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

Crespi Alessandro, Gerosa Davide, Boschini Matteo

July, 2024

Contents

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin luctus finibus euismod. Quisque purus mauris, mollis sed tellus quis, congue hendrerit eros. Aliquam tempus suscipit risus non viverra. Ut pharetra mollis ante, sit amet vestibulum augue laoreet eget. Nunc tincidunt ex sit amet rutrum euismod. Maecenas feugiat, mi aliquam semper rutrum, purus justo imperdiet massa, sed feugiat leo libero sit amet sapien. Mauris sit amet sem rhoncus, hendrerit nunc sed, sagittis felis.

1 *Introduction*

BREAK INTO 2 SENTENCES

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)[7] proposal for computer-driven experiments has led to spread up a broad variety of Computer-based simulation methods.

USED TO PERFORM

CAN YOU EXPLAIN THIS BETTER? ARE YOU DESCRIBING SIMULATIONS OR SIMULATION DESIGN? MAYBE YOU HAVE A DIFFERENT DEFINITION BUT, IN MY MIND, A SIMULATION AIMS AT MIMICKING A PROCESS TO STUDY ITS EVOLUTION AND EVENTUALLY PRE I AN OUTCOME. IT'S NOT MANDATORY TO HAVE A RANDOM SAMPLING STAGE. WHAT YOU ARE DESCRIBING IT'S A MONTE CARLO SIMULATION (IF SO SAY IT)

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.

MODELING

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.

WHAT DO YOU MEAN BY BEHAVIOR?

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

COVERAGE OVER THE ENTIRE HYPERSPACE OF INTEREST

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~~d~~ 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)* [4]) 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 multiple 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 section 2.4 in section 2.4.1. 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

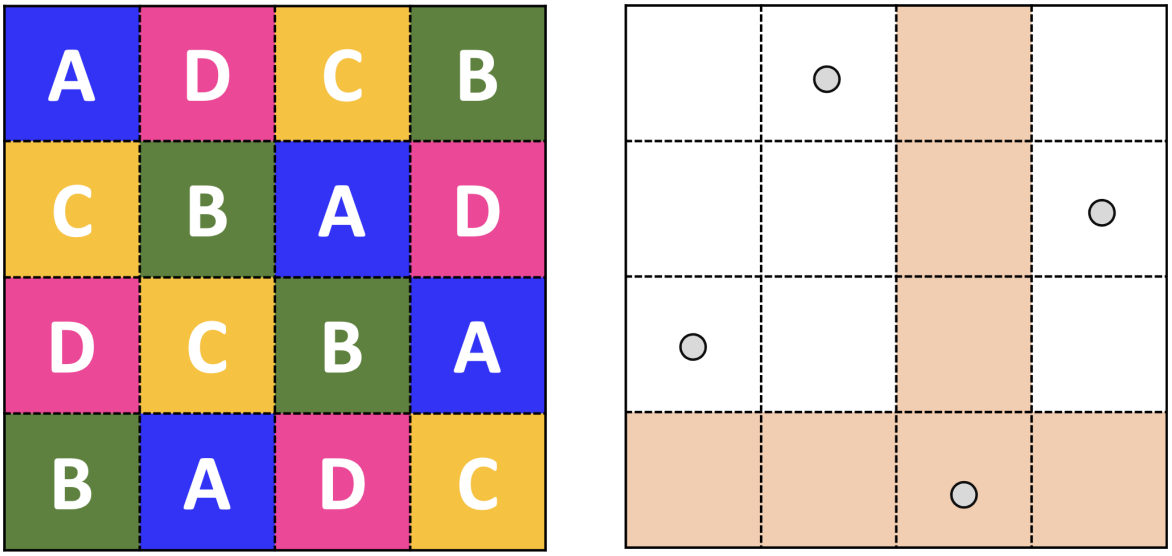
samples in the same parameter hyperspace which are suppose to maximize the non-collapsing property of the samples altogether. The original sample set is referred as ”starting set” and the add-on samples as ”expansion set”. The whole sample set joined together by both is called ”expanded ~~simulation~~”.

The paper is structured as follows: section 2 yields a Latin Hypercube Sampling brief history and formal definition; in section 3 is shown the research results and discussion of the expansion task issued; section 5 [...].

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 [9] and using Latin letters as symbols to fill the square up with. See Fig.1 for an example with 4 objects.



(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 from *Sheikholeslami and Razavi et al. (2017)* [8]

~~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 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:

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

for further usage, the right term $\frac{j+1}{N}$ of the interval was called *frontier of the j-th interval* which is shared with the left hand term of I_{j+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.[3]. 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

AGAIN I UNDERSTAND THE MEANING, BUT THIS IS NOT CLEAR. IN THE PREVIOUS PAGE YOU DEFINED A USING j (WHICH FOR THE PREVIOUS LINE $\in [0,P)$). SO THE SAMPLES ARE NOT ON THE DIAGONAL OF THE HYPERCUBE $[0,1)^P$. MOREOVER, IF YOU ADD A $u_{ij} \in U[0,1)$ THEY WON'T FOR SURE. AND YOU DON'T WANT SAMPLE ON THE DIAGONAL, BUT ON THE DIAGONAL OF INTERVALS, THEN YOU SHUFFLE IT → SAY IT MATHEMATICALLY ONCE AND FOR ALL

(starting from zero), its purpose will be clearer soon.

PLEASE REWRITE THIS

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:

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

AT SOME POINT CAN YOU WRITE / SAY THAT A SAMPLE SET IS CHARACTERIZED BY N AND P
=> $gr(S)$ MEANS $gr(S(N,P))$

2.3 Grade of a Sample Set

For the purpose of this research, the authors introduced a tool ~~meant~~ to measure~~s~~ 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~~ IS designed to ~~reduce the sampling~~ ASSIGN TO A SAMPLE SET S of N elements ~~to~~ IN P DIMENSIONS 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.

WHAT?? PLEASE SIMPLIFY AND SAY WHAT HAPPENS AT 1

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

CHOOSE THE INTERVALS FOR i, j ie

where $\mathbf{1}$ is the indicator function (see Appendix 6.1), the variable q is another way to represent the sorted index matrix $A = \{a_{ij} = j\}$ for 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 samples lie in q. Hence, the grade formula ignores the overlapping samples onto the same interval, hereafter only *overlaps*.

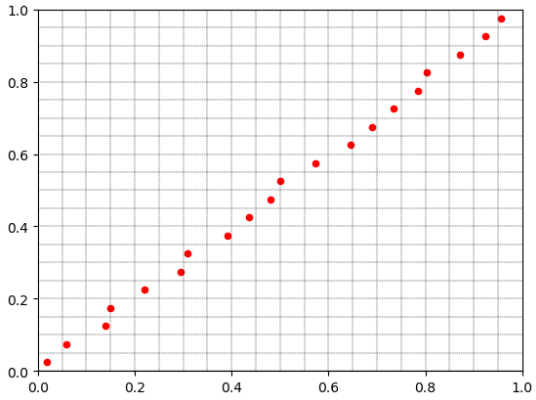
YOU SHOULD SAY THAT $gr(S) = 1 \iff S \equiv$ LHS

2.4 Additional properties (Work in progress)

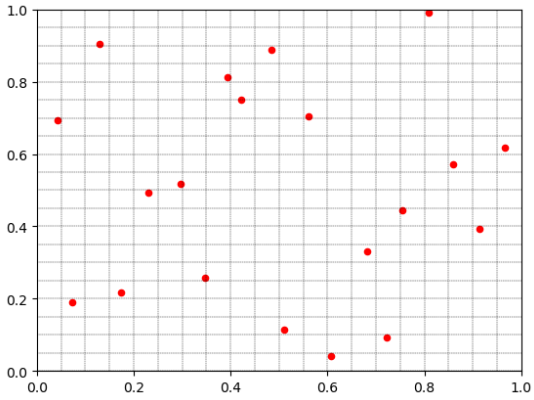
ARE YOU SURE THAT ACCURACY IS THE RIGHT WORD?

In this section, ~~the reader will deal with the idea of adding~~ I/WE/THE AUTHORS PRESENT SOME ADDITIONAL properties that could improve the accuracy - or the quality in other terms - of the surrogate model produced ~~when~~ BY 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~~ USUALLY, 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. [6]

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.**??**. 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*

The LHS paradigm's ability to implement several criteria over a rigorous grid, which forms the basis for the sequential creation of samples, has made it frequently utilized in engineering environments for surrogate manufacture of complex systems - see also section 2.4. For example, the LHS is used in hyperparameter tuning/optimization of Machine Learning models (*Koch et al. et al. (2018)* [2]), environmental and water system analysis (*Sheikholeslami and Razavi et al. (2017)* [8]), structural reliability analysis (*Olsson, Sandberg, and Dahlblom et al. (2003)* [5]) and many others.

As happens with every helpful technology, researchers and scientists have attempted to improve the LHS capability. As previously stated, add criterions first. Then, by rethinking the foundations of the algorithm. The reader has to acquaint that our Monte Carlo pseudo-random distribution, namely the Latin Hypercube technique, widely implemented, was labeled in literature as a *one-stage* algorithm. The fact that all samples are distributed and assessed "on the first run" bestows this adjective. It's relevant to clear out that the actual creation and propagation of points is not properly implemented as the resulting of a lone drawing of Eq.3 random variables, but instead it's a sequential drawing of points - might be a parallel drawing of several ones for optimization reasons, but it hasn't been a topic of this research - which the newest one has to be pulled out from a pool of optimal candidates in order to improve the criteria applied, given a desired N_1 number of samples, such as Maximin space-filling distance. The feature "one-stage" highlights that it could be possible to have many more "stages" of the algorithm. It's a game of perspectives: inside a current stage, the policy for drawing a fixed number of data points is aimed at pulling out point given the other ones; instead, a multistage policy is related to a more evolutionary approach, aiming to enhance the sample set stage by stage. Manipulate an already instantiated LHS may sounds difficult because LHS hasn't been

designed ~~for adding~~ points - or, at least, ~~it's~~ not supposed to consider it. Indeed, it is reasonable to assess that by adding points, one by one, over a full grid of N_1 intervals of an LHS, and then reshaping such a grid in $N_2 > N_1$ intervals, ~~should soon~~ lead to collisions (overlaps), which represents an issue ~~that scientists had to deal with which suggest noticeable solutions - see section 3.1 for more~~.

The process that embodies the evolution from a precursor LHS to his next-stage has been called *Expansion* by the authors; the resultant of the expansion process is the so-called *expanded set*. Please refer to **??** for the visual explanation of expansion.

Along this section, the authors used to refer several times to the expansion process without precising how the new samples are going to be placed, the actual process is described in section **??**.

3.1 The task of multistaging sampling

Concerns regarding the consistency of the one-projection property, which is valid for one-staged setups, ~~are raised by~~ the multistage approach. Fig.3 depicts an experiment carried out upon scipy's X with LHS distribution configured with $N_1 = 10$ samples displayed over its appropriate grid of N_1 intervals per dimension. The experiment consists in evaluating the behavior of the fixed X while the grid is "growing" - intended as creating a brand new grid with a greater number of intervals - one by one for three times. Light grey-colored rows and columns are vacant, meaning that no projections are located there; rows and columns marked in red indicate the locations of overlaps. After the first add-on (Fig.3a), the grid shows two overlaps in row #3 and in column number #3 as well; the next stage (Fig.3b) shows overlaps too, four in total, more than before. In the end, Fig.3c has no overlaps at all. ~~The read might have had an hunch about~~ the distribution of occurrences of overlaps ~~if kept growing~~ the grid, ~~such a distribution should be based~~ on the initial sample set. The authors discuss about this topic later in [section does not exist yet].

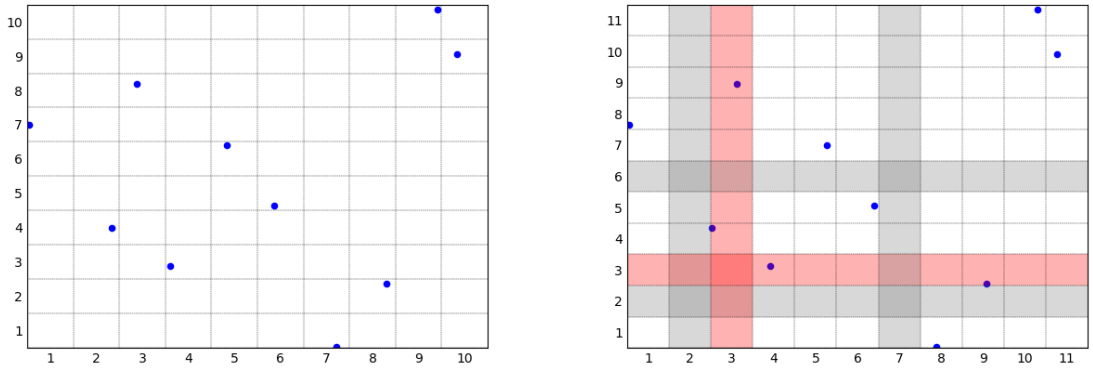
Regarding the LHS directive, the multistage experiment proposed shows that growing the interval grid could lead to a general downgrade the one-project (Fig.3b and Fig.3c) or ~~could~~ give a new opportunity to fill the new empty space with a set of brand new samples without compromise the whole set optimality.

The resulting sample set Z of the expansion process which takes place upon a well-formed LHS of N elements - hereby *perfect LHS* - over a growing $M+N$ grid is called *perfect expansion* if and only if Z reaches the *upper expanded grade limit* Eq.5.

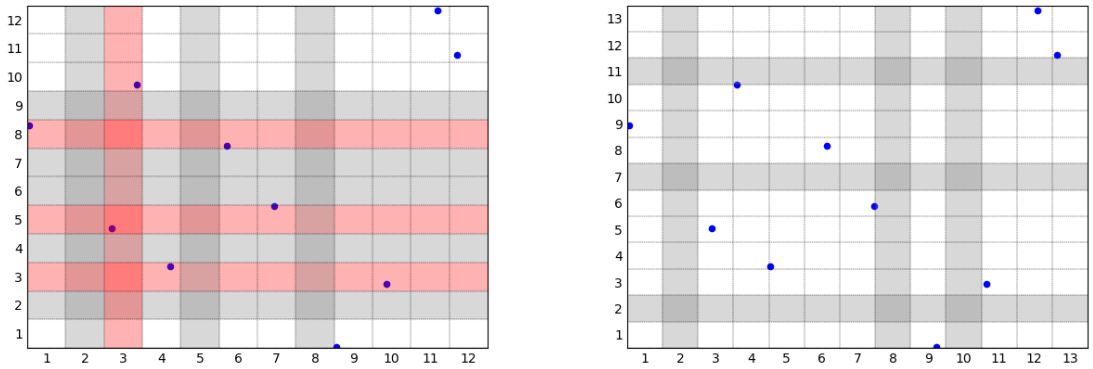
3.2 Grade of an Expansion

A precise heuristic that expresses how near the current expansion is to a perfect expansion, given the desired expansion magnitude M and the beginning state, is necessary for developing a new

YOU CAN AVOID SINGLE CAPTION AND SPECIFY N IN A BOX INSIDE THE PLOT OR AS A TITLE => YOU SAVE SPACE
YOU CAN MAKE PLOTS BIGGER AND SQUARED



(a) N = 10 (b) N = 11



(c) N = 12 (d) N = 13

Figure 3: To demonstrate the behavior of the grid, a S LHS sample set is first created in a ~~squared space~~ ~~($P = 2$ dimensional hypercube)~~ and projected over the grid as it grows one by one. Rows and columns marked in red indicate the locations of overlaps, while light grey-colored rows and columns are vacant. In (a) ~~a fixed~~ $N = 10$ LHS samples are ~~generated and~~ displayed over a ~~square~~ GRID of N interval per axis; (b) shows the first step of the growing grid: two overlaps occur in THE ~~interval~~ horizontal ~~#3~~ and vertical ~~#3~~ INTERVALS, two intervals per dimension are vacant one of which derives from the expansion vacancy and the other one is caused by the overlap; (c) samples displayed over $N + 2$ grid: on the vertical axis, 3 overlaps and 2 (from the expansion) + 3 (from the overlaps) vacancies; (d) the ~~grown~~ $N + 3$ grid have no collisions, then S is a perfect expansion ~~on~~ WITH $M = 3$ expansion magnitude [la sovrapposizione dei colori tra le evidenziature è fuorivante?]
FOR ME IT'S FINE, BUT YOU CAN MODIFY THE COLORS IF YOU DON'T LIKE IT

sample set. Since the metric in Eq.4 is no longer sufficient, the variation *expanded grade* has been offered. The purpose of the *expanded grade* is to convert the sample set S of N elements into a ~~percentage~~ index when the sample set is compared against a P-dimensional hypercube grid of $N+M$ intervals per axis - this is where it deviates from Eq.4. Here it follows:

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

Each I interval contributes to Eq.5 value with a share of:

$$share_I = \begin{cases} \frac{1}{P\cdot(N+M)} & if\ any\ x \in I \\ 0 & otherwise \end{cases} \tag{6}$$

which makes Eq.5 a discrete quantity that ranges from $\frac{1}{N+M}$, if every sample of a non-empty set lies in one single interval per axis, to 1, which represents the perfect LHS expansion ~~and it's a multiple of Eq.6~~.

~~Considering the~~ $N+M$ space grid of a perfect expansion ~~from~~ S of N samples over ~~M new intervals, given that in such a space would be covered by~~ N samples in N distinct intervals per dimension by definition of LHS, a number M of intervals per dimension ~~is left~~ empty. Then, the shares of the empty intervals are none, against the positive shares given by the N others. From this statement, we can provide an upper limit for the expanded grade by starting from 1 - the maximum possible index of a perfect LHS sample set - minus the total weight lost during the growing of the grid, which corresponds to the total number of *voids* times Eq.6. The total number of *voids* is given by $M\cdot P$ because they are evenly spread across each dimension. So, Z is a *perfect expansion* of S ~~on~~ M ~~new samples~~ if and only if:

$$gr_{max}(S,M) = 1 - \frac{M}{N+M} \tag{7}$$

3.3 The expansion process

The expansion task was initially handed at the very beginning of section 3.1 as a evolutionary process which augments the S starting sample set to a next-stage state with increased number of elements M, placed as better as it can to maximize criteria. We ~~do also said~~ at section 3.2 that an expansion may demote the non-collapsing property of the resultant expanded set Z, which can be measured with the metric Eq.5. In this section the authors propose how to place the new samples to achieve maximum "LHS-ness" over any M expansion magnitude needed and introduce its potentialities for future researches - ~~an interested reader should see section 5~~.

In first place, the proposal is delivered by studying the basic case of a perfect expansion in section 3.3.1, then extend to the general case with any non-perfect expansion outcome.

3.3.1 State case - Perfect Expansion

Referring to section 3.2, ~~from~~ an initial instanced sample set S of N elements, ~~the experiments~~ can be perfect-expanded if and only if S after the M-th step growing of the P-hypercube grid has max grade Eq.7. Furthermore, Fig.3d, which is a perfect expansion of S with $M = 3$, well depicts the situation which the experimenters would expect to encounter before setting down the expansion set E. We notice the best candidate extent is likely the empty intervals (in light grey) because they do not interfere with other well-formed intervals. ~~It's~~ also imperative that the ~~scattered~~ new ~~points don't stupidly~~ overlap ~~onto~~ each other over the vacant space, hereby *vacancies* or *voids* for sake of simplicity.

First of all, it's necessary to trace each void's index. Here it comes into use again the ~~(permuted)~~ index matrix, previously used in ~~section~~ 2.2 to scatter the samples across the hyper parameter space. ~~By the way~~, the matrix of voids $P \times M$ is composed by the row vectors:

$$\mathbf{V}_j = \left(z : \nexists\, x \in S_{ij}\ s.t.\ x \in \left[\frac{z}{N+M}, \frac{z+1}{N+M} \right) \right) \quad , \quad \forall j = 1..,P \tag{8}$$

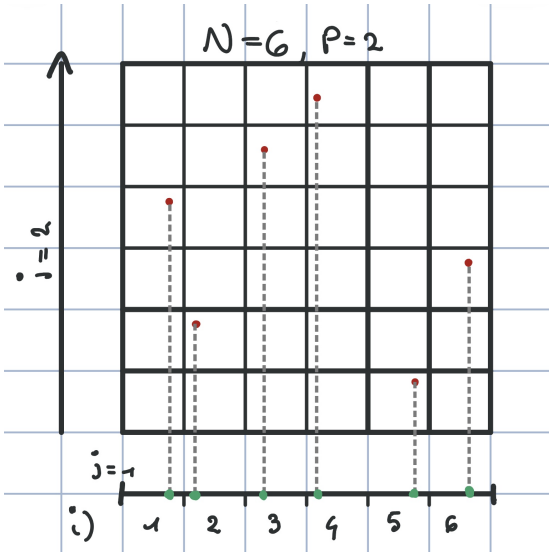
which should be column-permuted along each row to prevent Fig.2a diagonalized situation. Then, the expansion set E collects all the newly generated samples which are a distribution likewise Eq.3 but ~~based~~ using V_{ji}.

In the end, the initial S set is concatenated with expansion set E, originated from the voids of the $N+M$ growing grid, in a perfect expansion, that guarantees the fulfilling of the non-collapsing property (grade = 1).

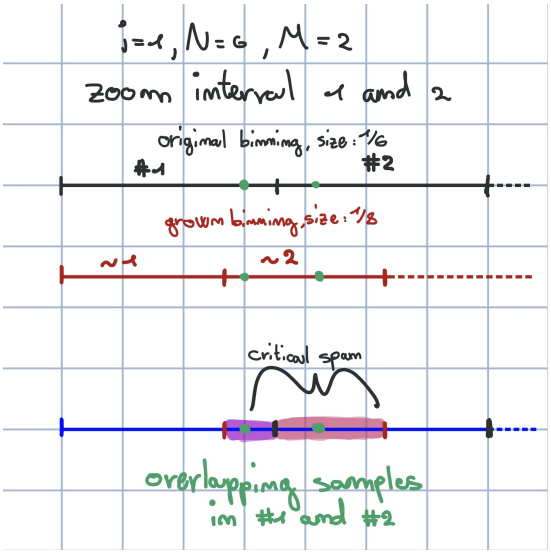
A perfect expansion is a intuitively rare event: ~~no~~ samples ~~have to~~ overlap ~~with anyone else, which can be excluded~~ if two sample projections are spaced more than $\frac{1}{N+M}$ apart (new size of the intervals). Furthermore, for any M close to N, a *critical span* exists across the right border of the i-th interval - namely *i-th frontier* (see right after Eq.1 for definition) - where two samples might be placed in, during the first stage of LHS sampling. An example of how can a couple of collapsing samples can occur is shown in Fig.4. The critical span is identified by the new interval created across two old ones, which the first sample is in Fig.4b has been generated in the critical area ~ 2 intersected with #1 original interval. On its hand, the second sample lies upon the intersection between ~ 2 and #2 original subspace.

3.3.2 State case - General Expansion

Because of Eq.5 complexity, it's an harsh to predict which samples are going to overlap after the re-binning of the hypercube and/or which M will produce a perfect expansion - and so a perfect expanded LHS, see section 3.3.1. During the upscale of S ~~sample set~~ from N to $N+M$ number of elements, it comes along with an O number of overlaps. In ~~section~~ 3.2 ~~was noted~~ the relation between ~~the number of vacancies~~ row vector $\mathbf{V}_j$, ~~the magnitude of the expansion~~ M,



(a) An LHS of $N = 6$ samples expanded for $M = 2$. on the horizontal axis are projected the horizontal component of all samples. [SKETCH - ne farò in digitale di migliori]



(b) Zooming in #1 and #2 interval, it's show how both shares a critical area (on the bottom) whereof each one may have been distributed in it. [SKETCH - ne farò in digitale di migliori]

Figure 4

and the amount of collisions in the j-th dimension. Specifically:

$$\| \mathbf{V}_j \| = O_j + M \tag{9}$$

↳ YOU HAVEN'T DEFINED O PER DIMENSION

whereof the overlaps count equals to zero, the expansion has been perfect (see section 3.1). In a general case of expansion, the overlaps amount is most likely not equal to zero for some j-th dimensions.

DEPENDS ON HOW YOU DEFINE V

The case study implies that the creation of E expansion set is not trivial anymore because of the irregularity of the vacancies set. By stressing what Eq.9 states, V would probably be no matrix at all. The number of $\mathbf{V}_j$ interval indexes would likely be more than M sample's projections to commit. The expansion algorithm has to pick up a reasonable subset of M void entries, and thus to discard an amount of intervals equal to the number of overlaps O_j. Therefore, given the sub-hyperspace settled by the joined selected voids, in order to plot an M amount of new samples, it will pick up an $P \times M$ submatrix (that mimics Eq.8) of V set. The submatrix should be handled being aware that it would effect the samples layout which may better improve another coherent criteria chosen (such as low-discrepancy or Maximin space-filling).

JUST SAY THAT YOU AN CHOOSE THE "OPTIMAL" SUBH THAT MAXIMIZES CRITERIA

The selection process of vacancies from an irregular V voids set is described by a function $\sigma : N \times P,\ M \to P \times M$, namely *perfectify* or *vacancy reduction* for the matter of giving names to anything. -_-

In this section, σ reduce function trivially picks up M intervals randomly per dimension and

build up a permuted Eq.8 vacancies matrix, which will be plugged into Eq.2 to produce an ~~E~~ expansion set.

In other words, σ ~~criterion~~ extracts from each j-th axis vacancies set $\mathbf{V}_j$ a fixed M number of elements which are going to compose the sub-hyperspace where M samples will be placed ~~in~~, using at least the non-collapsing property. The ~~reduce~~ function σ discards O_j number of intervals (Eq.9), then creating an amount of voids of the same quantity. Hence, the number of overlaps O_j determines the number of void intervals after the expansion is consumed.

In this paper, the term *quasi-LHS* refers to a non full-graded ~~(grade equals to 1) n-th stage of~~ expansion ~~descended from~~ a proper LHS.

3.3.3 The eLHS algorithm

The LHS expansion algorithm, namely *eLHS*, push a starting Latin Hypercube S to the next stage $Z = eLHS(S, M)$ ~~which best~~ maximize ~~(maintain)~~ the non-collapsing property, along with other eventual criterions.

1. Instance a V vacancies set of P tuples - which may have different lengths (Eq.8) because every dimension has an arbitrary number of voids (Eq.9). The list of all indexes of the $N+M$ grid is filtered accordingly with Eq.8.
   Visually, the algorithm re-bins the original S grid ~~(see Fig.5b and Fig.5c)~~ to archive $N+M$ ~~subdivisions~~. By plotting it against S, the reader can visualize where samples ~~collide (overlap)~~ and where there are ~~none (void)~~.

2. Reduce V vancancies set to a suitable indexes matrix $V' \in Matrix(P, M)$ by extracting from each $\mathbf{V}_j$ tuple M elements - which are going to compose the expansion binning grid - using σ reduction criteria (see section 3.3.1). If there are no overlaps (meaning S has maximum expanded grade Eq.7) then V is implicitly equal to matrix V', so no reduction criterion is required to be applied.

3. Generate new points over the sub-hyperspace outlined by the permuted V' indexes matrix. Currently, Scipy ~~hasn't implement~~ instancing of a LHS over a discontinuous space yet. To archive that, the algorithm sequentially draws M samples from an optimal sample pool which have been determined by the others criterion applied (such as maximin space-filling, see section 2.4) where every samples are placed according with V'

4 *Experiments*

...

5 Conclusions

...

6 APPENDIX

6.1 Indicator function

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

$$\mathbf{1}_A(x) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases} \tag{10}$$

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:

$$H(x) := \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{11}$$

So the indicator function can be also produced:

$$\mathbf{1}_{[a,b)}(x) = H(x-a) \cdot H(b-x) \tag{12}$$

References

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

[2] Patrick Koch et al. "Autotune: A Derivative-free Optimization Framework for Hyperparameter Tuning". In: KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, pp. 443–452. ISBN: 9781450355520. DOI: 10.1145/3219819.3219837. URL: https://doi.org/10.1145/3219819.3219837.

[3] Xiangshun Kong, Mingyao Ai, and Kwok Leung Tsui. "Design for Sequential Follow-Up Experiments in Computer Emulations". In: *Technometrics* 60.1 (Apr. 2017), pp. 61–69. DOI: 10.1080/00401706.2016.1258010. URL: https://doi.org/10.1080/00401706.2016.1258010.

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

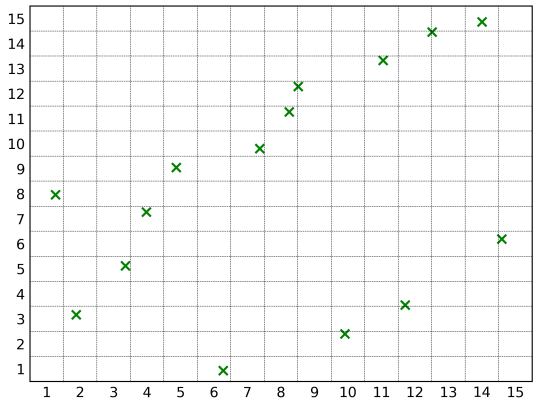
[5] A. Olsson, G. Sandberg, and O. Dahlblom. “On Latin hypercube sampling for structural reliability analysis”. In: *Structural Safety* 25.1 (2003), pp. 47–68. ISSN: 0167-4730. DOI: https://doi.org/10.1016/S0167-4730(02)00039-5. URL: https://www.sciencedirect.com/science/article/pii/S0167473002000395.

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

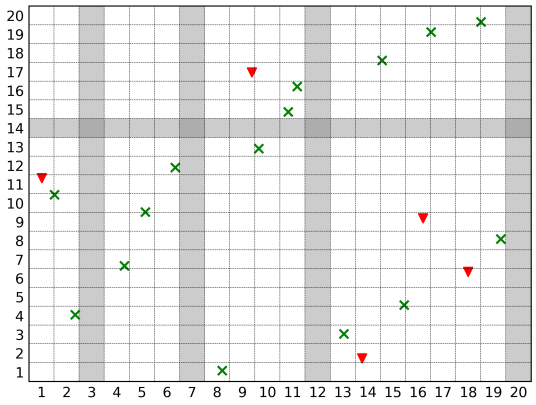
[7] L. A. Stone R. W. Kennard. “Computer Aided Design of Experiments”. In: *Technometrics* 11.1 (1969).

[8] Razi Sheikholeslami and Saman Razavi. “Progressive Latin Hypercube Sampling: An efficient approach for robust sampling-based analysis of environmental models”. In: *Environmental Modelling & Software* 93 (2017), pp. 109–126. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2017.03.010. URL: https://www.sciencedirect.com/science/article/pii/S1364815216305096.

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

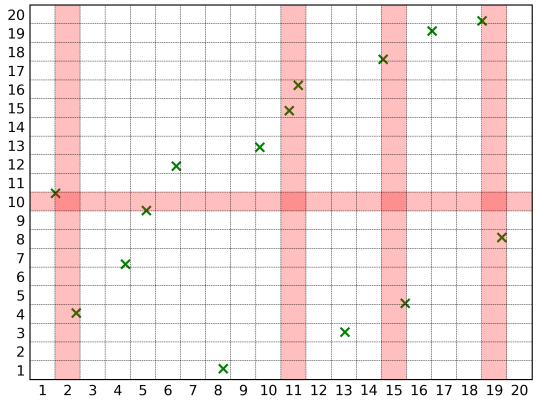


(a) First-stage LHS of $N = 15$ samples in $P = 2$ generated with scipy's qmc library.

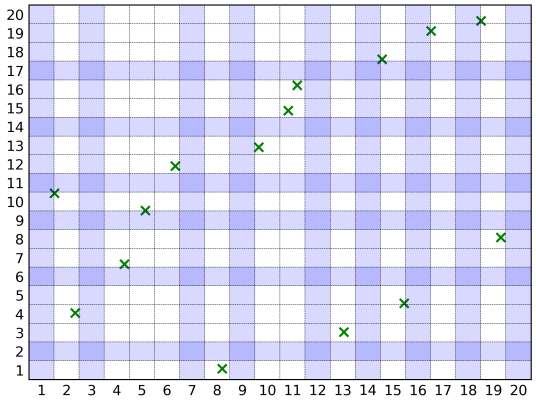


(b) Expansion of (a)'s LHS with section 3.3.3 eLHS algorithm given $M = 5$ new samples. Note that the light grey marked intervals are empty and, according to section 3.3.2, they are related with the overlaps distribution in (c)

Above is shown a two-staged quasi-LHS. (a) is the first original LHS and (b) is next stage expansion of it.



(c) Red intervals have two sample's projections in it and break the non-collapsing property.



(d) Blue intervals are empty. They represents the best candidate spots to place new LHS samples. Every interval all together has been referred as the sub-hyperspace of vacancies.

Re-binned (a)'s grids with $N + M$ intervals and plotted against the starting LHS.

Figure 5