Introduction to Probabilistic Machine Learning

Ralf Herbrich

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Gaussian Processes

Overview

1. Basic Concepts
2. Gaussian Processes for Regression
   - Weight-Space View
3. Gaussian Processes for Classification
4. Evidence Maximization for Gaussian Processes

Overview

1. **Basic Concepts**
2. Gaussian Processes for Regression
   - Weight-Space View
3. Gaussian Processes for Classification
4. Evidence Maximization for Gaussian Processes

Gaussian Processes

- So far, we assumed that we have a set of **parameterized** functions $f(x; \boldsymbol{w})$ and parameterized a distribution over f via the vector space $\mathbb{R}^M \ni \boldsymbol{w}$
  - Required *explicit* definition of M basis functions $\phi: \mathcal{X} \to \mathbb{R}$
  - The most expensive operation was $O(M^3)$ (matrix inversion)
- **Gaussian Process**. A Gaussian process is a probability distribution over functions $f: \mathcal{X} \to \mathbb{R}$ that has the property that for any $x_1, \ldots, x_n \in \mathcal{X}$

$$p\left(\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}\right) = \mathcal{N}\left(\boldsymbol{f}; \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{bmatrix}, \begin{bmatrix} C(x_1, x_1) & \ldots & C(x_1, x_n) \\ \vdots & \ddots & \vdots \\ C(x_n, x_1) & \ldots & C(x_n, x_n) \end{bmatrix}\right)$$

  - Fully parameterized by a mean *function* $m: \mathcal{X} \to \mathbb{R}$ and covariance *function* $C: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  - Mean function is arbitrary, but the covariance function must be positive definite!
- **Gaussian processes** have a long history in engineering ("Kriging")
  - In 1950, Danie G. Krige was studying gold locations predicted from a few boreholes
  - In 1960, Georges Matheron rediscovered Krige's work and formalized it for statistics
  - In 1999, Prof. David MacKay rediscovered Gaussian Processes for machine learning

Danie G. Krige
(1919 – 2013)

Georges Matheron
(1930 – 2000)

Sir David JC MacKay
(1967 – 2016)

Gaussian Process: Example

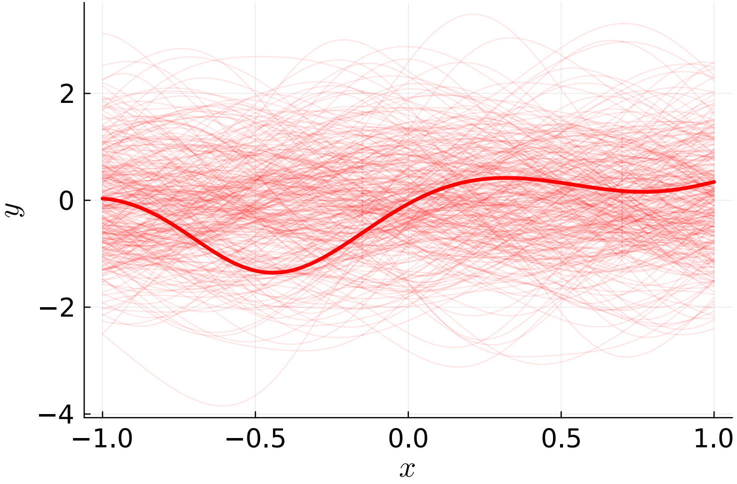
$$m(x) = 0$$

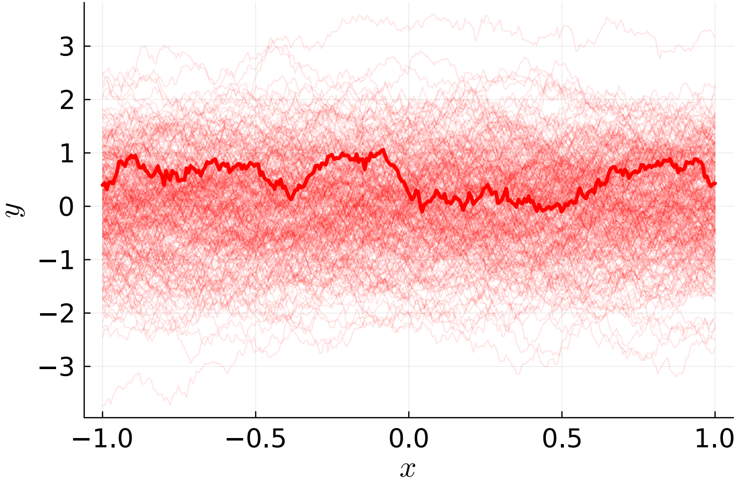
RBF Kernel

$$C(x, x') = \exp\left(-\frac{(x - x')^2}{\lambda^2}\right)$$

Ornstein-Uhlenbeck Kernel

$$C(x, x') = \exp\left(-\frac{|x - x'|}{\lambda}\right)$$





Overview

Partitioned Multivariate Normal Distribution

- **Partitioned Gaussians.** *Given a joint distribution* $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ *and a partitioning*

$$\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}$$

we have the following for the marginal $p(\boldsymbol{x}_a)$ *and the conditional* $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$

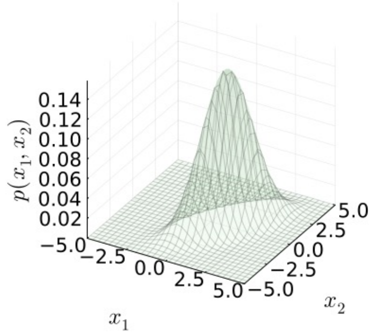
$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a;\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_{aa})$$

$$p(\boldsymbol{x}_b|\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_b;\boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a), \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab})$$

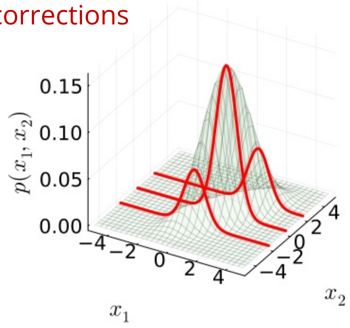
additive corrections

$$\boldsymbol{\mu} = \begin{bmatrix}1\\0\end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix}1 & 1\\ 1 & 2\end{bmatrix}$$



$$p(x_1) = \mathcal{N}(x_1; 1,1)$$



$$p(x_1|x_2) = \mathcal{N}\left(x_1; 1 + \frac{1}{2}x_2, \frac{1}{2}\right)$$

Introduction to Probabilistic Machine Learning

Unit 10 – Gaussian Processes

Bayes' Theorem for Normal Distributions

- **Conjugate Gaussians.** *Given a normally distributed variable*

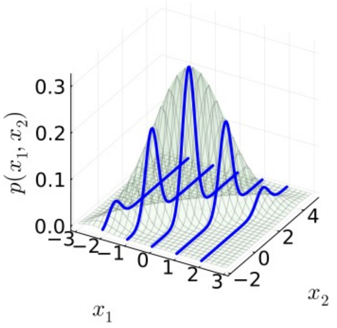
$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and a conditional distribution for $\boldsymbol{y}$ *given* $\boldsymbol{x}$ *such that* $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{S})$ *we have the following for the marginal* $p(\boldsymbol{y})$ *and the "inverse" conditional* $p(\boldsymbol{x}|\boldsymbol{y})$

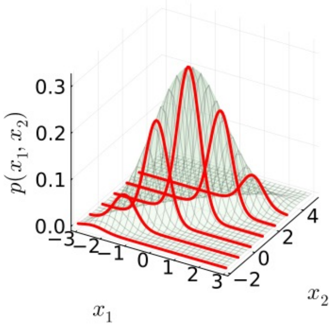
$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{S} + \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}\right)$$

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{G}\left(\boldsymbol{x}; \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}^{-1}(\boldsymbol{y} - \boldsymbol{b}), \boldsymbol{\Sigma}^{-1} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}^{-1}\boldsymbol{A}\right),$$

$$p(x_1) = \mathcal{N}(x_1; 0,1)$$



$$p(x_2|x_1) = \mathcal{N}\left(x_2; x_1 + 1, \frac{1}{2}\right)$$



$$p(x_1|x_2) = \mathcal{N}\left(x_1; \frac{2}{3}(x_2 - 1), \frac{1}{3}\right)$$

Inference in a Gaussian Processes (GP)

1. **Observation**: *The only part of our data that is predicted is the targets $\boldsymbol{y}$ and everything is conditioned on the locations $\{x_1, \dots, x_n\} =: X$.*

$$p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|\boldsymbol{f}, X) \cdot \overbrace{p(\boldsymbol{f}|X)}^{\mathrm{GP}(\boldsymbol{f};0,C)} \, d\boldsymbol{f}$$

$$p(\boldsymbol{y}|X) = \int \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \sigma^2 \boldsymbol{I}) \cdot \mathcal{N}(\boldsymbol{f}; \boldsymbol{0}, \boldsymbol{C}_X) d\boldsymbol{f}$$

$$p(\boldsymbol{y}|X) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{C}_X)$$

2. **Observation**: *The marginal distribution of data over all functions is also a GP with the covariance/kernel function modified by an identity function!*

$$\mathcal{N}(\boldsymbol{y}; \boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{C}_X) = \mathrm{GP}\big(\boldsymbol{y}; 0, \underbrace{C + \sigma^2 \mathbb{I}(x = x')}_{\tilde{C}}\big)$$

- **Inference** of the predictive distribution at a new point $x^* \in \mathcal{X}$ in a Gaussian Process can be done by conditioning!

$$p(y^*|\boldsymbol{y}, x^*, X) = \frac{p(y^*, \boldsymbol{y}|x^*, X)}{p(\boldsymbol{y}|X)}$$

$$\boxed{\begin{aligned} &p(y^*|x^*, X, \boldsymbol{y}) = \mathcal{N}\left(y^*; \boldsymbol{k}^{\mathrm{T}} \widetilde{\boldsymbol{C}}^{-1} \boldsymbol{y}, C(x^*, x^*) + \sigma^2 - \boldsymbol{k}^{\mathrm{T}} \widetilde{\boldsymbol{C}}^{-1} \boldsymbol{k}\right) \\ &\text{where } \boldsymbol{k} := \left(\tilde{C}(x_1, x^*), \dots, \tilde{C}(x_n, x^*)\right)^{\mathrm{T}} \text{ and } \widetilde{\boldsymbol{C}}_{ij} := \tilde{C}(x_i, x_j) \end{aligned}}$$

Properties of Gaussians

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{f}, \boldsymbol{S})$$

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{S} + \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathrm{T}}\right)$$

Partitioned Gaussians

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

$$p(\boldsymbol{x}_b|\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_b; \boldsymbol{m}, S)$$

$$\boldsymbol{m} = \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)$$

$$\boldsymbol{S} = \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}$$

Gaussian Process: Example

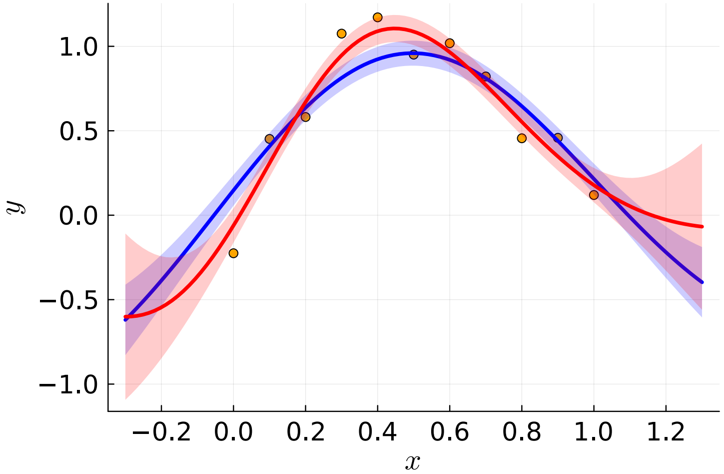
$$m(x) = 0$$

RBF Kernel

$$C(x, x') = \exp\left(-\frac{(x - x')^2}{\lambda^2}\right)$$

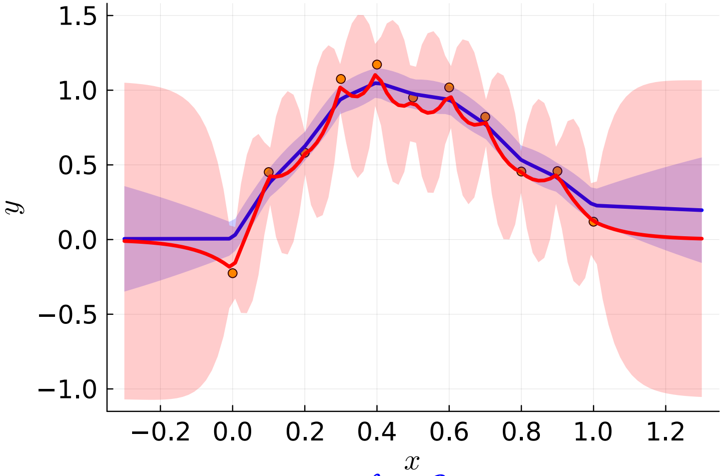
Ornstein-Uhlenbeck Kernel

$$C(x, x') = \exp\left(-\frac{|x - x'|}{\lambda}\right)$$



$\lambda = 1$

$\lambda = 0.5$



$\lambda = 2$

$\lambda = 0.1$

Overview

Bayesian Linear Regression Revisited

- **Bayesian Linear Regression**: Linear basis function model $f(x; \boldsymbol{w}) := \boldsymbol{w}^T \boldsymbol{\phi}(x)$
  - Likelihood: $p(\boldsymbol{y}|\boldsymbol{w}, X) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\Phi w}, \sigma^2 \boldsymbol{I})$
  - Prior: $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \boldsymbol{I})$
  - Marginal Likelihood: $p(\boldsymbol{y}|X) = \int p(\boldsymbol{y}|\boldsymbol{w}, X) \cdot p(\boldsymbol{w}) d\boldsymbol{w}$

$$p(\boldsymbol{y}|X) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{0}, \sigma^2 \boldsymbol{I} + \boldsymbol{\Phi\Phi}^{\mathrm{T}}\right) = \mathrm{GP}\left(\boldsymbol{y}; 0, \boldsymbol{\phi}(x)^{\mathrm{T}} \boldsymbol{\phi}(x') + \sigma^2 \mathbb{I}(x = x')\right)$$

- **Equivalence**: *A Gaussian Process is equivalent to the marginal likelihood of a linear basis function model with the prior* $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{0}, \boldsymbol{I})$ *and the covariance function* $C(x, x') = \boldsymbol{\phi}(x)^T \boldsymbol{\phi}(x')$.

	Number of Basis Functions	Computational Cost
Linear Basis Function Model	M	$O(M^3)$
Gaussian Process	∞	$O(n^3)$

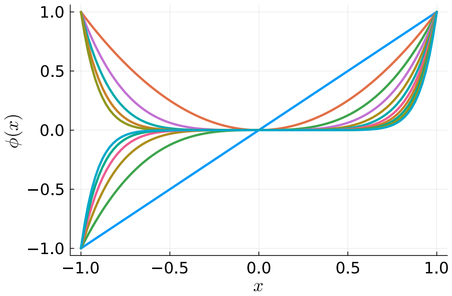
Properties of Gaussians

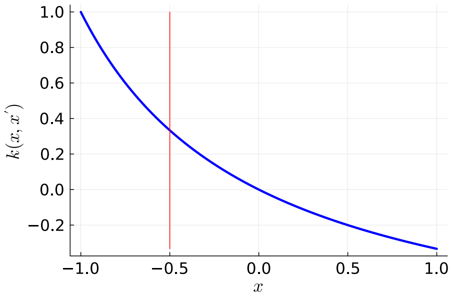
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$p(\boldsymbol{y}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{v}; \boldsymbol{\Phi w}, \boldsymbol{S})$$
$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\Phi\mu}, \boldsymbol{S} + \boldsymbol{A}\Sigma\mathrm{A}^{\mathrm{T}}\right)$$

Constructing Covariance Functions

Polynomial Basis

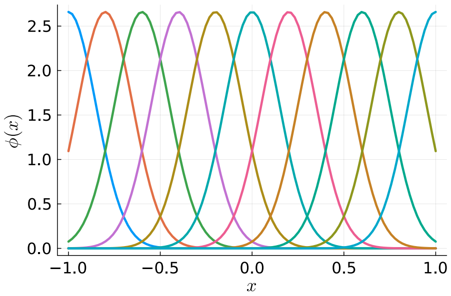
$$\phi_j(x) = x^j$$

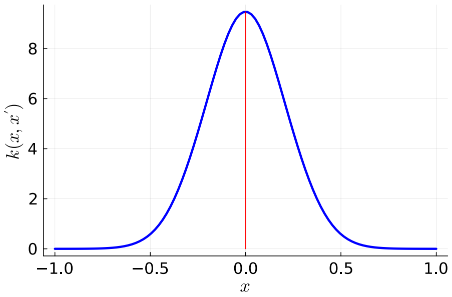




Gaussian Basis

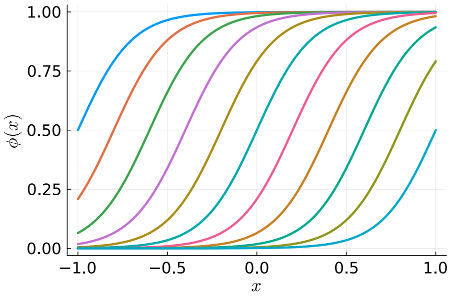
$$\phi_j(x) = \mathcal{N}\left(x; \frac{j}{5} - 1, 0.15^2\right)$$

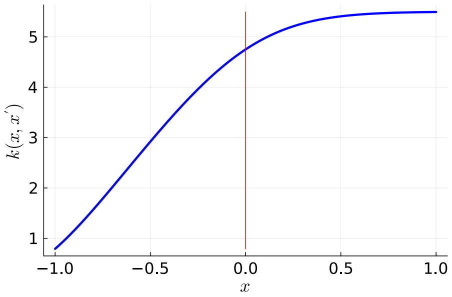




Sigmoid Basis

$$\phi_j(x) = \frac{\exp(x - 0.2j + 1)}{1 + \exp(x - 0.2j + 1)}$$





Overview

1. Basic Concepts
2. Gaussian Processes for Regression
   - Weight-Space View
3. **Gaussian Processes for Classification**
4. Evidence Maximization for Gaussian Processes

Gaussian Processes with Logit for Classification

- A **Gaussian Process (GP) prior** is a prior $p(f)$ over functions f such that for any $x_1, \dots, x_n \in \mathcal{X}$

$$p(\boldsymbol{f}|X) = p\left(\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}\right) = \mathcal{N}\left(\boldsymbol{f}; \boldsymbol{0}, \begin{bmatrix} C(x_1,x_1) & \dots & C(x_1,x_n) \\ \vdots & \ddots & \vdots \\ C(x_n,x_1) & \dots & C(x_n,x_n) \end{bmatrix}\right) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{0}, \boldsymbol{C}_X)$$

- **Problem**: The data model $p(y_i|f_i) = \mathrm{Ber}(y_i; g(f_i))$ is not conjugate to the Gaussian resulting a non-Gaussian $p(\boldsymbol{f}|X,\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{f}) \cdot p(\boldsymbol{f}|X)$!
- **Predictions**. By marginalization over $p(f^*|x^*,X,\boldsymbol{y})$ we can get the predictive distribution over $y^* \in \{0,1\}$ at a new point $x^* \in \mathcal{X}$

$$p(y^*|x^*,X,\boldsymbol{y}) = \int_{-\infty}^{+\infty} p(y^*|f^*) \cdot p(f^*|x^*,X,\boldsymbol{y})\, df^*$$

$$p(f^*|x^*,X,\boldsymbol{y}) = \int p(f^*|x^*,X,\boldsymbol{f}) \cdot p(\boldsymbol{f}|X,\boldsymbol{y})\, d\boldsymbol{f}$$

$$p(f^*|x^*,X,\boldsymbol{f}) = \frac{p(f^*,\boldsymbol{f}|x^*,X)}{p(\boldsymbol{f}|X)} = \mathcal{N}\left(f^*; \boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_X^{-1}\boldsymbol{f}, C(x^*,x^*) - \boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_X^{-1}\boldsymbol{k}\right)$$

Partitioned Gaussians

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

$$p(\boldsymbol{x}_b|\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_b; \boldsymbol{m}, S)$$

$$\boldsymbol{m} = \boldsymbol{\mu}_b + \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)$$

$$\boldsymbol{S} = \boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}$$

Approximate Gaussian Processes for Classification

- **Idea**: We approximate $p(\boldsymbol{f}|X,\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{f}) \cdot p(\boldsymbol{f}|X)$ with a Gaussian $\mathcal{N}(\boldsymbol{f};\boldsymbol{m},\boldsymbol{A})$.

$$p(y^*|x^*,X,\boldsymbol{y}) = \int_{-\infty}^{+\infty} p(y^*|f^*) \cdot \mathcal{N}\left(f^*; \boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_X^{-1}\boldsymbol{m}, C(x^*,x^*) - \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{C}_X^{-1} - \boldsymbol{C}_X^{-1}\boldsymbol{A}\boldsymbol{C}_X^{-1})\boldsymbol{k}\right) df^*$$

- **Laplace Approximation**: Similar to Bayesian Linear Logit Regression, we use the Laplace approximation on the distribution $p(\boldsymbol{f}|X,\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{f}) \cdot p(\boldsymbol{f}|X)$!

1. **Initialize:** A the latent function values vector $\boldsymbol{m} = \boldsymbol{0}$ and compute the covariance matrix $\boldsymbol{C} \in \mathbb{R}^{n\times n}$ once
2. **Iterate** until convergence
   - Compute $g_i = \frac{\exp(m_i)}{1+\exp(m_i)}$ and $\boldsymbol{R} = \mathrm{diag}\left(\begin{bmatrix} -g_1(1-g_1) \\ \vdots \\ -g_n(1-g_n) \end{bmatrix}\right)$
   - Update $\boldsymbol{m} \leftarrow \boldsymbol{m} - (\boldsymbol{R} - \boldsymbol{C}^{-1})^{-1}\left((\boldsymbol{y}-\boldsymbol{g}) - \boldsymbol{C}^{-1}\boldsymbol{m}\right)$
3. **Set** the covariance $\boldsymbol{A}$ of Gaussian approximation to $\boldsymbol{A} = (\boldsymbol{C}^{-1} - \boldsymbol{R})^{-1}$

Properties of Gaussians

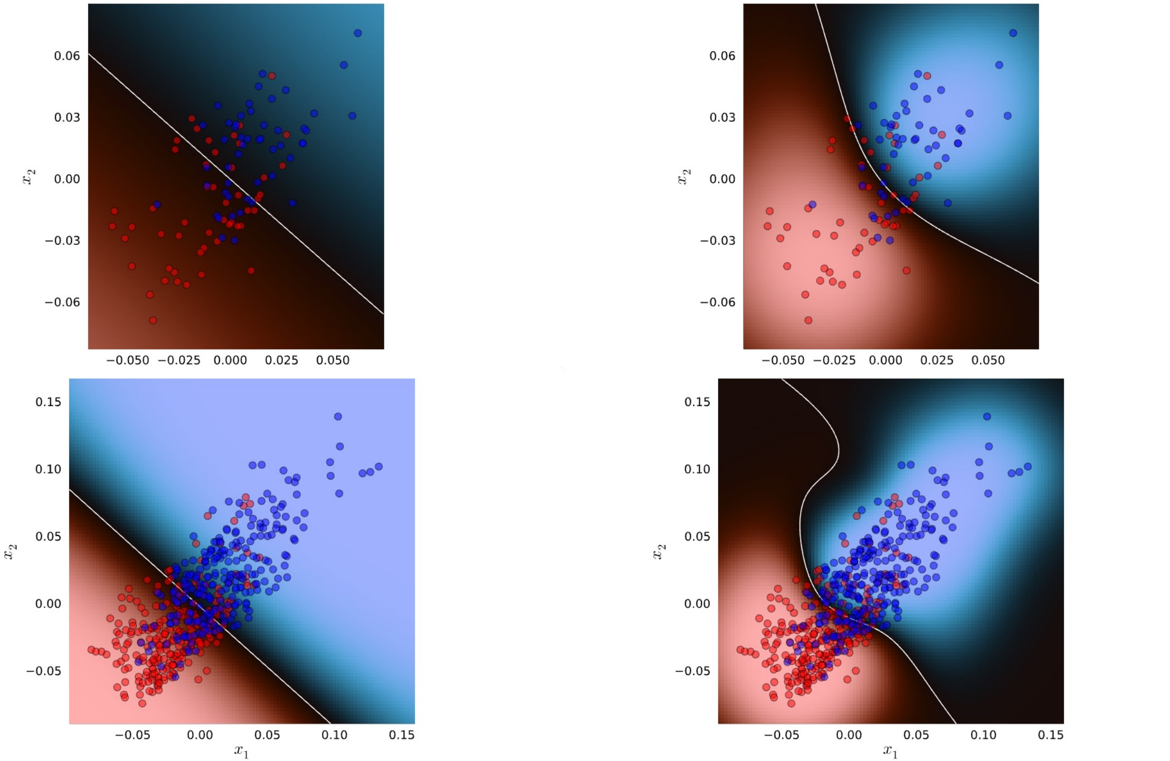
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w};\boldsymbol{\mu},\boldsymbol{\Sigma})$$
$$p(\boldsymbol{v}|\boldsymbol{w}) = \mathcal{N}(\boldsymbol{v};\boldsymbol{A}\boldsymbol{w},\Xi)$$
$$p(\boldsymbol{v}) = \mathcal{N}\left(\boldsymbol{v};\boldsymbol{A}\boldsymbol{\mu},\Xi + \boldsymbol{A}\Sigma\mathrm{A}^{\mathrm{T}}\right)$$

Gaussian Process Logistic Regression in Pictures

$\lambda = 0.5$ | $\lambda = 0.05$

$n = 100$

$n = 500$



Overview

1. Basic Concepts
2. Gaussian Processes for Regression
   - Weight-Space View
3. Gaussian Processes for Classification
4. **Evidence Maximization for Gaussian Processes**

Model Averaging and Selection

- All our inference algorithms have **assumed a known and fixed** set of basis functions as well as prior variance τ^2 and data noise variances σ^2 (i.e., **model**)
- **Model**. *A model $\mathcal{M}$ is assumed probability distribution over data, $p(D|\mathcal{M})$.*
  - Bayesian regression: $p(D|\mathcal{M}) = p(\mathbf{y}|X, \mathcal{M}) = \int p(\mathbf{y}|X, \mathbf{w}, \mathcal{M}) \cdot p(\mathbf{w}|\mathcal{M}) d\mathbf{w}$
  - Gaussian Processes: $p(D|\mathcal{M}) = p(\mathbf{y}|X, \mathcal{M}) = \int p(\mathbf{y}|\boldsymbol{f}, \mathcal{M}) \cdot p(\boldsymbol{f}|X, \mathcal{M}) d\boldsymbol{f}$
- **Model Averaging**. Given a set $\{\mathcal{M}\}$ of models, the predictive distribution for a new example x is obtained via

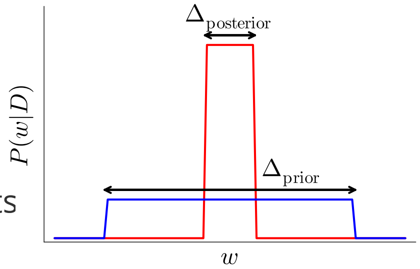
$$p(y|x, D) = \sum_{\mathcal{M}} p(y|x, D, \mathcal{M}) \cdot p(\mathcal{M}|D)$$

  - Often too difficult to compute so instead, approximate with a single best model $\mathcal{M}$
- **Model Selection**. Given a set $\{\mathcal{M}\}$ of models and a dataset D, select the best model

$$\mathcal{M}(D) = \mathrm{argmax}_{\mathcal{M}} p(\mathcal{M}|D) = \mathrm{argmax}_{\mathcal{M}} p(D|\mathcal{M}) \cdot p(\mathcal{M})$$

Model Selection: Intuition

- **Model Evidence**. *The probability of the data given a fixed model, $p(D|\mathcal{M})$, is called the* model evidence.
  - It's also called ***marginal likelihood*** because $p(D|\mathcal{M}) = \int p(D|f,\mathcal{M})p(f)df$ and $p(D|f,\mathcal{M})$ is called the likelihood of the function f (actually, really a misnomer!).
  - The **negative logarithm (base 2) of the model evidence** specifies the number of bits that the data can be compressed into without loss given the model $\mathcal{M}$.



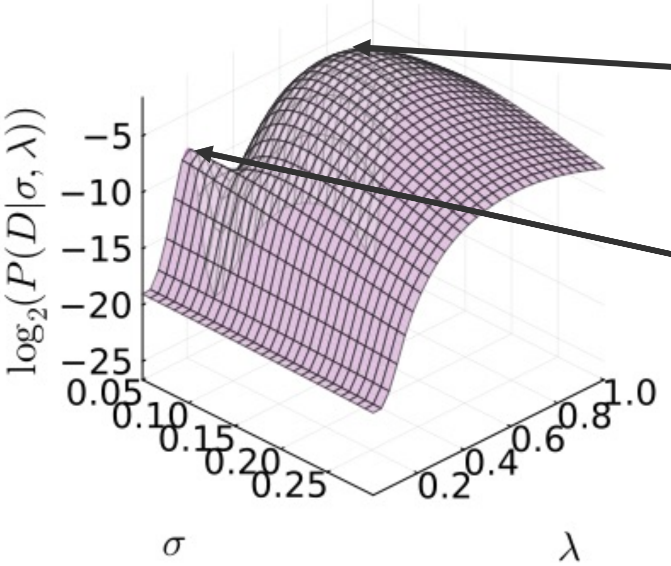
- **Approximation**. Assume that f has one parameter w, that $p(w)$ is uniform and that the posterior for a given dataset is also uniform. Then

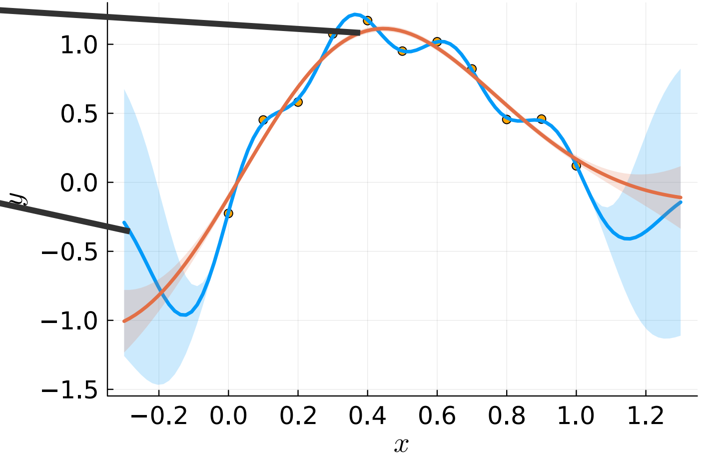
$$p(D|\mathcal{M}) = \int p(D|w,\mathcal{M}) \cdot p(w|\mathcal{M})dw = p(D|w_{\text{MAP}},\mathcal{M}) \cdot \frac{\Delta_{\text{posterior}}}{\Delta_{\text{prior}}}$$

$$\log_2 p(D|\mathcal{M}) = \underbrace{\log_2 p(D|w_{\text{MAP}},\mathcal{M})}_{\text{fit of the model to data}} - \underbrace{\log_2\left(\frac{\Delta_{\text{prior}}}{\Delta_{\text{posterior}}}\right)}_{\substack{\text{penalty} \\ \text{for} \\ \text{richness of model}}}$$

Gaussian Process Evidence Maximization

$$\log\left(p(\boldsymbol{y}|X)\right) = \log\left(\mathcal{N}\left(\boldsymbol{y}; \boldsymbol{0}, \sigma^2\boldsymbol{I} + \boldsymbol{C}_X\right)\right)$$
$$= -\frac{1}{2}\left[\log(|\sigma^2\boldsymbol{I} + \boldsymbol{C}_X|) + \boldsymbol{y}^{\mathrm{T}}(\sigma^2\boldsymbol{I} + \boldsymbol{C}_X)^{-1}\boldsymbol{y} + n \cdot \log(2\pi)\right]$$





Summary

1. **Gaussian Processes**
   - Distribution over space of functions rather than function parameters
   - However, Gaussian Process models are equivalent to linear basis function models with a different parameterization (covariance functions instead of basis functions!)
2. **Gaussian Process Classification**
   - Since the likelihood is no longer Gaussian, we have to approximate the posterior Gaussian process
   - Many approximation schemes exist; we introduced the Laplace approximation (Kuss et al., 2005)
   - Also possible to use approximate message passing
3. **Bayesian Model Comparison and Selection**
   - Model evidence as the key criterion: The probability of the data given a fixed model, $p(D \mid \mathcal{M})$.
   - Negative log-model evidence equals the compression length of the data (measured in bits): the further we can compress the target values, the better the model (see next lecture!)

See you next week!