Numerical Analysis of Hierarchical Gaussian Process Regression

Aretha Teckentrup

School of Mathematics, University of Edinburgh
Alan Turing Institute, London

Joint work with:

Andrew Stuart (Caltech)

SIAM UQ '18 - April 17, 2018

THE UNIVERSITY *of* EDINBURGH
School of Mathematics

The Alan Turing Institute

Outline

Gaussian Process Regression

Main idea

- Gaussian process emulators (also known as *kriging*) can be viewed as a Bayesian version of interpolation.
- We are given f at design points $D_N = \{u^n\}_{n=1}^N$, obtaining function values $\{f(u^n)\}_{n=1}^N$.

Gaussian Process Regression

Main idea

- Gaussian process emulators (also known as *kriging*) can be viewed as a Bayesian version of interpolation.
- We are given f at design points $D_N = \{u^n\}_{n=1}^N$, obtaining function values $\{f(u^n)\}_{n=1}^N$.
- We interpolate f by a random function f_N, where f_N is conditioned such that $f_N(u^n) \equiv f(u^n)$, for $n = 1, \ldots, N$.
- Choosing the distribution of f_N as a Gaussian process, we obtain a Gaussian process emulator.
- The distribution of f_N is chosen to reflect the smoothness and typical length scales of f.

Gaussian Process Regression

Simple Derivation [Rasmussen, Williams '06]

- We assign a prior probability distribution to f: a Gaussian process on $U \subseteq \mathbb{R}^{d_u}$, with mean $m : U \to \mathbb{R}$ and covariance kernel $k : U \times U \to \mathbb{R}$:

$$f \sim \mathsf{GP}(m(u), k(u, u'))$$

For every $u \in U$, $f(u)$ is a Gaussian random variable with $\mathbb{E}(f(u)) = m(u)$ and $\mathbb{C}\mathrm{ov}(f(u), f(u')) = k(u, u')$.

Gaussian Process Regression

Simple Derivation [Rasmussen, Williams '06]

- We assign a prior probability distribution to f: a Gaussian process on $U \subseteq \mathbb{R}^{d_u}$, with mean $m : U \to \mathbb{R}$ and covariance kernel $k : U \times U \to \mathbb{R}$:

$$f \sim \mathsf{GP}(m(u), k(u, u'))$$

For every $u \in U$, $f(u)$ is a Gaussian random variable with $\mathbb{E}(f(u)) = m(u)$ and $\mathbb{C}\mathrm{ov}(f(u), f(u')) = k(u, u')$.

- Conditioning the prior on the given function values $\{f(u^n)\}_{n=1}^N$ leads to the posterior distribution $f_N \sim \mathsf{GP}(m_N^f(u), k_N(u, u'))$, with

$$m_N^f(u) = m(u) + k_*(u)^T K_*^{-1}(f_* - m_*),$$
$$k_N(u, u') = k(u, u') - k_*(u)^T K_*^{-1} k_*(u'),$$

and $(k_*(u))_n = k(u, u^n), (K_*)_{nm} = k(u^n, u^m)$, $(f_*)_n = f(u^n)$ and $(m_*)_n = m(u^n)$.

Gaussian Process Regression

Approximation properties

- We have $m_N^f(u^n) = f(u^n)$ and $k_N(u^n, u^n) = 0$, for $n = 1, \ldots, N$.
  $\Rightarrow f_N(u^n) \equiv m_N^f(u^n) = f(u^n)$, for $n = 1, \ldots, N$.

Gaussian Process Regression

Approximation properties

- We have $m_N^f(u^n) = f(u^n)$ and $k_N(u^n, u^n) = 0$, for $n = 1, \ldots, N$.
  $\Rightarrow f_N(u^n) \equiv m_N^f(u^n) = f(u^n)$, for $n = 1, \ldots, N$.

- The predictive mean m_N^f is an interpolant of f, and the emulator f_N is a random interpolant of f, reflecting the uncertainty in f away from the design points D_N.

Gaussian Process Regression

Approximation properties

- We have $m_N^f(u^n) = f(u^n)$ and $k_N(u^n, u^n) = 0$, for $n = 1, \ldots, N$.
  $\Rightarrow f_N(u^n) \equiv m_N^f(u^n) = f(u^n)$, for $n = 1, \ldots, N$.

- The predictive mean m_N^f is an interpolant of f, and the emulator f_N is a random interpolant of f, reflecting the uncertainty in f away from the design points D_N.

- Under certain regularity assumptions on the design points D_N and the functions f and f_N, we have

$$\|f - m_N^f\|_{L^2(U)} \to 0, \qquad \text{and} \quad \|k_N^{\frac{1}{2}}\|_{L^2(U)} \to 0,$$

  as $N \to \infty$.

Gaussian Process Regression

Choice of mean and covariance kernel

- The mean function m is typically chosen as a polynomial: $m(u) = \sum_{q=1}^{Q} \beta_q \, p_q(u)$.

Gaussian Process Regression

Choice of mean and covariance kernel

- The mean function m is typically chosen as a polynomial: $m(u) = \sum_{q=1}^{Q} \beta_q \, p_q(u)$.
- Covariance kernels frequently used are
  - the family of Matèrn covariances

$$k_{\mathrm{Mat}}(u, u') = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\|u - u'\|}{\lambda} \right)^{\nu} B_\nu \left(\frac{\|u - u'\|}{\lambda} \right),$$

    with smoothness parameter $\nu > 0$, marginal variance $\sigma^2 > 0$ and correlation length $\lambda > 0$.

$$\nu = 1/2 \ : \ \sigma^2 \exp\big(- \tfrac{\|u-u'\|}{\lambda} \big), \ \nu = \infty \ : \ \exp\big(- \tfrac{\|u-u'\|^2}{\lambda^2} \big).$$

  - the family of separable Matèrn covariances

$$k_{\mathrm{sepMat}}(u, u') = \prod_{i=1}^{d_u} k_{\mathrm{Mat}}(u_i, u_i').$$

Gaussian Process Regression

Choice of mean and covariance kernel

- The mean function m is typically chosen as a polynomial: $m(u) = \sum_{q=1}^{Q} \beta_q \, p_q(u)$.
- Covariance kernels frequently used are
  - the family of Matèrn covariances

$$k_{\mathrm{Mat}}(u, u') = \frac{\sigma^2}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\|u - u'\|}{\lambda} \right)^{\nu} B_\nu \left(\frac{\|u - u'\|}{\lambda} \right),$$

    with smoothness parameter $\nu > 0$, marginal variance $\sigma^2 > 0$ and correlation length $\lambda > 0$.

$$\nu = 1/2 \ : \ \sigma^2 \exp\left(- \tfrac{\|u-u'\|}{\lambda} \right), \ \nu = \infty \ : \ \exp\left(- \tfrac{\|u-u'\|^2}{\lambda^2} \right).$$

  - the family of separable Matèrn covariances

$$k_{\mathrm{sepMat}}(u, u') = \prod_{i=1}^{d_u} k_{\mathrm{Mat}}(u_i, u_i').$$

- The hyper-parameters θ are unknown a-priori.

Gaussian Process Regression

Empirical Bayes'

- We use an empirical Bayes' (or plug-in) approach, where we estimate values of the hyper-parameters from $\{f(u^n)\}_{n=1}^N$ and plug these into the posterior distribution f_N.
- This gives a sequence of estimates $\widehat{\theta}_N$, which can be found via maximum likelihood estimation, maximum a-posteriori estimation, cross validation, ...
- We assume that there is a true parameter value θ_0, defined in a suitable way.

Convergence bounds

Matèrn kernels: convergence as $N \to \infty$

With design points $D_N = \{u^n\}_{n=1}^N$, define:

$$\text{fill distance} \quad h_{D_N} = \max_{u \in U} \min_{u^n \in D_N} \|u - u^n\|, \qquad h_{D_N} \sim N^{-1/d_u},$$

$$\text{mesh ratio} \quad \rho_{D_N} = \frac{\max_{u \in U} \min_{u^n \in D_N} \|u - u^n\|}{\frac{1}{2} \min_{n \neq l} \|u^n - u^l\|}, \qquad \rho_{D_N} \geq 1.$$

Theorem [Stuart, ALT in prep.]

Under certain regularity conditions, with covariance kernel k_{Mat}, we have

$$\|f - m_N^f(\widehat{\theta}_N)\|_{L^2(U)} \leq C(\widehat{\theta}_N) h_{D_N}^{\min\{\tilde{\tau}, \widehat{\nu}_N + \frac{K}{2}\}} \rho_{D_N}^{\max\{\widehat{\nu}_N + \frac{K}{2} - \tilde{\tau}, 0\}} \left(\|f\|_{H^{\tilde{\tau}}(U)} + \|m(\widehat{\theta}_N)\|_{H^{\tilde{\tau}}(U)} \right),$$

with C independent of f. Furthermore,

$$\|k_N^{\frac{1}{2}}\|_{L^2(U)} \leq C(\widehat{\theta}_N) h_{D_N}^{\min\{\tilde{\tau} - \frac{K}{2}, \widehat{\nu}_N\}} \rho_{D_N}^{\max\{\widehat{\nu}_N + \frac{K}{2} - \tilde{\tau}, 0\}}.$$

Convergence bounds

Separable Matèrn kernels: convergence as $N \to \infty$

Theorem [Stuart, ALT in prep.]

Under certain regularity conditions, with covariance kernel k_{sepMat} and

- tensor product domain $U = \prod_{k=1}^{K} U_k$,
- D_N chosen as a Smolyak sparse grid,

we have

$$\|f - m_N^f(\widehat{\theta}_N)\|_{L^2(U)} \leq C(\widehat{\theta}_N) N^{-\alpha(\widehat{\nu}_N)} |\log N|^{\tilde{\alpha}(\widehat{\nu}_N, K)} \left(\|f\|_{H^{\{\tilde{r}_k\}}_{\otimes^K}(U)} + \|m(\widehat{\theta}_N)\|_{H^{\{\tilde{r}_k\}}_{\otimes^K}(U)} \right),$$

with C independent of f. Furthermore,

$$\|k_N^{\frac{1}{2}}\|_{L^2(U)} \leq C(\widehat{\theta}_N) N^{-\alpha(\widehat{\nu}_N)+\frac{1}{2}} |\log N|^{\tilde{\alpha}(\widehat{\nu}_N, K)}.$$

Convergence bounds

Convergence as $\widehat{\theta}_N \to \theta_0$

Theorem [Stuart, ALT in prep.]

Under certain regularity conditions, with covariance kernel k_{Mat} or k_{sepMat}, we have for fixed $N \in \mathbb{N}$ and $\theta \to \theta_0$

$$\|m_N^f(\theta) - m_N^f(\theta_0)\|_{H^\kappa(U) \,/\, H^{\{\kappa_k\}}_{\otimes K}(U)} \to 0,$$

$$\|k_N^{1/2}(\theta) - k_N^{1/2}(\theta_0)\|_{L^2(U)} \to 0,$$

for all κ / $\{\kappa_k\}$ sufficiently small.

Application in Bayesian inverse problems

Bayesian posterior distribution

- We are interested in $\mu^y(u)$ being the posterior distribution in a Bayesian inverse problem (parameter identification problem):

$$\frac{d\mu^y}{d\mu_0}(u) \propto e^{-\|y-F(u)\|^2_{\Gamma^{-1}}}, \qquad \left(\pi^y(u) \propto e^{-\|y-F(u)\|^2_{\Gamma^{-1}}}\ \pi_0(u)\right).$$

- This arises from
  - incorporating knowledge on u in a prior distribution μ_0 (with density π_0),
  - observing data $y = F(u) + \eta$, with noise $\eta \sim N(0, \Gamma)$,
  - conditioning μ_0 on y, resulting in the posterior distribution μ^y (with density π^y).

Application in Bayesian inverse problems

Approximation with Gaussian process emulators

- The map F is often very expensive to simulate, e.g. involving the solution to a differential equation.
- Approximating the data log-likelihood $\Phi(u) = \|y - F(u)\|^2_{\Gamma^{-1}}$ (or directly $F(u)$) with a Gaussian process emulator results in an approximate posterior distribution μ_N^y.

Application in Bayesian inverse problems

Approximation with Gaussian process emulators

- The map F is often very expensive to simulate, e.g. involving the solution to a differential equation.
- Approximating the data log-likelihood $\Phi(u) = \|y - F(u)\|^2_{\Gamma^{-1}}$ (or directly $F(u)$) with a Gaussian process emulator results in an approximate posterior distribution μ^y_N.
- The error between μ^y and μ^y_N (measured in the Hellinger distance) can be bounded in terms of $\|\Phi - m^\Phi_N\|_{L^2_{\mu_0}(U)}$ and $\|k_N^{1/2}\|_{L^2_{\mu_0}(U)}$.
- For more details, see [Stuart, ALT '18].

References

- C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, (2006).
- A. M. Stuart, *Inverse Problems: A Bayesian Perspective*, Acta Numerica, 19 (2010), pp. 451–559.
- A. M. Stuart and A. L. Teckentrup, *Numerical Analysis of Hierarchical Gaussian Process Regression and Applications in Bayesian Inverse Problems*.
  In preparation.
- ——, *Posterior consistency for Gaussian process approximations of Bayesian posterior distributions*, Mathematics of Computation, 87 (2018), pp. 721–753.
- H. Wendland, *Scattered Data Approximation*, Cambridge University Press, 2005.