UNIVERSITY OF CAMBRIDGE

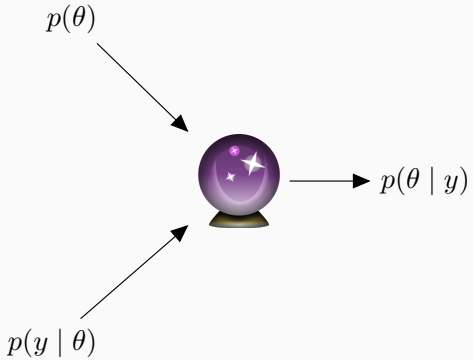
Modulated Surrogates Models for Bayesian Optimisation

Carl Henrik Ek - che29@cam.ac.uk

March 29, 2023

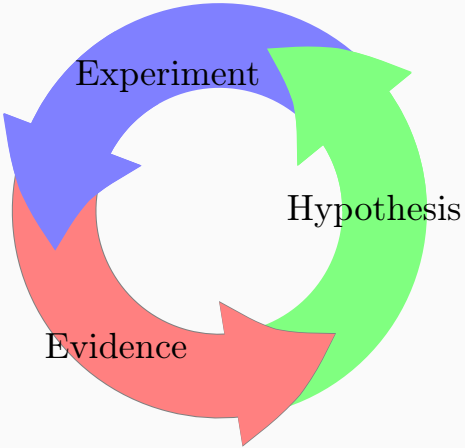
http://carlhenrik.com

The world I lived in[1]

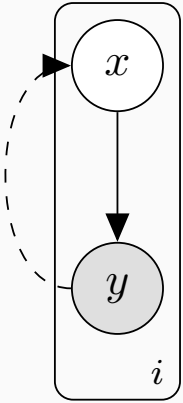


[1]Stolen from John Cunningham

Me on my high horse



"The Bayesian norm enforces hygiene between modelling and decision making."
– Hennig, Osborne, and Kersting



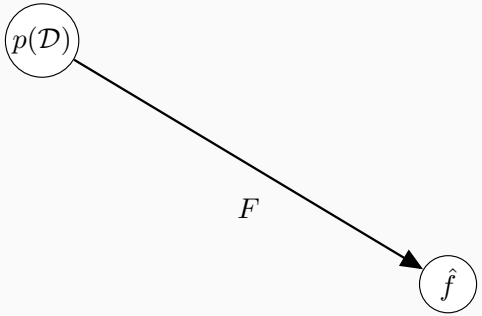
$$w := w - \eta \nabla \mathcal{L}_i(w) + \alpha \Delta w$$

- Philipp Hennig et al. (July 2015). "Probabilistic numerics and uncertainty in computations". In: *Proc. R. Soc. A* 471.2179, p. 20150142
- Jon Cockayne et al. (2017). "Bayesian Probabilistic Numerical Methods". In: *CoRR*. arXiv: `1702.03673 [stat.ME]`
- C. J. Oates et al. (2019). "A Modern Retrospective on Probabilistic Numerics". In: *CoRR*. arXiv: `1901.04457 [math.NA]`

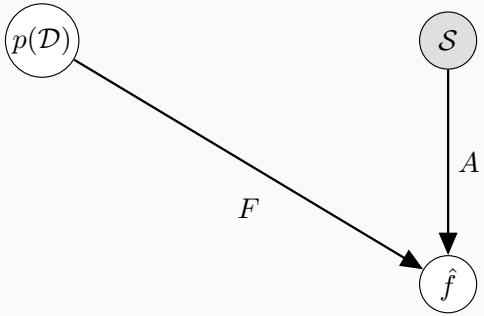
$$L(A_{h\in\mathcal{H}}(S)) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$


Come on!
It's Good for you!!
Emerald High ©
Friends 2011

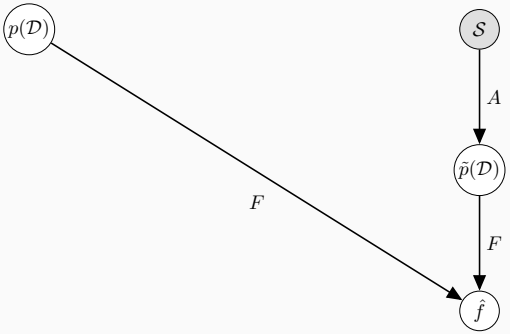
Formalisation Oates et al., **2019**



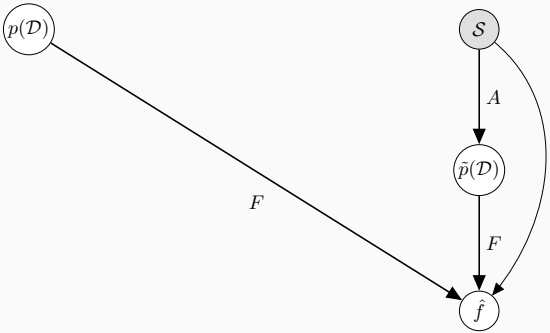
Formalisation Oates et al., 2019



Formalisation



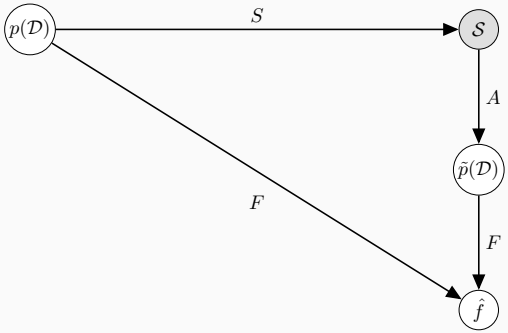
Formalisation



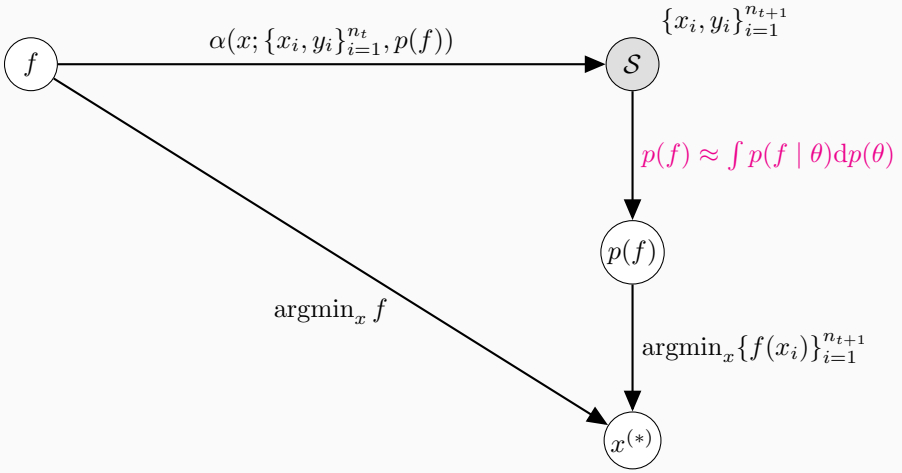
"Modelling (or inference) used to be thought of as a passive mathematical map, from data to estimate. But machine learning often views a model as an agent in autonomous interaction with its environment, most explicitly in reinforcement learning. This view of algorithms as agents is, as above, central to pn."

– Hennig, Osborne, and Kersting

Formalisation



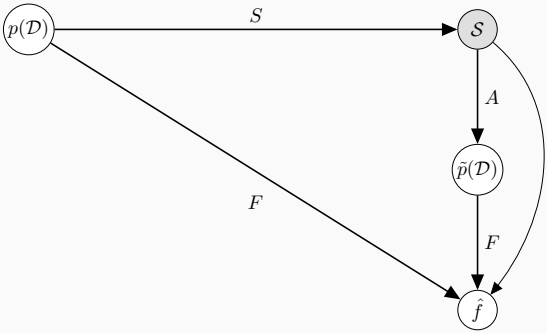
Bayesian Optimisation



The Book Loss

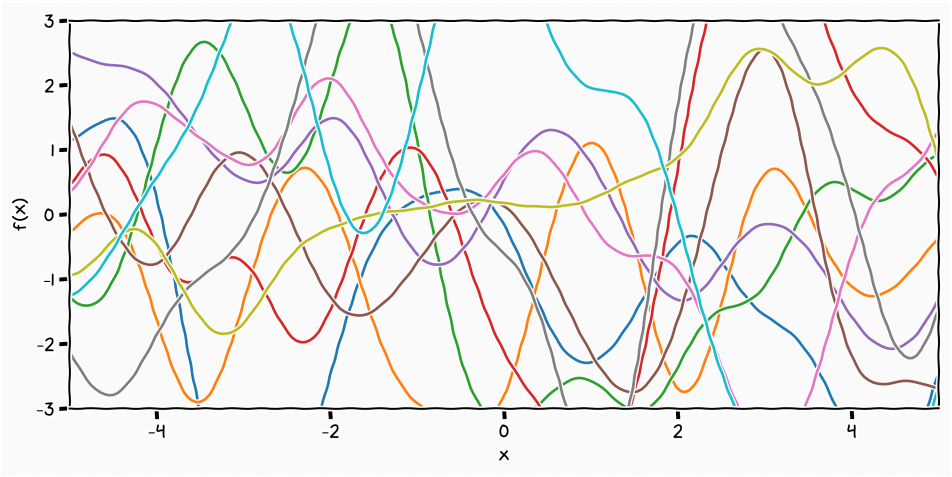
"This can be conceived as letting some loss function on computation dictate which elements of the prior can be incorporated"
– Hennig, Osborne, and Kersting

Formalisation

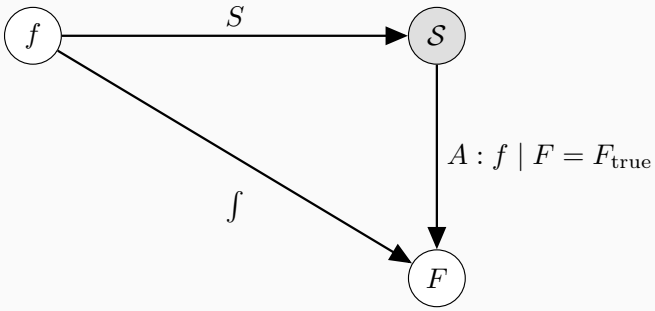


". . . . in considering an additional computation cost on a model, we must consider whether it is justified in improving performance for the given numerical task: this performance is measured by a loss function."
– Hennig, Osborne, and Kersting

GPs are not flexible models

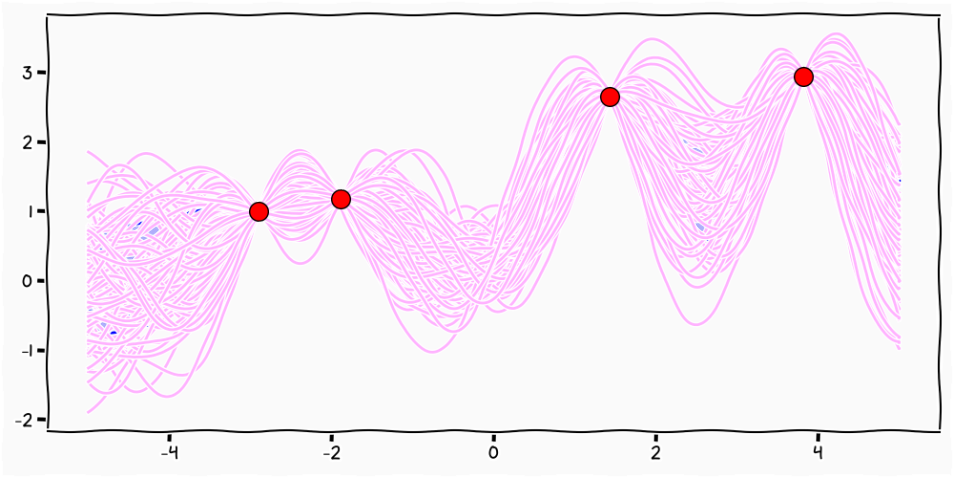


"All models are wrong" [2]



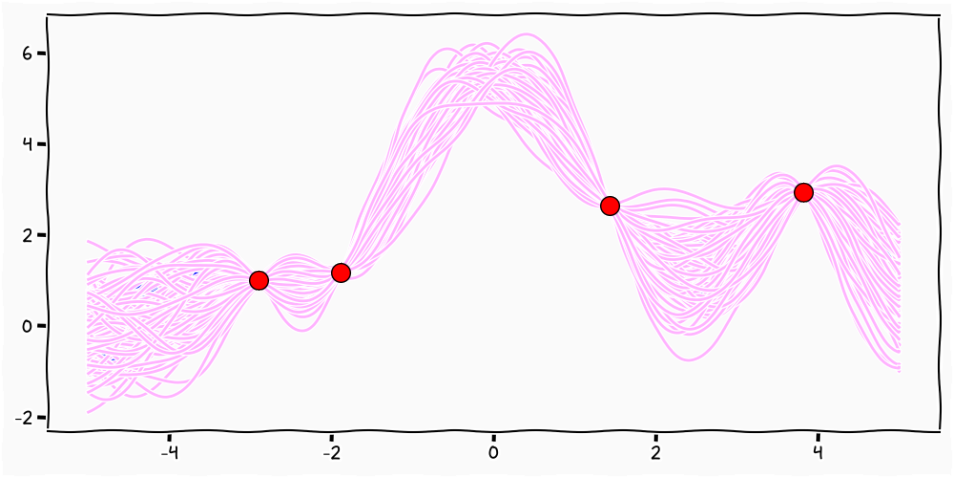
[2] Box, 1976

Quadrature [3]



[3] O'Hagan, 1991

Quadrature [4]



[4] O'Hagan, 1991

March 29, 2023

Modulated Surrogates for Bayesian Optimisation

Set-up

- Acknowlege that we will have model mismatch

Set-up

- Acknowlege that we will have model mismatch
- Uncertainties are to inform the agent (information operator)

Set-up

- Acknowlege that we will have model mismatch
- Uncertainties are to inform the agent (information operator)
- We only care about the latent variable up to the equivivalence class defined by the quantity of interest

Set-up

- Acknowlege that we will have model mismatch
- Uncertainties are to inform the agent (information operator)
- We only care about the latent variable up to the equivivalence class defined by the quantity of interest
- We cannot specify a model over the quantity of interest

Idea

- Specify prior over "structure" of the function that gives me the most efficient search

Idea

- Specify prior over "structure" of the function that gives me the most efficient search
- Explain away variance that is not helpful to solve the task

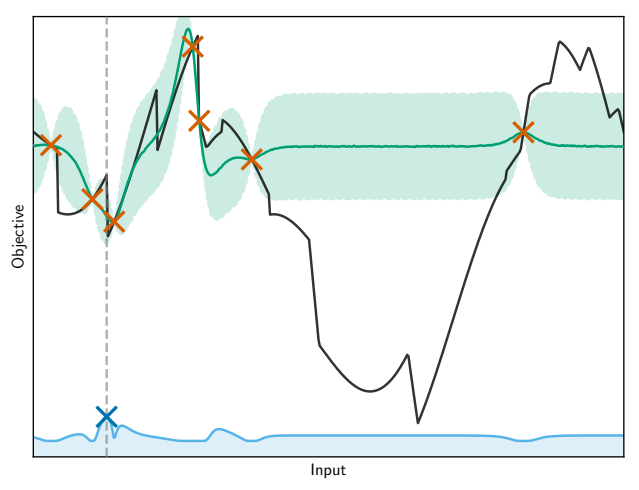
Idea

- Specify prior over "structure" of the function that gives me the most efficient search
- Explain away variance that is not helpful to solve the task
- All data is not equally informative of the extremum and the information content changes over time.

Idea

- Specify prior over "structure" of the function that gives me the most efficient search
- Explain away variance that is not helpful to solve the task
- All data is not equally informative of the extremum and the information content changes over time.
- The uncertainty about the function is only a "weak" proxy for our ignorance about its minima.

Irrelevant structures Bodin et al., 2020

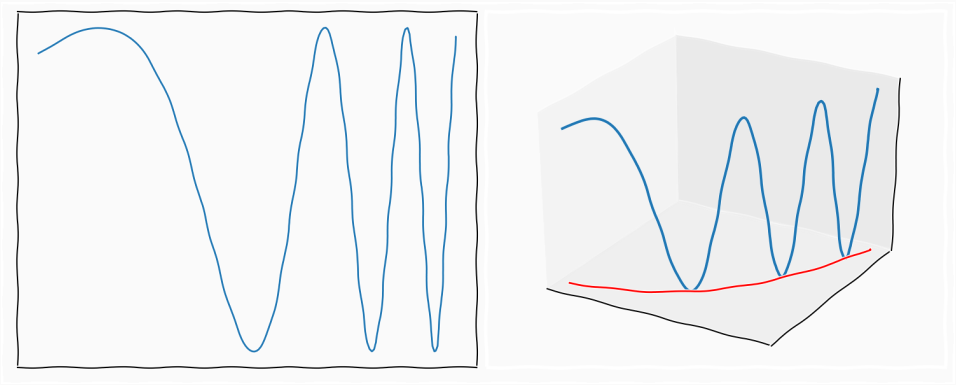


$$
\begin{aligned}
f(\mathsf{x}) &= g(\mathsf{x}, \mathsf{h}) \\
g &\sim \mathcal{GP}(0, k(\cdot, \cdot)) \\
k : (\mathcal{X} \times \mathcal{H}) \times (\mathcal{X} \times \mathcal{H}) &\to \mathbb{R} \\
\mathsf{h} &\sim \mathcal{N}(0, \sigma^2 \mathsf{I})
\end{aligned}
$$

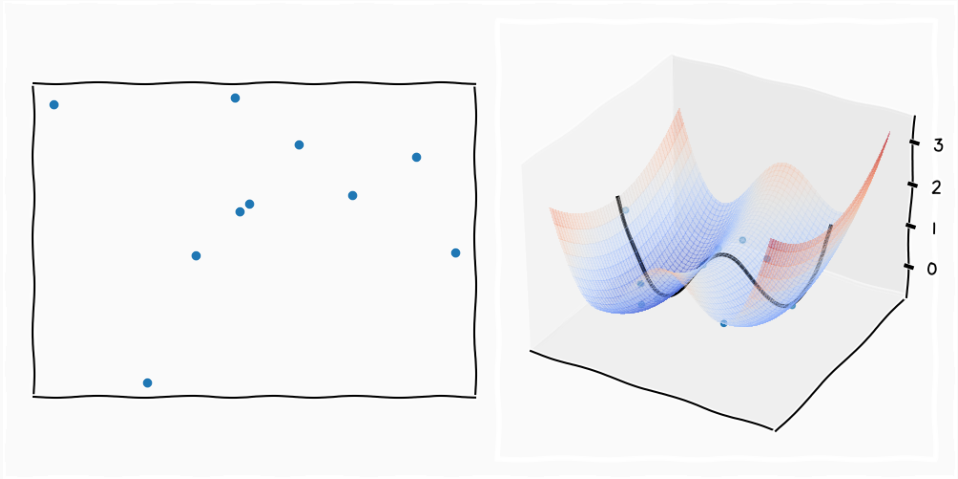
- treat h as a random variable that modulates the input to make $g(\cdot)$ well behaved[5]

[5] as in informative for the search

Latent Gaussian Process Regression Bodin et al., **2017**



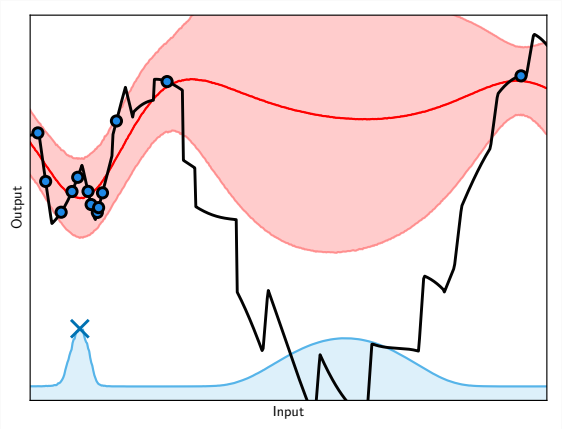
Latent Gaussian Process Regression Bodin et al., 2017



Predictive Posterior

$$p(\mathsf{f}_* \mid \mathsf{x}_*, \mathsf{X}, \mathsf{F}) = \int p(\mathsf{f}_* \mid \mathsf{x}_*, \mathsf{h}_*, \mathsf{X}, \mathsf{F}, \mathsf{H}, \boldsymbol{\theta}) \, p(\mathsf{H}, \theta \mid \mathsf{X}, \mathsf{F}) p(\mathsf{h}_*) \mathrm{d}\mathsf{H} \mathrm{d}\theta \mathrm{d}\mathsf{h}_*$$

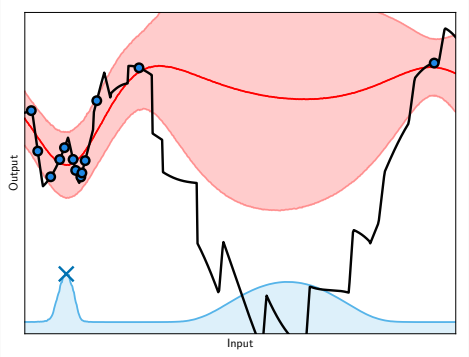
Marginalisation

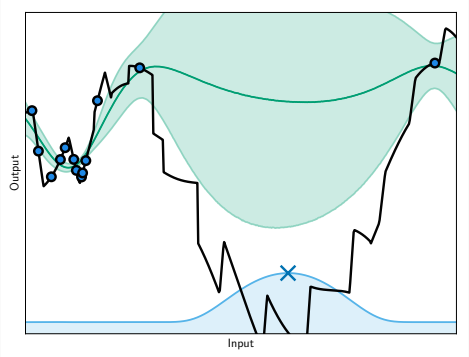


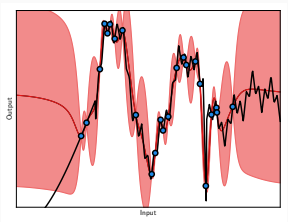
"Predictive Posterior"

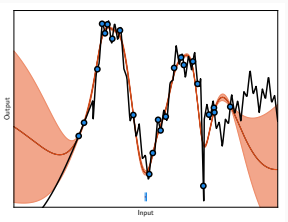
$$
\begin{aligned}
\hat{p}(\mathsf{f}_* \mid \mathsf{x}_*, \mathsf{X}, \mathsf{F}) &= \int p(\mathsf{f}_* \mid \mathsf{x}_* \mathsf{h}_*, \mathsf{X}, \mathsf{F}, \mathsf{H}, \theta) \\
&\quad p(\mathsf{H}, \theta \mid \mathsf{X}, \mathsf{F}) \hat{p}(\mathsf{h}_*) \mathrm{d}\theta \mathrm{d}\mathsf{H} \mathrm{d}\mathsf{h}_* \\
&= \int p(\mathsf{f}_* \mid \mathsf{x}_*, \mathsf{h}_* = 0, \mathsf{X}, \mathsf{F}, \mathsf{H}, \theta) \\
&\quad p(\mathsf{H}, \theta \mid \mathsf{X}, \mathsf{F}) \mathrm{d}\theta \mathrm{d}\mathsf{H} \\
\hat{p}(\mathsf{h}_*) &= \delta(\mathsf{h}_*)
\end{aligned}
$$

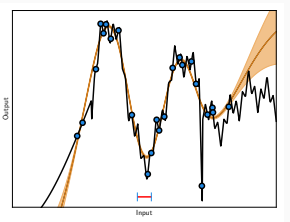
Marginal

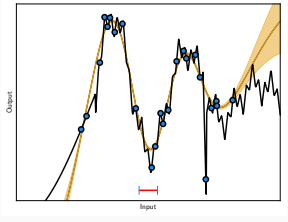


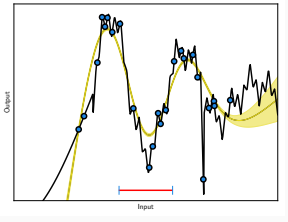


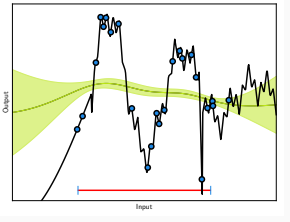

Output
Input


Output
Input

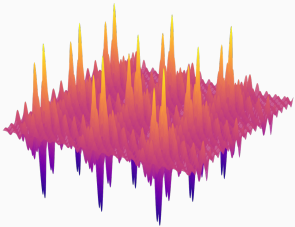

Output
Input

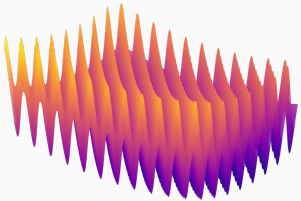

Output
Input

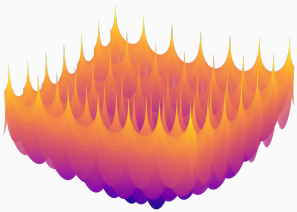

Output
Input


Output
Input

The Debacle of Empirical Evaluation

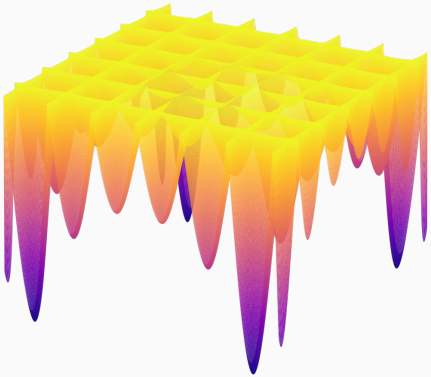


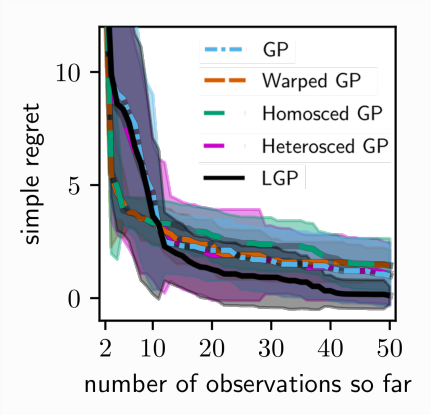




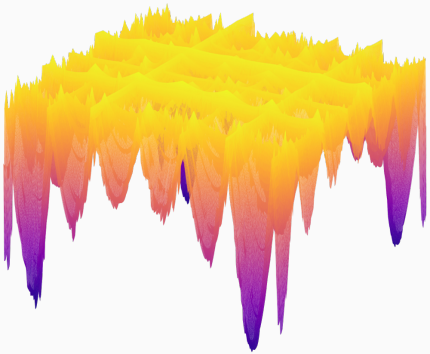
Benchmark	Evals	Dim	Properties	GP	Warped GP	Homosced GP	Heterosced GP	LGP
Hartmann	50	6	boring	**0.959**	0.537	0.881	**0.973**	**0.937**
Griewank	50	2	oscillatory	**0.914**	0.493	0.752	**0.913**	**0.897**
Shubert	50	2	oscillatory	0.378	0.158	**0.378**	**0.480**	**0.593**
Ackley $[-10, 30]^d$	50	2	complicated, oscillatory	**0.924**	0.274	**0.892**	**0.912**	**0.927**
Cross In Tray	50	2	complicated, oscillatory	**0.954**	0.385	0.929	**0.977**	**0.945**
Holder table	50	2	complicated, oscillatory	0.939	0.896	0.900	0.931	**0.993**
Corrupted Holder Table	50	2	complicated, oscillatory	0.741	0.798	0.826	0.729	**0.896**
Branin01	100	2	none	**1.000**		**1.000**	**1.000**	**1.000**
Branin02	100	2	none	**0.991**		0.964	**0.990**	**0.981**
Beale	100	2	boring	**0.987**		**0.982**	**0.987**	**0.988**
Hartmann	100	6	boring	**0.987**		0.947	**0.984**	**0.979**
Griewank	100	2	oscillatory	**0.967**		0.875	**0.969**	**0.946**
Levy	100	2	oscillatory	**0.997**		**0.999**	**0.998**	**0.998**
Deflected Corrugated Spring	100	10	oscillatory	0.347		**0.840**	0.406	0.697
Shubert $[-10, 10]^d$	100	2	oscillatory	0.510		0.511	**0.672**	**0.877**
Weierstrass	100	8	complicated	0.600		**0.704**	0.577	0.625
Cross In Tray	100	2	complicated, oscillatory	**1.000**		0.995	**1.000**	**1.000**
Holder Table	100	2	complicated, oscillatory	0.971		0.963	**0.964**	**1.000**
Ackley $[-10, 30]^d$	100	2	complicated, oscillatory	**0.971**		0.914	**0.980**	**0.974**
Ackley $[-10, 30]^d$	100	6	complicated, oscillatory	0.459		**0.789**	0.442	**0.712**
Corrupted Holder Table	100	2	complicated, oscillatory	0.844		0.889	0.822	**0.918**
Corrupted Exponential	100	8	complicated, oscillatory	0.580		**0.847**	0.581	**0.806**
HPO: NN Boston	100	9	unknown	**0.720**		**0.761**	**0.810**	**0.770**
HPO: NN Climate Model Crashes	100	9	unknown	0.629		**0.717**	**0.683**	**0.678**
Active learning: Robot Pushing	100	4	unknown	0.877		0.745	**0.907**	**0.932**

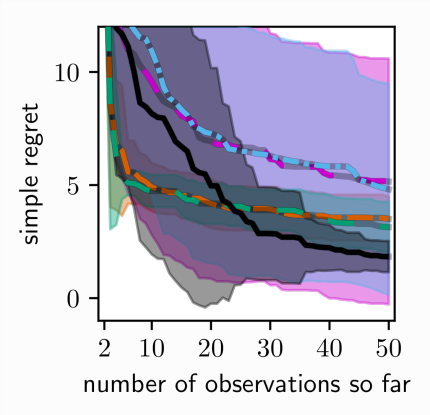
Holder Table





Corrupted Holder Table





Conclusion

Conclusion

- The loss function specifies an equvivalence class of functions

Conclusion

- The loss function specifies an equvivalence class of functions
- What is this prior?

Conclusion

- The loss function specifies an equvivalence class of functions
- What is this prior?
- BO has more to do with the bookeeping of linear algebra than the modelling of the quantity of interest as in quadrature

Conclusion

- The loss function specifies an equvivalence class of functions
- What is this prior?
- BO has more to do with the bookeeping of linear algebra than the modelling of the quantity of interest as in quadrature
- We need to acknowledge model mismatch, the non-parametric formulation means observations can lead to uncertainties which are deterimental for our agent to make decisions

Conclusion

- The loss function specifies an equvivalence class of functions
- What is this prior?
- BO has more to do with the bookeeping of linear algebra than the modelling of the quantity of interest as in quadrature
- We need to acknowledge model mismatch, the non-parametric formulation means observations can lead to uncertainties which are deterimental for our agent to make decisions
- This work probably isn't PN but its thinking was inspired by

PN equates data and compute

"Data is computations that the universe have already done for us"
– Neil D. Lawrence

eof

References


Bodin, Erik, Neill D. F. Campbell, and Carl Henrik Ek (2017). *Latent Gaussian Process Regression*. arXiv: 1707.05534v1 [stat.ML].


Bodin, Erik et al. (2020). "Modulating Surrogates for Bayesian Optimization.". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2019, 12-18 July 2020, Virtual*.


Box, George E. P. (1976). "Science and Statistics". In: *Journal of the American Statistical Association* 71.356, pp. 791–799. DOI: 10.1080/01621459.1976.10480949.


Cockayne, Jon, Chris Oates, Tim Sullivan, and Mark Girolami (2017). "Bayesian Probabilistic Numerical Methods". In: *CoRR*. arXiv: 1702.03673 [stat.ME].


Hennig, Philipp, Michael A Osborne, and Mark Girolami (July 2015). "Probabilistic numerics and uncertainty in computations". In: *Proc. R. Soc. A* 471.2179, p. 20150142.

References ii


Hennig, Philipp, Michael A. Osborne, and Hans P. Kersting (2022). *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press. DOI: 10.1017/9781316681411.


Oates, C. J. and T. J. Sullivan (2019). "A Modern Retrospective on Probabilistic Numerics". In: *CoRR*. arXiv: 1901.04457 [math.NA].


O'Hagan, A. (Nov. 1991). "Bayes-Hermite quadrature". In: *Journal of Statistical Planning and Inference* 29.3, pp. 245–260.