No Free Lunch Theorem: A Review

Stavros P. Adam, Stamatios-Aggelos N. Alexandropoulos, Panos M. Pardalos, and Michael N. Vrahatis

Abstract The "No Free Lunch" theorem states that, averaged over all optimization problems, without re-sampling, all optimization algorithms perform equally well. Optimization, search, and supervised learning are the areas that have benefited more from this important theoretical concept. Formulation of the initial No Free Lunch theorem, very soon, gave rise to a number of research works which resulted in a suite of theorems that define an entire research field with significant results in other scientific areas where successfully exploring a search space is an essential and critical task. The objective of this paper is to go through the main research efforts that contributed to this research field, reveal the main issues, and disclose those points that are helpful in understanding the hypotheses, the restrictions, or even the inability of applying No Free Lunch theorems.

1 Introduction

Optimization problems occurring in various fields of science, computing, and engineering depend on the number of parameters, the size of the solution space and, mainly, on the objective function whose definition is critical as it largely determines

S. P. Adam
Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece

Computational Intelligence Laboratory – CILab, Department of Mathematics, University of Patras, Patras, Greece
e-mail: adamsp@upatras.gr

S.-A. N. Alexandropoulos · M. N. Vrahatis (✉)
Computational Intelligence Laboratory – CILab, Department of Mathematics, University of Patras, Patras, Greece
e-mail: alekst@math.upatras.gr; vrahatis@math.upatras.gr

P. M. Pardalos
Department of Industrial & Systems Engineering, University of Florida, Gainesville, FL, USA
e-mail: pardalos@ufl.edu

I. C. Demetriou, P. M. Pardalos (eds.), *Approximation and Optimization*,
Springer Optimization and Its Applications 145,
https://doi.org/10.1007/978-3-030-12767-1_5

the level of difficulty of the problem. Hence, defining and solving an optimization problem is sometimes an extremely difficult and demanding task. Researchers from various fields have been involved in solving optimization problems either as this constitutes part of their main research or because the problem they face can be tackled by an optimization one. The research efforts on this matter have permitted the elaboration of numerous methods and techniques, built on solid mathematical concepts, whose application produced significantly good results.

However, contrary to any opposite claim, none of these methods has proven to be successful to all types of the problems it was applied. This argument has been the objective of important theoretical work carried out by David Wolpert which gave rise to the well-known *No Free Lunch* (NFL) theorem. Briefly, the NFL theorem states that: *"averaged over all optimization problems, without re-sampling all optimization algorithms perform equally well."* Besides optimization, the NFL theorem has been successfully used to tackle important theoretical issues pertaining supervised learning in machine learning systems. Actually, the NFL theorem has become a suite of theorems which has given significant results in various scientific fields where searching for some optimal solution is an important issue.

The NFL theorems constitute an important theoretic development which marked the limits of the range of successful application for a number of search, optimization, and supervised learning algorithms. At the same time the formulation of these theorems has provoked controversial discussions [4, 36, 44, 45] regarding the possibility to invent and effectively use general purpose algorithms in various fields where only a limited view of the real-world problem exists.

In this paper we aim at presenting a review on the most sound research work published by several researchers on this matter including its impact on the most important fields, that is, optimization and supervised learning. Other existing fields of interest such as user interface design [24], network calculus [8] are worth of merit but they are out of the scope of this review. The emphasis of this review will be, mainly, on the critical questions which promoted the development of NFL theorems as well as on the issues that proved to be important: namely for (a) *optimization*, (b) *searching*, and (c) *supervised learning*.

The rest of this paper is structured as follows. Section 2 provides a review of the early concepts and constructs that underpinned the definition of the NFL theorems. Section 3 covers the main research efforts of Wolpert establishing NFL for optimization and search. In Section 4 we survey the more recent work of Wolpert which clarifies older concepts while offering some new results on this field. Next, Section 5 is dedicated to the main research carried out by several researchers on NFL for optimization and evolutionary algorithms. Part of the research surveyed concerns the cases where NFL theorems do not apply and researchers have proved the existence of *"Free Lunches."* In Section 6 we describe the main research efforts on NFL theorems for supervised learning. The paper ends in Section 7 with a synopsis and some concluding remarks.

2 Early Developments

As noted by David Wolpert [56], the first attempt to underline the limits of inductive inference was made by the Scottish philosopher David Hume in 1740 in his seminal work *"A treatise of human nature"* [26, 27]. Hume wrote that:

> Even after the observation of the frequent conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience.

In the machine learning context this can be stated as follows:

> It is not reasonable to believe that the generalization error of a classifier-generalizer on test data drawn off the training set correlates with its performance on the training set itself by simply considering a priori information on the real world.

Wolpert based his theoretical work on earlier developments elaborated in his paper *"On the connection between in-sample testing and generalization error"* [55]. In this paper the generalization error is taken as the *off-training set* (OTS) error and the question addressed concerns its correlation with the error produced using in-sample testing. Moreover, Wolpert tackles the question of how *"...to take into account the probability distribution of target functions in the real world"* as any theory of generalization is irrelevant concerning its applicability on real- world problems if it does not tackle the previous problem. Some, but not all, of the important issues arising in this paper are:

(a) *"Can one prove inductive inference from first principles?"* In other words, given the performance of a learning algorithm on the training data set is it possible to obtain information on its ability to provide an exact representation of the target function for examples outside the data set?
(b) If one cannot answer the previous question then, what are the assumptions on the distribution of real-world data (the target function) can help with the generalization for training algorithms, such as back-propagation, which aim to minimize the error on the training data?
(c) Is there a mathematical basis of estimating when over-training occurs and proceed in modifying the learning algorithm in order to bound the effects of such over-training?
(d) Is it possible to express in mathematical terms the ability of a training set to faithfully represent the distribution over the entire data space?
(e) What are the hypotheses under which non-parametric statistics techniques, such as cross-validations, which are designed to choose between learning algorithms, succeed to diminish the generalization error?

In addressing these matters, the formalism proposed seems to extend the classical Bayesian formalism using the hypothesis function, i.e., the distribution of the data set as learned by the generalizer. The mathematical formalism adopted proposes a way to match the degree to which the distribution derived by the learning algorithm matches the distribution of the training data and it can be used to tackle various

generalization issues such as over-training and minimum number of parameters for the model. From another point of view this formalism is proposed with the aim to express in mathematical terms the assumptions made by a generalizer so that the used model best fits the training set representing the real world. As a result the elaboration of important theoretical proofs proposes a solid basis for tackling several issues in machine learning and gives rise to the development of concepts such as the NFL theorems.

The first and foremost contributions of Wolpert concerning NFL theorems were presented in the papers [56, 57]. In this set of two papers, namely:

(i) *"The lack of a priori distinctions between learning algorithms"* and
(ii) *"The existence of a priori distinctions between learning algorithms,"*

Wolpert develops his theory and formulates the NFL theorems. In the former, he discusses the hypothesis that given any two learning algorithms one cannot claim having any prior information that these algorithms are distinct as far as the performance of these algorithms on specific class of problems is concerned. In the latter paper, Wolpert unfolds the arguments concerning the inverse assumption, i.e., there are prior distinctions regarding the performance of any two algorithms. These two papers deal with supervised learning but the theoretical constructs were applied to multiple domains where two different algorithms compete as for which performs better for a class of problems and associated error functions.

Focusing on supervised learning, in the first of the previously mentioned papers the concept of *"off-training set"* (OTS) is defined and the associated performance measure of the supervised learning algorithm is proposed. The mathematical formalism used is based on the so-called *extended Bayesian formalism* and is refined in order to take into account the generalization error, the cost function, and their relation to the learning algorithm while providing the necessary hypotheses for the training sets and the targets. In the sequel the probability of some cost "c" of the learning algorithm associated with the loss function is proposed as follows:

$$P\,(c|d) = \int df\, dh\, P\,(h|d)\;\; P\,(f\,|d)\;\; M_{c,d}\,(f,h)\,,$$

which is considered to be the inner product between the infinite dimensional vectors $P(f|d)$ and $P(h|d)$ representing the target and the hypothesis functions, respectively. This inner product quantity is maximized if the target function f and the hypothesis function h given the training data d are close enough to each other, i.e., they are aligned. Given two learning algorithms (generalizers) A and B, an important question to be answered deals with the comparison of these two algorithms in terms of how the set F_1 of target functions f for which A beats B compares with the corresponding set F_2 of the target functions f for which algorithm B outperforms A. As it is stated in [56]: *"in order to analyze this issue it is proposed to compare the average over f of f-conditioned probability distributions for algorithm A to the same average for algorithm B. Then the relationship between these two averages is used to compare the sets F_1 and F_2 ."*

In the second paper: *"The existence of a priori distinctions between learning algorithms"* [57] Wolpert, besides revisiting the theorems and some examples of the first paper, examines the NFL theorems with respect to cross-validation and the so-called *head-to-head minimax* behavior that is the case where for an algorithm A there exist comparatively few target functions for which A is slightly worse than algorithm B and comparatively few target functions in which algorithm A is superior to algorithm B. Moreover, he develops an extension of his theory by considering averaging over generalizers rather than targets. This means that instead of characterizing two algorithms by averaging over targets, namely f, ϕ, $P(f)$, or $P(\phi)$, holding the hypothesis, $P(h|d)$, fixed it is tentative to consider alternative results where one holds one of the entities concerning the targets, fix and average over the hypothesis entities. For this case, Wolpert formulates some additional theorems and finally he examines the case when the loss function $L(\cdot|\cdot)$ is non-homogenous and thus the NFL theorems do not apply as one can make a priori distinctions between algorithms.

As a conclusion it is stated in [57] that these two papers investigate some of the behavior of OTS error. In particular, they formalize and investigate the concept that *"if you make no assumptions concerning the target, then you have no assurances about how well you generalize."*

3 No Free Lunch for Optimization and Search

Another direction of research for applying the ideas of the NFL theorems, as presented above, concerns the domain of optimization. The work *"No free lunch theorems for optimization"* [62] published by Wolpert and McReedy deals with this matter based on two technical reports produced by the authors at the Santa Fe Institute. The first technical report published in [35] with the title *"What makes an optimization problem hard?"* raises the question: *"Are some classes of combinatorial optimization problems intrinsically harder than others, without regard to the algorithm one uses, or can difficulty be assessed only relative to a particular algorithm?"* The second technical report [61], entitled: *"No free lunch theorems for search"* focuses on proving that all algorithms searching for an optimum of an optimization problem, i.e., an extremum of an objective function, performs exactly the same, no matter the performance measure used, when taking the average over all possible objective functions.

The work of Wolpert and McReedy *"No free lunch theorems for optimization"* [62], sets up a formalism for investigating the relation of the effectiveness of optimization algorithms and the problems they are solving. The NFL theorems developed in the paper establish that the successful performance of any optimization algorithm on one class of problems is counterbalanced by its degraded performance on another class of problems. A geometric interpretation is provided concerning the meaning of the fitness of an algorithm to cope with some optimization problem. Moreover, as mentioned in the previous technical reports the authors examine

applications of NFL theorems to information-theoretic aspects of optimization as well as to defining measures of performance for optimization benchmarks.

Given the multitude of black-box optimization techniques available, the authors try to provide the formalism for tackling the following problem: *"is there a relationship between how well an algorithm performs and the optimization problem on which it is run?"* This problem can be cast in several other such as:

(a) What are the mathematical constituents of optimization theory one needs to know before deciding on the necessary probability distributions to be applied?
(b) Are information theory and Bayesian analysis suitable for understanding the previous issues?
(c) Given the performance results of a certain algorithm on a certain class of problems can one provide a priori generalization of these results on other classes of problems?
(d) Is there a suitable measure of such generalization? Can one evaluate the performance of algorithms on problems so that he is able to compare those algorithms?

The formalism developed by the authors is articulated around the following concepts:

(i) A sample of size m is a set of m distinct points visited by the algorithm and is denoted by

$$d_m = \left\{\left(d_m^x(1),\, d_m^y(1)\right),\ \left(d_m^x(2),\, d_m^y(2)\right),\ \ldots,\ \left(d_m^x(m),\, d_m^y(m)\right)\right\},$$

where $d_m^x(i)$ denotes the $\mathscr{X}$ value of the i th element of the sample and $d_m^y(m)$ is the associated cost, i.e., the $\mathscr{Y}$ value.

(ii) An optimization algorithm α is a mapping from previously visited sets of points to a single new point in $\mathscr{X}$, i.e.,

$$\alpha :\ d \in \mathscr{D} \rightarrow \left\{x \mid x \notin d^x\right\},$$

where $\mathscr{D}$ denotes the space of all (m-sized) samples and α is deterministic in the sense that every sample maps to a unique new point.

(iii) The performance of an algorithm after m iterations is a function $\Phi(d_m^y)$ of the sample.

(iv) Given the space of all cost functions, i.e., optimization problems $\mathscr{F}$ the distribution:

$$P(f) = P\left(f(x_1),\ f(x_2),\ \ldots,\ f(x_{|\mathscr{X}|})\right),$$

defined over $\mathscr{F}$ gives the probability that each $f \in \mathscr{F}$ is the actual optimization problem at hand.

(v) The performance of an optimization algorithm α on a cost function f after m iterations is measured with $P(d_m^y \mid f, m, \alpha)$.

Let us consider the problem:

> Suppose that $F_1 \subseteq \mathscr{F}$ is the set of problems for which an algorithm α_1 performs better than algorithm α_2 and $F_2 \subseteq \mathscr{F}$ denotes the set for which α_2 performs better than α_1. How can one compare these two sets?

The answer provided by the authors relies on the sum of $P(d_m^y \mid f, m, \alpha_1)$ and the sum of $P(d_m^y \mid f, m, \alpha_2)$ over all f, i.e., over all problems. The following theorem as formulated in this paper addresses the previous problem.

Theorem 1 *For any pair of algorithms α_1 and α_2,*

$$\sum_f P\left(d_m^y \mid f, m, \alpha_1\right) = \sum_f P\left(d_m^y \mid f, m, \alpha_2\right).$$

In the theorem the problem is considered to be fixed over time. If the cost function is time-dependent in the sense that, while the problem is initially expressed with some cost function f_1 which is present when sampling the first value in $\mathscr{X}$, then this function is deformed before any subsequent iteration of the optimization algorithm. If deformation is represented with the mapping $T : \mathscr{F} \times \mathscr{N} \to \mathscr{F}$, and $T = T_i$, then $f_{i+1} = T_i(f)$ and the following theorem can be formulated:

Theorem 2 *For all d_m^y, D_m^y, $m > 1$, algorithms α_1 and α_2, and initial cost functions f_1*

$$\sum_T P(d_m^y \mid f_1, T, m, \alpha_1) = \sum_T P(d_m^y \mid f_1, T, m, \alpha_2),$$

and

$$\sum_T P(D_m^y \mid f_1, T, m, \alpha_1) = \sum_T P(D_m^y \mid f_1, T, m, \alpha_2).$$

One of the implications of the NFL theorems discussed by the authors deals with the geometric perspective of NFL. In this perspective consider the space $\mathscr{F}$ of all cost functions and the probability of obtaining a certain d_m^y defined by the relation:

$$P(d_m^y \mid m, \alpha) = \sum_f P(d_m^y \mid m, \alpha, f)\, P(f),$$

with $P(f)$ being the prior probability that the optimization problem at hand has cost function f. As noted by the authors the previous sum can be considered as an inner product in $\mathscr{F}$. Hence, if we define the vectors $\overrightarrow{v}_{d_m^y,\alpha,m}$ and $\overrightarrow{p}$ by their f components, respectively:

$$\overrightarrow{v}_{d_m^y,\alpha,m}(f) \equiv P(d_m^y \mid m, \alpha, f), \quad \text{and} \quad \overrightarrow{p} \equiv P(f),$$

then it holds that:

$$P(d_m^y \mid m, \alpha) = \overrightarrow{v}_{d_m^y,\alpha,m} \cdot \overrightarrow{p} \ .$$

The authors note that this equation provides a geometric interpretation of the optimization process. Hence, d_m^y represents the desired sample and m is taken as a measure of the computational effort needed for the algorithm. Moreover, if the vector $\overrightarrow{p}$ represents the prior which includes all knowledge about the cost functions, then the last equation formulates in mathematical terms that: "*the performance of an algorithm is determined by the magnitude of its projection on* $\overrightarrow{p}$ *or in other words by how aligned* $\overrightarrow{v}_{d_m^y,\alpha,m}$ *is with the problem's vector* $\overrightarrow{p}$." With respect to the geometric view the NFL result that $\sum_f P(d_m^y \mid f, m, \alpha)$ is independent of α means that for any particular d_m^y and m, all algorithms have the same projection onto the uniform $P(f)$ represented by the diagonal vector $\overrightarrow{1}$.

Moreover, the authors investigate the relationship of the above results with information-theoretic aspects of optimization and provide measures of performance for assessing the efficacy of a certain optimization algorithm. Finally, minimax distinctions between search algorithms are discussed and some performance measures for search algorithms are provided.

4 More Recent Work of Wolpert

The work of Köppen, Wolpert, and McReedy *"Remarks on a recent paper on the No Free Lunch Theorems"* [33] is a letter reconsidering a previous work of Köppen [32] with the title *"Some technical remarks on the proof of the No Free Lunch theorem."* In this letter the authors, following suggestions made in [32], provide a short proof of the NFL theorems while correcting a wrong claim made in [32] about circular reasoning of the original proof of the NFL theorems in [61, 62].

Hereafter, let us give some details on this theorem, as presented in [61, 62]; its proof is important for many papers on NFL theorems. First, consider two finite sets X and Y together with the set of all cost functions $f : X \to Y$. Moreover, for a positive integer m such that $m < |X|$ let $d_m = \{(d_m^x(i), d_m^y(i) = f(d_m^x(i)))\}$ i.e., the points sampled by the algorithm in m steps, with $i = 1, 2, \ldots, m$ $d_m^x(i) \in X$ $\forall x$ and for any i, j it holds that $d_m^x(i) \neq d_m^x(j)$. Let a denote the search algorithm of interest, which is a deterministic "blind" algorithm assigning to every possible d_m an element of X which is not already in the d_m^x. This means that,

$$d_{m+1}^x(m+1) = a[d_m] \notin \{d_m^x\}.$$

Let $Y(f, a, m)$ denote the sequence of the m values of Y produced by the algorithm a to f after m successive steps and $\delta(\cdot, \cdot)$ is the Kronecker delta function giving 1 if its arguments are identical and 0 otherwise. Then the following lemma holds:

Lemma 1 *For any algorithm a and any* d_m^y,

$$\sum_f \delta\big(d_m^y,\ Y(f,m,a)\big) \ = \ |Y|^{|X|-m}.$$

Thus, if $c(\cdot)$ is some performance measure assigning a real value to any set d_m^y and $k \in \mathbb{R}$ is a performance value, then the theorem in question is:

Theorem 3 *For any two deterministic algorithms a and b, any value* $k \in \mathbb{R}$, *and any performance measure* $c(\cdot)$,

$$\sum_f \delta\Big(k,\ c\big(Y(f,m,a)\big)\Big) \ = \ \sum_f \delta\Big(k,\ c\big(Y(f,m,b)\big)\Big).$$

Besides considering the proof of this theorem, in this letter, the authors take the chance to defend NFL theorems against what they call a rather nihilistic view that algorithms of universal applicability would not exist. NFL theorems should be considered as a research topic and not as simply some convenient or inconvenient result. Hence, they propose that a more open minded view should prevail in order to investigate the limits of NFL theorems as well as the potential issues arising by their application in various domains.

It is worth noting, here, a relatively more recent work of Wolpert [60] entitled: *"What the No Free Lunch Theorems Really Mean? How to Improve Search Algorithms?"* In this research report the author reconsiders the main ideas of his work on NFL as far as search algorithms are concerned. Wolpert insists on analyzing the issue that while the NFL theorems have strong implications whenever a uniform distribution of the cost function over the optimization problems is adopted, this is not meant to support the use of such a distribution when one has to solve an optimization problem. Trying to clarify what the NFL really mean in order to improve search algorithms, Wolpert analyzes some kind of *"deep formal relationship between supervised learning and searching."* As a result of the analysis of this relationship there are NFL theorems for both search and supervised learning and so there are various ways of reusing techniques first developed in supervised learning for guiding search. A number of experiments are presented which confirm the effectiveness of search algorithms built upon these concepts.

5 NFL for Optimization and Evolutionary Algorithms

5.1 No Free Lunches and Evolutionary Algorithms

The NFL theorems have attracted the interest of the scientific community and keep this interest unchanged. On the other hand, there has been occasionally the bone of contention between some researchers. Such conflicting positions are listed in

Perakh's essay [42]. In particular, one may note the position of Orr [37] regarding the NFL theorems, which was presented on the occasion of the publication of William Dembski's book [10]. Orr stated that:

> ...NFL theorems compare the effectiveness of evolutionary algorithms and look at how often such an algorithm can detect the target, within a certain number of steps...

Orr underlined some very useful observations regarding NFL theorems in relation with Darwinian theory and this has been the essence of the difference between Orr and Dembski. More precisely, Orr claims that *evolution* according to Darwin's theory cannot be seen as a search process and therefore, contrary to Dembski, one cannot claim that Darwinism constitutes a search algorithm. It is evident that NFL theorems do not exclude evolutionary process defined according to the Darwinian theory. Hence, evolutionary algorithms can be appropriate used for search and they are capable to overcome a random search algorithm.

In [42] Perakh gave a popularized interpretation of NFL theorems. This interpretation is presented hereafter along with some useful comments and remarks as given by the author. Suppose that A and B are two search algorithms, exploring the same search space. The algorithms explore the search space by moving from one point to another, selecting points either randomly or following a specific order. Each algorithm performs a certain number of moves. At any point visited the algorithm computes the value of the fitness function and so after, say, k steps the algorithm provides k measurements, which constitute what is called a sample.

In essence, this sample is nothing more than a table in which the values of the fitness function are recorded for each search point. However, an important question arises: *"could two algorithms return the same sample, given that they have selected the same number of points*?" Obviously, the samples computed by two arbitrarily chosen algorithms are not expected to be the same. This argument can be easily understood if one considers it in terms of probabilities.

Specifically, Perakh notes that:

> The probability of a sample (i.e., a table), of size k, produced by an algorithm, say A, differs from the probability that the same sample is produced by another algorithm B, for the same number of steps of the algorithms.

However, the first NFL theorem states that if the search results of the two algorithms are not compared for a particular fitness space but averaged over all possible search spaces, then the above probabilities of obtaining the same sample are equal for any pair of algorithms.

It is worth to underline that the NFL theorems are valid regardless how many times the algorithms are used to complete a search of the underlying problem space or which fitness function values are returned by the different search points. Another point that is worth paying attention is that NFL theorems make no claim about the relative performance of the algorithms, as defined in [42], for a particular search

space. As a result, in terms of performance, any algorithm could be much better than any other "competitor."

Despite the fact that NFL theorems are valid for evolutionary algorithms, in [58] it is argued that this may not stand for the case of co-evolutionary algorithms and so *"Free Lunches"* are possible. The NFL framework for the case of co-evolutionary algorithms as described in [49] is given next.

The statement relative to "*the average performance of the algorithms*," mentioned in the previous paragraphs and references therein, is meaningless without the definition of how this performance is measured. In other words, a very important issue is to define the metrics that one should use in order to effectively measure and compare the performance of the algorithms.

In addition, some other important questions may arise, such as:

(a) *Are there any classes of co-evolution for which there exist NFL theorems?*
(b) *For which co-evolution classes there can be Free Lunches?*

According to the literature [63] these questions are difficult to be answered and they still remain open problems.

Some recent research efforts regarding NFL theorems and black-box optimization have shown that there are co-evolutionary problems with No Free Lunches while Free Lunches are present in the context of other co-evolutionary problems. More precisely, in their work [49], Service and Tauritz present a NFL framework for classes of co-evolutionary algorithms. What is important in this work is the classification of co-evolutionary algorithms based on the solutions they seek. In the co-evolutionary algorithms framework defined in this work, the type of the solutions, or the corresponding individuals, that are effectively considered as solutions to the problem, depend, exclusively, on the type of the problem for which the co-evolutionary algorithm is designed. Note that the different solution concepts are related to the cooperative co-evolution case, the Nash equilibrium case, the maxmin case, etc.

The authors define the so-called weak preference relation which is a relatively simple way of measuring the performance of co-evolutionary algorithms and so it constitutes a metric. This metric is different than the one originally defined by Wolpert and Macready in their work *"Co-evolutionary Free Lunches"* [63].

The framework developed by Service and Tauritz can be considered as a combination of concepts and definitions originating from two theoretical frameworks. The first framework deals with the original NFL theorems [62] for search algorithms and the other for concerns co-evolutionary algorithms [18]. This fusion is done with respect to the consistency of both frameworks. Moreover, in [49] the authors showed that in co-evolution there are Free Lunches. In consequence, the important question that remains to be answered is: *"in which classes of co-evolutionary algorithms there are Free Lunches?"* and further studies are needed to explore additional classes of co-evolution.

5.2 No Free Lunches and Meta-Heuristic Techniques

It is well known that particle swarm optimization methods [41] have greatly contributed to the field of mathematical optimization. These swarm- based methods consist of a number of individuals who guide the optimization process through their collective behavior in order to attain an optimal solution. A great advantage of these methods is that, under suitable conditions and assumptions, they are capable to avoid local minima and ensure convergence of the algorithm to some globally optimal solution. However, convergence analysis of swarm optimization algorithms still remains an active research areas. The most important schemes that have been define and used include: *"Particle Swarm Optimization"* (PSO) [14], *"Ant Colony Optimization"* (ACO) [11], *"Firefly Algorithm"* (FA) [64], *"Artificial Bee Colony algorithm"* (ABC) [28], *"Bat Algorithm"* (BA) [65], *"Cuckoo Search"* (CS) [67], among others.

These methods, also called meta-heuristic techniques, involve exploration and exploitation; two specific search processes which under appropriate conditions "control" the swarm in order to avoid local minima of the fitness function. The applications of the above swarm-based schemes are many and belong to different scientific fields. More details on these can be found in [19], especially concerning engineering and industrial applications. In recent years, application of meta-heuristic techniques has constantly increased and has entered the field of art [1, 2, 12, 47, 52]. More specifically, meta-heuristic techniques have been applied in the tasks of Crowd simulation, Human swarming, and Swarmic art.

Meta-heuristic techniques or meta-models, such as those proposed in [38–41, 43, 50], are used in many cases of evolutionary computing techniques [15–17, 20, 34], in order to create faster optimization algorithms. Especially, in cases where data sets are incomplete or imbalanced or the objective function is computational costly, the meta-heuristic procedures provide alternative, effective, and efficient solution to the optimization problem. Specifically, these techniques are high-level heuristic processes that aim at choosing or creating meta-heuristic search models to resolve more efficiently optimization problems. As noted in [4], under some mild conditions with respect to objective functions, the surrogate algorithms achieve global convergence [5]. In addition, these meta-models are not plagued by damaging features of classic optimization methods, such as the calculation of derivatives. As a consequence, meta-models outperform classic methods, enabling them to be effectively and efficiently deployed in a variety of applications, such as [25, 30, 53].

An important discussion concerning NFL for meta-heuristics is proposed by Yang [66]. In this work the author notes that NFL theorems deal with the average performance of optimization algorithms on all existing problems. Nevertheless, in many real problems this does not hold, as the theoretical requirements are strict and they cannot be applied, in practice. As a consequence, this situation results in getting Free Lunches, and what needs to be determined is the performance of specific algorithms in particular classes of problems. Hence, in such cases there

may exist algorithms that are significantly better than others for a particular class of problems. This phenomenon, i.e., the non-validity of NFL theorems, often occurs when applying meta-heuristic approaches, as the primary NFL theorems concern algorithms searching for individual solutions while population-based meta-heuristic approaches explore simultaneously different parts of the search space and, in this sense, they are considered dealing with sets of solutions. As an example one may consider the cases of genetic algorithms or PSO. A similar situation is encountered in multi-objective optimization, where some algorithms are found to outperform others on specific problems, thus, giving rise to Free Lunches [9].

The theoretical results of NFL theorems while being very important for mathematical optimization with significant theoretical impact, however, incite a number of questions related with practical applications such as: *"What is the position and the opinion of optimization algorithm designers on the practical validity and the applicability of NFL theorems?"*

Yang [66] provides an answer to this question and classifies developers of optimization algorithms in three groups:

(a) A large part of researchers believe that the conditions set by NFL theorems cannot be applied in practice and therefore they do not accept them.
(b) Researchers in the second category accept the validity of NFL theorems but they believe that for specific types of problems there exist optimal algorithms. So, they focus on finding such algorithms for particular classes of problems.
(c) The last group claims that NFL theorems do not hold for continuous problems or for problems belonging to the NP-hard class. Therefore, they focus on discovering problems for which NFL theorems do not apply and hence on defining Free Lunches.

The appeal and the controversies caused by NFL theorems led a large part of the scientific community to re-examine these theorems and restate them in several equivalent forms. The studies resulting from this trend have led to the creation of many frameworks for black-box search algorithms such as the framework proposed by Schumacher et al. [48]. These authors studied the length of the problem description and they concluded that the NFL results as initially formulated by Wolpert are valid not only for the set of all functions but even for smaller sets. Hence, NFL results are independent of whether the set of functions is compressible or not. Finally, the authors conclude that the results of NFL theorems are best maintained in the case of the permutation closure of a single function.

The variety of scientific fields where NFL theorems have been applied made more and more researchers study and apply these theorems which led to the proposition of various extensions of NFL theorems. It is worth mentioning that Auger and Teytaud [4] proposed extensions of NFL theorems related to infinite spaces both countable and uncountable. In addition, they studied the design of optimal heuristic optimization models. According to the original work of Wolpert and Macready [62], the NFL theorems for optimization concern finite search spaces. So, in order to extend the theorems to infinite search spaces, stochastic terms and procedures are introduced. The authors demonstrated that in the case of infinite

countable spaces, the physical extension of the NFL theorems does not hold. In addition, for their proof they defined some distributions of the fitness functions, which lead to equal performance for all heuristic search techniques.

The above proof resulted in Free Lunch theorems based on a random fitness function and involves random search spaces. An additional contribution made in [4] deals with designing optimal algorithms for random fitness functions regarding a black-box optimization framework. In particular, the authors presented an optimal algorithm based on the Bellman's decomposition principle [6], for a certain number of algorithm iterations and a given distribution of fitness. Moreover, for the design procedure and the experiments conducted, the *"Monte-Carlo planning"* algorithm [31] and the *"Upper Confidence Tree"* algorithm [51] were used. Following these research results one may, reasonably, put forward the question: *"Is the improvement proposed by Auger and Teytaud just of theoretical importance or it can be applied in practical situations in acceptable computational time?"*

Hereafter, in order to present some of the results of Auger and Teytaud [4], more formally, we recall the necessary notation adopted in [4]. Let $\mathscr{X}$ denote the search space and $\mathscr{Y}$ its codomain for a given objective function f. For any integer $m \in \{1, 2, \ldots, |\mathscr{X}|\}$ let $(x_1, x_2, \ldots, x_m)$ be the vector of the first m iterates of a search algorithm and let $(f(x_1), f(x_2), \ldots, f(x_m))$ be the vector of the associated objective values. The performance of an algorithm a after m iterations is given by measuring the vector of cost values denoted by $Y(f, m, a) = \langle f(x_1), f(x_2), \ldots, f(x_m)\rangle$.

Using the previous notation, NFL theorems imply the following results for $\mathscr{X}$ any finite domain, $\mathscr{Y}$ its codomain, two search algorithms a and b, any number of iterations m and, finally, any objective function f and p any random permutation uniformly distributed (among all permutations) over $\mathscr{X}$: the random vectors:

$$Y(f \circ p,\ m,\ a) = \langle f \circ p(x_1),\ f \circ p(x_2),\ \ldots,\ f \circ p(x_m)\rangle,$$

and

$$Y(f \circ p,\ m,\ b) = \langle f \circ p(x_1),\ f \circ p(x_2),\ \ldots,\ f \circ p(x_m)\rangle,$$

follow the same distribution.

Moreover, let $\mathscr{X}$ be a countably infinite space and without loss of generality let $\mathscr{X} = \mathbb{N}$. If one is able to provide a non-trivial measurable objective function f, then the following proposition holds:

Proposition 1 *Assume that $\mathscr{NFL}(\mathbb{N}, p, f)$ is a No Free Lunch, and*

$$f(i) = (-1)^{i+1}\, i\,, \quad \forall\ i\ \in \mathbb{N}.$$

Then there is no random permutation p such that $\mathscr{NFL}(\mathbb{N}, p, f)$ holds. Consequently, the $\mathscr{NFL}(\mathbb{N}, f)$ does not hold.

Table 1 Number of citations of the references related to the NFL theorems for Optimization and Evolutionary Algorithms issues presented in Section 5

Contribution	Total citations	Citations per year
Schumacher et al. [48]	190	11.18
Droste et al. [13]	150	9.38
Perakh [42]	6	0.40
Griffiths and Orponen [23]	9	0.69
Service and Tauritz [49]	4	0.40
Poli and Graff [44]	35	3.18
Auger and Teytand [4]	70	8.75
Yang [66]	34	5.67

Among different theoretical results, one may stick to the following Continuous Free Lunch theorem which is considered to be the main result of Auger and Teytaud in [4].

Theorem 4 (Continuous Free Lunch) *Assume that f is a random fitness function with values in $\mathbb{R}^{[0,1]}$. Then $\mathcal{GNFL}([0,1], f)$ does not hold.*

In the above theorem $\mathcal{X}$ is considered to be a continuous domain and without loss of generality $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \mathbb{R}$. Moreover, the notation $\mathcal{GNFL}$ is used for a weaker NFL which does not restrict the fitness function to the compositional form $f \circ p$.

Remark 1 Let f be a random fitness. Then $\mathcal{GNFL}(\mathcal{X}, f)$ holds if and only if for any $m \in \mathbb{N}$ (smaller than $|\mathcal{X}|$ when $\mathcal{X}$ is finite) and any two optimization algorithms a and b, $Y(f, m, a)$ and $Y(f, m, b)$ follow the same distribution.

In Table 1, we provide information about the number of citations[1] received by the most significant contributions concerning the field of the NFL theorems for Optimization and Evolutionary Algorithms. This citation analysis can be considered as an additional information about the importance and contribution of these works in the field of the NFL theorems.

Designing an optimization algorithm that will be more effective than other optimization schemes is a very difficult process and requires a number of conditions. *"Multidisciplinary Design Optimization"* is a problem that is based on the best architecture selection. In such a context, it can be easily understood that obtaining the most efficient design scheme requires testing and may lead to errors. However, the trial-and-error procedure is not appropriate as it is a costly computational process. Vanaret et al. [54] proposed a general design process that avoids the above problem, as well as the inherent complexity that exists in such applications.

[1] Source: Google Scholar.

In most cases of Multidisciplinary Design Optimization, having efficient optimization algorithms is seriously restricted by the complexity of the objective function which is due to the fact that several different architectures are used for the design task. So, it is of primary importance to dispose a methodology that can be applied in all design cases and alleviate this disadvantage. In [54] this is accomplished through a replacement function that can be calculated much more easily than the original one. The authors propose a scalable replacement model through which the architectures can be evaluated easier and thus the choice made more appropriate. Through their experimental results it is clear that the performance of an architecture model depends significantly on the dimension of the original problem. Therefore, as stated by the NFL theorems, there is no architecture that is significantly more efficient than all the others, when dealing with problems of the same dimension. The authors adopt the *"Multidisciplinary Feasible"* and the *"Individual Disciplinary Feasible"* architecture models as the more representative among different architecture models for Multidisciplinary Design Optimization problems. Nevertheless, more architecture models need to be explored in the future.

Kimbrough et al. [29] studied several cases of optimization with constraints using population-based optimization algorithms. In particular, in their research they used genetic algorithms regarding two populations, those of feasible solutions and those of non-feasible ones. Theoretically, in a simple, typical scheme of genetic evolution, the individuals evaluated as feasible solutions would be the only ones that would take part in the evolution of the population and so, in the final formulation of the solution. However, this theoretical provision is not valid in [29] as in this case Kimbrough et al. use feasible solutions in improving the values of the objective function, while non-feasible solutions are used to correct the penalties caused by their constrains.

In order to ensure the smoothness of the optimization process, namely the evolution of the populations, the authors defined a metric distance between the two populations, both among the individuals and the populations' centroids. An important detail that has to be underlined is that the centroids of the two populations are approaching each other during the evolution.

At first sight, it might seem strange to maintain a whole population of infeasible solutions. However, a closer look reveals the usefulness of this position as this population is free to move to space areas, where the feasible solutions cannot, and thus to explore the limited search areas. The authors studied specific problems and spaces and showed that the conclusions of the NFL theorems regarding the equivalence of the black-box search algorithms do not hold. Furthermore, they shown that the NFL theorems do not hold for problems with constraints and specifically in many practical problems where the restrictions are fixed.

The evolutionary computing scheme adopted by Kimbrough et al. [29] is an elegant mechanism which permits to show that there exist constraint optimization problems for which NFL results do not hold. The interested reader is invited to refer to the work [29].

As a supplement of the above research works along with both the theoretical arguments and the conclusions of the specific problem classes mentioned above, Droste et al. in [13] provide some realistic remarks based on the computational complexity of heuristic optimization algorithms. The authors claim that NFL theorems are not possible in the case of heuristic optimization. However, an *"(Almost) No Free Lunch"* ((A)NFL) theorem shows that for each function that can be efficiently optimized by a heuristic search, many other related functions can be constructed where the same heuristic is bad. Consequently, heuristic search methods use some a priori known information, a kind of "ideas," of how to search for good solutions and so they can be successful only for functions that give the appropriate "help."

Theorem 5 *Assume that* **S** *is a randomized search strategy and let* f *be a function,* $f \in \{0,1\}^n$ *and the output range is* $\{0, 1, 2, \ldots, N-1\}$. *Then there are at least* $N^{2^{n/3}-1}$ *number of functions, let* g, $g : \{0,1\} \to \{0, 1, 2, \ldots, N\}$ *in agreement with* f *on all but at most* $2^{n/3}$ *inputs such that* **S** *does find the optimum of* g *within* $2^{n/3}$ *steps with a probability bounded above by* $2^{-n/3}$.

This theorem suggests that heuristic methods cannot succeed in all existing problems. This is because the effectiveness of these techniques is largely based on a good "guess." If this guess is correct, these methods can be very efficient. If not, the search time can reach exponential levels and this constitutes a serious disadvantage of this family of methods.

In [23] Griffiths and Orponen studied optimization strategies for a given finite set of functions. Specifically, they investigated the conditions that need to be satisfied for the functions under consideration in order to have the same performance for a uniform distribution of functions. The result of this research is related to some non-trivial Boolean functions and bounded search algorithms. An important conclusion of this research is that the relationship of NFL theorems and the closed under permutation conditions does not always hold. This happens when we consider functions used to maximize the performance of bounded length searches.

Closing this section, it is worth to mention the contribution of Poli and Graff [44] concerning the NFL theorems and hyper-heuristic techniques. Their conclusions further support the previously referenced works as far as the non-validity of the NFL theorems and the existence of Free Lunches are concerned. The NFL theorems guarantee that this phenomenon happens to hyper-heuristic techniques and high-level hyper-heuristics, if all the problems of interest are closed under permutation. For many real applications the corresponding optimization problems do not satisfy this condition and so, in these cases, there is a Free Lunch for hyper-heuristic techniques. Note that this happens provided that at each level of the search hierarchy the heuristics are evaluated using performance measures that reveal the differences in immediately lower level. The fact that the results of NFL theorems may not hold over heuristic searching techniques does not mean that the existing hyper-heuristic methods are good enough. This may need to be proven and so it requires to be further investigated. Finally, whenever implementation of the NFL theorems

is not possible, one should see the opportunity to try finding some new and more powerful, effective, and efficient hyper-heuristic algorithms, including techniques that are based on genetic programming and genetic algorithms.

6 NFL for Supervised Learning

Revisiting his initial work on supervised learning Wolpert in his work entitled: *"The supervised learning no-free-lunch theorems"* [59] analyzes the main issues underlying his theory on NFL. Wolpert criticizes conventional testing methods for supervised learning as they do not account for out-of-sample testing which is more important for the behavior of supervised learning algorithms. Actually, despite any opposite claim it is common practice in established supervised learning approaches to perform testing with test set that overlap training sets. Thus, conventional frameworks are bound with specific application fields of supervised learning and not with the very problems of the domain.

To cope with this inability of conventional frameworks and deal with the off-training-set error he proposes the so-called *Extended Bayesian Framework* (EBF) which besides offering an extension to classical Bayesian analysis it, also, has the major advantage that it encompasses the conventional frameworks. Based on the EBF, Wolpert develops the set of No Free Lunch theorems which *"bound how much one can infer concerning the (off-training-set) generalization error probability distribution without making relatively strong assumptions concerning the real world. They serve as a broad context in which one should view the claims of any supervised learning framework."*

All aspects of supervised learning are modeled by means of probability distributions. Wolpert provides definitions for those points that are ill defined and they are assumed to constitute defaults for conventional approaches which deal with generalization. Hence, according to Wolpert's notation, if

$$d = \{d_X(i),\, d_Y(i)\}, \quad \forall\, 1 \leqslant i \leqslant m\,,$$

denotes the training data, "f" is the function giving the probability $P(y \mid x,\, f) = f_{x,y}$ and "h" is the x-conditioned probability distribution over values y which is produced by the learning algorithm in response to training data d, $P(y|x, h) = h_{x,y}$ then the generalization error function typically used in supervised learning can be expressed by the expectation value $E(C \mid h,\, f, d)$ for some cost "C" induced by the learning algorithm. So, for the "average misclassification rate error," one may set:

$$E(C \mid h,\, f, d) = E(C \mid h,\, f) = \sum_x \pi(x)\left[1 - \delta(f(x),\, h(x))\right].$$

This is the average number of times across all $x \in X$ that h and f differ relatively to the sampling distribution $\pi(x)$ which produced the training data.

In the sequel, the following two theorems are formulated. These theorems are known as *"No Free Lunch theorems for supervised learning."*

Theorem 6 *$E(C \mid d)$ can be written as a (non-Euclidean) inner product between the distributions $P(h \mid d)$ and $P(f \mid d)$:*

$$E(C \mid d) = \sum_{h,f} Er(h, f, d)\, P(h|d)\, P(f|d),$$

where $Er(h, f, d)$ denotes the error function.

The following meanings are given by Wolpert to this theorem:

(a) An answer to how well a learning algorithm does on some problem is determined by how "aligned" the algorithm $P(h|d)$ is with the posterior $P(f \mid d)$.
(b) One cannot prove anything regarding how well a particular learning algorithm generalizes as one is, generally, unable to prove that $P(h|d)$ is aligned with $P(f \mid d)$ unless $P(f \mid d)$ has a certain form.

The impossibility to prove that $P(f \mid d)$ has a certain form is formalized by the following theorem.

Theorem 7 *Consider the off-training-set error function. Let "$E_i(\cdot)$" indicate an expectation value evaluated using learning algorithm "i." Then for any two learning algorithms $P_1(h \mid d)$ and $P_2(h \mid d)$, independent of the sampling distribution*

(i) Uniformly averaged over all f,
$E_1(C \mid f, m) - E_2(C \mid f, m) = 0$;
(ii) Uniformly averaged over all f, for any training set d,
$E_1(C \mid f, d) - E_2(C \mid f, d) = 0$;
(iii) Uniformly averaged over all $P(f)$,
$E_1(C \mid m) - E_2(C \mid m) = 0$;
(iv) Uniformly averaged over all $P(f)$, for any training set d,
$E_1(C \mid d) - E_2(C \mid d) = 0$.

Remark 2 Given that the quantities $E(C \mid d)$, $E(C \mid m)$, $E(C \mid f, d)$, or $E(C \mid f, m)$ denote different measures of risks, the theorem states that for any of these measures any two algorithms on average perform equally well. Actually, Algorithm 1 is superior to Algorithm 2 for as many problems as Algorithm 2 is superior to Algorithm 1.

The examples given by Wolpert are about cross-validation and Bayesian inference. Moreover, some variants of Theorem 7 are presented and the intuitive ideas of Theorem 7 are analyzed. These ideas gave rise to the following important research efforts concerning two critical issues of supervised learning, namely *early stopping* and *cross-validation*.

6.1 No Free Lunch for Early Stopping

Iterative methods, such as gradient descent, train a learner by updating its free parameters in order to make it better fit the training data and improve the performance of the learner on data outside the training set. Up to some point this is a successful task but beyond that point further training leads to over-fitting the training data while failing to deal with out-of-sample data, thus, increasing the generalization error of the learner. Regularization techniques including early stopping are used to avoid over-fitting.

The *"early stopping"* provides rules on how to conduct training and when to stop iterations in order to avoid over-fitting. In machine learning the early stopping has been used in many contexts and has been supported with various mathematical tools. A well-known and widely used technique is to guide validation of a training procedure with early stopping by monitoring the increase of the generalization error on validation data.

In [7] Cataltepe et al. aim at bringing the idea of NFL into the framework of early stopping. The method of choosing a model using the early stopping approach relies on a uniform selection of the model among the models giving the same training error. This approach is claimed to be similar to the *"Gibbs algorithm."* The uniform probability of selection around the training error minimum is equivalent to the isotropic distributions of Amari et al. [3], while it differs from this work as it does not assume a very large number of training examples. In addition to general linear models in [7] it is presumed that the probability of selection of models is symmetric only around the training error minimum.

This symmetry hypothesis is a weaker requirement than uniformity. The authors analyze early stopping for some training error minimum. If the training set constitutes all the information that one has about the target, then one should minimize the training error as much as possible to achieve lower generalization error. Moreover, the authors demonstrate that when additional information is available, early stopping can help.

6.2 No Free Lunch for Cross-Validation

In machine learning and, generally, in statistical learning theory, *"cross-validation"* is a model evaluation method used when a predictive modeling procedure or any learner is asked to make new predictions for data it has not already seen. This data constitutes the model validation set. Therefore, instead of using mathematical analysis cross-validation is a generally applicable method used to assess the performance of a model. Specific methods of cross-validation can be either of *"exhaustive"* (such as leave p-out, leave-one-out) or *"non-exhaustive"* type (such as k-fold, hold out, repeated random sub-sampling) and they are able to give meaningful results provided that the training set and the validation set are drawn from the same population, i.e., the same distribution.

Cross-validation is a statistical technique which constitutes an objective approach to compare different learning procedures as it does not rely on in-sample error rates. Thus, it was long widely believed that it can be successful, despite of the prior knowledge available on the problem at hand. Zhu and Rohwer [68] provide a numerical counter-example which, despite the fact that it is an artificial one, constitutes a minimal proof that cross-validation is not a *"universally beneficial method."* The problem consists in selecting the unbiased estimator of the expectation of a Gaussian distribution between two estimators, namely: (a) an unbiased and (b) a highly biased one. The authors apply the leave-one-out scheme and make an attempt to show that this method is inefficient even in small problems. Hence, cross-validation cannot defy the theoretical result of the NFL theorem, that is, *"no algorithm can be good for any arbitrary prior."*

Moreover, the authors carry out further experiments and give a detailed analysis with the aim to provide answers to any criticism against the main issue tackled by the paper which is *"as with any other algorithm, cross-validation and, in this sense, a number of other approaches such as bootstrap cannot solve equally good any kind of problem."* Hence, if some prior knowledge is used for an algorithm, then this should be communicated to any interested user so that he can decide whether to use it or not.

Goutte published a more elaborated approach on this matter in his work [22] entitled *"Note on free lunches and cross-validation."* In this paper the aforementioned approach of Zhu and Rohwer on cross-validation and NFL theorem is revisited by further elaborating on the numerical example. The author, also, applies the leave-one-out and the m-fold cross-validation schemes on the numerical result used by Zhu and Rohwer and performs a more detailed mathematical description. Analysis of the results obtained supports the argument that there is *"No Free Lunch for cross-validation"* and though the method is not the best approach for evaluating performance of learners, however, it is capable to give very good results in a number of practical situations.

Further to the above research, Rivals and Personnaz [46] took over the work of Goutte and by using probabilistic analysis they applied leave-one-out cross-validation on measures of model quality. The leave-one-out scores obtained show that the conclusions of Goutte are optimistic as they deal with a trivial problem for which any reasonable method is not prone to make a wrong choice. In addition, a comparison between leave-one-out cross-validation and statistical tests for the selection of linear models is performed. The numerical results obtained by a specific illustrative example show that for linear estimators with large number of samples, leave-one-out cross-validation does not perform well as compared to statistical tests. This leads to the conclusion that it is unlikely that this method is able to perform well in the case of nonlinear estimators such as neural networks. Hence, an important result is stressed, that is, statistical tests should be preferred to leave-one-out cross-validation *"provided that the (linear or nonlinear) model has the properties required for the statistical tests to be valid."*

6.3 Real-World Machine Learning Classification and No Free Lunch Theorems: An Experimental Approach

The majority of the research concerning NFL and supervised learning seems to be more or less theoretical. Unlike the previously reported work Gómez and Rojas published in [21] the results of a number of machine learning experiments with the aim to help understanding the impact of NFL on real-world problems. At the same time the authors attempted to provide sufficient experimental evidence on the validity of NFL.

The set of machine learning algorithms used in these experiments comprise:

(a) Naive Bayes classifiers,
(b) C4.5 decision trees,
(c) Neural networks,
(d) k-nearest neighbors classifiers,
(e) Random C4.5 forest,
(f) AdaBoost.M1,
(g) Stacking.

The performance of these approaches was examined in terms of average accuracy over six data sets taken from the UCI machine learning repository. These data sets are: "Audiology," "Column," "Breast cancer," "Multiple features (Fourier)," "German credit," and "Nursery." To a great extent, the results obtained are consistent with previous research. On the other hand, according to the NFL theorem the tested algorithms should expose the same degree of accuracy. However, this is valid when a sufficiently large number of data sets are available. The authors underline that some common assumptions pertain the data sets. These common assumptions concern the Occam's razor and the independent identical distribution of the samples as well as, mainly, the data-dependent structural properties found in the data sets, that is, determinism and the Pareto principle. Based on these last properties they explain the peculiarities of the data sets and the results concerning the accuracy. Then, it is clear that not all the algorithms perform equally well on all problems.

In addition to the above, the authors perform a number of experiments using kernel machines and especially support vector machines (SVM) as well as deep learning networks. The results obtained show that SVM outperform the other learning algorithms while the performance of deep learning on these small and relatively simple problems is disappointing. In fact while these architectures are designed to handle complex data sets which have inherent abstraction layers they seem to be incapable to cope with simpler data sets with possibly lower data abstraction. This shows that NFL applies even in the case of deep learning which is also subject to limitations as other machine learning algorithms.

In terms of conclusion the authors state that: *"While evaluating the average accuracy ranking for the six data sets, they noticed the effect of the NFL theorem and how assumptions are key to performance."* Comparing with similar research work they conclude that: *"the data and its pre-processing are as important as, if*

not more so than, the algorithm itself in determining the quality of the model. Data visualization or statistical techniques such as feature selection can be crucial to provide a better fit and obtain simpler and better models."

7 Synopsis and Concluding Remarks

In this paper we surveyed some of the most sound research works concerning No Free Lunch (NFL) theorems and their results in search, optimization, and supervised learning. Starting from the earlier work of David H. Wolpert, where the essential concepts underpinning NFL theorems were defined, we went through the research efforts that contributed to the formulation of the most relevant frameworks for applying NFL theorems. Moreover, we presented those research works which show when NFL theorems do not hold and so there are Free Lunches, i.e., algorithms that significantly outperform other algorithms on specific classes of problems. One of the objectives set for this survey was to make clear which are the hypotheses and the restrictions for applying NFL results, or on the contrary, to pinpoint the conditions under which there are Free Lunches, as defined by researchers in their respective papers.

One of the most relevant conclusion is that important research needs to be carried out in order to delineate those classes of problems for which NFL theorems apply and those for which they don't. NFL theorems do not put any obstacle on continuing research for developing more efficient algorithms which apply to even larger classes of problems. They just seem to make clear that there are limits for these algorithms. Finally, this does not mean that for some specific problems one is not able to design an algorithm performing better than its competitors.

Closing this review we hope that this work will assist all those who are interested in NFL theorems.

Acknowledgements S.-A. N. Alexandropoulos is supported by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY). P. M. Pardalos is supported by the Paul and Heidi Brown Preeminent Professorship at ISE (University of Florida, USA), and a Humboldt Research Award (Germany).

References

1. Al-Rifaie, M.M., Bishop, J.M.: Swarmic paintings and colour attention. In: International Conference on Evolutionary and Biologically Inspired Music and Art, pp. 97–108. Springer, Berlin (2013)
2. Al-Rifaie, M.M., Bishop, J.M., Caines, S.: Creativity and autonomy in swarm intelligence systems. Cogn. Comput. **4**(3), 320–331 (2012)

3. Amari, S., Murata, N., Muller, K.R., Finke, M., Yang, H.H.: Asymptotic statistical theory of overtraining and cross-validation. IEEE Trans. Neural Netw. **8**(5), 985–996 (1997)
4. Auger, A., Teytaud, O.: Continuous lunches are free plus the design of optimal optimization algorithms. Algorithmica **57**(1), 121–146 (2010)
5. Auger, A., Schoenauer, M., Teytaud, O.: Local and global order 3/2 convergence of a surrogate evolutionary algorithm. In: Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, pp. 857–864. ACM, New York (2005)
6. Bellman, R.: Dynamic Programming. Princeton University Press, Princeton (1957)
7. Cataltepe, Z., Abu-Mostafa, Y.S., Magdon-Ismail, M.: No free lunch for early stopping. Neural Comput. **11**(4), 995–1009 (1999)
8. Ciucu, F., Schmitt, J.: Perspectives on network calculus: no free lunch, but still good value. ACM SIGCOMM Comput. Commun. Rev. **42**(4), 311–322 (2012)
9. Corne, D., Knowles, J.: Some multiobjective optimizers are better than others. In: IEEE Congress on Evolutionary Computation (CEC 2003), vol. 4, pp. 2506–2512. IEEE, Piscataway (2003)
10. Dembski, W.A.: No Free Lunch: Why Specified Complexity Cannot be Purchased Without Intelligence. Rowman & Littlefield, Langham (2006).
11. Dorigo, M., Birattari, M.: Ant colony optimization. In: Encyclopedia of Machine Learning, pp. 36–39. Springer, Boston (2011)
12. Drettakis, G., Roussou, M., Reche, A., Tsingos, N.: Design and evaluation of a real-world virtual environment for architecture and urban planning. Presence Teleop. Virt. **16**(3), 318–332 (2007)
13. Droste, S., Jansen, T., Wegener, I.: Optimization with randomized search heuristics – the (A)NFL theorem, realistic scenarios, and difficult functions. Theor. Comput. Sci. **287**(1), 131–144 (2002)
14. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the IEEE Sixth International Symposium on Micro Machine and Human Science, 1995, MHS'95, pp. 39–43. IEEE, Piscataway (1995)
15. Epitropakis, M.G., Plagianakos, V.P., Vrahatis, M.N.: Evolutionary adaptation of the differential evolution control parameters. In: Proceedings of the IEEE Congress on Evolutionary Computation, 2009, CEC'09, pp. 1359–1366. IEEE, Piscataway (2009)
16. Epitropakis, M.G., Tasoulis, D.K., Pavlidis, N.G., Plagianakos, V.P., Vrahatis, M.N.: Enhancing differential evolution utilizing proximity-based mutation operators. IEEE Trans. Evol. Comput. **15**(1), 99–119 (2011)
17. Epitropakis, M.G., Plagianakos, V.P., Vrahatis, M.N.: Evolving cognitive and social experience in particle swarm optimization through differential evolution: a hybrid approach. Inf. Sci. **216**, 50–92 (2012)
18. Ficici, S.G.: Solution Concepts in Coevolutionary Algorithms. PhD thesis, Brandeis University Waltham, Waltham (2004)
19. Floudas, C.A., Pardalos, P.M.: Encyclopedia of Optimization. Springer Science & Business Media B.V., Dordrecht (2008)
20. Georgiou, V.L., Malefaki, S., Parsopoulos, K.E., Alevizos, Ph.D., Vrahatis, M.N.: Expeditive extensions of evolutionary Bayesian probabilistic neural networks. In: Third International Conference on Learning and Intelligent Optimization (LION3 2009). Lecture Notes in Computer Science, vol. 5851, pp. 30–44. Springer, Berlin (2009)
21. Gómez, D., Rojas, A.: An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. Neural Comput. **28**(1), 216–228 (2015)
22. Goutte, C.: Note on free lunches and cross-validation. Neural Comput. **9**(6), 1245–1249 (1997)
23. Griffiths, E.G., Orponen, P.: Optimization, block designs and no free lunch theorems. Inf. Process. Lett. **94**(2), 55–61 (2005)
24. Ho, Y.C.: The no free lunch theorem and the human-machine interface. IEEE Control. Syst. **19**(3), 8–10 (1999)

25. Hopkins, D.A., Thomas, M.: Neural network and regression methods demonstrated in the design optimization of a subsonic aircraft. Structural Mechanics and Dynamics Branch 2002 Annual Report, p. 25 (2003)
26. Hume, D. (Introduction by Mossner, E.C.): A Treatise of Human Nature. Classics Series. Penguin Books Limited, London (1986)
27. Hume, D.: A Treatise of Human Nature. The Floating Press Ltd., Auckland (2009). First published in 1740
28. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. J. Glob. Optim. **39**(3), 459–471 (2007)
29. Kimbrough, S.O., Koehler, G.J., Lu, M., Wood, D.H.: On a feasible–infeasible two-population (FI-2Pop) genetic algorithm for constrained optimization: distance tracing and no free lunch. Eur. J. Oper. Res. **190**(2), 310–327 (2008)
30. Kleijnen, J.P.C.: Sensitivity analysis of simulation experiments: regression analysis and statistical design. Math. Comput. Simul. **34**(3–4), 297–315 (1992)
31. Kocsis, L., Szepesvari, C.: Bandit-based Monte-Carlo planning. In: European Conference on Machine Learning (ECML 2006). Lecture Notes in Computer Science, vol. 4212, pp. 282–293. Springer, Berlin (2006)
32. Köppen M.: Some technical remarks on the proof of the no free lunch theorem. In: Proceedings of the Fifth Joint Conference on Information Sciences (JCIS), vol. 1, pp. 1020–1024. Atlantic City (2000)
33. Köppen, M., Wolpert, D.H., Macready, W.G.: Remarks on a recent paper on the "No Free Lunch" theorems. IEEE Trans. Evol. Comput. **5**(3), 295–296 (2001)
34. Laskari, E.C., Parsopoulos, K.E., Vrahatis, M.N.: Utilizing evolutionary operators in global optimization with dynamic search trajectories. Numer. Algorithms **34**(2–4), 393–403 (2003)
35. Macready, W.G., Wolpert, D.H.: What makes an optimization problem hard? Complexity **1**(5), 40–46 (1996)
36. Marshall, J.A.R., Hinton, T.G.: Beyond no free lunch: Realistic algorithms for arbitrary problem classes. In: IEEE Congress on Evolutionary Computation, pp. 1–6. IEEE, Piscataway (2010)
37. Orr, H.A.: Review of no free lunch by William A Dembski. Boston Review. Available on-line at http://bostonreview.net/BR27, 3 (2002)
38. Parsopoulos, K.E., Vrahatis, M.N.: Recent approaches to global optimization problems through particle swarm optimization. Nat. Comput. **1**(2–3), 235–306 (2002)
39. Parsopoulos, K.E., Vrahatis, M.N.: On the computation of all global minimizers through particle swarm optimization. IEEE Trans. Evol. Comput. **8**(3), 211–224 (2004)
40. Parsopoulos, K.E., Vrahatis, M.N.: Parameter selection and adaptation in unified particle swarm optimization. Math. Comput. Model. **46**(1–2), 198–213 (2007)
41. Parsopoulos, K.E., Vrahatis, M.N.: Particle Swarm Optimization and Intelligence: Advances and Applications. Information Science Publishing (IGI Global), Hershey (2010)
42. Perakh, M.: The No Free Lunch Theorems and Their Application to Evolutionary Algorithms (2003)
43. Petalas, Y.G., Parsopoulos, K.E., Vrahatis, M.N.: Memetic particle swarm optimization. Ann. Oper. Res. **156**(1), 99–127 (2007)
44. Poli, R., Graff, M.: There is a free lunch for hyper-heuristics, genetic programming and computer scientists. In: Proceedings of the 12th European Conference on Genetic Programming, EuroGP '09, pp. 195–207. Springer, Berlin (2009)
45. Poli, R., Graff, M., McPhee, N.F.: Free lunches for function and program induction. In: Proceedings of the Tenth ACM SIGEVO Workshop on Foundations of Genetic Algorithms, FOGA '09, pp. 183–194. ACM, New York (2009)
46. Rivals, I., Personnaz, L.: On cross validation for model selection. Neural Comput. **11**(4), 863–870 (1999)
47. Rosenberg, L.B.: Human swarms, a real-time paradigm for collective intelligence. Collective Intelligence (2015)

48. Schumacher, C., Vose, M.D., Whitley, L.D.: The no free lunch and problem description length. In: Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation, pp. 565–570. Morgan Kaufmann Publishers Inc., Burlington (2001)
49. Service, T.C., Tauritz, D.R.: A no-free-lunch framework for coevolution. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation, pp. 371–378. ACM, Piscataway (2008)
50. Sotiropoulos, D.G., Stavropoulos, E.C., Vrahatis, M.N.: A new hybrid genetic algorithm for global optimization. Nonlinear Anal. Theory Methods Appl. **30**(7), 4529–4538 (1997)
51. Teytaud, O., Flory, S.: Upper confidence trees with short term partial information. In: European Conference on the Applications of Evolutionary Computation, pp. 153–162. Springer, Berlin (2011)
52. Thalmann, D.: Crowd Simulation. Wiley Online Library (2007)
53. Van Grieken, M.: Optimisation pour l'apprentissage et apprentissage pour l'optimisation. PhD thesis, Université Paul Sabatier-Toulouse III, Toulouse (2004)
54. Vanaret, C., Gallard, F., Martins, J.: On the consequences of the "No Free Lunch" theorem for optimization on the choice of an appropriate MDO architecture. In: 18th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, pp. 3148 (2017)
55. Wolpert, D.H.: On the connection between in-sample testing and generalization error. Complex Syst. **6**(1), 47–94 (1992)
56. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. Neural Comput. **8**(7), 1341–1390 (1996)
57. Wolpert, D.H.: The existence of a priori distinctions between learning algorithms. Neural Comput. **8**(7), 1391–1420 (1996)
58. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: Soft Computing and Industry, pp. 25–42. Springer, London (2002)
59. Wolpert, D.H.: The Supervised Learning No-Free-Lunch Theorems, pp. 25–42. Springer, London (2002)
60. Wolpert, D.H.: What the no free lunch theorems really mean; how to improve search algorithms. SFI working paper: 2012–10-017. Santa Fe Institute, Santa Fe (2012)
61. Wolpert, D.H., Macready, W.G.: No Free Lunch Theorems for Search. Tech. Rep. SFI-TR-95-02-010. Santa Fe Institute, Santa Fe (1995)
62. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997)
63. Wolpert, D. H., Macready, W.G.: Coevolutionary free lunches. IEEE Trans. Evol. Comput. **9**(6), 721–735 (2005)
64. Yang, X.S.: Firefly algorithm, stochastic test functions and design optimization. Int. J. Bio-Inspired Comput. **2**(2), 78–84 (2010)
65. Yang, X.S.: A new metaheuristic bat-inspired algorithm. In: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pp. 65–74. Springer, Berlin (2010)
66. Yang, X.S.: Swarm-based metaheuristic algorithms and no-free-lunch theorems. In: Theory and New Applications of Swarm Intelligence. InTech, London (2012)
67. Yang, X.S., Deb, S.: Cuckoo search via Lévy flights. In: Proceedings of the World Congress on Nature & Biologically Inspired Computing, 2009, NaBIC 2009. pp. 210–214. IEEE, Piscataway (2009)
68. Zhu, H., Rohwer, R.: No free lunch for cross-validation. Neural Comput. **8**(7), 1421–1426 (1996)