A Whirlwind Tour of Gaussian Process Models and Applications

Vincent Dutordoir
University of Cambridge
Secondmind

vd309@cam.ac.uk
November 9th 2021

Introduction to Gaussian Processes

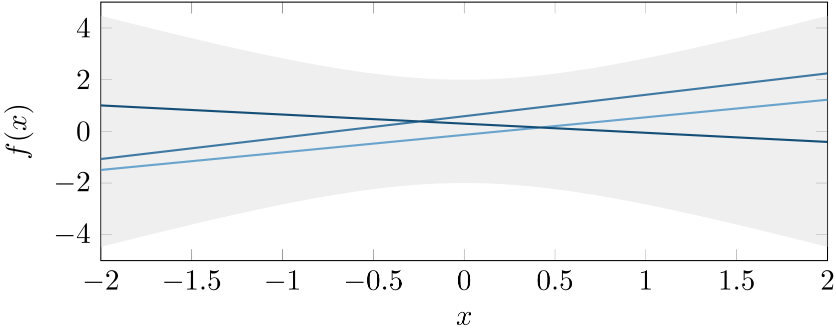
Distribution over functions

Gaussian processes are distributions over functions:

$$f(x) \;=\; ax + b, \qquad a \;\sim \mathcal{N}(0,1), \qquad b \;\sim\; \mathcal{N}(0,1), \tag{1}$$

with

$$m(x) = \mathbb{E}[f(x)] = \mathbb{E}[a]x + \mathbb{E}[b] = 0 \quad \text{and} \quad \sigma^2(x) = x^2 + 1 \tag{2}$$



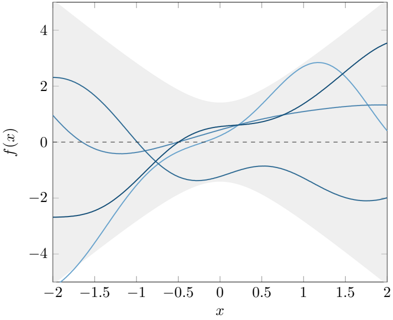
Gaussian Processes

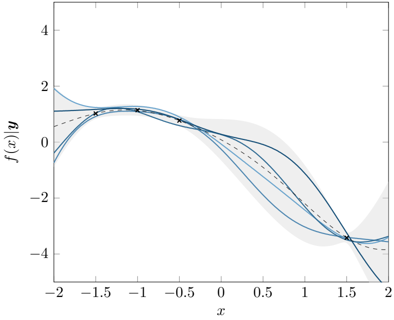
Prior.

$$y_i = f(x_i) + \varepsilon_i, \qquad \text{where} \qquad f \sim \mathcal{GP}(m_{\text{prior}}, k_{\text{prior}}) \quad \text{and} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

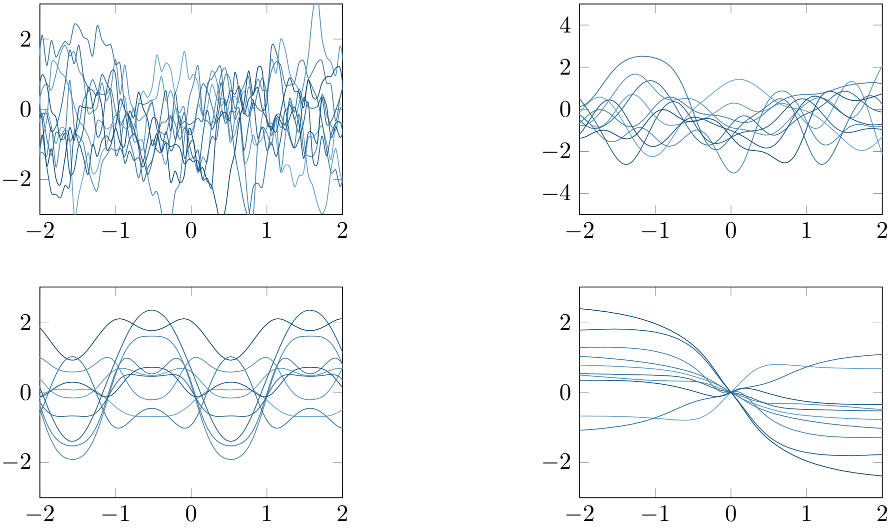
Posterior. Which yields an analytic posterior:

$$f \mid \boldsymbol{y} \sim \mathcal{GP}\left(m_{\text{post}}, k_{\text{post}}\right) \tag{4}$$





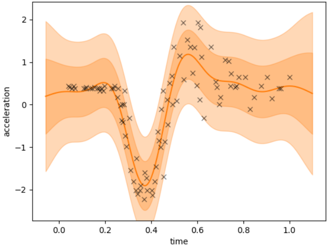
Prior beliefs



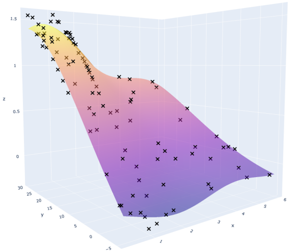
Gaussian Process deficiencies

1. Gaussian marginals,
2. Choosing the kernel à-priori is hard,
3. Simple kernels cannot effectively model ‘complex’ data,
4. Expressive kernels either require domain knowledge or need to be inferred.

Motorcycle dataset



Rocket Booster data

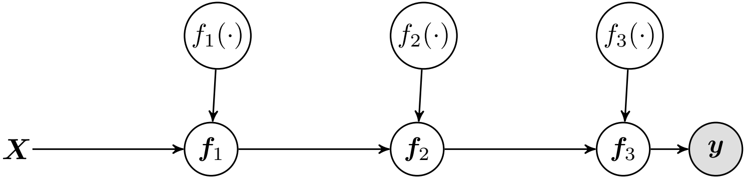


Advanced Gaussian Processes

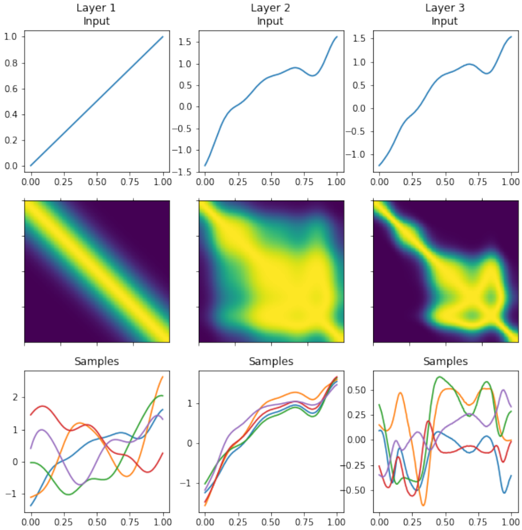
Deep Gaussian Processes

Deep Gaussian Process Hierarchical model by GP composition

$$y = (f_L \circ f_{L-1} \circ \ldots \circ f_1)(x) + \varepsilon, \quad \text{where} \quad f_\ell \sim \mathcal{GP}(0, k_\ell) \tag{5}$$



Deep Gaussian Processes

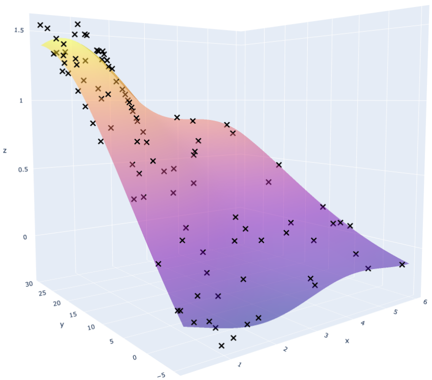


Motivations

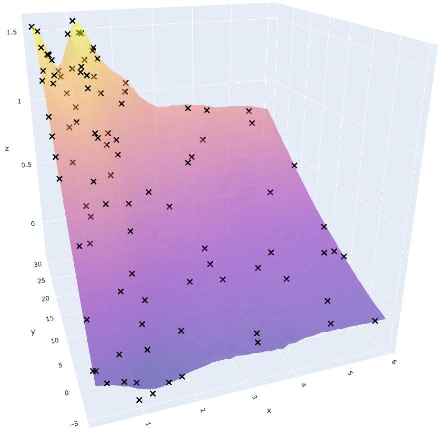
1. Deep learning has shown to work well,
2. More flexible priors,
3. Deep and complex Bayesian model.

Example DGPs: LGBB[1]

Single-layer GP



Two-layer DGP

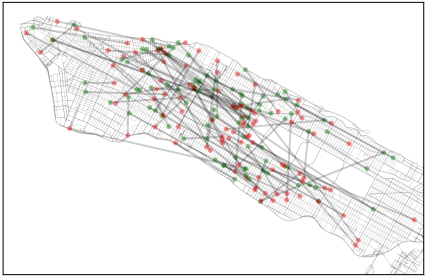


[1] Langley Glide-Back Booster (LGBB), see https://bobby.gramacy.com/surrogates/

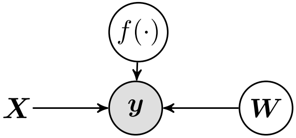
Conditional Density Estimation

- The mean $\mathbb{E}[f(x^*)]$ is not always informative enough due to multi-modality, asymmetry or heteroscedasticity.
- We are interested in learning the full conditional distribution $p(f(x^*) \mid x^*)$.

Examples



Latent Variable GP models



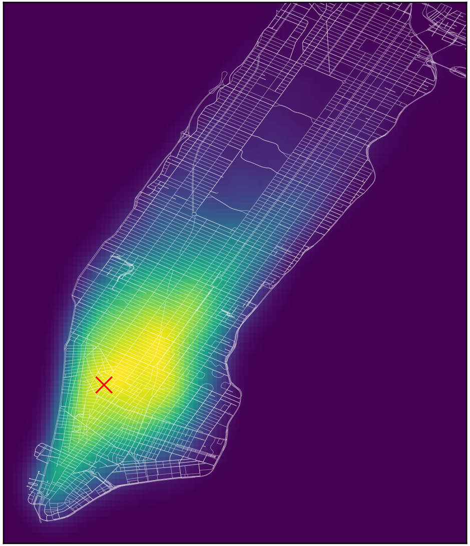
Model.

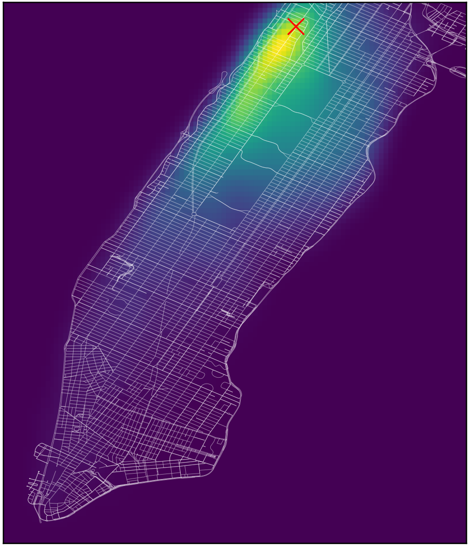
$$y_i = f([x_i, w_i]) \qquad f \sim \mathcal{GP} \text{ and } w_i \sim \mathcal{N}(0, 1).$$

Posterior. We need to learn the posterior of the GP and the latent variables $p(f, \{w\}_i^n \mid \mathcal{D})$. In practice, we learn a mean-field approximation

$$p(f, \{w\}_i^n \mid \mathcal{D}) \approx q(f) \prod_i^n q(w_i) \tag{6}$$

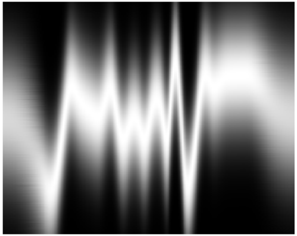
Example: Manhattan Taxi Drop-off



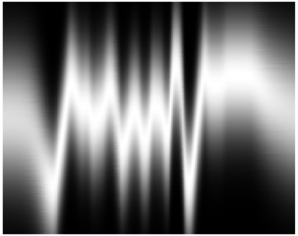


Example: 'DGP' letters

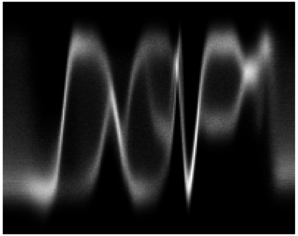
GP



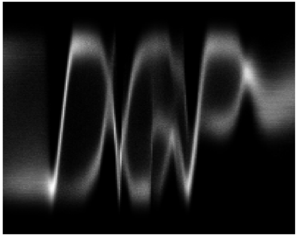
GP-GP



LV-GP

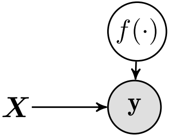


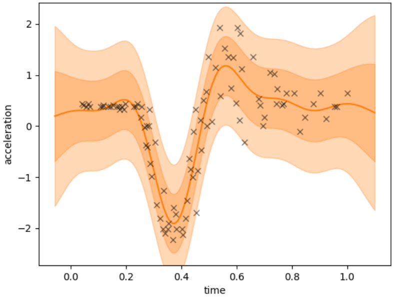
LV-GP-GP



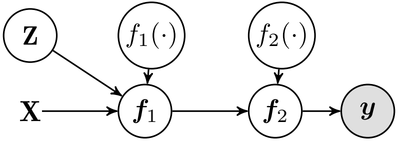
Deep Latent Variable GP: Motorcycle

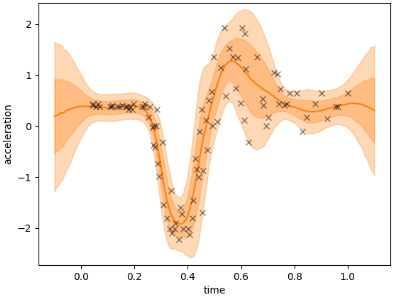
Single Layer model





Deep Latent GP





GPs on Manifolds

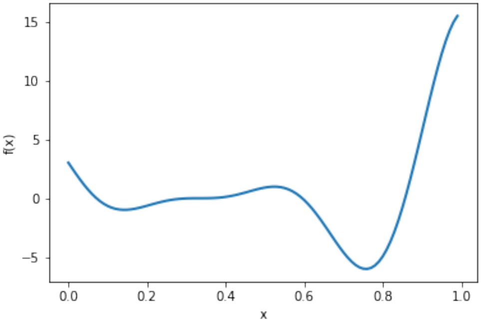
Most kernels we know are defined on $\mathbb{R}^d$. Some problems are more naturally defined in other space.

Images courtesy of Alexander Terenin

Decision Making: Bayesian Optimisation

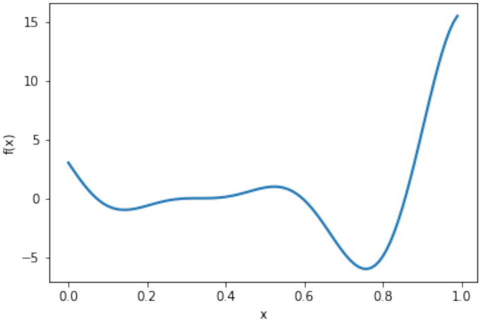
Demo

Lets try to find the minimum of $f(x) = (6x-2)^2 \sin(12x-4)$



Demo

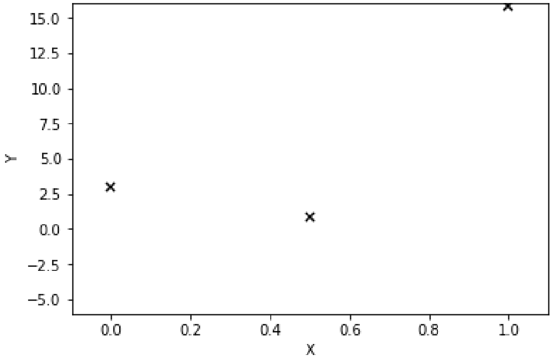
Lets try to find the minimum of $f(x) = (6x-2)^2 \sin(12x-4)$



Using as few function evaluations as possible!

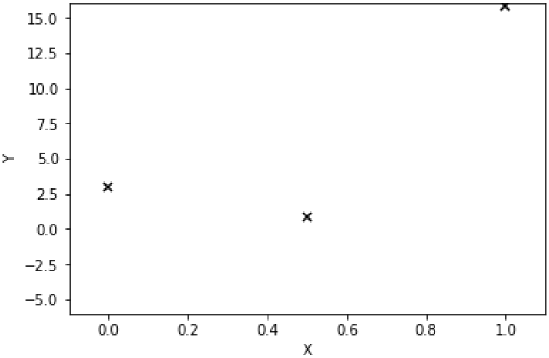
Demo

Suppose we make evaluations at 0, 0.5 and 1



Demo

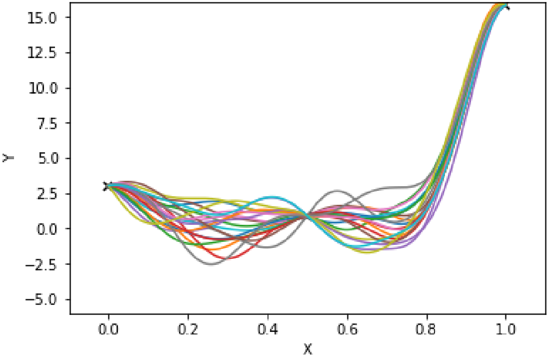
Suppose we make evaluations at 0, 0.5 and 1



Where should we evaluate next? Why is this an exploration-exploitation problem?

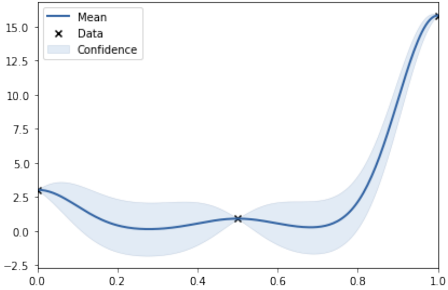
Possible functions

Possible functions that pass through the observed points



A Gaussian process model

We can summarise this belief by fitting a Gaussian process

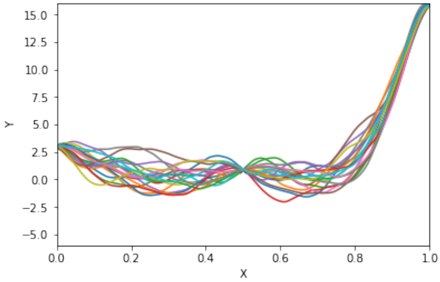


Predictive distribution at x is Gaussian $g(x) \sim \mathcal{GP}(m, k)$

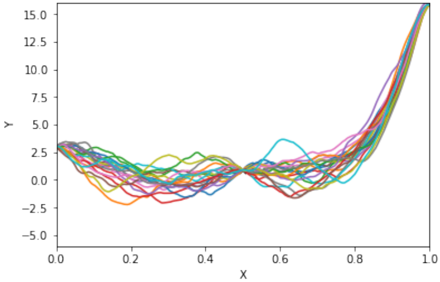
Choice of kernel function

Represents prior knowledge about function shape

Squared Exponential



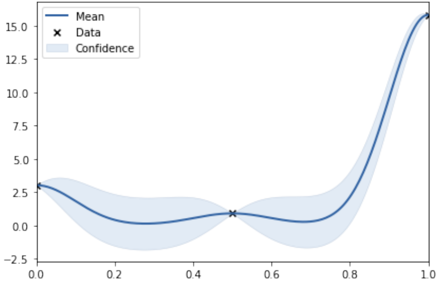
Matérn-$\frac{3}{2}$



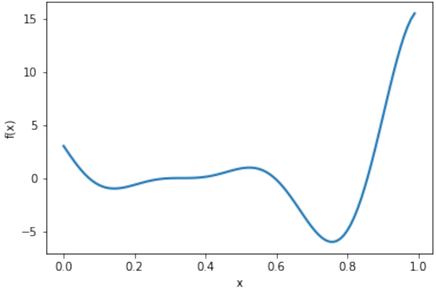
Model vs. truth

Compare our statistical model with the truth

Learned GP



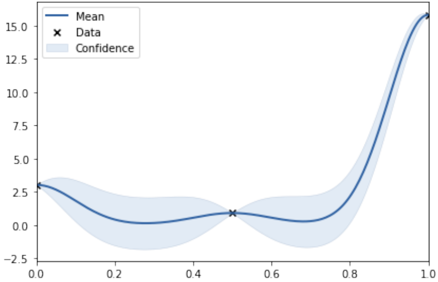
Truth



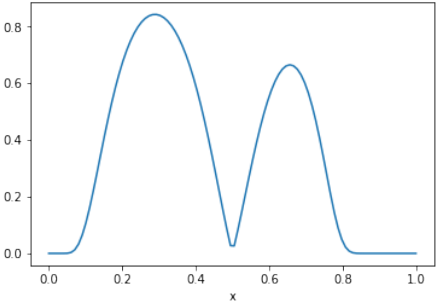
Where do we evaluate next?

We measure the utility of a potential evaluation with an **acquisition function**.

Learned GP

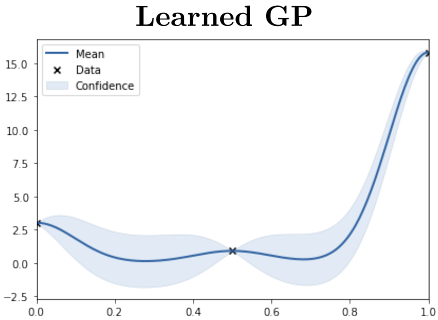


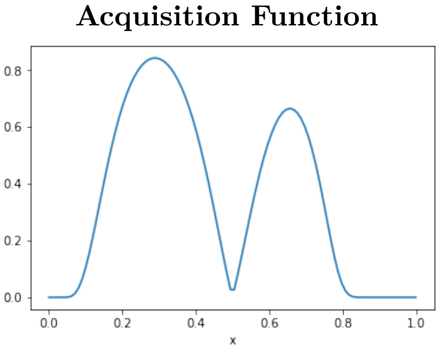
Acquisition Function



Where do we evaluate next?

We measure the utility of a potential evaluation with an **acquisition function**.

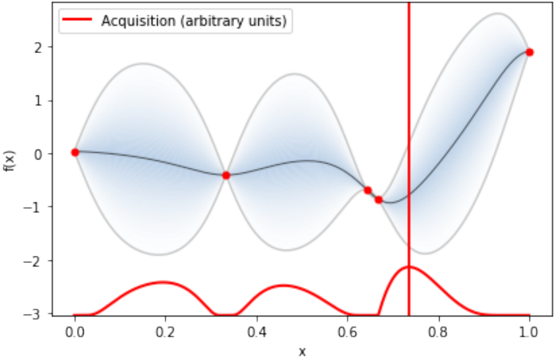




Lower Confidence Bound acquisition function: $\alpha(x) = \mu(x) - \beta\sigma(x)$

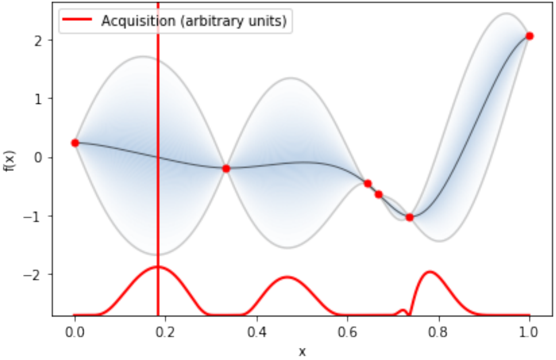
Step 1

Model after 4 initial points and 1 evaluation chosen by Bayesian optimisation



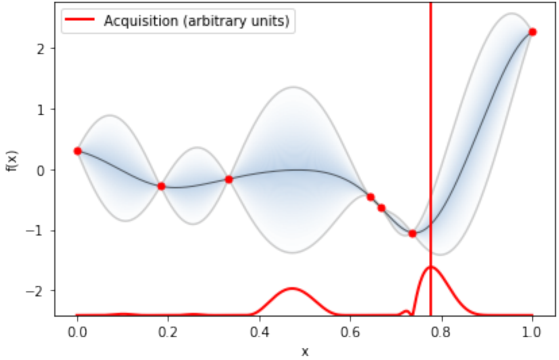
Step 2

Model after 4 initial points and 2 evaluations chosen by Bayesian optimisation



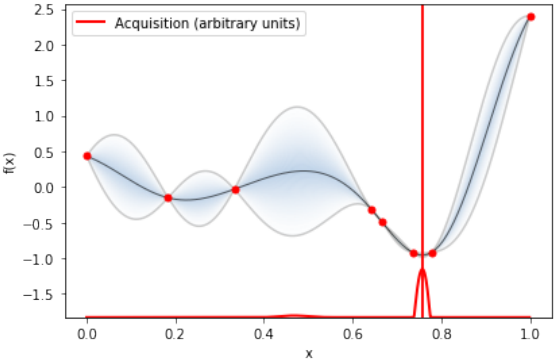
Step 3

Model after 4 initial points and 3 evaluations chosen by Bayesian optimisation



Bayesian optimisation demo: Step 4

Model after 4 initial points and 4 evaluations chosen by Bayesian optimisation



Take-away messages

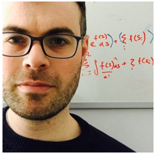
1. Gaussian processes are a framework for modelling unknown functions.
2. The classic framework can be extended in many ways to model more complex problems
3. Caveat: these models typically don't work out-of-the-box. A lot of expertise (read: trial and error) is needed to fit them satisfactory.
4. Ongoing topic of research in terms of scalability (faster and larger datasets) and accuracy (richer approximate posteriors)
5. Allow for principled decision-making (e.g., Bayesian optimisation)

Co-authors

Hugh Salimbeni

Felix Leibfried

Henry Moss



James Hensman

Nicolas Durrande

Software

- GP models in TensorFlow:
  `https://github.com/GPflow/GPflow`
- Kernels (for GPs) on interesting spaces:
  `https://github.com/GPflow/GeometricKernels`
- Deep Gaussian processes and Latent Variable models:
  `https://github.com/Secondmind-Labs/GPflux`
- Bayesian Optimisation:
  `https://github.com/Secondmind-Labs/Trieste`

References I

Damianou, Andreas and Neil D. Lawrence (2013). "Deep Gaussian Processes". In: *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS).*

Dutordoir, Vincent, Hugh Salimbeni, Eric Hambro, et al. (2021). "GPflux: A Library for Deep Gaussian Processes". In: *Proceedings of the 3th International Conference on Probabilistic Programming.*

Dutordoir, Vincent, Hugh Salimbeni, James Hensman, et al. (2018). "Gaussian Process Conditional Density Estimation". In: *Advances in Neural Information Processing Systems 31 (NeurIPS).* Vol. 31.

Salimbeni, Hugh and Marc P. Deisenroth (2017). "Doubly Stochastic Variational Inference for Deep Gaussian Processes". In: *Advances in Neural Information Processing Systems 30 (NeurIPS).*

Salimbeni, Hugh, Vincent Dutordoir, et al. (2019). "Deep Gaussian Processes with Importance-Weighted Variational Inference". In: *Proceedings of the 36th International Conference on Machine Learning (ICML).*

Terenin, Alexander (2021). "Gaussian Processes and Statistical Decision-making in Non-Euclidean Spaces". PhD thesis. Imperial College London, U.K.

Thank you

Feel free to e-mail me if you have any questions: `vd309@cam.ac.uk`