Sauro Pierucci, Flavio Manenti, Giulia Bozzano, Davide Manca (Eds.)
Proceedings of the 30th European Symposium on Computer Aided Process Engineering
(ESCAPE30), May 24-27, 2020, Milano, Italy. © 2020 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/B978-0-12-823377-1.50327-X

Nested Sampling Strategy for Bayesian Design Space Characterization

Kennedy P. Kusumo[a], Lucian Gomoescu[a,b,], Radoslav Paulen[c], Salvador García-Muñoz[d], Costas C. Pantelides[a,b], Nilay Shah[a], Benoit Chachuat[a,*]

[a] *Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, UK*
[b] *Process Systems Enterprise Ltd, London, UK*
[c] *Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Slovakia*
[d] *Small Molecule Design and Development, Eli Lilly & Company, Indianapolis, USA*
b. chachuat@imperial.ac.uk

Abstract

Design space is a key concept in pharmaceutical quality by design, providing better understanding of manufacturing processes and enhancing regulatory flexibility. It is of paramount importance to develop computational techniques for providing quantitative representations of a design space, in accordance with the ICH Q8 guideline. The focus is on Bayesian approaches to design space characterization, which rely on a process model to determine a feasibility probability that is used for measuring reliability and risk. The paper presents three improvements over an existing nested sampling method: two-phase strategy with the first phase using a cheap sorting function based on nominal model parameters; dynamic sampling strategy to refine the target design space; and vectorization to evaluate costly functions in parallel. These improvements are implemented as part of the python package DEUS and demonstrated on an industrial case study.

Keywords: pharmaceutical processes, quality-by-design, design space, nested sampling

1. Introduction

The quality-by-design (QbD) initiative, through the ICH Q8 guideline (Reklaitis et al., 2017), introduced the concept of design space (DS) to improve regulatory flexibility of processes in the pharmaceutical industry. Given a set of critical quality attributes (CQA), the DS represents a set of critical process parameters (CPP) that result in on-spec pharmaceutical production. Peterson (2008) defined the probabilistic DS in terms of feasibility probabilities, a concept akin to stochastic flexibility (Straub and Grossmann, 1990). As the use of mathematical models to support DS characterization is becoming more common in industrial practice (García-Muñoz et al., 2015), the uncertainty related to model parameters and structure needs to be considered and efficient computational tools are of great interest.

Existing computational approaches to probabilistic DS characterization differ in how they account for process model uncertainty and how they approximate the DS itself. Process model uncertainty may be represented as a sampled distribution (e.g. a joint posterior from Bayesian estimation) or a joint confidence region (e.g. a frequentist confidence ellipsoid). Sampling methods seek to produce a set of CPP values that belong to the DS at a desired reliability level, whereas optimization-based methods seek to inscribe a simple shape (e.g. box or ellipsoid) within the DS. For instance, Laky et al. (2019) proposed two optimization-based strategies akin to the classic feasibility test and index formulations that exploit confidence ellipsoids for the model parameters. Monte Carlo and Bayesian techniques have also been used to propagate the uncertainty to the CQAs and estimate a feasibility probability (Peterson et al., 2017; Bano et al., 2018). These techniques have proven effective in practice, but they are computationally expensive and mainly tractable for low-dimensional DS at present.

Recently, Kusumo et al. (2019) presented a sampling strategy based on an adaptation of the nested sampling (NS) algorithm (Skilling, 2004). The algorithm maintains a given set of live points through regions with increasing probability feasibility until reaching a desired reliability level. It leverages efficient strategies from Bayesian statistics for generating replacement proposals during the search and is applicable to problems with disjoint DS or black-box models. This paper presents three ideas to further improve the computational performance of nested sampling for DS characterization. These improvements are demonstrated on a comparative study of the Suzuki coupling reaction.

2. Background

Consider a manufacturing process for a pharmaceutical product that has its quality defined by some CQAs, denoted by $\boldsymbol{s} \in \mathbb{R}^{n_s}$. Assume that a mathematical model of the process (either knowledge- or data-driven) is available that predicts the CQAs corresponding to the CPPs, denoted by $\boldsymbol{d} \in \mathrm{K}$ within the knowledge space $\mathrm{K} \subset \mathbb{R}^{n_d}$:

$$\boldsymbol{s} = \boldsymbol{f}(\boldsymbol{d}, \boldsymbol{\theta}) \tag{1}$$

The model parameters, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ represent uncertain quantities e.g., physical constants, coefficients in a regression model, or disturbances that affect the CQAs. Feasibility of the process is defined by the CQA limits alongside other process constraints:

$$\boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\theta}) := \boldsymbol{g}(\boldsymbol{d}, \boldsymbol{f}(\boldsymbol{d}, \boldsymbol{\theta})) \leq \boldsymbol{0} \tag{2}$$

The mappings $\boldsymbol{f}$ and $\boldsymbol{G}$ need not be given in closed-form but could be implicitly defined via a DAE model or a CFD simulation. Given a set of nominal model parameters $\boldsymbol{\theta}_{\mathrm{nom}}$, the nominal DS is defined as:

$$\mathrm{D}_{\mathrm{nom}} := \left\{ \boldsymbol{d} \in \mathrm{K} : \boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\theta}_{\mathrm{nom}}) \leq \boldsymbol{0} \right\}. \tag{3}$$

However, the value of $\boldsymbol{\theta}$ is inherently uncertain by nature of the modelling exercise. A Bayesian framework considers $\boldsymbol{\theta}$ as random variables with a joint distribution $p(\boldsymbol{\theta})$ that describes the belief on the value of $\boldsymbol{\theta}$. In this framework the model is used to predict the probability that the manufacturing process is feasible for a given $\boldsymbol{d} \in \mathrm{K}$:

$$\mathrm{P}\left[\boldsymbol{G}(\boldsymbol{d}, \cdot) \leq \boldsymbol{0} \mid p(\boldsymbol{\theta}) \right] := \int_{\{\boldsymbol{\theta} : \boldsymbol{G}(\boldsymbol{d}, \cdot) \leq \boldsymbol{0}\}} p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \tag{4}$$

This paper focuses on characterizing the probabilistic DS given by:

$$\mathrm{D}_{\alpha} := \left\{ d \in \mathrm{K} : \mathrm{P}\left[\boldsymbol{G}(\boldsymbol{d}, \cdot) \leq \boldsymbol{0} \mid p(\boldsymbol{\theta}) \right] \geq \alpha \right\} \tag{5}$$

where $0 < \alpha \leq 1$ is the reliability value.

3. Improved nested sampling for design space characterization

The NS algorithm for DS characterization (Algorithm 1) starts with N_L *live points* $\boldsymbol{d}_i \in \mathrm{K}$, sampled uniformly within the knowledge space. These live points are sorted according to their estimated feasibility probabilities, evaluated as:

$$\hat{\mathrm{P}}\left[\boldsymbol{G}(\boldsymbol{d}_i,\boldsymbol{\theta}_j) \le \mathbf{0} \mid \mathrm{S}_\theta\right] := \sum_{(\boldsymbol{\theta}_j, w_j)\in \mathrm{S}_\theta} \mathbf{1}\left(G\left(d_i,\theta_j\right)\right) w_j, \quad \text{with } \mathbf{1}(\cdot) := \begin{cases} 1, & \text{if } g_k \le 0,\ \forall k \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where S_θ is a set of model parameter scenarios θ_j and its weight w_j , sampled from $p(\boldsymbol{\theta})$. The feasibility probability of a point $\boldsymbol{d}_\mathrm{i} \in \mathrm{K}$ is denoted by P_i below for brevity. Each iteration generates N_P *proposal points* $\boldsymbol{d}_\mathrm{k}$, for instance by sampling within an enlarged ellipsoid enclosing the current live points (Mukherjee et al., 2006). Following the same order as per their generation, each proposal $\boldsymbol{d}_\mathrm{k}$ will replace $\boldsymbol{d}_\mathrm{min}$ – the live point with the lowest P – when $\mathrm{P}_\mathrm{min} < \mathrm{P}_\mathrm{k}$. The replaced point $\boldsymbol{d}_\mathrm{min}$ – called a *dead point* – is recorded alongside its feasibility probability, while the point $\boldsymbol{d}_\mathrm{min}$ and its feasibility probability P_min are updated. A stop criterion is checked after each iteration, for instance testing if all live points belong to design space with target reliability value α^*. Three improvement strategies over this basic algorithm are described next:

Strategy 1 – Two-phase nested sampling. A first phase, called *nominal*, is added to Algorithm 1 (lines 5-12), whereby the feasibility is only evaluated at $\boldsymbol{\theta}_\mathrm{nom}$ - a much cheaper test than evaluating N_θ scenarios to estimate P (Eq. 6). This phase continues until all live points are in D_nom (Eq. 3). Though it is possible that D_nom may exclude parts of D_{α^*} , we observe that $\mathrm{D}_{\alpha^*} \subset \mathrm{D}_\mathrm{nom}$ when the target reliability α^* is close to unity and $\boldsymbol{\theta}_\mathrm{nom}$ is chosen as the maximum likelihood estimate or the mode or mean of the model parameter's posterior distribution. The second phase (lines 17-38), called *probabilistic*, is initialized with the live points from the first phase, after computing and sorting their feasibility probabilities (lines 13-16).

Strategy 2 – Dynamic live point population. The term *dynamic* is in reference to the strategy of increasing the number of live points N_L and proposals N_P over the course of the algorithm in order to generate a denser sample at the target reliability level (lines 27-34). N_L is increased every time the feasibility probability of the current nest, P_min gets larger than a predefined threshold, according to a user-specified top-up schedule S_T . The number of proposals N_P may also be adjusted when N_L is increased.

Strategy 3 – Vectorized function evaluations. Evaluations of the feasibility probability of the live points, replacement proposals, and top-up proposals are carried out in parallel using Python's multiprocessing to exploit multiple processors in modern computers.

An implementation of these algorithmic improvements is available in the Python package DEUS, which can be obtained from: https://github.com/omega-icl/deus. The input file for the case study below can also be retrieved from this link. In DEUS the candidate points are generated by sampling in a single ellipsoid enclosing the current live points (Mukherjee et al., 2006). In addition to setting the numbers of live points, replacement proposals, and the top-up schedule, other tuning parameters include the initial enlargement factor of the ellipsoid (default: 0.1), and the shrinking rate of that enlargement factor at each iteration (default: 0.3).

Algorithm 1 Nested sampling tailored to design space characterization

```
Inputs: $S_\theta = \{(\boldsymbol{\theta}_j, w_j) \sim p(\boldsymbol{\theta}) : j = 1,\ldots,N_\theta\}$, $S_T = \{(P_t^+, N_{L,t}^+, N_{P,t}^+) : t = 1,\ldots,N_T\}$,
    $K$, $\alpha^*$, $\boldsymbol{\theta}_{nom}$, $N_L$, and $N_P$.
Initialization
$S_L \leftarrow \{\boldsymbol{d}_i \in K : i = 1,\ldots,N_L\}$
$DS \leftarrow \varnothing$
$t \leftarrow 0$
Phase I: Nominal
while $\exists \boldsymbol{d}_i \in S_L : \boldsymbol{G}(\boldsymbol{d}_i, \boldsymbol{\theta}_{nom})$ ’ $\boldsymbol{0}$ do
    $S_P \leftarrow \{d_k^+ \in K : k = 1,\ldots,N_P\}$
    for all $\boldsymbol{d}_k^+ \in S_P$ do
        if $\boldsymbol{G}(\boldsymbol{d}_k^+, \boldsymbol{\theta}_{nom}) \le \boldsymbol{0}$ then
            $S_L \leftarrow S_L \cup \{\boldsymbol{d}_k^+\} \setminus \{\boldsymbol{d}_i\}$
        end if
    end for
end while
Reinitialization
for all $\boldsymbol{d}_i \in S_L$ do
    $DS \leftarrow DS \cup \{(\boldsymbol{d}_i ; P[\boldsymbol{G}(\boldsymbol{d}_i, \cdot) \le \boldsymbol{0} \mid S_\theta])\}$
end for
Phase II: Probabilistic
while $\exists \boldsymbol{d}_i \in S_L : P\lfloor \boldsymbol{G}(\boldsymbol{d}_i, \cdot) \le \boldsymbol{0} \mid S_\theta \rfloor \le \alpha^*$ do
    $S_P \leftarrow \{\boldsymbol{d}_k^+ \in K : k = 1,\ldots,N_P\}$                ➢ propose replacements
    for all $\boldsymbol{d}_k^+ \in S_P$ do                                        ➢ decide accepted replacements
        $P_{min} \leftarrow \min\{P[\boldsymbol{G}(\boldsymbol{d}_i, \cdot) \le \boldsymbol{0} \mid S_\theta] : \boldsymbol{d}_i \in S_L\}$
        $\boldsymbol{d}_{min} \leftarrow \arg\min\{P[\boldsymbol{G}(\boldsymbol{d}_i, \cdot) \le \boldsymbol{0} \mid S_\theta] : \boldsymbol{d}_i \in S_L\}$
        if $P\lfloor \boldsymbol{G}(\boldsymbol{d}_k^+, \cdot) \le \boldsymbol{0} \mid S_\theta \rfloor > P_{min}$ then
            $S_L \leftarrow S_L \cup \{\boldsymbol{d}_k^+\} \setminus \{\boldsymbol{d}_{min}\}$
            $DS \leftarrow DS \cup \{(\boldsymbol{d}_{min}, P_{min})\}$
        end if
    end for
    if $P_{min} \ge P_{t+1}^+$ then                                              ➢ top-up live points if needed
        $t \leftarrow t + 1$
        while $N_L < N_{L,t}^+$ do
            $N_L \leftarrow N_L + 1$
            $S_L \leftarrow S_L \cup \{\boldsymbol{d}_{N_L} \in K : P[\boldsymbol{G}(\boldsymbol{d}_{N_L}, \cdot) \le \boldsymbol{0} \mid S_\theta] > P_{min}\}$    ➢ top-up proposal
        end while
        $N_P \leftarrow N_{P,t}^+$
    end if
end while
for all $\boldsymbol{d}_i \in S_L$ do                                              ➢ add current live points
    $DS \leftarrow DS \cup \{(\boldsymbol{d}_i, P[\boldsymbol{G}(\boldsymbol{d}_i, \cdot) \le \boldsymbol{0} \mid S_\theta])\}$
end for
return DS
```

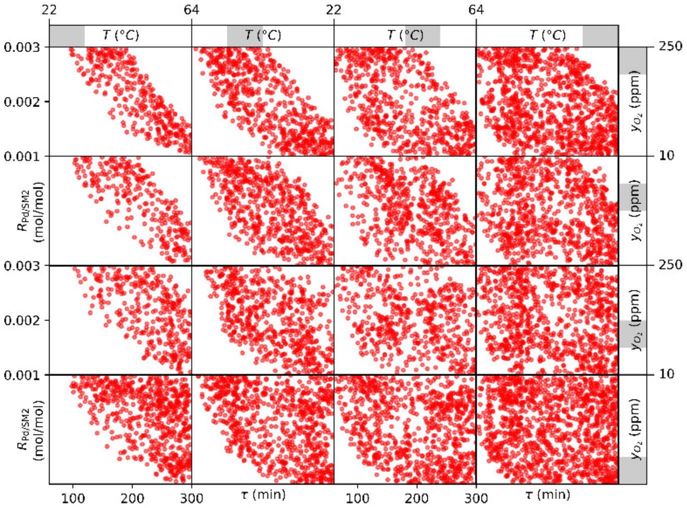

Figure 1: Probabilistic DS at a reliability $\alpha^* = 0.85$ for the Suzuki coupling reaction computed using Algorithm 1 with 1,000 uncertainty scenarios and 10,000 live points.

4. Case study: Suzuki coupling reaction

This case study investigates the Suzuki coupling reaction between a boronic ester (SM1) and an organohalide (SM2) to produce a desired pharmaceutical intermediate (P1) and a dimeric impurity (Imp1) related to SM1. The reaction is biphasic and conducted in batch mode. The gaseous phase is inert with traces of O_2 in vapor-liquid equilibrium with a liquid phase containing 17 chemical species dissolved in a mixture of water and THF. The active species participate in 12 reactions, 3 of which are reversible and 1 is considered instantaneous. The uncertain parameters are the 14 pre-exponential factors of the reactions. The DS is the set of (i) batch durations $\tau \in [75,\ 300]$ (min), (ii) catalyst equivalents $R_{\mathrm{Pd|SM2}} \in [0.001,\ 0.003]$ (mol/mol), (iii) temperatures $T \in [22,\ 64]$ (°C), and (iv) O_2 molar fractions in the head space $y_{O_2} \in [10,\ 250]$ (ppm) such that the product has (i) unreacted SM2 less than 0.001 mol/mol, and (ii) Imp1 less than 0.0015 mol/mol. Full details about the case study can be found in Kusumo et al. (2019).

For comparison with Kusumo et al. (2019) we compute the probabilistic DS at a reliability $\alpha^* = 0.85$ using Algorithm 1 with and without parallelization, considering two cases: (i) $N_\theta = 200$ and $N_L = 5{,}000$; (ii) $N_\theta = 1{,}000$ and $N_L = 10{,}000$. Case (i) was initialized with $N_L = N_P = 100$ and schedule $S_T = \{(0.01,\ 1000,\ 200),\ (0.20,\ 2000, 200),\ (0.85,\ 5000,\ 500)\}$; case (ii) uses the same with double N_L and N_P. We opted for larger N_P to more efficiently parallelize proposal computations with Python's multiprocessing package. Computational statistics are presented in Table 1. Applying Strategies 1+2 decreases the number of evaluations by nearly 20% in both cases and saves 16 hours of computational time for case 2, but only minor savings for case 1. Parallelization results in approximately 4 to 5-fold reduction in CPU time. The returned 4-dimensional samples for case 2 are visualized with a trellis chart in Figure 1, where only the target DS at reliability $\alpha^* = 0.85$ is shown since the improvement strategies are designed to describe that set quicker. The comparison with Kusumo et al. (2019) confirms that no part of the target DS is missed.

Table 1: Computational statistics for Suzuki coupling reaction using Algorithm 1.

		Original		With strategies 1–2		
N_L	N_θ	# eval.[1]	Serial (hr.)[2]	# eval.[1]	Serial (hr.)	Parallel (hr.)[2]
5,000	200	1.45	9.6	1.18	9.3	2.3
10,000	1,000	14.54	112.4	11.80	96.5	21.8

[1] Number of model evaluations (in million).

[2] CPU times (in hours) obtained on AMD Ryzen 5 2600X processor with 6 cores.

5. Conclusions

To address the need for efficient tools for probabilistic DS characterization, Kusumo et al. (2019) proposed a tailored nested-sampling algorithm. We have presented three improvement strategies to the algorithm, namely a two-phase strategy to exploit information from nominal model parameters, a dynamic sampling strategy to delineate the target design space faster, and a vectorization strategy to evaluate costly functions in parallel. These improvements were demonstrated on an industrial case study, leading to a four-fold reduction in CPU time on a typical desktop computer.

Acknowledgements: This work is supported by Eli Lilly and Company through the Pharmaceutical Systems Engineering Lab (PharmaSEL) program. Radoslav Paulen gratefully acknowledges the contribution of the European Commission under grant 790017 (GuEst) and the contribution of the Slovak Research and Development Agency (project APVV 15-0007). The authors are grateful to Carla Vanesa Luciani for her assistance with the case study.

References

G. Bano, P. Facco, F. Bezzo, M. Barolo, 2018. Probabilistic design space determination in pharmaceutical product development: A Bayesian/latent variable approach. AIChE Journal 64 (7), 2438–2449.

S. García-Muñoz, C. V. Luciani, S. Vaidyaraman, K. D. Seibert, 2015. Definition of design spaces using mechanistic models and geometric projections of probability maps. Organic Process Research & Development 19 (8), 1012–1023.

K. P. Kusumo, L. Gomoescu, R. Paulen, S. García-Muñoz, C. C. Pantelides, N. Shah, B. Chachuat, 2019. Bayesian approach to probabilistic design space characterization: A nested sampling strategy. Industrial & Engineering Chemistry Research (in press, doi: 10.1021/acs.iecr.9b05006).

D. Laky, S. Xu, J. S. Rodriguez, S. Vaidyaraman, S. García Muñoz, C. Laird, 2019. An optimization-based framework to define the probabilistic design space of pharmaceutical processes with model uncertainty. Processes 7 (2), 96.

P. Mukherjee, D. Parkinson, A. R. Liddle, 2006. A nested sampling algorithm for cosmological model selection. The Astrophysical Journal Letters 638 (2), L51.

J. J. Peterson, 2008. A Bayesian approach to the ICH Q8 definition of design space. Journal of Biopharmaceutical Statistics 18 (5), 959–975.

J. J. Peterson, M. Yahyah, K. Lief, N. Hodnett, 2017. Predictive distributions for constructing the ICH Q8 design space. In: G. V. Reklaitis, C. Seymour, S. García-Muñoz (Eds.), Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture. Wiley & Sons, Ch. 4, pp. 55–70.

G. V. Reklaitis, C. Seymour, S. García-Muñoz (Eds.), 2017. Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture. Wiley & Sons.

J. Skilling, 2004. Nested sampling. AIP Conference Proceedings 735 (1), 395–405.

D. A. Straub, I. E. Grossmann, 1990. Integrated stochastic metric of flexibility for systems with discrete state and continuous parameter uncertainties. Computers & Chemical Engineering 14 (9), 967–985.