Gaussian Processes and Uncertainty Quantification

Nicholas Krämer

May 17, 2019

Institute for Numerical Simulation, University of Bonn

Outline

1. Gaussian Process Regression
2. Uncertainty Quantification and Inverse Problems
3. Emulators for Inverse Problems
4. Numerical Linear Algebra with Covariances

Gaussian Process Regression

Gaussian Processes

Definition (Gaussian process)

A Gaussian process $Z \sim \mathrm{GP}$ is a random field, for which any finite number of evaluations has multivariate Gaussian distribution.

- Generalisation of multivariate Gaussian distribution
- Uniquely defined by mean function $m = m(\cdot)$ and covariance function $k = k(\cdot, \cdot)$
- Write: $Z \sim \mathrm{GP}(m, k)$

Gaussian Processes

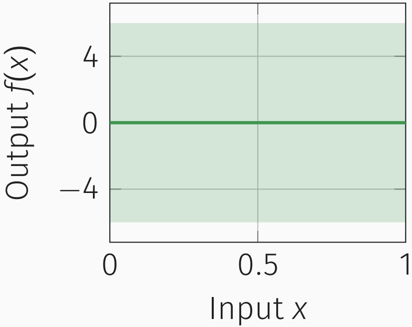


Figure 1: Gaussian process (prior distribution) with constant mean.

Gaussian Process Regression

- Goal: recover a function $f : \mathcal{X} \to \mathbb{R}$ from evaluations $f(\mathcal{X}_N) = (f(x_1), \dots, f(x_N))$ at $\mathcal{X}_N$
- Condition law of GP on attaining measurements
- Result: $\hat{Z} \sim \mathrm{GP}(m_f, k_f)$ with $m_f(x) = \sum_{i=1}^N c_i k(x, x_i)$
- Here, $k(\mathcal{X}_N, \mathcal{X}_N)c = f(\mathcal{X}_N)$.

Gaussian Process Regression

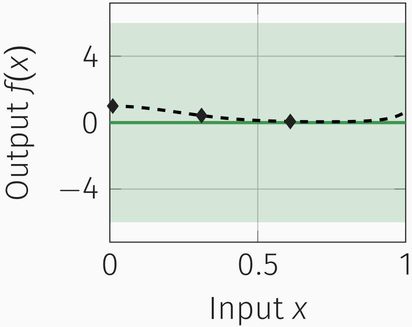


Figure 2: Gaussian process prior distribution with constant mean (green), true function (black; dashed), observations.

Gaussian Process Regression

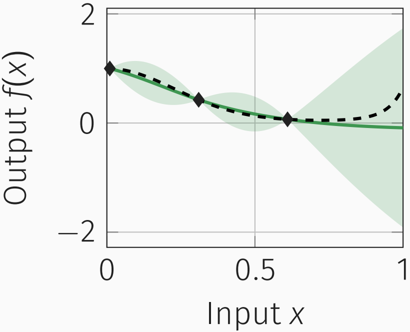


Figure 3: Gaussian process posterior distribution (green), true function (black; dashed), observations.

Error estimates

- Use Matérn kernel $k_\nu(x, y) = \varphi_\nu(\|x - y\|)$, $\varphi_\nu(s) \sim s^\nu K_\nu(s)$
- Repr. kernel Hilbert space $\mathcal{H}_k \cong H^{\nu+d/2}(\mathcal{X})$, $d = \dim(\mathcal{X})$
- Approximation order: $\|f - m_f\|_{L^2} \lesssim h^{\nu+d/2}$
- h is fill distance of $\mathcal{X}_N$

F. J. Narcowich and J. D. Ward and H. Wendland
Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions.
Constr. Approx., 24, pp. 175–186, 2006

Uncertainty Quantification and Inverse Problems

Uncertainty Quantification

- Goal: reduce uncertainty arising in computations (uncertain modelling, uncertain algorithms, ...)
- Inverse problem
- Given $\mathcal{G} : A \to Y$ and (noisy) measurements $y \in Y$, find $a \in A$ from

$$y = \mathcal{G}(a) + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2), \quad \sigma > 0$$

Inverse Problems: Differential Equation

Find $a \in [-1, 1]$ such that for

$$-\operatorname{div}(e^{1+0.5\sin(ax)}\nabla u(x)) = 1, \quad u(0) = u(1) = 0$$

the measurements satisfy

$$u(1/3) = 1.1241, \quad u(1/2) = 1.34235, \quad u(2/3) = 1.87.$$

Bayesian Approach

- Knowledge is probability distribution
- Statistical estimators using that distribution

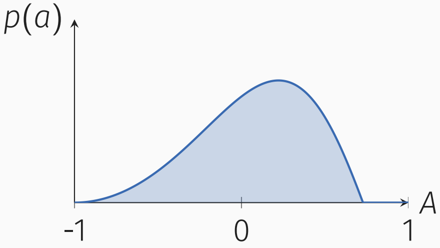


Figure 4: Bayesian answer to: “Which input parameter?”

Mathematically

- Prior distribution μ_0 on A
- Posterior distribution μ^y on A via Radon–Nikodym derivative

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(a) \propto e^{-\Phi(a)}, \qquad \Phi(a) = \|y - \mathcal{G}(a)\|_2^2/(2\sigma^2)$$

- Estimators: mean of μ^y, mode of μ^y, … EXPENSIVE

A. M. Stuart
Inverse problems: A Bayesian perspective.
Acta Numerica, 2010

Emulators for Inverse Problems

Statistical Estimators

- Quadrature for conditional mean estimator:

$$\hat{x}_{\mathrm{CM}} = \int_A s \, \mathrm{d}\mu^y(s) \approx \sum_{m=1}^{M} \omega_m \xi_m \frac{\mathrm{d}\mu^y}{\mu_0}(\xi_m)$$

- ***Challenge:*** evaluations of R.–N. derivative need evaluation of $\mathcal{G}$ which tends to be expensive
- Solution: use an approximation of $\mathcal{G}$
- Gaussian process: $m_{\mathcal{G}} \approx \mathcal{G}$

Approximate probability distributions

- “New” posterior distribution via Radon–Nikodym derivative:

$$\frac{\mathrm{d}\mu_N^y}{\mathrm{d}\mu_0}(a) \propto \exp\left(-\frac{\|y - m_{\mathcal{G}}(a)\|_2^2}{2\sigma^2}\right)$$

- Hope: $\mu_N^y \approx \mu^y$ as long as $m_{\mathcal{G}} \approx \mathcal{G}$

Posterior consistency

Theorem (Stuart & Teckentrup (2018); simplified)

Assuming $\mathcal{X}$ *compact and* $\mathcal{G} \in H^s(\mathcal{X})$, $s > \dim(\mathcal{X})/2$,

$$d_{Hell}(\mu^y, \mu_N^y) \leq C_2 \|\mathcal{G} - m^{\mathcal{G}}\|_{L^2}$$

A. M. Stuart and A. L. Teckentrup
Posterior consistency for Gaussian process approximations of Bayesian posterior distributions.
Mathematics of Computation, 2018

Convergence rates

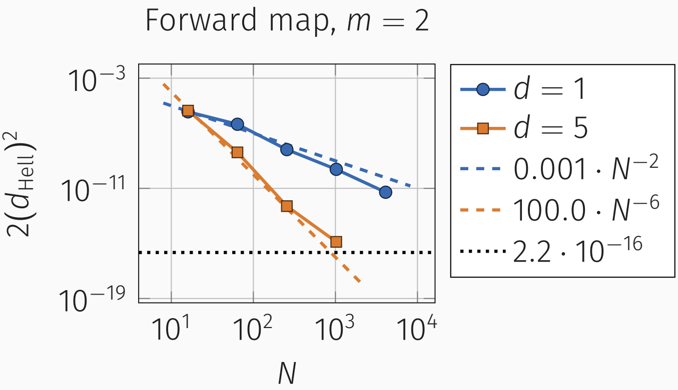


Figure 5: Approximation of Hellinger distance and expected convergence rates for a model problem, $s \in \{1, 5\}$ unknown inputs.

What we could have covered as well...

- Approximation of $\Phi(a) = \|y - \mathcal{G}(a)\|_2^2/(2\sigma^2)$ instead of $\mathcal{G}$
  $\rightarrow$ Same rates
- Use sample paths of $\mathrm{GP}(m_f, k_f)$ instead of m_f to approximate $\mathcal{G}$ or Φ
  $\rightarrow$ slightly lower rates (h^{ν} instead of $h^{\nu+m/2}$)

A. M. Stuart and A. L. Teckentrup
Posterior consistency for Gaussian process approximations of Bayesian posterior distributions.
Mathematics of Computation, 2018

Likelihood Functions

- Denote the likelihood function $\ell(a) := e^{-\Phi}(a)$
- Approximation of likelihood function gives same results!
- Let $m_\ell \approx \ell$ be a GP approx.

Theorem ($\rightarrow$ this thesis; simplified)

Assuming $\mathcal{X}$ compact and $\mathcal{G} \in H^s(\mathcal{X})$, $s > \dim(\mathcal{X})/2$,

$$d_{Hell}(\mu^y, \mu^y_{N,\ell}) \leq C_2 \|\ell - m_\ell\|_{L^2}$$

Essential to the proof

Lemma (→ this thesis; simplified)

There exist constants $C_1, C_2 > 0$ and $N_0 \in \mathbb{N}$ such that

$$C_1 \leq \mathbb{E}[m_{\ell,n}] \leq C_2$$

holds for all $n \geq N_0$.

Use additionally: Use $(a-b)^2 = (a^2-b^2)^2/(a+b)^2$ as well as straightforward inequalities

Convergence rates

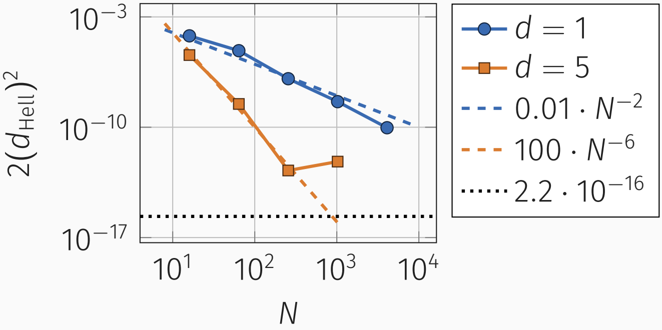


Figure 6: Approximation of Hellinger distance and expected convergence rates for a model problem, $s \in \{1, 5\}$ unknown inputs.

Furthermore

More rates:

- If using sample path–approximations, we may recover the same rate h^ν as for an approximation of $\mathcal{G}$ or Φ
- See the thesis for proofs and more simulations

Connection to Bayesian quadrature:

- Integrals of Gaussian processes are Gaussian processes
- If m_ℓ is available, $\mathbb{E}[m_\ell]$ is free

Conclusion (so far)

- Approximate posterior distributions are as good as the approximation of the forward map/potential
- The same can be shown for the likelihood function
- The rates can be observed in practice

Further work

- Slightly weaken assumptions for consistency estimates
- Emulation of vector–valued functions (in high dimension)

Numerical Linear Algebra with Covariances

Gaussian process regression

- Gaussian process regression requires solving the linear system $k(\mathcal{X}_N, \mathcal{X}_N)c = f(\mathcal{X}_N)$
- $K := k(\mathcal{X}_N, \mathcal{X}_N)$ is symmetric and positive definite
- It is also not sparse and ill–conditioned

Covariances are ill-conditioned

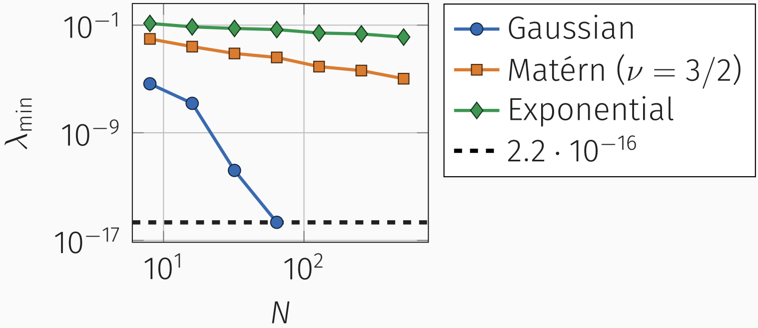
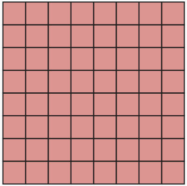


Figure 7: Smallest eigenvalue $\lambda_{\min}$ of covariance matrices constructed on Halton points in $d = 2$.

Hierarchical Matrices

- Approximate blocks of K with low rank approximations
- $K \approx H_K$ tree of low rank approximations
- Storage and MVM in $O(N \log N)$



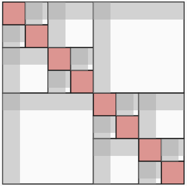


Figure 8: ***Left:*** full matrix, ***right:*** hierarchical matrix

Inverting Hierarchical Matrices

N	Relative error
400	$5.28 \cdot 10^{-8}$
800	$5.96 \cdot 10^{-7}$
1600	$5.63 \cdot 10^{-6}$
3200	$6.28 \cdot 10^{-5}$

Table 1: Relative error of solving $Kc = f$ with LU decomposition: $\mathcal{H}^2$–matrix and full matrix

MVM is more stable

N	Relative error
512	$3.9 \cdot 10^{-13}$
1024	$3.5 \cdot 10^{-13}$
2048	$6.3 \cdot 10^{-13}$
4096	$6.9 \cdot 10^{-13}$
8192	$8.3 \cdot 10^{-13}$

Table 2: Relative error of matrix–vector multiplication (MVM) using compression accuracy $\epsilon = 10^{-12}$.

Krylov Solvers & Preconditioners

- MVM is “reliable” and cheap ($O(N \log N)$)
- This motivates use of Krylov solvers: CG, GMRES, ...
- Problem: extreme ill–conditioning of covariance matrices
- Solution: Preconditioners
    - Nyström, incomplete Cholesky, ...
    - Localized Kernel spaces

Localised Kernel Spaces

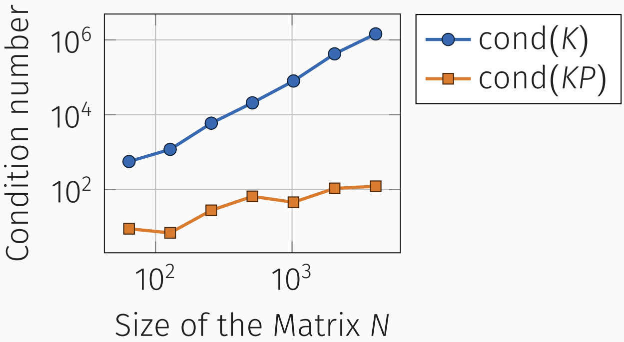


Figure 9: Condition numbers cond(K) and cond(KP) where the preconditioner is constructed with localised bases on $X = \mathbb{S}^2$.

Localised Kernel Spaces

N	n	Iter.	Runtime GMRES	RMSE of Interp.
4096	91	7	$<$ 1.0 sec	$5.2 \cdot 10^{-10}$
8192	111	11	1.0 sec	$4.8 \cdot 10^{-9}$
16384	124	12	2.0 sec	$3.9 \cdot 10^{-9}$
32768	142	8	3.0 sec	$3.3 \cdot 10^{-9}$
65536	162	13	15.0 sec	$8.4 \cdot 10^{-9}$
131072	183	8	13.0 sec	$1.4 \cdot 10^{-9}$

Table 3: GMRES with localised kernel spaces on the sphere $X = \mathbb{S}^2$; thin–plate spline kernel.

Bounded domains

N	n	Iterations	RMSE of Interp.
512	88	8	$1.3 \cdot 10^{-4}$
1024	108	13	$4.4 \cdot 10^{-5}$
2048	131	16	$1.4 \cdot 10^{-5}$
4096	156	44	$4.9 \cdot 10^{-6}$
8192	183	127	$1.2 \cdot 10^{-6}$

Table 4: GMRES with localised kernel spaces on $X = [0, 1]^2$ for the Matérn covariance.

Further work

Algorithm:

- Good *preconditioners* for covariance matrices
- Efficient assembly of hierarchical matrices

Hierarchical matrices and covariances

- Establish asymptotic smoothness of covariance functions (only done for very few)
- Error estimates w.r.t. interpolation error

Further Reading

Localised Bases for Kernel Spaces:

Localised Bases for Kernel Spaces on the Unit Sphere

E. Fuselier, T. Hangelbroek, F.J. Narcowich, J.D. Ward, G.B. Wright, SIAM Journal Numerical Analysis, 2013

ASKIT:

ASKIT: An Efficient, Parallel Library for High-Dimensional Kernel Summations

W. B. March, B. Xiao, C. D. Yu, G. Biros, SIAM Journal Scientific Computing, 2016

Further Reading

Gaussian processes, hierarchical matrices, Krylov methods:

Approximating Gaussian processes with H^2–matrices.

S. Börm, J. Garcke, ECML, 2007

Hierarchical Matrices and RBF Interpolation:

Hierarchical Matrix Approximation for Kernel-Based Scattered Data Interpolation.

A. Iske, S. Le Borne, M. Wende, SIAM Journal Scientific Computing, 2017