Hypervolume-based Multi-objective Bayesian Optimization with Student-t Processes

Joachim van der Herten   **Ivo Couckuyt**

Tom Dhaene
IDLab
Ghent University - imec
iGent Tower - Department of Electronics and Information Technology (INTEC)
Technologiepark-Zwijnaarde 15,
B-9052 Ghent, Belgium
joachim.vanderherten@ugent.be

Abstract

Student-t processes have recently been proposed as an appealing alternative non-parameteric function prior. They feature enhanced flexibility and predictive variance. In this work the use of Student-t processes are explored for multi-objective Bayesian optimization. In particular, an analytical expression for the hypervolume-based probability of improvement is developed for independent Student-t process priors of the objectives. Its effectiveness is shown on a multi-objective optimization problem which is known to be difficult with traditional Gaussian processes.

1 Introduction

The use of Bayesian models and acquisition functions to guide the optimization of expensive, noisy, black-box functions (Bayesian Optimization) has become more popular over the years, and has recently been applied to a wide variety of problems in several fields. The next candidate for evaluation of the computationally expensive black-box function is selected by optimizing an acquisition function relying on Bayesian model(s) approximating the previously observed responses of the black-box function(s). Within Machine Learning, it has for instance been applied to optimize model hyperparameters (Snoek et al., 2012; Swersky et al., 2014), as model training involving a lot of data typically makes use of traditional numeric optimization infeasible.

Within the field of engineering, Jones et al. (1998) introduced the combination of Expected Improvement (EI) and Kriging models for optimization of computer simulations which can take up to several days to perform a single run. This situation is commonly encountered in product design involving, i.a., Computational Fluid Dynamics (CFD) and Finite Element Methods (FEM). Multi-objective (or multi-task) optimization has gained a lot of attention in engineering optimization as product design inherently involves trade-offs as typically several (conflicting) aspects are involved. Frequently used are hypervolume-based acquisition functions such as Hypervolume Expected Improvement (Emmerich et al., 2006) or Hypervolume Probability of Improvement (HvPoI) (Couckuyt et al., 2014), assuming a Gaussian process (GP) prior for each objective. More recently, a Multi-objective version of the Predictive Entropy Search has been proposed (Hernández-Lobato et al., 2016).

Naturally, the correctness of the approximation of the objective(s) is crucial to perform succesful optimization. Erroneous model fits lead to selection of new evaluations based on false beliefs making the discovery of optima unlikely, especially when the input space is large. While GPs have received much attention both as a modeling strategy and within Bayesian optimization, recently Student-t

process priors (TP) have been proposed (O'Hagan, 1991; O'Hagan et al., 1998; Rasmussen, 2006; Shah et al., 2014) for use with EI. In this contribution we consider the TP prior in the context of multi-objective Bayesian optimization, and develop an analytical expression of the HvPoI acquisition function for it accordingly. TPs have shown to be promising, and their properties such as additional flexibility and enhanced predictive variance seem to be appealing properties for Bayesian optimization. A brief overview of TPs is given in Section 2, and the formulation of HvPoI assuming TP priors for the objectives is given in Section 3. The performance of this approach is then compared to HvPoI with GP priors in Section 4 on a multi-objective problem.

2 Student-t Processes

Given a d-dimensional input space $\mathcal{X} \subset \mathbb{R}^d$, f is a Student-t process with degrees of freedom $\nu > 2$, a continuous mean function ϕ and a parametrized kernel function k. For any set $X_n \subset \mathcal{X}$ of n inputs $\boldsymbol{x}$, the mapping of these inputs by f is distributed according to a multivariate Student-t distribution: $\boldsymbol{y} = (f(\boldsymbol{x}_1), ... f(\boldsymbol{x}_n)) \sim \mathrm{MVT}_n(\nu, \boldsymbol{\phi}, K)$ with $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The likelihood corresponds to the probability density function of an MVT:

$$p(\boldsymbol{y}|X_n, \nu, \boldsymbol{\theta}) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{((\nu-2)\,\pi)^{\frac{n}{2}} \Gamma\left(\frac{\nu}{2}\right)} |K|^{-1/2} \left(1 + \frac{\beta}{\nu-2}\right)^{-\frac{\nu+n}{2}}, \tag{1}$$

with $\beta = (\boldsymbol{y}-\boldsymbol{\phi})^T K (\boldsymbol{y}-\boldsymbol{\phi})$. Shah et al. (2014) have shown that considering $\boldsymbol{y}|\sigma \sim \mathcal{GP}\left(\phi, (\nu-2)\sigma\right)$ and marginalizing σ out assuming an *inverse Wishart process* prior, recovers Equation 1.

For an arbitrary $\boldsymbol{x}^\star \in \mathcal{X}$ the predictive distribution is also a MVT:

$$p(f(\boldsymbol{x}^\star)|X_n, \boldsymbol{y}, \boldsymbol{\theta}, \nu) \sim \mathrm{MVT}_1\left(\nu+n, \mu(\boldsymbol{x}^\star), \tilde{s}(\boldsymbol{x}^\star)^2\right), \tag{2}$$

$$\tilde{s}(\boldsymbol{x}^\star)^2 = \frac{\nu+\beta-2}{\nu+n-2} s(\boldsymbol{x}^\star)^2. \tag{3}$$

The quantities μ and s^2 are identical to the predictive mean and variance of a GP (assuming the same kernel and parameters). Recent work also shows marginalizing the output scale also yields a related MVT predictive distribution Gramacy and Apley (2015); Montagna and Tokdar (2016). This differs from non-analytical marginalization of the kernel lengthscales with Markov chain Monte Carlo methods as applied frequently in Bayesian optimization (see van der Herten et al. (2016) for a comparison of the latter with traditional maximum likelihood estimates).

A fundamental difference is observed in Equation 3: the variance prediction includes the observed responses, as opposed to GP which only considers the space between inputs. This allows a TP to anticipate changes in covariance structure. Furthermore it was proven that a GP is a special case of a TP, with $\nu \to \infty$. However, the approach applied for GPs to include noise as part of the likelihood can not be applied for TPs, as the sum of two independent MVTs is not analytically tractable. Instead, a diagonal white noise kernel is added to allow approximation of noisy observations.

3 Hypervolume-based Probability of Improvement

Given a multi-objective (or multi-task) optimization problem, each evaluated input $\boldsymbol{x}_i$ has p observed reponses $\boldsymbol{y}_i = (f_1(\boldsymbol{x}_i), ..., f_p(\boldsymbol{x}_i))$, together forming a matrix $Y \in \mathbb{R}^{n \times p}$. The rows of this matrix correspond to points in the p-dimensional objective space. Of interest are the non-dominated solutions forming the Pareto set $\mathcal{P} \subset Y$. Ideally, we like to find a point $\boldsymbol{x}^\star \in \mathcal{X} : \boldsymbol{p} = f(\boldsymbol{x}^\star)$:

$$\max_{\boldsymbol{x}^\star \in \mathcal{X}} I(\boldsymbol{p}, \mathcal{P}) \tag{4}$$

where $I(.)$ is the improvement function which is defined in this work using the hypervolume indicator as,

$$I(\boldsymbol{p}, \mathcal{P}) = \begin{cases} \mathcal{H}(\mathcal{P} \cup \boldsymbol{p}) - \mathcal{H}(\mathcal{P}) & \boldsymbol{p} \in A \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

with A the non-dominated section of the objective space, $\mathcal{H}(.)$ defined as the hypervolume of the section of the objective space dominated by the Pareto front (bounded by a reference point $\boldsymbol{f}^{\max}$).
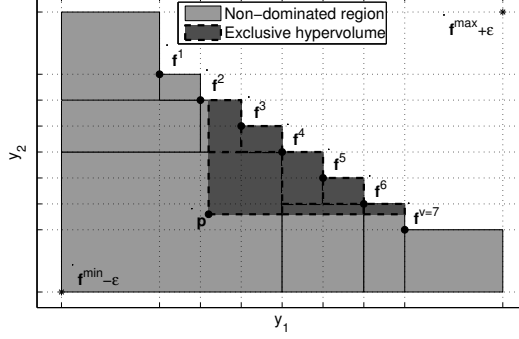
Figure 1: Illustration of a Pareto set (members illustrated by $\boldsymbol{f}^i$) with two objective functions. $\boldsymbol{f}^{\min}$ and $\boldsymbol{f}^{\max}$ denote the ideal and anti-ideal point respectively. The shaded areas (both light and dark) represent the non-dominated region and is decomposed into q cells by a binary partitioning procedure. These cells provide integration bounds to compute $I(\boldsymbol{p}, \mathcal{P})$. Courtesy of (Couckuyt et al., 2014)

The situation is illustrated in Figure 1: the exclusive (or contributing) hypervolume corresponds to $\mathcal{H}(\mathcal{P} \cup \boldsymbol{p}) - \mathcal{H}(\mathcal{P})$. Because $\boldsymbol{p}$ is a (black-box) mapping of p objective functions f of an unknown $\boldsymbol{x}^\star$, and because each evaluation is expensive, direct application of traditional numerical optimization methods is infeasible. Instead, we approximate each f and optimize an acquisition function incorporating the information provided by the predictive distributions of the approximations of the objectives. The optimum of the acquisition yields a candidate $\boldsymbol{x}^\star$ to be evaluated on all f.

We propose the *Hypervolume Probability of Improvement* as proposed earlier by Couckuyt et al. (2014) as it is tractable and scales to a higher number of objectives, however we assume each $f_i \sim \mathcal{TP}$ instead of a GP. Formally, this acquisition function is defined as

$$P_{\mathrm{hv}}(\boldsymbol{x}) = I(\boldsymbol{\mu}, \mathcal{P}) P[\boldsymbol{x} \in A] \tag{6}$$

$$\boldsymbol{\mu} = [\mu_1(\boldsymbol{x}), ..., \mu_p(\boldsymbol{x})]. \tag{7}$$

The latter term of the multiplication represents the probability a new point is located in A and, hence, requires an integration over that region. Exact computation of this integral is performed by decomposing A into q cells spanned by upper and lower bounds $[\boldsymbol{l}^k, \boldsymbol{u}^k]$. This decomposition can be done using a binary partition algorithm (which scales poorly as the number of objectives grows) as illustrated in Figure 1, or by applying faster algorithms such as the Walking Fish Group (While et al., 2012). We then make use of the predictive distribution of the TPs:

$$P[\boldsymbol{x} \in A] = \sum_{k=1}^{q} \prod_{j=1}^{p} \left(\Phi_{\nu+n} \left(\frac{u_j^k - \mu_j(\boldsymbol{x})}{\tilde{s}_j(\boldsymbol{x})} \right) - \Phi_{\nu+n} \left(\frac{l_j^k - \mu_j(\boldsymbol{x})}{\tilde{s}_j(\boldsymbol{x})} \right) \right) \tag{8}$$

Φ_ν represents the cumulative density function of a $\mathrm{MVT}_1(\nu, 0, 1)$. In addition, we can simply compute the volume of the exclusive volume using the existing q cells with no extra computation as follows (assuming $\boldsymbol{\mu}$ is non-dominated):

$$\mathcal{H}(\mathcal{P} \cup \boldsymbol{\mu}) - \mathcal{H}(\mathcal{P}) = \sum_{k=1}^{q} \prod_{j=1}^{p} \left(u_j^k - \left(\max l_j^k, \mu_j(\boldsymbol{x}) \right) \right) \tag{9}$$

4   Illustration

We illustrate the effectiveness of the TP prior on the DTLZ1 function, including 6 input parameters and 3 output parameters. The function itself is computed analytically, with some mild Gaussian noise added. Couckuyt et al. (2014) report difficulties approximating the first objective, hence we try the traditional HvPoI in combination with GP priors, and compare it with the modified version as introduced in Section 3 with TP priors. The initial set of data points consists of an optimized Latin Hypercube of 10 points. The acquisition function is then permitted to select an additional 30 data points for evaluation. For both TP and GP, the RBF kernel was used, and the hyperparameters $\boldsymbol{\theta}$ and ν were optimized with multi-start Sequential Quadratical Programming. Note that the optimization
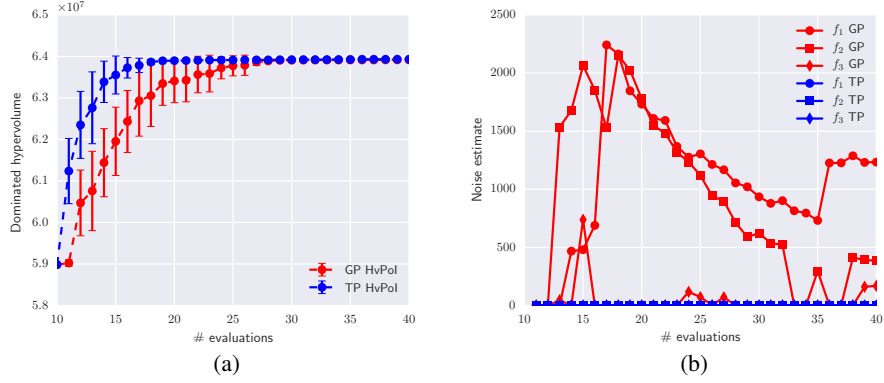
3

Figure 2: (a) Comparison of the growth of the dominated hypervolume for the DTLZ1 function, for 10 experiments using both GP and TP priors for the objectives. The mean and 95% confidence intervals are shown. (b) The noise parameter for all three objectives approximated by GP and TPs. For GPs, the noise is part of the likelihood whereas for TPs a diagonal matrix was added to the kernel matrix. Clearly, the TPs are more flexible and do not consider the evaluated data noisy.

can result in a very large value ν, causing the TP to become a GP. Hence, we expect better or equal performance, not worse. Both experiments were repeated 10 times.

As performance metric, the hypervolume indicator (size of the dominated hypervolume with respect to a fixed reference point $\boldsymbol{f}^{\max} = [400, 400, 400]$) is recorded after every function evaluation. The average hypervolume and 95% confidence intervals were computed and plotted in Figure 2a. Clearly, the runs using the TP approximations of the objectives obtain larger hypervolumes faster. The GP experiments lag behind although they also eventually manage to obtain the same hypervolume indicator performance after additional evaluations. In the end, TPs are able to find a decent hypervolume in about 30% of the function evaluations needed by the GPs for the same hypervolume indicator performance.

Closer investigation reveals the GP approximations for some of the objective functions have large noise levels, varying significantly as more evaluations are added, whereas the TPs do not as illustrated in Figure 2b. It seems the GP is not flexible enough to approximate the objective functions and has to increase the noise variance to avoid ill-conditioning of the kernel matrix. The TPs compensate for this by decreasing the degrees of freedom, which also affects the prediction variance resulting in better selection of evaluation candidates.

5   Conclusion

Student-t processes present themselves as an appealing alternative for Gaussian Processes in the context of Bayesian Optimization. Their robustness was proven earlier by Shah et al. (2014) and their enhanced prediction variance can make them more informative for acquisition functions leading to faster discovery of optima. We demonstrated this on a multi-objective optimization problem, using an adapted Hypervolume Probability of Improvement (HvPoI) criterion.

To make better use of the enhanced prediction variance we aim to adapt the Hypervolume Expected Improvement in further work, as the HvPoI acquisition function does not consider the improvement part of the integration (Couckuyt et al., 2014). In addition other acquisition functions can be modified to be used with TPs, although for some of the more complex acquisition functions the Student-t distribution might introduce tractability challenges. We will be looking at multivariate TPs for multi-objective optimization as in Hernández-Lobato et al. (2016). Objectives can then be evaluated indepedently depending on the expected information gain of each.

4

References

Ivo Couckuyt, Dirk Deschrijver, and Tom Dhaene. Fast calculation of multiobjective probability of improvement and expected improvement criteria for Pareto optimization. *Journal of Global Optimization*, 60(3):575–594, 2014. ISSN 0925-5001.

Michael TM Emmerich, Kyriakos C Giannakoglou, and Boris Naujoks. Single-and multiobjective evolutionary optimization assisted by gaussian random field metamodels. *IEEE Transactions on Evolutionary Computation*, 10(4):421–439, 2006. ISSN 1089-778X.

Robert B Gramacy and Daniel W Apley. Local Gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578, 2015.

Daniel Hernández-Lobato, José Miguel Hernández-Lobato, Amar Shah, and Ryan P Adams. Predictive Entropy Search for Multi-objective Bayesian Optimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, pages 1492–1501. JMLR Workshop and Conference Proceedings, 2016.

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. ISSN 0925-5001.

Silvia Montagna and Surya T Tokdar. Computer Emulation with Nonstationary Gaussian Processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):26–47, 2016.

Anthony O'Hagan. Bayes–hermite quadrature. *Journal of statistical planning and inference*, 29(3): 245–260, 1991.

Anthony O'Hagan, JM Bernardo, JO Berger, AP Dawid, AFM Smith, et al. Uncertainty analysis and other inference tools for complex computer codes. 1998.

Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

Amar Shah, Andrew Gordon Wilson, and Zoubin Ghahramani. Student-t Processes as Alternatives to Gaussian Processes. In *AISTATS*, pages 877–885, 2014.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014.

Joachim van der Herten, Ivo Couckuyt, Dirk Deschrijver, and Tom Dhaene. Fast Calculation of the Knowledge Gradient for Optimization of Deterministic Engineering Simulations. *arXiv preprint arXiv:1608.04550*, 2016.

Lyndon While, Lucas Bradstreet, and Luigi Barone. A fast way of calculating exact hypervolumes. *IEEE Transactions on Evolutionary Computation*, 16(1):86–95, 2012. ISSN 1089-778X.