On Gaussian Processes and their Interpretability

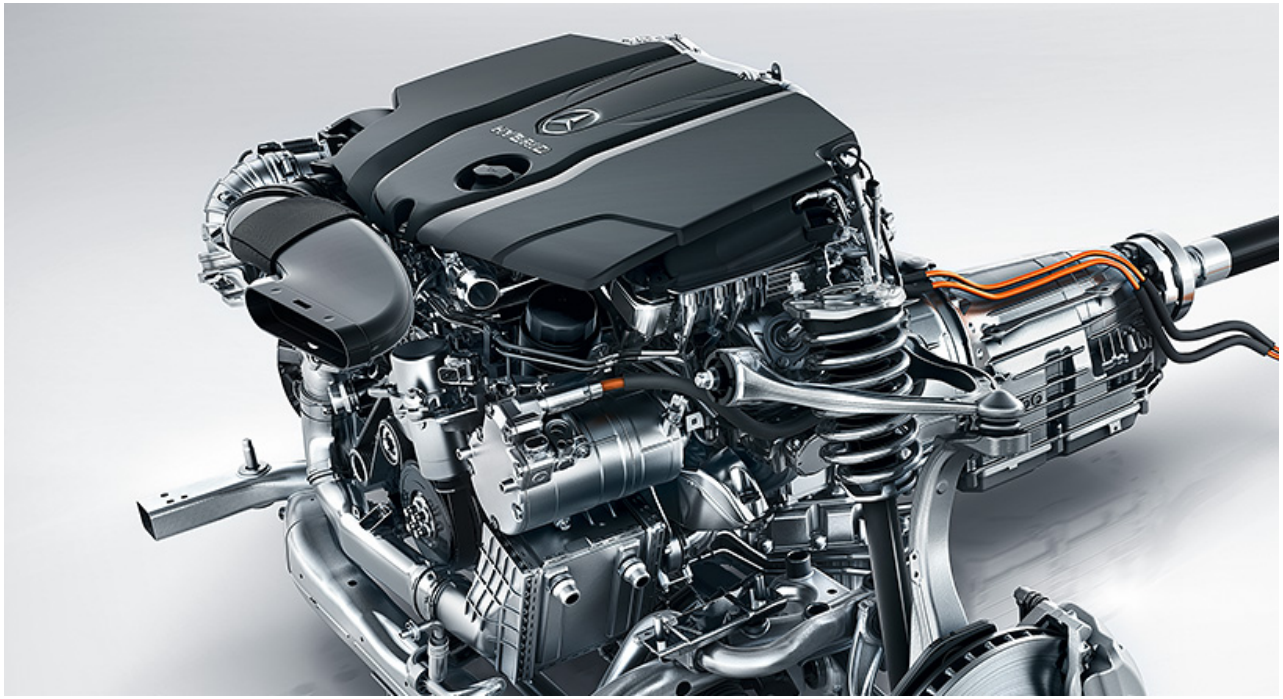
Markus Lange-Hegermann (inIT, TH OWL)

2022/03/25, Spring School DataNinja

Motivation: Diesel engine calibration

Processes Combustion Calibration Diesel Engine

measure engine → math model → optimize



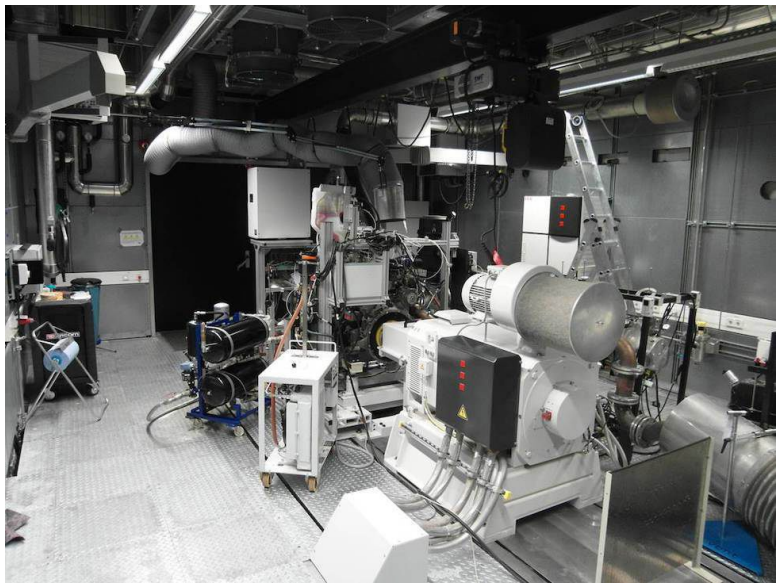
http://www.mercedes-benz.com.au

Processes Combustion Calibration Diesel Engine

measure engine → math model → optimize

Measure engine

- Measurements at engine test bench
- Measurement **costly** in time and money
- Plan measurements carefully (design of experiments, adaptive)



http://www.mechatronics.rwth-aachen.de

Processes Combustion Calibration Diesel Engine

measure engine → math model → optimize

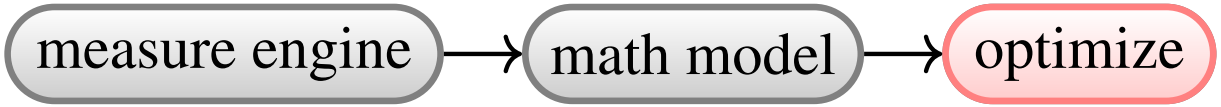
Use **measurement data** to create a **model**

$$\left\{\begin{array}{l} \textit{engine speed} \\ \textit{torque} \\ \text{main injection timing} \\ \text{rail pressure} \\ \text{turbo boost} \\ \text{exhaust gas recirculation} \\ \text{pre-/postinjection timings\&quantities} \\ \ldots \end{array}\right\} \rightarrow \left\{\begin{array}{l} \text{fuel consumption} \\ \text{NO}_x \text{ emissions} \\ \text{particulate emissions} \\ \text{noise} \\ \text{temperatures} \\ \text{pressures} \\ \text{actuators} \\ \ldots \end{array}\right.$$

Such models are almost universally Gaussian processes:

- A model class with strong inductive biases
- Work well with few data points
- Allow inclusion of expert knowledge (engineers want to tinker)

Processes Combustion Calibration Diesel Engine



Optimize

This is another topic...

Table of contents

- What is a GP?
- Why do GPs appear in applications with little data?
- Why do GPs appear in physical applications?

Interpretability of GPs

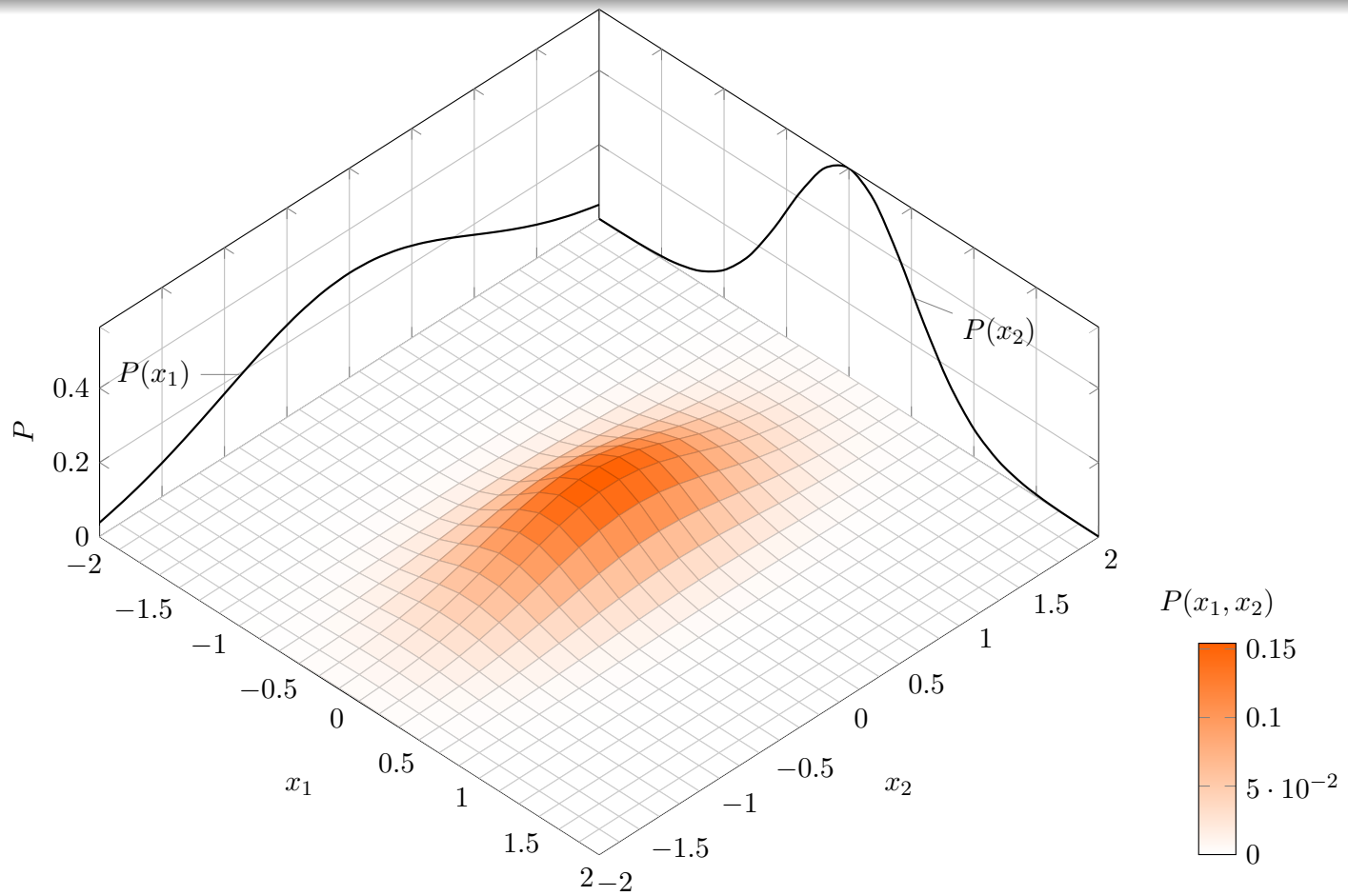
- Dictionary: inductive bias of GPs $\leftrightarrow$ math
- Inductive bias can be specified by domain experts
- Inductive bias can depend on parameters, which can be learned
- Well calibrated model uncertainties
- And more

Math asides will be typeset small

Feel free to aks questions anytime!

Reminder: Gaussian Distributions

Gaussian Distribution on $\mathbb{R}^n$



Density: $\frac{(2\pi)^{-\frac{1}{2}n}}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

Why is the Gaussian distribution so ubiquitous?

Gaussian Distribution: Properties

Theorem

The **Gaussian distribution maximizes the entropy** among all probability distributions on $\mathbb{R}^n$ with fixed mean and (co)variance.

Maximum entropy prior (Jaynes)

Known/suspected mean and (co)variance: take **Gaussian prior**.

Gaussian distribution: properties

Corollaries (colloquial)

"Everything" described by the **first two moments/cumulants**.

- uncorrelated $\implies$ independent.
- Central limit theorem

  iid random variables with finite mean and variance converge to a Gaussian distribution by averaging.
- Closed under marginal distributions
  Trivial to **compute**: drop the marginalized part
- Closed under conditional distributions
  Compute via linear algebra:
  $\mu_{x_1|x_2=a} = \mu_{x_1} + \Sigma_{x_1,x_2}\Sigma^{-1}_{x_2,x_2}(a - \mu_{x_2})$
  $\Sigma_{x_1,x_1|x_2=a} = \Sigma_{x_1,x_1} - \Sigma_{x_1,x_2}\Sigma^{-1}_{x_2,x_2}\Sigma_{x_2,x_1}$
- Sampling is possible
  Compute via linear algebra: diagonalize covariance

Questions?

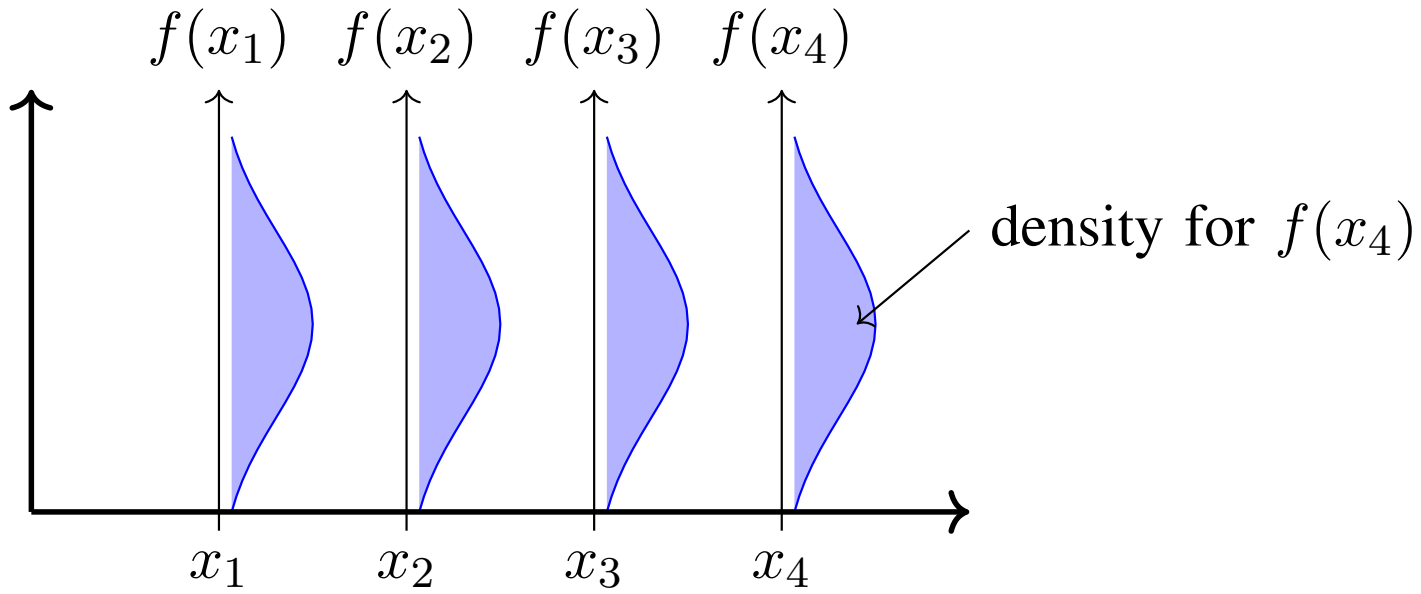
Questions?

Gaussian processes

Regression: Gaussian Processes

Idea

Assume **Gaussian function values** of the regression function f.
Marginalization: only consider finitely many function evaluations.



Definition: **Gaussian process**

A distribution on functions s.t. the evaluations $f(x_1), \ldots, f(x_n)$ at any $x_1, \ldots, x_n$ are (jointly) Gaussian.

Characterizing Gaussian Processes

Gaussian distribution		Gaussian process
1D	finite dimensional	
$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{N}(\mu, \Sigma)$	$\mathcal{GP}(\mu(x), k(x_1, x_2))$
mean	mean vector	mean function
μ	μ	$\mu(x)$
variance	covariance matrix	covariance function
σ^2	Σ	$k(x_1, x_2)$
higher moments/cumulants irrelevant/zero		

Set mean function to the constant zero function (normalize data).

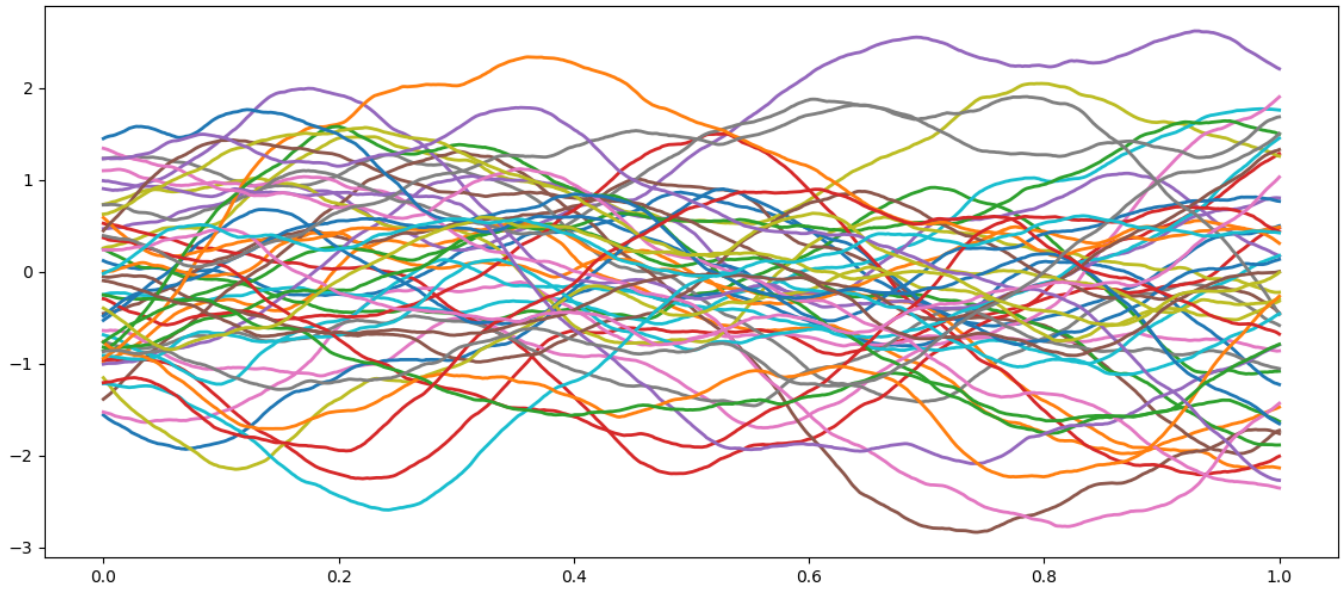
It remains to...

...encode information in the covariance function.

Plot GPs: sample a Gaussian distribution one dimension per pixel.

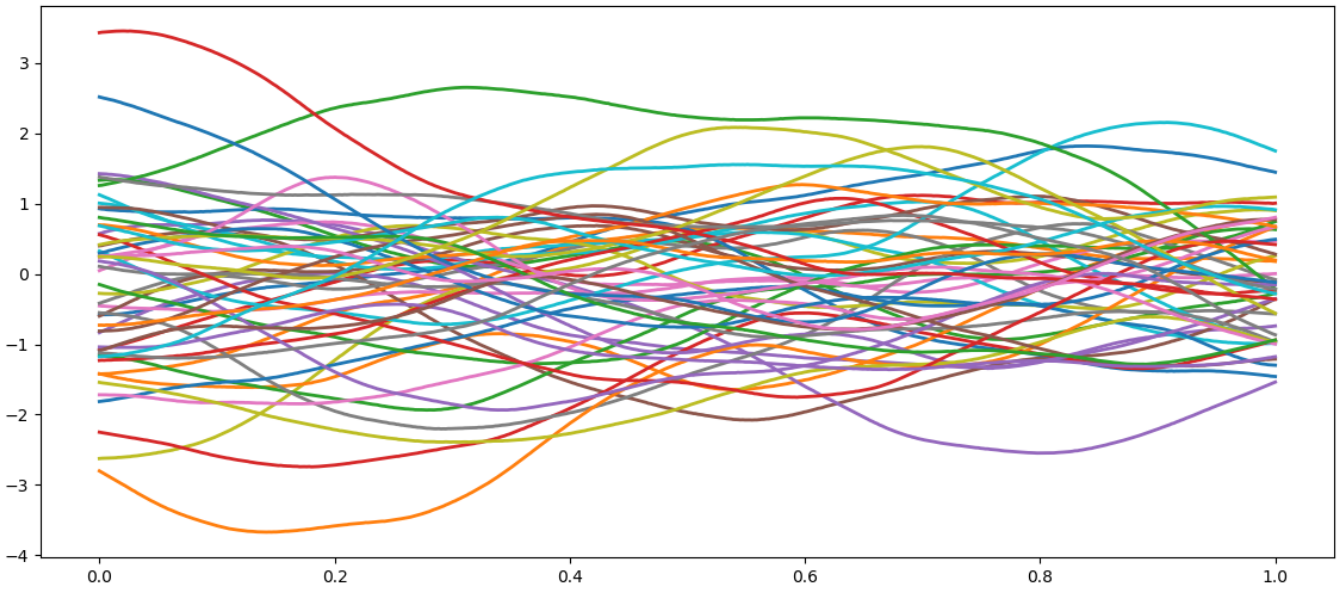
Covariance: Interdependence of Function Evaluations

C^1: continuously differentiable



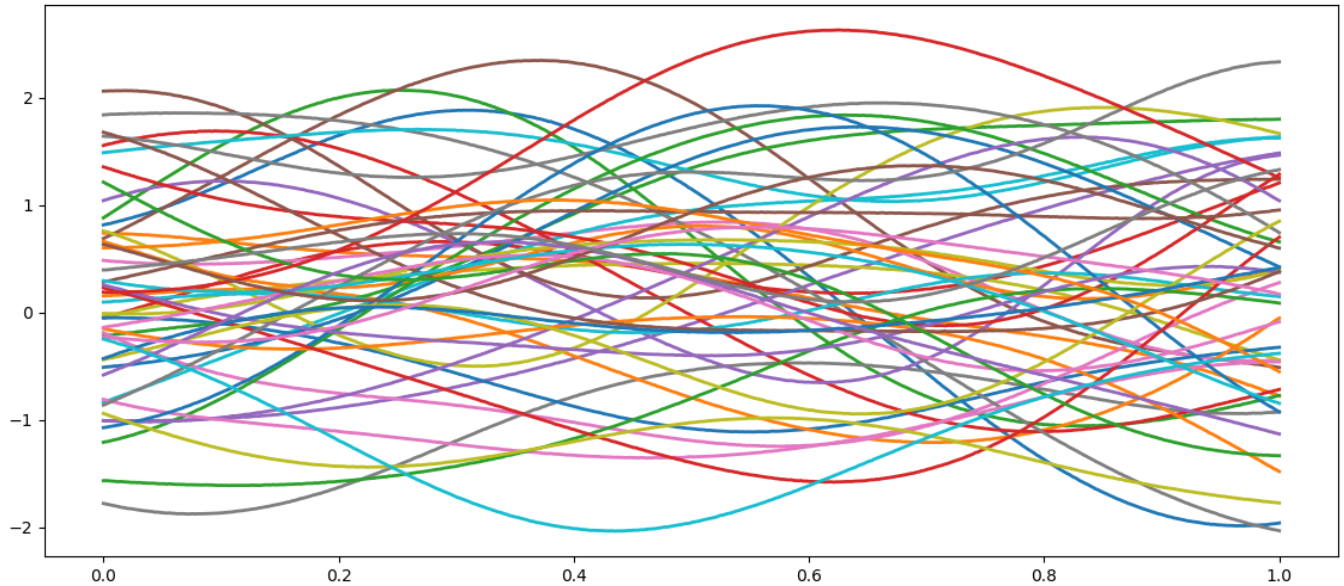
Covariance: Interdependence of Function Evaluations

C^2: twice continuously differentiable



Covariance: Interdependence of Function Evaluations

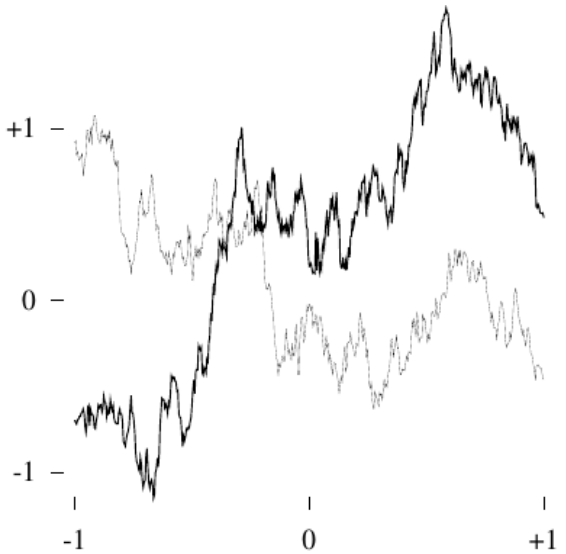
C^∞: smooth



Central limit theorem

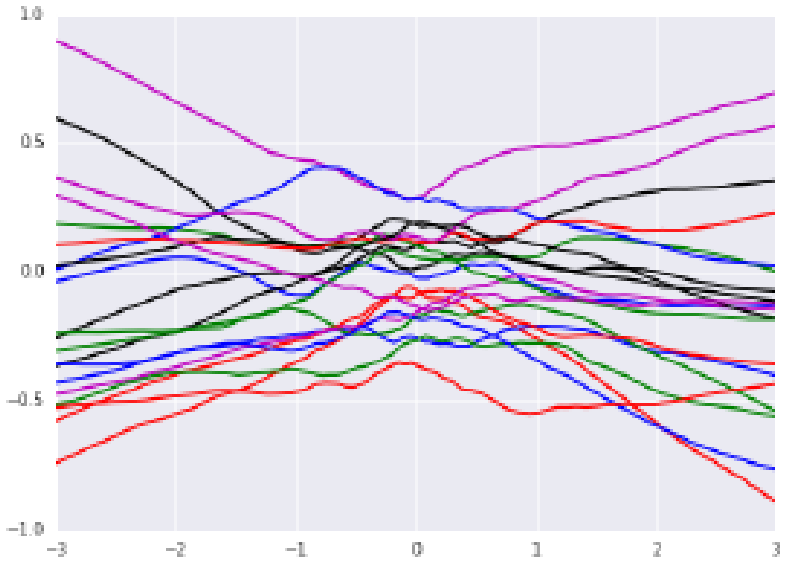
Neural nets with infinite width converge to Gaussian processes

iid parameters, controlling mean and variance



Step function, depth 1

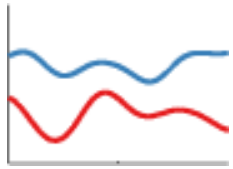
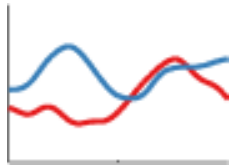
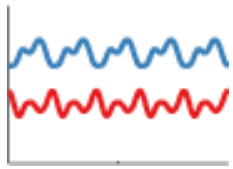
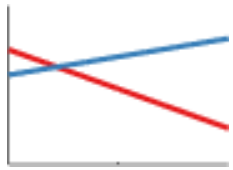
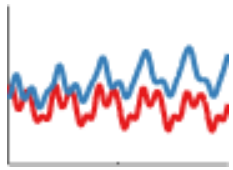
Neal, *Bayesian learning for neural networks*



ReLU, depth 10

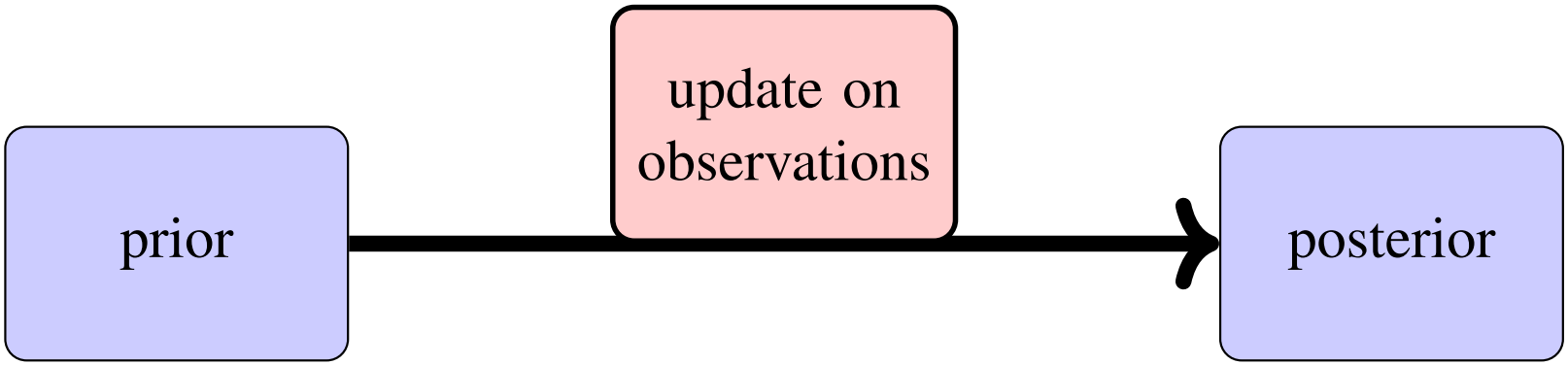
Lee et.al, *Deep Neural Networks as Gaussian Processes*

Kernel Cookbook

squared exponential	$\sigma^2 \exp\left(-\frac{1}{2}\frac{(x-x')^2}{\ell^2}\right)$	
rational quadratic	$\sigma^2\left(1+\frac{1}{2\alpha}\frac{(x-x')^2}{\ell^2}\right)^{-\alpha}$	
periodic	$\sigma^2 \exp\left(-2\frac{\sin^2(\frac{\pi}{p}\lvert x-x'\rvert)}{\ell^2}\right)$	
linear	$a^2 + b^2 x x'$	
local periodic	$\sigma^2 \exp\left(-2\frac{\sin^2(\frac{\pi}{p}\lvert x-x'\rvert)}{\ell^2} - \frac{1}{2}\frac{(x-x')^2}{\ell^2}\right)$	

David Duvenaud, *Kernel Cookbook*, `http://www.cs.toronto.edu/~duvenaud/cookbook/`

Bayesian approach



- Prior: domain knowledge or uninformative
- Condition on data
- Goal: posterior

Good prior: stable and decent models for few data points.

(Due to their simplicity: GPs are the standard functional prior in Bayesian ML&Stats.)

Gaussian process regression: math

Reminder: Gaussian process $g = \mathcal{GP}(\mu, k)$

A distribution on $\mathbb{R}^d \to \mathbb{R}^\ell$ s.t. $g(x_1), \ldots, g(x_n)$ are Gaussian.
Data structure: $\mu : \mathbb{R}^d \to \mathbb{R}^\ell$ and $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{\ell\times\ell}_{\geq 0}$.

Regression model

Assume $\mu = 0$. Condition on $\{(x_i, y_i) \in \mathbb{R}^{1\times(d+\ell)} \mid i = 1, \ldots, n\}$.

$$\mathcal{GP}\Big(\quad x \mapsto \mathbf{y}\mathbf{k}(\mathbf{X},\mathbf{X})^{-1}\mathbf{k}(\mathbf{X},\mathbf{x}),$$
$$(x, x') \mapsto \mathbf{k}(\mathbf{x},\mathbf{x}') - \mathbf{k}(\mathbf{x},\mathbf{X})\mathbf{k}(\mathbf{X},\mathbf{X})^{-1}\mathbf{k}(\mathbf{X},\mathbf{x}')\Big)\,.$$

$$k(X,X) = \begin{bmatrix} k(x_1,x_1) & \ldots \\ \vdots & \ddots \end{bmatrix} \in \mathbb{R}^{\ell n\times\ell n}_{\geq 0},$$

$k(x, X) = \begin{bmatrix} k(x,x_1) & \ldots \end{bmatrix} \in \mathbb{R}^{\ell\times\ell n}$, and $y = \begin{bmatrix} y_1 & \ldots \end{bmatrix} \in \mathbb{R}^{1\times\ell n}$.

Gaussian process regression: classical algorithm

Inputs: $X \in \mathbb{R}^{n \times d}$ (inputs), $y \in \mathbb{R}^{n \times \ell}$ (outputs), covariance function k

Output: posterior GP

1 $L := \text{Cholesky}(k(X, X))$ (hence, $k(X, X) = L^T L$, precompute, $\mathcal{O}(n^3)$)

2 $\alpha := y/L/L^T$ (precompute, $\mathcal{O}(n^2)$)

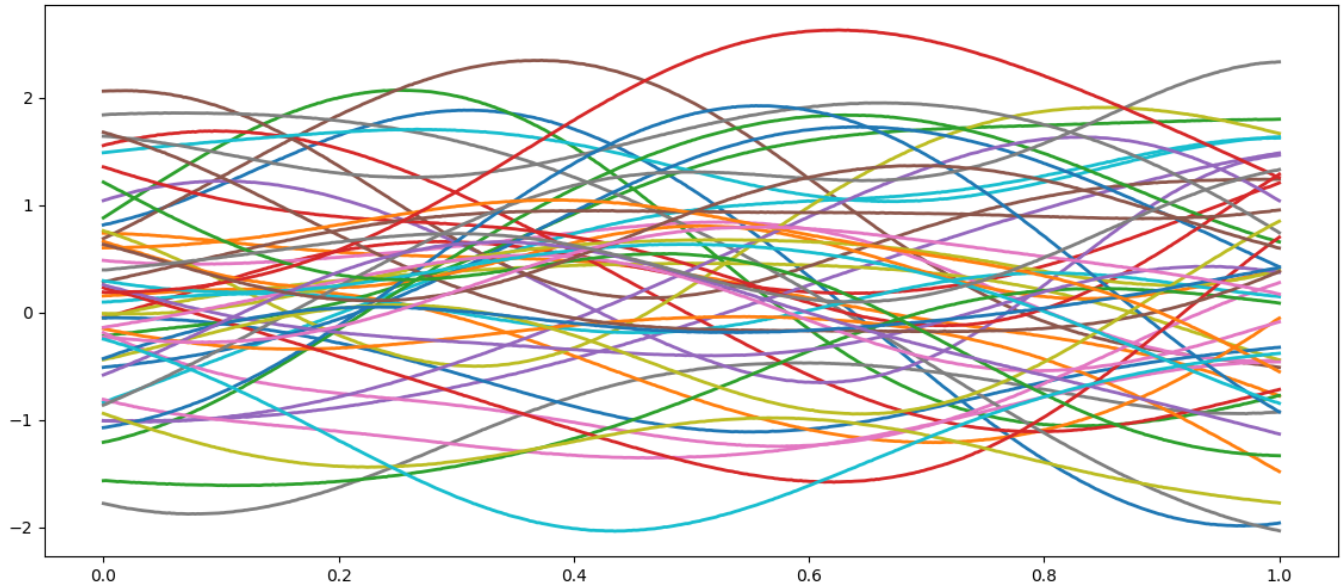
3 return

$$\mathcal{GP}(\quad x \mapsto \alpha \cdot k(X, x),$$
$$(x, x') \mapsto k(x, x') - (k(x, X)/L^T) \cdot (L \backslash k(x', X)))$$

$(\mathcal{O}(n)), \mathcal{O}(n^2)))$

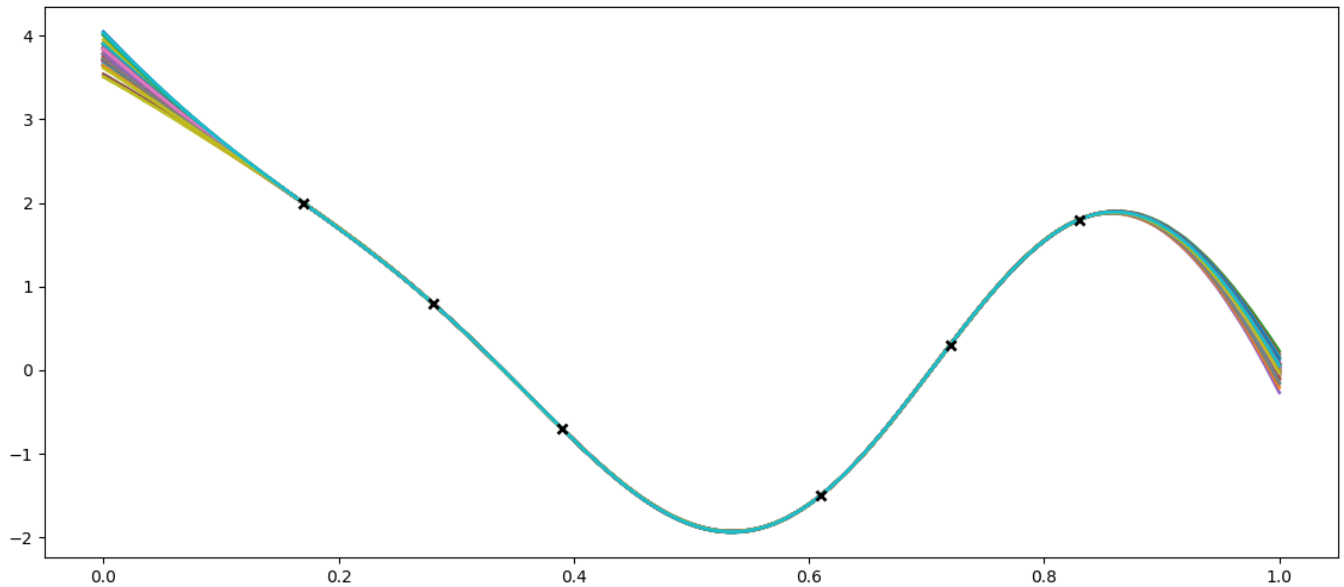
Include Measurement Data

C^∞: smooth



Include Measurement Data

C^∞, conditioned on data



Gaussian process regression: noise

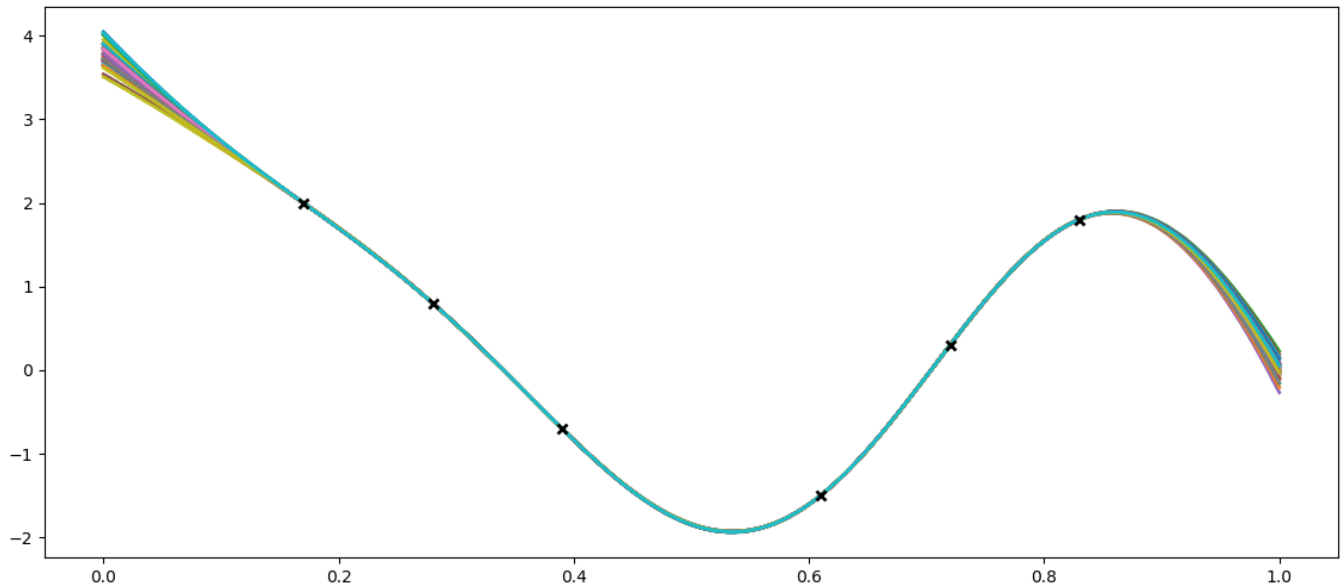
- Take a maximum entropy prior on the behavior unexplained by g:
  Add Gaussian white **noise** ε (works well enough if noise is not strictly Gaussian).
- Replace covariance $k(X, X)$ by $k(X, X) + \mathrm{var}(\varepsilon) I_{\ell n}$.
  (more variance, no new correlations)
- Posterior:

$$\mathcal{GP}\Big(\quad x \mapsto y(k(X,X) + \mathrm{var}(\varepsilon) I_{\ell n})^{-1} k(x,X)^T,$$
$$(x, x') \mapsto k(x, x') - k(x,X)(k(X,X) + \mathrm{var}(\varepsilon) I_{\ell n})^{-1} k(x', X)^T \Big)$$

- Noise makes computations **numerically stable** decreases condition number.
- Possible: set noise individual for data points or data channels.

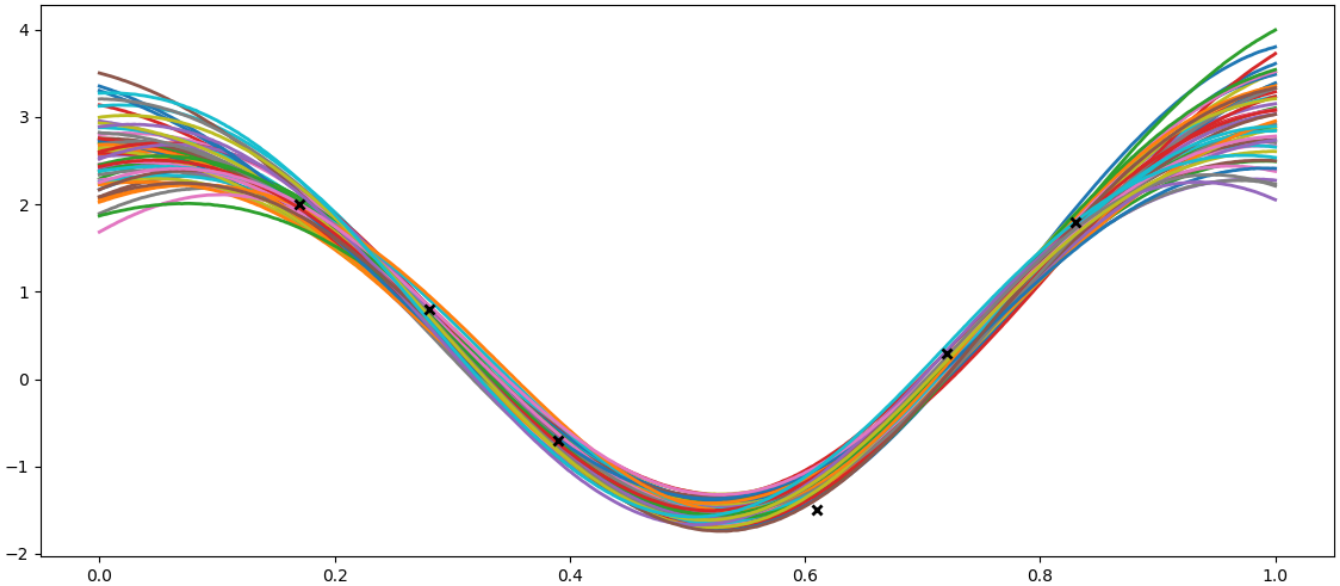
Include Measurement Data

C^∞, conditioned on data



Include Measurement Data

C^∞, conditioned on noisy data



Gaussian process regression: hyperparameters

- Hyperparameters in the priors:
  - length scales ℓ
  - signal variance σ
  - noise ε
  - period p
  - etc.
- **Optimal hyperparameters**: optimize the (log-)likelihood.

$$\log p(y|X) = \underbrace{-\frac{1}{2} y^T K^{-1} y}_{\text{data fit}} \underbrace{-\frac{1}{2}\log(\det(K))}_{\text{model complexity}} - \frac{n}{2}\log 2\pi$$

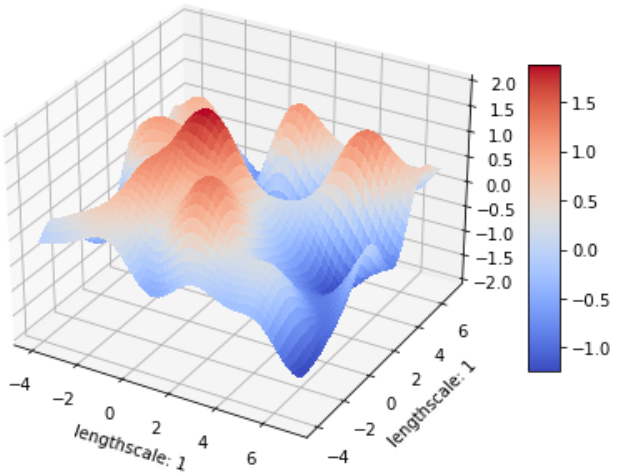
Computable via linear Algebra (including gradients)

Optimization instead of integration: potential overfitting? Part of GAIA (tandem in Dataninja) to reduce this problem.

- Hyperparameters in GPs are interpretable and learnable
  E.g. learn a period in your data.

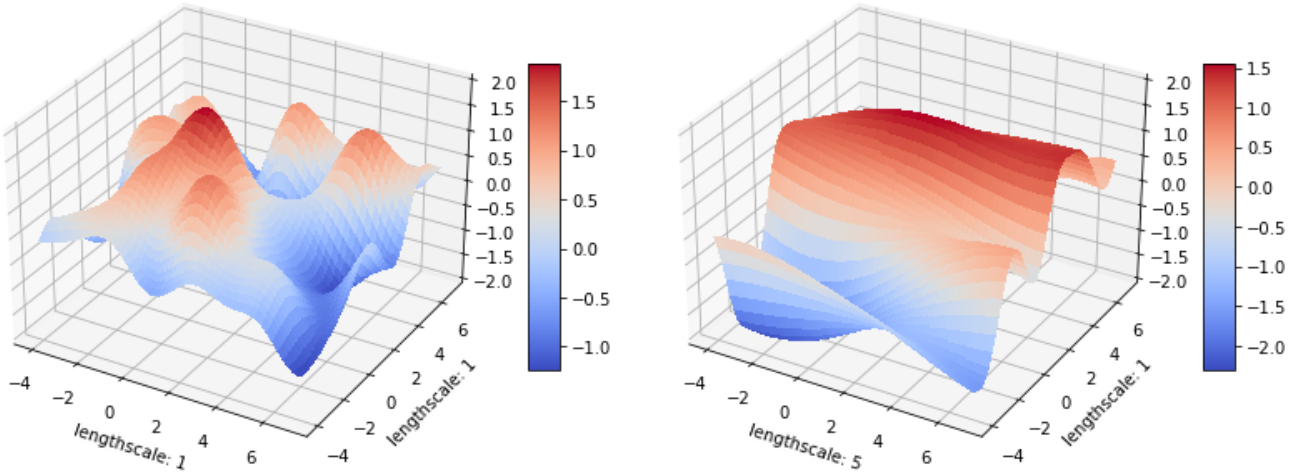
Automatic relevance determination

Hyperparameters in GPs are interpretable and learnable
E.g. lengthscales learn which inputs are relevant.



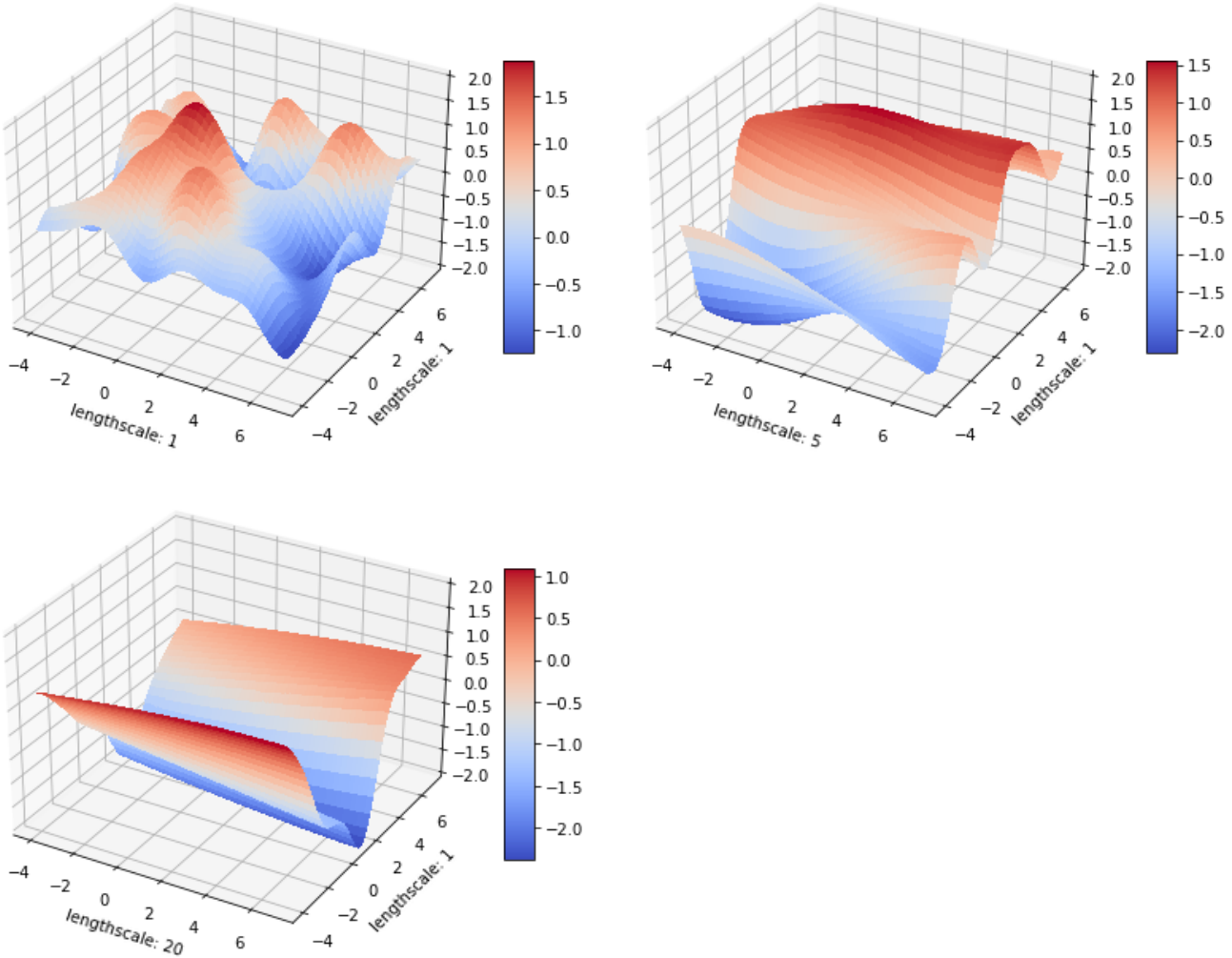
Automatic relevance determination

Hyperparameters in GPs are interpretable and learnable
E.g. lengthscales learn which inputs are relevant.



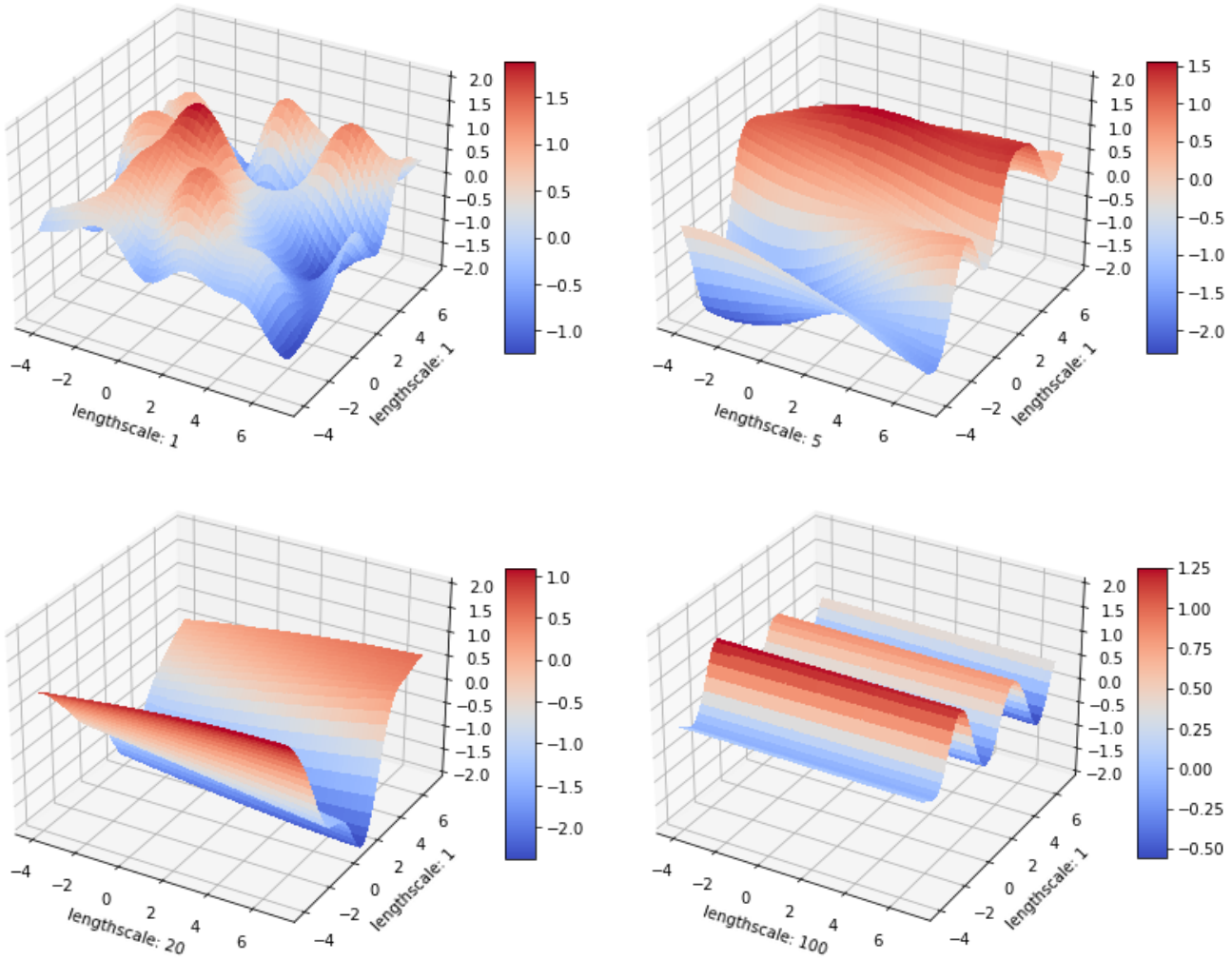
Automatic relevance determination

Hyperparameters in GPs are interpretable and learnable
E.g. lengthscales learn which inputs are relevant.

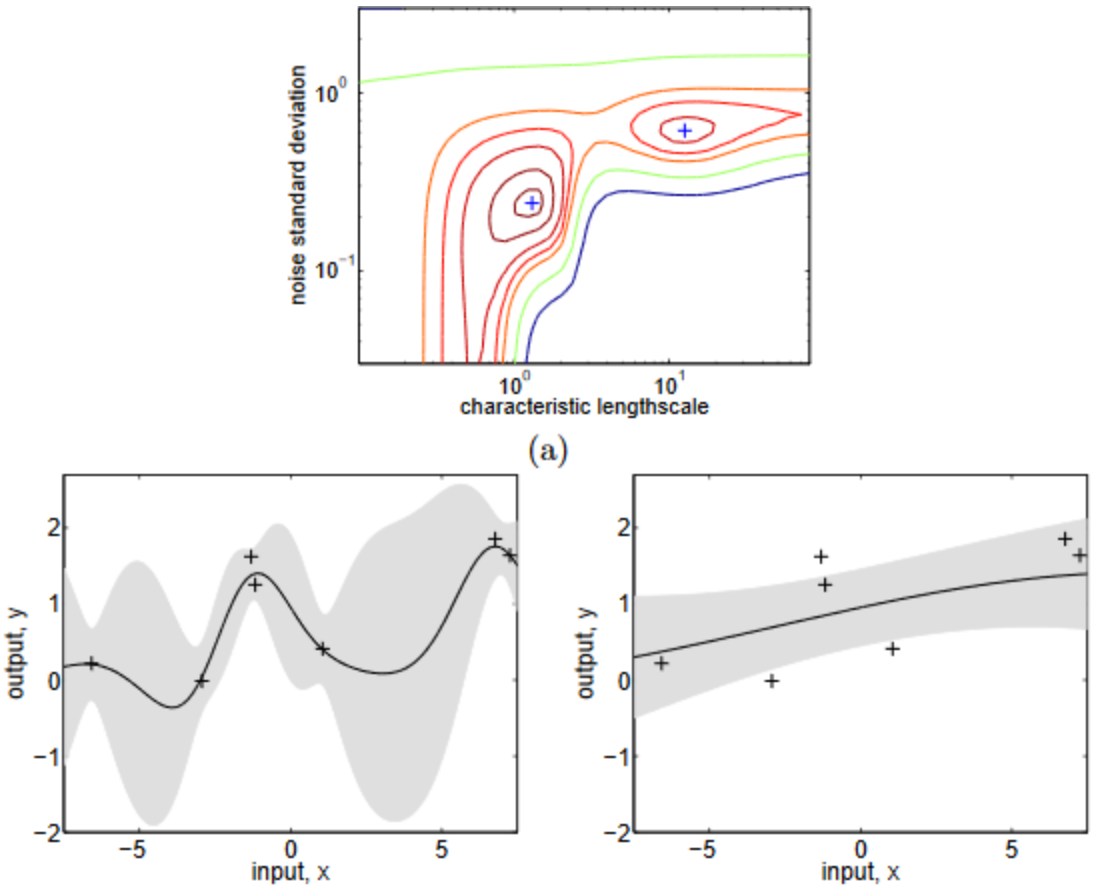


Automatic relevance determination

Hyperparameters in GPs are interpretable and learnable
E.g. lengthscales learn which inputs are relevant.



Hyperparameter estimation: multiple explanations of data



Rasmussen/Williams, Gaussian processes for Machine Learning

Typical GP Workflow

1. Get data
2. Construct a covariance function, including hyperparameters
3. Get best hyperparameters by optimizing the log-likelihood
4. Carefully inspect the predictions and hyperparameter values
5. If inspection is not good, go to step 1 or 2.
6. ???
7. Profit

Default setting:

- squared exponential covariance $k(t, t') = \exp\left(\frac{1}{2}(t - t')^2\right)$
- ARD length scales
- unified noise hyperparameter
- zero mean

Automating this workflow is part of GAIA.

Typical Problems

- Cholesky fails: the covariance matrix is not positive definite
  - Matrix not symmetric or eigenvalues $\ll 0$: no covariance function
  - Minor negative eigenvalues: numerical problems
    - (Iteratively) add more noise
    - Use float64
- Predictions are bad
  - Data is bad (duh)
  - Data does not fit the prior
  - Look at the trained hyperparameters. Are they reasonable?
  - Dictionaries below interpret what your model is doing wrong.
- Everything is slow
  - Remember $\mathcal{O}(n^3)$
  - (L-)BFGS>SGD
  - Take an approximation to GPs which make SGD applicable

    start with the 2009 AISTATS paper from Titsias and continue with the 2013 UAI paper from Hensman et. al.

Message for interpretability

Interpretability of basic GPs

- Construct suitable covariance functions for each application
- Learn interpretable parameters: noise, relevance of inputs, periods etc. (for physical constants see below)
- One model class can give several interpretations of the data
- Interpretable parameters can be changed or set manually

Questions?

Questions?

Use case: GDI timing

Calibration of gasoline engines for particulate reduction

http://www.autonews.com/article/20051031/SUB/51102026/

Optimization goal

Find optimal injection strategy

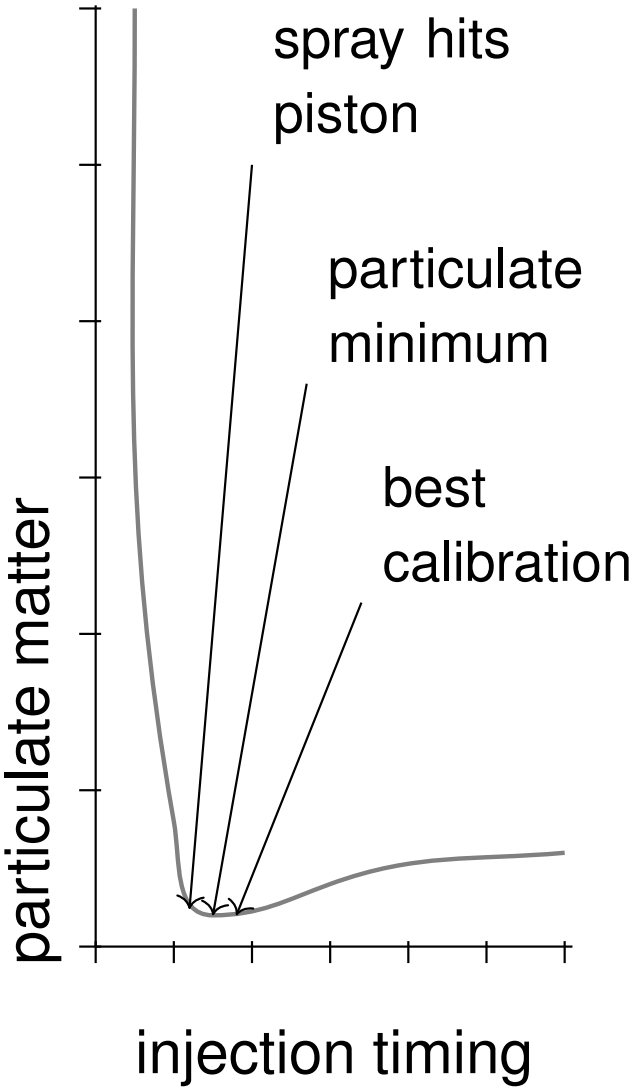
- (rail) pressure
- timing
- amount

for 1-5 injection such that particulate matter is minimal.

Challenge

- Inject early to allow mixing, without hitting the piston.
- If piston is hit, sudden incline of particulate matter.

Particulate matter measurements are not Gaussian



Particulate matter measurements are not Gaussian

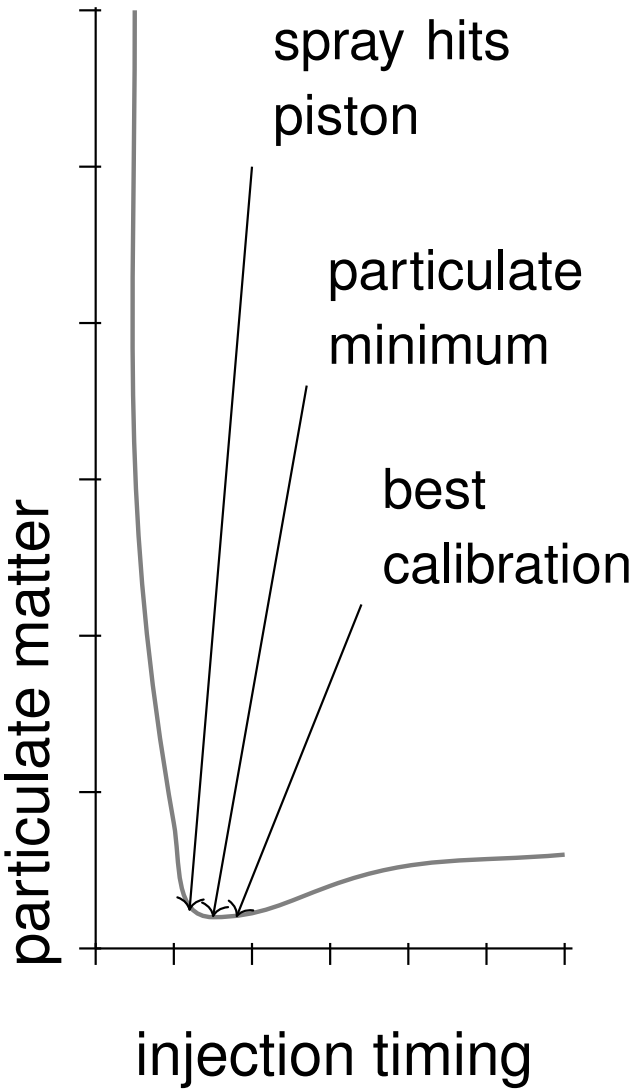
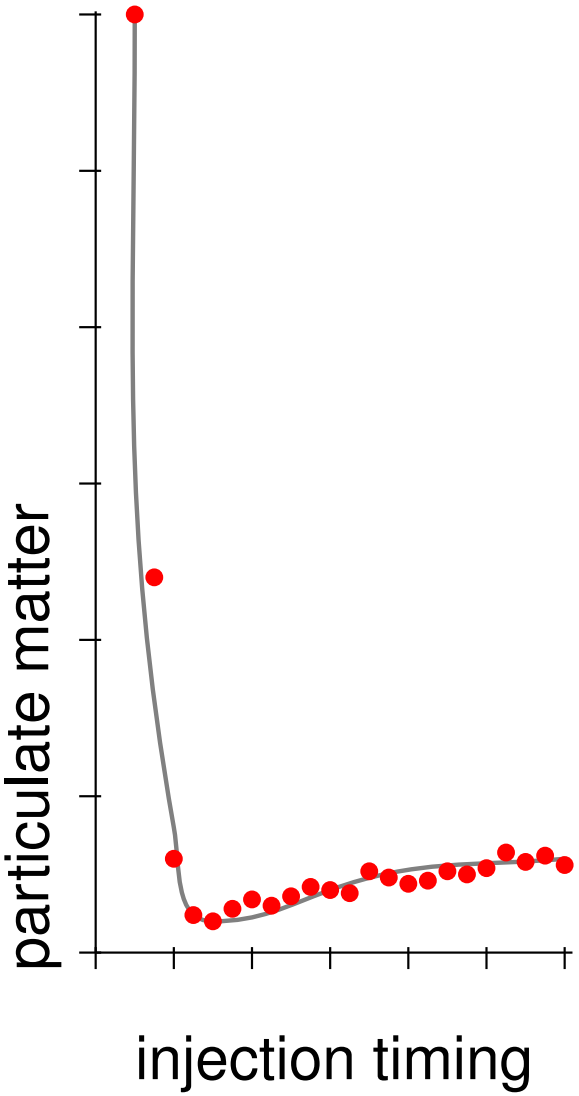


Diagram depends on

- speed
- load
- air pressure
- intake valve timing
- exhaust valve timing
- rail pressure
- (ignition timing)
- (timings of later injections)
- (quantities in later injections)
- etc.

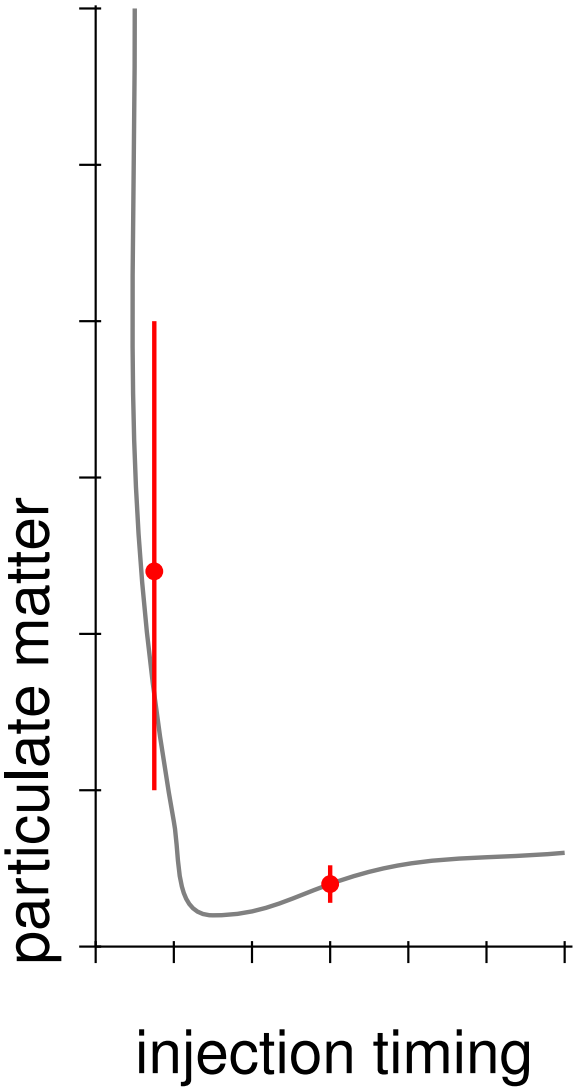
Particulate matter measurements are not Gaussian



Violated default assumptions

- Gaussian distribution of data

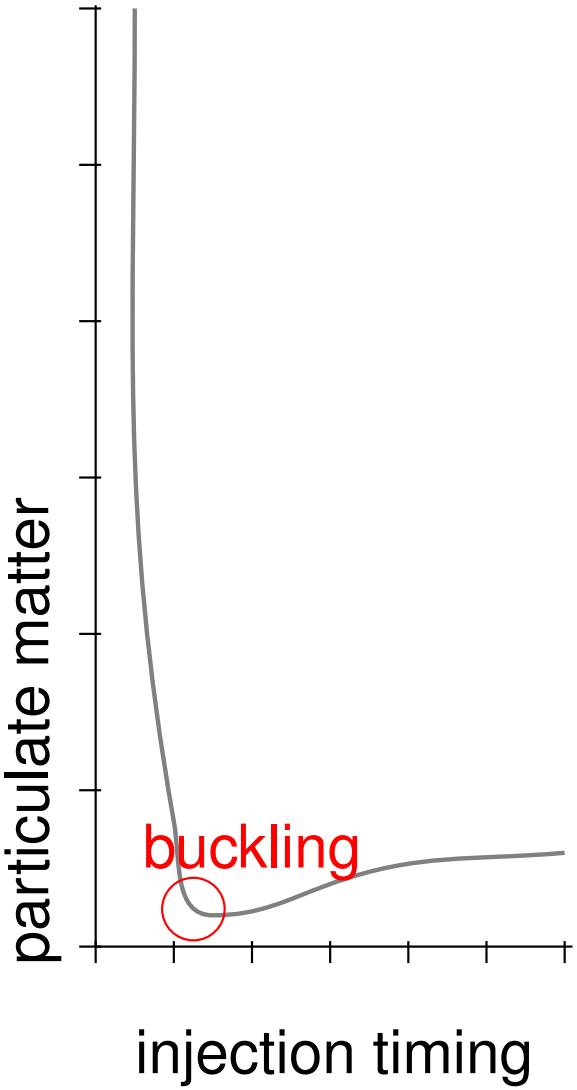
Particulate matter measurements are not Gaussian



Violated default assumptions

- Gaussian distribution of data
- Constant noise

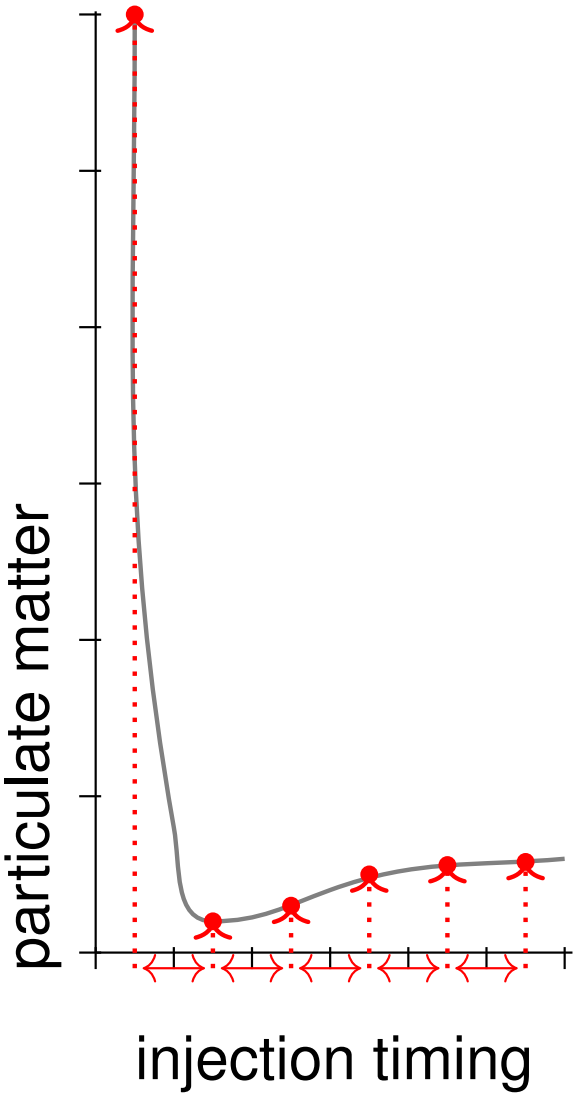
Particulate matter measurements are not Gaussian



Violated default assumptions

- Gaussian distribution of data
- Constant noise
- Smooth behavior

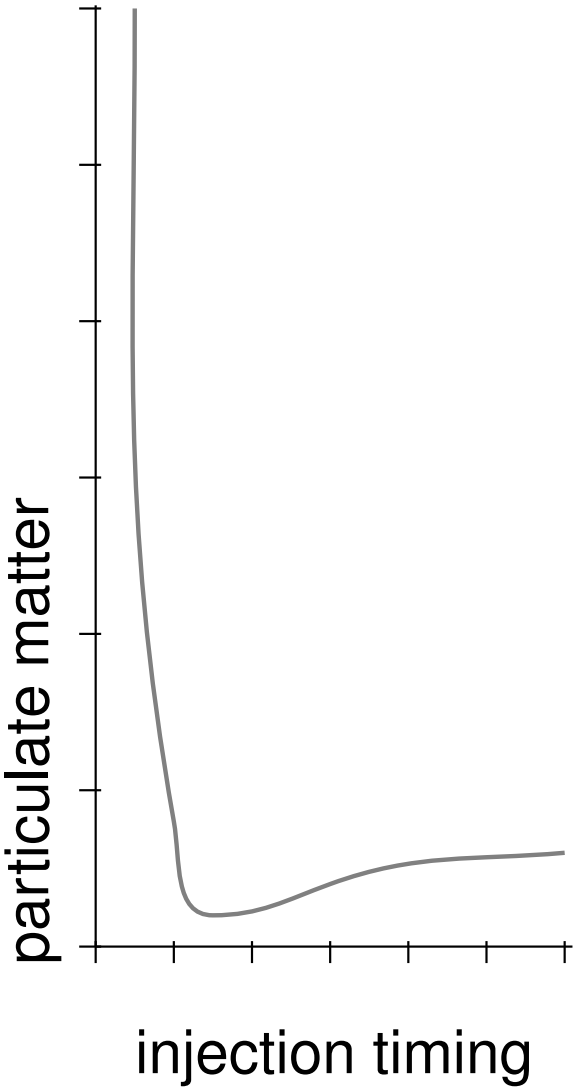
Particulate matter measurements are not Gaussian



Violated default assumptions

- Gaussian distribution of data
- Constant noise
- Smooth behavior
- Covariance function stationary

Particulate matter measurements are not Gaussian



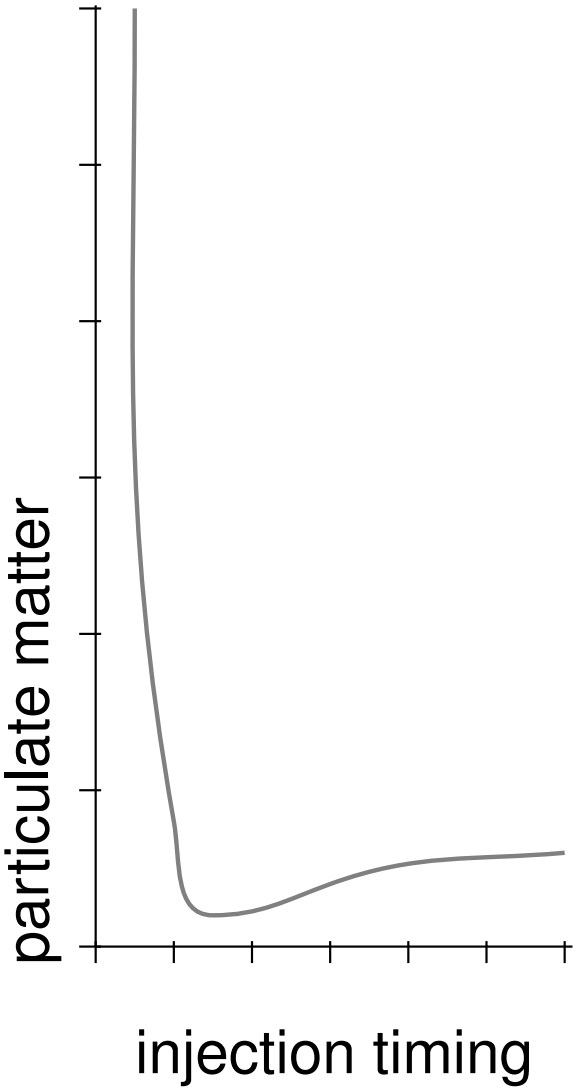
Violated default assumptions

- Gaussian distribution of data
- Constant noise
- Smooth behavior
- Covariance function stationary

Approach

- Output transformation

Particulate matter measurements are not Gaussian



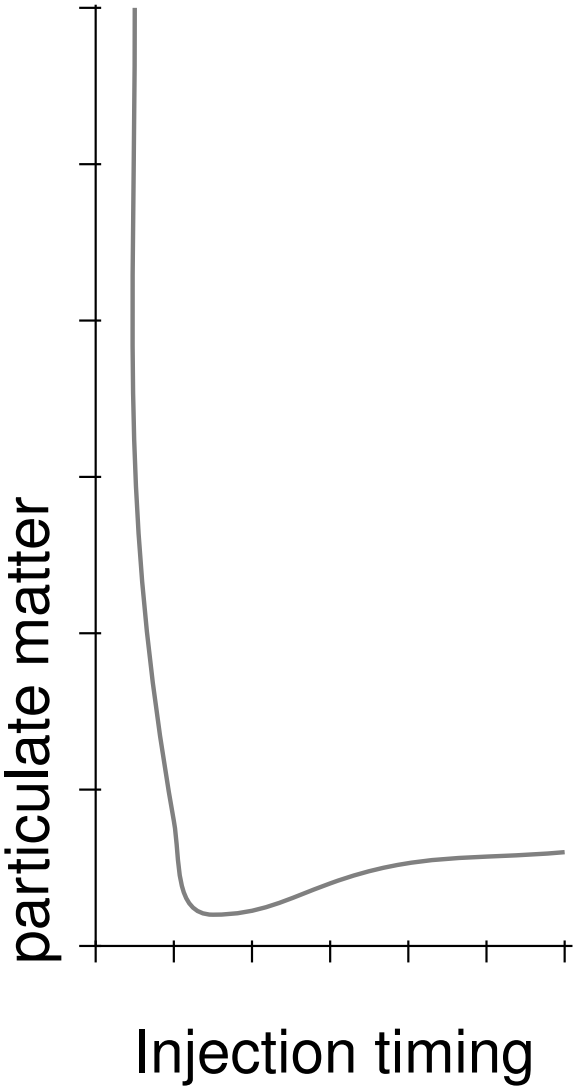
Violated default assumptions

- Gaussian distribution of data
- Constant noise
- Smooth behavior
- Covariance function stationary

Approach

- Output transformation
- Input transformation

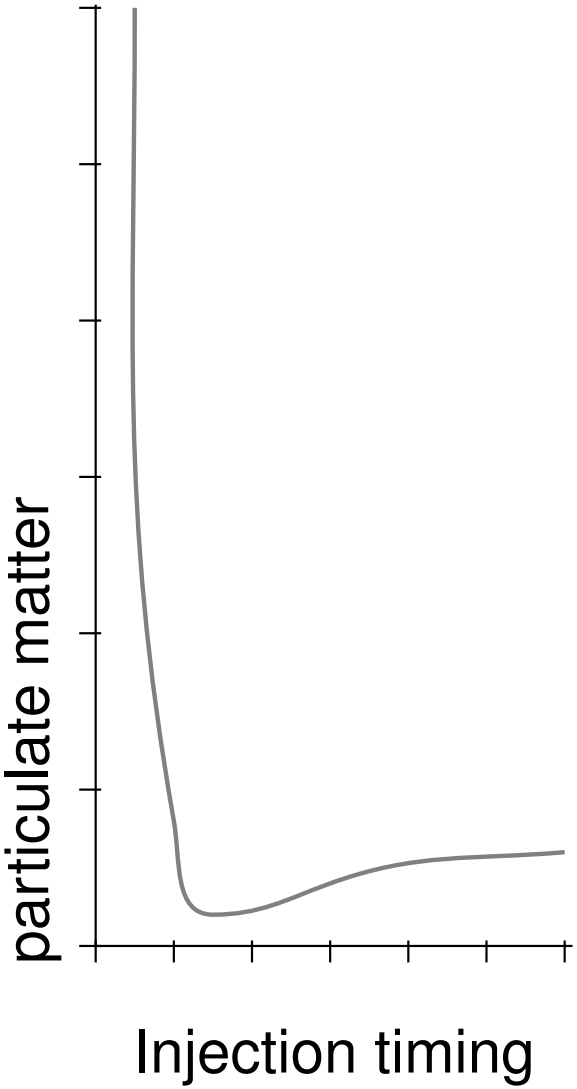
Warped and manifold Gaussian processes



Warped Gaussian processes

- transformation of the output space
- goal: make data more Gaussian
- goal: make noise more even

Warped and manifold Gaussian processes



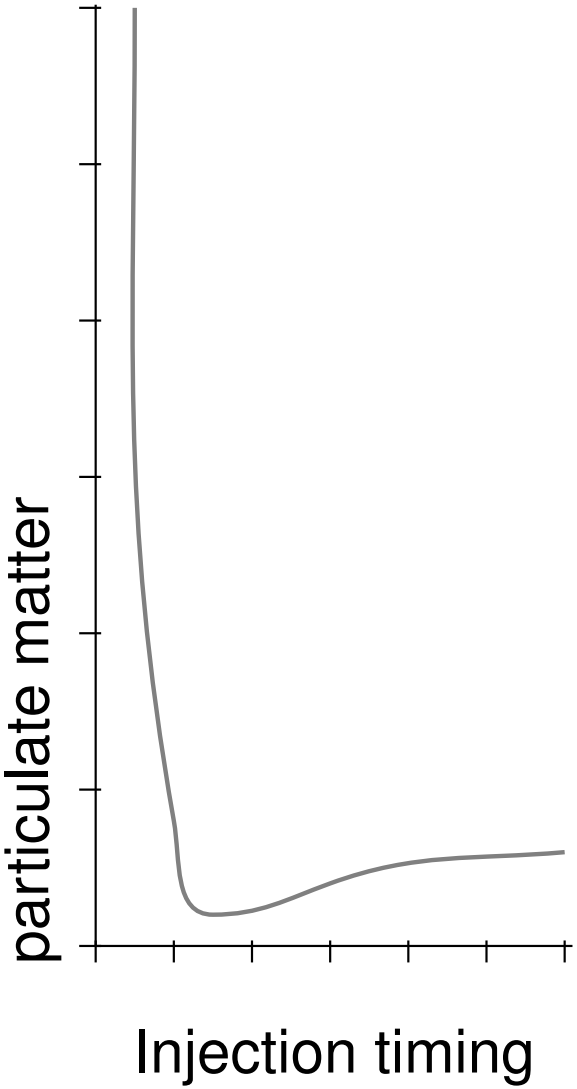
Warped Gaussian processes

- transformation of the output space
- goal: make data more Gaussian
- goal: make noise more even

Manifold Gaussian processes

- transformation of the input space
- goal: model unsmooth behavior
- goal: allow non-stationary behavior

Warped and manifold Gaussian processes



Warped Gaussian processes

- transformation of the output space
- goal: make data more Gaussian
- goal: make noise more even

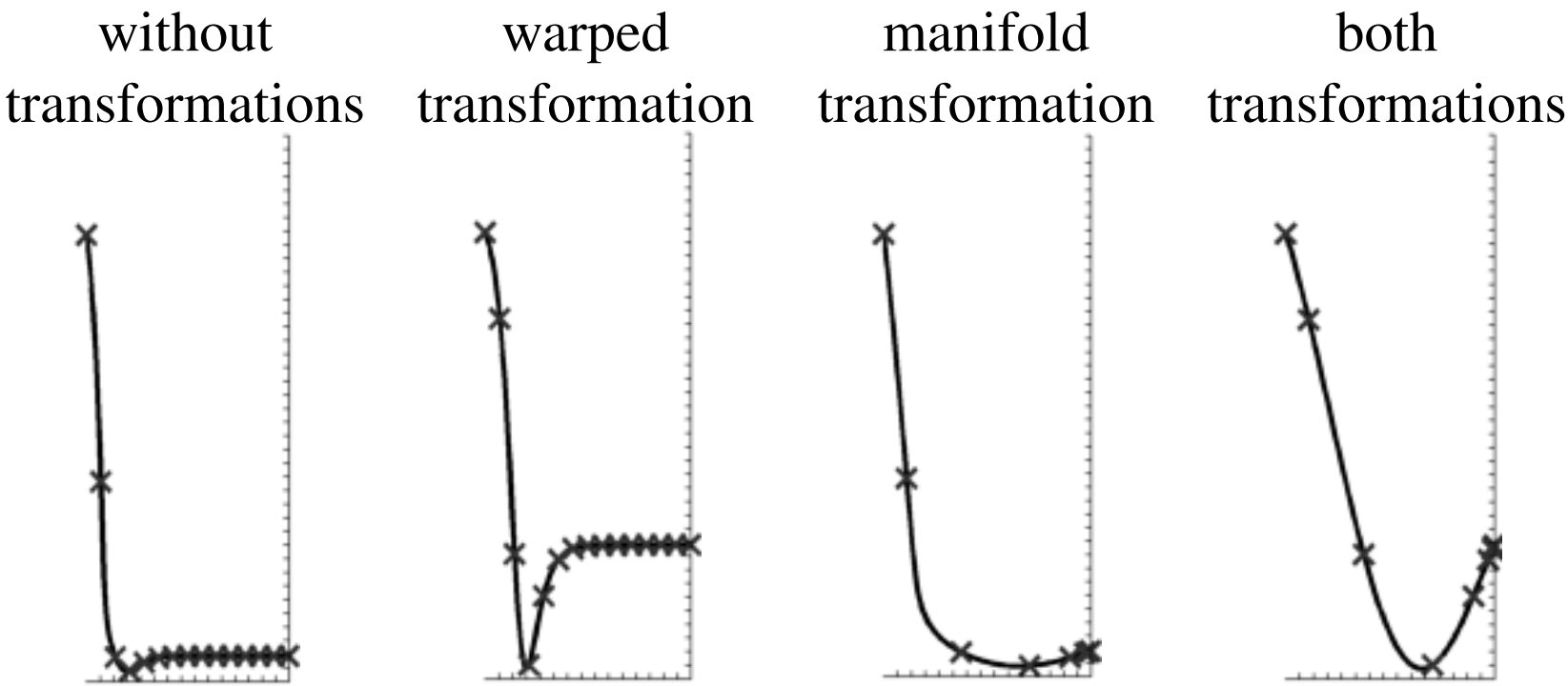
Manifold Gaussian processes

- transformation of the input space
- goal: model unsmooth behavior
- goal: allow non-stationary behavior

Warped and manifold Gaussian processes

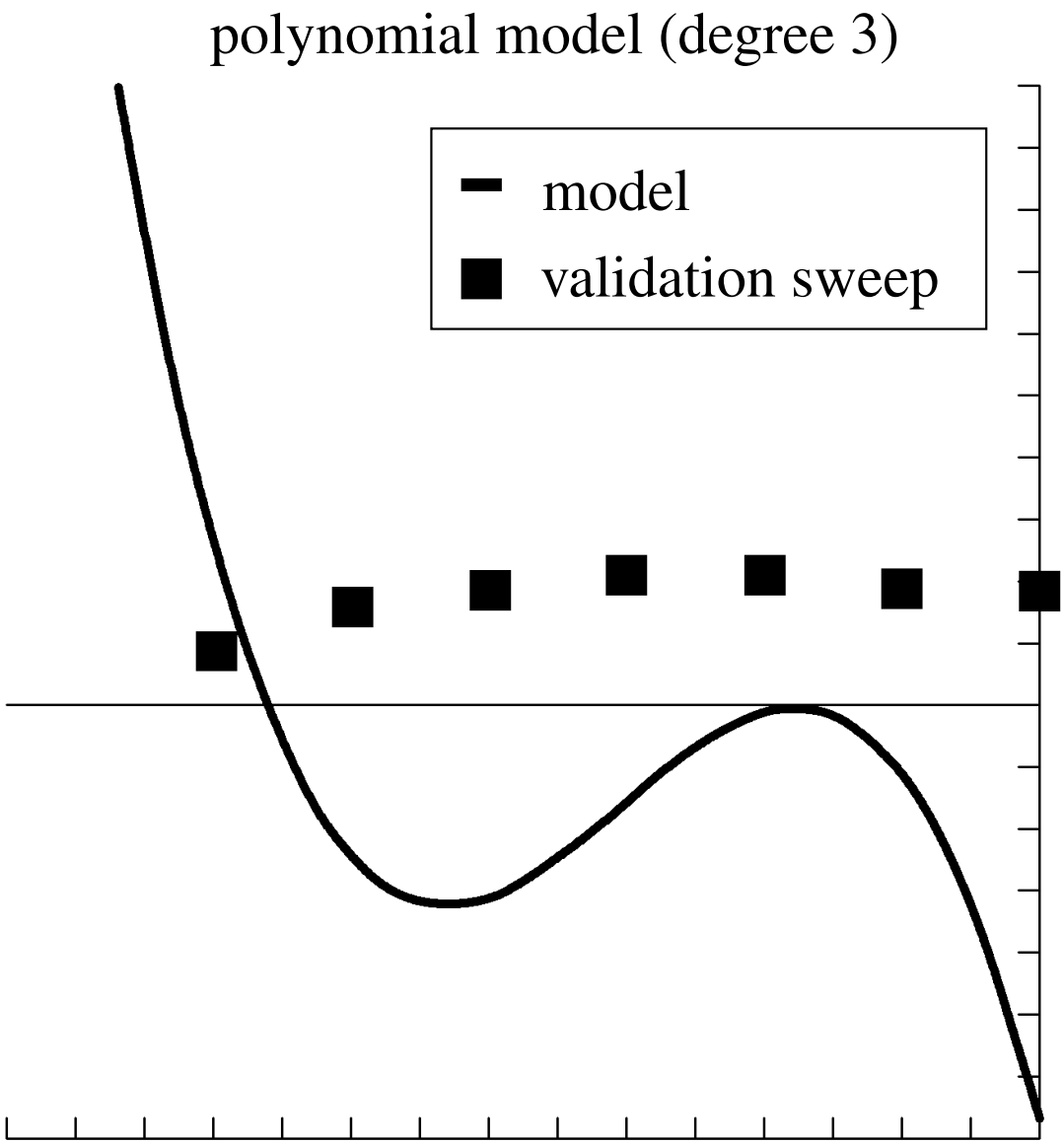
- transformation and covariance parameter chosen jointly

Effect of warped and manifold Gaussian processes



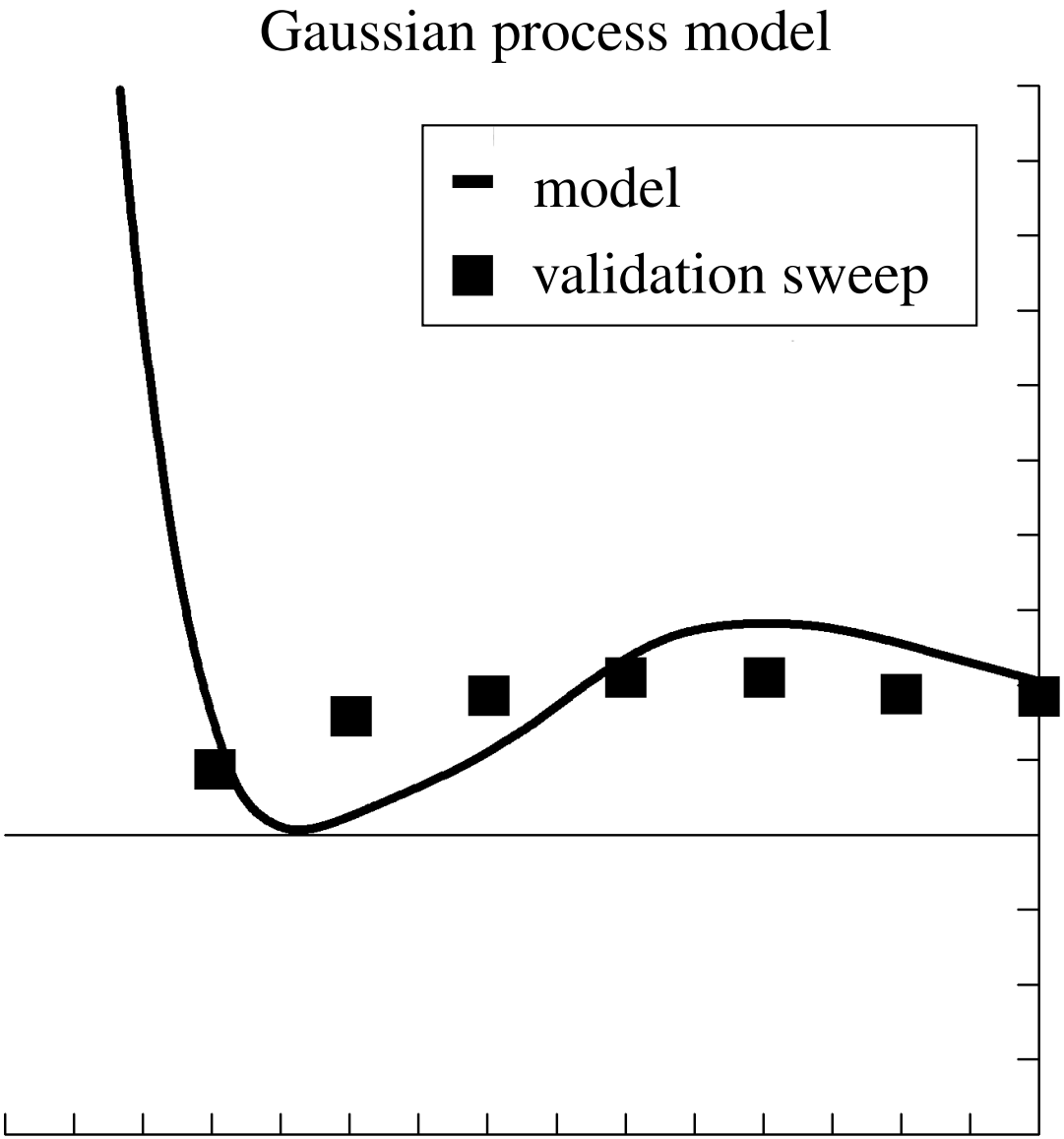
Thewes, Lange-Hegermann, Reuber, Beck, *Erweiterte Modellierungstechniken für Gaussche Prozessmodelle*

Model comparison to validation sweep



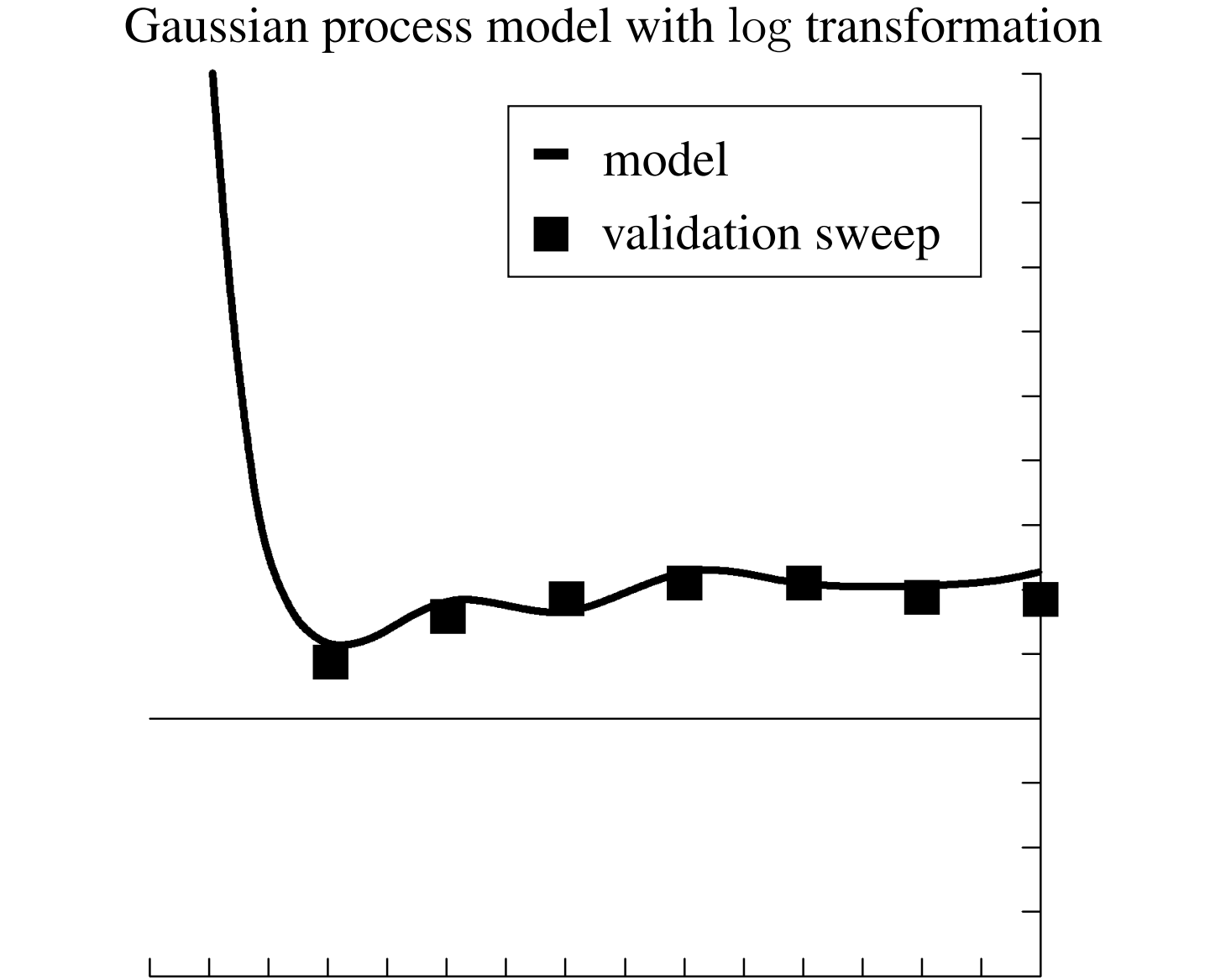
Thewes, Lange-Hegermann, Reuber, Beck, *Erweiterte Modellierungstechniken für Gaussche Prozessmodelle*

Model comparison to validation sweep



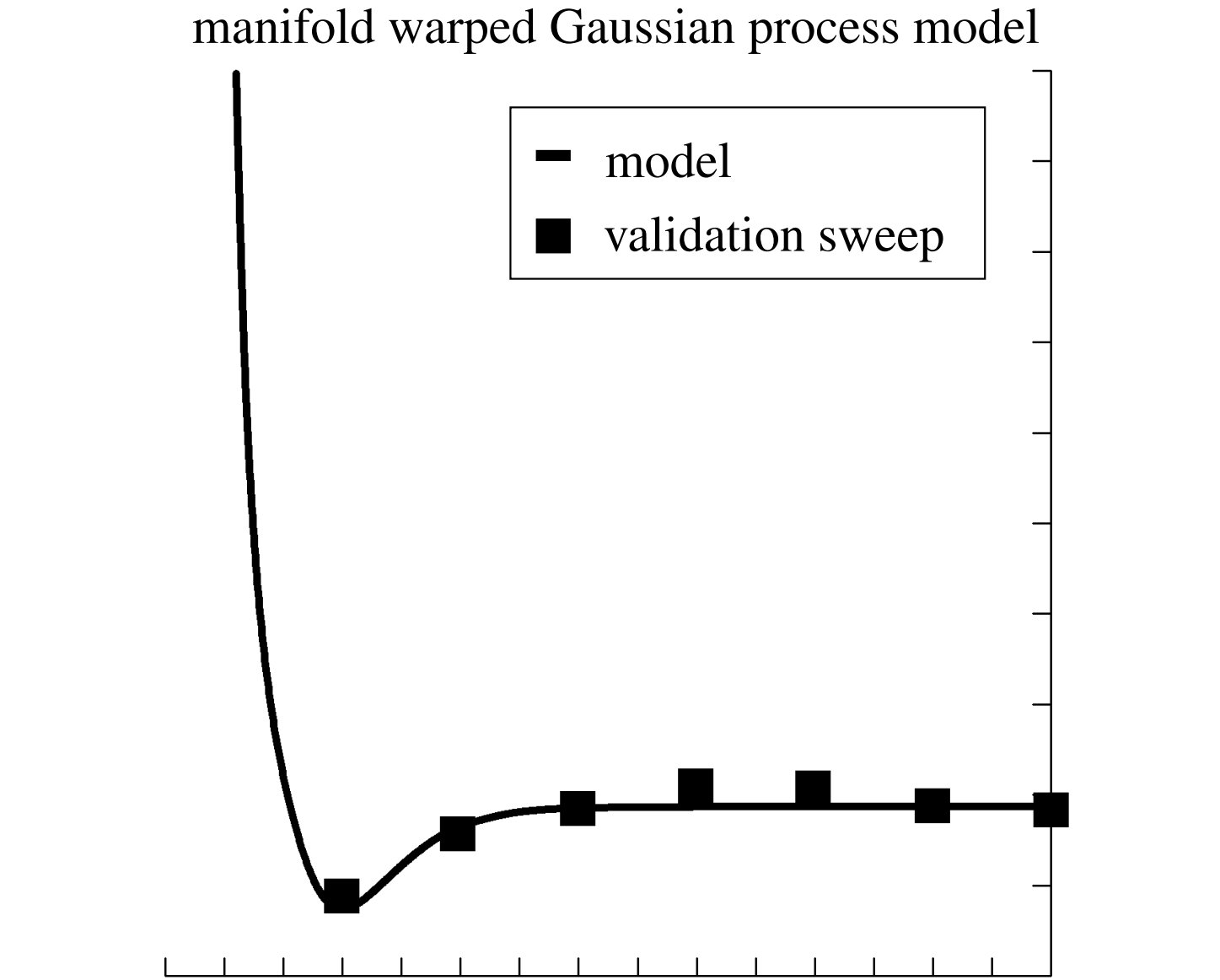
Thewes, Lange-Hegermann, Reuber, Beck, *Erweiterte Modellierungstechniken für Gaussche Prozessmodelle*

Model comparison to validation sweep



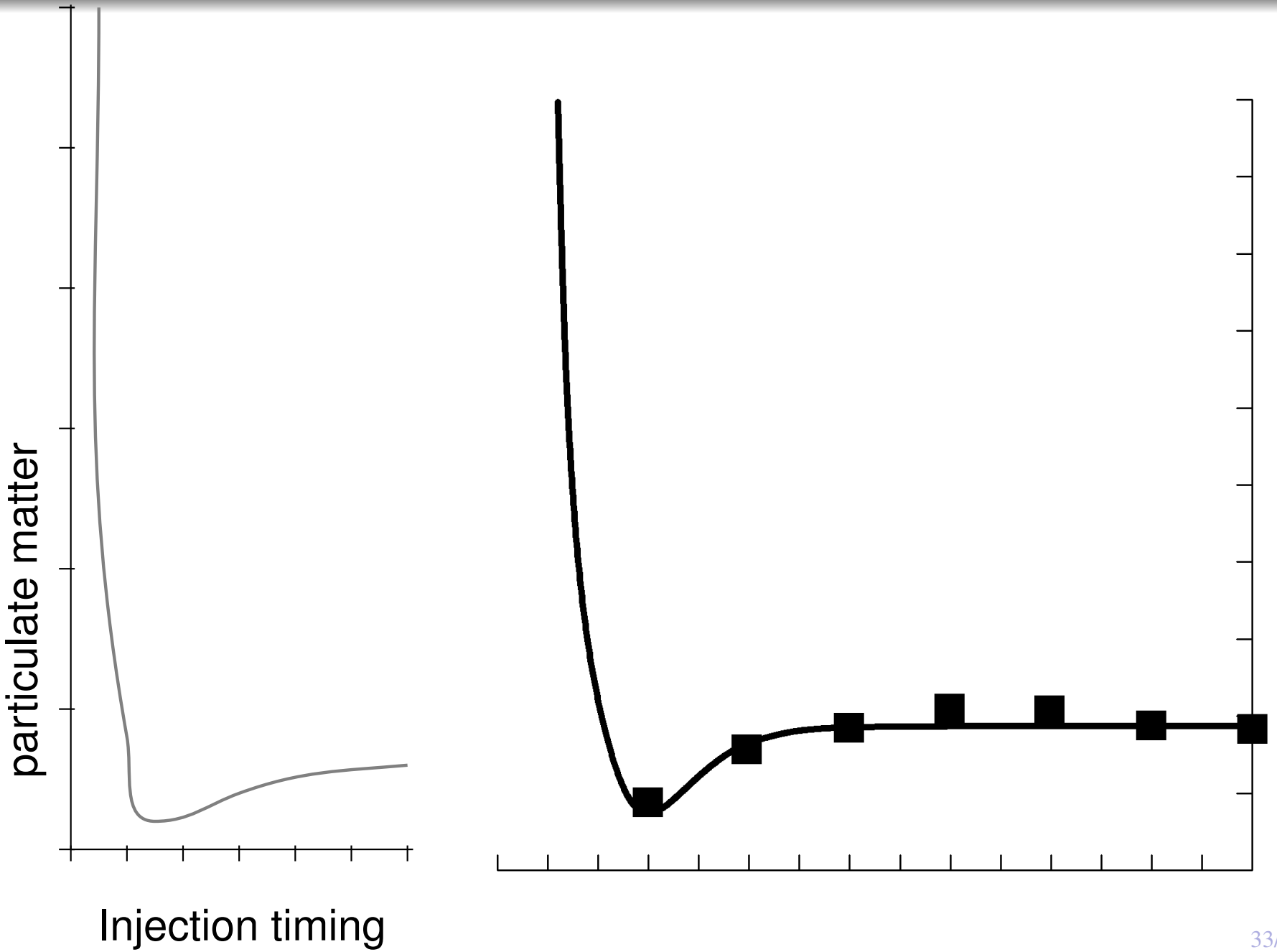
Thewes, Lange-Hegermann, Reuber, Beck, *Erweiterte Modellierungstechniken für Gaussche Prozessmodelle*

Model comparison to validation sweep



Thewes, Lange-Hegermann, Reuber, Beck, *Erweiterte Modellierungstechniken für Gaussche Prozessmodelle*

Model comparison to initial drawing



Message for interpretability

Interpretability of transformations

- Build in specific expert knowledge via transformations
- Automatic learning of (parameters of) transformations

Questions?

Questions?

Understanding and Manipulating the Prior

Reproducing Kernel Hilbert Spaces (RKHS)

A Hilbert space of functions s.t. their evaluations are continuous.

- Continuity (or even differentiability) of the model evaluation is typically required for model training.
- Hence, most ML-models can be described by an RKHS.
  This is very obvious for kernel methods (GPs, SVMs, ...).
- RKHS are an important tool in understanding neural networks.

  (via neural tangent kernel (NTK) or infinite width limit)

Let $g = \mathcal{GP}(0, k)$.

The $x \mapsto k(x_i, x)$ for $x_i \in \mathbb{R}^d$ generate the pre-Hilbert space $\mathcal{H}^0(g)$, which we endow with scalar product $\langle k(x_i, -), k(x_j, -) \rangle := k(x_i, x_j)$. Its closure w.r.t. $\langle \cdot, \cdot \rangle$ $\mathcal{H}(g)$ is the RKHS of g.

Theorem

Any RKHS is of this form, i.e. has a so-called **reproducing kernel** k.
In particular, there is a 1-1-correspondence: kernels $\leftrightarrow$ RKHS.

RKHS encode the covariance functions. They allow to interpret GPs.

$\mathcal{H}^0(g)$ is the space of posterior mean functions.

In many settings, the RKHS $\mathcal{H}(g)$ is the Cameron-Martin Space of the Gaussian measure induced by g.

Support and realizations

No Gaussian measure on $\mathcal{H}(g)$ if it is infinite dimensional.

GP g induces a Gaussian measure on a space of functions $\mathcal{F} \hookleftarrow \mathcal{H}(g)$

(e.g., abstract Weiner space) under mild assumptions on the topology of $\mathcal{F}$.

The realizations (samples) of g are the support of this measure.

This is the closure $\overline{\mathcal{H}(g)}$ of $\mathcal{H}(g)$ in $\mathcal{F}$.

Trivial example

The linear covariance function $k(t, t') = t \cdot t'$ induces a GP with realizations equal to the space $\mathcal{H}(k) = \mathbb{R} \cdot (t \mapsto t)$ of linear functions.

Non-trivial example

The squared exponential covariance function

$$k(t, t') = \exp\left(\frac{1}{2}(t - t')^2\right)$$

induces a GP with realizations dense (Fréchet topology) in $C^\infty(\mathbb{R}, \mathbb{R})$.

Universal Approximation Theorem

The previous slide shows examples of universal approximation theorems of GPs.

Let me go back and explain...

Theorems

Universal Approximation Theorem

The previous slide shows examples of universal approximation theorems of GPs.

Let me go back and explain...

Representer Theorem

Given data, the posterior GP mean of the prior $\mathcal{GP}(0, k)$ is the function of lowest $\mathcal{H}(k)$-norm interpolating this data.

This is why Frequentists use GPs, Bayesian do not care and say that of course Bayesian update has nice properties.

Similar theorems explain the double descent curve of neural nets.

Idea: Constructing covariance functions

- A GP can be interpreted via its RKHS.
- Many constructions for covariance functions are possible.
- Interpret these constructions via their RKHS.
- Knowledge about suitable regression functions yields a kernel.
- We have seen base kernels above.
- Now: build new covariance functions from old ones.
- Later: some interesting kernels.

Sums

Theorem

Let $g_1 = (0, k_1)$ and $g_2 = (0, k_2)$ GPs and $g = (0, k_1 + k_2)$. Then,

$$\mathcal{H}(g) = \mathcal{H}(g_1) + \mathcal{H}(g_2).$$

(For a suitable choice of the scalar product in the sum, which is usually not direct.)

- Explain an effect as a sum of two causes
  - E.g. smooth plus periodic
- Extrapolation when adding 1D-kernels in high dimensions.
  - Might weaken the curse of dimensionality
  - Might overfit when extrapolating
  - Decent for very few data points
- Learnable and interpretable hyperparameters for weighting.
- Special case from above: one summand is the noise.
- In practice: use a summand for “unexplained behavior”.

Products

Theorem

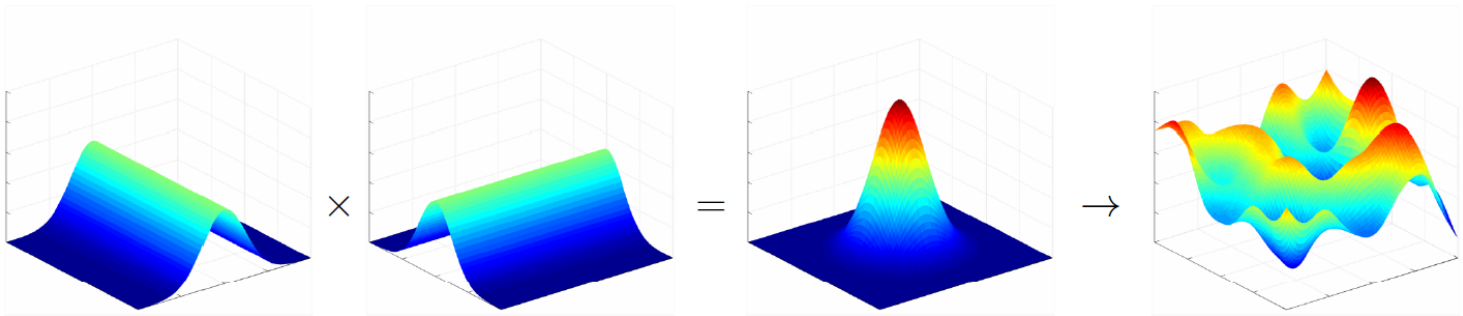
Let $g_1 = (0, k_1)$ and $g_2 = (0, k_2)$ GPs and $g = (0, k_1 \cdot k_2)$. Then,

$$\mathcal{H}(g) = \mathcal{H}(g_1) \otimes \mathcal{H}(g_2).$$

The Hilbert space $\otimes$ is the completion of the vector space $\otimes$. The fun begins when tensoring the spaces of realizations.

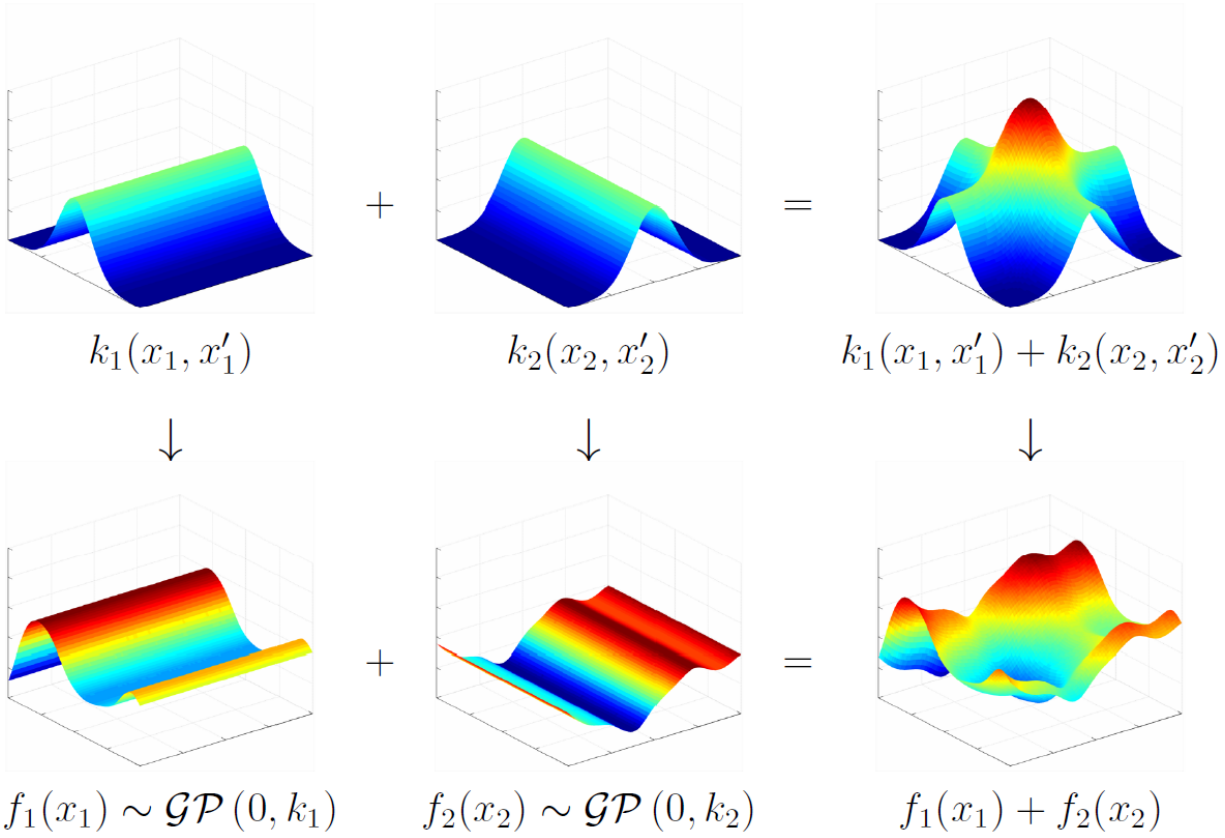
- All causes are needed for an effect.
  - E.g. locally periodic behavior
- Product over input dimensions.
- No hyperparameters necessary.
- Special case: multiplying constant covariance leads to scaling.
- Special case: automatic relevance determination

Sums and Products



from David Duvenaud's PhD thesis.

Sums and Products



from David Duvenaud's PhD thesis.

Theorem

Let V be a closed subspace of $\mathcal{H}(g)$ with corresponding orthogonal projection $\pi : \mathcal{H}(g) \to V$. Let $g_V := (0, \pi_1 k \pi_2))$. Then,

$$\mathcal{H}(g_V) = \pi(\mathcal{H}(g)) =: \pi_* \mathcal{H}(g).$$

Example: V closed subspace of symmetric functions with

$$\pi : f \mapsto \left(x \mapsto \frac{1}{2}(f(x) + f(-x)) \right).$$

Kernel (can be simplified for stationary covariances):

$$\frac{1}{4}(k(t,t') + k(-t,t') + k(t,-t') + k(-t,-t'))$$

This demonstrates how to use Reynold's operator to construct covariance functions for invariants under finite Groups.

If you can solve the necessary integral, this also works for Lie groups.

If π is only epic but not orthogonal, then the theorem holds as sets but not as Hilbert spaces.

Input transformations (Manifold GPs)

Let $g = \mathcal{GP}(\mu, k)$ defined on $\mathbb{R}^d$ and $f : \mathbb{R}^c \to \mathbb{R}^d$. Then

$$f^* g := \mathcal{GP}(x \mapsto \mu(f(x)), (x, y') \mapsto k(f(x), f(x')))$$

is defined on $\mathbb{R}^c$ with

$$\mathcal{H}(f^* g) = f^* \mathcal{H}(g).$$

Using these constructions to explain data

Automatic finding of explanations

- Train several GPs on a dataset.
- Take the best fitting one.
- Its covariance function interpretes the data.
- Iterative procedures possible.

Used in GAIA to explain anomalies.

Message for interpretability

Interpretability of RKHS

- Covariance functions and GPs can be interpreted mathematically.
- Various constructions and combinations of covariance functions.
- Communicate such interpretations to laypeople?
- Laypeople teach their knowledge to AIs?
- Automatically choosing a covariance function interprets data.

Questions?

Questions?

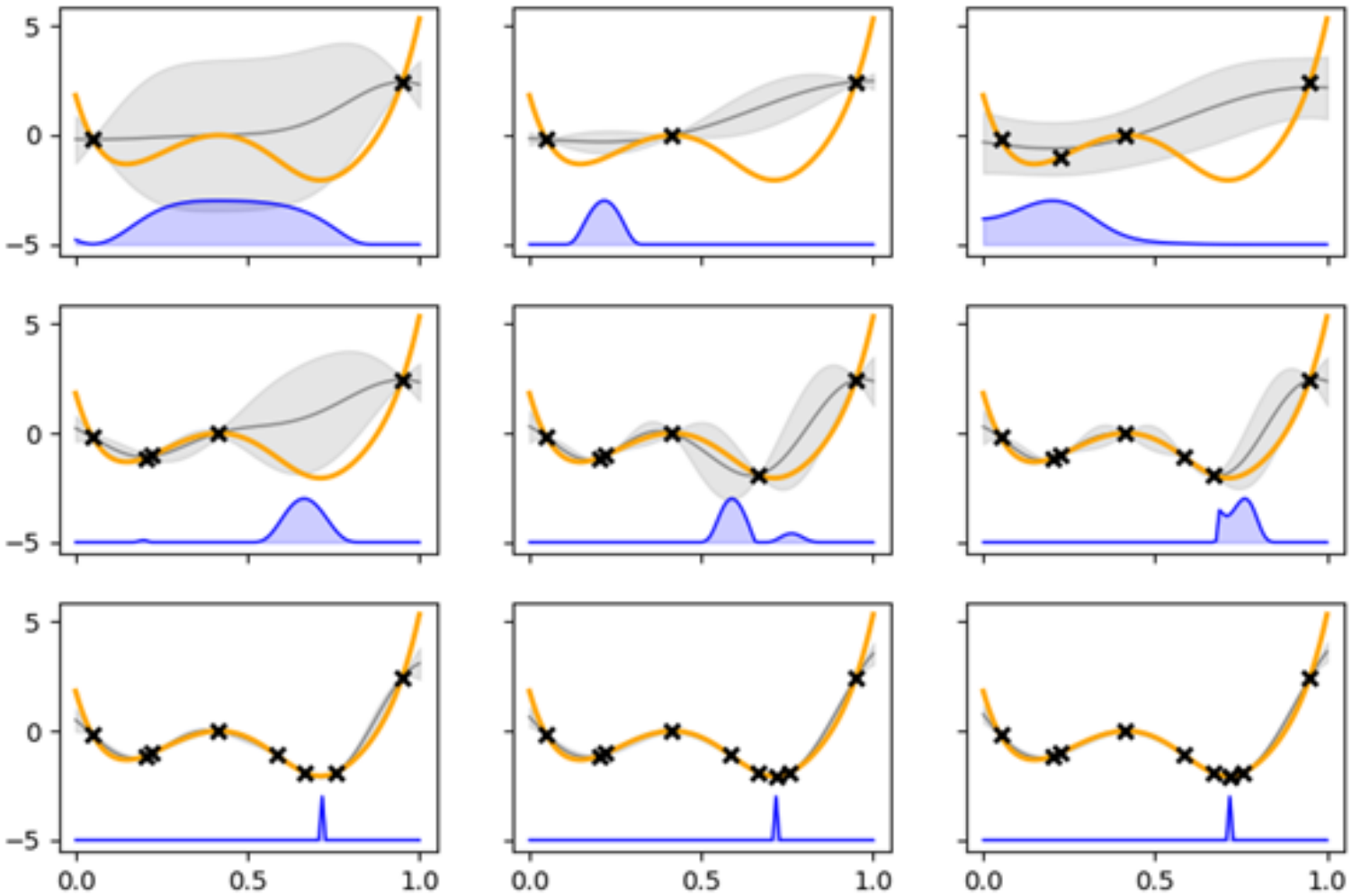
Using variances of Gaussian processes

Variances

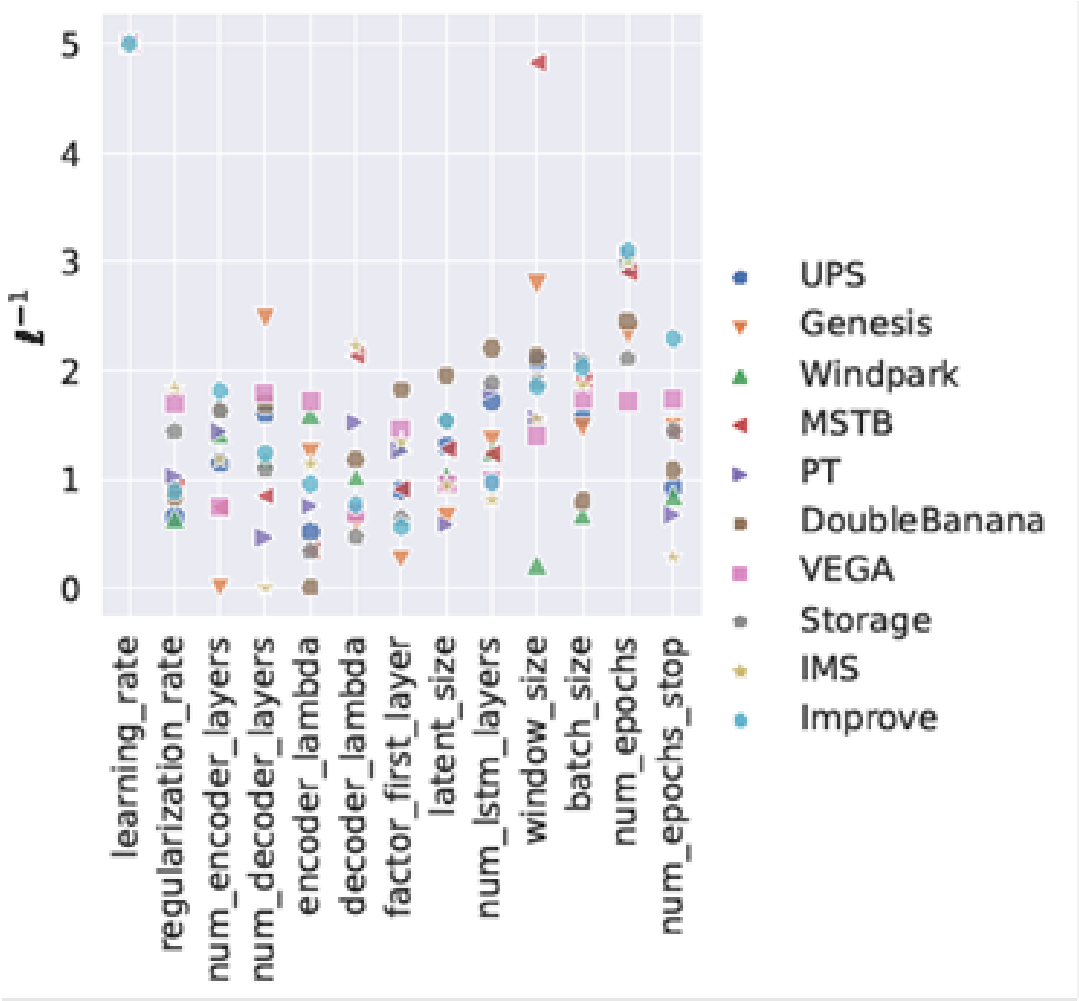
- A posterior GP is a GP, which comes with variances.
- This is a way to quantify uncertainty, even to humans.
- Are there enough data points near predictions? (Depends on length scales)
- Compare/calibrate model uncertainties with measurement noise.
- Heterogenous noise: Warping, individual noise levels, . . .

Bayesian optimization

- Minimize (expensive to evaluate function) f.
- Approximate f by a Gaussian process using few data points.
- Sample f at promising point (high variance, low prediction)

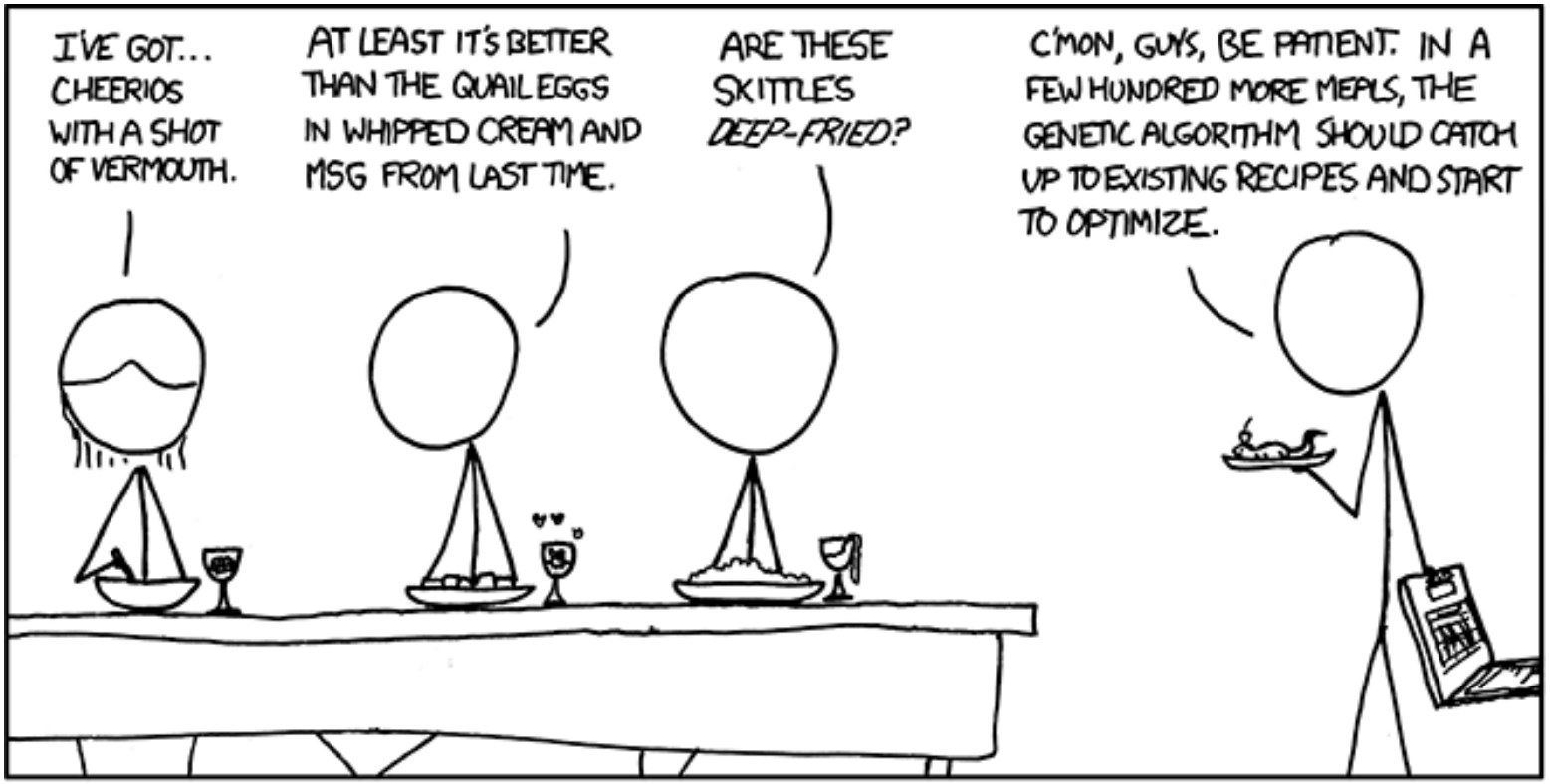


Example: Bayesian optimization for hyperparameter tuning



joint with Alissa Müller and Alexander von Birgelen.

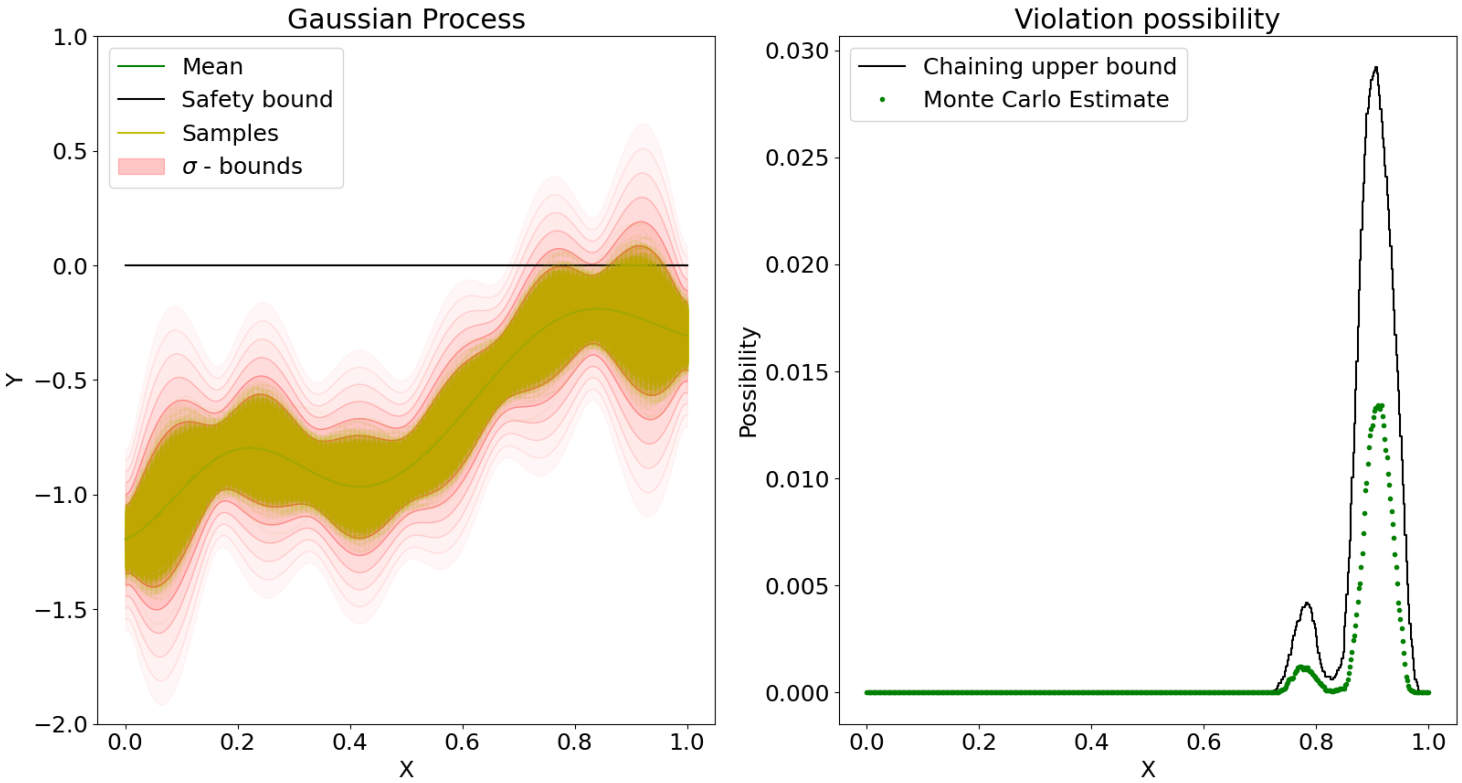
Safe Optimization



Safety bounds

Probability that a GP surpass a safety bound:

- Approximate via MCMC (expensive to compute)
- Upper bound via chaining (guaranteed safety)



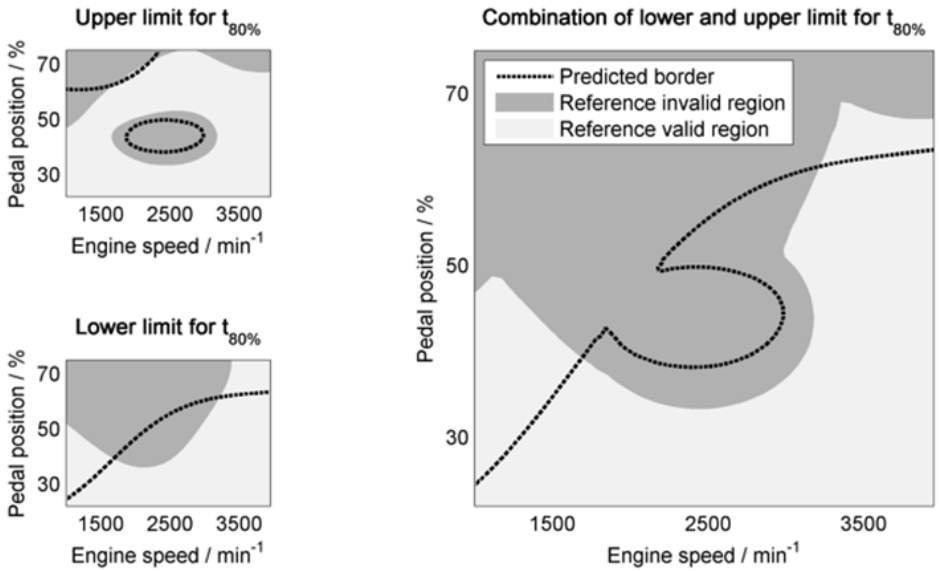
joint with Fabian Mies, Jörn Tebbe, and XXX.

“Online DoE” for drivability calibration

- Optimize torque demand response to the gas pedal in a sports car.
- Controllable Inputs: time constant T_1, amplification factor K_d.
- Uncontrollable Inputs: engine speed, pedal position.
- Avoid uncomfortably large or unsporty small jerks (VDV) and undynamic response (t_{80}): useless, expensive, uncomfortable.
- Design of experiment (DoE), adaptable online.

“Online DoE” for drivability calibration

- Optimize torque demand response to the gas pedal in a sports car.
- Controllable Inputs: time constant T_1, amplification factor K_d.
- Uncontrollable Inputs: engine speed, pedal position.
- Avoid uncomfortably large or unsporty small jerks (VDV) and undynamic response (t_{80}): useless, expensive, uncomfortable.
- Design of experiment (DoE), adaptable online.

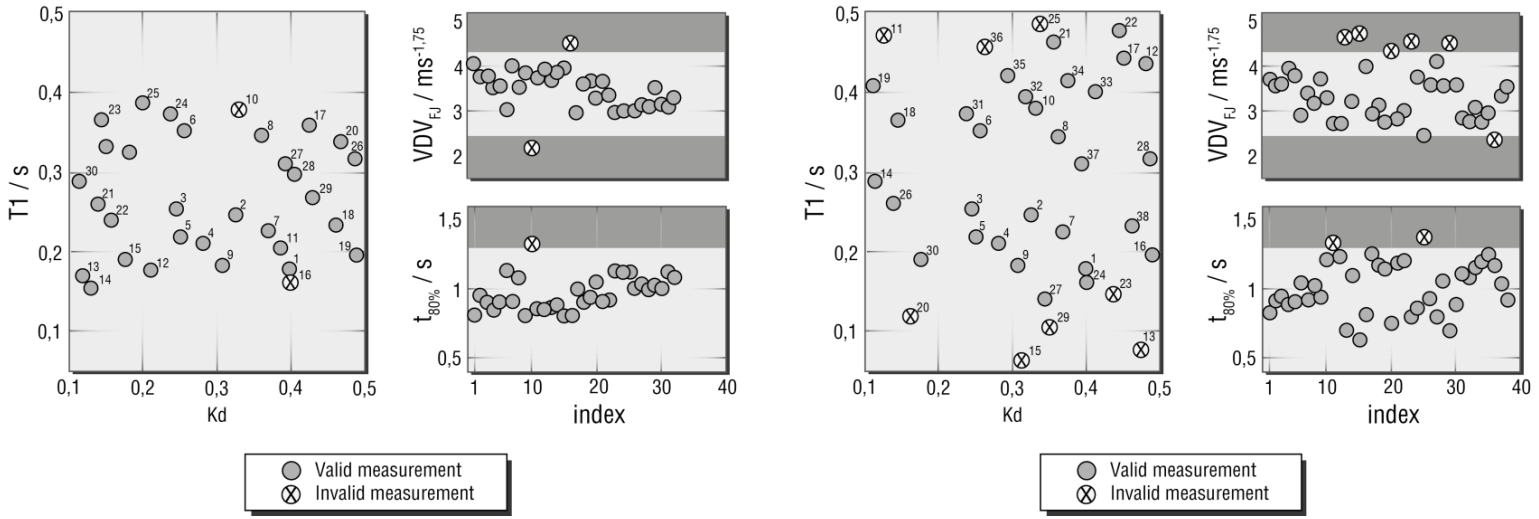


Prediction of the border of the valid design space at $K_d = 0.1$ and $T_1 = 0.5s$ after 46 measurements. Reference boundaries after 300 measurements.

Thewes, Krause, Reuber, Lange-Hegermann, Dziadek, Rebbert, *Efficient in-vehicle calibration by the usage of automation and enhanced online DoE approaches*

“Online DoE” for drivability calibration

- Optimize torque demand response to the gas pedal in a sports car.
- Controllable Inputs: time constant T_1, amplification factor K_d.
- Uncontrollable Inputs: engine speed, pedal position.
- Avoid uncomfortably large or unsporty small jerks (VDV) and undynamic response (t_{80}): useless, expensive, uncomfortable.
- Design of experiment (DoE), adaptable online.



conservative setting aggressive setting

Thewes, Krause, Reuber, Lange-Hegermann, Dziadek, Rebbert, *Efficient in-vehicle calibration by the usage of automation and enhanced online DoE approaches*

Message for interpretability

Interpretability of variances

- Variances give model uncertainty
- Variances can be used for optimization
- Variances address safety constraints
- Variances guides where additional data is needed

Questions?

Questions?

Gaussian processes and Linear Systems (in the sense of linear algebra)

Linear Systems (in the sense of linear algebra)

For $\mathcal{F} = C^\infty(\mathbb{R}, \mathbb{R})$ and $A = \begin{bmatrix} 2 & -3 \end{bmatrix}$ consider

$$\mathrm{sol}_{\mathcal{F}}(A) := \left\{ \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} \in \mathcal{F}^{2\times 1} \,\middle|\, A \cdot \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = 0 \right\}$$

Linear Systems (in the sense of linear algebra)

For $\mathcal{F} = C^\infty(\mathbb{R}, \mathbb{R})$ and $A = \begin{bmatrix} 2 & -3 \end{bmatrix}$ consider

$$\mathrm{sol}_{\mathcal{F}}(A) := \left\{ \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} \in \mathcal{F}^{2\times 1} \,\middle|\, A \cdot \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = 0 \right\}$$

Use $B = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ as parametrization:

$$\mathrm{sol}_{\mathcal{F}}(A) = B \cdot \mathcal{F} = \{B \cdot f(x) \mid f(x) \in \mathcal{F}\}$$

Linear Systems (in the sense of linear algebra)

For $\mathcal{F} = C^\infty(\mathbb{R}, \mathbb{R})$ and $A = \begin{bmatrix} 2 & -3 \end{bmatrix}$ consider

$$\mathrm{sol}_{\mathcal{F}}(A) := \left\{ \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} \in \mathcal{F}^{2\times 1} \,\middle|\, A \cdot \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = 0 \right\}$$

Use $B = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ as parametrization:

$$\mathrm{sol}_{\mathcal{F}}(A) = B \cdot \mathcal{F} = \{B \cdot f(x) \mid f(x) \in \mathcal{F}\}$$

Taking a GP prior $g = \mathcal{GP}(0, k)$ for $g \in \mathcal{F}$ yields a GP prior

$$B_* g = \mathcal{GP}(0, BkB^T)$$

for $\mathrm{sol}_{\mathcal{F}}(A)$.

Message for interpretability

Interpretability of linear algebra

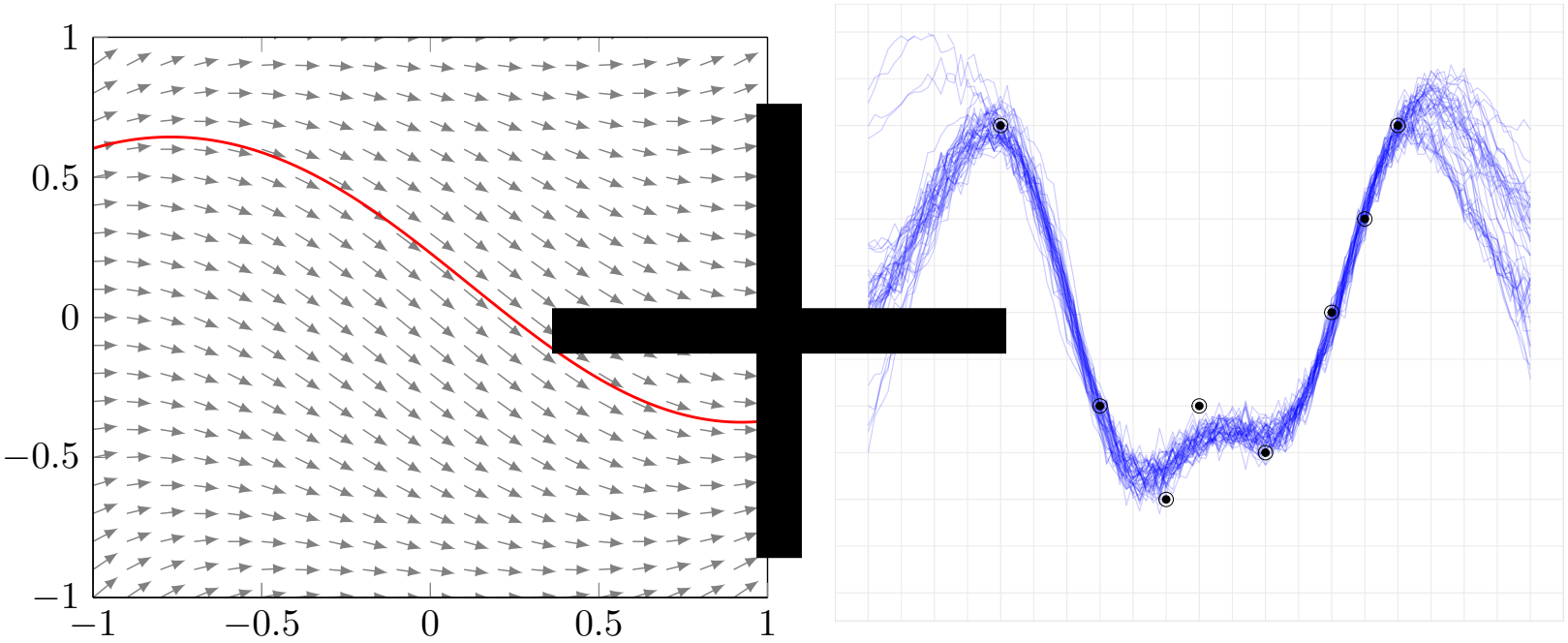
- GPs play nice with linear algebra
- GPs allow to encode linear dependencies

Questions?

Questions?

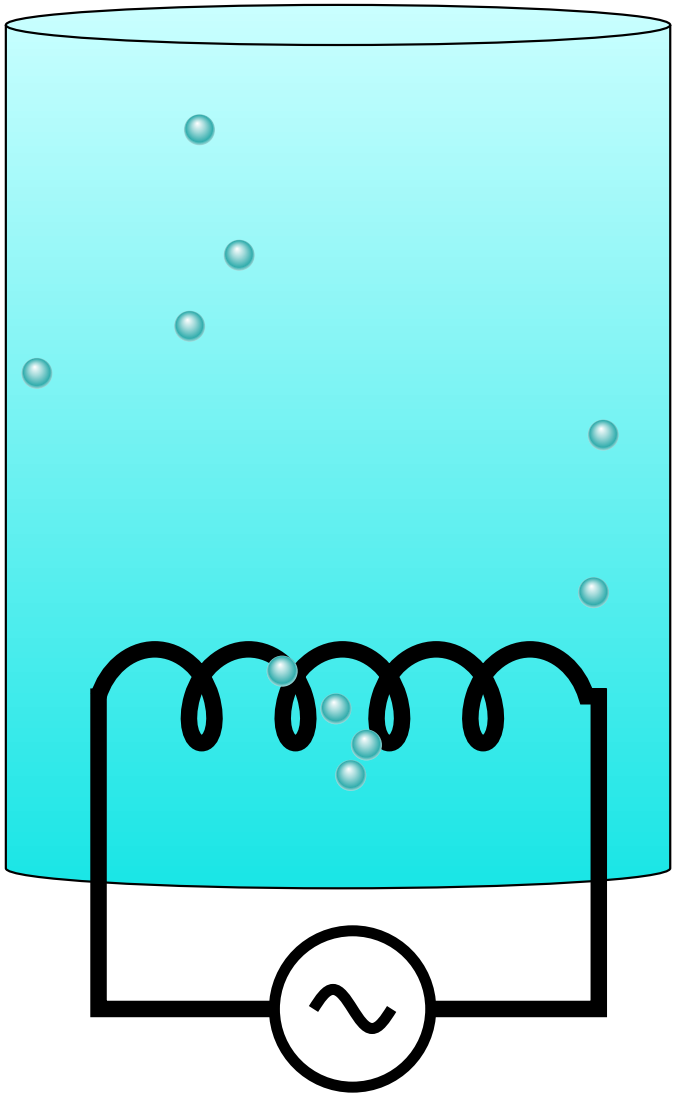
Gaussian processes and linear ordinary differential equations

Combination of Gaussian processes with operator equations

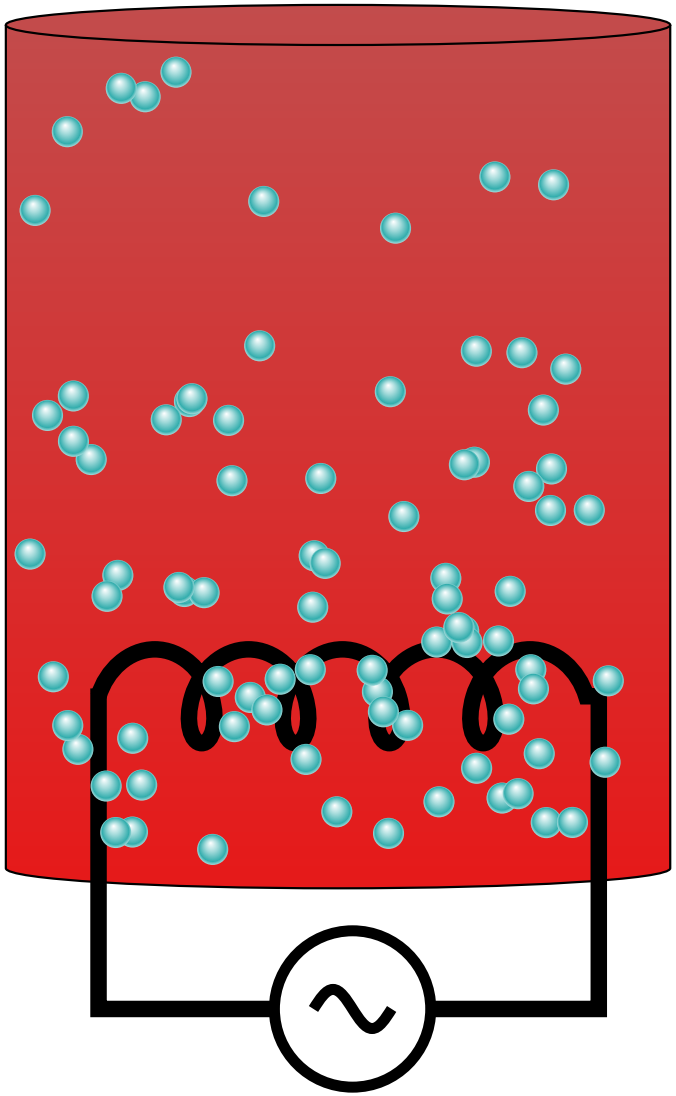


- Combine **strict, global information** from differential equations with **noisy, local information** from observations.
- Incorporate justified assumptions: use the **full information** of the observations for a precise regression model.

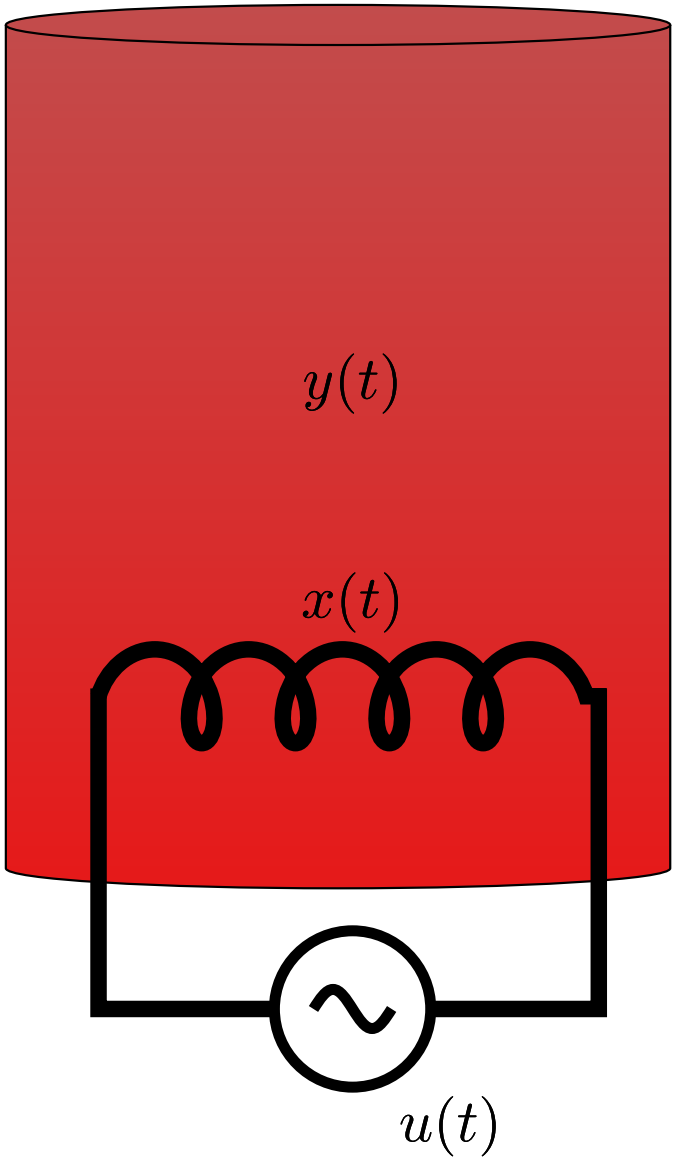
Example



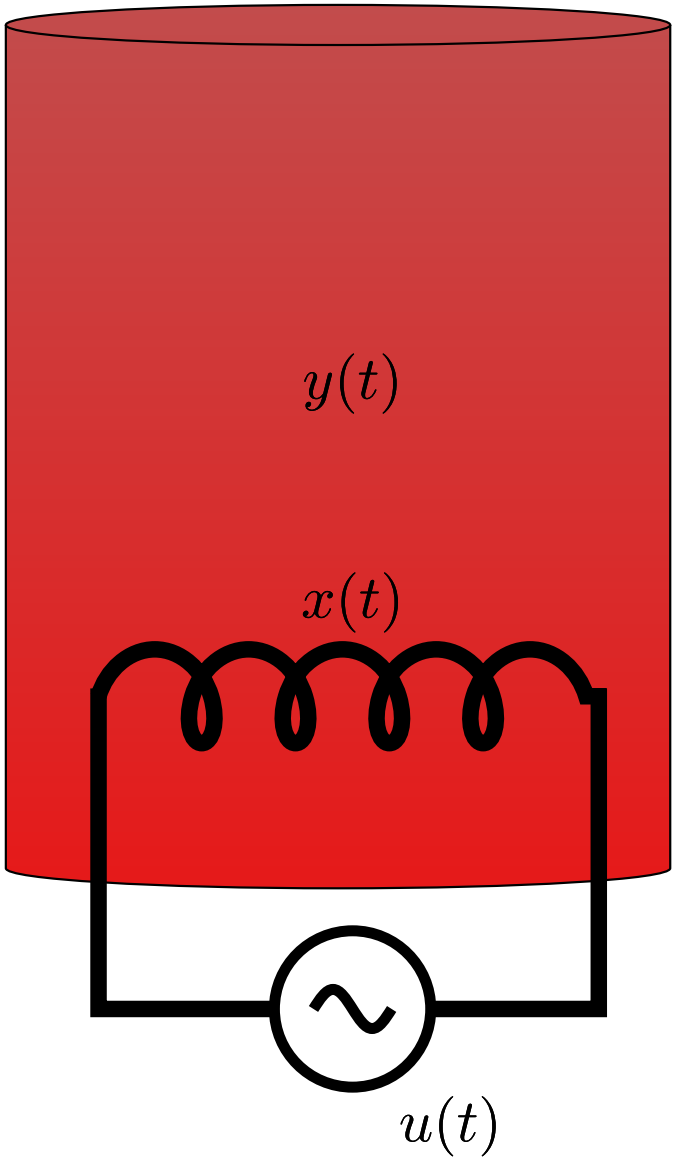
Example



Example

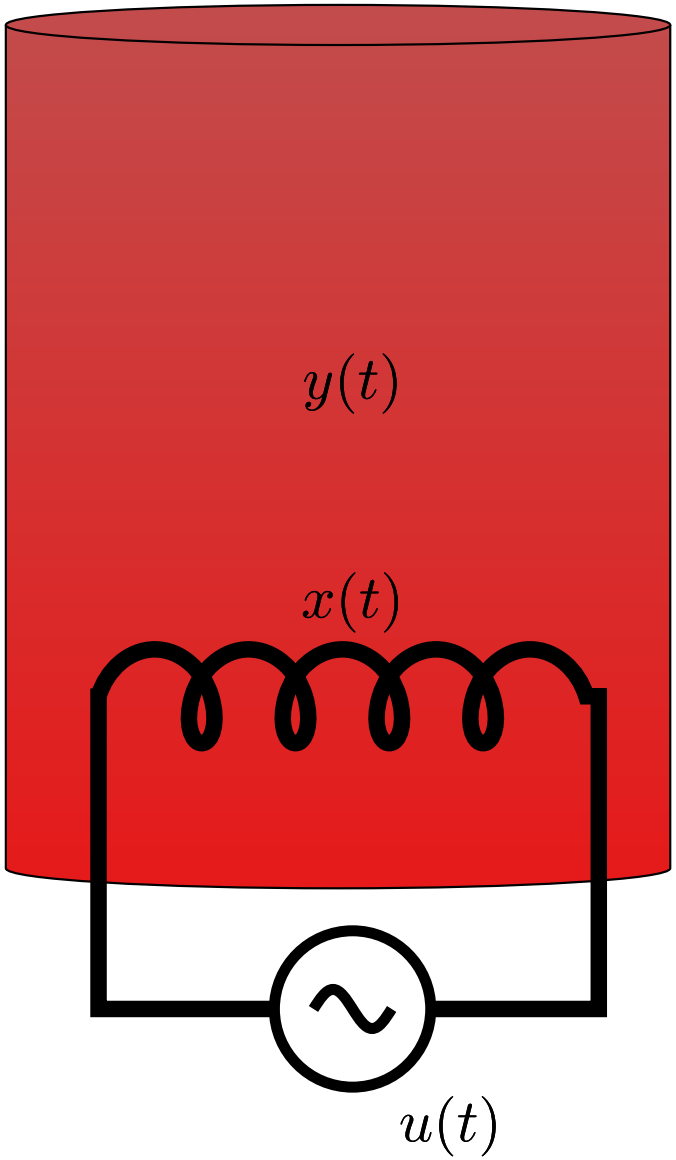


Example

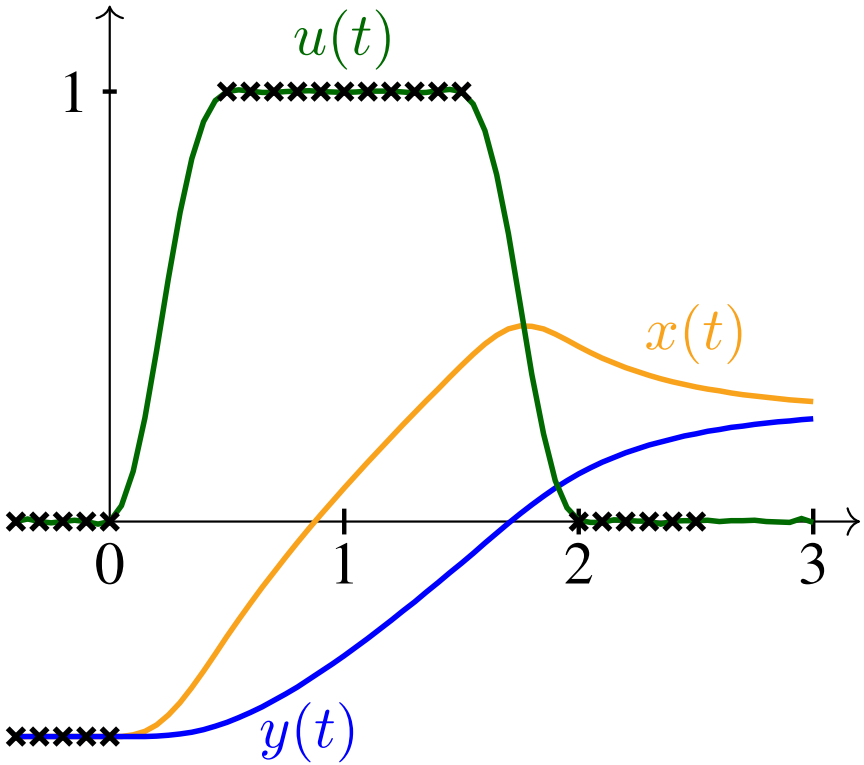


$$\partial_t x(t) = -(x(t) - y(t)) + u(t)$$
$$\partial_t y(t) = +(x(t) - y(t))$$

Example



$$\partial_t x(t) = -(x(t) - y(t)) + u(t)$$
$$\partial_t y(t) = +(x(t) - y(t))$$



Constructing suitable covariance functions

$$\partial_t x(t) = -(x(t) - y(t)) + u(t) \quad \Leftrightarrow 0 = -\partial_t x(t) - x(t) + y(t) + u(t)$$
$$\partial_t y(t) = +(x(t) - y(t)) \quad \Leftrightarrow 0 = x(t) - \partial_t y(t) - y(t)$$

Is the same as

$$\underbrace{\begin{bmatrix} -\partial_t - 1 & 1 & 1 \\ 1 & -\partial_t - 1 & 0 \end{bmatrix}}_{=:A} \cdot \underbrace{\begin{bmatrix} x(t) \\ y(t) \\ u(t) \end{bmatrix}}_{=:f} = 0$$

Here, A is an operator Matrix. Its entries are polynomials in ∂_t, i.e. in the polynomial ring $R = \mathbb{R}[\partial_t]$.

Smith normal form

Given a matrix A (over a PID), there are invertible matrices S and T s.t.

$$SAT = D$$

where D is a matrix with non-zero entries only on the diagonal.

(D can be made unique by demanding that each diagonal entry divides the next one.)

Everything can be computed in polynomial time (as long as the PID is Euclidean).

Using the Smith normal form

$$Af = 0 \Leftrightarrow SAT \underbrace{T^{-1} f}_{=:h} = 0$$
$$\Leftrightarrow Dh = 0$$

If we get a GP prior for $h = T^{-1} f$, we have one for $f = Th$.

Prior for h

$$\begin{bmatrix} 1 & & & & \\ & \partial_t - 1 & & & \\ & & \partial_t^2 + 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} h_1(t) \\ h_2(t) \\ h_3(t) \\ h_4(t) \\ h_5(t) \end{bmatrix} = 0$$

Since we can easily solve such ODEs:

$h_1(t) = 0$	$k_1(t_1, t_2) = 0$
$h_2(t) = c \cdot \exp(t)$	$k_2(t_1, t_2) = \exp(t_1)\exp(t_2)$
$h_3(t) = c_1 \sin(t) + c_2 \cos(t)$	$k_3(t_1, t_2) = \cos(t_1 - t_2)$
$h_4(t)$ arbitrary (smooth)	$k_4(t_1, t_2) = \exp(-\frac{1}{2}(t_1 - t_2)^2)$
$h_5(t)$ arbitrary (smooth)	$k_5(t_1, t_2) = \exp(-\frac{1}{2}(t_1 - t_2)^2)$

joint with Andreas Besginow.

Let $T \in R^{\ell \times m}$ and $g = \mathcal{GP}(\mu, k)$.

Lemma

Assume that T commutes w.r.t. expectation of the relevant measures.

- (Pushforward is again a Gaussian process)
  $T_* g = \mathcal{GP}(T\mu(x), Tk(x, x')(T')^T)$ where T' operates on x'.
- (Realizations behave reasonable)
  For $g = \mathcal{GP}(0, k)$ with zero mean function, $\mathcal{H}(T_* g) = T\mathcal{H}(g)$.

Example: Simple Control System

Time dependent system $\partial_t x(t) = t^3 u(t)$.

(We need the Jacobson form instead of the Smith form, since we are over a Weyl algebra)

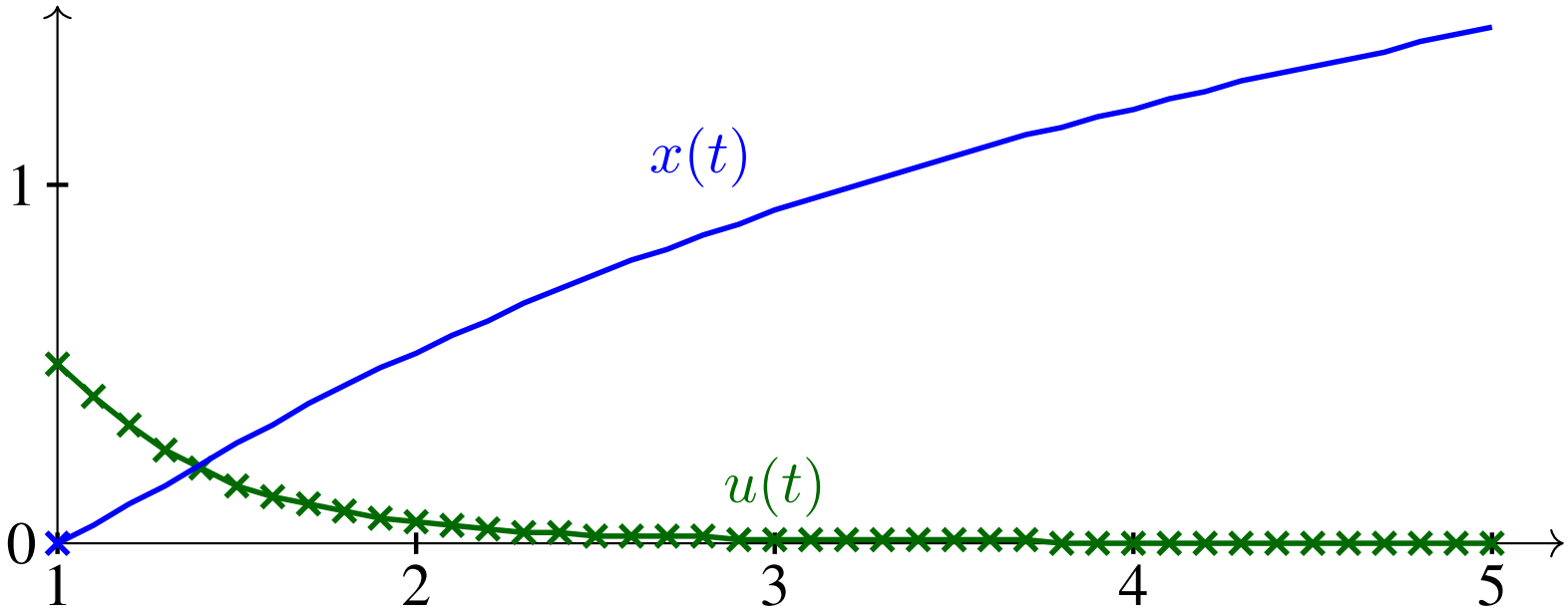
Example: Simple Control System

Time dependent system $\partial_t x(t) = t^3 u(t)$.

(We need the Jacobson form instead of the Smith form, since we are over a Weyl algebra)

Set an input $u(t)$ to **influence** a state $x(t)$.

Set $x(1) = 0$ and $u(t) = \frac{1}{t^4+1}$ for $t \in \{1, \frac{11}{10}, \frac{12}{10}, \ldots, 5\}$.



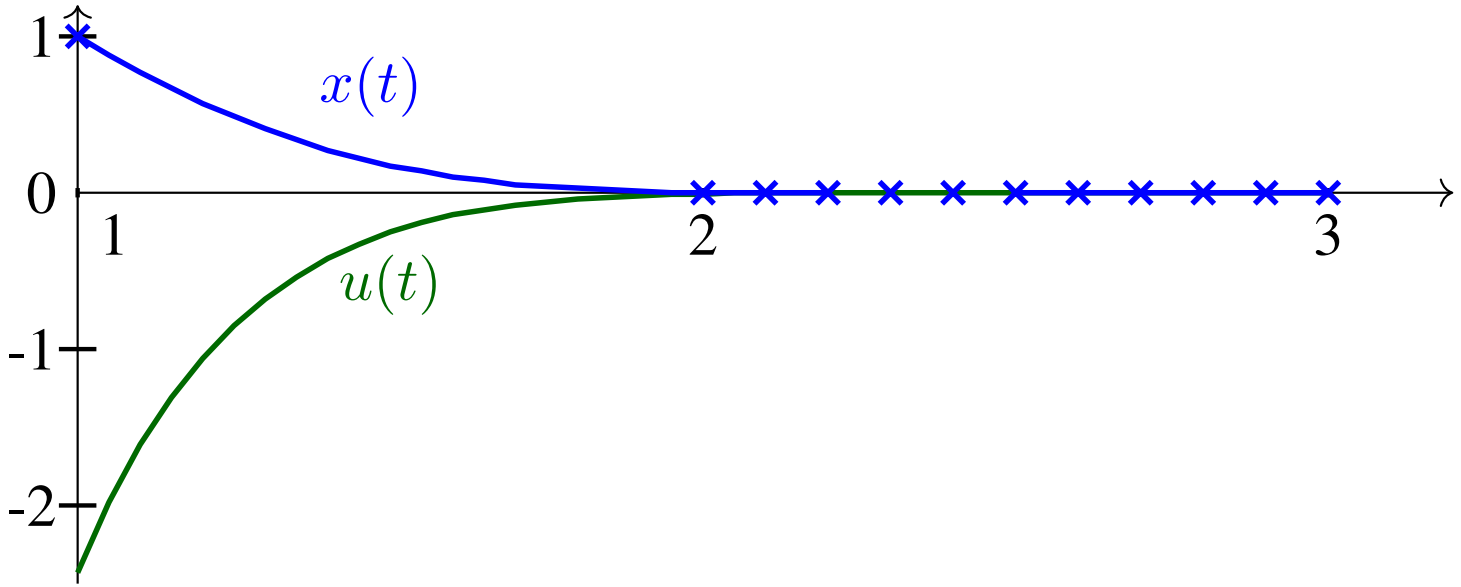
Model: $x(5) \approx 1.436537$, close to $\int_1^5 \frac{t^3}{t^4+1}\,\mathrm{d}t \approx 1.436551$.

Example: Simple Control System

Time dependent system $\partial_t x(t) = t^3 u(t)$.

(We need the Jacobson form instead of the Smith form, since we are over a Weyl algebra)

Prescribe a state $x(t)$. Automatically **construct** an input $u(t)$.



Message for interpretability

Interpretability of ODEs

- GPs play nice with linear ODEs
- Can be used to learn/understand systems
- Can be used to control systems
- Can be used as a strong inductive bias
- Can be used as a very strong inductive bias

Questions?

Questions?

Gaussian processes and linear partial differential equations

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. We can compute with operators
2. Functions yield enough solutions
3. Gaussian processes describe functions

4. Operators and topology are compatible
5. Gaussian processes and topology are compatible

6. Gaussian processes and operators are compatible

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. We can compute with operators: R allows a Gröbner basis algorithm.
2. Functions yield enough solutions: $\mathcal{F}$ is an injective R-module.
3. Gaussian processes describe functions: There is a scalar $g = \mathcal{GP}(0, k)$ s.t. its RKHS $\mathcal{H}(g)$ is dense in $\mathcal{F}$ and its set of realizations is contained (a.s.) in $\mathcal{F}$.
4. Operators and topology are compatible: R acts continuously on $\mathcal{F}$.
5. Gaussian processes and topology are compatible: GPs in $\mathcal{F}$ are $1:1$ with Gaussian measures on $\mathcal{F}$ w.r.t. the Borel σ-algebra.
6. Gaussian processes and operators are compatible: the operation of R on $\mathcal{H}(g)$ commutes with expectation (g induces measure).

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. We can compute with operators: R allows a Gröbner basis algorithm.
2. Functions yield enough solutions: $\mathcal{F}$ is an injective R-module.
3. Gaussian processes describe functions: There is a scalar $g = \mathcal{GP}(0, k)$ s.t. its RKHS $\mathcal{H}(g)$ is dense in $\mathcal{F}$ and its set of realizations is contained (a.s.) in $\mathcal{F}$.
4. Operators and topology are compatible: R acts continuously on $\mathcal{F}$.
5. Gaussian processes and topology are compatible: GPs in $\mathcal{F}$ are $1:1$ with Gaussian measures on $\mathcal{F}$ w.r.t. the Borel σ-algebra.
6. Gaussian processes and operators are compatible: the operation of R on $\mathcal{H}(g)$ commutes with expectation (g induces measure).

Theorem

Assumptions hold for $R = \mathbb{R}[\partial_{x_1}, \ldots, \partial_{x_d}]$, $\mathcal{F} = C^\infty(\mathbb{R}^d, \mathbb{R})$ with Fréchet topology, and g with SE covariance.

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. We can compute with operators: R allows a Gröbner basis algorithm.
2. Functions yield enough solutions: $\mathcal{F}$ is an injective R-module.
3. Gaussian processes describe functions: There is a scalar $g = \mathcal{GP}(0, k)$ s.t. its RKHS $\mathcal{H}(g)$ is dense in $\mathcal{F}$ and its set of realizations is contained (a.s.) in $\mathcal{F}$.
4. Operators and topology are compatible: R acts continuously on $\mathcal{F}$.
5. Gaussian processes and topology are compatible: GPs in $\mathcal{F}$ are $1:1$ with Gaussian measures on $\mathcal{F}$ w.r.t. the Borel σ-algebra.
6. Gaussian processes and operators are compatible: the operation of R on $\mathcal{H}(g)$ commutes with expectation (g induces measure).

Proposition

Assumptions hold for $R = \mathbb{R}(t)\langle\partial_t\rangle$, $\mathcal{F} = C^\infty(D, \mathbb{R})$ with Fréchet topology, g with SE covariance and $D \subseteq \mathbb{R}$.

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. **We can compute with operators**: R allows a Gröbner basis algorithm.
2. **Functions yield enough solutions**: $\mathcal{F}$ is an injective R-module.
3. **Gaussian processes describe functions**: There is a scalar $g = \mathcal{GP}(0, k)$ s.t. its RKHS $\mathcal{H}(g)$ is dense in $\mathcal{F}$ and its set of realizations is contained (a.s.) in $\mathcal{F}$.
4. **Operators and topology are compatible**: R acts continuously on $\mathcal{F}$.
5. **Gaussian processes and topology are compatible**: GPs in $\mathcal{F}$ are $1:1$ with Gaussian measures on $\mathcal{F}$ w.r.t. the Borel σ-algebra.
6. **Gaussian processes and operators are compatible**: the operation of R on $\mathcal{H}(g)$ commutes with expectation (g induces measure).

Remark

Assumptions hold for $R = \mathbb{R}[x_1, \ldots, x_n]$, $\mathcal{F} = C^\infty(D, \mathbb{R})$ with Fréchet topology, g with SE covariance and $D \subseteq \mathbb{R}$.

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. We can compute with operators: R allows a Gröbner basis algorithm.
2. Functions yield enough solutions: $\mathcal{F}$ is an injective R-module.
3. Gaussian processes describe functions: There is a scalar $g = \mathcal{GP}(0, k)$ s.t. its RKHS $\mathcal{H}(g)$ is dense in $\mathcal{F}$ and its set of realizations is contained (a.s.) in $\mathcal{F}$.
4. Operators and topology are compatible: R acts continuously on $\mathcal{F}$.
5. Gaussian processes and topology are compatible: GPs in $\mathcal{F}$ are $1:1$ with Gaussian measures on $\mathcal{F}$ w.r.t. the Borel σ-algebra.
6. Gaussian processes and operators are compatible: the operation of R on $\mathcal{H}(g)$ commutes with expectation (g induces measure).

Remark

Assumptions hold for $R = \mathbb{R}[\sigma_1, \ldots, \sigma_n]$, $\mathcal{F} = C^\infty(\mathbb{R}^n, \mathbb{R})$ with Fréchet topology and g with SE covariance, where $\sigma_i(x_j) = x_i + \delta_{ij}$.

Assumptions

Let R be an $\mathbb{R}$-algebra, a *ring of linear operators*, and $\mathcal{F}$ an R-module of functions $\mathbb{R}^d \to \mathbb{R}$ with topology.

Assume:

1. We can compute with operators: R allows a Gröbner basis algorithm.
2. Functions yield enough solutions: $\mathcal{F}$ is an injective R-module.
3. Gaussian processes describe functions: There is a scalar $g = \mathcal{GP}(0, k)$ s.t. its RKHS $\mathcal{H}(g)$ is dense in $\mathcal{F}$ and its set of realizations is contained (a.s.) in $\mathcal{F}$.
4. Operators and topology are compatible: R acts continuously on $\mathcal{F}$.
5. Gaussian processes and topology are compatible: GPs in $\mathcal{F}$ are $1:1$ with Gaussian measures on $\mathcal{F}$ w.r.t. the Borel σ-algebra.
6. Gaussian processes and operators are compatible: the operation of R on $\mathcal{H}(g)$ commutes with expectation (g induces measure).

Theorem

Under the above assumptions, we can construct a GP prior for controllable systems.

Compute a Parametrization of a System

Let $M = \operatorname{coker}(A)$ be a torsionless R-module, i.e. there is a mono

$$M \xhookrightarrow{B} R^{1\times \ell''} .$$

Algorithm

- Compute $\hom_R(M, R)$ and a free hull $\hom_R(M, R) \twoheadleftarrow R^{\ell''\times 1}$.
- Gives embedding
$$\hom_R(\hom_R(M, R), R) \hookrightarrow R^{1\times \ell''} .$$
- If $M \to \hom_R(\hom_R(M, R), R)$ is monic, then
$$M \hookrightarrow \hom_R(\hom_R(M, R), R) \hookrightarrow R^{1\times \ell''} .$$

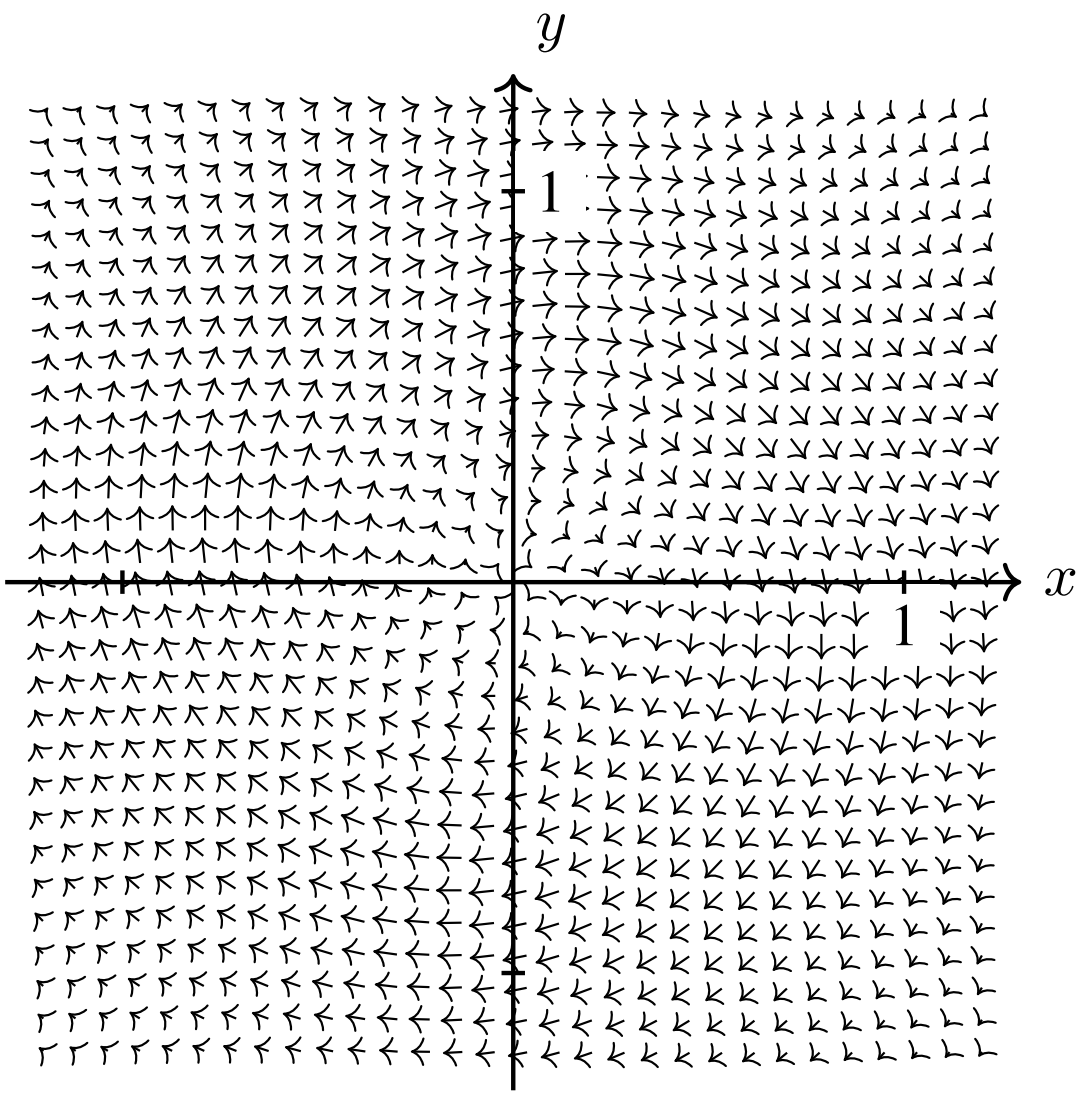
All steps are possible using Gröbner bases.

Parametrize the system by applying the exact (since $\mathcal{F}$ is injective) functor $\hom_R(-, \mathcal{F})$:

$$\operatorname{sol}_{\mathcal{F}}(A) = \hom_R(M, \mathcal{F}) \xleftarrow{B} \mathcal{F}^{\ell''\times 1}$$

(More is possible if certain Ext's vanish.)

Example: Maxwell's Equations



Example: the Koszul Complex

The matrix $A = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$ yields tangents of a sphere.

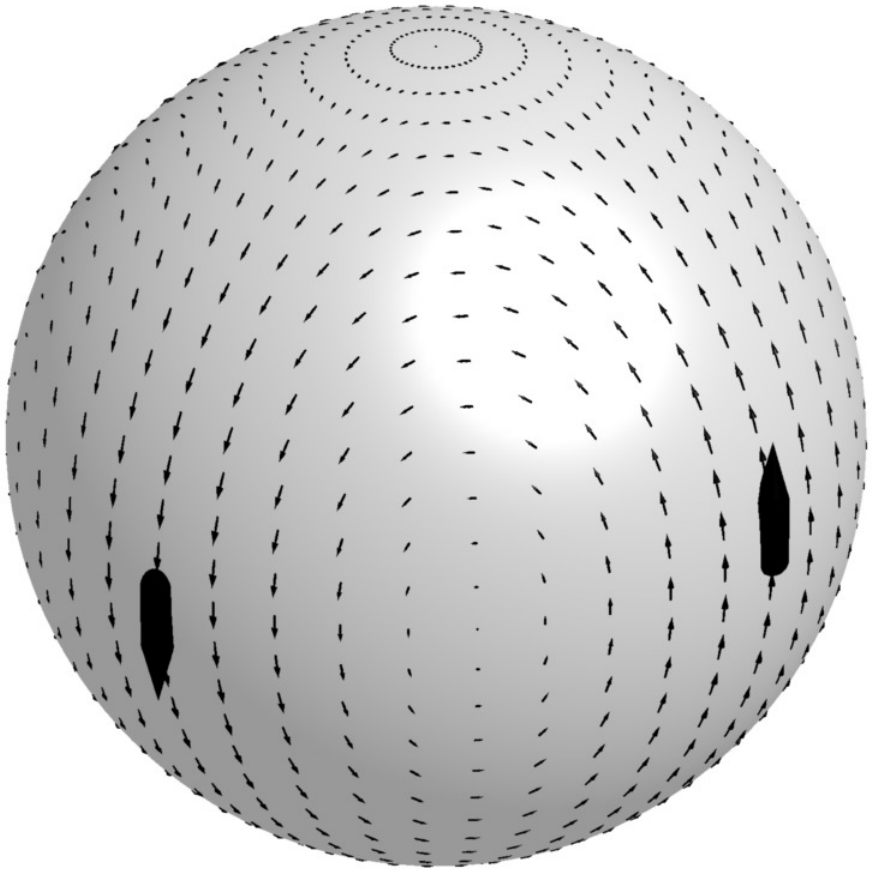
Parametrized by $B = \begin{bmatrix} 0 & x_3 & -x_2 \\ -x_3 & 0 & x_1 \\ x_2 & -x_1 & 0 \end{bmatrix}$.

Covariance function for tangential fields on the sphere:

$$\begin{bmatrix} -y_1y_2 - z_1z_2 & y_1x_2 & z_1x_2 \\ x_1y_2 & -x_1x_2 - z_1z_2 & z_1y_2 \\ x_1z_2 & y_1z_2 & -x_1x_2 - y_1y_2 \end{bmatrix} \cdot k$$

Example: the Koszul Complex

Smooth field, conditioned at **4 points** at the equator, neighboring tangent vectors point into opposed directions (north/south).



Example: Intersecting two Koszul Complexes

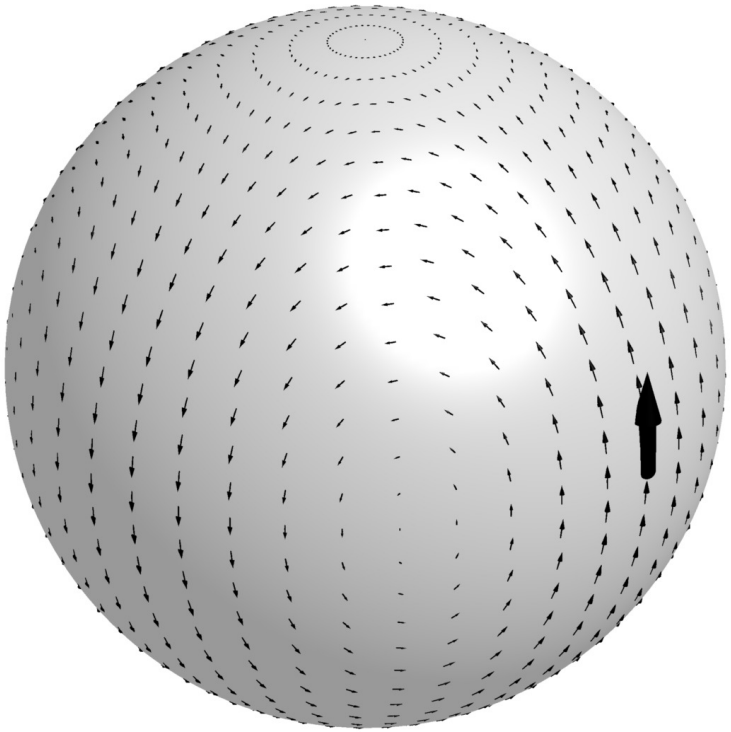
The matrix $A = \begin{bmatrix} \partial_1 & \partial_2 & \partial_3 \end{bmatrix}$ represents the divergence and its kernel is the rotation $B = \begin{bmatrix} 0 & \partial_3 & -\partial_2 \\ -\partial_3 & 0 & \partial_1 \\ \partial_2 & -\partial_1 & 0 \end{bmatrix}$.

Intersecting parametrizations

We can intersect parametrizations via a pullback under suitable assumptions

Example: Intersecting two Koszul Complexes

Intersection of tangent fields with divergence free fields.
Data: 2 points opposed at the equator with tangents pointing north:



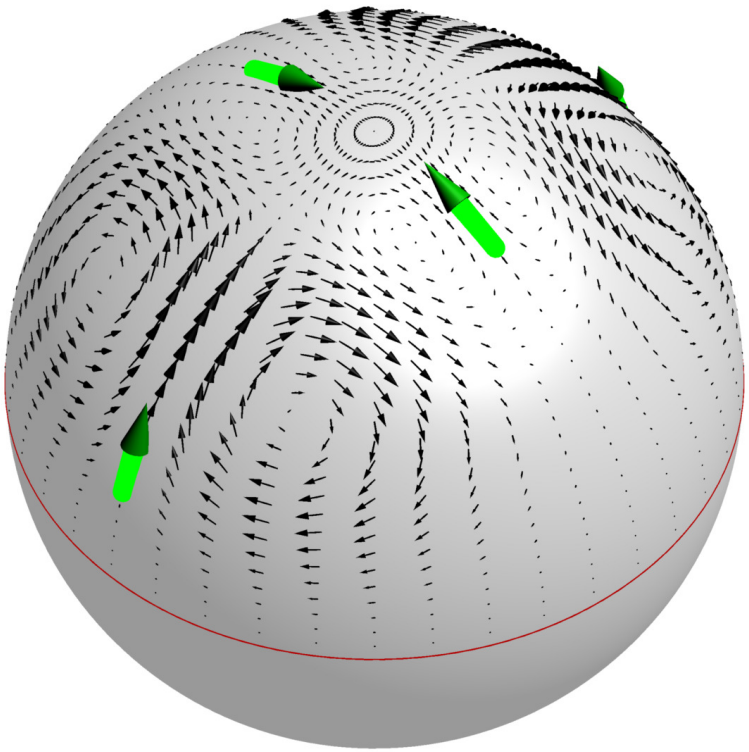
Dirichlet Boundary Conditions

Parametrization of Dirichlet boundary conditions

Functions vanishing on hyperplane $z = 0$: $\langle z \rangle \trianglelefteq \mathcal{F} = C^\infty(\mathbb{R}^d, \mathbb{R})$.

Intersect parametrizations via pullback.

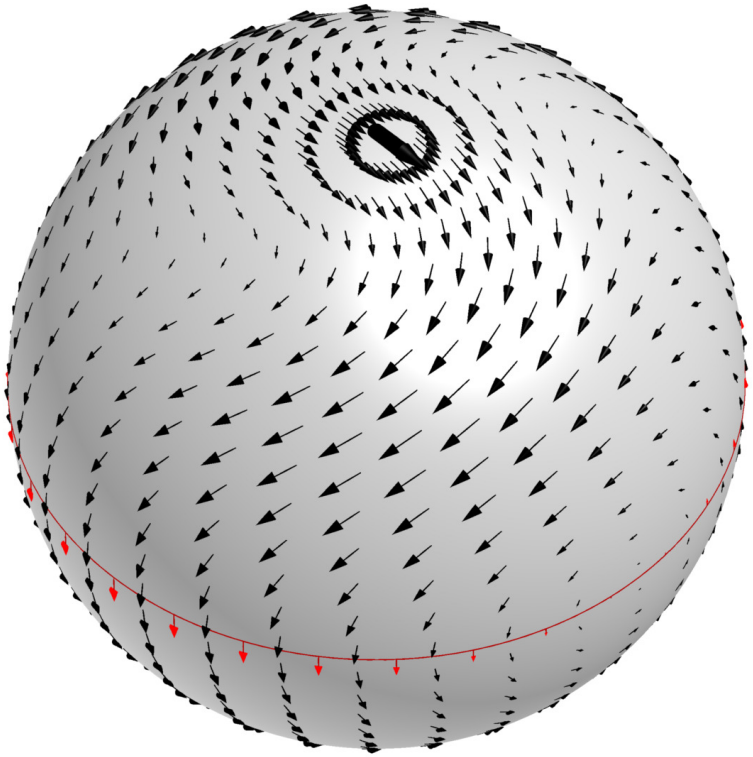
Example: Dirichlet Boundary Conditions and two Koszul Complexes



Inhomogeneous boundary conditions

Smooth divergence free fields f on the sphere and inhomogeneous boundary condition $f_3(x, y, 0) = y$.

Take particular solution $\mu = \begin{bmatrix} 0 & -z & y \end{bmatrix}^T$ as mean.



Message for interpretability

Interpretability of Operators

- GPs play nice with linear operators (in particular PDEs)
- Can be used to learn/understand systems
- Can be used as a very, *very* strong inductive bias

Questions?

Questions?

Summary

There is more to say about Gaussian processes...

There is more to say...

- Larger datasets (approximations, tricks from linear algebra)
- More constructions of covariances on manifolds or graphs
- More about Bayesian optimiations
- Deep Gaussian processes (various schools)
- Stochastic differential equations (finance, control)
- Mathematical foundations
- GPs for Classification
- Unsupervised GP models

Summary

GPs might be suitable for you, if you have. . .

- . . . a small dataset
- . . . a lot of expert knowledge
- . . . safety constraints
- . . . want to interpret your model

Summary

Interpretability of GPs

- Covariance functions and GPs can clearly be interpreted mathematically via a dictionary $k \leftrightarrow$ RKHS
- Encoding expert knowledge in covariance functions is possible
- Automatically choosing a covariance function interprets data
- Covariance functions can be combined
- Interpretable parameters can be learned
- Build in (learnable) transformations
- Model quantifies uncertainty (optimization, safety)
- Include linear operators (ODEs, PDEs, shifts, boundaries)
- Sampling, conditioning and controlling of systems

Summary

Thx!
Questions?