

ADVANCED BAYESIAN LEARNING

GAUSSIAN PROCESSES

SPRING 2014

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

TOPIC OVERVIEW

- ▶ Gaussian process regression
 - ▶ Recall: Bayesian inference for Gaussian linear/nonlinear regression.
 - ▶ Gaussian processes for nonparametric regression
 - ▶ Covariance kernels
 - ▶ Properties of GPs
 - ▶ Selecting the kernel and hyperparameters
- ▶ Gaussian process classification
 - ▶ Flexible classification using GPs
 - ▶ Laplace approximation of the posterior
- ▶ Main literature: Rasmussen and Williams (2006). *Gaussian Processes for Machine Learning*.

FLEXIBLE NONLINEAR REGRESSION

- ▶ **Linear regression**

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

$$f(\mathbf{x}) = \mathbf{x}^T \cdot \mathbf{w}$$

and $\epsilon \sim N(0, \sigma_n^2)$ and iid over observations.

- ▶ The weights \mathbf{w} are called regression coefficients (β) in statistics.

- ▶ **Polynomial regression:** $\mathbf{x} = (1, x, x^2, x^3, \dots, x^k)^T$.

- ▶ **Spline regression:**

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \cdot \mathbf{w}$$

where $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_N(\mathbf{x}))^T$ for N basis functions

- ▶ Example: **thin plate splines** with N knots $\kappa_1, \dots, \kappa_N$ in \mathbf{x} -space

$$\phi_k(\mathbf{x}) = \ln(\|\mathbf{x} - \kappa_k\|) \|\mathbf{x} - \kappa_k\|^2$$

BAYESIAN LINEAR REGRESSION - INFERENCE

- ▶ \mathbf{w} is unknown. σ_n is assumed known.
- ▶ **Prior** [note: RW do *not* use $\Sigma_p = \sigma_n^2 \Omega$]

$$\mathbf{w} \sim N(0, \Sigma_p)$$

- ▶ **Posterior** [note: \mathbf{X} is $D \times n$]

$$\mathbf{w} | \mathbf{X}, \mathbf{y} \sim N(\bar{\mathbf{w}}, \mathbf{A}^{-1})$$

$$\mathbf{A} = \sigma_n^{-2} \mathbf{X} \mathbf{X}^T + \Sigma_p^{-1}$$

$$\bar{\mathbf{w}} = \sigma_n^{-2} \mathbf{A}^{-1} \mathbf{X} \mathbf{y} = \left(\mathbf{X} \mathbf{X}^T + \sigma_n^2 \Sigma_p^{-1} \right)^{-1} \mathbf{X} \mathbf{y}$$

- ▶ Recall: **Posterior precision = Data Precision + Prior Precision.**
- ▶ Marginal posterior of \mathbf{w} is multivariate student- t when σ_n^2 is unknown with $\text{Inv-}\chi^2$ prior (and $\Sigma_p = \sigma_n^2 \Omega$).

BAYESIAN LINEAR REGRESSION - PREDICTION

- **Predictive density for mean** $f(\mathbf{x}_*)$ at new location \mathbf{x}_*

$$f(\mathbf{x}_*)|\mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim N\left(\mathbf{x}_*^T \bar{\mathbf{w}}, \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{x}_*\right)$$

- Proof: $f(\mathbf{x}_*) = \mathbf{x}_*^T \mathbf{w}$ and \mathbf{w} has a normal posterior. Linear combs of normals is normal.

- **Predictive density for new response** y_*

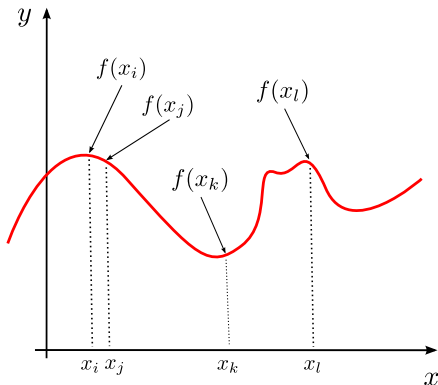
$$y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y} \sim N\left(\mathbf{x}_*^T \bar{\mathbf{w}}, \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{x}_* + \sigma_n^2\right)$$

- Replace \mathbf{X} with $\Phi(\mathbf{X})$ in the above for the case with basis expansion (e.g. splines).

NON-PARAMETRIC REGRESSION

- ▶ **Non-parametric regression:** avoiding a parametric form for $f(\cdot)$. Treat $f(\mathbf{x})$ as an unknown parameter for every \mathbf{x} .
- ▶ **Weight space view**
 - ▶ Restrict attention to a grid of (ordered) x -values: x_1, x_2, \dots, x_k .
 - ▶ Put a joint prior on the k function values: $f(x_1), f(x_2), \dots, f(x_k)$.
- ▶ **Function space view**
 - ▶ Treat f as an **unknown function**.
 - ▶ Put a **prior over a set of functions**.
- ▶ Kolmogorov's existence theorem for stochastic processes equates the two views. Just make sure that the set of finite dimensional distributions are consistent: adding or deleting variables does not change the marginal of the original variable set.

NONPARAMETRIC = ONE PARAMETER FOR EVERY x !



THE MULTIVARIATE NORMAL DISTRIBUTION

- ▶ The **density function** of a p -variate normal vector $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$f(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{p/2} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- ▶ Example: **Bivariate normal** ($p = 2$)

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- ▶ **Linear combinations.** Let $\mathbf{y} = \mathbf{B}\mathbf{x} + \mathbf{b}$, where \mathbf{x} is $p \times 1$ and \mathbf{B} is a $m \times p$ constant matrix. Then

$$\mathbf{y} \sim N(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$$

THE MULTIVARIATE NORMAL DISTRIBUTION, CONT.

- ▶ Let $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$ where \mathbf{x}_1 is $p_1 \times 1$ and \mathbf{x}_2 is $p_2 \times 1$ ($p_1 + p_2 = p$).
- ▶ Partition μ and Σ accordingly as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

- ▶ **Marginals are normal.** Let $\mathbf{x} \sim N(\mu, \Sigma)$, then

$$\mathbf{x}_1 \sim N(\mu_1, \Sigma_{11})$$

- ▶ **Conditionals are normal.** Let $\mathbf{x} \sim N(\mu, \Sigma)$, then

$$\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{x}_2^* \sim N \left[\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{x}_2^* - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right]$$

- ▶ Life is beautiful ...

GAUSSIAN PROCESS REGRESSION

- ▶ Weight-space view. GP assumes

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

- ▶ But how do we specify the $k \times k$ **covariance matrix** \mathbf{K} ?

$$\text{Cov}(f(x_p), f(x_q))$$

GAUSSIAN PROCESS REGRESSION

- ▶ Weight-space view. GP assumes

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

- ▶ But how do we specify the $k \times k$ **covariance matrix** \mathbf{K} ?

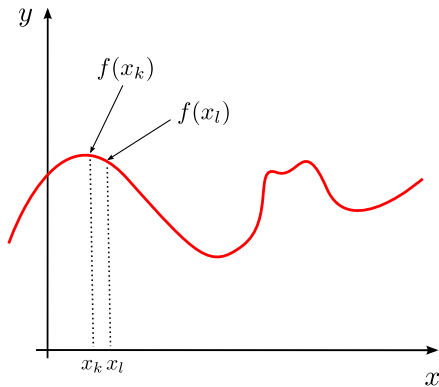
$$\text{Cov}(f(x_p), f(x_q))$$

- ▶ **Squared exponential covariance function**

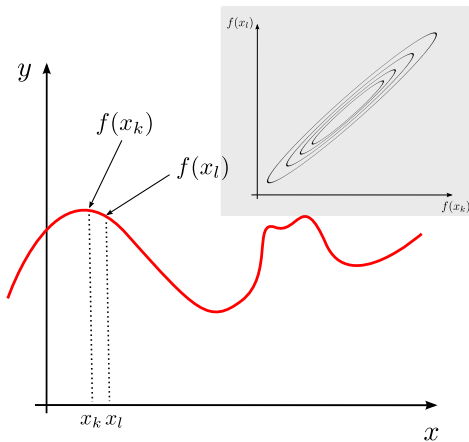
$$\text{Cov}(f(x_p), f(x_q)) = k(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{1}{2} (x_p - x_q)^2\right)$$

- ▶ Nearby x 's have highly correlated function ordinates $f(x)$.
- ▶ We can compute $\text{Cov}(f(x_p), f(x_q))$ for *any* x_p and x_q .
- ▶ Extension to multiple covariates: $(x_p - x_q)$ replaced by $\|\mathbf{x}_p - \mathbf{x}_q\|$.

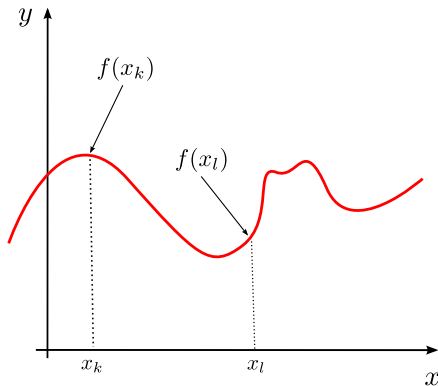
SMOOTH FUNCTION - POINTS NEARBY



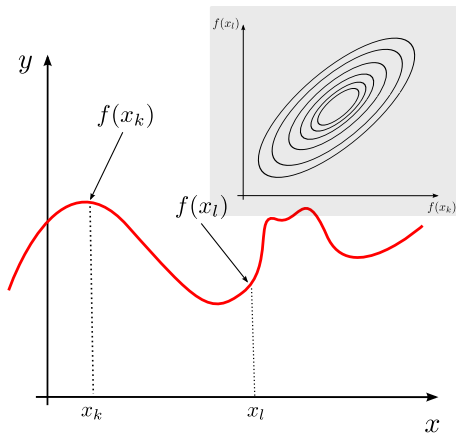
SMOOTH FUNCTION - POINTS NEARBY



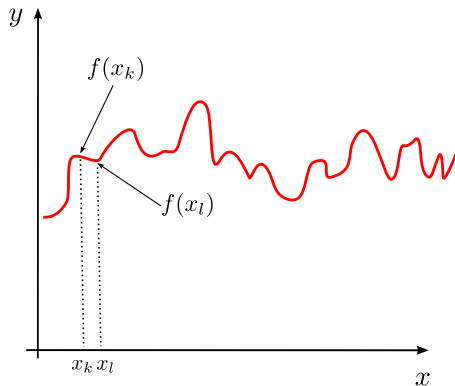
SMOOTH FUNCTION - POINTS FAR APART



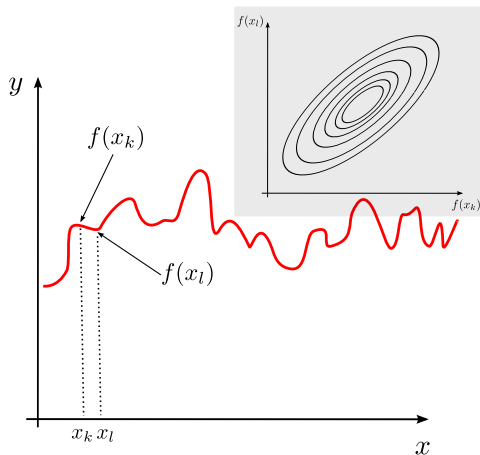
SMOOTH FUNCTION - POINTS FAR APART



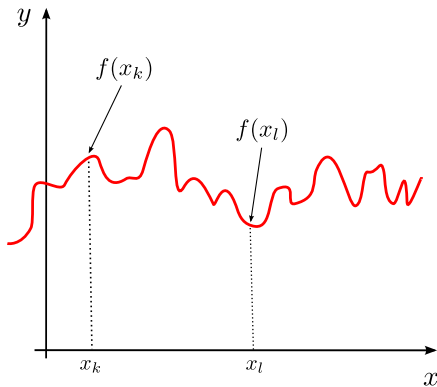
JAGGED FUNCTION - POINTS NEARBY



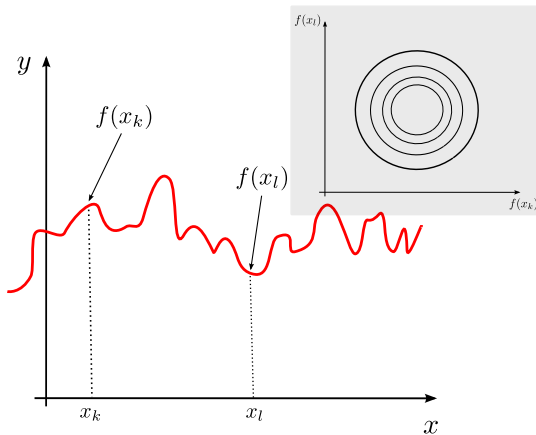
JAGGED FUNCTION - POINTS NEARBY



JAGGED FUNCTION - POINTS FAR APART



JAGGED FUNCTION - POINTS FAR APART



GAUSSIAN PROCESS REGRESSION, CONT.

DEFINITION

A **Gaussian process (GP)** is a collection of random variables, any finite number of which have a multivariate Gaussian distribution.

- ▶ A Gaussian process is really a **probability distribution over functions** (curves). No need for a grid!
- ▶ A GP is completely specified by a **mean** and a **covariance function**

$$m(x) = E[f(x)]$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x')))]$$

for any two inputs x and x' (note: this is *not* the transpose here).

- ▶ A **Gaussian process** is denoted by

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

- ▶ **Bayesian:** $f(x) \sim GP$ encodes **prior beliefs** about the unknown $f(\cdot)$.

A SIMPLE GP EXAMPLE

- ▶ Example:

$$m(x) = \sin(x)$$

$$k(x, x') = \sigma_f^2 \exp \left(-\frac{1}{2} \left(\frac{x - x'}{\ell} \right)^2 \right)$$

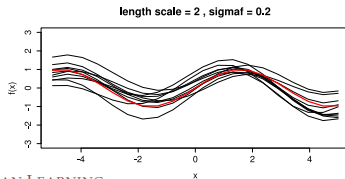
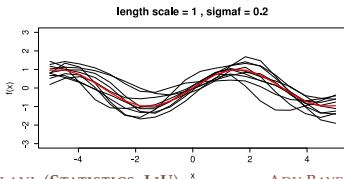
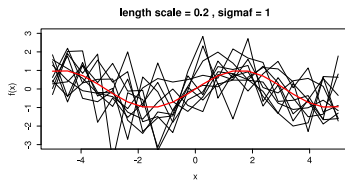
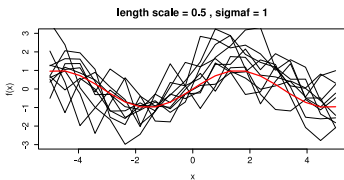
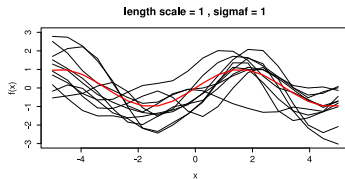
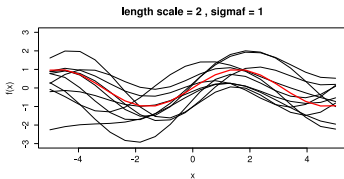
where $\ell > 0$ is the length scale.

- ▶ Larger ℓ gives more smoothness in $f(x)$.
- ▶ Simulate draw from $f(x) \sim GP(m(x), k(x, x'))$ over a grid $\mathbf{x}_* = (x_1, \dots, x_n)$ by using that

$$f(\mathbf{x}_*) \sim N(m(\mathbf{x}_*), K(\mathbf{x}_*, \mathbf{x}_*))$$

- ▶ Note that the **kernel** $k(x, x')$ produces a **covariance matrix** $K(\mathbf{x}_*, \mathbf{x}_*)$ when evaluated at the vector \mathbf{x}_* .

SIMULATING A GP - SINE MEAN AND SE KERNEL



SIMULATING A GP

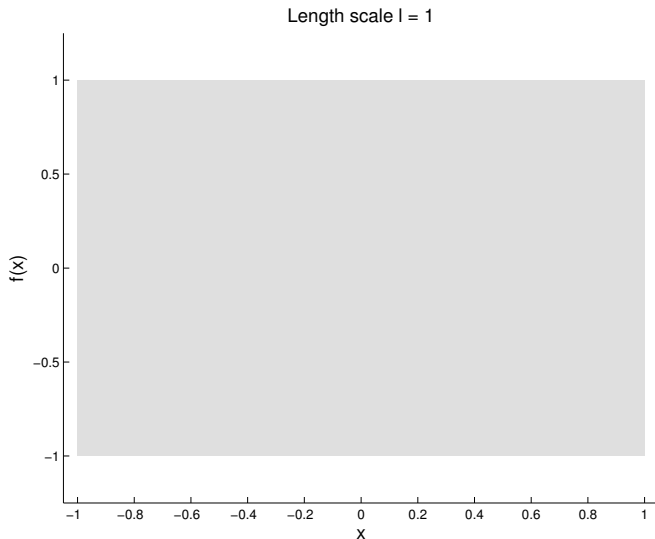
- ▶ The joint way: Choose a grid x_1, \dots, x_k . Simulate the k -vector

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_k) \end{pmatrix} \sim N(\mathbf{m}, \mathbf{K})$$

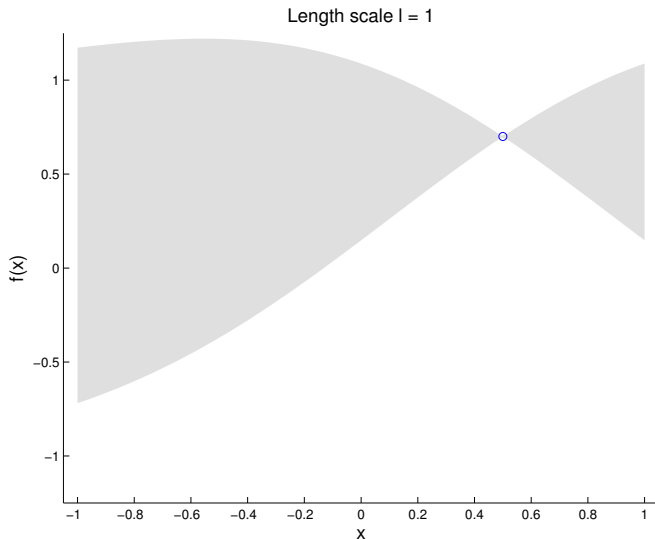
- ▶ More intuition from the conditional decomposition

$$\begin{aligned} p(f(x_1), f(x_2), \dots, f(x_k)) &= p(f(x_1)) p(f(x_2)|f(x_1)) \cdots \\ &\quad \times p(f(x_k)|f(x_1), \dots, f(x_{k-1})) \end{aligned}$$

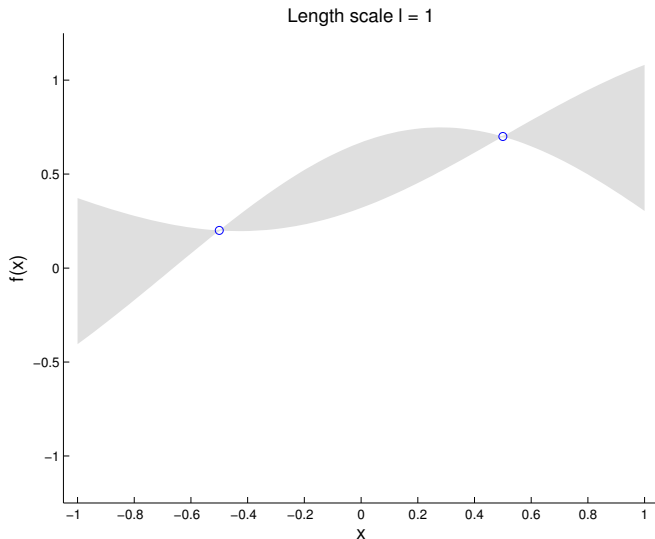
DENSITY BEFORE FIRST DRAW



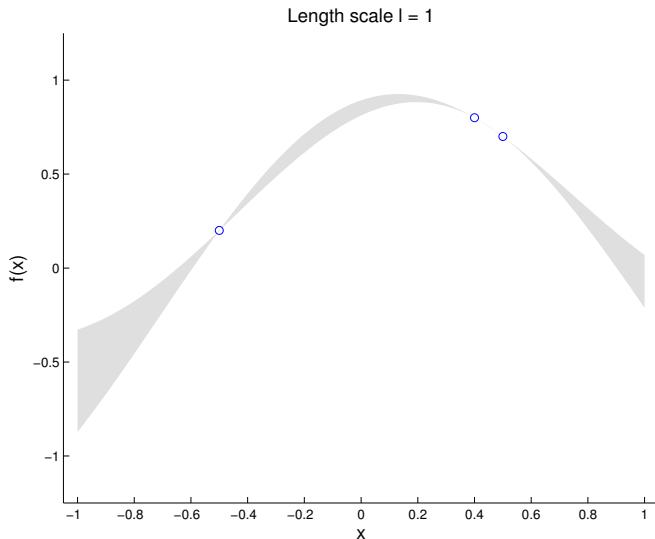
DENSITY BEFORE SECOND DRAW



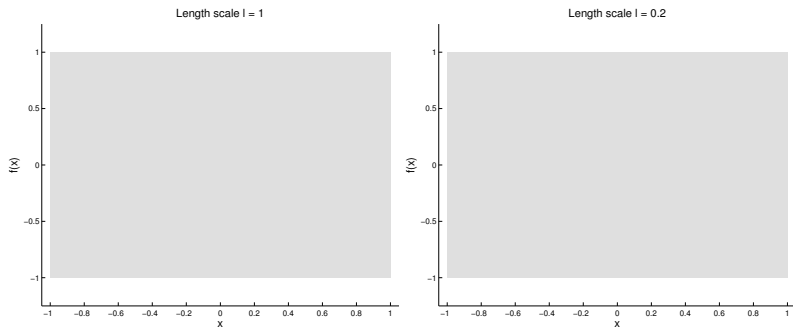
DENSITY BEFORE THIRD DRAW



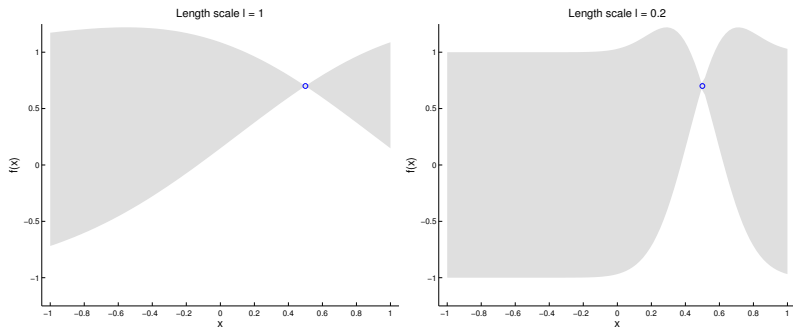
DENSITY BEFORE FOURTH DRAW



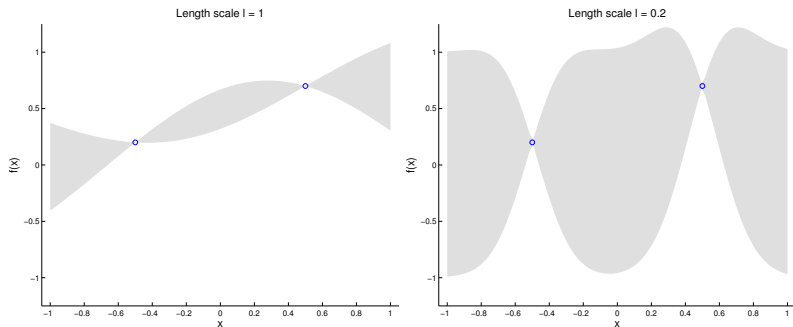
SIMULATION FROM $L=1$ VS $L=0.2$. BEFORE FIRST DRAW.



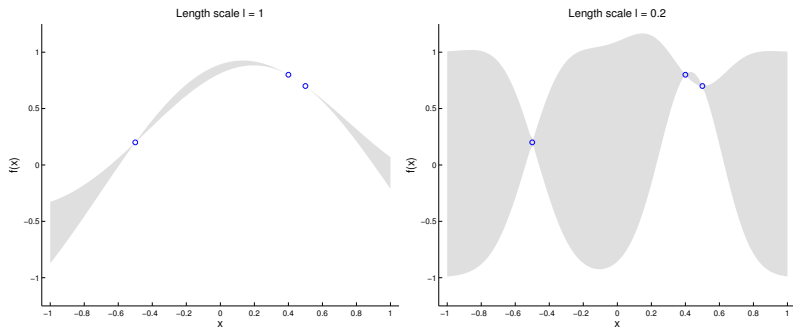
SIMULATION FROM $L=1$ VS $L=0.2$. BEFORE SECOND DRAW.



SIMULATION FROM $L=1$ VS $L=0.2$. BEFORE THIRD DRAW.



SIMULATION FROM $L=1$ VS $L=0.2$. BEFORE FOURTH DRAW.



THE POSTERIOR FOR A GPR

- ▶ **Model**

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

- ▶ **Prior**

$$f(x) \sim GP(0, k(x, x'))$$

- ▶ You have observed the data: $\mathbf{x} = (x_1, \dots, x_n)'$ and $\mathbf{y} = (y_1, \dots, y_n)'$.
- ▶ Goal: the posterior of $f(\cdot)$ over a grid of x -values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.

THE POSTERIOR FOR A GPR

► Model

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

► Prior

$$f(x) \sim GP(0, k(x, x'))$$

- You have observed the data: $\mathbf{x} = (x_1, \dots, x_n)'$ and $\mathbf{y} = (y_1, \dots, y_n)'$.
- Goal: the posterior of $f(\cdot)$ over a grid of x -values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.
- Intermediate step: joint distribution of \mathbf{y} and \mathbf{f}_*

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma^2 I & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right\}$$

THE POSTERIOR FOR A GPR

► Model

$$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

► Prior

$$f(x) \sim GP(0, k(x, x'))$$

► You have observed the data: $\mathbf{x} = (x_1, \dots, x_n)'$ and $\mathbf{y} = (y_1, \dots, y_n)'$.

► Goal: the posterior of $f(\cdot)$ over a grid of x -values: $\mathbf{f}_* = \mathbf{f}(\mathbf{x}_*)$.

► Intermediate step: joint distribution of \mathbf{y} and \mathbf{f}_*

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) + \sigma^2 I & K(\mathbf{x}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{x}) & K(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix} \right\}$$

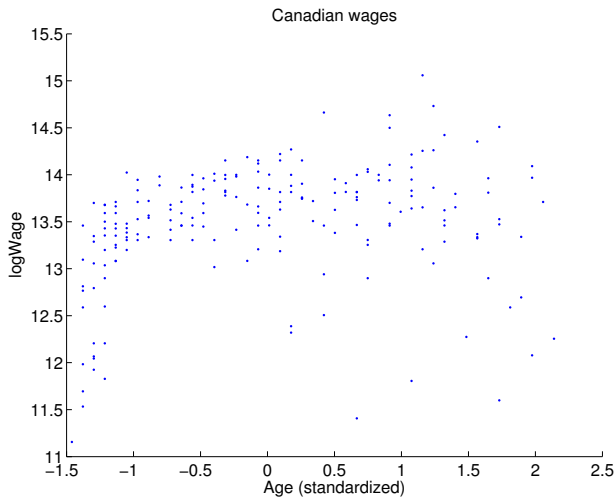
► The posterior

$$\mathbf{f}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_* \sim N(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*))$$

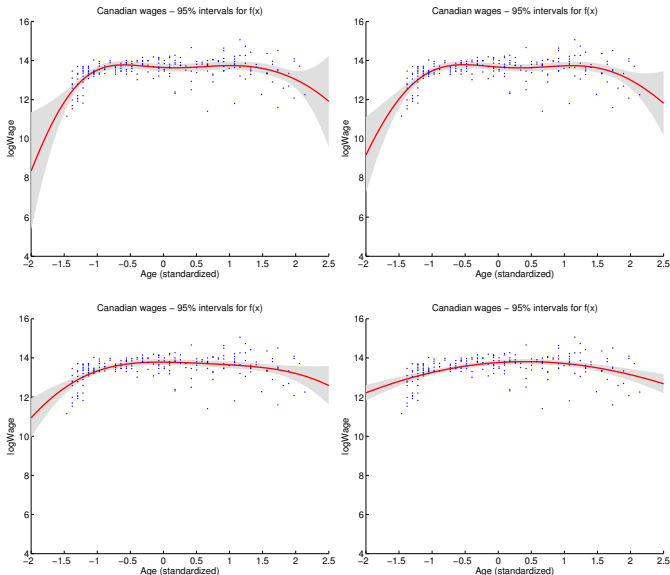
$$\bar{\mathbf{f}}_* = K(\mathbf{x}_*, \mathbf{x}) [K(\mathbf{x}, \mathbf{x}) + \sigma^2 I]^{-1} \mathbf{y}$$

$$\text{cov}(\mathbf{f}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{x}) [K(\mathbf{x}, \mathbf{x}) + \sigma^2 I]^{-1} K(\mathbf{x}, \mathbf{x}_*)$$

EXAMPLE - CANADIAN WAGES



POSTERIOR OF F - $\ell = 0.2, 0.5, 1, 2$



PREDICTION AND DECISION

- ▶ Predicting a new set of y -values $\mathbf{y}_* = f(\mathbf{x}_*) + \epsilon$ is easy

$$\mathbf{y}_* | \mathbf{x}, \mathbf{y}, \mathbf{x}_* \sim N(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*) + \sigma_n^2 \mathbf{I})$$

- ▶ Choosing a point prediction \mathbf{y}_{guess} by maximizing expected utility

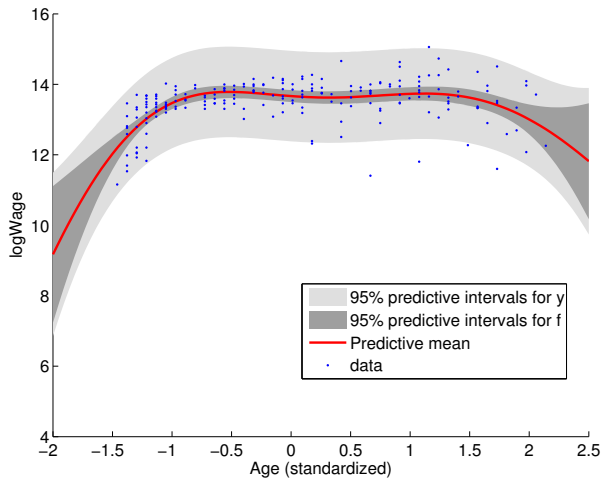
$$\bar{\mathcal{U}}(\mathbf{y}_{guess} | \mathbf{x}_*) = \int \mathcal{U}(\mathbf{y}_*, \mathbf{y}_{guess}) p(\mathbf{y}_* | \mathbf{x}_*, \mathbf{y}, \mathbf{x}) d\mathbf{y}_*$$

- ▶ Have to make a decision $a \in \mathcal{A}$ whose consequences (utility) depends on the uncertain \mathbf{f}_* (or \mathbf{y}_*)? Just maximize expected utility

$$\bar{\mathcal{U}}(a) = \int \mathcal{U}(a, \mathbf{f}_*) p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{y}, \mathbf{x}) d\mathbf{f}_*$$

where $\mathcal{U}(a, \mathbf{f}_*)$ is the utility of action $a \in \mathcal{A}$ if \mathbf{f}_* turns out to be the “true state of the world”.

CANADIAN WAGES - PREDICTION WITH $\ell = 0.5$



STATIONARY PROCESSES AND SMOOTHNESS

- ▶ A stochastic process (field) $\{f(\mathbf{x}), \mathbf{x} \in \mathbb{R}^D\}$ is **weakly stationary** if $E(f(\mathbf{x})) = \mu$ and its covariance function $k(\mathbf{x}, \mathbf{x}')$ is a function of $\mathbf{t} = \mathbf{x} - \mathbf{x}'$

$$k(\mathbf{x}, \mathbf{x}') = \text{Cov} [f(\mathbf{x}), f(\mathbf{x}')] = k(\mathbf{t}).$$

- ▶ The covariance function is **isotropic** if it only depends on the distance $t = \|\mathbf{x} - \mathbf{x}'\|$ (invariant to directions)

$$k(\mathbf{x}, \mathbf{x}') = \text{Cov} [f(\mathbf{x}), f(\mathbf{x}')] = k(t).$$

- ▶ The **smoothness** of a stationary process is determined by the smoothness of the covariance function.
- ▶ A stationary (isotropic) process is **continuous in quadratic mean**

$$E \left(|f(\mathbf{x} + t) - f(\mathbf{x})|^2 \right) \rightarrow 0 \text{ as } t \rightarrow 0$$

iff $k(t)$ is continuous at $t = 0$.

- ▶ A little more is required to guarantee **continuous sample paths** (continuous $f(\mathbf{x} + t, \omega)$ for any $\mathbf{x} \in \mathbb{R}^D$ and $\omega \in \Omega$).

KERNELS AND SPECTRAL DENSITIES

- ▶ $k(\mathbf{x}, \mathbf{x}')$ is a **covariance function** (i.e. positive definite) \rightarrow the $n \times n$ **Gram matrix** $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1,\dots,n}$ is a **covariance matrix**.
- ▶ **Bochner's theorem**: A complex valued function $k(\cdot)$ on \mathbb{R}^D is the covariance function of a weakly stationary continuous complex-valued stochastic process on \mathbb{R}^D iff

$$k(\mathbf{t}) = \int_{\mathbb{R}^D} e^{2\pi i \mathbf{s} \cdot \mathbf{t}} S(\mathbf{s}) d\mathbf{s}$$

- ▶ $S(\mathbf{s})$ is the **spectral density**. $S(\mathbf{s})$ is the energy allocated to the complex exponential basis function $e^{2\pi i \mathbf{s} \cdot \mathbf{t}}$ at frequency \mathbf{s} .
- ▶ For real-valued processes, think of $e^{2\pi i \mathbf{s} \cdot \mathbf{t}}$ as a multi-dimensional sine wave with frequency \mathbf{s} and amplitude $S(\mathbf{s})$.
- ▶ Spectral density \iff Covariance function of stationary process
 \iff Smoothness properties of the process.

COMMONLY USED COVARIANCE KERNELS

- ▶ Let $r = \|x - x'\|$. All kernels can be scaled by $\sigma_f > 0$.
- ▶ **Squared exponential (SE)** ($\ell > 0$)

$$K_{SE}(r) = \exp\left(-\frac{r^2}{2\ell^2}\right)$$

- ▶ Spectral density $S(s) = (2\pi\ell^2)^{D/2} \exp(-2\pi^2\ell^2s^2)$. Higher frequencies tail off like a Gaussian (i.e. quickly).
- ▶ Infinitely mean square differentiable. Very smooth.
- ▶ **Matérn** ($\ell > 0, \nu > 0$)

$$K_{Matern}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{\ell}\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}r}{\ell}\right)$$

- ▶ Spectral density behaves like a student- t density with 2ν degrees of freedom. For $\nu = 1/2$, $S(s)$ is Cauchy. Much weight on high frequency. Very rough.
- ▶ $\nu = 3/2$ and $\nu = 5/2$ most useful for ML. As $\nu \rightarrow \infty$, Matérn's kernel approaches SE kernel.

COMMONLY USED COVARIANCE KERNELS, CONT.

- ▶ **γ -exponential** ($\ell > 0$, $0 < \gamma \leq 2$)

$$K_{\gamma}(r) = \exp \left[- \left(\frac{r}{\ell} \right)^{\gamma} \right]$$

- ▶ Mean square differentiable only when $\gamma = 2$ (SE).

- ▶ **Rational quadratic** ($\ell > 0$, $\alpha > 0$)

$$K_{RQ}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2} \right)^{-\alpha}$$

- ▶ Scale mixture of SE covariance functions with different length-scales.
 - ▶ $K_{RQ}(r)$ approaches the SE kernel as $\alpha \rightarrow \infty$.
- ▶ $k(r) = \int \exp(-r^2/2\ell^2) p(\ell) d\ell$ is the most general representation of an isotropic kernel with a valid covariance function in all dimensions D .

MORE ON KERNELS

- ▶ Anisotropic version of isotropic kernels by setting $r^2(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')$ where \mathbf{M} is positive definite.
- ▶ **Automatic Relevance Determination (ARD):**
 $\mathbf{M} = \text{Diag}(\ell_1^{-2}, \dots, \ell_D^{-2})$ is diagonal with different length scales.
- ▶ **Factor kernels:** $M = \Lambda \Lambda^T + \Psi$, where Λ is $D \times k$ for low rank k .
- ▶ Length-scales $\ell(\mathbf{x})$ that vary with \mathbf{x} . Non-trivial to make positive definite, but see Gibbs kernel in Eq. 4.32.
- ▶ Kernels are often combined into **composite kernels**. Sum of kernels is a kernel. Product of kernels is a kernel.
- ▶ Kernels can be used for non-vectorial inputs by defining distance function between objects (e.g. words). String kernels for text analysis. Fisher kernels.

BAYESIAN INFERENCE FOR HYPERPARAMETERS

- ▶ Kernel depends on hyperparameters θ . Example SE kernel $[\theta = (\sigma_f, \ell)^T]$

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp \left(-\frac{1}{2} \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\ell^2} \right)$$

- ▶ If the hyperparameters are unknown, just compute the posterior

$$p(\theta | \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} | \mathbf{X}, \theta) p(\theta | \mathbf{X}).$$

- ▶ We need to compute

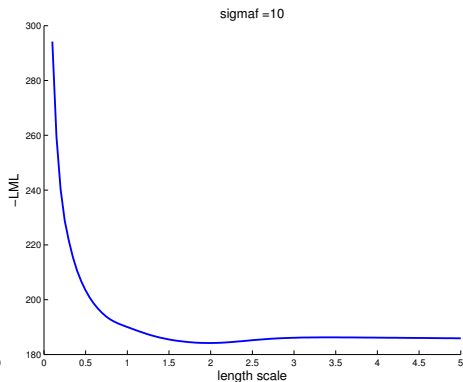
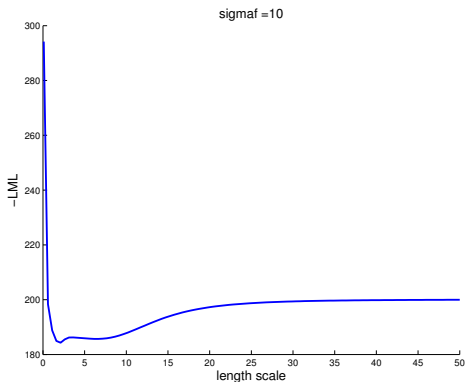
$$p(\mathbf{y} | \mathbf{X}, \theta) = \int p(\mathbf{y} | \mathbf{X}, \mathbf{f}, \theta) p(\mathbf{f} | \mathbf{X}, \theta) d\mathbf{f}$$

where $\mathbf{f} = f(\mathbf{X})$ is a vector with function values in the training data.

- ▶ For Gaussian process regression we can actually do this analytically [since $\mathbf{y} | \mathbf{X}, \theta \sim N(0, K + \sigma_n^2 I)$]

$$\log p(\mathbf{y} | \mathbf{X}, \theta) = -\frac{1}{2} \mathbf{y}^T (K + \sigma_n^2 I)^{-1} \mathbf{y} - \frac{1}{2} \log |K + \sigma_n^2 I| - \frac{n}{2} \log(2\pi)$$

CANADIAN WAGES - LML DETERMINATION OF ℓ



CLASSIFICATION

- ▶ **Classification: binary response** $y \in \{-1, 1\}$ (or multi-class $y \in \{1, 2, \dots, C\}$) explained/predicted by covariates/features \mathbf{x} .
- ▶ Aim: posterior distribution $p(y|\mathbf{x})$
- ▶ **Generative approach**

$$p(y|\mathbf{x}) \propto p(\mathbf{x}|y)p(y)$$

Need to model the class-conditional distributions $p(\mathbf{x}|y)$ for each y .

- ▶ **Discriminative approach** models the posterior $p(y|\mathbf{x})$ directly.
Example: linear logistic regression

$$\Pr(y = 1|\mathbf{x}) = \lambda(\mathbf{x}^T \mathbf{w})$$

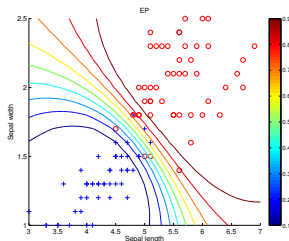
where $\lambda(z)$ is the logistic **link function**

$$\lambda(z) = \frac{1}{1 + \exp(-z)}$$

- ▶ $\lambda(z)$ 'squashes' the linear prediction $\mathbf{x}^T \mathbf{w} \in \mathbb{R}$ into $\lambda(\mathbf{x}^T \mathbf{w}) \in [0, 1]$.
- ▶ Can also use the normal CDF $\Phi(z)$ for squashing. **Probit regression**.

CLASSIFICATION

- ▶ The posterior of the label $p(y|\mathbf{x})$ is clearly a probabilistic classifier.
- ▶ Visual representation: level contours of $p(y|\mathbf{x})$ over \mathbf{x} -space.



- ▶ Decision boundaries.
- ▶ **Zero-one loss** \Rightarrow allocate to class c which maximizes $Pr(y = c|\mathbf{x})$.
- ▶ **Reject option.** Refuse to classify when $Pr(y = c|\mathbf{x})$ are small for all c .

GP CLASSIFICATION

- ▶ **Linear logistic regression**

$$Pr(y = 1|\mathbf{x}) = \lambda(\mathbf{x}^T \mathbf{w})$$

has linear decision boundaries (conditional on \mathbf{w}).

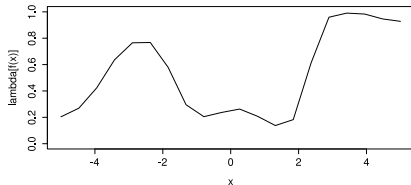
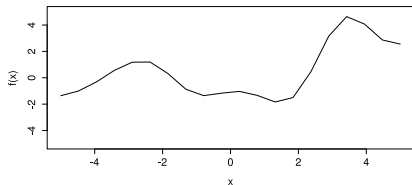
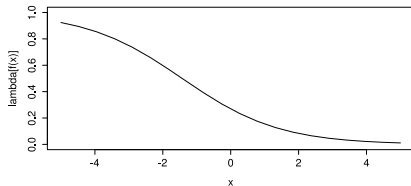
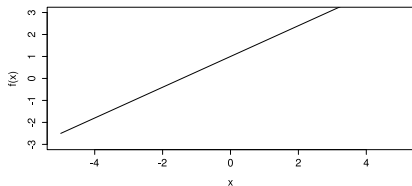
- ▶ Obvious **GP extension**: replace $\mathbf{x}^T \mathbf{w}$ by

$$f(\mathbf{x}) \sim GP(0, k(\mathbf{x}, \mathbf{x}'))$$

and squash f through logistic function (or normal CDF)

$$Pr(y = 1|\mathbf{x}) = \lambda(f(\mathbf{x}))$$

SQUASHING F



GP CLASSIFICATION - INFERENCE

- Prediction for a test case \mathbf{x}_* :

$$Pr(y_* = +1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int \sigma(f_*)p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y})df_*$$

where $\sigma(f_*)$ is some sigmoidal function (logistic, normal CDF...) and f_* is the latent f at the test input \mathbf{x}_* .

- The predictive distribution of f_* is

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

where

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})$$

is the posterior of \mathbf{f} from the training data.

- Note that $p(\mathbf{y}|\mathbf{f})$ is no longer Gaussian in classification problems. Posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ is not analytically tractable.

THE LAPLACE APPROXIMATION

- ▶ Approximates $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ with $N(\hat{\mathbf{f}}, \mathbf{A}^{-1})$, where $\hat{\mathbf{f}}$ is the posterior mode and \mathbf{A} is the negative Hessian of the log posterior at $\mathbf{f} = \hat{\mathbf{f}}$.
- ▶ The log posterior is (proportional to)

$$\begin{aligned}\Psi(\mathbf{f}) &= \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|\mathbf{X}) \\ &= \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \frac{1}{2}\log |K| - \frac{n}{2}\log 2\pi\end{aligned}$$

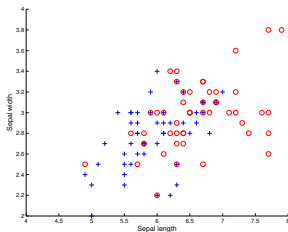
- ▶ Differentiating wrt \mathbf{f}

$$\begin{aligned}\nabla \Psi(\mathbf{f}) &= \nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1}\mathbf{f} \\ \nabla \nabla \Psi(\mathbf{f}) &= \nabla \nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1} = -W - K^{-1}\end{aligned}$$

where W is a diagonal matrix since each y_i only depends on its f_i .

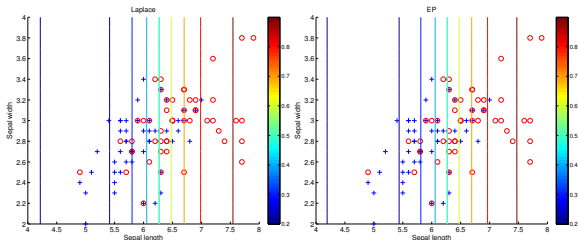
- ▶ Use **Newton's method** to iterate to the mode.
- ▶ The elements $\partial/\partial f_i \log p(y_i|f_i)$ and $\partial^2/\partial f_i^2 \log p(y_i|f_i)$ are given in the table on page 43 in RW for the logistic and probit cases.
- ▶ **Approximate predictions** of f_* are **possible**. Predictions of y_* require one-dimensional numerical integration.

IRIS DATA - SEPAL - SE KERNEL WITH ARD

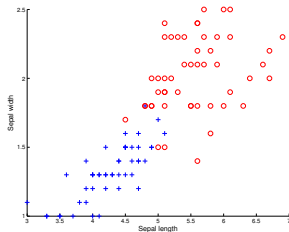


Laplace: $\hat{\ell}_1 = 1.7214, \hat{\ell}_2 = 185.5040, \sigma_f = 1.4361$

EP: $\hat{\ell}_1 = 1.7189, \hat{\ell}_2 = 55.5003, \sigma_f = 1.4343$

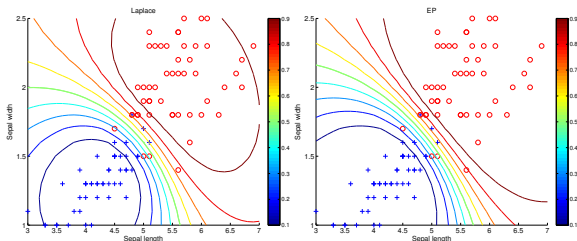


IRIS DATA - PETAL - SE KERNEL WITH ARD

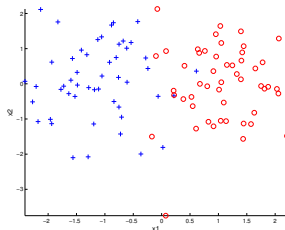


Laplace: $\hat{\ell}_1 = 1.7606, \hat{\ell}_2 = 0.8804, \sigma_f = 4.9129$

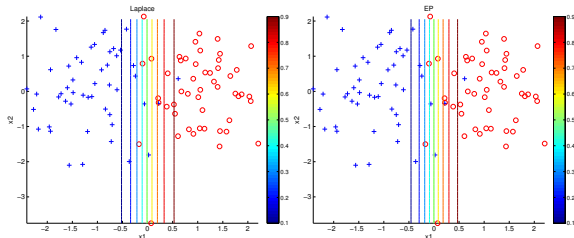
EP: $\hat{\ell}_1 = 2.1139, \hat{\ell}_2 = 1.0720, \sigma_f = 5.3369$



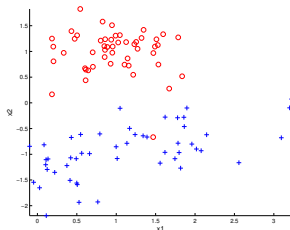
TOY DATA 1 - SE KERNEL WITH ARD



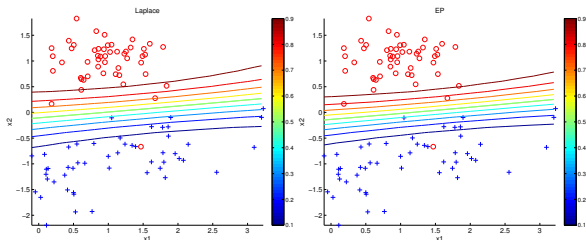
EP: $\hat{\ell}_1 = 2.4503, \hat{\ell}_2 = 721.7405, \sigma_f = 4.7540$



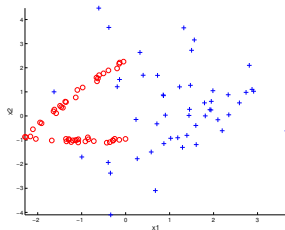
TOY DATA 2 - SE KERNEL WITH ARD



EP: $\hat{\ell}_1 = 8.3831, \hat{\ell}_2 = 1.9587, \sigma_f = 4.5483$



TOY DATA 3 - SE KERNEL WITH ARD



Laplace: $\hat{\ell}_1 = 0.7726, \hat{\ell}_2 = 0.6974, \sigma_f = 11.7854$

EP: $\hat{\ell}_1 = 1.2685, \hat{\ell}_2 = 1.0941, \sigma_f = 17.2774$

