Active learning for feasible region discovery

Nicolas Knudde*, Ivo Couckuyt†, Tom Dhaene‡
Department of Information Technology
Ghent University - imec, IDLab
Ghent, Belgium
*nicolas.knudde@ugent.be, †ivo.couckuyt@ugent.be, ‡tom.dhaene@ugent.be

Kohei Shintani
Toyota Motor Corporation
Aichi, Japan
kohei_shintani@mail.toyota.co.jp

***Abstract*—Often in the design process of an engineer, the design specifications of the system are not completely known initially. However, usually there are some physical constraints which are already known, corresponding to a region of interest in the design space that is called feasible. These constraints often have no analytical form but need to be characterised based on expensive simulations or measurements. Therefore, it is important that the feasible region can be modeled sufficiently accurate using only a limited amount of samples.**

This can be solved by using active learning techniques that minimize the amount of samples w.r.t. what we try to model. Most active learning strategies focus on classification models or regression models with classification accuracy and regression accuracy in mind respectively. In this work, regression models of the constraints are used, but only the (in)feasibility is of interest.

To tackle this problem, an information-theoretic sampling strategy is constructed to discover these regions. The proposed method is then tested on two synthetic examples and one engineering example and proves to outperform the current state-of-the-art.

***Index Terms*—active learning, feasible region, Gaussian Process**

I. Introduction

In a lot of practical engineering problems, the designing process depends on simulations and prototypes. These simulations have become increasingly important over the last decades, because of their low cost and the increasing available computational power. However, because of the increasing precision of these simulators, one simulation sometimes can take hours to days [1]. This computational cost can be a large bottleneck during the design process.

This restriction ignited the birth of surrogate-based designing, which uses a data-efficient machine learning model that is cheap to evaluate and mimics the original system (whether it be a simulation or measurements), called a surrogate model. This surrogate model is constructed using only a limited amount of evaluations of the original system, which makes it computationally faster. This model can be used for different purposes like optimization, sensitivity analysis and domain exploration of the system under study.

The choice of surrogate model depends on the availability of data, the type of data and the desired properties of the model. Examples of surrogate models include Random Forests [2], Support Vector Machines, Least-Squares Support Vector Machines [3], Multi-Layer Perceptrons (MLPs) [4; 5], Bayesian Neural Networks [6], Gaussian Processes (GPs) [7; 8] and linear regression. The GP will be used here because of its flexibility and predictive distribution.

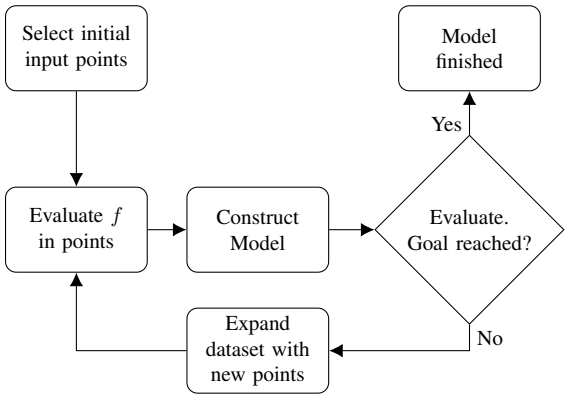


Fig. 1. Flowchart of active learning

The training data can be chosen according to different methodologies. For example, the samples can be chosen all at once at the start. Without any prior knowledge this leads to a space-filling design which is not very efficient as no information about the problem is taken into account. It is impossible to know upfront if there are too few samples to obtain an accurate model, or too many samples such that it is not efficient. On top of that there will be too many samples in regions where the system does not have an interesting behavior and too few in regions where it does. An alternative to this approach is Active Learning (AL), which intelligently and sequentially extends the dataset to suit the desired properties of the model (e.g., high accuracy).

In Bayesian AL the sampling, the sampling strategy is of paramount importance and usually depend on an acquisition function. The location of the maximum of this acquisition function determines the next sample. Depending on the purpose of modeling there are different choices of acquisition functions. For example when one is merely interested in the optimum of the function, the paradigm of AL is also called Bayesian Optimization (BO) and acquisition functions like Entropy Search (ES), Expected Improvement (EI), Upper Confidence Bound (UCB) and Probability of Improvement (PI) can be used. If information about the class regions in

a classification problem is of interest one can use Bayesian Active Learning by Disagreement (BALD). When the model will be used for extensive future studies, the global accuracy is more crucial and sampling can be done at the location with the largest predictive variance.

In this paper, a slightly different problem will be studied. Often in engineering design, initially the objectives that need to be optimized are not yet clearly defined. Often, however, it is already clear that some quantities have to be within certain ranges. These ranges define feasible regions, i.e., regions in which the function f at hand takes on admissible values $a < f(\mathbf{x}) < b$. It is useful to know these regions in order to limit the size of the design phase. To gain information about this region, often simulations have to be done, which are computationally expensive. In this work, an efficient AL method is developed that maximizes the information about this region using information-theoretic sampling employing GP modeling, extending optimization sampling schemes like [9; 10; 11]. This method is evaluated using benchmark functions and proves to perform well.

II. Gaussian Processes

There are a wide variety of supervised regression models that can be used as a surrogate model. The most important ones are Neural Networks [6], Kriging [12; 13], Gaussian Processes (GP) [7; 8] and Least Squares Support Vector Machines (LS-SVM) [3]. The latter three surrogate models are kernel-based regression methods, for which the prediction is a linear combination of kernel evaluations between the test and data points.

In data-efficient machine learning the most frequently used surrogate model is the Gaussian Process. This GP automatically guards against overfitting [7], is analytically tractable and provides a predictive distribution for any given point. The latter is important for active learning, since it provides a measure of model uncertainty, which makes it straightforward to explore the space taking the data into account. More formally, GPs are a powerful non-parametric Bayesian model which represents a distribution over functions $f : \mathcal{X} \to \mathbb{R}$. A Gaussian process is completely defined by a mean function $m : \mathcal{X} \to \mathbb{R}$ and a covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ in the sense that every finite set of function values $[f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_N)]$ is distributed according to a multivariate Gaussian with mean $\mathbf{m}$ and covariance $K_{\mathbf{xx}}$, where $\mathbf{m}_i = m(\mathbf{x}_i)$ and $(K_{\mathbf{xx}})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. We write this as $f \sim \mathcal{GP}(m, k)$. In this work, a mean function which is zero everywhere is used. The hyperparameters θ of the GP are optimized using the Maximum Likelihood Estimation (MLE):

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{f}|\theta) \tag{1}$$

$$= \arg\max_{\theta} -\frac{1}{2}\left(\log|2\pi K_{\mathbf{xx}}| + \mathbf{f}^T K_{\mathbf{xx}}^{-1}\mathbf{f}\right) \tag{2}$$

When considering the prediction of the model for new test points $X_\star = [\mathbf{x}_{\star 1}, ..., \mathbf{x}_{\star N_\star}]$ some notation conventions for the kernel matrices will be used:

$$(K_{\mathbf{xx}})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \tag{3}$$

$$(K_{\star\mathbf{x}})_{ij} = k(\mathbf{x}_{\star i}, \mathbf{x}_j), \tag{4}$$

$$(K_{\star\star})_{ij} = k(\mathbf{x}_{\star i}, \mathbf{x}_{\star j}). \tag{5}$$

Once the optimal hyperparameters have been determined the predictive distribution for new testing inputs $X_\star$ can be calculated and becomes a Gaussian distribution with the following moments [14]:

$$\mu(\mathbf{x}) = \mathbb{E}(\mathbf{f}_\star|X_\star, \mathcal{D}_n) = K_{\star\mathbf{x}}K_{\mathbf{xx}}^{-1}\mathbf{y}_n, \tag{6}$$

$$\sigma^2(\mathbf{x}) = \mathrm{Var}(\mathbf{f}_\star|X_\star, \mathcal{D}_n) = K_{\star\star} - K_{\star\mathbf{x}}K_{\mathbf{xx}}^{-1}K_{\star\mathbf{x}}^T. \tag{7}$$

All the kernels in this work are Squared Exponential (SE) also known as Radial Basis Function (RBF) kernels, which are of the following form:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_k^2 \exp\left(-\sum_{d=1}^{D}\frac{(\mathbf{x}_d - \mathbf{x}'_d)^2}{2\ell_d^2}\right). \tag{8}$$

This is the Automatic Relevance Detection (ARD) [7] version of the kernel, which means there is a separate lengthscale for every dimension. It is used to eliminate irrelevant input features, as the lengthscales increase for irrelevant dimensions [7].

III. Active Learning

A flowchart of the sequential sampling approach is given in Figure 1. First, the function of interest f is evaluated for an initial set of input points. This initial exploratory set of input points is selected to be as space filling as possible, as this will give an idea of the general behavior of the function. There are different possible choices of initial design such as factorial designs [15] and (optimal) Latin Hypercube designs [16].

In the next step a surrogate model is constructed, using the evaluated data $\mathcal{D}_n = \{\mathbf{x}_i, y_i\}_{i=1}^N$. This surrogate model mimics the behavior of the real, expensive-to-evaluate function, and thus provides a cheap alternative. GPs are arguably the most standard surrogate model.

Next, it is checked if our goal is reached, like having an accurate enough model, or having reached an optimal value. If the goal is not reached more data has to be gathered by running the simulator. To determine which data point has to be evaluated, a sampling strategy or acquisition function α has to be constructed. The next sampling location is then determined by the maximum of this acquisition function. This sampling strategy depends on the quantity that one is interested in as will be discussed later. The location where the acquisition function is maximum is used as an extra sample and evaluated. This data point is added to the original dataset the process can be repeated.

IV. Information-theoretic sampling for feasibility

In Bayesian active learning a possible approach is to maximize the loss in entropy of the posterior distribution of the quantity of interest g [17]:

$$\alpha(\mathbf{x}) = \mathbb{H}(p(g|\mathcal{D}_n)) - \mathbb{E}_{p(f|\mathcal{D}_n,\mathbf{x})}(\mathbb{H}(p(g|\mathcal{D}_n \cup \{(\mathbf{x}, f)\}))). \tag{9}$$

As this can be seen as the mutual information between g and f given D_n, and the mutual information is symmetric this can also be written as [18]:

$$\alpha(\mathbf{x}) = \mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x})) - \mathbb{E}_{p(g|\mathcal{D}_n)}(\mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x}, g))). \tag{10}$$

This concept has been used in active learning for purposes like preference learning [18], classification [18], single-objective optimization [11; 19; 20] and multi-objective optimization [10].

In this particular problem statement, the goal is to maximize the information about the regions of feasibility and infeasibility. A novel acquisition function is introduced that uses the information about each of the three regions defined by $b < f(\mathbf{x})$, $a < f(\mathbf{x}) < b$ and $f(\mathbf{x}) < a$. Using Equation 10, this becomes:

$$\begin{aligned}\alpha(\mathbf{x}) = 3\mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x})) &- \mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x}, f > b)) \\ &- \mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x}, a < f < b)) \\ &- \mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x}, f < a)).\end{aligned} \tag{11}$$

$p(f|\mathcal{D}_n, \mathbf{x}, f > b)$, $p(f|\mathcal{D}_n, \mathbf{x}, a < f < b)$ and $p(f|\mathcal{D}_n, \mathbf{x}, f < a)$ are all truncated Gaussian distributions, for which the entropy can be calculated analytically using straightforward properties of normal distributions. Hence, in general a single entropy term in Equation 11 becomes:

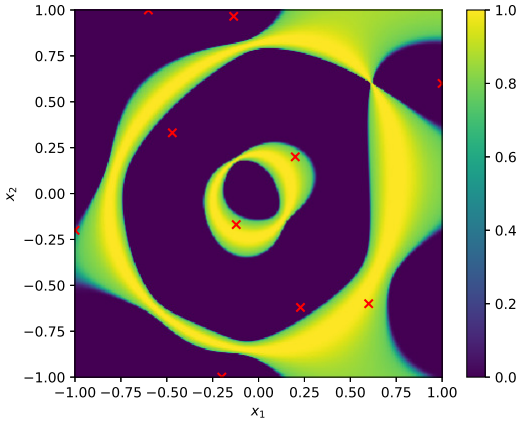
$$\mathbb{H}(p(f|\mathcal{D}_n, \mathbf{x}, \alpha < f < \beta)) \tag{12}$$

$$= \mathbb{H}\left(\frac{1}{Z}\mathcal{N}(f|\mu(\mathbf{x}), \sigma^2(\mathbf{x}))\mathbb{I}(\alpha < f < \beta)\right) \tag{13}$$

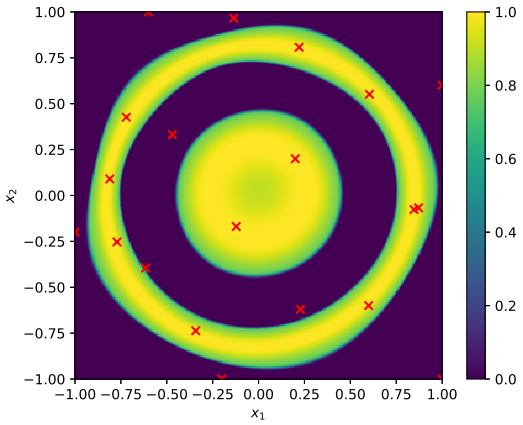
$$\begin{aligned}= \log\left(\sqrt{2\pi e\sigma^2(\mathbf{x})}Z\right) + \frac{1}{2Z}((\alpha - \mu(\mathbf{x}))\mathcal{N}(\mu|\alpha, \sigma^2(\mathbf{x})) \\ - (\beta - \mu(\mathbf{x}))\mathcal{N}(\mu|\beta, \sigma^2(\mathbf{x}))).\end{aligned} \tag{14}$$

Z is a normalization constant equal to $\Phi\left(\frac{\beta-\mu(\mathbf{x})}{\sigma(\mathbf{x})}\right) - \Phi\left(\frac{\alpha-\mu(\mathbf{x})}{\sigma(\mathbf{x})}\right)$, where $\Phi(\cdot)$ is the standard normal cumulative density function. The introduction of this acquisition function is the main contribution of this paper. It has a closed form expression, rather than using many approximations, often used in information-theoretic optimization approaches [9; 10].

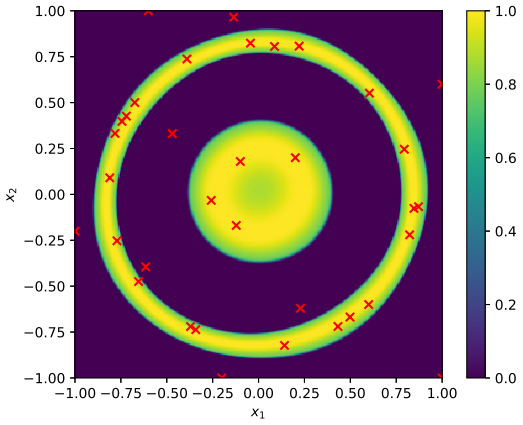
An example is shown in Figure 2 with a simple benchmark function. Initially, with a small amount of samples the acquisition function is smeared out more in space, which enhances exploration. The next evaluated sample is determined by the maximum of the acquisition function. After evaluating more samples, the model has a smaller predictive variance, which means that the acquisition function is more centered along the boundaries of the feasible region.



(a) $N = 10$



(b) $N = 20$



(c) $N = 35$

Fig. 2. Proposed acquisition function with different amount of samples N. The red crosses represent the sampled data. The function under consideration is $f(x_1, x_2) = x_1^2 + x_2^2$, while the upper and lower boundary of the feasible region are 0.7 and 0.04 respectively.

This is logical since intuitively maximizing the information about the (in)feasible regions, means finding the border of the region rather than exploring the region itself [21].

V. Experiments

The experiments start out with a Latin hypercube of size 6 as initial design. An RBF kernel with ARD is used for modeling the problem. The acquisition function and optimization process are implemented using GPflowOpt [22].

The results are compared to a state-of-the-art sampling strategy [21] using surrogate models that samples the inner region of the feasible region, this method will be referred to here as Probability of Feasibility (PoF). All the discussed examples are taken from [23], which uses Active Expansion Sampling (AES). The experiments are replicated 10 times with different random seeds and the median and the 98% confidence interval are shown.

The metric to evaluate the models is the F_1 score, which is used to account for class imbalance:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{15}$$

where

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{16}$$

and

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}. \tag{17}$$

The F_1 score always lies between zero and one, the best score being one. These are evaluated using a test set of 1000 points that is drawn uniformly from the problem domain.

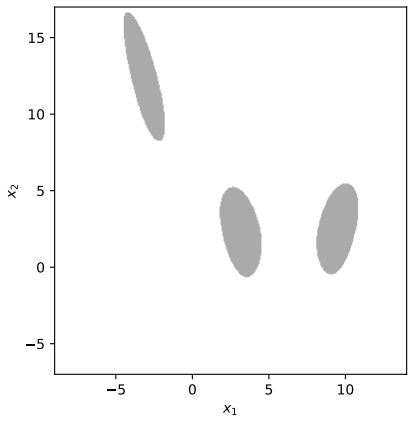
A. Branin

The first test function under consideration is the Branin function, often used in optimization and surrogate modeling benchmarks [13], and has a 2-dimensional input of domain $[-9, 14] \times [-7, 14]$:

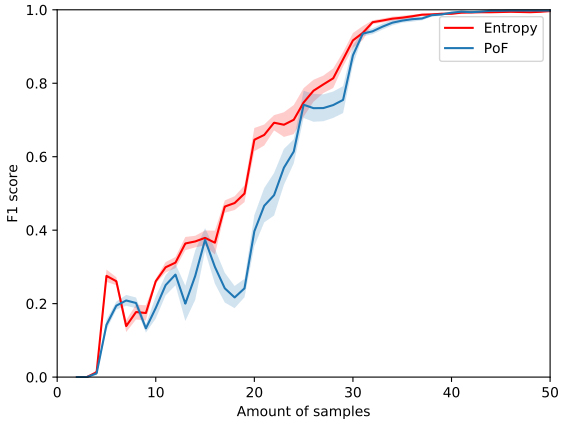
$$f(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 \\ 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10. \tag{18}$$

An upper bound is defined for the feasible region, which is 8 [23], which results in 3 disconnected feasible sub-regions, as seen in Figure 3(a).

The results are shown in Figure 3(b), and show that almost in the entire process, the proposed entropy method is superior. The problem is fairly easy to model and therefore does not show very large improvement.



(a) Feasible region



(b) Convergence Curve

Fig. 3. Branin example: a two-dimensional example where the feasible region consists of 3 disconnected areas of approximately the same size.

B. Hosaki

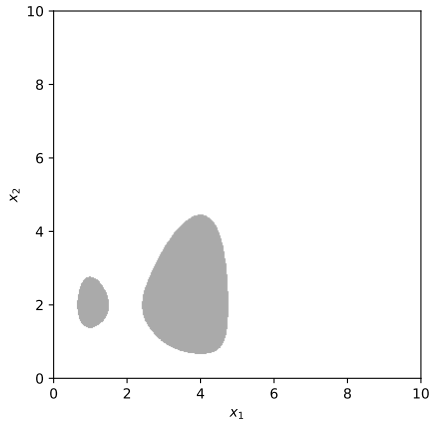
A second example is the Hosaki example [23]. Different to the Branin example the two disconnected feasible regions are of different size here. The domain is two dimensional $[0, 10] \times [0, 10]$ and the upper boundary of the feasible region is -1.

$$f(x_1, x_2) = \left(1 - 8x_1 + 7x_1^2 - \frac{7}{3}x_1^3 + \frac{1}{4}x_1^4\right)x_2^2 e^{-x_2} \tag{19}$$

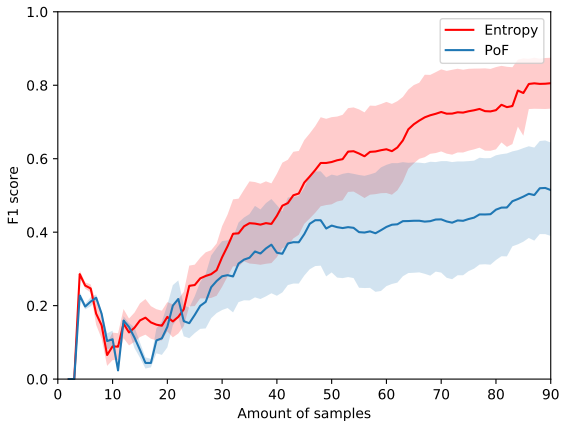
This problem is more complex to model using an RBF kernel and hence the difference is larger, shown Figure 4(b). The entropy method performs considerably better than PoF, since it only samples at the borders.

C. Nowacki Beam

The last example that is discussed is the Nowacki beam example [24], which is a real engineering problem. The problem involves a beam of length l and a load F that is exerted on the end of the beam. The design variables are



(a) Feasible region



(b) Convergence curve

Fig. 4. Hosaki example: a two-dimensional example where the feasible region consists of 2 disconnected areas of different size.

the breadth b and the height h. There are several feasibility constraints in this problem [23]:

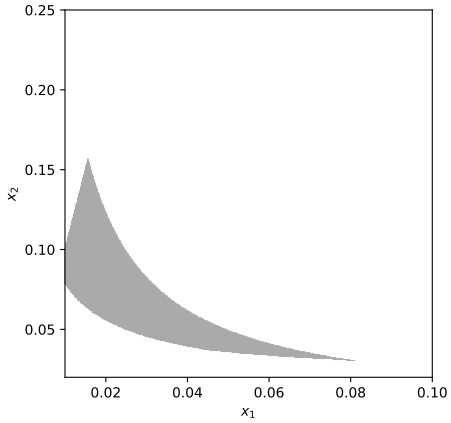
(a) the area must be limited: $bh \leq 0.0025\text{m}^2$,
(b) there is a maximum tip deflection: $\delta = Fl^3/(3EI_Y) \leq 5\text{mm}$,
(c) there is a maximum blending stress: $\sigma_B = 6Fl/(bh^2) \leq \sigma_Y$,
(d) there is a maximum shear stress: $\tau 1.5F/(bh) \leq \sigma_Y/2$,
(e) the load has to be smaller than the critical buckling force: $Ff \leq 4/l^2\sqrt{GI_T EI_Z/(1-\nu^2)}$.

Here, $I_Y = bh^3/12$, $I_Z = b^3h/12$, $I_T = I_Y + I_Z$ and f is a safety factor. And σ_Y,E,ν, and G represent the yield stress, Youngs modulus, Poissons ratio, and shear modulus respectively. Their values are summarized in Table I.

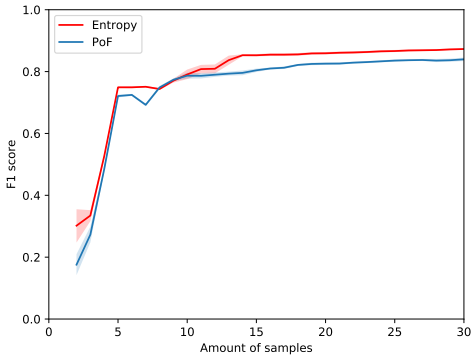
Since there are multiple constraints, each constraint is modeled independently from each-other. The acquisition function that is used is equal to the sum of the acquisition functions corresponding to each constraint individually. More advanced approaches could be taken here, like modeling the constraints jointly [25], resulting in a joint distribution at test time, but are not discussed in this work. This problem has a connected feasible region as shown in Figure 5(a).

TABLE I
VALUES ASSOCIATED WITH THE NOWACKI BEAM PROBLEM

σ_Y	240MPa
E	216.62GPa
ν	0.27
G	86.65GPa
l	0.5m
f	2
F	5kN



(a) Feasible region



(b) Convergence curve

Fig. 5. Nowacki Beam example: a two-dimensional engineering example considering a beam subject to a load. The feasible region is connected.

In Figure 5(b) is shown that the entropy approach performs better than PoF. The $F1$ score seems to saturate at a value of 0.88, like reported in [23]. This source using AES, however, claims that it only saturates when using 250 samples.

VI. CONCLUSION

A novel active learning approach for feasible region discovery was introduced based on information theoretic principles.

It is benchmarked on two synthetic problems and one engineering problem, on which it shows superiority to Probability of Feasibility [21] and Active Expansion Sampling [23]. When using more than one constraint, the constraints are modeled independently, which can be improved in future work.

REFERENCES

[1] D. Gorissen, I. Couckuyt, P. Demeester, T. Dhaene, and K. Crombecq, "A surrogate modeling and adaptive sampling toolbox for computer based design," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2051–2055, 2010.

[2] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Learning and Intelligent Optimization*, C. A. C. Coello, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 507–523.

[3] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[4] A.-C. Zvoianu, G. Bramerdorfer, E. Lughofer, S. Silber, W. Amrhein, and E. Peter Klement, "Hybridization of multi-objective evolutionary algorithms and artificial neural networks for optimizing the performance of electrical drives," *Eng. Appl. Artif. Intell.*, vol. 26, no. 8, pp. 1781–1794, Sep. 2013.

[5] A.-C. Zvoianu, E. Lughofer, W. Koppelsttter, G. Weidenholzer, W. Amrhein, and E. P. Klement, "Performance comparison of generational and steady-state asynchronous multi-objective evolutionary algorithms for computationally-intensive problems," *Knowledge-Based Systems*, vol. 87, pp. 47 – 60, 2015, computational Intelligence Applications for Data Science.

[6] D. Fonseca, D. Navaresse, and G. Moynihan, "Simulation metamodeling through artificial neural networks," *Engineering Applications of Artificial Intelligence*, vol. 16, no. 3, pp. 177–183, 2003.

[7] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. University Press Group Limited, 2006.

[8] D. Gorissen, "Grid-enabled adaptive surrogate modeling for computer aided engineering," Ph.D. Thesis, Dept. of Computer Science, Ghent University, 2010.

[9] P. Hennig and C. J. Schuler, "Entropy search for information-efficient global optimization," *Journal of Machine Learning Research*, vol. 13, no. Jun, pp. 1809–1837, 2012.

[10] D. Hernandez-Lobato, J. Hernandez-Lobato, A. Shah, and R. Adams, "Predictive entropy search for multi-objective bayesian optimization," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1492–1501.

[11] Z. Wang and S. Jegelka, "Max-value entropy search for efficient Bayesian optimization," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 3627–3635.

[12] T. J. Santner, B. J. Williams, and W. I. Notz, *The design and analysis of computer experiments*. Springer Science & Business Media, 2013.

[13] A. Forrester, A. Keane *et al.*, *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2008.

[15] D. C. Montgomery, *Design and analysis of experiments*. John wiley & sons, 2017.

[16] E. R. van Dam, B. Husslage, D. den Hertog, and H. Melissen, "Maximin latin hypercube designs in two dimensions," *Operations Research*, vol. 55, no. 1, pp. 158–169, 2007.

[17] D. J. C. MacKay, "Information-Based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, Jul. 1992.

[18] N. Houlsby, F. Huszar, Z. Ghahramani, and J. M. Hernández-lobato, "Collaborative gaussian processes for preference learning," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 2096–2104.

[19] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 918–926.

[20] B. Ru, M. A. Osborne, M. Mcleod, and D. Granziol, "Fast information-theoretic Bayesian optimisation," in *International Conference on Machine Learning (ICML)*, 2018, pp. 4384–4392.

[21] A. Kaintura, K. P. Foss, I. Couckuyt, T. Dhaene, O. Zografos, A. Vaysset, and B. Soree, "Machine learning for fast characterization of magnetic logic devices," in *2018 IEEE Electrical Design of Advanced Packaging and Systems (EDAPS) Symposium*, 2018, pp. 1–3.

[22] N. Knudde, J. van der Herten, T. Dhaene, and I. Couckuyt, "GPflowOpt: A Bayesian Optimization Library using TensorFlow," *arXiv preprint – arXiv:1711.03845*, 2017.

[23] W. Chen and M. Fuge, "Active expansion sampling for learning feasible domains in an unbounded input space," *Structural and Multidisciplinary Optimization*, vol. 57, no. 3, pp. 925–945, Mar 2018.

[24] H. Nowacki, "Modelling of design decision for cad," in *Computer Aided Design: Modelling, Systems Engineering, CAD-Systems - CREST Advanced Course*. London, UK, UK: Springer-Verlag, 1980, pp. 177–223.

[25] M. A. Álvarez and N. D. Lawrence, "Computationally efficient convolved multiple output gaussian processes," *J. Mach. Learn. Res.*, vol. 12, pp. 1459–1500, Jul. 2011.