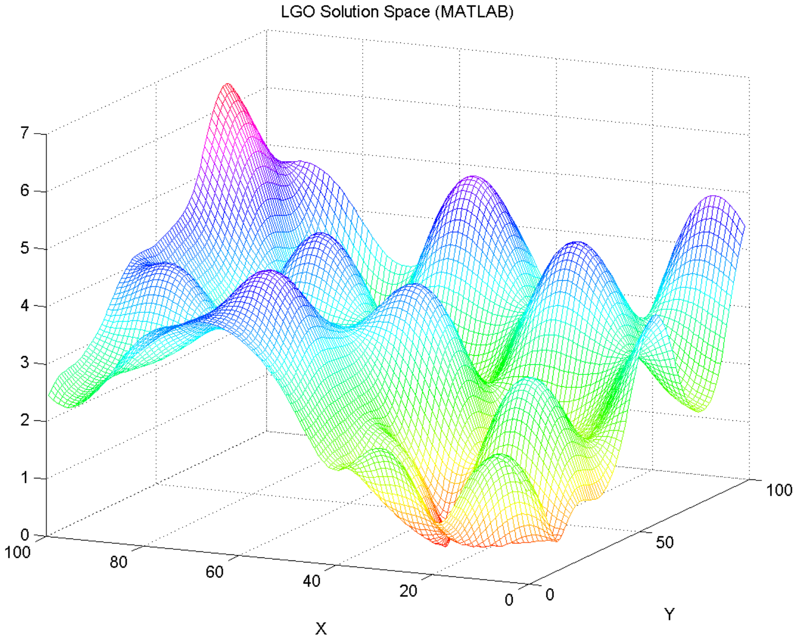
# Bayesian optimisation is probabilistic numerics

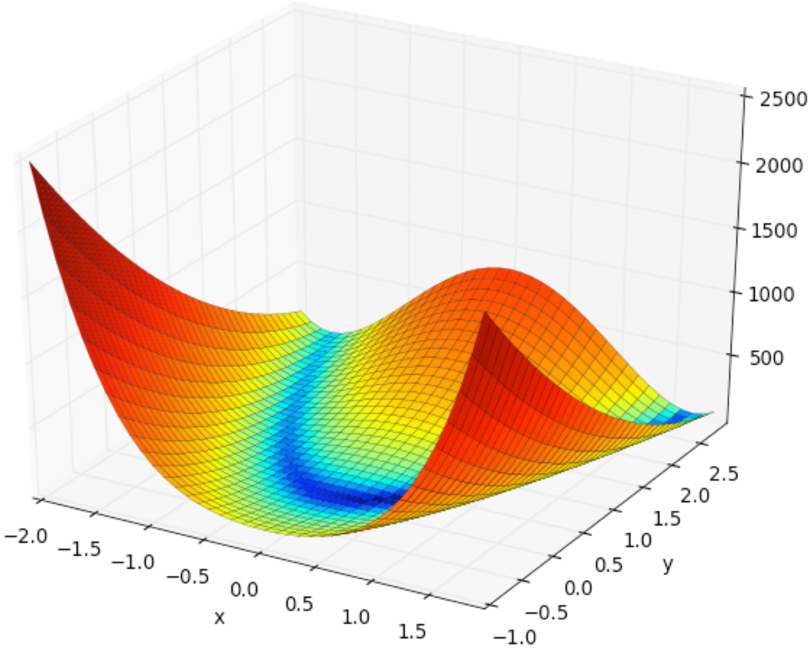
Mike Osborne @maosbot

# Global optimisation considers objective functions that are multi-modal and often expensive to evaluate.

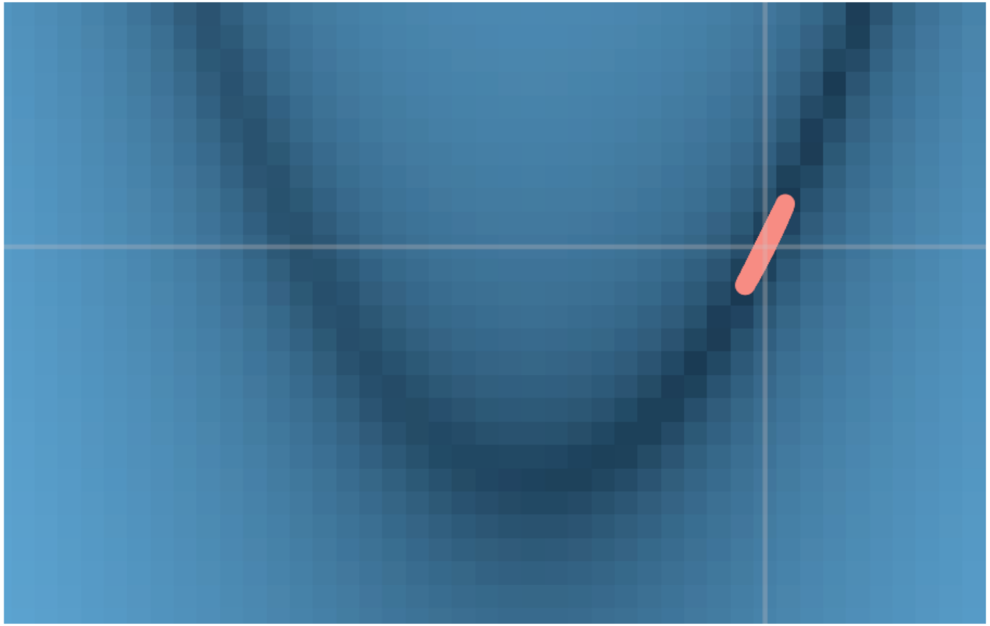


# The Rosenbrock is expressible in closed-form.

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2$$

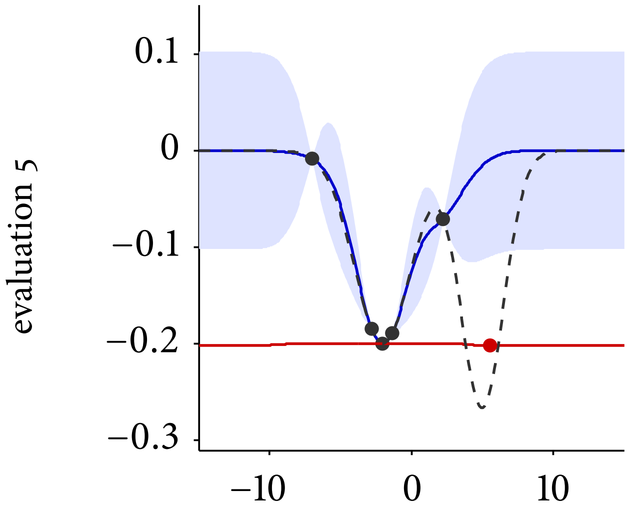


Computational limits form the core of the optimisation problem.



$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2$$

**Bayesian optimisation** is the approach of **probabilistically modelling** $f(x,y)$, and using decision theory to make **optimal use of computation.**



We define a loss function that is the lowest function value found after our algorithm ends.

Assuming that we have only one evaluation remaining, the loss of it returning value $y$, given that the current lowest value obtained is $\eta$, is

$$\lambda(y) \triangleq \begin{cases} y; & y < \eta \\ \eta; & y \geq \eta \end{cases}.$$

This loss function makes computing the expected loss simple: we'll take a myopic approximation and consider only the next evaluation.

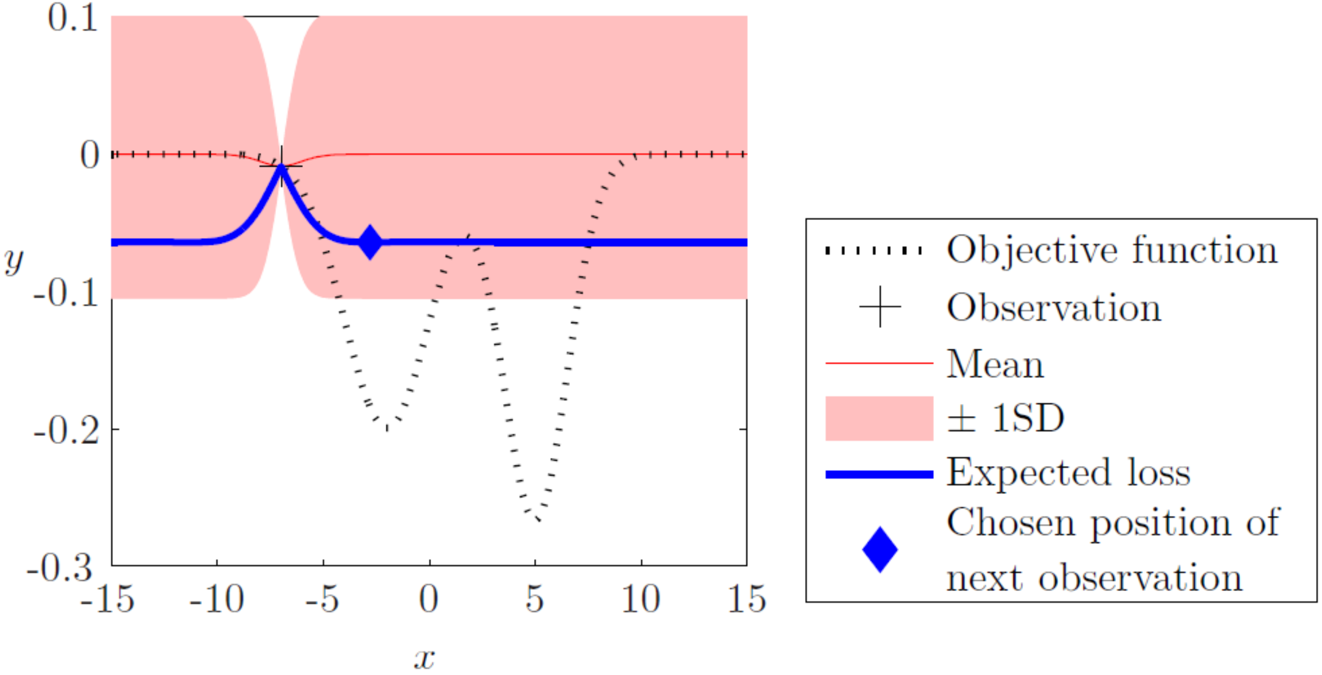
$$\boxed{\int \lambda(y)\, p(\, y \mid x,\, I_0\,)\ \mathrm{d}y}$$

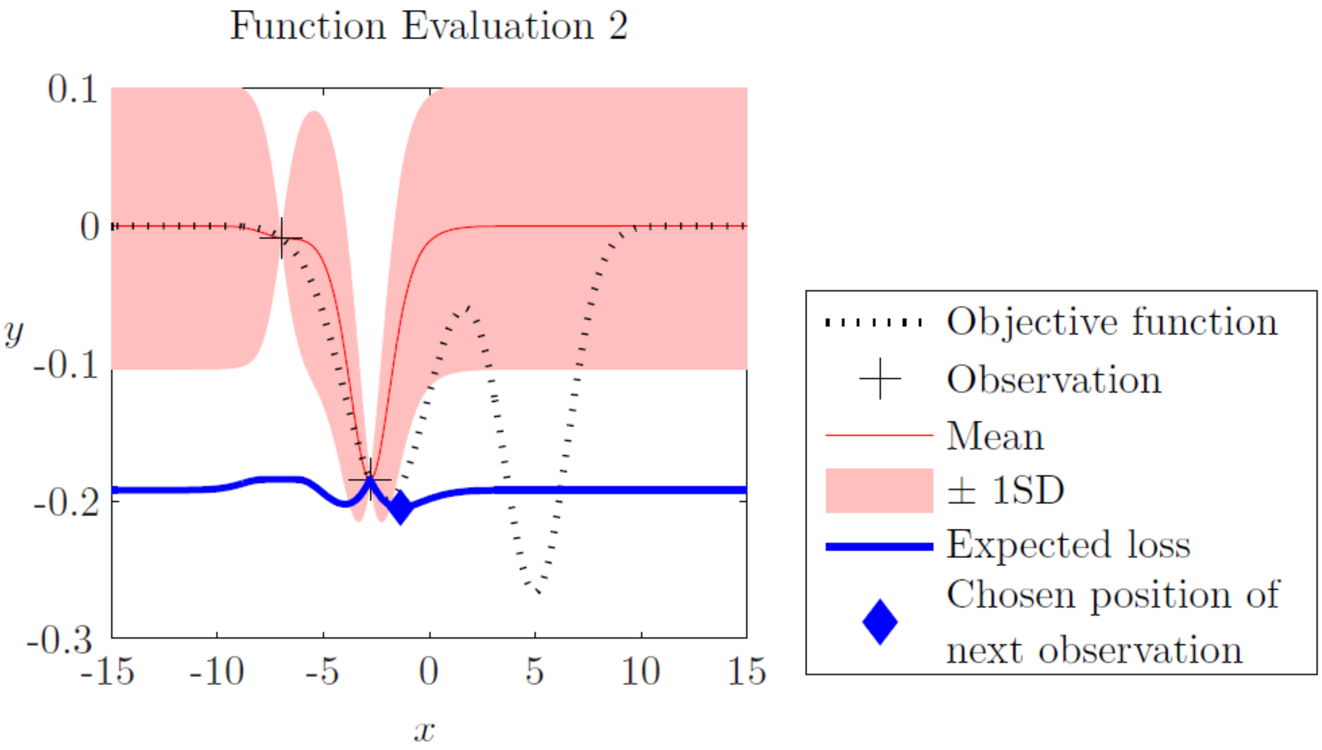
$I_0$ : All available information.

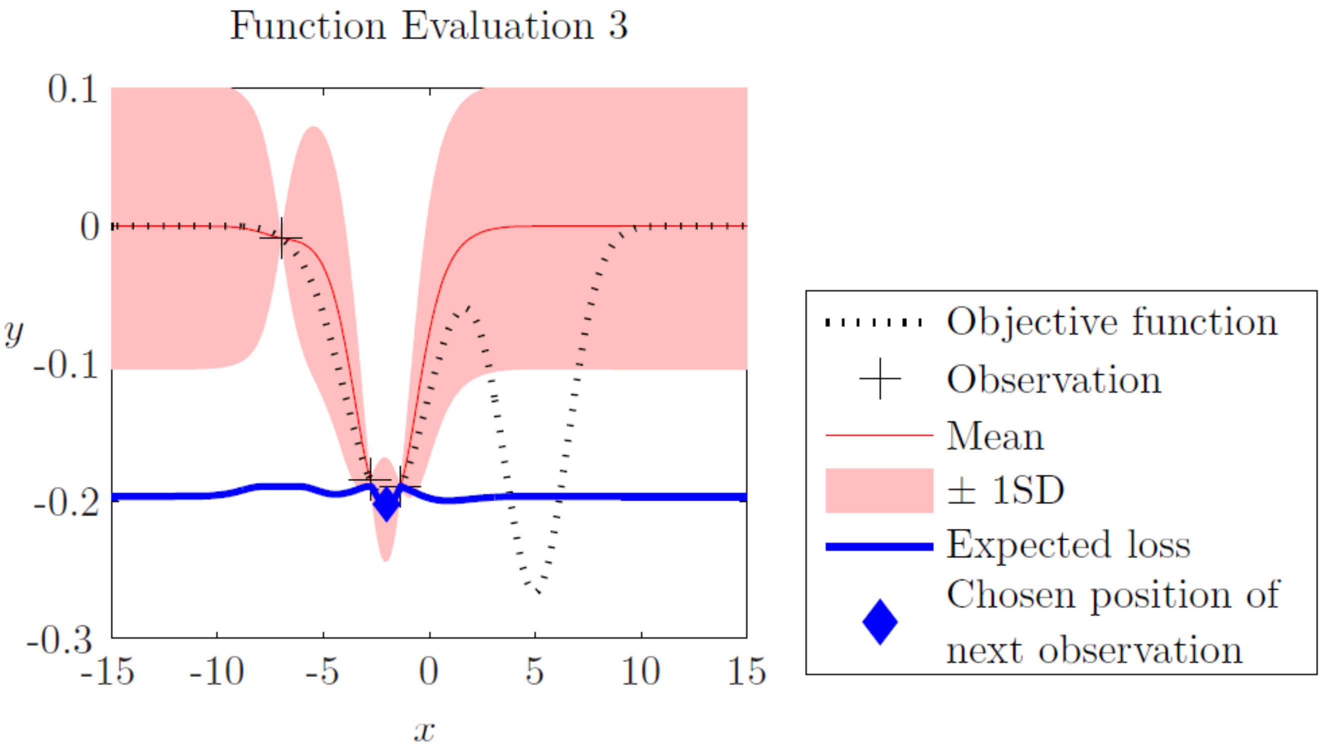
$x$ : Next evaluation location.

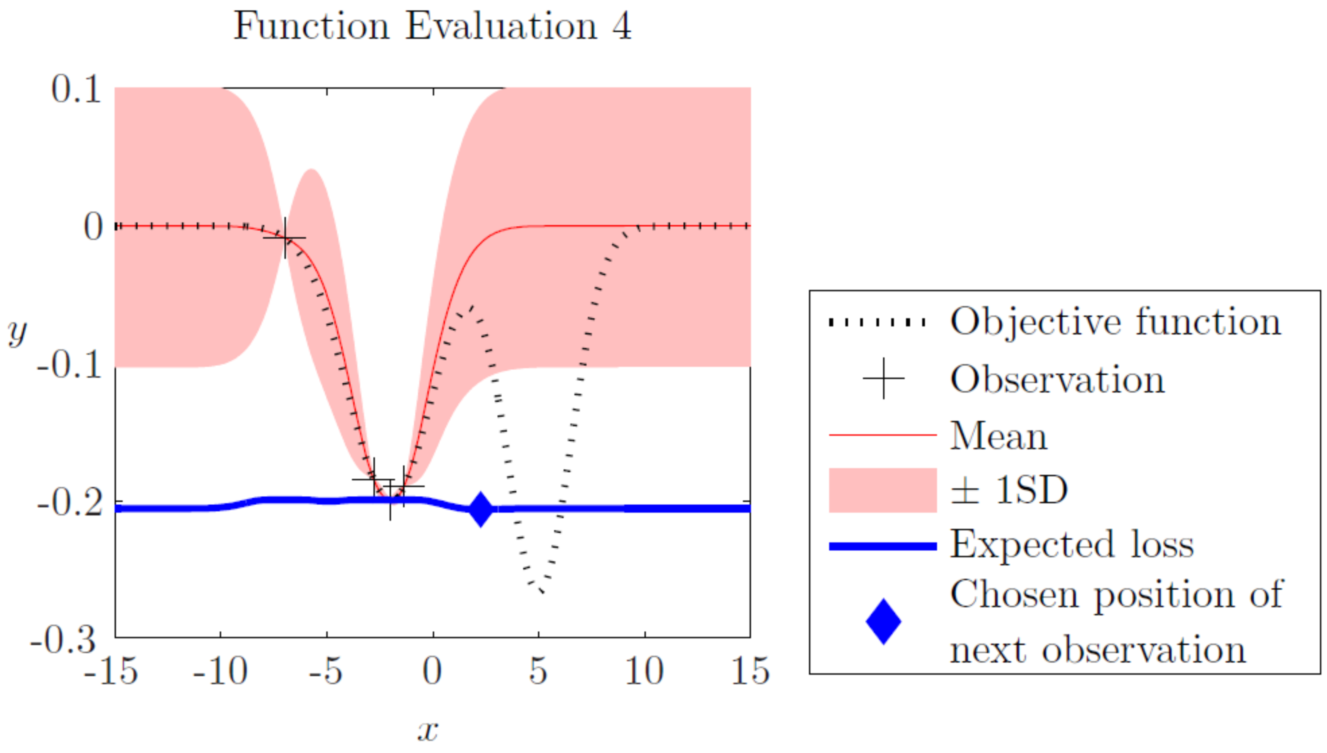
The expected loss is the expected lowest value of the function we've evaluated after the next evaluation.

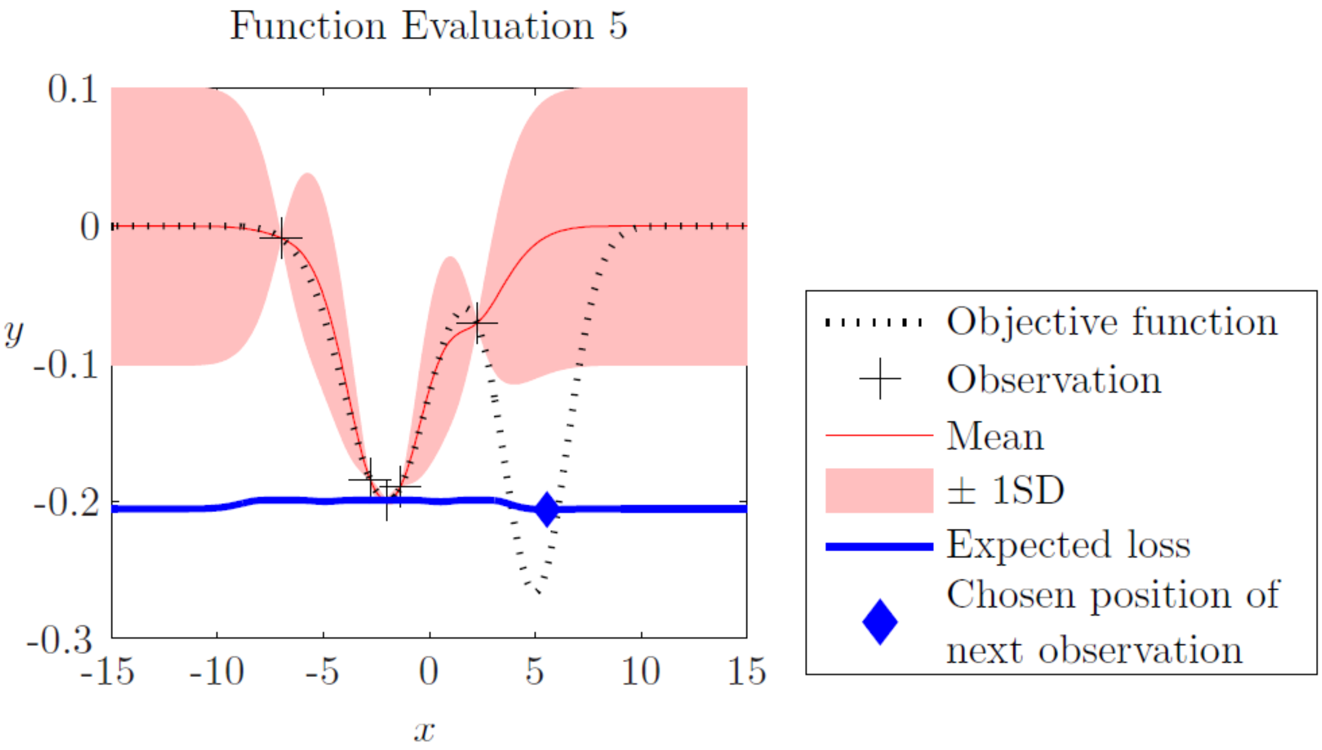
We choose a Gaussian process as the probability distribution for the objective function, giving a tractable expected loss.



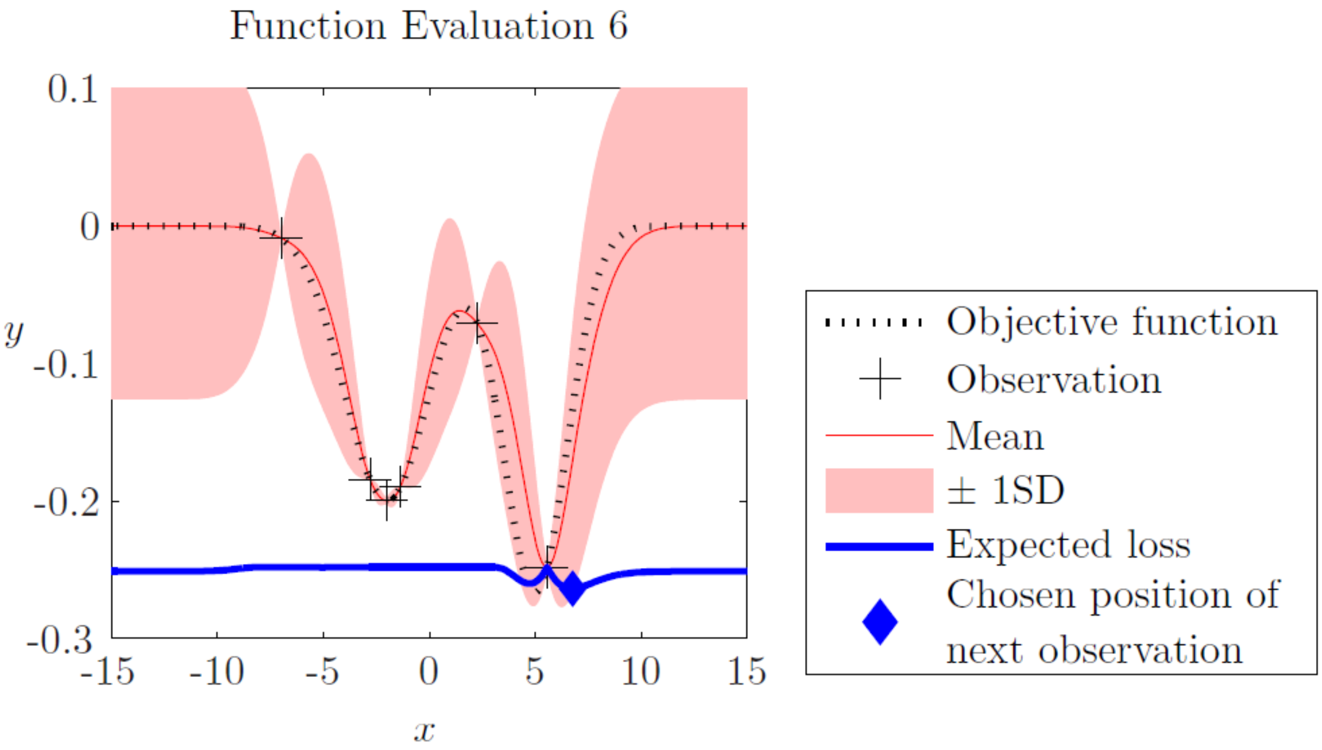

Function Evaluation 2
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of next observation


Function Evaluation 3
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of
next observation

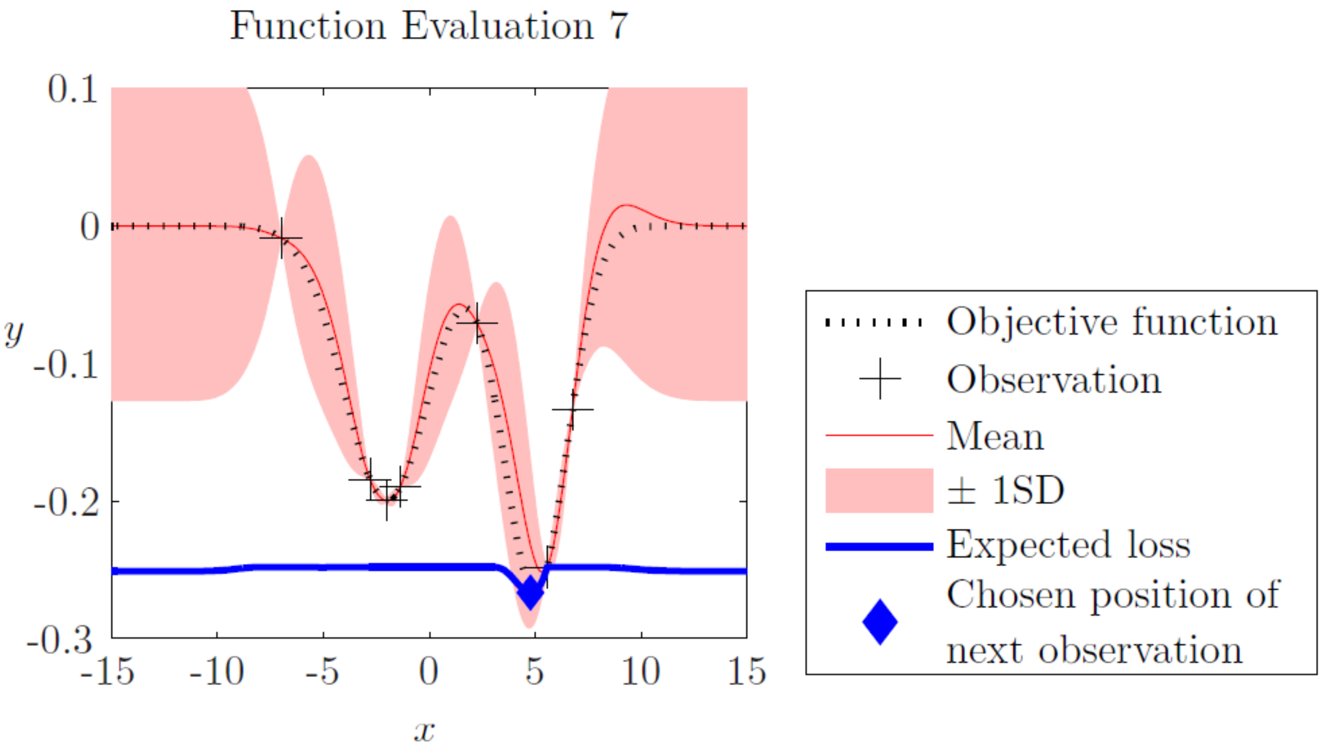

Function Evaluation 4
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of
next observation



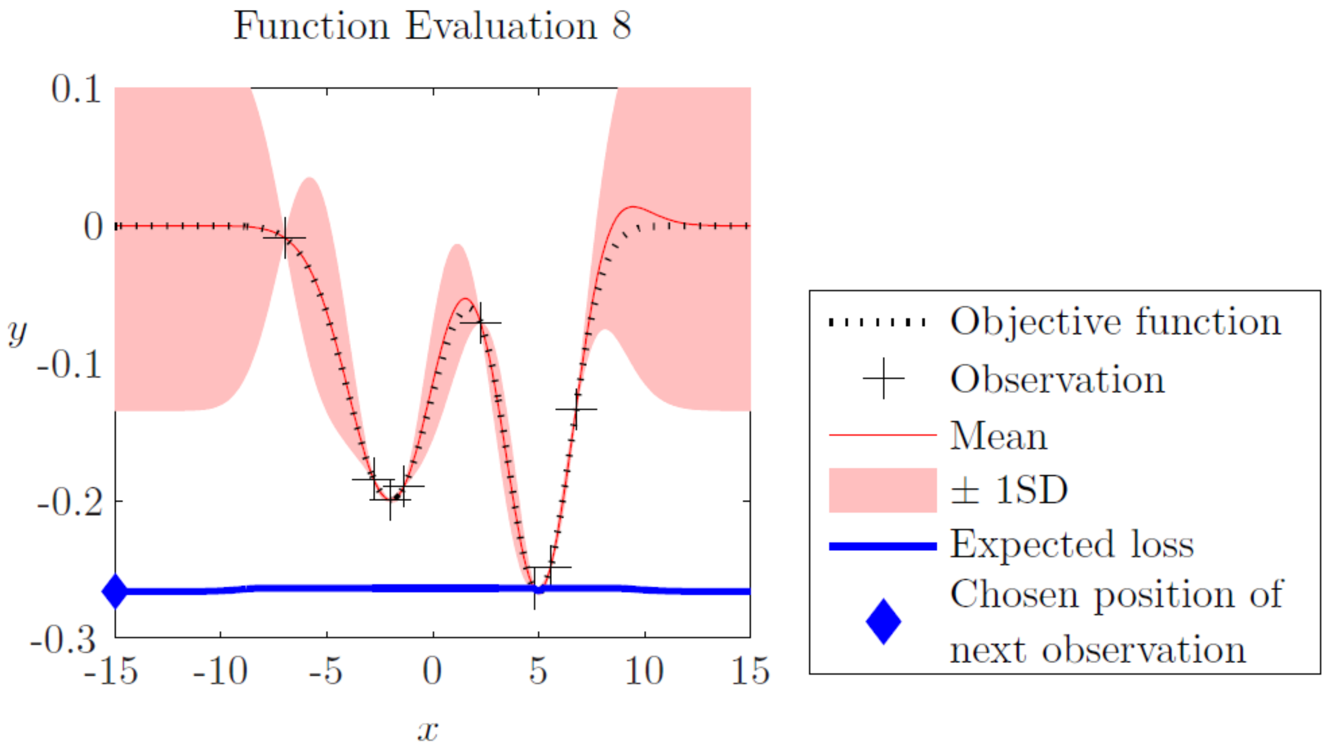
Function Evaluation 5
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of next observation



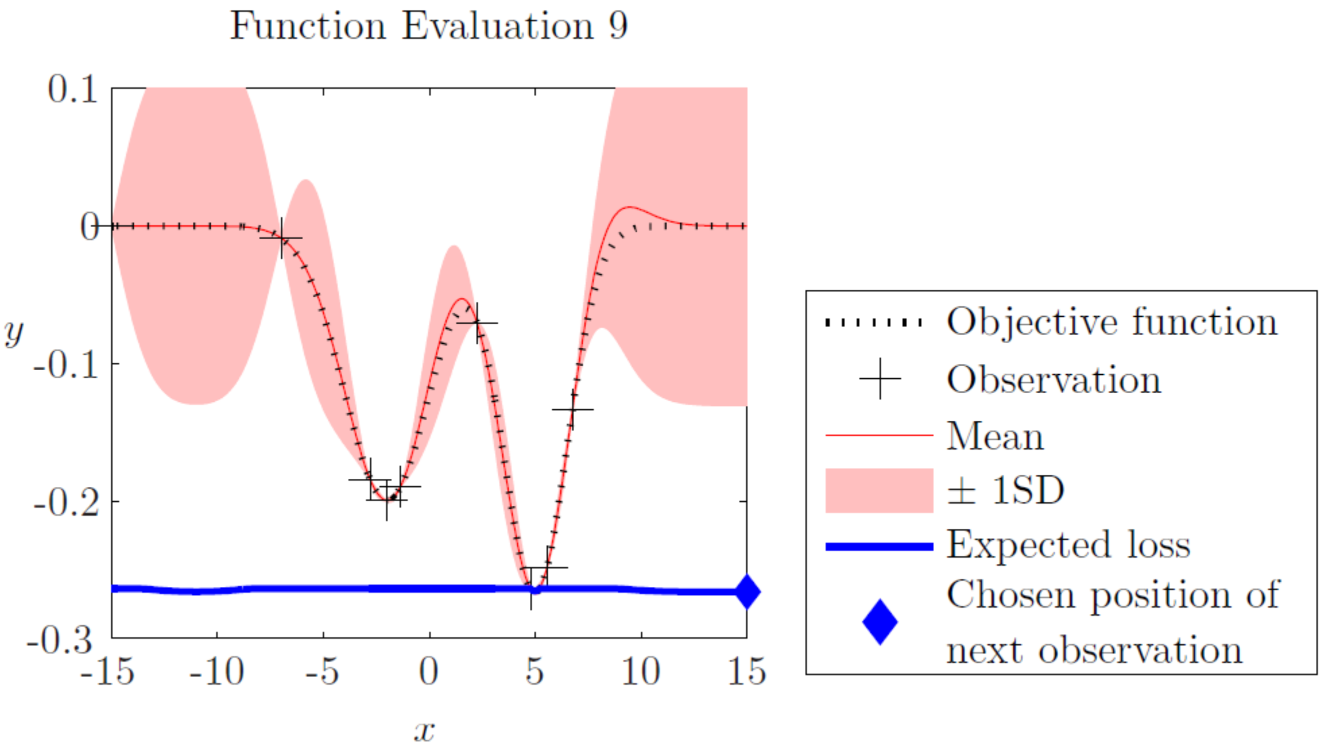
Function Evaluation 6
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of
next observation



Function Evaluation 7
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of
next observation

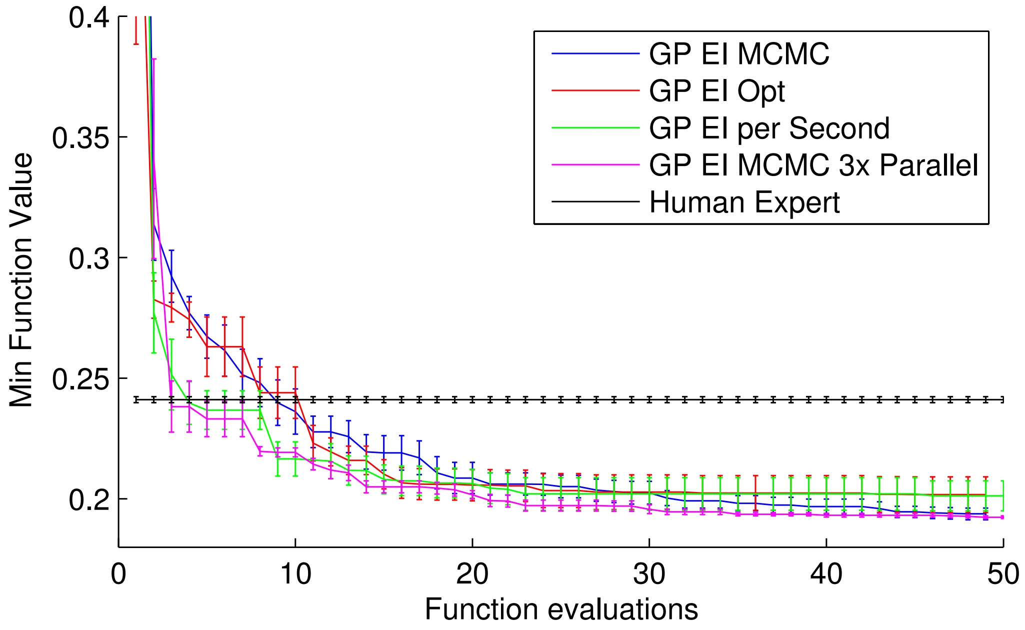


Function Evaluation 8
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of
next observation


Function Evaluation 9
0.1
0
-0.1
-0.2
-0.3
y
-15
-10
-5
0
5
10
15
x
Objective function
Observation
Mean
± 1SD
Expected loss
Chosen position of
next observation

Bayesian optimisation is the most impactful probabilistic numeric method.

# Snoek, Larochelle and Adams (2012) used Bayesian optimisation to tune expensive convolutional neural networks.



Previously, hyperparameters were often hand-tuned (!!): Bayesian optimisation hence offers automation.

No. of open positions for deep learning experts, according to Gartner: 41,000.

No. of deep learning experts, according to Yoshua Bengio: 50.

Practical **bayesian optimization** of machine learning algorithms

J Snoek, H Larochelle, RP Adams - Advances in neural information …, 2012 - papers.nips.cc

Abstract The use of machine learning algorithms frequently involves careful tuning of learning parameters and model hyperparameters. Unfortunately, this tuning is often a "black art" requiring expert experience, rules of thumb, or sometimes brute-force search. There is …

Cited by 1233 Related articles All 20 versions

[PDF] **Some Bayesian numerical analysis**

A O'Hagan - **Bayesian** statistics, 1992 - stat.duke.edu

'. SUMMARY Bayesian approaches to interpolation, quadrature and optimisation are discussed, based on representing prior information about the function in question in terms of a Gaussian process. Emphasis is placed on how different methods are appropriate when the …



[CITATION] **Bayesian numerical analysis**

P Diaconis - … decision theory and related topics IV, 1988 - Springer-Verlag, New York

# Bayesian optimisation has spawned many companies.

**Whetlab**

We make machine learning better and faster, automatically.

MIND FOUNDRY

Leading the way in in ML 2.0: machine learning becomes automated and operational

SIGOPT

Improve ML models 100x faster

**SigOpt's API tunes your model's parameters** through *state-of-the-art* Bayesian optimization.

- Exponentially faster and more accurate than grid search. **Faster, more stable, and easier to use than open source solutions.**
- **Extracts additional revenue and performance** left on the table by conventional tuning.

DataRobot

Become an AI-Driven Enterprise with Automated Machine Learning

# How did this happen?

# Firstly, numerics is central to machine learning performance.

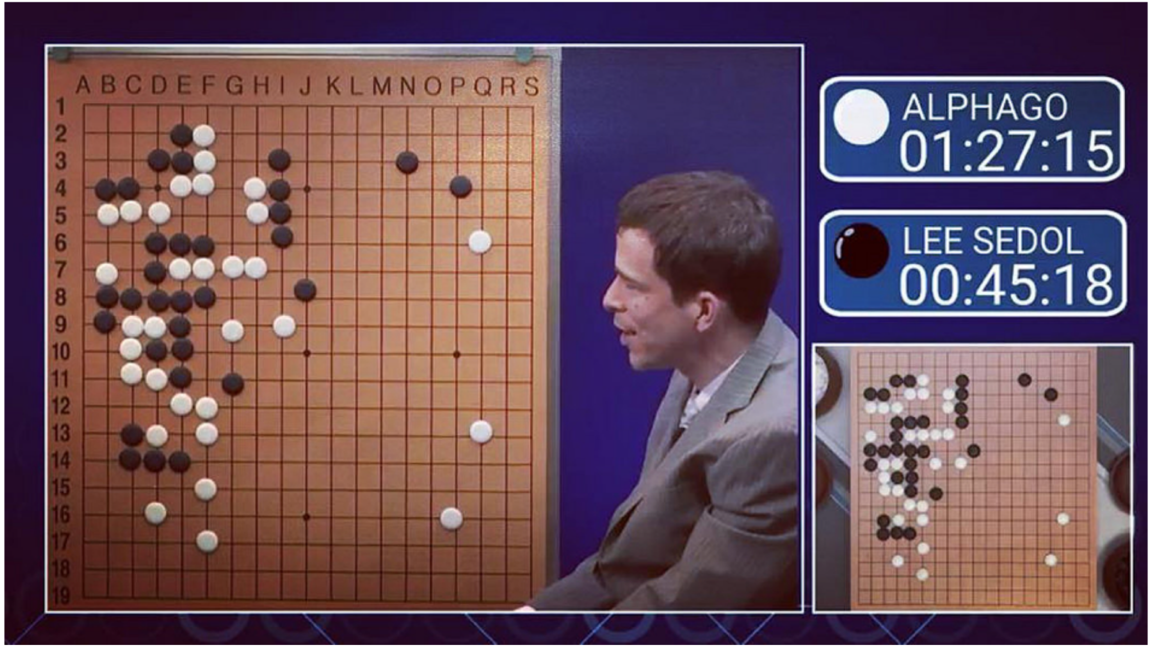
**Adam**: A method for stochastic optimization

DP Kingma, J Ba - arXiv preprint arXiv:1412.6980, 2014 - arxiv.org

… i=1 g1:T,i 2 ≤ dG∞ √T. Thus, limT→∞ R(T) T = 0. 5 RELATED WORK Optimization methods bearing a direct relation to **Adam** are **RMSProp** (Tieleman & Hinton, 2012; Graves, 2013) and AdaGrad (Duchi et al., 2011); these relationships are discussed below …

Cited by 7527 Related articles All 14 versions

## There have been real successes in machine learning enabled by Bayesian optimisation.




Source: Silver et al (2017); Melis, Dyer, and Blunsom (2017).

# Secondly, Bayesian optimisation built a community, agreed on goals and provided open-source software.

## BayesOpt 2017

NIPS Workshop on Bayesian Optimization
December 9, 2017
Long Beach, USA

Home
Schedule
Accepted Papers
Past Workshops

Special Issue

## Past Workshops

The BayesOpt workshop has become an integral
community in machine learning and has been helc
Below are links to previous workshops:

- 2016: Bayesian Optimization: Black-box Optimization and Beyond
- 2015: Bayesian Optimization: Scalability and Flexibility
- 2014: Bayesian Optimization in Academia and Industry
- 2013: Bayesian Optimization in Theory and Practice
- 2012: Bayesian Optimization & Decision Making
- 2011: Bayesian Optimization, Experimental Design and Bandits

## Towards an Empirical Foundation for Assessing Bayesian Optimization of Hyperparameters

**Katharina Eggensperger, Matthias Feurer, Frank Hutter**
Freiburg University
{eggenspk,feurerm,fh}@informatik.uni-freiburg.de

**James Bergstra**
University of Waterloo
james.bergstra@uwaterloo.ca

**Jasper Snoek**
Harvard University
jsnoek@seas.harvard.edu

**Holger H. Hoos and Kevin Leyton-Brown**
University of British Columbia
{hoos,kevinlb}@cs.ubc.ca

**Abstract**

Progress in practical Bayesian optimization is hampered by the fact that the only available standard benchmarks are artificial test functions that are not representative of practical applications. To alleviate this problem, we introduce a library of benchmarks from the prominent application of hyperparameter optimization and use it to compare Spearmint, TPE, and SMAC, three recent Bayesian optimization methods for hyperparameter optimization.

95 commits | 3 branches | 0 releases

naster ▾ | New pull request

lbart committed on 18 Dec 2016 Merge pull request #66 from midak/master ...

| | |
|---|---|
| ples | removed non-default grid size from config file of noisy function, add... |
| mint | solved issue #32: Simple Case of 1 Optimization Variable |
| nore | initial commit |
| TRIBUTING.rst | Update CONTRIBUTING.rst |
| NSE.md | Update LICENSE.md |
| ME.md | correct reference to `branin.py` |
| ibutors.md | Update contributors.md |
| .py | Fixed a couple of issues |

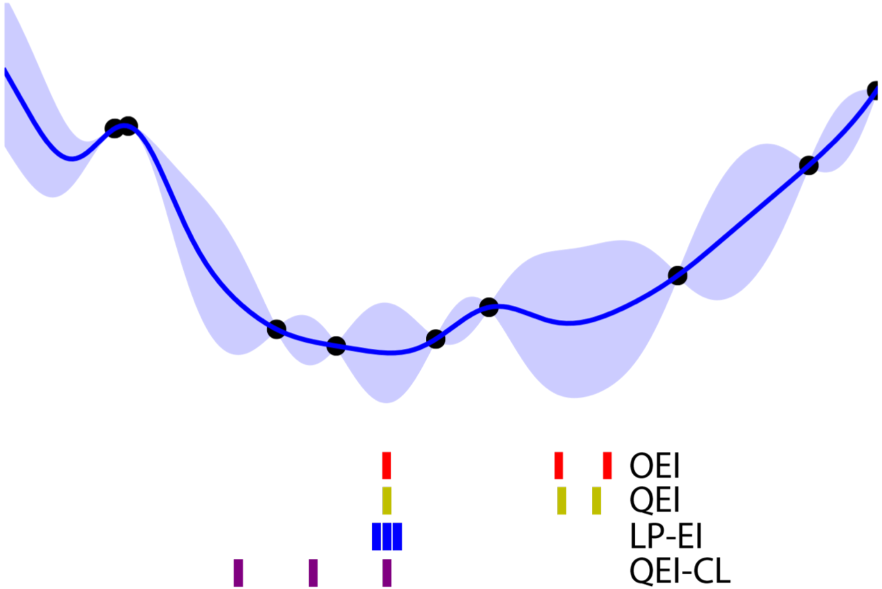
**ME.md**

## Spearmint

Spearmint is a software package to perform Bayesian optimization. The Software is designed t
experiments (thus the code name spearmint) in a manner that iteratively adjusts a number of p
some objective in as few runs as possible.

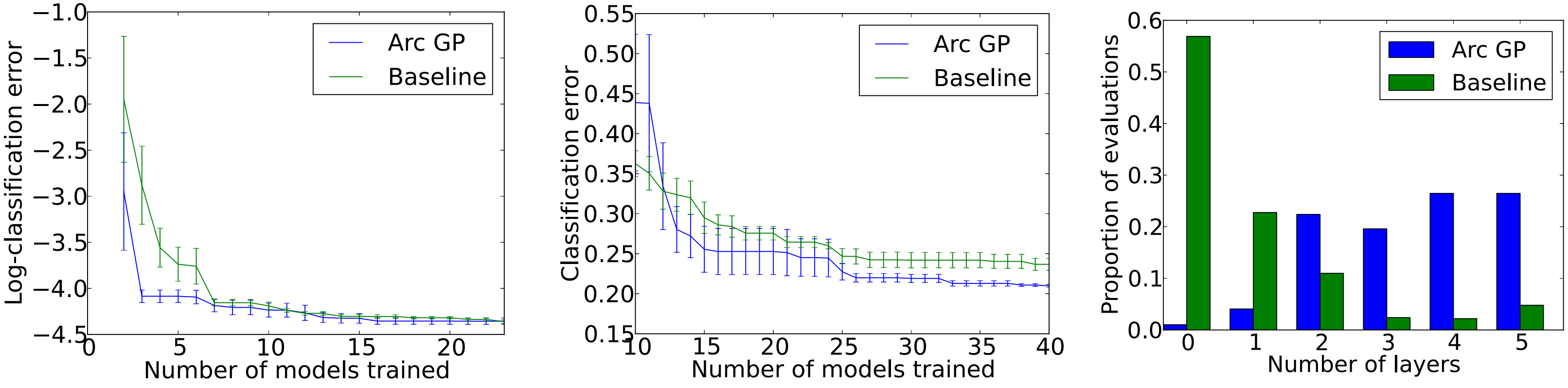
**IMPORTANT:** Spearmint is under an **Academic and Non-Commercial Research Use License.**

Thirdly, Bayesian optimisation adopted a ruthless focus on the real pain-points of users. A good example is batch Bayesian optimisation.




Source: Rontsis, Osborne & Goulart (2017)

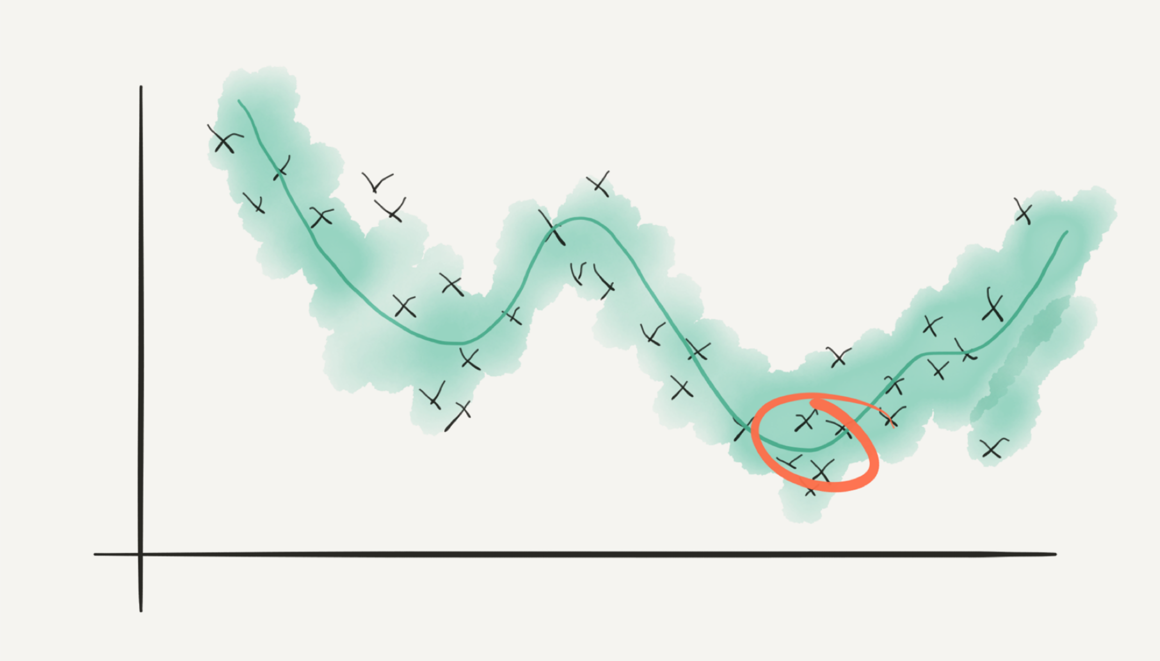
Bayesian optimisation is useful in automating structured search over # hidden layers, learning rates, dropout rates, # hidden units per layer & L2 weight constraints.



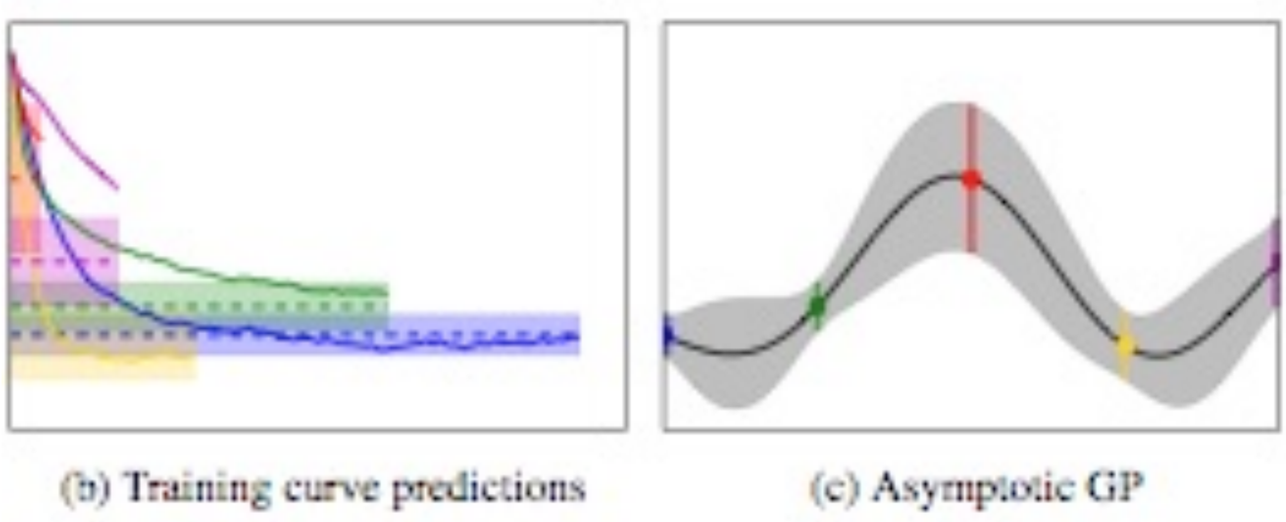

(a) MNIST (b) CIFAR-10 (c) Architectures searched

Source: Swersky, Duvenaud, Snoek, Hutter and Osborne (2013).

# Bayesian optimisation is well-suited to noisy optimisation.

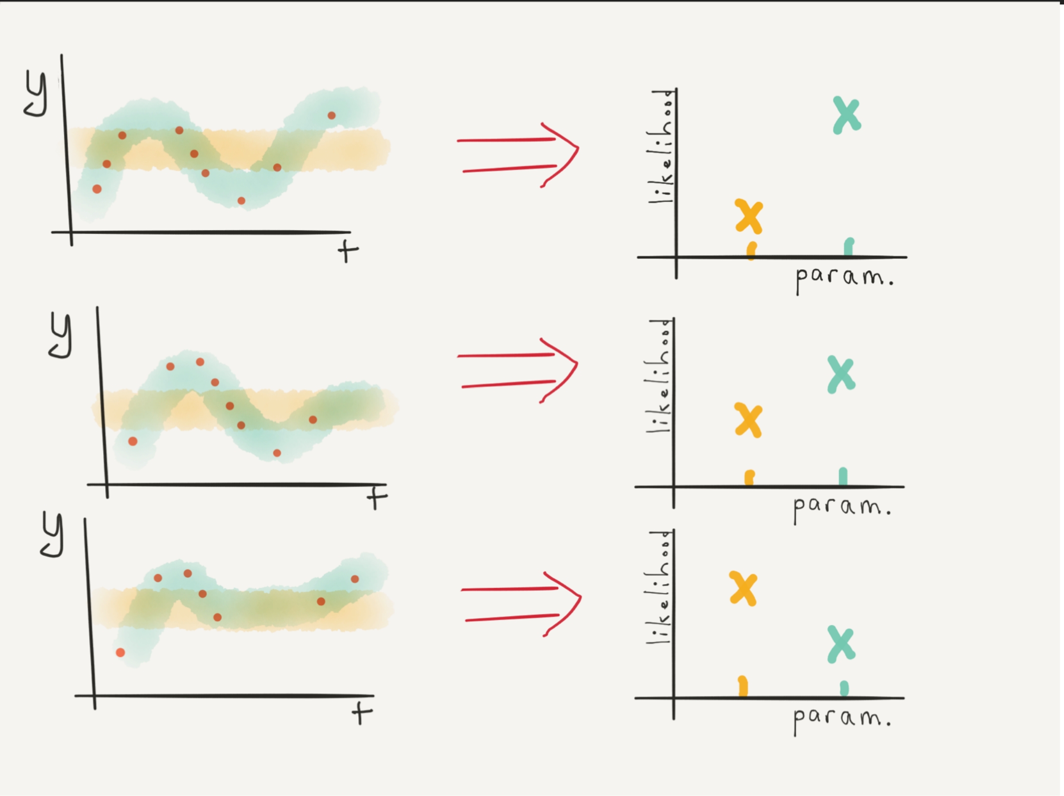


Freeze-thaw Bayesian optimisation unifies inner-loop (training) optimisation and outer-loop (hyperparameter) optimisation.

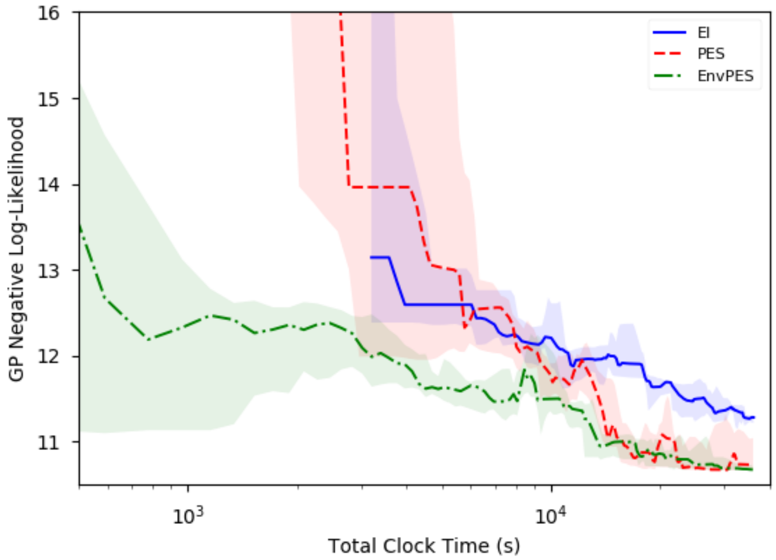



Source: Swersky, Snoek & Adams (2014).

Using only a subset of the data (a mini-batch) gives a noisy likelihood evaluation.



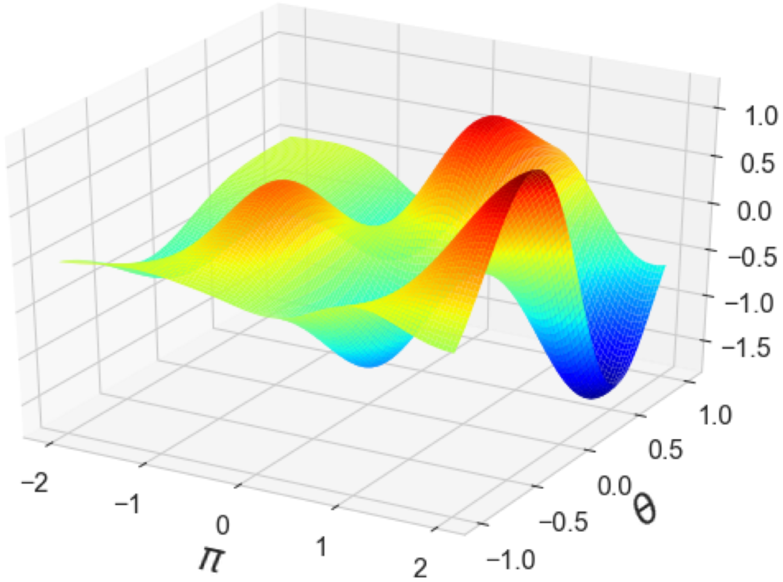
Lower-variance evaluations (on smaller subsets) are higher cost: let’s also Bayesian optimise over the fidelity of our evaluations!

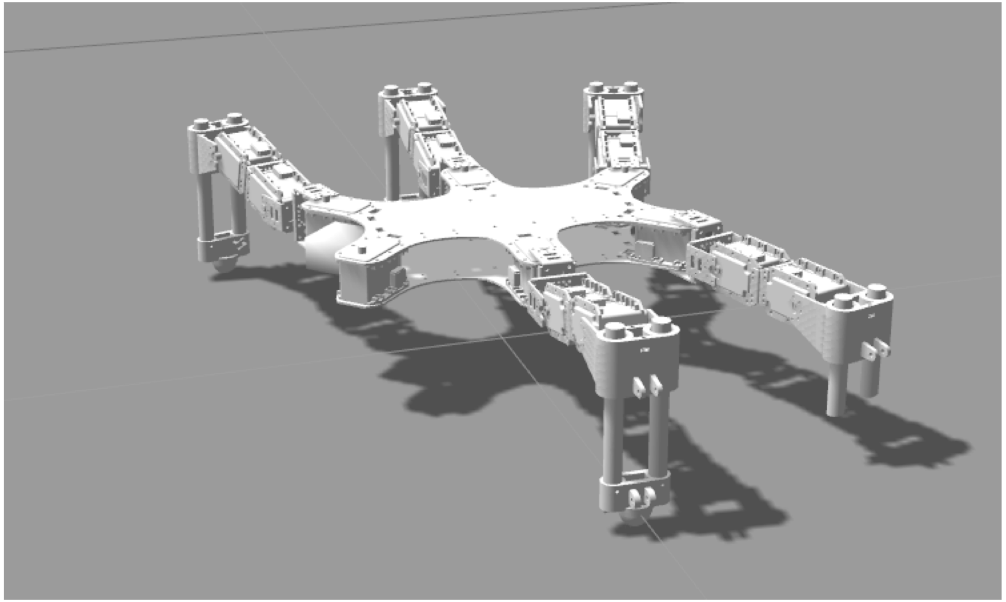



We tune the hyperparameters of a GP fitted to half hourly time series data for UK electricity demand for 2015, for which a full evaluation costs ten minutes.

Klein, Falkner, Bartels, Hennig & Hutter (2017);
McLeod, Osborne & Roberts (2017), `arxiv.org/abs/1703.04335`

Bayesian optimisation and Bayesian quadrature can be combined to optimize in the presence of environmental variables.

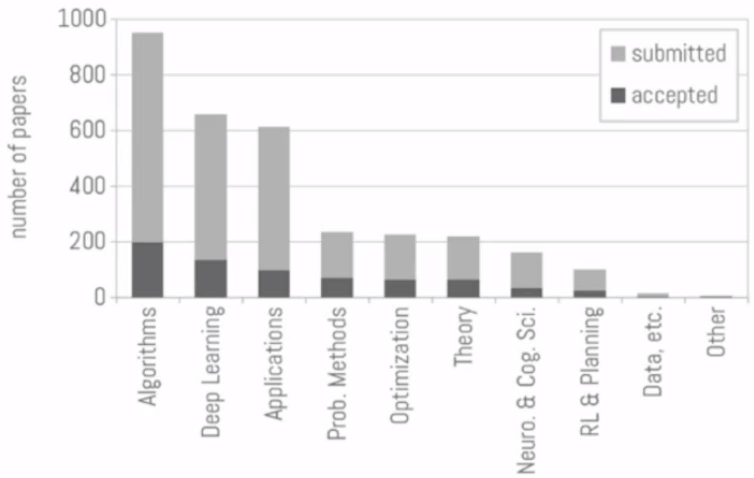


Source: Paul, Chatzilygeroudis, Ciosek, Mouret, Osborne and Whiteson (2018).

What are the lessons for (other) probabilistic numeric methods?

## Firstly, machine learning is a rich source of important numeric problems.




These problems are often exotic (e.g. combined optimisation and quadrature), structured, noisy and expensive, all motivating bespoke probabilistic numeric methods.

Source: NIPS 2017 statistics; https://is.gd/Abw95A

Secondly, probabilistic numerics needs to to cohere around recurring workshops, performance metrics and open-source software.


PROBABILISTIC-NUMERICS.ORG

ABOUT/NEWS BLOG MEETINGS RESEARCH LITERATURE

## COMMUNITY MEETINGS AND EVENTS

This page lists past and future meetings of the Probabilistic Numerics community.

**2018**

- *16 - 19 April*
  Minisymposium on Probabilistic Numerical Methods for Quantification of Discretization Error at the SIAM Conference on Uncertainty Quantification in Garden Grove, California. Organized by Tim Sullivan, Philipp Hennig, Chris Oates and Mark Girolami
- *11 - 13 April*
  SAMSI-Lloyds-Turing Workshop on Probabilistic Numerics at the Alan Turing Institute in London, UK. Organized by Chris Oates and Tim Sullivan

**2017**

- *June 18 - 23*
  Dobbiaco Summer School on Probabilistic Numerics at the Hotel Union in Dobbiaco, Italy. Organized by Alfredo Bellen, Stefano Maset and Marino Zennaro (University of Trieste) and Alexander Ostermann (University of Innsbruck).
  Taught by Philipp Hennig & Mark Girolami
- *June 5 - 9*
  Seminar on Probabilistic Scientific Computing: Statistical inference approaches to numerical analysis and algorithm design
  at ICERM (the Institute for Computational and Experimental Research in Mathematics), Brown University, Providence, Rhode Island.
  Organized by Philipp Hennig, George Em Karniadakis, Michael A Osborne, Houman Owhadi and Paris Perdikaris

**2016**

- *18 August*
  Probabilistic Numerics @ MCQMC 2016
  at Stanford University, California
  organized by Mark Girolami and François-Xavier Briol
- *7 January*
  Probabilistic Numerics: Integrating Inference With Integration @ MCMSki
  in Lenzerheide, Switzerland
  organized by Michael Osborne, Chris Oates and François-Xavier Briol

Thirdly, there's a movement to make machine learning more rigorous and interpretable: it's time for rigorous and interpretable numerics.

Probabilistic numerics should provide rigorous solutions to the diverse needs of machine learning.