Gaussian random fields in machine learning

Viacheslav Borovitskiy (Slava)

St. Petersburg State University
St. Petersburg Department of Steklov Mathematical Institute

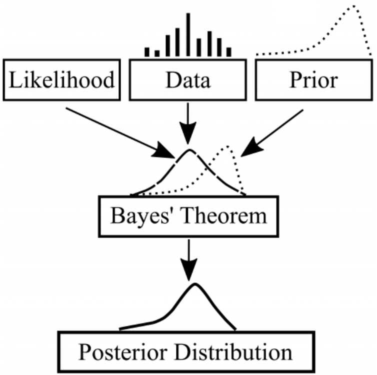
19 November 2020

Talk structure

Talk structure

Gaussian processes in machine learning

Bayesian learning paradigm:



Gaussian processes (GPs) — non-parametric prior over functions.

GPs are indeed useful

Bayesian Optimization in AlphaGo

Yutian Chen, Aja Huang, Ziyu Wang, Ioannis Antonoglou, Julian Schrittwieser, David Silver & Nando de Freitas

DeepMind, London, UK
`yutianc@google.com`

They used GPs to model target function and guide decision (optimization) process.

Talk structure

Gaussian process regression

GP — distribution over functions.

Bayesian inference for GPs:

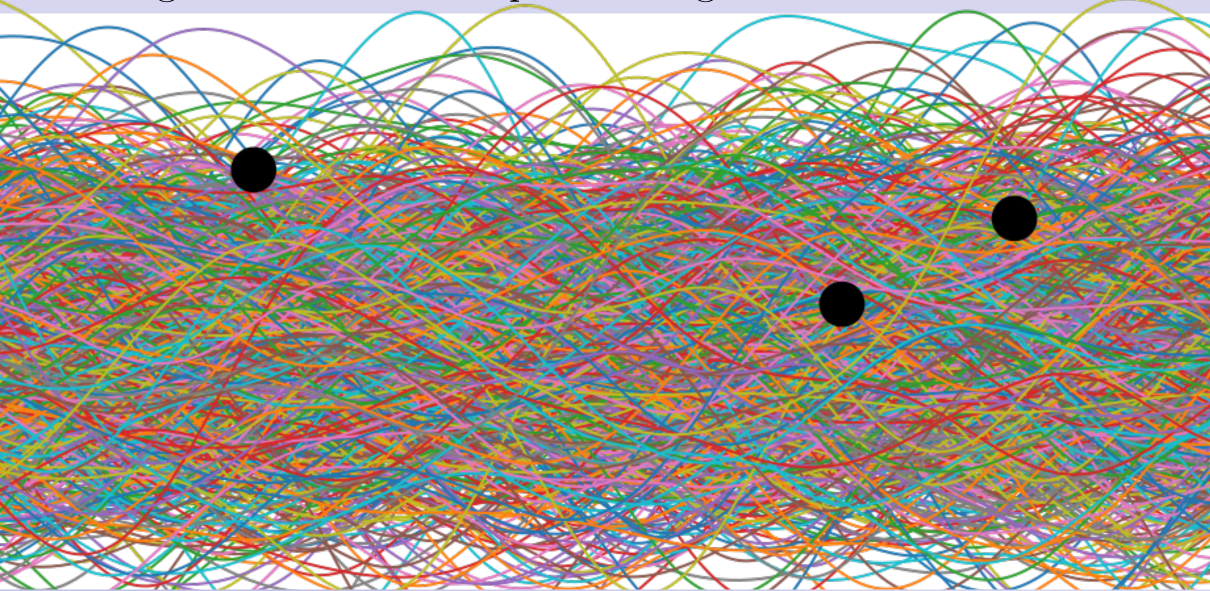
- prior: hand-picked GP
- data: noisy evaluations of the function
- likelihood: induced by Gaussian noise assumption
- posterior: another GP

Let us explore this visually . . .

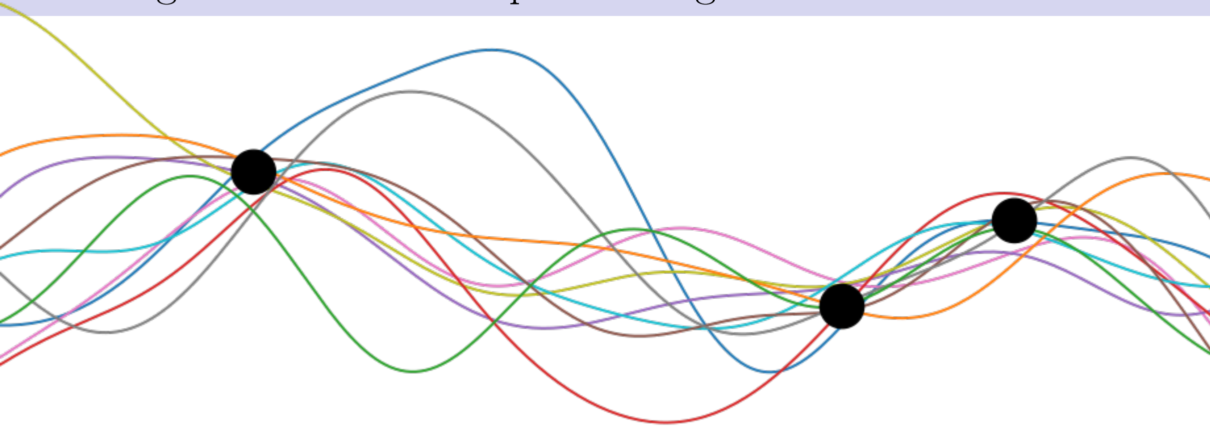
Visual guide to Gaussian process regression

Visual guide to Gaussian process regression

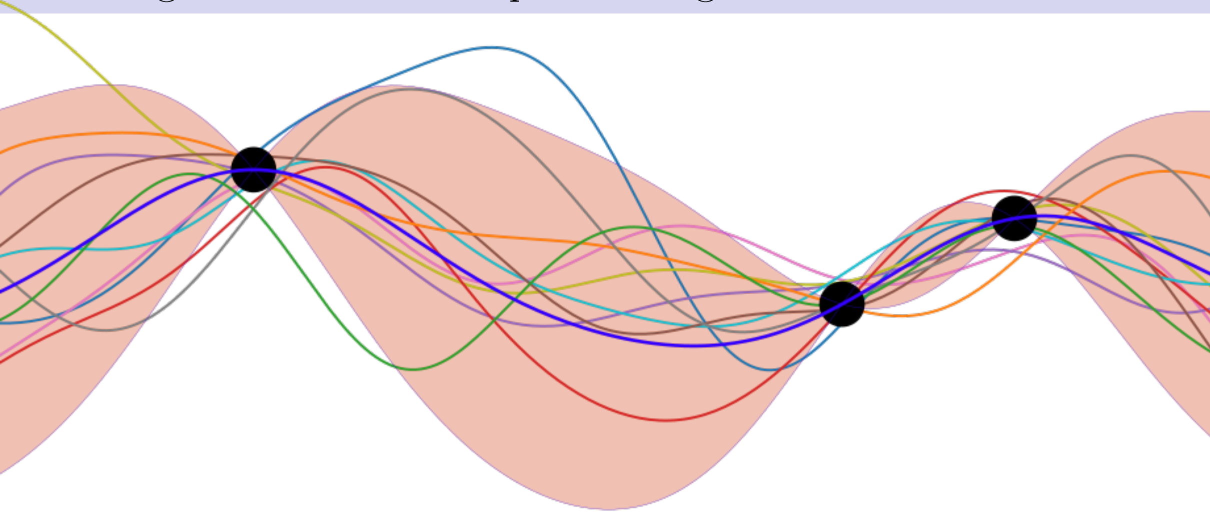
Visual guide to Gaussian process regression



Visual guide to Gaussian process regression



Visual guide to Gaussian process regression



What is a Gaussian process?

Gaussian random variable

- distribution over $\mathbb{R}$, denoted by $\mathrm{N}(\mu, \sigma^2)$,
- determined by two numbers: mean μ and variance σ^2.

Multivariate Gaussian random variable

- distribution over $\mathbb{R}^d$, denoted by $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,
- determined by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

Gaussian process

- distribution over functions from X to $\mathbb{R}$, denoted by $\mathrm{GP}(m, k)$,
- determined by two functions $m : X \to \mathbb{R}$ (mean) and $k : X \times X \to \mathbb{R}$ (covariance).

Gaussian processes are appealing in practice due to their simplicity (among other stochastic processes).

Bayesian inference for GPs

Bayesian inference for GPs <u>takes in</u>

- a prior distribution over functions of form $GP(m, k)$,
- noisy evaluations $y_1, .., y_n$ of the unknown function of interest at $x_1, .., x_n$.

and <u>returns</u> the distribution over functions of form

$$GP(\tilde{m}, \tilde{k}).$$

Given m and k, the functions $\tilde{m}$ and $\tilde{k}$ can be computed in a finite time. Specifically:

$$\tilde{m}(u) = m(u) + \mathbf{K}_{f(u)f(\boldsymbol{x})}\left(\mathbf{K}_{f(\boldsymbol{x})f(\boldsymbol{x})} + \sigma^2 I\right)^{-1}(\boldsymbol{y} - m(\boldsymbol{x}))$$

$$\tilde{k}(u, v) = k(u, v) - \underbrace{\mathbf{K}_{f(u)f(\boldsymbol{x})}}_{\text{vector } 1\times n} \underbrace{\left(\mathbf{K}_{f(\boldsymbol{x})f(\boldsymbol{x})} + \sigma^2 I\right)^{-1}}_{\text{matrix } n\times n} \underbrace{\mathbf{K}_{f(\boldsymbol{x})f(v)}}_{\text{vector } n\times 1}.$$

The Gaussian process regression algorithm

So how do we turn the data $(x_1, y_1), .., (x_n, y_n)$ into a reasonable stochastic model interpolating it?

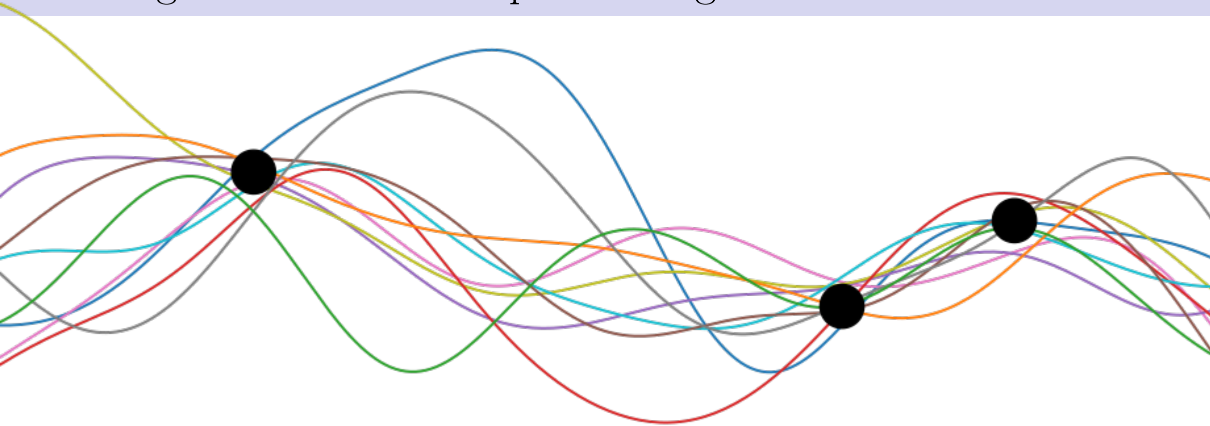
1. Come up with a parametric families $m_{\boldsymbol{\theta}}$ and $k_{\boldsymbol{\theta}}$ for prior mean and covariance functions.
2. Use maximum likelihood estimation to pick the optimal set of parameters $\boldsymbol{\theta}$ and the optimal noise value σ^2 from data $(x_1, y_1), .., (x_n, y_n)$.
3. Perform Bayesian inference with prior $GP(m_{\boldsymbol{\theta}}, k_{\boldsymbol{\theta}})$, data $(x_1, y_1), .., (x_n, y_n)$ and likelihood noise σ^2.
   As a result, obtain the posterior $\tilde{m}$ and $\tilde{k}$.
4. Use
   - $\mathrm{N}(\tilde{m}(u), \tilde{k}(u, u))$ as a stochastic prognosis at a new location u.
   - use samples of $GP(\tilde{m}, \tilde{k})$ as an ensemble of possible deterministic models.

Visual guide to Gaussian process regression

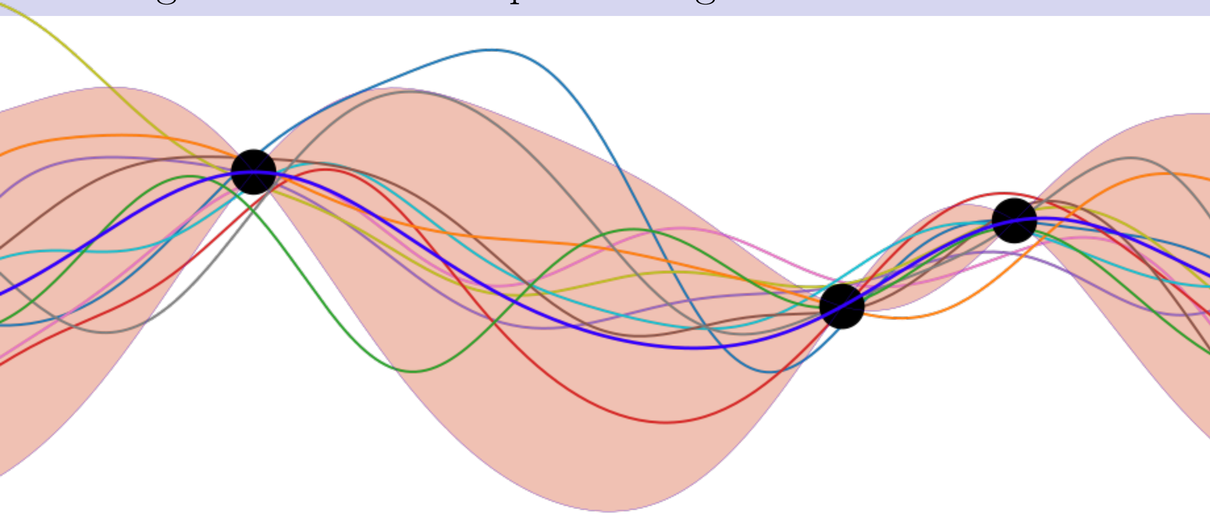
Visual guide to Gaussian process regression

Visual guide to Gaussian process regression

Visual guide to Gaussian process regression



Visual guide to Gaussian process regression

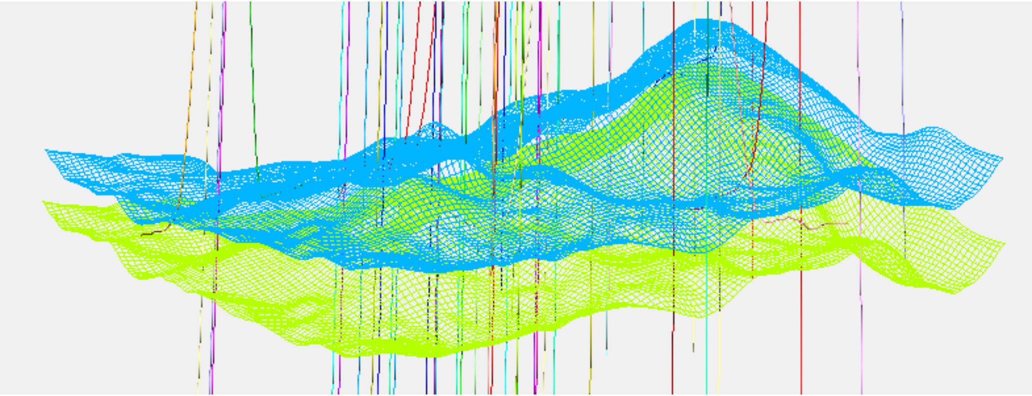


Talk structure

Geostatistical modeling of petroleum reservoirs

Problem: interpolate well data into the interwell space.

The data is very sparse, thus deterministic model is undesirable.

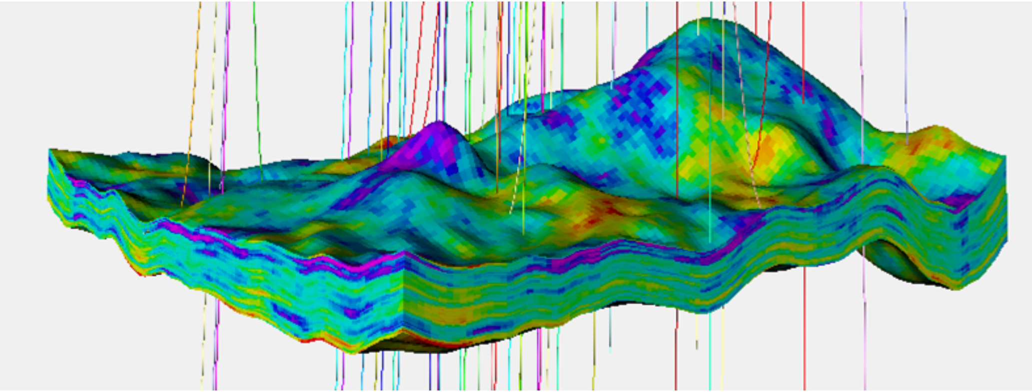


Reservoir structure, well locations.

Geostatistical modeling of petroleum reservoirs

Problem: interpolate well data into the interwell space.

The data is very sparse, thus deterministic model is undesirable.



A single sample of a Gaussian process model in the interwell space

Bayesian optimization of expensive black-box functions

Problem: minimize the target function $\phi : \mathbb{R}^d \to \mathbb{R}$.

At n'th step ϕ has already been evaluated at $x_1, .., x_n$. How do we choose x_{n+1}?

Build posterior GP f using data

$$x_1, .., x_n, \qquad\qquad \phi(x_1), .., \phi(x_n).$$

Choose

$$x_{n+1} = \arg\max_{x \in \mathbb{R}^d} \mathbb{P}(f(x) < \min_{i=1..n} \phi(x_i)). \qquad (MPI)$$

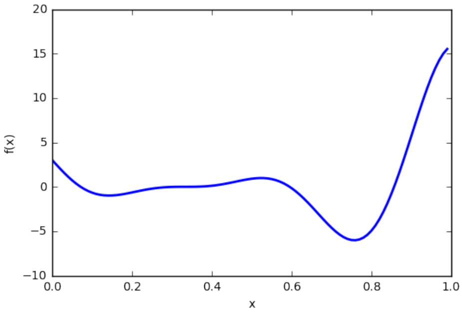
or

$$x_{n+1} = \arg\max_{x \in \mathbb{R}^d} \mathbb{E} \max(\min_{i=1..n} \phi(x_n) - f(x), 0). \qquad (EI)$$

Automatic exploration/exploitation trade-off.

Example

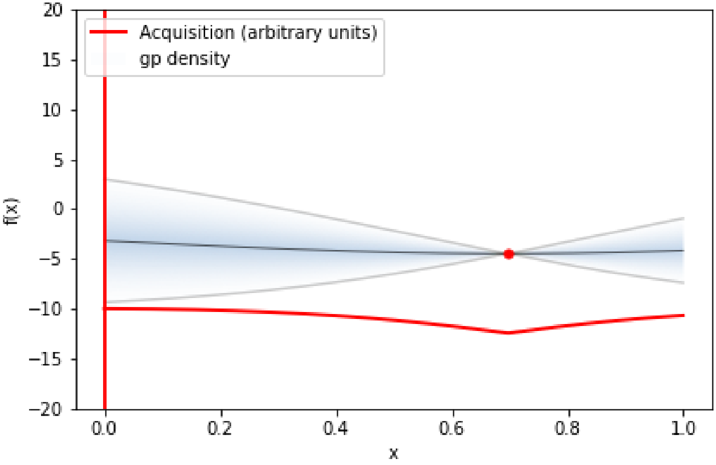
Let us minimize Forrester function $f(x) = (6x-2)^2 \sin(12x-4)$.



Choose some prior as $f_0 \sim \mathrm{GP}(?, ?)$.

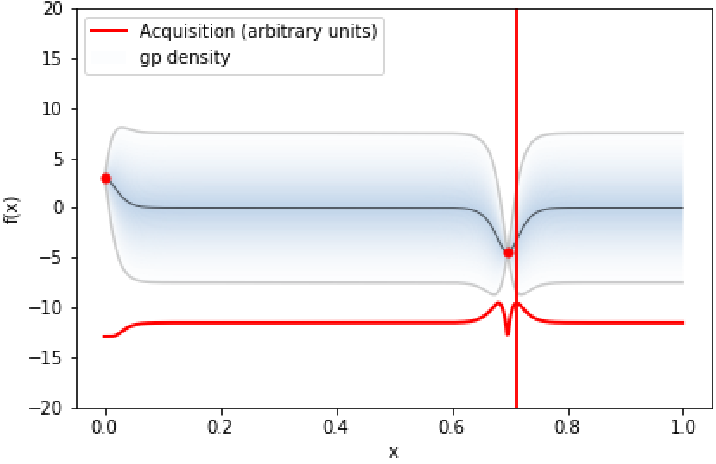
Example

Iteration 1.



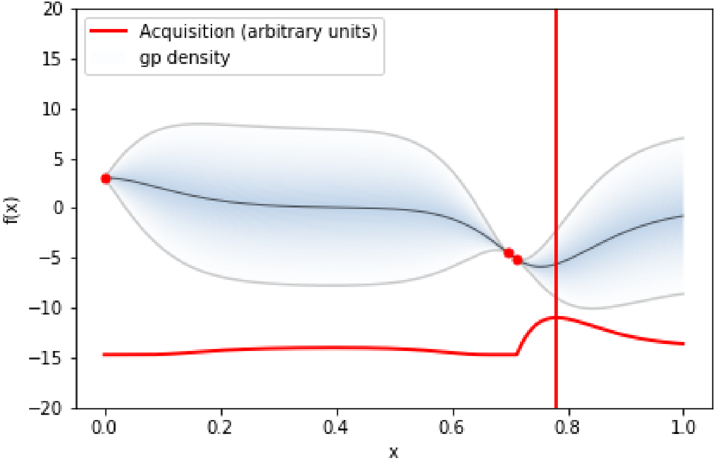
Example

Iteration 2.



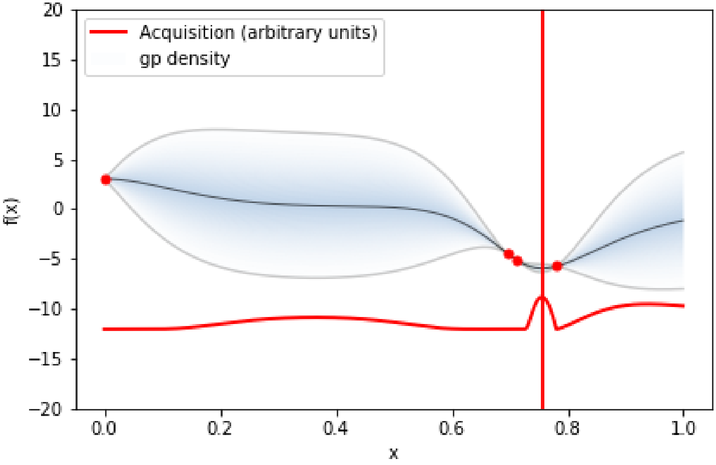
Example

Iteration 3.



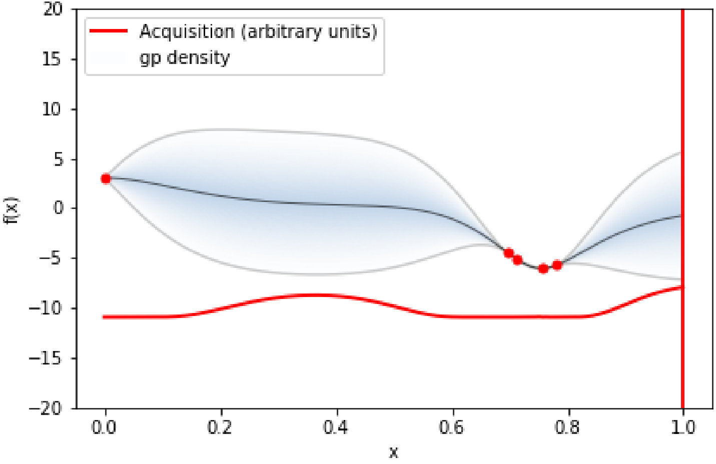
Example

Iteration 4.



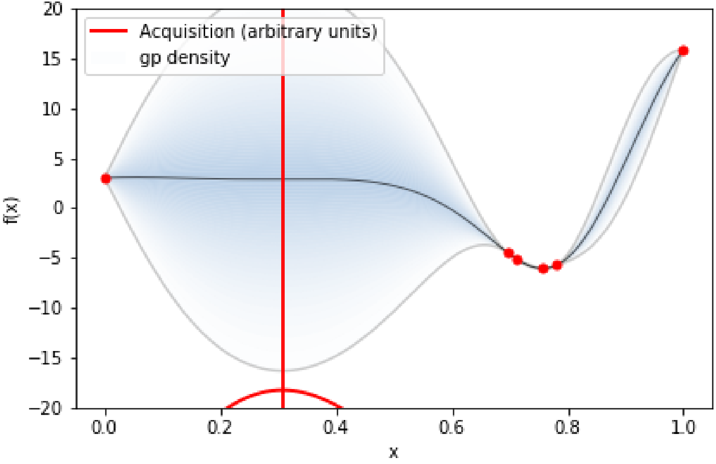
Example

Iteration 5.

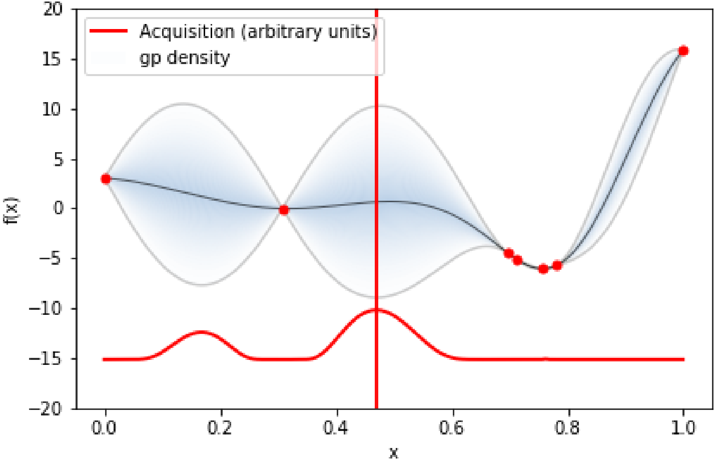


Example

Iteration 6.

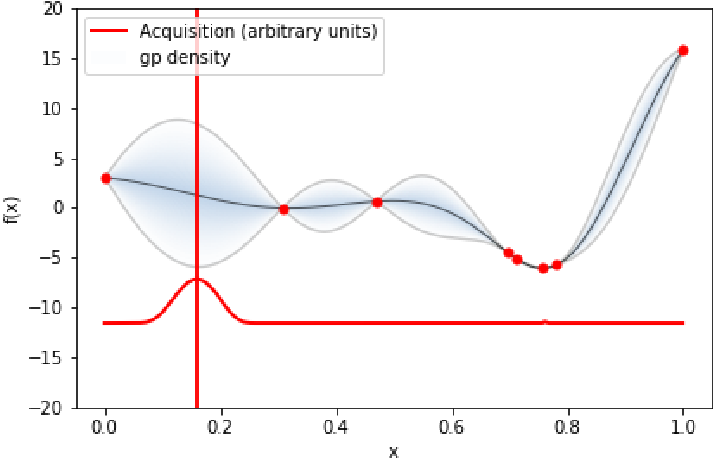


Iteration 7.


Acquisition (arbitrary units)
gp density
20
15
10
5
0
−5
−10
−15
−20
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
x

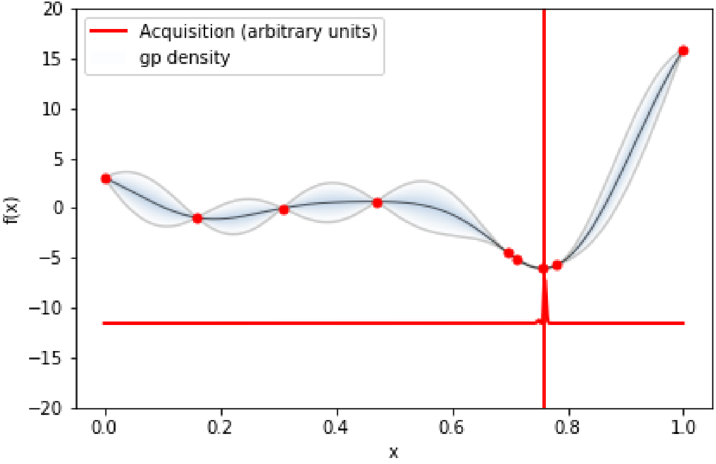
Example

Iteration 8.



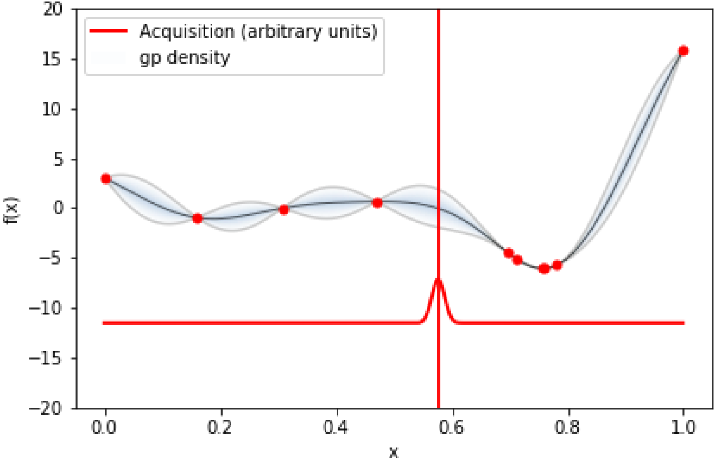
Example

Iteration 9.



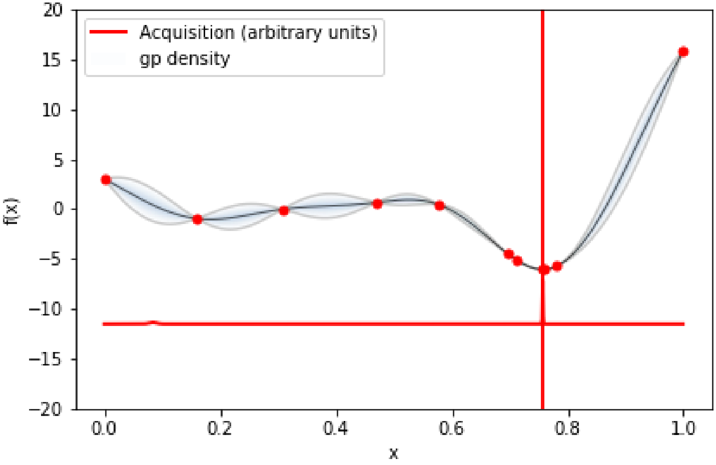
Example

Iteration 10.



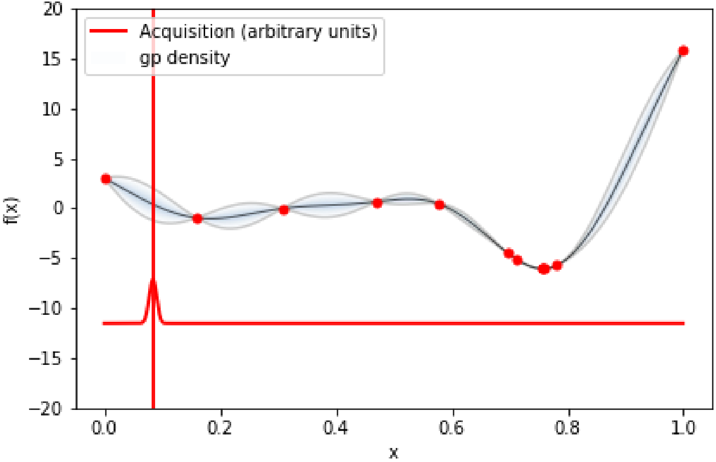
Example

Iteration 11.



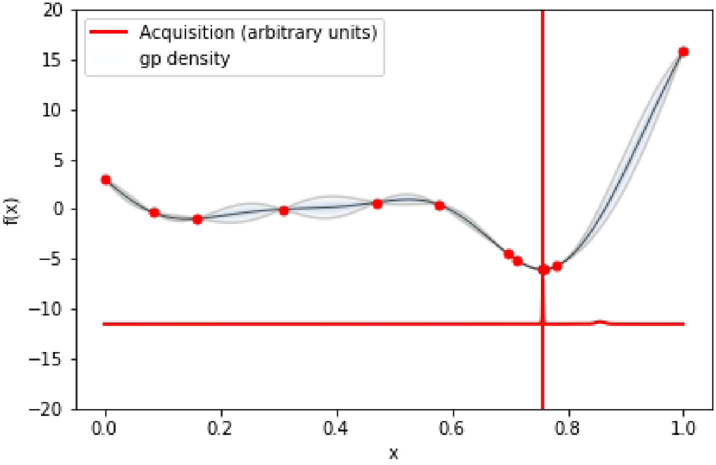
Example

Iteration 12.



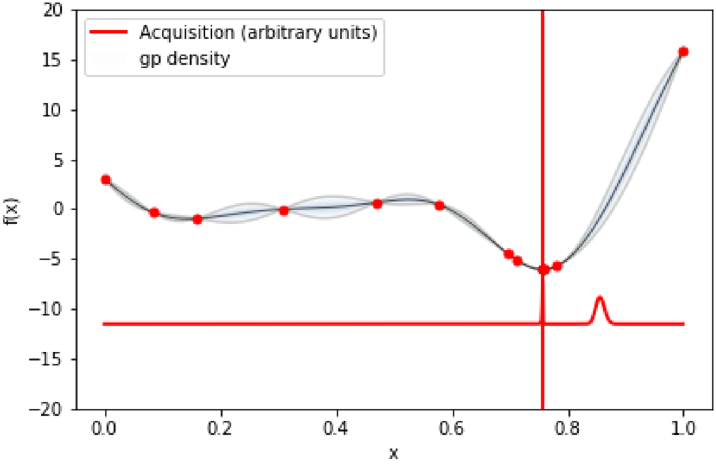
Example

Iteration 13.



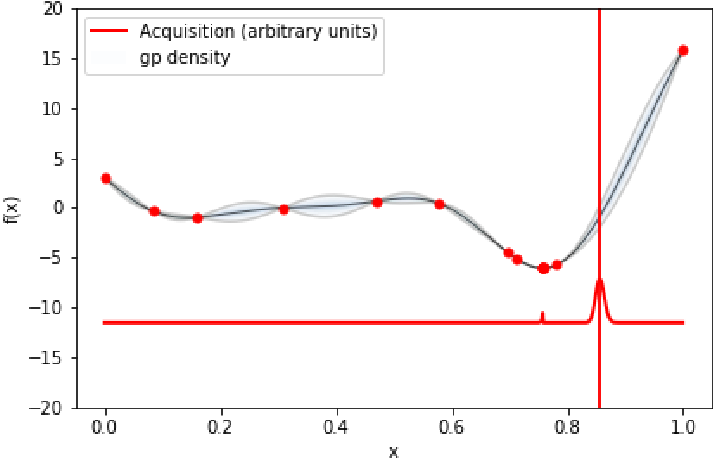
Example

Iteration 14.



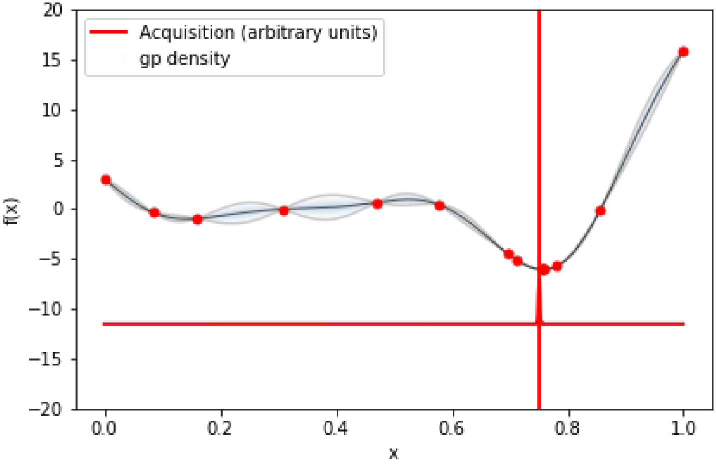
Example

Iteration 15.



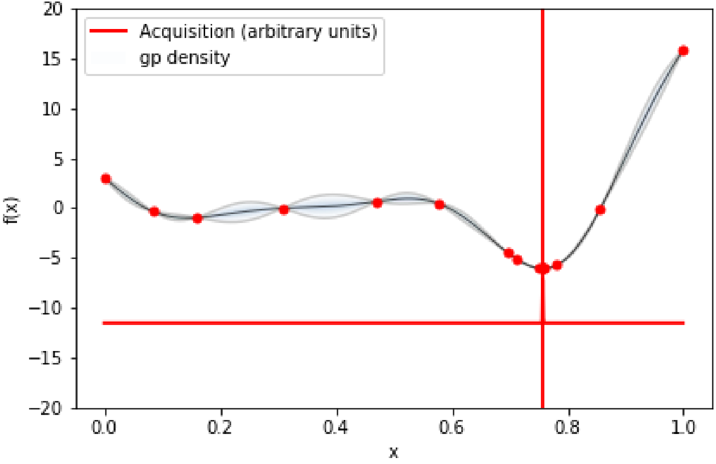
Example

Iteration 16.



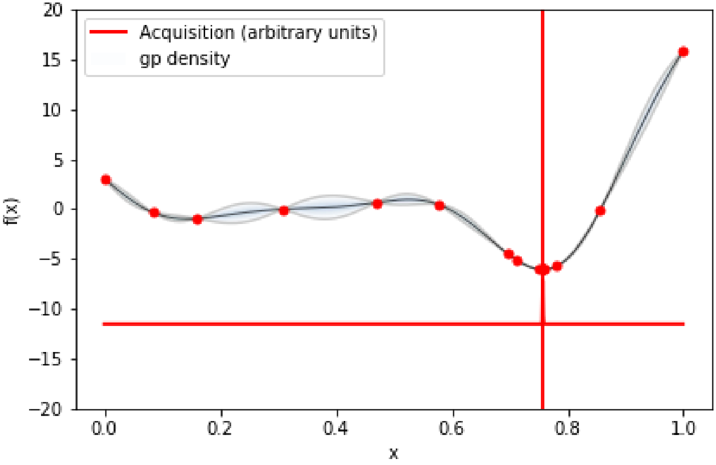
Example

Iteration 17.



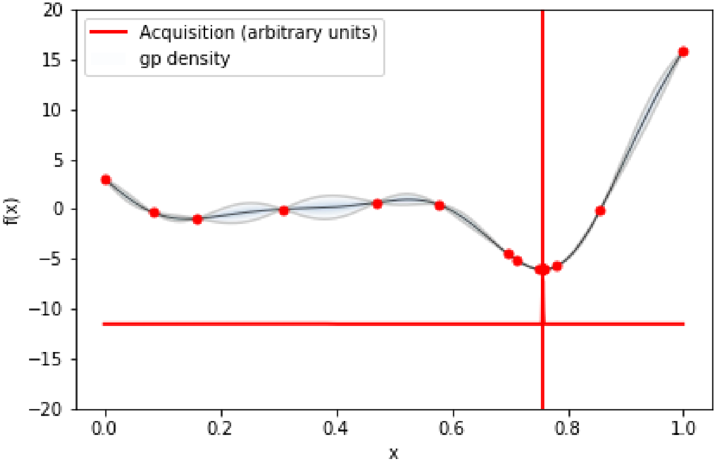
Example

Iteration 18.



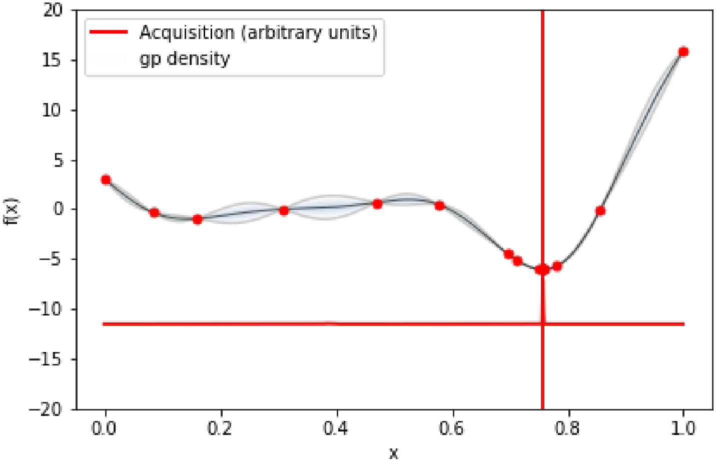
Example

Iteration 19.



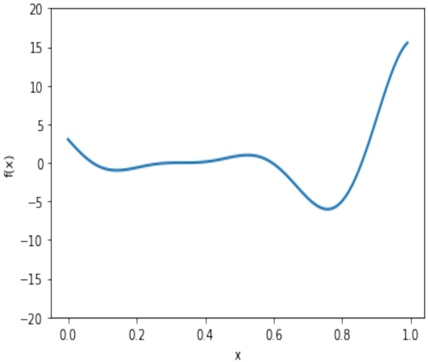
Example

Iteration 20.

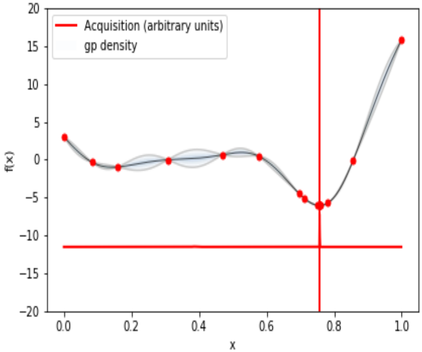


Example

Let us compare the model after 20 iterations with the target function.



(a) Target function



(b) GP model

Robotics and control

Classical control problem: physics is known, find optimal control.

Reinforcement learning control problem: physics is unknown, try to learn physics from data and on the go build the optimal control.

Second approach is supposed to bring us the cheap robots, for which

- we don't indeed know the physics (it deviates too much from the "ideal"),
- learning this physics by hand is of course possible, but it increases the price.

PILCO for robotics and control

PILCO (Probabilistic Inference for Learning COntrol) — an approach that uses GPs to model the unknown physics.

PILCO: A Model-Based and Data-Efficient Approach to Policy Search

---

Marc Peter Deisenroth MARC@CS.WASHINGTON.EDU
Department of Computer Science & Engineering, University of Washington, USA

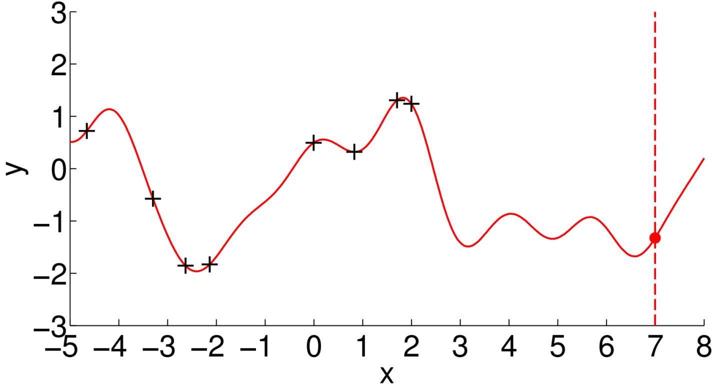
Carl Edward Rasmussen CER54@CAM.AC.UK
Department of Engineering, University of Cambridge, UK

The model can be described by $x_{t+1} = f(x_t, u_t) + w$, where

- x_t — trajectory,
- u_t — control,
- f models physics,
- $w \sim N(0, \sigma^2)$ — random noise.

PILCO for robotics and control

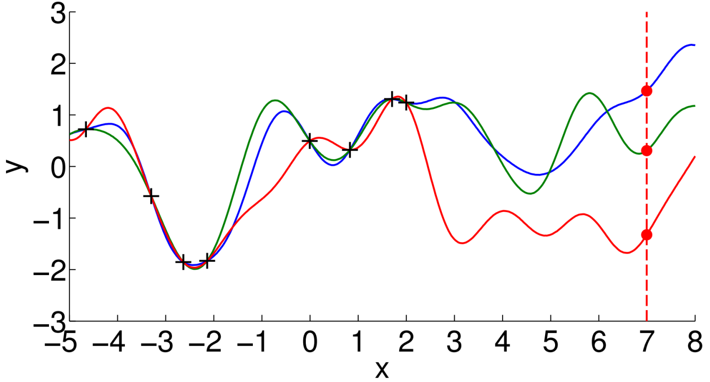
Imagine that f is modeled deterministically.



Consider a prognosis at $x = 7$.

PILCO for robotics and control

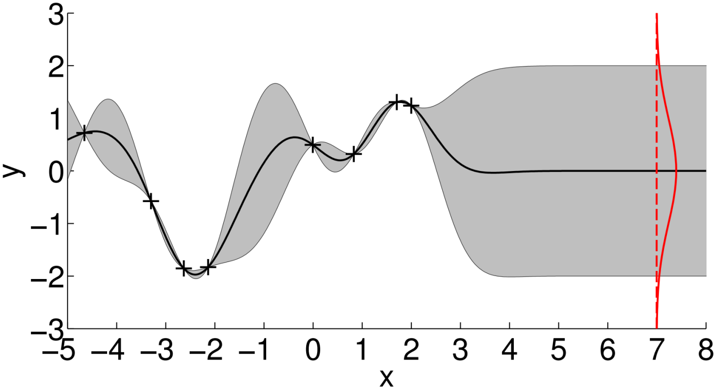
Imagine that f is modeled deterministically.



There exists a number of plausible models and thus a number of different predictions.

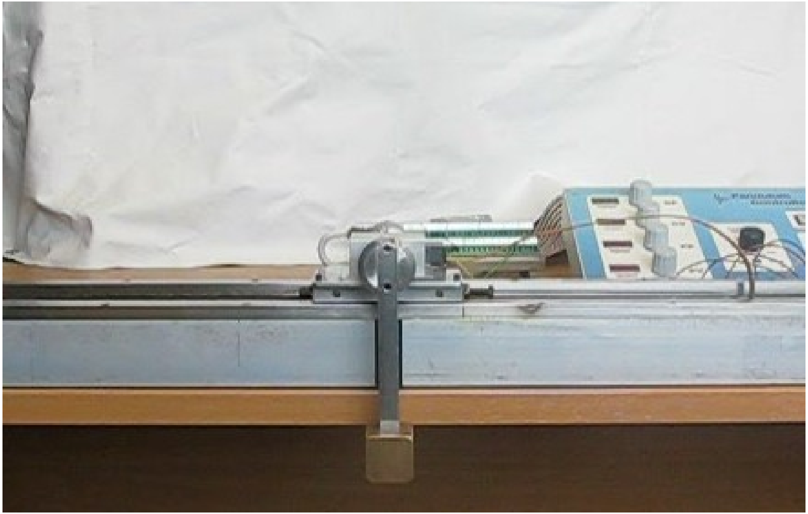
PILCO for robotics and control

What if we model f as a GP?

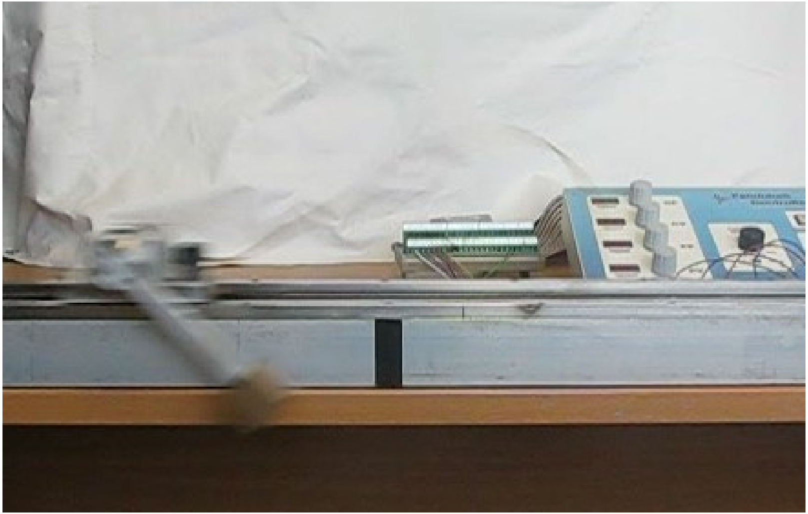


If we use GPs, we are able to use an infinite number of plausible models all at once.

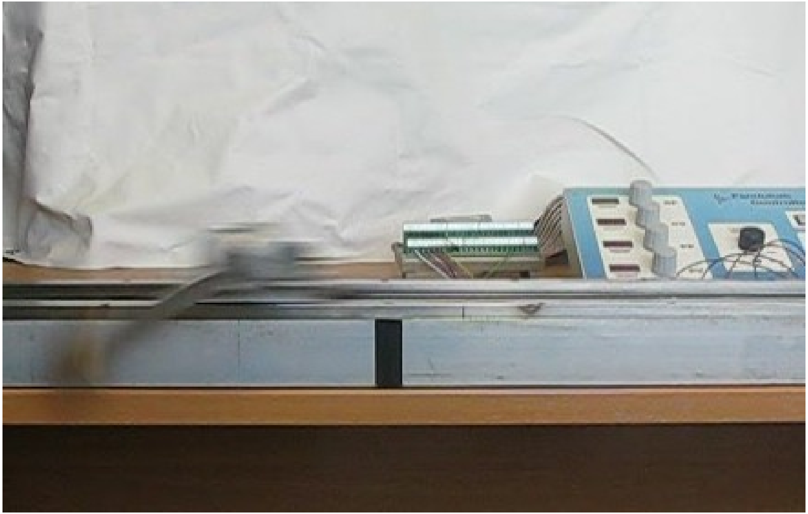
Example: learning to control a pendulum



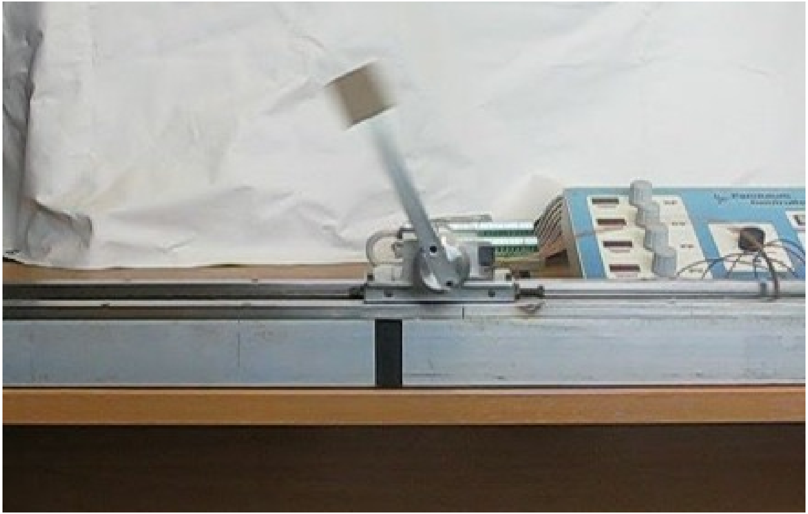
Example: learning to control a pendulum



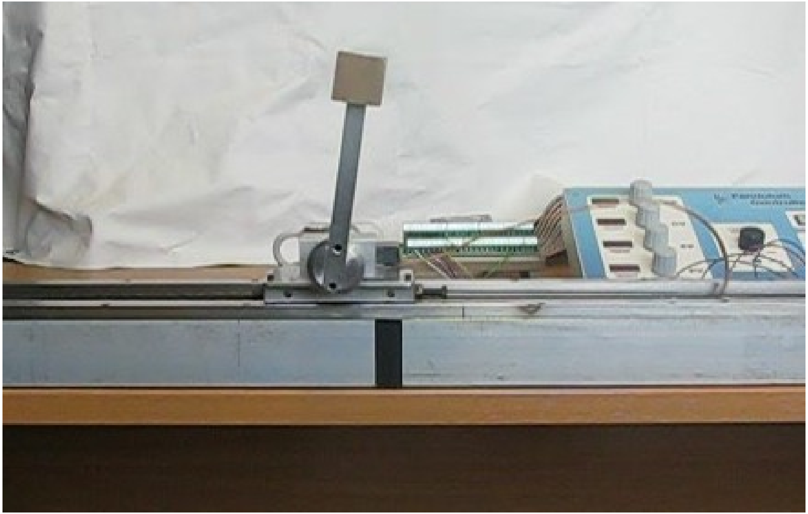
Example: learning to control a pendulum



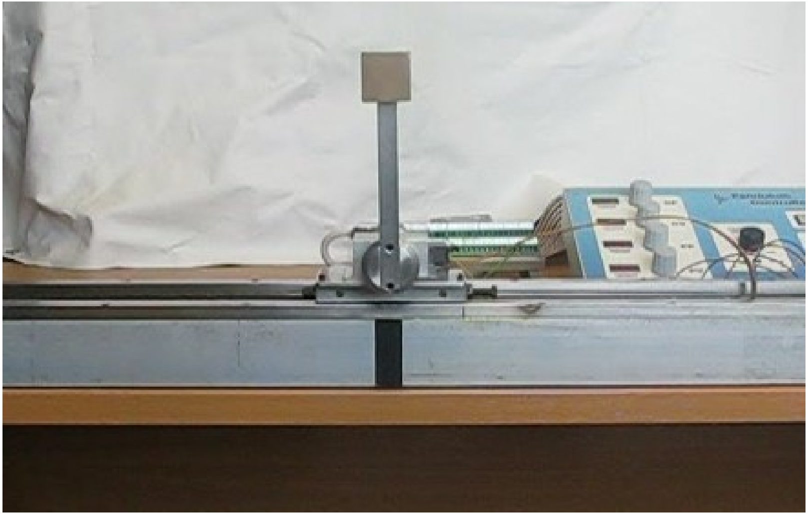
Example: learning to control a pendulum



Example: learning to control a pendulum



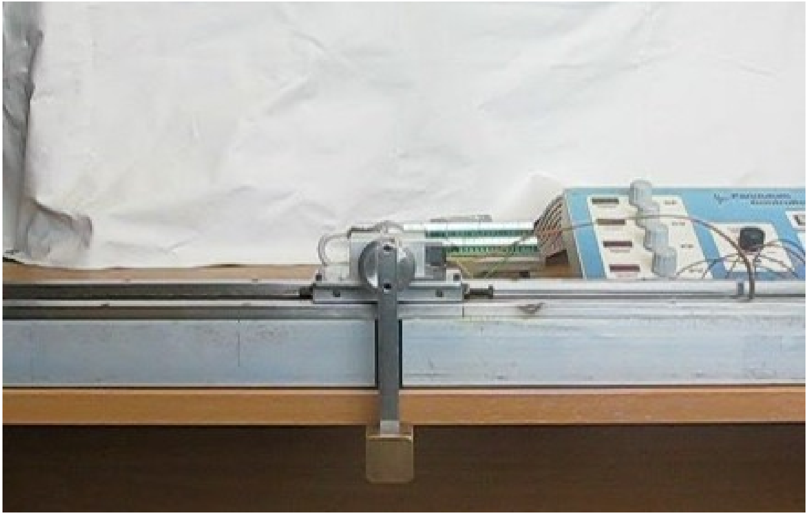
Example: learning to control a pendulum



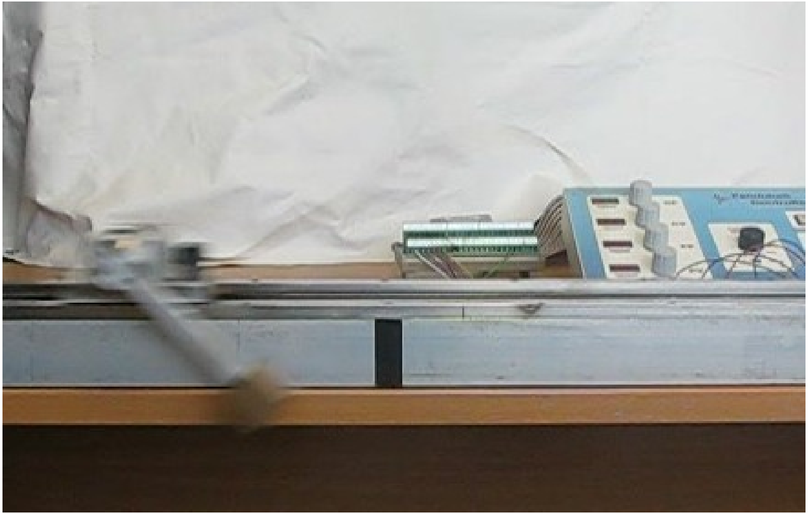
Example: learning to control a pendulum

Once more...

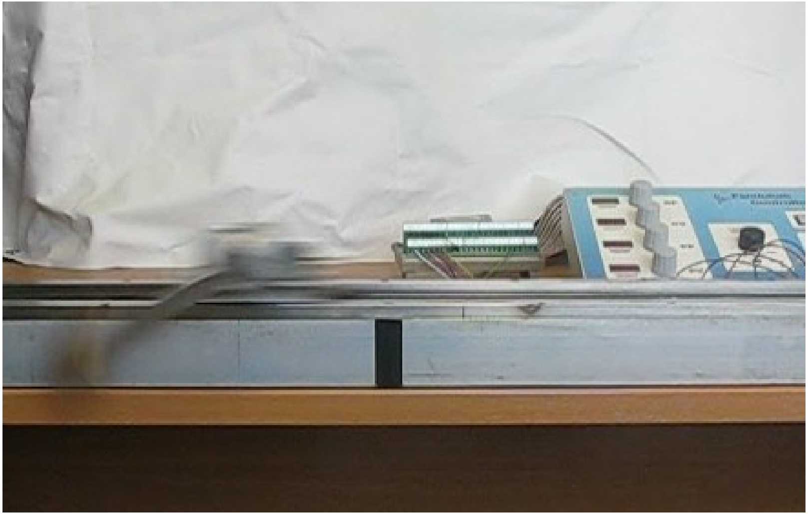
Example: learning to control a pendulum



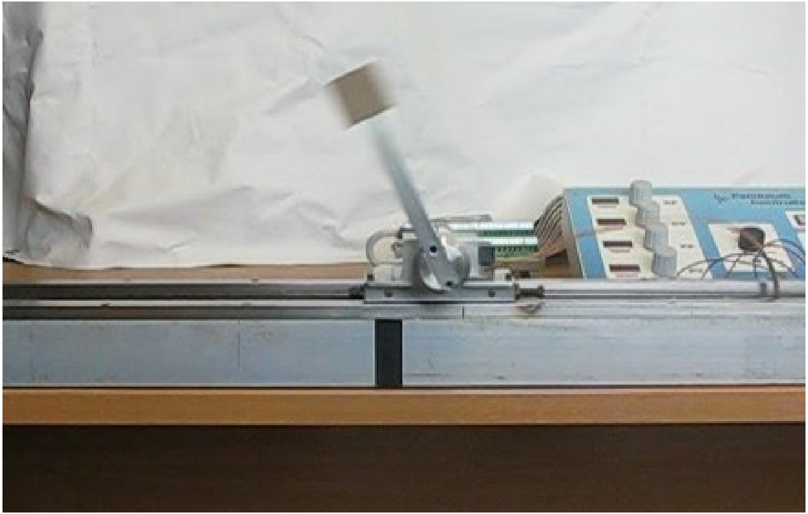
Example: learning to control a pendulum



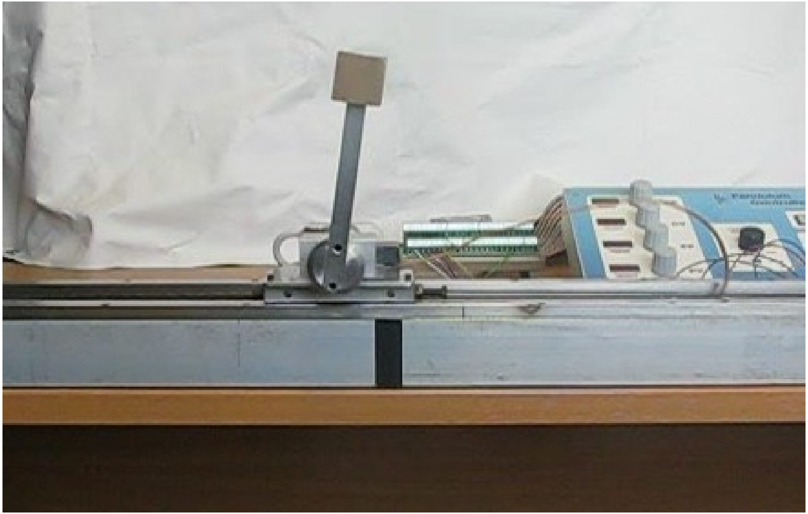
Example: learning to control a pendulum



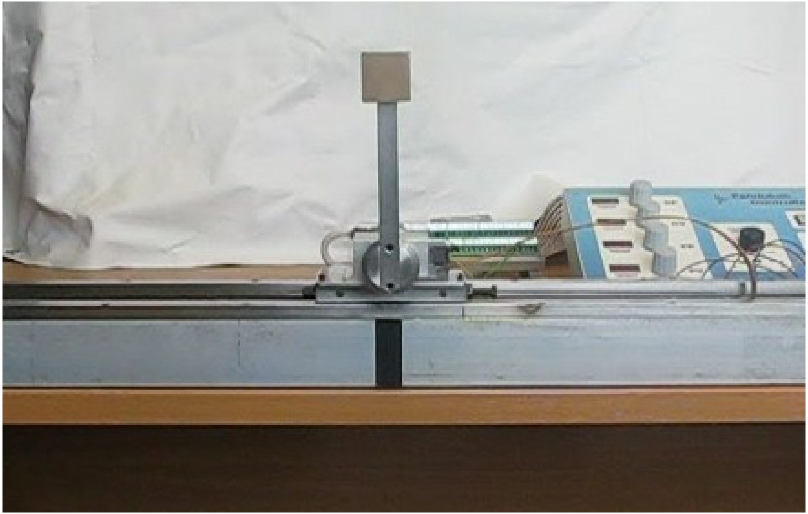
Example: learning to control a pendulum



Example: learning to control a pendulum



Example: learning to control a pendulum



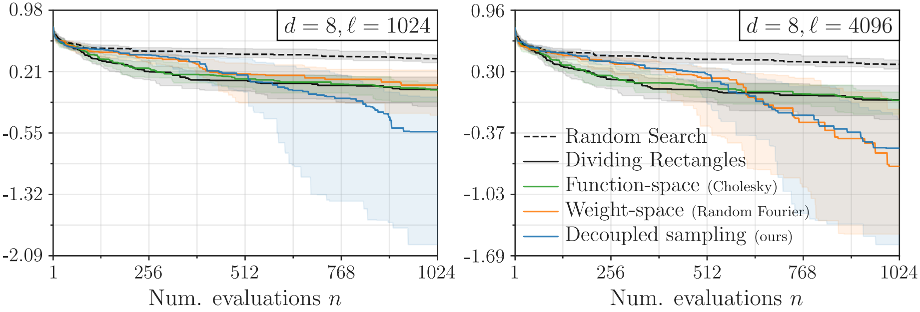
Talk structure

Efficient sampling from (approximate) posteriors

Efficiently Sampling Functions from Gaussian Process Posteriors

James T. Wilson [*1] **Viacheslav Borovitskiy** [*2 3] **Alexander Terenin** [*1]
Peter Mostowsky [*2] **Marc Peter Deisenroth** [4]

Outstanding Paper Honorable Mention Award at ICML 2020



✓ Improved performance owing to smaller error

Models in manifold setting

Matérn Gaussian processes on Riemannian manifolds

Viacheslav Borovitskiy[*1,4] Alexander Terenin[*2] Peter Mostowsky[*1] Marc Peter Deisenroth[3]

To be presented on NeurIPS 2020.

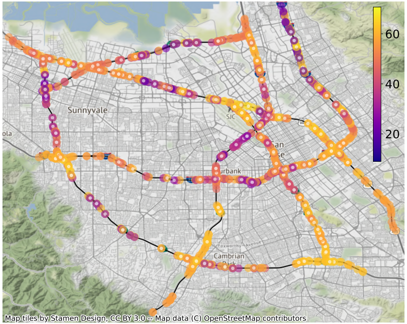
(a) Ground truth

(b) Posterior mean

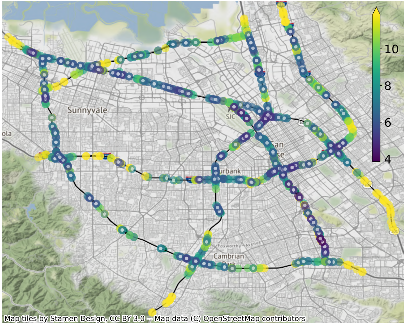
(c) Standard deviation

Matérn Gaussian Processes on Graphs

In review for AISTATS 2021



(a) Mean



(b) Standard deviation

Thank you for your attention!

viacheslav.borovitskiy@gmail.com

St Petersburg University

Mathematics & Computer Science department

Some figures were taken from: http://inverseprobability.com/talks/.