Two pitfalls in Gaussian process interpolation

Toni Karvonen

Department of Mathematics and Statistics
University of Helsinki, Finland

ProbNum & Friends
University of Tübingen

29 March 2023

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

Setting and overview

I consider Gaussian process interpolation:

- Let $f : \Omega \to \mathbb{R}$ be a data-generating function on a set $\Omega \subset \mathbb{R}^d$.
- Obtain *noiseless data* $\mathcal{D}_n = \{(\mathbf{x}_1, f(\mathbf{x}_1)), \dots, (\mathbf{x}_n, f(\mathbf{x}_n))\}$ at some pairwise distinct points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \Omega$.
- Model f as a Gaussian process $f_{\text{GP}} \sim \text{GP}(m, K)$.
- Compute the posterior $f_{\text{GP}} \mid \mathcal{D}_n$.

Gaussian process interpolation underlies underlies *Bayesian quadrature* and *Bayesian optimisation*.

This talk discusses for two pitfalls that are present in this setting:

1. Lengthscale estimation when a constant shift of m is observed.
2. The commonly used Gaussian kernel (i.e., squared exponential) is too smooth.

Table of contents

Gaussian processes

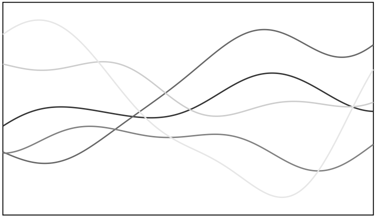
- Model f as a Gaussian process $f_{\mathrm{GP}} \sim \mathrm{GP}(m, K)$ with
  - a positive-definite covariance kernel $K \colon \Omega \times \Omega \to \mathbb{R}$ and
  - a mean function $m \colon \Omega \to \mathbb{R}$.

Under this model $[f_{\mathrm{GP}}(\mathbf{x}_1), \ldots, f_{\mathrm{GP}}(\mathbf{x}_n)] \in \mathbb{R}^n$ is a multivariate normal random variable with mean $\mathbf{m}_n$ and covariance matrix $\mathbf{K}_n$:

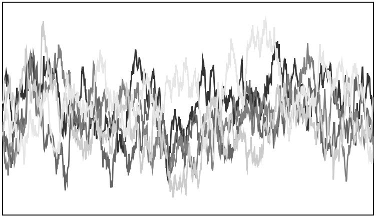
$$\begin{bmatrix} f_{\mathrm{GP}}(\mathbf{x}_1) \\ \vdots \\ f_{\mathrm{GP}}(\mathbf{x}_n) \end{bmatrix} \sim \mathrm{N}(\mathbf{m}_n, \mathbf{K}_n) = \mathrm{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right).$$

Gaussian process priors

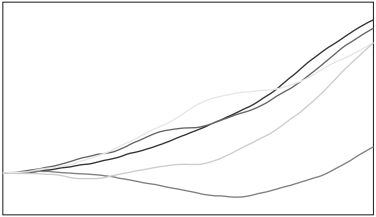
Gaussian: $K(x, y) = \mathrm{e}^{-(x-y)^2/2}$



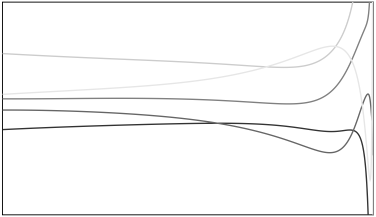
Matérn: $K(x, y) = \mathrm{e}^{-|x-y|}$



BM: $K(x, y) = \frac{\min\{x,y\}^3}{3} + \frac{|x-y|\min\{x,y\}^2}{2}$



Hardy: $K(x, y) = \frac{1}{1-xy}$



Gaussian process posterior

Recall that we have access to the *noiseless* data

$$\mathcal{D}_n = \left\{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, f(\mathbf{x}_n))\right\} \tag{1}$$

at some pairwise distinct points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \Omega$.

Conditional Gaussian process

The conditional process $f_{\mathrm{GP}} \mid \mathcal{D}_n$ is also a Gaussian process. Standard Gaussian conditioning formulae give

$$\mu_n(\mathbf{x}) = \mathbb{E}[f_{\mathrm{GP}}(\mathbf{x}) \mid \mathcal{D}_n] = m(\mathbf{x}) - \mathbf{k}_n(\mathbf{x})^{\mathsf{T}} \mathbf{K}_n^{-1} (\mathbf{f}_n - \mathbf{m}_n), \tag{2}$$

$$\mathbb{V}_n(\mathbf{x}) = \mathbb{V}[f_{\mathrm{GP}}(\mathbf{x}) \mid \mathcal{D}_n] = K(\mathbf{x}, \mathbf{x}) - \mathbf{k}_n(\mathbf{x})^{\mathsf{T}} \mathbf{K}_n^{-1} \mathbf{k}_n(\mathbf{x}). \tag{3}$$

Here

$$\mathbf{f}_n = \begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix}, \quad \mathbf{k}_n(\mathbf{x}) = \begin{bmatrix} K(\mathbf{x}, \mathbf{x}_1) \\ \vdots \\ K(\mathbf{x}, \mathbf{x}_n) \end{bmatrix}, \quad \mathbf{K}_n = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \cdots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_n, \mathbf{x}_1) & \cdots & K(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}.$$

Example from PN: Bayesian quadrature

We want to compute the integral $I_P(f) = \int_\Omega f(\mathbf{x})\,\mathrm{d}P(\mathbf{x})$.

Bayesian quadrature

Set $m \equiv 0$. Integration of the posterior GP yields Bayesian quadrature:

$$I_P(f_{\mathrm{GP}}) \mid \mathcal{D}_n \sim \mathrm{N}\big(\, Q_n^{\mathrm{BQ}},\ \mathbb{V}_n^{\mathrm{BQ}} \,\big), \tag{4}$$

where

$$Q_n^{\mathrm{BQ}} = I_P(\mu_n) = \mathbf{k}_{P,n}^\top \mathbf{K}_n^{-1} \mathbf{f}_n \quad \text{and} \quad \mathbb{V}_n^{\mathrm{BQ}} = K_{PP} - \mathbf{k}_{P,n}^\top \mathbf{K}_n^{-1} \mathbf{k}_{P,n}. \tag{5}$$

Here

$$\mathbf{k}_{P,n} = \begin{bmatrix} \int_\Omega K(\mathbf{x}, \mathbf{x}_1)\,\mathrm{d}P(\mathbf{x}) \\ \vdots \\ \int_\Omega K(\mathbf{x}, \mathbf{x}_n)\,\mathrm{d}P(\mathbf{x}) \end{bmatrix} \quad \text{and} \quad K_{PP} = \int_\Omega \int_\Omega K(\mathbf{x}, \mathbf{y})\,\mathrm{d}P(\mathbf{x})\,\mathrm{d}P(\mathbf{y}).$$

Illustration: Conditional distributions

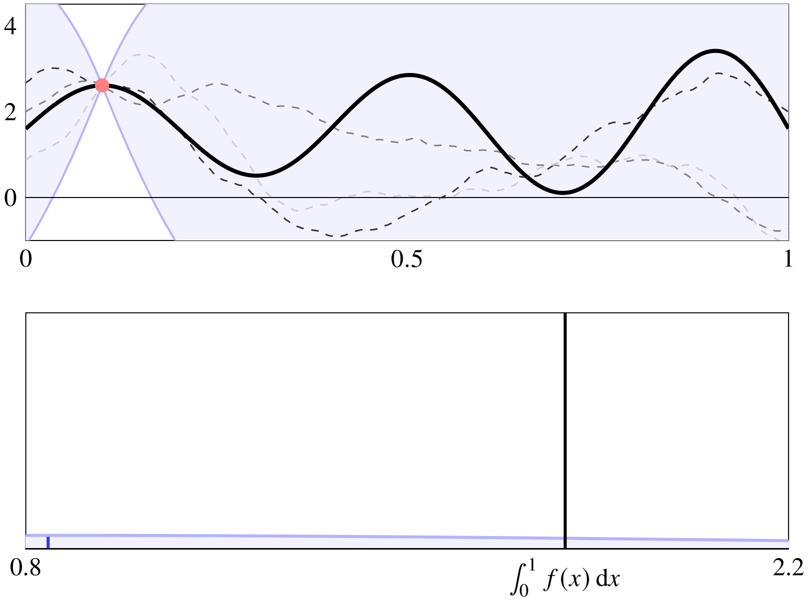
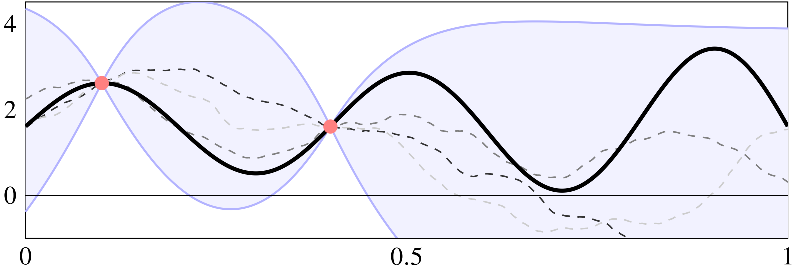


Illustration: Conditional distributions



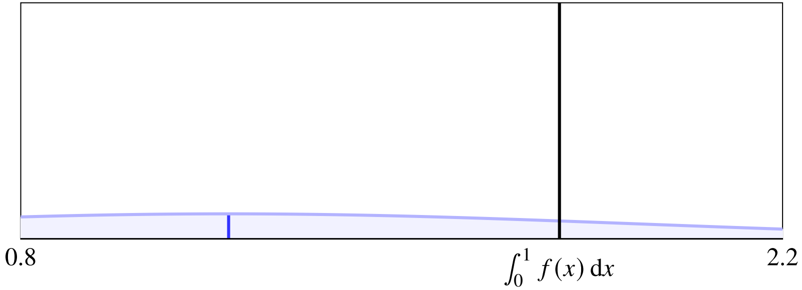
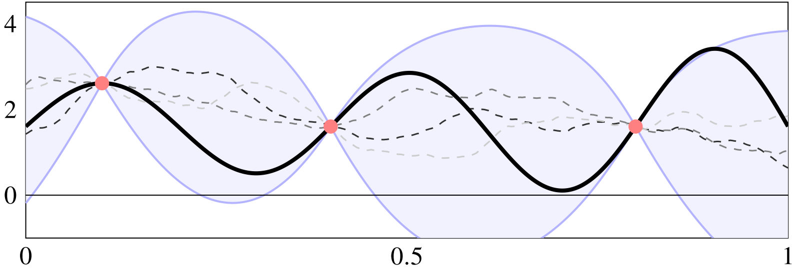


Illustration: Conditional distributions



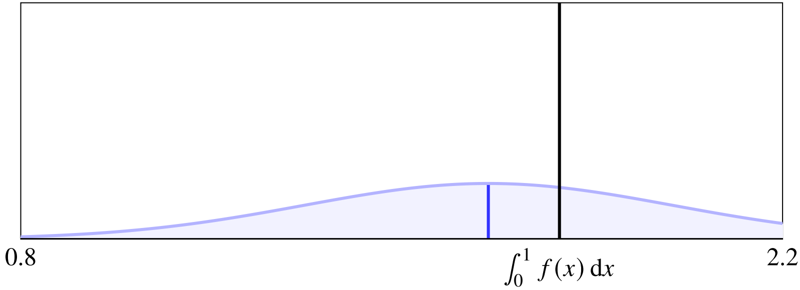


Illustration: Conditional distributions

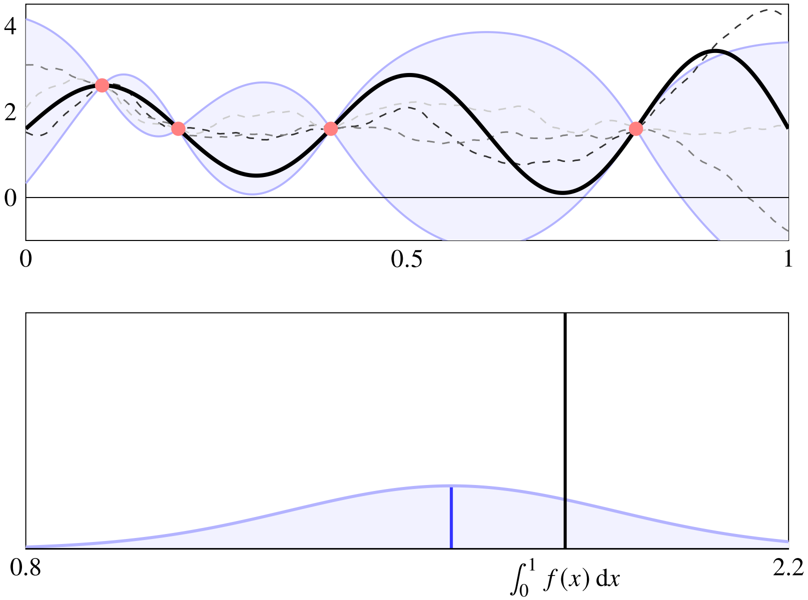


Illustration: Conditional distributions

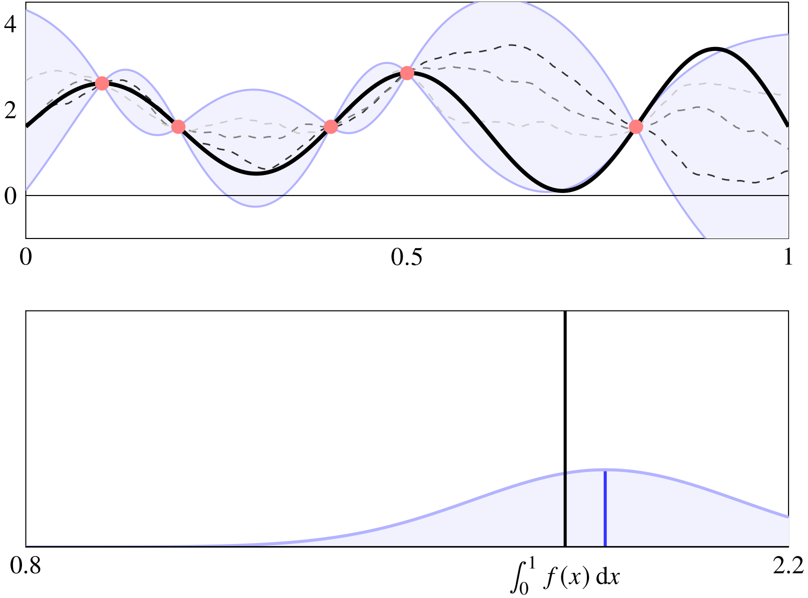
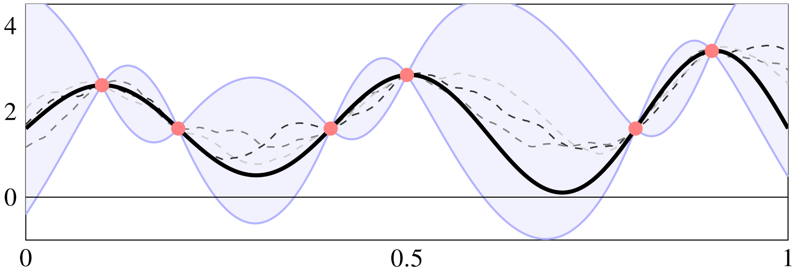


Illustration: Conditional distributions



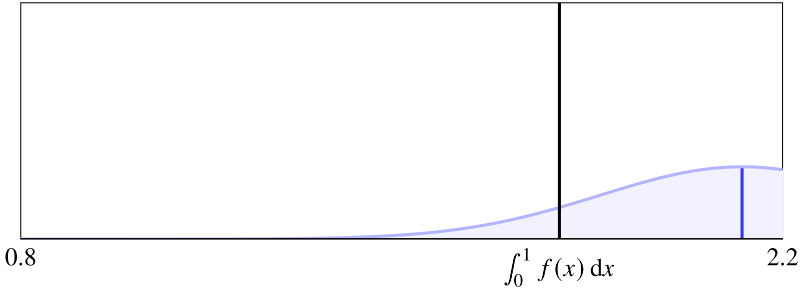


Illustration: Conditional distributions

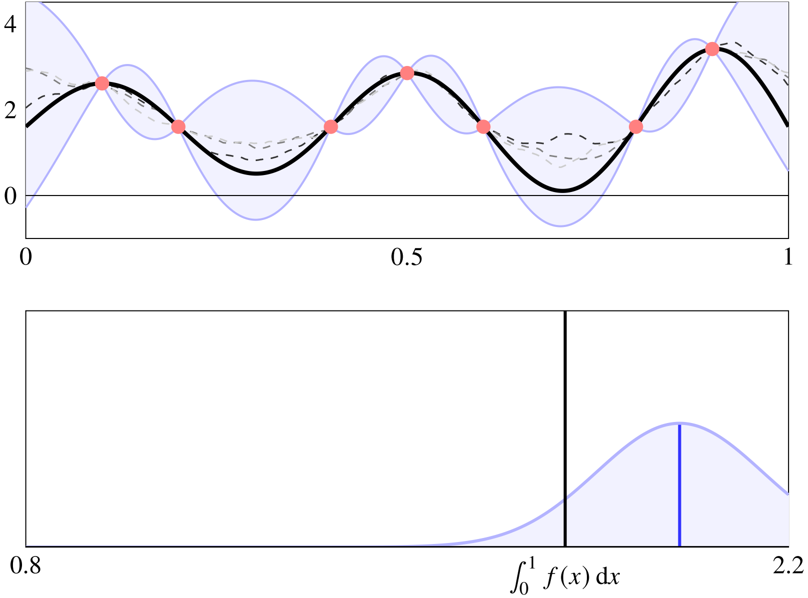


Illustration: Conditional distributions

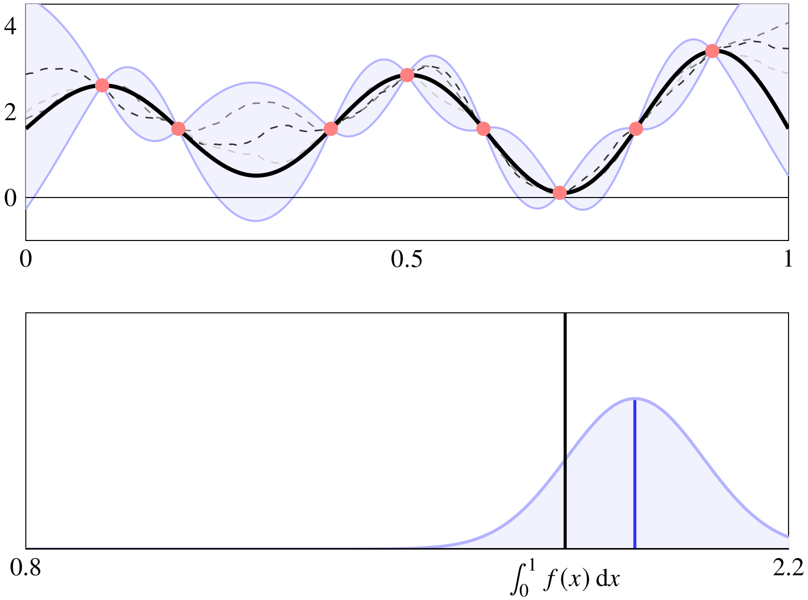


Illustration: Conditional distributions

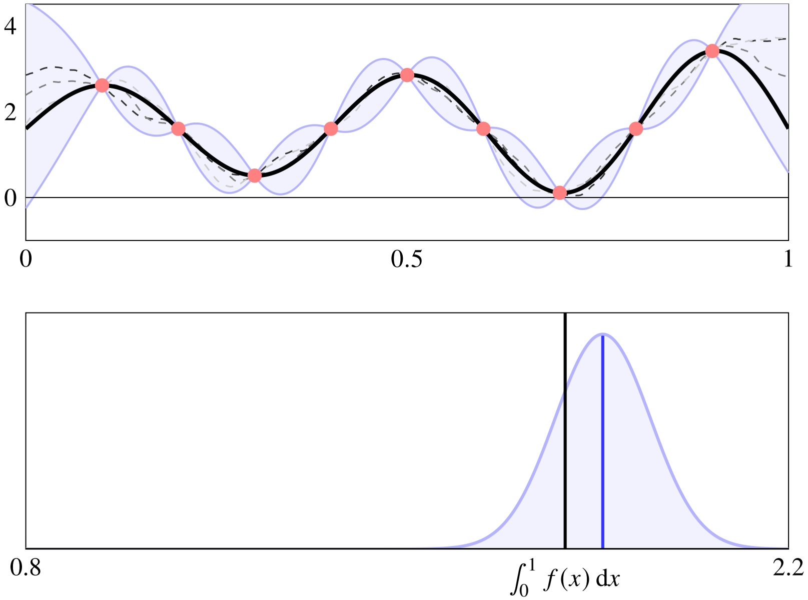
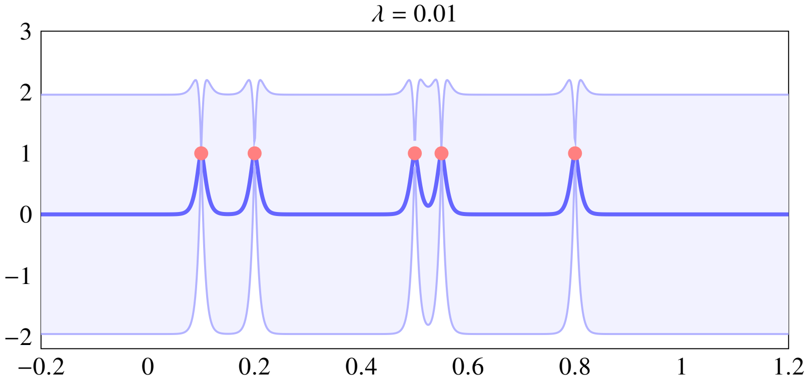


Table of contents

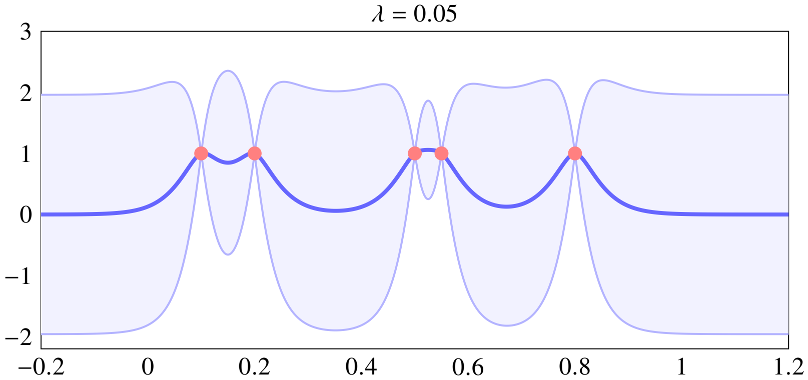
Effect of the lengthscale

$$K_{\color{red}\lambda}(x,y) = \left(1 + \frac{\sqrt{3}\,|x-y|}{\color{red}\lambda}\right)\exp\left(-\frac{\sqrt{3}\,|x-y|}{\color{red}\lambda}\right)$$



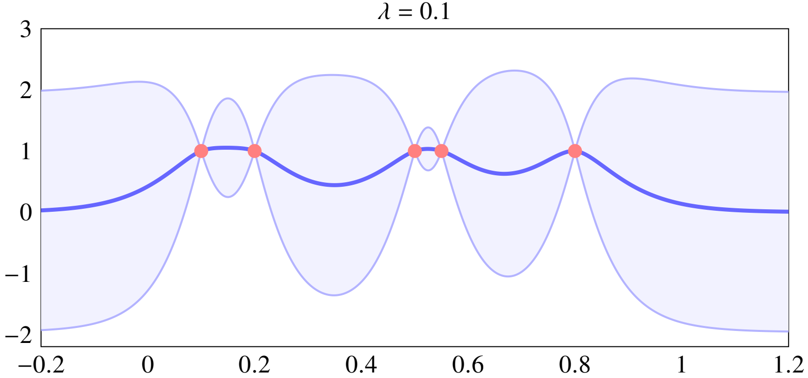
Effect of the lengthscale

$$K_{\lambda}(x, y) = \left(1 + \frac{\sqrt{3}\,|x - y|}{\lambda}\right) \exp\left(-\frac{\sqrt{3}\,|x - y|}{\lambda}\right)$$



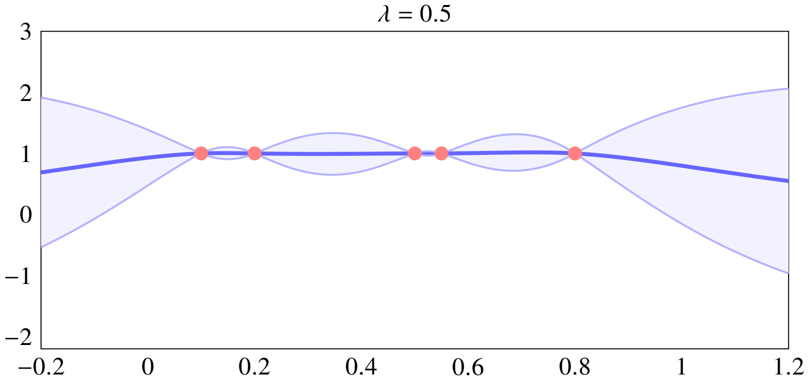
Effect of the lengthscale

$$K_{\lambda}(x, y) = \left(1 + \frac{\sqrt{3}\,|x - y|}{\lambda}\right) \exp\left(- \frac{\sqrt{3}\,|x - y|}{\lambda}\right)$$



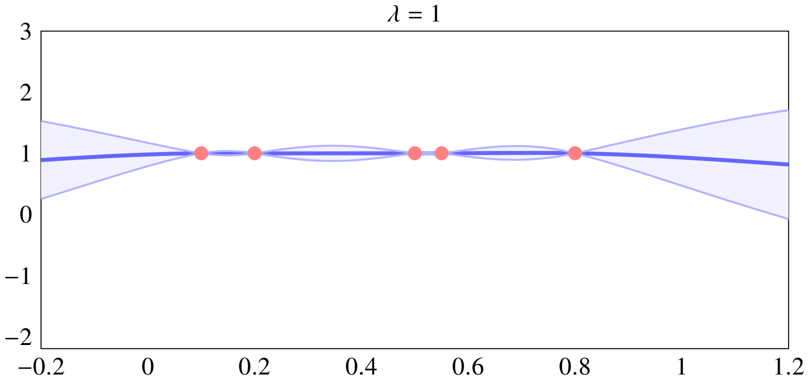
Effect of the lengthscale

$$K_{\lambda}(x, y) = \left(1 + \frac{\sqrt{3}\,|x - y|}{\lambda}\right) \exp\left(-\frac{\sqrt{3}\,|x - y|}{\lambda}\right)$$



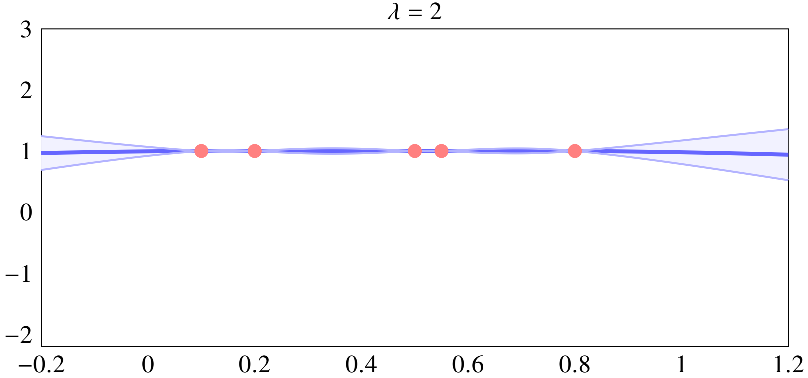
Effect of the lengthscale

$$K_{\lambda}(x, y) = \left(1 + \frac{\sqrt{3}\,|x - y|}{\lambda}\right) \exp\left(- \frac{\sqrt{3}\,|x - y|}{\lambda}\right)$$



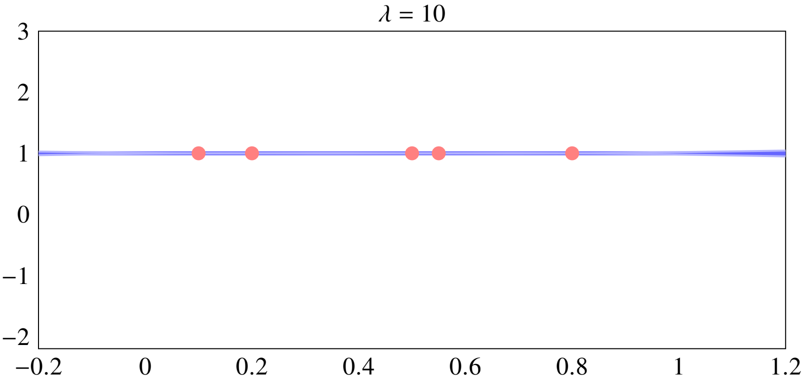
Effect of the lengthscale

$$K_{\lambda}(x, y) = \left(1 + \frac{\sqrt{3}\,|x - y|}{\lambda}\right) \exp\left(-\frac{\sqrt{3}\,|x - y|}{\lambda}\right)$$



Effect of the lengthscale

$$K_{\lambda}(x, y) = \left(1 + \frac{\sqrt{3}\,|x - y|}{\lambda}\right) \exp\left(- \frac{\sqrt{3}\,|x - y|}{\lambda}\right)$$



Matérn class

We consider kernels of the Matérn class.

Matérn class

Matérn kernel of smoothness $\nu > 0$ is

$$K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}) \tag{6}$$

where

$$\Phi(\mathbf{z}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, \|\mathbf{z}\|\right)^{\nu} \mathcal{K}_{\nu}\left(\sqrt{2\nu}\, \|\mathbf{z}\|\right). \tag{7}$$

For example, $\nu = 1/2$ and $\nu = 3/2$ give

$$\Phi_{\nu=1/2}(\mathbf{z}) = e^{-\|\mathbf{z}\|} \quad \text{and} \quad \Phi_{\nu=3/2}(\mathbf{z}) = \left(1 + \sqrt{3}\, \|\mathbf{z}\|\right) e^{-\sqrt{3}\|\mathbf{z}\|}. \tag{8}$$

[In fact, what follows applies to any stationary kernel $K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y})$ with Fourier transform $\widehat{\Phi}$ such that

$$C_1 (1 + \|\boldsymbol{\omega}\|^2)^{\alpha} \leq \widehat{\Phi}(\boldsymbol{\omega}) \leq C_2 (1 + \|\boldsymbol{\omega}\|^2)^{\alpha} \quad \text{for all} \quad \boldsymbol{\omega} \in \mathbb{R}^d.]$$

Maximum likelihood estimation

Let $\boldsymbol{\theta} \in \Theta$ be a kernel parameter vector. The log-likelihood function is

$$L(\boldsymbol{\theta}; \mathcal{D}_n) = -\frac{1}{2}\left[(\mathbf{f}_n - \mathbf{m}_n)^\mathsf{T} \mathbf{K}_{\boldsymbol{\theta},n}^{-1} (\mathbf{f}_n - \mathbf{m}_n) + \log\det \mathbf{K}_{\boldsymbol{\theta},n} + C\right], \quad \textbf{(9)}$$

where $\mathbf{f}_n = [f(\mathbf{x}_i)]_{i=1}^n$, $\mathbf{m}_n = [m(\mathbf{x}_i)]_{i=1}^n$ and $\mathbf{K}_{\boldsymbol{\theta},n} = [K_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n$.

Maximum likelihood estimation

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}_{\mathrm{ML}}(\mathcal{D}_n) = \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathcal{D}_n). \quad \textbf{(10)}$$

We are interested in

$$\lambda_{\mathrm{ML}}(\mathcal{D}_n) = \operatorname*{arg\,max}_{\lambda > 0} L(\lambda; \mathcal{D}_n) \quad \text{and} \quad K_\lambda(\mathbf{x}, \mathbf{y}) = K\left(\frac{\mathbf{x} - \mathbf{y}}{\lambda}\right) = \Phi\left(\frac{\mathbf{x} - \mathbf{y}}{\lambda}\right).$$

Failure (or not?) of maximum likelihood

We say that the data are m-constant if there is $c \in \mathbb{R}$ such that

$$\mathbf{f}_n - \mathbf{m}_n = [f(\mathbf{x}_1) - m(\mathbf{x}_1), \dots, f(\mathbf{x}_n) - m(\mathbf{x}_n)] = [c, \dots, c]. \quad \textbf{(11)}$$

Theorem (Karvonen & Oates, 2023)

Let $n \geq 2$ be **fixed** and K a Matérn kernel (isotropic or product). Then

$$\lambda_{\mathrm{ML}}(\mathcal{D}_n) = \infty \quad \text{and} \quad \lim_{\lambda \to \infty} L(\lambda; \mathcal{D}_n) = \infty \qquad \textbf{(12)}$$

if and only if the data are m-constant. Moreover, for every $\mathbf{x} \in \mathbb{R}^d$,

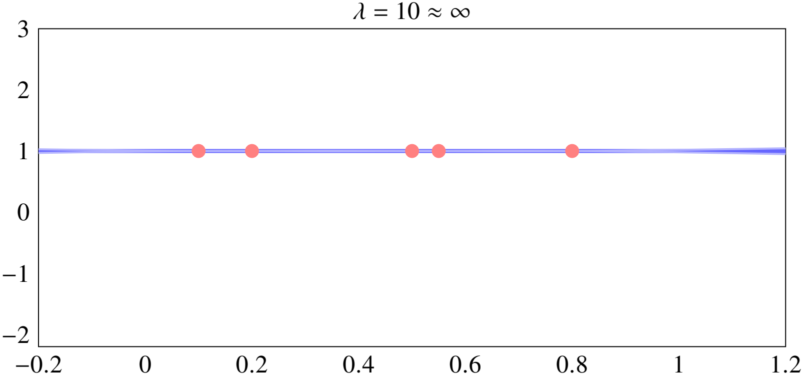
$$\lim_{\lambda \to \infty} \mu_{\lambda,n}(\mathbf{x}) = m(\mathbf{x}) + c \quad \text{and} \quad \lim_{\lambda \to \infty} \mathbb{V}_{\lambda,n}(\mathbf{x}) = 0. \qquad \textbf{(13)}$$

$\Longrightarrow$ If the data are m-constant, the posterior becomes degenerate.

Karvonen & Oates (2023). Maximum likelihood estimation in Gaussian process regression is ill-posed. *Journal of Machine Learning Research*. To appear.

Constant data and $\lambda \approx \infty$

$$K_{\lambda}(x,y) = \left(1 + \frac{\sqrt{3}\,|x-y|}{\lambda}\right) \exp\left(-\frac{\sqrt{3}\,|x-y|}{\lambda}\right)$$



Sketch of proof

$\mathcal{H}(K)$ = reproducing kernel Hilbert space (RKHS) of K.

1. If $g \in \mathcal{H}(K)$, then $\mathbf{g}_n^\top \mathbf{K}_n^{-1} \mathbf{g}_n \leq \|g\|_{\mathcal{H}(K)}^2$ for all $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$.

2. Use K_λ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$. $\iff$ Use K and $\frac{1}{\lambda}\mathbf{x}_1, \ldots, \frac{1}{\lambda}\mathbf{x}_n$.

3. If K is Matérn, $c \in \mathcal{H}(K)$. But $\mathbf{c}_n$ does not depend on $\mathbf{x}_1, \ldots, \mathbf{x}_n$!

4. We are to maximise

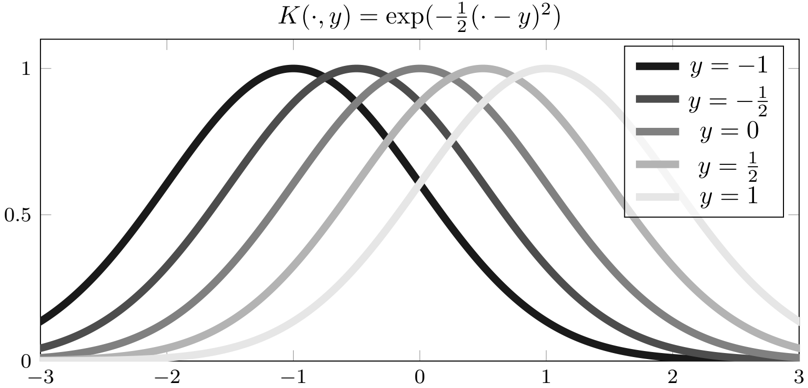
$$
\begin{aligned}
L(\lambda; \mathcal{D}_n) &= -(\mathbf{f}_n - \mathbf{m}_n)^\top \mathbf{K}_{\lambda,n}^{-1} (\mathbf{f}_n - \mathbf{m}_n) - \log\det \mathbf{K}_{\lambda,n} && \textbf{(14)} \\
&= -\mathbf{c}_n^\top \mathbf{K}_{\lambda,n}^{-1} \mathbf{c}_n - \log\det \mathbf{K}_{\lambda,n} && \textbf{(15)} \\
&\geq -\|c\|_{\mathcal{H}(K)}^2 - \log\det \mathbf{K}_{\lambda,n}. && \textbf{(16)}
\end{aligned}
$$

5. $\lim_{\lambda\to\infty} \mathbf{K}_{\lambda,n} = \mathbf{1}_n \mathbf{1}_n^\top \implies \log\det \mathbf{K}_{\lambda,n} \to -\infty$ as $\lambda \to \infty$.

Table of contents

Gaussian kernel

$$K(x, y) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\lambda^2}\right) \tag{17}$$



Variance decay — $d = 1$

Consider the Gaussian kernel

$$K(x, y) = \exp\left(-\frac{(x-y)^2}{2\lambda^2}\right) \quad \text{on} \quad \Omega = [-1, 1] \subset \mathbb{R}.$$

Theorem (Karvonen, 2022)

Let $x_1, \ldots, x_n \subset [-1, 1]$ be any pairwise distinct points. Then

$$C_1 \left(\frac{1}{4\lambda^2}\right)^n \frac{1}{n!} \leq \sup_{x \in [-1,1]} \mathbb{V}_n(x) \leq C_2\, n^{-1/4} e^{2\sqrt{n}/\lambda} \left(\frac{4}{\lambda^2}\right)^n \frac{1}{n!}. \quad \textbf{(18)}$$

$\Longrightarrow$ The variance decays everywhere with rate $(n!)^{-1} \approx n^{-n}$ regardless of how well $x_1, \ldots, x_n$ cover $[-1, 1]$.

$\Longrightarrow$ The magnitude of $\mathbb{V}_n(x)$ does not necessarily tell us much.

Karvonen (2022). Approximation in Hilbert spaces of the Gaussian and other weighted power series kernels. *arXiv:2209.12473v2*.

The uncertainty principle

In GP interpolation we need to work with the matrix $\mathbf{K}_n$.

Theorem (Schaback. 1995)

Let K be any positive-definite kernel. Then

$$\operatorname{cond}(\mathbf{K}_{n+1}) \geq \frac{1}{\mathbb{V}_n(\mathbf{x}_{n+1})}. \tag{19}$$

$\Longrightarrow$ Fast decay of $\mathbb{V}_n$ implies ill-conditioned $\mathbf{K}_n$.

Corollary

Let K be the Gaussian kernel on $[-1, 1]$ and $x_1, \ldots, x_{n+1} \in [-1, 1]$ any pairwise distinct points. Then

$$\operatorname{cond}(\mathbf{K}_{n+1}) \;\geq\; C_2^{-1} n^{1/4} e^{-2\sqrt{n}/\lambda} \left(\frac{\lambda^2}{4}\right)^n n! \tag{20}$$

$\Longrightarrow$ Must use nugget a nugget term: $\mathbf{K}_n \mapsto \mathbf{K}_n + \sigma^2 \mathbf{I}_n$.

Schaback (1995). Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264.

Variance decay — $d > 1$

Let $\widehat{\Phi}$ be the Fourier transform of $\Phi\colon \mathbb{R}^d \to \mathbb{R}$. Consider a kernel

$$K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x} - \mathbf{y}) \quad \text{such that} \quad \widehat{\Phi}(\boldsymbol{\omega}) \leq Ce^{-c\|\boldsymbol{\omega}\|}.$$

E.g., the Gaussian $\Phi(\mathbf{z}) = \exp\left(-\frac{\|z\|^2}{2\lambda^2}\right)$ has $\widehat{\Phi}(\boldsymbol{\omega}) \propto \exp\left(-\frac{\lambda^2\|\boldsymbol{\omega}\|}{2}\right)$.

Theorem

If the closure of $\{\mathbf{x}_i\}_{i=1}^{\infty} \subset [-1, 1]^d$ has non-empty interior, then

$$\sup_{\mathbf{x} \in [-1,1]^d} \mathbb{V}_n(\mathbf{x}) \to 0 \quad \text{as} \quad n \to \infty. \tag{21}$$

$\Longrightarrow$ Covering a part of $[-1, 1]^d$ well is enough for the variance to tend to zero uniformly on $[-1, 1]^d$.

$\Longrightarrow$ $\mathbb{V}_n \to 0$ even when the points are “badly” placed (e.g., cluster).

Nothing new here

Kolmogorov–Wiener prediction problem (1940s)

Kolmogorov (1941). Interpolation and extrapolation of stationary random sequences. *Izv. Akad. Nauk SSSR*.
Krein (1945). On a problem of extrapolation of A. N. Kolmogorov. *Dokl. Akad. Nauk SSSR*.
Wiener (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*.

— We shall see later that [...] when (1.795) holds, the future of the function f from which Φ is obtained is determinable completely in terms of its own past. [**p. 54**]

Stein (1999). ***Interpolation of Spatial Data: Some Theory for Kriging*.**

— That is, it is possible to predict $Z(t)$ perfectly for all $t > 0$ based on observing $Z(s)$ for all $s \in (-\varepsilon, 0]$ for any $\varepsilon > 0$. [**p. 30**]

— However, as I previously argued in the one-dimensional setting, random fields possessing these autocovariance functions are unrealistically smooth for physical phenomena. [**p. 55**]

— I strongly recommend not using autocovariance functions of the form Ce^{-at^2} to model physical processes. [**pp. 69–70**, in *More criticism of Gaussian autocovariance functions*]

Rasmussen & Williams (2006). ***Gaussian Processes for Machine Learning*.**

— Stein [1999] argues that such strong smoothness assumptions are unrealistic for modelling many physical processes [...]. However, the squared exponential is probably the most widely-used kernel within the kernel machines field. [**p. 83**]

Also: *No empty ball property* of **Vazquez & Bect (2010)**. [*J.Stat.Plan.Infer.*, 140(11):3088–3095.]

Conclusion

- Maximum likelihood estimation of the lengthscale parameter may yield a degenerate posterior.
- The Gaussian kernel is too smooth to be robust. Do not use as a default kernel in PN?

Thank you for your attention!