Expected Regret Minimization for Bayesian optimization with Student's-t Processes

Mr. Conor Clare
School of Computing, Ulster University
Newtownabbey, Northern Ireland
clare-c@ulster.ac.uk

Dr. Glenn Hawe
School of Computing, Ulster University
Newtownabbey, Northern Ireland
gi.hawe@ulster.ac.uk

Professor Sally McClean
School of Computing, Ulster University
Coleraine, Northern Ireland
si.mcclean@ulster.ac.uk

ABSTRACT

Student's-t Processes were recently proposed as a probabilistic alternative to Gaussian Processes for Bayesian optimization. Student's-t Processes are a generalization of Gaussian Processes, using an extra parameter ν, which addresses Gaussian Processes' weaknesses. Separately, recent work used prior knowledge of a global optimum f^*, to create a new acquisition function for Bayesian optimization called Expected Regret Minimization. Gaussian Processes were then combined with Expected Regret Minimization to outperform existing models for Bayesian optimization. No published work currently exists for Expected Regret Minimization with Student's-t Processes. This research compares Expected Regret Minimization for Bayesian optimization, using Student's-t Processes versus Gaussian Processes. Both models are applied to four benchmark problems popular in mathematical optimization. Our work enhances Bayesian optimization by showing superior training regret minimization for Expected Regret Minimization, using Student's-t Processes versus Gaussian Processes.

CCS CONCEPTS

• **Theory of computation → Theory and algorithms for application domains**; **Machine learning theory**; **Kernel methods**; **Gaussian processes**;

KEYWORDS

Bayesian Optimization, Supervised Machine Learning, Student's-t Processes, Expected Regret Minimization, Gaussian Processes, Hyperparameter Tuning, XGBoost Classification.

ACM Reference Format:
Mr. Conor Clare, Dr. Glenn Hawe, and Professor Sally McClean. 2020. Expected Regret Minimization for Bayesian optimization with Student's-t Processes. In *Proceedings of Under Review.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Bayesian optimization [2, 7, 14] uses supervised machine learning [11] to efficiently seek the global optimum $\mathbf{x}^*$ of a black-box, objective function $f(\mathbf{x})$ within a design-space χ [6, 7]:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\chi} f(\mathbf{x})$$

Bayesian optimization is widely-used in applications that have a computationally-expensive non-linear objective, such as hyperparameter tuning of machine learning algorithms [15, 18], aerostructural engineering [17] and nuclear science [3]. A probabilistic model is chosen to incorporate our prior beliefs about f. Bayesian optimization updates the prior with targets from $f(\mathbf{x})$, corresponding to locations $\mathbf{x}$, creating a posterior distribution that better approximates $f(\mathbf{x})$ [10].

There are two high-level modelling choices in Bayesian optimization - a probabilistic model and an acquisition function. The probabilistic model is also called the surrogate and uses a multivariate probability distribution. The surrogate models the joint-behaviour of the locations $\mathbf{x}$ [10]. Gaussian Processes (GPs) use the Gaussian multivariate distribution and are usually chosen as the Bayesian optimization surrogate. GPs are simply defined, using mean and covariance functions [6, 11].

The surrogate's posterior defines the acquisition function, determining the next iteration's location $\mathbf{x}$, for the corresponding target $f(\mathbf{x})$ [10]. Bayesian optimization combines the posterior mean for $\mathbf{x}$, with posterior uncertainty using an acquisition function, such as the popular Expected Improvement (EI) acquisition function [9, 19]. The acquisition function uses the surrogate's posterior mean to model exploitation - sampling in χ near the current, best $\mathbf{x}$ [2]. The acquisition function also uses the surrogate's posterior standard deviation to model exploration - sampling in χ where the uncertainty in $f(\mathbf{x})$ is highest [2].

Recently, [10] combined a GP surrogate with prior knowledge of a global optimum f^*. Since the objective function should be no better or worse than the optimal value f^*, than ideally neither should the GP posterior mean [10]. Two new acquisition functions, Confidence Bound Minimization (CBM) and Expected Regret Minimization (ERM) were derived, with both outperforming EI for Bayesian optimization [4, 10].

GPs have two known weaknesses [12, 13, 17]. First, low probability is assigned to remote outlier locations in $\mathbf{x}$, despite observed locations for an aerostructural engineering problem showing otherwise [17]. Secondly, the GP posterior covariance does not depend on the black-box function's y_k-targets. Instead, only the location of the training set $\mathbf{x}_k \in \mathcal{D}_N$ determines GP posterior covariance [12, 13, 17], where the training set of observations $\mathcal{D}_N$ is $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ for $k = 1, \ldots, N$ iterations [17].

One recently proposed solution to the weaknesses of GPs are Student-t Processes (STPs), which use the multivariate Student's-t distribution [12, 13, 17]. STPs generalize the multivariate Gaussian distribution. STPs have an additional scalar parameter ν ($\nu > 2$), which defines the 'degrees-of-freedom' of the STP [12, 13, 17] and controls STP kurtosis, influencing the size of the tails and hence,

Under Review,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.

the probability of outliers [1]. This addresses the first weakness of GPs, regarding low probability of outliers. Further, unlike the GP posterior, the STP posterior covariance does depend on the black-box function's y_k-targets [12, 13, 17], which addresses the second weakness of GPs.

Bayesian optimization with STPs is currently under-explored, with existing work mainly focused on the EI acquisition function using STPs [12, 13, 17]. Research on other acquisition functions using STPs is still embryonic, with no publications on the STP ERM acquisition function. Motivated by this knowledge gap, the main contributions of this paper are:

(1) exploiting prior knowledge of a global optimum for Bayesian optimization with Student's-t Processes;
(2) a derivation of the Expected Regret Minimization acquisition function for Student's-t Processes;
(3) comparing Expected Regret Minimization, using Student's-t Processes versus Gaussian Processes, on four benchmark problems popular in mathematical optimization.

2 BAYESIAN OPTIMIZATION SURROGATES

2.1 Gaussian Processes

A stochastic process $f(\mathbf{x})$ is Gaussian when observations jointly sampled have a multivariate Gaussian probability distribution [2, 11]. GPs are parameterized by two functions. The first is the mean function, $m(\mathbf{x})$, defining the expected value of a location, $\mathbf{x}$. The second is the kernel function $k(\mathbf{x}, \mathbf{x}')$, which calculates the covariance between two different locations $\mathbf{x}$ and $\mathbf{x}'$ [11]:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

The GP posterior covariance $\hat{\Sigma}_{GP}$ is given by [11, 17]:

$$\hat{\Sigma}_{GP} = K_{\mathbf{x}_*,\mathbf{x}_*} - K_{\mathbf{x}_*,\mathbf{x}} K^{-1}_{\mathbf{x},\mathbf{x}} K_{\mathbf{x},\mathbf{x}_*}$$

where: $K_{\mathbf{x},\mathbf{x}}$ is the covariance defined by the kernel between the observed training locations, $\mathbf{x}_k \in \mathcal{D}_N$; $K_{\mathbf{x}_*,\mathbf{x}}$ is the covariance of the kernel between the unobserved prediction locations and observed training locations; and $K_{\mathbf{x}_*,\mathbf{x}_*}$ is the covariance of the unobserved prediction locations [17]. As can be seen, the GP posterior covariance does not depend on the black-box function's y-targets [11].

2.2 Student's-t Processes

One recently proposed solution to these GP weaknesses is to instead use Student-t Processes (STPs), which uses the multivariate Student's-t probability distribution [12, 13, 17]. STPs are parameterized by two functions and a third scalar parameter, ν ($\nu > 2$). As with GPs, the mean function, $m(\mathbf{x})$, defines the expected value of a location, $\mathbf{x}$. The kernel function $k(\mathbf{x}, \mathbf{x}')$ calculates the covariance between two different locations $\mathbf{x}$ and $\mathbf{x}'$ [17]. A stochastic process $f(\mathbf{x})$ is Student's-t when observations jointly sampled have a multivariate Student's-t probability distribution [12, 13, 17]:

$$f(\mathbf{x}) \sim \mathcal{STP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'), \nu\right)$$

The STP posterior covariance $\hat{\Sigma}_{STP}$ is given by [17]:

$$\hat{\Sigma}_{STP} = \frac{\nu + y^T K^{-1}_{\mathbf{x},\mathbf{x}} y - 2}{\nu + |D_N| - 2} \left(K_{\mathbf{x}_*,\mathbf{x}_*} - K_{\mathbf{x}_*,\mathbf{x}} K^{-1}_{\mathbf{x},\mathbf{x}} K_{\mathbf{x},\mathbf{x}_*}\right)$$

where: $y^T K^{-1}_{\mathbf{x},\mathbf{x}} y$ is the squared Mahalanobis distance of the training locations $\mathbf{x}_k$ using their covariance [17]. As can be seen, the STP posterior covariance depends on the black-box function's y-targets.

Common kernels widely-used in Bayesian optimization include the squared-exponential covariance function [11] and the Matérn class of covariance functions e.g. Matérn 3/2 and Matérn 5/2 [11]. Both GPs and STPs can use these kernels.

3 EXPLOITING PRIOR KNOWLEDGE OF A GLOBAL OPTIMUM

3.1 ERM for Gaussian Processes

[10] combined prior knowledge about a global optimum f^* with a GP surrogate to enhance Bayesian optimization. Their work developed the GP ERM acquisition function as $\alpha_n^{ERM+f^*}(\mathbf{x})$:

$$\alpha_n^{ERM+f^*}(\mathbf{x}) = \hat{\sigma}_{GP}(\mathbf{x})\phi(z) + [f^* - \hat{\mu}_{GP}(\mathbf{x})]\Phi(z)$$

where: $\hat{\mu}_{GP}(\mathbf{x})$ and $\hat{\sigma}_{GP}(\mathbf{x})$ are the respective GP posterior mean and GP posterior standard deviation; $z = \frac{f^* - \hat{\mu}_{GP}(\mathbf{x}) - r(\mathbf{x})}{\hat{\sigma}_{GP}(\mathbf{x})}$; with $\phi(z)$ and $\Phi(z)$ the respective probability density function (PDF) and cumulative distribution function (CDF) of a univariate, standard normal random variable, z. The exploration of the GP ERM acquisition function is modelled by: $\hat{\sigma}_{GP}(\mathbf{x})\phi(z)$, with GP ERM exploitation modelled by: $[f^* - \hat{\mu}_{GP}(\mathbf{x})]\Phi(z)$.

3.2 Expected Likelihood of Regret: Student's-t

Now consider the univariate Student's-t PDF, with mean μ, standard deviation σ and degrees-of-freedom ν. For simplicity, define C as [17]:

$$C = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}$$

The Student's-t PDF becomes [17]:

$$\mathcal{T}(\mu, \sigma, \nu) = \frac{C}{\sigma} \times \left(1 + \frac{[(y - \mu)/\sigma]^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad -\infty < y < +\infty$$

$\mathbb{E}[r(\mathbf{x})]$ is the expected likelihood of regret and is defined for STPs as [10]:

$$= \int_0^\infty \frac{rC}{\hat{\sigma}_{STP}(\mathbf{x})} \left(1 + \frac{[(f^* - \hat{\mu}_{STP}(\mathbf{x}) - r(\mathbf{x}))/\hat{\sigma}_{STP}(\mathbf{x})]^2}{\nu}\right)^{-\frac{\nu+1}{2}} dr(\mathbf{x}) \tag{1}$$

3.3 ERM for Student's-t Processes

The STP ERM acquisition function, denoted $\alpha_n^{ERM+f^*}(\mathbf{x})$, minimizes Eq. 1 and is defined as [10, 17]:

$$\hat{\sigma}_{STP}(\mathbf{x}) \left(\frac{\nu}{\nu - 1}\right) \left(1 + \frac{z_s^2}{\nu}\right) \phi_s(z_s) + [f^* - \hat{\mu}_{STP}(\mathbf{x})]\Phi_s(z_s) \tag{2}$$

where: $\hat{\mu}_{STP}(\mathbf{x})$ and $\hat{\sigma}_{STP}(\mathbf{x})$ are respectively the STP posterior mean and STP posterior standard deviation; $z_s = \frac{f^* - \hat{\mu}_{STP}(\mathbf{x}) - r(\mathbf{x})}{\hat{\sigma}_{STP}(\mathbf{x})}$; with $\phi_s(z_s)$ and $\Phi_s(z_s)$ the respective PDF and CDF of a univariate, standard Student's-t random variable, z_s. STP ERM exploration is modelled by: $\hat{\sigma}_{STP}(\mathbf{x}) \left(\frac{\nu}{\nu-1}\right) \left(1 + \frac{z_s^2}{\nu}\right) \phi_s(z_s)$, with STP ERM exploitation modelled by: $[f^* - \hat{\mu}_{STP}(\mathbf{x})]\Phi_s(z_s)$. Algorithm 1 defines

Bayesian optimization [2, 7, 14], with the surrogate trained at each iteration using [11].

Algorithm 1: Bayesian optimization [2, 7, 14]:

```
(1) Input: black-box objective function $f(\mathbf{x})$, $n$ random-initialization iterations, $N$ post-initialization iterations.
(2) Construct $\mathcal{D}_0$, a randomly-sampled, location-target pairs' set $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \chi$, $y_i = f(\mathbf{x}_i)$, $i = 1 \ldots n$
(3) for: $k = 1, \ldots, N$ iterations do
(4)     Train surrogate using $\mathcal{D}_{k-1}$ [11]
(5)     select: $\mathbf{x}_k = \arg\max_{\mathbf{x} \in \chi} \alpha_{k-1}(\mathbf{x})$ (use Eq. 2)
(6)     query the objective $f$ at $\mathbf{x}_k$ to obtain $y_k$
(7)     augment data: $\mathcal{D}_k = \mathcal{D}_{k-1} \cup \{(\mathbf{x}_k, y_k)\}$
(8) end for
(9) Return: $\mathbf{x}_k = \arg\max_{\mathbf{x}_k \in D_k} y_k$
```

4 BENCHMARK OPTIMIZATION PROBLEMS

Four Bayesian optimization experiments are programmed in the Python language, using the 'pyGPGO' package [8]. Three synthetic functions popular in mathematical optimization are chosen, namely SixHumpCamel, Rosenbrock and Hartmann3 [16]. The fourth problem is hyperparameter tuning for XGBoost classification [10]. Each uses [11] to train a surrogate to estimate $f(\mathbf{x})$. The difference between a global optimum f^* and the best y-sampled value, defines training regret at each iteration of Bayesian optimization. The natural logarithm of training regret is then calculated and used for comparison between different Bayesian optimization models [17]. Algorithm 1 can efficiently seek a global minimum (rather than a global maximum), by multiplying both $f(\mathbf{x})$ and f^* by -1.

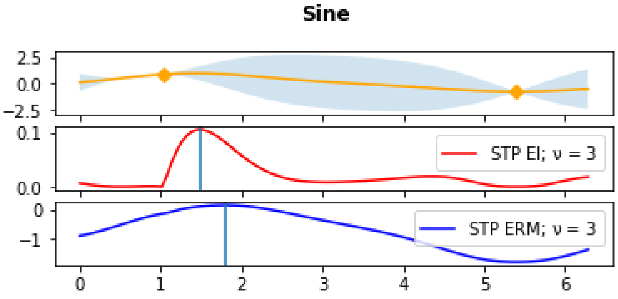

Figure 1: Optimizing the Sine function using Algorithm 1: The surrogate is STP (ν = 3) [4] and the kernel is squared-exponential [11]. The model is randomly-initialized using n = 2 locations (top). The first location x after random-initialization is shown (light-blue, vertical line), using two acquisition functions: STP (ν = 3) EI (middle) versus STP (ν = 3) ERM (bottom). This is the first iteration using lines 4-7 of Algorithm 1.

Each problem's global optimum is sought by Bayesian optimization with ERM, using STPs versus GPs. Both use a squared-exponential covariance kernel [11]. ν = 5 [17] is chosen for each STP surrogate. Each model is randomly-initialized with n = 5 iterations [5]. There are N = 100 post-initialization iterations for the three synthetic functions [17] and N = 30 post-initialization iterations for XGBoost hyperparameter tuning [4, 10], which uses a logistic objective function [4, 10].

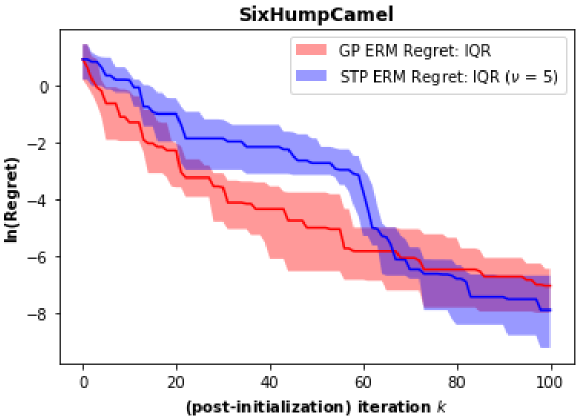

Figure 2: Comparing the SixHumpCamel function [16]: The training regret IQR for STP (ν = 5) [17] ERM is lower than the IQR of GP ERM.

The results are shown in Figures 2 - 5, with experiments repeated 20 different times for each problem. The natural logarithm of training regret ('ln(Regret)') is shown on the y-axis, with total iterations N shown on the $\mathbf{x}$-axis. The interquartile range (IQR) for the natural logarithm of training regret is shaded red for GP ERM and blue for STP (ν = 5) ERM. The red curved lines represent the median for GP ERM, while the blue curved lines show the median for STP (ν = 5) ERM. The 25th and 75th percentiles are the upper and lower bounds of the red shaded area for GP ERM and blue shaded area for STP (ν = 5) ERM. For each experiment, the training regret IQR for STP (ν = 5) ERM is lower than the IQR for GP ERM.

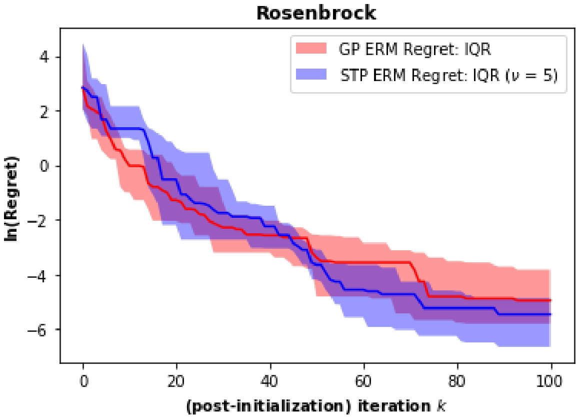

Figure 3: Comparing the Rosenbrock function [16]: The training regret IQR for STP (ν = 5) [17] ERM is lower than the IQR of GP ERM.

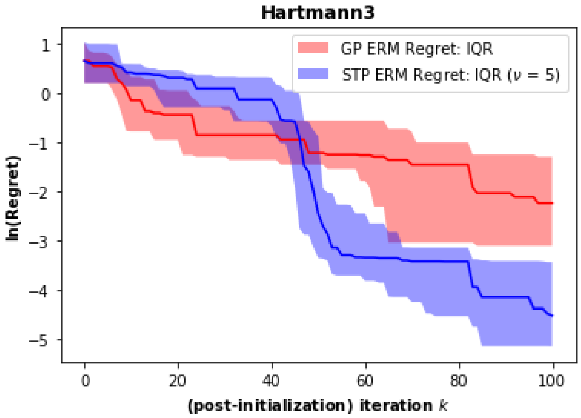

Figure 4: Comparing the Hartmann3 function [16]: The training regret IQR for STP (ν = 5) [17] ERM is lower than the IQR of GP ERM.

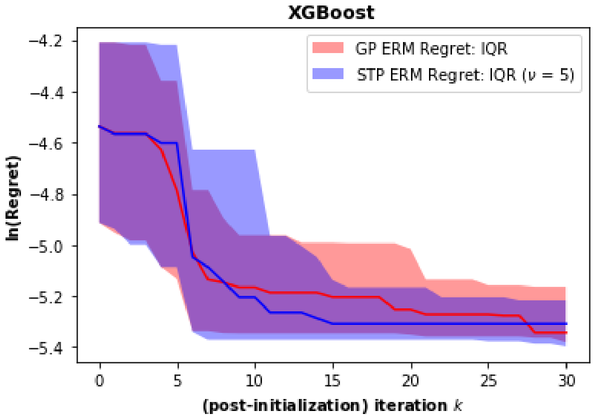

Figure 5: Comparing hyperparameter tuning for XGBoost classification training accuracy [4, 10]: The training regret IQR for STP (ν = 5) [17] ERM is lower than the IQR of GP ERM.

5 CONCLUSIONS

This paper exploits prior knowledge of a global optimum to derive the STP ERM acquisition function and compares Bayesian optimization with ERM, using STPs (ν = 5) versus GPs. Our work enhances Bayesian optimization by showing STP (ν = 5) ERM outperforms GP ERM on four problems [16] popular in mathematical optimization. Rather than choosing ν = 5 [17], future work will consider STP ERM for different ν values, with training regret minimization compared to GP ERM. Alternatively, a prior distribution for ν will be chosen.

ACKNOWLEDGMENTS

Mr. Conor Clare acknowledges the Northern Ireland Executive's Department for the Economy ('DfE') in funding this research.

REFERENCES

[1] Richard G. Brereton. 2015. The t-distribution and its relationship to the normal distribution. *Journal of Chemometrics* 29, 9 (2015), 481–483. https://doi.org/10.1002/cem.2713

[2] Eric Brochu, Vlad M. Cora, and Nando de Freitas. 2010. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. arXiv:arXiv:1012.2599

[3] Mark Alan Chilenski. 2016. *Experimental Data Analysis Techniques for Validation of Tokamak Impurity Transport Simulations.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[4] Conor Clare, Glenn Hawe, Zhiwei Lin, and Sally McClean. 2020. Confidence Bound Minimization for Bayesian Optimization with Student's-t Processes. In *3rd International Conference on Applications of Intelligent Systems.*

[5] Javier González, Michael Osborne, and Neil Lawrence. 2016. GLASSES: Relieving The Myopia Of Bayesian Optimisation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Vol. 51. PMLR, 790–799. http://proceedings.mlr.press/v51/gonzalez16b.html

[6] Phillip Hennig and Christian J. Schuler. 2012. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research* 13 (June 2012), 1809–1837.

[7] José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. 2014. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'14).* 918–926.

[8] José Jiménez and Josep Ginebra. 2017. pyGPGO: Bayesian Optimization for Python. *The Journal of Open Source Software* 2 (11 2017), 431. https://doi.org/10.21105/joss.00431

[9] J Mockus, Vytautas Tiesis, and Antanas Zilinskas. 1978. *The application of Bayesian methods for seeking the extremum.* Vol. 2. 117–129.

[10] Vu Nguyen and Michael A. Osborne. 2020. Knowing The What But Not The Where in Bayesian Optimization. arXiv:arXiv:1905.02685

[11] Carl. E. Rasmussen and Christopher K. I. Williams. 2006. Gaussian Processes for Machine Learning. In *Gaussian Processes for Machine Learning.* MIT Press.

[12] Amar Shah, Andrew G. Wilson, and Zoubin Ghahramani. 2013. Bayesian Optimization using Student-t Processes. *NIPS Workshop on Bayesian Optimization* (2013).

[13] Amar Shah, Andrew G. Wilson, and Zoubin Ghahramani. 2014. Student-t Processes as Alternatives to Gaussian Processes. *The Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS), 2014* (02 2014).

[14] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. 2016. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* 104, 1 (Jan 2016), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

[15] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'12).* 2951–2959.

[16] Sonja Surjanovic and Derek Bingham. 2019. Virtual Library of Simulation Experiments: Test Functions and Datasets. http://www.sfu.ca/~ssurjano.

[17] Brendan Tracey and David Wolpert. 2018. Upgrading from Gaussian Processes to Student's-t Processes. 2018 AIAA Non-Deterministic Approaches Conference. https://doi.org/10.2514/6.2018-1659

[18] Jian Wu, Matthias Poloczek, Andrew G. Wilson, and Peter Frazier. 2017. Bayesian Optimization with Gradients. In *Advances in Neural Information Processing Systems 30.* 5267–5278.

[19] Antanas Zilinskas. 1992. A review of statistical models for global optimization. *Journal of Global Optimization* 2 (06 1992), 145–153. https://doi.org/10.1007/BF00122051