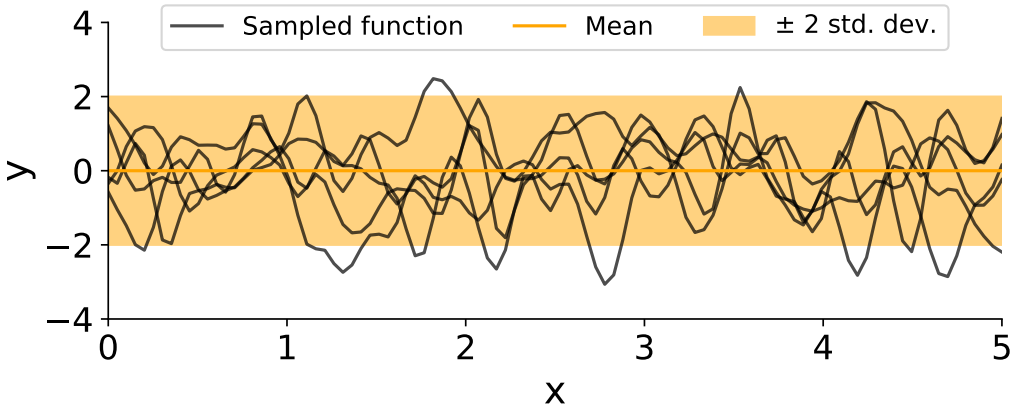
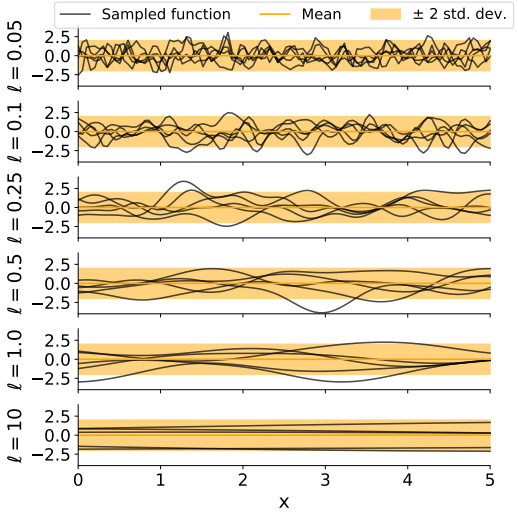
RBF kernel

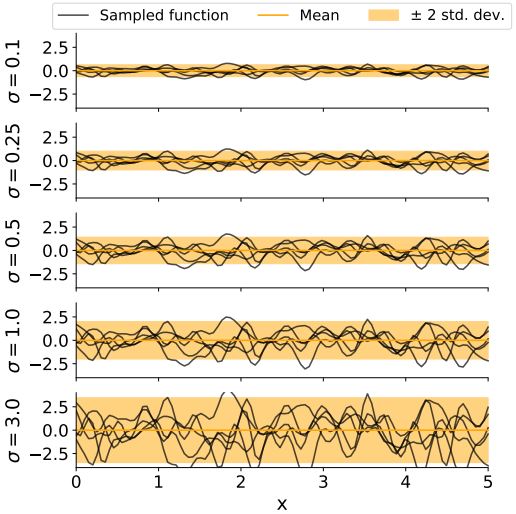
RBF kernel - samples from prior



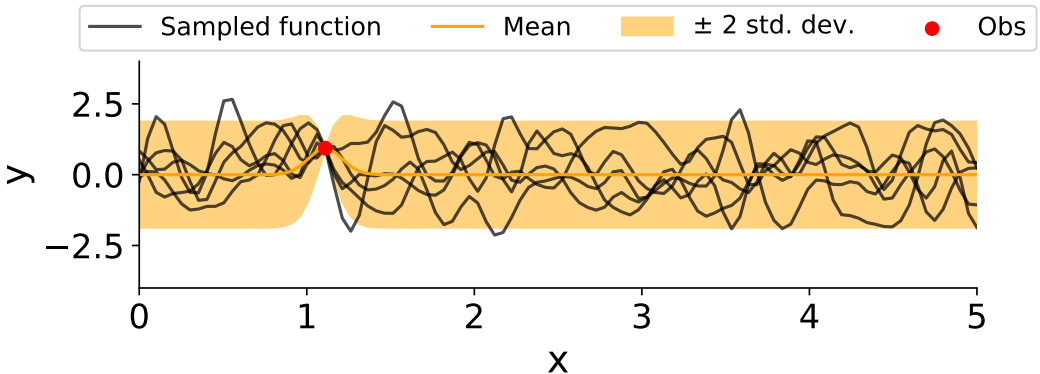
RBF kernel - samples from prior



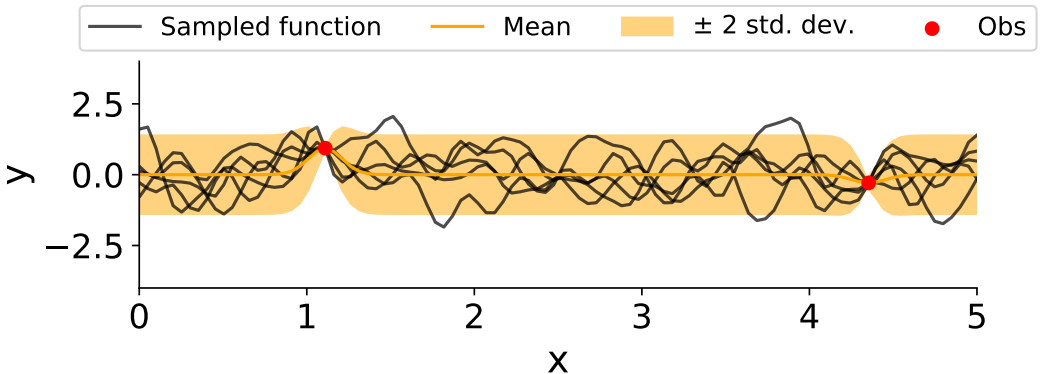
RBF kernel - samples from prior



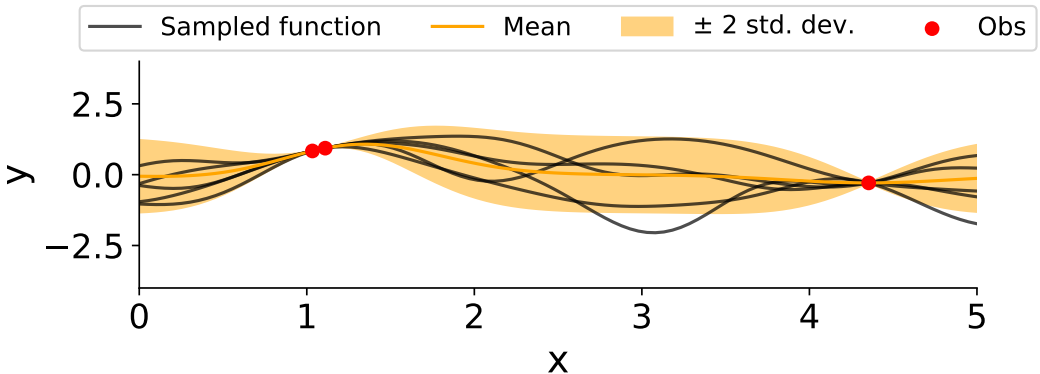
RBF kernel - samples from posterior



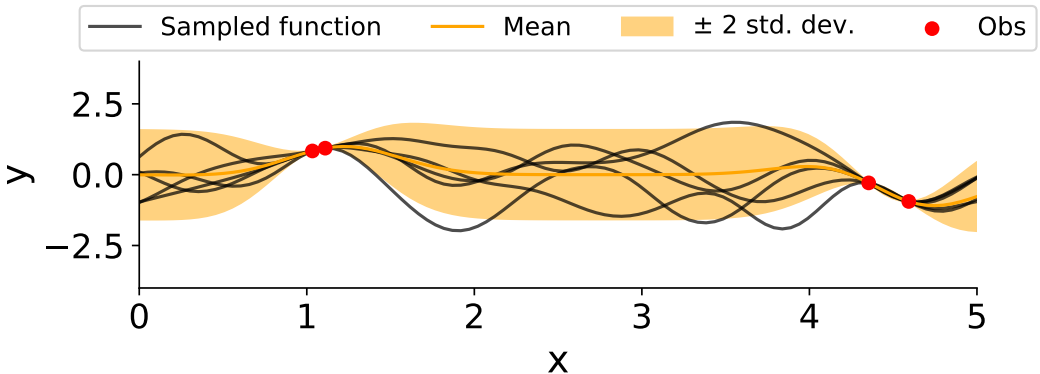
RBF kernel - samples from posterior



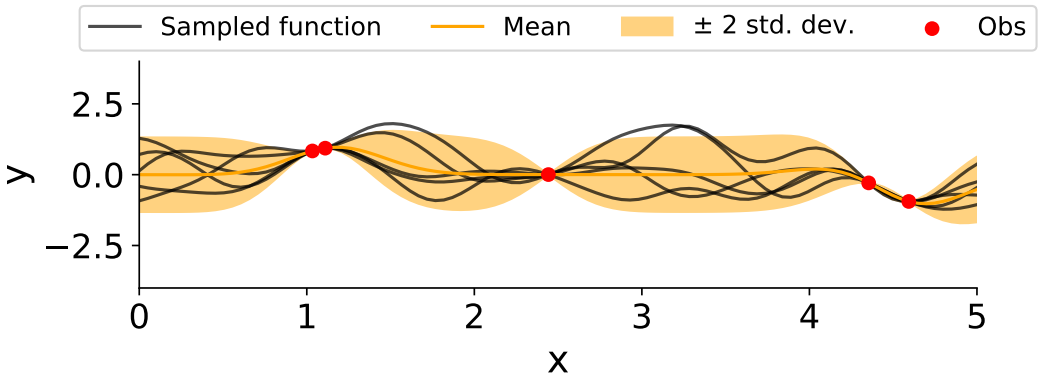
RBF kernel - samples from posterior



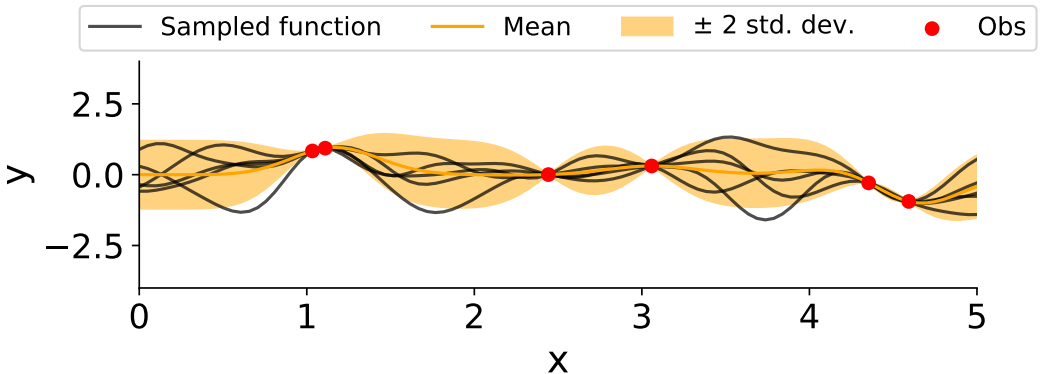
RBF kernel - samples from posterior



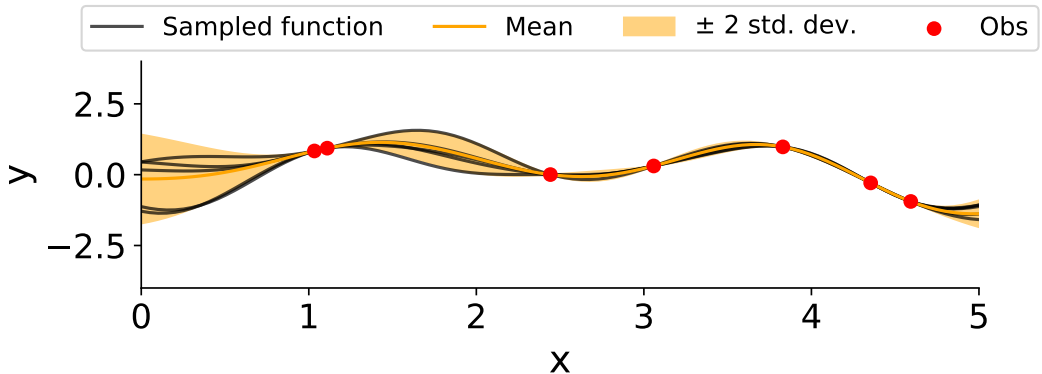
RBF kernel - samples from posterior



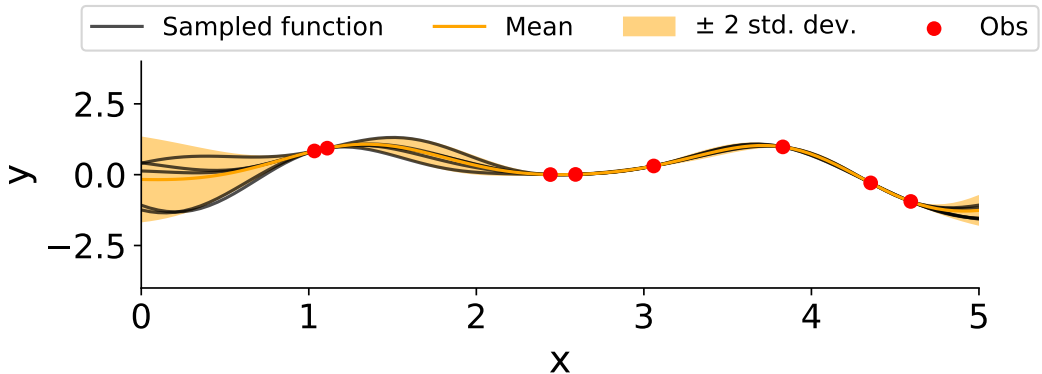
RBF kernel - samples from posterior



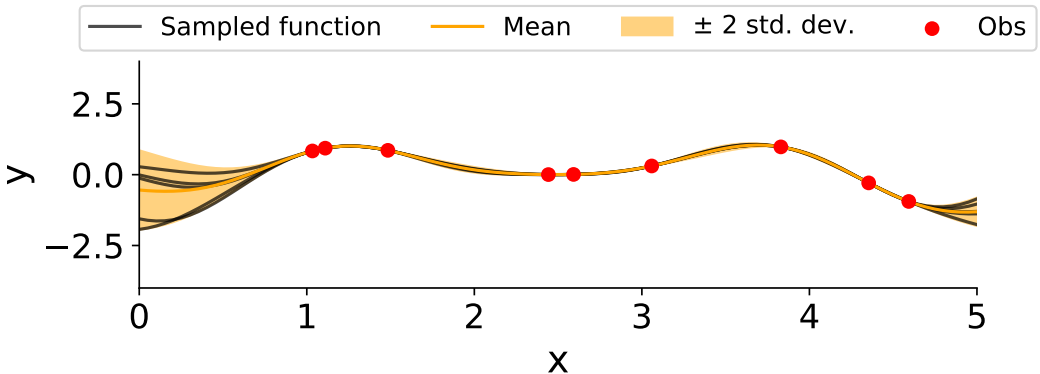
RBF kernel - samples from posterior



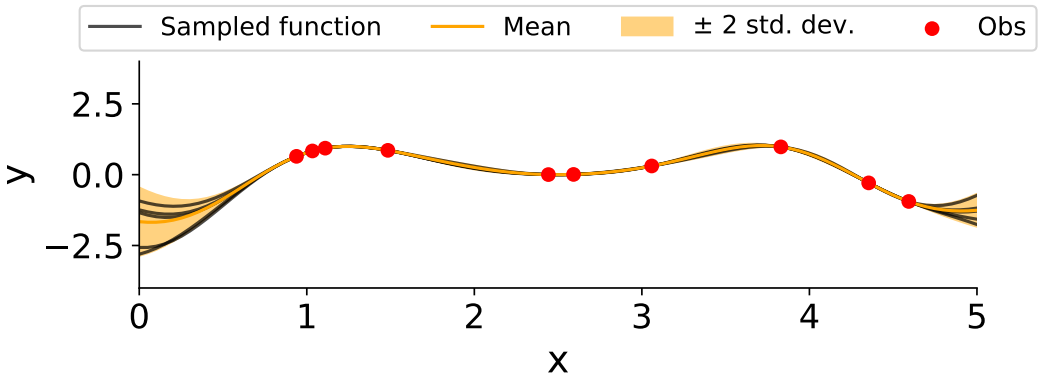
RBF kernel - samples from posterior



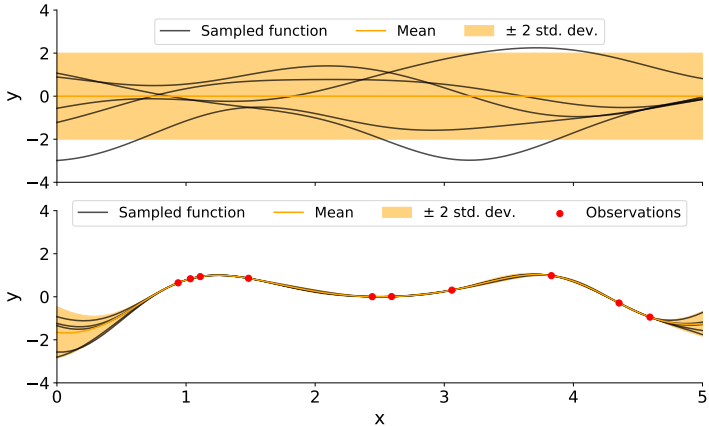
RBF kernel - samples from posterior



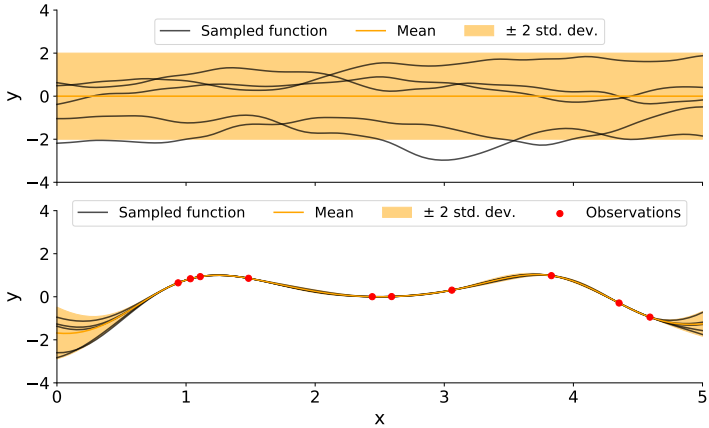
RBF kernel - samples from posterior



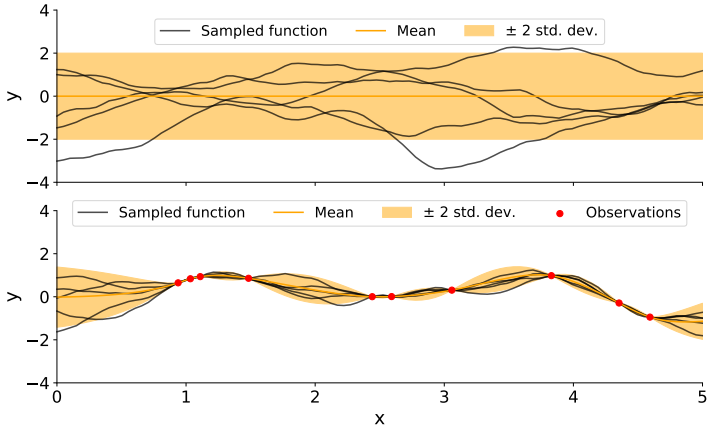
RBF kernel - prior / posterior



Rational-quadratic kernel - prior / posterior



Matern kernel - prior / posterior



Kernels in high dimensions

SE kernel:

$$K_y\left(\mathbf{x}_p, \mathbf{x}_q\right) = \sigma_f^2 \exp\left(-\frac{1}{2}\left(\mathbf{x}_p - \mathbf{x}_q\right)^T \mathbf{M}\left(\mathbf{x}_p - \mathbf{x}_q\right)\right) + \sigma_y^2 \delta_{pq}$$

Estimating the kernel parameters

$$K_y(x_p, x_q) = {\sigma_f}^2 \exp\left(-\frac{1}{2\ell^2}(x_p - x_q)^2\right) + {\sigma_y}^2 \delta_{pq}$$

We can maximize the marginal likelihood:

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

Estimating the kernel parameters

We can maximize the marginal likelihood:

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

where:

$$p(\mathbf{f} \mid \mathbf{X}) = \mathcal{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K})$$

$$p(\mathbf{y} \mid \mathbf{f}) = \prod_i \mathcal{N}\left(y_i \mid f_i, \sigma_y^2\right)$$

Estimating the kernel parameters

The marginal likelihood is given by:

$$\begin{aligned}\log p(\mathbf{y} \mid \mathbf{X}) &= \log \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \mathbf{K}_y\right) \\ &= \underbrace{-\frac{1}{2}\mathbf{y}\mathbf{K}_y^{-1}\mathbf{y}}_{\text{data fit}} \underbrace{-\frac{1}{2}\log\left|\mathbf{K}_y\right|}_{\text{model complexity}} \underbrace{-\frac{N}{2}\log(2\pi)}_{\text{constant}}\end{aligned}$$

Estimating the kernel parameters

Let the kernel parameters (also called hyper-parameters) be denoted by $\boldsymbol{\theta}$.

One can show that:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} \log p(\mathbf{y} \mid \mathbf{X}) &= \frac{1}{2}\mathbf{y}^T \mathbf{K}_y^{-1} \frac{\partial \mathbf{K}_y}{\partial \theta_j} \mathbf{K}_y^{-1}\mathbf{y} - \frac{1}{2}\,\mathrm{tr}\left(\mathbf{K}_y^{-1}\frac{\partial \mathbf{K}_y}{\partial \theta_j}\right) \\ &= \frac{1}{2}\,\mathrm{tr}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^T - \mathbf{K}_y^{-1}\right)\frac{\partial \mathbf{K}_y}{\partial \theta_j}\right)\end{aligned}$$

where $\boldsymbol{\alpha} = \mathbf{K}_y^{-1}\mathbf{y}$.