function with unit characteristic length scale and unit variance:



Matérn covariance functions II

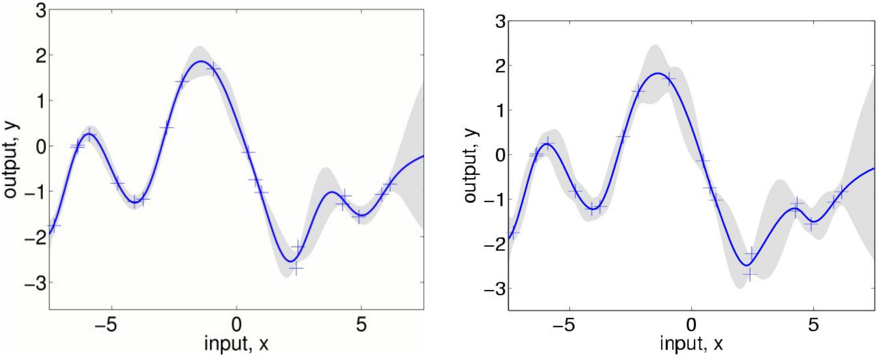
It is possible that the most interesting cases for machine learning are $\nu = 3/2$ and $\nu = 5/2$, for which

$$k_{\nu=3/2}(r) = \Big(1 + \frac{\sqrt{3}r}{\ell}\Big) \exp\Big(-\frac{\sqrt{3}r}{\ell}\Big),$$
$$k_{\nu=5/2}(r) = \Big(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\Big) \exp\Big(-\frac{\sqrt{5}r}{\ell}\Big),$$

Other special cases:

- $\nu = 1/2$: Laplacian covariance function, sample functions: stationary Browninan motion
- $\nu \to \infty$: Gaussian covariance function with smooth (infinitely differentiable) sample functions

A Comparison



Left, SE covariance function, log marginal likelihood −15.6, and right Matérn covariance function with $\nu = 3/2$, marginal likelihood −18.0.

GP regression recap

We use a Gaussian process prior for the latent function:

$$\mathbf{f}|X, \theta \sim \mathcal{N}(\mathbf{0}, K)$$

The likelihood is a factorized Gaussian

$$\mathbf{y}|\mathbf{f} \sim \prod_{i=1}^{m} \mathcal{N}(y_i|f_i, \sigma_n^2)$$

The posterior is Gaussian

$$p(\mathbf{f}|\mathcal{D}, \theta) = \frac{p(\mathbf{f}|X, \theta)\, p(\mathbf{y}|\mathbf{f})}{p(\mathcal{D}|\theta)}$$

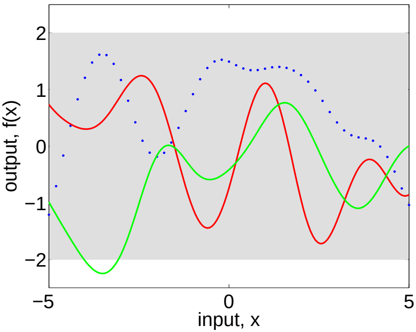
The latent value at the test point, $f(\mathbf{x}^*)$ is Gaussian

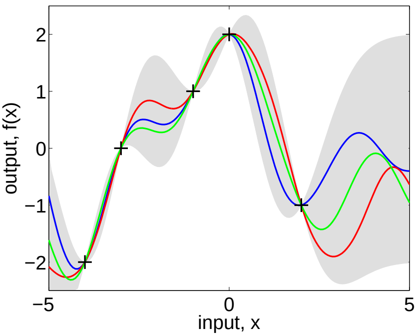
$$p(f_*|\mathcal{D}, \theta, \mathbf{x}_*) = \int p(f_*|\mathbf{f}, X, \theta, \mathbf{x}_*) p(\mathbf{f}|\mathcal{D}, \theta) d\mathbf{f},$$

and the predictive class probability is Gaussian

$$p(y_*|\mathcal{D}, \theta, \mathbf{x}_*) = \int p(y_*|f_*) p(f_*|\mathcal{D}, \theta, \mathbf{x}_*) df_*.$$

Prior and posterior





Predictive distribution:

$$p(y^*|x^*, \mathbf{x}, \mathbf{y}) \;\sim\; \mathcal{N}\big(\mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{y},$$
$$k(x^*, x^*) + \sigma^2_{\text{noise}} - \mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{k}(x^*, \mathbf{x})\big)$$

The marginal likelihood

To chose between models $M_1, M_2, \ldots$, compare the posterior for the models

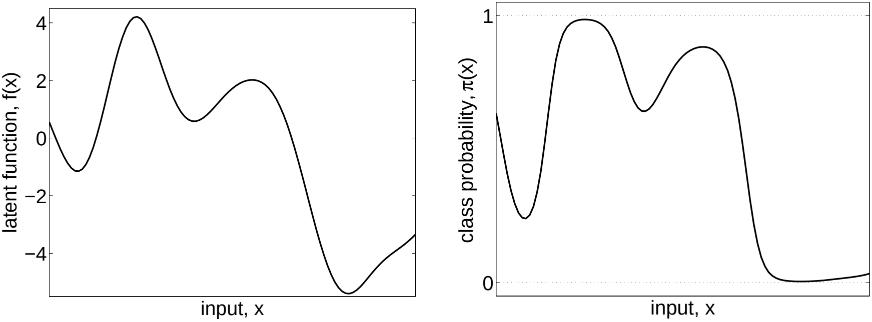
$$p(M_i|\mathcal{D}) \;=\; \frac{p(\mathbf{y}|\mathbf{x}, M_i)p(M_i)}{p(\mathcal{D})}.$$

Log marginal likelihood:

$$\log p(\mathbf{y}|\mathbf{x}, M_i) \;=\; -\frac{1}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi)$$

is the combination of a data fit term and complexity penalty. Occam's Razor is automatic.

Binary Gaussian Process Classification



The class probability is related to the *latent* function through:

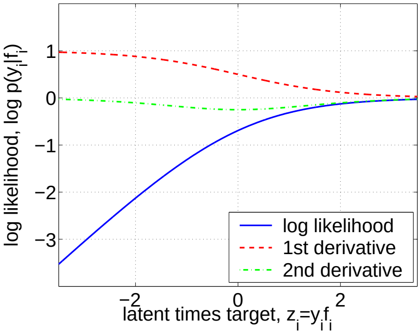
$$p(y = 1|f(\mathbf{x})) \;=\; \pi(\mathbf{x}) \;=\; \Phi\big(f(\mathbf{x})\big).$$

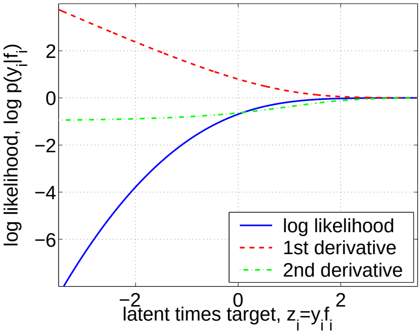
Observations are independent given f, so the likelihood is

$$p(\mathbf{y}|\mathbf{f}) \;=\; \prod_{i=1}^{n} p(y_i|f_i) \;=\; \prod_{i=1}^{n} \Phi(y_i f_i).$$

Likelihood functions

The logistic $(1 + \exp(-y_i f_i))^{-1}$ and probit $\Phi(y_i f_i)$ and their derivatives:





Exact expressions

We use a Gaussian process prior for the latent function:

$$\mathbf{f}|X,\theta \sim \mathcal{N}(\mathbf{0},\ K)$$

The posterior becomes:

$$p(\mathbf{f}|\mathcal{D},\theta) = \frac{p(\mathbf{f}|X,\theta)\, p(\mathbf{y}|\mathbf{f})}{p(\mathcal{D}|\theta)} = \frac{\mathcal{N}(\mathbf{f}|\mathbf{0},\ K)}{p(\mathcal{D}|\theta)} \prod_{i=1}^{m} \Phi(y_i f_i),$$

which is non-Gaussian.
The latent value at the test point, $f(\mathbf{x}^*)$ is

$$p(f_*|\mathcal{D},\theta,\mathbf{x}_*) = \int p(f_*|\mathbf{f},X,\theta,\mathbf{x}_*) p(\mathbf{f}|\mathcal{D},\theta) d\mathbf{f},$$

and the predictive class probability becomes

$$p(y_*|\mathcal{D},\theta,\mathbf{x}_*) = \int p(y_*|f_*) p(f_*|\mathcal{D},\theta,\mathbf{x}_*) df_*,$$

both of which are intractable to compute.

Gaussian Approximation to the Posterior

We approximate the non-Gaussian posterior by a Gaussian:

$$p(\mathbf{f}|\mathcal{D},\theta) \;\simeq\; q(\mathbf{f}|\mathcal{D},\theta) \;=\; \mathcal{N}(\mathbf{m},\ A)$$

then $q(f_*|\mathcal{D},\theta,\mathbf{x}_*) = \mathcal{N}(f_*|\mu_*,\sigma_*^2)$, where

$$\begin{aligned}\mu_* \;&=\; \mathbf{k}_*^\top K^{-1}\mathbf{m}\\ \sigma_*^2 \;&=\; k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top (K^{-1} - K^{-1}AK^{-1})\mathbf{k}_*.\end{aligned}$$

Using this approximation:

$$q(y_*=1|\mathcal{D},\theta,\mathbf{x}_*) \;=\; \int \Phi(f_*)\,\mathcal{N}(f_*|\mu_*,\sigma_*^2)df_* \;=\; \Phi\Big(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\Big)$$

What Gaussian?

Some suggestions:

- local expansion: Laplace's method
- optimize a variational lower bound (using Jensen's ineqality):

$$\log p(\mathbf{y}|X) \;=\; \log \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f} \;\geq\; \int \log\big(\frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{q(\mathbf{f})}\big)q(\mathbf{f})d\mathbf{f}$$

- the Expectation Propagation (EP) algorithm

Expectation Propagation

Posterior:

$$p(\mathbf{f}|X,\mathbf{y}) \;=\; \frac{1}{Z}p(\mathbf{f}|X)\prod_{i=1}^{n}p(y_i|f_i),$$

where the normalizing term is the marginal likelihood

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

Exact likelihood:

$$p(y_i|f_i) \;=\; \Phi(f_i y_i)$$

which makes inference intractable. In EP we use a *local likelihood approximation*

$$p(y_i|f_i) \;\simeq\; t_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2) \;\triangleq\; \tilde{Z}_i\mathcal{N}(f_i|\tilde{\mu}_i,\tilde{\sigma}_i^2),$$

where the *site parameters* are $\tilde{Z}_i$, $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$, such that:

$$\prod_{i=1}^{n}t_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2) \;=\; \mathcal{N}(\tilde{\boldsymbol{\mu}},\tilde{\Sigma})\prod_i\tilde{Z}_i.$$

Expectation Propagation II

We approximate the posterior by:

$$q(\mathbf{f}|X,\mathbf{y}) \triangleq \frac{1}{Z_{\text{EP}}} p(\mathbf{f}|X) \prod_{i=1}^{n} t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\boldsymbol{\mu}, \Sigma),$$

$$\text{with } \boldsymbol{\mu} = \Sigma\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}, \text{ and } \Sigma = (K^{-1} + \tilde{\Sigma}^{-1})^{-1},$$

How do we choose the site parameters?

Key idea: iteratively update each site in turn, based on approximation so far.

The approximate posterior for f_i contains three kinds of terms:

1. the prior $p(\mathbf{f}|X)$
2. the approximate likelihoods t_j for all cases $j \neq i$
3. the exact likelihood for case i, $p(y_i|f_i)$.

The Cavity distribution

The *cavity* distribution

$$q_{-i}(f_i) \;\propto\; \int p(\mathbf{f}|X) \prod_{j \neq i} t_j(f_j|\tilde{Z}_j, \tilde{\mu}_j, \tilde{\sigma}_j^2) df_j,$$

can be found by "removing" one term from the posterior:

$$q(f_i|X, \mathbf{y}) \;=\; \mathcal{N}(f_i|\mu_i, \sigma_i^2)$$

to get:

$$\begin{aligned} q_{-i}(f_i) \;&\triangleq\; \mathcal{N}(f_i|\mu_{-i}, \sigma_{-i}^2), \\ \text{where} \;\; \mu_{-i} \;&=\; \sigma_{-i}^2(\sigma_i^{-2}\mu_i - \tilde{\sigma}_i^{-2}\tilde{\mu}_i), \;\; \text{and} \;\; \sigma_{-i}^2 \;=\; (\sigma_i^{-2} - \tilde{\sigma}_i^{-2})^{-1}. \end{aligned}$$

Now, find $\hat{q}(f_i)$ which matches the desired:

$$\hat{q}(f_i) \;\triangleq\; \hat{Z}_i \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2) \;\simeq\; q_{-i}(f_i) p(y_i|f_i).$$

by matching moments.

Expectation Propagation III

The desired moments can be computed in closed form:

$$\hat{Z}_i = \Phi(z_i), \qquad \hat{\mu}_i = \mu_{-i} + \frac{y_i \sigma^2_{-i} \mathcal{N}(z_i)}{\Phi(z_i)\sqrt{1+\sigma^2_{-i}}},$$

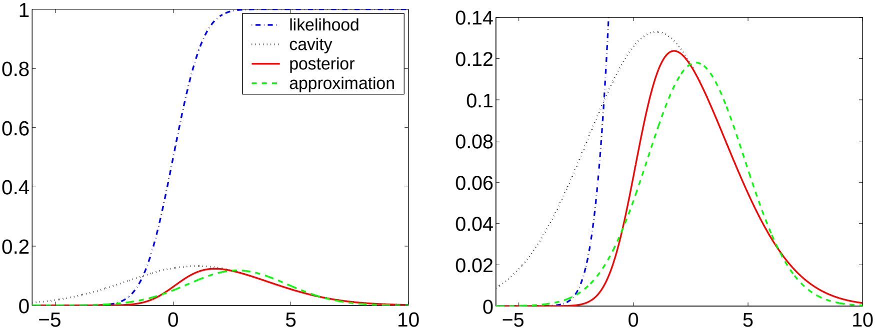
$$\hat{\sigma}^2_i = \sigma^2_{-i} - \frac{\sigma^4_{-i}\mathcal{N}(z_i)}{(1+\sigma^2_{-i})\Phi(z_i)}\Big(z_i + \frac{\mathcal{N}(z_i)}{\Phi(z_i)}\Big), \quad \text{where} \quad z_i = \frac{y_i \mu_{-i}}{\sqrt{1+\sigma^2_{-i}}}.$$

These moments are achieved by setting the site parameters to:

$$\tilde{\mu}_i = \tilde{\sigma}^2_i(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma^{-2}_{-i}\mu_{-i}), \qquad \tilde{\sigma}^2_i = (\hat{\sigma}_i^{-2} - \sigma^{-2}_{-i})^{-1},$$

$$\tilde{Z}_i = \hat{Z}_i\sqrt{2\pi}\sqrt{\sigma^2_{-i} + \tilde{\sigma}^2_i}\exp\big(\tfrac{1}{2}(\mu_{-i} - \tilde{\mu}_i)^2/(\sigma^2_{-i} + \tilde{\sigma}^2_i)\big),$$

The EP approximation



Predictive distribution

The latent predictive mean:

$$\mathbb{E}_q[f_*|X,\mathbf{y},\mathbf{x}_*] \;=\; \mathbf{k}_*^\top K^{-1}\boldsymbol{\mu} \;=\; \mathbf{k}_*^\top K^{-1}(K^{-1}+\tilde{\Sigma}^{-1})^{-1}\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\mu}} \;=\; \mathbf{k}_*^\top(K+\tilde{\Sigma})^{-1}\tilde{\boldsymbol{\mu}}.$$

and variance:

$$\mathbb{V}_q[f_*|X,\mathbf{y},\mathbf{x}_*] \;=\; k(\mathbf{x}_*,\mathbf{x}_*) - \mathbf{k}_*^\top(K+\tilde{\Sigma})^{-1}\mathbf{k}_*,$$

which can be plugged into the class probability equation:

$$q(y_*=1|\mathcal{D},\theta,\mathbf{x}_*) \;=\; \int \Phi(f_*)\,\mathcal{N}(f_*|\mu_*,\sigma_*^2)df_* \;=\; \Phi\Big(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\Big)$$

Marginal Likelihood

The EP approximation for the marginal likelihood:

$$Z_{\mathrm{EP}} \;=\; q(\mathbf{y}|X) \;=\; \int p(\mathbf{f}) \prod_{i=1}^{n} t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) d\mathbf{f}.$$

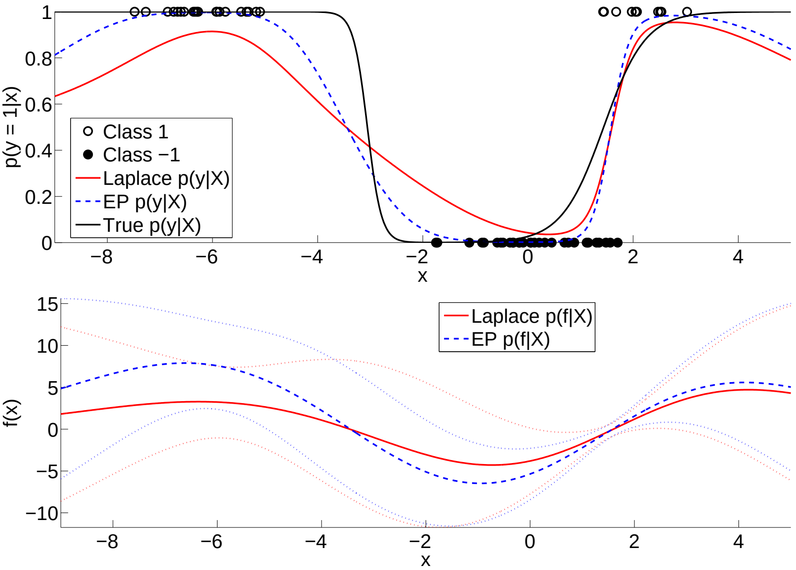
which evaluates to:

$$\begin{aligned}\log(Z_{\mathrm{EP}}|\boldsymbol{\theta}) \;=\; & -\frac{1}{2}\log|K+\tilde{\Sigma}| - \frac{1}{2}\tilde{\boldsymbol{\mu}}^\top (K+\tilde{\Sigma})^{-1}\tilde{\boldsymbol{\mu}} \\ & + \sum_{i=1}^{n} \log \Phi\Big(\frac{y_i \mu_{-i}}{\sqrt{1+\sigma_{-i}^2}}\Big) + \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_{-i}^2 + \tilde{\sigma}_i^2) + \sum_{i=1}^{n}\frac{(\mu_{-i}-\tilde{\mu}_i)^2}{2(\sigma_{-i}^2+\tilde{\sigma}_i^2)},\end{aligned}$$

which has a nice interpretation.

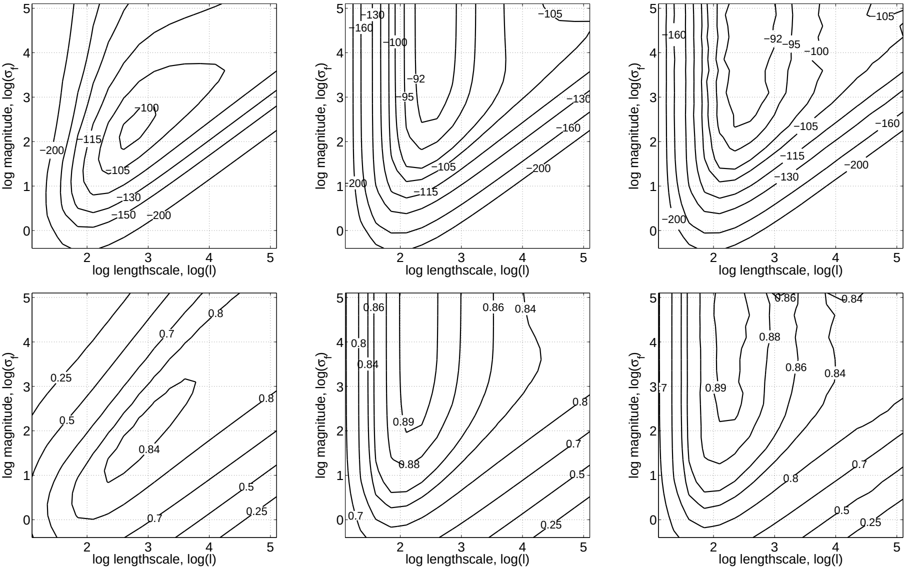
It is possible to analytically evaluate the derivatives of the estimated log marginal likelihood w.r.t. the hyperparameters.

Example

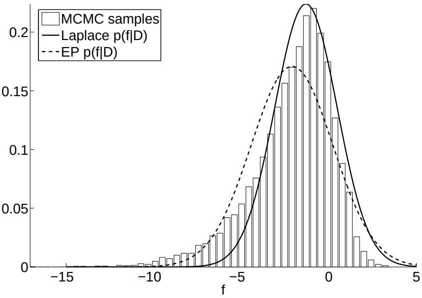


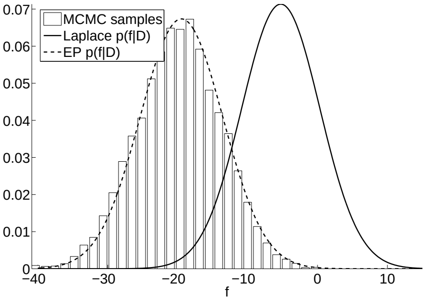
1
0.8
0.6
0.4
0.2
0
p(y = 1|x)
Class 1
Class −1
Laplace p(y|X)
EP p(y|X)
True p(y|X)
−8
−6
−4
−2
0
2
4
x
15
10
5
0
−5
−10
f(x)
Laplace p(f|X)
EP p(f|X)
x

USPS Digits, 3s vs 5s

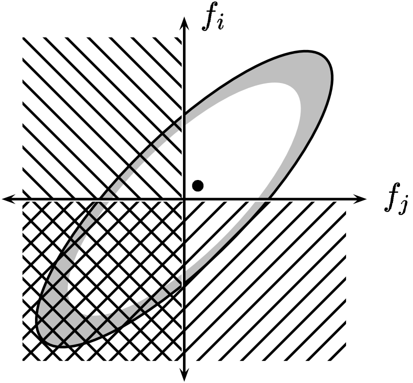


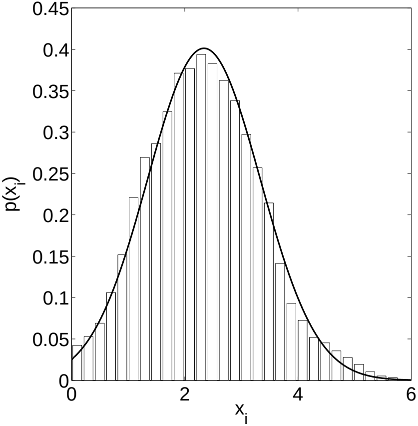
USPS Digits, 3s vs 5s





The Structure of the posterior





Conclusions

Covariance functions for Gaussian processes

- encodes useful information about the functions
- can be *learnt* from the data

Whereas inference for regression with Gaussian noise can be done in closed form

- non-Gaussian likelihoods (as eg in classification) cannot
- (many) good approximations exist

For the details: Rasmussen and Williams 'Gaussian Processes for Machine Learning', the MIT Press 2006.

For the (matlab) code `www.GaussianProcess.org/gpml`.