05 Gaussian Process & Bayesian Models

Advanced Machine Learning

Malte Schilling, Neuroinformatics Group, Bielefeld University

Probabilities and Bayesian Reasoning

Gaussian (normal) distribution

Is characterized by mean μ and variance σ. The probability distribution is given as

$$p(X = x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

The multivariate Gaussian for D dimensions is given as

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}(det\ \Sigma)^{1/2}} exp\ (-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

For Visual Exploration of Covariance and GP

Bayes' rule

… tells us how to invert conditional probabilities:

$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$
$$\Rightarrow p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Here,

- $p(B)$ is the *a priory probability*, or the prior,
- $p(A|B)$ is the *likelihood of B for a fixed A*,
- and $p(B|A)$ is the *a posteriori probability* of B given A.

Gaussian Process – Parametric View

Bayesian Inference

Our goal is to establish inferences between inputs and targets. This is the conditional distribution of the targets given the input.

Our training set $\mathcal{D}$ consists of n observations:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$$

which we can collect in the design matrix.

(Rasmussen and Williams 2006)

A prior on parameters

In a parametric model $\mathcal{M}$, the model is defined by the structure and the parameters:

$$f_w(\mathbf{x}) = \sum_{m=0}^{M} w_m \phi_m(\mathbf{x})$$

We can define a prior $p(\mathbf{w}|\mathcal{M})$ for the parameters of the model – this determines the functions the model can generate.

- First, we are selecting a structure.
- Secondly, we are selecting a probability distribution for the parameters.

Bayesian Analysis of Linear Regression

We do regression on a function $t(\mathbf{x}) = \mathbf{x}^T \mathbf{w}$ with added Gaussian noise.

This leads to observation

$$y = f(\mathbf{x}) + \varepsilon, \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma_n^2)$$

We can calculate the likelihood of the data (due to i.i.d.):

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

A prior on the parameters is required and we use a zero mean Gaussian with covariance matrix Σ_p:

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

Inference in Bayesian linear model

We are looking for the posterior distribution over the weights which we get through Bayes' rule:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}},\ p(\mathbf{w}|\mathbf{y},\ X) = \frac{p(\mathbf{y}|X, w)p(\mathbf{w})}{p(\mathbf{y}|X)}$$

Parametric View

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

Reminder Gaussian probability distribution:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \mathcal{N}(\mu, \sigma^2)$$

Likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y_i - \mathbf{x}^T\mathbf{w})^2}{2\sigma_n^2}}$$

$$= \frac{1}{(2\pi\sigma_n^2)^{n/2}} e^{-\frac{1}{2\sigma_n^2}|\mathbf{y} - \mathbf{x}^T\mathbf{w}|^2} = \mathcal{N}(\mathbf{X}^T\mathbf{w}, \sigma_n^2\mathbf{I})$$

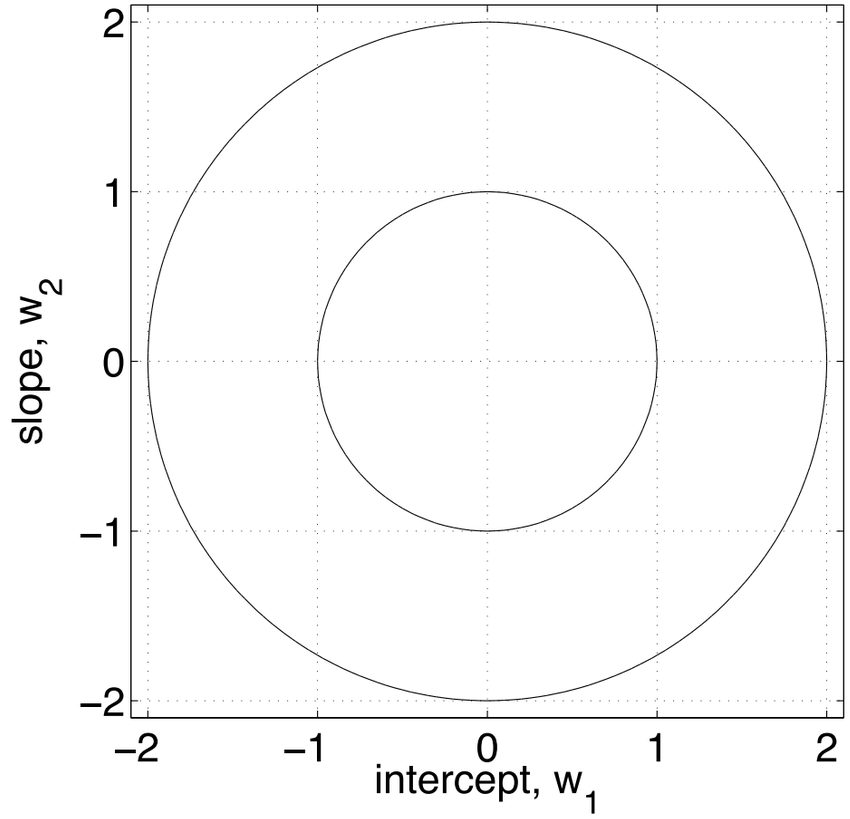
Setting the prior

Use a zero mean Gaussian as prior on parameters:

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

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}},$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$



Contours of the prior distribution (1 and 2 standard deviation equi-probability lines) for

Deriving the posterior

Importantly, the marginal likelihoodis independent of the weights and acts as a normalizing constant which does not affect the search for the best weights.

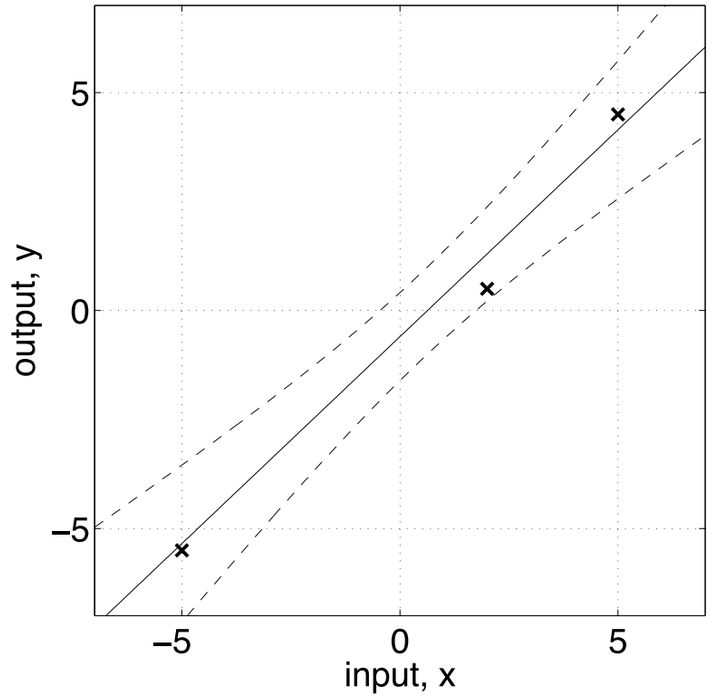
$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$

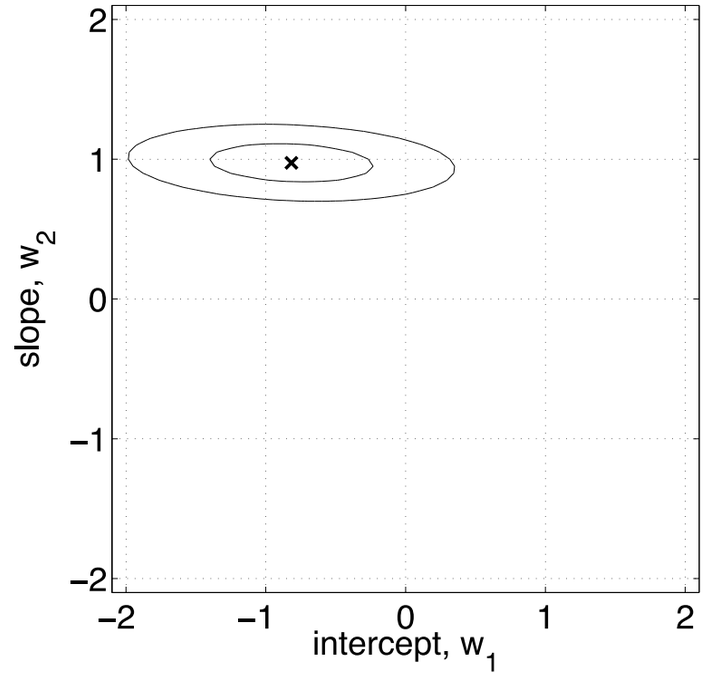
$$\begin{aligned} p(\mathbf{w}|\mathbf{y}, \mathbf{X}) &\propto e^{-\frac{1}{2\sigma_n^2}(\mathbf{y}-\mathbf{X}^T\mathbf{w})^T(\mathbf{y}-\mathbf{X}^T\mathbf{w})} e^{-\frac{1}{2}\mathbf{w}^T\Sigma_p^{-1}\mathbf{w}} \\ &\propto e^{-\frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}})^T(\frac{1}{\sigma_n^2}\mathbf{X}\mathbf{X}^T+\Sigma_p^{-1})(\mathbf{w}-\bar{\mathbf{w}})}, \bar{\mathbf{w}} = \sigma_n^{-2}(\sigma_n^{-2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1})^{-1}\mathbf{X}\mathbf{y} \end{aligned}$$

The form of the posterior distribution is again Gaussian (recognize the form) with mean $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{A}^{-1}$:

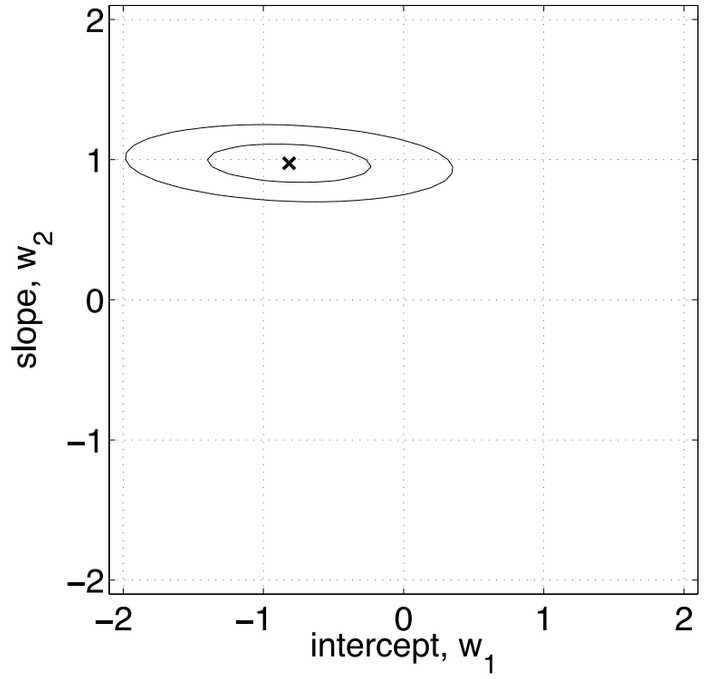
$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \sim \mathcal{N}(\bar{\mathbf{w}} = \frac{1}{\sigma_n^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{y}, \mathbf{A}^{-1}),\ \mathbf{A} = \sigma_n^{-2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1}$$

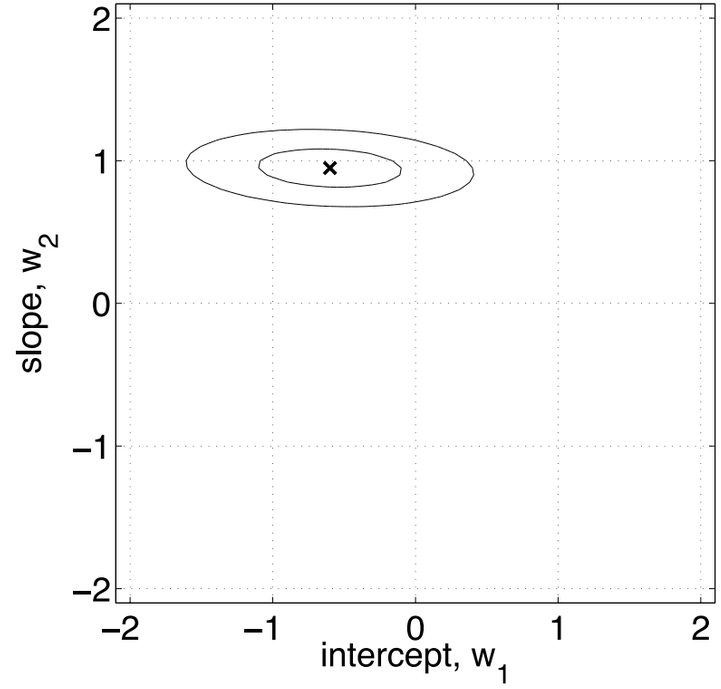
Example of Bayesian linear model: Condition on data





Example of Bayesian linear model: Condition on data





Predictive Distribution

We are not choosing (as we would in non-Bayesian schemes, MAP) a specific weight. Instead, we work with the distribution over parameters which is a distribution over functions.

For prediction, we average over all possible parameters. This gives us a predictive distribution f_* for a test case $\mathbf{x}_*$

$$
\begin{aligned}
p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int p(f_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{y}) d\mathbf{w} \\
&= \mathcal{N}(\frac{1}{\sigma_n^2} \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{X}\mathbf{y}, \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{x}_*).
\end{aligned}
$$

This predictive distribution is again Gaussian.

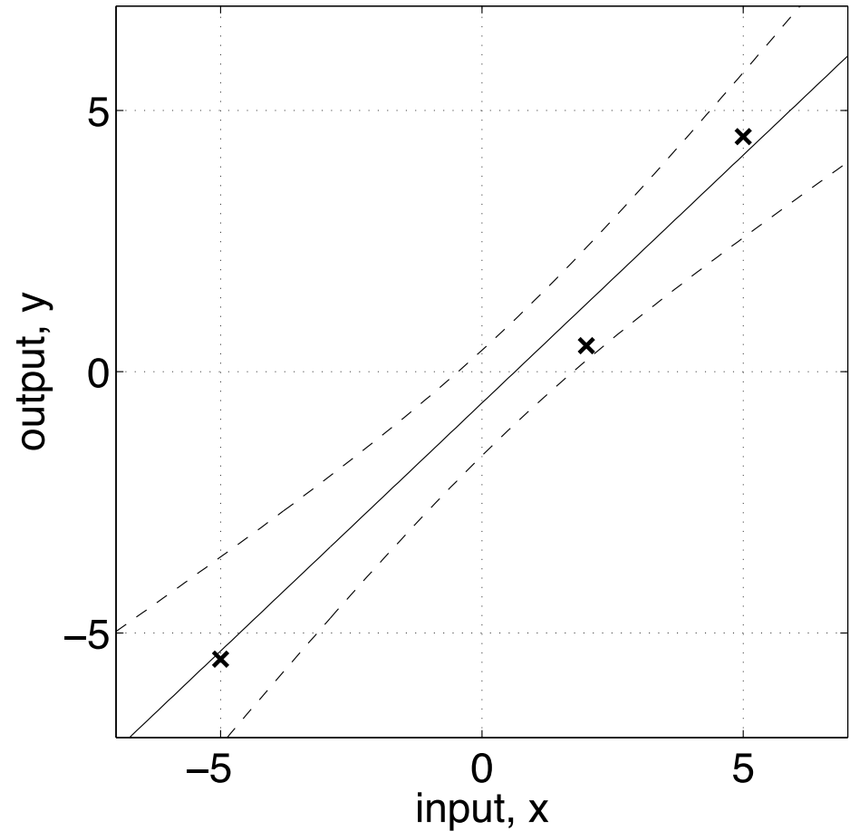
Example of Bayesian linear model: Predcition

Superimposed on the data is the predictive mean plus contours for two standard deviations of the (noise-free) predictive distribution

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}).$$

which is a Gaussian probability distribution for every x_* (see last slide):

$$\mathcal{N}(\frac{1}{\sigma_n^2}\mathbf{x}_*^T\mathbf{A}^{-1}\ \mathbf{X}\mathbf{y}, \mathbf{x}_*^T\mathbf{A}^{-1}\mathbf{x}_*).$$



Three training data points.

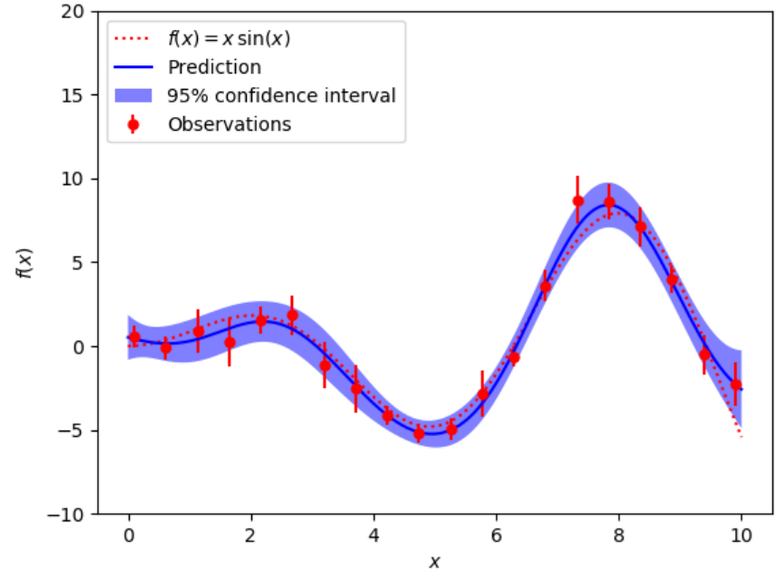
Gaussian Processes – Bayesian Inference

Prior

Posterior

Gaussian Processes Overview

- aware of uncertainty of the fitted GP that increases away from the training data,
- let you incorporate expert knowledge,
- are non-parametric,
- need to take into account the whole training data for prediction.



Three random function drawn from the posterior that includes example points.

Further reading: (Rasmussen and Williams 2006).

Two Bayesian Perspectives on Functions

Create Gaussian Distribution for each variable – distribute these through your space. Informally such an infinite long vector constitutes a function.

Prior

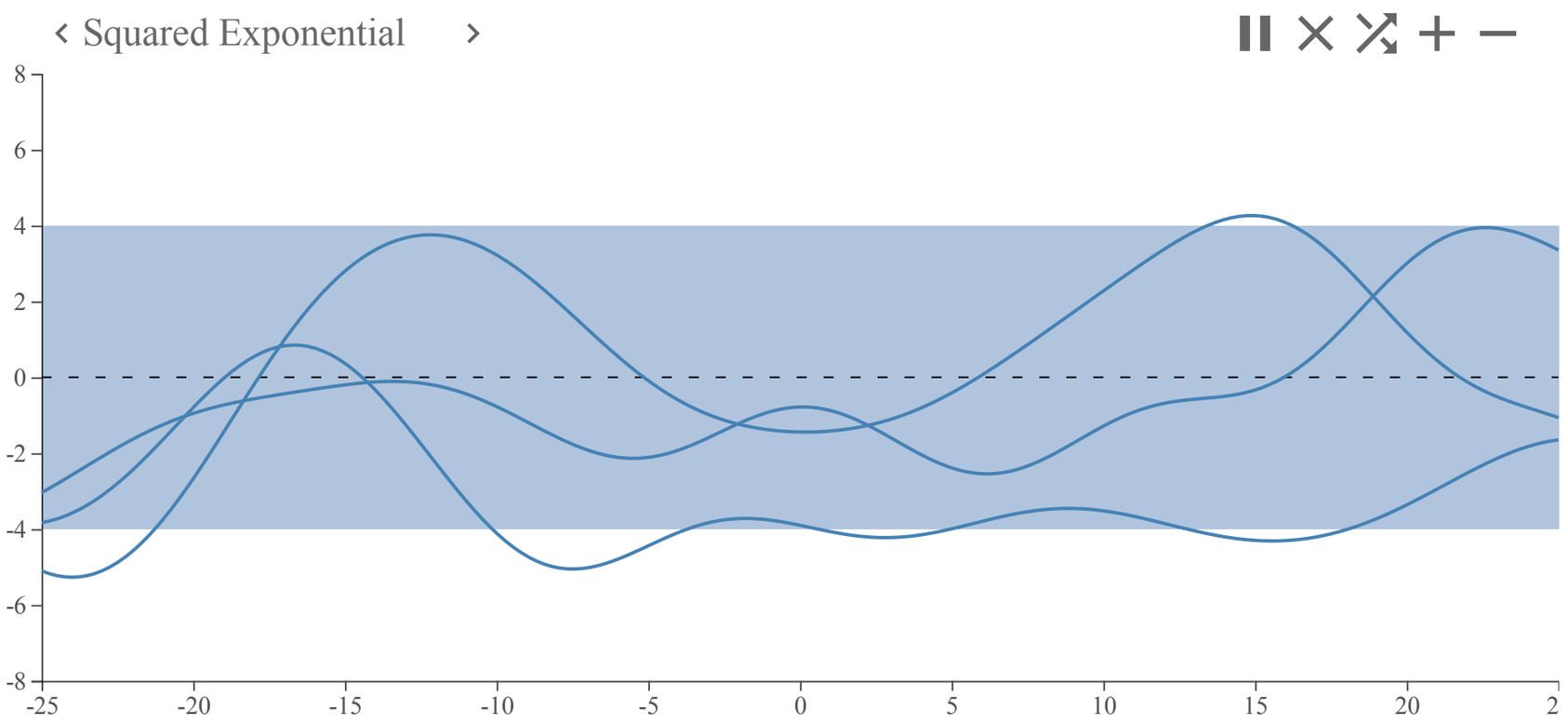
Posterior

UPPSALA
UNIVERSITET

Visualization of the Gaussian Process

by Johan Wågberg 2019. Try different kernels, change hyperparameters and add observations by clicking in the figure!

The technical idea on how to smoothly loop over Gaussian process samples (as done in this animation) are described by this document.



$$\kappa(t,s) = \sigma^2 e^{-\frac{(t-s)^2}{2\ell^2}}$$

$\sigma = 2$

$$y \mid f(x) \sim \mathcal{N}(f(x), \sigma^2)$$

$\sigma^2 = 0.1$

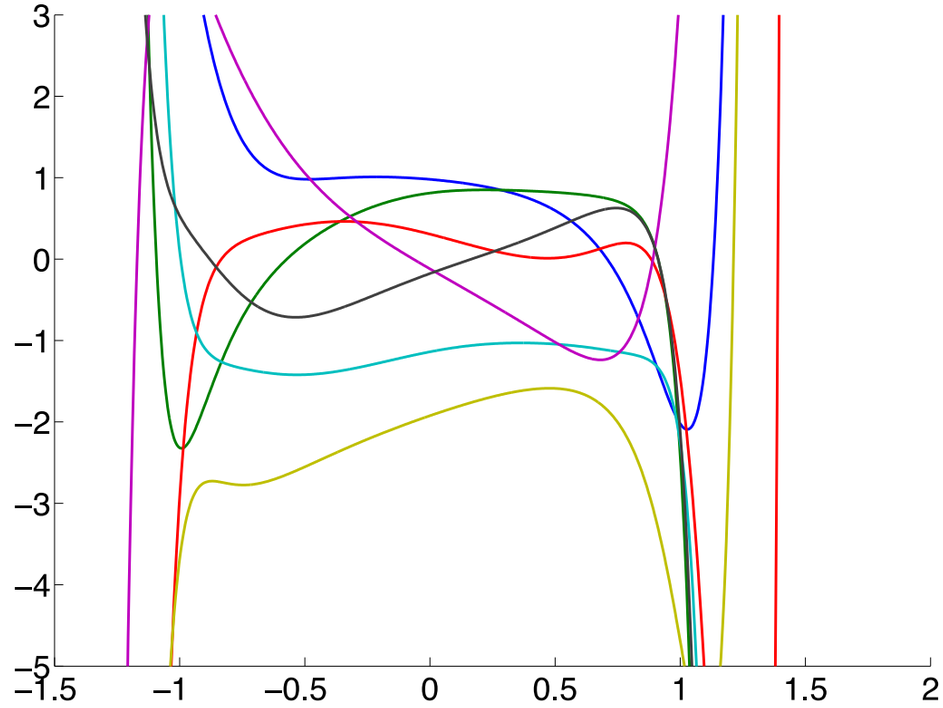
Gaussian Process – Distribution over Parameters

Example: A prior distribution over functions

As an example,

- we choose a polynomical model with $M = 17$: $\phi_m(\mathbf{x}) = \mathbf{x}^m$
- as a prior for the parameter distribution we choose a normal distribution:

$$p(w_m) = \mathcal{N}(w_m|\mu, \sigma_w^2)$$



Shown is one example for which we sampled all the parameters from the normal distribution.

Distribution over functions

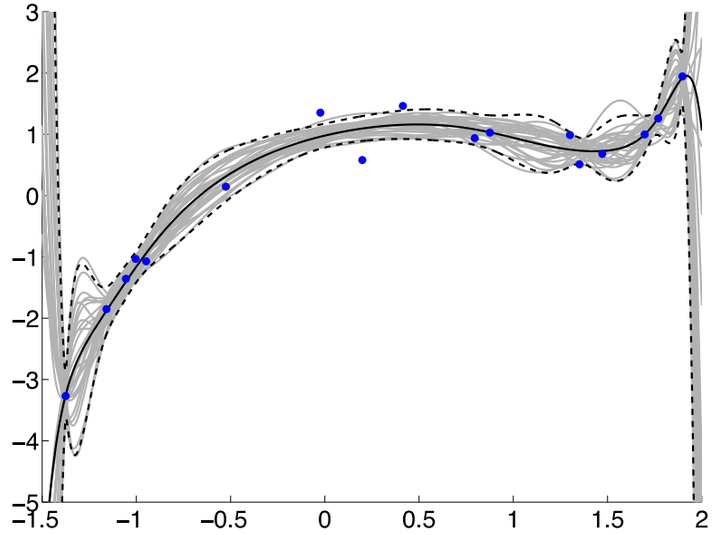
- We have seen now an algorithm for building a model through selecting the model type and sample parameters.

- But we are interested in predictions of the model and not the parameters as such.
- Secondly, we want to work directly in the space of functions. This becomes possible as a distribution over parameters induces a distribution over functions $p(\mathbf{f}|\mathcal{M})$.
- This would be simpler and allow for more efficient inference.

Posterior probabilities for a function

Our goal is to use our functions $\mathbf{f}$ to make predictions for novel inputs. But until now, we have only looked at the prior for these functions $p(\mathbf{f}|\mathcal{M})$.
We are interested in the posterior distribution of the function – that is which is conditioned on our evidence:

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}$$



Sample from the posterior (Rasmussen 2016)

Drawback of polynomials as priors for functions

Drawback of sampling over parameters

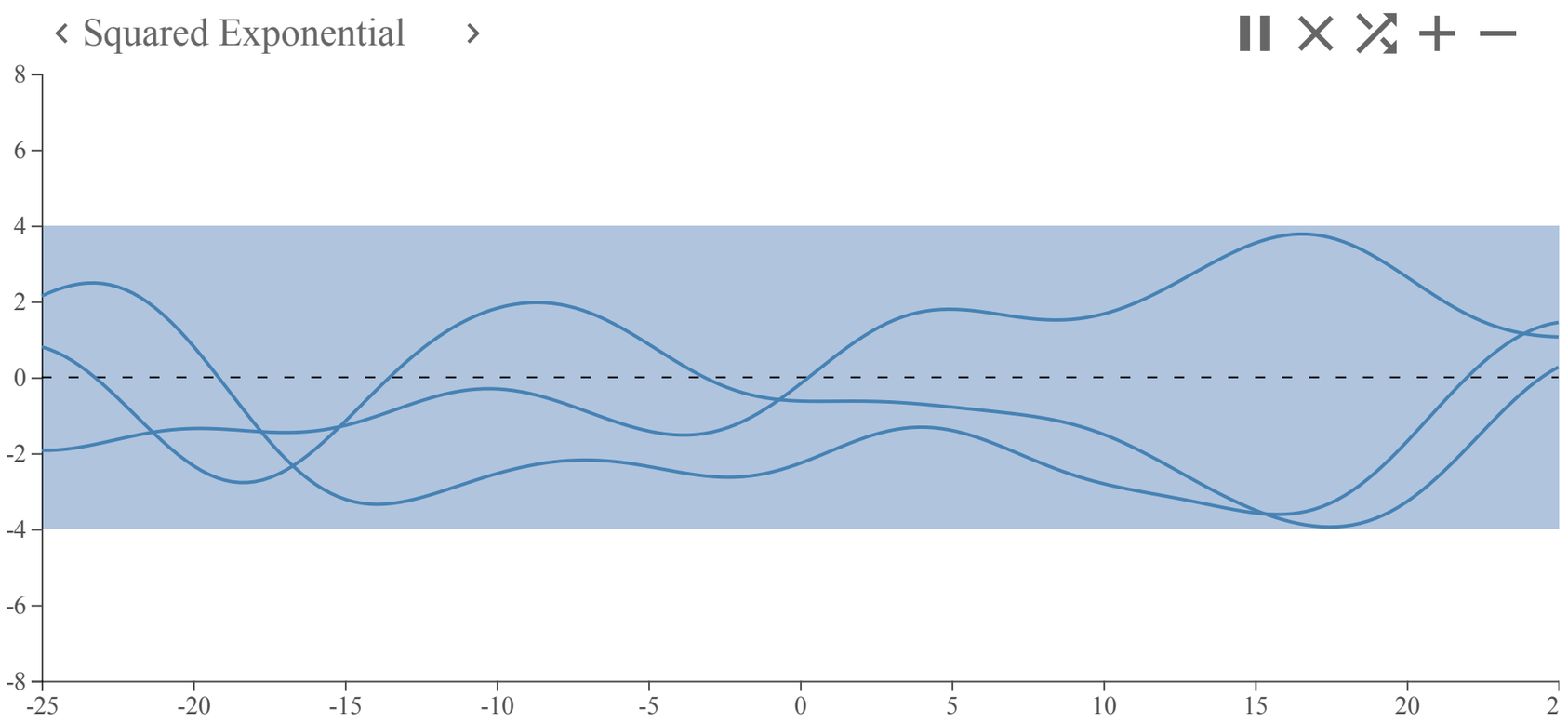
- Distributions over parameters induce distribution over functions.
- But sampling over parameter space and using priors over functions might not lead to good results (see example for polynomials).
- Therefore, we want to work directly on priors and probability distributions over functions.
- This leads to the question of how probability distribution over functions look like and how they could be specified.

UPPSALA
UNIVERSITET

Visualization of the Gaussian Process

by Johan Wågberg 2019. Try different kernels, change hyperparameters and add observations by clicking in the figure!

The technical idea on how to smoothly loop over Gaussian process samples (as done in this animation) are described by this document.



$$\kappa(t,s) = \sigma^2 e^{-\frac{(t-s)^2}{2\ell^2}}$$

σ =2

$$y \mid f(x) \sim \mathcal{N}(f(x), \sigma^2)$$

σ^2 =0.1

Gaussian Processes

Prior

Posterior

References

Rasmussen, Carl Edward. 2016. “Probabilistic Machine Learning.” Lecture Notes, University of Cambridge.

Rasmussen, CE., and CKI. Williams. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: Biologische Kybernetik; Max-Planck-Gesellschaft; MIT Press.