4 Datasets

It is a capital mistake to theorize
before one has data.

Sherlock Holmes (Conan Doyle)

In practice when one desires to use a Machine Learning methodology to model a certain function, one is faced with two unavoidable issues. The issue considered hereafter is the construction of a dataset (the second issue discussed in the next Chapter is the training of the parameters).
It is possible to consider that datasets are generated in three different ways.

- **The dataset is given.** This is the standard situation in data science and statistical studies. There exists banks[1] of datasets where one finds data of any kind, exoplanets, medical surveys, sports data, financial studies, environment data, ...
- **The dataset is designed for analytical or algorithmic studies.** For example one can considers the Neural Networks representations of polynomials functions illustrated in Figures 2.7 and 2.5 and design the corresponding dataset. For the function $f^{\text{obj}}(x) = x^2$ a dataset obtained by uniform sampling over the unit interval could be

  $$\mathcal{D} = \left\{ \left(x_i, f^{\text{obj}}(x_i) \right) \mid x_i = \frac{i}{N}, \quad i = 0, 1, \dots, N. \right\}.$$

  The interest of this dataset is of course null for the modeling of real phenomenas. However it can be used for evaluation of the performance of training algorithms on elementary situations where analytical solutions are known (for example it is used in the numerical Section 6.5).
- **The dataset is designed for the modeling of a numerical inverse problem.** The situation is of course not new and has a long history in engineering and statistical science [7], and in many other different disciplines.
  To go beyond these historical references, one notices that this situation is more and more encountered in modern applications of Neural Networks and Machine Learning to the numerical modeling of physical processes and Scientific Computing. A recent example for the calculation of phase equilibriums (known as flash calculations) in the oil industry is in [81], another example is the modeling of flame fronts for turbulence calculations [50]. The calculation of trouble cell

[1] A popular bank of datasets with many examples is at https://www.kaggle.com/datasets.

https://doi.org/10.1515/9783110218091-005

indicators for the improvement of non linear hyperbolic solvers is addressed in [84]. One observes that there is nowadays a fierce (exponentially-growing is more adequate) activity in numerical and applied sciences and in the industry on the the new possibilities offered by neural network representations for the numerical modeling of inverse problems.

The third situation is the one analyzed in the sequel. A convenient framework starts from the general situation (1.1-(1.2)-(1.3) and makes the additional hypothesis that there exists a variable denoted as

$$\theta \in \mathbb{R}^p$$

and two explicit functions denoted as $u : \mathbb{R}^p \to \mathbb{R}^m$ and $v : \mathbb{R}^p \to \mathbb{R}^n$.
The variable θ may be called a hidden variable. The function u is maps the hidden variable θ to the input variable x, while the function v maps the hidden variable θ to the output variable y. What is meant by the fact that u and v are explicit functions is the fact that one can implement in the computer a numerical approximation of u and v with good accuracy.
From these premises, one can decide of some sampling in the space of hidden variable. It yields a finite list

$$\theta_i \in \mathbb{R}^p, \quad i = 1, 2, \ldots, N. \tag{4.1}$$

Because u and v are explicit functions, one can calculate pairs (x_i, y_i)

$$x_i = u(\theta_i) \text{ and } y_i = v(\theta_i), \qquad 1 \le i \le N.$$

It defines the dataset

$$\mathcal{D} = \{(x_i, y_i) = (u(\theta_i), v(\theta_i)), \quad 1 \le i \le N\}.$$

Let us remind the reader that the design principle of Machine Learning, as it was presented in this text, is to establish a implementable function f between the x_i's and the y_i's.

Lemma 4.0.1. *If the function u admits a left inverse $u^{-1} : \mathbb{R}^m \to \mathbb{R}^p$, then the objective function is*

$$f^{\text{obj}} = v \circ u^{-1} : \mathbb{R}^m \to \mathbb{R}^n.$$

Proof. A left inverse u^{-1} satisfies $u^{-1} \circ u = I$ where the identity operator operates in $\mathbb{R}^m$. So $f^{\text{obj}}(x_i) = v \circ u^{-1}(u(\theta_i)) = v \circ u^{-1} \circ u(\theta_i) = v(\theta_i) = y_i$. □

For many problems, the existence of the hidden variable and the existence of a left inverse are of course entirely hypothetical. In the next, we evaluate 4 different ways to construct the sampling (4.1).

4.1 The curse of dimension and uniform sampling

The curse of dimension corresponds either to $p \gg 1$ or to to $m \gg 1$ (for many applications, n is moderate). That is either the dimension of the hidden space is large or the dimension of the space of inputs is large. Then a natural issue is the ability of a finite number of points to sample adequately a region of such a space of high dimensionality. Because this is a well known and extremely difficult issue to tackle, it is coined the **curse of dimension**.

A first illustration of the curse of dimension consists to imagine that a finite number of samples of the hidden variable has to sample the hypercube $[0,1]^p \subset \mathbb{R}^p$. The naive method consists to use **uniform sampling**, that is one samples uniformly with $s \geq 1$ values per dimension. With this strategy the number of samples is

$$\#(\mathcal{D}) = s^p.$$

This number grows extremely fast with the dimension (one says there is a blow up with respect to p). For example make the hypothesis that one takes $s = 10$ points per direction which seems moderate at first sight. Then $\#(\mathcal{D}) = 10^p$ which is rapidly enormous if p is too large. It is useful to compare with the number of atoms per mole (equal to the Avogadro number 6×10^{23}), the number of bacteria on earth ($\approx 10^{30}$), the number of different games at the bridge game ($4^{52} \approx 10^{32}$) or the number of particles in the universe ($\approx 10^{80}$). These numbers are huge. In summary, the curse of dimension has the consequence that a brute force sampling in a space of large dimension is useless.

A second illustration of the curse of dimension consists to evaluate the distance between the points with respect to some given a priori error bound (this part is borrowed from Mallat [65]). The argument is explained in the space of inputs, but the analysis is the same in the hidden space (or in any other space of large dimension). Consider a function $g \in \mathrm{Lip}(\alpha)$ where the space of Lipschitz functions is

$$\mathrm{Lip}(\alpha) = \left\{ g \in W^{1,\infty}[0,1]^m, \quad \|\nabla g\|_{L^\infty[0,1]^m} \leq \alpha \right\}, \quad \alpha > 0.$$

A finite dataset is set up where the inputs (x_i) are given and

$$\mathcal{D} = \left\{ (x_i, y_i) \mid x_i \in [0,1]^m, y_i = g(x_i),\ i = 1, \dots, N \right\}.$$

We consider a reconstruction $\widetilde{g} : [0,1]^m \to \mathbb{R}^n$ and we make the hypothesis that the reconstruction is the best one in the sense that

$$\widetilde{g}(x) = y_i = g(x_i), \qquad 1 \leq i \leq N.$$

where i is the index of one of the closest neighbor of x, that is

$$|x - x_i| \leq |x - x_j| \quad j = 1, \dots, N.$$

Here x_i is not necessarily unique, but it is not important since if there is more than one closest neighbors, one can take the one x_i with the smallest index i.

The question is to evaluate the worst possible difference between $g \in \mathrm{Lip}(\alpha)$ and his best reconstruction. The answer is that it is the product of the Lipschitz constant times the distance between samples which is defined as

$$\varepsilon = \sup_{x\in[0,1]^m} \min_{1\le i\le I} |x - x_i|. \tag{4.2}$$

Lemma 4.1.1. *One has* $\sup_{g\in\mathrm{Lip}(\alpha)} \|g - \widetilde{g}\|_{L^\infty[0,1]^p} = \alpha\varepsilon.$

Proof. • By definition

$$|g(x) - \widetilde{g}(x)| = |g(x) - g(x_i)| \le \|\nabla g\|_{L^\infty[0,1]^m} |x - x_i| \le \alpha \min_i |x - x_i| \le \alpha\varepsilon.$$

So $\|g - \widetilde{g}\|_{L^\infty[0,1]^m} \le \alpha\varepsilon$.

• It remains to show the bound is reached. The function $\varphi : x \mapsto \min_{1\le i\le I} \|x - x_i\|$ is continuous. Since $[0,1]^m$ is compact, there exists $x^* \in [0,1]^m$ such that $\varphi(x^*) = \max_{x\in[0,1]^m} \varphi(x) = \varepsilon$. Denoting a best neighbor of x^* as x_i^*, one has

$$\min_{1\le i\le I} |x - x_i| \le |x^* - x_i^*| = \varepsilon \le |x^* - x_i|, \qquad \forall x \in [0,1]^m,\ \forall i = 1,\dots,N.$$

Consider $g^*(x) = \alpha|x - x_i^*|$. One shows that $g^* \in \mathrm{Lip}(\alpha)$, which yields of course that $\|g^* - \widetilde{g}^*\|_{L^\infty[0,1]^p} \le \alpha\varepsilon$. Moreover

$$|g^*(x^*) - \widetilde{g}^*(x^*)| = \alpha|x^* - x_i^*| - \alpha|x_i^* - x_i^*| = \alpha|x^* - x_i^*| = \alpha\varepsilon.$$

Therefore $\|g^* - \widetilde{g}^*\|_{L^\infty[0,1]^m} \ge \alpha\varepsilon$. The double inequality yields the claim. □

The definition (4.2) of ε means that all points $x \in [0,1]^m$ in the domain belong to at least one ball centered on x_i, and that all these balls have a radius less or equal to ε. It is equivalent to say that $[0,1]^m$ is covered by N balls of radius ε, that is $[0,1]^m \subset \bigcup_{i=1}^N B(x_i, \varepsilon)$. The question is the relation between N and ε, in dimension m.

For a given distance ε, one is interested with covering the domain with the less possible number of balls. So one considers

$$N^*(\varepsilon) = \min N$$

where $[0,1]^m$ is covered by N balls of of radius less of equal to ε. With respect to the result of Lemma 4.1.1, the best strategy for the distribution of the x_i is to cover $[0,1]^m$ with balls of radius $\varepsilon > 0$, trying at the same time to have the minimal number $N^*(\varepsilon)$ of such balls.

Lemma 4.1.2. *One has* $N^*(\varepsilon) \leq \left(\frac{\sqrt{m}}{2\varepsilon}\right)^m$.

Proof. By definition, $N^*(\varepsilon) \leq N$ where N is obtained from an uniform sampling in all directions of hypercubes of size $\Delta x > 0$, each of them **exactly embedded** in balls of radius ε. One has $N\Delta x^m = 1$. Also the corner is at distance $l = \sqrt{m\frac{\Delta x^2}{4}} = \frac{\sqrt{m}}{2}\Delta x$, so one takes $l = \varepsilon$. It yields the proof. □

Lemma 4.1.3. *One has*

$$\left[\left(\frac{2}{\pi e}\right)^{\frac{m}{2}} \sqrt{\pi m}\left(1 + O(m^{-1})\right)\right]\left(\frac{\sqrt{m}}{2\varepsilon}\right)^m \leq N^*(\varepsilon).$$

Proof. Consider an optimal distribution. Since $[0,1]^m \subset \bigcup_i B(x_i, \varepsilon)$, then

$$1 = \text{Vol}\left([0,1]^m\right) \leq N^*(\varepsilon)\text{Vol}\left(B(0,\varepsilon)\right).$$

One has $\text{Vol}\left(B(0,\varepsilon)\right) = \frac{\pi^{\frac{m}{2}}}{\Gamma(m/2+1)}\varepsilon^m$ where the Γ-function is in the denominator. If m is even then $\Gamma(m/2+1) = (m/2)!$. The Stirling formula for large $z > 1$ yields $\Gamma(z+1) = \sqrt{2\pi z}\left(\frac{z}{e}\right)^z\left(1 + O(z^{-1})\right)$. So

$$\frac{1}{\text{Vol}\left(B(0,\varepsilon)\right)} = \left(\frac{\sqrt{m}}{2\varepsilon}\right)^m \left(\frac{2}{\pi e}\right)^{\frac{m}{2}} \sqrt{\pi m}\left(1 + O(m^{-1})\right)$$

which is the low bound of the claim. □

One can rewrite the above results as

$$\left[\sqrt{\frac{2}{\pi e}}\left(1 + O\left(\frac{\log m}{m}\right)\right)\right]\frac{\sqrt{m}}{2}N^*(\varepsilon)^{-\frac{1}{m}} \leq \varepsilon \leq \frac{\sqrt{m}}{2}N^*(\varepsilon)^{-\frac{1}{m}}.$$

These inequalities explain that the optimal packing is still extremely sensitive to the dimension m. For large dimension $m \gg 1$, the cost in terms of the number of samples to obtain a small accuracy $\varepsilon > 0$ contains a term proportional to $\frac{\varepsilon}{-m}$. In summary the accuracy cannot be taken as small as desired with this simple approach. This is the curse of dimension.

4.2 The Monte-Carlo method

The result of Lemma 4.1.1 shows that the approximation of functions is difficult in spaces of large dimensionality. The Monte-Carlo method by Metropolis and Ulam [67] adopts a different angle of analysis, by considering that what is important is the approximation of integrals. The main result is that the Monte-Carlo method does not suffer from the curse of dimensionality (even it has other drawbacks). The presentation below focuses on basic notions which are useful for the implementation of datasets in the computer. To go beyond this elementary presentation, the reader is strongly advised to refer to classical textbooks on Machine Learning and Probability theory such as [74].

With the notation (1.27) and considering the least square approach, our cost function can take the form

$$J(W) = \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} |f(x_i, W) - y_i|^2 \tag{4.3}$$

where $N = \#(\mathcal{D})$ is the number of pairs in the dataset. It is very natural to consider that $J(W)$ is a numerical approximation of an integral

$$\int_{x\in[0,1]^m} |f(x, W) - y(x)|^2 \, dx \tag{4.4}$$

where $[0, 1]^m \ni x \mapsto y(x)$ is a function such that $y_i = y(x_i)$ for all $1 \le i \le N$. The question is now to evaluate the error between the discrete sum (4.3) and the integral (4.4) and to evaluate how much the approximation procedure is prone to the curse of dimension.

The basic problem can be formulated as follows. Let us consider the mean value of a given function f

$$\mu = \int_{[0,1]^m} f(x)dx \tag{4.5}$$

and the variance

$$\sigma = \int_{[0,1]^m} |f(x) - \mu|^2 dx = \int_{[0,1]^m} f(x)^2 dx - \mu^2. \tag{4.6}$$

The mean value can a priori be approximated by a sum like

$$J_N(x_1, \dots, x_N) = \frac{1}{N} \sum_{i=1}^{N} f(x_i). \tag{4.7}$$

The whole point is to construct the x_i's such that $|J_N - \mu|$ is small in a sense to determine.

Definition 4.2.1 (Monte-Carlo method)**.** *The Monte-Carlo method writes as follows: choose x_1 at random uniformly, then x_2 uniformly, and so on and so forth.*

Remark 4.2.2. *In a computer, it is sufficient to use the pre-defined function* **random** *to implement the Monte-Carlo procedure.*

To formalize the Monte-Carlo method, it is necessary to give a rigorous meaning to the term *at random uniformly* in the definition. Hereafter we favor a very elementary presentation. One makes the assumption that there exists another function defined from an abstract space Ω equipped with a measure such that $\int_\Omega d\omega = 1$. Writing

$$x = X(\omega) \tag{4.8}$$

means that the value of $x \in [0,1]^m$ is chosen accordingly to the choice of $\omega \in \Omega$ via some function X called a random variable. If $\Omega = [0,1]^m$ and $X = I$ is the identity map, then $x = \omega$. The probability that $x = X(\omega) \in A \subset [0,1]^m$ is calculated accordingly to

$$P\{X \in A\} = \int\limits_{\omega \text{ such that } X(\omega)\in A} d\omega.$$

With this formalism one considers the quantity

$$J_N(X(\omega_1), \ldots, X(\omega_N)) = \frac{1}{N}\sum_{i=1}^{N} f(X(\omega_i)), \tag{4.9}$$

where $\omega_i \in \Omega$ for all $1 \le i \le N$. Because of the transformation $X : \Omega \to [0,1]^m$, one redefines the mean value

$$\mu_X = \int\limits_\Omega f(X(\omega))d\omega \tag{4.10}$$

and the variance

$$\sigma_X^2 = \int\limits_\Omega |f(X(\omega)) - \mu|^2 d\omega = \int\limits_\Omega f(X(\omega))^2 d\omega - \mu^2. \tag{4.11}$$

If $X = I$ is the identity map, then (4.5) and (4.10) are identical, as well as (4.6) and (4.11). To evaluate how much $|J_N - \mu_X|$ is small, the method is to consider its integral

$$\sigma_N^2 = \int\limits_{\Omega^N} |J_N - \mu_X|^2 d\omega_1 \ldots d\omega_N. \tag{4.12}$$

The fundamental result is that σ_N goes to zero as $N \to \infty$ provided σ_X is finite.

Lemma 4.2.3. *One has* $\sigma_N = \frac{1}{\sqrt{N}}\sigma_X$.

Proof. The proof is by direct calculation, where one uses $\int_\Omega d\omega = 1$. One has

$$\sigma_N^2 = \int\limits_{\Omega^N} \left| \frac{1}{N}\sum_{i=1}^{N} (f(X(\omega_i)) - \mu_X) \right|^2 d\omega_1 \ldots d\omega_N$$

$$
\begin{aligned}
&= \frac{1}{N^2} \sum_{1\le i\le N} \int\limits_{\Omega} |f(X(\omega_i)) - \mu_X|^2 \, d\omega_i \\
+\frac{1}{N^2} \sum_{1\le i\ne j\le N} &\int\limits_{(\omega_i,\omega_j)\in\Omega^2} (f(X(\omega_i)) - \mu)\,(f(X(\omega_i j) - \mu_X)\, d\omega_i d\omega_j \\
&= \frac{1}{N} \int\limits_{\Omega} |f(X(\omega)) - \mu_X|^2 \, d\omega
\end{aligned}
$$

because

$$
\int\limits_{\Omega^2} (f(X(\omega_i)) - \mu)\,(f(X(\omega_i j) - \mu_X)\, d\omega_i d\omega_j = \left(\int\limits_{\Omega} (f(X(\omega)) - \mu_X)\, d\omega \right)^2 = 0.
$$

So $\sigma_N^2 = \frac{1}{N}\sigma_X^2$ which is equivalent to the claim. □

One immediately notices three features of the formula $\sigma_N = \frac{1}{\sqrt{N}}\sigma_X$.

- The convergence is relatively low with respect to N.
- The convergence is in the integral sense (in probability sense), which is an indication that the convergence may be noisy.
- The convergence is independent of the dimension m, which explains that the Monte-Carlo does not suffer from the curse of dimension. It is possible to derive intervals of confidence [2].

Practical consequence are explained below on an example. Let us take the objective function

$$
f^{\text{obj}}(x) = sin(x^1) + e^{-x^2} + x^1(x^2)^2(x^3)^3, \quad x = (x^1, x^2, x^3) \in [0,1]^3.
$$

[2] Intervals of confidence are consequence of the Tchebycheff inequality.

Lemma 4.2.4 (Tchebycheff inequality)**.** *Take $\lambda > 0$. Then $P\{|f(X) - \mu_X| \ge \lambda\} \le \frac{\sigma_X^2}{\lambda^2}$.*

Proof. One has $\sigma_X^2 = \int_\Omega |f(X(\omega)) - \mu|^2 d\omega \ge \int_{|f(X(\omega))-\mu_X|\ge\lambda} |f(X(\omega)) - \mu|^2 d\omega$. So $\sigma_X^2 \ge \lambda^2 \int_{|f(X(\omega))-\mu_X|\ge\lambda} d\omega = \lambda^2 P\{|f(X) - \mu_X| < \lambda\}$ which is the claim. □

Theorem 4.2.5. *One has the interval of confidence $P\{|J_N - \mu_X| < \lambda\} \ge 1 - \frac{\sigma_X}{\sqrt{N}\lambda}$.*

Proof. One firstly apply the Tchebycheff inequality to the function $(\omega_1, \dots, \omega_N) \mapsto J_N(X(\omega_1), \dots, X(\omega_N))$. It yields $P\{|J_N - \mu_X| \ge \lambda\} \le \frac{\sigma_N^2}{\lambda^2} = \frac{\sigma_X^2}{N\lambda^2}$. One has of course that $P\{|J_N - \mu_X| \ge \lambda\} + P\{|J_N - \mu_X| < \lambda\} = 1$. Therefore $P\{|J_N - \mu_X| < \lambda\} \ge 1 - \frac{\sigma_X^2}{N\lambda^2}$ which is the claim. □

For the simplicity, we consider a least square approximation

$$f(x : W, b) = Wx + b, \qquad (W, b) \in \mathcal{M}_{1,3}(\mathbb{R}) \times \mathbb{R}.$$

We construct two datasets.

The first dataset $\mathcal{D}_1$ is obtained by uniform sampling of x in $[0,1]^3$. It corresponds to using $X = I$ in (4.8). A practical definition is

$$\mathcal{D}_1 = \left\{ (x_i, y_i) = (x_i, f^{\text{obj}}(x_i)) \mid x_i^j \text{ randomized in } [0,1] \text{ for } 1 \le i \le N \text{ and } 1 \le j \le 3 \right\}.$$

Then Theorem 4.2.5 explains that the sum (4.3) is an approximation in the sense of probability of the integral

$$I_1(W, b) = \int_{[0,1]^3} \left| f^{\text{obj}}(x) - f(x : W, b) \right|^2 dx. \tag{4.13}$$

The second dataset $\mathcal{D}_2$ is constructed in a slightly different way. Let us consider the function

$$\begin{array}{cccc} \varphi : & [0,1] & \to & [0,1], \\ & u & \mapsto & v = \sin^2 \frac{\pi u}{2}. \end{array} \tag{4.14}$$

It is a diffeomorphism and

$$dv = \pi \sin \frac{\pi u}{2} \cos \frac{\pi u}{2} du = \pi \sqrt{v(1-v)} du. \tag{4.15}$$

Let us consider the transformation

$$\begin{array}{cccc} X : & [0,1]^3 & \to & [0,1]^3, \\ & \omega = (\omega^1, \omega^2, \omega^3) & \mapsto & x = (\varphi(\omega^1), \varphi(\omega^2), \varphi(\omega^3)). \end{array}$$

It defines a second dataset

$$\begin{aligned} \mathcal{D}_2 = \ & \Big\{ (x_i, y_i) = (X(\omega_i), f^{\text{obj}}(X(\omega_i)) \mid \\ & \omega_i^j \text{ randomized in } [0,1] \text{ for } 1 \le i \le N \text{ and } 1 \le j \le 3 \Big\}. \end{aligned}$$

Then Theorem 4.2.5 explains that the sum (4.3) is an approximation in the sense of probability of the integral

$$I_2(W, b) = \int_{[0,1]^3} \left| f^{\text{obj}}(X(\omega)) - f(X(\omega) : W, b) \right|^2 d\omega.$$

The differential equivalence (4.15) yields the other expression

$$I_2(W, b) = \frac{1}{\pi^3} \int_{[0,1]^3} \left| f^{\text{obj}}(x) - f(x : W, b) \right|^2 \frac{dx}{\sqrt{x^1(1-x^1)x^2(1-x^2)x^3(1-x^3)}} \tag{4.16}$$

where $dx = dx^1 dx^2 dx^3$. There is a weight in the integral, which is singular at the boundary of the hypercube $[0,1]^3$. It gives more importance of values where x is near the boundary. This is due to the reenforced sampling near the boundary.

Remark 4.2.6. *The comparison of (4.13) and (4.16) shows that the sampling of the Monte-Carlo method (which can be modified with a convenient transformation X) has an influence on the weight which shows up in the integrals.*

4.3 The Latin hypercube method

The Latin hypercube method was invented by researchers in scientific computing [66] and it has been analyzed in the context of probability theory in [57]. Improvements with optimal orthogonal-array-based latin hypercubes can be found in [55]. This can be understood as an intermediate between uniform sampling and the Monte-Carlo method. An illustration is in Figure 4.1. The Latin hypercube method can be analyzed with simple numerical analysis arguments as shown below, without any probability techniques. Instead the Jackson Theorem will be used to evaluate the truncation error of a function to its Jackson-Fourier approximation. However, the final error formula has an interest mainly in the context of probability interpretation.

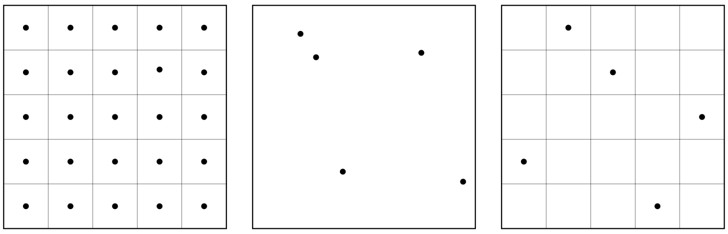

Fig. 4.1: Different sampling of the unit 2D square: a) uniform sampling with N^2 points on the left; b) Monte-Carlo sampling with N points in the center; c) Latin hypercube sampling with N points on the right. Here $N = 5$.

Let us consider that $x \in [0,1]^m$ and $m \gg 1$ is large (once again the method can be presented in the space of hidden parameters $\theta \in [0,1]^p$). One decides a priori of a certain number of samples $N \geq 1$. Then one divides the m segments which are the directions of the hypercube $[0,1]^m$ in N intervals of equal length. The N^m hypercubes of volume N^{-m} are referred to with the multi-index

$$\mathbf{j} = (j_1, \ldots, j_m) \in \{0, 1, \ldots, N-1\}^m.$$

An hypercube of volume N^{-m} is referred to as $c(\mathbf{j}) \subset [0,1]^m$. The center of the small hypercube $c(\mathbf{j})$ is

$$x(\mathbf{j}) = \left(j_1 + \frac{1}{2}, \ldots, j_m + \frac{1}{2}\right) \in c(\mathbf{j}). \tag{4.17}$$

Definition 4.3.1. *The Latin hypercube method writes as follows: choose randomly N hypercubes of volume N^{-m}, with the constraint that*

$$c(\mathbf{j}) \neq c(\mathbf{j}') \Longleftrightarrow j_1 \neq j_1' \text{ and } j_2 \neq j_2' \dots \text{ and } j_m \neq j_m'.$$

Since there are exactly N small hypercubes, it means that there is exactly one hypercube such that $j_1 = 0$, one hypercube such that $j_1 = 1$ until one hypercube such that $j_1 = N - 1$, and the same for the other directions. The collection of these small hypercubes will be called a Latin hypercube $\mathcal{L}_N$ of size N. Let us note S_N the group of permutations in $\{0, \dots, N-1\}$. There is bijection between all Latin hypercubes and S_N^{m-1} since all possible lists of small hypercubes can be described as

$$\mathcal{L}_N = \left\{ \mathbf{L}_N = \left(c\left(j, \sigma_2(j), \dots, \sigma_m(j)\right)\right)_{0 \le j \le N-1} \text{ for } (\sigma_2, \dots, \sigma_m) \in S_N^{m-1} \right\}. \tag{4.18}$$

A Latin hypercube is stratified for all input dimensions since the list of indices for the first dimension is exactly $\{0, \dots, N-1\}$, the list of indices for the second dimension is exactly $\{\sigma_2(0), \dots, \sigma_2(N-1)\} = \{0, \dots, N-1\}$, and the same for all input dimensions. The cardinality of all Latin hypercubes of size N in $[0,1]^m$ is

$$\#\left(\mathcal{L}_N\right) = (N!)^{m-1}.$$

Remark 4.3.2. *In the description (4.18) of all Latin hypercubes there is no permutation applied on the first axis, or in other terms the permutation on the first axis is the identity $\sigma_1 = I_{S_N}$. The choice of all axises is of course arbitrary in this description and it can be changed. This will be used to simplify some proofs below.*

Once an Latin hypercube $\mathbf{L}_N \in \mathcal{L}_N$ is chosen, then the dataset is

$$\mathcal{D} = \left\{ (x, y) = \left(x(\mathbf{j}), f^{\text{obj}}(x(\mathbf{j})) \right) \mid \mathbf{j} = (j, \sigma_2(j), \dots, \sigma_m(j)) \text{ for } 0 \le j \le N-1 \right\},$$

and the cost function is calculated accordingly to (4.3).

Assume $f^{\text{obj}} \in L^2([0,1]^m)$. Then one has the Fourier representation

$$f^{\text{obj}}(x) = \sum_{\mathbf{k} \in \mathbb{Z}^m} f^{\text{obj}}_{\mathbf{k}} e^{2\mathbf{i}\pi\langle \mathbf{k}, x\rangle}, \qquad f^{\text{obj}}_{\mathbf{k}} = \int_{[0,1]^m} f^{\text{obj}}(x) e^{-2\mathbf{i}\pi\langle \mathbf{k}, x\rangle} dx$$

where the series converges in the quadratic sense and

$$\|f\|^2_{[0,1]^m} = \sum_{\mathbf{k} \in \mathbb{Z}^m} \left| f^{\text{obj}}_{\mathbf{k}} \right|^2 < \infty.$$

Here we consider a particular case where the objective function is just one single Fourier mode

$$f^{\text{obj}}(x) = e^{2\mathbf{i}\pi\langle \mathbf{k}, x\rangle}, \qquad \mathbf{k} = (k_1, \dots, k_m) \in \mathbb{Z}^m. \tag{4.19}$$

The general case is examined in the next Section. The Fourier mode is periodic in all directions. Then the sum $J(\mathbf{L}_N) = \frac{1}{N}\sum_{x\in\mathbf{L}_N} f^{\mathrm{obj}}(x)$ reduces to

$$J(\mathbf{L}_N) = \frac{1}{N}\sum_{j=0}^{N-1} e^{\frac{2\mathbf{i}\pi}{N}(k_1 j + k_2\sigma_2(j)+\dots k_m\sigma_m(j))}. \tag{4.20}$$

An important question is to analyze the dependence of $J(\mathbf{L}_N)$ with respect to N. It is expected that $J(\mathbf{L}_N)$ should be close to the integral

$$\int\limits_{[0,1]^m} f^{\mathrm{obj}}(x)dx = 1 \text{ if } k=0, \quad \int\limits_{[0,1]^m} f^{\mathrm{obj}}(x)dx = 0 \text{ otherwise.}$$

The case of general functions is treated [57], where it is shown that $J(\mathbf{L}_N)$ converges in the sense of probability. We provide hereafter an elementary proof for Fourier modes (4.19).

Lemma 4.3.3. *One has the equality*

$$\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_N} |J(\mathbf{L}_N)|^2 = \frac{1}{N}\left(1+(N-1)\Pi_{\alpha=1}^m\left(\delta_{k_\alpha}^N - \frac{\mu_{k_\alpha}^N}{N-1}\right)\right) \tag{4.21}$$

where the notation is $\delta_{k_\alpha}^N = 1$ if $k_\alpha \equiv 0$ mod N and $\delta_{k_\alpha}^N = 0$ if $k_\alpha \not\equiv 0$ mod N, and the complementary quantity is $\mu_{k_\alpha}^N = 1-\delta_{k_\alpha}^N$.

Remark 4.3.4. *If all $\delta_{k_\alpha}^N$ vanish except one, then $\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_N}|J(\mathbf{L}_N)|^2 = 0$ which means that the corresponding Fourier modes are exactly integrated. This property is due to the fact that the projection of the grid on all axis is uniform, as visible in Fig. 4.1.*
An example of a Fourier mode $e^{2\mathbf{i}\pi\langle k,x\rangle}$ such that all $\delta_{k_\alpha}^N$ vanish except one is

$$\mathbf{k} = (k_1, 0, \dots, 0) \text{ with } k_1 \notin N\mathbb{Z},$$

that is $e^{2\mathbf{i}\pi\langle k,x\rangle}$ is constant with respect to the variable $x_2, \dots, x_m$.
Since the projection of the grid on the first axis x_1 is uniform, it is expected that $J(\mathbf{L}_N) = 0$ for all $\mathbf{L}_N \in \mathcal{L}_N$. A direct verification is

$$\sum_{j_1=0}^{N-1} e^{\frac{2\mathbf{i}\pi}{N}k_1(j_1+\frac{1}{2})} = e^{\frac{\mathbf{i}\pi k_1}{N}}\sum_{j_1=0}^{N-1} e^{2\mathbf{i}\pi k_1 j_1} = e^{\mathbf{i}\pi k_1}\frac{1-e^{\frac{2\mathbf{i}}{N}\pi k_1 N}}{1-e^{\frac{2\mathbf{i}}{N}\pi k_1}} = e^{\mathbf{i}\pi k_1}\frac{(=0)}{(\neq 0)} = 0.$$

The condition $k_1 \notin N\mathbb{Z}$ insures that the denominator does not vanish.
The property is the same for all axis.

Proof. The result is by a direct calculation where one reminds that $\#(\mathcal{L}_N) = (N!)^{m-1}$. One has that

$$|J(\mathbf{L}_N)|^2 = \frac{1}{N}$$

$$+\frac{1}{N^2} \underbrace{\sum_{0 \le j \ne j' \le N-1} e^{\frac{2i\pi}{N}\left(k_1(j-j')+k_2(\sigma_2(j)-\sigma_2(j'))+\cdots+k_m(\sigma_m(j)-\sigma_m(j'))\right)}}_{=\mathrm{A}(\sigma_2,\ldots,\sigma_m)}$$

where $\mathrm{A}(\sigma_2,\ldots,\sigma_m)$ corresponds to the double products in the calculation of $|J(\mathbf{L}_N)|^2$. The sum of the double products over all possible Latin hypercubes can be written as

$$\sum_{\mathbf{L}_N \in \mathcal{L}_N} \mathrm{A}(\sigma_2,\ldots,\sigma_m) = \sum_{0 \le j \ne j' \le N-1} \left(e^{\frac{2i\pi}{N}k_1(j-j')} \right. \tag{4.22}$$

$$\left. \times \left(\sum_{\sigma_2} e^{\frac{2i\pi}{N}k_2(\sigma_2(j)-\sigma_2(j'))} \times \left(\cdots \times \left(\sum_{\sigma_m} e^{\frac{2i\pi}{N}k_m(\sigma_m(j)-\sigma_m(j'))} \right) \cdots \right)\right)\right).$$

The point is that all sums can be calculated and their value does not dependent on j and j'. Let us concentrate for example on $\Lambda(k_m, j, j') = \sum_{\sigma_m} e^{\frac{2i\pi}{N}k_m(\sigma_m(j)-\sigma_m(j'))}$, where only the numerical values of $\sigma_m(j)$ and $\sigma_m(j')$ matter. By construction $\sigma_m(j)$ take exactly N values which are $\sigma_m(j) = 0, 1, \ldots, N-1$. Then $\sigma_m(j')$ take values in the same list but $\sigma_m(j') \ne \sigma(j)$, so it yields $N-1$ different values. For $\alpha \ne \beta \in \{0, \ldots, N-1\}$, the number of permutations such that $\sigma_m(j) = \alpha$ and $\sigma_m(j') = \beta$ is equal to $N-2)!$. Therefore one has

$$\Lambda(k_m, j, j') = (N-2)! \sum_{\alpha} e^{\frac{2i\pi}{N}k_m\alpha} \sum_{\beta \ne \alpha} e^{-\frac{2i\pi}{N}k_m\beta}.$$

The discussion is as follows.

- The first case is $k_m \not\equiv 0 \bmod N$. Then $u = e^{-\frac{2i\pi}{N}k_m}$ is a non trivial root of unity (that is $u^N = 1$ and $u \ne 1$). So $0 = \sum_\beta e^{-\frac{2i\pi}{N}k_m\beta} = \sum_{\beta \ne \alpha} e^{-\frac{2i\pi}{N}k_m\beta} + e^{-\frac{2i\pi}{N}k_m\alpha}$ and

$$\sum_{\beta \ne \alpha} e^{-\frac{2i\pi}{N}k_m\beta} = -e^{-\frac{2i\pi}{N}k_m\alpha}.$$

Then

$$\Lambda(k_m, j, j') = -(N-2)! \sum_{\alpha} e^{\frac{2i\pi}{N}k_m\alpha} e^{-\frac{2i\pi}{N}k_m\alpha} = -N(N-2)!.$$

- The second case is $k_m =\equiv 0 \bmod N$. It yields $\Lambda(k_m, j, j') = N!$.
- Both cases can be written as $\Lambda(k_m, j, j') = \delta^N_{k_m} N! - \mu^N_{k_m} N(N-2)!$.
- The result is therefore independent of j and j', and it is extended to other terms

$\Lambda(k_2, j, j')$, ..., $\Lambda(k_{m-1}, j, j')$ in the product (4.22).

• A similar analysis yields that

$$\sum_{0\le j\neq j'\le N-1} e^{\frac{2i\pi}{N}k_1(j-j')} = \frac{1}{(N-2)!}\sum_{\sigma_1} e^{\frac{2i\pi}{N}k_1(\sigma_1(j)-\sigma_1(j'))}$$

$$= \frac{1}{(N-2)!}\left(\delta_{k_1}^N N! - \mu_{k_1}^N N(N-2)!\right).$$

• One obtains the formula

$$\sum_{\mathbf{L}_N\in\mathcal{L}_N} |J(\mathbf{L}_N)|^2 = \frac{1}{N}(N!)^{m-1} + \frac{1}{N^2(N-2)!}\Pi_{\alpha=1}^m\left(\delta_{k_\alpha}^N N! - \mu_{k_\alpha}^N N(N-2)!\right)$$

which is reorganized as (4.21).

• Finally let us consider that all $\delta_{k_\alpha}^N$ vanish except one. For example □

The interpretation of formula (4.21) depends on k. We distinguish four cases.

•For all $\mathbf{k}\in\mathbb{Z}^m$, one has the evident bound $J(\mathbf{L}_N)\le 1$.

• The second case is $\mathbf{k}=0_{\mathbb{Z}^m}$. Then $J(\mathbf{L}_N)=1$ by construction for all $\mathbf{L}_N\in\mathcal{L}_N$ and this is recovered by the right hand side of (4.21).

• The third case is when $\mathbf{k}\neq 0_{\mathbb{Z}^m}$ stays in the box

$$\left\{\mathbf{k}\in\mathbb{Z}^m \mid -N<k_j<N \text{ for } 1\le j\le m\right\}. \tag{4.23}$$

Then (4.21) is implies

$$\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_N}|J(\mathbf{L}_N)|^2 \le \frac{2}{N} \tag{4.24}$$

which can be rewritten as [3]

$$\left(\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_N}|J(\mathbf{L}_N)|^2\right)\le\frac{C}{N^{\frac{1}{2}}},\qquad C=\sqrt{2}.$$

• The fourth case is when $\mathbf{k}$ is in the box (4.23) and all its components vanish except one. Then the right hand side of (4.21) is zero. It is due to the fact that the

[3] As in the Monte-Carlo method, the constant C is independent of the dimension m of the inputs space. In this situation, the coefficient $\frac{1}{(N!)^{m-1}}$ is the probability of drawing a particular Latin hypercube in the ensemble $\mathcal{L}_N$ (the number of elements is $\#(\mathcal{L}_N)=(N!)^{m-1}$). The situation is similar to the probability of drawing one facet of a dice. The probability is 1/6 because the number of facets is 6. Taking the square root one obtains the classical $N^{-\frac{1}{2}}$ rate of convergence zero characteristics of the Monte-Carlo method (see Lemma 4.2.3), at the price of having to take into account all results in the sense of probability in the discrete space $\mathcal{L}_N$.

Fourier mode reduces to a purely monodimensional Fourier mode for which the uniform distribution yields perfect integration.

The next result extends the result of Lemma 4.3.3 to the calculation of the product of $J(\mathbf{L}_N)$ against

$$\widetilde{J}(\mathbf{L}_N) = \frac{1}{N}\sum_{j=0}^{N-1} e^{\frac{2\mathrm{i}\pi}{N}\left(\widetilde{k}_1 j+\widetilde{k}_2\sigma_2(j)+\ldots\widetilde{k}_m\sigma_m(j)\right)} \tag{4.25}$$

where the Fourier numbers are different

$$\mathbf{k} = (k_1,\ldots,k_m) \neq \widetilde{\mathbf{k}} \equiv (\widetilde{k}_1,\ldots,\widetilde{k}_m) \bmod N \tag{4.26}$$

which means there exists $1 \le i \le m$ such that $k_i \not\equiv \widetilde{k}_i \bmod N$.

Lemma 4.3.5. *Assume (4.26). Then the quantities $J(\mathbf{L}_N)$ and $\widetilde{J}(\mathbf{L}_N)$ satisfy the orthogonality relation*

$$\sum_{\mathbf{L}_N\in\mathcal{L}_N} J(\mathbf{L}_N)\overline{\widetilde{J}(\mathbf{L}_N)} = 0. \tag{4.27}$$

Proof. One has

$$N^2 J(\mathbf{L}_N)\overline{\widetilde{J}(\mathbf{L}_N)} = \underbrace{\sum_{0\le j\le N-1} e^{\frac{2\mathrm{i}\pi}{N}\left((k_1-\widetilde{k}_1)j+(k_2-\widetilde{k}_2)\sigma_2(j)+\cdots+(k_m-\widetilde{k}_m)\sigma_m(j)\right)}}_{=\mathrm{C}(\sigma_2,\ldots,\sigma_m)}$$

$$+\underbrace{\sum_{0\le j\neq j'\le N-1} e^{\frac{2\mathrm{i}\pi}{N}\left(k_1 j-\widetilde{k}_1 j'+k_2\sigma_2(j)-\widetilde{k}_2\sigma_2(j')+\cdots+k_m\sigma_m(j)-\widetilde{k}_m\sigma_m(j')\right)}}_{=\mathrm{D}(\sigma_2,\ldots,\sigma_m)}.$$

One has that

$$\sum_{\mathbf{L}_N\in\mathcal{L}_N} \mathrm{C}(\sigma_2,\ldots,\sigma_m) = \sum_j e^{\frac{2\mathrm{i}\pi}{N}(k_1-\widetilde{k}_1)j}\sum_{\sigma_2} e^{\frac{2\mathrm{i}\pi}{N}(k_2-\widetilde{k}_2)\sigma_2(j)}\ldots\sum_{\sigma_m} e^{\frac{2\mathrm{i}\pi}{N}(k_m-\widetilde{k}_m)\sigma_m(j)}.$$

The internal term is

$$\sum_{\sigma_m} e^{\frac{2\mathrm{i}\pi}{N}(k_m-\widetilde{k}_m)\sigma_m(j)} = (N-1)!\left(\sum_{j''=0}^{N-1} e^{\frac{2\mathrm{i}\pi}{N}(k_m-\widetilde{k}_m)j''}\right). \tag{4.28}$$

As noticed in Remark 4.3.2, it is always possible to use an arbitrary ordering of the axis, that is one can assume that $k_m - \widetilde{k}_m \not\equiv 0 \bmod N$. One obtains $\sum_{\mathbf{L}_N\in\mathcal{L}_N} \mathrm{C}(\sigma_2,\ldots,\sigma_m) = 0$.

The other term is analyzed in a similar way. One has

$$\sum_{\mathbf{L}_N\in\mathcal{L}_N} \mathrm{D}(\sigma_2,\ldots,\sigma_m) = \sum_{j\neq j'} e^{\frac{2\mathrm{i}\pi}{N}(k_1 j'-\widetilde{k}_1 j')}\ldots\sum_{\sigma_m} e^{\frac{2\mathrm{i}\pi}{N}(k_m\sigma_m(j)-\widetilde{k}_m\sigma_m(j'))}.$$

The internal term is

$$\sum_{\sigma_m} e^{\frac{2\mathrm{i}\pi}{N}(k_m\sigma_m(j)-\widetilde{k}_m\sigma_m(j'))} = (N-2)!\sum_{\alpha} e^{\frac{2\mathrm{i}\pi}{N}k_m\alpha}.$$

Let us take as well that $k_m - \widetilde{k}_m \not\equiv 0 \bmod N$. So either $k_m \neq 0$ or $\widetilde{k}_m \not\equiv 0 \bmod N$. Because of the symmetry

$$\sum_{\alpha} e^{\frac{2\mathrm{i}\pi}{N}k_m\alpha}\sum_{\beta\neq\alpha} e^{-\frac{2\mathrm{i}\pi}{N}\widetilde{k}_m\beta} = \sum_{\alpha} e^{\frac{2\mathrm{i}\pi}{N}\widetilde{k}_m\alpha}\sum_{\beta\neq\alpha} e^{-\frac{2\mathrm{i}\pi}{N}k_m\beta},$$

it is sufficient to consider the case $\widetilde{k}_m \not\equiv 0 \bmod N$. Then $\sum_{\beta\neq\alpha} e^{-\frac{2\mathrm{i}\pi}{N}\widetilde{k}_m\beta} = -e^{-\frac{2\mathrm{i}\pi}{N}\widetilde{k}_m\alpha}$ and $\sum_{\alpha} e^{\frac{2\mathrm{i}\pi}{N}k_m\alpha}\sum_{\beta\neq\alpha} e^{-\frac{2\mathrm{i}\pi}{N}\widetilde{k}_m\beta} = -\sum_{\alpha} e^{\frac{2\mathrm{i}\pi}{N}(k_m-\widetilde{k}_m)\alpha}$, so one can use the result already obtained for (4.28) which vanishes under the condition $k_m - \widetilde{k}_m \not\equiv 0 \bmod N$. Therefore $\sum_{\mathbf{L}_N\in\mathcal{L}_N} \mathrm{D}(\sigma_2,\ldots,\sigma_m) = 0$ which yields the claim. □

For further development, we investigate a different Fourier representation where the mode are cosine functions. That is we consider

$$f^{\mathrm{obj}}(x) = \cos\left(\pi\left\langle \mathbf{k}, x\right\rangle\right) \tag{4.29}$$

and

$$I(\mathbf{L}_N) = \frac{1}{N}\sum_{j=0}^{N-1}\cos\left(\frac{\pi}{N}\left(k_1\left(j+\frac{1}{2}\right)+k_2\left(\sigma_2(j)+\frac{1}{2}\right)+\ldots k_m\left(\sigma_m(j)+\frac{1}{2}\right)\right)\right). \tag{4.30}$$

The right hand is expressed with the cosine function and a difference with (4.20) is the factor π instead of 2π.

Lemma 4.3.6. *One has the equality*

$$\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_N}|I(\mathbf{L}_N)|^2 = \frac{1}{N}\left(1+(N-1)\Pi_{\alpha=1}^{m}\left(\delta_{k_\alpha}^{2N}-\frac{\mu_{k_\alpha}^{2N}}{N-1}\right)\right) \tag{4.31}$$

with the same notations for $\delta_{k_\alpha}^{2N}$ and $\mu_{k_\alpha}^{2N}$ as in the previous Lemma.

Remark 4.3.7. *The interpretation of (4.31) is similar to the interpretation of (4.21). In particular both sides vanish if all $\delta_{k_\alpha}^{2N}$ vanish but one. It comes from the fact that the numerical integration of a cosine mode $k_1 \notin 2N\mathbb{Z}$ on a 1D uniform grid is exact.*

Indeed one has the identity

$$\sum_{j_1=0}^{N-1}\cos\left(\frac{\pi k_1}{N}\left(j_1+\frac{1}{2}\right)\right) = \frac{1}{2}\sum_{j_1=-N}^{N-1} e^{\mathbf{i}\frac{\pi k_1}{N}(j_1+\frac{1}{2})}$$

$$= \frac{e^{\mathbf{i}\pi k_1 \frac{1-2N}{2N}}}{2} \sum_{j_1=0}^{2N-1} e^{\mathbf{i}\frac{\pi k_1}{N} j_1} = \frac{e^{\mathbf{i}\pi k_1 \frac{1-2N}{2N}}}{2} \frac{1 - e^{\mathbf{i}\frac{\pi k_1}{N} 2N}}{1 - e^{\mathbf{i}\frac{\pi k_1}{N}}} = 0.$$

Proof. The method is very similar to the proof of the previous Lemma. One has

$$\begin{aligned} I(\mathbf{L}_N) = \quad & \frac{1}{2N} \sum_{j=0}^{N-1} \Big(e^{\frac{\mathbf{i}\pi}{N}\left(k_1\left(j+\frac{1}{2}\right)+k_2\left(\sigma_2(j)+\frac{1}{2}\right)+\ldots k_m\left(\sigma_m(j)+\frac{1}{2}\right)\right)} \\ & \qquad + e^{-\frac{\mathbf{i}\pi}{N}\left(k_1\left(j+\frac{1}{2}\right)+k_2\left(\sigma_2(j)+\frac{1}{2}\right)+\ldots k_m\left(\sigma_m(j)+\frac{1}{2}\right)\right)} \Big) . \end{aligned} \tag{4.32}$$

So

$$\begin{aligned} |I(\mathbf{L}_N)|^2 = & \frac{1}{2N} \\ & + \frac{1}{4N^2} \Bigg(\sum_{0 \le j \ne j' \le N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j-j')+k_2(\sigma_2(j)-\sigma_2(j'))+\ldots k_m(\sigma_m(j)-\sigma_m(j'))\right)} \\ & \quad + \sum_{0 \le j \ne j' \le N-1} e^{-\frac{\mathbf{i}\pi}{N}\left(k_1(j-j')+k_2(\sigma_2(j)-\sigma_2(j'))+\ldots k_m(\sigma_m(j)-\sigma_m(j'))\right)} \\ & \quad + \sum_{0 \le j, j' \le N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+j'+1)+k_2(\sigma_2(j)+\sigma_2(j')+1)+\ldots k_m(\sigma_m(j)+\sigma_m(j')+1)\right)} \\ & \quad + \sum_{0 \le j, j' \le N-1} e^{-\frac{\mathbf{i}\pi}{N}\left(k_1(j+j'+1)+k_2(\sigma_2(j)+\sigma_2(j')+1)+\ldots k_m(\sigma_m(j)+\sigma_m(j')+1)\right)} \Bigg) . \end{aligned} \tag{4.33}$$

For all $2 \le r \le m$, the permutation σ_r is a bijection from $\{0, \ldots, N-1\}$ into itself. We extend it as a permutation in $\{0, \ldots, 2N-1\}$ by setting

$$\begin{aligned} \widehat{\sigma}_r(j'') &= \sigma_r(j'') && \text{for } j'' \in \{0, \ldots, N\}, \\ \widehat{\sigma}_r(j'') &= 2N - 1 - \sigma_r(2N-1-j'') && \text{for } j'' \in \{N, \ldots, 2N-1\}. \end{aligned}$$

By construction the restriction of $\widehat{\sigma}_r$ to $\{0, \ldots, N\}$ is a permutation, and the restriction $\widehat{\sigma}_r$ to $\{N, \ldots, 2N-1\}$ is also a permutation.

Consider the third term in the parenthesis and make the change of index

$$j'' = 2N - 1 - j' \in \{N, \ldots, 2N-1\}.$$

So

$$e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+j'+1)\right)} = e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+2N-j'')\right)} = e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j-j'')\right)}.$$

Similarly one has for all $2 \le r \le m$

$$e^{\frac{\mathbf{i}\pi}{N}\left(k_r(\sigma_r(j)+\sigma_r(j')+1)\right)} = e^{\frac{\mathbf{i}\pi}{N}\left(k_r(\sigma_r(j)+\sigma_r(2N-1-j'')+1)\right)}$$

$$= e^{\frac{\mathbf{i}\pi}{N}\left(k_r(\sigma_r(j)+\sigma_r(2N-1-j'')+1)\right)} = e^{\frac{\mathbf{i}\pi}{N}\left(k_r(\widehat{\sigma}_r(j)+2N-\widehat{\sigma}_r(j''))\right)} = e^{\frac{\mathbf{i}\pi}{N}\left(k_r(\widehat{\sigma}_r(j)-\widehat{\sigma}_r(j''))\right)}.$$

One has

$$\sum_{0\leq j,j'\leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+j')+k_2(\sigma_2(j)+\sigma_2(j'))+\ldots k_m(\sigma_m(j)+\sigma_m(j'))\right)}$$

So one can group the first term and the third terms in the parenthesis. One obtains

$$\sum_{0\leq j\neq j'\leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j-j')+k_2(\sigma_2(j)-\sigma_2(j'))+\ldots k_m(\sigma_m(j)-\sigma_m(j'))\right)}$$

$$+\sum_{0\leq j,j'\leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+j'+1)+k_2(\sigma_2(j)+\sigma_2(j')+1)+\ldots k_m(\sigma_m(j)+\sigma_m(j')+1)\right)}$$

$$=\underbrace{\sum_{j=0}^{N-1}\sum_{j''=0(j''\neq j)}^{2N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j-j'')+k_2(\sigma_2(j)-\widehat{\sigma}_2(j'''))+\ldots k_m(\sigma_m(j)-\widehat{\sigma}_m(j'''))\right)}}_{\text{renamed as } B(\sigma_2,\ldots,\sigma_m)}.$$

It means that (4.33) can be written in the compact form

$$|I(\mathbf{L}_N)|^2=\frac{1}{2N}+\frac{1}{2N^2}\;Re\left(B(\widehat{\sigma}_2,\ldots,\widehat{\sigma}_m)\right).$$

Now one can sum over all Latin hypercubes. One gets

$$\sum_{\mathbf{L}_N} B(\widehat{\sigma}_2,\ldots,\widehat{\sigma}_m)=\sum_{j=0}^{N-1}\sum_{j''=0(j''\neq j)}^{2N-1}\Bigg(e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j-j'')\right)}$$

$$\times\left(\sum_{\sigma_2} e^{\frac{\mathbf{i}\pi}{N}k_2(\sigma_2(j)-\widehat{\sigma}_2(j'))}\times\left(\cdots\times\left(\sum_{\sigma_m} e^{\frac{\mathbf{i}\pi}{N}k_m(\sigma_m(j)-\widehat{\sigma}_m(j'))}\right)\cdots\right)\right)\Bigg).$$

The analysis of these terms follows the same path as in the proof of (4.21). For example the internal term is

$$\begin{aligned}
\textstyle\sum_{\sigma_m} e^{\frac{\mathbf{i}\pi}{N}k_m(\widehat{\sigma}_m(j)-\widehat{\sigma}_m(j'))} &= (N-2)!\textstyle\sum_{\alpha=0}^{N-1} e^{\frac{\mathbf{i}\pi}{N}k_m\alpha}\textstyle\sum_{\beta=0(\beta\neq\alpha)}^{2N-1} e^{-\frac{\mathbf{i}\pi}{N}k_m\beta}\\
&= \begin{cases} -(N-2)!N & \text{if } k_m\not\equiv 0 \bmod 2N,\\ N! & \text{if } k_m\equiv 0 \bmod 2N,\end{cases}\\
&= N!\left(\delta_{k_m}^{2N}-\tfrac{1}{N-1}(\mu_{k_m}^{2N}\right)().
\end{aligned}$$

The rest of the proof is left to the reader. □

To establish the orthogonality of the cosine modes, we consider another mode with $\mathbf{k}=(\widetilde{k}_1,\ldots,\widetilde{k}_m)$

$$\widetilde{I}(\mathbf{L}_N)=\frac{1}{N}\sum_{j=0}^{N-1}\cos\left(\frac{\pi}{N}\left(\widetilde{k}_1\left(j+\frac{1}{2}\right)+\widetilde{k}_2\left(\sigma_2(j)+\frac{1}{2}\right)+\ldots\widetilde{k}_m\left(\sigma_m(j)+\frac{1}{2}\right)\right)\right).$$

We make the assumption

$$\mathbf{k} = (k_1, \dots, k_m) \neq \widetilde{\mathbf{k}} \equiv (\widetilde{k}_1, \dots, \widetilde{k}_m) \bmod 2N, \tag{4.34}$$

which means similarly to (4.26) that there exists $1 \leq i \leq m$ such that $k_i \not\equiv \widetilde{k}_i$ mod $2N$.

Lemma 4.3.8. *Assume (4.34). Then one has the orthogonality relation*

$$\sum_{\mathbf{L}_N \in \mathcal{L}_N} I(\mathbf{L}_N)\widetilde{I}(\mathbf{L}_N) = 0. \tag{4.35}$$

Proof. Using the expansion (4.32) and and the same one mutatis mutandis for $\widetilde{I}(\mathbf{L}_N)$, one obtains

$$\begin{aligned}
& 4N^2 I(\mathbf{L}_N)\widetilde{I}(\mathbf{L}_N) = \\
& \left(e^{\frac{\mathbf{i}\pi}{N}\left(k_1\left(j+\frac12\right)+k_2\left(\sigma_2(j)+\frac12\right)+\dots k_m\left(\sigma_m(j)+\frac12\right)\right)} + e^{-\frac{\mathbf{i}\pi}{N}\left(k_1\left(j+\frac12\right)+k_2\left(\sigma_2(j)+\frac12\right)+\dots k_m\left(\sigma_m(j)+\frac12\right)\right)}\right) \\
\times \quad & \left(e^{-\frac{\mathbf{i}\pi}{N}\left(\widetilde{k}_1\left(j+\frac12\right)+\widetilde{k}_2\left(\sigma_2(j)+\frac12\right)+\dots \widetilde{k}_m\left(\sigma_m(j)+\frac12\right)\right)} + e^{\frac{\mathbf{i}\pi}{N}\left(\widetilde{k}_1\left(j+\frac12\right)+\widetilde{k}_2\left(\sigma_2(j)+\frac12\right)+\dots \widetilde{k}_m\left(\sigma_m(j)+\frac12\right)\right)}\right) \\
= \quad & \sum_{0 \leq j \leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left((k_1-\widetilde{k}_1)(j+\frac12)+(k_2-\widetilde{k}_2)(\sigma_2(j)+\frac12)+\dots\right)} \qquad (= A(\mathbf{L}_N)) \\
+ \quad & \sum_{0 \leq j \neq j' \leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+\frac12)-\widetilde{k}_1(j'+\frac12)+k_2(\sigma_2(j)+\frac12)-\widetilde{k}_2(\sigma_2(j')+\frac12)+\dots\right)} \qquad (= B(\mathbf{L}_N)) \\
+ \quad & \sum_{0 \leq j \neq j' \leq N-1} e^{-\frac{\mathbf{i}\pi}{N}\left(k_1(j+\frac12)-\widetilde{k}_1(j'+\frac12)+k_2(\sigma_2(j)+\frac12)-\widetilde{k}_2(\sigma_2(j')+\frac12)+\dots\right)} \qquad (= C(\mathbf{L}_N)) \\
+ \quad & \sum_{0 \leq j \leq N-1} e^{-\frac{\mathbf{i}\pi}{N}\left((k_1-\widetilde{k}_1)(j+\frac12)+(k_2-\widetilde{k}_2)(\sigma_2(j)+\frac12)+\dots\right)} \qquad (= D(\mathbf{L}_N)) \\
+ \quad & \sum_{0 \leq j, j' \leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+\frac12)+\widetilde{k}_1(j'+\frac12)+k_2(\sigma_2(j)+\frac12)+\widetilde{k}_2(\sigma_2(j')+\frac12)+\dots\right)} \qquad (= E(\mathbf{L}_N)) \\
+ \quad & \sum_{0 \leq j, j' \leq N-1} e^{-\frac{\mathbf{i}\pi}{N}\left(k_1(j+\frac12)-\widetilde{k}_1(j'+\frac12)+k_2(\sigma_2(j)+\frac12)-\widetilde{k}_2(\sigma_2(j')+\frac12)+\dots\right)} \qquad (= F(\mathbf{L}_N)).
\end{aligned}$$

One has that

$$\sum_{\mathbf{L}_N \in \mathcal{L}_N} (A(\mathbf{L}_N) + D(\mathbf{L}_N)) = \sum_{-N \leq j \leq N-1} e^{\frac{\mathbf{i}\pi}{N}(k_1-\widetilde{k}_1)(j+\frac12)}$$

$$\times \left(\sum_{\sigma_2} e^{\frac{\mathbf{i}\pi}{N}(k_2-\widetilde{k}_2)(\sigma_2(j)+\frac12)} \times \left(\dots \times \left(\sum_{\sigma_m} e^{\frac{\mathbf{i}\pi}{N}(k_m-\widetilde{k}_m)(\sigma_m(j)+\frac12)}\right) \dots\right)\right).$$

The hypothesis (4.34) and a convenient ordering (see Remark 4.3.2) yield that the internal term vanishes. One gets a first identity $\sum_{\mathbf{L}_N \in \mathcal{L}_N} (A(\mathbf{L}_N) + D(\mathbf{L}_N)) = 0$.

Next one has

$$\sum_{\mathbf{L}_N \in \mathcal{L}_N} (B(\mathbf{L}_N) + C(\mathbf{L}_N) + E(\mathbf{L}_N) + F(\mathbf{L}_N))$$

$$= \sum_{-N \leq j \neq j' \leq N-1} e^{\frac{\mathbf{i}\pi}{N}\left(k_1(j+\frac{1}{2}) - \widetilde{k}_1(j'+\frac{1}{2})\right)} \left(\sum_{\sigma_2} e^{\frac{\mathbf{i}\pi}{N}\left(k_2(\sigma_2(j)+\frac{1}{2}) - \widetilde{k}_2(\sigma_2(j')+\frac{1}{2})\right)}\right.$$

$$\left.\left(\dots\left(\sum_{\sigma_m} e^{\frac{\mathbf{i}\pi}{N}\left(k_m(\sigma_m(j)+\frac{1}{2}) - \widetilde{k}_m(\sigma_m(j')+\frac{1}{2})\right)}\right)\dots\right)\right).$$

Consider the internal term and assume that $k_m \not\equiv \widetilde{k}_m$ mod $2N$. Due to the symmetry between j and j', one can as well assume that $\widetilde{k}_m \not\equiv 0$ mod $2N$. Then

$$\sum_{\sigma_m} e^{\frac{\mathbf{i}\pi}{N}\left(k_m(\sigma_m(j)+\frac{1}{2}) - \widetilde{k}_m(\sigma_m(j')+\frac{1}{2})\right)} = (N-2)! \sum_{\alpha} e^{\frac{\mathbf{i}\pi}{N} k_m(\alpha+\frac{1}{2})} \sum_{\beta \neq \alpha} e^{\frac{\mathbf{i}\pi}{N} \widetilde{k}_m(\beta+\frac{1}{2})}$$

$$= -(N-2)! \sum_{\alpha} e^{\frac{\mathbf{i}\pi}{N} k_m(\alpha+\frac{1}{2})} e^{\frac{\mathbf{i}\pi}{N} \widetilde{k}_m(\alpha+\frac{1}{2})} = 0.$$

It yields the claim by summation with the previous identity. □

4.4 Comparison of uniform sampling versus Latin hypercube method

To compare the efficiency of the two methods, we consider the integral

$$I = \int_{[0,1]^m} g(x)dx$$

where the function g has a certain smoothness. We will consider the smoothness condition

$$\|\nabla g\|_{L^\infty([0,1]^m)} < \infty, \tag{4.36}$$

Our main example is

$$g(x) = \left| f(x, W) - f^{\mathrm{obj}}(x) \right|^2.$$

If one takes $(x, W) \mapsto f(x, W)$ to be predicted by a Neural Network architecture with the ReLU function, then the derivative of f is also bounded is maximal norm. For the objective function f^{obj}, the smoothness assumption (4.36) is an hypothesis.

We will need the notion of a Jackson kernel [24] over the torus $\mathbb{T} = [-1, 1]_{\mathrm{per}}$. The non negative Jackson kernel [24][page 203] is defined by

$$K_n(t) = \lambda_n \left(\frac{\sin m\pi t/2}{\sin \pi t/2}\right)^4, \quad m = [n/2] + 1, \quad \int_{\mathbb{T}} K_n(t)dt = 1.$$

It is an even 2-periodic trigonometric polynomial of degree $\leq n$

$$K_n(t) = \sum_{0 \leq k \leq n} \alpha_{n,k} \cos(k\pi t), \quad \alpha_{n,0} = 1. \tag{4.37}$$

The Jackson integral of a 2-periodic function h is defined as the convolution

$$J_n(t) = K_n \star h(\mu) = \int_{\mathbb{T}} K_n(t - \mu) h(\mu) d\mu.$$

The convolution commutes with the differentiation, so if $\|h'\|_{L^\infty(\mathbb{T})} < \infty$, then one has the bound

$$\|J_n'\|_{L^\infty(\mathbb{T})} \leq \|h'\|_{L^\infty(\mathbb{T})}. \tag{4.38}$$

Theorem 4.4.1 (Jackson, see [24] page 204). *For all 2-periodic function $h \in W^{1,\infty}(\mathbb{T})$ one has the inequality*

$$\|h - J_n\|_{L^\infty(\mathbb{T})} \leq \frac{C\|h\|_{L^\infty(\mathbb{T})}}{n}$$

where the constant $C > 0$ is independent of h and n.

Consider a function g and define the function

$$h(x) = g(\lfloor |x| \rfloor), \quad x \in \mathbb{T}.$$

Then $g \in W^{1,\infty}([0,1])$ if and only if $h \in W^{1,\infty}(\mathbb{T})$. We will consider the convolution over the interval $[0,1]$

$$\overline{J_n}(\mu) = \overline{K_n} \star g = \int_0^1 \overline{K_n}(\mu, t) g(t) dt$$

where the kernel is

$$\overline{K_n}(\mu, t) = K_n(\mu - t) + K_n(\mu + t).$$

By construction, one has that $\overline{K_n} \star g(\mu) = K_n \star h(\mu)$ (for $\mu \in [0,1]$) and

$$\int_0^1 \overline{K_n} \star g(\mu) d\mu = \int_0^1 g(\mu) d\mu. \tag{4.39}$$

Definition 4.4.2. *The multivariate convolution of $g \in L^\infty([0,1]^m)$ is*

$$\overline{J^N}(x) = \int_{x \in [0,1]^m} \overline{K_N}(x^1 - t^1) \dots) \overline{K_N}(x^m - t^m)) g(t^1, \dots, t^m) dt$$

where $x = (x^1, \dots, x^m)$ and $\omega = (\omega^1, \dots, \omega^m)$, also written as $\overline{J^N} = \overline{K_N^1} \star \dots \overline{K_N^m} \star g$.

Corollary 4.4.3 (of the Jackson Theorem). *Take $g \in W^{1,\infty}([0,1]^m)$. Then*

$$\left\| \overline{J^N} - g \right\|_{L^\infty([0,1]^m)} \leq \frac{Cm\|\nabla g\|_{L^\infty[0,1]^m}}{N}.$$

Proof. We consider the two cases $m = 1$ then $m > 1$.

• **First case** $m = 1$**.** One has

$$\overline{J}^N(x) - g(x) = J_N(x) - g(x) = K_B \star h(x) - h(x) \qquad \text{for } x \in [0,1].$$

So the result follows from the Jackson Theorem.

• **Second case** $m > 1$**.** One has the telescopic decomposition

$$\begin{array}{rlll} \overline{J}^N - g = & & K_N^1 \star g_2 - g_2, & g_2 = K_N^2 \star \ldots K_N^m \star g \\ & + & K_N^2 \star g_3 - g_3, & g_3 = K_N^3 \star \ldots K_N^m \star g \\ & + & \ldots & \\ & + & K_N^{m-1} \star g_m - g_m, & g_m = K_N^m \star g \\ & + & K_N^m \star g - g. & \end{array}$$

All functions g_i have derivatives bounded by (4.38) that is

$$\max_{i=2}^{m} \|\nabla g_i\|_{L^\infty([0,1]^m)} \leq \|\nabla g\|_{L^\infty([0,1]^m)}.$$

Then one applies the result for $m = 1$ separately on each line. □

We now have the tools to evaluate the ratio accuracy/cost of the uniform sampling method. For $\mathbf{j} = (j_1, \ldots, j_m) \in \{0, \ldots, N-1\}^m$, we will write

$$\omega(\mathbf{j}) = \left(j_1 + \frac{1}{2}, \ldots, j_m + \frac{1}{2}\right) \in c(\mathbf{j}).$$

Theorem 4.4.4 (Accuracy of the uniform sampling method). *Assume $g \in W^{1,\infty}([0,1]^m)$. Then one has the inequality*

$$\left| I - \frac{1}{N^m} \sum_{\mathbf{j} \in \{0,\ldots,N-1\}^m} g(x(\mathbf{j})) \right| \leq \frac{2Cm\|\nabla g\|_{L^\infty([0,1]^m)}}{N}$$

where $x(\mathbf{j})$ is defined in (4.17).

Proof. One has the decomposition $I = I_1 + I_2$ where

$$I_1 = \int\limits_{[0,1]^m} \overline{J^{2N}}(x)dx \text{ and } I_2 = \int\limits_{[0,1]^m} (g - \overline{J^{2N}})(x)dx. \tag{4.40}$$

Also $\frac{1}{N^m} \sum_{\mathbf{j} \in \{0,\ldots,N-1\}^m} g(x(\mathbf{j})) = I_3 + I_4$ where

$$I_3 = \frac{1}{N^m} \sum_{\mathbf{j} \in \{0,\ldots,N-1\}^m} \overline{J^{2N}}(x(\mathbf{j})) \text{ and } I_4 = \frac{1}{N^m} \sum_{\mathbf{j} \in \{0,\ldots,N-1\}^m} (g - \overline{J^{2N}})(x(\mathbf{j})).$$

The corollary of the Jackson Theorem yields that

$$|I_3| + |I_4| \leq 2Cm\|\nabla g\|_{L^\infty[0,1]}/N.$$

The uniform sampling of a non trivial ($\delta_k^{2N} \neq 0$) cosine mode $e_k(x) = \cos(\pi \langle k, x\rangle)$ vanishes, that is $\frac{1}{N^m}\sum_{\mathbf{j}\in\{0,\dots,N-1\}^m} e_k(x(\mathbf{j})) = 0$. Then, with (4.37), one obtains $I_1 = I_2$ which ends the proof. □

The convergence of the Latin hypercube method is obtained in the next result, from which the convergence result of Loh [57][Theorem 3 page 2065] can be derived.

Theorem 4.4.5 (Accuracy of the Latin hypercube method)**.** *Assume* $g \in W^{1,\infty}([0,1]^m)$*. Then one has the inequality*

$$\left(\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_n}\left|I - \frac{1}{N}\sum_{x\in\mathbf{L}_N} g(x)\right|^2\right)^{\frac{1}{2}} \leq \frac{C}{\sqrt{N}}\|g\|_{L^\infty[0,1]^m} + \frac{2Cm}{N}\|\nabla g\|_{L^\infty[0,1]^m}.$$

Proof. • One already has the decomposition $I = I_1 + I_2$, see (4.40). Also $\frac{1}{N}\sum_{x\in\mathbf{L}_N} g(x) = I_5 + I_6$ where

$$I_5 = \frac{1}{N}\sum_{x\in\mathbf{L}_N}\overline{J^{2N}}(x) \text{ and } I_6 = \frac{1}{N}\sum_{x\in\mathbf{L}_N}(g - \overline{J^{2N}})(x).$$

• The corollary of the Jackson Theorem yields the estimate

$$|I_5| + |I_6| \leq 2Cm\|\nabla g\|_{L^\infty[0,1]}/N. \tag{4.41}$$

• Next one has $I_1 - I_5 = -\frac{1}{N}\sum_{x\in\mathbf{L}_N}(\overline{J^{2N}}(x) - I_1)$ where

$$\overline{J^{2N}}(x) - I_1 = \sum_{0\neq\mathbf{k}\in\{0,\dots,2N-1\}^m} \alpha_{\mathbf{k}}\cos(\pi\langle\mathbf{k}, x]),$$

where the Fourier mode $\mathbf{k} = 0$ is not present because the Jackson convolution is exact for constants. Using the orthogonality of the cosine functions

$$\int_{[0,1]^m}\cos(\pi\langle\mathbf{k}, x])\cos(\pi\langle\mathbf{k}', x])dx = 0, \qquad \mathbf{k}\neq\mathbf{k}',$$

one has the bound

$$\sum_{0\neq\mathbf{k}\in\{0,\dots,2N-1\}^m}|\alpha_{\mathbf{k}}|^2\int_{[0,1]^m}\cos(\pi\langle\mathbf{k}, x])^2dx \leq \int_{[0,1]^m}\overline{J^{2N}}(x)^2dx \leq \int_{[0,1]^m} f^{\text{obj}}(x)^2dx.$$

The condition $0 \not\equiv \mathbf{k} \in \{0,\dots,2N-1\}^m$ is interpreted modulo $2N$. Since one always has $\int_{[0,1]^m}\cos(\pi\langle\mathbf{k}, x])^2dx \geq \frac{1}{2}$, one obtains $\sum_{0\not\equiv\mathbf{k}\in\{0,\dots,2N-1\}}|\alpha_{\mathbf{k}}|^2 \leq$

$2\int_{[0,1]^m} g(x)^2 dx$.

- One can now use the orthogonality (4.35) of the cosine mode for the scalar product defined on the Latin structure and the bound (4.31)

$$\sum_{\mathbf{L}_N\in\mathcal{L}_n} |\cos(\pi\langle\mathbf{k},x])|^2 \le \frac{2}{N}, \quad 0\neq\mathbf{k}\in\{0,\dots,2N-1\}^m.$$

It yields

$$\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_n}\left|\frac{1}{N}\sum_{x\in\mathbf{L}_N}\sum_{0\neq\mathbf{k}\in\{0,\dots,2N-1\}^m}\alpha_{\mathbf{k}}\cos(\pi\langle\mathbf{k},x])\right|^2 \tag{4.42}$$

$$\le \frac{C}{N}\sum_{0\neq\mathbf{k}\in\{0,\dots,2N-1\}}|\alpha_{\mathbf{k}}|^2 \le \frac{2C}{N}\int\limits_{[0,1]^m} g(x)^2 dx \le \frac{2C}{N}\|g\|^2_{L^\infty([0,1]^m)}.$$

- From (4.41) and (4.42), one finally obtains

$$\left(\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_n}\left|I-\frac{1}{N}\sum_{x\in\mathbf{L}_N}g(x)\right|^2\right)^{\frac{1}{2}} \le \left(\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_n}|(I_1-I_5)+(I_2-I_6)|^2\right)^{\frac{1}{2}}$$

$$\le \left(\frac{1}{(N!)^{m-1}}\sum_{\mathbf{L}_N\in\mathcal{L}_n}|(I_1-I_5)+(I_2-I_6)|^2\right)^{\frac{1}{2}} + \frac{2Cm}{N}\|\nabla g\|_{L^\infty[0,1]^m}$$

$$\le \frac{2C}{N} + \|g\|^2_{L^\infty([0,1]^m)} + \frac{2Cm}{N}\|\nabla g\|_{L^\infty[0,1]^m}$$

which is the first claim. □

Lemma 4.4.6. *Assume g is a function of just one variable, for example $g(\mathbf{x}) = h(x_1)$, with the regularity $g\in W^{r,\infty}([0,1]^m)$ with $r=1$ or $r=2$. Then there exists $C>0$ such that*

$$\left|I-\frac{1}{N}\sum_{x\in\mathbf{L}_N}g(x)\right| \le \frac{C}{N^r}\|\nabla g\|_{L^\infty[0,1]^m} \quad \forall\mathbf{L}_N\in\mathcal{L}_N.$$

Proof. For $r=1$, this is a corollary of Lemma 4.4.4 for the uniform sampling in dimension $m=1$. For $r=2$, the function g is twice differentiable in L^∞. The points of quadrature (4.17) are at the barycenter of the small hypercubes, so one can apply the standard error bound for the mi-point rule integrator, which explain the $1/N^{-2}$ in the result.

□

A way to evaluate the result of Lemma 4.4.4 is by considering the cost of implementation of the dataset set obtained from uniform sampling. It is proportional to the cardinality of the dataset, that is $\#\left(\mathcal{D}_{\text{samp}}\right) = O(N^m)$. The accuracy is $\varepsilon = O(N^{-1})$. Therefore one has the scaling

$$\#\left(\mathcal{D}_{\text{samp}}\right) = O\left(\varepsilon^{-m}\right)$$

which shows once again that the method of uniform sampling is plagued with the curse of dimension.

On the other hand the implementation cost of one Latin hypercube scales as $\#\left(\mathcal{D}_{\text{Latin}}\right) = O(N)$. It is reasonable to think that the accuracy is similar to the mean accuracy, that is $\varepsilon = O(N^{-\frac{1}{2}})$. Therefore one has the scaling

$$\#\left(\mathcal{D}_{\text{Latin}}\right) = O\left(\varepsilon^{-2}\right).$$

Since this scaling is independent of the dimension m, it shows that the Latin hypercube method does not suffer from the curse of dimension.

4.5 Lattice rules

Remarks 4.3.4 and 4.3.7 and Lemma 4.4.5 explain that the Latin hypercube sampling method behaves as uniform sampling for one-dimensional functions with a $O(N^{-1})$ rate of convergence, while it behaves as the Monte-Carlo method for general functions with a $O(N^{-1/2}) \gg O(N^{-1})$ rate of convergence. It is then a natural question to determine if there exists sampling methods which are optimal for the numerical integration of two-dimensional or three-dimensional functions, and which offer the rate of convergence of the Monte-Carlo method for general functions. It seems to be a difficult question to answer in full generality. Instead we show how **Lattice rules**, which are due to Sloan [39, 36, 91] based on ideas from number theory by Korobov, can be used to construct such samplings, at least when N is a prime number reasonably large with respect to m which is not a restriction for practical uses. It might seem strange that number theory is evoked in relation with the construction of datasets motivated by Machine Learning methods, but this fact just illustrates the profound unity of mathematics.

Once again we favor an elementary presentation. A preliminary remark is that some permutations $\sigma \in S_N$ can be constructed with a linear representation. Indeed take two integer numbers $a, b \in \mathbb{Z}$ and consider

$$\begin{array}{rccc} \sigma: & \{0,\dots,N-1\} & \to & \{0,\dots,N-1\} \\ & i & \mapsto & \sigma(i) \equiv ai + b \bmod N. \end{array} \tag{4.43}$$

Since the equality holds modulo N, we will also use, depending on the context, the notation $\mathbb{Z}/N\mathbb{Z}$ instead of $\{0, \ldots, N-1\}$. It is a basic fact in number theory that if a and N are prime to each other, then the Bezout identity holds: there exists $\alpha, \beta \in \mathbb{Z}$ such that $\alpha a + \beta N = 1$ which means that a is invertible in $\mathbb{Z}/N\mathbb{Z}$ (the inverse is α mod N). Still with the assumption that a and N are prime to each other, a consequence is that σ defined by (4.43) is a permutation. Indeed $\sigma(i) \equiv \sigma(i')$ mod N is equivalent to

$$ai + b \equiv ai' + b \bmod N \Leftrightarrow a(i - i') = 0 \bmod N \Leftrightarrow i - i' = 0 \bmod N.$$

So σ is injective in a finite set, so it is bijective and $\sigma \in S_N$.

A generalization in dimension $r \geq 1$ considers

$$\begin{array}{cccc} \sigma: & (\mathbb{Z}/N\mathbb{Z})^r & \to & (\mathbb{Z}/N\mathbb{Z})^r, \\ & \mathbf{j} = (i_1, \ldots, i_r) & \mapsto & \sigma(\mathbf{j}) \equiv A\mathbf{j} + \mathbf{b} \bmod N \end{array} \tag{4.44}$$

where $A \in \mathcal{M}_r(\mathbb{Z})$ and $\mathbf{b} \in \mathbb{Z}^r$. The equality modulo N is interpreted component wise, that is in $(\mathbb{Z}/N\mathbb{Z})^r$.

Lemma 4.5.1. *Assume det(A) and N are prime to each other. Then σ is a permutation in $(\mathbb{Z}/N\mathbb{Z})^r$.*

Proof. In the finite set $(\mathbb{Z}/N\mathbb{Z})^r$, it is sufficient to check that σ is injective to obtain the claim. Consider $\mathbf{j}, \mathbf{j}' \in (\mathbb{Z}/N\mathbb{Z})^r$ such that $\sigma(\mathbf{j}) = \sigma(\mathbf{j}')$, that is

$$A\mathbf{j} + \mathbf{b} \equiv A\mathbf{j}' + \mathbf{b} \bmod N \Leftrightarrow A(\mathbf{j} - \mathbf{j}') \equiv 0.$$

From the Cramer's rule, the transpose of the comatrix $B \in \mathcal{M}_r(\mathbb{Z})$ verifies $BA = \det(A)I_r$ where I_r is the identity matrix. One obtains $\det(A)(\mathbf{j}-\mathbf{j}') \equiv BA(\mathbf{j}-\mathbf{j}') \equiv 0$ mod N. Since $\det(A)$ is prime with N, then it has an inverse which implies that $\mathbf{j} - \mathbf{j}' \equiv 0$ mod N. □

Following Sloan, we define Lattice rules in the m-dimensional hypercube .

Definition 4.5.2 (Lattice rules)**.** *Let $1 \leq r \leq m$. One considers the affine transformation*

$$\begin{array}{cccc} L: & (\mathbb{Z}/N\mathbb{Z})^r & \to & (\mathbb{Z}/N\mathbb{Z})^m, \\ & \mathbf{j} & \mapsto & L\mathbf{j} \equiv A\mathbf{j} + \mathbf{b} \textit{ mod } N \end{array}$$

where $A \in \mathcal{M}_{mr}(\mathbb{Z})$ and $\mathbf{b} \in \mathbb{Z}^m$.

Another notation [62] with the column vectors of the matrix A is

$$A\mathbf{j} + \mathbf{b} = \mathbf{z}_0 + j_1\mathbf{z}_1 + \cdots + j_r\mathbf{z}_r \tag{4.45}$$

where the column vectors are $\mathbf{z}_i \in \mathbb{Z}^m$ for $1 \leq i \leq r$ and $\mathbf{z}_0 = \mathbf{b}$.

Theorem 4.5.3. *Assume N and the determinants of all minor matrices of size r extracted from the matrix A are prime to each other. Then Span$(L) \subset \{0, \ldots, N-1\}^m$ is a list of N^r points in the hypercube $\{0, \ldots, N-1\}^m$ such that the projection of these points on all hyperfaces of dimension r is full (that is all points of the hyperface are reached).*

Remark 4.5.4. *The assumptions that all minor matrices have a determinant which is a prime with respect to N brings a restriction.*

If m is too large with respect to N, there might be no solutions. Indeed the matrix A a m line vectors of size 2. The numbers of different elements/vectors in $\{0, \ldots, N-1\}^2$ is equal to N^2. So if $m \geq N^2 + 1$, then necessarily two vectors will be equal so that the corresponding 2×2 determinant will be zero.

However the restriction is not severe in most practical situations. For example let us consider the case $m = 7$ and $r = 2$ and assume N is a sufficiently large prime number. One has to check $(m-1)(m-2) = 30$ primality conditions. Just picking A at random will most probably produce an admissible matrix. We refer to [64] for early considerations on the construction/implementation of good lattice rules.

Proof of the Theorem. By construction one has of course that $\#\left(\text{Span}(L)\right) \leq N^r$. Let us consider the projection of Span(L) on one hyperface of dimension r of the hypercube of dimension m. The hypothesis and Lemma 4.5.1 yields that the projection is made with exactly N^r different points (it shows in passing that $\#\left(\text{Span}(L)\right) = N^r$). It is the same for all hyperfaces with the same dimension. □

We will say that the rank of the Lattice is r. Now we consider the composition of a Lattice rule with a component-wise permutation

$$\sigma \circ L : \quad (\mathbb{Z}/N\mathbb{Z})^r \to (\mathbb{Z}/N\mathbb{Z})^m \tag{4.46}$$

where L is a Lattice rule and $\sigma \in S_N^m$. That is $\sigma = (\sigma_1, \ldots, \sigma_m)$ and all permutations σ_j are applied component wise to $L\mathbf{j}$ for all $\mathbf{j} \in (\mathbb{Z}/N\mathbb{Z})^r$. We will refer to (4.46) as a Lattice rule with randomization [48][page 755], or as a randomized Lattice rule.

Corollary 4.5.5 (of Theorem 4.5.3)**.** *Make the same assumptions as in the Theorem. Then Span$(\sigma \circ L) \subset \{0, \ldots, N-1\}^m$ is a list of N^r points in the hypercube $\{0, \ldots, N-1\}^m$ such that the projection of these points on all hyperfaces of dimension r is full (that is all points of the hyperface are reached).*

Proof. Indeed Span$(\sigma \circ L) = \sigma\left(\text{Span}(L)\right)$ so the result follows from the characterization of Span(L) provided by Theorem 4.5.3. □

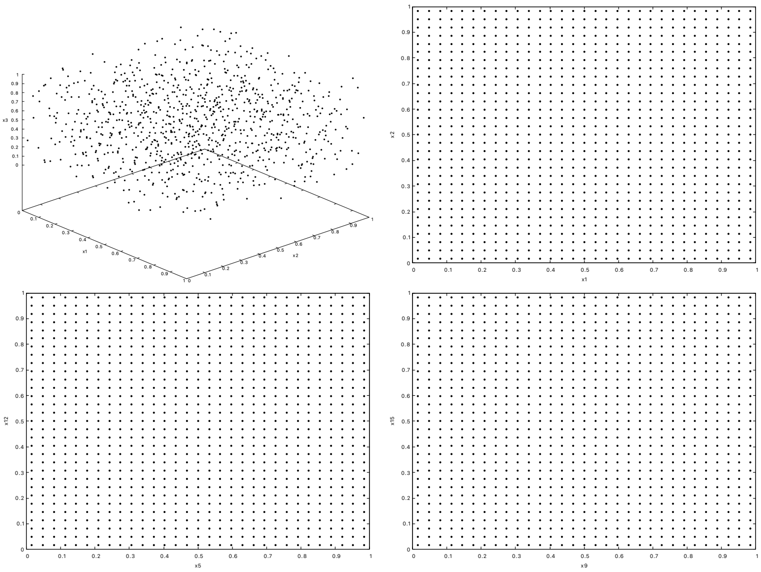

Fig. 4.2: Plot of the points generated by a randomized Lattice rule $\sigma \circ L$ in dimension $m = 15$ with $r = 2$ and $N = 31$. Top left, the plot is displayed along x_1, x_2, x_3. Top right, bottom left and bottom right: the plots arbitrarily displayed along (x_1, x_2), (x_5, x_{12}) or (x_9, x_{15}) are cartesian 2D grids

Lemma 4.5.6. *Take $r = 1$ and make the same assumptions as in Theorem 4.5.3. Then Span$(\sigma \circ L) \in \mathcal{L}_N$, that is $\sigma \circ L$ defines a Latin hypercube (4.18).*

Proof. With Sloan notations (4.45), all components of the vector $\mathbf{z}_1$ are non zero (module N). One can take as well $\mathbf{z}_0 = 0$, $\mathbf{z}_1 = (1, \ldots, 1)$ and $\sigma = (I, \sigma_2, \ldots, \sigma_m)$ where I is the identity permutation, then one obtains the parametrization (4.18). There is of course no restriction on m like the ones in Remark 4.5.4 which concerns $r \geq 2$. □

Considering Theorem 4.4.5, the accuracy of Lattice rules is less immediate to discuss. In particular the generalization of the orthogonality properties (4.21)-(4.27) of Fourier modes are not known for Lattice rules. Similarly the orthogonality properties (4.31)-(4.35) of cosine modes are not known. The difficulty is the establishment of the list of all possible Lattice rules. It is nevertheless possible to generalize Lemma 4.4.6. Assume g is a function of just two variables, for example $g(\mathbf{x}) = h(x_1, x_2)$, with the regularity $g \in W^{r,\infty}([0,1]^m)$ with $r = 1$ or $r = 2$. Then there exists $C > 0$

such that

$$\left| I - \frac{1}{N^2} \sum_{x \in \mathrm{Span}(\sigma \circ L)} g(x) \right| \leq \frac{C}{N^r} \|\nabla g\|_{L^\infty[0,1]^m}. \tag{4.47}$$

The constant C is independent of $\sigma \in S_N^m$ and of the affine function L, because the projection of points on a face is a 2D cartesian grid (corollary 4.5.5).

The accuracy of Lattice rules and randomized Lattice rules can also be discussed in relation with the smoothness of the terms of the ANOVA decomposition of a function, refer to [10, 35].

4.6 Numerical application

To illustrate the approximation properties of the different sampling methods, we consider two different objective functions.
The first one $f_1^{\mathrm{obj}} \in C^\infty([0,1]^m)$ is defined by

$$f_1^{\mathrm{obj}}(\mathbf{x}) = e^{x_1 + \cdots + x_m} = e^{x_1} \ldots e^{x_m}$$

The exact value of the integral is obtained by separation of variables $I_1 = \int_{[0,1]^m} f_1^{\mathrm{obj}}(\mathbf{x}) dx = \left(\int_0^1 e^y dy \right)^m = (e-1)^m$.
The second function is less regular $f_2^{\mathrm{obj}} \in W^{1,\infty}([0,1]^m)$

$$f_2^{\mathrm{obj}}(\mathbf{x}) = w(\mathbf{x}) \, |x_1 + \cdots + x_m - 1|$$

where w is piecewise constant: $w(\mathbf{x}) = \alpha$ if $x_1 + \cdots + x_m \leq 1$, and $w(\mathbf{x}) = \beta$ if $1 < x_1 + \cdots + x_m$. It can be assembled with ReLU. The integral is noted $I_2 = \int_{[0,1]^m} f_2^{\mathrm{obj}}(\mathbf{x}) dx$.

Lemma 4.6.1. $I_2 = \frac{\alpha+\beta}{(m+1)!} + \beta \left(\frac{m}{2} - 1 \right)$.

Proof. One decomposes $I_2 = \alpha U_m + \beta V_m$ where $U_m = \int_{\{\sum_{i=1}^m x_i \leq 1\} \cap [0,1]^m} (1 - \sum_{i=1}^m x_i) dx$ and $V_m = \int_{\{1 < \sum_{i=1}^m x_i\} \cap [0,1]^m} (\sum_{i=1}^m x_i - 1) dx$.
One has

$$U_m = \int_{x_1=0}^{1} \left(\int_{\{\sum_{i=2}^m x_i \leq 1 - x_1\} \cap [0,1]^{m-1}} (1 - x_1 - \sum_{i=2}^m x_i) dx_2 \ldots dx_m \right) dx_1$$

A change a variable $x_i = (1 - x_1) y_i$ for $2 \leq i \leq m$ yields

$$U_m = \int_{x_1=0}^{1} \left((1 - x_1)^m \int_{\{\sum_{i=2}^m y_i \leq 1\} \cap [0,1]^{m-1}} (1 - \sum_{i=2}^m y_i) dy_2 \ldots dy_m \right) dx_1,$$

that is $U_m = \int_{x_1=0}^{1}(1-x_1)^m dx_1 U_{m-1} = \frac{1}{m+1}U_{m-1}$. Since $U_1 = \frac{1}{2}$, one gets $U_m = \frac{1}{(m+1)!}$.
A relation is $V_m - U_m = \int_{[0,1]^m}(x_1 + \cdots + x_m - 1)dx = \frac{m}{2} - 1$, from which one obtains $V_m = \frac{m}{2} - 1 + \frac{1}{(m+1)!}$. The claim is obtained by linear combination. □

An illustration of the accuracy of the various algorithms is displayed in Figure 4.3. We plot the density function which is smooth version of the histogram representation. The plots reveal the dispersion of the result around the exact value I_1 or I_2 which is the abscissa at the peak. We consider four algorithms which the Monte-Carlo method, the Monte-Carlo method where the points are projected on the grid, the Latin hypercube method (LH) and the randomized Lattice rule method with $r = 2$ (as in Figure 4.2) where we pick a matrix A and a vector $\mathbf{b}$ at random, followed with another randomization on all axis. The value of N is $N = 71$ (a prime number) for the randomized Lattice rule method, and $N = 71^2$ for both Monte-Carlo methods and the Latin hypercube method, so that the total number of quadrature points is the same.

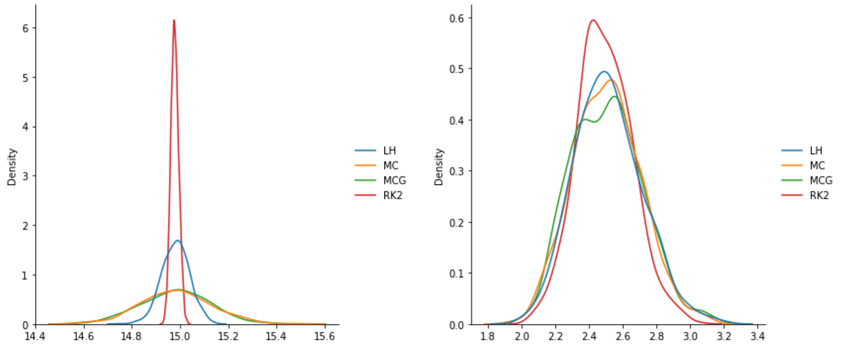

Fig. 4.3: Plots of density function of the numerical evaluation of $I_1 \approx 14.96$ (on the left) and $I_2 = 2.5$ (on the right with $\alpha = (m+1)! - 1$ and $\beta = 1$). The dimension $m = 5$. The four algorithms are the Monte-Carlo method (MC), the Monte-Carlo method projected on the grid (MCG), the Latin hypercube method (LH) and the randomized Lattice rule method of rank $r = 2$ (RK2).

For the smooth objective function f_1^{obj}, the randomized Lattice rule method provide much less dispersion around the exact value compared with the three other methods. It can be explained with inequality (4.47) which yields that the low complexity part of f_1^{obj} is calculated with the accuracy provided by a 2D cartesian grid, while the remaining part seems to be calculated with the accuracy similar to a Monte-Carlo method. The same remark holds for the results provided by the

Latin hypercube method which in between the randomized Lattice rule method and the two Monte-Carlo methods. For the other less smooth objective function $f_2^{\rm obj}$, the results are much less pronounced in favor of the randomized Lattice rule method, which nevertheless performs the best since the dispersion is less.

Remark 4.6.2. *The problem of deciding of the best method to use for a given practical problem is difficult to address. Generally the choice is made on a combination of theoretical and practical considerations. The conclusion in [11] is a good example of such a discussion.*

4.7 Summary of the chapter

The framework considered in this chapter is the construction of datasets for the modeling of numerical inverse problems.The practical issue is the sampling of inputs and outputs in spaces of large dimension. Then a possibility is to interpret the cost function is the approximation of an integral in $[0,1]^m$ with $m \gg 1$.

Four methods were discussed. Uniform sampling is plagued with the curse of dimension and is of not real use. The Monte-Carlo method, justified by probability arguments, is a standard way to circumvent the curse of dimension. The Latin hypercube method and randomized Lattice rules are discussed as way to sample with a methodology which is in between the uniform sampling method and the Monte-Carlo method.

Next chapter is dedicated to the remaining practical issue which is the optimization of the numerical parameters, which is called the training.