Optimal experimental design for process optimization with stochastic binary outcomes

Estefanía Colombo, Martín Luna, and Ernesto Martínez

INGAR (CONICET-UTN), Avellaneda 3657, Santa Fe S3002 GJC, Argentina.
{estefania.colombo, martinluna, ecmarti}@santafe-conicet.gob.ar

Abstract. Effective control of the end-use properties in order to guarantee product quality is of paramount importance. This is especially valuable for products such as medicament, polymers and nanomaterials. There are plenty of innovative processes whose principles of operation are unknown and for which is not possible to develop reliable models due to time and cost. Despite this, it is important to know the optimum operating zone where the process should work to ensure that the end product meets the required properties. To address this issue, a run-to-run optimization approach is proposed to find the reduced region of operation which guarantees obtaining end-use properties with high probability when a first-principles model for the batch process is not available *a priori*. In order to evaluate the effectiveness of the methodology to find optimal policies for runs involving stochastic binary outcomes, the well-known example of the emulsion polymerization of styrene has been addressed. Results obtained demonstrated that the proposed method is a powerful tool for optimal design of experiments aiming to guarantee end-use product properties.

Keywords: optimal experimental design, optimization, end-use properties.

1 Introduction and problem statement

Most real-world optimization problems for batch processes are related to consistently complying with end-point specifications so as to minimize the variability in product quality and process performance. For example, in many practical applications including foods, cosmetics, pharmaceuticals, etc., lipids are emulsified in an aqueous phase such that their end-use properties (rheological behavior, stability, color, etc.) completely define the values for an end-user of resulting products[1]. Emulsion polymerization processes are also representative examples of the importance of guaranteeing reproducibility and tight control of end-use properties such as tensile strength and melt index by properly choosing the operating policy[2].

Proper setting of key process variables are critical for consistently complying with end-point specifications, including end-use product properties[3–5]. Ideally, a performance model for end-point conditions would be first established, and then the best settings can be searched off-line via an optimization method. However, the performance in a run is only observed through a stochastic binary outcome –success or failure of the experiment– which is dependent of both the values for *controllable* inputs and a number of uncontrollable (and possibly unknown) factors. As a result, the performance model is necessarily of a probabilistic nature and correlates the probability of success of an experiment with the alternative settings for controllable input factors. For this, optimal design of experiments is an important tool because of the increasing need to reduce the resource requirement for achieving end-use properties[6].

Fiordalis and Georgakis[7] have proposed a data-driven experimental design of dynamic experiments as a means of developing a response surface model that can be used to optimize complex batch processes for which it is difficult to develop a knowledge-driven model. Unfortunately, in their approach, the probability of complying with end-point specifications is not explicitly modeled. To achieve this goal it is necessary to develop a design of experiments that includes a compound criterion accounting for the dual purpose of obtaining efficient estimation parameters and furthermore maximizing the probability of a particular event. McGree et al.[8] have proposed different approaches for combining parameter estimation with opposing design criteria for nonlinear models, and particularly McGree and Eccleston[9] have developed a probability-based optimal design which achieves both objectives by simultaneously optimizing a design with respect to the D-optimality criterion as well as maximizing a function of the probability of observing an outcome.

The search for a reduced region of operating conditions having a high probability of success can be formally stated as follows: given an input (parameter) space $\mathfrak{X} \in \Re$ and an unknown function $\pi\colon \mathfrak{X} \to [0,1]$ which represents the binomial probability of success of an experiment, a short sequence of experiments should be generated such that an operating policy x^* having a high probability of success is found after a small number of runs. The ultimate goal of the optimization procedure is to recommend, after a rather small number of experiments, an operating policy x_{r} which minimizes the (typically unknown) error, or simple regret, $\max_{\mathrm{x}\in\mathfrak{X}} \pi(x) - \pi(x_{\mathrm{r}})$, which is equivalent to $\max_{x_{\mathrm{r}}\in\mathfrak{X}} \pi(x_{\mathrm{r}})$. To this aim, the sequence of runs should be generated so as to exploit what is already known to maximize the probability of success, and simultaneously, to explore the input space for efficient parameter estimation in modeling the response surface for $\pi(x)$.

2 Probability-Based Optimal Design

Data-driven modeling of the probability of success $\pi(x)$ typically resorts to Generalized Linear Models (GLMs)[9, 10]. GLMs use a response function σ to convert a linear model with a range of $(-\infty, +\infty)$ to an output that lies within $[0; 1]$ (i.e., a valid probability). Therefore, given a generalized regression model $\eta(x; \theta)$, the predicted success probability $\pi(x_i)$ at x_i is $\sigma(\eta(x; \theta))$. The choice of θ for the latent linear regression model is typically accomplished via maximizing the likelihood of the data given the model.

For a given experimental design ξ consisting of the vectors of covariates $x_i \in \mathrm{X}, i = 1, \dots, n$ for which the corresponding observations (responses) y_i have been obtained, GLMs are defined by three components: a distribution of the response, a linear predictor and a link function $g(\bullet)$ that relates the mean of the response to the linear predictor. If the binary response variable has a Bernoulli distribution with success probability $\pi(x_i) = \sigma(y_i) = g^{-1}(y_i)$, a convenient link function is the logistic function[10]:

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \qquad (1)$$

If the logistic link function is used, gives rise to $\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i\theta$, where π_i is the expectation of y_i, namely the probability of success for x_i.

The aim of an optimum experimental design is to find a design ξ^* that maximizes a particular optimality criterion of interest. In this work, the goal is to generate a short sequence of runs whose outcomes for different operating policies help achieving the dual goal of providing information about the model parameters θ and increasingly maximizing the probability of success π. To this aim, McGree and Eccleston[9] have proposed a compound criterion for experimental design which balances two opposing criteria: D-optimality, which provides efficient estimates of the model parameters and P-optimality, which maximizes the probability of observing a success.

D-optimal designs provide efficient estimates of the model parameters, as they minimize the volume of the ellipsoidal confidence regions around an estimate. For a Bernoulli data with a logistic link function, the expected Fisher information matrix can be expressed as[9]:

$$M(\theta, \xi) = X^T W X \qquad (2)$$

where the matrix W is $\mathrm{diag}\big(\pi_1(1-\pi_1), \dots, \pi_n(1-\pi_n)\big)$. The D-efficiency of any design ξ is defined respect to D-optimal design ξ_D^* like:

$$D_{eff} = \left(\frac{|M(\theta, \xi)|}{|M(\theta, \xi_D^*)|}\right)^{1/q} \qquad (3)$$

where q is the number of parameters in the response surface model $\pi(\theta, x)$.

P-optimality maximizes the function of the probability of getting a specific outcome. In this work is specifically employed the *maximin* criterion where the minimum probability of success is maximized. The P-efficiency of a given design ξ with respect to the P-optimal design ξ_P^* is defined as:

$$P_{eff} = \left(\frac{\min\{\pi_i(\theta,\xi)\}}{\min\{\pi_i(\theta,\xi_P^*)\}}\right) \qquad (4)$$

Compound Criterion considers making a compromise between the two criteria described above, as the product of the D- and P-efficiencies of an experimental design ξ, weighted by a mixing constant $0 \leq \alpha \leq 1$. That is[9]:

$$\Phi^{(DP)} = \left[D_{eff}(\xi)\right]^{\alpha}\left[P_{eff}(\xi)\right]^{1-\alpha} \qquad (5)$$

A mathematical nonlinear program (NLP) for optimal design of experiments using the compound criterion can be state as follows:

$$\max_{\xi} \Phi^{(DP)}(\xi) \qquad (6)$$

$$\log\left(\Phi^{(DP)}\right) = (\alpha/q)\log|M(\theta,\xi)| + (1-\alpha)\log(\min\{\pi_i(\theta,\xi)\}), for: i = 1, \dots, n(7)$$

$$\pi(\xi) = \frac{exp(X\theta)}{1+exp(X\theta)} \qquad (8)$$

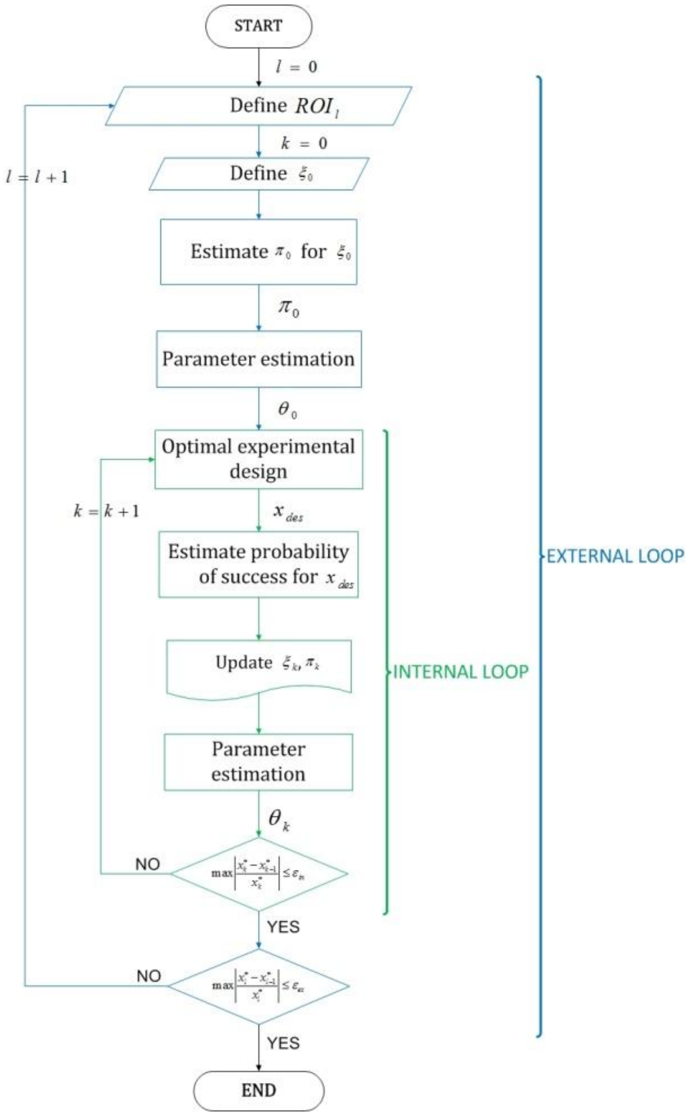
3 Optimization with stochastic binary outcomes

A run-to-run optimization methodology is proposed to pinpoint a reduced region of operating conditions where there is a high probability of satisfying end-point specifications. By increasingly biasing operating conditions so as to increase the parametric precision for the response surface $\pi(\theta, x)$ towards the region where the probability of obtaining a desired outcome is high, a sequence of experiments is defined. The run-to-run optimization algorithm uses two nested loops as shown in Fig. 1.

The first step to begin with the technique consists in establishing the region of interest (*ROI*) for the process, that is, to define the lower bounds (LB) and upper bounds (UB) for all input variables in the vector x. These bounds are defined taking into account available data and prior knowledge about the process under study. Once defined these initial limits, the algorithm starts with the external loop $l = 0$ by choosing points from a central composite design (CCD)[11] to decide where to experiment first in the *ROI*. Each design consists of a factorial design (the corners of a cube) together plus center and star points that allow estimating second-order effects. For a full quadratic model with n factors, CCDs have more than the required design points to estimate the coefficients in a full quadratic model. For the sake of economy, only a fraction of points in the initial CCD are used for the each iteration of the external loop. Thus, to begin each iteration experimental points are chosen as follows: the center and star points are always chosen, whereas factorial points are randomly chosen up to the necessary number for parameter estimation.

For any external loop iteration l, at each iteration of the internal loop (see Fig. 1) an additional point corresponding to the current estimated optimum of the response surface model for the probability of success $\pi(\theta, x)$ is added to ξ_{k-1}. The corresponding experimental design ξ_k includes all data points from the first iteration ($k = 0$) up to the current iteration. For any point in ξ_k, the probability of success is estimated using a number of experimental replicas.

Fig. 1. Run-to-run optimization using probability-based optimal experimental design.


START
$l = 0$
Define ROI_l
$k = 0$
$l = l + 1$
Define ξ_0
Estimate π_0 for ξ_0
π_0
Parameter estimation
θ_0
Optimal experimental design
$k = k + 1$
x_{des}
Estimate probability of success for x_{des}
EXTERNAL LOOP
Update ξ_k, π_k
INTERNAL LOOP
Parameter estimation
θ_k
NO
$\max\left|\frac{x_k^* - x_{k-1}^*}{x_k^*}\right| \le \varepsilon_{in}$
YES
NO
$\max\left|\frac{x_l^* - x_{l-1}^*}{x_l^*}\right| \le \varepsilon_{ex}$
YES
END

Using vectors ξ_k and π_k the response surface of the probability of success is modeled as a GLM with the logistic function as the link function. The corresponding vector of model parameters for the k^{th} internal iteration is θ_k. Accordingly, the model for the probability of success over the *ROI* for the l^{th} iteration of the external loop is:

$$\pi_k = \frac{exp\left(\eta(\theta_k;x)\right)}{1+exp\left(\eta(\theta_k;x)\right)}, x \in ROI \qquad (9)$$

Using the model in Eq. (9) with θ_{k-1} from the previous iteration, in the k^{th} internal iteration a new experimental point x_{des} is added by solving the mathematical program in Eqs. (6, 7 and 8) using the compound objective previously described in Eq. (5), where the vector $\xi_k = [\xi_{k-1}, x_{des}]$. As soon as the probability of success at x_{des} is estimated by doing a number of experimental replicas, a new vector of model parameters θ_k is obtained. Later on, the predicted optimum x_k^* that maximizes the success probability is obtained and compared with the one from the previous iteration to check convergence of the internal stopping criterion.

When the internal stopping condition is satisfied, the optimum for the current iteration of the external loop is calculated as $x_l^* = x_k^*$ and stopping condition for the external loop is checked to assess convergence.

For obtaining the point with the maximum probability of success $\pi(\theta_k, x)$ in the response surface model, the following mathematical nonlinear program (NLP) is solved in the k^{th} internal loop iteration

$$\max_x \pi(\theta_k, x) \qquad (10)$$

$$LB_w^l \leq x_w \leq UB_w^l, w = 1, \ldots, dim(x) \quad (11)$$

where LB_w^l and UB_w^l are the lower and upper bounds of the w^{th} explanatory variables in the l^{th} iteration of the external loop.

Whenever the external stopping criterion is not fulfilled, a new iteration of the external loop begins such that the *ROI* is reduced to a fraction of the previous iteration. In our implementation, the size of each new *ROI* is chosen to be a quarter of the previous one, centered at the optimum x_{l-1}^* found in the previous iteration; the new *ROI* is always chosen within the initial *ROI*. If the optimum is on one boundary of the initial *ROI* or near it, the new *ROI* is conveniently defined so as to be within the initial *ROI* and with a center which is the closest to x_{l-1}^*.

Every time a new external iteration begins with a reduced *ROI*, the first internal iteration begins with fractional CCD based on the corresponding ROI_l. Additional iterations in the internal loop increasingly add new experimental points using probability-based optimal design (see Eqs. 6, 7 and 8). When both loops converge, the last optimal point obtained is chosen as the optimal operating point of the process to work with high probability of ensuring the end-use properties of the product.

4 Case Study: Emulsion Polymerization of Styrene

In the field of the polymers, there exists a special interest in controlling process variables to ensure product fit because there are significant properties that the product needs to have in order to comply with specifications for each particular use. About this, Hinchliffe et al.[12] have developed a model to analyze the influence of simultaneous process conditions and resin characteristics on a set of end-use properties of the polymers.

In this work, the emulsion polymerization of styrene is addressed as a representative example of a process where complying with end-use properties is mandatory. Mathematical models for this process have been discussed in Liotta[13] and Li and Brooks[14] as well as in more recent works[15–17]. Taking into account this available information of the process, a probabilistic simulation model of the process is developed and used to describe the effect of intrinsic variability and modeling errors in end-point specifications. A random variability effect using the percentage of variability of $SP = 10\%$ in the process dynamics is added to the initial amounts of each population of seeded particles as follows:

$$N_i = N_i^{\circ}\left(1 + (rand(1,1) - 0.5).\frac{SP}{100}\right)(12)$$

where N_i° is the initial amount of each specie i.

The kinetic model and conservation equations for the process accounts for the following key equations: volume balance for the reaction medium, volume balance of particles in each population, molar balance of monomer, balance of the average number of radicals per particle for each population, polymerization rate for each population, addition rate of the chain transfer agent (CTA), molar balance of the initiator, molar balance for chain transfer agent. All constitutive equations and balances that made up the dynamic model have been developed in previous works[13–17].

For polymerization processes, it is of special concern that the final product obtained satisfies two end-use properties: the melt flow index (*MI*) and the tensile strength (σ). These properties may be reliably estimated using correlations which are dependent on the weight average molecular weight (MW_w) and the number average molecular weight (MW_n), respectively[2]:

$$MI = \frac{30}{(MW_w^{3.4} \times 10^{-18} - 0.2)} \qquad (13)$$

$$\sigma = 7390 - 4.5110^8\left(\frac{1}{\mathrm{MW_n}}\right) \qquad (14)$$

where MW_w and MW_n can be calculated by using the information arrived with balances of moments for the live and dead polymer[2, 16, 18].

To apply the run-to-run optimization methodology in Fig. 1, the monomer feed rate and the chain transfer agent feed rate to polymerization reactor, both of them in moles per seconds, were chosen as the input variables to guarantee end-use properties. Also, the desired values of the end-point specification for the resulting product are:

$$1.25 \times 10^{-4} < MI < 7.5 \times 10^{-4} [g/min]$$

$$6900 < \sigma < 7200 [psi]$$

To implement the run-to-run optimization methodology is necessary to define first the following set of hyper-parameters: lower and upper operating bounds of each input variable (LB and UB), the number of designed experimental points in each internal loop iteration (*ep*), number of replicates to perform in each operating point (*r*), maximum number of internal loops (*il*), maximum number of external loops (*el*), internal and external stop criteria (ε_{in} and ε_{ex}). Values chosen for hyper-parameters for this case study are shown in Table 1. A quadratic model was chosen to approximate the surface response by GLM and model simulations were performed with the MATLAB software version R2013a. The average CPU time per trial is 15 minutes.

Table 1. Hyper-parameter values for emulsion polymerization of styrene.

Parameter	**Value**
LB	[0.0184,1x10^{-4}]
UB	[0.026,8x10^{-4}]
ep	1
r	10
il	6
el	4
ε_{in}	0.5x10^{-2}
ε_{ex}	1x10^{-3}

This case study constitutes a challenging problem for the proposed methodology. Fig. 2 and Fig. 3 depict the response surface model for the probability of success obtained for the initial *ROI* chosen. Based on 100 independent trials of the run-to-run optimization strategy, the probability of success of the resulting optimal policy has been categorized in Table 2. Results obtained demonstrated that using the proposed methodology, the optimal policy found has a probability of success of more than 90% in 47% of the trials. Moreover, the obtained optimal policy has a probability of success higher than 80% in 73% of the trials. Finally, only a 7% of the trials provide an optimal policy with a probability of success of less than 60%.

Fig. 4 depicts, for a given trial, the systematic reduction of the *ROI* as external loop iterations in the proposed methodology in Fig. 1 are carried out. Table 3 provides a summary of the results obtained, including upper/lower bounds for the reduced *ROI* for external loop iterations.

Fig. 2. Contour levels of probability of success over the initial ROI for the emulsion polymerization of styrene with $SP = 10\%$.

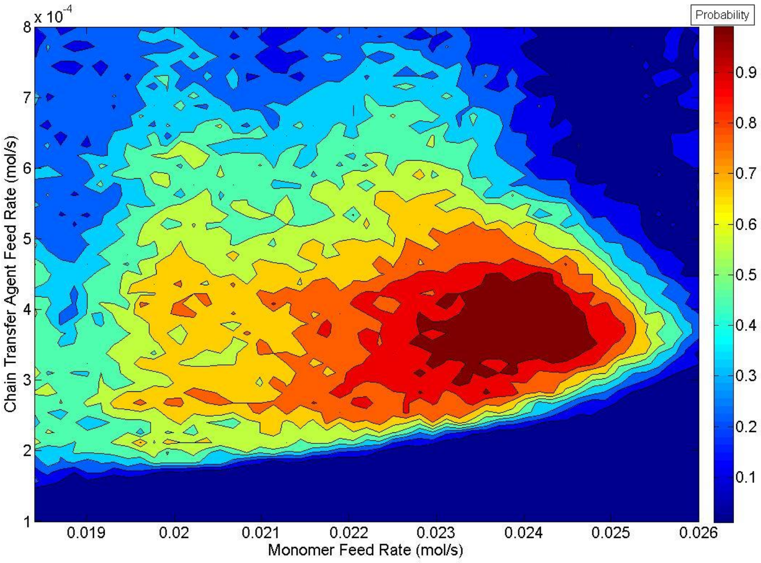


Fig. 3. Response surface of probability of success over the initial ROI for the emulsion polymerization of styrene with $SP = 10\%$.

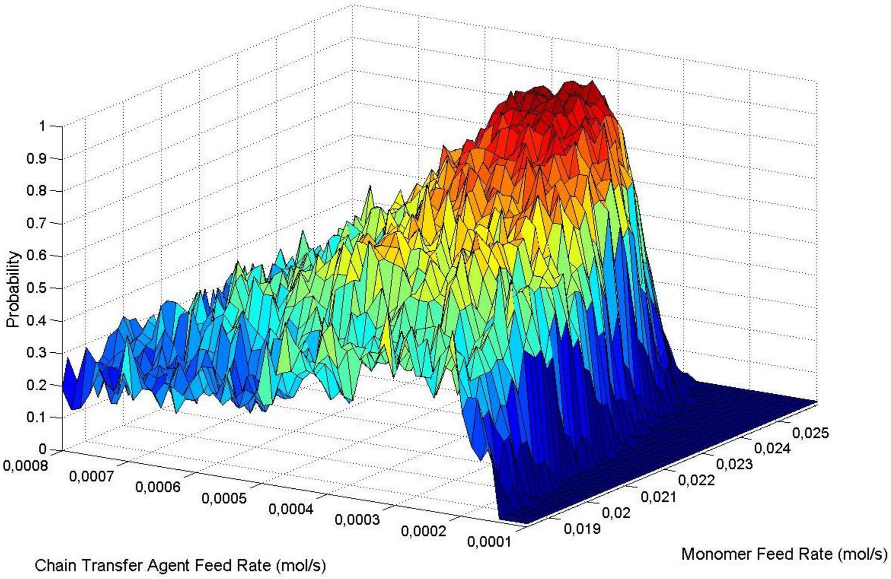


Table 2. Percentage of trials which found an operating policy that complies with the end-point specifications for different levels of success ($SP = 10\%$ and $\alpha = 0.25$).

Minimum probability of success required	Percentage of trials which reach a zone with a given level of success
95%	15%
90%	47%
85%	61%
80%	73%
70%	86%
60%	93%

With the implementation of this case study, it was possible to demonstrate that by embedding data-driven modeling of the probability of success in the proposed run-to-run optimization approach is feasible to define a reduced ROI where end-use properties can be obtained with high probability. Results obtained proved that for complex processes where there are no reliable models for policy optimization, the proposed methodology is an appealing tool to obtain the desired operating region by performing only a small amount of experiments.

Fig. 4. A complete trial (until convergence of both loops) of the run-to-run optimization methodology in Fig. 1 for the emulsion polymerization of styrene ($SP = 10\%$ and $\alpha = 0.25$).

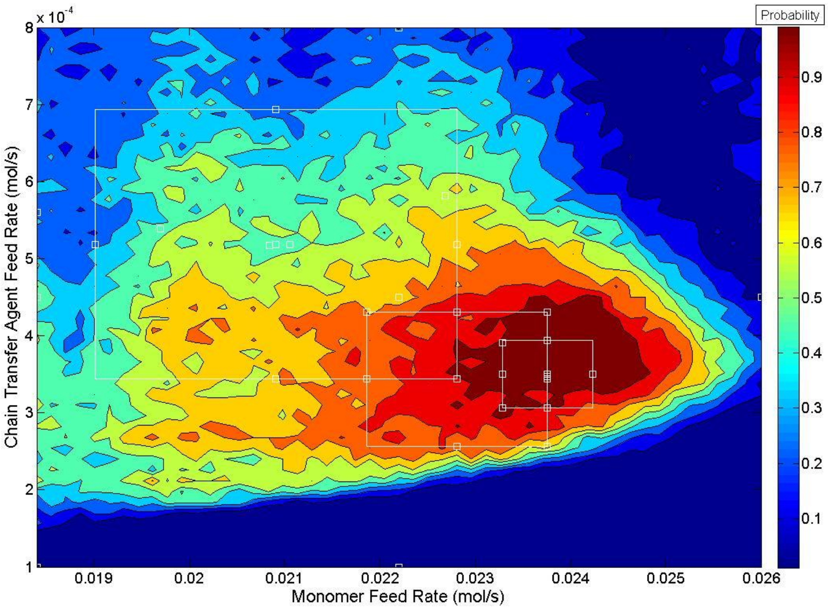


Table 3. A summary of results of the trial depicted in Fig. 4 ($SP = 10\%$ and $\alpha = 0.25$)

External Loop	Lower bound monomer feed rate (mol/s)	Upper bound monomer feed rate (mol/s)	Lower bound CTA feed rate (mol/s)	Upper bound CTA Feed Rate (mol/s)	Optimal monomer feed rate (mol/s)	Optimal CTA Feed Rate (mol/s)	Probability of success	Total of experimental conditions required
1	0.0184	0.0260	1x10^{-4}	8x10^{-4}	0.0209	5x10^{-4}	0.5100	10
2	0.0190	0.0228	3.25x10^{-4}	6.75x10^{-4}	0.0228	3x10^{-4}	0.8400	9
3	0.0219	0.0238	2.13x10^{-4}	3.88x10^{-4}	0.0238	4x10^{-4}	0.9300	9
4	0.0233	0.0243	3.56x10^{-4}	4.44x10^{-4}	0.0242	4x10^{-4}	0.9700	7

5 Concluding remarks

A novel approach to probability-based design of experiments for batch process optimization with end-use properties constraints has been proposed. It aims to define an optimal operating policy with high probability of obtaining the desired outcome. The proposed methodology is data-driven which makes very appealing for innovative processes for which first-principles knowledge or a detailed model is not available. The methodology resorts to a compound criterion for optimal experimental design in order to balance parametric precision with the probability of success. As a result, the optimization strategy in Fig. 1 increasingly bias operating conditions towards a reduced region of operation with higher success probabilities while the response surface model is improved.

The proposed method has been tested *in silico* in a well-known case study to check its efficacy. Emulsion polymerization of styrene has been addressed with encouraging results. For this case study, significant levels of intrinsic variability have been added to simulate uncertainty regarding process behavior. It is worth noting that the proposed methodology can found an optimal policy with a high probability of success after a few runs. More specifically, based on information from 35 experimental points it was possible to identify a reduced area of operation having a probability of success of 97% in satisfying the required specifications for end-use properties. This is encouraging given the complexity of the polymerization process and the high variability involved. Moreover, it was demonstrated that the methodology provides adequate experimental points for parameter estimation of the response surface, and using it the optimal operating zone can be defined. As it can be expected, whenever the intrinsic variability increases results obtained gracefully degrades by lowering the probability of success of the optima policy found.

References

1. Leal-calderon, F.: Emulsified lipids: formulation and control of end-use properties. Doss. Fonct. DES HUILES. 19, 111–119 (2012).
2. Valappil, J., Georgakis, C.: Nonlinear Model Predictive Control of End-Use Properties in Batch Reactors. Am. Inst. Chem. Eng. 48, 2006–2021 (2002).
3. Kong, X., Yang, Y., Chen, X., Shao, Z., Gao, F.: Quality Control via Model-Free Optimization for a Type of Batch Process with a Short Cycle Time and Low Operational Cost. Ind. Eng. Chem. Res. 50, 2994–3003 (2011).
4. Zhao, F., Lu, N., Lu, J.: Quality Control of Batch Processes Using Natural Gradient Based Model-Free Optimization. Ind. Eng. Chem. Res. (2014).
5. Georgakis, C.: Design of Dynamic Experiments: A Data-Driven Methodology for the Optimization of Time-Varying Processes. Ind. Eng. Chem. Res. 52, 12369–12382 (2013).
6. Kolhatkar, A.G., Jamison, A.C., Litvinov, D., Willson, R.C., Lee, T.R.: Tuning the Magnetic Properties of Nanoparticles. Int. J. Mol. Sci. 14, 15977–16009 (2013).
7. Fiordalis, A., Georgakis, C.: Data-driven, using design of dynamic experiments, versus model-driven optimization of batch crystallization processes. J. Process Control. 23, 179–188 (2013).
8. McGree, J.M., Eccleston, J.A., Duffull, S.B.: Compound optimal design criteria for nonlinear models. J. Biopharm. Stat. 18, 646–661 (2008).
9. McGree, J.M., Eccleston, J.A.: Probability-Based Optimal Design. Aust. N. Z. J. Stat. 50, 13–28 (2008).
10. Woods, D.C., Lewis, S.M., Eccleston, J.A., Russell, K.G.: Designs for generalized linear models with several variables and model uncertanly. Southampt. Stat. Sci. Res. Inst. Methodol. Work. Pap. M06/01. (2006).
11. Montgomery, D.C., Runger, G.C.: Applied Statistics and Probability for Engineers. John Wiley & Sons, Inc. (2002).
12. Hinchliffe, M., Montague, G., Willis, M., Burke, A.: Correlating polymer resin and end-use properties to molecular-weight distribution. AIChE J. 49, 2609–2618 (2003).
13. Liotta, V., Georgakis, C., Sudol, E.D., El-Aasser, M.S.: Manipulation of Competitive Growth for Particle Size Control in Emulsion Polymerization. Ind. Eng. Chem. Res. 36, 3252–3263 (1997).
14. Li, B.-G., Brooks, B.W.: Prediction of the Average Number of Radicals per Particle for Emulsion Polymerization. J. Poliymer Sci. Part A Polym. Chem. 31, 2397–2402 (1993).
15. Liotta, V., Sudol, E.D., El-Aasser, M.S., Georgakis, C.: On-line Monitoring, Modeling, and Model Validation of Semibatch Emulsion Polymerization in an Automated Reactor Control Facility. J. Polym. Sci. Part A Polym. Chem. 36, 1553–1571 (1998).
16. Crowley, T.J., Choi, K.Y.: Calculation of Molecular Weight Distribution from Molecular Weight Moments in Free Radical Polymerization. Ind. Eng. Chem. Res. 36, 1419–1423 (1997).
17. Meyer, T., Keurentjes, J.: Handbook of Polymer Reaction Engineering. (2005).
18. Katz, S., Saldel, G.M.: Moments of the Size Distribution in Radical Polymerization. Am. Inst. Chem. Eng. 13, 319–326 (1967).