Gaussian Process Approximations in Bayesian Inverse Problems

Nicholas Krämer

Master Thesis Colloquium

December 6, 2018

Outline

1. Radial basis functions and Gaussian process regression
2. Bayesian approach to inverse problems
3. Gaussian process approximations in Bayesian inverse problems
4. Current and upcoming work

Radial basis functions and Gaussian process regression

Interpolation with radial basis functions

- $\star$ Recover $f : \Omega \to \mathbb{R}$ from values $y = f(X)$ on $X = \{x_1, ..., x_N\}$

Interpolation with radial basis functions

Data

Interpolation with radial basis functions

- $\star$ Recover $f : \Omega \to \mathbb{R}$ from values $y = f(X)$ on $X = \{x_1, ..., x_N\}$
- $\star$ Kernel: radial basis functions (RBF) $K(x, y) = \varphi(\|x - y\|)$

Interpolation with radial basis functions

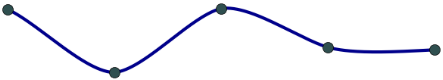
- $\star$ Recover $f : \Omega \to \mathbb{R}$ from values $y = f(X)$ on $X = \{x_1, ..., x_N\}$
- $\star$ Kernel: radial basis functions (RBF) $K(x, y) = \varphi(\|x - y\|)$
- $\star$ Approximate with kernel evaluations

$$f(x) \approx \mathcal{P}_f(x) = \sum_{i=1}^{N} c_i K(x, x_i)$$

Interpolation with radial basis functions

Data

Interpolation with radial basis functions



Data and RBF interpolant

Interpolation with radial basis functions

- ⋆ Recover $f : \Omega \to \mathbb{R}$ from values $y = f(X)$ on $X = \{x_1, ..., x_N\}$
- ⋆ Kernel: radial basis functions (RBF) $K(x, y) = \varphi(\|x - y\|)$
- ⋆ Approximate with kernel evaluations

$$f(x) \approx \mathcal{P}_f(x) = \sum_{i=1}^{N} c_i K(x, x_i)$$

- ⋆ Matrix notation $\mathcal{P}_f(X_{\text{new}}) = K(X_{\text{new}}, X)(K(X, X))^{-1} y$

Interpolation with radial basis functions

- ⋆ Recover $f : \Omega \to \mathbb{R}$ from values $y = f(X)$ on $X = \{x_1, ..., x_N\}$
- ⋆ Kernel: radial basis functions (RBF) $K(x, y) = \varphi(\|x - y\|)$
- ⋆ Approximate with kernel evaluations

$$f(x) \approx \mathcal{P}_f(x) = \sum_{i=1}^{N} c_i K(x, x_i)$$

- ⋆ Matrix notation $\mathcal{P}_f(X_{\text{new}}) = K(X_{\text{new}}, X)(K(X, X))^{-1} y$
- ⋆ "Works well" in reproducing kernel Hilbert space (RKHS)

Interpolation with radial basis functions

- Recover $f : \Omega \to \mathbb{R}$ from values $y = f(X)$ on $X = \{x_1, ..., x_N\}$
- Kernel: radial basis functions (RBF) $K(x, y) = \varphi(\|x - y\|)$
- Approximate with kernel evaluations

$$f(x) \approx \mathcal{P}_f(x) = \sum_{i=1}^{N} c_i K(x, x_i)$$

- Matrix notation $\mathcal{P}_f(X_{\text{new}}) = K(X_{\text{new}}, X)(K(X, X))^{-1} y$
- "Works well" in reproducing kernel Hilbert space (RKHS)
- For example Matérn kernel $\varphi_\nu(z) \sim z^\nu K_\nu(z)$

Different point of view

- ⋆ We still try to recover f from its values on X

Different point of view

- $\star$ We still try to recover f from its values on X
- $\star$ We assume f is a Gaussian process: $f \sim \mathrm{GP}(m(\cdot), K(\cdot, \cdot))$

Different point of view

- $\star$ We still try to recover f from its values on X
- $\star$ We assume f is a Gaussian process: $f \sim \mathrm{GP}(m(\cdot), K(\cdot, \cdot))$
- $\star$ Assume $m \equiv 0$

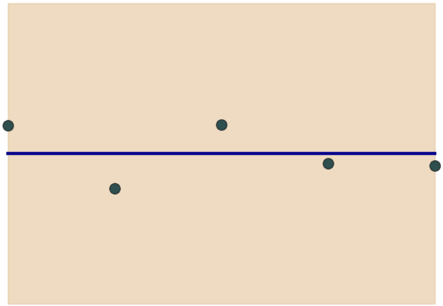
Definition (Gaussian process)

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution

Different point of view

Data

Different point of view



Mean and standard deviation of Gaussian process

Different point of view

- $\star$ We still try to recover f from its values on X
- $\star$ We assume f is a Gaussian process: $f \sim \mathrm{GP}(m(\cdot), K(\cdot,\cdot))$,
- $\star$ Assume $m \equiv 0$

Definition (Gaussian process)

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution

- $\star$ $K(\cdot,\cdot)$ symmetric positive definite - covariances are kernels

Different point of view

- $\star$ We still try to recover f from its values on X
- $\star$ We assume f is a Gaussian process: $f \sim \mathrm{GP}(m(\cdot), K(\cdot, \cdot))$,
- $\star$ Assume $m \equiv 0$

Definition (Gaussian process)

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution

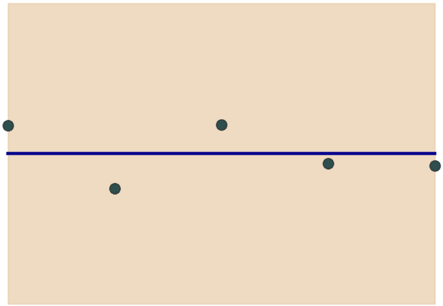
- $\star$ $K(\cdot, \cdot)$ symmetric positive definite - covariances are kernels
- $\star$ We want to reproduce f at points X_{new}

- ⋆ Condition joint distribution $(f(X), f(X_{\text{new}}))$ on hitting y

Different point of view

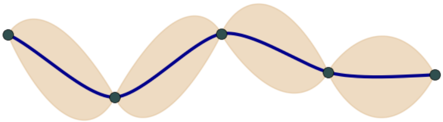
Data

Different point of view



Mean and standard deviation of Gaussian process

Conditioning Gaussian processes



Predictive mean and standard deviation of Gaussian process

Conditioning Gaussian processes

- ⋆ Condition joint distribution $(f(X), f(X_{\text{new}}))$ on hitting y
- ⋆ Conditioning suggests $f(X_{\text{new}}) \sim \mathcal{N}(m_{\text{new}}, K_{\text{new}})$ with predictive mean

$$m_{\text{new}}(X_{\text{new}}) = K(X_{\text{new}}, X)(K(X, X))^{-1} f(X)$$

Conditioning Gaussian processes

- Condition joint distribution $(f(X), f(X_{\text{new}}))$ on hitting y
- Conditioning suggests $f(X_{\text{new}}) \sim \mathcal{N}(m_{\text{new}}, K_{\text{new}})$ with predictive mean

$$m_{\text{new}}(X_{\text{new}}) = K(X_{\text{new}}, X)(K(X, X))^{-1} f(X)$$

- The predictive mean m_{new} is the RBF interpolant!

Bayesian approach to inverse problems

Example for an inverse problem

Differential Equation on $(0,1)$

Find $a = (a_1, a_2) \in [-1,1]^2$ such that for

$$-\operatorname{div}((\sin(a_1 x) + \cos(a_2 x))\nabla u(x)) = 1, \quad u(0) = u(1) = 0$$

the measurements satisfy

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

Example for an inverse problem

Differential Equation on $(0,1)$

Find $a = (a_1, a_2) \in [-1,1]^2$ such that for

$$-\operatorname{div}((\sin(a_1 x) + \cos(a_2 x))\nabla u(x)) = 1, \quad u(0) = u(1) = 0$$

the measurements satisfy

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

- ⋆ Operator $\mathcal{G} : \mathbb{R}^2 \to \mathbb{R}^3,\ (a_1, a_2) \mapsto (u(1/3), u(1/2), u(2/3))$

Example for an inverse problem

Differential Equation on $(0,1)$

Find $a = (a_1, a_2) \in [-1,1]^2$ such that for

$$-\operatorname{div}((\sin(a_1 x) + \cos(a_2 x))\nabla u(x)) = 1, \quad u(0) = u(1) = 0$$

the measurements satisfy

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

- ⋆ Operator $\mathcal{G} : \mathbb{R}^2 \to \mathbb{R}^3,\ (a_1, a_2) \mapsto (u(1/3), u(1/2), u(2/3))$
- ⋆ How can we find a from $y = \mathcal{G}(a) +$ "measurement error"

Our setting:

Inverse problems

Our setting:

- $\star$ Space of input parameters $\mathcal{A} \subseteq \mathbb{R}^m$ compact, $m \in \mathbb{N}$

Inverse problems

Our setting:

- ⋆ Space of input parameters $\mathcal{A} \subseteq \mathbb{R}^m$ compact, $m \in \mathbb{N}$
- ⋆ Parameter-to-observation operator $\mathcal{G} : \mathcal{A} \to \mathbb{R}^n$

Inverse problems

Our setting:

- $\star$ Space of input parameters $\mathcal{A} \subseteq \mathbb{R}^m$ compact, $m \in \mathbb{N}$
- $\star$ Parameter-to-observation operator $\mathcal{G} : \mathcal{A} \to \mathbb{R}^n$
- $\star$ Noisy measurements $y \in \mathbb{R}^n$, noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I_n)$

Inverse problems

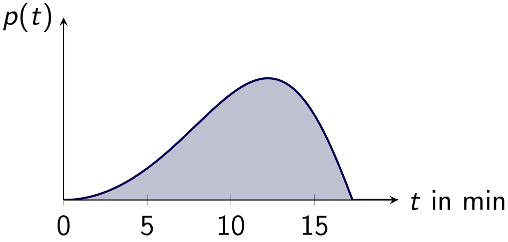
Our setting:

- $\star$ Space of input parameters $\mathcal{A} \subseteq \mathbb{R}^m$ compact, $m \in \mathbb{N}$
- $\star$ Parameter-to-observation operator $\mathcal{G} : \mathcal{A} \to \mathbb{R}^n$
- $\star$ Noisy measurements $y \in \mathbb{R}^n$, noise $\eta \sim \mathcal{N}(0, \sigma_\eta^2 I_n)$
- $\star$ Find input $a \in \mathcal{A}$ such that $y = \mathcal{G}(a)$ or $y = \mathcal{G}(a) + \eta$

Bayesian statistics

- ⋆ Knowledge is probability distribution

⋆ Knowledge is probability distribution



Bayesian answer to: "When will he finish?"

Bayesian statistics

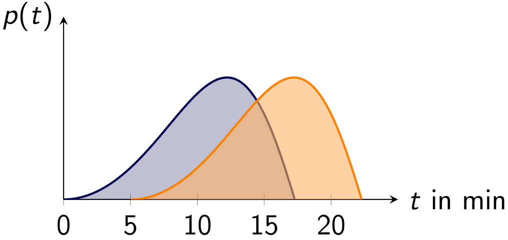
- ⋆ Knowledge is probability distribution
- ⋆ Initial belief: prior distribution μ_0 on $\mathcal{A}$

Bayesian statistics

- ⋆ Knowledge is probability distribution
- ⋆ Initial belief: prior distribution μ_0 on $\mathcal{A}$
- ⋆ Condition unknown to match collected data: posterior distribution μ^y

Bayesian statistics

- Knowledge is probability distribution
- Initial belief: prior distribution μ_0 on $\mathcal{A}$
- Condition unknown to match collected data: posterior distribution μ^y



Updated answer after collecting data ("How long did 2 bullet pts. take?")

How do we find the posterior distribution?

Theorem (Bayes, simplified)

Let $\mathcal{G} \in C(\mathcal{A}; \mathbb{R}^n)$ *and* $\mu_0(\mathcal{A}) = 1$. *Then* $\mu^y \ll \mu_0$ *with density*

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(a) \propto \exp\left(-\frac{1}{2\sigma_\eta^2}\|y - \mathcal{G}(a)\|^2\right)$$

How do we find the posterior distribution?

Theorem (Bayes, simplified)

Let $\mathcal{G} \in C(\mathcal{A}; \mathbb{R}^n)$ and $\mu_0(\mathcal{A}) = 1$. Then $\mu^y \ll \mu_0$ with density

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(a) \propto \exp\left(-\frac{1}{2\sigma_\eta^2}\|y - \mathcal{G}(a)\|^2\right)$$

Exploit this density to extract information

How do we find the posterior distribution?

Theorem (Bayes, simplified)

Let $\mathcal{G} \in C(\mathcal{A}; \mathbb{R}^n)$ and $\mu_0(\mathcal{A}) = 1$. Then $\mu^y \ll \mu_0$ with density

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(a) \propto \exp\left(-\frac{1}{2\sigma_\eta^2}\|y - \mathcal{G}(a)\|^2\right)$$

Exploit this density to extract information

- ⋆ Conditional mean and higher moments → numerical cubature

How do we find the posterior distribution?

Theorem (Bayes, simplified)

Let $\mathcal{G} \in C(\mathcal{A}; \mathbb{R}^n)$ and $\mu_0(\mathcal{A}) = 1$. Then $\mu^y \ll \mu_0$ with density

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(a) \propto \exp\left(-\frac{1}{2\sigma_\eta^2}\|y - \mathcal{G}(a)\|^2\right)$$

Exploit this density to extract information

- ⋆ Conditional mean and higher moments $\rightarrow$ numerical cubature
- ⋆ Maximum-a-posteriori estimator (mode)

Gaussian process approximations in Bayesian inverse problems

Computational bottleneck

- ⋆ What does $\mathcal{G}$ actually do at each evaluation?

Computational bottleneck

⋆ What does $\mathcal{G}$ actually do at each evaluation?

$$\mathcal{G} : \mathcal{A} \xrightarrow{\text{(solve PDE)}} V \xrightarrow{\text{(evaluate sol.)}} \mathbb{R}^n$$

Computational bottleneck

- What does $\mathcal{G}$ actually do at each evaluation?

$$\mathcal{G} : \mathcal{A} \xrightarrow{\text{(solve PDE)}} V \xrightarrow{\text{(evaluate sol.)}} \mathbb{R}^n$$

- Evaluating $\mathcal{G}$ is expensive!

Computational bottleneck

- What does $\mathcal{G}$ actually do at each evaluation?

$$\mathcal{G} : \mathcal{A} \xrightarrow{\text{(solve PDE)}} V \xrightarrow{\text{(evaluate sol.)}} \mathbb{R}^n$$

- Evaluating $\mathcal{G}$ is expensive!
- $\mathcal{A} \subseteq \mathbb{R}^m$, hence $\mathcal{G}$ is essentially a map from $\mathbb{R}^m$ to $\mathbb{R}^n$

Computational bottleneck

- $\star$ What does $\mathcal{G}$ actually do at each evaluation?

$$\mathcal{G} : \mathcal{A} \xrightarrow{\text{(solve PDE)}} V \xrightarrow{\text{(evaluate sol.)}} \mathbb{R}^n$$

- $\star$ Evaluating $\mathcal{G}$ is expensive!
- $\star$ $\mathcal{A} \subseteq \mathbb{R}^m$, hence $\mathcal{G}$ is essentially a map from $\mathbb{R}^m$ to $\mathbb{R}^n$
- $\star$ $\mathcal{G}$ has certain regularity (high at least for "simple" PDEs)

Computational bottleneck

- What does $\mathcal{G}$ actually do at each evaluation?

$$\mathcal{G} : \mathcal{A} \xrightarrow{\text{(solve PDE)}} V \xrightarrow{\text{(evaluate sol.)}} \mathbb{R}^n$$

- Evaluating $\mathcal{G}$ is expensive!
- $\mathcal{A} \subseteq \mathbb{R}^m$, hence $\mathcal{G}$ is essentially a map from $\mathbb{R}^m$ to $\mathbb{R}^n$
- $\mathcal{G}$ has certain regularity (high at least for “simple” PDEs)
- **Good setting for approximations!**

Approximations in Bayesian inverse problems

- Replace $\mathcal{G}$ by its GP/RBF approximation

$$\mathcal{G}(a) \approx m^{\mathcal{G}}(a) = \sum_{i=1}^{N} c_i K(a, a_i)$$

Approximations in Bayesian inverse problems

- $\star$ Replace $\mathcal{G}$ by its GP/RBF approximation

$$\mathcal{G}(a) \approx m^{\mathcal{G}}(a) = \sum_{i=1}^{N} c_i K(a, a_i)$$

- $\star$ Solve $K(X, X)c = f(X)$ **once**

Approximations in Bayesian inverse problems

- ⋆ Replace $\mathcal{G}$ by its GP/RBF approximation

$$\mathcal{G}(a) \approx m^{\mathcal{G}}(a) = \sum_{i=1}^{N} c_i K(a, a_i)$$

- ⋆ Solve $K(X, X)c = f(X)$ **once**
- ⋆ Replace μ^y by approximate posterior distribution μ^y_{app}

- Replace $\mathcal{G}$ by its GP/RBF approximation

$$\mathcal{G}(a) \approx m^{\mathcal{G}}(a) = \sum_{i=1}^{N} c_i K(a, a_i)$$

- Solve $K(X, X)c = f(X)$ **once**
- Replace μ^y by approximate posterior distribution μ^y_{app}
- Is $\mu^y_{\text{app}} \approx \mu^y$ for $\mathcal{G} \approx m^{\mathcal{G}}$?

Approximations in Bayesian inverse problems

- Replace $\mathcal{G}$ by its GP/RBF approximation

$$\mathcal{G}(a) \approx m^{\mathcal{G}}(a) = \sum_{i=1}^{N} c_i K(a, a_i)$$

- Solve $K(X, X)c = f(X)$ **once**
- Replace μ^y by approximate posterior distribution μ^y_{app}
- Is $\mu^y_{\text{app}} \approx \mu^y$ for $\mathcal{G} \approx m^{\mathcal{G}}$?
- New approximation results (Stuart, Teckentrup (2018))

Theoretical preliminaries

Assumptions

1. $\mathcal{G} \in H^{\nu+m/2}(\mathcal{A}; \mathbb{R}^n)$ for some $\nu > 0$ (recall $\mathcal{A} \subseteq \mathbb{R}^m$)
2. $\lim_{N\to\infty} \sup_{u\in\mathcal{A}} \|\mathcal{G}(u) - m^{\mathcal{G}}(u)\| = 0$
3. $\sup_{u\in\mathcal{A}} \|\mathcal{G}(u)\| \leq C_{\mathcal{G}} < \infty$

Theoretical preliminaries

Assumptions

1. $\mathcal{G} \in H^{\nu+m/2}(\mathcal{A}; \mathbb{R}^n)$ for some $\nu > 0$ (recall $\mathcal{A} \subseteq \mathbb{R}^m$)
2. $\lim_{N\to\infty} \sup_{u\in\mathcal{A}} \|\mathcal{G}(u) - m^{\mathcal{G}}(u)\| = 0$
3. $\sup_{u\in\mathcal{A}} \|\mathcal{G}(u)\| \leq C_{\mathcal{G}} < \infty$

Hellinger distance:

$$d_{\mathrm{Hell}}(\mu_1, \mu_2) = \left(\frac{1}{2} \int_{\mathcal{A}} \left(\sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_0}} - \sqrt{\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_0}} \right)^2 \, \mathrm{d}\mu_0 \right)^{1/2}$$

Main results

Theorem (Stuart, Teckentrup (2018))

Under the previous assumptions, there exists C_2 independent of X and N such that

$$d_{Hell}(\mu^y, \mu^y_{app}) \leq C_2 \|\mathcal{G} - m^{\mathcal{G}}\|_{L^2_{\mu_0}(\mathcal{A},\mathbb{R}^n)}$$

Main results

Theorem (Stuart, Teckentrup (2018))

Under the previous assumptions, there exists C_2 independent of X and N such that

$$d_{Hell}(\mu^y, \mu^y_{app}) \leq C_2 \|\mathcal{G} - m^{\mathcal{G}}\|_{L^2_{\mu_0}(\mathcal{A},\mathbb{R}^n)}$$

One can also...

Theorem (Stuart, Teckentrup (2018))

Under the previous assumptions, there exists C_2 independent of X and N such that

$$d_{Hell}(\mu^y, \mu^y_{app}) \leq C_2 \|\mathcal{G} - m^{\mathcal{G}}\|_{L^2_{\mu_0}(\mathcal{A},\mathbb{R}^n)}$$

One can also...

- ...use the full process instead of only the mean,

Main results

Theorem (Stuart, Teckentrup (2018))

Under the previous assumptions, there exists C_2 independent of X and N such that

$$d_{Hell}(\mu^y, \mu^y_{app}) \leq C_2 \|\mathcal{G} - m^{\mathcal{G}}\|_{L^2_{\mu_0}(\mathcal{A},\mathbb{R}^n)}$$

One can also...

- ...use the full process instead of only the mean, e.g. sample a random approximation from $\mathrm{GP}(m^{\mathcal{G}}(\cdot), K^{\mathcal{G}}(\cdot,\cdot))$

Theorem (Stuart, Teckentrup (2018))

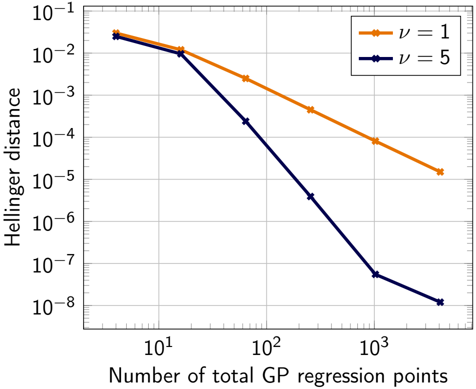
Under the previous assumptions, there exists C_2 independent of X and N such that

$$d_{Hell}(\mu^y, \mu^y_{app}) \leq C_2 \|\mathcal{G} - m^{\mathcal{G}}\|_{L^2_{\mu_0}(\mathcal{A}, \mathbb{R}^n)}$$

One can also...

- ⋆ ...use the full process instead of only the mean, e.g. sample a random approximation from $\mathrm{GP}(m^{\mathcal{G}}(\cdot), K^{\mathcal{G}}(\cdot, \cdot))$
- ⋆ ...do all of this with $\Phi(a) = \frac{1}{2\sigma_\eta^2}\|y - \mathcal{G}(a)\|^2$ instead of $\mathcal{G}$

Hellinger distance for problem with $m = 2$ inputs



Different regularities of GP approximation on uniform tensor grid

Takeaway messages

1. RBF-interpolants and predictive means are the same

Takeaway messages

1. RBF-interpolants and predictive means are the same
2. Error estimates from RBF-interpolation come from the regularity of the kernel

Takeaway messages

1. RBF-interpolants and predictive means are the same
2. Error estimates from RBF-interpolation come from the regularity of the kernel
3. Forward maps from Bayesian inverse problems are expensive to evaluate

Takeaway messages

1. RBF-interpolants and predictive means are the same
2. Error estimates from RBF-interpolation come from the regularity of the kernel
3. Forward maps from Bayesian inverse problems are expensive to evaluate
4. They are easy to approximate with radial basis functions

Current and upcoming work

Currently: “Tidying up” some theory (for myself)

- Some approximation errors seem neglected
  - Forward model is only approximately available (FEM for PDE)
  - Numerical error in $\cong$ “noise” in evaluations
- Which error has which influence...
  - ... on the conditional mean?
  - ... on the hellinger distance?
- Some quantities seem arbitrary
  - Why the Matérn kernel–which parameters?
  - **Which pointset for GP approximation?**

Soon: Optimising the choice of GP locations

- Pick design points intelligently
- *Bayesian optimisation*
- Non-adaptively: *experimental design*
- Adaptively: *sequential design*
- Make computations a little bit more efficient
- ...while trying not to blow up the runtime with unnecessary optimisations
- More about this next time

Further readings on radial basis functions

RBF interpolation:
Scattered Data Approximation
H. Wendland, Cambridge University Press, 2004

Relationship between GP regression and RBF interpolation:
Interpolation of spatial data–a stochastic or a deterministic problem?
M. Scheuerer, R. Schaback, M. Schlather, European Journal of Applied Mathematics, 2013

Further readings on Bayesian inverse problems

Bayesian approach to inverse problems:
The Bayesian approach to inverse problems
M. Dashti, A. M. Stuart, Handbook of Uncertainty Quantification, Springer, 2017

GP approximations in Bayesian inverse problems:
Posterior consistency for Gaussian process approximations of Bayesian posterior distributions
A. M. Stuart, A.L. Teckentrup, Mathematics of Computation, 2018

Thanks!