On classes of functions for which No Free Lunch results hold

Christian Igel *, Marc Toussaint

Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany

Received 19 December 2002
Communicated by A.E. Eiben

Keywords: Combinatorial problems; No Free Lunch; Optimization; Randomized algorithms

1. Introduction

The No Free Lunch (NFL) theorems for combinatorial optimization state, roughly speaking, that all search algorithms have the same average performance over all possible objective functions \( f : \mathcal{X} \rightarrow \mathcal{Y} \), where the search space \( \mathcal{X} \) as well as the cost-value space \( \mathcal{Y} \) are finite sets [6]. However, it has been argued that in practice one does not need an algorithm that performs well on all possible functions, but only on a subset that arises from the constraints of real-world problems. For example, it has been shown that for pseudo-Boolean functions restrictions of the complexity can lead to subsets of functions on which some algorithms perform better than others (in [5] complexity is defined in terms of the number of local minima and in [1,2] the complexity is defined based on the size of the smallest ordered binary decision diagram, OBDD, representing the function).

Recently, a sharpened version of the NFL theorem has been proven that states that NFL results hold for any subset \( F \) of the set of all possible functions if and only if \( F \) is closed under permutation (c.u.p.) [4]. Based on this important result, we can derive classes of functions where NFL does not hold simply by showing that these classes are not c.u.p. This leads to the main result of this paper: It is proven that the fraction of subsets that are c.u.p. is negligibly small. Arguments are given why we think that classes of objective functions resulting from important classes of real-world problems are likely not to be c.u.p.

In the following section, we give some basic definitions and concisely restate the sharpened NFL theorem given in [4]. Then we derive the number of subsets c.u.p. Finally, we discuss some observations regarding structured search spaces and closure under permutation.

2. Preliminaries

We consider a finite search space \( \mathcal{X} \) and a finite set of totally ordered cost values \( \mathcal{Y} \). Let \( \mathcal{F} = \mathcal{Y}^{\mathcal{X}} \) be the set of all objective functions \( f : \mathcal{X} \rightarrow \mathcal{Y} \) to be optimized (also called fitness, energy, or cost functions). NFL theorems are concerned with nonrepeating black-box search algorithms (referred to simply as algorithms for brevity) that choose a new exploration...
point in the search space depending on the complete history of prior explorations: Let the sequence \( T_m = ((x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m))) \) represent \( m \) nonrepeating explorations \( x_i \in \mathcal{X}, \forall i, j: x_i \neq x_j \) and their cost values \( f(x_i) \in \mathcal{Y} \). An algorithm \( A \) appends a pair \( (x_{m+1}, f(x_{m+1})) \) to this sequence by mapping \( T_m \) to a new point \( x_{m+1}, \forall i: x_{m+1} \neq x_i \). Generally, the performance of an algorithm \( A \) after \( m \) iterations with respect to a function \( f \) depends on the sequence \( Y(f, m, A) = (f(x_1), f(x_2), \ldots, f(x_m)) \) of cost values the algorithm has produced. Let the function \( c \) denote a performance measure mapping sequences of \( \mathcal{Y} \) to the real numbers (e.g., in the case of function minimization a performance measure that returns the minimum \( \mathcal{Y} \) value in the sequence could be a reasonable choice).

Let \( \pi: \mathcal{X} \rightarrow \mathcal{X} \) be a permutation of \( \mathcal{X} \). The set of all permutations of \( \mathcal{X} \) is denoted by \( \Pi(\mathcal{X}) \). A set \( F \subseteq \mathcal{F} \) is said to be closed under permutation (c.u.p.) if for any \( \pi \in \Pi(\mathcal{X}) \) and any function \( f \in F \) the function \( f \circ \pi \) is also in \( F \).

**Theorem 1** (Sharpened NFL). For any two algorithms \( A \) and \( B \), any \( k \in \mathbb{R} \), any \( m \in \{1, \ldots, |\mathcal{X}|\} \), and any performance measure \( c \)

\[
\sum_{f \in F} \delta(k, c(Y(f, m, A))) = \sum_{f \in F} \delta(k, c(Y(f, m, B)))
\]

iff \( F \) is c.u.p.

Herein, \( \delta \) denotes the Kronecker function \( \delta(i, j) = 1 \) if \( i = j \), \( \delta(i, j) = 0 \) otherwise. A proof of Theorem 1 is given in [4]. Note that the summation means *uniformly* averaging over all functions \( f \in \mathcal{F} \). This theorem implies that for any two algorithms \( A \) and \( B \) and any function \( f_A \in F \), where \( F \) is c.u.p., there is a function \( f_B \in F \) on which \( B \) has the same performance as \( A \) on \( f_A \).

**Theorem 2.** The number of nonempty subsets of \( \mathcal{Y}^{\mathcal{X}} \) that are c.u.p. is given by

\[2^{|\mathcal{Y}|(|\mathcal{X}|−1)} − 1\]

and therefore the fraction of nonempty subsets c.u.p. is

\[
\frac{(2^{|\mathcal{Y}|(|\mathcal{X}|−1)} − 1)}{(2^{|\mathcal{Y}|(|\mathcal{X}|−1)} − 1)}.
\]

The proof is given in Appendix A. The fraction decreases for increasing \( |\mathcal{X}| \) as well as for increasing \( |\mathcal{Y}| \). Already for small \( |\mathcal{X}| \) and \( |\mathcal{Y}| \) the fraction almost vanishes, e.g., for Boolean functions \( \{0, 1\}^3 \rightarrow \{0, 1\} \) the fraction is \( \approx 10^{-74} \).

Using the bounds \( \binom{n}{m} \leq n^m/(m!) \) and

\[
\sqrt{2\pi} m^{m+1/2}e^{-m} \cdot e^{(12m+1)}^{-1}
\]

for \( n, m \in \mathbb{N} \) [3, p. 54] we have

\[
(2^{|\mathcal{Y}|(|\mathcal{X}|−1)} − 1)/\{2^{|\mathcal{Y}|(|\mathcal{X}|−1)} − 1\}
\]

\[
< 2^{e+\epsilon(|\mathcal{Y}|(|\mathcal{X}|−e)/(|\mathcal{X}|−e)|\mathcal{X}|−|\mathcal{Y}|)}.
\]

For \( |\mathcal{Y}| > e/|\mathcal{X}|(|\mathcal{X}|−e) \) and \( |\mathcal{X}| > 2 \) this expression is not larger than

\[
2^{e+\epsilon(|\mathcal{Y}|(|\mathcal{X}|−e)/|\mathcal{X}|−|\mathcal{Y}|)}
\]

and converges to zero double exponentially fast with increasing \( |\mathcal{X}| \).

4. Search spaces with neighborhood relations

In the previous section, we have shown that the fraction of subsets c.u.p. is close to zero already for small search and cost-value spaces. Still, the absolute number of subsets c.u.p. grows rapidly with increasing \( |\mathcal{X}| \) and \( |\mathcal{Y}| \). What if these classes of functions are the “important” ones, i.e., those we are dealing with in practice? In this section, we define some quite general constraints on functions important in practice that induce classes of functions that are not c.u.p.

We believe that two assumptions can be made for most of the functions we are dealing with in real-world optimization: First, the search space has some topological structure. Second, the set of objective functions we are interested in fulfills some constraints based on
Proof. It holds $\forall x_i, x_j \in X$ $x_i \neq x_j \land n(x_i, x_j) = 1$.
We call a neighborhood nontrivial iff $\exists x_i, x_j \in X$: $x_i \neq x_j \land n(x_i, x_j) = 0$. It holds:

Theorem 3. A nontrivial neighborhood on $X$ is not invariant under permutations of $X$.

Proof. It holds $\exists x_i, x_j, x_k, x_l \in X$: $x_i \neq x_j \land x_k \neq x_l \land n(x_i, x_j) = 0 \land n(x_k, x_l) = 0$. For any permutation $\pi$ that maps $x_i$ and $x_j$ onto $x_k$ and $x_l$, respectively, the invariance property, $\forall a, b \in X$: $n(x_a, x_b) = n(\pi(x_a), \pi(x_b))$, is violated.

Remark 1. Assume the search space $X$ can be decomposed as $X = X_1 \times \cdots \times X_l$, $l > 1$, and let on one component $X_l$ exist a nontrivial neighborhood $n_l: X_1 \times \cdots \times X_l \to [0, 1]$. This neighborhood induces a nontrivial neighborhood on $X$, where two points are neighbors iff their $i$th components are neighbors with respect to $n_l$. Thus, the constraints discussed below need only refer to a single component.

Remark 2. The neighborhood relation need not be a canonical one (e.g., Hamming distance for Boolean search spaces). Instead, it can be based on “phenotypic” (i.e., functional) properties (e.g., if integers are encoded by bit-strings, then the bit-strings can be defined as neighbors iff the corresponding integers are).

Now we describe some constraints that are defined with respect to a neighborhood relation and are—to our minds—relevant in practice. For this purpose, we assume a metric $d_X: X \times X \to \mathbb{R}$ on $X$, e.g., in the typical case of real-valued fitness function $Y \subseteq \mathbb{R}$ the Euclidean distance.

First, we show how a constraint on steepness (closely related to the concept of strong causality) leads to a set of functions that is not c.u.p. Based on a neighborhood relation on the search space, we can define a simple measure of maximum steepness of a function $f \in F$ by

$$s_{\max}^\pi(f) = \max_{x_i, x_j \in X} d_Y(f(x_i), f(x_j)).$$

Further, for a function $f \in F$, we define the diameter of its range as

$$d_{\max}(f) = \max_{x_i, x_j \in X} d_Y(f(x_i), f(x_j)).$$

Corollary 1. If the maximum steepness $s_{\max}^\pi(f)$ of every function $f$ in a nonempty subset $F \subset F$ is constrained to be smaller than the maximal possible $\max_{f \in F} d_{\max}(f)$, then $F$ is not c.u.p.

Proof. Let $g = \arg\max_{f \in F} d_{\max}(f)$ be a function with maximal range and let $x_i$ and $x_j$ be two points with property $d(g(x_i), g(x_j)) = d_{\max}(g)$. Since the neighborhood on $X$ is nontrivial there exist two neighboring points $x_k$ and $x_l$. There exists a permutation $\pi$ that maps $x_i$ and $x_j$ on $x_k$ and $x_l$. If $F$ is c.u.p., the function $g \circ \pi$ is in $F$. This function has steepness $s_{\max}^\pi(g \circ \pi) = d_{\max}(g) = \max_{f \in F} d_{\max}(f)$, which contradicts the steepness constraint.

As a second constraint, we consider the number of local minima, which is often regarded as a measure of complexity [5]. For a function $f \in F$ a point $x \in X$ is a local minimum iff $f(x) < f(x_i)$ for all neighbors $x_i$ of $x$. Given a function $f$ and a neighborhood relation on $X$, we define $l_{\max}(f)$ as the maximal number of minima that functions with the same $Y$-histogram as $f$ can have (i.e., functions where the number of $X$-values that are mapped to a certain $Y$-value are the same as for $f$, see Appendix A).

As an example, consider pseudo-Boolean functions $\{0, 1\}^n \to R \subset \mathbb{R}$ and let two points be neighbors iff they have Hamming-distance one. Then the maximum number of local minima is $2^{n-1}$ (e.g., the $n$-dimensional parity function, which is 1 if the number of ones in the input bitstring is even and 0 otherwise, has $2^{n-1}$ different global minima).

In Appendix A we prove that for any two functions $f, g$ with the same $Y$-histogram there exists a permutation $\pi \in \Pi(X)$ with $f \circ \pi = g$. Thus, it follows:
Corollary 2. If the number of local minima of every function $f$ in a nonempty subset $F \subset \mathcal{F}$ is constrained to be smaller than the maximal possible $\max_{f \in F} n_{\text{max}}(f)$, then $F$ is not c.u.p.

5. Conclusion

Based on the results in [4], we have shown that the statement “I’m only interested in a subset $F$ of all possible functions, so the precondition of the NFL theorem is not fulfilled” is true with a probability close to one (if $F$ is chosen uniformly and $\mathcal{Y}$ and $\mathcal{X}$ have reasonable cardinalities). Further, the statements “In my application domain, functions with maximum number of local minima are not realistic” and “For some components, the objective functions under consideration will not have the maximal possible steepness” lead to scenarios where the precondition of the NFL theorem is not fulfilled.

The fact that the precondition of the NFL theorem is violated does not say much about the performance of a particular algorithm averaged over the considered set of functions for a given performance measure. In particular, our results do not quantify any differences. That a problem class is not c.u.p. does not lead to a “free lunch”, but ensures the possibility of a “free appetizer”, e.g., we know that there exists a performance when averaged over all the considered objective functions.

Acknowledgements

We thank Hannes Edelbrunner for fruitful discussions and Thomas Jansen, Stefan Wiegand, and Michael Hüsken for their comments on the manuscript. We thank the second anonymous reviewer for his comments, especially those concerning the bound on the fraction of subsets c.u.p. This work was supported by the DFG, grant Solesys, number SE251/41-1.

Appendix A. Proof of Theorem 2

For the proof, we use the concepts of $\mathcal{Y}$-histograms: We define a $\mathcal{Y}$-histogram (histogram for short) as a mapping $h: \mathcal{Y} \to \mathbb{N}_0$ such that $\sum_{y \in \mathcal{Y}} h(y) = |\mathcal{X}|$. The set of all histograms is denoted $\mathcal{H}$. With any function $f: \mathcal{X} \to \mathcal{Y}$ we associate the histogram $h(y) = |f^{-1}(y)|$ that counts the number of elements in $\mathcal{X}$ that are mapped to the same value $y \in \mathcal{Y}$ by $f$. Herein, $f^{-1}(y)$, $y \in \mathcal{Y}$ returns the preimage $\{x \mid f(x) = y\}$ of $y$ under $f$. Further, we call two functions $f, g \in \mathcal{F}$ $h$-equivalent iff they have the same histogram. We call the corresponding $h$-equivalence class $B_h \subseteq \mathcal{F}$ containing all function with histogram $h$ a basis class.

Before we prove Theorem 2, we consider the following lemma that gives some basic properties of basis classes.

Lemma 1.

(a) There exist \[
\left(\frac{|\mathcal{X}| + |\mathcal{Y}| - 1}{|\mathcal{X}|}\right)
\]
pairwise disjoint basis classes and \[
\bigcup_{h \in \mathcal{H}} B_h = \mathcal{F}.
\]

(b) Two functions $f, g \in \mathcal{F}$ are $h$-equivalent iff there exists a permutation $\pi$ of $\mathcal{X}$ such that $f \circ \pi = g$.

(c) $B_h$ is equal to the permutation orbit of any function $f$ with histogram $h$, i.e., \[
B_h = \bigcup_{\pi \in \Pi(\mathcal{X})} \{f \circ \pi\}.
\]

(d) Any subset $F \subseteq \mathcal{F}$ that is c.u.p. is uniquely defined by a union of pairwise disjoint basis classes.

Proof. (a) The number $|\mathcal{H}|$ of different histograms is given by $\left(\binom{|\mathcal{X}| + |\mathcal{Y}| - 1}{|\mathcal{X}|}\right)$, i.e., the number of distinguishable distributions (e.g., [3, p. 38]). Two basis classes $B_h$ and $B_{h_1}$, $h_1 \neq h_2$, are disjoint because functions in different basis classes have different histograms. The union $\bigcup_{h \in \mathcal{H}} B_h = \mathcal{F}$ because every function in $\mathcal{F}$ has a histogram.

(b) Let $f, g \in \mathcal{X}$ be two functions with the same histogram $h$. Then, for any $y \in \mathcal{Y}$, $f^{-1}(y)$ and $g^{-1}(y)$ are equal in size and there exists a bijective function $\pi_y$ between these two subsets. Then the bijection $\pi(x) = \pi_y(x)$, where $y = f(x)$, defines a permutation such that $f \circ \pi = g$. Thus, $h$-equivalence implies existence of a permutation. On the other hand, the histogram of
a function is invariant under permutation since for any \( y \in Y \) and \( \pi \in \Pi(X) \) it holds
\[
| (f \circ \pi)^{-1} (y) | = \sum_{x \in X} \delta (y, f(\pi(x))) = \sum_{x \in X} \delta (y, f(x)) = | f^{-1} (y) |,
\]
because \( \pi \) is bijective and the addends can be reordered. Thus, existence of a permutation implies \( h \)-equivalence.

(c) For a function \( f \) with histogram \( h \), let \( O_f = \bigcup_{\pi \in \Pi(X)} \{ f \circ \pi \} \) be the orbit of \( f \) under permutations \( \pi \). By (b), all functions in \( O_f \) have the same histogram and thus \( O_f \subseteq B_h \). On the other hand, for any functions \( g \in B_h \) there exists by (b) a permutation \( \pi \) such that \( f \circ \pi = g \) and thus \( B_h \subseteq O_f \).

(d) For a c.u.p. subset \( F \subseteq \mathcal{F} \), let \( F_h = B_h \cap F \) (i.e., \( F_h \) contains all functions in \( F \) with the same histogram \( h \)). By (a), all \( F_h \) are pairwise disjoint and \( F = \bigcup_{h \in \mathcal{H}} F_h \). Suppose \( F_h \neq \emptyset \). Since \( F \) is c.u.p. there exists a function \( f \in F_h \) that spans the orbit \( B_h \). Thus \( B_h \subseteq F \) and therefore \( F_h = B_h \). Because basis classes are disjoint, the union \( F = \bigcup_{h \in \mathcal{H}}: h \in \mathcal{H} \land F_h \neq \emptyset B_h \) is unique. \( \square \)

**Proof of Theorem 2.** By Lemma 1(a), the number of different basis classes is given by \( \binom{|X|+|Y|-1}{|X|} \). The number of different, nonempty unions of basis classes (equal to the cardinality of the power set of the set of all basis classes minus one for the empty set) is given by \( 2^{(|X|+|Y|-1)} - 1 \). By Lemma 1(d), this is the number of nonempty subsets of \( \mathcal{F} \) that are c.u.p. \( \square \)

**References**


