# Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters

**Ziyu Wang**[1], **Nando de Freitas**[1,2,3]
[1]University of Oxford, [2]Google DeepMind
[3]Canadian Institute for Advanced Research
{ziyuw, nando}@cs.ox.ac.uk

## Abstract

Bayesian optimisation has gained great popularity as a tool for optimising the parameters of machine learning algorithms and models. Somewhat ironically, setting up the hyper-parameters of Bayesian optimisation methods is notoriously hard. While reasonable practical solutions have been advanced, they can often fail to find the optima. Surprisingly, there is little theoretical analysis of this crucial problem in the literature. To address this, we derive a cumulative regret bound for Bayesian optimisation with Gaussian processes and unknown kernel hyper-parameters in the stochastic setting. The bound, which applies to the expected improvement acquisition function and sub-Gaussian observation noise, provides us with guidelines on how to design hyper-parameter estimation methods. A simple simulation as well as experiments on standard benchmarks demonstrate the importance and effectiveness of following these guidelines.

## 1 Introduction

Bayesian optimisation has become an important area of research and development in the field of machine learning, as evidenced by recent media coverage [8] and a blossoming range of applications to interactive user-interfaces [4], robotics [16, 19], environmental monitoring [18], information extraction [28], combinatorial optimisation [13, 29], automatic machine learning [3, 23, 25, 26, 12], sensor networks [9, 24], adaptive Monte Carlo [17], experimental design [1] and reinforcement learning [5].

In Bayesian optimisation, Gaussian processes are one of the preferred priors for quantifying the uncertainty in the objective function [5]. However, estimating the hyper-parameters of the Gaussian process kernel with very few objective function evaluations is a daunting task, often with disastrous results as illustrated by a simple example in [2]. The typical estimation of the hyper-parameters by maximising the marginal likelihood [22, 15] can easily fall into traps; as shown in [6]. Several authors have proposed to integrate out the hyper-parameters using quadrature and Monte Carlo methods [21, 4, 23]. Despite the advantages brought in by this more sophisticated treatment of uncertainty, Bayesian optimisation can still fall in traps, as illustrated with a simple simulation example in this paper. To the best of our knowledge, the work of Bull [6] provides the only known regret bound for Bayesian optimisation when the hyper-parameters are unknown. Here, we extend the work of Bull to stochastic objective functions with sub-Gaussian observation noise, *e.g.*, symmetric Gaussian, Bernoulli, or uniform noise. We derive an algorithm that is inspired by the theory and show that not only is it more robust to misspecification of hyper-parameters but also is competitive against the state of the art approaches on standard benchmarks.

## 2 Bayesian optimisation

We consider a sequential decision approach to global optimisation of smooth functions $f(\cdot) : \mathcal{X} \mapsto \mathbb{R}$ over an index set $\mathcal{X} \subset \mathbb{R}^d$. At the $t$-th decision round, we select an input $\mathbf{x}_t \in \mathcal{X}$ and observe the

**Algorithm 1** Bayesian optimisation with Hyper-parameter optimisation.

```
input Threshold $t_\sigma > 0$, percentage of reduction parameter $p \in (0,1)$, and $c_2 > c_1 > 0$.
input Lower and upper bounds $\boldsymbol{\theta}^L$, $\boldsymbol{\theta}^U$ for the hyper-parameters.
input Initial length scale hyper-parameter $\boldsymbol{\theta}^L \le \boldsymbol{\theta}_1 \le \boldsymbol{\theta}^U$.
Initialize $E = 0$
for $t = 1, 2, \ldots$ do
    Select $\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha^{\text{EI}}_{\boldsymbol{\theta}_t}(\mathbf{x}|\mathcal{D}_{t-1})$
    if $\sigma^2_{t-1}(\mathbf{x}_t; \theta_t) < t_\sigma \sigma^2$ then
        $E = E + 1$
    else
        $E = 0$
    end if
    Augment the data $\mathcal{D}_t = \mathcal{D}_{t-1} \cup (\mathbf{x}_t, y_t)$
    if $E = 5$ then
        Restrict $\boldsymbol{\theta}^U$ such that $\theta_i^U = \max\left\{\min\left[p\max_j\{\theta_j^U\}, \theta_i^U\right], \theta_i^L\right\}$
        $E = 0$
    end if
    Choose hyper-parameters $\boldsymbol{\theta}_{t+1}$ such that $\boldsymbol{\theta}^L \le \boldsymbol{\theta}_{t+1} \le \boldsymbol{\theta}^U$.
    Choose hyper-parameter $\nu_t^{\boldsymbol{\theta}_{t+1}}$ such that $c_1\xi_{t+1}^{\boldsymbol{\theta}_{t+1}} \le \nu_{t+1}^{\boldsymbol{\theta}_{t+1}} \le c_2\xi_{t+1}^{\boldsymbol{\theta}_{t+1}}$, where $\xi_t^{\boldsymbol{\theta}_t}$ is defined in [27].
end for
```

value of a *black-box* reward function $f(\mathbf{x}_t)$. The returned value may be deterministic, $y_t = f(\mathbf{x}_t)$, or stochastic, $y_t = f(\mathbf{x}_t) + \epsilon_t$. Our goal is to approach optimiser $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ rapidly.

Although the function is unknown, we assume that it is smooth and introduce a Gaussian process prior to encode our beliefs over the smoothness of the function. We derive a posterior distribution $p(f(\cdot)|\mathcal{D}_t)$ from which we can carry out inference. A Gaussian processes (GP) offer a flexible and relatively simple way of placing priors over functions. Such priors are completely characterised by a mean function $m(\cdot)$ and a covariance kernel $k(\cdot,\cdot)$. In particular, given any finite collection of inputs $\mathbf{x}_{1:t}$ the outputs are jointly Gaussian, $f(\mathbf{x}_{1:t})|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{m}(\mathbf{x}_{1:t}), \mathbf{K}^{\boldsymbol{\theta}}(\mathbf{x}_{1:t}, \mathbf{x}_{1:t}))$, where $\mathbf{K}^{\boldsymbol{\theta}}(\mathbf{x}_{1:t}, \mathbf{x}_{1:t})_{ij} = k^{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j)$ is the covariance matrix (parametrised by $\boldsymbol{\theta}$) and $\mathbf{m}(\mathbf{x}_{1:t})_i = m(\mathbf{x}_i)$ the mean vector. For convenience, we assume a zero-mean prior. We consider the following types of covariance kernels: $k^{\boldsymbol{\theta}}_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2}r^2)$ and $k^{\boldsymbol{\theta}}_{\text{Matérn(5/2)}}(\mathbf{x}, \mathbf{x}') = \exp(-\sqrt{5}r)(1 + \sqrt{5}r + \frac{5}{3}r^2)$ where $r = (\mathbf{x} - \mathbf{x}')^{\mathsf{T}}\text{diag}(\boldsymbol{\theta}^2)^{-1}(\mathbf{x} - \mathbf{x}')$. We assume that the observations of the function at any point $\mathbf{x}_t$ are corrupted by $\sigma$-sub-Gaussian noise $y_t = f(\mathbf{x}_t) + \epsilon_t$. It is important to note that one does not need to implement GP differently because of the sub-Gaussianity assumption and theoretical results in this paper follows from standard implementations of GPs.

Having specified a prior distribution, we turn our attention to the problem of selecting an acquisition function $\alpha(\cdot|\mathcal{D}_t)$ for choosing the next query point, $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x}|\mathcal{D}_t)$. Although many acquisition functions have been proposed (see for example [20, 14, 11, 10, 23, 12]), the expected improvement (EI) criterion remains a default choice in popular Bayesian optimisation packages, such as SMAC and Spearmint [13, 23]. If we let $\mathbf{x}_t^+ = \arg\max_{i\le t} f(\mathbf{x}_i; \boldsymbol{\theta})$ denote the current *incumbent*, the EI acquisition function can be written in closed form as $\alpha^{\text{EI(f)}}_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{D}_t) = \mathbb{E}[\max\{0, f(\mathbf{x}) - f(\mathbf{x}^+)\}|\mathcal{D}_t] = \sigma_t(\mathbf{x}; \boldsymbol{\theta})[a\Phi(a) + \phi(a)]$ with $a = \frac{\mu_t(\mathbf{x};\boldsymbol{\theta}) - f(\mathbf{x}^+)}{\sigma_t(\mathbf{x};\boldsymbol{\theta})}$, and $\phi$ and $\Phi$ are the standard normal PDF and CDF respectively. In the special case of $\sigma_t(\mathbf{x}; \boldsymbol{\theta}) = 0$, we set $\alpha^{\text{EI(f)}}_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{D}_t) = 0$. While the above member is reasonable for deterministic optimisation, observation noise could cause it to be brittle in the stochastic case. In the stochastic setting, the improvement over the best mean value $\mu^+_{\boldsymbol{\theta}} = \max_{\mathbf{x}\in\mathcal{X}} \mu_t(\mathbf{x}; \boldsymbol{\theta})$ seems to be a more reasonable alternative. In this paper, we will consider a re-scaled version of this criterion: $\alpha^{\text{EI}}_{\boldsymbol{\theta}}(\mathbf{x}|\mathcal{D}_t) = \mathbb{E}[\max\{0, f(\mathbf{x}) - \mu^+_{\boldsymbol{\theta}}\}|\mathcal{D}_t] = \nu\sigma_t(\mathbf{x}; \boldsymbol{\theta})[\frac{u}{\nu}\Phi(\frac{u}{\nu}) + \phi(\frac{u}{\nu})]$ where $u = \frac{\mu_t(\mathbf{x};\boldsymbol{\theta}) - \mu^+_{\boldsymbol{\theta}}}{\sigma_t(\mathbf{x};\boldsymbol{\theta})}$ and $\nu$ is a parameter to be estimated. Intuitively, this parameter enables us to rescale the kernel. In the deterministic case, it plays an equivalent role to multiplying the kernel by an unknown coefficient $\nu$.

## 3 Theoretical analysis

Our theoretical analysis uses *regret* to measure convergence and *information gain* to measure how informative the samples are about $f(\cdot)$. It assumes that the noise process $\epsilon_t$ is *sub-Gaussian*, and that

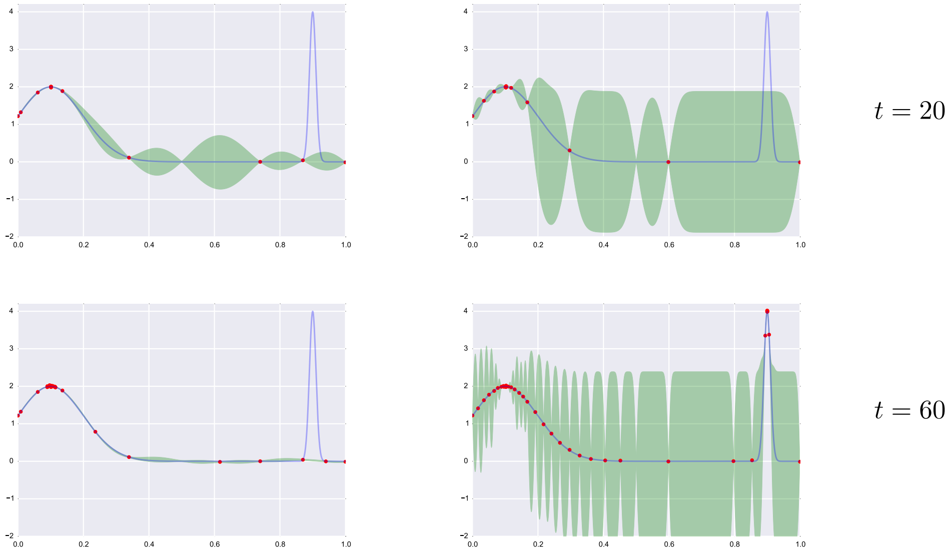

Figure 1: Convergence of EI with slice sampling over the kernel hyper-parameters [**left**] and EI using Algorithm 1 [**right**] at two function evaluation steps ($t$). The objective function (in blue) was constructed so that it has a trap. Unless EI with slice sampling hits the narrow optimum by random chance, it becomes too confident and fails to converge after 60 evaluations. In contrast, the confidence bounds for Algorithm 1 can increase enabling it to sample the function in a more reasonable way and thus find the optimum.

the function $f(\cdot)$ is smooth according to the *reproducing kernel Hilbert space (RKHS)* associated with the GP kernel $k^{\boldsymbol{\theta}}(\cdot,\cdot)$. As in [24], we will measure the performance of the Bayesian optimisation algorithm using *regret*. The cumulative regret after $T$ iterations is $R_T = \sum_{t=1}^{T} f(\mathbf{x}^*) - f(\mathbf{x}_t)$. For more informative introductions of the background material, please refer to the longer version of this paper [27]. Our theorem assumes bounds on the kernel hyper-parameters of the form $\boldsymbol{\theta}^L \leq \boldsymbol{\theta}_t \leq \boldsymbol{\theta}^U$ for all $t \geq 1$ with $f(\cdot) \in \mathcal{H}_{\boldsymbol{\theta}^U}(\mathcal{X})$ ($f$ is an element of the RKHS defined by parameters $\boldsymbol{\theta}^U$). While we could recall all the conditions on the kernel function necessary for our theorem to apply, we simply restrict the family of kernels to one that satisfies the conditions detailed in [6]. Without loss of generality, we assume that $k(\mathbf{x},\mathbf{x})$= 1. Our theorem characterising the growth in the cumulative regret $R_T$ with the number of function evaluations $T$ follows.

**Theorem 1.** *Let* $C_2 := \prod_{i=1}^{d} \frac{\theta_i^U}{\theta_i^L}$. *Suppose* $\boldsymbol{\theta}^L \leq \boldsymbol{\theta}_t \leq \boldsymbol{\theta}^U$ *for all* $t \geq 1$ *and* $f(\cdot) \in \mathcal{H}_{\boldsymbol{\theta}^U}(\mathcal{X})$. *If* $\left(\nu_t^{\boldsymbol{\theta}}\right)^2 = \boldsymbol{\Theta}\left(\gamma_{t-1}^{\boldsymbol{\theta}} + \log^{1/2}(2t^2\pi^2/3\delta)\sqrt{\gamma_{t-1}^{\boldsymbol{\theta}}} + \log(t^2\pi^2/3\delta)\right)$ *for all* $t \geq 1$. *Then with probability at least* $1-\delta$, *the cumulative regret obeys the following rate:*

$$R_T = \mathcal{O}\left(\beta_T \sqrt{\gamma_T^{\boldsymbol{\theta}^L} T}\right), \tag{1}$$

*where* $\beta_T = 2\log\left(\frac{T}{\sigma^2}\right)\gamma_{T-1}^{\boldsymbol{\theta}^L} + \sqrt{8}\log\left(\frac{T}{\sigma^2}\right)\log^{1/2}(4T^2\pi^2/6\delta)\left(\sqrt{C_2}\|f\|_{\mathcal{H}_{\boldsymbol{\theta}^U}(\mathcal{X})} + \sqrt{\gamma_{T-1}^{\boldsymbol{\theta}^L}}\right) + C_2\|f\|^2_{\mathcal{H}_{\boldsymbol{\theta}^U}(\mathcal{X})}$.

For a proof of the theorem, please refer to [27]. Our result is analogous to Theorem 3 of [24] which proves convergence rates for the GP-UCB algorithm in the agnostic setting. Their result, however, does not allow for the estimation of hyper-parameters. In addition, EI does not require explicit knowledge of the RKHS norm of the objective function while GP-UCB does require this.

## 4 An algorithm inspired by the theory

For Theorem 1 to hold, it is necessary that there exist element-wise upper-bounds $\boldsymbol{\theta}^U$ on the hyper-parameters $\boldsymbol{\theta}$, such that the objective function $f(\cdot) \in \mathcal{H}_{\boldsymbol{\theta}^U}(\mathcal{X})$. In practice, it is difficult to assess

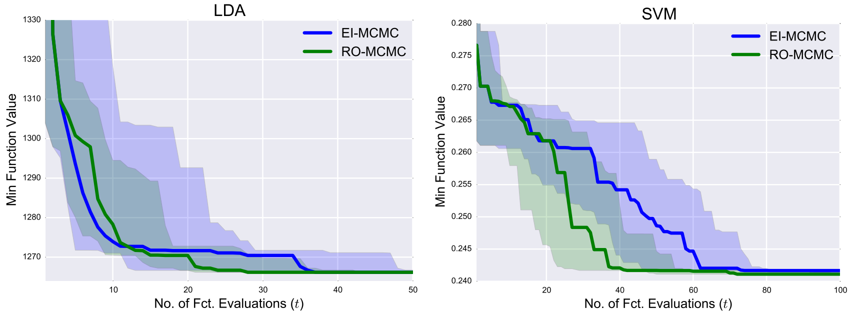

Figure 2: Comparison of the proposed approach (RO-MCMC) with the Spearmint package [23] (EI-MCMC) on standard benchmarks [7] (lower is better). Here, the proposed outperforms the state of the art.

this condition. To surmount this difficulty, we draw inspiration from [29], and propose to reduce the upper bound of the length scales $\boldsymbol{\theta}^U$ when the algorithm becomes overconfident. In particular, we adaptively reduce $\boldsymbol{\theta}^U$ whenever the model repeatedly samples points of low posterior variance in comparison to the noise variance $\sigma^2$. Once the algorithm optimises to the precision of the noise variance, it suffers from a slower convergence rate. As $\boldsymbol{\theta}^U$ is successively decreased, after a finite number of iterations, we can ensure that $f(\cdot) \in \mathcal{H}_{\boldsymbol{\theta}^U}(\mathcal{X})$ as long as there exists $\boldsymbol{\theta} \geq \boldsymbol{\theta}^L$ such that $f(\cdot) \in \mathcal{H}_{\boldsymbol{\theta}}(\mathcal{X})$. We advocate a conservative choice of $\boldsymbol{\theta}^L$ whenever we have little knowledge of the range of possible values of $\boldsymbol{\theta}$. In practice, we could use a number of strategies for estimating the hyper-parameters, provided they fall within the bounds set by Theorem 1. In particular, we could use maximum likelihood or MCMC (taking only one sample) to estimate the hyper-parameters in this constrained space. The full algorithm is summarised in algorithm 1. The astute reader would have noticed the parameters $t_\sigma$, $p$, $c_2$ and $c_1$ in the algorithm. If we want to achieve an accuracy comparable to the noise variance, we should set $t_\sigma = 1$. The other parameters simply determine how fast the algorithm converges and should be set to reasonable fixed values, e.g. $p = 0.5$, $c_2 = 1$ and $c_1 = 0.001$. Provided $t_\sigma > 0$, $p \in (0, 1)$ and $c_2 > c_1 > 0$, the theory is satisfied.

If we have strong beliefs about our GP prior model, it may seem unnecessary to estimate our parameters with Algorithm 1. When our prior belief is misplaced, however, we could fail to converge if we were to follow the traditional probabilistic approach. We provide an illustration of this effect by optimising the following stochastic function:

$$f(x) = 2\mathbf{k}_{SE}^{\theta_1}(x_1, x) + 4\mathbf{k}_{SE}^{\theta_2}(x_2, x) + \epsilon$$

over the interval $[0, 1]$, where $\theta_1 = 0.1$, $\theta_2 = 0.01$, $x_1 = 0.1$, $x_2 = 0.9$, and $\epsilon$ is zero-mean Gaussian with $10^{-2}$ standard deviation. Figure 1 compares Algorithm 1 against standard Bayesian optimisation with the same EI function, but using slice sampling to infer the kernel hyper-parameters (without imposing the theoretical bounds on the hyper-parameters). We see that, in the absence of reasonable prior beliefs, conditions like the ones detailed in our theoretical results are necessary to guarantee reasonable sampling of the objective function. While heteroskedastic GP approaches could mitigate this problem, there are no theoretical results to guarantee this to the best of our knowledge.

To further evaluate the effectiveness of the proposed approach, we have also applied the proposed algorithm to standard benchmarks in BO (please refer to [7] for more detailed descriptions). The results are summarized in Figure 2. In these examples, we used slice sampling to optimise the hyper-parameters albeit taking only the last sample. Despite using fewer samples, the proposed approach is competitive against current state of the art approaches.

## 5 Conclusion

Despite the rapidly growing literature on Bayesian optimisation and the proliferation of software packages that learn the kernel hyper-parameters, to the best of our knowledge, only Bull [6] and us have attacked the question of convergence of GP-based Bayesian optimisation with unknown hyper-parameters. Bull's results focused on deterministic objective functions. Our new results apply to the abundant class of noisy objective functions.

## References

[1] J. Azimi, A. Jalali, and X.Z. Fern. Hybrid batch bayesian optimization. In *ICML*, 2012.

[2] R. Benassi, J. Bect, and E. Vazquez. Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion. In *Learning and Intelligent Optimization*, pages 176–190. Springer, 2011.

[3] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *NIPS*, pages 2546–2554, 2011.

[4] E. Brochu, T. Brochu, and N. de Freitas. A Bayesian interactive optimization approach to procedural animation design. In *ACM SIGGRAPH / Eurographics SCA*, pages 103–112, 2010.

[5] E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report UBC-2009-23 and arXiv:1012.2599v1, 2009.

[6] A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12:2879–2904, 2011.

[7] K. Eggensperger, M. Feurer, F. Hutter, J. Bergstra, J. Snoek, H. Hoos, and K. Leyton-Brown. Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. In *NIPS Workshop on Bayesian Optimization in Theory and Practice*, 2013.

[8] K. Finley. Netflix is building an artificial brain using Amazons cloud, February, Wired.com 2014.

[9] R. Garnett, M. A. Osborne, and S. J. Roberts. Bayesian optimization for sensor set selection. In *ACM/IEEE IPSN*, pages 209–219. ACM, 2010.

[10] P. Hennig and C.J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13:1809–1837, 2012.

[11] M.W. Hoffman, E. Brochu, and N. de Freitas. Portfolio allocation for Bayesian optimization. In *UAI*, pages 327–336, 2011.

[12] M.W. Hoffman, B. Shahriari, and N. de Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *AIStats*, pages 365–374, 2014.

[13] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *LION*, pages 507–523, 2011.

[14] D.R. Jones. A taxonomy of global optimization methods based on response surfaces. *J. of Global Optimization*, 21(4):345–383, 2001.

[15] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *J. of Global optimization*, 13(4):455–492, 1998.

[16] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans. Automatic gait optimization with Gaussian process regression. In *IJCAI*, pages 944–949, 2007.

[17] N. Mahendran, Z. Wang, F. Hamze, and N. de Freitas. Adaptive MCMC with Bayesian optimization. In *AIStats*, pages 751–760, 2012.

[18] R. Marchant and F. Ramos. Bayesian optimisation for intelligent environmental monitoring. In *IROS*, pages 2242–2249, 2012.

[19] R. Martinez-Cantin, N. de Freitas, A. Doucet, and J. A Castellanos. Active policy learning for robot planning and exploration under uncertainty. *RSS*, 2007.

[20] J. Močkus. The Bayesian approach to global optimization. In *Systems Modeling and Optimization*, volume 38, pages 473–481. Springer, 1982.

[21] M. A. Osborne, R. Garnett, and S. J. Roberts. Gaussian processes for global optimisation. In *LION*, 2009.

[22] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[23] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *NIPS*, pages 2951–2959, 2012.

[24] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, pages 1015–1022, 2010.

[25] K. Swersky, J. Snoek, and R. P. Adams. Multi-task Bayesian optimization. In *NIPS*, pages 2004–2012, 2013.

[26] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown. Auto-WEKA: Combined selection and hyper-parameter optimization of classification algorithms. In *KDD*, pages 847–855, 2013.

[27] Z. Wang and N. de Freitas. Theoretical analysis of Bayesian optimisation with unknown gaussian process hyper-parameters. 2014.

[28] Z. Wang, B. Shakibi, L. Jin, and N. de Freitas. Bayesian multi-scale optimistic optimization. In *AIStats*, pages 1005–1014, 2014.

[29] Z. Wang, M. Zoghi, D. Matheson, F. Hutter, and N. de Freitas. Bayesian optimization in high dimensions via random embeddings. In *IJCAI*, pages 1778–1784, 2013.