Bayesian Optimization using Student-t Processes

Amar Shah
Department of Engineering
Cambridge University
as793@cam.ac.uk

Andrew Gordon Wilson
Department of Engineering
Cambridge University
agw38@cam.ac.uk

Zoubin Ghahramani
Department of Engineering
Cambridge University
zoubin@eng.cam.ac.uk

Abstract

Finding the global minimum of a function is often difficult. We consider efficiently minimizing functions which are computationally expensive to evaluate. A Bayesian approach to the global function optimization problem places a prior distribution on the function and chooses where to evaluate the function based on its posterior distribution given a set of observations. While many recent applications use Gaussian processes as a prior for the objective function, here we show that a Student-t process is an ideal prior for such a problem, as it is also nonparametric, but naturally models heavy tailed behaviour and has a predictive covariance which explicitly depends on observations.

1 The Student-t Process

We begin by deriving the Student-t process[1], and its marginal likelihood and predictive distribution, starting from a hierarchical Gaussian process model. We define a prior over continous functions using the following generative model

$$r^{-1} \sim \Gamma\Big(\frac{\nu}{2}, \frac{\rho}{2}\Big) \qquad y|r \sim \mathcal{GP}\big(0, r(\nu-2)k/\rho\big), \tag{1}$$

where $\nu > 2$, $\rho > 0$ and $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is a kernel function. If we marginalize over r, y is a scaled mixture of Gaussian processes. Suppose $\boldsymbol{y} = (y_1, ..., y_N)$ is a finite collection of observations at input points $\boldsymbol{x}_1, ..., \boldsymbol{x}_N \in \mathbb{R}^D$ and let K be the Gram matrix such that $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. We can compute the marginal probability of these observations under the generative prior above as follows

$$\begin{aligned} p(\boldsymbol{y}) = \int p(\boldsymbol{y}|r)p(r)dr &= \frac{(\pi(\nu-2))^{-N/2}\big(\rho/2\big)^{\frac{\nu+N}{2}}}{|K|^{1/2}\Gamma(\nu/2)} \int r^{-(\nu+N)/2-1} \exp\Big(-\frac{\rho}{2r}\Big(1+\frac{\boldsymbol{y}^\top K^{-1}\boldsymbol{y}}{\nu-2}\Big)\Big)dr \\ &= (\pi(\nu-2))^{-N/2}\frac{\Gamma((\nu+N)/2)}{\Gamma(\nu/2)}|K|^{-1/2}\Big(1+\frac{\boldsymbol{y}^\top K^{-1}\boldsymbol{y}}{\nu-2}\Big)^{-(\nu+N)/2}. \end{aligned} \tag{2}$$

[1] The Student-t process [O'Hagan, 1991, O'Hagan et al., 1999] has been used in a number of applications [Yu et al., 2007, Zhang and Yeung, 2010, Xu et al., 2011]. Our parameterization differs slightly from previous constructions.

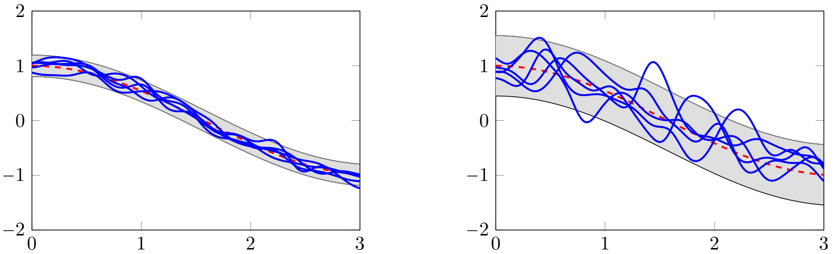

Figure 1: Five samples (blue solid) from $\mathrm{GP}(h, \kappa)$ (left) and $\mathrm{TP}(\nu, h, \kappa)$ (right), with $\nu = 3$, $h(x) = \cos(x)$ (red dashed) and $\kappa(x_i, x_j) = 0.01 \exp(-20(x_i - x_j)^2)$. The shaded areas are 95% predictive intervals. Despite having the same mean and marginal variance, the Student-t process has more extreme excursions.

Thus $\boldsymbol{y}$ is marginally multivariate Student-t distributed, with mean $\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\mathbb{E}[\boldsymbol{y}|r]] = 0$ and covariance $\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] = \mathbb{E}[\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top|r]] = \mathbb{E}[r(\nu-2)K/\rho] = K$. We write $\boldsymbol{y} \sim \mathrm{MVT}_N(\nu, 0, K)$. Notice the redundancy in the parameter ρ; without loss of generality we set $\rho = 1$. It is clear from this generative process that a subset of $\{y_1, ..., y_N\}$ will be multivariate Student-t distributed, which motivates defining a Student-t process.

Definition 1. *f is a Student-t process with scale parameter $\nu > 2$, mean function μ and kernel k if for any finite collection of inputs $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$ the vector $(f_1, ..., f_N)^\top$ is multivariate Student-t distributed with scale ν, mean $(\mu(\boldsymbol{x}_1), ..., \mu(\boldsymbol{x}_N))^\top$ and covariance K where $K_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. We write $f \sim \mathrm{TP}(\nu, \mu, k)$.*

Samples from a Student-t process are shown in Figure 1. We now derive the MVT conditional distribution.

Lemma 2. *The conditional distribution for a multivariate Student-t is also multivariate Student-t.*

Proof. Suppose $\boldsymbol{y} \sim \mathrm{MVT}_N(\nu, 0, K)$ and let $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ represent the first N_1 and remaining N_2 entries of $\boldsymbol{y}$ respectively. Let $\beta_1 = \boldsymbol{y_1}^\top K_{11}^{-1} \boldsymbol{y_1}$ and $\beta_2 = (\boldsymbol{y_2} - \tilde{\boldsymbol{\phi}}_{\boldsymbol{2}})^\top \tilde{K}_{22}^{-1} (\boldsymbol{y_2} - \tilde{\boldsymbol{\phi}}_{\boldsymbol{2}})$, where $\tilde{\boldsymbol{\phi}}_{\boldsymbol{2}} = K_{21} K_{11}^{-1} \boldsymbol{y_1}$ and $\tilde{K}_{22} = K_{22} - K_{21} K_{11}^{-1} K_{12}$. Note that $\beta_1 + \beta_2 = (\boldsymbol{y} - \boldsymbol{\phi})^\top K^{-1} (\boldsymbol{y} - \boldsymbol{\phi})$. We have

$$p(\boldsymbol{y_2}|\boldsymbol{y_1}) = \frac{p(\boldsymbol{y_1}, \boldsymbol{y_2})}{p(\boldsymbol{y_1})} \propto \left(1 + \frac{\beta_1 + \beta_2}{\nu - 2}\right)^{-\frac{\nu+N}{2}} \left(1 + \frac{\beta_1}{\nu - 2}\right)^{\frac{\nu+N_1}{2}} \propto \left(1 + \frac{\beta_2}{\beta_1 + \nu - 2}\right)^{-\frac{\nu+N}{2}} \tag{3}$$

and hence that $\boldsymbol{y_2}|\boldsymbol{y_1} \sim \mathrm{MVT}_{N_2}\left(\nu + N_1, \tilde{\boldsymbol{\phi}}_{\boldsymbol{2}}, \frac{\nu+\beta_1-2}{\nu+N_1-2} \times \tilde{K}_{22}\right)$. □

The predictive mean is the same as for the Gaussian process, but the predictive covariance now explicitly depends on the observations whilst the Gaussian process predictive covariance does not. This difference between the GP and the TP makes the TP particularly well suited to Bayesian optimization, an application where both accurate predictive means and predictive variances are highly desirable for determining where to evaluate a function sequentially.

2 Bayesian Optimization with Student-t Processes

Our goal is to find the minimum of a function $f(\boldsymbol{x})$ on some compact subset of $\mathbb{R}^D$. A Bayesian approach to such a problem would be to place a prior on the unknown function f and make decisions about where to evaluate the function next while integrating over any uncertainty. An overview of Bayesian optimization methods can be found in Brochu et al. [2010]. Our approach is to place a Student-t process prior on the unknown function f.

2.1 Acquisition Functions

The acquisition function, which we denote as $a : \mathbb{R}^D \to \mathbb{R}_+$, determines where we should next evaluate the function f by choosing the point $\boldsymbol{x}_{\text{next}} = \text{argmax } a(\boldsymbol{x})$ and scoring the utility of evaluating the unknown function at a given input location. The acquisition function will depend on previous observations and the hyperparameters of the Student-t process; we denote this dependence $a(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the Student-t process hyperparameters. We denote the current best value as $\boldsymbol{x}_{\text{best}} = \text{argmin}_{\boldsymbol{x}_n} f(\boldsymbol{x}_n)$, the predictive mean function as $\mu(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta})$ and the predictive variance function as $\sigma(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta})$. We let the λ_ν and Λ_ν denote the probability density and distribution functions respectively of a $\text{MVT}_1(\nu, 0, 1)$ distribution. In this work we use the expected improvement criterion for choosing sequential points to query.

Expected Improvement A sensible strategy is to maximize the expected improvement over the current minimum. This also has a closed form solution under a Student-t process prior

$$\begin{aligned} a_{\text{EI}}(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta}) &= \mathbb{E}[\max\left(f(\boldsymbol{x}_{\text{best}}) - f(\boldsymbol{x}), 0\right) | \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta}] \\ &= \int_{-\infty}^{f(\boldsymbol{x}_{\text{best}})} dy \frac{(f(\boldsymbol{x}_{\text{best}}) - y)}{\sigma} \lambda_{\nu+N}\left(\frac{y-\mu}{\sigma}\right) \\ &= \gamma\sigma\Lambda_{\nu+N}(\gamma(\boldsymbol{x})) + \sigma\left(1 + \frac{\gamma(\boldsymbol{x})^2 - 1}{\nu + N - 1}\right)\lambda_{\nu+N}(\gamma(\boldsymbol{x})), \end{aligned} \tag{4}$$

where $\gamma(\boldsymbol{x}) = \sigma(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta})^{-1}[f(\boldsymbol{x}_{\text{best}}) - \mu(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta})]$. This acquisition function is similar to that of a GP prior [Snoek et al., 2012]; however, the added scale parameter in the Student-t process can have a significant impact on where the acquisition function is maximized, as illustrated in Figure 2. The argmax of the acquisition function changes drastically as ν varies.

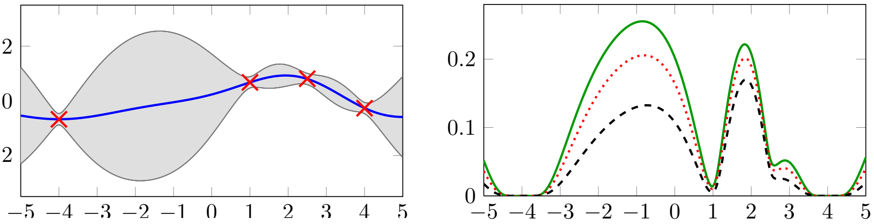

Figure 2: Posterior distribution of a function to maximize under a GP prior (left) and acquisition functions (right). The solid line is the acquisition function for a GP, the dotted and dashed lines are for TP priors with $\nu = 15$ and $\nu = 5$ respectively. All other hyperparameters are kept the same.

2.2 Kernel choice and integrating out uncertainty

We choose to work with a kernel function which is the sum of a ARD Matérn 5/2 kernel and a delta function kernel. As discussed in Snoek et al. [2012], this kernel is a better choice than a squared exponential which can be unrealistically smooth. The ARD Matérn 5/2 is twice differentiable and is defined as

$$K_{\text{M52}}(\boldsymbol{x}, \boldsymbol{x}') = \theta_0\left(1 + \sqrt{5r^2(\boldsymbol{x}, \boldsymbol{x}')}\right)\exp\left(-\sqrt{5r^2(\boldsymbol{x}, \boldsymbol{x}')}\right), \tag{5}$$

where $r^2(\boldsymbol{x}, \boldsymbol{x}') = \sum_{d=1}^{D}(x_d - x'_d)^2/\theta_d^2$. Ideally we would like to analytically integrate out the uncertainty in the hyperparameters by placing a prior on them and working directly with

$$\hat{a}(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}) = \int a(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{6}$$

Since this integral is intractable, our approach, analogous to Snoek et al. [2012], is to sample hyperparameters $\{\boldsymbol{\theta}_h\}_{h=1}^H$ from their posterior distributions and maximize $\tilde{a}(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}) = \frac{1}{H}\sum_{h=1}^{H} a(\boldsymbol{x}; \{\boldsymbol{x}_n, y_n\}, \boldsymbol{\theta}_h)$ to find the next point to query the function being optimized.

3 Experiments

We compare a Student-t process prior with a Matérn plus a delta function kernel to a Gaussian process prior with the same kernel, for Bayesian optimization. To integrate away uncertainty we slice sample the hyperparameters [Neal, 2003]. We consider 3 functions: a 1-dim synthetic sinusoidal, 2-dim Branin-Hoo and a 6-dim Hartmann function. We place uniform priors on the length scales, lognormal priors on the amplitude, noise and $\nu - 2$ and a Gaussian prior on the constant mean. All results are shown in Figure 3.

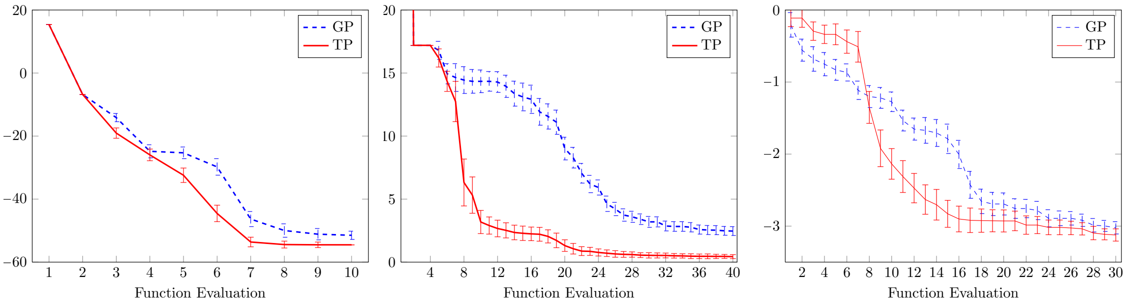

Figure 3: Function evaluations for the synthetic function (left), Branin-Hoo function (centre) and the Hartmann function (right). Evaluations under a Student-t process prior (solid line) and a Gaussian process prior (dashed line) are shown. Error bars represent the standard deviation of 50 runs. In each panel we are minimizing an objective function. The vertical axis represents the running minimum function value.

Sinusoidal synthetic function In this experiment we aimed to find the minimum of $f(x) = -(x - 1)^2 \sin(3x + 5x^{-1} + 1)$ in the interval $[5, 10]$. The function has 2 local minima in this interval. TP optimization clearly outperforms GP optimization in this problem; the TP was able to come to within 0.1% of the minimum in 8.1 ± 0.4 iterations whilst the GP took 10.7 ± 0.6 iterations.

Branin-Hoo function This function is a popular benchmark for optimization methods [Jones, 2001] and is defined on the set $\{(x_1, x_2) : 0 \leq x_1 \leq 15, -5 \leq x_2 \leq 15\}$. We initialized the runs with 4 initial observations, one for each corner of the square on which the function is defined.

Hartmann function This is a function with 6 local minima in $[0, 1]^6$ [Picheny et al., 2013]. The runs are initialised with 6 observations at corners of the unit cube in $\mathbb{R}^6$. Notice that the TP tends to behave more like a step function whereas the Gaussian process' rate of improvement is somewhat more constant. The reason for this behaviour is that the TP tends to more thoroughly explore any modes which it has found, before moving away from these modes. This phenomenon seems more prevelant in higher dimensions.

4 Conclusions

We proposed to use Student-t process priors over functions we wish to optimize, and demonstrate that the TP can outperform the GP for Bayesian optimization in many cases. The added scale parameter in the TP gives it the ability to learn heavy tailed function behaviour. The fact that the predictive covariance for the TP explicitly depends on the data is a useful property which the Gaussian process lacks. The TP therefore seems to have all the benefits of the GP for Bayesian optimization, e.g. it has consistent marginals, an analytically representable conditional distribution and closed form representations of popular acquisition functions, but the TP also appears to add significant flexibility.

Acknowledgments

We wish to thank Kevin Swersky and Jasper Snoek for helpful discussions.

References

E. Brochu, M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Applications to Active User Modeling and Hierarchical Reinforcement Learning. *arXiv*, 2010. URL `http://arxiv.org/abs/1012.2599`.

D. R. Jones. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization*, 21(4):345–383, 2001.

R. M. Neal. Slice Sampling. *Annals of Statistics*, 31(3):705–767, 2003.

A. O'Hagan. Bayes-Hermite Quadrature. *Journal of Statistical Planning and Inference*, 29:245–260, 1991.

A. O'Hagan, M. C. Kennedy, and J. E. Oakley. *Uncertainty Analysis and other Inference Tools for Complex Computer Codes*. Oxford University Press, 1999.

V. Picheny, T. Wagner, and D. Ginsbourger. A Benchmark of Kriging-Based Infill Criteria for Noisy Optimization. *Structural and Multidisciplinary Optimization*, 48(3):607–626, 2013.

J. Snoek, H. Larochelle, and R. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, 2012.

Z. Xu, F. Yan, and Y. Qi. Sparse Matrix-Variate t Process Blockmodel. *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, 2011.

S. Yu, V. Tresp, and K. Yu. Robust Multi-Task Learning with t-Processes. *Proceedings of the 24th International Conference on Machine Learning*, 2007.

Y. Zhang and D. Y. Yeung. Multi-Task Learning using Generalized t Process. *Proceedings of the 13th Conference on Artificial Intelligence and Statistics*, 2010.