Computer Aided Chemical Engineering (Elsevier)

Volume 43, 2018, Pages 943-948

Anton Friedl, Jiří J. Klemeš, Stefan Radl, Petar S. Varbanov, Thomas Wallek (Eds.)
Proceedings of the 28th European Symposium on Computer Aided Process Engineering
June 10th to 13th, 2018,Graz,Austria.© 2018 Elsevier B.V. All rights reserved.

Sequential Bayesian Experimental Design for Process Optimization with Stochastic Binary Outcomes

Martin F. Luna, Ernesto C. Martínez*

INGAR (CONICET-UTN), Avellaneda 3657, Santa Fe S3002 GJC, Argentina

ecmarti@santafe-conicet.gov.ar

Abstract

For innovative products, the issue of reproducibly obtaining their desired end-use properties at industrial scale is the main problem to be addressed and solved in process development. Lacking a reliable first-principles process model, a Bayesian optimization algorithm is proposed. On this basis, a short of sequence of experimental runs for pinpointing operating conditions that maximize the probability of successfully complying with end-use product properties is defined. Bayesian optimization is able to take advantage of the full information provided by the sequence of experiments made using a probabilistic model (Gaussian process) of the probability of success based on a one-class classification method. The metric which is maximized to decide the conditions for the next experiment is designed around the expected improvement for a binary response. The proposed algorithm's performance is demonstrated using simulation data from a fed-batch reactor for emulsion polymerization of styrene.

Keywords: Bayesian optimization, end-use product properties, Gaussian processes, one-class classification, scale-up.

1. Introduction

For innovative products, the issue of reproducibly obtaining their desired end-use properties is the main problem to be addressed in process development (Colombo et al., 2016). For example, emulsion polymerization processes are well-known examples of the importance of guaranteeing reproducibility of end-use properties such as tensile strength and melt index by properly choosing the operating policy (Valappil and Georgakis, 2002). Similar problems are encountered in production of high-quality graphene sheets, single-walled carbon nanotubes and functionalized polymer nanofibers. The main drawback in the development of this type of innovative, high-value products is lacking a reliable first-principles model to predict the binary outcome (success/failure) of a production run for a given setting of the controlled variables in the face of variability because of uncontrolled factors on end-use properties. Also, even at bench scale, experiments are time consuming and/or expensive, which demand fast pinpointing operating conditions where the probability of success is maximum.

The problem of sequential experimental design for process optimization with stochastic binary outcomes is addressed by combining one-class classification with Gaussian processes (Xiao et al., 2015) and Bayesian optimization (Shahriari, et al., 2016). It is assumed that errors (failures) incurred in the sequence of runs are not punished, but instead it is of major concern the final recommendation for operating conditions once the available budget (in time and/or money) for experimental optimization is over.

2. Problem statement

Given an initial Region of Interest (ROI) $\mathbb{X} \subset \mathbb{R}^d$ for the controlled inputs, an unknown objective function $\pi\colon \mathbb{X} \rightarrow [0, 1]$ descriptive of the probability of complying with product end-use properties and a maximum budget of n experiments, the problem of sequentially making decisions $\mathbf{X}_i = [x_1, x_2, \dots, x_i]^T$ which are rewarded by a "success" with probability $\pi(x)$ and "failure" with probability $1-\pi(x)$, is to recommend, after n experiments, the operating conditions x^* that maximizes π. Note that the choice of the operating conditions for each experiment x_i in the sequence is based solely on knowledge of the binary outcomes $\mathbf{y}_i = [y_1, y_2, \dots, y_i]^T$ from previous runs. The observations at x_i are considered to be drawn from a Bernoulli distribution with a success probability $p(y = 1|x_i)$.The probability of success is related to a latent function $f(x)\colon \mathbb{R}^d \rightarrow \mathbb{R}$ that is mapped to a unit interval by a sigmoid transformation. The transformation used is the *probit* function $p(y = 1|x_i) = \Phi\big(f(x_i)\big)$, where Φ denotes the cumulative probability function of the standard Normal density.

As there not exist correct examples of the success probability π over ROI but evaluative feedback from binary outcomes {-1, +1}, using Gaussian processes (GPs) for one-class classification is a more appropriate choice for probabilistic modelling of the objective function being maximized. At the observed inputs, the latent variables $\mathbf{f} = \{f(x_i)\}_{i=1}^n$ have a Gaussian prior distribution. Given a training set D = (**X**, **y**), the probabilistic model chosen $p(y_*|D, x_*)$ aims to predict the target value y_* for a new sample x_* by computing the posterior probability $p = (\mathbf{f}|\mathbf{X}) = N\big(\mathbf{f}\big|\mu, \mathbf{K}_{\mathrm{f,f}}\big)$, where $\mathbf{K}_{\mathrm{f,f}}$ is the covariance matrix and μ is the mean function. Since neither of the class labels is considered more probable, the prior mean is often set to zero. As a GP generate an output z in the range $(-\infty, \infty)$, a monotonically increasing response function $\sigma(z)$is used convert the GP outputs to values within [-1, 1] which can be interpreted as class probabilities (Rasmussen and Williams, 2006). In particular, the latent GP $\hat{f}$ defines a Gaussian probability density function p_f^x for an input $x \in \mathbb{X}$. At any given x, the corresponding probability density for the positive class (success) is defined as p_π^x.

The inference step for conditioning the posterior GP on sampled observations **X** and **y** require computing the following integral to determine the posterior $\hat{f}$ at any x_* over $\mathbb{X}$:

$$p\big(\hat{f}_*\big|\mathbf{x}, \mathbf{y}, x_*\big) = \int p\big(\hat{f}_*\big|\mathbf{x}, \mathbf{y}, x_*\big)\, p(\mathbf{f}_*|\mathbf{x}, \mathbf{y}) d\mathbf{f}_* \qquad (1)$$

In this equation, $\mathbf{f}_*$ represents the GP prior on the latent function at x_*.The main idea is to use a mean of the prior with a smaller value than our positive class labels (i.e., y=1), such as a zero mean. This restricts the space of probable latent functions to those whose values gradually decrease when being far away from observed points. By choosing a smooth covariance function such as the simple squared exponential

$$k\big((\mathbf{x}_r, \mathbf{x}_m|\zeta, \ell)\big) = \zeta^2 \exp\left(-\frac{||\mathbf{x}_r - \mathbf{x}_m||^2}{2\ell^2}\right) \qquad (2)$$

an important subset of latent functions is obtained. The parameter ℓ defines its characteristic length scale whose value must be optimized for improved discriminatory power, and ζ^2 is the magnitude parameter. The GP mean μ_* typically decreases for inputs distant from the training data and can be directly utilized as a measure of membership for the positive class. Conversely, the variance σ_*^2 of the prediction is always increasing for distant inputs, which suggests that the negative variance value can serve as an alternative criterion for discriminating operating conditions for successes from those end-use properties are obtained. As it is shown by Rasmussen and Williams (2006), if σ is the Gaussian cumulative density function, the expected value of the probability of success (posterior) at x_* can be approximated by

$$\mathbb{E}[p_\pi^x] = \bar{\pi}(x) = \Phi\left(\frac{\mu_*}{\sqrt{1+\sigma_*^2}}\right) \tag{3}$$

3. Bayesian optimization

Mathematically speaking, given the problem of finding a global maximum of the unknown objective function $\bar{\pi}$ over $\mathbb{X}$ which is defined based on a priori knowledge

$$x_{best} = \arg max\, \bar{\pi}(x), x\ \in \mathbb{X} \tag{4}$$

The sequential Bayesian experimental design algorithm in Fig. 1 selects, at each iteration *i*, the operating conditions x_{i+1} for the next experiment and observe the binary outcome y_{i+1}. After *n* experiments, the algorithm makes a final recommendation x^* which represents the algorithm's best estimate (based on the available experimental budget) of the operating conditions for which the probability of success is the global maximum.

The Bayesian optimization algorithm (see Fig. 1 for details) resorts to a selection metric (often known as *acquisition function*) that allows selecting the next experiment to be made using a trade-off between exploration and exploitation. In this work, the expected improvement for stochastic binary outcomes proposed by Tesch et al. (2013) is used. By querying the GP posterior at each point in **X**, and letting

$$\tilde{\pi}_{max} = max\ \ \bar{\pi}(x), x \in \mathbf{X} \tag{5}$$

The improvement I_π for stochastic binary outcomes at any $x \in\ \mathbb{X}$ is defined as follows

$$I_\pi\big(\pi(x)\big) = max\ (\pi(x) - \tilde{\pi}_{max}, 0), x \in\ \mathbb{X} \tag{6}$$

The corresponding expectation for I_π over $\mathbb{X}$ is

$$\mathbb{E}I_\pi(\tilde{\pi}_{max}) = max\ \int_{\sigma^{-1}(\tilde{\pi}_{max})}^{\infty} (\sigma(z) - \tilde{\pi}_{max})\, p_f^x(z)\, dz \tag{7}$$

4. Case study

4.1. Process description

The proposed methodology is tested by simulation of the emulsion polymerization of styrene. The reactor operates in fed-batch mode, with two inlets: the monomer feed (*x*$_1$) and the chain transfer agent (CTA) feed (*x*$_2$), both measured in mol/s. The reactor is initially charged with solvent and an initiator, and is seeded with particles of three different sizes. As the monomer is fed, polymer chains growth unevenly, giving rise to a distribution of chain lengths (and molecular weights). The CTA modifies the length of the chains. The end-use properties of the product depend on the distribution of molecular weights, both in weight (*MW_w*) and in number (*MW_n*).

```
Algorithm: Bayesian optimization
• Inputs: $n_0$, $n$, $D_0 = \{\mathbf{X}_0, \mathbf{y}_0\}$
▷ For $i = n_0 + 1$ to $n$ do
    • Select new $x_i$ by optimizing the expected improvement $\mathbb{E}I_\pi(x)$
          $x_{i+1} = \arg max\ \mathbb{E}I_\pi(x, D_i), x \in \mathbb{X}$
    • Do the next experiment at $x_{i+1}$ and observe $y_{i+1}$
    • Augment dataset $D_{i+1} = \{D_i, (x_{i+1}, y_{i+1})\}$
    • Update statistical model $\hat{f}$
▷ End for
• $x^* = \arg max\ \bar{\pi}(x), x \in \mathbf{X}_n$
• Output: $x^*$
```

Figure 1. Bayesian optimization algorithm for stochastic binary outcomes.

In this work, the melt flow index (*MI*) and the tensile strength (*TS*) are the end-use properties of interest. A batch is considered successful if both properties are kept within their desired intervals:

$$1.25 \times 10^{-4} < MI \leq 7.5 \times 10^{-4}\ [g/min];\ 6900 < TS \leq 7200\ [psi] \tag{8}$$

These end-use properties are correlated with the molecular weights as follows:

$$MI = \frac{30}{\left(MW_w^{3.4} \times 10^{-18} - 0.2\right)} \tag{9}$$

$$TS = 7390 - 4.51 \times 10^{8}\left(\frac{1}{MW_n}\right) \tag{10}$$

The variability due to uncontrolled factors is introduced in the simulation as a 5% perturbation (normally distributed) in the initial charge of seeded particles and the initiator concentration. Details of the stochastic simulation model for the polymerization process can be found in Colombo et al. (2016). The contour lines for the probability of success in the ROI are shown in Fig. 2.

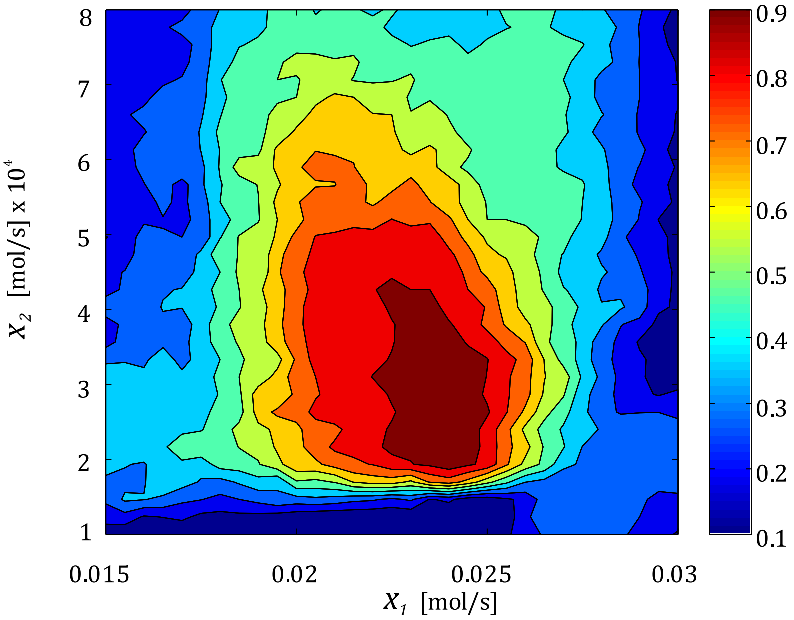

Figure 2. Contour plots for the probability of success in the case study.

4.2. Results

The proposed algorithm is tested for the case study described in Section 4.1 using the stochastic simulation model. In each trial, 50 experiments are performed before the method selects the optimal operation condition x^* that maximizes π. Experiments are divided between an initial sampling set (with n_0 points) and experiments designed based on Bayesian optimization. Thus, the algorithm in Fig. 1 is tested using 100 independent sequences generated from different number and selection of the initial experiments. The initial n_0 points where chosen using Latin hypercube sampling. A simple squared exponential covariance with fixed hyper-parameters (length scale of $e^{0.75}$ and signal variance of e^5) was used. Results obtained are shown in Table I.

Fig. 3 depicts the GP approximation $\bar{\pi}(x_1, x_2)$ of the probability of success, after 50 experiments in one the independent trials made, obtained using the proposed algorithm in Fig. 1. It is worth noting that for the process simulation model the (assumed unknown) maximum probability of success is 0.97.

5. Concluding remarks

The role of Bayesian optimization in sequentially making decisions regarding operating conditions aiming at maximizing the probability of success in achieving the desired end-use properties has been discussed. The proposed algorithm is based on expected improvement for stochastic binary outcomes and one-class classification using Gaussian processes. Simulations results are promising bearing in mind the level of variability considered and that Bayesian optimization does not require a first-principles model.

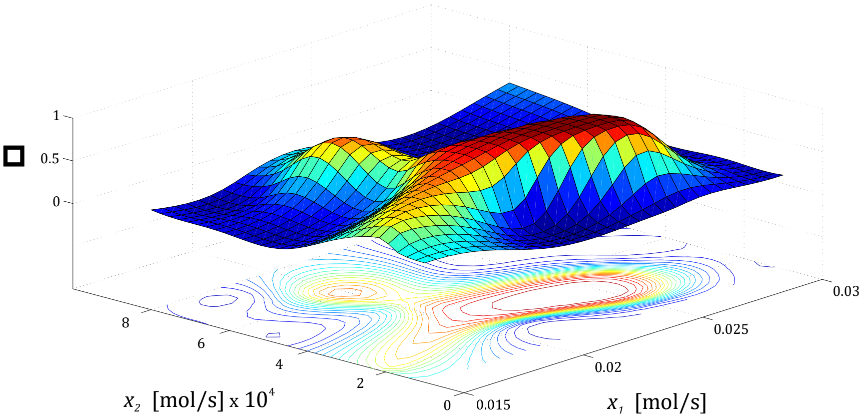

Figure 3. Response surface and contour plots for the estimated $\bar{\pi}(x_1, x_2)$ in a trial.

Table I. Bayesian optimization based on 100 independent trials of 50 runs.

n_0	$\bar{x}_1^*$ [mol/s]	$\bar{x}_2^*$ [mol/s] x 10^4	$\bar{\pi}$	$\sigma^2(\pi)$	$\pi(\bar{x}_1^*, \bar{x}_2^*)$
5	0,0229	3,679	0,860	0,015	0,927
10	0,0229	3,522	0,879	0,013	0,939
15	0,0230	3,602	0,867	0,018	0,930
20	0,0228	3,522	0,843	0,020	0,915
25	0,0227	3,630	0,840	0,024	0,915
30	0,0225	3,695	0,801	0,029	0,881
45	0,0229	3,840	0,812	0,037	0,916

References

Colombo, E., M. F. Luna, M., E. C. Martínez, 2016, Probability-Based Design of Experiments for Batch Process Optimization with End-Point Specifications, Ind. Eng. Chem. Res., 55, 1254−1265.

C. E. Rasmussen , C. K. Williams, 2006, Gaussian Processes for Machine Learning, MIT Press.

Shahriari, B. et al., 2016, Taking the Human Out of the loop: A Review of Bayesian Optimization, Proceedings of the IEEE, 104, 148-175.

M. Tesch. J. Schneider, H. Choset, 2013, Expensive Function Optimization with Stochastic Binary Outcomes, Proceedings of the 30th International Conference on Machine Learning.

J. Valappil, C. Georgakis, 2002, Nonlinear Model Predictive Control of End-Use Properties in Batch Reactors, AIChE J., 48, 9, 2006-2021.

Y. Xiao, H. Wang, W. Xu, 2015, Hyperparameter Selection for Gaussian Process One-Class Classification, IEEE Trans. on Neural Networks and Learning Systems, 26, 9, 2182-2187.