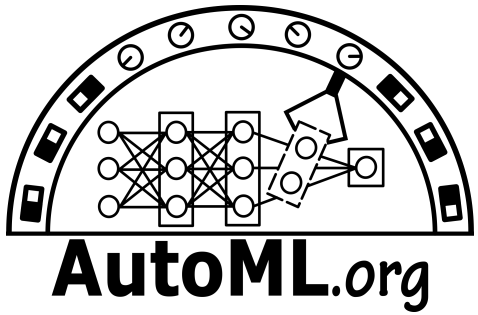
Automated Machine Learning
(AutoML)

M. Lindauer F. Hutter

University of Freiburg



Lecture 6:
Bayesian linear regression and Gaussian processes

Where are we? The big picture

- Introduction
- Background
  - Design spaces in ML
  - Evaluation and visualization

→ Hyperparameter optimization (HPO)

  → Bayesian optimization
  - Other black-box techniques
  - More details on Gaussian processes

- Pentecost (Holiday) – no lecture
- Architecture search I + II
- Meta-Learning
- Learning to learn & optimize
- Beyond AutoML: algorithm configuration and control
- Project announcement and closing

Today's Learning Goals

After today's lecture, you can ...

- Derive Bayesian linear regression
- Derive Gaussian processes
- Explain the relationship between Bayesian linear regression and Gaussian processes

Resources

Today's material is based on:

- Chapter 2 of
  Gaussian Processes for Machine Learning, Carl Edward Rasmussen and Christopher K. I. Williams[1]
- Another excellent resource: Philipp Hennig's Gaussian process tutorials[2]

[1] `http://www.gaussianprocess.org/gpml/`
[2] `http://ei.is.tuebingen.mpg.de/person/phennig`

Lecture Overview

1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

The univariate form

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

Named after Carl Friedrich Gauss (1777-1855)

Figure: Wikipedia, public domain

The univariate form

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

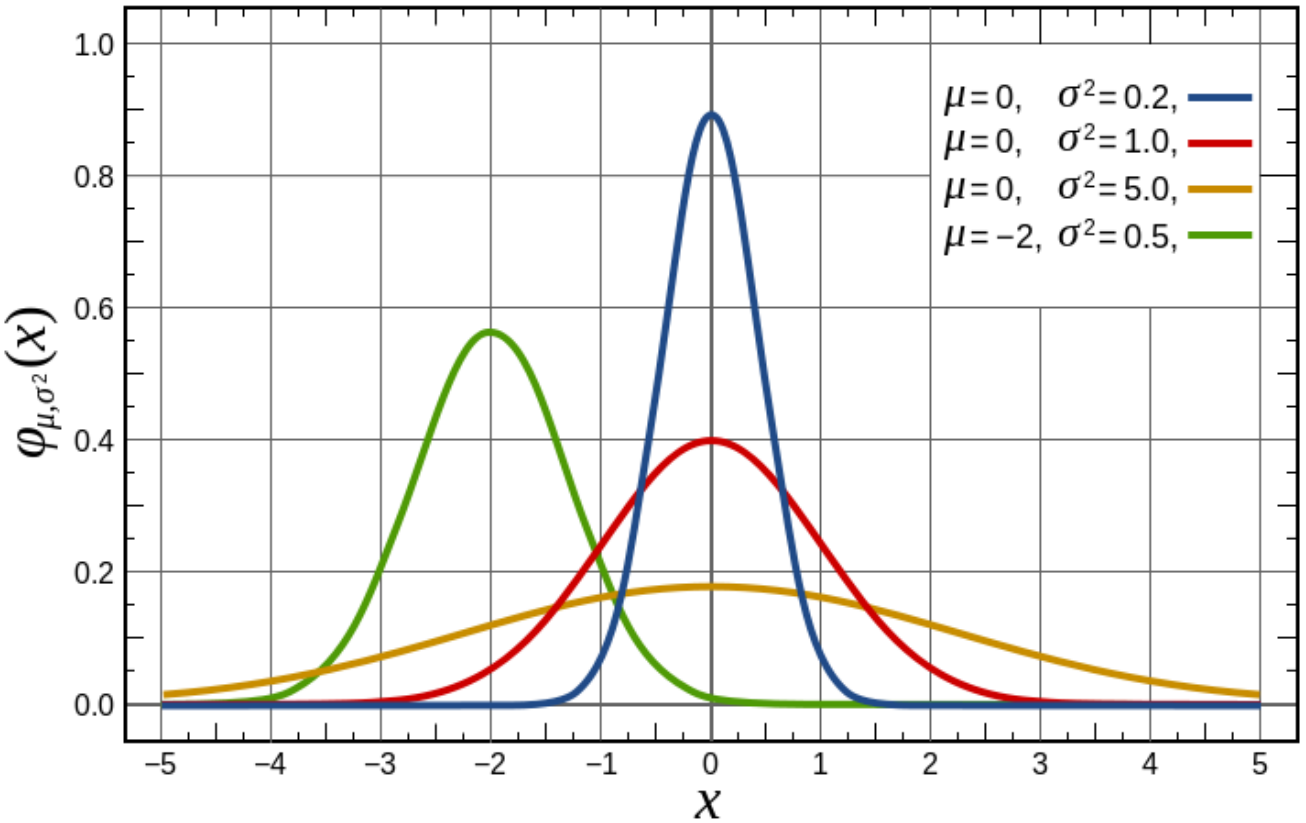


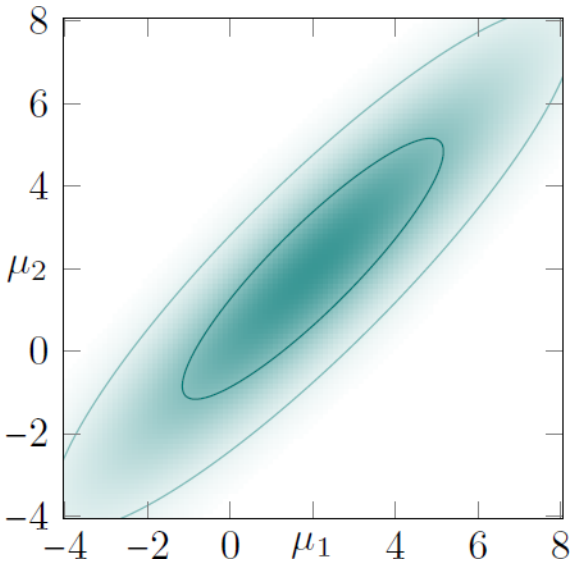
Figure: Wikipedia, public domain

The multivariate Gaussian (in N dimensions)

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are $N \times 1$ (column) vectors
- $\boldsymbol{\Sigma}$ is an $N \times N$ matrix

An example in 2 dimensions:



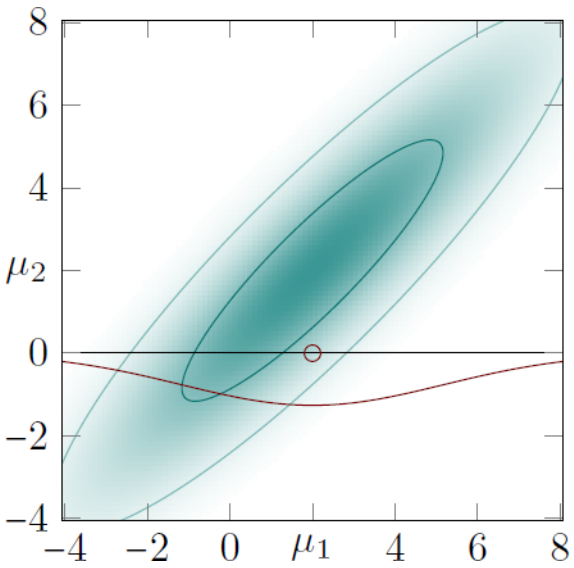
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}$$

Property 1: Closure under marginalization

Gaussian distributions are closed under marginalization

Let $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be jointly Gaussian distributed:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^{\mathsf{T}} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$ Then $x \sim \mathcal{N}(\mu_1, \boldsymbol{\Sigma}_{11})$.



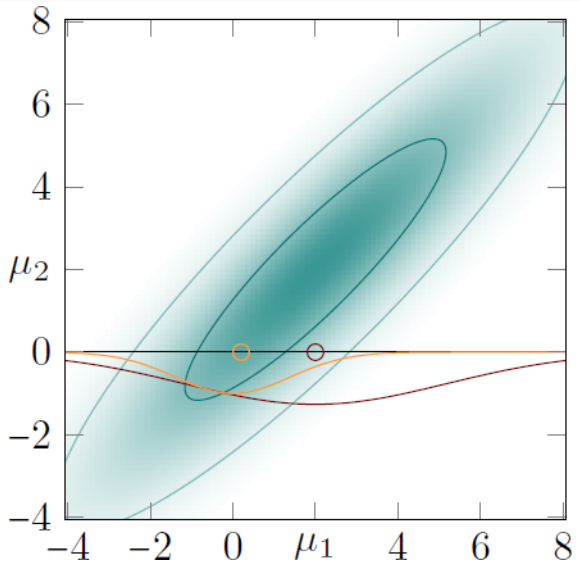
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{pmatrix}$$

Property 2: Closure under conditioning

Gaussian distributions are closed under conditioning

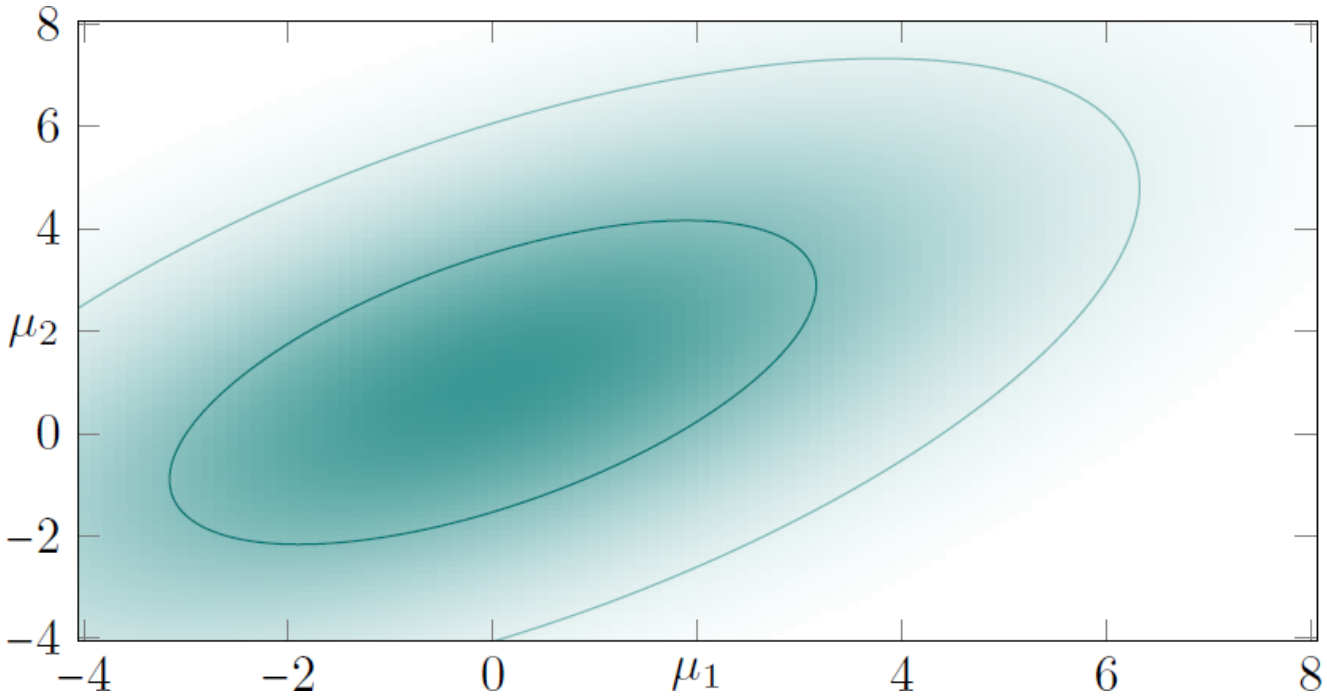
Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be jointly Gaussian distributed: $\begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \begin{bmatrix} A & C \\ C^\intercal & B \end{bmatrix} \right)$

Then $\boldsymbol{b}|\boldsymbol{a} \sim \mathcal{N}\left(\boldsymbol{\mu}_b + C^\intercal A^{-1}(\boldsymbol{a} - \mu_a), B - C^\intercal A^{-1} C\right)$.



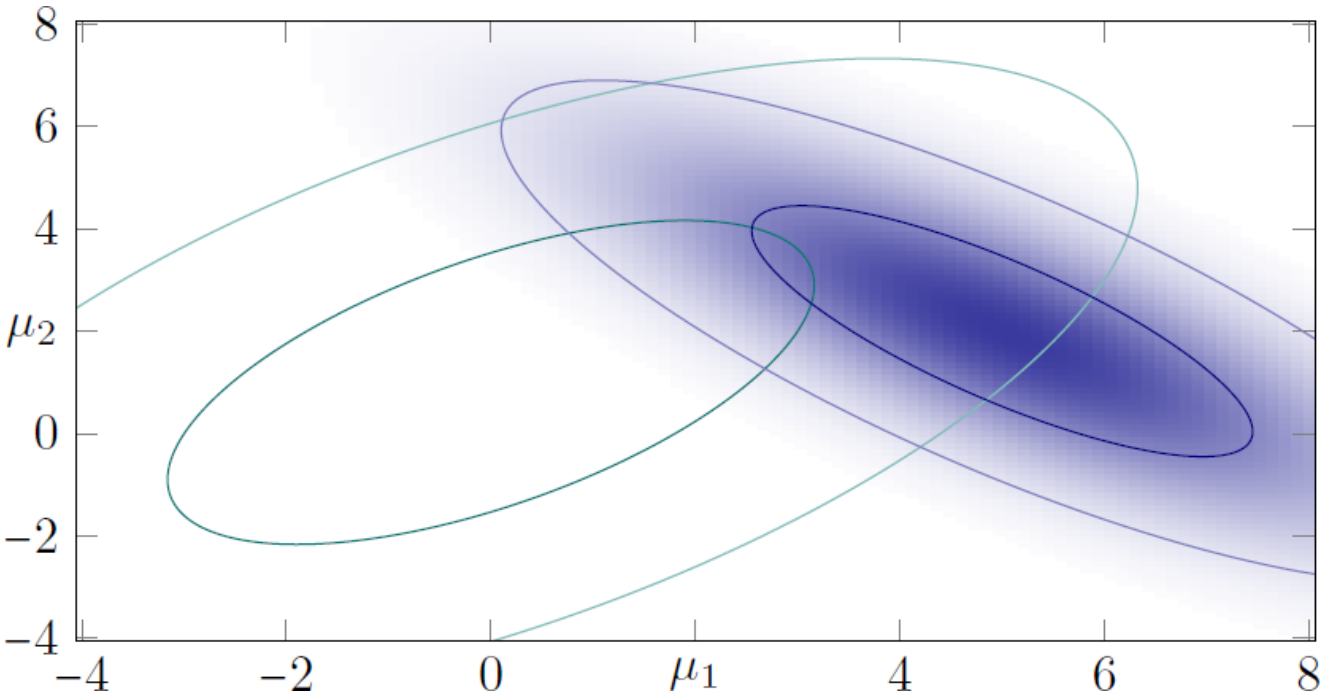
Property 3: Closure under multiplication

$$\mathcal{N}(x; a, A)$$



Property 3: Closure under multiplication

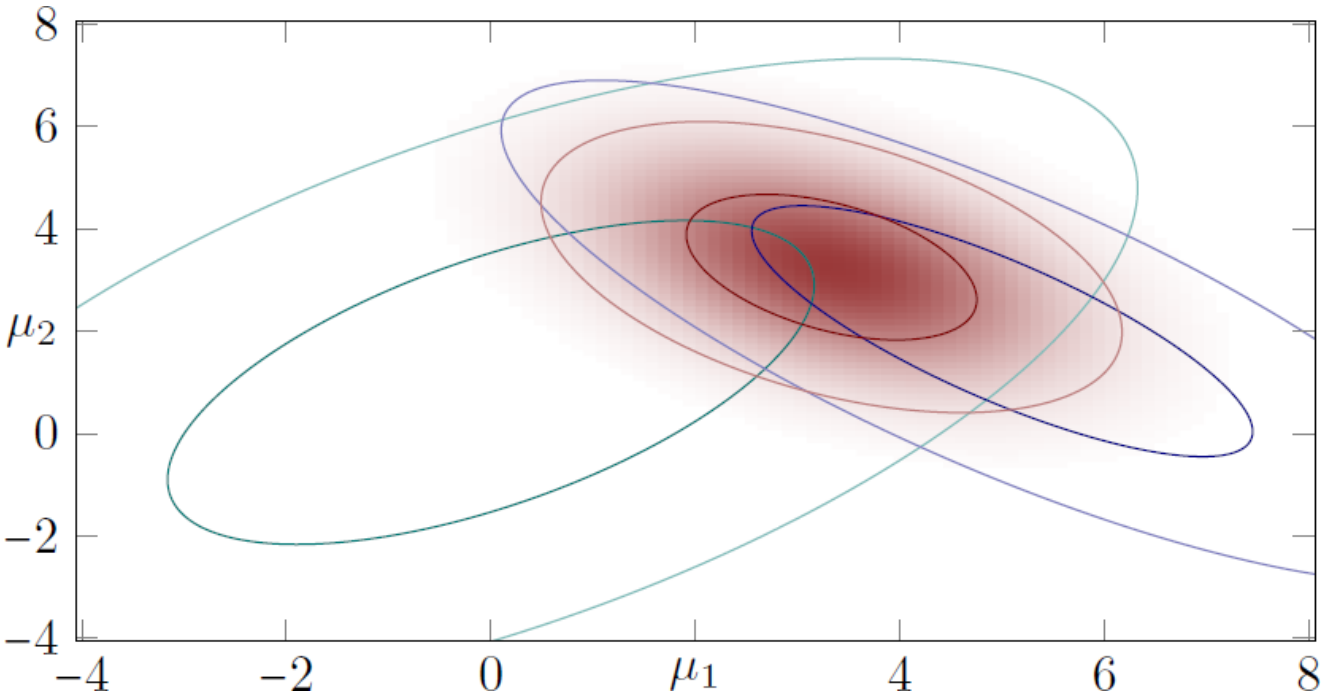
$$\mathcal{N}(x; b, B)$$



Property 3: Closure under multiplication

$$\mathcal{N}(x; a, A)\mathcal{N}(x; b, B) \propto \mathcal{N}(x; c, C)$$

$$C := (A^{-1} + B^{-1})^{-1} \quad c := C(A^{-1}a + B^{-1}b)$$



Why Gaussians?

- The central limit theorem tells us that means of large populations are distributed according to a Gaussian
- However, individual measurements are very rarely Gaussian-distributed
- Rather, Gaussians are mostly a mathematical convenience that lets us solve integrals in closed form!

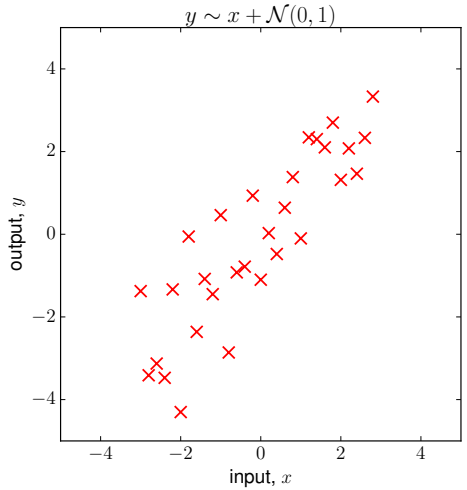
Lecture Overview

1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

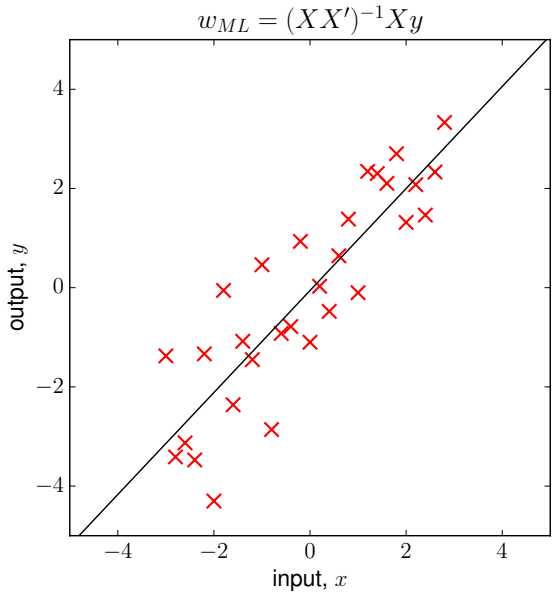
3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

Regression



- We have a dataset of n i.i.d. observations, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i | i = 1, \ldots, n\}$
- Each $\boldsymbol{x}_i$ denotes an input (column) vector of dimension D and each y_i is a scalar output or target
- We collect all inputs $\boldsymbol{x}_i$ into the $D \times n$ design matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n]$ and the targets into the $n \times 1$ vector $\boldsymbol{y}$.
- Note: these slides are using the notation of Rasmussen's and Williams' book; $\boldsymbol{X}$ is transposed from our usual use

Standard Linear Regression Model



$$f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{w}, \qquad y = f(\boldsymbol{x}) + \epsilon \tag{2.1}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{2.2}$$

The predictive distribution with fixed $\boldsymbol{w}$

$$
\begin{aligned}
p(\boldsymbol{y}|X, \boldsymbol{w}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) \\
&= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi}\sigma_n)} \exp\left(-\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{w}_i)^2}{2\sigma_n^2}\right) \\
&= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}|\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{w}|^2\right) \\
&= \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}^\top \boldsymbol{w}, \sigma_n^2 I)
\end{aligned} \tag{2.3}
$$

How does each of these steps follow?

The Bayesian way of setting $\boldsymbol{w}$

- Define a prior distribution over $\boldsymbol{w}$ quantifying our uncertainty
- Integrate the likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ of the training data across all possible values of $\boldsymbol{w}$, weighted by the prior $p(\boldsymbol{w})$

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w} \tag{2.6}$$

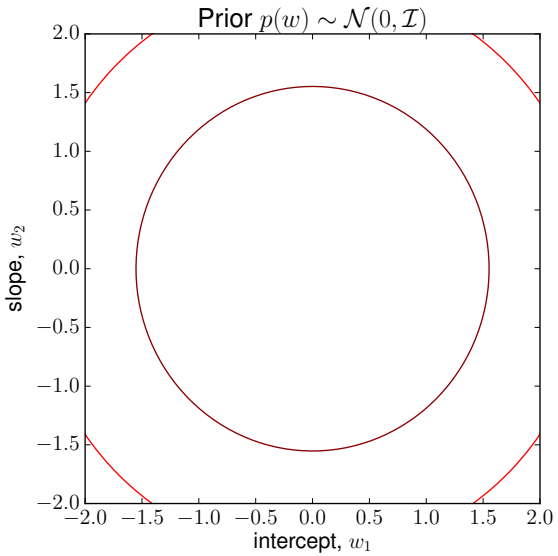
- We don't commit to a single $\boldsymbol{w}$, but integrate over an infinite number of possibilities

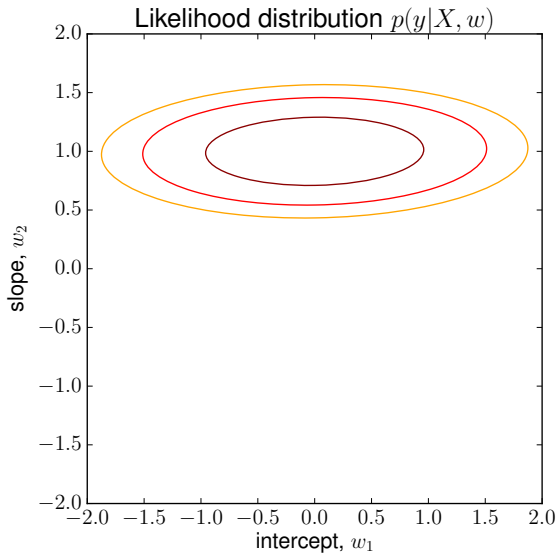
The Bayesian way of setting $\boldsymbol{w}$

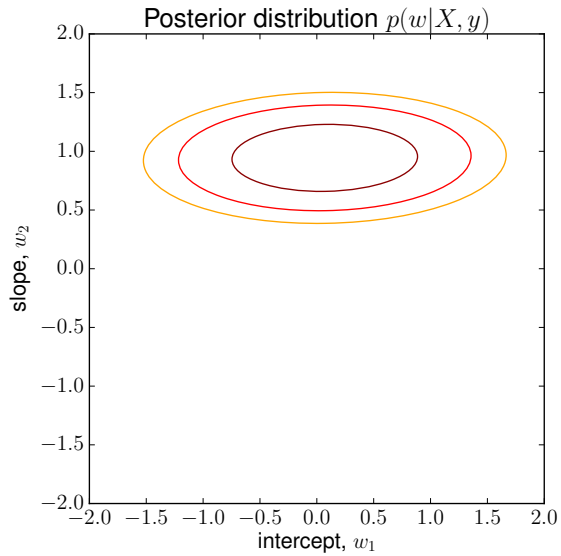
We combine the prior and the data likelihood using Bayes' rule:

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

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{X})} \tag{2.5}$$







Posterior of $\boldsymbol{w}$ under a Gaussian prior on $\boldsymbol{w}$

In general, the posterior $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$ has no closed-form solution, but it simplifies if we choose a prior $p(\boldsymbol{w})$ that is conjugate to the likelihood; in this case another Gaussian:

$$\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_p) \tag{2.4}$$

In that case, the posterior is Gaussian again:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \sim \mathcal{N}\left(\bar{\boldsymbol{w}} = \frac{1}{\sigma_n^2} A^{-1} \boldsymbol{X}\boldsymbol{y}, A^{-1}\right), \tag{2.7}$$

with $A = \frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^\intercal + \Sigma_p^{-1}$.

We'll do the core of the derivation for this on the next two slides.

Key part of the derivation: completing the square (1)

- The data likelihood is $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}^\top \boldsymbol{w}, \sigma_n^2 \boldsymbol{I})$.
- Written as a function of $\boldsymbol{w}$, this turns out to be proportional to a Gaussian $\mathcal{L}(w; \boldsymbol{y}, \boldsymbol{X}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \propto \mathcal{N}(\boldsymbol{w}; \boldsymbol{w'}, \boldsymbol{\Sigma}^{-1})$
- We show this by matching coefficients of $\boldsymbol{w}$ and $\boldsymbol{w}^\top \boldsymbol{w}$.

Key part of the derivation: completing the square (1)

- The data likelihood is $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}^\mathsf{T}\boldsymbol{w}, \sigma_n^2 \boldsymbol{I})$.
- Written as a function of $\boldsymbol{w}$, this turns out to be proportional to a Gaussian $\mathcal{L}(w; \boldsymbol{y}, \boldsymbol{X}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \propto \mathcal{N}(\boldsymbol{w}; \boldsymbol{w'}, \boldsymbol{\Sigma}^{-1})$
- We show this by matching coefficients of $\boldsymbol{w}$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{w}$.

$$
\begin{aligned}
\mathcal{L}(w; \boldsymbol{y}, \boldsymbol{X}) &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})\right) \\
&= \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{X}\boldsymbol{y} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{w})\right)
\end{aligned}
$$

Key part of the derivation: completing the square (1)

- The data likelihood is $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{X}^\mathsf{T}\boldsymbol{w}, \sigma_n^2 \boldsymbol{I})$.
- Written as a function of $\boldsymbol{w}$, this turns out to be proportional to a Gaussian $\mathcal{L}(w; \boldsymbol{y}, \boldsymbol{X}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) \propto \mathcal{N}(\boldsymbol{w}; \boldsymbol{w'}, \boldsymbol{\Sigma}^{-1})$
- We show this by matching coefficients of $\boldsymbol{w}$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{w}$.

$$
\begin{aligned}
\mathcal{L}(w; \boldsymbol{y}, \boldsymbol{X}) &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}^\mathsf{T}\boldsymbol{w})\right) \\
&= \exp\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{X}\boldsymbol{y} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{w})\right) \\
p(\boldsymbol{w}|\boldsymbol{w'}, \boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w'})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{w} - \boldsymbol{w'})\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w'} + \boldsymbol{w'}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w'})\right)
\end{aligned}
$$

Key part of the derivation: completing the square (2)

Let's match the coefficients of $\boldsymbol{w}$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{w}$...

$$\mathcal{L}(\boldsymbol{w}; \boldsymbol{y}, \boldsymbol{X}) \propto \exp\left(-\tfrac{1}{2\sigma_n^2}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{X}\boldsymbol{y} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{w}\right)\right)$$

$$p(\boldsymbol{w}|\boldsymbol{w'}, \boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w'} + \boldsymbol{w'}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w'})\right)$$

Key part of the derivation: completing the square (2)

Let's match the coefficients of $\boldsymbol{w}$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{w}$...

$$\mathcal{L}(\boldsymbol{w};\boldsymbol{y},\boldsymbol{X}) \propto \exp\left(-\tfrac{1}{2\sigma_n^2}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{X}\boldsymbol{y} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{w}\right)\right)$$

$$p(\boldsymbol{w}|\boldsymbol{w'},\boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w'} + \boldsymbol{w'}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w'})\right)$$

This directly yields:

$$\begin{aligned}\boldsymbol{w'} &= \frac{1}{\sigma_n^2}\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{y}\\ \boldsymbol{\Sigma} &= \left(\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}\end{aligned}$$

Key part of the derivation: completing the square (2)

Let's match the coefficients of $\boldsymbol{w}$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{w}$...

$$\mathcal{L}(\boldsymbol{w};\boldsymbol{y},\boldsymbol{X}) \propto \exp\left(-\tfrac{1}{2\sigma_n^2}\left(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{X}\boldsymbol{y} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{w})^\mathsf{T}\boldsymbol{X}^\mathsf{T}\boldsymbol{w}\right)\right)$$

$$p(\boldsymbol{w}|\boldsymbol{w}',\boldsymbol{\Sigma}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}' + \boldsymbol{w}'^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{w}')\right)$$

This directly yields:

$$\begin{aligned}\boldsymbol{w}' &= \frac{1}{\sigma_n^2}\boldsymbol{\Sigma}\boldsymbol{X}\boldsymbol{y}\\ \boldsymbol{\Sigma} &= \left(\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}\end{aligned}$$

Thus: $\mathcal{L}(\boldsymbol{w};\boldsymbol{y},\boldsymbol{X}) \propto \mathcal{N}\left(\frac{1}{\sigma_n^2}(\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\boldsymbol{X}\boldsymbol{y}, (\frac{1}{\sigma_n^2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T})^{-1}\right)$.
Then, simply multiply two Gaussians to get the posterior in (2.7)

The posterior predictive distribution

Recall: in linear regression (so far without basis functions), each weight vector $\boldsymbol{w}$ corresponds to a linear model:

$$f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{w} \tag{2.1}$$

We just derived the posterior $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$ over the weights $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \sim \mathcal{N}\left(\bar{\boldsymbol{w}} = \frac{1}{\sigma_n^2} A^{-1}\boldsymbol{X}\boldsymbol{y}, A^{-1}\right). \tag{2.7}$$

The posterior predictive distribution

Recall: in linear regression (so far without basis functions), each weight vector $\boldsymbol{w}$ corresponds to a linear model:

$$f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{w} \tag{2.1}$$

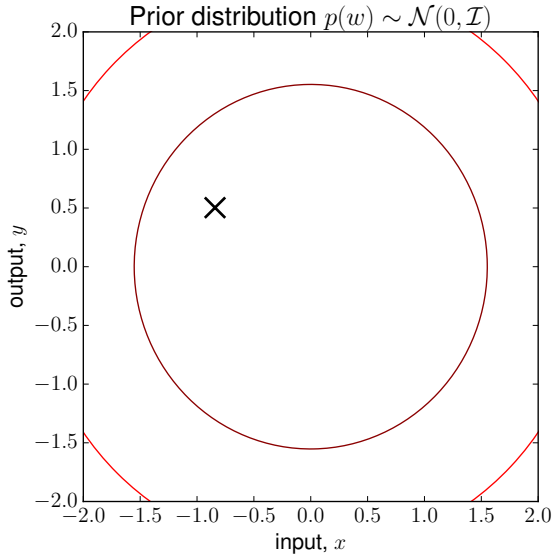
We just derived the posterior $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y})$ over the weights $\boldsymbol{w}$:

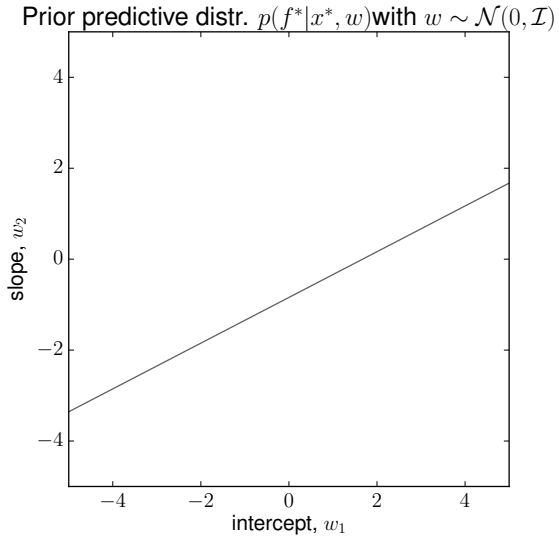
$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \sim \mathcal{N}\left(\bar{\boldsymbol{w}} = \frac{1}{\sigma_n^2} A^{-1}\boldsymbol{X}\boldsymbol{y}, A^{-1}\right). \tag{2.7}$$

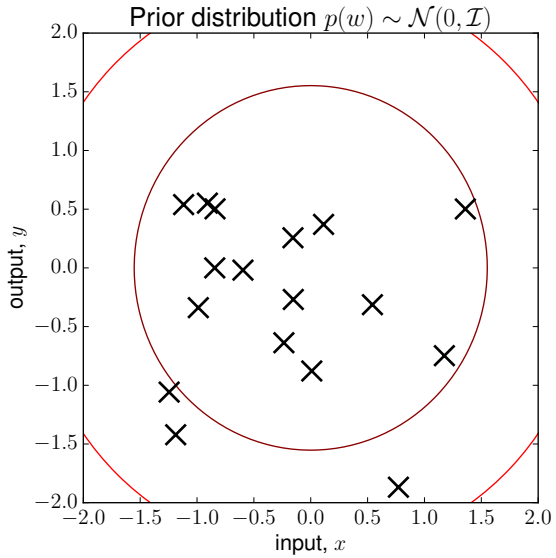
Gaussians are closed under linear maps. Thus, the posterior predictive distribution at new input $\boldsymbol{x}_\star$ given training data $\boldsymbol{X}$, $\boldsymbol{y}$ is:

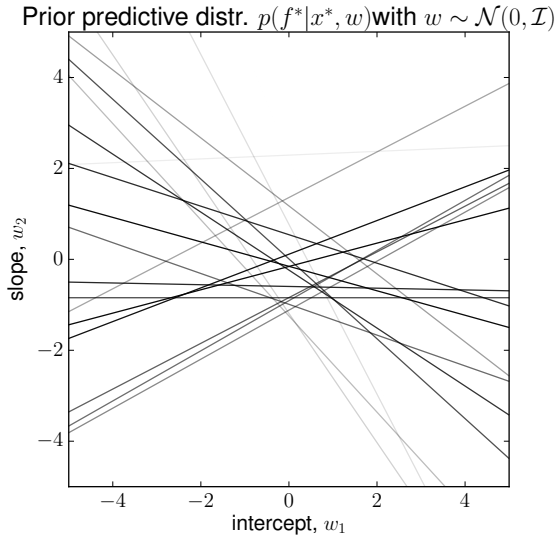
$$\begin{aligned} p(f_\star|\boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y}) &= p(\boldsymbol{x}_\star^\mathsf{T}\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \\ &= \mathcal{N}(f_\star; \frac{1}{\sigma_n^2}\boldsymbol{x}_\star^\mathsf{T} A^{-1}\boldsymbol{X}\boldsymbol{y}, \boldsymbol{x}_\star^\mathsf{T} A^{-1}\boldsymbol{x}_\star) \end{aligned} \tag{2.9}$$

Weights correspond to functions (1)

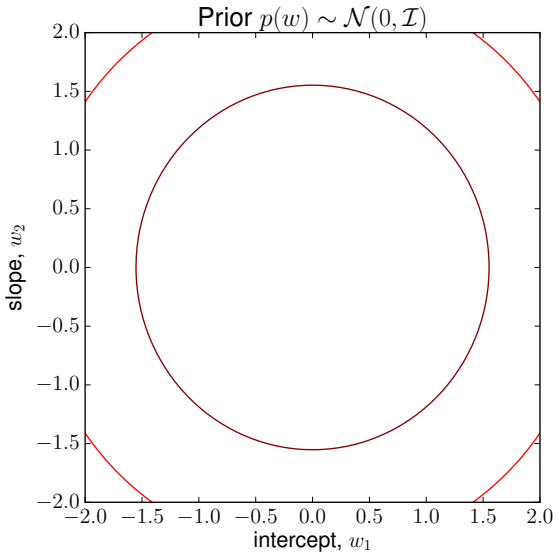


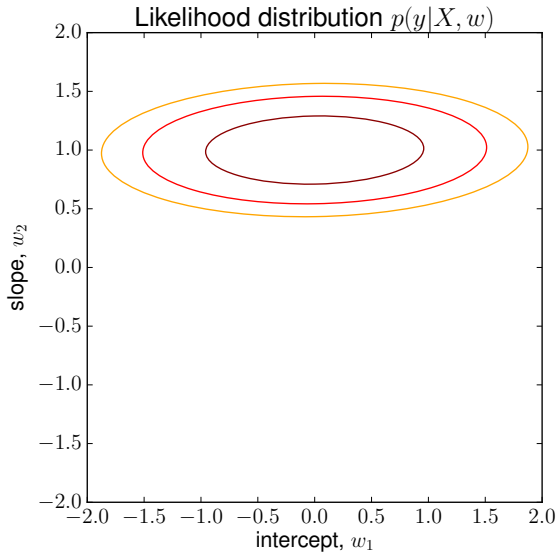


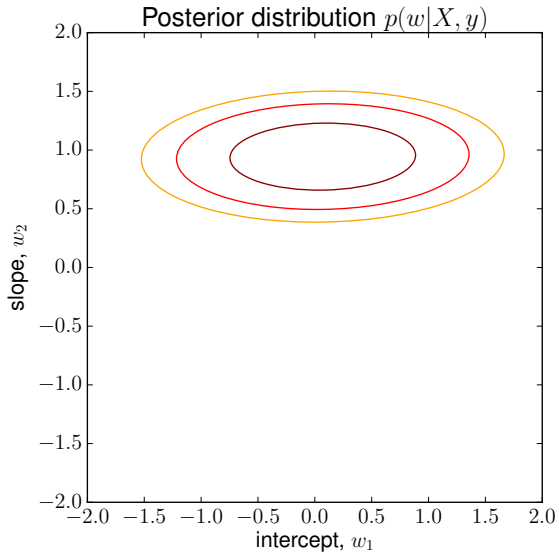


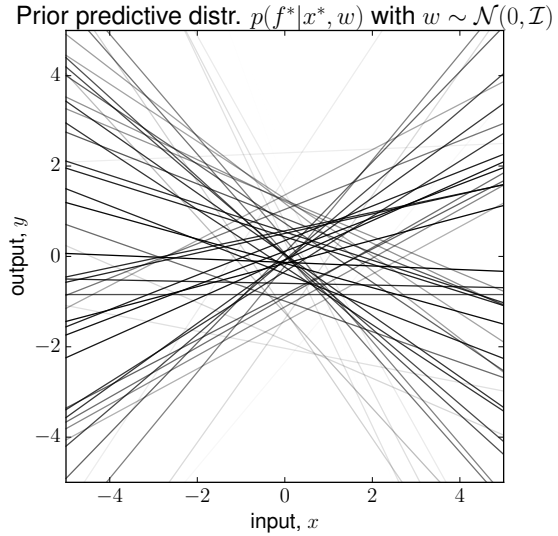


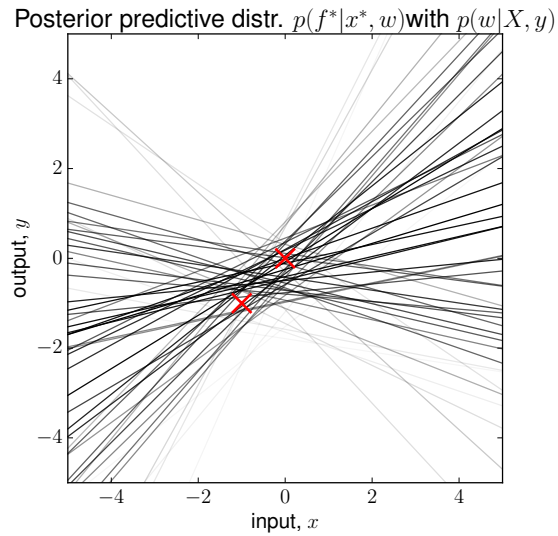
Weights correspond to functions (2)



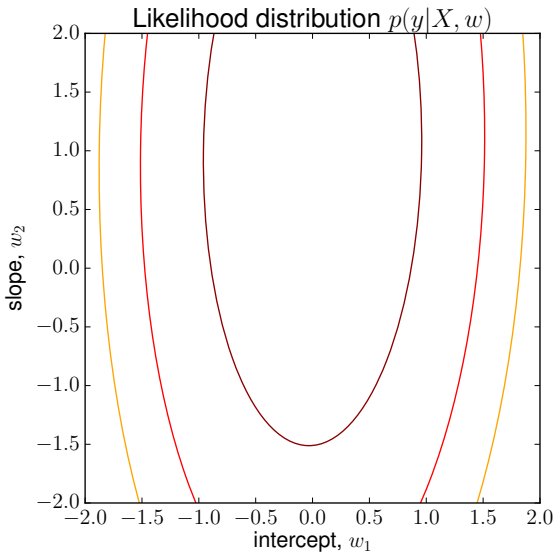


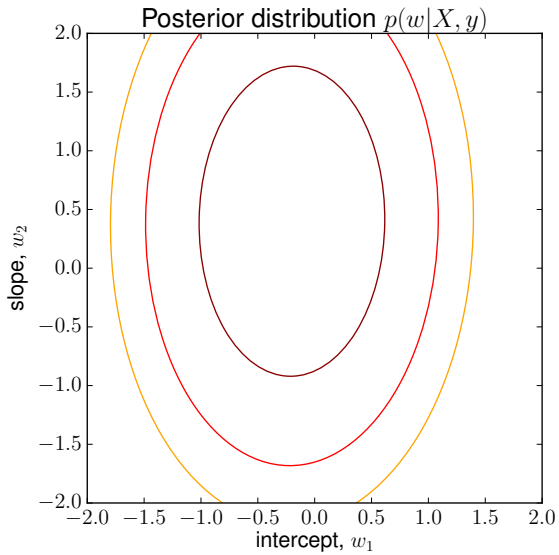


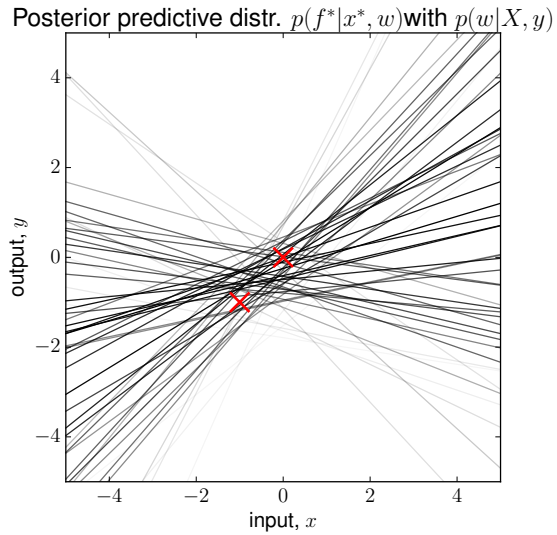


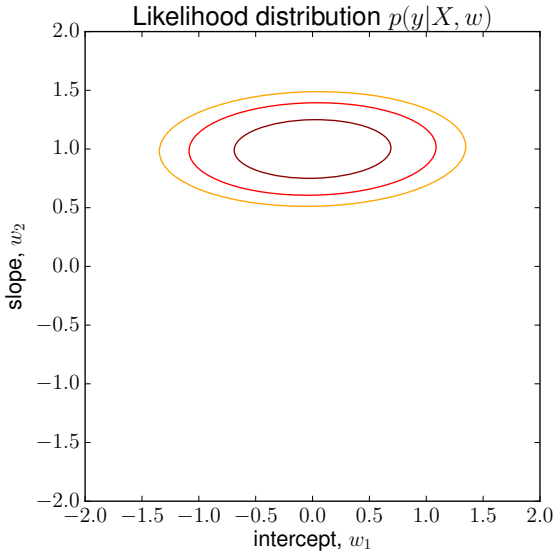


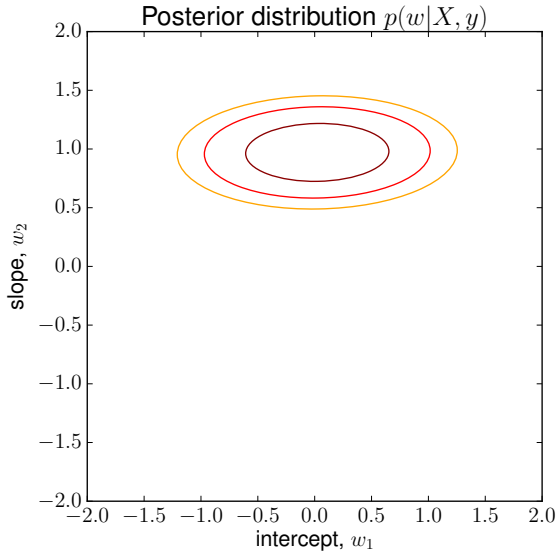
Enough data overwhelms the prior

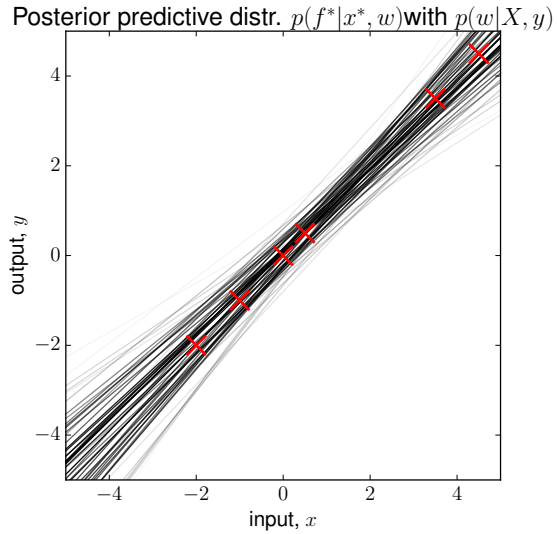












1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

Bayesian linear regression with basis functions

The same analysis goes through if we replace x with some so-called basis function of x, e.g., powers of x

$$\phi(x) = (1, x, x^2, x^3, \ldots)^\mathsf{T}$$

Basis function linear regression is then:

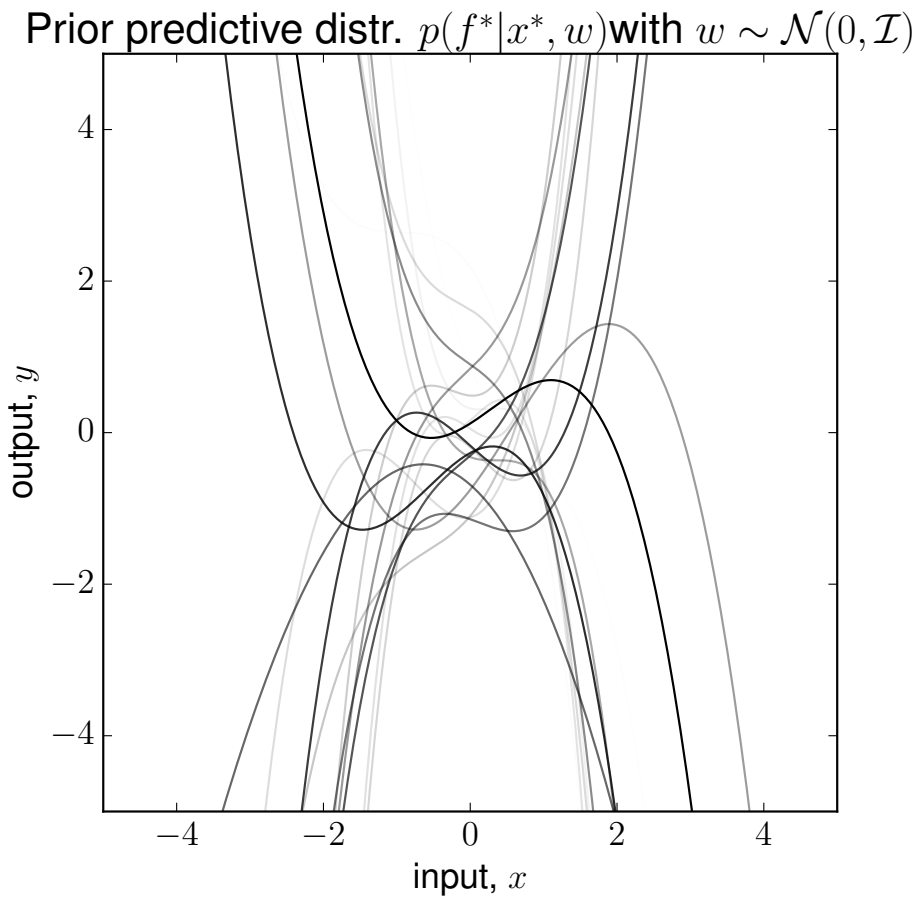
$$f(\boldsymbol{x}) = \phi(\boldsymbol{x})^\mathsf{T}\boldsymbol{w} \tag{2.10}$$

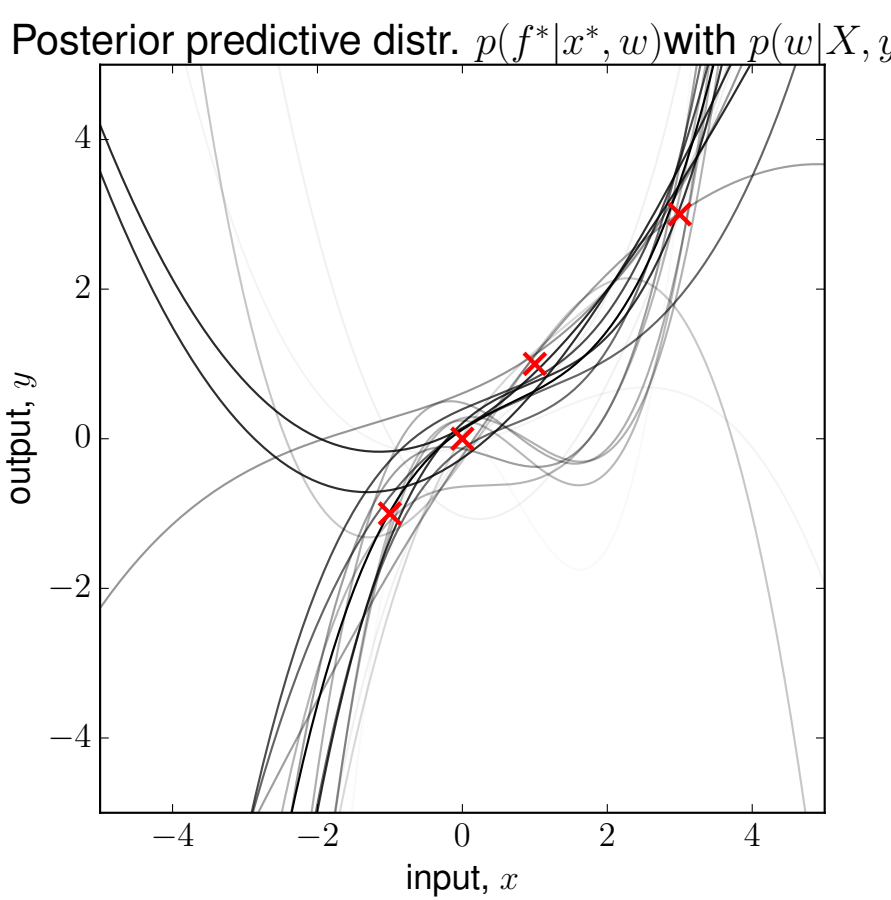
And the equivalent of 2.9 is:

$$f_\star|\boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}\left(\frac{1}{\sigma_n^2}\phi(\boldsymbol{x}_\star)^\mathsf{T}A^{-1}\Phi\boldsymbol{y}, \phi(\boldsymbol{x}_\star)^\mathsf{T}A^{-1}\phi(\boldsymbol{x}_\star)\right) \tag{2.11}$$

$$\text{with } \Phi = \phi(\boldsymbol{X}) \text{ and } A = \sigma_n^{-2}\Phi\Phi^\mathsf{T} + \Sigma_p^{-1}$$

Bayesian Cubic Regression





Rewriting this equation using a kernel (1)

In (2.11), we have to invert the $N \times N$ matrix A
(where N = number of dimensions, or basis functions):

$$f_\star | \boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\frac{1}{\sigma_n^2} \phi(\boldsymbol{x}_\star)^\intercal A^{-1} \Phi \boldsymbol{y}, \phi(\boldsymbol{x}_\star)^\intercal A^{-1} \phi(\boldsymbol{x}_\star))$$

We can rewrite this equation to instead have to invert a $n \times n$ matrix
(where n = number of data points):

$$\begin{aligned} f_\star | \boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(&\phi_\star^\intercal \Sigma_p \Phi (\Phi^\intercal \Sigma_p \Phi + \sigma_n^2 I)^{-1} \boldsymbol{y}, \\ &\phi_\star^\intercal \Sigma_p \phi_\star - \phi_\star^\intercal \Sigma_p \Phi (\Phi^\intercal \Sigma_p \Phi + \sigma_n^2 I)^{-1} \Phi^\intercal \Sigma_p \phi_\star), \end{aligned} \tag{2.12}$$

with shorthand notation $\phi_\star$.
You will derive this equivalence as part of this week's exercise.

Rewriting this equation using a kernel (2)

$$f_\star | \boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(\phi_\star^\mathsf{T} \Sigma_p \Phi (\Phi^\mathsf{T} \Sigma_p \Phi + \sigma_n^2 I)^{-1} \boldsymbol{y}, \tag{2.12}$$
$$\phi_\star^\mathsf{T} \Sigma_p \phi_\star - \phi_\star^\mathsf{T} \Sigma_p \Phi (\Phi^\mathsf{T} \Sigma_p \Phi + \sigma_n^2 I)^{-1} \Phi^\mathsf{T} \Sigma_p \phi_\star),$$

with shorthand notation $\phi_\star = \phi(\boldsymbol{x}_\star)$.

Note that the features $\phi(\boldsymbol{x}_\star)$ only enter this equation in the forms of $\phi_\star^\mathsf{T} \Sigma_p \Phi$, $\phi_\star^\mathsf{T} \Sigma_p \phi_\star$, and $\Phi^\mathsf{T} \Sigma_p \Phi$.
We can replace all of these occurrences by expressions of the form

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \phi(\boldsymbol{x}_1)^\mathsf{T} \Sigma_p \phi(\boldsymbol{x}_2)$$

This is known as the kernel trick. Then, the equation becomes:

$$f_\star | \boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(k_\star^\mathsf{T} (K + \sigma_n^2 I)^{-1} \boldsymbol{y}, \tag{2.12}$$
$$k_{\star\star} - k_\star^\mathsf{T} (K + \sigma_n^2 I)^{-1} k_\star.$$

Lecture Overview

1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

Gaussian process

Definition (Gaussian process)

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

A GP is completely specified by its mean and covariance function:

$$\begin{aligned} m(\boldsymbol{x}) =& \mathbb{E}\left[f(\boldsymbol{x})\right], \\ k(\boldsymbol{x}, \boldsymbol{x}') =& \mathbb{E}\left[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))\right] \end{aligned}$$

The GP is then:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{2.14}$$

The Bayesian linear regression model is a GP

The prior Bayesian linear model (before seeing any data $\boldsymbol{X}, \boldsymbol{y}$) is:

$$\begin{aligned} \mathbb{E}\left[f(\boldsymbol{x})\right] =& \phi(\boldsymbol{x})^T \mathbb{E}\left[\boldsymbol{w}\right] = 0 \\ \mathbb{E}\left[f(\boldsymbol{x}) f(\boldsymbol{x}')\right] =& \phi(\boldsymbol{x})^\mathsf{T} \mathbb{E}\left[\boldsymbol{w}\boldsymbol{w}^\mathsf{T}\right] \phi(\boldsymbol{x}') = \phi(\boldsymbol{x})^T \Sigma_p \phi(\boldsymbol{x}') \end{aligned} \tag{2.15}$$

Thus $f(\boldsymbol{x})$ and $f(\boldsymbol{x}')$ are jointly Gaussian with mean zero and covariance function $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\mathsf{T} \Sigma_p \phi(\boldsymbol{x}')$

The squared exponential (SE) covariance

Many different covariance functions (or kernel functions) are possible; a prominent one is the squared exponential (SE) covariance function:

$$cov(f(\boldsymbol{x}_p), f(\boldsymbol{x}_q)) = k(\boldsymbol{x}_p, \boldsymbol{x}_q) = \exp\left(-\frac{1}{2}|\boldsymbol{x}_p - \boldsymbol{x}_q|^2\right) \tag{2.16}$$

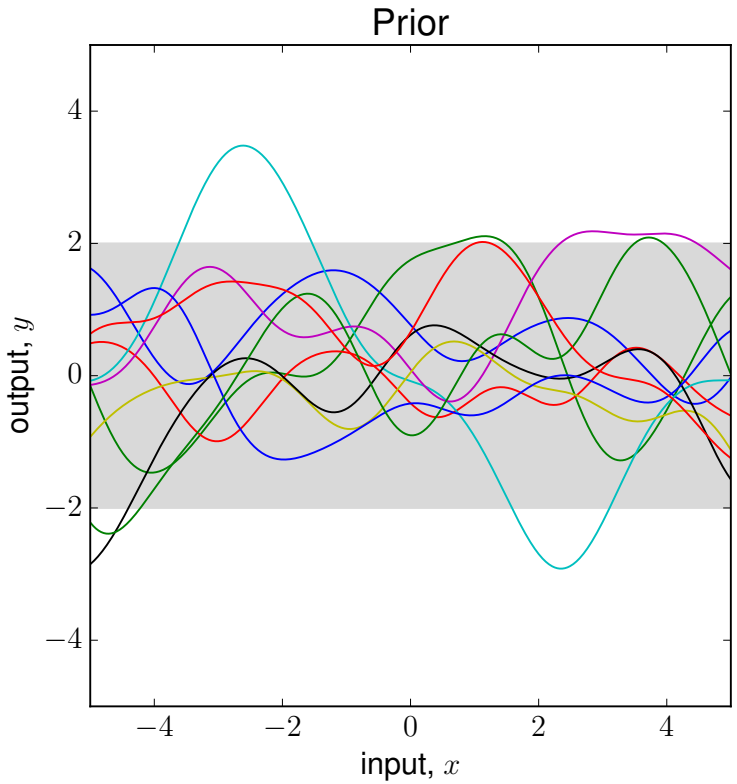
- Points that are close to each other have correlation ≈ 1
- Points that are far away from each other have correlation ≈ 0
- This gives rise to very smooth functions

Samples from the GP prior

Samples for a finite number of points $\boldsymbol{X}_\star$

$$\boldsymbol{f}_\star \sim \mathcal{N}\left(\mathbf{0}, K(\boldsymbol{X}_\star, \boldsymbol{X}_\star)\right)$$

Thus, we're just sampling from a multivariate Gaussian distribution. For visualization, we connect the dots.



1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

Prediction with Noise-free Observations

Observed function values f and new function values $f_\star$ are jointly Gaussian-distributed

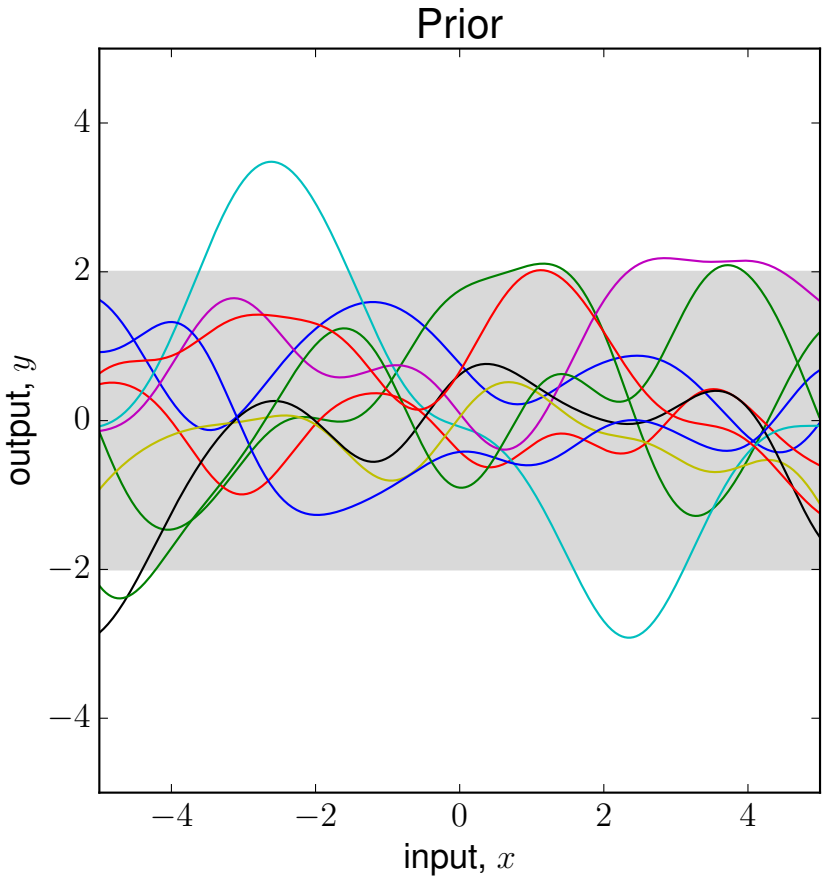
$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{X}, \boldsymbol{X}_\star) \\ K(\boldsymbol{X}_\star, \boldsymbol{X}) & K(\boldsymbol{X}_\star, \boldsymbol{X}_\star) \end{bmatrix}\right)$$

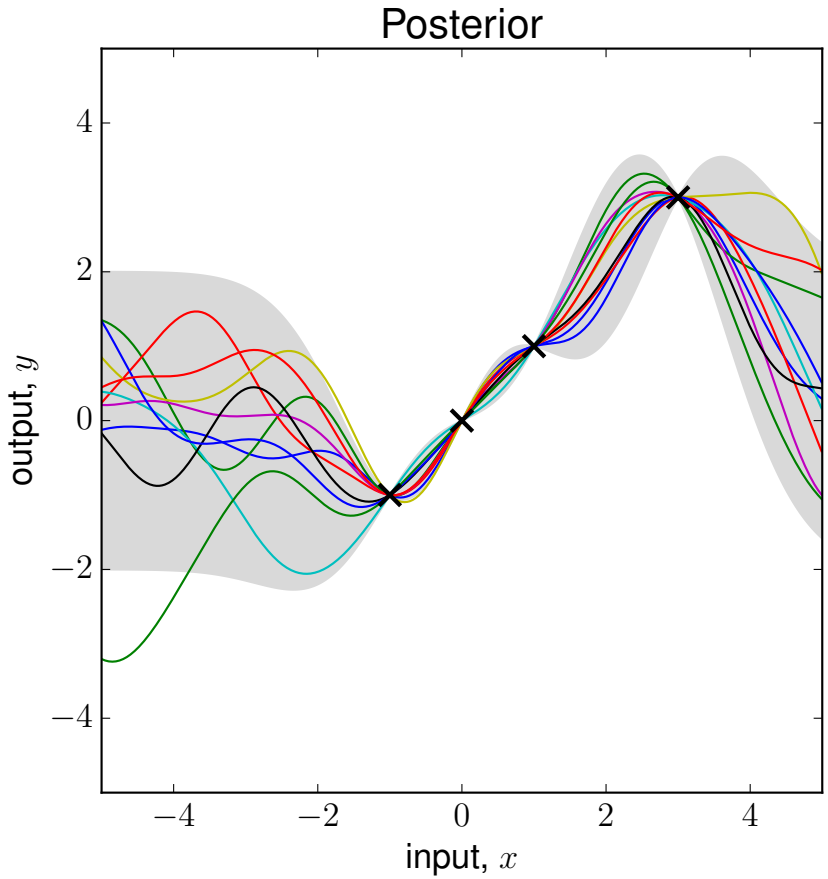
We simply apply the formula for conditioning a Gaussian to compute the predictive distribution $\boldsymbol{f}_\star | \boldsymbol{f}$:

$$\begin{aligned} \boldsymbol{f}_\star | \boldsymbol{X}_\star, \boldsymbol{X}, \boldsymbol{f} \sim \mathcal{N}(&K(\boldsymbol{X}_\star, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{f}, \\ &K(\boldsymbol{X}_\star, \boldsymbol{X}_\star) - K(\boldsymbol{X}_\star, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{X}, \boldsymbol{X}_\star)) \end{aligned} \tag{2.19}$$

Samples from prior and posterior

Intuitively, we draw many (infinitely many) samples from the prior and discard those that don't agree perfectly with the noise-free data points





1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

Now we have noisy observations

As in the case of Bayesian liner regression:
we'll assume i.i.d Gaussian noise with variance σ_n^2

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

The prior on the noise observations is:

$$cov(y_p, y_q) = k(\boldsymbol{x}_p, \boldsymbol{x}_q) + \sigma_n^2 \delta_{pq}$$
$$cov(\boldsymbol{y}) = K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I \tag{2.20}$$

Thus:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I & K(\boldsymbol{X}, \boldsymbol{X}_\star) \\ K(\boldsymbol{X}_\star, \boldsymbol{X}) & K(\boldsymbol{X}_\star, \boldsymbol{X}_\star) \end{bmatrix}\right) \tag{2.21}$$

Predictions under noisy observations

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I & K(\boldsymbol{X}, \boldsymbol{X}_\star) \\ K(\boldsymbol{X}_\star, \boldsymbol{X}) & K(\boldsymbol{X}_\star, \boldsymbol{X}_\star) \end{bmatrix}\right) \tag{2.21}$$

Using the same equation for conditioning Gaussians as before, we derive the posterior predictive distribution:

$$f_\star | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}_\star \sim \mathcal{N}(\bar{\boldsymbol{f}}_\star, cov(f_\star)), \text{ where} \tag{2.22}$$

$$\bar{\boldsymbol{f}}_\star \triangleq [f_\star | \boldsymbol{X}, y, \boldsymbol{X}_\star] = K(\boldsymbol{X}_\star, \boldsymbol{X}) \left[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I\right]^{-1} y, \tag{2.23}$$

$$cov(\boldsymbol{f}_\star) = K(\boldsymbol{X}_\star, \boldsymbol{X}_\star) - K(\boldsymbol{X}_\star, \boldsymbol{X}) \left[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 I\right]^{-1} K(\boldsymbol{X}, \boldsymbol{X}_\star) \tag{2.24}$$

The predictive distribution

Using more compact notation, we have:

$$\bar{f}_\star = \boldsymbol{k}_\star^\mathsf{T}(K + \sigma_n^2 I)^{-1}\boldsymbol{y}, \tag{2.25}$$

$$\mathbb{V}\left[f_\star\right] = k(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - \boldsymbol{k}_\star^\mathsf{T}(K + \sigma_n^2 I)^{-1}\boldsymbol{k}_\star, \tag{2.26}$$

Note that with covariance function $\phi(\boldsymbol{x})^\mathsf{T}\Sigma_p\phi(\boldsymbol{x}')$ this is exactly the same as Bayesian linear regression:

$$\begin{aligned} f_\star|\boldsymbol{x}_\star, \boldsymbol{X}, \boldsymbol{y} \sim \mathcal{N}(&\phi_\star^\mathsf{T}\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\boldsymbol{y}, \\ &\phi_\star^\mathsf{T}\Sigma_p\phi_\star - \phi_\star^\mathsf{T}\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\Phi^\mathsf{T}\Sigma_p\phi_\star) \end{aligned} \tag{2.12}$$

The predictive distribution

$$\bar{f}_\star = \boldsymbol{k}_\star^\mathsf{T}(K + \sigma_n^2 I)^{-1}\boldsymbol{y}, \tag{2.25}$$

$$\mathbb{V}\left[f_\star\right] = k(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - \boldsymbol{k}_\star^\mathsf{T}(K + \sigma_n^2 I)^{-1}\boldsymbol{k}_\star, \tag{2.26}$$

Some properties of the mean

- A linear combination of response values y
- A linear combination of n kernel functions, each one centered on a training data point:

$$f(\bar{\boldsymbol{x}}_\star) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}_\star) \tag{2.27}$$

$$\text{where } \boldsymbol{\alpha} = (K + \sigma_n^2 I)^{-1}\boldsymbol{y}$$

1. The Gaussian distribution

2. Gaussian processes: the weight space view
   - Bayesian linear regression
   - Bayesian linear regression with basis functions

3. Gaussian processes: the function space view
   - The Case of Noise-free Observations
   - The Case of Noisy Observations
   - Marginal likelihood and kernel hyperparameters

The marginal likelihood

The marginal likelihood is the integral of the likelihood times the prior:

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X}) p(\boldsymbol{f}|\boldsymbol{X}) d\boldsymbol{f}. \tag{2.28}$$

It marginalizes (integrates out) the actual function values.
We know that this is distributed according to a Gaussian:

$$\boldsymbol{y} \sim \mathcal{N}(0, K + \sigma_n^2 I)$$

Usually, one works with the log marginal likelihood:

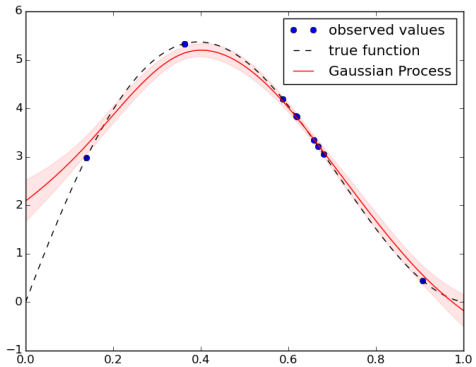
$$\log p(\boldsymbol{y}|\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^\mathsf{T}(K + \sigma_n^2 I)^{-1}\boldsymbol{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \tag{2.30}$$

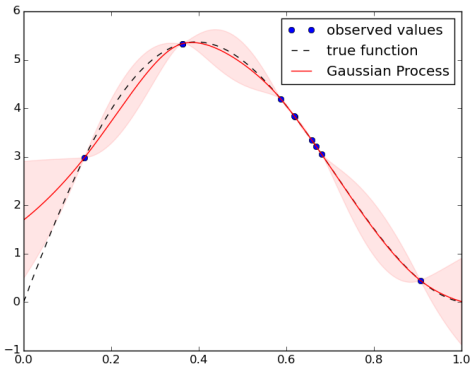
The effect of the kernel hyperparameters

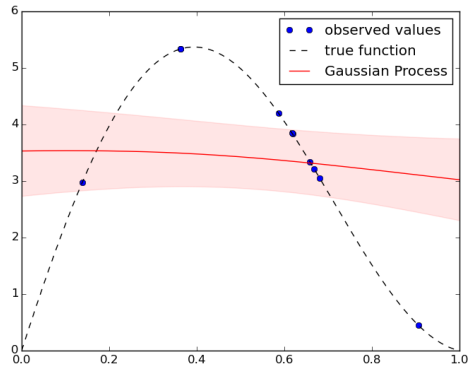
The SE covariance function in 1-d:

$$k_y(x_p, x_q) = \sigma_f^2 \exp(-\frac{1}{2l^2}(x_p - x_q)^2) + \sigma_n^2 \delta_{pq} \tag{2.31}$$

The marginal likelihood is highest for the left plot. It trades off data fit and "complexity" of the function







Summary by Learning Goals

Now (after digesting the material and doing the exercise), you can . . .

- Derive Bayesian linear regression
- Derive Gaussian processes
- Explain the relationship between Bayesian linear regression and Gaussian processes