Data, Estimation & Inference

Gaussian Processes & Bayesian Deep Learning

Michaelmas Term, 2020

University of Oxford

Tim G. J. Rudner

tim.rudner@cs.ox.ac.uk

Syllabus: https://www.notion.so/Data-Estimation-and-Inference-Part-2-d177bdb310e74f9aafd034c35652e4f3

Overview

Syllabus:
https://www.notion.so/Data-Estimation-and-Inference-Part-2-d177bdb310e74f9aafd034c35652e4f3

Today:

- Weight-space function-space duality
- Basics of Gaussian processes (GPs)
- Implementing GPs (if time permits)

Bayesian Inference in Machine Learning

Bayesian Inference in Machine Learning

How to represent (stochastic/random) processes in the real world probabilistically?

Bayesian Inference in Machine Learning

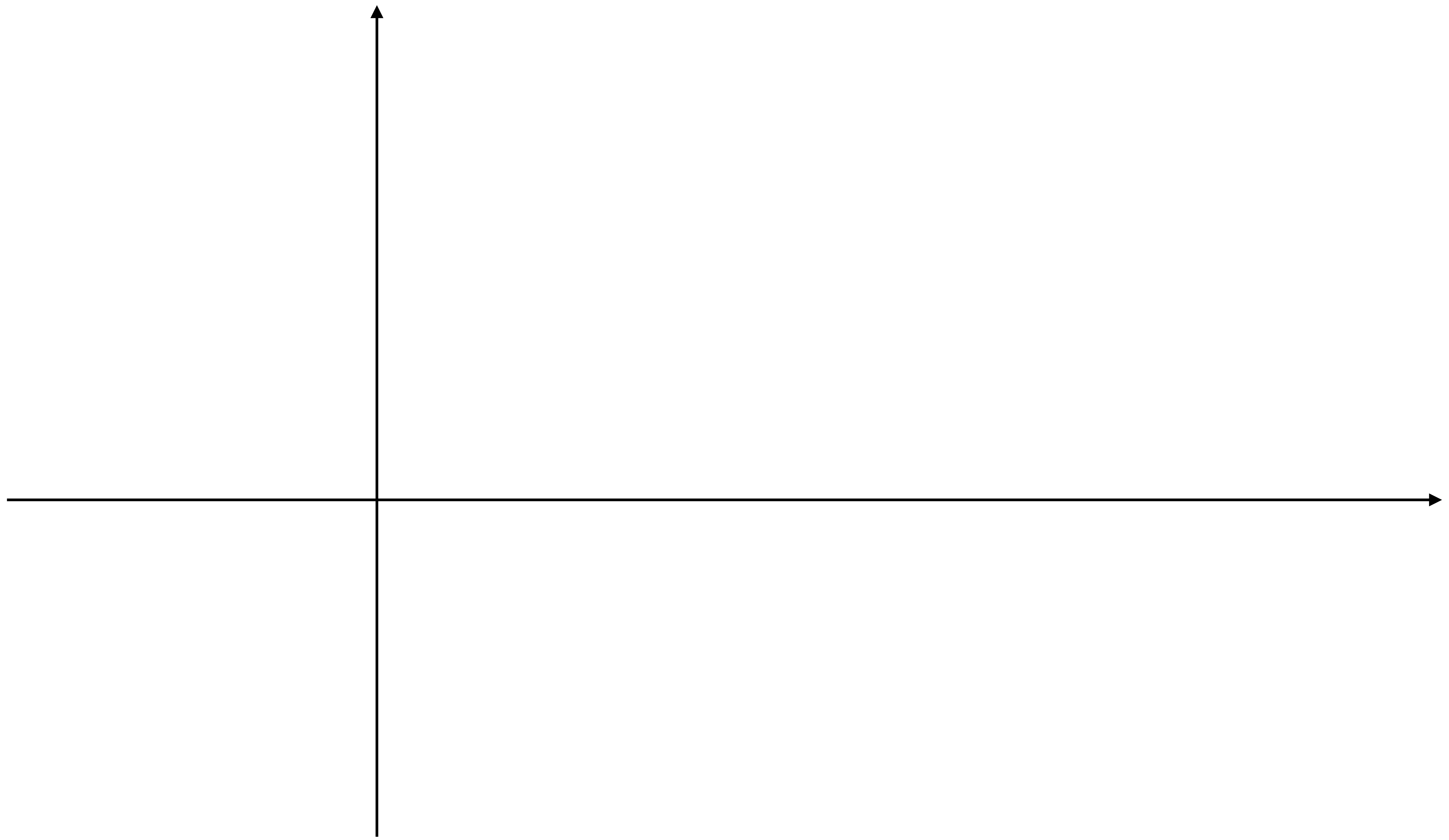
How to represent (stochastic/random) processes in the real world probabilistically?

Distributions over Functions

Distributions over Functions

Q: What is a stochastic function?

Q: What is a distribution over functions?



Consider

$$f(x; w) = wx, \quad w \sim \mathcal{N}(w|0, 1)$$

Link to visualization

Weight-space Function-space Duality

Weight-space Function-space Duality

Probabilistic model:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon,$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Bayesian inference:

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

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

Weight-space Function-space Duality

Probabilistic model:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon,$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Bayesian inference:

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

Posterior over weights:

$$p(\mathbf{w}|X, \mathbf{y}) \propto \exp\big(-\frac{1}{2\sigma_n^2}(\mathbf{y} - X^\top \mathbf{w})^\top (\mathbf{y} - X^\top \mathbf{w})\big) \exp\big(-\frac{1}{2}\mathbf{w}^\top \Sigma_p^{-1} \mathbf{w}\big)$$

$$\propto \exp\big(-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^\top \big(\frac{1}{\sigma_n^2} X X^\top + \Sigma_p^{-1}\big)(\mathbf{w} - \bar{\mathbf{w}})\big)$$

$$\bar{\mathbf{w}} = \sigma_n^{-2}(\sigma_n^{-2} X X^\top + \Sigma_p^{-1})^{-1} X \mathbf{y}$$

Weight-space Function-space Duality

Probabilistic model:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon,$$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Posterior over weights:

$$p(\mathbf{w}|X, \mathbf{y}) \propto \exp\big(-\frac{1}{2\sigma_n^2}(\mathbf{y} - X^\top \mathbf{w})^\top (\mathbf{y} - X^\top \mathbf{w})\big) \exp\big(-\frac{1}{2}\mathbf{w}^\top \Sigma_p^{-1} \mathbf{w}\big)$$

$$\propto \exp\big(-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^\top \big(\frac{1}{\sigma_n^2} X X^\top + \Sigma_p^{-1}\big)(\mathbf{w} - \bar{\mathbf{w}})\big)$$

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

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

Weight-space Function-space Duality

Probabilistic model: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}, \qquad y = f(\mathbf{x}) + \varepsilon$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Posterior over weights:

$$\bar{\mathbf{w}} = \sigma_n^{-2}(\sigma_n^{-2}XX^\top + \Sigma_p^{-1})^{-1}X\mathbf{y} \quad A = \sigma_n^{-2}XX^\top + \Sigma_p^{-1}$$

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

Posterior predictive distribution: **Let** $f_* \triangleq f(\mathbf{x}_*)$ at $\mathbf{x}_*$

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

Weight-space Function-space Duality

Probabilistic model: $f(\mathbf{x}) = \phi(\mathbf{x})^\top \mathbf{w}$

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p) \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n^2)$$

Posterior predictive distribution:

$$f_*|\mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}\Big(\frac{1}{\sigma_n^2}\phi(\mathbf{x}_*)^\top A^{-1}\Phi\mathbf{y},\ \phi(\mathbf{x}_*)^\top A^{-1}\phi(\mathbf{x}_*)\Big)$$

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

After application of Woodbury matrix inversion lemma:

$$\boxed{\begin{aligned} f_*|\mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}\big(&\phi_*^\top\Sigma_p\Phi(K+\sigma_n^2 I)^{-1}\mathbf{y}, \\ &\phi_*^\top\Sigma_p\phi_* - \phi_*^\top\Sigma_p\Phi(K+\sigma_n^2 I)^{-1}\Phi^\top\Sigma_p\phi_*\big)\end{aligned}}$$

$$\phi(\mathbf{x}_*) = \phi_* \qquad K = \Phi^\top\Sigma_p\Phi$$

WEIGHT-SPACE FUNCTION-SPACE DUALITY

Gaussian Processes:

Definition 2.1 *A* Gaussian process *is a collection of random variables, any finite number of which have a joint Gaussian distribution.* □

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$$

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

Bayesian linear model:

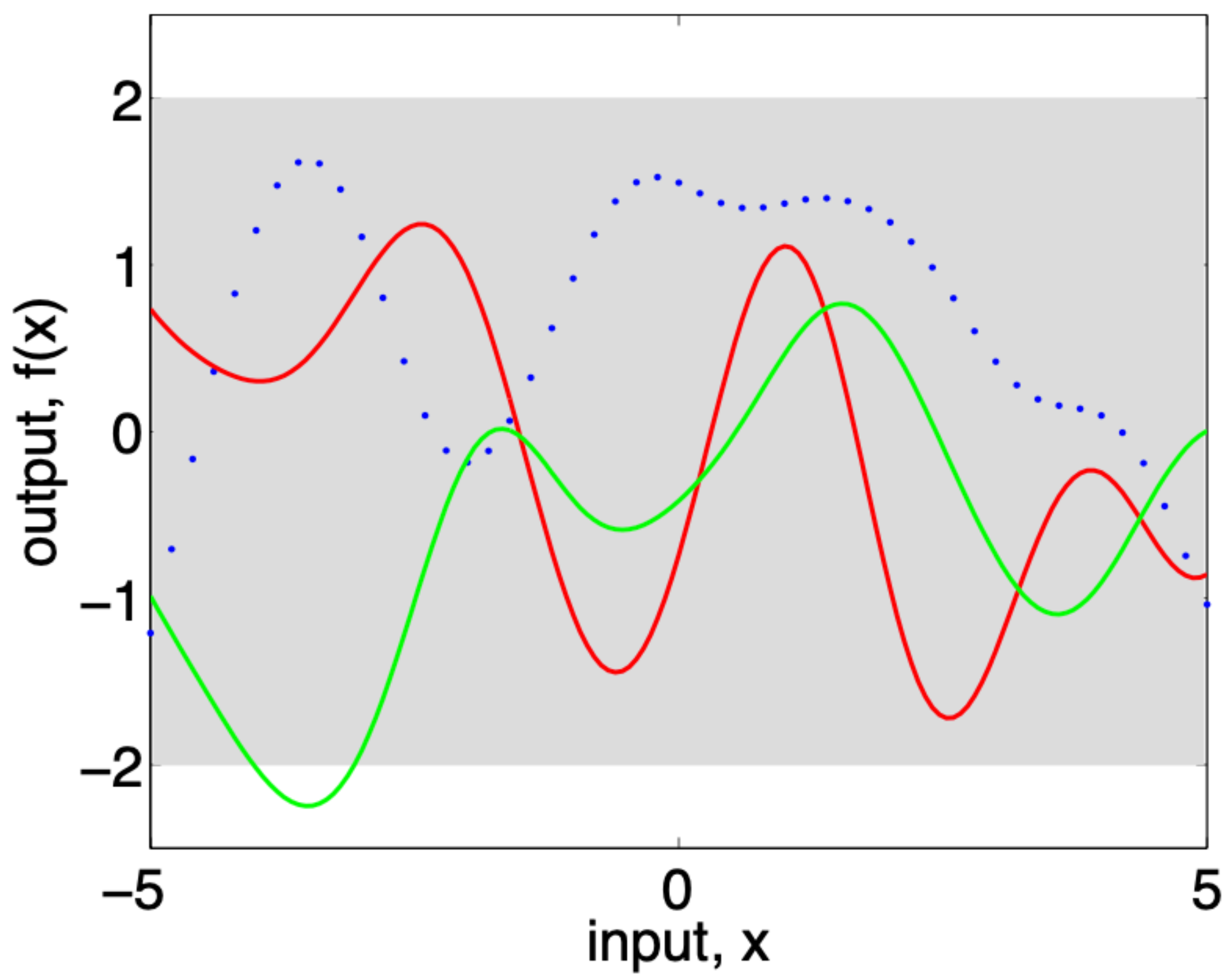
$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w} \qquad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$$

$$\begin{aligned} \mathbb{E}[f(\mathbf{x})] &= \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}[\mathbf{w}] = 0, \\ \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] &= \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}[\mathbf{w}\mathbf{w}^\top]\boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \Sigma_p \boldsymbol{\phi}(\mathbf{x}') \end{aligned}$$

WEIGHT-SPACE FUNCTION-SPACE DUALITY

Gaussian Process prior:

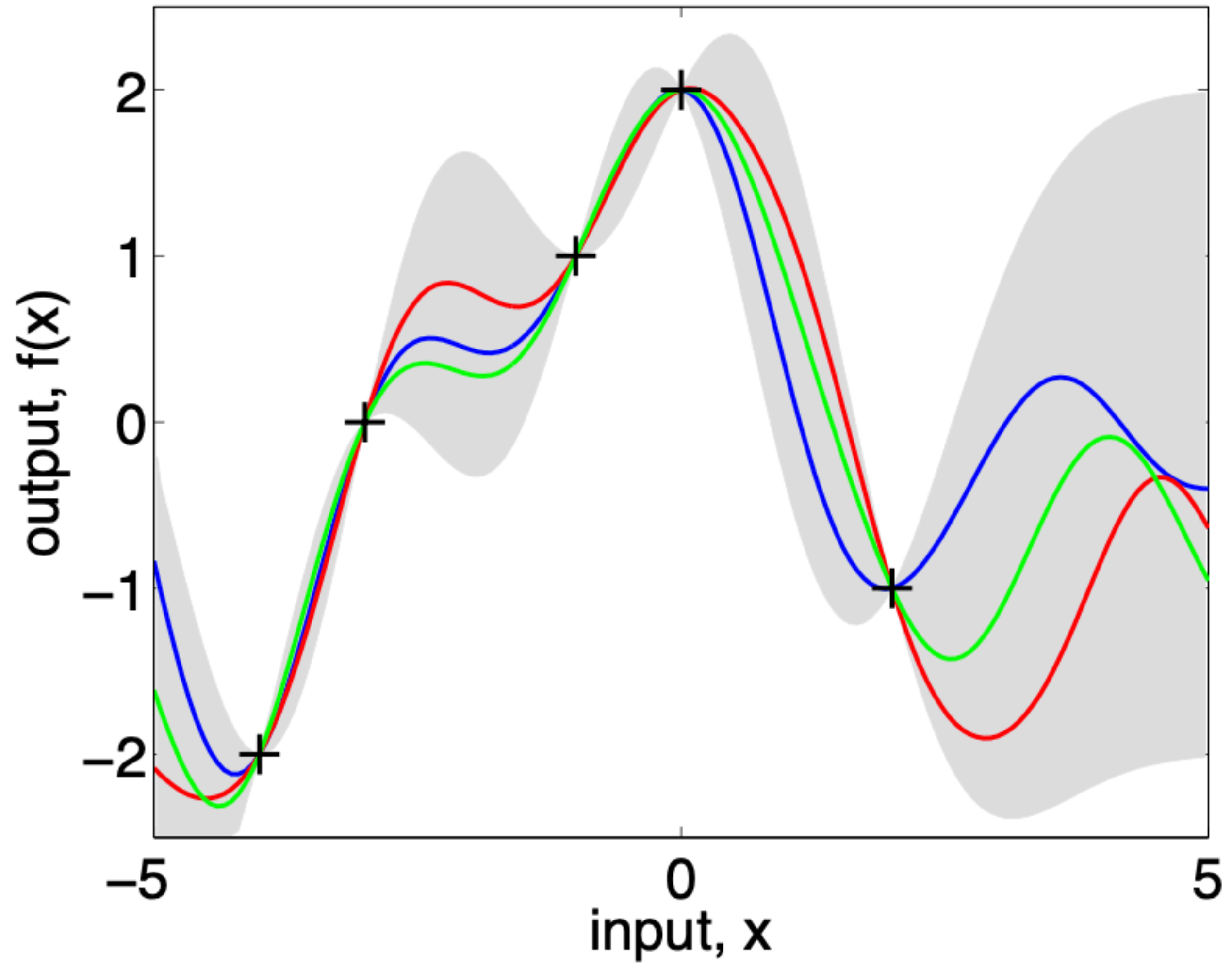
$$\mathbf{f}_* \sim \mathcal{N}(\mathbf{0}, K(X_*, X_*))$$



(a), prior

Gaussian Process joint:

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

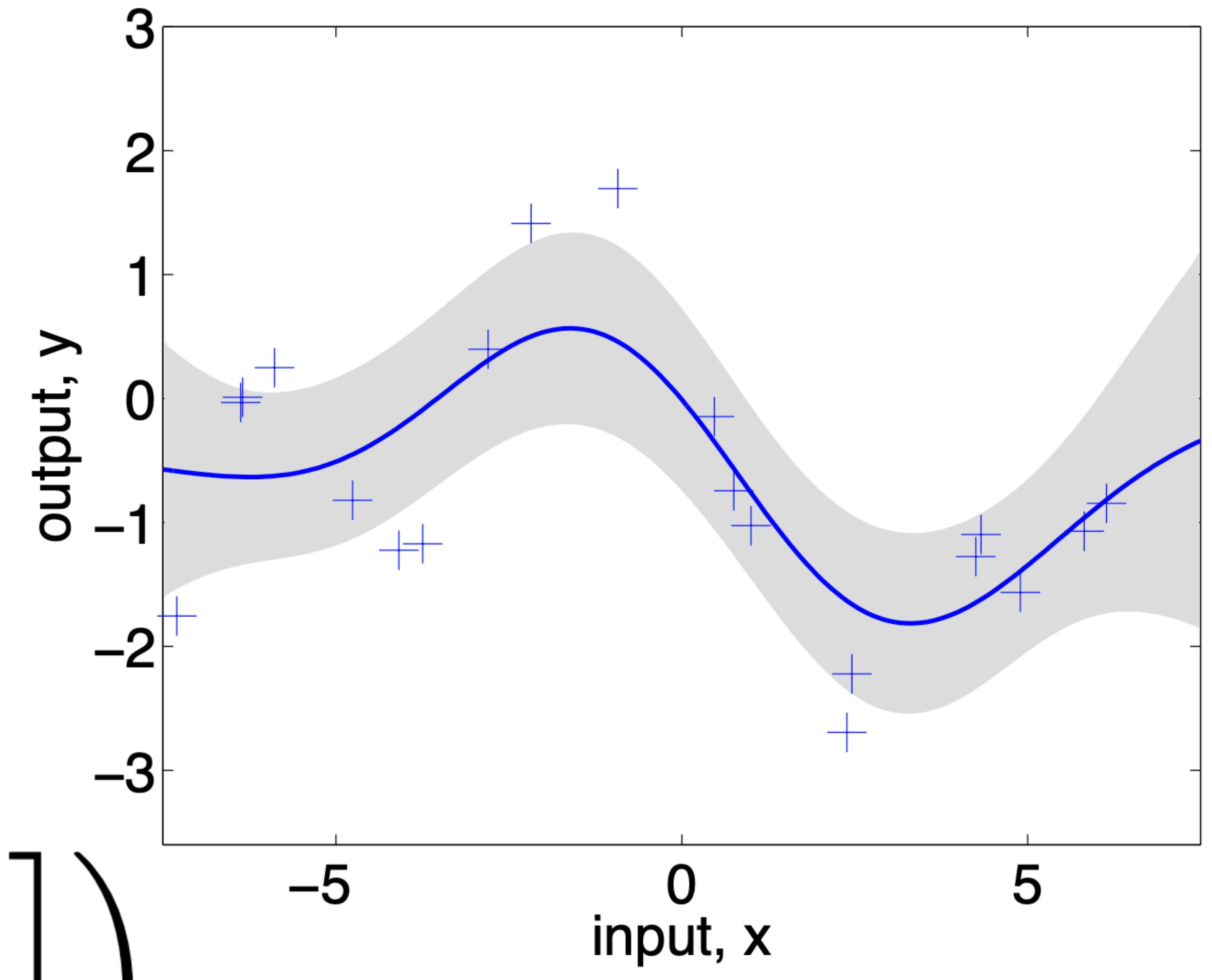


(b), posterior

Gaussian Process posterior:

$$\mathbf{f}_* | X_*, X, \mathbf{f} \sim \mathcal{N}(K(X_*,X)K(X,X)^{-1}\mathbf{f},$$
$$K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*))$$

Weight-space Function-space Duality



Gaussian Process prior:

$$\mathbf{f}_* \sim \mathcal{N}\big(\mathbf{0}, K(X_*, X_*)\big)$$

Gaussian Process joint:

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

Gaussian Process posterior:

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)\big), \quad \text{where}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X,X) + \sigma_n^2 I]^{-1}\mathbf{y},$$

$$\operatorname{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X,X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

Weight-space Function-space Duality

Gaussian Process posterior predictive distribution:

$$\mathbf{f}_* | X, \mathbf{y}, X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \ \mathrm{cov}(\mathbf{f}_*)\big), \quad \text{where}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_* | X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y},$$

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

Bayesian linear regression posterior predictive distribution:

$$f_* | \mathbf{x}_*, X, \mathbf{y} \sim \mathcal{N}\big(\phi_*^\top \Sigma_p \Phi (K + \sigma_n^2 I)^{-1}\mathbf{y},$$
$$\phi_*^\top \Sigma_p \phi_* - \phi_*^\top \Sigma_p \Phi (K + \sigma_n^2 I)^{-1} \Phi^\top \Sigma_p \phi_*\big)$$

$$\phi(\mathbf{x}_*) = \phi_* \quad K = \Phi^\top \Sigma_p \Phi$$

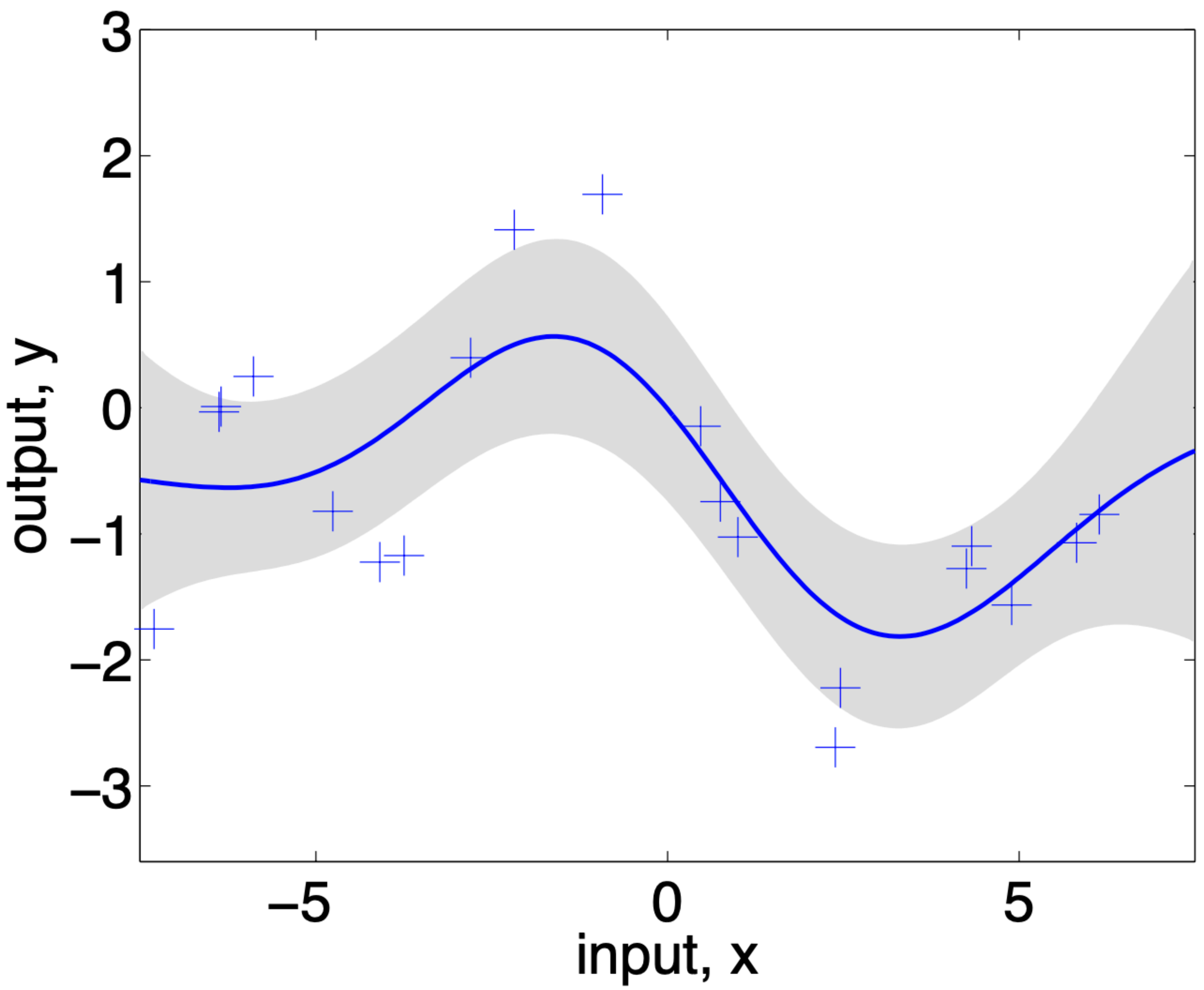
Implementing Gaussian Process Regression

Implementing Gaussian Process Regression

Gaussian Process prior:

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$$

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})],$$
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$



Gaussian Process posterior:

$$\mathbf{f}_*|X, \mathbf{y}, X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \ \mathrm{cov}(\mathbf{f}_*)\big), \quad \text{where}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_*|X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y},$$

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*)$$

Implementing Gaussian Process Regression

Marginal likelihood:

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

$$\log p(\mathbf{f}|X) = -\tfrac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} - \tfrac{1}{2}\log|K| - \tfrac{n}{2}\log 2\pi$$

$$\log p(\mathbf{y}|X) = -\tfrac{1}{2}\mathbf{y}^\top(K + \sigma_n^2 I)^{-1}\mathbf{y} - \tfrac{1}{2}\log|K + \sigma_n^2 I| - \tfrac{n}{2}\log 2\pi$$

input: X (inputs), $\mathbf{y}$ (targets), k (covariance function), σ_n^2 (noise level), $\mathbf{x}_*$ (test input)

$L := \text{cholesky}(K + \sigma_n^2 I)$

$\boldsymbol{\alpha} := L^\top \backslash (L \backslash \mathbf{y})$ } predictive mean eq. (2.25)

$\bar{f}_* := \mathbf{k}_*^\top \boldsymbol{\alpha}$

$\mathbf{v} := L \backslash \mathbf{k}_*$ } predictive variance eq. (2.26)

$\mathbb{V}[f_*] := k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^\top \mathbf{v}$

$\log p(\mathbf{y}|X) := -\tfrac{1}{2}\mathbf{y}^\top \boldsymbol{\alpha} - \sum_i \log L_{ii} - \tfrac{n}{2}\log 2\pi$ eq. (2.30)

return: $\bar{f}_*$ (mean), $\mathbb{V}[f_*]$ (variance), $\log p(\mathbf{y}|X)$ (log marginal likelihood)

Implementing Gaussian Process Regression

Implementation checklist:

1) GP prior, including kernel function

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$$

2) Posterior predictive distribution:

$$\mathbf{f}_*|X, \mathbf{y}, X_* \sim \mathcal{N}\big(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)\big), \quad \text{where}$$

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[\mathbf{f}_*|X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y},$$

$$\operatorname{cov}(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*)$$

3) Marginal likelihood:

$$\log p(\mathbf{y}|X) = -\tfrac{1}{2}\mathbf{y}^\top (K + \sigma_n^2 I)^{-1}\mathbf{y} - \tfrac{1}{2}\log|K + \sigma_n^2 I| - \tfrac{n}{2}\log 2\pi$$

LAB ASSIGNMENT

Course page: `http://www.robots.ox.ac.uk/~mosb/aims_cdt`

Goal: predict missing sensor measurements using GP regression