Automated Machine Learning using Stochastic Algorithm Tuning

Thomas Nickson, Michael A Osborne, Steven Reece and Stephen Roberts
MLRG
Department of Engineering Science
University of Oxford
{tron,mosb,reece,sjrob}@robots.ox.ac.uk

Abstract

More often than not, the critical tuning of ML algorithm parameters has relied on domain expertise, along with laborious hand-tuning, exhaustive search or lengthy sampling runs. Against this background, Bayesian optimisation is finding increasing use in automating parameter tuning, making ML algorithms accessible even to non-experts. The state of the art in Bayesian optimisation is incapable of scaling to the large number of evaluations of algorithm performance required to fit realistic models to complex data. We solve this problem with a stochastic, sparse Bayesian optimisation strategy, using a sparse Gaussian process (GP) and many thousands of noisy evaluations of algorithm performance to train previously intractable models.

1 Introduction

In this paper we present a novel extension of Bayesian optimisation (BO) to intractable optimisation problems made possible by two techniques:

Scalable BO: Firstly, we propose a new BO algorithm that makes use of advances in sparse GPs to deliver more benign scaling. This enables the tackling of challenging algorithm configuration tasks and provides non-trivial acceleration on generic BO problems, while retaining the powerful probabilistic machinery of the GP, which provides better performance than the empirical mean and variance of random forest regressors (RFRs) [1, 2].

Stochastic BO: Our second contribution is introducing the use of BO strategies to configure algorithms where performance evaluations are either uncertain or stochastic. This contribution is significant when considering the fitting of models to large amounts of data. The state of the art for such tasks relies upon evaluations of model fit on stochastically selected subsets of data, giving rise to algorithms including stochastic optimisation [3], stochastic gradient Langevin dynamics [4] and stochastic variational inference [5]. Within our BO framework, such evaluations on subsets are simply treated as evaluations of a latent performance curve corrupted by noise. Unlike existing stochastic approaches, our BO strategy uses these noisy objective evaluations to construct an explicit surrogate model even when *gradients are unavailable* (as is often the case for black-box algorithms) or are excessively expensive.

Crucially, the sparse GPs used in scalable BO scale as $\mathcal{O}(N)$ (where N is the number of observations) as compared to $\mathcal{O}(N^3)$ for the full GP, making them comparable to the efficient RFR. This allows very large numbers of observations to be made to mitigate uncertainty introduced by the noisy measurements of the objective in stochastic BO.

2 Bayesian Optimisation with Noisy Observations

We use a GP as the surrogate for the latent performance curve. Coupled with this probabilistic surrogate is an acquisition function $\lambda(x \mid \mathcal{D})$ (often interpretable as an expected loss function), used as a means of choosing each successive function evaluation.

There are a wide range of options for the choice of acquisition function [6]. In this work, given our goal of managing noisy likelihood evaluations, we use the noise-tolerant expected improvement acquisition function of [7]. Specifically, defining $y_+ = y(x_+)$, $f_* = f(x_*)$ and ν as an appropriately small threshold, we choose

$$\lambda(x_+ \mid \mathcal{D}) \coloneqq \eta \int_{\eta}^{\infty} p(y_+ \mid \mathcal{D})\, \mathrm{d}y_+ + \int_{\infty}^{\eta} y_+\, p(y_+ \mid \mathcal{D})\, \mathrm{d}y_+, \eta \coloneqq \min_{x_* : \mathbb{V}\left[p(f_* \mid \mathcal{D})\right] < \nu^2} \mathbb{E}\left[p(f_* \mid \mathcal{D})\right]. \tag{1}$$

That is, our acquisition function is the expected lowest function value about which we are sufficiently confident (with confidence specified by ν) after evaluating at x_+.

3 Sparse Bayesian Optimisation on a Mixture of Gaussians

We tested a full GP, a Laplacian-GP [8], a spectral approximation to the covariance k [9], the fully independent training conditional (FITC) approximation [10, 11] and the RFR on a toy mixture of gaussians with 5 local minima and one global minimum in $[-5, 5]$. We "primed" each regressor with two randomly sampled points. The convergence is shown in Figure 1. This is the mean for 5 sample runs.

The posterior of the GP (and its various sparse approximations) allows the regressor to be bootstrapped from a minimal starting set of sample points. The initial uncertainty strongly promotes exploration. This is in contrast to the RFR, which can be excessively certain in sparsely sampled or flat regions. Incorrect estimates of the hyperparameters can cause poor fitting of the function in all of the GP models, causing exploration or exploitation to proceed incorrectly. The FITC approximation is most affected by this due to the additional optimisation required for the inducing inputs which may cause it to under- or over-estimate the variance, as shown by its slower convergence in Figure 1.

We have found little difference in the performance between the FITC, Laplacian and spectral approximations (the Laplacian model is slightly more efficient in one dimension, while the FITC is more prone to incorrectly estimating the variance if the inducing inputs are incorrectly located, causing the delayed convergence shown in Figure 1). We will concentrate on the spectral approximation henceforth, due to its better scaling with dimensionality and more reliable variance.

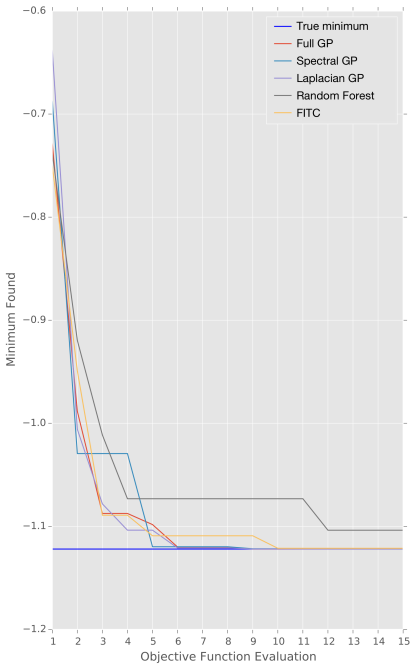

Figure 1: Convergence of BO using different regressors as surrogates to the true minimum.

4 Stochastic Bayesian Optimisation

Stochastic inference is a technique where the likelihood is evaluated on subsets of the data. Methods in this family include stochastic optimisation [3], stochastic gradient Langevin dynamics [4] and stochastic variational inference [5, 12]. These methods inspire *Stochastic Bayesian Optimisation,* henceforth known as STOchastic Algorithm Tuning (STOAT). Here, we make noisy observations of the likelihood of a large machine learning (ML) model by evaluating it on subsets of the data. We

use the probabilistic power of the GP and the large-data capabilities of the spectral approximation to make many observations of the likelihood with different subsets of the data at each step of the BO algorithm. Unlike existing stochastic optimisation approaches, we do not make use of gradient observations, allowing us to consider real, black-box algorithm configurations for which gradients are unavailable or excessively expensive. Our approach also permits the global exploration of complex likelihood surfaces, reducing the risk presented by local minima.

4.1 Optimisation of the noise corrupted Branin function

To test the hypothesis that we can perform global optimisation (GO) with noisy observations, we optimised the Branin function with the evaluations corrupted by Gaussian noise with $\sigma^2 = 5$, which is very large compared to the range of function values around the Branin's three minima. Each experiment was limited to 1000 seconds of wall-clock time. At each step we made 50 observations of the noisy Branin function. These were passed to the GP. With 50 noisy evaluations per BO evaluation, and 20-30 seconds per evaluation, the GP gathered between 1,500 and 2,500 samples. In addition, we found it beneficial to "prime" the GP with 400 points from a Sobol sequence, to allow it to concentrate more closely on exploring low regions of the space and learn hyper-parameters from a larger initial set. In the most extreme case STOAT efficiently performed BO with nearly 3,000 data points.

We used the 'Gap' measure of performance to compare performance of STOAT and standard BO [13]: $G \coloneqq \left(y\left(x^{\text{init}}\right) - y\left(x^{\text{best}}\right)\right) / \left(y\left(x^{\text{init}}\right) - y\left(x^{\text{opt}}\right)\right)$. Table 1 shows the maximum gap, mean gap, and the best minimum found by STOAT and the standard BO algorithm. We also tested the covariance matrix adaptive evolution strategy (CMAES) [14] and dividing rectangles (DIRECT) [15] methods, however, we found that these did not converge on this very noisy objective.

Method	Max Gap	Mean Gap	Best minimum found *(true minima)*
BO	0.926	0.831	2.15 *(0.398)*
Stochastic BO	**0.997**	**0.965**	**0.472** *(0.398)*

Table 1: Performance of the algorithms minimising the Branin function with observation noise.

5 Stochastic Bayesian Configuration of Model Parameters on Energy Data

AgentSwitch [16] is a project that aims to assist people in selecting the most economical energy tariff for their expected electricity use. A GP model is used to predict the energy that will be used by a household. The posterior prediction of this GP is used to inform group bidding for energy tariffs, to ensure that a user pays the lowest price for their power. The posterior variance is particularly useful here, because it allows the group to understand the risk of going for a cheaper, fixed use contract compared to a more expensive flexible plan. To replicate the work in [16] we used household power use data from UCI[1].

The dataset contains seven features describing the average power use in the house in different ways, in addition to the date and time. The data was gathered at a resolution of one sample per minute. We down-sampled this to hourly averages, and selected just the 'global active power' (power actually used throughout the household). In total we had 34,166 observations, with a single input dimension. We retained the last 5,000 entries for testing, and did not use these for learning the hyper-parameters. Our training set consisted of 29,166 samples. Computing the full likelihood on this data is impractical on reasonable computing hardware, requiring 50 gigabytes of RAM simply to *store* the Gram matrix.

The authors of [16] used a periodic GP with a exponentiated quadratic (EQ) kernel, and fixed the periodicity hyper-parameter a-priori to one day and fit the other parameters using multistart type 2 maximum likelihood estimation (MLE-II) on subsets of the data. From inspection of the log-likelihood surface and the spectrum, we can see that there is a strong periodic component around

[1] http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

one year, however, the likelihood for smaller periods is highly multi-modal. To adequately explore this surface would require multiple restarts at different initial hyper-parameter values, which quickly approaches the computational burden of true BO. The authors of [16] chose not to optimise their period with MLE-II for this reason and set it to a fixed number; STOAT automatically balances exploration and exploitation, and can learn the dual periodic model on this in comparable time to a multi-start MLE-II optimisation *learning only the non-periodic hyper-parameters*.

We used STOAT to learn the periods of a sum of periodic squared exponential (SE) kernels on this data. We constrained the periods to lie within $[0.1, 10] \times [10, 1000]$. Our noisy likelihood observations were evaluated on 1,000 samples from the training data. At each step we made 10 of these stochastic observations (each on a different random subset of the training data without replacement). We ran the experiments on a laptop computer with a 2.3 GHz Intel i7 CPU. In addition to the 10 stochastic samples at each iteration, before beginning our BO sequence we generated 600 space filling points from a Sobol sequence and evaluated the noisy likelihood at each of these.

The pre-sampling allowed the GP to quickly become certain about the hyper-parameters, which reduced wasted exploration steps. The pre-sample step took around 5 minutes, however once it was complete the larger period quickly converged to 382 days, approximately one year. The shorter period oscillated between 1.3 and 1.5 days. After one hour, the optimiser had converged to 1.5. For comparison to [16], we also optimised a single-periodic kernel using STOAT. After the pre-sample, this converged after a few steps to a period of 378 days. Marginalisation of the hyper-parameters as recommended in [7] (Bayesian Quadrature (BQ)) or [17] (Markov chain Monte-Carlo (MCMC)) may reduce the need to pre-sample, at the expense of additional computational load. Each iteration of the BO algorithm took between 20 and 30 seconds to complete. Including the pre-sample, the number of measurements made was between 1,500 and 2,500 with no noticeable slow down as N grew.

Method	Test data log-likelihood
STOAT learned double periodic	**-7.25**
STOAT learned single periodic	-7.39
AgentSwitch a-priori single periodic	-7.40
A-periodic GP	-9.22

Table 2: Log-likelihood on real held out electricity use data of models learned using our method, a-priori setting of periods and naïve a-periodic GP.

To test our results, we compared the predictive log-likelihoods (on the held out test-data) of our dual periodic and single periodic kernels, [16]'s single periodic kernel and a simple, a-periodic SE kernel. The results in Table 2 show the model using the parameters tuned by our algorithm outperforming both the model from [16] and the aperiodic GP. Additionally, our algorithm tuner's ability to quickly find the second period at 1.5 days substantially improves the predictive performance when compared to simpler models, even when searching the highly multi-modal first dimension.

6 Conclusion

Using STOAT on a consumer grade laptop, we have quickly optimised the parameters of an ML algorithm of such computational complexity that *we cannot evaluate the likelihood on the full data*. On real, noisy data our algorithm quickly converges to the large global optimum in one dimension, and in two dimensions is able to find a second optimal location amongst many nearby local optima.

We extend the principled exploration of expensive functions developed in the BO and Sequential Model Based Optimisation (SMBO) literature to allow noisy observations of an objective function. In real machine learning problems with computationally intractable likelihoods, we are able to find the global optimum by evaluating the likelihood on subsets of the data, relying on the information handling properties of our sparse GP to allow for this additional noise.

Acknowledgments

This work is supported by the EPSRC funded ORCHID Project EP/I011587/1.

References

[1] Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. "Sequential Model-Based Optimization for General Algorithm Configuration". In: *Proc. of LION-5*. 2011, pp. 507–523.

[2] Robert B Gramacy, Matt Taddy, Stefan M Wild, et al. "Variable selection and sensitivity analysis using dynamic trees, with an application to computer code performance tuning". In: *The Annals of Applied Statistics* 7.1 (2013), pp. 51–80.

[3] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". English. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407. ISSN: 00034851.

[4] Max Welling and Yee W Teh. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.

[5] Matthew D Hoffman et al. "Stochastic variational inference". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 1303–1347.

[6] Eric Brochu, Vlad M Cora, and Nando De Freitas. "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning". In: *arXiv preprint arXiv:1012.2599* (2010).

[7] Michael A Osborne, Roman Garnett, and Stephen J Roberts. "Gaussian processes for global optimization". In: *3rd international conference on learning and intelligent optimization (LION3)*. 2009, pp. 1–15.

[8] Arno Solin and Simo Sarkka. "Hilbert Space Methods for Reduced-Rank Gaussian Process Regression". In: (2014). arXiv: arXiv:1401.5508v1.

[9] Steven Reece et al. "Efficient State-Space Inference of Periodic Latent Force Models". In: *Journal of Machine Learning Research* (2014). arXiv: 1319.6319v2.

[10] A Naish-Guzman and SB Holden. "The Generalized FITC Approximation." In: *NIPS* (2007).

[11] Edward Snelson and Z Ghahramani. "Sparse Gaussian processes using pseudo-inputs". In: (2006).

[12] James Hensman, Nicolo Fusi, and Neil D Lawrence. "Gaussian Processes for Big Data". In: *UAI* (2013).

[13] D. Huang et al. "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models". In: *Journal of Global Optimization* 34.3 (Mar. 2006), pp. 441–466. ISSN: 0925-5001. DOI: 10.1007/s10898-005-2454-3.

[14] Nikolaus Hansen and Andreas Ostermeier. "Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation". In: *ICEC* (1996), pp. 312–317.

[15] D. R. Jones, C. D. Perttunen, and B. E. Stuckmann. "Lipschitzian optimization without the lipschitz constant". In: *Journal of Optimization Theory and Application* 79 (1993), pp. 157 –181.

[16] SD Ramchurn et al. "Agentswitch: Towards smart energy tariff selection". In: *Autonomous Agents and Multi-Agent Systems* (2013).

[17] Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams. "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. 2012.