Data, Estimation & Inference

Gaussian Processes & Bayesian Deep Learning

Michaelmas Term, 2020

University of Oxford

Tim G. J. Rudner

tim.rudner@cs.ox.ac.uk

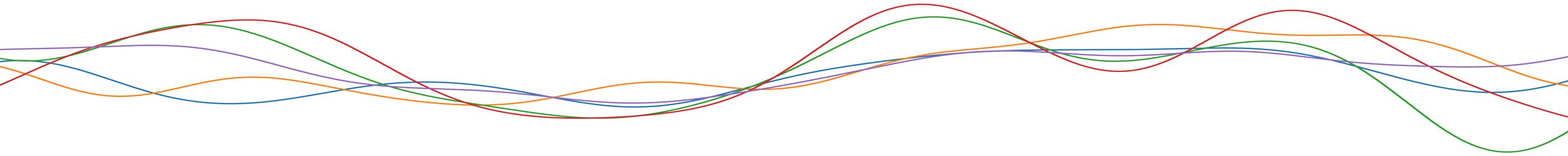
Syllabus: https://www.notion.so/Data-Estimation-and-Inference-Part-2-d177bdb310e74f9aafd034c35652e4f3

Overview

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

Today:

- Recap
- Hyperparameter optimization for GPs
- Uncertainty quantification



Recap: Gaussian Process Regression 101

Recap: Gaussian Process Regression 101

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:

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

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

$1 \times p \quad p \times p \quad p \times 1$

$\Leftrightarrow K(x, x')$

$\phi : X \to Z$

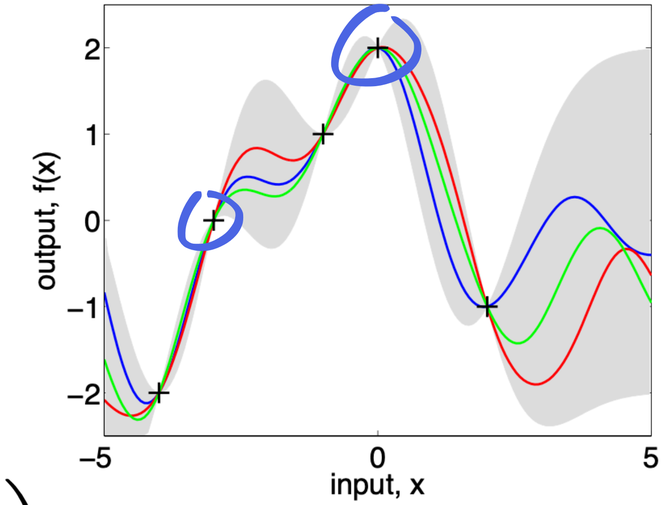
Example:

$\phi(x) \triangleq [x, x^2]^\top$

Recap: Gaussian Process Regression 101

Gaussian Process prior:

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



(b), posterior

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:

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

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

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

Recap: Gaussian Process Regression 101

duality

Gaussian Process posterior predictive distribution:

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

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

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

= 2

Bayesian linear regression posterior predictive distribution:

move on this later

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

$$\boldsymbol{\phi}_*^\top\Sigma_p\boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top\Sigma_p\Phi(K+\sigma_n^2 I)^{-1}\Phi^\top\Sigma_p\boldsymbol{\phi}_*\big)$$

= 2

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

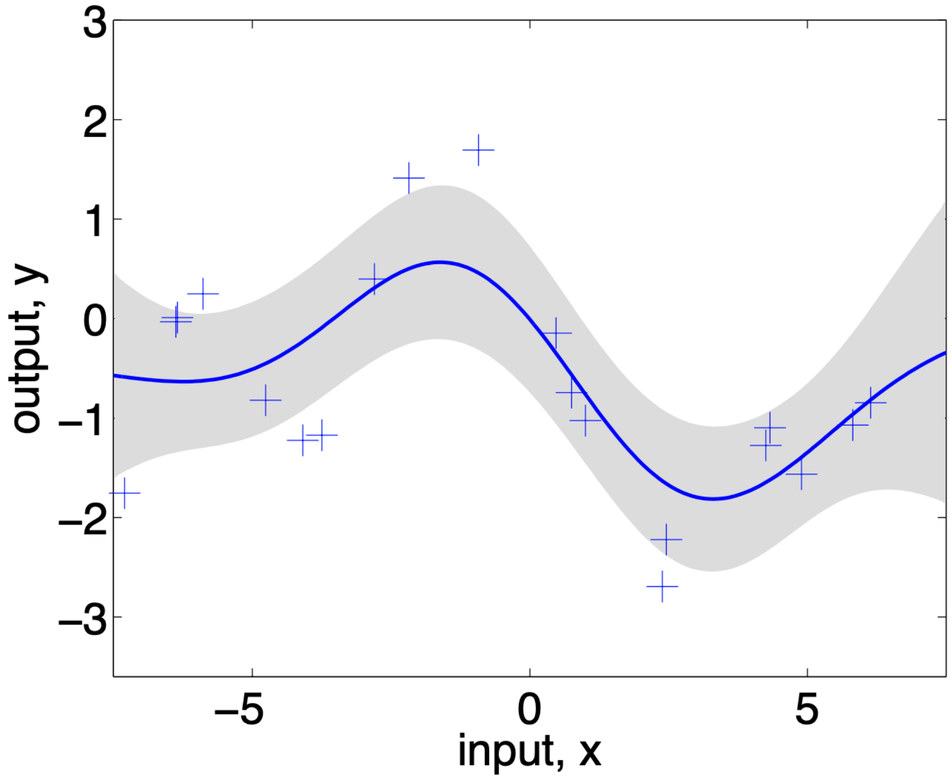
Implementing Gaussian Process Regression

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:

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

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

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

Implementing Gaussian Process Regression

Marginal likelihood:

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

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

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

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

$\boldsymbol{\alpha} := L^\top \backslash (L \backslash \mathbf{y})$

4: $\bar{f}_* := \mathbf{k}_*^\top \boldsymbol{\alpha}$ } predictive mean eq. (2.25)

$\mathbf{v} := L \backslash \mathbf{k}_*$

6: $\mathbb{V}[f_*] := k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^\top \mathbf{v}$ } predictive variance eq. (2.26)

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

8: **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:

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

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

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

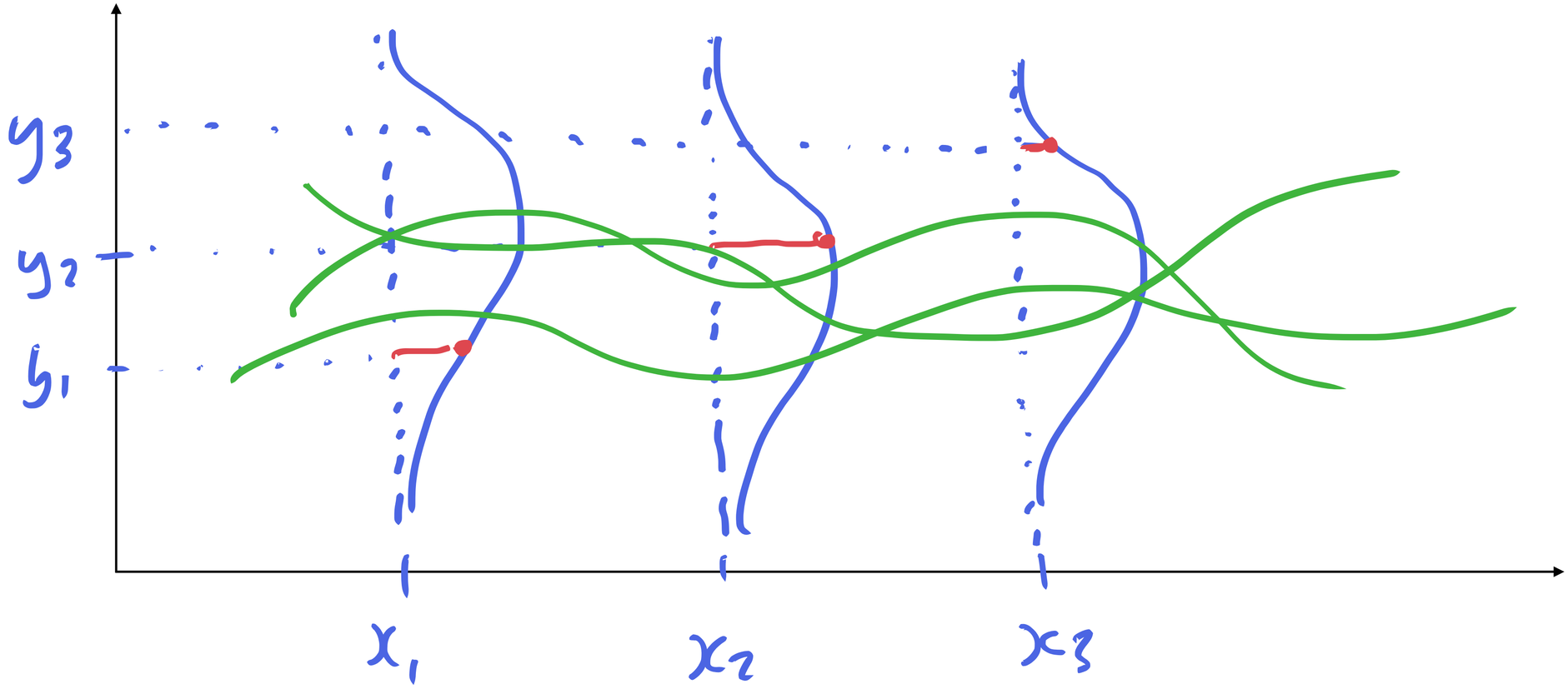
3) Marginal likelihood:

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

Implementing Gaussian Process Regression

Understanding the marginal likelihood

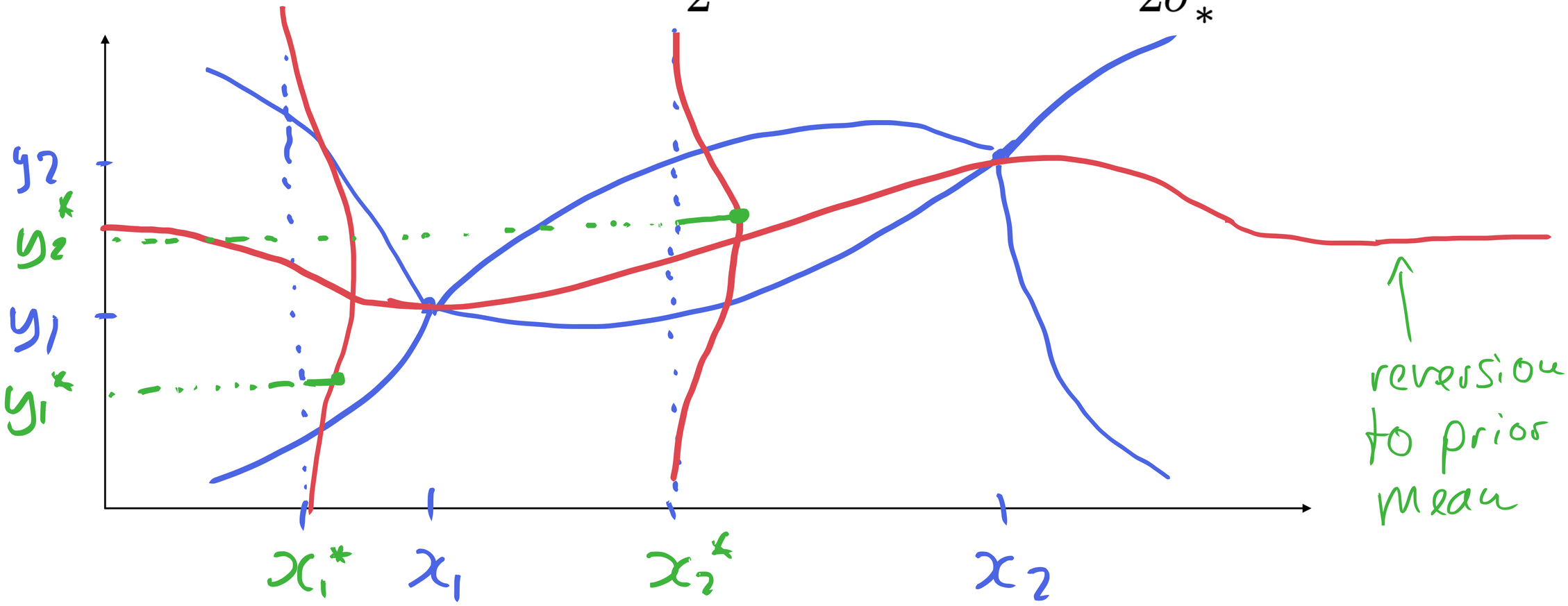
$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)\, d\mathbf{f}$$



Implementing Gaussian Process Regression

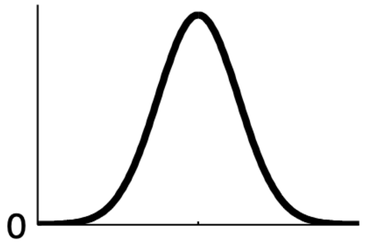
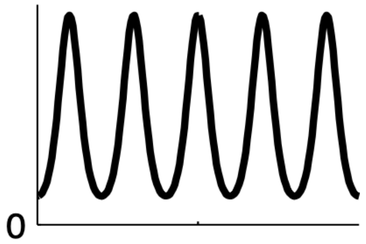
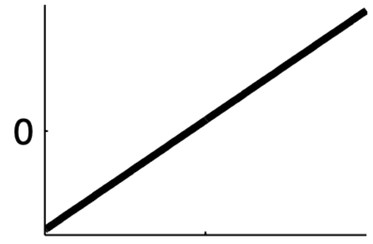
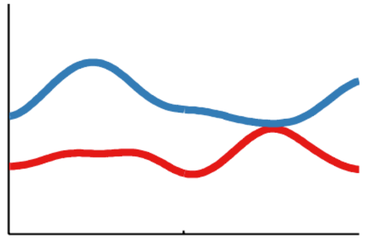
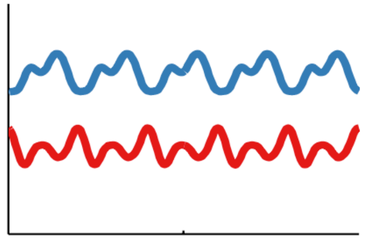
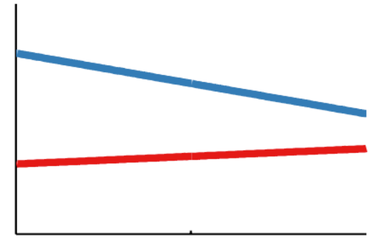
Assessing performance

$$-\log p(y_*|\mathcal{D}, \mathbf{x}_*) = \frac{1}{2}\log(2\pi\sigma_*^2) + \frac{(y_* - \bar{f}(\mathbf{x}_*))^2}{2\sigma_*^2}$$



Implementing Gaussian Process Regression

Covariance (~kernel) functions

Kernel name:	Squared-exp (SE)	Periodic (Per)	Linear (Lin)
$k(x, x') =$	$\sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$	$\sigma_f^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\pi\frac{x-x'}{p}\right)\right)$	$\sigma_f^2(x-c)(x'-c)$
Plot of $k(x, x')$:			
	$x - x'$	$x - x'$	x (with $x' = 1$)
Functions $f(x)$ sampled from GP prior:			
	x	x	x
Type of structure:	local variation	repeating structure	linear functions

$X^T X$

Bayesian linear regression

$\phi(x)^T \phi(x)$

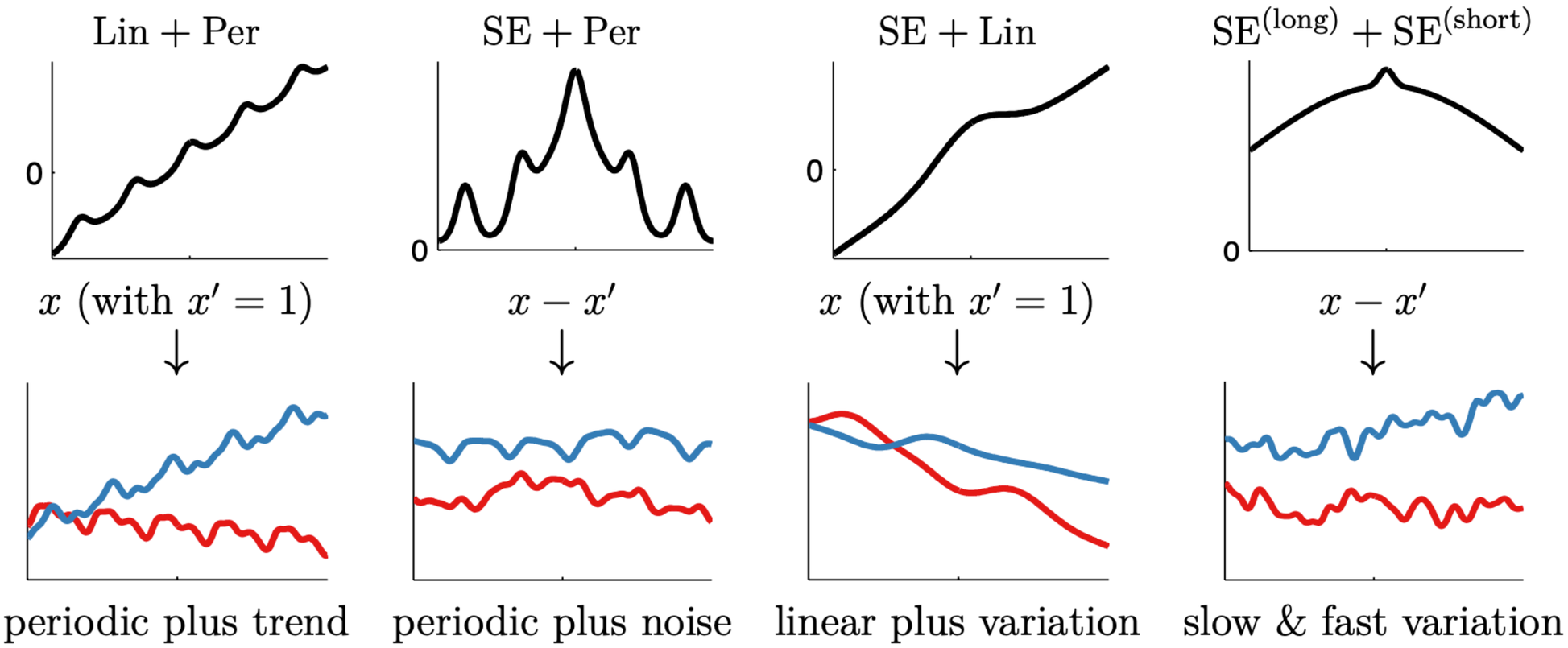
$n \times p$ $p \times 1$

Implementing Gaussian Process Regression

Combining covariance functions through addition

$$k_a + k_b = k_a(\mathbf{x}, \mathbf{x}') + k_b(\mathbf{x}, \mathbf{x}')$$

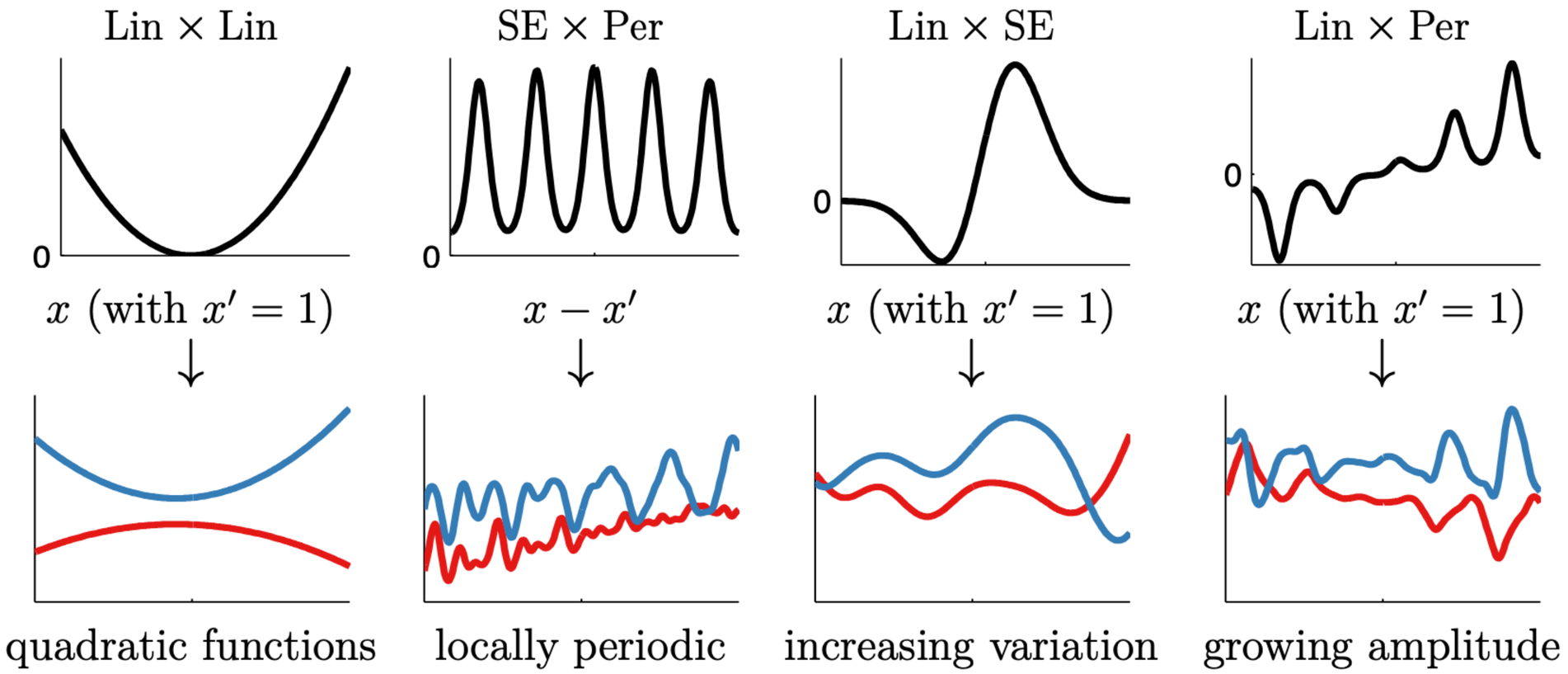
Link to visualization



Implementing Gaussian Process Regression

Combining covariance functions through multiplication

$$k_a \times k_b = k_a(\mathbf{x}, \mathbf{x}') \times k_b(\mathbf{x}, \mathbf{x}')$$

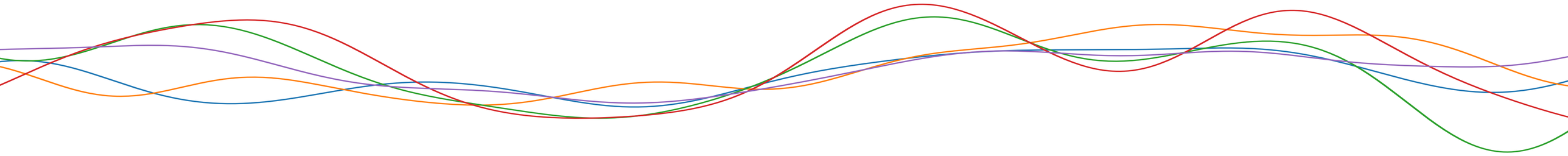


Uncertainty Quantification

Uncertainty Quantification

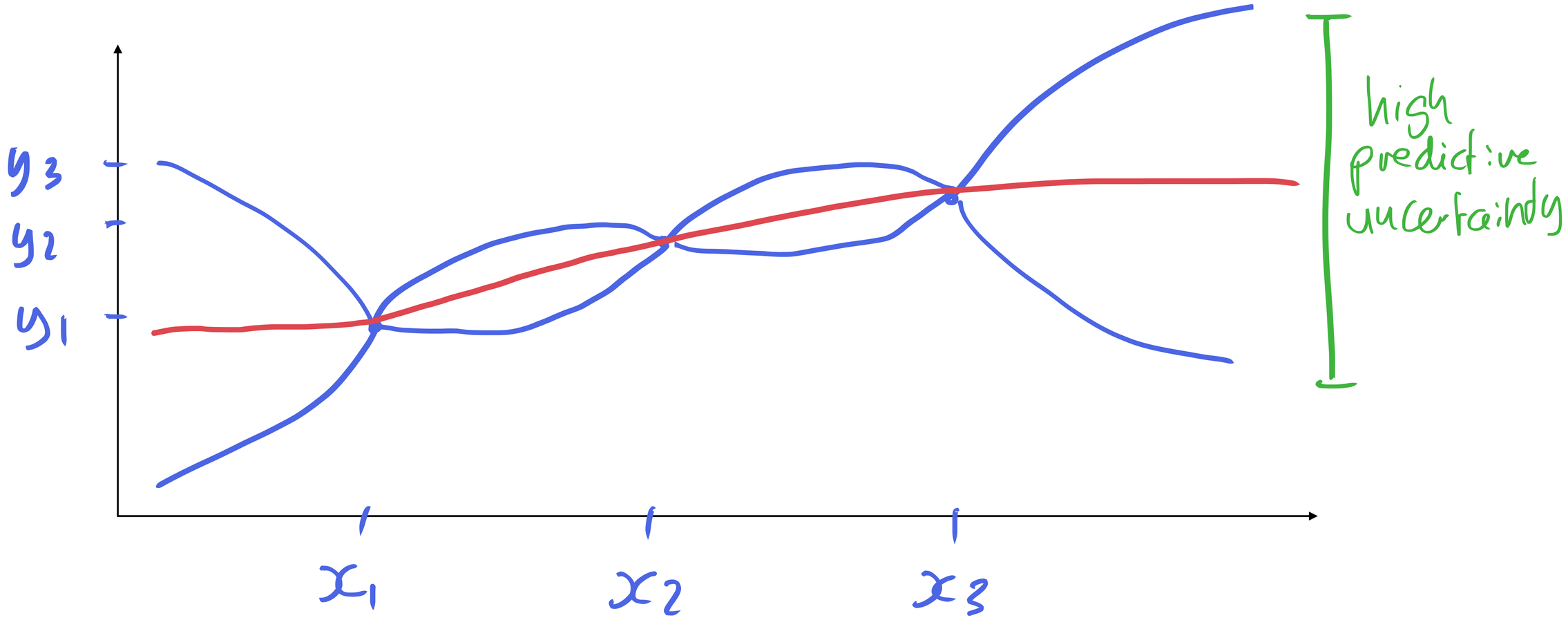
Why Bayesian uncertainty quantification?

- Know what you don’t know
- AI safety
- Efficient exploration
- Mathematically-principled foundations



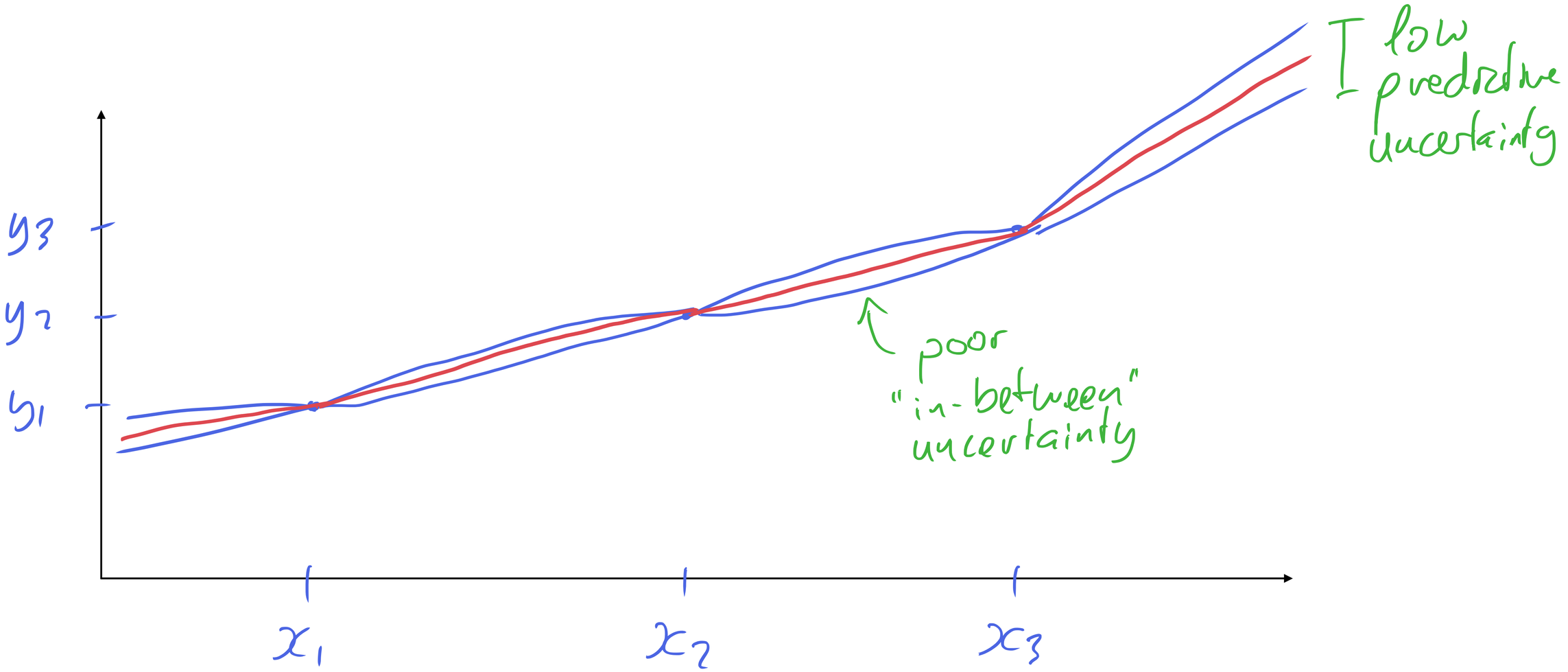
Uncertainty Quantification

What predictive uncertainty is good predictive uncertainty?



Uncertainty Quantification

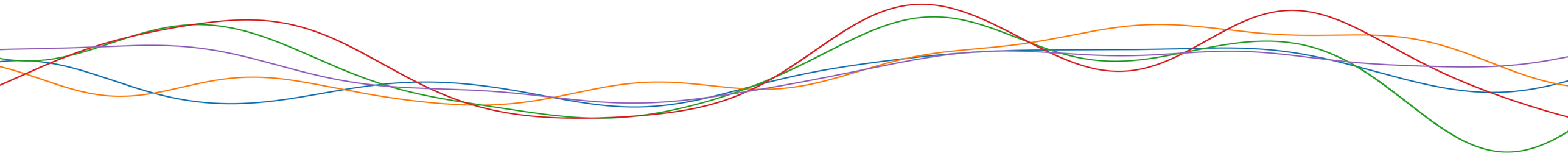
What predictive uncertainty is bad predictive uncertainty?



Uncertainty Quantification

Desirable properties in predictive uncertainty estimates

- Low predictive uncertainty close to observed data
- Reversion to the prior away from observed data
- A suitable prior predictive distribution
- Avoid confident but wrong predictions

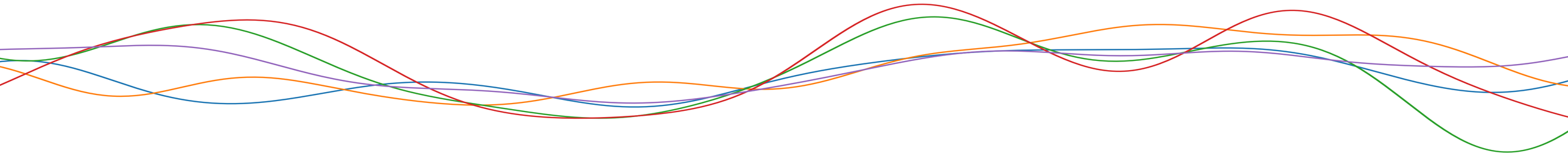


Uncertainty Quantification

What models reliably provide good predictive uncertainty estimates?

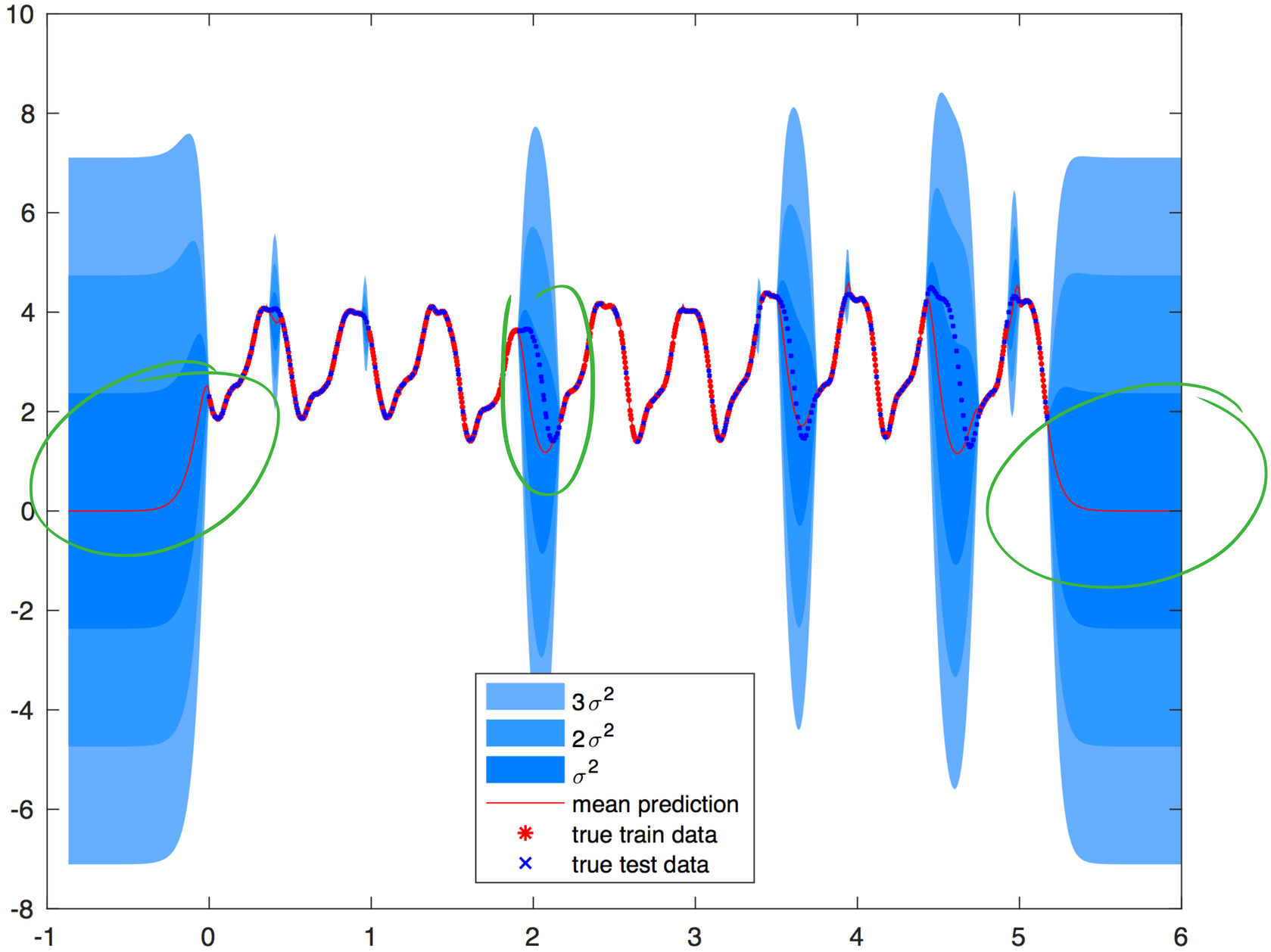
- Gaussian Processes?
- Bayesian linear models?
- Bayesian Neural Networks?

It depends!



Lab Assignment

Lab Assignment



Lab Assignment