European Journal of Operational Research 284 (2020) 201–211

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Versatile sequential sampling algorithm using Kernel Density Estimation

Pamphile T. Roy[a,*], Lluís Jofre[b], Jean-Christophe Jouhaud[a], Bénédicte Cuenot[a]

[a] CFD Team, CERFACS, 42 Avenue Gaspard Coriolis, Toulouse cedex 1 31057, France
[b] Center for Turbulence Research, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 28 April 2019
Accepted 30 November 2019
Available online 16 December 2019

Keywords:
Stochastic processes
Design of experiments
Discrepancy
Optimal design
Uncertainty quantification

ABSTRACT

Understanding the physical mechanisms governing scientific and engineering systems requires performing experiments. Therefore, the construction of the Design of Experiments (DoE) is paramount for the successful inference of the intrinsic behavior of such systems. There is a vast literature on one-shot designs such as low discrepancy sequences and Latin Hypercube Sampling (LHS). However, in a sensitivity analysis context, an important property is the stochasticity of the DoE which is partially addressed by these methods. This work proposes a new stochastic, iterative DoE – named KDOE – based on a modified Kernel Density Estimation (KDE). It is a two-step process: (i) candidate samples are generated using Markov Chain Monte Carlo (MCMC) based on KDE, and (ii) one of them is selected based on some metric. The performance of the method is assessed by means of the C^2-discrepancy space-filling criterion. KDOE appears to be as performant as classical one-shot methods in low dimensions, while it presents increased performance for high-dimensional parameter spaces. It is a versatile method which offers an alternative to classical methods and, at the same time, is easy to implement and offers customization based on the objective of the DoE.

1. Introduction

One of the main objectives when performing numerical, or real experiments, is to understand the variation of a Quantity of Interest (QoI) with respect to the variation of some input parameters (Sacks, Welch, Mitchell, & Wynn, 1989). Each experiment, or sample, corresponds to a particular set of input parameters x_k with $k \in [1, \ldots, d]$, where d is the number of dimensions. The group of N_s samples, or Design of Experiments (DoE), is noted as $\mathbf{X}_d^{N_s}$. From exploratory phases to more advanced analyses, such as Uncertainty Quantification (UQ) and robust optimization, DoE aims at helping better understanding the physical mechanisms governing the problem of interest (Saltelli et al., 2007). Therefore, the objective of efficient DoE is to maximize the coverage of input space, i.e., space filling, with the aspiration of capturing most of the underlying physics. Such analyses typically require large number of experiments to converge the statistical moments of the QoIs. In this regard, many studies have focused on reducing the computational cost. However, depending on the required quality of the analysis, the complexity of the experiment, or its return time, the total number of experiments may be limited. Thus, the objective of this work is to optimize the space-filling properties for a given computational budget.

Different metrics are commonly used to assess the space filling of a DoE. They can be categorized into (i) geometrical and (ii) uniformity criteria. Among the most used geometrical criteria are the *maximin* and *minimax* (Pronzato, 2017). They, respectively, maximize the minimum distance between all points or minimize the maximum distance between any location in space and all points of the sample. A similar criterion is found by using a *minimum spanning tree* (Franco, Vasseur, Corre, & Sergent, 2009) in which the best design corresponds to a maximization of the mean distance between all connections among samples and the minimization of the variance in these distances. The uniformity criterion, instead, measures how the spread of the points deviates from a uniform distribution. The central discrepancy is commonly used (Damblin et al., 2013; Fang, Li, & Sudjianto, 2006) to measure the uniformity.

There are three main methodologies for building a DoE: (i) *Monte Carlo* (MC), (ii) *Latin Hypercube Sampling* (LHS) and (iii) Quasi-*Monte Carlo* (QMC) methods (Cavazzuti, 2013; Garud, Karimi, & Kraft, 2017). Kucherenko, Albrecht, and Saltelli (2015) have recently compared MC and LHS against the well-established low dis-

* Corresponding author.
E-mail address: roy@cerfacs.fr (P.T. Roy).

https://doi.org/10.1016/j.ejor.2019.11.070
0377-2217/

crepancy sequence of *Sobol'*. They concluded that LHS and QMC both offer superior integration performance over MC. LHS-based sampling methods are one-shot design strategies (Fang et al., 2006; Mckay, Beckman, & Conover, 1979). Their utilization requires the practitioner to set *a priori* the total number of samples contained in the DoE. Although, there have been some attempts to construct progressive LHS, they still require an initial design to work properly (Sheikholeslami & Razavi, 2017). On the other hand, low discrepancy sequences are iterative designs which can be continued without compromising the discrepancy. The practitioner is then able to increase the number of samples afterwards for quality reasons, for instance, or if other experiments can be afforded. Liu, Ong, and Cai (2018) recently reviewed iterative DoE in a metamodeling context. In their study, it is shown that most iterative methods need an initial design as a starting point. By using an initial design, the physics information from the system can be used to further guide the construction of the DoE. In this case, such iterative methods are called adaptive methods. Except for some work in Crombecq, Laermans, and Dhaene (2011), and to the best of the authors' knowledge, the number of iterative methods not requiring an initial design is limited. This can be explained both by the quality of the initial designs (using LHS of *Sobol'* sequence) and by the performance of the refinement algorithm. There are even fewer options if the iterative design cannot take advantage of the output of the experiments. Low discrepancy sequences are an example of such methods. But in some context, stochastic methods may be required – to compute sensitivity indices for instance Saltelli et al. (2010). Scrambling the sequences (Owen, 1998) can avoid this pitfall but then the method is no longer iterative.

Our work proposes a new methodology to stochastically sample the input parameter space iteratively allowing, at the same time, to take into account any constraint, such as non-rectangular DoE (Lekivetz & Jones, 2015), sensitivity indices or even constraints on the quality of particular subprojections as in Joseph, Gul, and Ba (2015). Starting from an initial design, or with a single random sample, a Kernel Density Estimation (KDE) is used to infer possible new samples. A second step selects the new sample to be added to the DoE. This two-step DoE is versatile and shows good discrepancy properties compared to standard one-shot designs.

The paper is organized as follows. Section 2 describes first a representative example of a complex case requiring an efficient DoE. A solution utilizing the iterative technique proposed in this paper is presented next in Section 3. Section 4 demonstrates the performance of the method by considering an exhaustive set of numerical examples. Finally, conclusions and future work are drawn in Section 6.

2. Motivation example: predictive studies of multiphysics turbulent flows

The number of uncertainties involved in the study, design and optimization of complex, multiphysics turbulent flows is typically large due to (i) the modeling assumptions required to mathematically describe the different physics and their couplings, i.e., epistemic uncertainty, and (ii) the aleatoric incertitude resulting, for instance, from the lack of detailed evidence regarding the initial and boundary conditions. Therefore, numerical analyses based on single deterministic realizations for a particular set of input parameters cannot be deemed predictive (Roache, 1997). A solution to this problem is to consider the system under study stochastic and analyze the relation between input and output probability distributions by means of efficient statistical methods. In this regard, the field of uncertainty quantification (UQ) applied to computational sciences and engineering has remarkably grown over the last decades, e.g., Najm (2009), Chernatynskiy, Phillpot, and LeSar (2013), Beran, Stanford, and Schrock (2017), Masquelet et al. (2017),

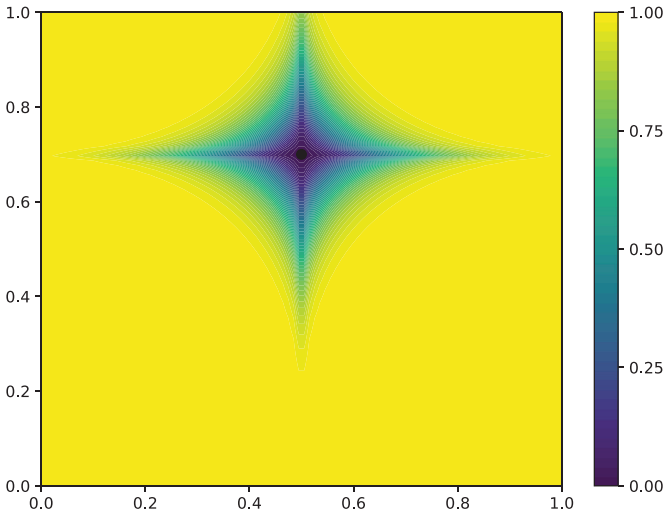


Fig. 1. Probability density function of presence in a 2-dimensional parameter space. *Dot* represent a sample already drawn.

and it is now extensively accepted that the potential of estimating and minimizing uncertainties, in combination with numerical verification and physics validation (V&V), is crucial for augmenting the confidence in the numerical predictions.

As an example, in the field of solar energy engineering (Ho, 2017), the physics-based modeling of irradiated, particle-laden turbulent flow, and its numerical investigation, are difficult tasks that intrinsically require several model assumptions, selection of coefficient and parameter values, and characterization of initial and boundary conditions (Jofre, Geraci, Fairbanks, Doostan, & Iaccarino, 2017). These steps, even if performed carefully, result in sources of uncertainty that can impact the quantities of interest (QoI). Some examples encompass the incomplete description of particle diameters (Rahmani, Geraci, Iaccarino, & Mani, 2018) and thermal radiative properties (Frankel & Iaccarino, 2017), variability of the incident radiation and its complex interaction with boundaries, and model-form incertitude (Jofre, Domino, & Iaccarino, 2018; 2019). As a result, the number of uncertainties involved in such studies is typically in the order of $\mathcal{O}(10^1 - 10^2)$. In addition, accurate representation of the underlying physical phenomena mandates the utilization of expensive high-fidelity (HF) computations based on point particle direct numerical simulations (Esmaily, Jofre, Mani, & Iaccarino, 2018; Jofre et al., 2017). Hence, characterization of the stochastic output by means of naively running hundreds, or thousands, of HF realizations with different input values easily exceeds the resources of the largest computing facilities available.

In this regard, efficient DoE strategies that reduce the number of samples required, like the one presented in this work, are paramount for accelerating the calculation of these type of studies within a feasible computing budget.

3. Presentation of the method

In its basis form, our sequential sampling strategy consists in adding a point far from the existing points in the parameter space. The notion of distance corresponds to a measure of discrepancy. However, instead of considering the whole hypercube, the proposed technique only focuses on empty regions defined using an Exclusion Field (EF). This exclusion field describes the probability of selecting a new point depending on its position and allows to generate new samples that are located preferentially in these empty regions. Then, out of the n_{gen} generated samples from the EF, the one that leads to the best value of some criterion is selected. It is to be noted that there is no optimization process in the

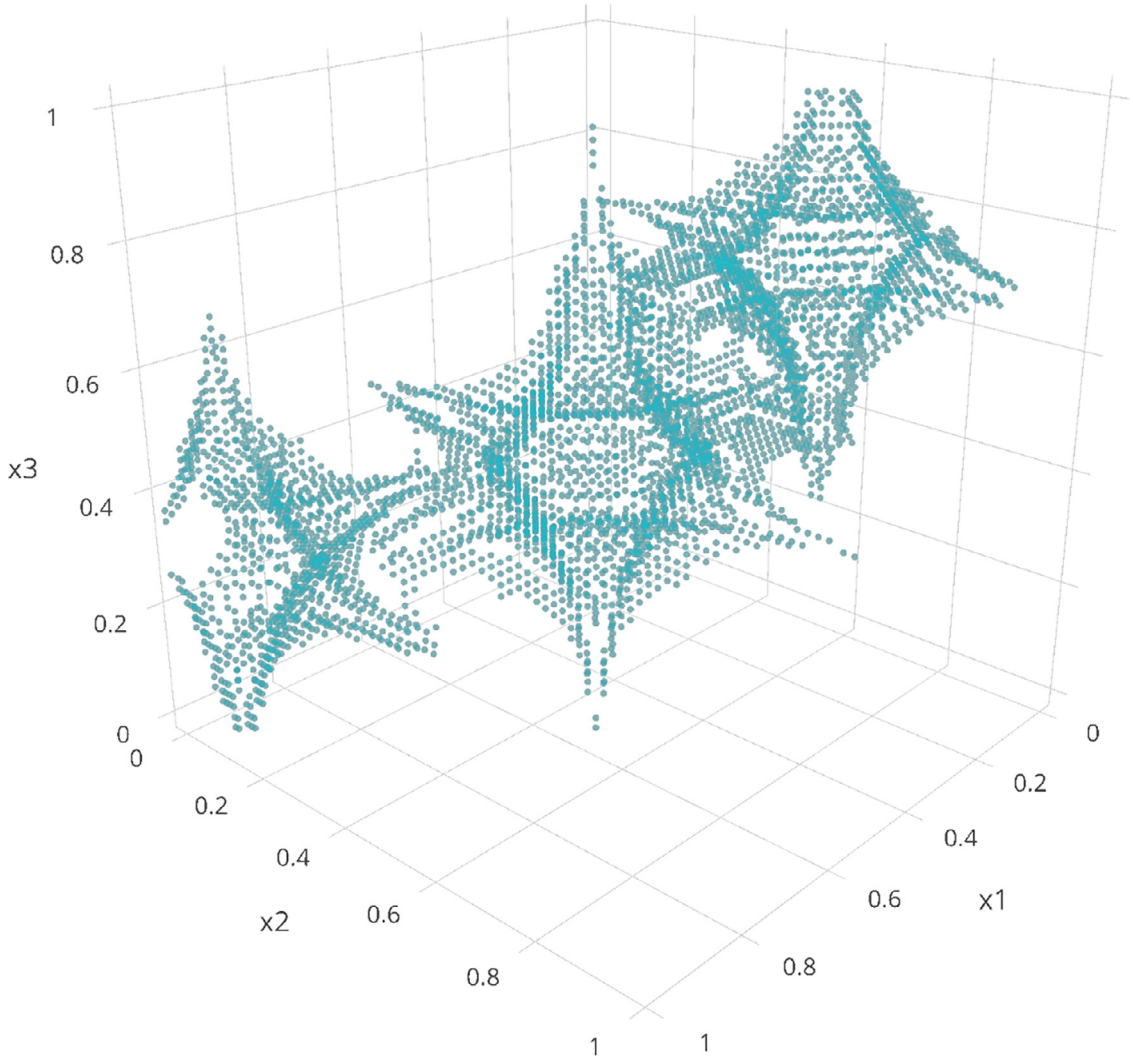


Fig. 2. Scatter plot representation of a 3-dimensional PDF with three points already set in the parameter space. Points represents an iso value of probability.

sense that it is just a selection process based on probable samples of the EF. The whole process ensures randomness in the generation of samples in the parameter space.

Section 3.1 introduces the EF, and Section 3.2 describes the sampling procedure from the EF. Finally, Section 3.3 gives an overview of the method. In the following, it is referred to as the Kernel-DoE (KDOE) method.

3.1. Determination of the exclusion field

Assumed that N_s samples have already been selected. The spatial probability density function used to draw a new sample is given by

$$f(\mathbf{x}) = 1 - \sum_{i=1}^{N_s} K\left(\mathbf{x}, \mathbf{x}^{(i)}\right). \tag{1}$$

The N_s samples that have already been chosen are denoted $\mathbf{x}^{(i)}$, with i between 1 and N_s. The dimension is noted d and K is a kernel expressed by

$$K\left(\mathbf{x}, \mathbf{x}^{(i)}\right) = \exp\left(-\frac{D\left(\mathbf{x}, \mathbf{x}^{(i)}\right)^2}{2h^2}\right), \tag{2}$$

D is a distance function that will be expressed later. The general idea is to lower the probability of selecting a new point close to the samples already drawn. Hence a zone of exclusion is created around each of the already selected points, with a width parameterized by h, set here at $h = \sigma / N_s^{1/d}$ with $\sigma = 0.3$. This in particular allows to have a width of exclusion that decreases as the number of samples increases. In addition, the probability is set to 0 outside of the unit hypercube in order to prevent sampling outside of the region of interest. We also ensure that the probability is always greater than or equal to 0. Note that this probability is not normalized. It will be shown in the next section that normalization is not required for the sampling procedure

Various expressions of the distance function D allow to generate many different shapes. In the present case, alignment of samples on each axis should be avoided as done with Latin Hypercube Sampling designs. This is achieved by using a Minkowsky distance (Cha, 2007) for D

$$D\left(\mathbf{x}, \mathbf{x}^{(i)}\right) = \left(\sum_{j=1}^{d} |x_j - x_{ij}|^p\right)^{1/p}, \tag{3}$$

where p is the order of the distance. Setting $p < 1$ leads to a *star shape* for the PDF f, as shown in Fig. 1 starting from one already selected sample ($N_s = 1$), and using $p = 0.5$. In Fig. 2, three samples were already selected in a 3-dimensional parameter space. The star shape is visible in all dimensions, its branches interact with each

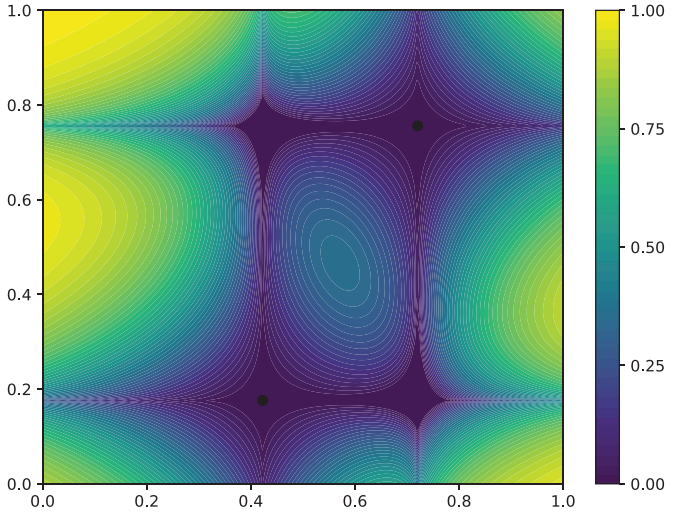


Fig. 3. Cumulative effects on the probability density function in a 2-dimensional parameter space. *Dots* represent 2 existing samples.

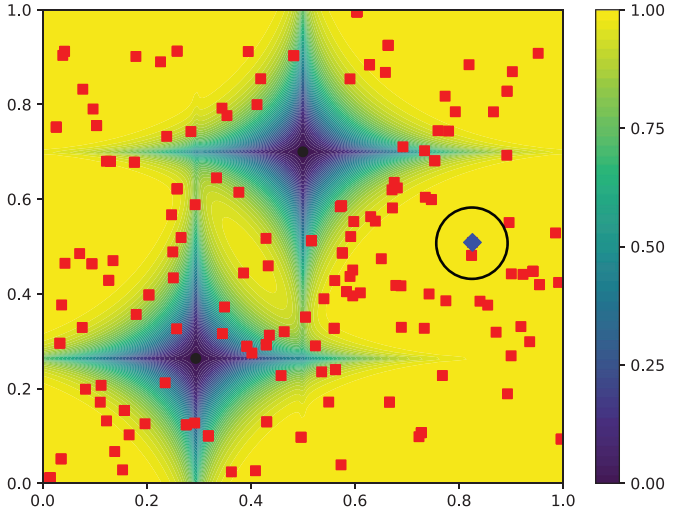


Fig. 4. Probability density in a 2-dimensional parameter space. *Dots* represent the samples already drawn, *squares* are the result of the Metropolis–Hasting sampling and *circled-diamond* is the sample selected based on the resulting centered discrepancy.

other as shown in Fig. 3 clearly illustrating the cumulative property of the PDFs. In this last case $\sigma = 0.8$ to highlight this property.

3.2. Sampling and selection procedures

The classical way to sample from a PDF is to use the inverse transform sampling method. However, finding the inverse cumulative distribution function of a complex PDF can be computationally intensive – the cost increases with dimensionality. Here the *Metropolis–Hasting* (Hastings, 1970) Markov Chain Monte Carlo (MCMC) algorithm was selected. Contrary to methods such as HMC or NUTS (Hoffman & Gelman, 2014), it does not require the calculation of the gradient of the log-probability density function, which is a costly operation. This algorithm provides a random walk of the parameter space that converges toward the target PDF.

Fig. 4 shows an example with two initial points already selected in the hypercube $[0, 1]^2$. Based on these two points (*dots*), f is built and new samples are drawn (*squares*) using the MCMC method.

The next step consists in choosing a new sample from these candidates. Any metric can be chosen here depending on the final

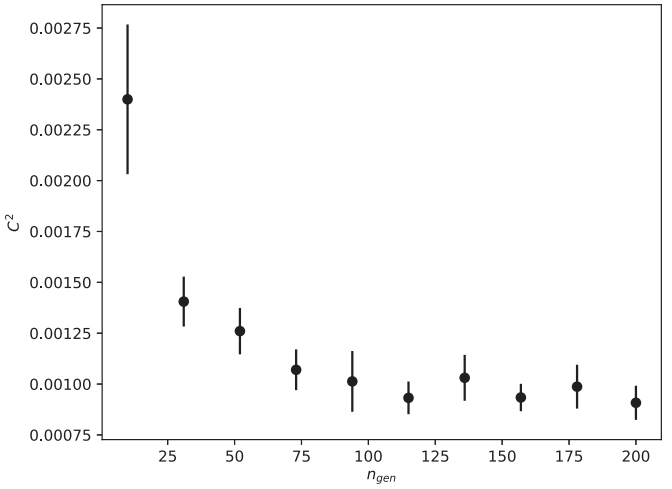


Fig. 5. Convergence of the C^2-discrepancy function of n_{gen}, the size of sample using Metropolis–Hasting MCMC for X_2^{40}.

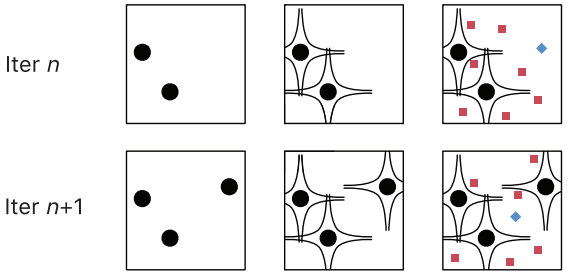


Fig. 6. Schema representing two iterations of the algorithm with the construction of the exclusion field, the sampling and selection of a new candidate. *Dots* represent the samples already drawn, *squares* are the result of the Metropolis–Hasting sampling and *diamonds* are the sample selected based on a metric.

objective. In the following, focus is made on the uniformity of the DoE. Hence, the centered discrepancy C^2 is used (Fang et al., 2006). It writes

$$C^2(\mathbf{X}_d^{N_s}) = \left(\frac{13}{12}\right)^d - \frac{2}{N_s}\sum_{i=1}^{N_s}\prod_{k=1}^{d}\left(1+\frac{1}{2}\mid x_k^{(i)}-0.5\mid -\frac{1}{2}\mid x_k^{(i)}-0.5\mid^2\right) + \frac{1}{N_s^2}\sum_{i,j=1}^{N_s}\prod_{k=1}^{d}\left(1+\frac{1}{2}\mid x_k^{(i)}-0.5\mid +\frac{1}{2}\mid x_k^{(j)}-0.5\mid -\frac{1}{2}\mid x_k^{(i)}-x_k^{(j)}\mid\right). \tag{4}$$

Since the lowest values of C^2 result in most uniform samples, the sample that minimizes C^2 is chosen (*circled-diamond*).

The whole procedure is then restarted with an added point to the initial set. This results in an iterative procedure which acts like an optimizer on the C^2-discrepancy where the candidates are not drawn totally randomly but with the knowledge of the existing samples.

Thanks to the MCMC method, the procedure is not deterministic which is useful to generate a new independent set of experiments. Indeed, to compute sensitivity indices of *Sobol'*, two independent samples are required (Saltelli et al., 2010). As stated in Saltelli et al. (2010), quasi-random sequences such as *Sobol'* are classically used but, as they are deterministic, they cannot generate independent samples. In order to get two independent samples, a sample of shape $X_{2d}^{N_s}$ is generated. Splitting the matrix column wise-like ensures independence of the two resulting samples. However, as the dimensionality increases, the quality of the sequence deteriorates ($d > 10$). Hence, this technique is limited to a

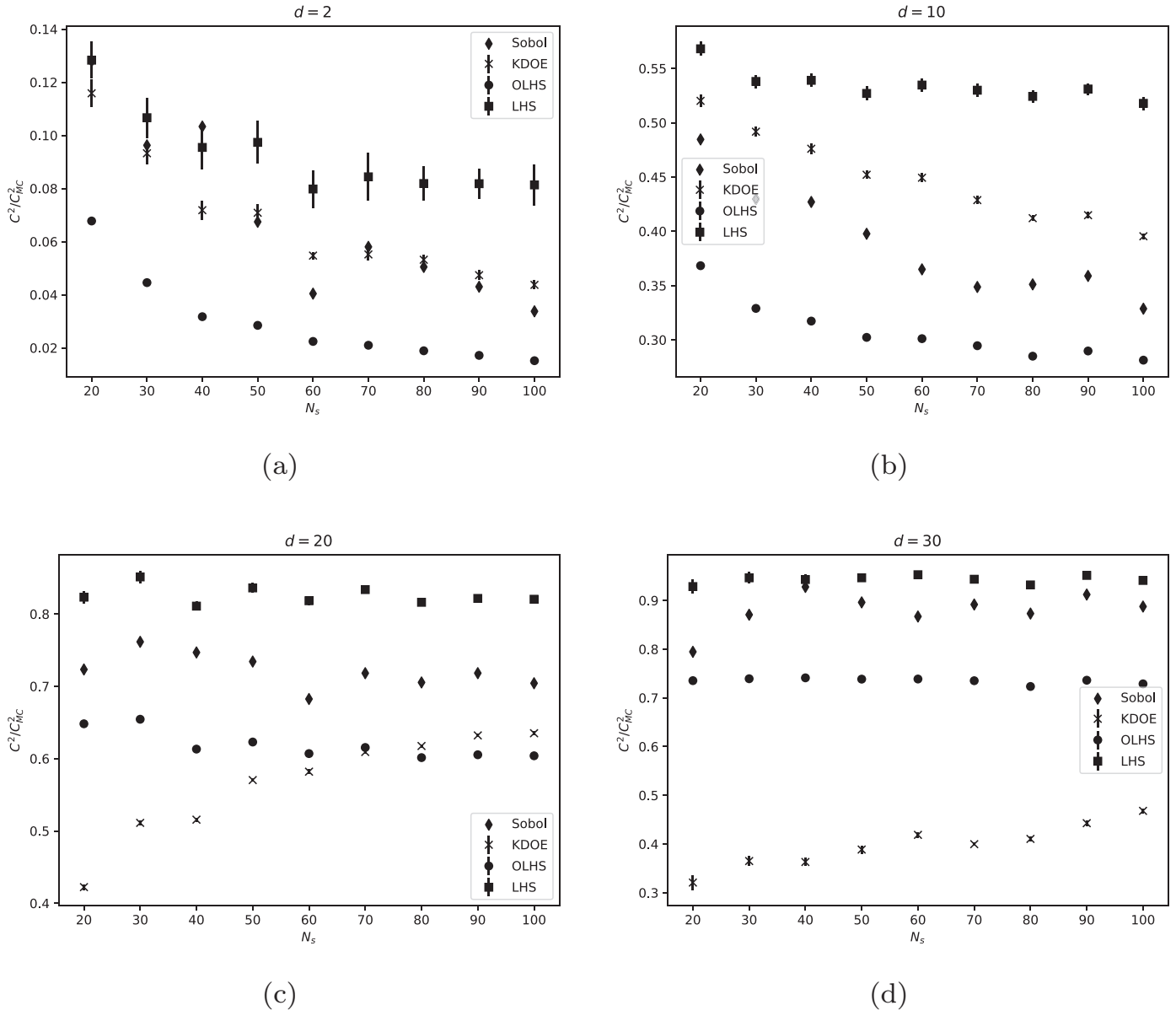


Fig. 7. Normalized C^2-discrepancy function of the number of dimensions d of the parameter space and of the size N_s of the design for various DoE methods.

small number of dimensions. The method proposed in this paper overcomes this limitation.

As stated, n_{gen} candidate samples are generated through MCMC. Fig. 5 presents a convergence analysis of the quality (via C^2) of the final design X_2^{40} depending on the size of the MCMC sample at each iteration. Confidence intervals are calculated using 100 realizations of the same parameterization. The discrepancy converges to its final value above $n_{gen} = 100$. This allows to control the computational cost required to generate the DoE. Various configurations of $X_d^{N_s}$ have been tested and results are similar. In the rest of this paper, n_{gen} is fixed to 100.

3.3. Algorithm

Thus, the sequential strategy is described in Algorithm 1 and in Fig. 6 where two iterations are schematized. From a given DoE, an exclusion field is constructed. Then points are sampled using a Metropolis–Hasting strategy, and finally a new candidate is selected based on a metric. This process is repeated until the desired size of the DoE is reached.

Algorithm 1 Sampling Strategy: Kernel-DoE.

Require: $\mathbf{X}_d^{N_s}$, N_{max}, N_s, n_{gen} ▷ Start from a sample $\mathbf{X}_d^{N_s}$ composed of N_s samples in dimension d

1: **while** $N_s < N_{max}$ **do**
2: $f \leftarrow$ Construction of the Exclusions Field from $\mathbf{X}_d^{N_s}$
3: $\mathbf{Y}_d^{n_{gen}} \leftarrow$ pick n_{gen} samples using Metropolis–Hasting
4: $\mathbf{Y}_d^{j} \leftarrow$ point in N_{gen} which minimize the discrepancy
5: $\mathbf{X}_d^{N_s+1} \leftarrow \left\{\mathbf{X}_d^{N_s}, \mathbf{Y}_d^{j}\right\}$
6: **end while**

4. Results

4.1. Uniformity of the design

As stated previously, the uniformity of the DoE is paramount to ensure that the physics of interest are well captured. Fig. 7 presents a convergence study of the KDOE method ver-

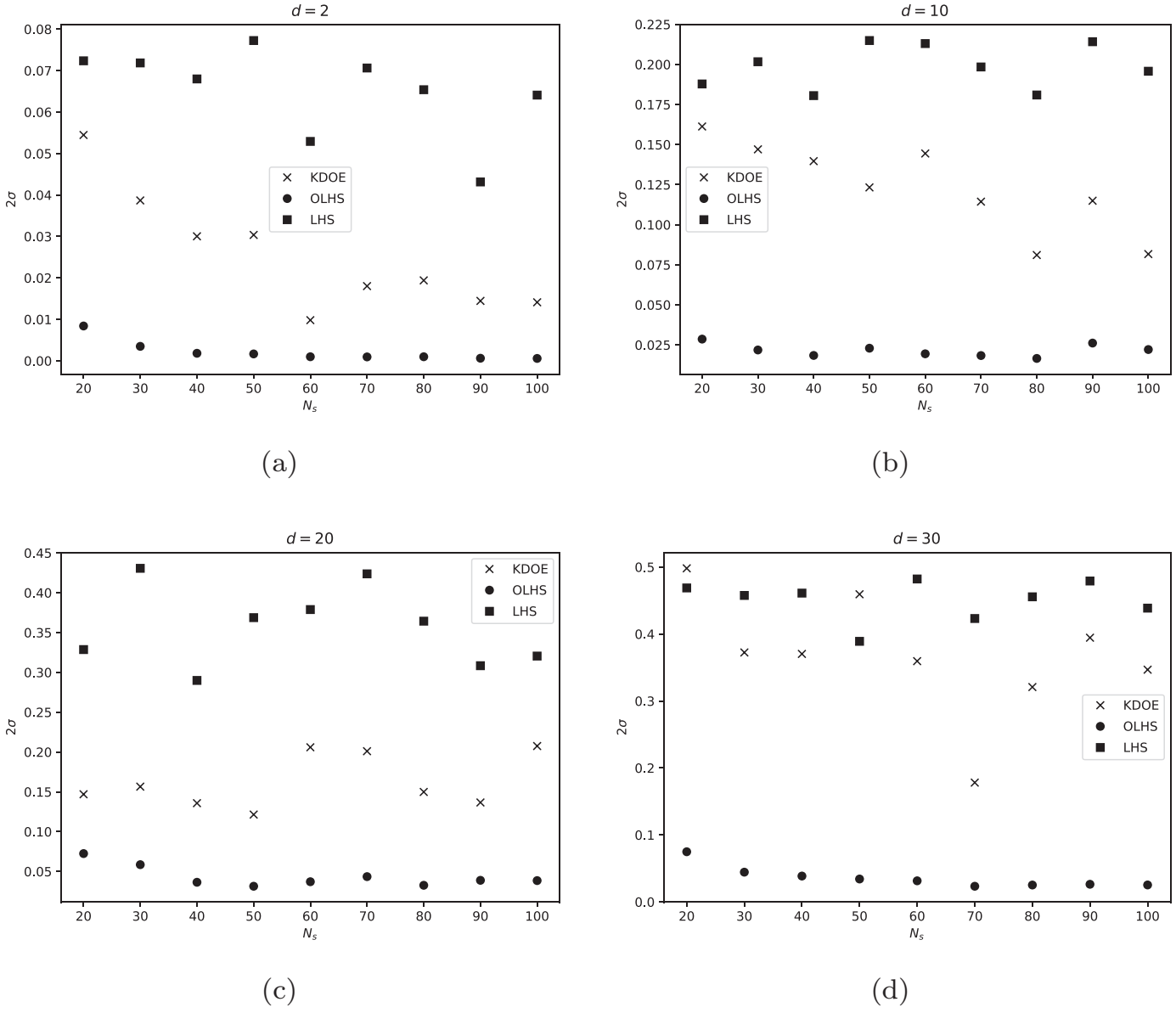


Fig. 8. Normalized deviation at 2σ on the C^2-discrepancy function of the number of dimensions d of the parameter space and of the size N_s of the design for various DoE methods.

sus *Sobol'* sequences (Sobol', 1967), classical LHS (Mckay et al., 1979), and optimized LHS as proposed in Baudin, Dutfoy, Iooss, and Popelin (2015). Each point corresponds to a given sample size N_s for a given number of dimensions d. Due to the stochastic nature of the LHS algorithms and of the KDOE, confidence intervals are computed based on 100 realizations. To measure the improvement of a method with respect to the other, the C^2-discrepancy is used (Androulakis, Drosou, Koukouvinos, & Zhou, 2016; Fang et al., 2006) and values are normalized by crude *Monte Carlo* (MC) results. This transformation shows that a uniform improvement factor is obtained in comparison to MC. Looking for instance at $d = 20$, LHS enables a 20% improvement in terms of C^2-discrepancy over MC, *Sobol'* sequence gives 30% and both OLHS and KDOE roughly give 40%.

The obtained hierarchy between LHS, OLHS and *Sobol'* is quite stable. OLHS is the best method followed by *Sobol'* sequence and finally LHS. For KDOE, it performs closely to *Sobol'* sequence up to $d \lesssim 20$. For $d \gtrsim 20$, KDOE performs better than all the other methods tested.

Moving on to two standard deviation (2σ) – see Fig. 8 –, the results of KDOE always lies between LHS's and OLHS's.

Fig. 9 presents the convergence analysis of the C^2-discrepancy as function of the number of dimensions for $N_s = 100$. When the dimensionality increases, the gain with both LHS and *Sobol'* sequences is close to zero. On the contrary, OLHS seems to stabilize around a 30% improvement. Regarding KDOE, it performs equally with other methods up to $d \lesssim 20$, while for $d \gtrsim 20$ it becomes more performant. It can be seen that the method has not yet reached its minimum at $d = 40$.

In terms of C^2-discrepancy, KDOE appears to perform better with respect to crude *Monte Carlo*, LHS, OLHS and *Sobol'* sequence. Fig. 10 shows an example of a sample of size $N_s = 50$ in dimension $n_{dim} = 30$. The subprojection x_{20}/x_8 is represented. Fig. 10(d) depicts the principal challenge with classical *Sobol'* sequences. In high dimensional parameter space, clear patterns may appear in some subprojections. This behavior was not observed with KDOE. In this case, the result of the KDOE may not appear optimized for 2-dimensional subprojections. This is due to the fact that the

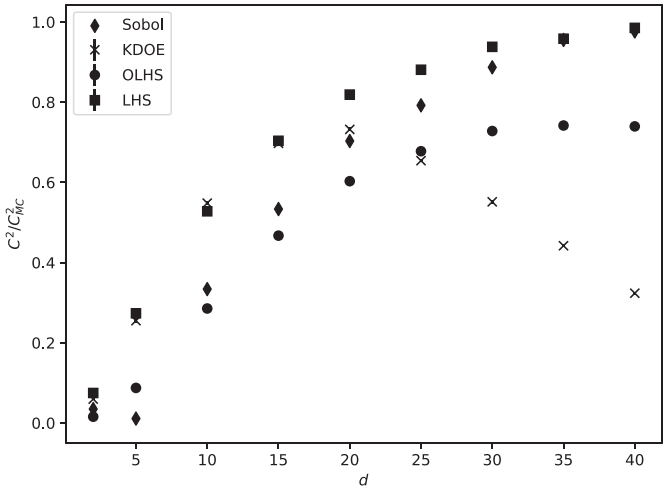


Fig. 9. Normalized C^2-discrepancy function of the number of dimensions n_{dim} of the parameter space with a design of size $N_s = 100$ for various DoE methods.

objective is to optimize the total discrepancy of the parameter space.

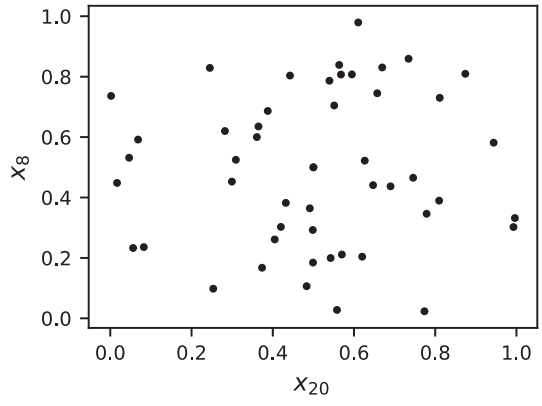
4.2. Integration convergence

Even if this method is not designed for integral evaluation, its performance is evaluated on small numbers of samples up to 512. The number of evaluations has been restricted as the purpose of the method is to generate a small design in high dimensions. Also, the use of an iterative method to generate such sample can be questioned due to the resulting computational cost. Moreover, although this method can be used to continue an existing design created using another technique, such possibility was not evaluated in the following. in Kucherenko et al. (2015), convergence plots are presented in order to assess the performance of *Sobol'* sequence versus LHS and *Monte-Carlo* sampling. The functions used are categorized into types A, B and C. These categories state how the variables are important with respect to the function output:

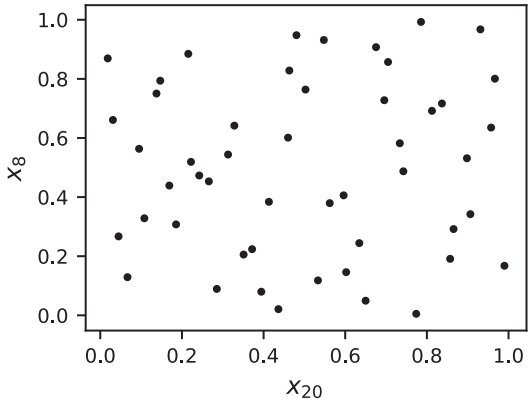
Type A: Functions with a low number of important variables,
Type B: Functions with almost equally important variables but with low interactions with each other,
Type C: Functions with almost equally important variables and with high interactions with each other.

Type C functions represent the most challenging case. In this work, one function per group is considered as detailed in Table 1. The theoretical integral for all these functions in the unit hypercube is 1. Quality of the integration is computed using the Root Mean Square Error (RMSE) defined as

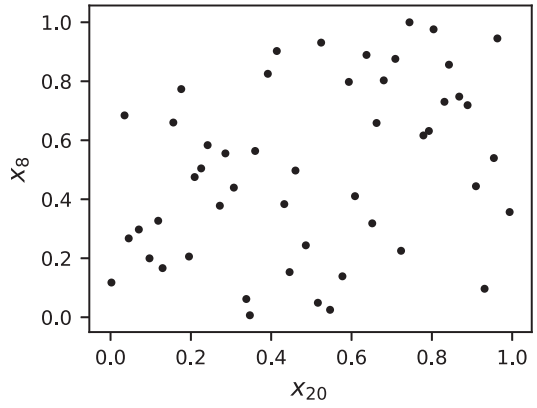
$$\epsilon = \left(\frac{1}{K} \sum_{k=1}^{K} \left(I[y] - I^k_{N_s}[y] \right)^2 \right)^{1/2}, \tag{5}$$



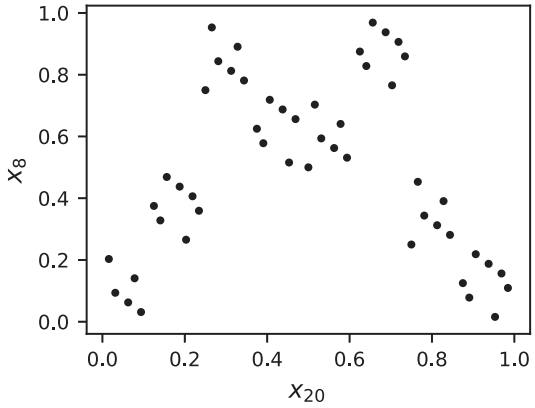
(a) KDOE—$C^2 = 2.18$



(b) OLHS—$C^2 = 3.43$

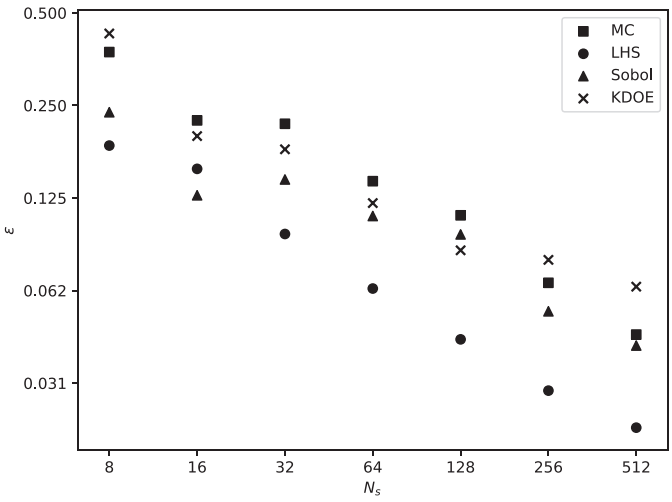


(c) LHS—$C^2 = 3.81$

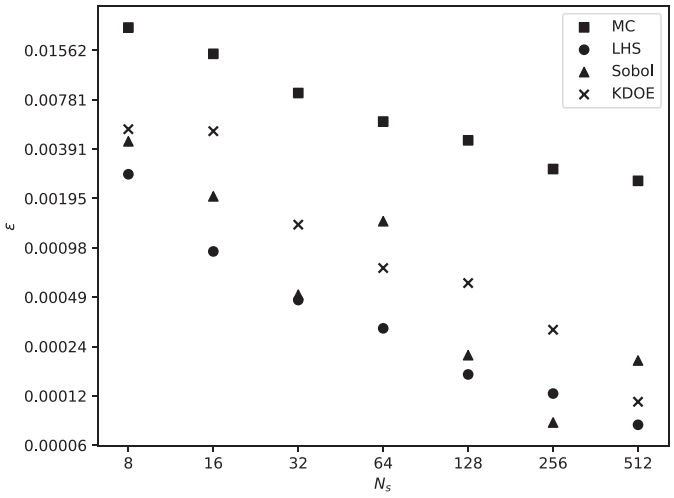


(d) Sobol'—$C^2 = 3.76$

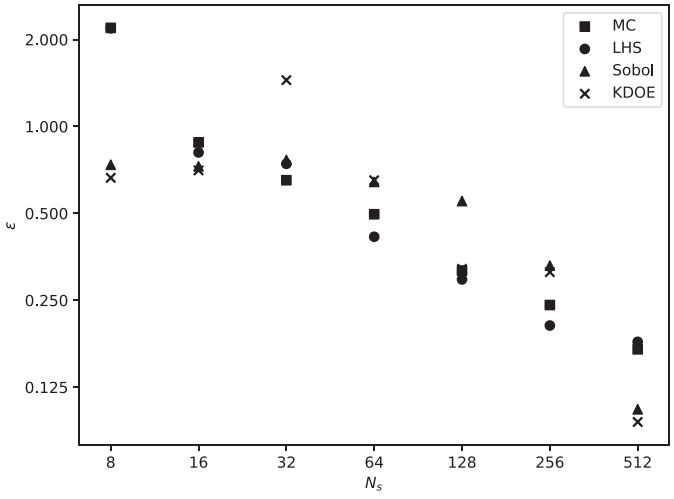
Fig. 10. Example of a 2-dimensional subprojection of the sample of size $N_s = 50$ in dimension $d = 30$ with various DoE methods.



(a) Type A

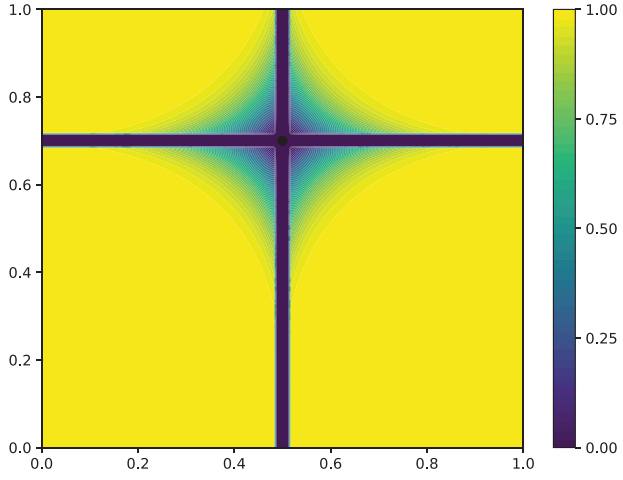


(b) Type B

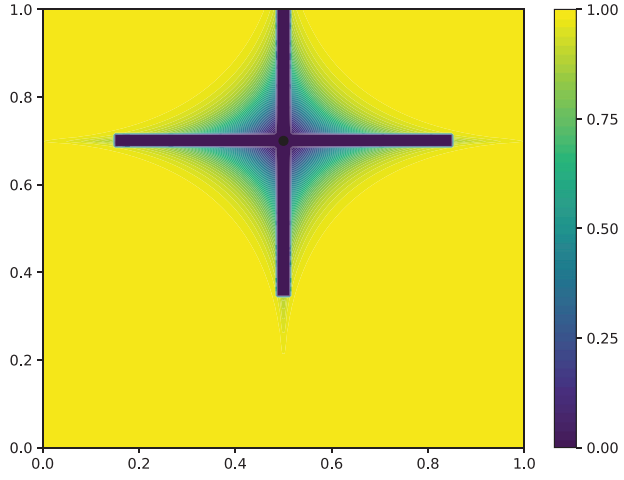


(c) Type C

Fig. 11. RMSE function of the sample size N_s for type A, B and C functions.



(a) Inverse Minkowsky distance with LHS properties



(b) Inverse Minkowsky distance with LHS properties and constraint

Fig. 12. Probability density in a 2-dimensional parameter space. *Dot* represents the sample used to construct the KDE.

Table 1
Type A, B and C functions used in the convergence analysis.

Type	Function $y(x)$	Dim d
A	$\prod_{i=1}^{d} \frac{\mid 4x_i - 2 \mid + a_i}{1 + a_i}$	30
B	$\prod_{i=1}^{d} \frac{d - x_i}{d - 0.5}$	30
C	$2^d \prod_{i=1}^{d} x_i$	10

with y the function to integrate and $K = 50$ the number of independent trials and the estimate integral defined as

$$I^k_{N_s}[y] = \frac{1}{N_s} \sum_{i=1}^{N_s} y(\mathbf{X}^i_d), \tag{6}$$

Fig. 11 presents the convergence study. KDOE is not the best method but seems to compare well to both LHS and *Sobol'* sequence. The convergence rates are correct for every function type.

5. Perspectives

Depending on the property sought, the combination of a Kernel and a metric allows an infinite number of possible customizations of the method.

Using the Minkowsky distance as a metric, the LHS constraint is not strict which can be useful when dealing with discrete parameters. Indeed, strict LHS would prevent having more than one sample per discrete parameter. In Fig. 12(a), an additional constraint is added to strongly limit the probability to 0 when the $L_{-\infty}$-norm is inferior to a threshold. This limitation can be restricted to a domain of influence using an additional L_2-norm constraint (Fig. 12(b)). Hence, the presented method acts as an iterative LHS strategy.

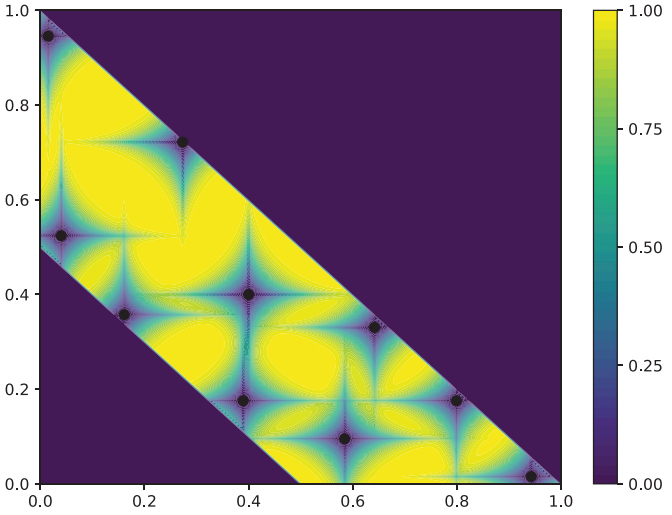


Fig. 13. Probability density in a non-rectangular 2-dimensional parameter space. The 10 *dots* represent the samples used to fit the KDE.

Using this method, it is also possible to consider non-rectangular domains (Lekivetz & Jones, 2015). This example presents a 2-dimensional domain with the constraint $0.5 < x_1 + x_2 < 1$. In this case, the selection of the point criterion has to be changed as the C^2-discrepancy assumes rectangular domains. Fig. 13 shows a sampling of the aforementioned constrained design using a *maximin* criterion (Fang et al., 2006). This criterion

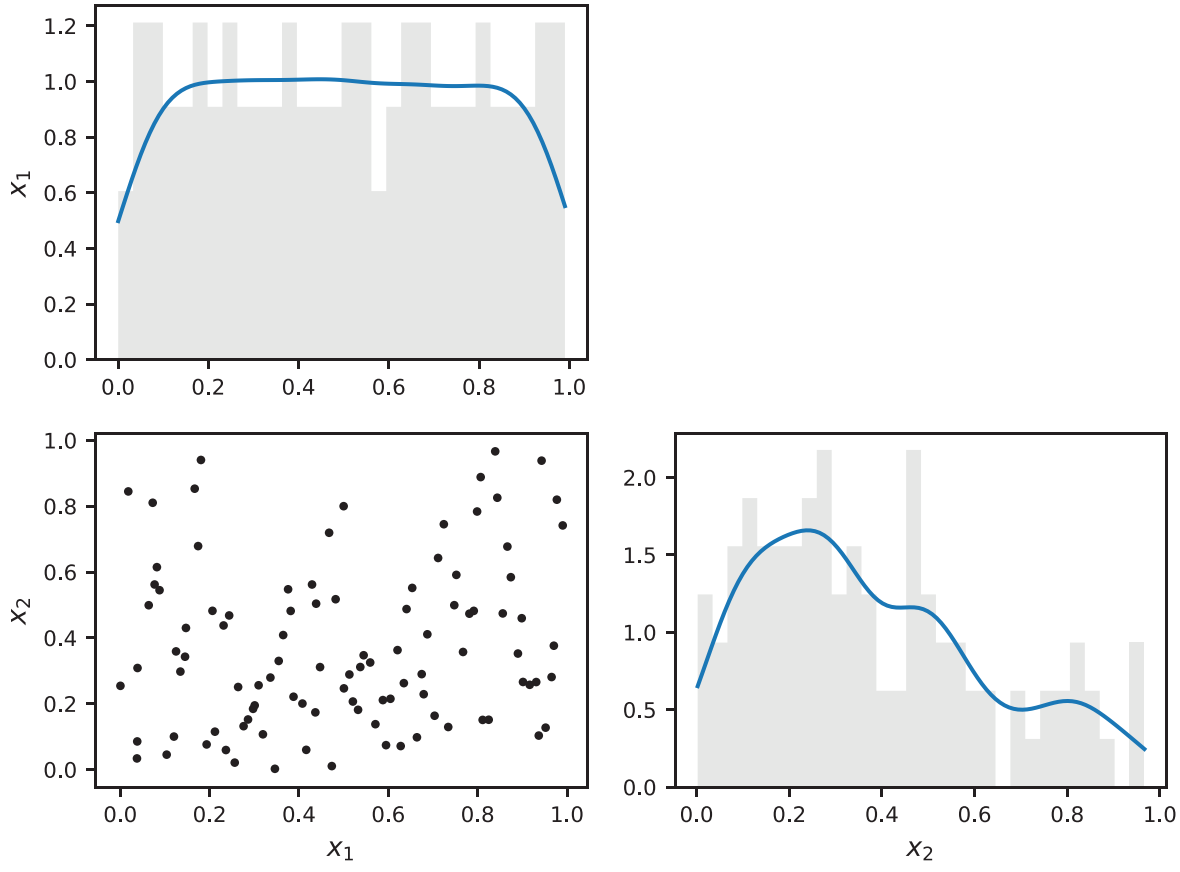


Fig. 14. 2-dimensional parameter space with x_0 the highest. *Dots* represent the sample. Sample distributions for each parameter are plotted along the diagonal.

only considers the points of the sample resulting in an optimal sphere packing problem. The criterion seeks to maximize the minimum distance between the new point and the existing samples. This adaptation is to ensure that the new point is not penalized by existing samples that would be ill positioned in the parameter space.

The ability to change the selection criterion is even more useful. With a prior knowledge on the sensitivity of the parameters to the quantity of interest (Saltelli et al., 2007), it is possible to bias the design. Considering a 2-dimensional space – as the example in Fig. 14 –, if the parameter x_2 is known to have a small impact, it might be more interesting to optimize the C^2-discrepancy on the parameter x_1. More complicated things can be performed if one wants to optimize a particular subprojection as in Joseph et al. (2015). This is referred to as *Maximum Projection Design*.

Last but not least, this method can be used to generate designs by mixing continuous and discrete variables. The star shape of the kernel does not forbid the presence of a new sample along a given axis, it lower its probability of being sampled up to a certain distance. In this case, a Gaussian kernel might be more appropriate in order to relax some constraint on the axes. Another option would be to modify the kernel to limit the point influence along the discrete axis.

The ability to play both with the kernel and the selection criteria is really powerful as it allows to manage most of the challenges in constrained optimization problems, use sensitivity information, and sample by means of following individual PDFs for each parameter.

6. Conclusion

This work proposes a new method to stochastically and iteratively sample a parameter space, referred to as KDOE method. This is a two-step process: (i) through a Kernel Density Estimation (KDE) some candidates are generated, then (ii) the best candidate is selected based on a criterion. This method does not take into account the physics of the problem of interest as adaptative strategies, but is purely iterative and case independent.

Compared to LHS and low discrepancy sequences, KDOE is totally iterative and stochastic. The space-filling properties of the new designs based on the C^2-discrepancy are assessed and show good behavior in high dimensions with small sample sizes. Moreover, it shows similar capabilities for numerical integration compared to classical methods. The KDOE method is versatile in the sense that it can be easily adapted to take into account constraints in the parameter space, both discrete and continuous parameters can be used and sensitivity indices of the parameters can be incorporated. This ability comes from the two-step process which can be independently tuned.

The quality of the design is of prime importance as it determines the quality of the analysis of the experiments. The proposed method provides an alternative to classical one-shot methods to generate initial designs and to continue existing ones. Its versatility and performance allow the analysis of expensive and high-dimensional cases to be within affordable budgets.

Acknowledgments

The authors acknowledge Dr Sergei Kucherenko from Imperial College London and Dr Jean-François Parmentier for helpful discussions and OpenTURNS' development team for their support. The financial support provided by all the CERFACS shareholders (AIRBUS Group, Cnes, EDF, Météo-France, ONERA, SAFRAN and TOTAL) is greatly appreciated and we thank them for enabling the achievement of such research activities.

References

Androulakis, E., Drosou, K., Koukouvinos, C., & Zhou, Y. D. (2016). Measures of uniformity in experimental designs: A selective overview. *Communications in Statistics – Theory and Methods, 45*(13), 3782–3806. doi:10.1080/03610926.2014.966843.

Baudin, M., Dutfoy, A., Iooss, B., & Popelin, A.-L. (2015). OpenTURNS: An industrial software for uncertainty quantification in simulation https://arxiv.org/pdf/1501.05242.pdf.

Beran, P., Stanford, B., & Schrock, C. (2017). Uncertainty quantification in aeroelasticity. *Annual Review Fluid Mechanics, 49*, 361–386.

Cavazzuti, M. (2013). Design of experiments. In *Optimization methods: From theory to design* (pp. 13–42). Berlin Heidelberg: Springer. doi:10.1007/978-3-642-31187-1_2.

Cha, S.-h. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences, 1*(4), 300–307. doi:10.1007/s00167-009-0884-z.

Chernatynskiy, A., Phillpot, S. R., & LeSar, R. (2013). Uncertainty quantification in multiscale simulation of materials: a prospective. *Annual Review Fluid Mechanics, 43*, 157–182.

Crombecq, K., Laermans, E., & Dhaene, T. (2011). Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. *European Journal of Operational Research, 214*(3), 683–696. doi:10.1016/j.ejor.2011.05.032.

Damblin, G., Couplet, M., & Iooss, B. (2013). Numerical studies of space-filling designs: optimization of Latin Hypercube Samples and subprojection properties. *Journal of Simulation, 7*(4), 276–289. doi:10.1057/jos.2013.16.

Esmaily, M., Jofre, L., Mani, A., & Iaccarino, G. (2018). A scalable geometric multigrid solver for nonsymmetric elliptic systems with application to variable-density flows. *Journal of Computational Physics, 357*, 142–158.

Fang, K.-T., Li, R. Z., & Sudjianto, A. (2006). *Design and modeling for computer experiments*. Chapman & Hall/CRC.

Franco, J., Vasseur, O., Corre, B., & Sergent, M. (2009). Minimum Spanning Tree: A new approach to assess the quality of the design of computer experiments. *Chemometrics and Intelligent Laboratory Systems, 97*(2), 164–169. doi:10.1016/j.chemolab.2009.03.011.

Frankel, A., & Iaccarino, G. (2017). Efficient control variates for uncertainty quantification of radiation transport. *Journal of Quantitative Spectroscopy & Radiative Transfer, 189*, 398–406.

Garud, S. S., Karimi, I. A., & Kraft, M. (2017). Design of computer experiments: A review. *Computers & Chemical Engineering, 106*(182), 71–95. doi:10.1016/j.compchemeng.2017.05.010.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97. doi:10.2307/2334940.

Ho, C. K. (2017). Advances in central receivers for concentrating solar applications. *Solar Energy, 152*, 38–56.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*, 1593–1623. http://jmlr.org/papers/v15/hoffman14a.html.

Jofre, L., Domino, S. P., & Iaccarino, G. (2018). A framework for characterizing structural uncertainty in large-eddy simulation closures. *Flow Turbulence and Combustion, 100*(2), 341–363.

Jofre, L., Domino, S. P., & Iaccarino, G. (2019). Eigensensitivity analysis of subgrid-scale stresses in large-eddy simulation of a turbulent axisymmetric jet. *International Journal of Heat and Fluid Flow, 7*, 314–335.

Jofre, L., Geraci, G., Fairbanks, H. R., Doostan, A., & Iaccarino, G. (2017). *Multi-fidelity uncertainty quantification of irradiated particle-laden turbulence* (pp. 21–34). Annual Research Briefs, Center for Turbulence Research, Stanford University.

Joseph, V. R., Gul, E., & Ba, S. (2015). Maximum projection designs for computer experiments. *Biometrika, 102*(2), 371–380. doi:10.1093/biomet/asv002.

Kucherenko, S., Albrecht, D., & Saltelli, A. (2015). Exploring multi-dimensional spaces: A comparison of latin hypercube and quasi Monte Carlo sampling techniques. In *Proceedings of the 8th IMACS seminar on Monte Carlo methods* (pp. 1–32). doi:10.1016/j.ress.2017.04.003.

Lekivetz, R., & Jones, B. (2015). Fast flexible space-filling designs for nonrectangular regions. *Quality and Reliability Engineering International, 31*(5), 829–837. doi:10.1002/qre.1640.

Liu, H., Ong, Y. S., & Cai, J. (2018). A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. *Structural and Multidisciplinary Optimization, 57*(1), 393–416. doi:10.1007/s00158-017-1739-8.

Masquelet, M., Yan, J., Dord, A., Laskowski, G., Shunn, L., Jofre, L., & Iaccarino, G. (2017). Uncertainty quantification in large eddy simulations of a rich-dome aviation gas turbine. In *Proceedings of the ASME turbo expo 2017* (pp. 1–11). gt2017-64835

Mckay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics, 21*(2), 239–245.

Najm, H. N. (2009). Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics. *Annual Review Fluid Mechanics, 41*, 35–52.

Owen, A. B. (1998). Scrambling Sobol' and Niederreiter–Xing points. *Journal of Complexity, 14*(4), 466–489. doi:10.1006/jcom.1998.0487.

Pronzato, L. (2017). Minimax and maximin space-filling designs : some properties and methods for construction. *Journal de la Société Française de Statistique, 158*(1), 7–36.

Rahmani, M., Geraci, G., Iaccarino, G., & Mani, A. (2018). Effects of particle polydispersity on radiative heat transfer in particle-laden turbulent flows. *International Journal of Multiphase Flow, 104*, 42–59.

Roache, P. J. (1997). Quantification of uncertainty in computational fluid dynamics. *Annual Review Fluid Mechanics, 29*, 123–160.

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science, 4*(4), 409–423. doi:10.1214/ss/1177012413.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications, 181*(2), 259–270. doi:10.1016/j.cpc.2009.09.018.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., … Tarantola, S. (2007). *Global sensitivity analysis. The primer*. John Wiley & Sons, Ltd. doi:10.1002/9780470725184.

Sheikholeslami, R., & Razavi, S. (2017). Progressive latin hypercube sampling: An efficient approach for robust sampling-based analysis of environmental models. *Environmental Modelling & Software, 93*, 109–126. doi:10.1016/j.envsoft.2017.03.010.

Sobol', I. M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics, 7*(4), 86–112. doi:10.1016/0041-5553(67)90144-9.