# 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 $\nu$, which addresses Gaussian Processes' weaknesses. Separately, recent work used prior knowledge of a black-box function's 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 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

**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 Redacted*. ACM, New York, NY, USA, 5 pages.

## 1 INTRODUCTION

Bayesian optimization [2, 7, 16] uses supervised machine learning [13] to efficiently seek the global optimum $\mathbf{x}^*$ of a black-box, objective function $f(\mathbf{x})$ within a design-space $\chi$ [6, 7]:

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


*Redacted, December 31 2020, Redacted*


Bayesian optimization is widely-used in applications that have a computationally-expensive non-linear objective, such as hyperparameter tuning of machine learning algorithms [17, 21] and aerostructural engineering [19]. 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 $f(\mathbf{x})$ [12].

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 $\mathbf{x}$ [12]. 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, 13].

Bayesian optimization uses an acquisition function at each iteration to determine where to sample next in design space $\chi$. Acquisition functions can combine the surrogate posterior mean with the surrogate posterior standard deviation, to balance exploitation and exploration. A common acquisition function combined with GPs is Expected Improvement (EI), introduced first by [11] and popularized by [10].

In some optimization problems, we have prior knowledge [12, 18] of what the objective function value $f^* = f(\mathbf{x}^*)$ is at the global optimum, even though we do not know where ($\mathbf{x}^*$) it occurs in design space $\chi$. For example, for some classification problems we may know in advance that the optimum F-score is 1 [12] (optimal precision and recall), but we do not know what algorithm settings (e.g. hyperparameter values) give this performance. In this case, $\mathbf{x}^*$ represents the unknown algorithm settings and $f^* = f(\mathbf{x}^*)$ has a value of 1 [12] (the known optimal F-score).

Jones investigated this setting almost 20 years ago [9], using so-called "one-stage" approaches to locate $\mathbf{x}^*$ based on the credibility of the Gaussian Processes that pass through ($\mathbf{x}^*$,$f^*$). More recently, [12] utilised knowledge of the value $f^*$ to ensure the Gaussian Process posterior mean did not exceed $f^*$ (in the case of maximization problems) and derived two new acquisition functions, Confidence Bound Minimization (CBM) and Expected Regret Minimization (ERM) for use with a bounded GP surrogate. GP CBM and GP ERM both outperformed GP EI for Bayesian optimization [12].

GPs have two known weaknesses [14, 15, 19]. First, low probability is assigned to remote outlier locations in $\mathbf{x}$, despite some applications, such as aerostructural engineering design problems [19], indicating otherwise. 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 [14, 15, 19], 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 [19].

One recently proposed solution to the weaknesses of GPs are Student-t Processes (STPs), which use the multivariate Student's-t distribution [14, 15, 19]. STPs generalize the multivariate Gaussian distribution. STPs have an additional scalar parameter $\nu$ ($\nu > 2$), which defines the 'degrees-of-freedom' of the STP [14, 15, 19] and controls STP kurtosis, influencing the size of the tails and hence, 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 [14, 15, 19], 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 [14, 15, 19]. 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 $f^*$ 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 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, 13]. GPs are simply defined 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 $\mathbf{x}'$ [13]:

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

The GP posterior covariance $\hat{\Sigma}_{GP}$ is given by [13, 19]:

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

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 [19]. As can be seen, the GP posterior covariance does not depend on the black-box function's $y$-targets [13].

### 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 [14, 15, 19]. Like GPs, STPs are simply defined by two functions and a third scalar parameter, $\nu$ ($\nu > 2$). 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 $\mathbf{x}'$ [19]. A stochastic process $f(\mathbf{x})$ is Student's-t when observations jointly sampled have a multivariate Student's-t probability distribution [14, 15, 19]:

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

The STP posterior covariance $\hat{\Sigma}_{STP}$ is given by [19]:

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

where: $y^T K_{\mathbf{x},\mathbf{x}}^{-1} y$ is the squared Mahalanobis distance of the training locations $\mathbf{x}_k$ using their covariance [19]. 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 [13] and the Matérn class of covariance functions e.g. Matérn 3/2 and Matérn 5/2 [13]. Both GPs and STPs can use these kernels.

## 3 EXPLOITING PRIOR KNOWLEDGE OF A GLOBAL OPTIMUM

### 3.1 ERM for Gaussian Processes

With the optimum objective function value $f^*$ known, we can define the regret of evaluating $\mathbf{x}$ as $r(\mathbf{x}) = f^* - f(\mathbf{x})$. Therefore the goal of optimization is achieved if we minimize regret, i.e. find $\mathbf{x}^*$ such that $f(\mathbf{x}^*) = f^*$, so that $r(\mathbf{x}^*) = 0$. Nguyen and Osborne combined prior knowledge about a global optimum $f^*$ with a GP surrogate, to enhance Bayesian optimization by minimizing the expected regret $\mathbb{E}[r(\mathbf{x})]$ [12]. The surrogate's posterior mean is now closer to the known $f^*$ and has low variance to ensure the surrogate's estimation at the chosen $\mathbf{x}$ is correct [12]. ERM selects $\mathbf{x}$ to minimize expected regret - in contrast, EI chooses $\mathbf{x}$ to balance exploration and exploitation [10, 11]. The GP ERM acquisition function $\alpha_{GP}^{ERM}(\mathbf{x})$ is [12]:

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

where: $\hat{\mu}_{GP}(\mathbf{x})$ and $\hat{\sigma}_{GP}(\mathbf{x})$ are the respective GP posterior mean and GP posterior standard deviation; $z = \frac{f^* - \hat{\mu}_{GP}(\mathbf{x})}{\hat{\sigma}_{GP}(\mathbf{x})}$; with $\phi(z)$ and $\Phi(z)$ the standard normal probability density function (PDF) and cumulative distribution function (CDF) respectively. The $[f^* - \hat{\mu}_{GP}(\mathbf{x})]\Phi(z)$ term is low for (i.e. favours) $\mathbf{x}$ for which $f(\mathbf{x})$ is predicted to be close to the known optimum $f^*$, whilst the $\hat{\sigma}_{GP}(\mathbf{x})\phi(z)$ term is low for (again, favours) $\mathbf{x}$ for which the uncertainty in $f(\mathbf{x})$ is low.

### 3.2 ERM for Student's-t Processes

Now consider the univariate Student's-t PDF, with mean $\mu$, standard deviation $\sigma$ and degrees-of-freedom $\nu$. For simplicity, define $C$ as [19]:

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

The Student's-t PDF becomes [19]:

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

Define the STP expected likelihood of regret as [12]:

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

The STP ERM acquisition function $\alpha^{ERM}_{STP}(\mathbf{x})$ minimizes the STP expected regret in Eq. 1 and is [12, 19] [1]:

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

where: $\hat{\mu}_{STP}(\mathbf{x})$ and $\hat{\sigma}_{STP}(\mathbf{x})$ are respectively the STP posterior mean and STP posterior standard deviation; $z_s = \frac{f^*-\hat{\mu}_{STP}(\mathbf{x})}{\hat{\sigma}_{STP}(\mathbf{x})}$; with $\phi_s(z_s)$ and $\Phi_s(z_s)$ the standard Student's-t PDF and CDF respectively. As for GP ERM, the $[f^*-\hat{\mu}_{STP}(\mathbf{x})]\Phi_s(z_s)$ term is low for (again, favours) $\mathbf{x}$ for which $f(\mathbf{x})$ is predicted to be close to the known optimum $f^*$, whilst the $\hat{\sigma}_{STP}(\mathbf{x})\left(\frac{\nu}{\nu-1}\right)\left(1+\frac{z_s^2}{\nu}\right)\phi_s(z_s)$ term favours $\mathbf{x}$ for which the uncertainty in $f(\mathbf{x})$ is low. Algorithm 1 defines Bayesian optimization [2, 7, 16], with the surrogate trained at each iteration using Algorithm 2.1 of [13].

**Algorithm 1:** Bayesian optimization [2, 7, 16]:

(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 $\mathcal{D}_{k-1}$ [13]
(5) select: $\mathbf{x}_k = \arg\min_{\mathbf{x}\in\chi} \alpha(\mathbf{x})$
(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 EXPERIMENTS

Four Bayesian optimization experiments are programmed in the Python language, using the 'pyGPGO' package [8] [2]. Each uses [13] 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 [19]. Algorithm 1 can efficiently seek a global minimum (rather than a global maximum), by multiplying both $f(\mathbf{x})$ and $f^*$ by -1.

### 4.1 Synthetic Functions

Three synthetic functions popular in mathematical optimization are chosen, namely SixHumpCamel, Rosenbrock and Hartmann3 [18]. Each problem's global optimum is sought by Bayesian optimization with ERM, using STPs versus GPs. Both use a squared-exponential covariance kernel [13]. $\nu = 5$ [19] is chosen for each STP surrogate.

[1] https://www.researchgate.net/publication/342201306_Appendix_Derivation_of_Expected_Regret_Minimization_for_Bayesian_Optimization_with_Student's-t_Processes
[2] https://github.com/CPJClare/ERM-for-BayesOpt-with-STPs

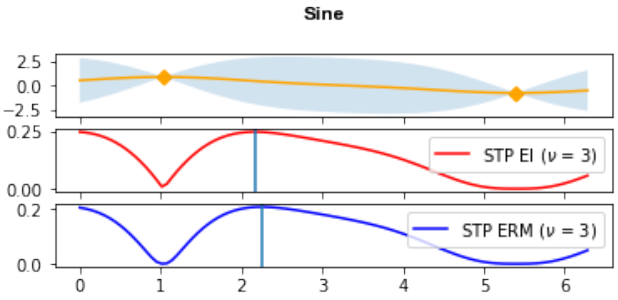


**Figure 1: Optimizing the Sine function using Algorithm 1: The surrogate is STP ($\nu$ = 3) [4] and the kernel is squared-exponential [13]. The model is randomly-initialized using n = 2 locations (top). The first location x after random-initialization is shown (light-blue, vertical line), using two acquisition functions: STP ($\nu$ = 3) EI (middle) versus STP ($\nu$ = 3) ERM (bottom). This is the first iteration using lines 4-7 of Algorithm 1.**

Each model is randomly-initialized with n = 5 iterations [5] and N = 100 post-initialization iterations [19] for each of the three synthetic functions.

The results are shown in Figures 2 - 4, with experiments independently repeated 20 times [12] 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 ($\nu$ = 5) ERM. The red curved lines represent the median for GP ERM, while the blue curved lines show the median for STP ($\nu$ = 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 ($\nu$ = 5) ERM. For each experiment, the training regret IQR for STP ($\nu$ = 5) ERM is lower than the IQR for GP ERM.

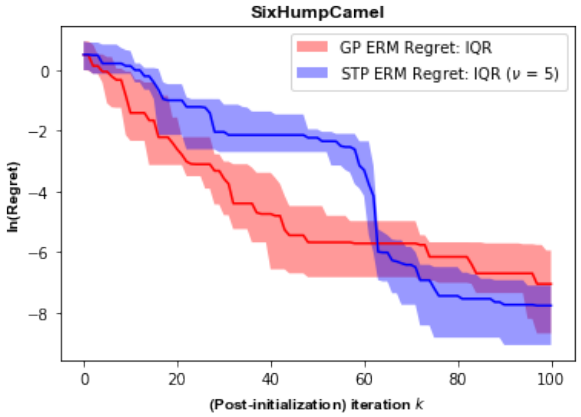


**Figure 2: Comparing the SixHumpCamel function [18]: The training regret IQR for STP ($\nu$ = 5) [19] ERM is lower than the IQR of GP ERM.**

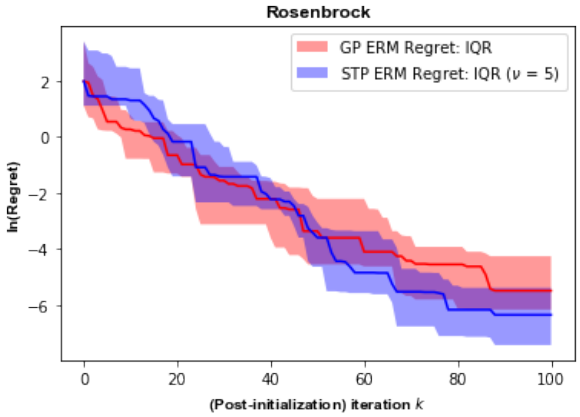


**Figure 3: Comparing the Rosenbrock function [18]: The training regret IQR for STP ($\nu = 5$) [19] ERM is lower than the IQR of GP ERM.**

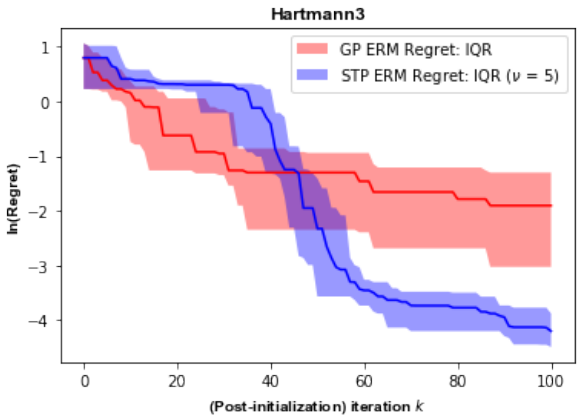


**Figure 4: Comparing the Hartmann3 function [18]: The training regret IQR for STP ($\nu = 5$) [19] ERM is lower than the IQR of GP ERM.**

## 4.2 Application: XGBoost Hyperparameter Tuning

Recently, [12] applied Bayesian optimization with GP ERM to hyperparameter tuning [12] for XGBoost classification [3]. The data was "Skin Segmentation" [3]. Our work enhances Bayesian optimization by comparing STP ($\nu = 5$) ERM versus GP ERM [12]. The data is split 85/15 between training and testing [4]. 3-fold cross-validation of the XGBoost classifier is averaged to measure $y$ [4, 12]. We use a logistic objective function [4, 12], with 5 random initializations [5] and N = 30 post-initialization iterations [4, 12]. The surrogacy training results are shown in Figure 5 and independently repeated 20 times [12] for both STP ($\nu = 5$) ERM and GP ERM. XGBoost classification hyperparameters chosen using STP ($\nu = 5$) ERM outperform GP ERM. The training regret IQR for STP ($\nu = 5$) ERM (blue shading) is lower than the IQR for GP ERM (red shading).

[3]https://archive.ics.uci.edu/ml/datasets/Skin+Segmentation

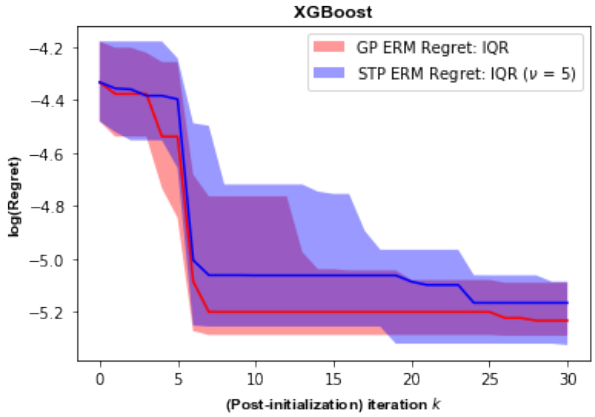


**Figure 5: Hyperparameter tuning for XGBoost classification [3] training accuracy [4, 12]. The training regret IQR for STP ($\nu = 5$) [19] ERM is lower than the IQR of GP ERM [12].**

## 5 CONCLUSIONS

This paper exploits prior knowledge of a global optimum $f^*$ [12] to derive the STP ERM acquisition function and compares Bayesian optimization with ERM, using STPs ($\nu = 5$) versus GPs. Our work enhances Bayesian optimization by showing STP ($\nu = 5$) ERM outperforms GP ERM on three popular synthetic problems [18] and one real-world application [12] in mathematical optimization. Rather than choosing $\nu = 5$ [19], future work will consider STP ERM with prior $\nu$ chosen using Kullback-Leibler divergence [20].

## 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] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *CoRR* abs/1603.02754 (2016). arXiv:1603.02754 http://arxiv.org/abs/1603.02754
[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] Donald R. Jones. 2001. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *J. of Global Optimization* 21, 4 (Dec. 2001), 345–383.

https://doi.org/10.1023/A:1012771025575

[10] Donald R. Jones, Matthias Schonlau, and William J. Welch. 1998. Efficient Global Optimization of Expensive Black-Box Functions. *J. of Global Optimization* 13, 4 (Dec. 1998), 455–492. https://doi.org/10.1023/A:1008306431147

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

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

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

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

[15] 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).

[16] 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

[17] 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.

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

[19] 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

[20] Cristiano Villa and Francisco J. Rubio. 2017. Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula. arXiv:stat.ME/1701.05638

[21] 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.