Properties and Implementation of Sequential Expansion of Latin Hypercube Sampling for Simulation Design

Crespi Alessandro, Gerosa Davide, Boschini Matteo

July, 2024

1 Introduction

Simulation design is a branch of Statistics that studies ways to build better simulations, intended as tools for enhancing the comprehension of phenomena, which have become widely used in mathematics, physics, economics, mechanics, and every scientific field as experimental tools for proofing theories, interpolating real sampled values, and, more generally, generating predictive models that try to explore uncharted traits, features, and peculiarities, perhaps intuitively or roundly developed in the early stages of the study of a specific problem.

Since the advent of hybrid mechanical-electrical programmable calculators such as IBM's machines, the first of their kind to be really useful in engineering, scientists have them involved in heavy computations for experiments. Following, in 1969, the Kennar and Stone's Computer Aided Design for Experiments (CADEX)[4] proposal for computer-driven experiments has led to spread up a broad variety of Computer-based simulation methods.

Eventually, computer-based simulations are a set of strategies that benefit from mathematical modelling techniques based on discrete known points placed in a limited parameter space, hereafter samples, and the computer programmability advantage has been used to design, shape and enhance a specific subset of samples that satisfy the desired properties, hereafter sample sets.

The general concept of computer simulation has been defined in the past few decades, it's based on the following key ideas: taking into account a desired behavior F the experimenters have an interest in, F has to be explored through its N real parameter space; the algorithm takes randomly samples from a standard N-dimensional hyperspace Ω and arranges the sample set for the simulation; afterwards, the simulation is carried out by evaluating F over the sample set and eventually producing a so-called surrogate model. The class of algorithms meant to implement this abstraction is commonly named as Sampling Methods.

A critical consideration when evaluating sampling methods is the trade-off between exploration and exploitation. An exploration-oriented sample set maximizes the simulation skill of seeking key features over the studied behavior. Exploration has been depicted as a model-free practice, so that it does not base its own actions on the model (behavior) evolution or any other on-site response. On the other hand, exploitation is an auxiliary mechanism that aims to better assist the simulation by deploying samples in strategic placements that prevent exploration from exceeding the prediction surrogate made upon a key region (such as overshooting an optima or mismatching a discontinuity for a steep slope).

The most iconic sample method for simulations is obviously the pseudo-random sampling, lightened of every other criteria, namely the Monte Carlo Sampling or MCS (*Metropolis et al. 1987* [2]) which it has been proposed as fundamental design of sampling methods. Quasi-Monte Carlo methods are a class of sampling algorithms based on Monte Carlo, indeed, but without a proper random drawing of sample points from the parameter space, instead, points are sequentially extracted in order to satisfy one or multiples criteria as best as the computational time required remains acceptable.

Many criteria have been theorized and tested; each of them has its own best scenario, which it'd be better to apply to. An updated, summarized list of the most remarkable ones is shown and commented in section **??** in table 1. As per the interest of this paper, the reader is going to learn about the space-filling class of criteria and the one-projection property (also known as non-collapsing property or projective property); the latter has been widely known because of the Latin Hypercube Sampling (LHS), topic of this research and explained afterwards. The space-filling design of points measures the quality of a sample set to be spread evenly across an hyperspace; the way space-filling is defined determines the final aspect of the sampling. On the other hand, a sample set, in P-dimensional space, admits the one-projection property if and only if each projection of every sample x on a specific axis does not overlap onto each other's interval I_j. These I_j intervals are well-known fixed-width slices of the limited volume of parameter space taken into examination (e.g., a volume $[0,1)^N$) and the number of intervals is equal to the number of samples taken. So, the non-collapsing property prevents samples to fall into a busy "private space" (which has been occupied by another sample). Furthermore, given that the number of intervals and the size of the sample set is equal, it does ensure there are no empty intervals across the parameter space.

The authors used to work with sampling methods for simulation design, more likely LHS designs, in Astrophysics related experiments, such as simulating black hole binary systems collapsing, which obviously requires a massive amount of computational time and many different parameters. The authors often experienced difficulties predicting how long it would take to run a full simulation given N known sampling points in a P-dimensional space. This situation forces them to reserve more machine time on a shared company supercomputer for experiments than they really need. In order to better spend the reserved machine time left after the execution of

the first run of sampling points, the authors designed an algorithm that adds up points to the previous LHS' sample set and another set of samples that has been drawn by the expansion algorithm in order to preserve the non-collapsing property of both sample sets together.

The authors of this paper propose a algorithm called "Expansion of an LHS" that, indeed, takes a already existing Latin Hypercube Sampling' sample set and propose a new set of points samples in the same parameter hyperspace which are suppose to maximize the non-collapsing property of the samples altogether. The original sample set is referred as "starting set" and the add-on samples as "expansion set". The whole sample set joined together by both is called "expanded simulation".

The paper is structured as follows: 1. Latin Hypercube Sampling [...]; 2. Expansion algorithm [...]; 3. .....; X. Conclusion .

2 LATIN HYPERCUBE SAMPLING

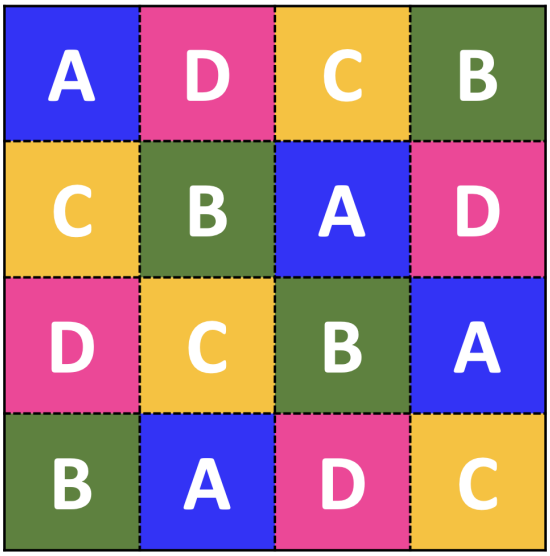
2.1 What is an LHS?

According to the Handbook of Computational Designs (2nd edition, 2006) [1], the first appearance in history of the "Latin Square" has to be attributed to the Korean mathematician Choi Seok-jeong, who described it, using modern terminologies, as a $N \times N$ matrix (the square) with N distinct symbols, appearing N times each but precisely once per type for each row and column. The suffix "Latin" has been inspired by the efforts that Leonhard Euler has put into this topic while defining a general theory for Latin Squares [5] and using Latin letters as symbols to fill the square up with. See Fig.1 for an example with 4 objects.

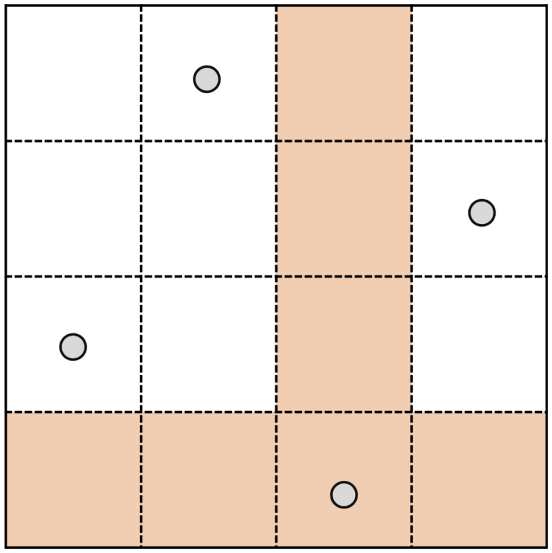
Per the interest of this paper, which doesn't aim to study how the symbols behave with each other and several algebraic combinatorics considerations, and for sake of clarity, the N number of distinct symbols factor is not taken into account anymore. Instead, it has been considered a sole symbol, hereafter "sample", which has to obey the "do not cross into each other's row and column", hereafter non-collapsing property. The sample must occur in the hyper-matrix exactly N times.

Generally, we can speak of "Hypercubes" by transposing the classic 2-dimensional matrix concept (depicted as a grid in Fig.1a) into a more complex multidimensional matrix where N samples are placed such that each one lies exactly once on each fiber (in literature, a "fiber" of a multidimensional matrix is the general term for a one-dimensional substructure in any dimension (mode) of a tensor. e.g. In a 2D matrix, the fibers of the first and second dimensions are respectively "rows" and "columns").

The modern approach of LHS is not supposed to place symbols in matrices but, instead, attempts to place N samples on hyperspaces that represent the examined parameter space of the interested model, where each axis is associated with a different parameter. The parameter



(a) A Latin 4x4 Square with 4 distinct symbols, both letters or colors either, arranged so that no letter occurs more than once in a row or a column



(b) The skeleton of (a) Latin Square highlighting only the D (or PINK) symbol positions, it's easily noticeable how the positions does not overlap onto each other's row and column both

Figure 1: The images were kindly taken [inserting sheikholeslami2017 throws an error?]

hyperspace has a limited span, which is the hyper-volume $[0,1)^P$, commonly used in literature, which represents, again, a hypercube. However, some texts use a different standard, and they let the parameter space take place in a $[-1,1)^P$ hypercube. Every parameter axis is sliced into smaller consecutive intervals of width $\frac{1}{N}$ that consequently depict the sub-region where each coordinate of the points is sampled randomly. In anticipation for further definitions, we provide a indexing system for intervals; in any dimension of the hypercube with N intervals, for each j from 0 to $N-1$, the j-th interval's boundaries are:

$$I_j = [\frac{j}{N}, \frac{j+1}{N}) \tag{1}$$

With no loss of generality, the examples and considerations designed by the authors of this paper assess the random distribution to be uniform in all intervals. The uniformly distributed samples can be transformed by associated transformation functions for any other distribution (e.g., a Gaussian distribution).

2.2 How to build a Latin Hypercube Sample Set

In this section it has been marked out mathematically the construction of an LHS sample. This description is widely used for introducing the topic on many textbooks and lectures and it takes inspiration from the work of X.Kong at al. [6]. Let $S = \{S_1, S_2, ..., S_N\}$ be the Latin Hypercube sample set with N number of samples, where for each S_i its cardinality is P number of dimensions. It is comfortable to use the matrix representation S_{ij} whereof rows are the i-th sample

and columns, instead, represents the projection of every sample on each j-th dimension. Then, we introduce the sorted index matrix $A = \{A_{ij} = j - 1\}$ as a tool for trace the intervals index (starting from zero), its purpose will be clearer soon.
Given an A index matrix, the preliminary design matrix S is given by:

$$S_{ij} = \frac{a_{ij} + u_{ij}}{N} \tag{2}$$

where u_{ij} is a uniform distributed variable U[0,1). This preliminary design has the peculiarity to have the samples always placed on the diagonal of the hypercube $[0,1)^P$. Now the index matrix come into use, typically shuffling the original A we attain $B = \{b_{ij}\}$ random permutation, which plugging it into eq. (2) describe the uniform random variable S_{ij} living inside the statistical bin identified by the interval index b_{ij}. We explicitly write down the uniform random variable S_{ij} involving eq. (1) boundaries for the ij-th interval as well:

$$S_{ij} \sim U[\frac{b_{ij}}{N}, \frac{b_{ij}+1}{N}) \tag{3}$$

2.3 Grade of a Sample Set

For the purpose of this research, the authors introduced a tool meant to measures how much a Monte Carlo sample set is close to a Latin Hypercube one, namely "grade of a sample set". This metric Eq.4 has been designed to reduce the sampling S of N elements to an index between 0 and 1 (percentage), when it's compared against the P-dimensional hypercube sectioned for the LHS that would have covered it if S was generated to fulfill the one-projection property.

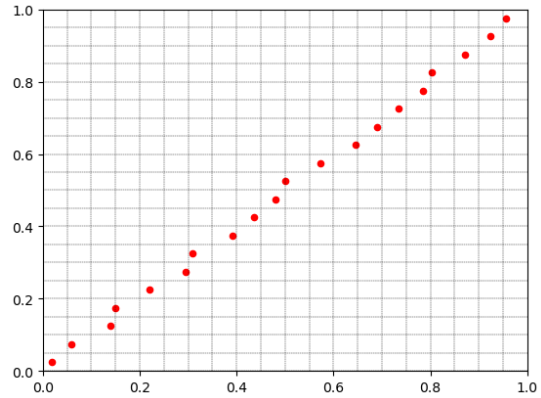
$$gr(S) = \frac{\sum_{j=1}^{P} \sum_{q=1}^{N} min(\sum_{i=1}^{N} \boldsymbol{1}_{[\frac{q-1}{N}, \frac{q}{N})}(S_{ij}), 1)}{P \cdot N} \tag{4}$$

where $\boldsymbol{1}$ is the indicator function (see Appendix 4.1), the variable q is another way to represent the sorted index matrix $A = \{a_{ij} = j\}$ in a more immediate comprehension. The formula has meant to compute the arithmetical average of the presence (with numerical value 1) of the projection of every sample in each interval for each dimension. The use of the *min* operator states that the presence of several samples' projections in a specific interval q doesn't weight up the whole term, which would be at most 1 even if multiple sample lies in q. Thus, it ignores the overlapping samples onto the same interval, hereafter only *overlaps*.

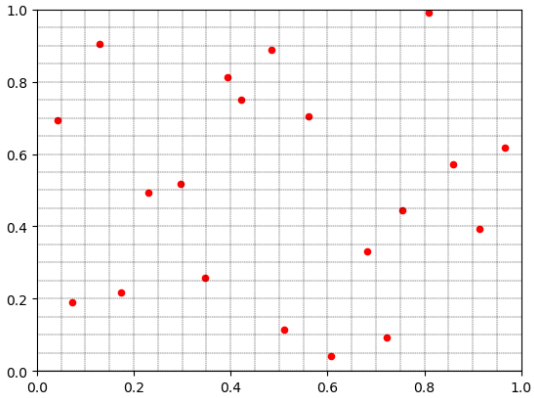
2.4 Additional properties (working on)

In this section, the reader will deal with the idea of adding properties that could improve the accuracy - or the quality in other terms - of the surrogate model produced when the simulation is consumed. A well known issue with LHS designs is well depicted in Fig.2a. It's obvious that

it must be considered a poor simulation design for the most of the experiments. In fact, every really distributed implementation of LHS adds some peculiar properties to the design. We can highlight two superclasses of properties which LHS may be joined with: model-free properties and model-based properties.



(a) A poor Latin Hypercube sampling in 2 dimension despite it fulfill the one-projection rule. This experiment shows how simple LHS can be insufficient to achieve a good simulation.



(b) This sampling has been generated using scipy LHSSampler utility class. The sample set bears random permutations of coordinates to lower the centered discrepancy. [3]

Figure 2

2.4.1 Model-free and model-based simulation designs

As the names may suggest, the latter implies the experimental design to involve some the peculiarities of the function to evaluate e.g. the shape that is expected, the initial or boundaries conditions, any well known critical region. Given that this approach is knowledge-driven, ”What do we already know about?”, can knock down pretty much the computational time required for the whole experiment with eventually high accuracy, but it can be equally easy that results may be deformed, based on assumptions that may prove wrong later, or make hard any further effort from other scientists to retake the experiment and confirm the results if the assumptions are not completely clear. On the other hand, the model-free is the formal way to depict a fully independent sampling of the parameter space from the experiment which it has been designed for. Because of not many criteria have been left, all of the properties in this class are based on inter-element relationships, which convey that for each of them has a heuristic meant to quantify how much each sample is well-placed among the others. The shorthand for such a class of criteria is *space-filling properties*.

Beside the many interesting considerations and suggestions that the model-based class of criteria has to offer to mathematicians, the space-filling properties has highly excited the experimenter community through the decades, producing creative and curious features that the samples could experience among each others. A very short list of the most representative ones

is shown in Tab.1. The most simple one is the [L2? phi di P?] criterion where the assumed metric is the Euler distance between each point in the hyperspace. The non-trivial issue comes along with the computational effort necessary to run through the search tree generated by the maximin (or minimax?) algorithm.

By the way, the characteristic non-collapsing property of the LHS itself is actually a space-filling property. It ensures that, for example, the average distance between the consecutive projections of a group of 3 samples on the same axis is at least 1/2 distance units and at most 3/2 d.u.

Authors	**Year**	**Algorithm**	**Criteria**
Audze and Eglajs	1977	Coordinates exchange	Potential energy

Table 1: Riempio la tabella piano piano che leggo i paper. Ho visto che molte ricerche su LHS usano dare una cronologia sull'utilizzo dei criteri utilizzati throught history

3 Expansion of a Latin Hypercube Sampling

Given the problem outlined, it's necessary an heuristic which express how much the expanded issue is far away from a perfect expansion, given the starting state and the expansion magnitude desired. The Eq.4 metric is not sufficient to achieve this, then it has been proposed the variant Eq.5. The "expanded grade" has been designed to reduce the sample set S of N elements to an percentage index when- here it differs from the previous definition of grade - the sample set is compared against a P-dimensional hypercube sectioned for an hypothetical LHS set of $N+M$ samples. Here it follows:

$$gr(S,M) := \frac{\sum_{j=1}^{P}\sum_{q=1}^{N+M} min(\sum_{i=1}^{N} \mathbf{1}_{[\frac{q-1}{N+M},\frac{q}{N+M})}(S_{ij}),1)}{P\cdot(N+M)} \tag{5}$$

4 APPENDIX

4.1 Indicator function

The indicator function $\mathbf{1}$ of a set **A** indicates whether the input belongs to A or not, specifically:

$$\mathbf{1}_A(x) := \begin{cases} 1 & if\, x \in A \\ 0 & if\, x \notin A \end{cases} \tag{6}$$

As in the matter of sectioning a space into continuous intervals in the shape of [a, b), it is useful to redefine the indicator function as an operation that occurs with the boundaries of A using the Heaviside step function which does not involve set operators but only logical ones.

It's important to remark that it doesn't matter what happens precisely on the boundaries. The Heaviside function is defined:

$$H(x) := \begin{cases} 1 & if\, x \geq 0 \\ 0 & if\, x < 0 \end{cases} \tag{7}$$

So the indicator function can be also produced:

$$\mathbf{1}_{[a,b)}(x) = H(x-a) \cdot H(b-x) \tag{8}$$

DRAFTBOX

Let's introduce the Latin Hypercube Sampling, a popular experimental sampling method, model-free.... In this paper we'll consider and discuss the scipy's LHS implementation for python $\geq$ v3 as it does implement the expected non-collapsing property and an optimize sequential space-filling design based on the Minimax rule [6] (?)

In this paper, authors have used different notable sampling methods for comparison purposes, here follows each of them along with a brief not of they prominent characteristics:
• Sobol' sequence as low-discrepancy sequence, when the percentage of a sequence's points that fall into an arbitrary set Z is nearly proportionate to the measure of Z, the sequence so-called low-discrepancy;
• ...

- each sample in a P-dimensional hypercube identifies P sub-hyperplanes (in $(P-1)$ dimensions) that pass through it and are parallel to one of the axes

- Talk about problems with space-filling, search trees ecc.

- why Granularity [1] ? Discrepancy

- Why has this paper been proposed ? The Least Computational Time Problem

- PLHS [2] ? no, I suppose

- LHS and the problem of the so-called "expansion"

- In this paper the authors propose the implementation of the expansion algorithm for the LHS that can maximize space-filling properties and the projective property.

+ [remember to say that we are generating the sample sets using python scipy lhs implmenetation)

References

[1] Charles J. Colbourn. *Handbook of Combinatorial Designs*. 2nd. CRC Press, 2006.

[2] N. Metropolis. "The beginning of the Monte Carlo Method". In: *Los Alamos Science* Special Issue (1987).

[3] *Python 3.x - Scipy Documentation*. URL: `docs.scipy.org/doc/scipy/reference/generated/scipy.stats.qmc.LatinHypercube.html`.

[4] L. A. Stone R. W. Kennard. "Computer Aided Design of Experiments". In: *Technometrics* 11.1 (1969).

[5] Wikipedia. *Latin Square*. URL: `en.wikipedia.org/wiki/Latin_square`.

[6] Mingyao Ai Xiangshun Kong. "Design for sequential follow-up experiments in computer emulations". In: ().