
Automated Machine Learning on Big Data 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

We introduce a means of automating machine learning (ML) for big data tasks, by performing scalable stochastic Bayesian optimisation of ML algorithm parameters and hyper-parameters. More often than not, the critical tuning of ML algorithm parameters has relied on domain expertise from experts, along with laborious hand-tuning, brute 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. However, 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, big data. We here describe a stochastic, sparse, Bayesian optimisation strategy to solve this problem, using many thousands of noisy evaluations of algorithm performance on subsets of data in order to effectively train algorithms for big data. We provide a comprehensive benchmarking of possible sparsification strategies for Bayesian optimisation, concluding that a Nyström approximation offers the best scaling and performance for real tasks. Our proposed algorithm demonstrates substantial improvement over the state of the art in tuning the parameters of a Gaussian Process time series prediction task on real, big data.

1 Introduction

The performance of many machine learning algorithms is highly sensitive to the learning of model structure, parameters and hyperparameters [1]. Examples include parameters specifying model capacity (numbers of layers), learning rates and kernel parameters (bandwidths). Selecting these parameters has traditionally required deep domain expertise, brute-force search and/or laborious hand-tuning. As such, as the demand for machine learning algorithms grows faster than the supply of machine learning experts, there is an increasing need for methods to automatically configure algorithms to the task at hand, even where the task at hand is complex and algorithm performance (such as cross-validation error, or likelihood) expensive to evaluate. Ultimately, we aim to deliver off-the-shelf algorithms without compromising model quality: This is the challenge of *automated machine learning*.

One approach to automated machine learning is the use of a global optimisation algorithm to automatically optimise the performance of a supplied algorithm as a function of its parameters [2]. One popular choice of global optimiser for this purpose is Bayesian optimisation (BO) [3], providing a robust means of exploring complex, multi-modal, performance surfaces. BO has been employed successfully in the configuration of deep neural network hyperparameters [4, 2], a complex vision architecture [5], and in Auto-WEKA [6], a framework incorporating a diverse range of classification algorithms. Most applications of BO make use of a Gaussian process (GP) surrogate for the objective function, enabling flexible non-parametric modelling of performance that provides principled

representations of uncertainty to guide exploration. However, the $\mathcal{O}(N^3)$ scaling of GP inference in the number N of function evaluations has prevented the extension of these methods to many real problems of interest. Fitting complex models to real data often requires prohibitively large numbers of performance evaluations in order to effect the optimisation of model parameters. We make two central contributions towards these challenges.

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. We provide explicit comparisons against the BO state of the art for managing large numbers of evaluations, random forest regressors (RFRs) [7, 8]. Specifically, we present evidence that RFRs provide misleading uncertainty estimates that hinder exploratory optimisation relative to our algorithm.

Stochastic BO: Our second contribution is introducing the use of BO strategies for noisy optimisation to configure algorithms where performance evaluations are either uncertain or stochastic. This contribution is significant when considering the fitting of models to big 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 [9], stochastic gradient Langevin dynamics [10] and stochastic variational inference [11]. 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. Coupled with our scalable models, we can take sufficient stochastic evaluations to explore and optimise for real, multi-modal, algorithm configuration problems.

2 Bayesian Optimisation

We frame our central problem, algorithm configuration, as one of global optimisation (GO). That is, we view the performance of a machine learning algorithm (such as its cross-validation error, or likelihood) as an expensive function of the algorithm’s parameters, and apply an optimiser to this function in order to find the best settings of its parameters. GO is required to explore typically non-convex, complex parameter spaces using a parsimonious number of function evaluations.

Bayesian optimisation [3, 12] applies probabilistic modelling to global optimisation. Explicitly, we define the GO task as finding the minimum of an objective function $f(x)$ on some bounded set $\mathcal{X} \subset \mathbb{R}^d$, given only noise-corrupted observations of the objective, $y(x) = f(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We consider the setting in which the expense of evaluating $y(x)$, and the unavailability of gradients of y , motivate the use of sophisticated means of iteratively selecting function evaluations. Specifically, we use a probabilistic model $p(f, y)$ as a surrogate for f , allowing evaluations from across the domain \mathcal{X} , and the uncertainty inherent in y , to inform future selections. We define the set of information available after the n th function evaluation as $\mathcal{D}_n := \{x_i, y(x_i) \mid i = 1, \dots, n\}$.

Coupled with the 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, as $x_{i+1} = \arg \max_{x_+} \lambda(x_+ \mid \mathcal{D})$. While this introduces a new optimisation problem, acquisition functions are chosen to admit observations of the gradient and Hessian, and to be trivially cheap to evaluate relative to the cost of the objective itself. There are a wide range of options for the choice of acquisition function [3]. In this work, given our goal of managing noisy likelihood evaluations, we use the noise-tolerant expected improvement acquisition function of [13]. Specifically, defining $y_+ = y(x_+)$, $f_* = f(x_*)$ and ν as an appropriately small threshold, we choose

$$\lambda(x_+ \mid \mathcal{D}) := \eta \int_{\eta}^{\infty} p(y_+ \mid \mathcal{D}) dy_+ + \int_{\infty}^{\eta} y_+ p(y_+ \mid \mathcal{D}) dy_+; \quad \eta := \min_{x_*: \mathbb{V}[p(f_* \mid \mathcal{D})] < \nu^2} \mathbb{E}[p(f_* \mid \mathcal{D})]. \quad (1)$$

That is, our acquisition function is the expected lowest function value about which we are sufficiently confident in (with confidence specified by ν) after evaluating at x_+ . Below, we will describe our extension of BO to robustly manage many function evaluations, and to perform stochastic optimisation.

3 Sparse Regression

3.1 Sparse Models

The GP is widely used in regression, classification and other machine learning algorithms as a prior over functions due to its conceptual simplicity and full Bayesian treatment of uncertainty, providing not just a prediction of a value but also an associated estimate of the uncertainty. A GP's behaviour is defined by its covariance function $k(\mathbf{x}, \mathbf{x}')$. The function k for $x, x' \in \mathcal{X}$ maps $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, and is known as a kernel. For two input or feature matrices $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{X}' \in \mathbb{R}^{m \times d}$ the $n \times m$ covariance matrix $K(\mathbf{X}, \mathbf{X}')$ is the Gram matrix made by the pairwise mapping of $k(\cdot, \cdot)$ to each row in \mathbf{X} and \mathbf{X}' (each row corresponding to a feature vector). For Gaussian noise (with variance σ^2) corrupted values \mathbf{y} observed at \mathbf{X} from a function $f \sim \mathcal{GP}(\mathbf{0}, K)$, we can express the distribution over \mathbf{y} as $p(\mathbf{y} | \mathbf{X}) = \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K} + \mathbf{I}\sigma^2)$. Predictions \mathbf{f}_* at inputs \mathbf{X}_* using the GP are available in closed form [14, Appendix B]. The elegance of the mathematics underlying the prediction and marginalisation of the GP comes at computational cost. Evaluation of the likelihood of the model scales as $\mathcal{O}(n^3)$ for n samples, and prediction scales as $\mathcal{O}(pn^2)$ for p predictions.

There is much prior art on improving the scaling of a GP with n . Most are based on reducing the rank of the covariance matrix, rendering the matrix inversion $\mathcal{O}(m^3)$, where m is the new rank.

FITC: Perhaps the most widely used method is the fully independent training conditional (FITC) method [15]. This method uses an inducing matrix $\tilde{\mathbf{X}}$ (whose rows are feature vectors describing the locations of *inducing points*) with associated latent values \mathbf{u} to restrict the bandwidth of the kernel, forcing information exchange between the training and test data to pass through these points rather than the infinite bandwidth link of a full GP. In the implementation used in this paper, we select a set of inducing points on a linear grid, optimise the GP hyper-parameters, then jointly optimise the hyper-parameters and the inducing inputs. We found this to give superior results to performing the joint optimisation initially, and to be more accurate than simply fixing the inducing points.

Nyström-GP: A similar method to FITC is the Nyström approximation [16]. A kernel k can be expressed by an infinite weighted sum of orthonormal bases $k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{\infty} \mu_j^\phi \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$ where ϕ_j and μ_j are the j th eigenfunctions and eigenvalues (henceforth 'eigenpairs' when considered together) of the kernel under a distribution p such that $\int k(\mathbf{x}, \mathbf{x}') \phi_j(\mathbf{x}') p(\mathbf{x}') d\mathbf{x}' = \mu_j^\phi \phi_j(\mathbf{x})$. It is rarely feasible to evaluate an infinite sum, so we consider only the m most significant eigenpairs, which has the effect of reducing the frequency response of the kernel. Smoother kernels such as the squared exponential (SE) require fewer components than rougher ones such as the Matérn $\frac{3}{2}$.

Following the recommendations in [16], in the implementation used in this paper we maintain an area of interest \mathcal{S} in which we wish to approximate the kernel and a representative sample set $\mathbf{S} \in \mathcal{S}$ of cardinality L . For a stationary kernel as used in this paper uniform sampling is acceptable, however for non-stationary kernels denser sampling in regions with shorter lengthscales is advised. We use a uniform probability density over \mathcal{S} because we have no prior knowledge of the importance of different regions. We construct the Gram matrix by pairwise evaluation of the full kernel k on the elements of \mathbf{S} , and find the eigenpairs (μ and \mathbf{v}). The eigenpairs of the Gram are used to construct the approximate eigenfunctions $\tilde{\phi}_j(\mathbf{x}) = \sqrt{L/\mu_j} K(\mathbf{x}, \mathbf{S}) \mathbf{v}_j$ and values $\mu_j^\phi = \mu_j/L$. Using these approximate eigenpairs we can express the truncated approximation as $k(\mathbf{x}, \mathbf{x}') \approx \sum_{j=1}^m \mu_j^\phi \tilde{\phi}_j(\mathbf{x}) \tilde{\phi}_j(\mathbf{x}')$. We refer to this approximation as the *Nyström-GP*.

Laplacian-GP: A similar result to [16] is arrived at in [17] by finding the eigenfunctions of the Laplacian within a specific domain. They define a covariance operator $\mathcal{K}\phi = \int k(\cdot, \mathbf{x}') \phi(\mathbf{x}') d\mathbf{x}'$ and, for an isotropic function $k(\mathbf{x}, \mathbf{x}') := k(\|\mathbf{r}\|)$ in a compact set $\Omega \subset \mathbb{R}^d$, expand the operator \mathcal{K} into a series of Laplacian operators. They use this to generate a series expansion of the kernel k , leading (by analogous argument to that used in the Nyström GP) to the truncated approximation $k(\mathbf{x}, \mathbf{x}') \approx \sum_{j=1}^m S(\sqrt{\lambda_j}) \phi_j(\mathbf{x}) \phi_j(\mathbf{x}')$, where λ_j and ϕ_j are the eigenpairs of the Laplacian in a given domain and $S(\cdot)$ is the spectral density of the covariance function. In general, the eigenpairs of the Laplacian have simple closed form solutions. For example, in one dimensional Cartesian coordinates they correspond to the Fourier basis. In order to find the eigenpairs of the Laplacian, boundary conditions must be specified. Dirichlet boundary conditions force the response to zero

at the edges of Ω , leading to some distortion near the edges of the space. Solin et al. recommend defining an area of interest and setting the boundary conditions slightly beyond this. In this paper we refer to this approximation as the *Laplacian-GP*.

Both the Nyström-GP and the Laplacian-GP lead to an rank m approximation to the full covariance $\mathbf{K} \approx \Phi^T \Lambda \Phi$, allowing inversion in $\mathcal{O}(m^3)$ using the matrix inversion lemma.

Sparse spectrum GP: Lázaro-Gredilla et al. [18] provide an alternative method to select the eigenpairs by Monte-Carlo sampling of the kernel function’s spectral density. This method is found to compare poorly (for equivalent covariance rank m) to the methods presented in both [16] and [17] due to the random selection of spectral points.

Random forest regressors: RFRs are recommended as an alternative to the GP for BO by Hutter et al. [7]. They note the computational load of the GP, and the possibly sub-optimal management of changepoints. We found that the latter criticism is of limited importance for the real objective functions we consider, and the sparse GPs considered in this paper aims to rectify the former. In this paper, we use the RFR from scikit-learn [19] with 30 trees. The posterior mean and variance were taken as the mean and variance of the outputs of the regressors.

3.2 Model Accuracy and Computational Complexity

Figure 2 shows the posteriors of the Full GP and the Nyström, FITC and Laplacian approximations discussed above, along with a RFR. There is very little to distinguish the GP approximations given a suitable number of inducing points or eigenpairs. The regression forest shows reasonable behaviour, by reducing its variance in densely sampled regions or regions where the mean is flat, and increasing it in sparsely sampled or steep regions. This dependence on gradient may be problematic because sparsely sampled flat regions will have a spuriously low variance. The effects of the posterior inaccuracy on BO are quantitatively evaluated in Section 4.

The computational complexities of the 5 methods considered here are listed in Table 1. A point of note is that the *Laplacian approximation scales poorly with dimensionality* because it requires a dense spectrum. If m is the cardinality of the basis needed for a good representation in one dimension, for a d dimensional space one would need m^d bases. This presents problems in high dimensions: fourteen basis functions in five dimensions would lead to a matrix with rank 537,824.

The Nyström intelligently (if heuristically) selects the m most important eigenpairs for a given space, and the FITC approximation allows one to select the number of inducing points needed. In the case of the FITC approximation, there is no clear way to select the optimal set of inducing points. The method of optimisation used in this paper is unreliable with many inducing inputs or in high dimension, and a naïve grid of inducing points scales poorly with dimensionality. Incorrectly located inducing points can cause the posterior variance to become excessively large or small. The Nyström method has more benign scaling with dimensionality. Figure 1 shows the relationship between the number of eigenpairs and the lengthscale of the GP kernel. These values were generated in a space $[0, 1]^d$, with length scales between 1 and 0.1. For very high dimensional spaces it would seem impractical to use a length-scale much below the size of the space. It should be noted that we used the heuristic recommended in Reece et al. [16] and selected all eigenpairs with eigenvalues larger than $\max \text{ eig}/100$, where $\max \text{ eig}$ is the largest eigenvalue of the matrix.

We found that the logarithm of the number of basis functions needed scaled inversely with the logarithm of the length-scales of the GP. In the worst case we tested (11 dimensions with length scale 0.3 in $[0, 1]^{11}$) we needed 1,500 bases to adequately represent the kernel. As length-scales become very short in relation to the input space, the exponential scaling begins to reduce the viability of the this model, however we have not found this to be a problem in practical algorithm configuration. Figure 1 shows the log-log relationship between the lengthscale and the number of bases.

4 Sparse Bayesian Optimisation on a Mixture of Gaussians

We tested each of the five regressors above as the response surface within BO to optimise a toy mixture of Gaussians with 5 local minima and one global minimum in $[-5, 5]$. For the first test, we used two randomly sampled points to ‘prime’ the regressors. The full GP, Laplacian, FITC and

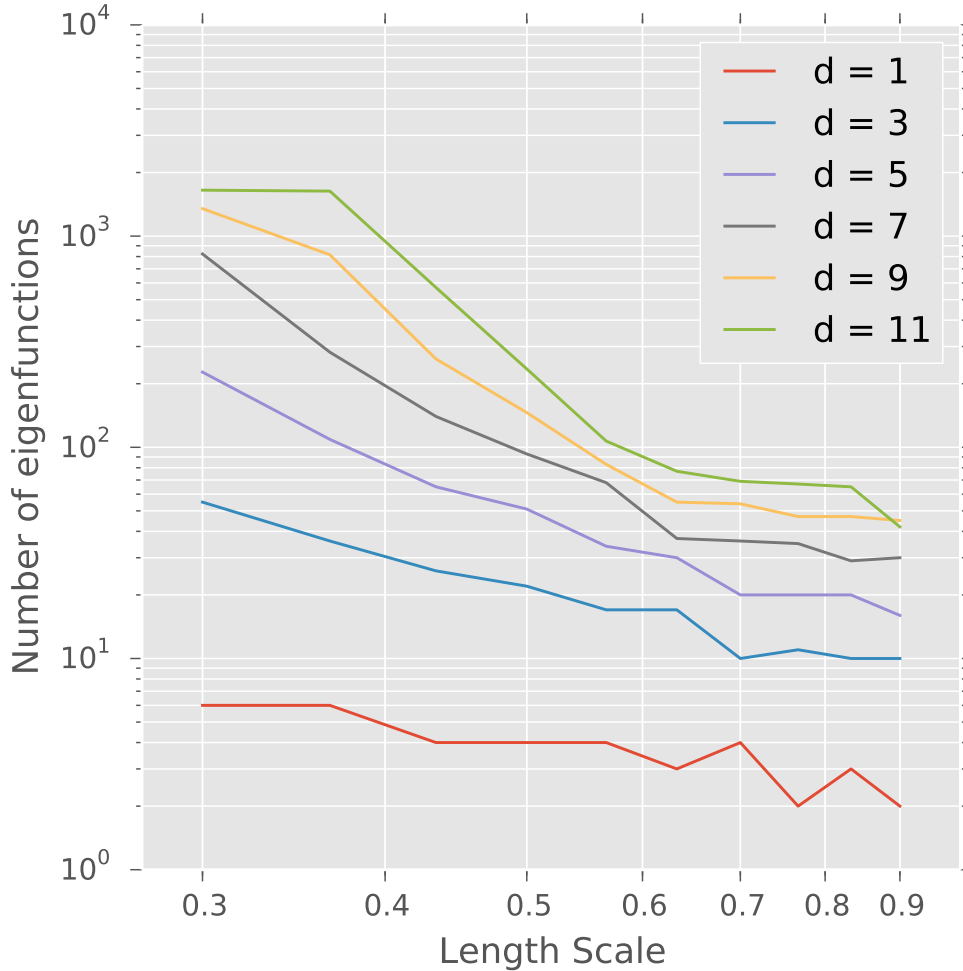


Figure 1: Scaling of number of Eigenfunctions with lengthscale and dimensionality.

Regressor	Prediction complexity	Learning complexity
GP	Initial $\mathcal{O}(n^3)$ then $\mathcal{O}(pn^2)$	$\mathcal{O}(hn^3)$
Nyström	Initial $\mathcal{O}(m^3)$ then $\mathcal{O}(pm^2)$	$\mathcal{O}(hs^3m^3)$
Laplacian	Initial $\mathcal{O}(m^3)$ then $\mathcal{O}(pm^2)$	Initial $\mathcal{O}(hnm^2)$ then $\mathcal{O}(hm^3)$
FITC	$\mathcal{O}(pm^2)$	$\mathcal{O}(nm^2)$
Regression Forest	$\mathcal{O}(ptc)$	$\mathcal{O}(dtn \log(n))$

Table 1: Computational complexity of regressors. n is the number of training points, p is the number of prediction points, s is the size of the sample set used to characterise the spectrum of the kernel, h is the number of training steps when optimising the hyper-parameters, t is the number of trees making up the regression forest, c is the number of decisions that must be made in a tree, d is the dimensionality of the input space and m is the rank of the low-rank covariance matrix. GP results from [20], Nyström from [16], Laplacian from [17], FITC from [15] and RFR from [21].

Nyström approximations all converged to the minimum within 10 iterations. The RFR approximation did not perform well. The use of the empirical mean and variance in the RFR caused a large amount of the function space to be given zero variance and so was effectively ignored as the regressor

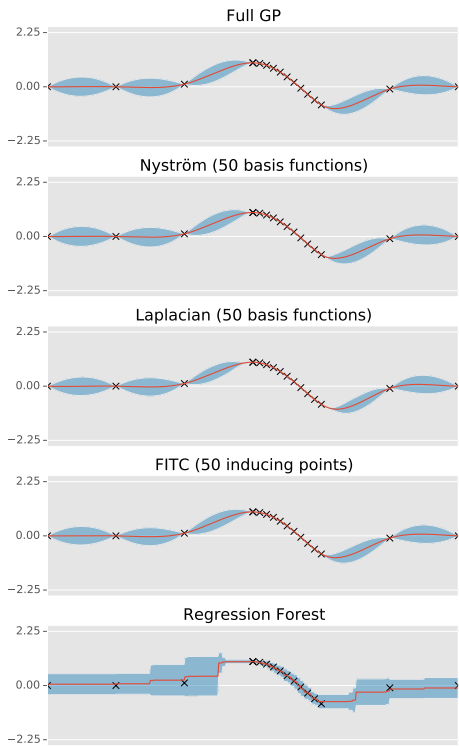


Figure 2: Mean and variance of the GP, spectral GPs and other methods

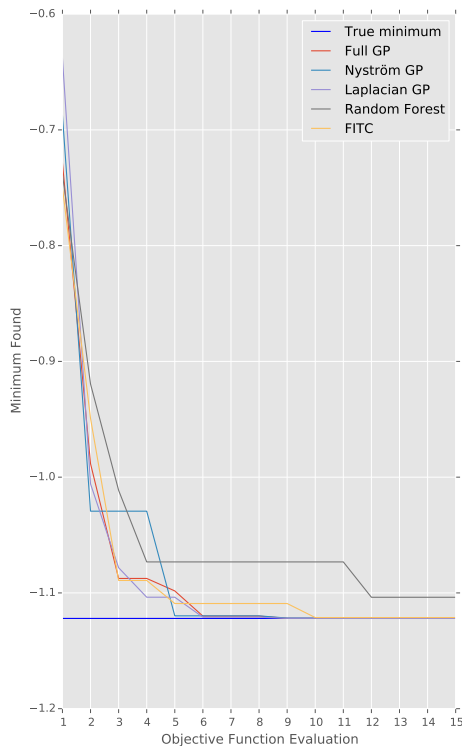


Figure 3: Convergence of different regressors to the true minimum.

thoroughly explored a local minima. As a further experiment, we primed the regressors with 200 randomly sampled points in $[-5, 5]$. With this density of samples all methods converged to the global minimum, however again the RFR took considerably longer.

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 hyper-parameters 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 by the inducing inputs which may cause it to under- or over-estimate the variance, as shown by its slower convergence in Figure 3.

We have found little difference in the performance between the FITC, Laplacian and Nyström 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 3). We will concentrate on the Nyström approximation henceforth, due to its better scaling with dimensionality and more reliable variance.

5 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 [9], stochastic gradient Langevin dynamics [10] and stochastic variational inference [11, 22]. 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 Nyström approximation to make many observations of the likelihood with different subsets of the data at each step of the

BO algorithm. Our approach provides a means of global exploration using stochastic likelihood evaluations that complements the local, gradient-driven, optimisation enabled by existing stochastic approaches. Unlike these approaches, we do not make use 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.

5.1 Optimisation of the Branin function

We tested the performance of both traditional BO and Nyström BO in optimising the Branin function [23], a standard function for testing machine learning algorithms. With 200 basis functions, the performance of the full GP and Nyström approximations were indistinguishable. By wall clock time, the full GP performed better initially (when the time taken to compute the Cholesky decomposition was less than the eigendecomposition of the Gram matrix), however the Nyström method showed better scaling as n increased. We also found that the Nyström method was more robust to conditioning errors caused by making multiple local samples.

To test the hypothesis that we can perform GO with noisy observations, we re-ran the experiment with the evaluations of the Branin function 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 either passed to the GP directly (‘full’ mode), or averaged and passed to the GP as a single less noisy observation (‘average’ mode). With 50 noisy evaluations per BO evaluation, and 20-30 seconds per evaluation, the ‘full’ mode gathered between 1,500 and 2,500 samples, while the ‘average’ mode gathered 30 to 50. In tests, we found the full mode to outperform the average mode with little computational overhead. In addition, we drew 400 points from a Sobol sequence, evaluated the noisy Branin on these and passed them to the GP as a pre-sample, 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 4,000 data points.

For each of the algorithms, we started a local minimiser at each of the three true minima of the Branin function and let them run to convergence, finding the local minima of the response surface nearest to each of the Branin minima (in all cases one of these three points was also the global minimum of the response surface). We take this global minimum of the response surface as our estimate of ν , as discussed in Section 2. We used the ‘Gap’ measure of performance to compare the methods [24]:

$$G := \frac{y(x^{\text{init}}) - y(x^{\text{best}})}{y(x^{\text{init}}) - y(x^{\text{opt}})} \tag{2}$$

Table 2 shows the maximum gap, mean gap, and the best minimum found by STOAT and the standard BO algorithm. We also tested covariance matrix adaptive evolution strategy (CMAES) [25] and dividing rectangles (DIRECT) [26] however found that these did not converge on this very noisy objective.

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

Table 2: Performance of the algorithms minimising the Branin function with observation noise.

6 Stochastic Bayesian Configuration of Model Parameters on Energy Data

AgentSwitch [27] 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

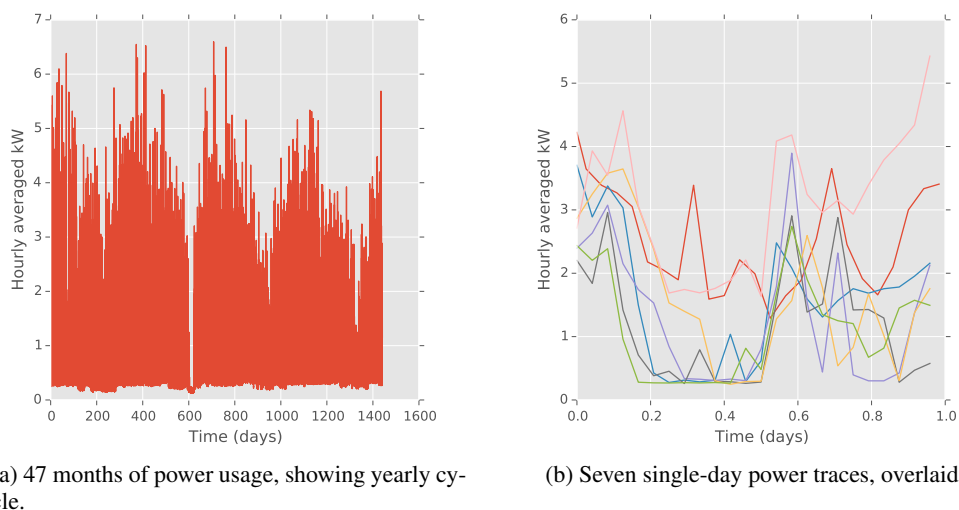


Figure 4: Hourly averaged household power use.

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 [27] we used on household power use data¹. *We will use our algorithm to fit the parameters of this complex model to big data.*

An inspection of Fourier transform of the power use data may allow one to find the major frequencies. There is a clear peak between 350 and 500 days (the resolution prohibits more accuracy) and a multitude of peaks at higher frequencies. It is not clear which is the best to select. In addition, the data is not equally spread in time, with some elements missing, making any Fourier transform an approximation of the true spectrum.

The authors of [27] fix the periodicity hyper-parameter a-priori to one day and fit the other parameters using type 2 maximum likelihood estimation (MLE-II) on subsets of the data. From inspection of the log-likelihood surface (and the spectrum computed above), we can see that there is a strong periodic component around one year, however the lower period minima is highly multi-modal. To adequately explore this surface would require multiple restart at different initial hyper-parameter values, which quickly approaches the computational burden of true BO. This authors of [27] 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 two periods on large, real data in comparable time to a multi-start optimisation of MLE-II *learning only the non-periodic hyper-parameters.*

The dataset contains seven features describing the average of 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 extremely impractical on reasonable computing hardware, requiring 50 gigabytes of RAM simply to *store* the Gram matrix.

We used STOAT to learn a the periods of a sum of periodic 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

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

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 [27], 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 [13] (Bayesian Quadrature (BQ)) or [4] (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 loglikelihood
STOAT learned double periodic	-7.25
STOAT learned single periodic	-7.39
AgentSwitch a-priori single periodic	-7.40
Aperiodic GP	-9.22

Table 3: Data log-likelihood on real 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 held out test-data of our dual periodic and single periodic kernels, [27]’s single periodic kernel and a simple, aperiodic SE kernel. The results in Table 3 show the model using our parameters tuned by our algorithm outperforming both [27] 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.

7 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 UK Research Council (EPSRC) funded ORCHID Project EP/I011587/1. Thanks to Chris Lloyd for his help with the FITC approximation. Thanks to the authors, packager and maintainers involved with the Julia language [28].

References

- [1] Frank Hutter, Holger Hoos, and Kevin Leyton-Brown. “An Efficient Approach for Assessing Hyper-parameter Importance”. In: *Proceedings of The 31st International Conference on Machine Learning*. 2014, pp. 754–762.
- [2] James Bergstra et al. “Algorithms for hyper-parameter optimization”. In: *Advances in Neural Information Processing Systems*. 2011.

- [3] 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).
- [4] Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams. “Practical Bayesian Optimization of Machine Learning Algorithms”. In: *Advances in Neural Information Processing Systems*. 2012.
- [5] James Bergstra, Daniel Yamins, and David Cox. “Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures”. In: *International Conference on Machine Learning*. 2013, pp. 115–123.
- [6] Chris Thornton et al. “Auto-WEKA: Combined Selection and Hyperparameter Optimization of Classification Algorithms”. In: *Proc. of KDD’13*. 2013, pp. 847–855.
- [7] 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.
- [8] 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.
- [9] Herbert Robbins and Sutton Monro. “A Stochastic Approximation Method”. English. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407. ISSN: 00034851.
- [10] 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.
- [11] Matthew D Hoffman et al. “Stochastic variational inference”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 1303–1347.
- [12] Donald R Jones. “A taxonomy of global optimization methods based on response surfaces”. In: *Journal of global optimization* 21.4 (2001), pp. 345–383.
- [13] 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.
- [14] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006. ISBN: 9780387310732.
- [15] A Naish-Guzman and SB Holden. “The Generalized FITC Approximation.” In: *NIPS* (2007).
- [16] Steven Reece et al. “Efficient State-Space Inference of Periodic Latent Force Models”. In: *Journal of Machine Learning Research* (2014). arXiv:1319.6319v2.
- [17] Arno Solin and Simo Sarkka. “Hilbert Space Methods for Reduced-Rank Gaussian Process Regression”. In: (2014). arXiv:arXiv:1401.5508v1.
- [18] Miguel Lázaro-Gredilla et al. “Sparse Spectrum Gaussian Process Regression”. In: *The Journal of Machine Research* 11 (2010), pp. 1865–1881.
- [19] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [20] C E Rasmussen and C K I Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN: 026218253X.
- [21] Leo Breiman et al. *Classification and Regression Trees*. Wadsworth, 1984.
- [22] James Hensman, Nicolo Fusi, and Neil D Lawrence. “Gaussian Processes for Big Data”. In: *UAI* (2013).
- [23] L. C. W. Dixon and G. P. Szegö. “The global optimization problem: an introduction”. In: *Towards Global Optimisation* 2. 1978, pp. 1–15.
- [24] 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.
- [25] Nikolaus Hansen and Andreas Ostermeier. “Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation”. In: *ICEC* (1996), pp. 312–317.
- [26] 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.
- [27] SD Ramchurn et al. “AgentSwitch: Towards smart energy tariff selection”. In: *Autonomous Agents and Multi-Agent Systems* (2013).
- [28] J. Bezanson et al. “Julia: A Fast Dynamic Language for Technical Computing”. In: *CoRR* (2012).