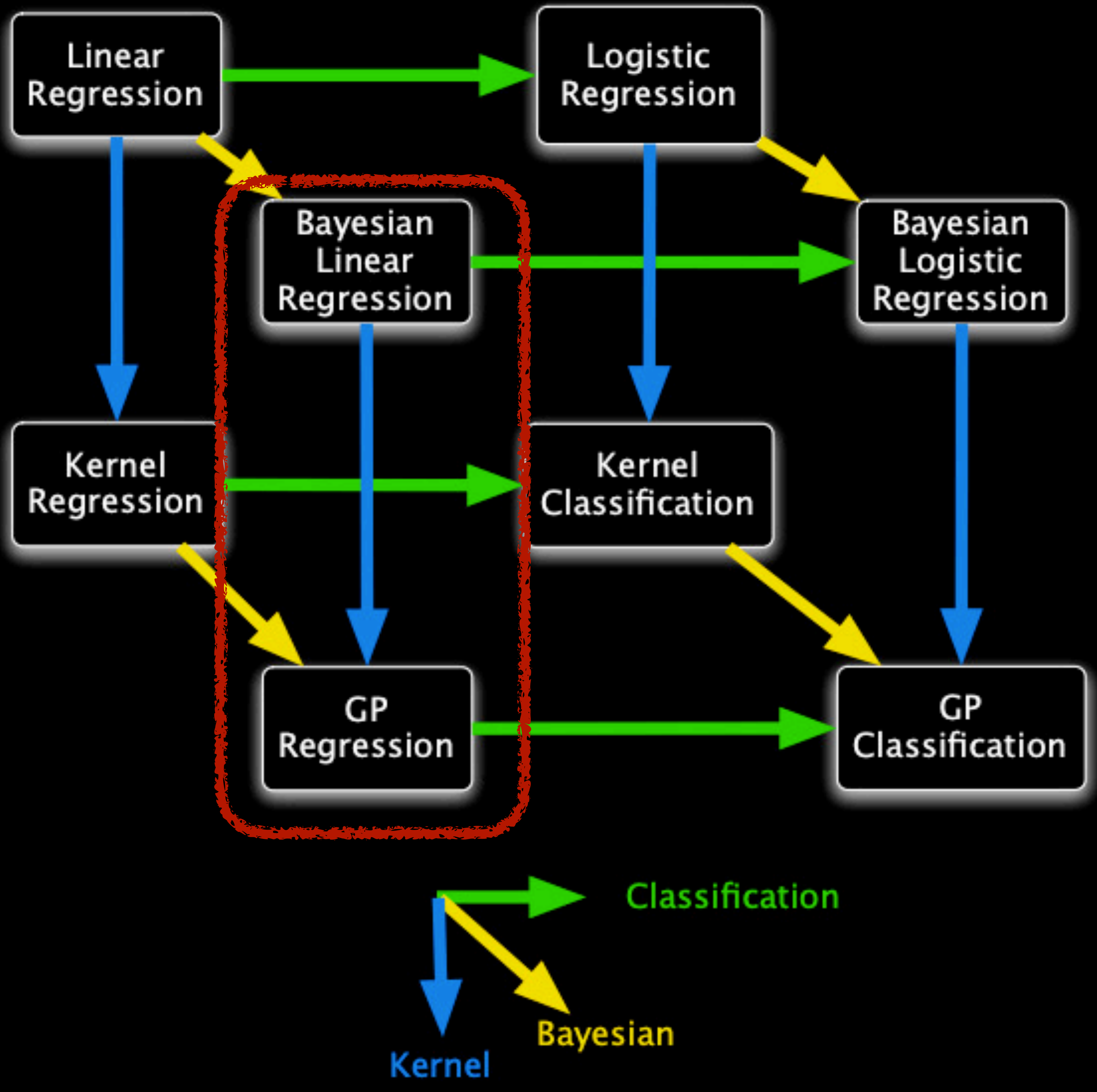
Gaussian process regression

Week 4b - Statistical ML / Thang Bui / ANU / 2025 S1 - Thanks to Yuan-Sen Ting for many slides

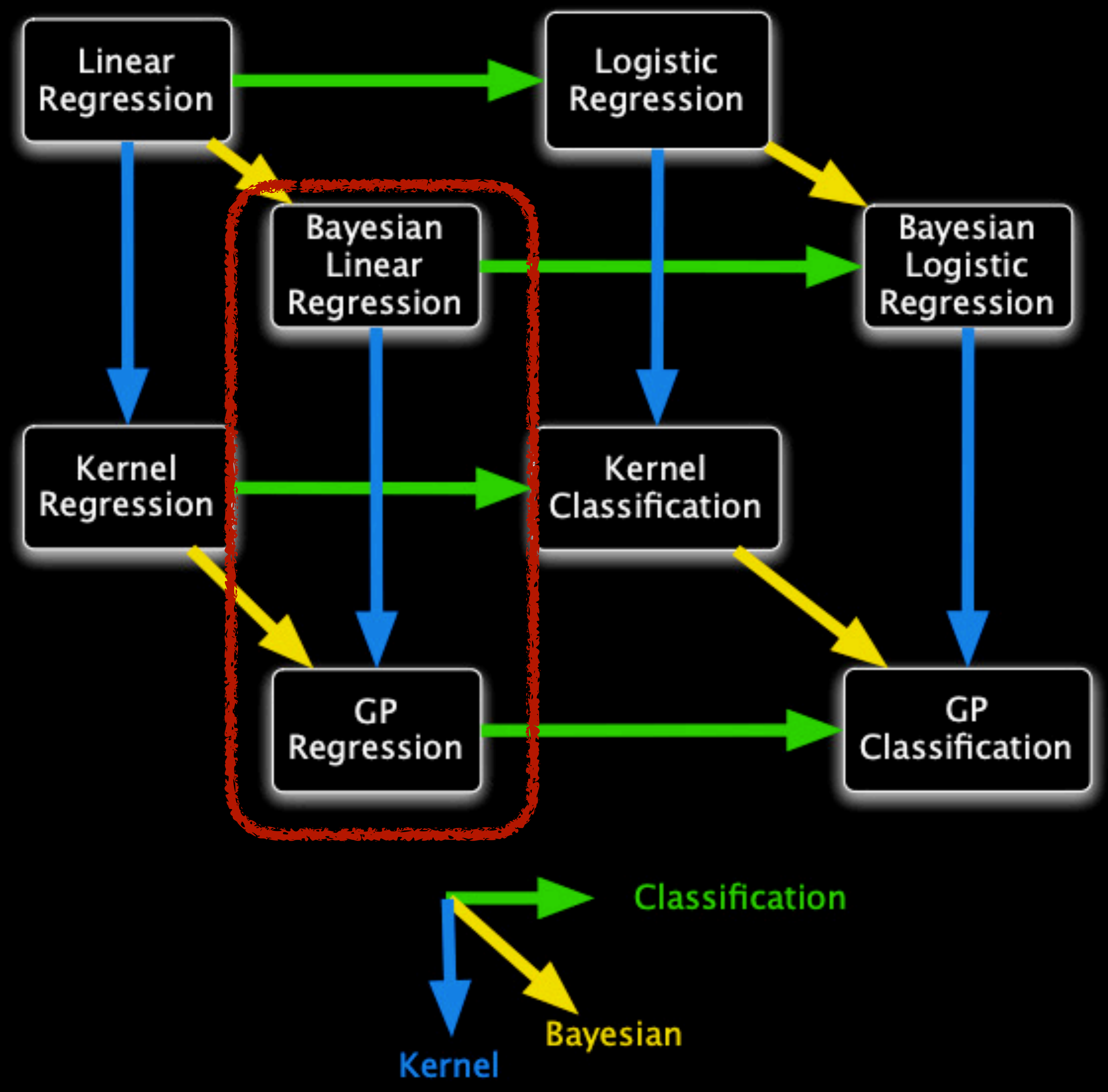
Gaussian Process - Weight-Space Perspective



Gaussian Process - Weight-Space Perspective

Gaussian Process =

Kernelising Bayesian
Linear Regression



Recap : Linear Regression

$\Phi \in \mathbb{R}^{N \times D}$: feature matrix

$\mathbf{y} \in \mathbb{R}^{N}$: target matrix

D : number of features

N : number of training data

Recap : Linear Regression

$\Phi \in \mathbb{R}^{N \times D}$: feature matrix
$\mathbf{y} \in \mathbb{R}^{N}$: target matrix

D : number of features
N : number of training data

Week 2a

Features

$$f(\mathbf{x}) = \underbrace{\phi(\mathbf{x})^{\mathrm{T}}}_{\mathbb{R}^D} \underbrace{\theta^*}_{\mathbb{R}^D}$$

Recap : Linear Regression

$\Phi \in \mathbb{R}^{N \times D}$: feature matrix
$\mathbf{y} \in \mathbb{R}^{N}$: target matrix

D : number of features
N : number of training data

Week 2a

$$f(\mathbf{x}) = \overbrace{\underbrace{\phi(\mathbf{x})^{\mathrm{T}}}_{\mathbb{R}^D}}^{\text{Features}} \underbrace{\theta^*}_{\mathbb{R}^D}$$

$$\theta^* = (\underbrace{\lambda \mathbf{I_D}}_{\text{Regularisation}} + \underbrace{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}}_{\mathbb{R}^{D \times D}})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Recap : Linear Regression

$\Phi \in \mathbb{R}^{N\times D}$: feature matrix
$\mathbf{y} \in \mathbb{R}^{N}$: target matrix

D : number of features
N : number of training data

Week 2a

$$f(\mathbf{x}) = \underbrace{\phi(\mathbf{x})^{\mathrm{T}}}_{\mathbb{R}^D} \underbrace{\theta^*}_{\mathbb{R}^D}$$

Features

$$\theta^* = (\underbrace{\lambda \mathbf{I_D}}_{\text{Regularisation}} + \underbrace{\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}}_{\mathbb{R}^{D\times D}})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}$$

Kernelised

Week 4a

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

Recap : Linear Regression

$\Phi \in \mathbb{R}^{N\times D}$: feature matrix
$\mathbf{y} \in \mathbb{R}^{N}$: target matrix

D : number of features
N : number of training data

Week 2a

$$f(\mathbf{x}) = \overbrace{\underbrace{\phi(\mathbf{x})^{\mathrm{T}}}_{\mathbb{R}^D}}^{\text{Features}} \underbrace{\theta^*}_{\mathbb{R}^D}$$

$$\theta^* = (\underbrace{\lambda \mathbf{I_D}}_{\text{Regularisation}} + \underbrace{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}}_{\mathbb{R}^{D\times D}})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Kernelised $\Rightarrow$

Week 4a

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

$$= k(\mathbf{x})^T(\lambda \mathbf{I_N} + \overbrace{\underbrace{\mathbf{K}}_{\mathbb{R}^{N\times N}}}^{\text{Kernel}})^{-1}\mathbf{y}$$

(Bishop eq 6.9)

Gram matrix (positive semi-definite)

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n) = \mathbf{\Phi}\,\mathbf{\Phi}^{\mathrm{T}}$$

(Bishop eq 6.10, 6.11)

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g.,
$$k(\mathbf{x_m}, \mathbf{x_n}) = \mathbf{x_m^T x_n}$$

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g., $$k(\mathbf{x_m}, \mathbf{x_n}) = \mathbf{x_m^T x_n} \quad \Rightarrow \quad \phi(\mathbf{x}) = \mathbf{x}$$

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = \mathbf{x_m^T x_n}$ ➡ $\phi(\mathbf{x}) = \mathbf{x}$

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = (\mathbf{x_m^T x_n})^2$

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = \mathbf{x_m^T x_n}$ ➡ $\phi(\mathbf{x}) = \mathbf{x}$

In 2D

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = (\mathbf{x_m^T x_n})^2$ ➡ $\phi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$

(Bishop eq 6.12)

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = \mathbf{x_m^T x_n}$ $\Rightarrow$ $\phi(\mathbf{x}) = \mathbf{x}$

In 2D

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = (\mathbf{x_m^T x_n})^2$ $\Rightarrow$ $\phi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$

(Bishop eq 6.12)

e.g., $$k(\mathbf{x_m}, \mathbf{x_n}) = \exp\left(-\frac{\|\mathbf{x_n} - \mathbf{x_m}\|_2^2}{2\sigma^2}\right)$$

(Bishop eq 6.23, GP Book eq 2.16)

Why Kernel ?

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = \mathbf{x_m^T x_n}$ $\Rightarrow$ $\phi(\mathbf{x}) = \mathbf{x}$

In 2D

e.g., $k(\mathbf{x_m}, \mathbf{x_n}) = (\mathbf{x_m^T x_n})^2$ $\Rightarrow$ $\phi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$

(Bishop eq 6.12)

e.g., $$k(\mathbf{x_m}, \mathbf{x_n}) = \exp\left(-\frac{\|\mathbf{x_n} - \mathbf{x_m}\|_2^2}{2\sigma^2}\right)$$

(Bishop eq 6.23, GP Book eq 2.16)

$\Rightarrow$ Infinite dimensional features

Why Kernel ?

Simplifying the process of coming up with “features”

$$K_{mn} \equiv k(\mathbf{x}_m, \mathbf{x}_n)$$

e.g., $k(\mathbf{x}_\mathrm{m}, \mathbf{x}_\mathrm{n}) = \mathbf{x}_\mathrm{m}^\mathrm{T}\mathbf{x}_\mathrm{n}$ ➡ $\phi(\mathbf{x}) = \mathbf{x}$

In 2D

e.g., $k(\mathbf{x}_\mathrm{m}, \mathbf{x}_\mathrm{n}) = (\mathbf{x}_\mathrm{m}^\mathrm{T}\mathbf{x}_\mathrm{n})^2$ ➡ $\phi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$

(Bishop eq 6.12)

e.g., $k(\mathbf{x}_\mathrm{m}, \mathbf{x}_\mathrm{n}) = \exp\left(-\frac{\|\mathbf{x}_\mathrm{n} - \mathbf{x}_\mathrm{m}\|_2^2}{2\sigma^2}\right)$

(Bishop eq 6.23, GP Book eq 2.16)

➡ Infinite dimensional features

Recap : Bayesian Linear Regression

Features

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

$$\theta^* = (\lambda \mathbf{I}_{\mathrm{D}} + \underbrace{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}}_{\mathbb{R}^{D\times D}})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Week 2a

Recap : Bayesian Linear Regression

Features

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

$$\theta^* = (\underbrace{\lambda \mathbf{I}_{\mathrm{D}} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}}_{\mathbb{R}^{D\times D}})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Week 2a

Week 2b

Posterior :

$$p(\theta\,|\,\underbrace{\mathbf{X}, \mathbf{y}}_{\text{Training data}})\ ?$$

→

Bayesian

Recap : Bayesian Linear Regression

Features

$f(\mathbf{x}) = \phi^{\mathbf{T}}(\mathbf{x})\,\theta^*$

$\theta^* = (\lambda \mathbf{I_D} + \underbrace{\mathbf{\Phi^T\Phi}}_{\mathbb{R}^{D\times D}})^{-1}\mathbf{\Phi^T y}$

Week 2a

Week 2b

Posterior :

$$p(\theta\,|\,\underbrace{\mathbf{X}, \mathbf{y}}_{\text{Training data}})\ ?$$

Likelihood :

Bayesian

$$p(y\,|\,\mathbf{X}, \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n; \theta^T \phi(\mathbf{x_n}), \sigma^2)$$

(Bishop eq 3.10, GP book eq 2.3)

Recap : Bayesian Linear Regression

Features

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

$$\theta^* = (\lambda \mathbf{I_D} + \underbrace{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}}_{\mathbb{R}^{D\times D}})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Week 2a

Bayesian

Week 2b

Posterior :

$$p(\theta \,|\, \underbrace{\mathbf{X}, \mathbf{y}}_{\text{Training data}}) \; ?$$

Likelihood :

$$p(y \,|\, \mathbf{X}, \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n; \theta^T \phi(\mathbf{x_n}), \sigma^2)$$

(Bishop eq 3.10, GP book eq 2.3)

(Conjugate) Prior :

$$p(\theta) = \mathcal{N}(\theta; 0, \sigma_0^2 \mathbf{I_D})$$

(Bishop eq 3.52, GP book eq 2.4)

Recap : Bayesian Linear Regression

Features

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

$$\theta^* = (\lambda \mathbf{I_D} + \underbrace{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}}_{\mathbb{R}^{D\times D}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Week 2a

Bayesian version

Week 2b

Posterior :

$$p(\theta \,|\, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\theta; \mu, \mathbf{\Sigma})$$

(Bishop eq 3.49)

$$\Sigma^{-1} = \sigma_0^{-2}\mathbf{I_D} + \sigma^{-2}\,\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$$

(Bishop eq 3.54, GP book eq 2.7, 2.8)

$$\mu = \sigma^{-2}\,\Sigma\,\mathbf{\Phi}^{\mathrm{T}}\,\mathbf{y}$$

$$= \sigma^{-2}\,(\sigma_0^{-2}\mathbf{I_m} + \sigma^{-2}\,\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\,\mathbf{\Phi}^{\mathrm{T}}\,\mathbf{y}$$

(Bishop eq 3.53, GP book eq 2.7, 2.8)

Recap : Bayesian Linear Regression

Features

$$f(\mathbf{x}) = \phi^{\mathrm{T}}(\mathbf{x})\,\theta^*$$

$$\theta^* = (\lambda \underbrace{\mathbf{I_D} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}}_{\mathbb{R}^{D\times D}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}$$

Week 2a

Bayesian version

Week 2b

Posterior :

$$p(\theta\,|\,\mathbf{X},\mathbf{y}) = \mathcal{N}(\theta;\mu,\mathbf{\Sigma})$$

(Bishop eq 3.49)

$$\Sigma^{-1} = \sigma_0^{-2}\mathbf{I_D} + \sigma^{-2}\,\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$$

(Bishop eq 3.54, GP book eq 2.7, 2.8)

$$\mu = \sigma^{-2}\,\Sigma\,\mathbf{\Phi}^{\mathrm{T}}\,\mathbf{y}$$

$$= \sigma^{-2}\,(\sigma_0^{-2}\mathbf{I_m} + \sigma^{-2}\,\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\,\mathbf{\Phi}^{\mathrm{T}}\,\mathbf{y}$$

(Bishop eq 3.53, GP book eq 2.7, 2.8)

Recap : Bayesian Linear Regression

Predictive distribution of $y*$, given a new data $\mathbf{x}*$

$$\boxed{p(y^* \mid \mathbf{x}^*, \mathbf{X}, y)} = \int p(y^* \mid \mathbf{x}^*, \theta)\, p(\theta \mid \mathbf{X}, y)\, \mathrm{d}\theta$$

Recap : Bayesian Linear Regression

Predictive distribution of y^*, given a new data $\mathbf{x}^*$

$$\boxed{p(y^* \mid \mathbf{x}^*, \mathbf{X}, y)} = \int \overbrace{p(y^* \mid \mathbf{x}^*, \theta)}^{\text{Gaussian}} \, \overbrace{p(\theta \mid \mathbf{X}, y)}^{\text{Gaussian}} \, \mathrm{d}\theta$$

Recap : Bayesian Linear Regression

Predictive distribution of y^*, given a new data $\mathbf{x}^*$

$$\boxed{p(y^* \mid \mathbf{x}^*, \mathbf{X}, y)} = \int \overbrace{p(y^* \mid \mathbf{x}^*, \theta)}^{\text{Gaussian}} \, \overbrace{p(\theta \mid \mathbf{X}, y)}^{\text{Gaussian}} \, \mathrm{d}\theta$$

$$= \mathcal{N}(y^*; \mu^T \phi(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

(Bishop eq 3.58, GP book eq 2.9)

$$\sigma^2(\mathbf{x}^*) = \underbrace{\sigma^2}_{\text{Observation noise}} + \phi(\mathbf{x}^*)^{\mathrm{T}} \boldsymbol{\Sigma} \, \phi(\mathbf{x}^*)$$

(Bishop eq 3.59, GP book eq 2.9)

Recap : Bayesian Linear Regression

Predictive distribution of y^*, given a new data $\mathbf{x}^*$

$$\boxed{p(y^* \mid \mathbf{x}^*, \mathbf{X}, y)} = \int \overbrace{p(y^* \mid \mathbf{x}^*, \theta)}^{\text{Gaussian}} \, \overbrace{p(\theta \mid \mathbf{X}, y)}^{\text{Gaussian}} \, \mathrm{d}\theta$$

$$= \mathcal{N}(y^*; \mu^T \phi(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

(Bishop eq 3.58, GP book eq 2.9)

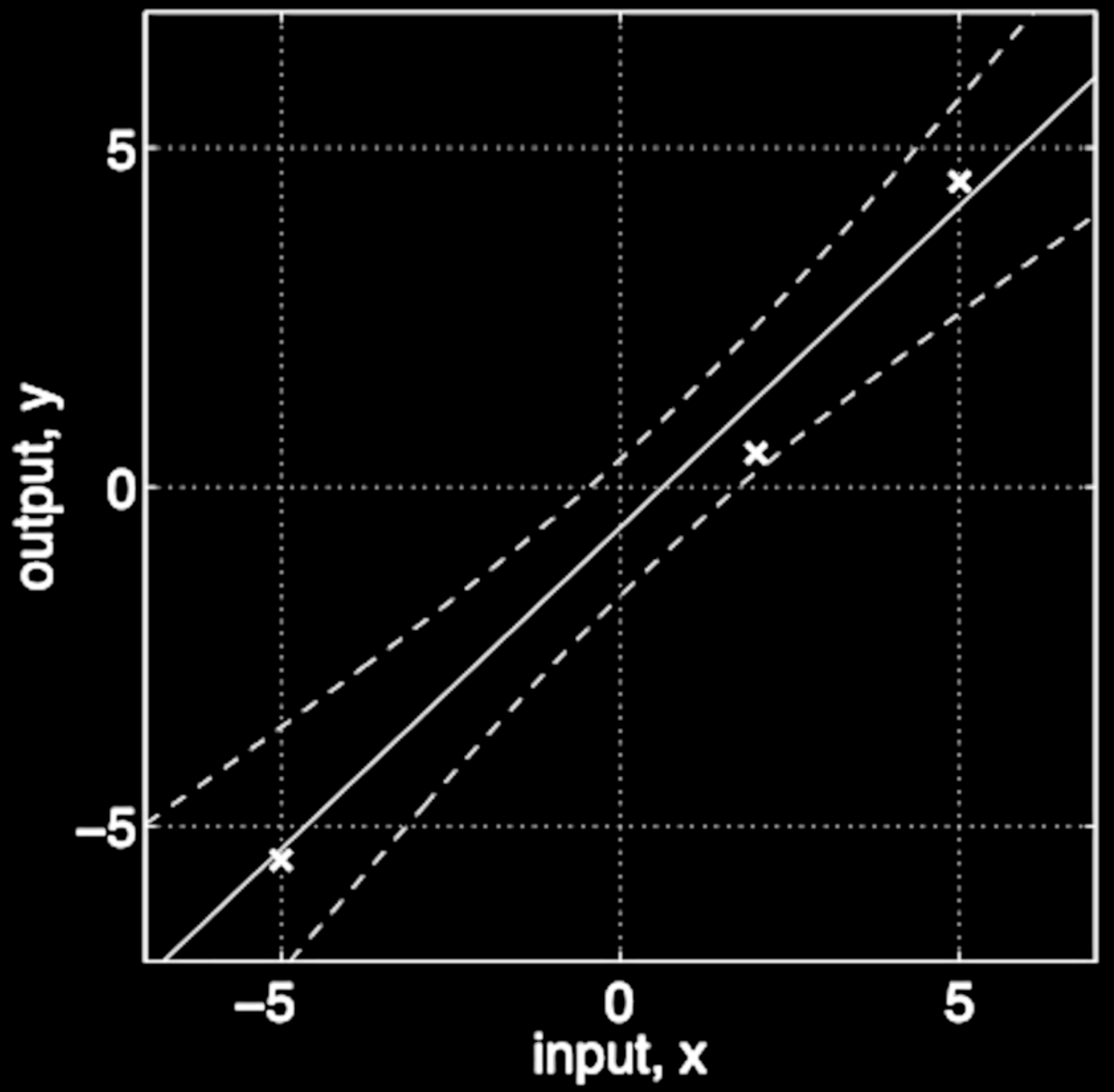
$$\sigma^2(\mathbf{x}^*) = \underbrace{\sigma^2}_{\text{Observation noise}} + \underbrace{\phi(\mathbf{x}^*)^{\mathrm{T}} \boldsymbol{\Sigma} \, \phi(\mathbf{x}^*)}_{\text{Uncertainty due to } \theta}$$

(Bishop eq 3.59,
GP book eq 2.9)

Why Full Bayesian Treatment Can Be Advantageous

Predictive distribution of y^*, given a new data $\mathbf{x}^*$

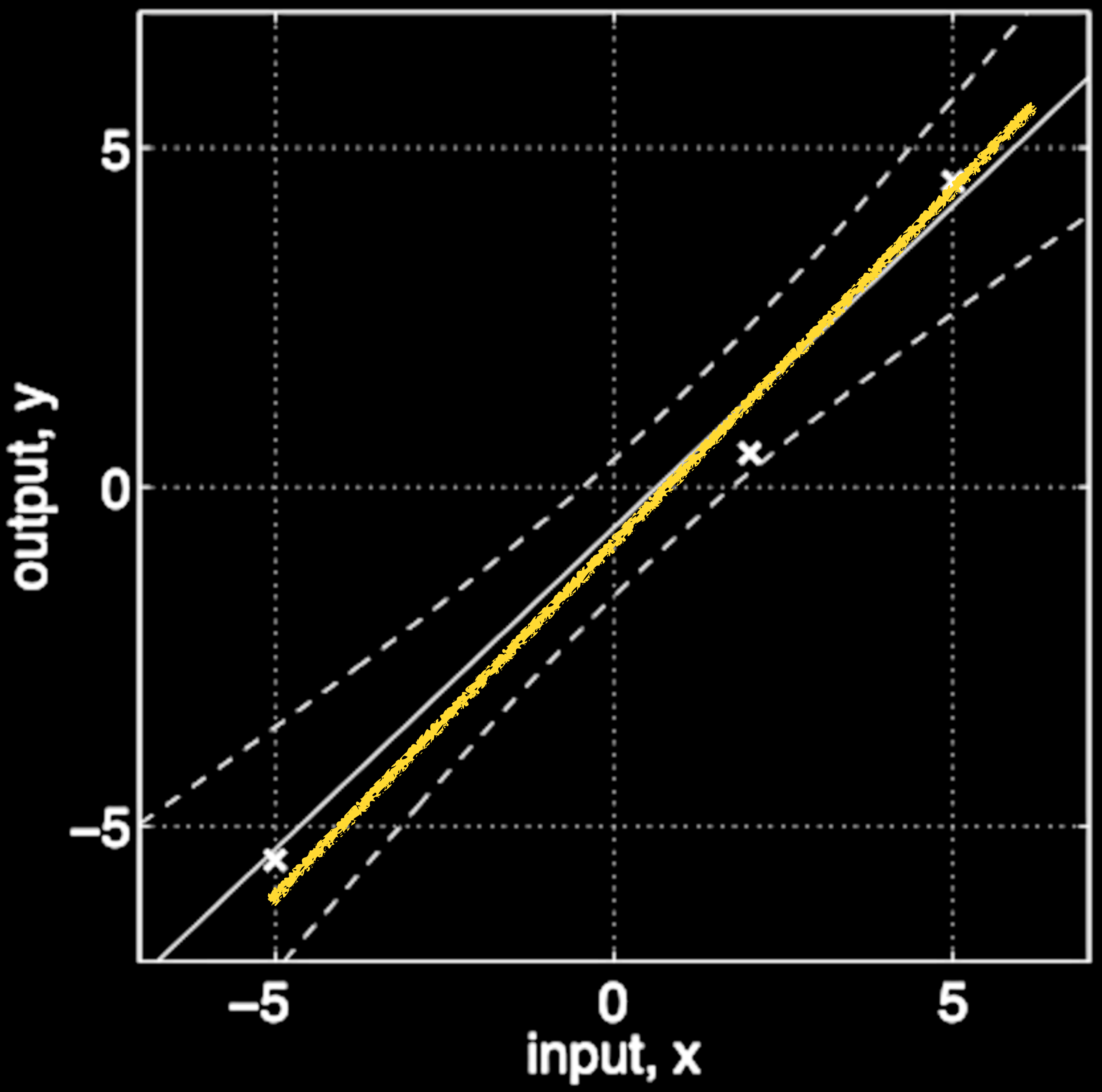
$$p(y^* \,|\, x^*, \mathbf{X}, \mathbf{y})$$



Why Full Bayesian Treatment Can Be Advantageous

Predictive distribution of $y*$, given a new data $\mathbf{x}*$

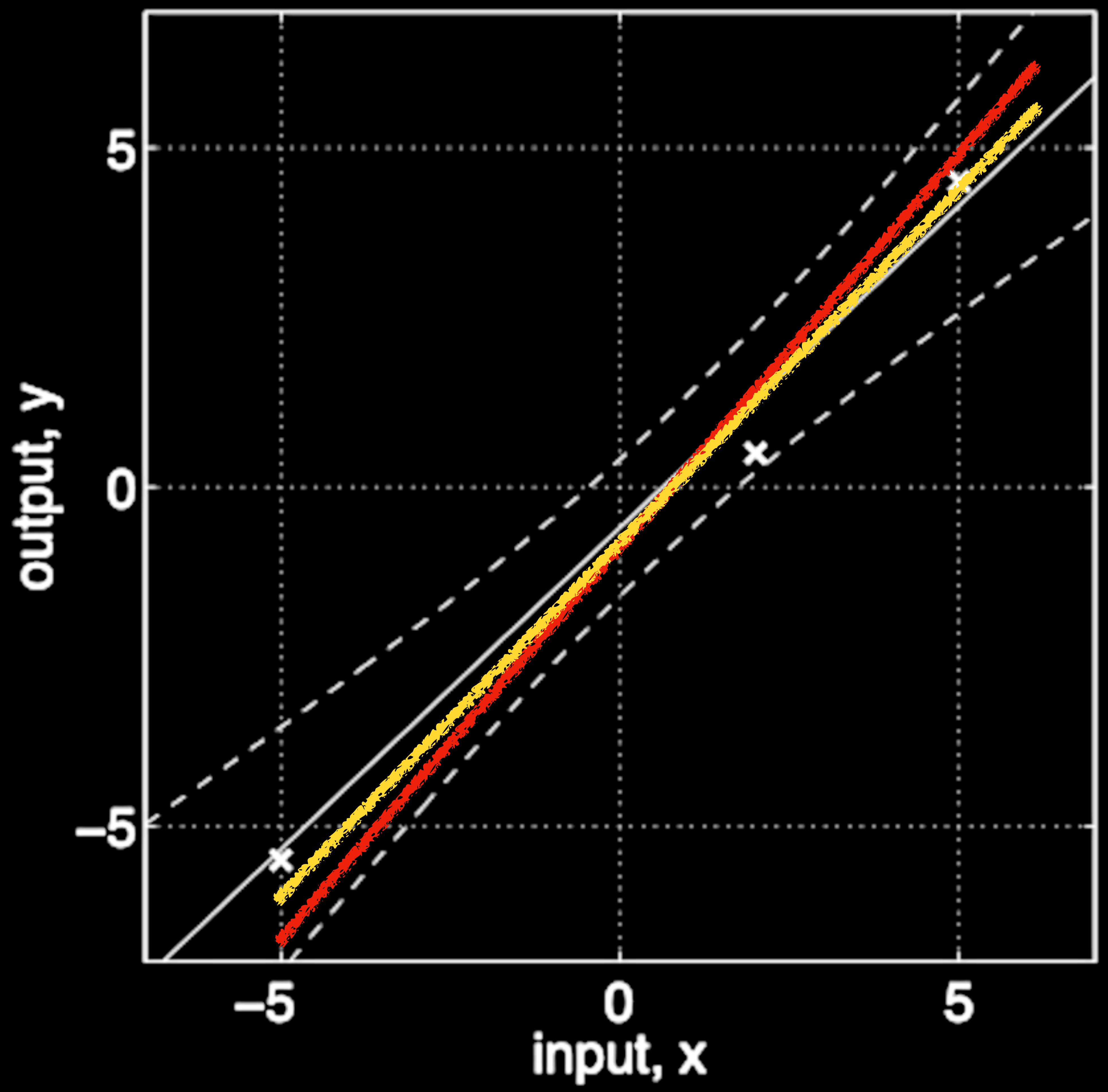
$$p(y* \,|\, x*, \mathbf{X}, \mathbf{y})$$



Why Full Bayesian Treatment Can Be Advantageous

Predictive distribution of $y*$, given a new data $\mathbf{x}*$

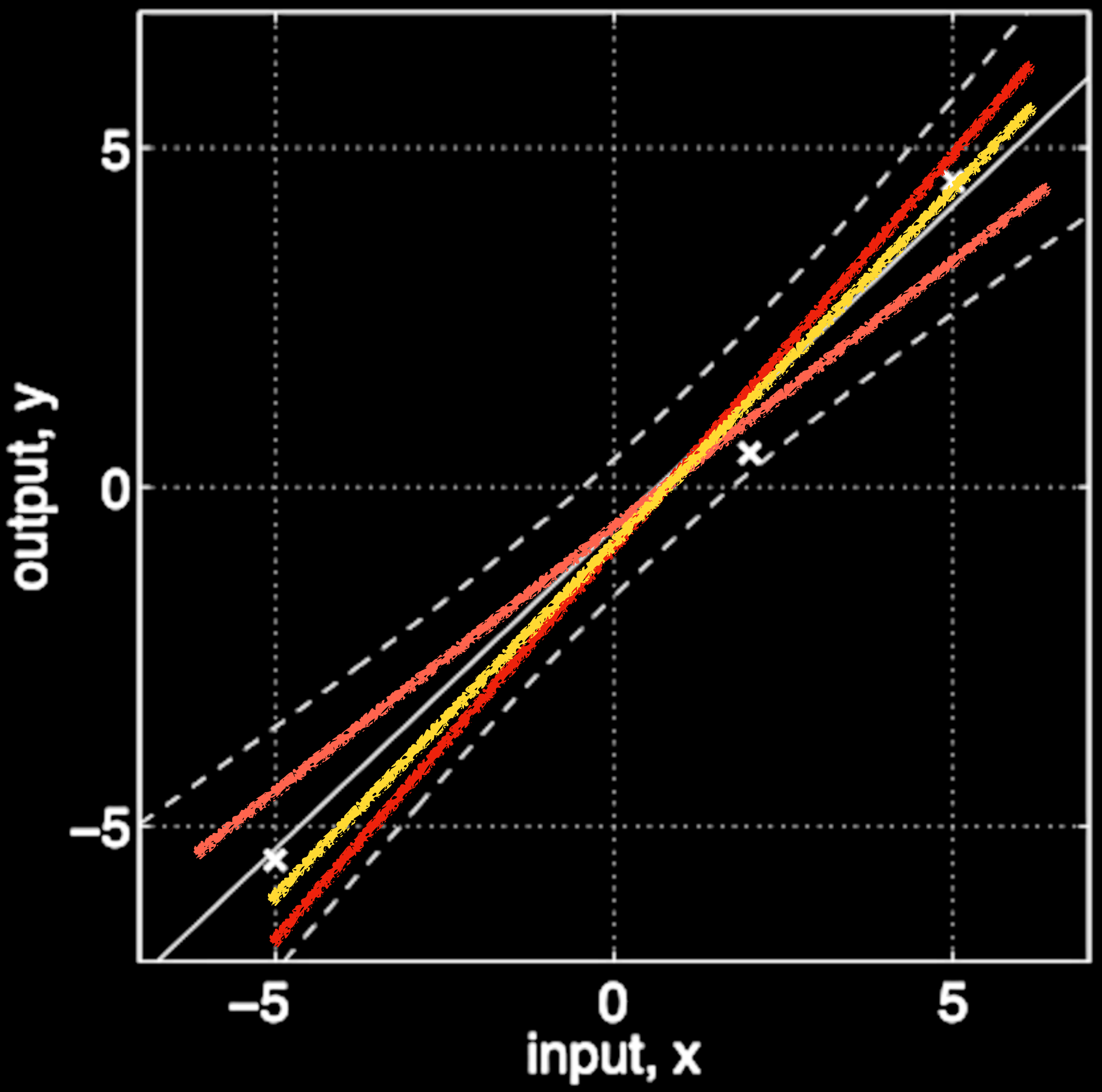
$$p(y* \mid x*, \mathbf{X}, \mathbf{y})$$



Why Full Bayesian Treatment Can Be Advantageous

Predictive distribution of $y*$, given a new data $\mathbf{x}*$

$$p(y* \mid x*, \mathbf{X}, \mathbf{y})$$

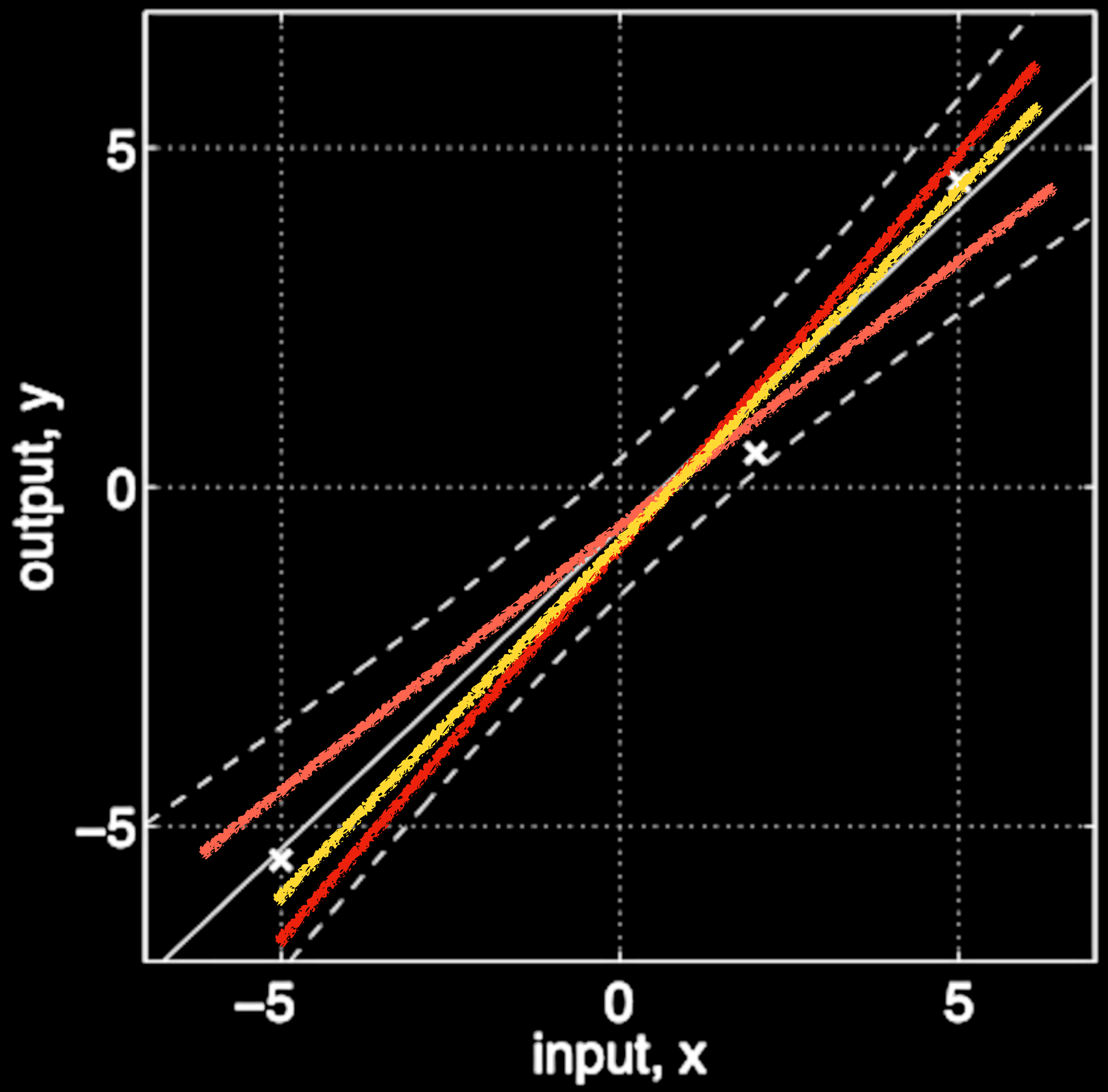


Why Full Bayesian Treatment Can Be Advantageous

Predictive distribution of y^*, given a new data $\mathbf{x}^*$

Integrating all possible θ

$$p(y^* \,|\, x^*, \mathbf{X}, \mathbf{y})$$



Yesterday:

	Primal	Dual
	parameters θ	weights α with $\theta = \sum_n \alpha_n \phi_n$
	optimal $\theta = (\Phi^\intercal \Phi + \lambda \mathrm{I}_D)^{-1} \Phi^\intercal \mathbf{y}$	optimal $\alpha = (\Phi \Phi^\intercal + \lambda \mathrm{I}_N)^{-1} \mathbf{y}$
	prediction $f(x_*) = \phi_*^\intercal \theta$	prediction $f(x_*) = \sum_n \alpha_n \phi_n^\intercal \phi_*$
	complexity $\mathcal{O}(D^3 + D^2 N)$	complexity $\mathcal{O}(N^3 + N^2 D)$

	Primal	Dual
Yesterday:	parameters θ	weights α with $\theta = \sum_n \alpha_n \phi_n$
	optimal $\theta = (\Phi^\intercal \Phi + \lambda \mathrm{I}_D)^{-1} \Phi^\intercal \mathbf{y}$	optimal $\alpha = (\Phi \Phi^\intercal + \lambda \mathrm{I}_N)^{-1} \mathbf{y}$
	prediction $f(x_*) = \phi_*^\intercal \theta$	prediction $f(x_*) = \sum_n \alpha_n \phi_n^\intercal \phi_*$
	complexity $\mathcal{O}(D^3 + D^2 N)$	complexity $\mathcal{O}(N^3 + N^2 D)$

Can we write down

$$p(y^* \mid x^*, \mathbf{X}, \mathbf{y})$$

In the form of $k(\mathbf{x}^*, \mathbf{X}), k(\mathbf{X}, \mathbf{X})$

Gaussian Process - Weight-Space Perspective

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

Gaussian Process - Weight-Space Perspective

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$m(\mathbf{x}^*) = \phi^{\mathrm{T}}(\mathbf{x}^*)\,\sigma^{-2}\,(\sigma_0^2\mathbf{I}_{\mathbf{D}} + \sigma^{-2}\,\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\,\boldsymbol{\Phi}^{\mathrm{T}}\,\mathbf{y}$$

Gaussian Process - Weight-Space Perspective

$$p(y^* \,|\, \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$m(\mathbf{x}^*) = \phi^{\mathrm{T}}(\mathbf{x}^*)\, \sigma^{-2} (\sigma_0^2 \mathbf{I}_{\mathrm{D}} + \sigma^{-2} \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi})^{-1} \mathbf{\Phi}^{\mathrm{T}} \mathbf{y}$$

Kernel trick
(see Lecture 4a)

$$m(\mathbf{x}^*) = \sigma_0^2\, \phi^{\mathrm{T}}(\mathbf{x}^*)\, \mathbf{\Phi}^{\mathrm{T}} (\sigma^2 \mathbf{I}_{\mathrm{N}} + \sigma_0^2\, \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}})^{-1} \mathbf{y}$$

(GP book eq 2.12)

Gaussian Process - Weight-Space Perspective

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$m(\mathbf{x}^*) = \phi^{\mathrm{T}}(\mathbf{x}^*)\, \sigma^{-2} (\sigma_0^2 \mathbf{I}_{\mathrm{D}} + \sigma^{-2}\, \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi})^{-1}\, \boldsymbol{\Phi}^{\mathrm{T}}\, \mathbf{y}$$

Kernel trick
(see Lecture 4a)

$$m(\mathbf{x}^*) = \underbrace{\boxed{\sigma_0^2\, \phi^{\mathrm{T}}(\mathbf{x}^*)\, \boldsymbol{\Phi}^{\mathrm{T}}}}_{k(\mathbf{x}^*, \mathbf{X})} (\sigma^2 \mathbf{I}_{\mathrm{N}} + \underbrace{\boxed{\sigma_0^2\, \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}}}_{k(\mathbf{X}, \mathbf{X})})^{-1}\, \mathbf{y}$$

(GP book eq 2.12)

Gaussian Process - Weight-Space Perspective

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + \phi(\mathbf{x}^*)^{\mathrm{T}} \boldsymbol{\Sigma}\, \phi(\mathbf{x}^*)$$

$k(\mathbf{x}^*, \mathbf{x}^*)$

$k(\mathbf{x}^*, \mathbf{X})$ $k(\mathbf{X}, \mathbf{X})$ $k(\mathbf{X}, \mathbf{x}^*)$

Gaussian Process - Weight-Space Perspective

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + \phi(\mathbf{x}^*)^{\mathrm{T}} \mathbf{\Sigma}\, \phi(\mathbf{x}^*)$$

Kernel trick
(see Lecture 4a)

$$= \sigma^2 + \underbrace{\sigma_0^2 \phi(\mathbf{x}^*)^{\mathrm{T}} \phi(\mathbf{x}^*)}_{k(\mathbf{x}^*, \mathbf{x}^*)}$$

(GP book eq 2.12)

$$- \underbrace{\sigma_0^2 \phi(\mathbf{x}^*)^{\mathrm{T}}\, \mathbf{\Phi}^{\mathrm{T}}}_{k(\mathbf{x}^*, \mathbf{X})} (\sigma^2 \mathbf{I}_{\mathrm{N}} + \underbrace{\sigma_0^2\, \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}}_{k(\mathbf{X}, \mathbf{X})})^{-1} \underbrace{\sigma_0^2 \mathbf{\Phi} \phi\, (\mathbf{x}^*)}_{k(\mathbf{X}, \mathbf{x}^*)}$$

Gaussian Process

Kernelised

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$$

$$= \mathcal{N}(y^*; m(x^*), \sigma^2(x^*))$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma \mathrm{I})^{-1}\,\mathbf{y}$$

(Bishop eq 6.66)

$$\sigma^2(x^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*)$$

$$-k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\,k(\mathbf{X}, \mathbf{x}^*)$$

(Bishop eq 6.67)

Bayesian Linear Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$$

$$= \mathcal{N}(y^*; \mu^T\,\phi(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$\mu = \sigma^{-2}\,(\sigma_0^{-2}\mathbf{I}_\mathbf{D} + \sigma^{-2}\,\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi})^{-1}\,\boldsymbol{\Phi}^\mathbf{T}\,\mathbf{y}$$

(Bishop eq 3.53)

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + \phi(\mathbf{x}^*)^\mathbf{T}\,\boldsymbol{\Sigma}\,\phi(\mathbf{x}^*)$$

$$\Sigma = (\sigma_0^{-2}\mathbf{I}_\mathbf{D} + \sigma^{-2}\,\boldsymbol{\Phi}^\mathbf{T}\boldsymbol{\Phi})^{-1}$$

(Bishop eq 3.54)

Gaussian Process

Kernelised

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$$

$$= \mathcal{N}(y^*; m(x^*), \sigma^2(x^*))$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X}) \boxed{(k(\mathbf{X}, \mathbf{X}) + \sigma \mathrm{I})^{-1}} \mathbf{y}$$

(Bishop eq 6.66)

$$\sigma^2(x^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*)$$

$$-k(\mathbf{x}^*, \mathbf{X}) \boxed{(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}} k(\mathbf{X}, \mathbf{x}^*)$$

(Bishop eq 6.67)

Inverse of an $\mathbb{R}^{N \times N}$ matrix

Bayesian Linear Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$$

$$= \mathcal{N}(y^*; \mu^T \phi(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$\mu = \sigma^{-2} (\sigma_0^{-2} \mathbf{I_D} + \sigma^{-2} \mathbf{\Phi}^\mathrm{T} \mathbf{\Phi})^{-1} \mathbf{\Phi}^\mathrm{T} \mathbf{y}$$

(Bishop eq 3.53)

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + \phi(\mathbf{x}^*)^\mathrm{T} \mathbf{\Sigma} \, \phi(\mathbf{x}^*)$$

$$\Sigma = (\sigma_0^{-2} \mathbf{I_D} + \sigma^{-2} \mathbf{\Phi}^\mathrm{T} \mathbf{\Phi})^{-1}$$

(Bishop eq 3.54)

Gaussian Process

Kernelised

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$$

$$= \mathcal{N}(y^*; m(x^*), \sigma^2(x^*))$$

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X}) \boxed{(k(\mathbf{X}, \mathbf{X}) + \sigma \mathrm{I})^{-1}} \mathbf{y}$$

(Bishop eq 6.66)

$$\sigma^2(x^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*)$$

$$- k(\mathbf{x}^*, \mathbf{X}) \boxed{(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}} k(\mathbf{X}, \mathbf{x}^*)$$

(Bishop eq 6.67)

Inverse of an $\mathbb{R}^{N \times N}$ matrix

Bayesian Linear Regression

$$p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$$

$$= \mathcal{N}(y^*; \mu^T \phi(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$\mu = \sigma^{-2} \boxed{(\sigma_0^{-2} \mathbf{I_D} + \sigma^{-2} \mathbf{\Phi}^\mathrm{T} \mathbf{\Phi})^{-1}} \mathbf{\Phi}^\mathrm{T} \mathbf{y}$$

(Bishop eq 3.53)

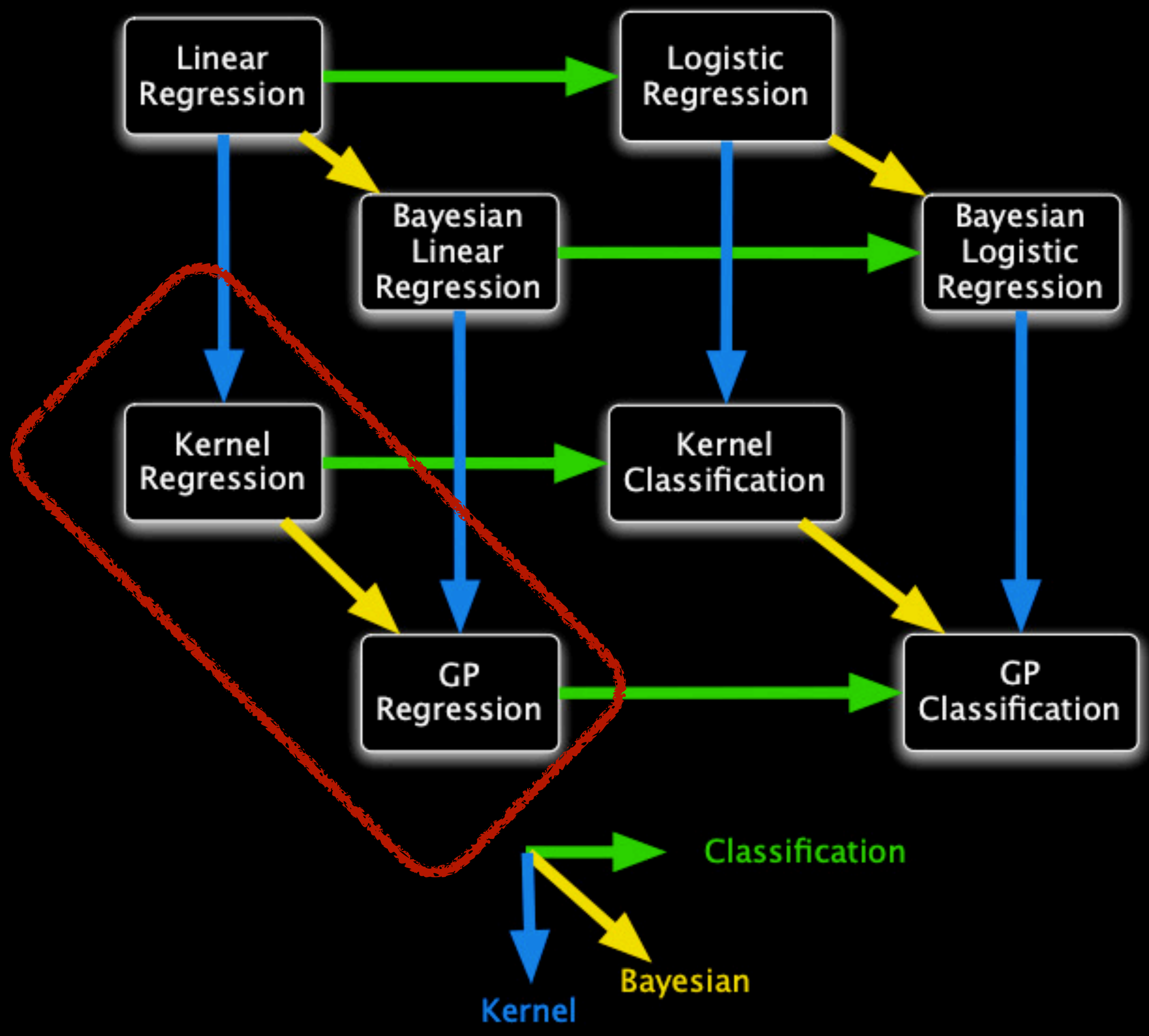
$$\sigma^2(\mathbf{x}^*) = \sigma^2 + \phi(\mathbf{x}^*)^\mathrm{T} \mathbf{\Sigma} \, \phi(\mathbf{x}^*)$$

$$\Sigma = \boxed{(\sigma_0^{-2} \mathbf{I_D} + \sigma^{-2} \mathbf{\Phi}^\mathrm{T} \mathbf{\Phi})^{-1}}$$

(Bishop eq 3.54)

Inverse of an $\mathbb{R}^{D \times D}$ matrix

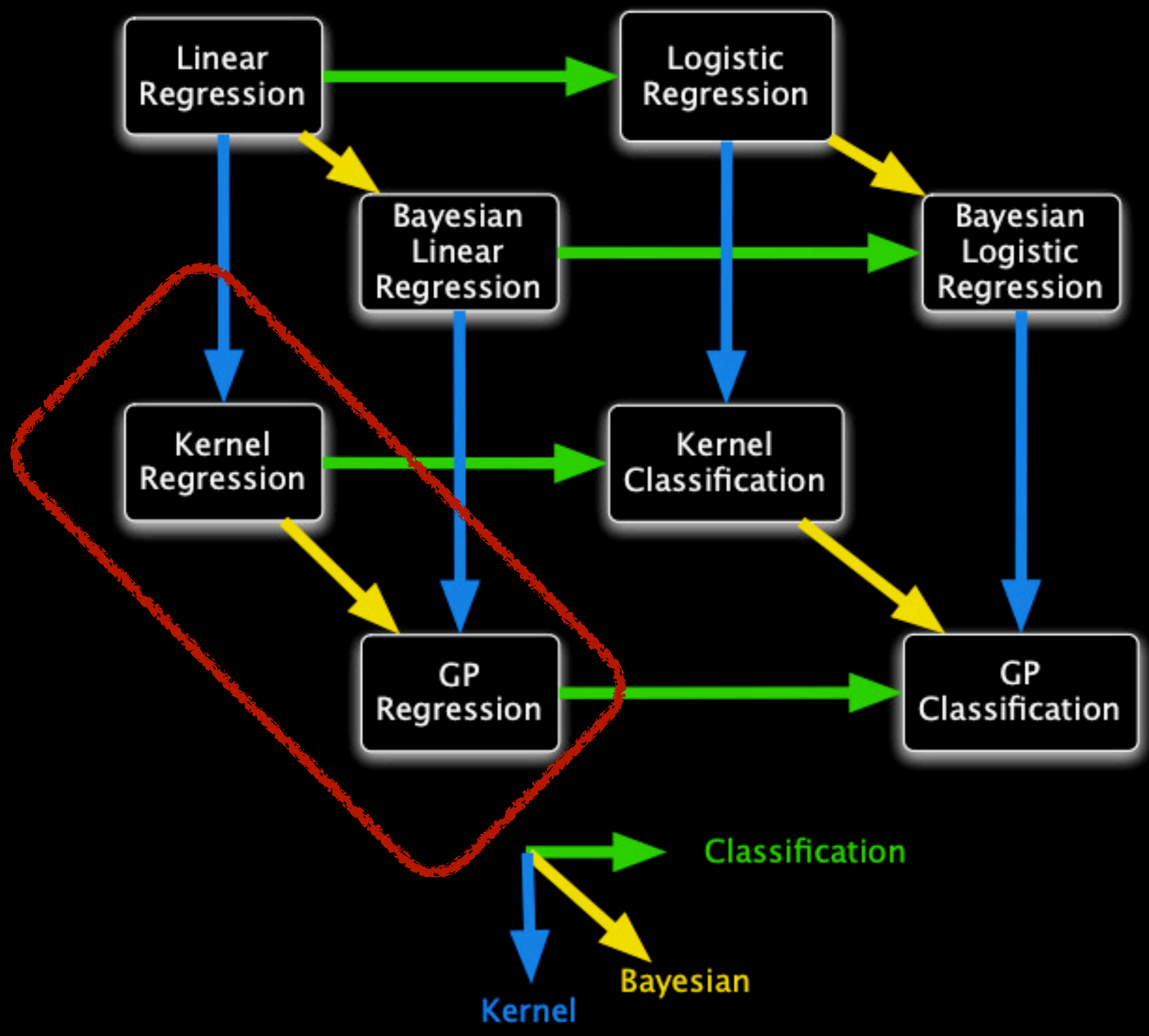
Gaussian Process - Function-Space Perspective



Gaussian Process - Function-Space Perspective

Gaussian Process =

Making Kernel Regression "Bayesian"



Gaussian Process - Function-Space Perspective

Gaussian Process - Function-Space Perspective

- A Gaussian Process is a probability distribution over functions $f(\mathbf{x})$

  such that $\forall(\mathbf{x_1}, \ldots, \mathbf{x_n}), p(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n})) = \mathcal{N}$

Gaussian Process - Function-Space Perspective

- A Gaussian Process is a probability distribution over functions $f(\mathbf{x})$

  such that $\forall(\mathbf{x_1}, \ldots, \mathbf{x_n}), p(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n})) = \mathcal{N}$

- Commonly set mean to zero, due to no prior knowledge of $f(x)$

Gaussian Process - Function-Space Perspective

- A Gaussian Process is a probability distribution over functions $f(\mathbf{x})$

  such that $\forall(\mathbf{x_1}, \ldots, \mathbf{x_n}), p(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n})) = \mathcal{N}$

- Commonly set mean to zero, due to no prior knowledge of $f(x)$

- Fully specified by their covariance function or kernel.

Gaussian Process - Function-Space Perspective

- A Gaussian Process is a probability distribution over functions $f(\mathbf{x})$

  such that $\forall(\mathbf{x_1}, \ldots, \mathbf{x_n}), p(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n})) = \mathcal{N}$

- Commonly set mean to zero, due to no prior knowledge of $f(x)$

- Fully specified by their covariance function or kernel.

➡ $$p\big(\mathbf{f}(\mathbf{x})\big) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$$

(Bishop eq 6.60, GP Book eq 2.17)

Gaussian Process - Function-Space Perspective

- A Gaussian Process is a probability distribution over functions $f(\mathbf{x})$

  such that $\forall(\mathbf{x_1}, \ldots, \mathbf{x_n}), p(f(\mathbf{x_1}), \ldots, f(\mathbf{x_n})) = \mathcal{N}$

- Commonly set mean to zero, due to no prior knowledge of $f(x)$

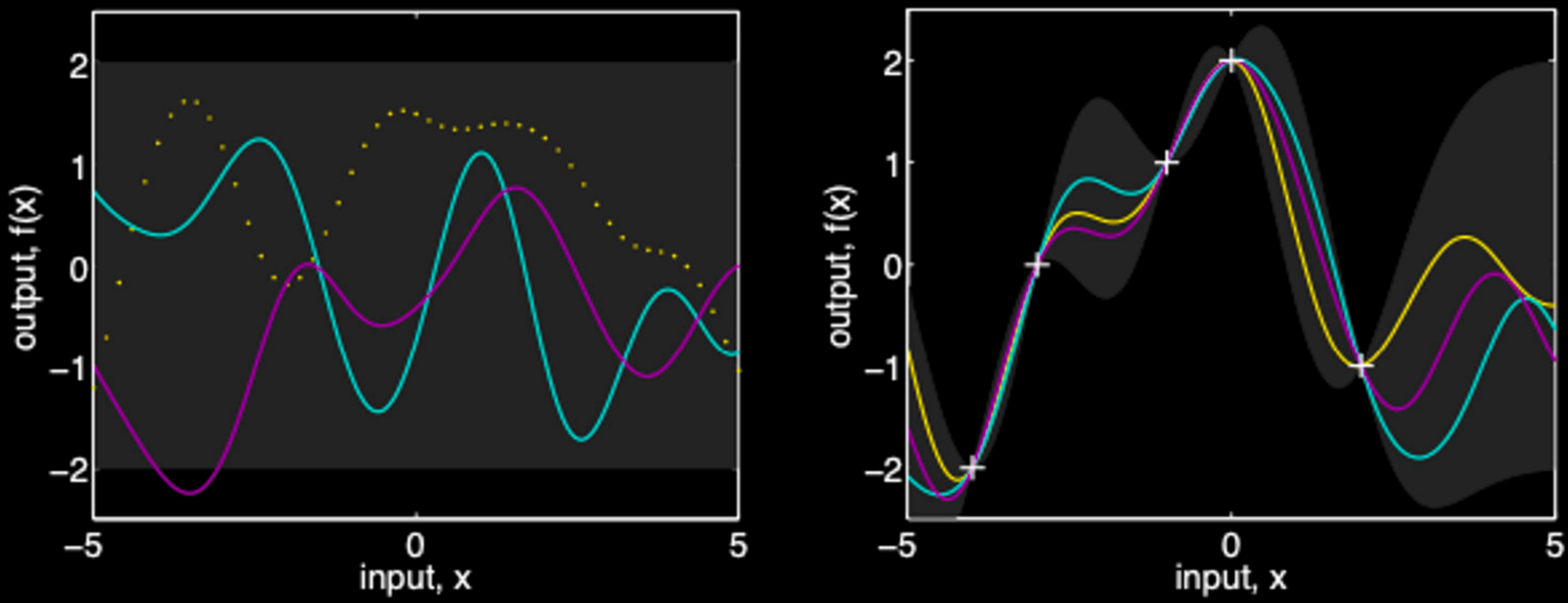
- Fully specified by their covariance function or kernel.

$$\Rightarrow \quad p\big(\mathbf{f}(\mathbf{x})\big) = \mathcal{N}(\mathbf{f}; \underbrace{\mathbf{0}}_{\substack{\text{Assuming the prior “weight”} \\ p(w) \text{ to have mean zero}}}, \mathbf{K})$$

(Bishop eq 6.60, GP Book eq 2.17)

Function-Space Perspective

Note : prior mean of zero



(a), prior (b), posterior

Posterior distribution of functions : “rejecting” functions that do not satisfy the constraints

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \underbrace{\mathbf{K}_{\mathbf{y}}}_{\in \mathbb{R}^{N+1,N+1}})$$ (Bishop eq 6.64)

$$K_{y,mn} = k(x_m, x_n) + \underbrace{\sigma^2 \delta_{mn}}_{\text{observation noise}}$$ (Bishop eq 6.62)

Our prior belief over the kind of functions (smoothness) we expect

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \underbrace{\mathbf{K_y}}_{\in \mathbb{R}^{N+1,N+1}})$$ (Bishop eq 6.64)

$$K_{y,mn} = k(x_m, x_n) + \underbrace{\sigma^2 \delta_{mn}}_{\text{observation noise}}$$ (Bishop eq 6.62)

Our prior belief over the kind of functions (smoothness) we expect

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \underbrace{\mathbf{K}_{\mathbf{y}}}_{\in \mathbb{R}^{N+1,N+1}})$$ (Bishop eq 6.64)

Also explains why we can deal with infinite-dimensional functional space in GP. We only study a "finite" distribution on data points that concerns us

$$K_{y,mn} = k(x_m, x_n) + \underbrace{\sigma^2 \delta_{mn}}_{\text{observation noise}}$$ (Bishop eq 6.62)

Joint “prior” distribution

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \mathbf{K_y})$$

Joint “prior” distribution

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \mathbf{K_y})$$

Conditional distribution of Gaussians is trivial

$$p(y^* \,|\, y) = p(y^* \,|\, \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

Joint "prior" distribution

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \mathbf{K_y})$$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

(Bishop eq 2.81-2.82)

Conditional distribution of Gaussians is trivial

$$p(y^* \,|\, y) = p(y^* \,|\, \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

(Bishop eq 2.81-2.82)

Joint “prior” distribution

$$p(y, y^*) = \mathcal{N}(\mathbf{0}, \mathbf{K}_\mathbf{y})$$

Conditional distribution of Gaussians is trivial

$$p(y^* \,|\, y) = p(y^* \,|\, \mathbf{x}^*, \mathbf{X}, y) = \mathcal{N}(y^*; m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$

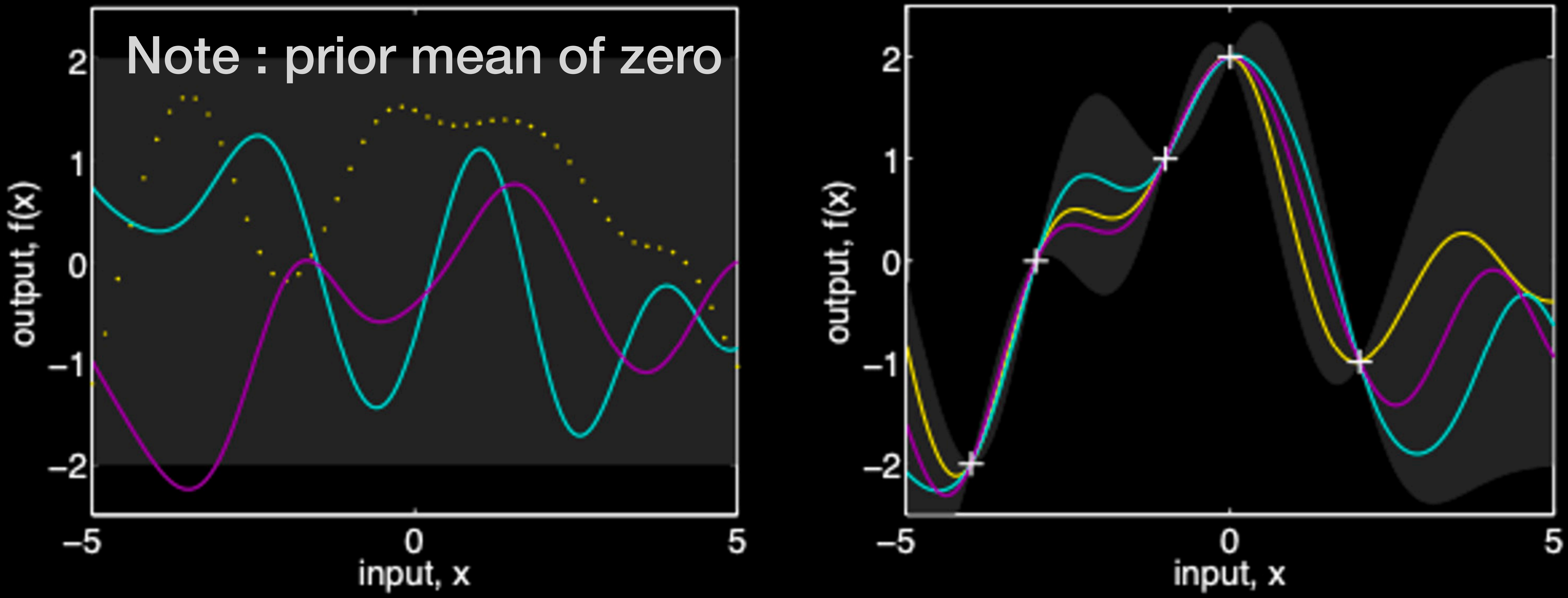
$$m(x^*) = k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, y$$ (Bishop eq 6.66, GP Book eq 2.19, 2.25)

$$\sigma^2(x^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

(Bishop eq 6.67, GP Book eq 2.19, 2.26)

Function-Space Perspective

$$p(y^* \mid y) = \mathcal{N}(m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$



(a), prior   (b), posterior

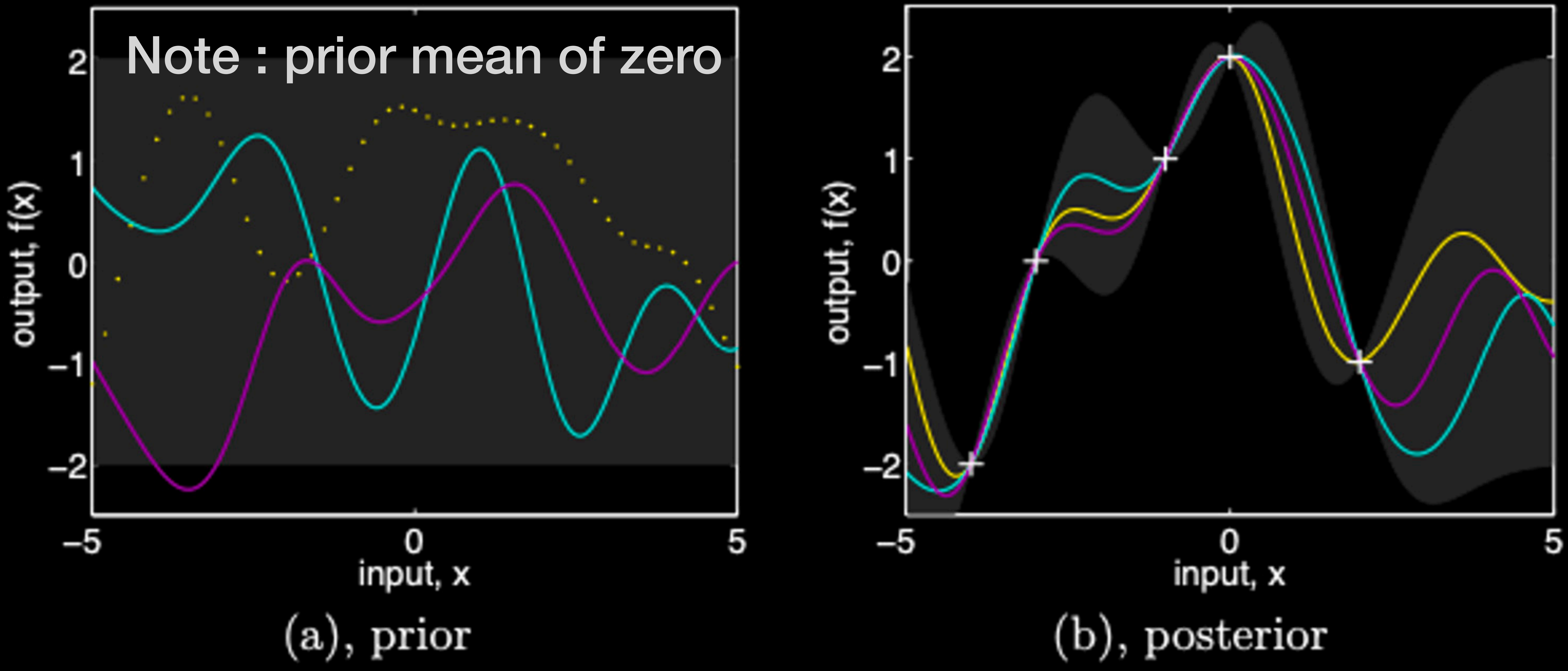
Posterior distribution of functions : “rejecting” functions that do not satisfy the constraints

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, y$$

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

Function-Space Perspective

$$p(y^* \mid y) = \mathcal{N}(m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$



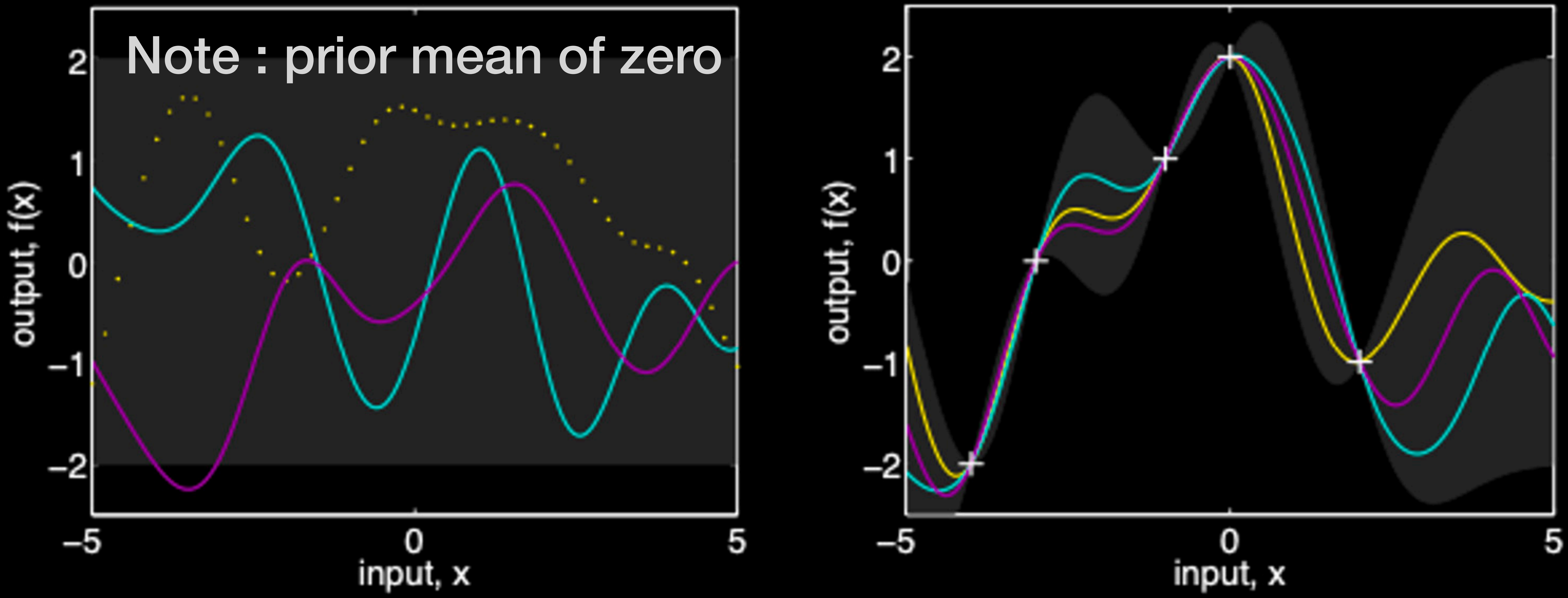
Posterior distribution of functions : “rejecting” functions that do not satisfy the constraints

$$m(\mathbf{x}^*) = \boxed{k(\mathbf{x}^*, \mathbf{X})}\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, y$$ The influence depends on the proximity

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)$$

Function-Space Perspective

$$p(y^* \mid y) = \mathcal{N}(m(\mathbf{x}^*), \sigma^2(\mathbf{x}^*))$$



(a), prior

(b), posterior

Posterior distribution of functions : “rejecting” functions that do not satisfy the constraints

$$m(\mathbf{x}^*) = \boxed{k(\mathbf{x}^*, \mathbf{X})}\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, y$$

The influence depends on the proximity

$$\sigma^2(\mathbf{x}^*) = \sigma^2 + k(\mathbf{x}^*, \mathbf{x}^*)\ \boxed{-\, k(\mathbf{x}^*, \mathbf{X})\,(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\, k(\mathbf{X}, \mathbf{x}^*)}$$

Uncertainty reduction given the training data, does not depend on y!

Visualising Kernels

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right) + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

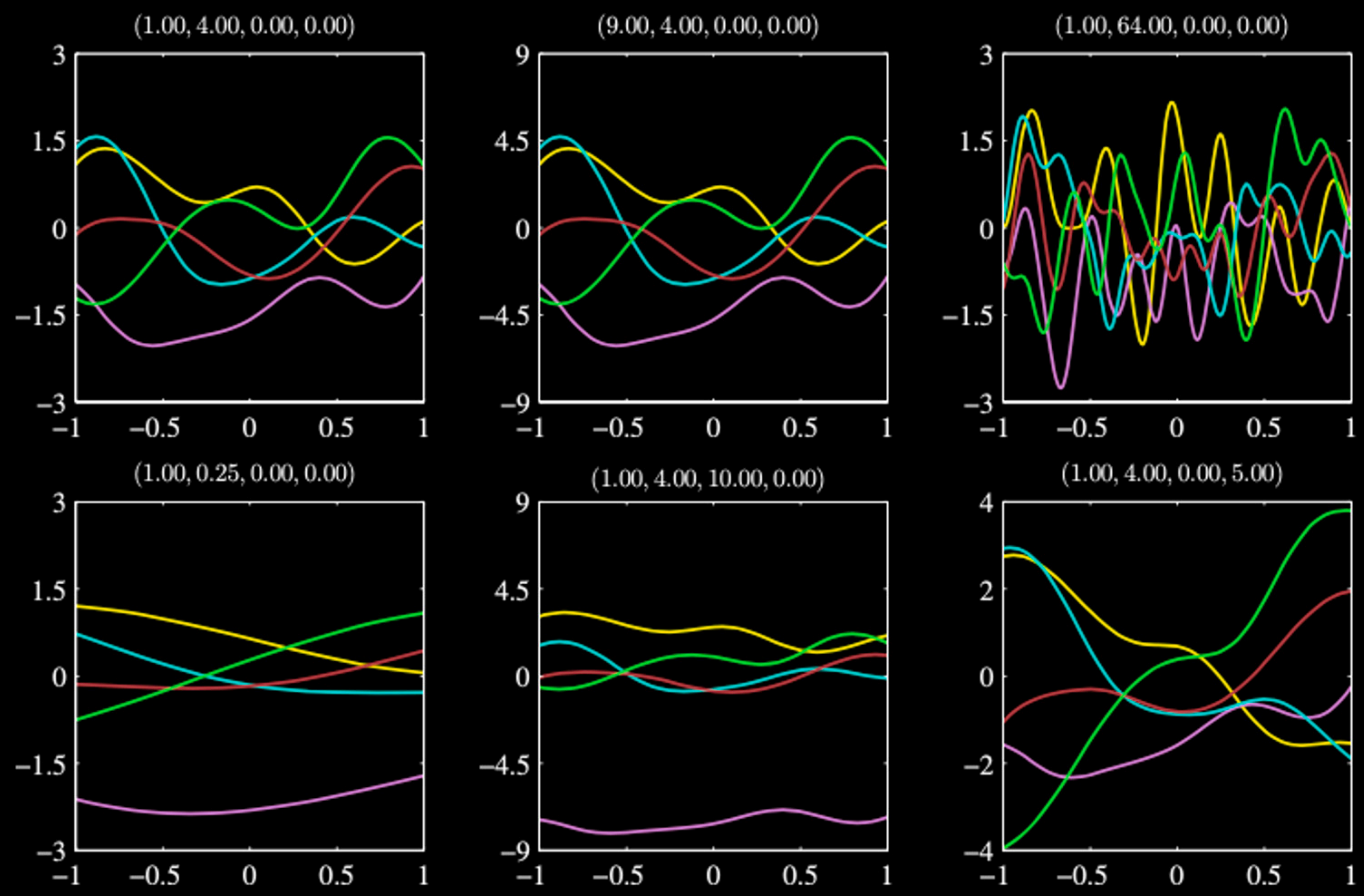


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right) + \theta_2 + \theta_3 \mathbf{x}_n^{T} \mathbf{x}_m$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

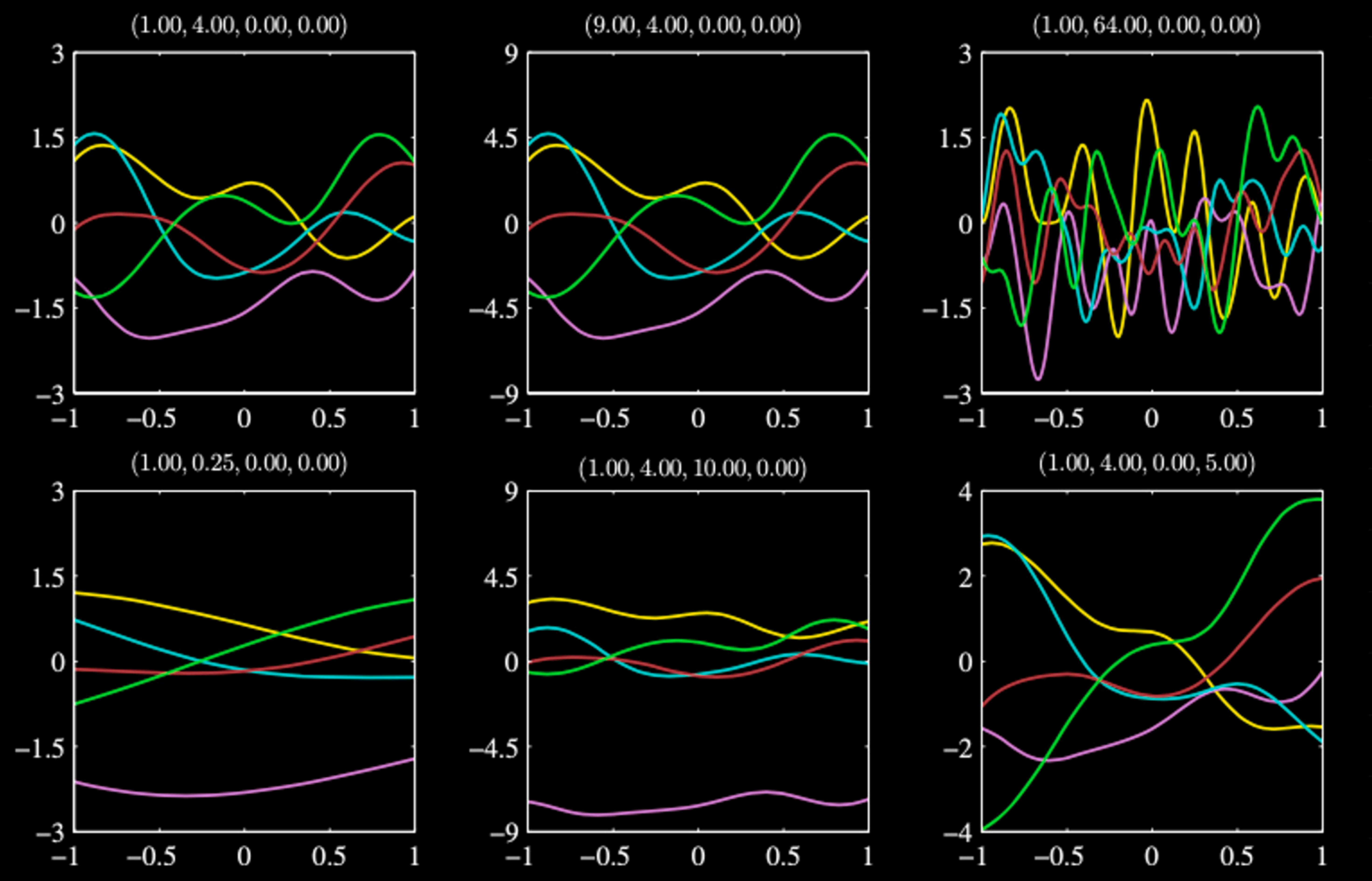


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n^T x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

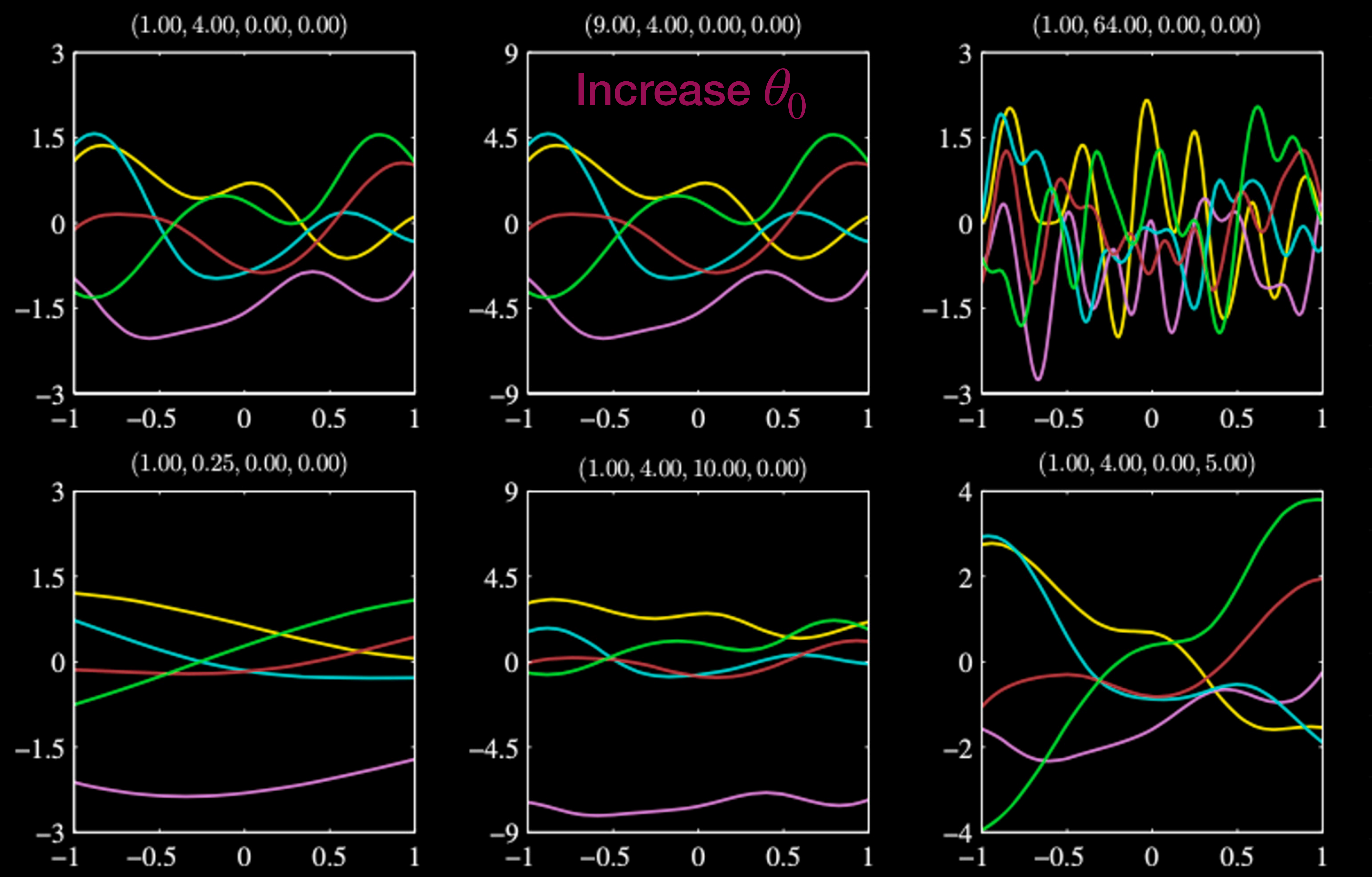


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n^T x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

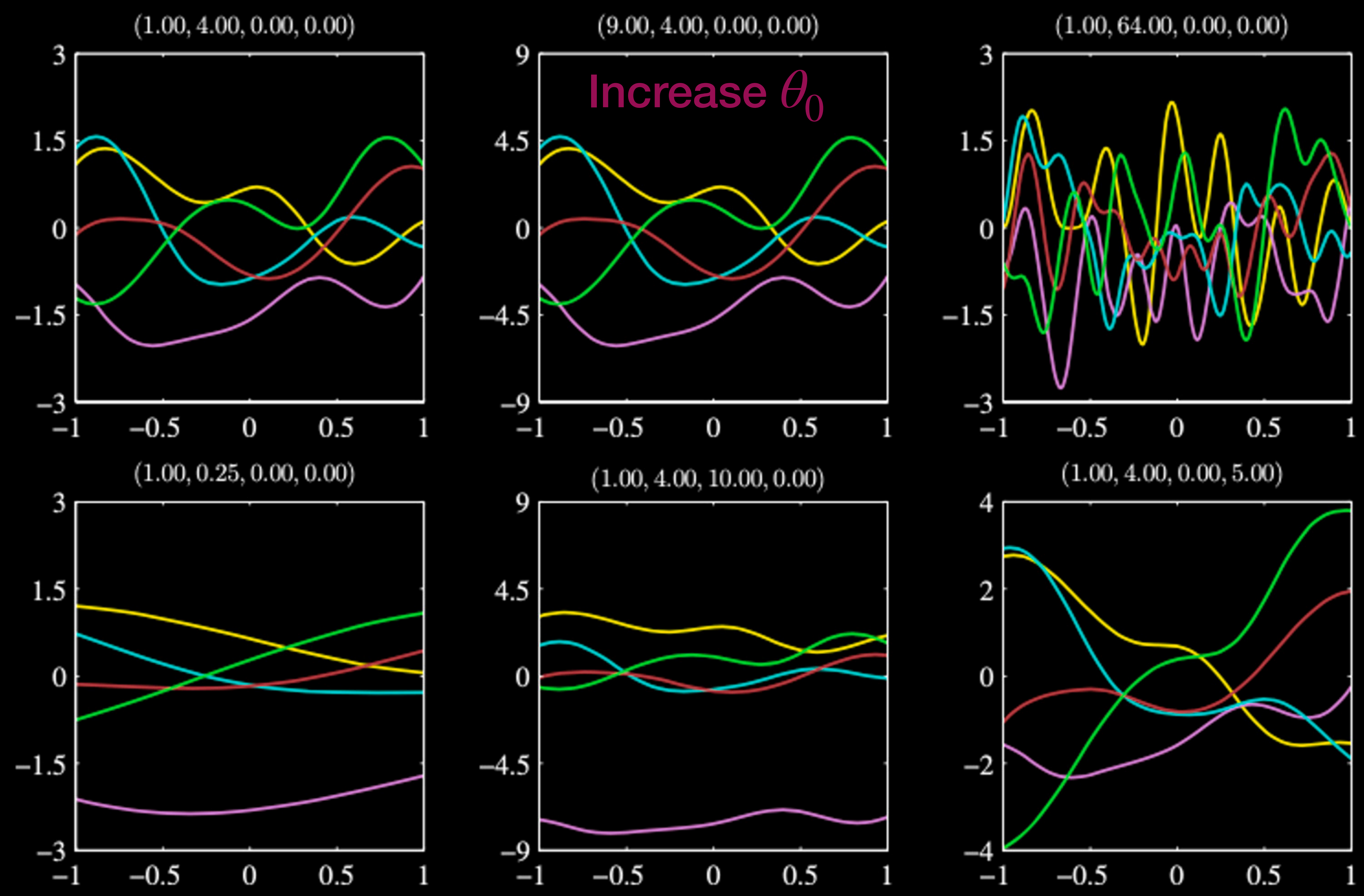


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n^T x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

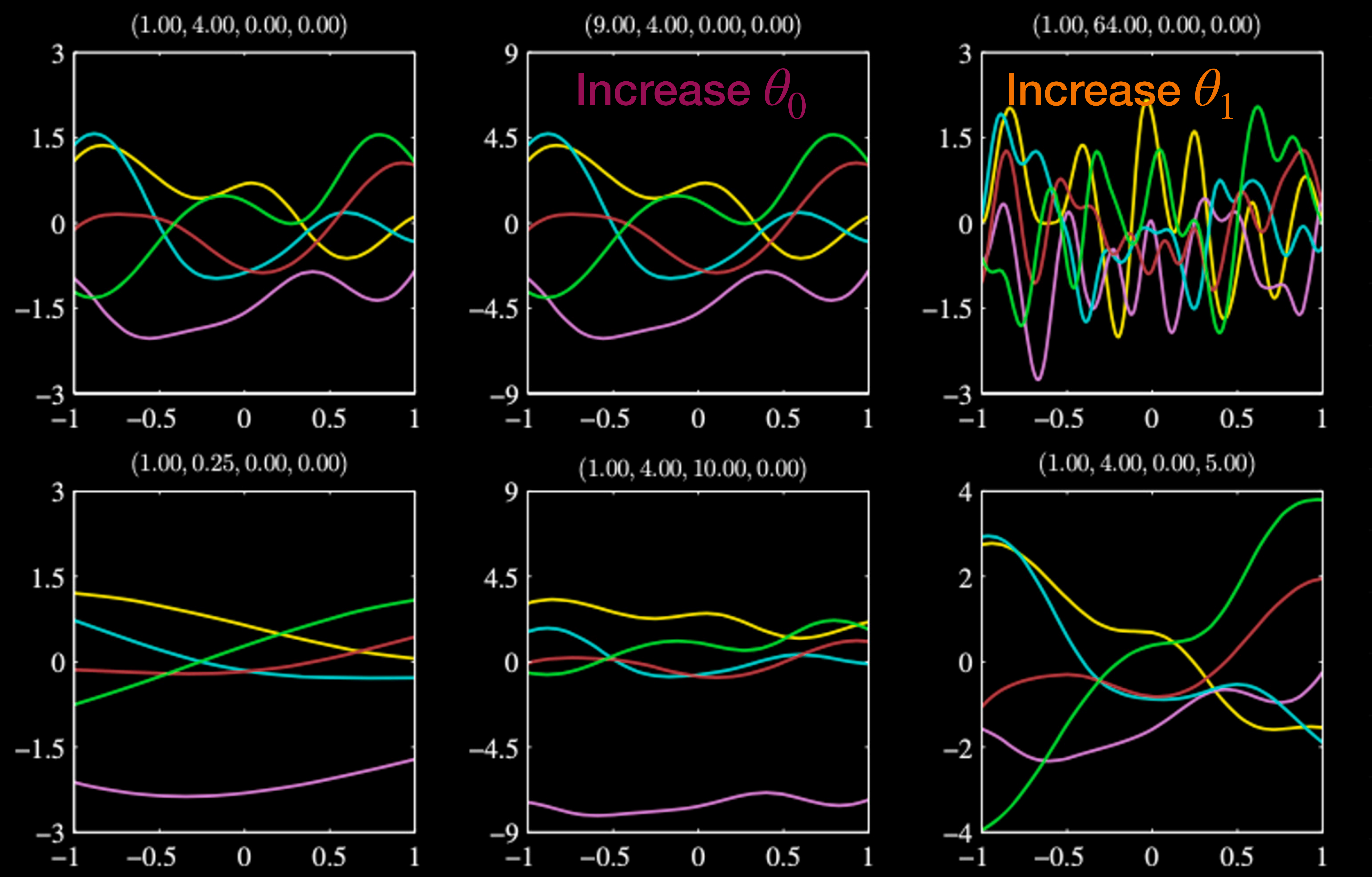


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n^T x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

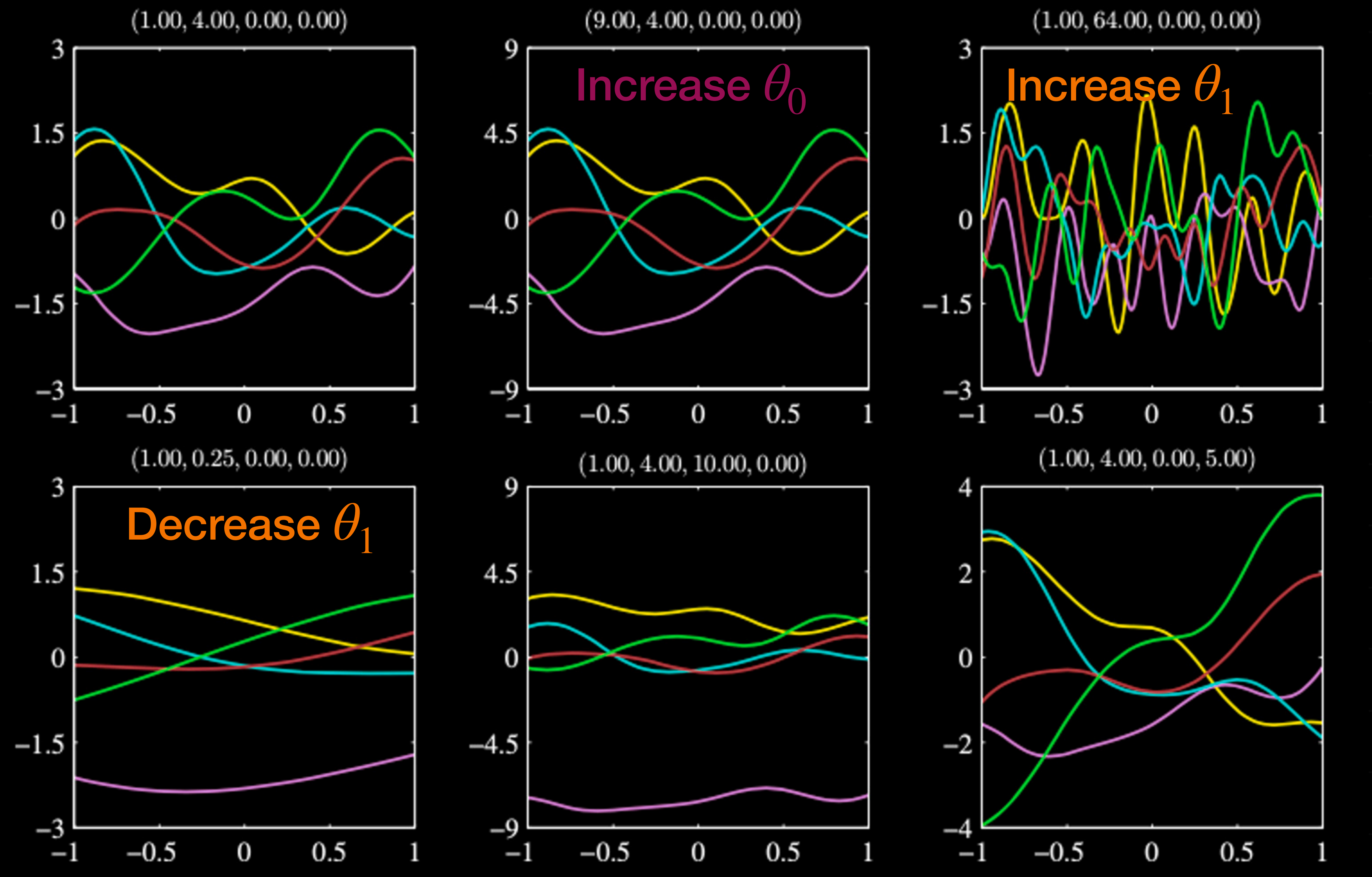


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n^T x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

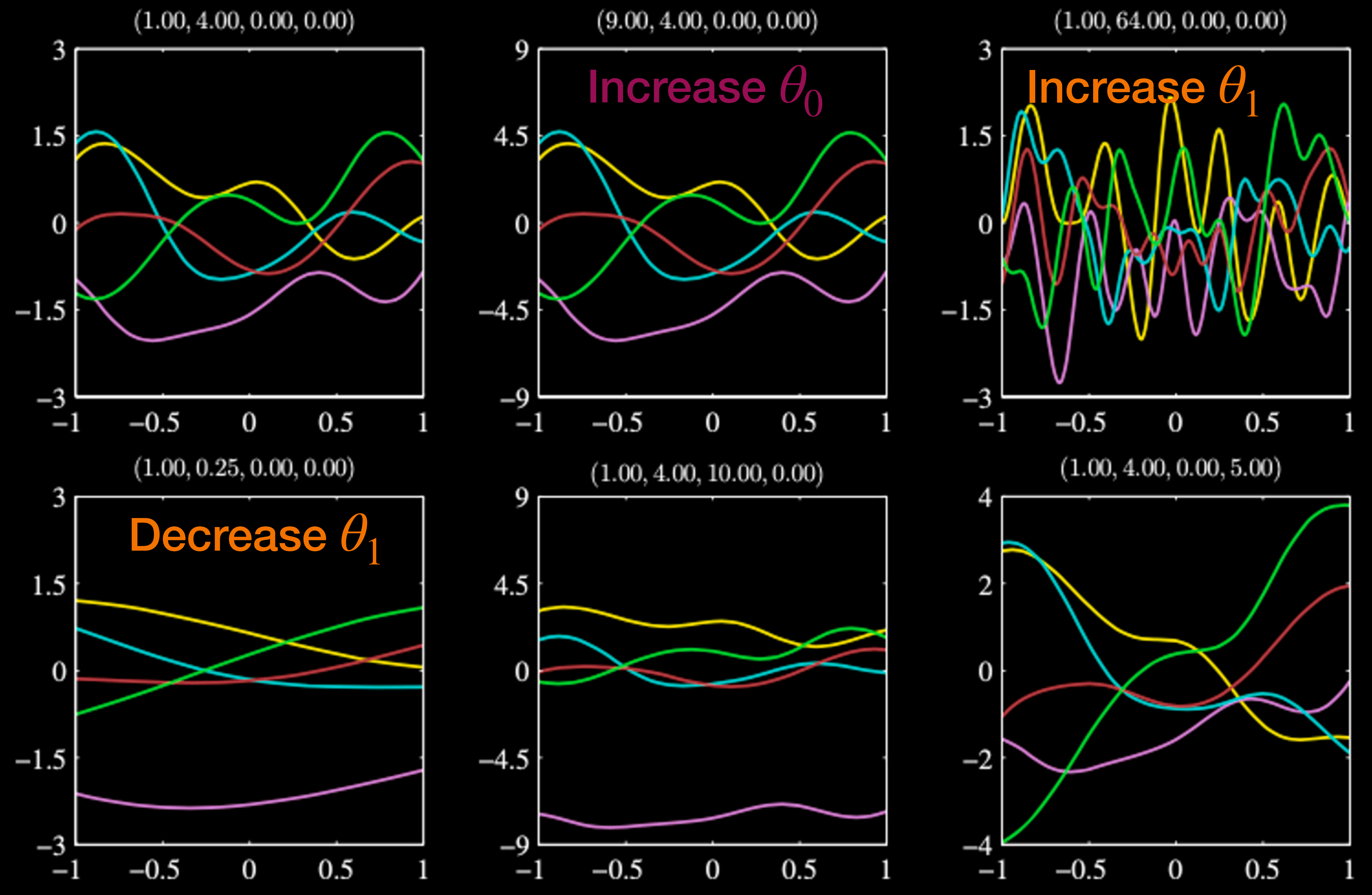


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n^T x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

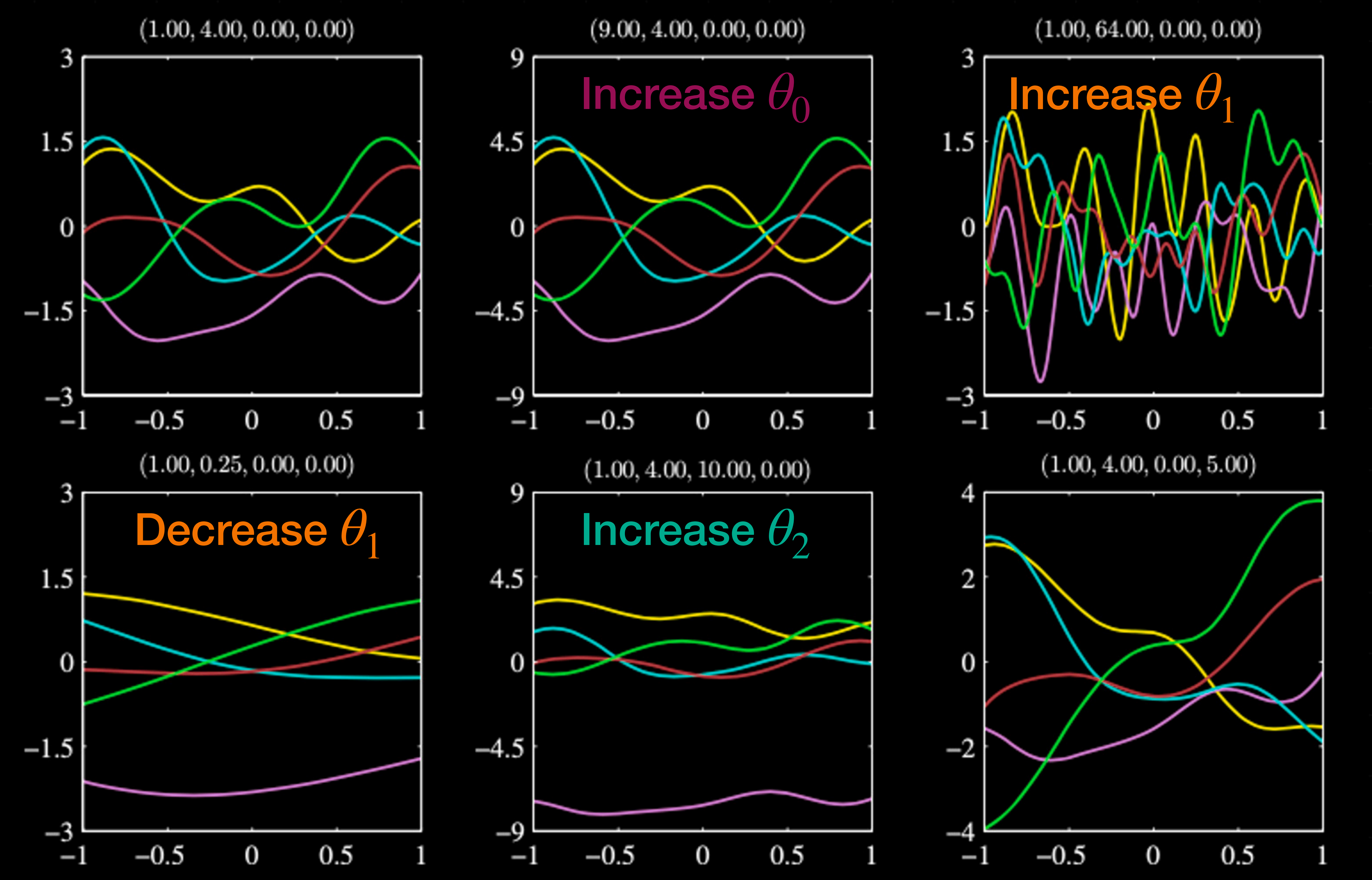


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n}^{\mathrm{T}}\mathbf{x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

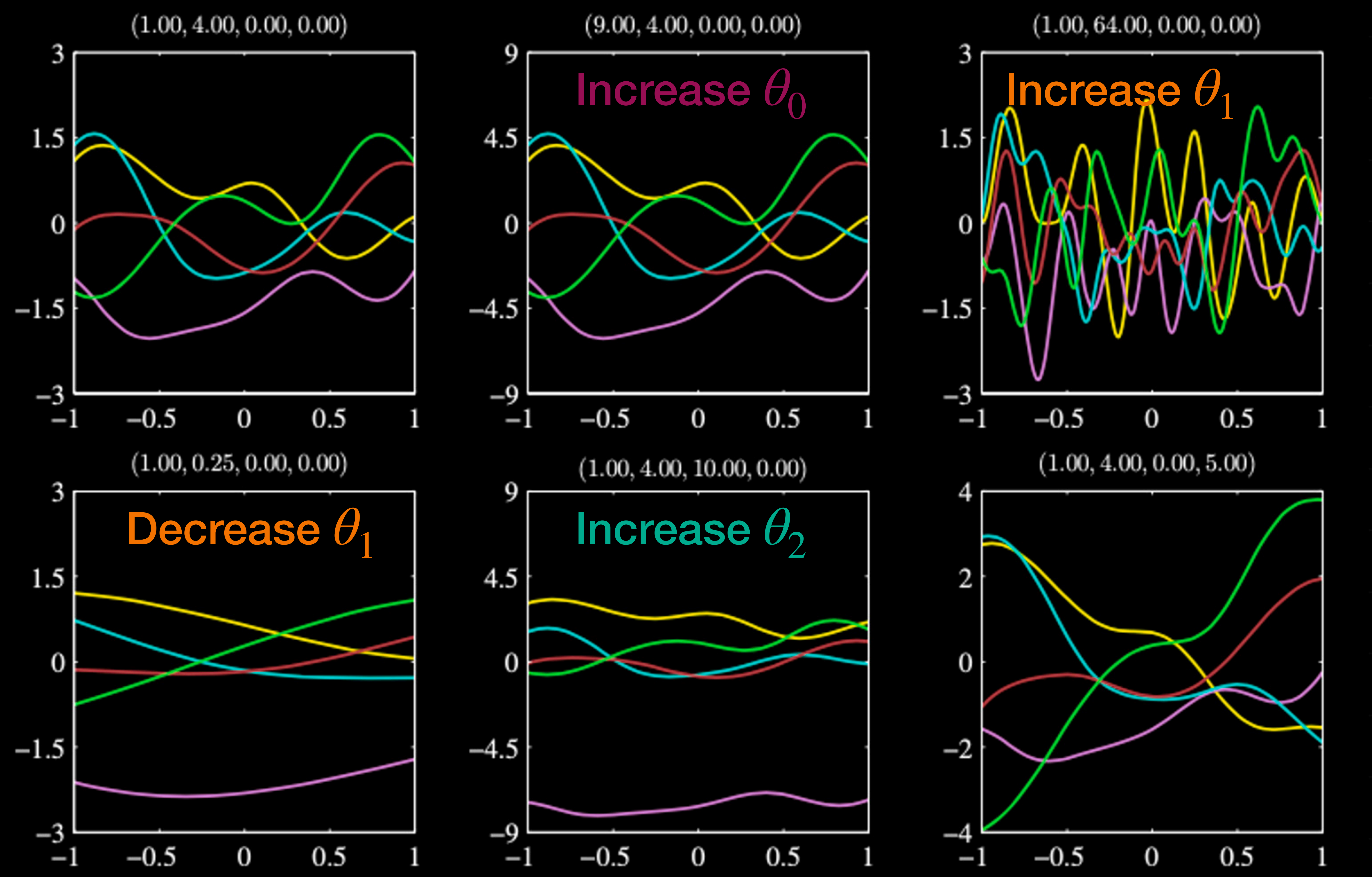


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\mathbf{x_n} - \mathbf{x_m}\|^2\right) + \theta_2 + \theta_3 \mathbf{x_n}^T \mathbf{x_m}$$

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

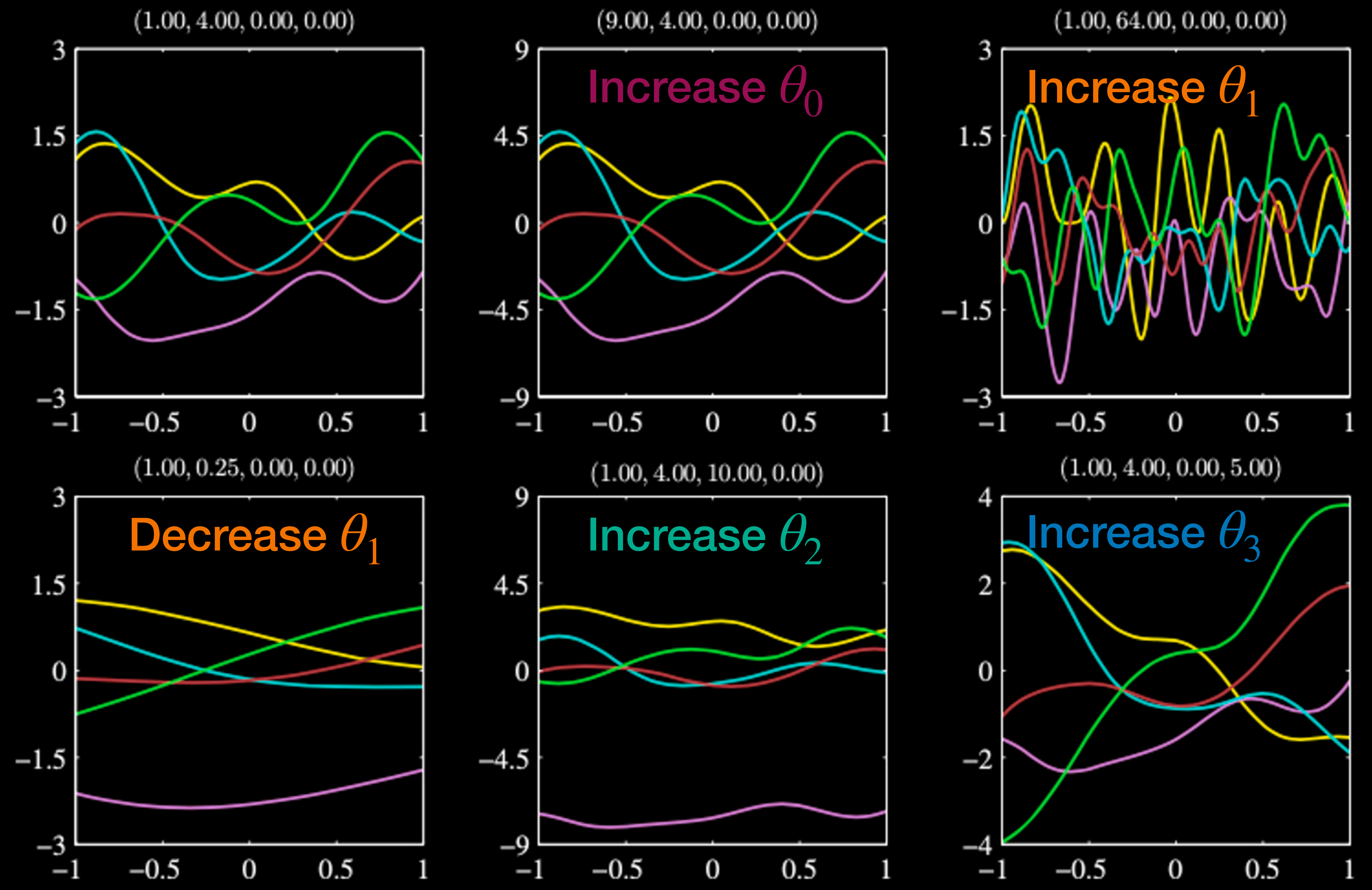


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

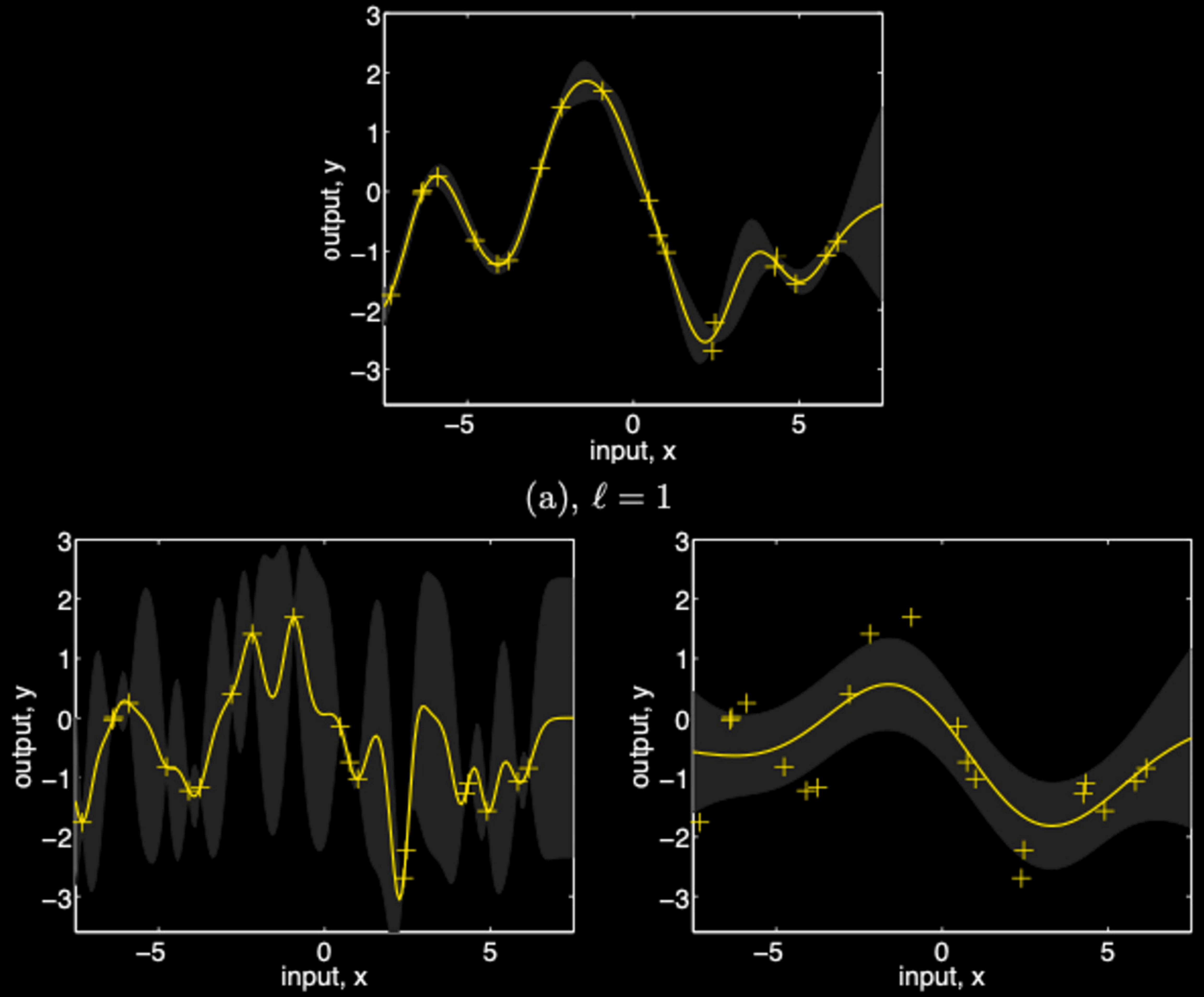
More examples: https://smlbook.org/GP/

Visualising Kernels

$$k(\mathbf{x_n}, \mathbf{x_m}) = \sigma_\mathbf{f}^2 \exp\left(-\frac{1}{2\ell^2}||\mathbf{x_n} - \mathbf{x_m}||^2\right) + \sigma_\mathbf{n}^2 \delta_\mathbf{nm}$$

Gaussian Process,
with constraints

(GP book)



(a), $\ell = 1$

(b), $\ell = 0.3$

(c), $\ell = 3$

Bayesian Linear Regression

Gaussian Process

Bayesian Linear Regression

- When $D \ll N$, faster training and inference

Gaussian Process

Bayesian Linear Regression

- When $D \ll N$, faster training and inference
- But restrict to the functional class we assume

Gaussian Process

Bayesian Linear Regression

- When $D \ll N$, faster training and inference
- But restrict to the functional class we assume

Gaussian Process

- Prior distribution assumed to be all possible functions (subjected to the kernel assumption and hyper parameters)

Bayesian Linear Regression

- When $D \ll N$, faster training and inference
- But restrict to the functional class we assume

Gaussian Process

- Prior distribution assumed to be all possible functions (subjected to the kernel assumption and hyper parameters)
- Implicitly assume that $D \gg N$ [depend on kernels], do not need to worry about lack of degrees of freedom to fit all the data

Bayesian Linear Regression

- When $D \ll N$, faster training and inference
- But restrict to the functional class we assume

Gaussian Process

- Prior distribution assumed to be all possible functions (subjected to the kernel assumption and hyper parameters)
- Implicitly assume that $D \gg N$ [depend on kernels], do not need to worry about lack of degrees of freedom to fit all the data
- Training limited N^3 complexity (extra reading - Chapter 8 of the GP book, GP approximations in Week 6)