Fully Bayesian Optimisation

Juan Ungredda

University of Warwick

Outline

Background & Motivation

Including Input Uncertainty over Hyperparameters

- Impact on Bayesian Optimisation
- Results
- Several (Markov Chain Monte Carlo) MCMC approximations

Background: Gaussian Process Approximation

magenta Given a random variable that represents the value of the function $f(\mathbf{x})$ at location $\mathbf{x}$. A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$m(x) = \mathbb{E}[f(\mathbf{x})]$$
$$k_\theta(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(x))(f(\mathbf{x}') - m(x'))]$$

Usually, $m(x) = \mathbf{0}$ as prior with a user-defined kernel $k(\mathbf{x}, \mathbf{x}')$.

Background: Kernels

$k(\mathbf{x}, \mathbf{x}')$ imposes stronger preferences for certain types of functions, i.e. smooth or stationary functions, or functions with certain lengthscales.

Common choices of kernels $k_\theta(\mathbf{x}, \mathbf{x}')$.

- Square Exponential: $\sigma_f^2 exp\{-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^T\Sigma^{-1}(\mathbf{x}-\mathbf{x}')\}$
- Matern 5/2 Kernel: $\alpha(1+\sqrt{5}r+\frac{5}{3}r^2)exp\{-\sqrt{5}r\}$; where $r=\frac{||\mathbf{x}-\mathbf{x}'||_2}{l}$

Background: Prediction

Consider the possible designs $x \in X$, and a function $f\colon X \to \mathbb{R}$.

$$y(x) = f(x) + \epsilon, \text{ where } \epsilon \sim N(0, \sigma_\epsilon^2)$$

The posterior distribution of latent variables is[1]

$$p(f, f^*|y, \theta) = \frac{\overbrace{p(y|f)}^{Likelihood}\ \overbrace{p(f, f^*|\theta)}^{prior}}{\underbrace{p(y|\theta)}_{MarginalLikelihood}}$$

[1] f^* is the result of evaluating $f(x)$ at new design x^*

Background: Prediction

The posterior predictive distribution is,[2]

$$p(f^*|y,\theta) \propto \int \underbrace{p(y|f)}_{\text{Likelihood}} \underbrace{p(f,f^*|\theta)}_{\text{Prior}} df$$

Commonly, $p(y|f) \sim N(y; f, \sigma_\epsilon^2)$ and $p(f, f^*|\theta) \sim N(f, f^*; \mathbf{0}, k_\theta(\mathbf{x}, \mathbf{x}'))$

$$\begin{aligned} p(f^*|y,\theta) &= N(\mu^n, \Sigma^n), \text{ where} \\ \mu^n &= k_{*,f} k_{y,y}^{-1} y \\ \Sigma^n &= k_{*,*} - k_{*,f} k_{y,y}^{-1} k_{f,*} \end{aligned}$$

[2] Simplified notation: $k_\theta(\mathbf{x}, \mathbf{x}') = k_{f,f}$ and $k_{y,y} = k_{f,f} + \sigma_\epsilon^2$

Background: Point Estimation for Hyperparameters

- Maximum Likelihood (ML):

$$\hat{\theta} = arg \max_{\theta \in \Theta} \{ log(p(y|\theta)) \}$$

- Maximum a Posteriori (MAP):

$$E = log(p(\theta|y)) = log(p(y|\theta)) + log(p(\theta))$$
$$\hat{\theta} = arg \max_{\theta \in \Theta} \{ E \}$$

Motivation

Potential issues of point estimations,

- Multimodality
- Deceptive functions
- Uncertainty Underestimation

Motivation: Multimodality

- Likelihood may be multimodal. Solutions may converge to poor local maxima.

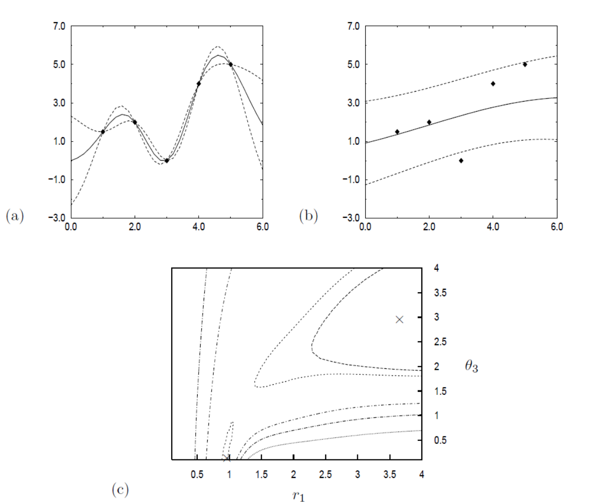


Figure 1: MacKay, D. (2002) "*Information Theory, Inference & Learning Algorithms*"

Motivation: Deceptive functions

- *Deceptive functions*: Describe functions that appear to be "flat" based on evaluation results.

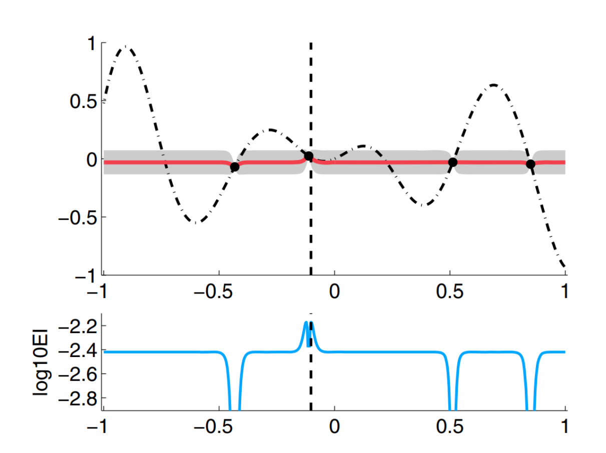
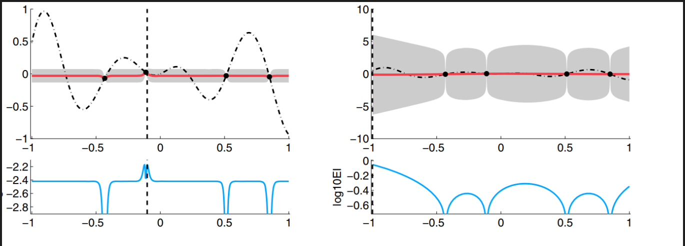


Figure 2: Benassi R., et al.(2011)

Motivation: Uncertainty Underestimation

- Standard deviation of the error of prediction is underestimated.



(a) parameters estimated by MLE (b) Bayesian approach for the parameters

Figure 3: Benassi R., et al.(2011)

Including Input Uncertainty over Hyperparameters

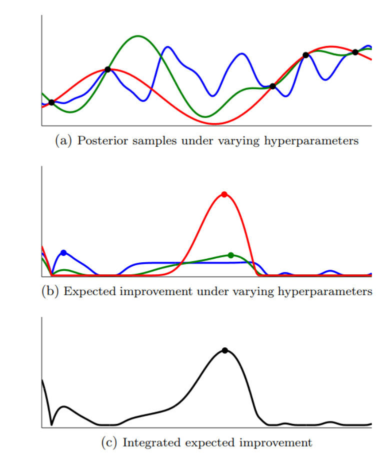


Figure 4: Snoek J., et al.(2012)

Impact on Bayesian Optimisation

Strategies for Optimisation:

- $EI(x)_{\theta ML}$: Use ML estimates for Expected Improvement.
- $EI(x)_{\theta True}$: Use True Hyperparameters.
- $\mathbb{E}_\theta[EI(x)]$: Marginalising Hyperparameters.

Performance metric:

$Opportunity\ Cost(OC) = \max\{f(x)\} - \max_{i=1,\ldots,n}\{y_i\}$
where,

- $f(x) =$ True function
- $y_i =$ sampled data

Results

content...

MCMC approximations

Algrthm	Parameters
Hamiltonian Monte Carlo (Y. Saatc, et al.(2010))	Leapfrog steps Leapfrog Δt
Slice Sampling (Murray, et al.(2010))	noise level S_θ
Sequential Monte Carlo (A. Svensson, et al. (2015))	Partition P MH-moves K Proposal distribution q
Bayesian Monte Carlo (Osborne M. A., et al (2008))	Hyperparameters of GP Approximation
Adaptive Importance Sampling (Petelin D., et al (2014))	Proposal distribution q

References

- Benassi R., Bect J., Vazquez E. (2011) "Robust Gaussian Process-Based Global Optimization Using a Fully Bayesian Expected Improvement Criterion. In: Coello C.A.C. (eds) Learning and Intelligent Optimization". LION 2011. Lecture Notes in Computer Science, vol 6683. Springer, Berlin, Heidelberg
- M. A. Osborne, S. J. Roberts, A. Rogers, S. D. Ramchurn, and N. R. Jennings, "Towards real-time information processing of sensor network data using computationally efficient multi-output Gaussian processes," in Proceedings of the 7th international conference on information processing in sensor networks, St. Louis, MO, USA, Apr. 2008, pp. 109–120.
- Murray, Iain and Adams, Ryan Prescott,"Slice sampling covariance hyperparameters of latent Gaussian models" Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2, 2010.
- Y. Saatc i, R. D. Turner, and C. E. Rasmussen, "Gaussian process change point models," in Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, Israel, Jun. 2010, pp. 927–934.
- D. Petelin, M. Gasperin, and V. Smidl, "Adaptive importance sampling for Bayesian inference in Gaussian process models," in Proceedings of the 19th IFAC World Congress, Cape Town, South Africa, Aug. 2014, pp. 5011–5015.

References

- A. Svensson, J. Dahlin and T. B. Schön, "Marginalizing Gaussian process hyperparameters using sequential Monte Carlo," 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Cancun, 2015, pp. 477-480.
- Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 2. Curran Associates Inc., USA, 2951-2959.