Why not to use the Gaussian kernel

Toni Karvonen

Department of Mathematics and Statistics
University of Helsinki, Finland

Prob Num 2022
London

29 March 2022

UNIVERSITY OF HELSINKI
FACULTY OF SCIENCE

Table of contents

Introduction

Reason 1: Analyticity

Reason 2: Ill-conditioning

Reason 3: Uncertainty quantification

Some remarks

Gaussian process interpolation

- Let $f\colon [-1, 1] \to \mathbb{R}$ be the data-generating function.
- Let $K\colon [-1, 1] \times [-1, 1] \to \mathbb{R}$ be a positive-definite covariance kernel.
- Let $x_1, \ldots, x_n \in [-1, 1]$ be distinct sampling points.

Model f as a Gaussian process $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K)$ and obtain the noiseless data

$$\mathcal{D}_n(f) = \{(x_1, f(x_1)), \ldots, (x_n, f(x_n))\}.$$

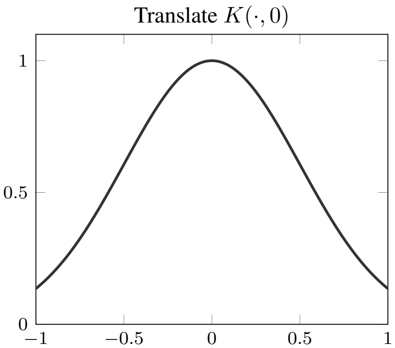
The conditional mean and variance are

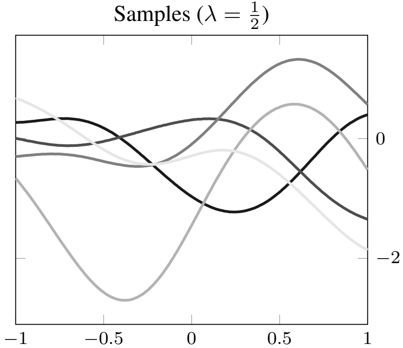
$$\mu_n(x) = \mathbf{K}_n(x)^\mathsf{T} \mathbf{K}_{n,n}^{-1} \mathbf{f}_n \quad \text{and} \quad \mathbb{V}_n(x) = K(x, x) - \mathbf{K}_n(x)^\mathsf{T} \mathbf{K}_{n,n}^{-1} \mathbf{K}_n(x). \quad (1)$$

Which kernel K to use?

Gaussian kernel

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





Common default kernel

`probnum/quad/solvers/bayesian_quadrature.py:`

```python
        # Select policy and belief update
        if kernel is None:
            kernel = ExpQuad(input_shape=(input_dim,))
```

Natural extension of the Matérn class

Let

$$K_\nu(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\,|x-y|}{\lambda}\right)^{\nu}\mathcal{K}_\nu\left(\frac{\sqrt{2\nu}\,|x-y|}{\lambda}\right)$$

be the Matérn kernel of order $\nu > 0$. Then

$$K_\nu(x,y) \to K(x,y) \quad \text{as} \quad \nu \to \infty.$$

The convergence to the Gaussian kernel occurs naturally:

Theorem [Karvonen 2022, Corollary 3.6]

Suppose that the points $\{x_i\}_{i=1}^n$ are sufficiently uniform on $[-1,1]$. If the data-generating function $f\colon [-1,1] \to \mathbb{R}$ is infinitely differentiable, then

$$\lim_{n\to\infty} \hat{\nu}_{\mathrm{ML}}(n) = \lim_{n\to\infty} \hat{\nu}_{\mathrm{LOO\text{-}CV}}(n) = \infty.$$

Karvonen (2022). Asymptotic bounds for smoothness parameter estimates in Gaussian process regression. *arXiv:2203.05400*.

The Gaussian kernel and its RKHS are interesting

- **Karvonen & Särkkä (2019)**. Gaussian kernel quadrature at scaled Gauss–Hermite nodes. *BIT Numerical Mathematics*, 59(4):877–902.
- **Karvonen & Särkkä (2020)**. Worst-case optimal approximation with increasingly flat Gaussian kernels. *Advances in Computational Mathematics*, 46:21.
- **Karvonen, Tanaka & Särkkä (2021)**. Kernel-based interpolation at approximate Fekete points. *Numerical Algorithms*, 87(1):445–468.
- **Karvonen, Oates & Girolami (2021)**. Integration in reproducing kernel Hilbert spaces of Gaussian kernels. *Mathematics of Computation*, 90(331):2209–2233.
- **Karvonen (2022)**. Small sample spaces for Gaussian processes. *Bernoulli*. To appear.

But please do not use it!

Nothing new here

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

"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 subsection "*More criticism of Gaussian autocovariance functions*"]

The Gaussian kernel is not robust

- The prior imposed by the Gaussian kernel is too strong.
- The prior is not only smooth, it is “super smooth”.

Implications:

1. Not robust with respect to sampling point placement.
2. Not numerically robust.
3. Non-robust uncertainty quantification. (Likely)

Table of contents

Introduction

Reason 1: Analyticity

Reason 2: Ill-conditioning

Reason 3: Uncertainty quantification

Some remarks

Analytic functions

Let $D \subseteq \mathbb{R}$ be an open interval.

Analytic function

A function $f\colon D \to \mathbb{R}$ is analytic on D if it is infinitely differentiable and equal to its Taylor series in the neighbourhood of every $a \in D$:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

for some $\varepsilon > 0$ and all $x \in D$ such that $|x - a| < \varepsilon$.

For analytic functions local information is global information.

The Gaussian kernel is "very analytic"

Necessary conditions for analyticity

A function $g : \mathbb{R} \to \mathbb{R}$ is analytic if either of the following holds:

1. $\sup_{x \in \mathbb{R}} |g^{(k)}(x)| \le C^k k!$ for some $C > 0$ and every $k \in \mathbb{N}_0$.
2. The function is integrable and there are $B > 0$ and $\alpha > 0$ such that

$$|\widehat{g}(\xi)| \le B \exp(-\alpha |\xi|) \quad \text{for all} \quad \xi \in \mathbb{R}.$$

Let $K(x, y) = \phi(x - y)$ for $\phi(r) = e^{-r^2/(2\lambda^2)}$. Then

$$\sup_{r \in \mathbb{R}} |\phi^{(k)}(r)| \le (2\ell^2)^{-k/2} \sqrt{\frac{(2k)!}{k!}} \le c_1(\lambda)^k \sqrt{k!} \tag{2}$$

and $[c_1(\lambda), c_2(\lambda) > 0]$

$$|\widehat{\phi}(\xi)| = c_2(\lambda) \exp\left(- \frac{\lambda^2}{2} |\xi|^2 \right). \tag{3}$$

Variance is weakly dependent on sampling points

Theorem [work in progress]

There are positive constants C_1 and C_2 such that

$$C_1(\lambda)\frac{1}{\sqrt{n}}\left(\frac{e}{4\lambda^2}\right)^n n^{-n} \le \sup_{x\in[-1,1]} \mathbb{V}_n(x) \le C_2(\lambda)\frac{1}{n}\left(\frac{8e}{\lambda^2}\right)^n n^{-n} \qquad (4)$$

for ***any sampling points*** $\{x_i\}_{i=1}^n$.

Variance decays to zero globally even if the sampling points do not cover the domain.

$\Longrightarrow$ The Gaussian kernel is not robust against badly placed sampling points.

Table of contents

Introduction

Reason 1: Analyticity

Reason 2: Ill-conditioning

Reason 3: Uncertainty quantification

Some remarks

Ill-conditioning

Condition number

The condition number $\kappa(\mathbf{A})$ of a symmetric matrix $\mathbf{A}$ is

$$\kappa(\mathbf{A}) = \left|\frac{\lambda_{\text{max}}(\mathbf{A})}{\lambda_{\text{min}}(\mathbf{A})}\right| = \left|\frac{\text{largest eigenvalue of } \mathbf{A}}{\text{smallest eigenvalue of } \mathbf{A}}\right|.$$

Large condition number = numerically unstable matrix inversion

The uncertainty principle

In GP interpolation we need to compute

$$\mathsf{K}_{n,n}^{-1}\mathsf{K}_n(x), \quad \text{where} \quad (\mathsf{K}_{n,n})_{ij} = K(x_i, x_j). \tag{5}$$

Theorem [Schaback 1995, Theorem 2.1]

Let K be any positive-definite kernel. Then

$$\kappa(\mathsf{K}_{n+1,n+1}) \geq \frac{1}{\mathbb{V}_n(x_{n+1})}.$$

Fast decay of conditional variance = ill-conditioned kernel matrix

[Of course, one can do something else than solve (5) directly.]

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

Condition number for the Gaussian kernel

Theorem [consequence of the uncertainty principle]

For the Gaussian kernel we have

$$\kappa(\mathsf{K}_{n+1,n+1}) \geq C_1(\lambda)\sqrt{n}\left(\frac{\lambda^2}{4e}\right)^n n^n$$

for ***any sampling points***.

In contrast, for K = Matérn-ν and sufficiently uniform points,

$$\kappa(\mathsf{K}_{n+1,n+1}) \geq C_2(\lambda)\, n^{2\nu-1}.$$

Table of contents

Introduction

Reason 1: Analyticity

Reason 2: Ill-conditioning

Reason 3: Uncertainty quantification

Some remarks

Uncertainty quantification

We want the conditional variance to reflect the approximation error:

- Ideally,

$$|f(x) - \mu_n(x)| \approx \mathbb{V}_n(x)^{1/2} \tag{6}$$

- Or at minimum,

$$|f(x) - \mu_n(x)| \leq a_n \mathbb{V}_n(x)^{1/2} \tag{7}$$

  for a sequence $(a_n)_{n=1}^{\infty}$ which does not grow "too fast" as $n \to \infty$.

For Matérn kernels there are result which say that, essentially,

$$|f(x) - \mu_n(x)| \leq C \sqrt{n}\, \hat{\sigma}_{\mathrm{ML}}(n)\, \mathbb{V}_n(x)^{1/2}. \tag{8}$$

Karvonen, Wynne, Tronarp, Oates & Särkkä (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958.

Misspecification and scale estimation

- Let $\hat{\sigma}(n)$ be any scale estimator of $\sigma > 0$ in the parametrisation $K_\sigma(x, y) = \sigma^2 K(x, y)$. For example,

$$\hat{\sigma}_{\mathrm{ML}}(n)^2 = \frac{\mathbf{f}_n^\top \mathbf{K}_{n,n}^{-1} \mathbf{f}_n}{n}. \tag{9}$$

- Let $f \colon [-1, 1] \to \mathbb{R}$ be a finitely smooth function such that

$$\sup_{x \in [-1,1]} |f(x) - \mu_n(x)| \approx n^{-\alpha} \quad \text{for} \quad \alpha > 0. \tag{10}$$

and recall that

$$\sup_{x \in [-1,1]} \mathbb{V}_n(x)^{1/2} \approx r^n \, n^{-n/2} \quad \text{for} \quad r > 0. \tag{11}$$

- To achieve, say,

$$|f(x) - \mu_n(x)| \approx \sqrt{n}\, \hat{\sigma}(n) \mathbb{V}_n(x)^{1/2} \tag{12}$$

we thus would need

$$\hat{\sigma}(n) \approx n^{-\alpha - 1/2} r^{-n} n^{n/2}. \tag{13}$$

Table of contents

Introduction

Reason 1: Analyticity

Reason 2: Ill-conditioning

Reason 3: Uncertainty quantification

Some remarks

The Cauchy kernel

What if you really want to use a smooth prior? The Cauchy kernel is

$$K(x,y) = \frac{1}{1+(x-y)^2/\lambda^2} = \phi(x-y). \tag{14}$$

Properties of the Cauchy kernel [quite easy to prove]

It holds that [results for the Gaussian in parentheses]

$$\sup_{x\in\mathbb{R}} |\phi^{(k)}(x)| \le \lambda^{-k} k!, \qquad \left[\ \le c_1(\lambda)^k \sqrt{k!}\ \right] \tag{15}$$

$$|\widehat{\phi}(\xi)| \le \frac{\lambda}{2}\exp(-\lambda|\xi|) \qquad \left[\ \le c_2(\lambda)\ \exp\left(-\frac{\lambda^2}{2}|\xi|^2\right)\right] \tag{16}$$

and

$$\mathbb{V}_n(x) \le C_1(\lambda)\frac{1}{\sqrt{n}}\left(\frac{16}{\lambda^2}\right)^n \qquad \left[\ \le C_2(\lambda)\frac{1}{n}\left(\frac{8e}{\lambda^2}\right)^n n^{-n}\ \right] \tag{17}$$

for any sampling points (RHS does not tend to zero if $\lambda < 4$).

Dette & Zhigljavsky (2021). Reproducing kernel Hilbert spaces, polynomials, and the classical moment problem. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1589–1614.

The role of sampling points

General principle in numerical analysis:

- Finitely smooth approximation (e.g., Matérn GPs) works with any sampling points.
- Infinitely smooth approximation does not. [e.g., Runge's phenomenon]

To approximate using an infinitely smooth functions the sampling points $\{x_i\}_{i=1}^n$ need to be selected carefully (e.g., Chebyshev nodes).

But this is typically not done in GP interpolation.

$\Longrightarrow$ Do not use infinitely smooth kernels if you are not willing to find "good" points!

Thank you for your attention!

Platte, Trefethen & Kuijlaars (2011). Impossibility of fast stable approximation of analytic functions from equispaced samples. *SIAM Review*, 53(2):308–318.