Heteroscedastic Bayesian Optimisation in Scientific Discovery

Ryan-Rhys Griffiths
Department of Physics
University of Cambridge
Cambridge, UK
rrg27@cam.ac.uk

Miguel Garcia-Ortegon
Department of Mathematics
University of Cambridge
Cambridge, UK
mg770@cam.ac.uk

Alexander A. Aldrick
Department of Physics
University of Cambridge
Cambridge, UK
av495@cam.ac.uk

Alpha A. Lee
Department of Physics
University of Cambridge
Cambridge, UK
aal44@cam.ac.uk

Abstract

We propose a Bayesian Optimisation problem in which the goal is to simultaneously optimise an objective whilst penalising aleatoric noise. Such problems are ubiquitous in the experimental natural sciences in areas such as drug and materials discovery. In our approach to the optimisation problem, we explicitly capture aleatoric uncertainty in the surrogate model using a heteroscedastic Gaussian Process. We extend the existing augmented expected improvement framework to the heteroscedastic setting and present a new acquisition function, aleatoric noise-penalised expected improvement (ANPEI) which applies a penalty to regions of the input space with large aleatoric noise. We demonstrate improvements over a standard Bayesian Optimisation scheme on one and two-dimensional toy experiments as well as a real-world problem where the goal is to identify soils with minimal phosphorus content.

1 Introduction

The design of Bayesian Optimisation schemes to find effective molecules and materials is an important problem in the natural sciences [1, 2]. In many settings, a secondary objective to a molecule or material's performance is its robustness to measurement noise (aleatoric noise) [3], a factor that can be crucial to enabling production at scale.

We present an approach to Bayesian Optimisation that explicitly attempts to minimise aleatoric uncertainty. Using the most likely heteroscedastic GP of [4] as the surrogate to model aleatoric uncertainty, we extend the augmented expected improvement (AEI) acquisition function of [5] to the heteroscedastic setting and introduce an alternative acquisition function, ANPEI, based on expected improvement that attempts to minimise aleatoric uncertainty in the sampled points. On toy examples as well as a real-world soil optimisation problem we demonstrate that our approach outperforms a naive implementation of Bayesian Optimisation based on a vanilla GP in conjunction with the expected improvement criterion. Lastly, we provide an open-source implementation of the most likely heteroscedastic GP of [4] as well as code for its use as a surrogate model for Bayesian Optimisation.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

2 Heteroscedastic Bayesian Optimisation

We wish to perform Bayesian Optimisation whilst minimising input-dependent aleatoric noise. In order to represent input-dependent aleatoric noise, a heteroscedastic surrogate model is required. We take the most likely heteroscedastic GP approach of [4], adopting the notation presented there for consistency. We have a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{n}$ in which the target values t_i have been generated according to $t_i = f(\boldsymbol{x}_i) + \epsilon_i$. We assume independent Gaussian noise terms $\epsilon_i \sim \mathcal{N}(0, \sigma_i)$ with variances given by $\sigma_i = r(\boldsymbol{x}_i)$ where r is the function that represents input-dependent noise. In order to perform Bayesian Optimisation, we wish to model the predictive distribution $P(\boldsymbol{t}^* \mid \boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_q^*)$ at the query points $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_q^*$. Placing a GP prior on f and taking $r(\mathbf{x})$ as the assumed noise rate function, the predictive distribution is multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}^*, \Sigma^*)$ with mean

$$\boldsymbol{\mu}^* = E[\boldsymbol{t^*}] = K^*(K + R)^{-1}\boldsymbol{t} \tag{1}$$

and covariance matrix

$$\Sigma^* = \text{var}[\boldsymbol{t^*}] = K^{**} + R^* - K^*(K + R)^{-1}K^{*T}, \tag{2}$$

where $K \in \mathbb{R}^{n\times n}$, $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $K^* \in \mathbb{R}^{q\times n}$, $K^*_{ij} = k(\boldsymbol{x}^*_i, \boldsymbol{x}_j)$, $K^{**} \in \mathbb{R}^{q\times q}$, $K^{**}_{ij} = k(\boldsymbol{x}^*_i, \boldsymbol{x}^*_j)$, $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^T$, $R = \text{diag}(\boldsymbol{r})$ with $\boldsymbol{r} = (r(\boldsymbol{x}_1), r(\boldsymbol{x}_2), \ldots, r(\boldsymbol{x}_n))^T$, and $R^* = \text{diag}(\boldsymbol{r}^*)$ with $\boldsymbol{r}^* = (r(\boldsymbol{x}^*_1), r(\boldsymbol{x}^*_2), \ldots, r(\boldsymbol{x}^*_q))^T$.

The Bayesian Optimisation problem may be framed as

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\chi}{\arg\min}\, f(\boldsymbol{x}), \tag{3}$$

where the black-box objective f, to be minimised has the form

$$f(\boldsymbol{x}) = g(\boldsymbol{x}) + s(\boldsymbol{x}). \tag{4}$$

$s(\boldsymbol{x})$ is, in this instance, the true noise rate function. We may write the expected improvement [6] in terms of the targets t and the incumbent best objective function value, η, found so far as

$$\text{EI}(\boldsymbol{x}) = \mathbb{E}\big[\,(\eta - t)_+\big] = \int_{-\infty}^{\infty} (\eta - t)_+\, p(t \,|\, \boldsymbol{x})\, dt \tag{5}$$

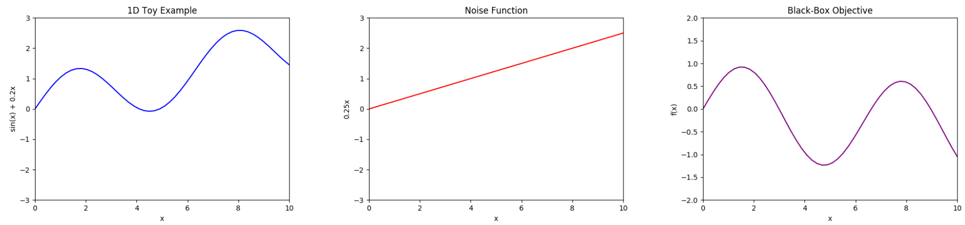
where $p(t \,|\, \boldsymbol{x})$ is the posterior predictive marginal density of the objective function evaluated at $\boldsymbol{x}$. $(\eta - t)_+ \equiv \max(0,\, \eta - t)$ is the improvement over the incumbent best objective function value η.

We propose two extensions to the expected improvement criterion. The first is an extension of the augmented expected improvement criterion of [5] to the heteroscedastic setting:

$$\text{het-AEI}(\boldsymbol{x}) = \mathbb{E}\big[(\eta - t)_+\big]\left(1 - \frac{\sqrt{r(\boldsymbol{x})}}{\sqrt{\text{var}[\mathbf{t}^*] + r(\boldsymbol{x})}}\right). \tag{6}$$

where $r(\boldsymbol{x})$ is the predicted aleatoric uncertainty at input $\boldsymbol{x}$ under the most likely heteroscedastic GP and var$[\mathbf{t}^*]$ is the predictive variance of the heteroscedastic GP incorporating both aleatoric and epistemic components of the uncertainty. We also propose a simple modification to the expected improvement acquisition function that explicitly penalises regions of the input space with large aleatoric uncertainty. We call this acquisition function aleatoric noise-penalised expected improvement (ANPEI) and denote it

$$\text{ANPEI} = \alpha\text{EI}(\boldsymbol{x}) - (1 - \alpha)\sqrt{r(\boldsymbol{x})} \tag{7}$$



(a) 1D Toy Objective (b) Noise Rate Function (c) Black-box Objective

Figure 1: Toy 1D Problem. The toy objective in a) is corrupted with heteroscedastic noise according to the function in b). The combined objective, that when optimised maximises the sin wave subject to the minimisation of aleatoric noise is given in c) and is obtained by subtracting the noise function from the 1D sinusoid.

where α is a scalarisation constant which we set to 0.5 for the experiments in this paper. In the multiobjective optimisation setting a particular value of α will correspond to a point on the Pareto frontier. We use both the modification to AEI (het-AEI) and ANPEI acquisition function in conjunction with the most likely heteroscedastic GP surrogate model in the experiments that follow.

3 Experiments

3.1 Implementation

Experiments were run using a custom implementation of Gaussian Process regression and most likely heteroscedastic Gaussian Process regression. The code is available at https://github.com/Ryan-Rhys/Heteroscedastic-BO. The squared exponential kernel was chosen as the covariance function for both the homoscedastic GP as well as G_1 and G_2 of the most likely heteroscedastic GP. The lengthscales, ℓ, of the homoscedastic GP were initialised to 1.0 for each input dimension across all toy problems after standardisation of the output values following the recommendation of [7]. The signal amplitude σ_f^2 was initialised to a value of 2.5. The lengthscale, ℓ, of G_2 of the most likely heteroscedastic GP [4] was initialised to 1.0, the initial noise level of G_2 was set to 1.0. The EM-like procedure required to train the most likely heteroscedastic GP was run for 10 iterations and the sample size required to construct the variance estimator producing the auxiliary dataset was 100. Hyperparameter values were obtained by optimising the marginal likelihood using the scipy implementation of the L-BFGS optimiser. The objective function in all cases is the principal objective $g(\boldsymbol{x})$ minus one standard deviation of the ground truth noise function $s(\boldsymbol{x})$.

3.2 1D Toy Objective with Linear Noise Rate Function

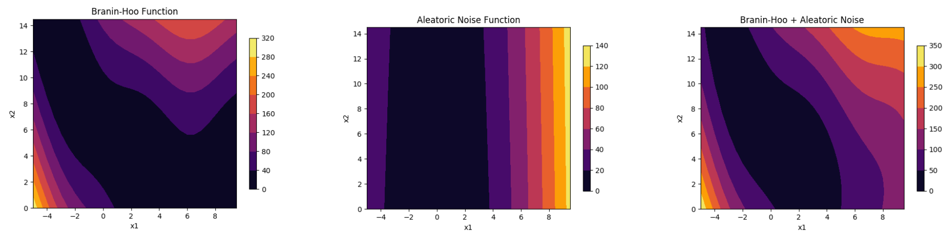
Referring to Equation 4 from section 2, in the first experiment we take a one-dimensional sin wave

$$g(x) = \sin(x) + 0.2(x) \tag{8}$$

with noise rate function $s(x) = 0.25x$. These functions as well as the black-box objective $f(x)$ are shown in Figure 1. The Bayesian Optimisation problem is designed such that the first maximum in 1(a) is to be preferred as samples from this region of the input space will have a smaller noise rate. The black-box objective in 1(c) illustrates this trade-off. In 3(a) we compare the performance of a Bayesian Optimisation scheme involving a vanilla GP in conjunction with the EI acquisition function with the most likely heteroscedastic GP in conjunction with the ANPEI and het-AEI acquisition functions. The experiment is designed to contrast the performance of a standard Bayesian Optimisation scheme against our approach in a situation where minimising aleatoric noise is desirable.

3.3 Branin-Hoo with Non-linear Noise Rate Function

In the second experiment, we consider the Branin-Hoo function as $g(\boldsymbol{x})$ with a non-linear noise rate function given by $s(\boldsymbol{x}) = 1.4{x_1}^2 + 0.3x_2$. Given that this example is a minimisation problem,



(a) 2D Branin-Hoo Function (b) Non-linear Noise Function (c) Black-box Objective

Figure 2: Toy 2D Problem. The Branin-Hoo objective function in a) is corrupted by the heteroscedastic noise function in b) $s(x_1, x_2) = 1.4{x_1}^2 + 0.3x_2$. The black-box objective function c) is obtained by summing the functions in a) and b). The sum is required to penalise regions of large aleatoric noise because the objective is being minimised.

the black-box objective consists of the sum of the Branin-Hoo function and the noise rate function. Contour plots of the functions are shown in Figure 2. A comparison, in terms of the best objective function values found, between the vanilla GP and EI acquisition function with the most likely heteroscedastic GP and ANPEI and AEI acquisition functions is given in 3(b).

3.4 Optimising the Phosphorus Fraction of Soil

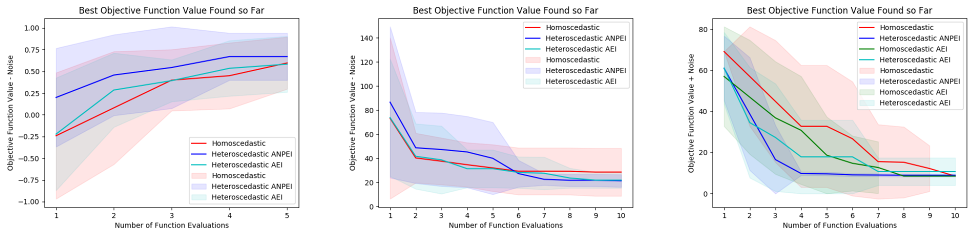
In this real-world problem, we apply heteroscedastic Bayesian optimisation to the task of optimising the phosphorus fraction of soil. Soil phosphorus is an essential nutrient for plant growth and is widely used as a fertiliser in agriculture. While the amount of arable land worldwide is declining, global population is expanding and so is food demand. As such, understanding the availability of plant nutrients that increase crop yield is essential. To this end, reference [8] shows a curated dataset on soil phosphorus, relating phosphorus content to variables including soil particle size, total nitrogen, organic carbon and bulk density. In this experiment, we study the relationship between bulk soil density and the phosphorus fraction, the goal being to minimise the phosphorus content of soil subject to heteroscedastic noise. We provide evidence that there is heteroscedasticity in the problem by comparing the fits of a homoscedastic GP and the most likely heteroscedastic GP in Figure 4 and provide a predictive performance comparison based on negative log predictive density values in the appendix. In this problem, we do not have access to a continuous-valued black-box function or a ground truth noise function. As such, the surrogate models were initialised with a subset of the data and the query locations selected by Bayesian Optimisation were mapped to the closest data points in the held-out data. The following kernel smoothing procedure was used to generate pseudo ground truth noise values:

1. Fit a homoscedastic GP to the full dataset.
2. At each point x_i, compute the corresponding square error $s_i^2 = (y_i - \mu(x_i))^2$.
3. Estimate variances by computing a moving average of the squared errors, where the relative weight of each s_i^2 was assigned with a Gaussian kernel.

The performances of homoscedastic Bayesian Optimisation using EI and AEI and heteroscedastic Bayesian Optimisation using ANPEI and AEI are compared in 3(c).

3.5 Discussion

In all experiments the most likely heteroscedastic GP and ANPEI combination/heteroscedastic AEI combination outperform the homoscedastic GP and EI. The fact that the homoscedastic GP has no knowledge of the heteroscedasticity of the noise rate function puts it at a serious disadvantage. In the sin wave problem, designed to highlight this point, the heteroscedastic Bayesian Optimisation scheme consistently and preferentially finds the first maximum as that which minimises aleatoric noise. In contrast, the homoscedastic GP finds it impossible to differentiate between the two maxima. The experiments provide strong evidence that modelling heteroscedasticity in Bayesian optimisation is a more flexible approach to assuming homoscedastic noise.



(a) 1D Toy Objective (b) 2D Branin-Hoo (c) Soil Phosphorus Fraction

Figure 3: Results of heteroscedastic and homoscedastic Bayesian Optimisation on the 3 problems considered. Error bars are computed using 10 random initialisations.

4 Conclusion

We have presented an approach for performing Bayesian Optimisation with the explicit goal of minimising aleatoric uncertainty in the suggestions. We posit that such an approach can prove useful for the natural sciences in the search for molecules and materials that are robust to experimental noise. We demonstrate concrete improvements on one and two-dimensional toy problems as well as a real-world optimisation problem and contribute an open-source implementation of the most likely heteroscedastic GP as a surrogate model for Bayesian optimisation. In future work, we plan to apply our approach to molecular property optimisation [9]. To this end, it would be useful for investigate the potential for replicates to improve the heteroscedastic BO scheme [10, 11, 12]. Furthermore, it may be possible to leverage recent advances in combinatorial Bayesian optimisation [13, 14, 15] in order to perform heteroscedastic Bayesian optimisation over molecular graphs.

References

[1] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[2] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design. *arXiv preprint arXiv:1709.05501*, 2017.

[3] Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha A. Lee. Dataset bias in the natural sciences: A case study in chemical reaction prediction and synthesis design. *ChemRxiv*, 2018.

[4] Kristian Kersting, Christian Plagemann, Patrick Pfaff, and Wolfram Burgard. Most likely heteroscedastic Gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning - ICML '07*, pages 393–400, Corvalis, Oregon, 2007. ACM Press. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273546.

[5] Deng Huang, Theodore T Allen, William I Notz, and Ning Zeng. Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of global optimization*, 34(3):441–466, 2006.

[6] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[7] Iain Murray. Introduction to Gaussian processes. 2008.

[8] Enqing Hou, Xiang Tan, Marijke Heenan, and Dazhi Wen. A global dataset of plant available and unavailable phosphorus in natural soils derived by Hedley method. *Scientific Data*, 5(1): 180166, December 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.166.

[9] Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric P Xing. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. *arXiv preprint arXiv:1908.01425*, 2019.

[10] Alexis Boukouvalas and Dan Cornford. Learning heteroscedastic Gaussian processes for complex datasets. *Technical report*, 2009.

[11] Mickael Binois, Robert B Gramacy, and Mike Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.

[12] Mickaël Binois, Jiangeng Huang, Robert B Gramacy, and Mike Ludkovski. Replication or exploration? sequential design for stochastic simulation experiments. *Technometrics*, 61(1): 7–23, 2019.

[13] Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *International Conference on Machine Learning*, pages 462–471, 2018.

[14] Poompol Buathong, David Ginsbourger, and Tipaluck Krityakierne. Kernels over sets of finite sets using RKHS embeddings, with application to Bayesian (combinatorial) optimization. *arXiv preprint arXiv:1910.04086*, 2019.

[15] Changyong Oh, Jakub M Tomczak, Efstratios Gavves, and Max Welling. Combinatorial Bayesian optimization using graph representations. *arXiv preprint arXiv:1902.00448*, 2019.

[16] Roberto Calandra. *Bayesian Modeling for Optimization and Control in Robotics*. PhD thesis, Technische Universität, 2017.

[17] Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic Gaussian process regression. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 841–848. Omnipress, 2011.

[18] Scott R Kuindersma, Roderic A Grupen, and Andrew G Barto. Variable risk control via stochastic optimization. *The International Journal of Robotics Research*, 32(7):806–825, 2013.

[19] John-Alexander M Assael, Ziyu Wang, Bobak Shahriari, and Nando de Freitas. Heteroscedastic treed Bayesian optimisation. *arXiv preprint arXiv:1410.7172*, 2014.

[20] Ryo Ariizumi, Matthew Tesch, Howie Choset, and Fumitoshi Matsuno. Expensive multiobjective optimization for robotics with consideration of heteroscedastic noise. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2230–2235. IEEE, 2014.

[21] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning*, pages 997–1005, 2015.

[22] Felix Berkenkamp, Andreas Krause, and Angela P. Schoellig. Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. *ArXiv*, abs/1602.04450, 2016.

[23] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.

[24] Benjamin Letham, Brian Karrer, Guilherme Ottoni, Eytan Bakshy, et al. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

A Heteroscedasticity of Soil Phosphorus Fraction Dataset

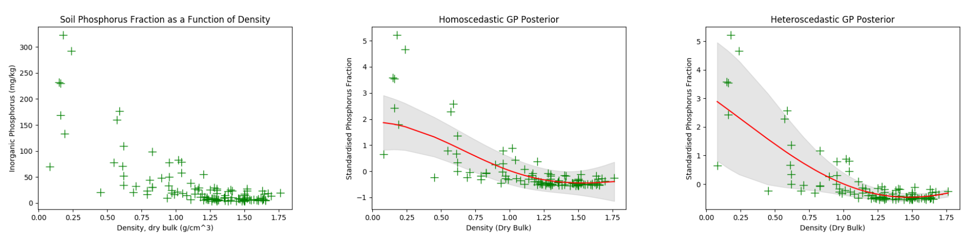
The purpose of Figure 4 and Table 1 is to demonstrate the efficacy of modelling the soil phosphorus fraction dataset using a heteroscedastic GP. The heteroscedastic GP outperforms the homoscedastic GP on prediction based on the metric of negative log predictive density (NLPD)

$$\text{NLPD} = \frac{1}{n}\sum_{i=1}^{n} -\log p(t_i|\boldsymbol{x_i}) \tag{9}$$

which penalises both over and under-confident predictions.

B Related Work

The most similar work to our own is that of [16] where experiments are reported on a heteroscedastic Branin-Hoo function using the variational heteroscedastic GP approach of [17] although to the best of our knowledge this work does not consider sequential evaluations. A modification to EI, expected risk improvement is introduced in [18] and is applied to problems in robotics where robustness to aleatoric noise is desirable. [19, 20] implement heteroscedastic Bayesian Optimisation but don't introduce an acquisition function that penalises aleatoric uncertainty. [21, 22] consider the related problem of safe Bayesian Optimisation through implementing constraints in parameter space. In this instance, the goal of the algorithm is to enforce a performance threshold for each evaluation of the black-box function and so is unrelated to our problem definition. In terms of acquisition functions, [23, 24] propose principled approaches to handling aleatoric noise in the homoscedastic setting that could be extended to the heteroscedastic setting. Our primary focus in this work however, is to highlight that heteroscedasticity in the surrogate model is beneficial and so an examination of a subset of acquisition functions is sufficient for this purpose.



(a) Dataset (b) Homoscedastic GP Fit (c) Heteroscedastic GP Fit

Figure 4: Comparison of Homoscedastic and Heteroscedastic GP Fits to the Soil Phosphorus Fraction Dataset.

Table 1: Comparison of NLPD values on the soil phosphorus fraction dataset. Standard errors are reported for 10 independent train/test splits.

Soil Phosphorus Fraction Dataset	**GP**	**Het GP**
NLPD	1.35 ± 1.33	1.00 ± 0.95