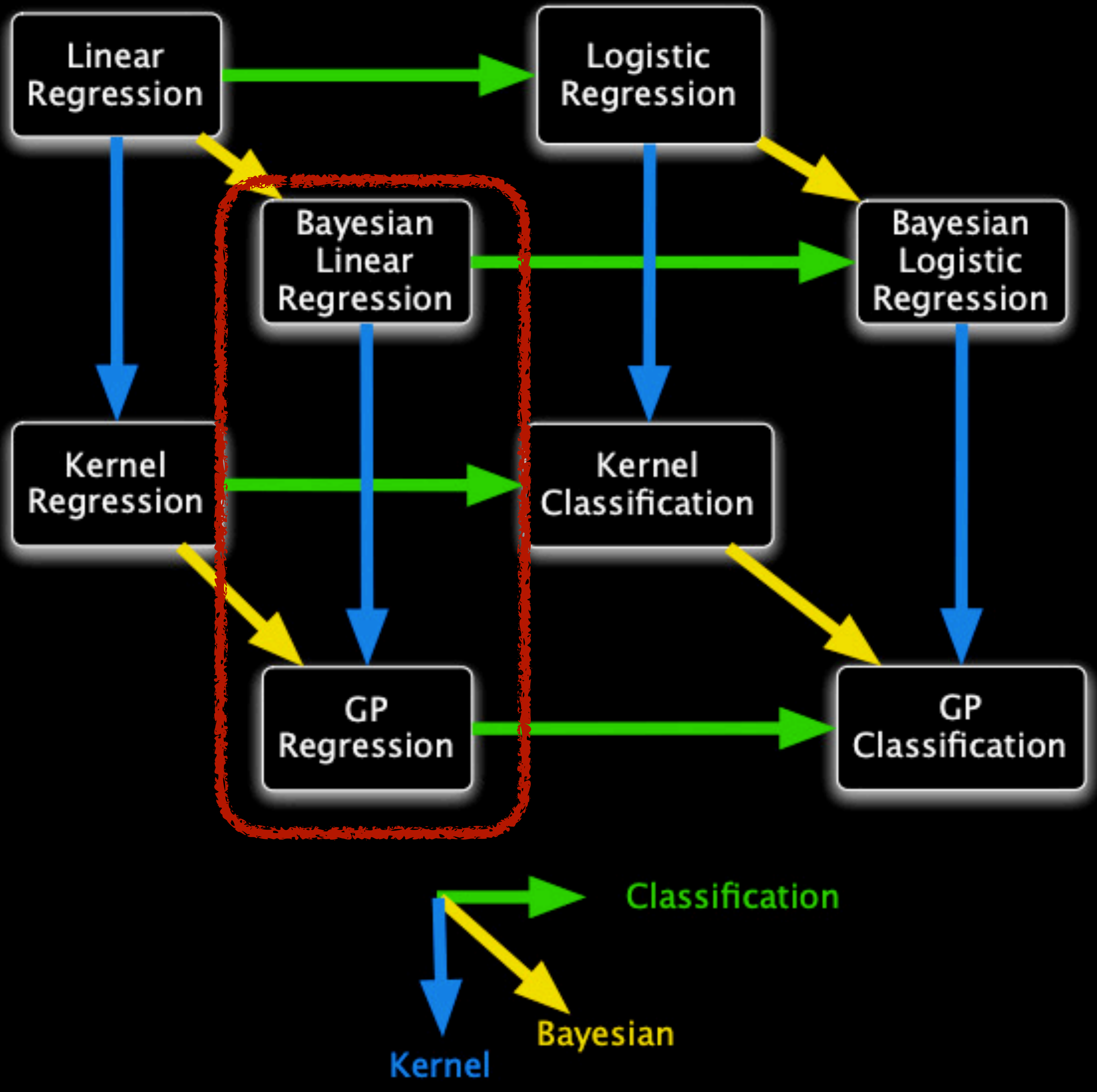
Gaussian process regression (part 2)

Week 5a - Statistical ML / Thang Bui / ANU / 2025 S1 - Thanks to Yuan-Sen Ting for many slides

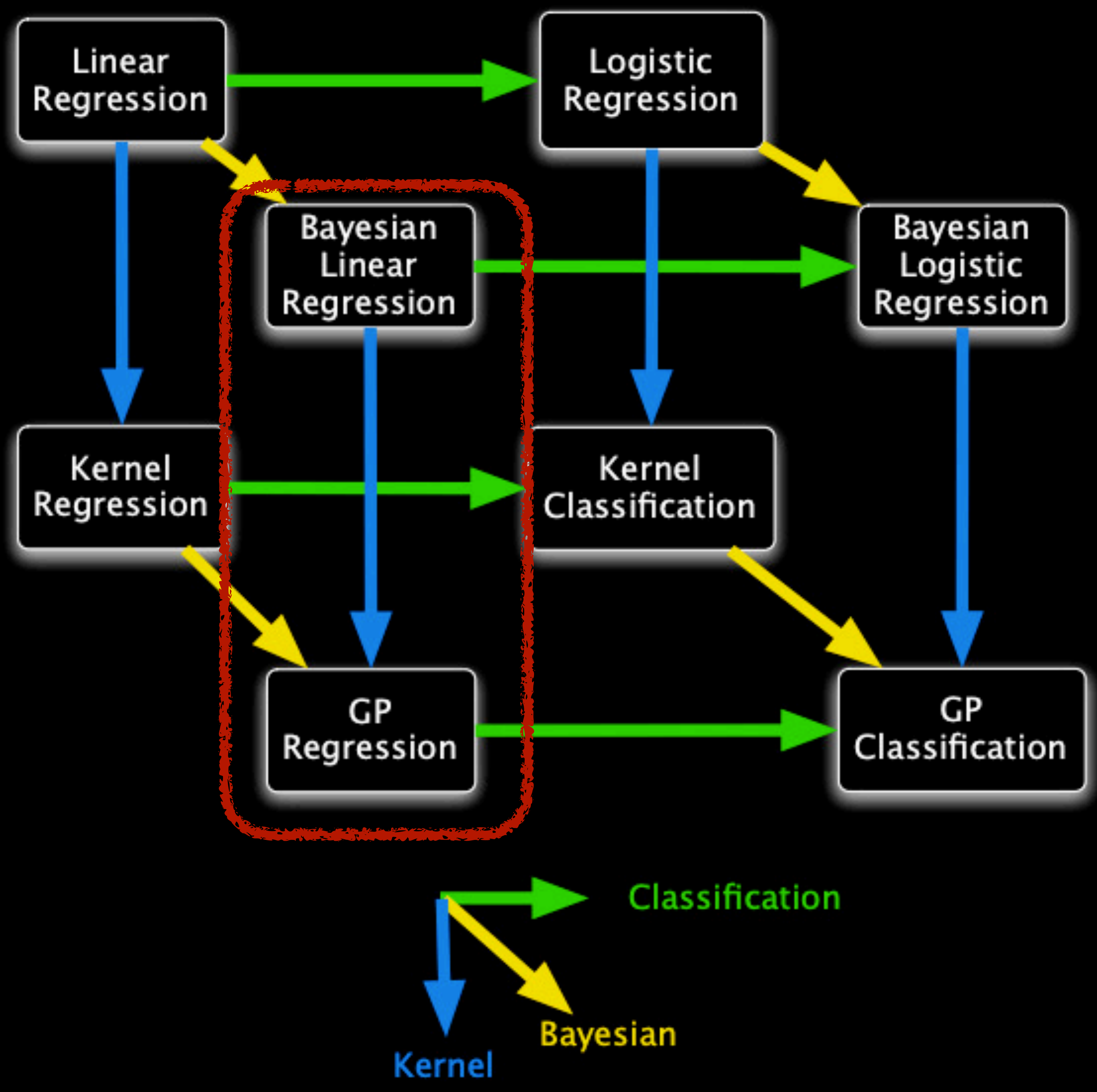
Gaussian Process - Weight-Space Perspective



Gaussian Process - Weight-Space Perspective

Gaussian Process =

Kernelising Bayesian
Linear Regression



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

Why Kernel ?

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

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

e.g., $k(\mathbf{x}_\mathrm{m}, \mathbf{x}_\mathrm{n}) = (\mathbf{x}_\mathrm{m}^\mathrm{T}\mathbf{x}_\mathrm{n})^2$

Why Kernel ?

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

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

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)

Why Kernel ?

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

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

In 2D

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

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

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)

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)

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

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)

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

(Bishop eq 6.23, GP Book eq 2.16)

➡ Infinite dimensional features

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

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

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

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

(Bishop eq 3.53)

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

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

(Bishop eq 3.54)

Gaussian Process

Kernelised

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

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

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

(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

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

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

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

(Bishop eq 3.53)

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

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

(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)

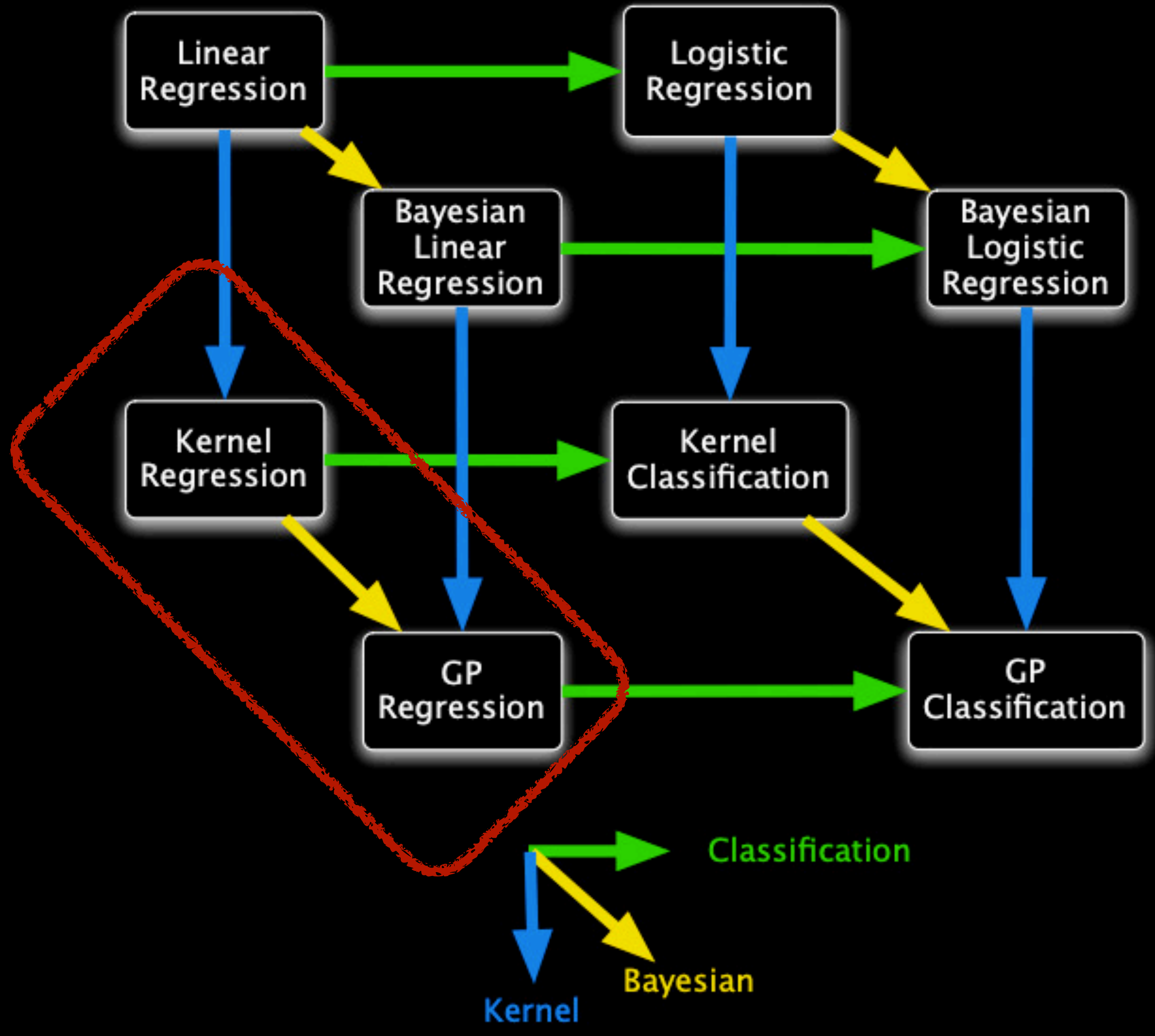
$$\sigma^2(\mathbf{x}^*) = \sigma^2 + \phi(\mathbf{x}^*)^{\mathrm{T}} \mathbf{\Sigma} \, \phi(\mathbf{x}^*)$$

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

(Bishop eq 3.54)

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

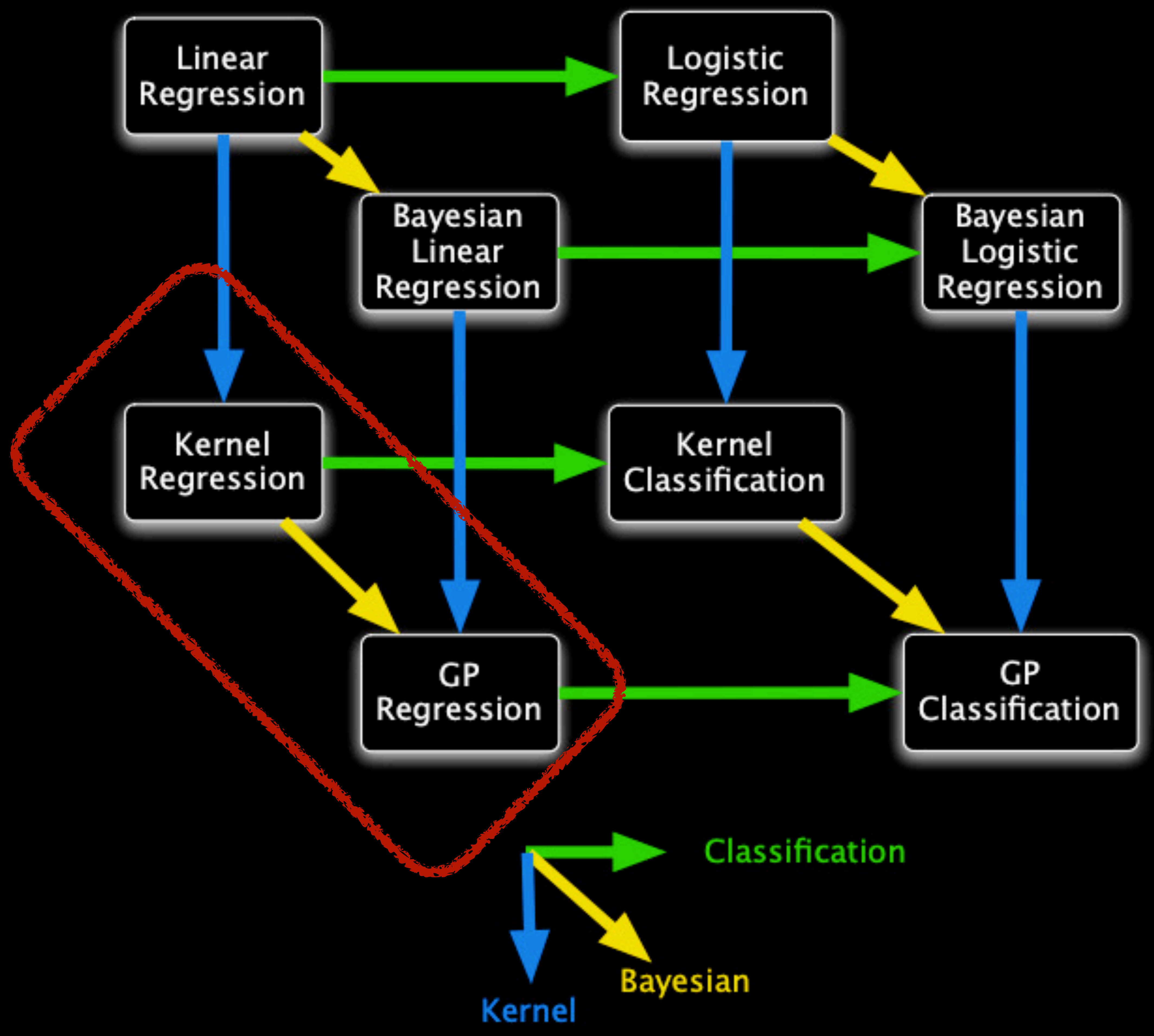
Gaussian Process - Function-Space Perspective



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.

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

(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.

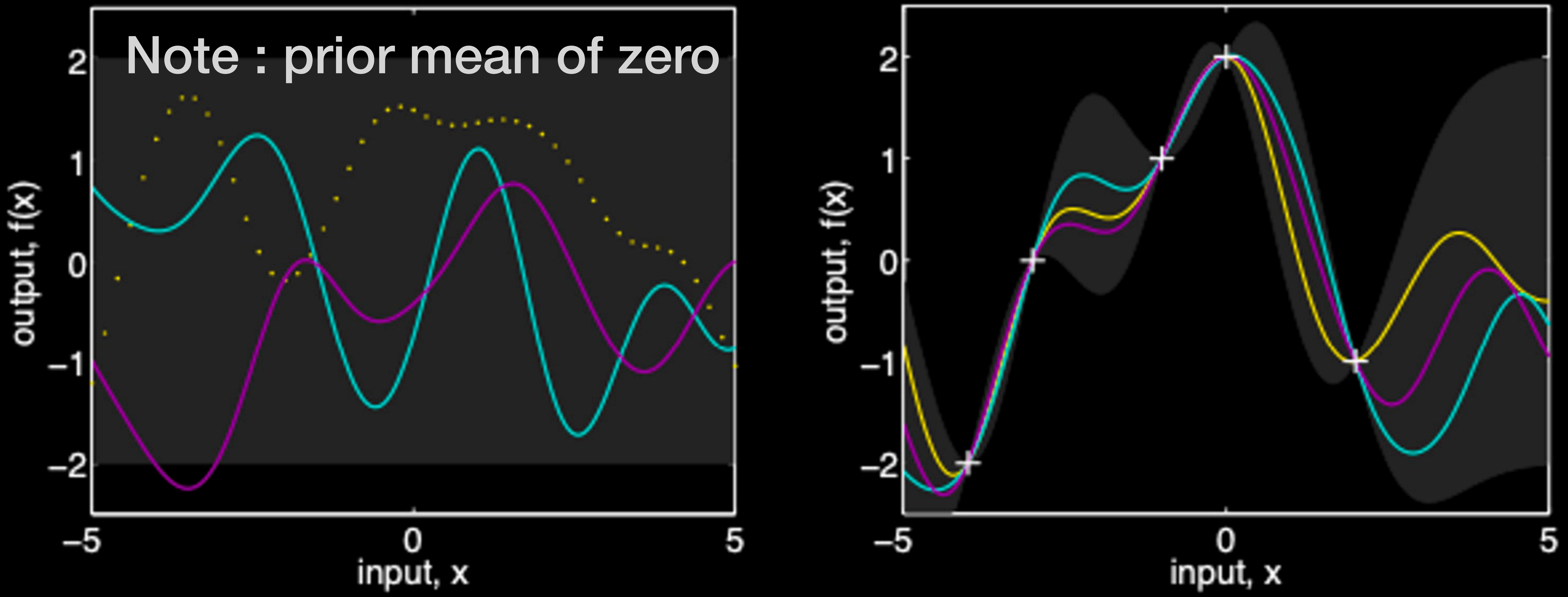
$$\Rightarrow \quad p\left(\mathbf{f}(\mathbf{x})\right) = \mathcal{N}(\mathbf{f}; \underbrace{\mathbf{0}}_{\text{Assuming the prior "weight" } p(w) \text{ to have mean zero}}, \mathbf{K})$$

(Bishop eq 6.60, GP Book eq 2.17)

Assuming the prior “weight”
$p(w)$ to have mean zero

Function-Space Perspective

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



(a), prior (b), posterior

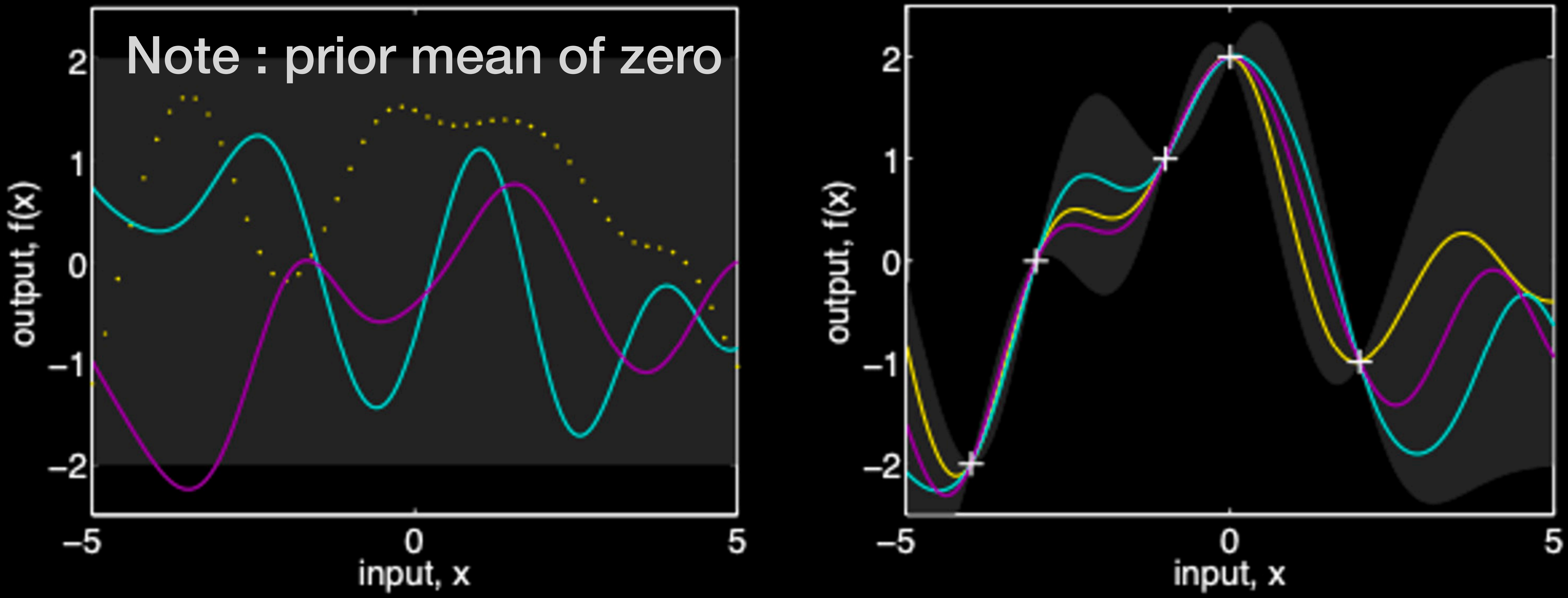
Posterior distribution of functions : “rejecting” functions that do not satisfy the constraints

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

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

Function-Space Perspective

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



(a), prior (b), posterior

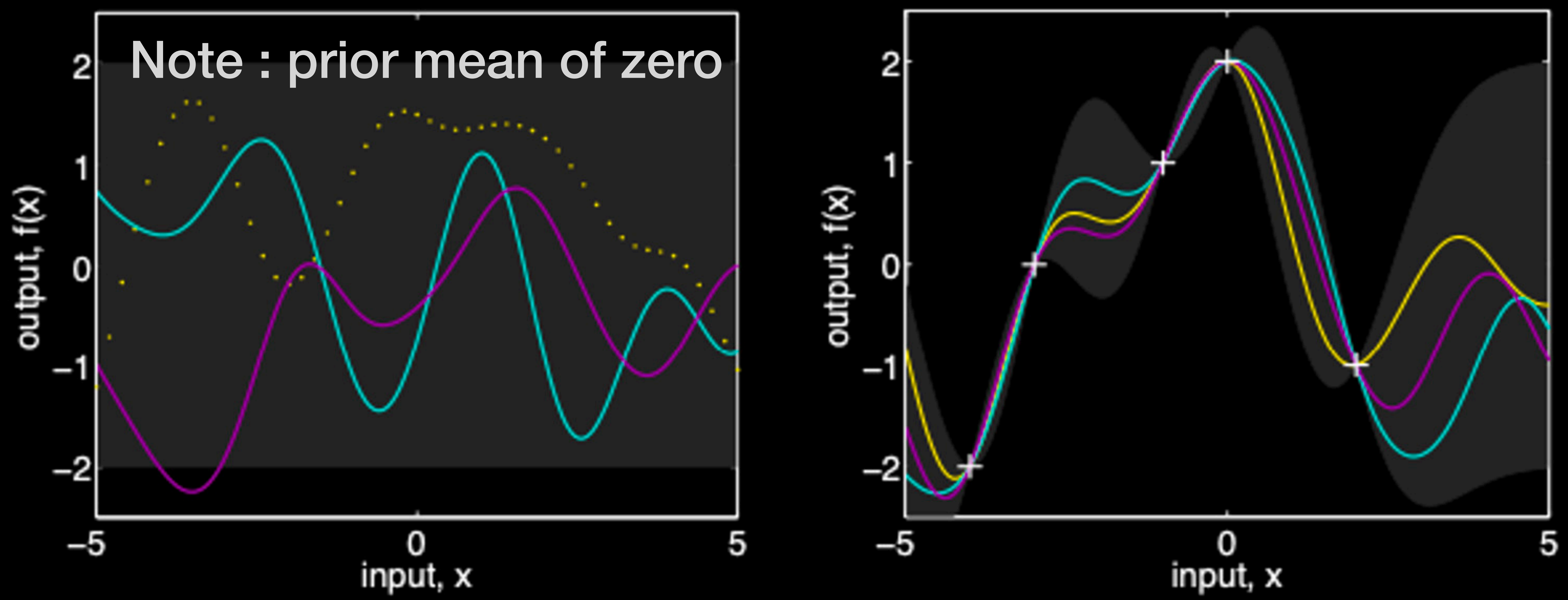
Posterior distribution of functions : “rejecting” functions that do not satisfy the constraints

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

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

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

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

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

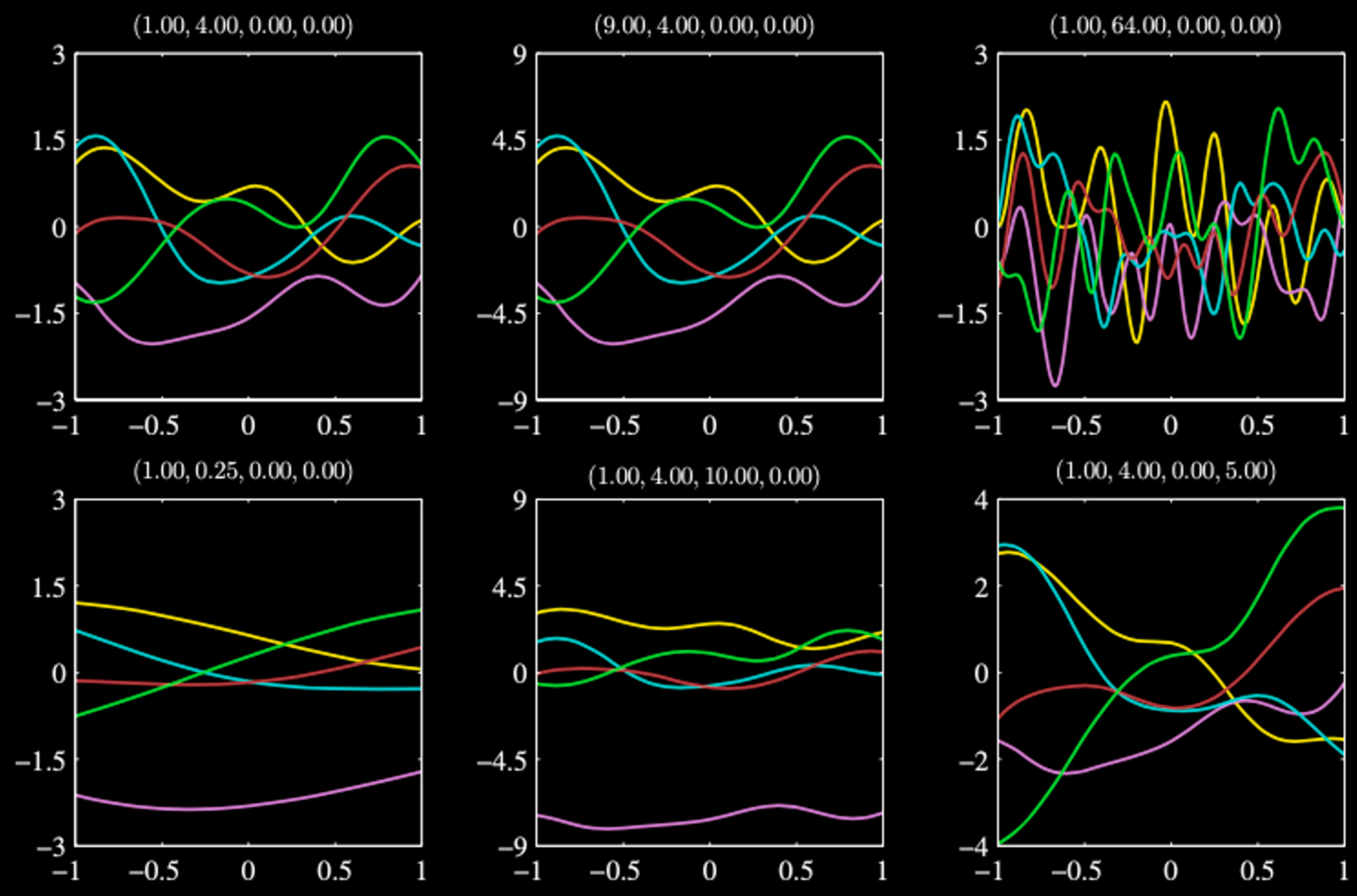


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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)

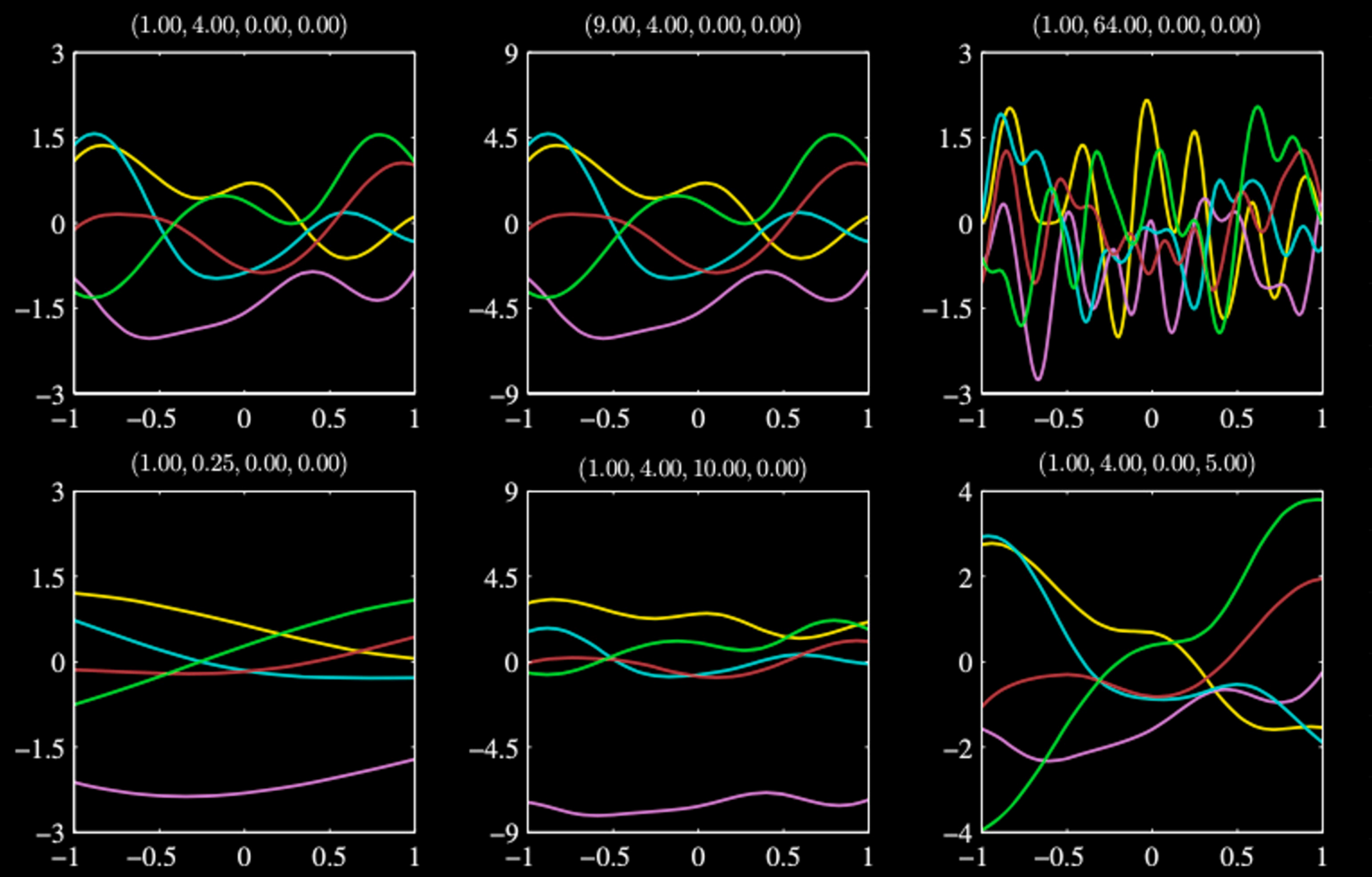


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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

Visualising Kernels

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

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

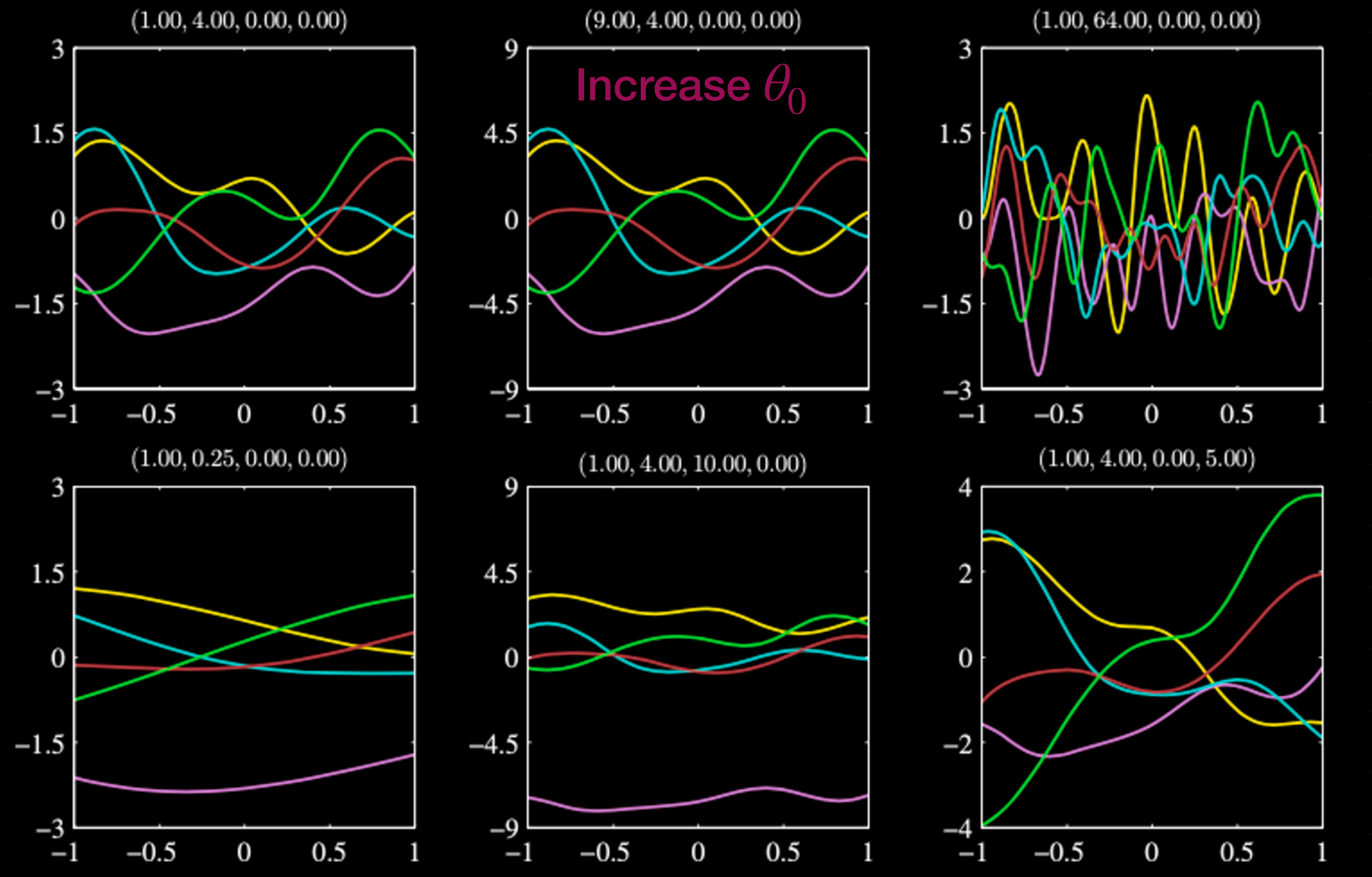


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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)

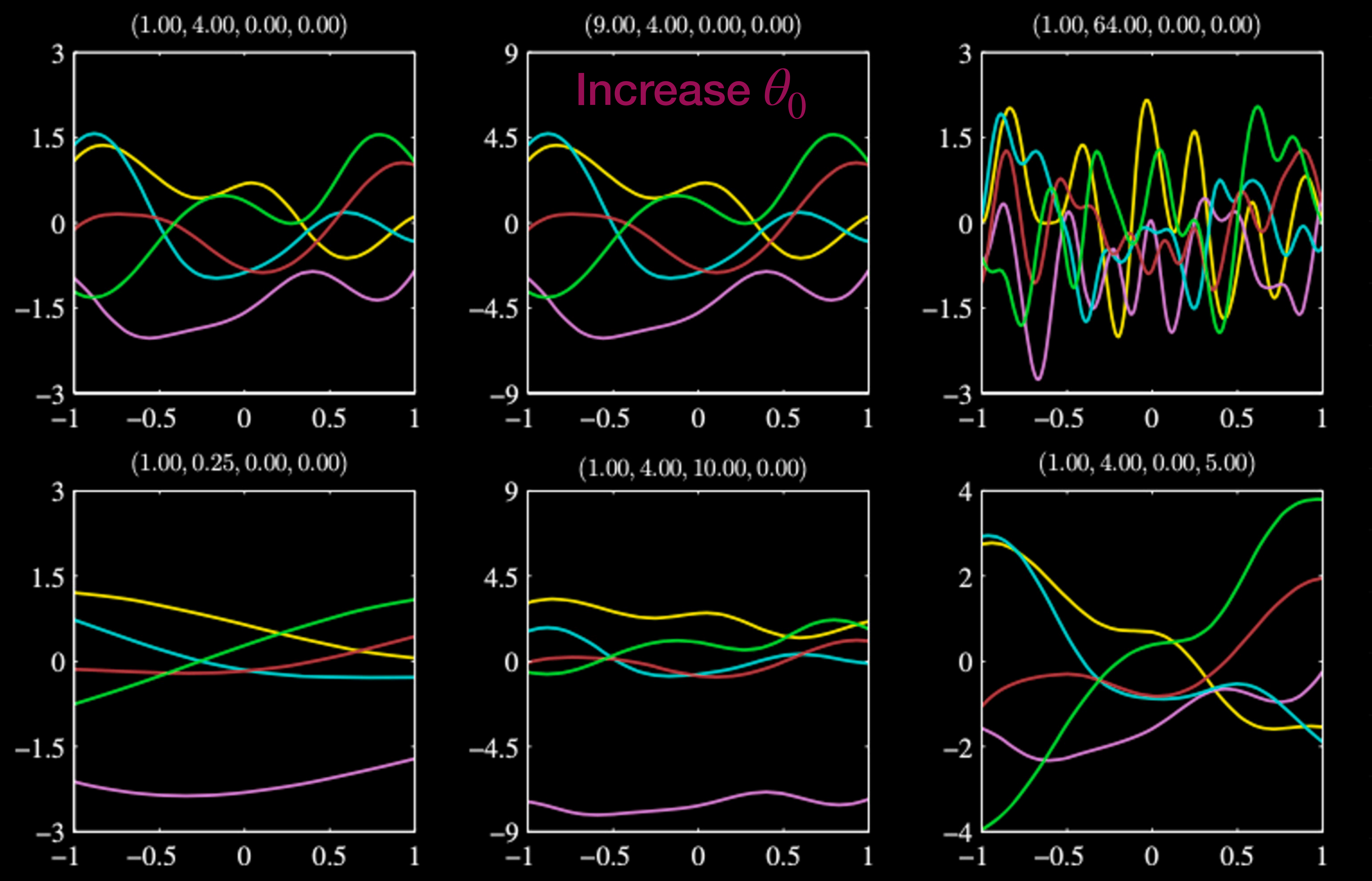


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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)

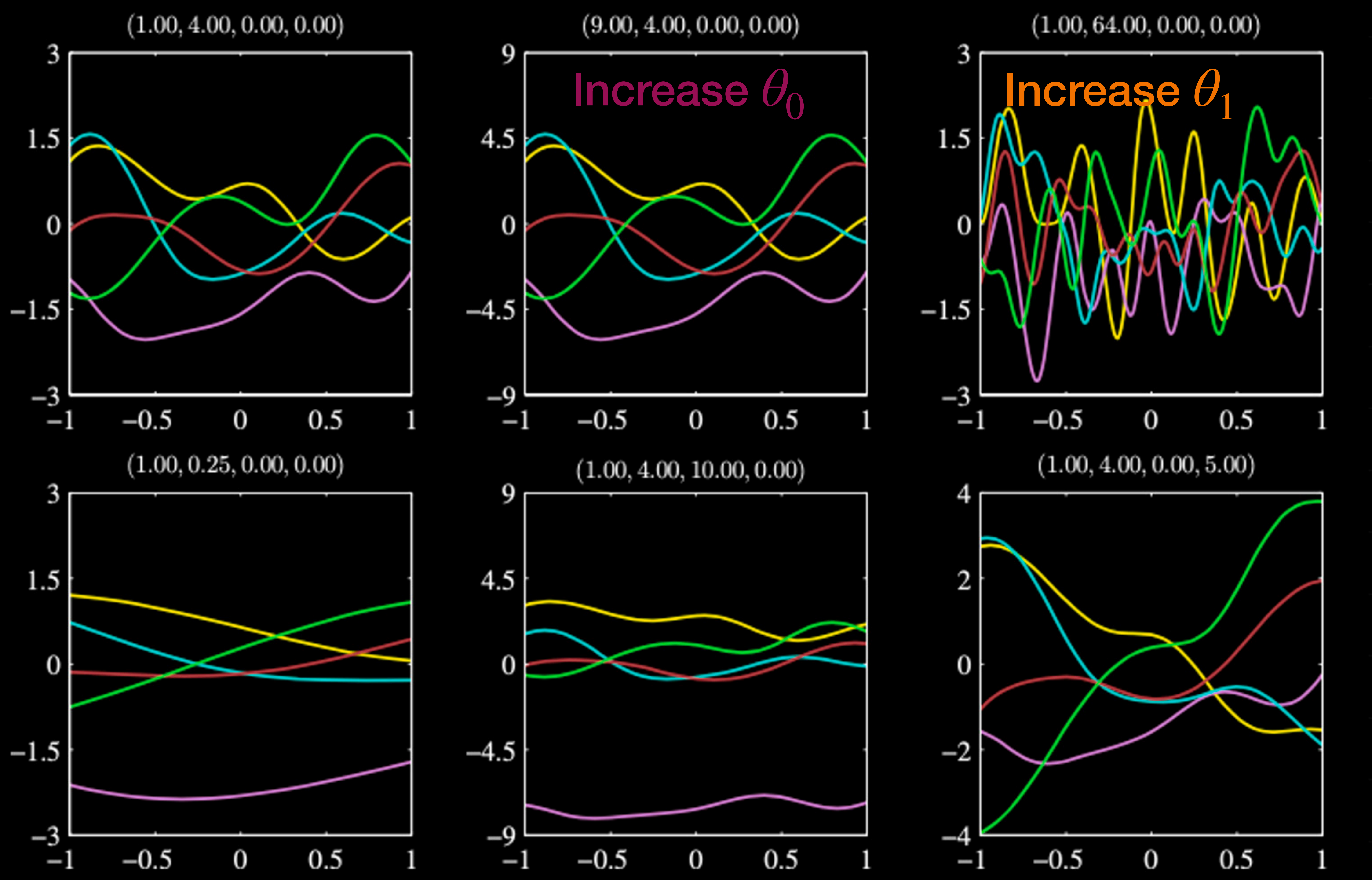


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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

Visualising Kernels

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

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

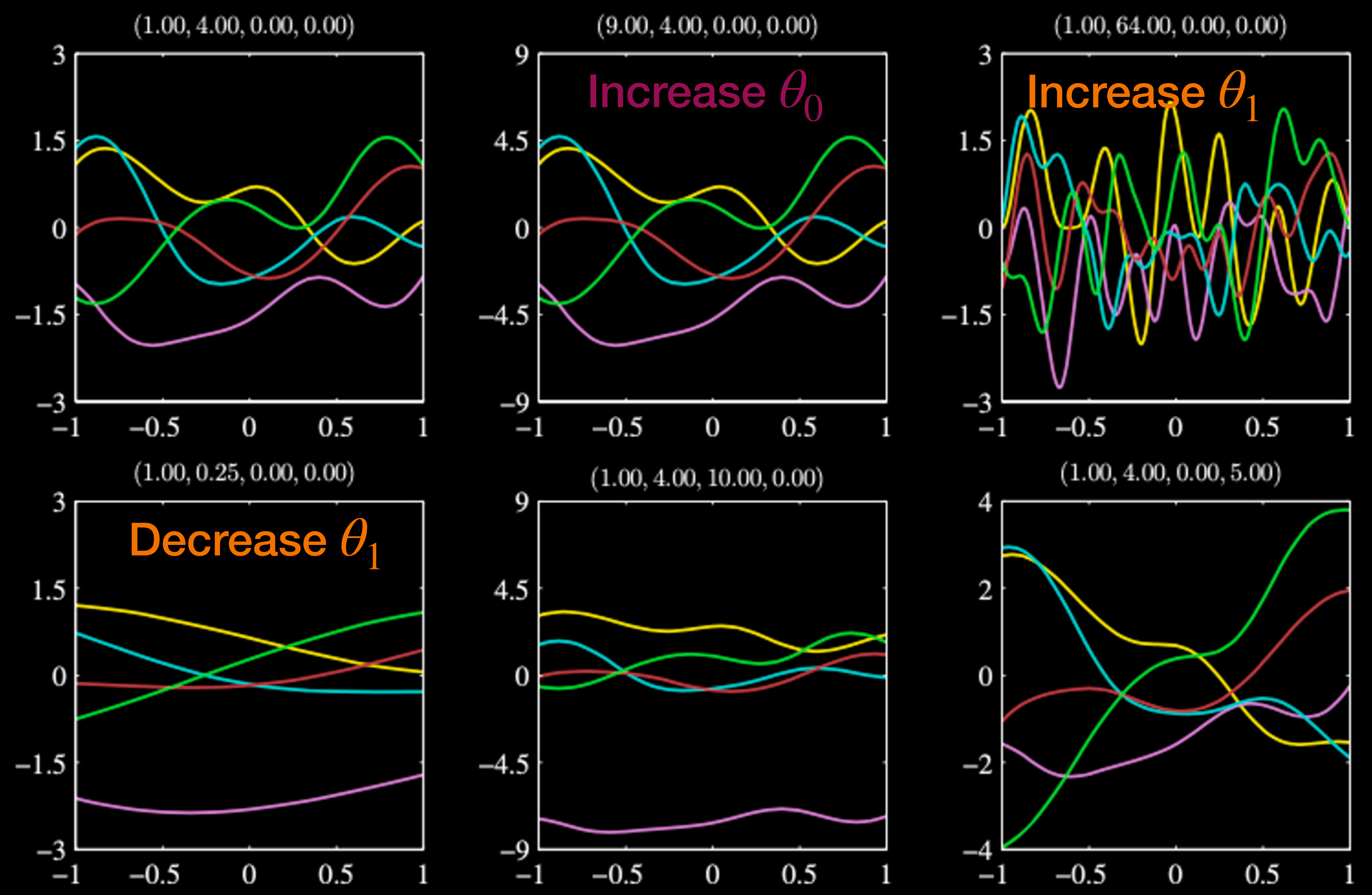


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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)

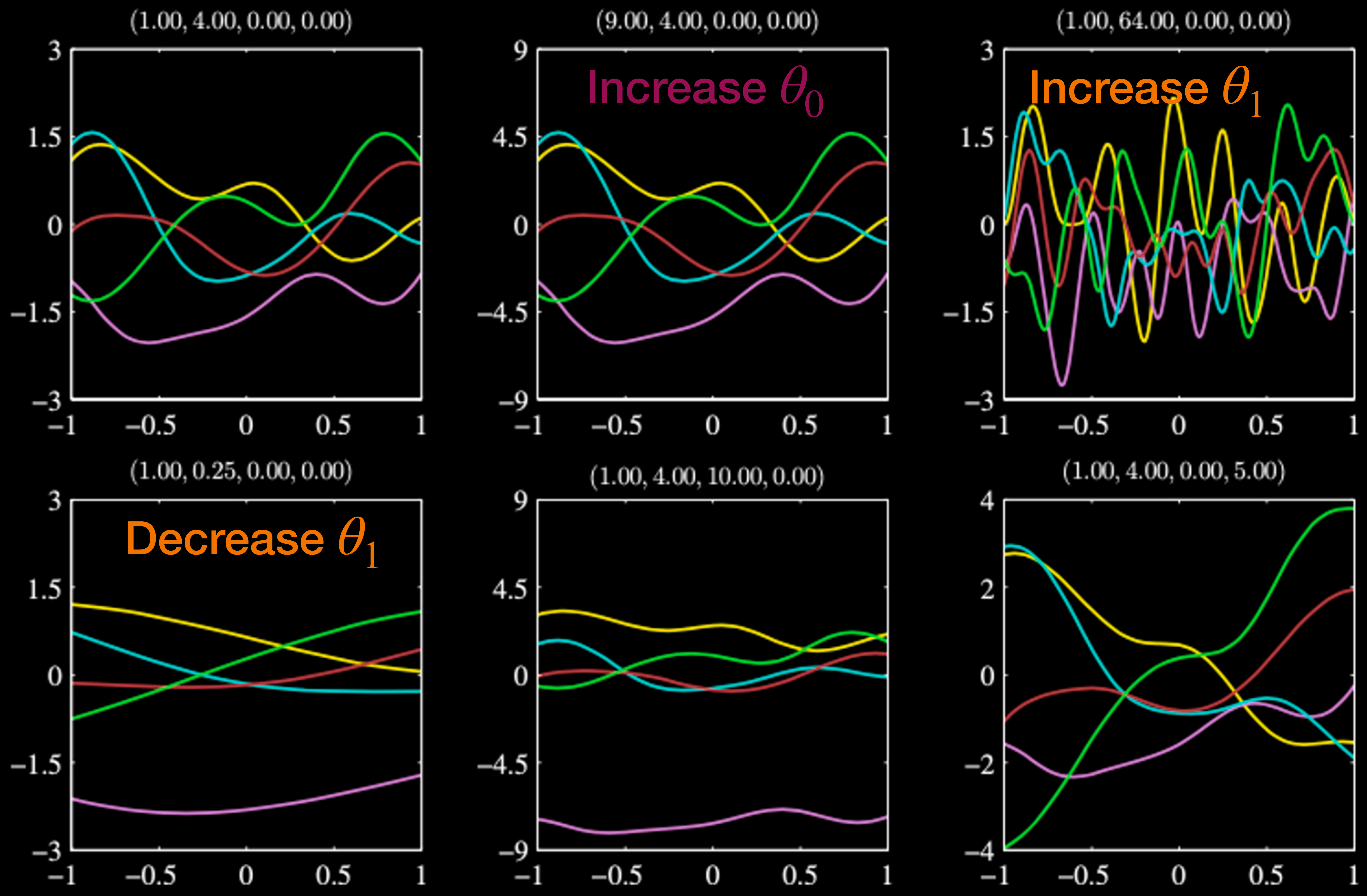


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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

Visualising Kernels

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

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

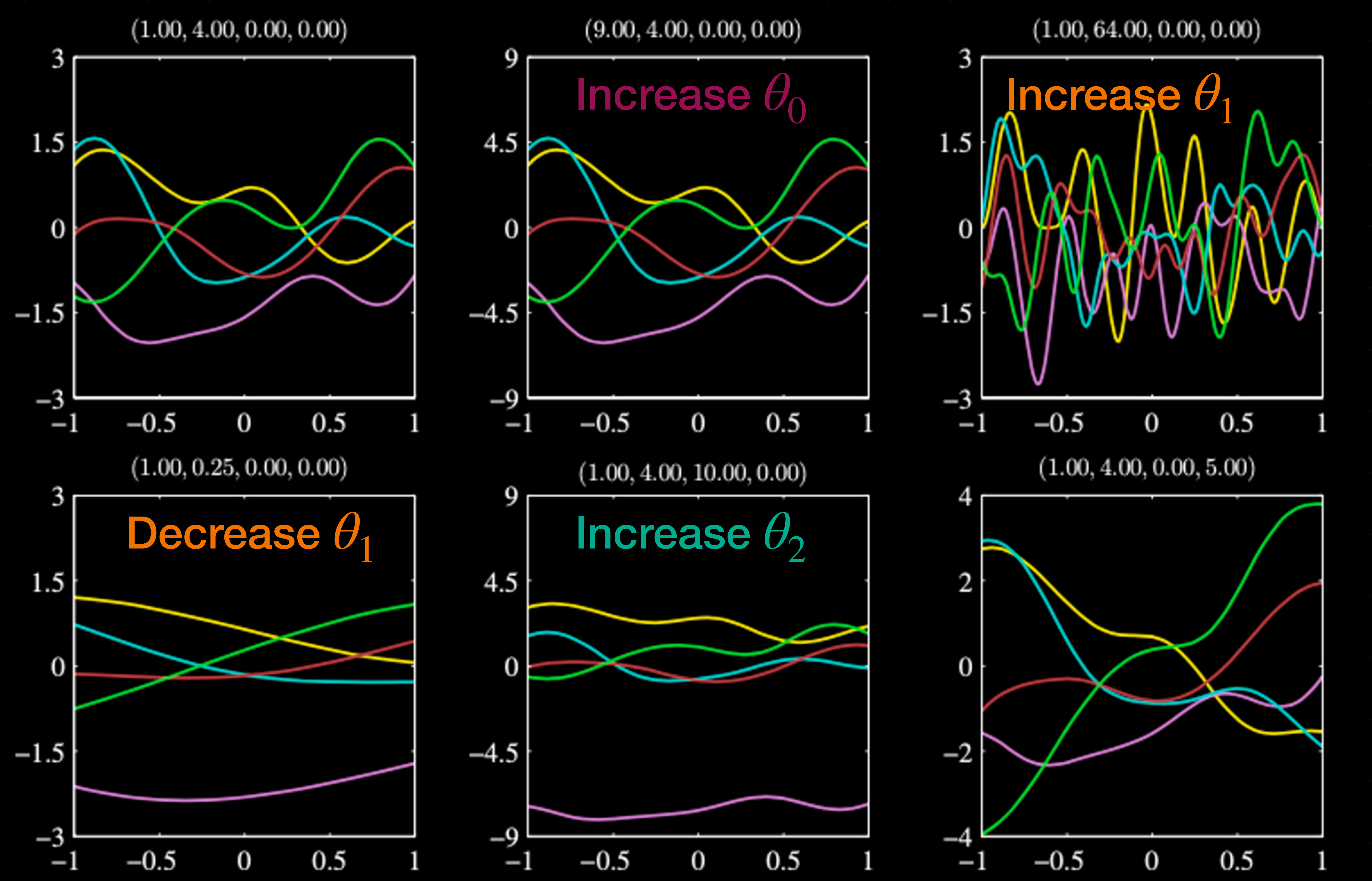


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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)

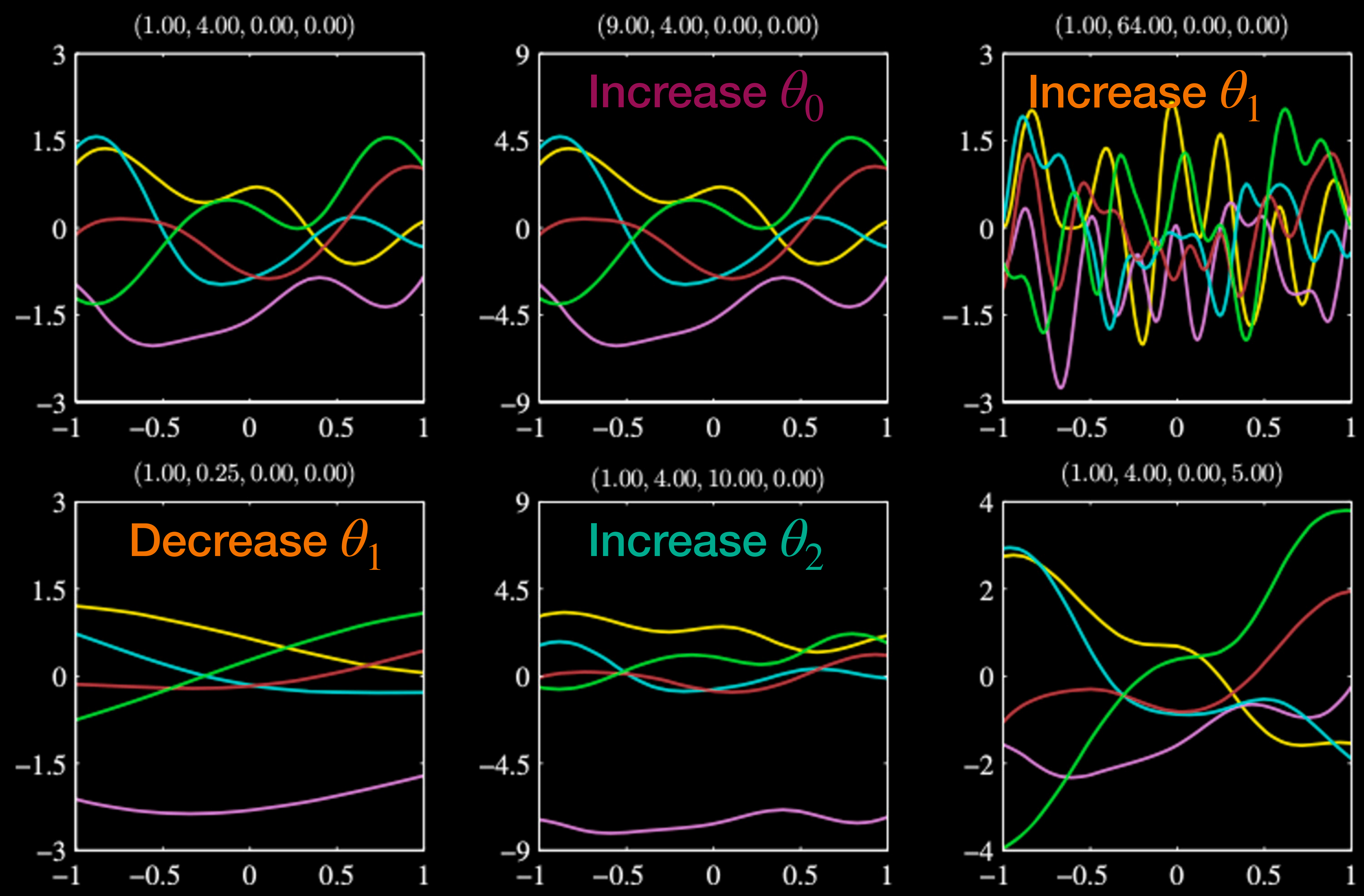


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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

Visualising Kernels

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

Randomly sampling from the Gaussian Process prior distribution

(Bishop textbook)

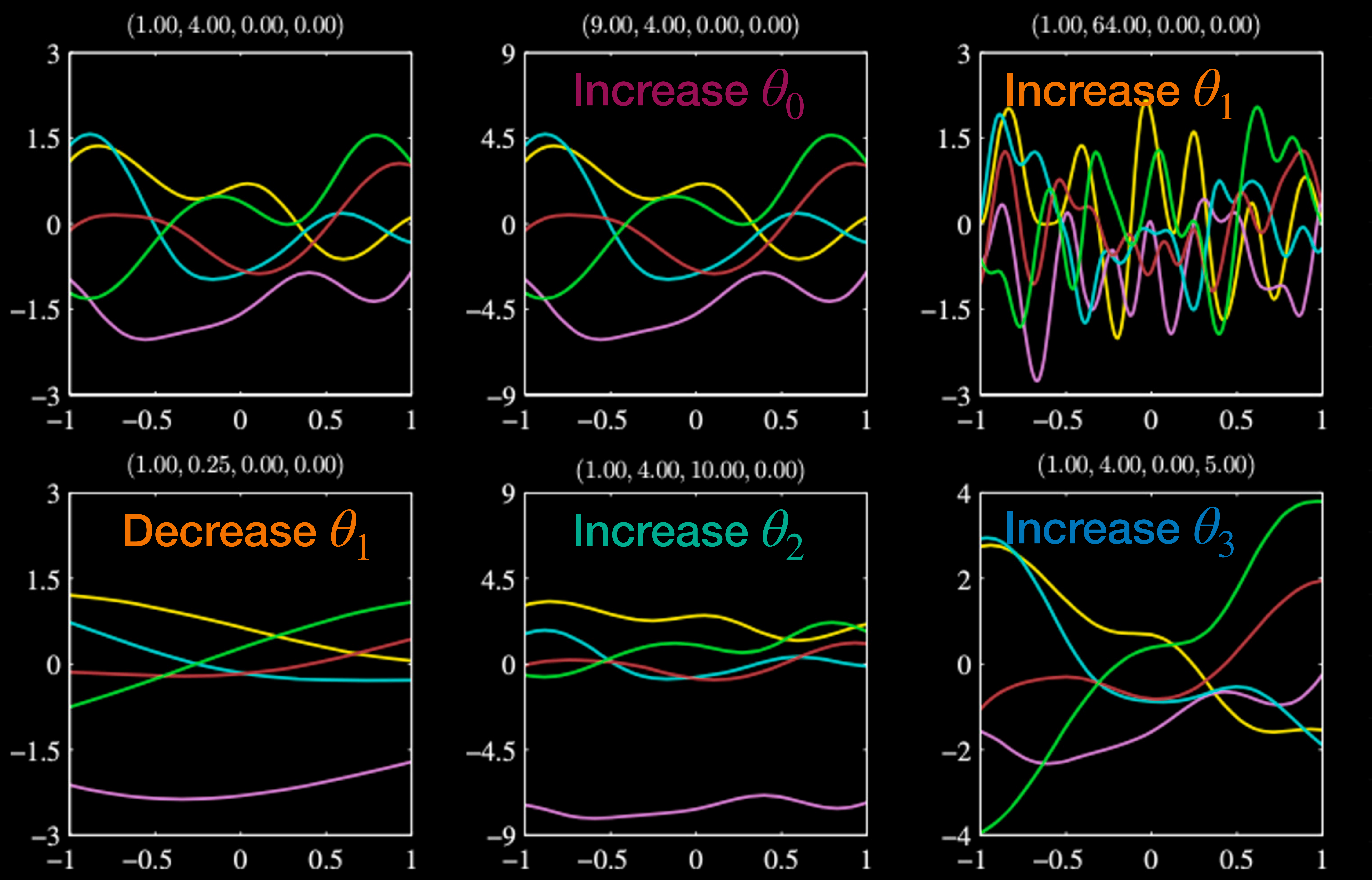


Figure 6.5 Samples from a Gaussian process prior defined by the covariance function (6.63). The title above each plot denotes $(\theta_0, \theta_1, \theta_2, \theta_3)$.

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

Algorithm

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

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

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

Algorithm

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

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

$$\mathcal{O}(N^3)$$

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

- $L =$ Cholesky $(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I}_N)$

$$k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I}_N = \mathbf{L} \cdot \mathbf{L}^T$$

$\mathbf{L}$ lower-triangular

Algorithm

$N^3/3$ operations

- $L = \text{Cholesky}\left(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I}_N\right)$

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

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

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

$\mathcal{O}(N^3)$

$$k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I}_N = \mathbf{L} \cdot \mathbf{L}^T$$

$\mathbf{L}$ lower-triangular

Why Cholesky?

Cost of some matrix factorizations and decompositions. A is $n \times n$, except for QR and SVD, where it is $m \times n$ ($m \geq n$).

Factorization/decomposition	Number of flops
LU factorization with partial pivoting ($PA = LU$)	$2n^3/3$
LU factorization with partial pivoting of upper Hessenberg matrix ($PA = LU$)	n^2
Cholesky factorization ($A = R^*R$)	$\boxed{n^3/3}$
Householder QR factorization ($A = QR$)	$2n^2(m - n/3)$ for R; $4(m^2n - mn^2 + n^3/3)$ for $m \times m$ Q; $2n^2(m - n/3)$ for $m \times n$ Q; $2np(2m - n)$ for QB with $m \times p$ B and Q held in factored form.
SVD[a] ($A = P\Sigma Q^*$)	$14mn^2 + 8n^3$ ($P(:, 1{:}n)$, Σ, and Q)[b] $6mn^2 + 20n^3$ ($P(:, 1{:}n)$, Σ, and Q)[c]
Hessenberg decomposition ($A = QHQ^*$)	$14n^3/3$ (Q and H), $10n^3/3$ (H only)
Schur decomposition[a] ($A = QTQ^*$)	$25n^3$ (Q and T), $10n^3$ (T only)
For Hermitian A:	
Tridiagonal reduction ($A = QTQ^*$)	$8n^3/3$ (Q and T), $4n^3/3$ (T only)
Spectral decomposition ($A = QDQ^*$)	$9n^3$ (Q and D), $4n^3/3$ (D only)

https://math.stackexchange.com/questions/2840755/what-is-the-computation-time-of-lu-cholesky-and-qr-decomposition

Algorithm

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

$$m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X}) \boxed{(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})^{-1}\,\mathbf{y}} \quad \mathcal{O}(N^3)$$

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

$N^3/3$ operations

- $L =$ Cholesky $(k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I})$
- $\boxed{\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \setminus (\mathbf{L} \setminus \mathbf{y})}$
- $m(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{X}) \cdot \boldsymbol{\alpha}$
- $\boxed{\mathbf{v} = \mathbf{L} \setminus k(\mathbf{X}, \mathbf{x}^*)}$
- $\sigma^2(\mathbf{x}^*) = k(\mathbf{x}^*, \mathbf{x}^*) - \boxed{\mathbf{v}^{\mathrm{T}}\mathbf{v}}$

$$k(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathrm{I} = \mathbf{L} \cdot \mathbf{L}^T$$

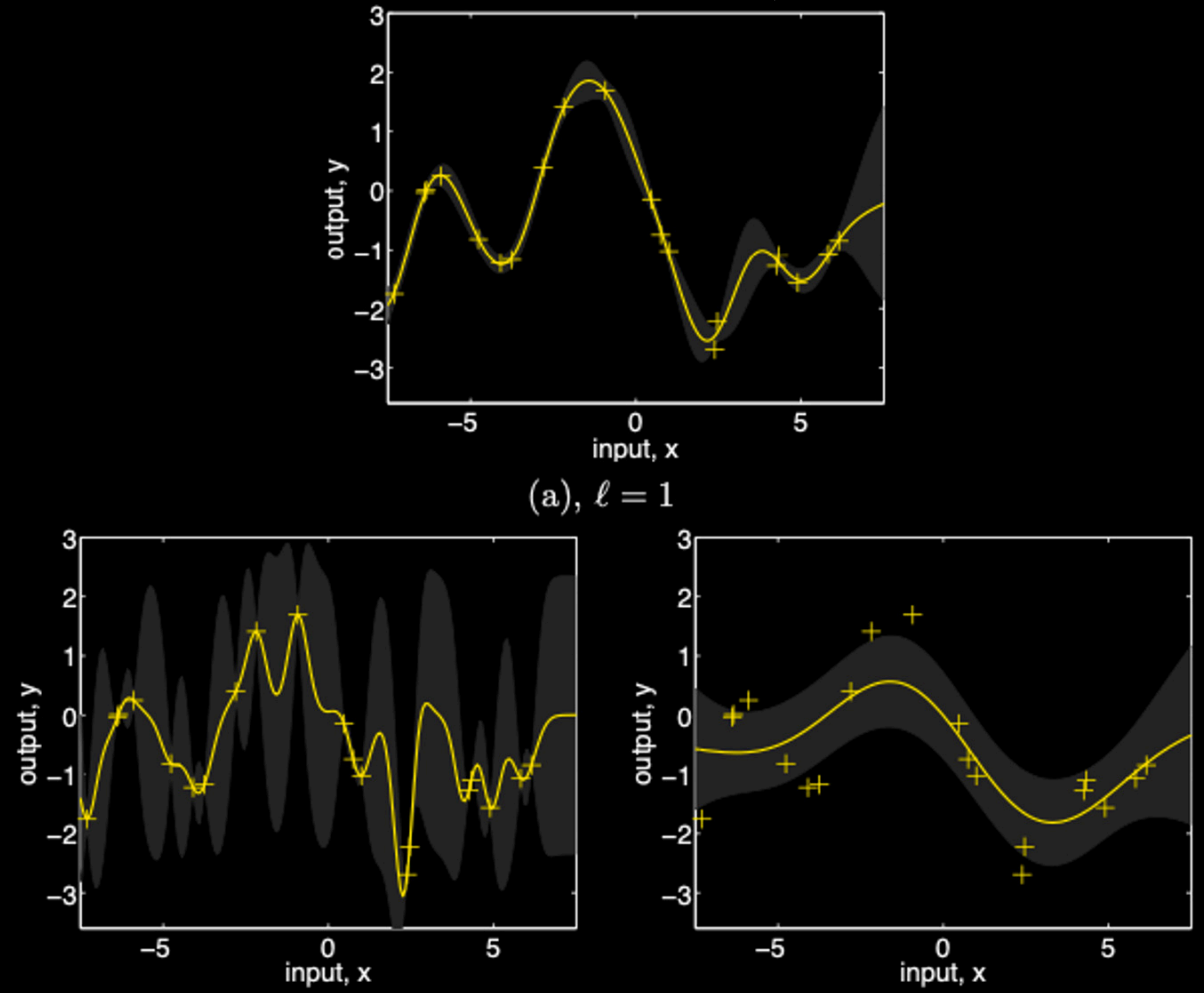
$\mathbf{L}$ lower-triangular

$$\mathbf{A}\mathbf{x} = \mathbf{b} \Rightarrow \mathbf{x} = \mathbf{A} \setminus \mathbf{b}$$

Learning Hyperparameters

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

How to choose the characteristic length / smoothness ?



(a), $\ell = 1$

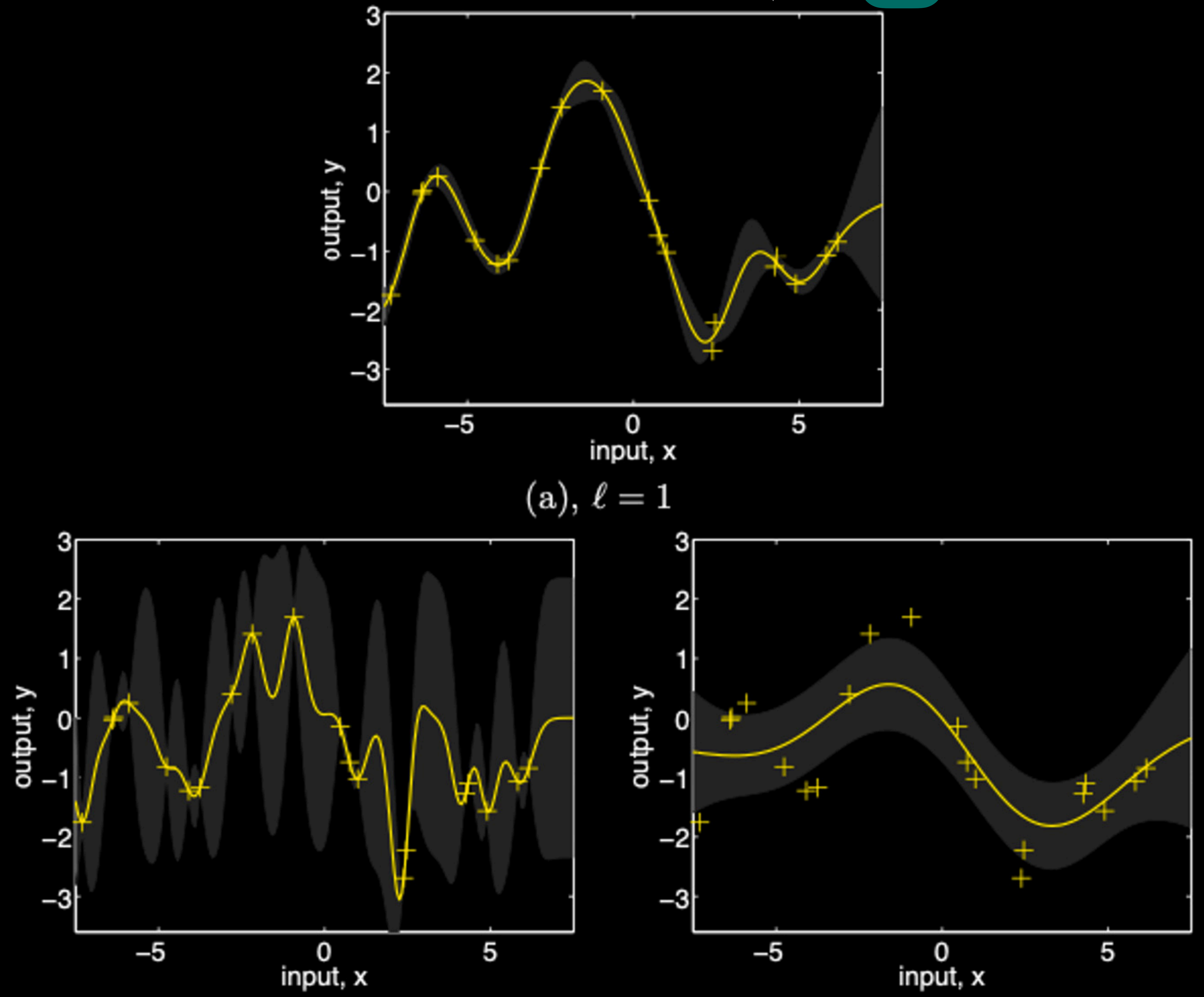
(b), $\ell = 0.3$

(c), $\ell = 3$

Learning Hyperparameters

$$k(\mathbf{x}_n, \mathbf{x}_m) = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right) + \sigma_n^2 \delta_{nm}$$

How to choose the characteristic length / smoothness ?



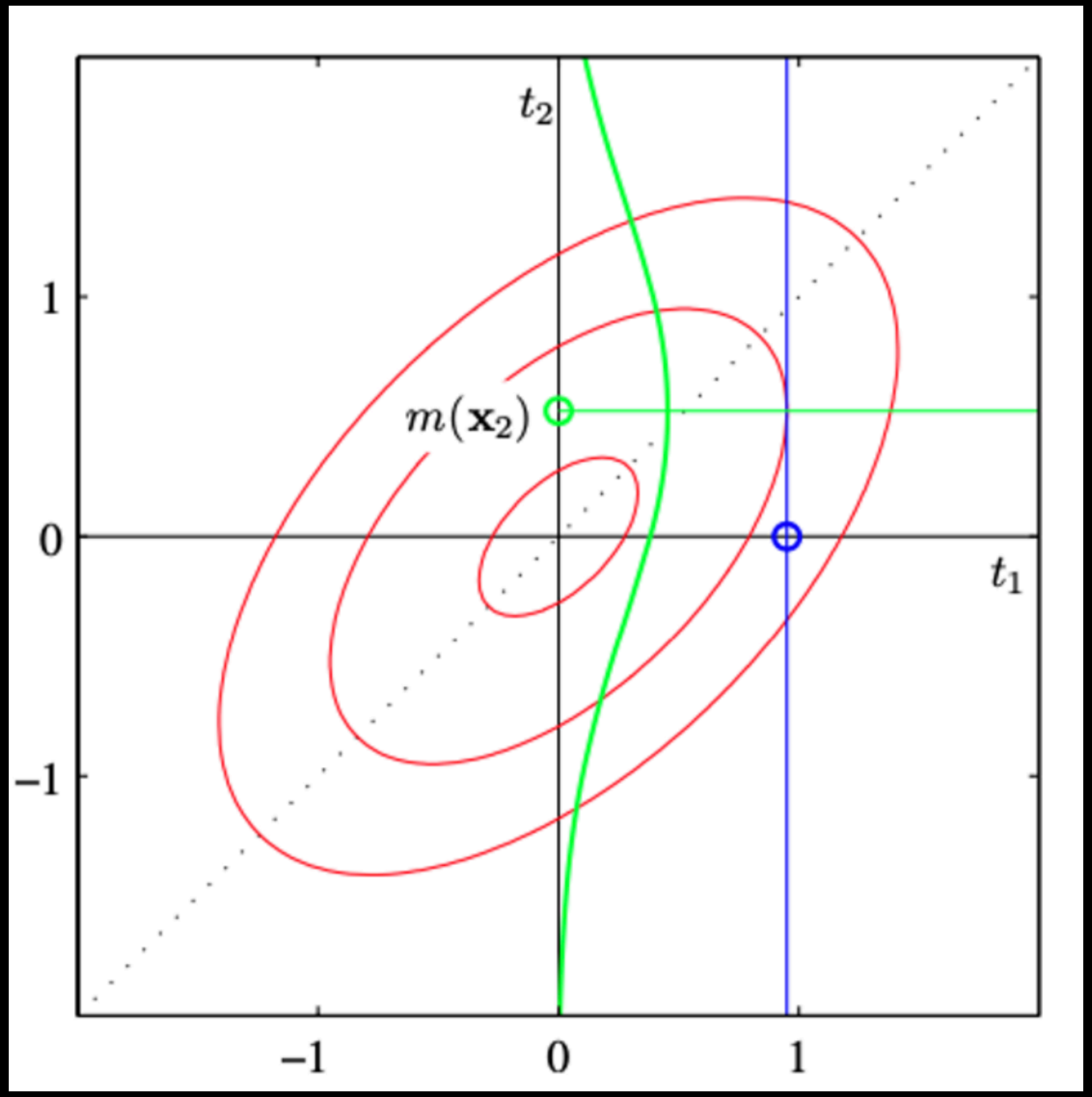
(a), $\ell = 1$

(b), $\ell = 0.3$

(c), $\ell = 3$

Learning Hyperparameters

$$\underbrace{p(\mathbf{y}\,|\,\mathbf{X},}_{\text{the marginal likelihood}}\overbrace{\theta}^{\text{Kernel hyperparameter}}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X},\mathbf{X})}_{\hat{\mathbf{K}}} + \sigma^2 \mathrm{I})$$

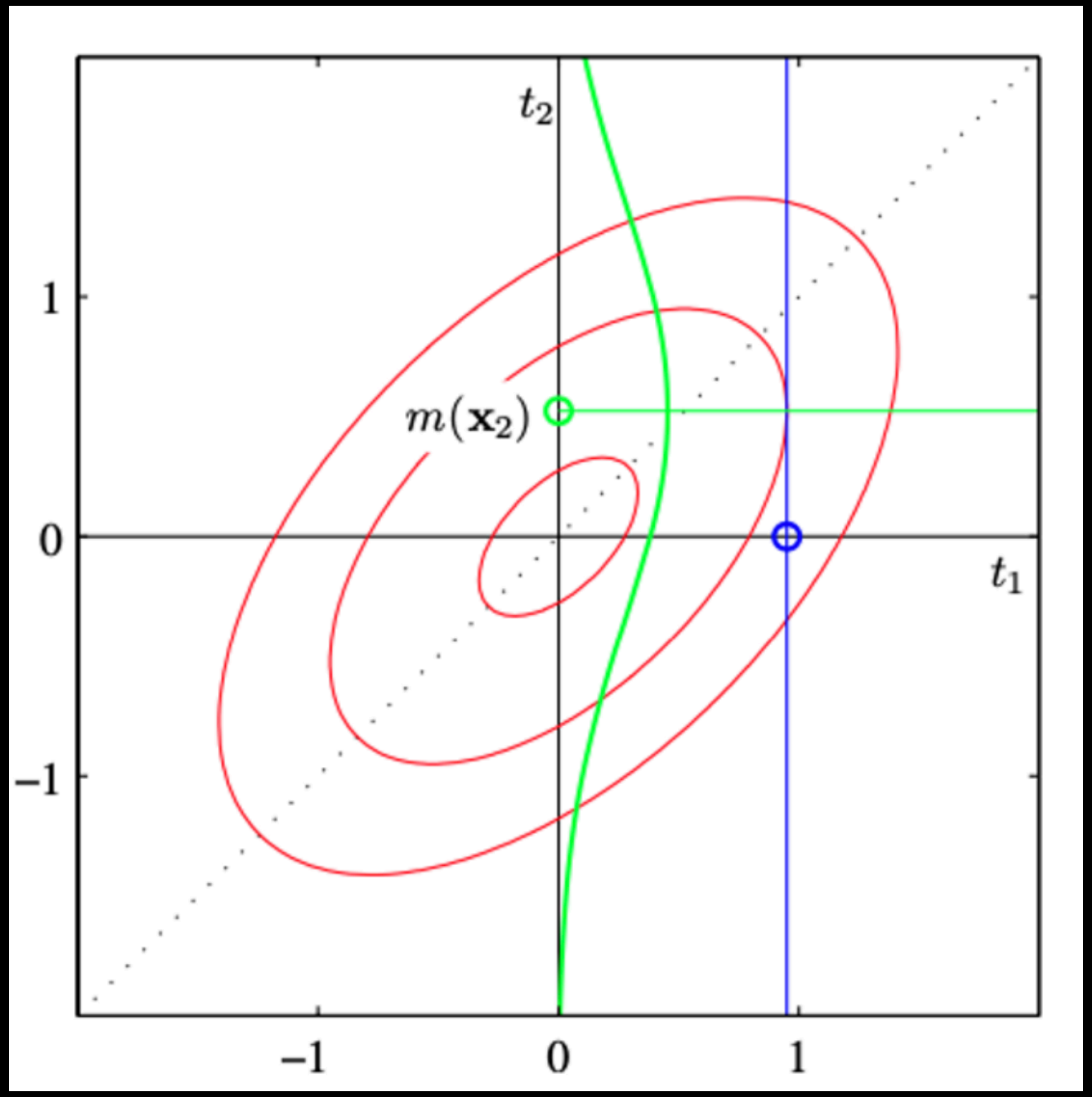


$$\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = -\sum_{\mathbf{i}} \ln \mathbf{L}_{\mathbf{ii}} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\alpha - \frac{\mathbf{N}}{2}\ln(2\pi)$$

$$L = \text{Cholesky } \hat{\mathbf{K}}$$
$$\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \backslash (\mathbf{L} \backslash \mathbf{y})$$

Learning Hyperparameters

$$\underbrace{p(\mathbf{y} \mid \mathbf{X}, \overbrace{\theta}^{\text{Kernel hyperparameter}})}_{\text{the marginal likelihood}} = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}}_{\hat{\mathbf{K}}})$$



$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \ln |\hat{\mathbf{K}}| - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \hat{\mathbf{K}}^{-1} \mathbf{y} - \frac{\mathrm{N}}{2} \ln(2\pi)$$

(Bishop eq 6.69)

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\sum_{\mathrm{i}} \ln \mathrm{L}_{\mathrm{ii}} - \frac{1}{2} \mathbf{y}^{\mathrm{T}} \alpha - \frac{\mathrm{N}}{2} \ln(2\pi)$$

$L =$ Cholesky $\hat{\mathbf{K}}$

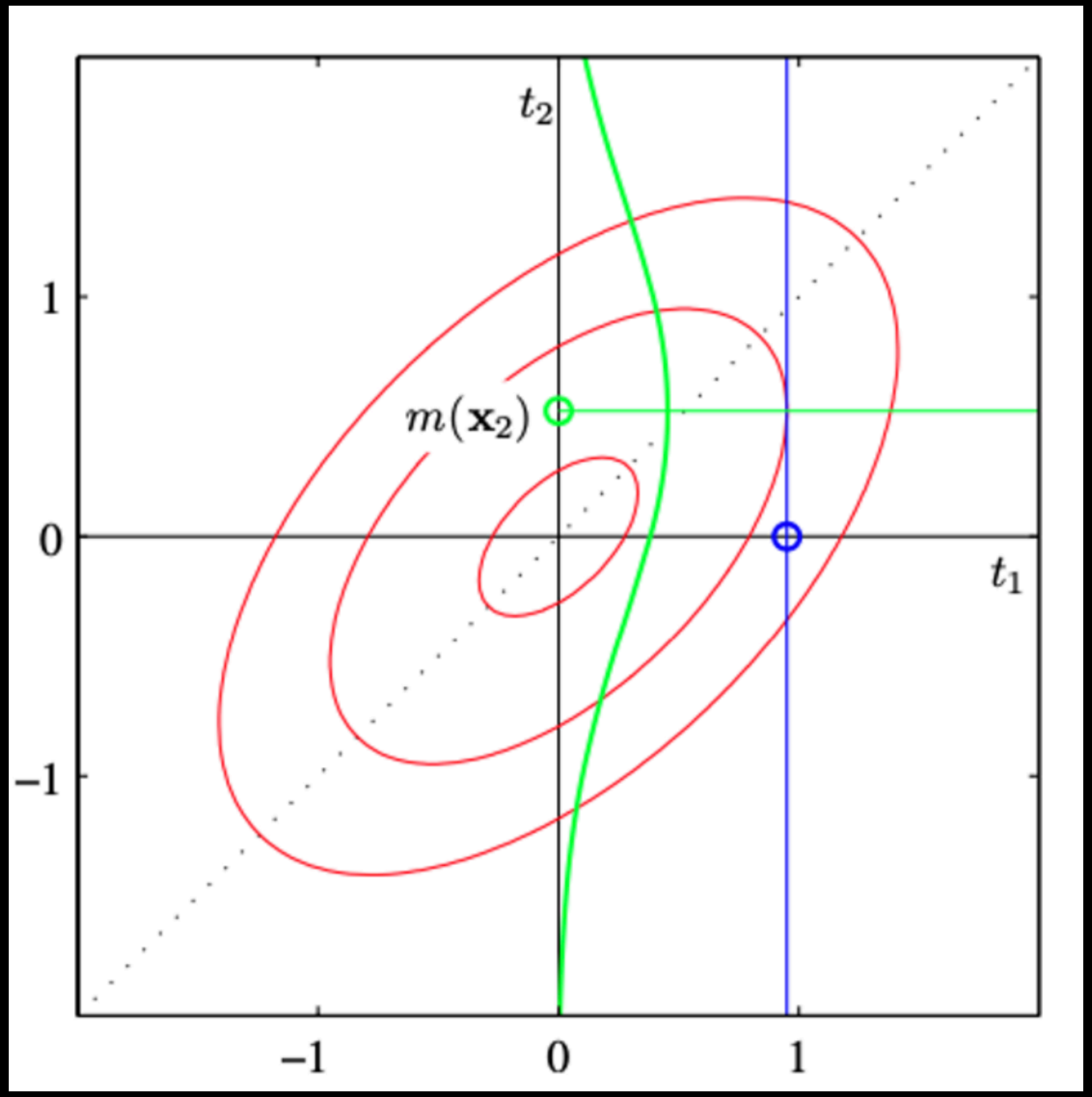
$\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \backslash (\mathbf{L} \backslash \mathbf{y})$

Learning Hyperparameters

$$\underbrace{p(\mathbf{y} \mid \mathbf{X}, \overbrace{\theta}^{\text{Kernel hyperparameter}})}_{\text{the marginal likelihood}} = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}}_{\hat{\mathbf{K}}})$$

$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \ln |\hat{\mathbf{K}}| - \frac{1}{2} \mathbf{y}^T \hat{\mathbf{K}}^{-1} \mathbf{y} - \frac{N}{2} \ln(2\pi)$$

(Bishop eq 6.69)



$$\frac{\partial}{\partial x} \ln |\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x}\right)$$

$$\frac{\partial}{\partial x} (\mathbf{A}^{-1}) = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1}$$

(Bishop eq C.21, C.22)

Learning through gradient descent

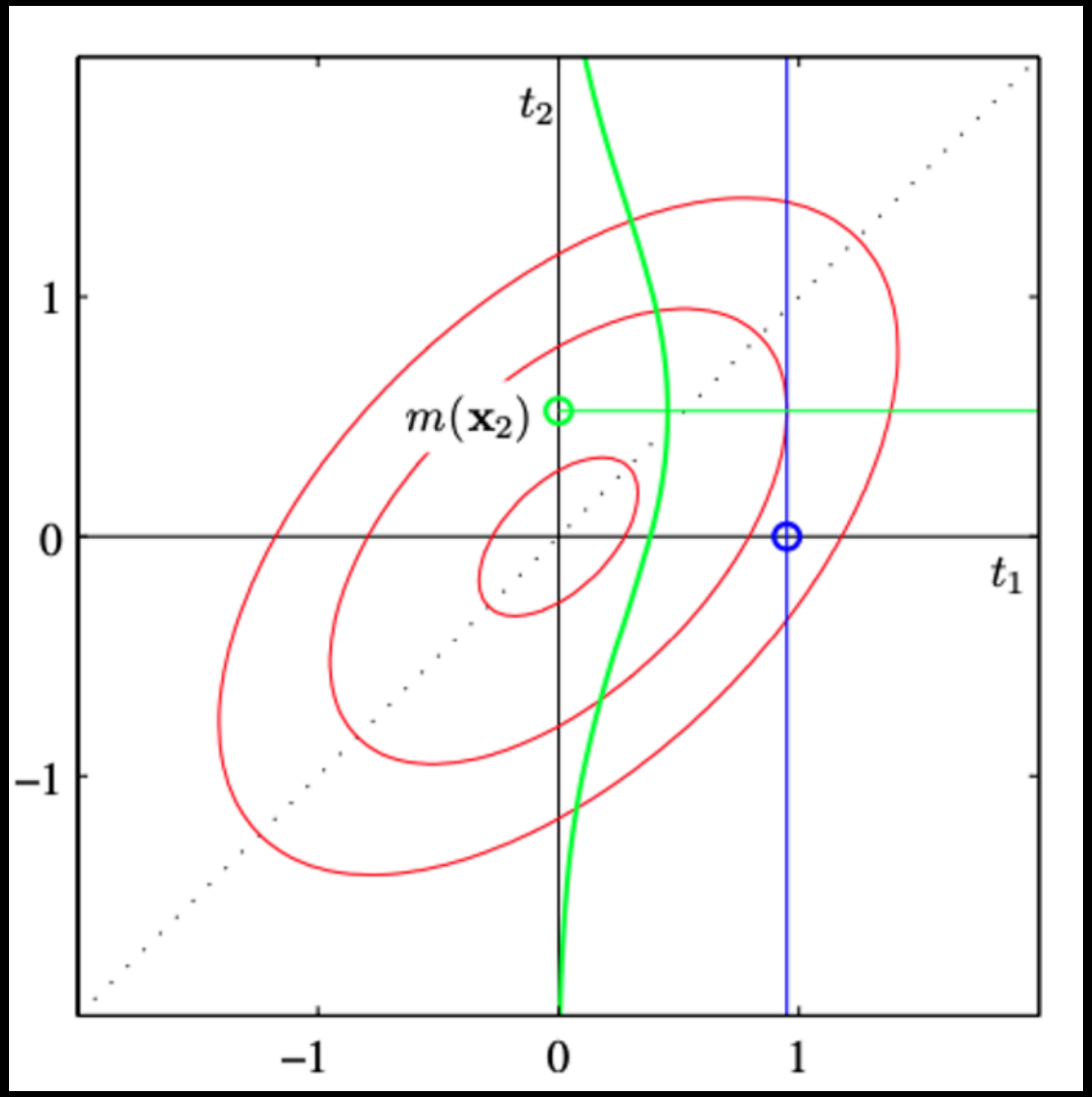
$$\ln p(\mathbf{y} \mid \mathbf{X}, \theta) = -\sum_{i} \ln \mathbf{L}_{ii} - \frac{1}{2} \mathbf{y}^T \alpha - \frac{N}{2} \ln(2\pi)$$

L = Cholesky $\hat{\mathbf{K}}$

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

Learning Hyperparameters

$$\underbrace{p(\mathbf{y}\,|\,\mathbf{X},\overbrace{\theta}^{\text{Kernel hyperparameter}})}_{\text{the marginal likelihood}} = \mathcal{N}(\mathbf{y}; \mathbf{0}, \underbrace{\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma^2 \mathrm{I}}_{\hat{\mathbf{K}}})$$



$$\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = -\frac{1}{2}\ln|\hat{\mathbf{K}}| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\hat{\mathbf{K}}^{-1}\mathbf{y} - \frac{\mathrm{N}}{2}\ln(2\pi)$$

(Bishop eq 6.69)

$$\frac{\partial}{\partial x}\ln|\mathbf{A}| = \mathrm{Tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\right)$$

$$\frac{\partial}{\partial x}(\mathbf{A}^{-1}) = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}$$

(Bishop eq C.21, C.22)

Learning through gradient descent

$$\frac{\partial}{\partial \theta_i}\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = -\frac{1}{2}\mathrm{Tr}\left(\hat{\mathbf{K}}^{-1}\frac{\partial \hat{\mathbf{K}}}{\partial \theta_\mathrm{i}}\right) + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\hat{\mathbf{K}}^{-1}\frac{\partial \hat{\mathbf{K}}}{\partial \theta_\mathrm{i}}\hat{\mathbf{K}}^{-1}\mathbf{y}$$

(Bishop eq 6.70)

$$\ln p(\mathbf{y}\,|\,\mathbf{X},\theta) = -\sum_\mathrm{i}\ln \mathrm{L}_\mathrm{ii} - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\alpha - \frac{\mathrm{N}}{2}\ln(2\pi)$$

$$L = \text{Cholesky } \hat{\mathbf{K}}$$

$$\boldsymbol{\alpha} = \mathbf{L}^{\mathrm{T}} \backslash (\mathbf{L} \backslash \mathbf{y})$$