Published in *Information Processing Letters*, **86**(6), pp. 317-321, 2003

On Classes of Functions for which No Free Lunch Results Hold

Christian Igel and Marc Toussaint
Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
{Christian.Igel, Marc.Toussaint}@neuroinformatik.ruhr-uni-bochum.de

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 $f : \mathcal{X} \to \mathcal{Y}$, 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 $f : \mathcal{X} \to \mathcal{Y}$ to be optimized (also called fitness, energy, or cost functions). NFL theorems are concerned with non-repeating 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 $T_m = \langle (x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m)) \rangle$ represent m non-repeating 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 $Y(f, m, A) = \langle f(x_1), f(x_2), \ldots, f(x_m) \rangle$ 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 $\pi : \mathcal{X} \to \mathcal{X}$ 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, δ 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.

3 Fraction of Subsets Closed under Permutation

Let $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$ be the set of all functions mapping $\mathcal{X} \to \mathcal{Y}$. There exist $2^{|\mathcal{Y}|^{|\mathcal{X}|}} - 1$ non-empty subsets of $\mathcal{F}$. We want to calculate the fraction of subsets that are c.u.p.

Theorem 2. *The number of non-empty subsets of $\mathcal{Y}^{\mathcal{X}}$ that are c.u.p. is given by*

$$2^{\binom{|\mathcal{X}|+|\mathcal{Y}|-1}{|\mathcal{X}|}} - 1$$

and therefore the fraction of non-empty subsets c.u.p. is

$$\left(2^{\binom{|\mathcal{X}|+|\mathcal{Y}|-1}{|\mathcal{X}|}} - 1\right) \Big/ \left(2^{|\mathcal{Y}|^{|\mathcal{X}|}} - 1\right) \quad .$$

The proof is given in the appendix. 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 $\{0, 1\}^3 \to \{0, 1\}$ the fraction is $\approx 10^{-74}$.

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

$$\sqrt{2\pi} m^{m+1/2} e^{-m} \cdot e^{(12m+1)^{-1}} < m! < \sqrt{2\pi} m^{m+1/2} e^{-m} \cdot e^{(12m)^{-1}}$$

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

$$\left(2^{\binom{|\mathcal{X}|+|\mathcal{Y}|-1}{|\mathcal{X}|}} - 1\right) \Big/ \left(2^{\binom{|\mathcal{Y}|^{|\mathcal{X}|}}{}} - 1\right) < 2^{(e+e|\mathcal{Y}|/|\mathcal{X}|-e/|\mathcal{X}|)^{|\mathcal{X}|}-|\mathcal{Y}|^{|\mathcal{X}|}} \quad .$$

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

$$2^{(e+e|\mathcal{Y}|/|\mathcal{X}|-e/|\mathcal{X}|-|\mathcal{Y}|)^{|\mathcal{X}|}} \quad ,$$

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 this structure. More formally, there exists a non-trivial neighborhood relation on $\mathcal{X}$ based on which constraints on the set of functions under consideration are formulated. For example, with respect to a neighborhood relation we can define concepts like ruggedness or local optimality and constraints like upper bounds on the ruggedness or on the maximum number of local minima. Intuitively, it is likely that in a function class c.u.p. there exists a function that violates such constraints.

We define a simple neighborhood relation on $\mathcal{X}$ as a symmetric function $n : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$. Two elements $x_i, x_j \in \mathcal{X}$ are called neighbors iff $n(x_i, x_j) = 1$. We call a neighborhood non-trivial iff $\exists x_i, x_j \in \mathcal{X} : x_i \neq x_j \wedge n(x_i, x_j) = 1$ and $\exists x_k, x_l \in \mathcal{X} : x_k \neq x_l \wedge n(x_k, x_l) = 0$. It holds:

Theorem 3. *A non-trivial neighborhood on $\mathcal{X}$ is not invariant under permutations of $\mathcal{X}$.*

Proof. It holds $\exists x_i, x_j, x_k, x_l \in \mathcal{X} : x_i \neq x_j \wedge x_k \neq x_l \wedge n(x_i, x_j) = 0 \wedge n(x_k, x_l) = 1$. For any permutation π that maps x_i and x_j onto x_k and x_l, respectively, the invariance property, $\forall a, b \in \mathcal{X} : n(x_a, x_b) = n(\pi(x_a), \pi(x_b))$, is violated. □

Remark 1. Assume the search space $\mathcal{X}$ can be decomposed as $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_l, l > 1$ and let on one component $\mathcal{X}_i$ exist a non-trivial neighborhood $n_i : \mathcal{X}_i \times \mathcal{X}_i \to \{0, 1\}$. This neighborhood induces a non-trivial neighborhood on $\mathcal{X}$, where two points are neighbors iff their i-th components are neighbors with respect to n_i. 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 $d_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ on $\mathcal{Y}$, e.g., in the typical case of real-valued fitness function $\mathcal{Y} \subset \mathbb{R}$ 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 $f \in \mathcal{F}$ by

$$s^{\max}(f) = \max_{x_i, x_j \in \mathcal{X} \wedge n(x_i, x_j) = 1} d_{\mathcal{Y}}(f(x_i), f(x_j)) \ .$$

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

$$d^{\max}(f) = \max_{x_i, x_j \in \mathcal{X}} d_{\mathcal{Y}}(f(x_i), f(x_j)) \ .$$

Corollary 1. *If the maximum steepness $s^{\max}(f)$ of every function f in a non-empty subset $F \subset \mathcal{F}$ 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 $\mathcal{X}$ is non-trivial there exist two neighboring points x_k and x_l. There exists a permutation π 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 $s^{\max}(g \circ \pi) = d^{\max}(g) = \max_{f \in F} d^{\max}(f)$, 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 $f \in \mathcal{F}$ a point $x \in \mathcal{X}$ 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 $\mathcal{X}$, we define $l^{\max}(f)$ as the maximal number of minima that functions with the same $\mathcal{Y}$-histogram as f can have (i.e.,

functions where the number of $\mathcal{X}$-values that are mapped to a certain $\mathcal{Y}$-value are the same as for f, see appendix).

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 the appendix we prove that for any two functions f, g with the same $\mathcal{Y}$-histogram there exists a permutation $\pi \in \Pi(\mathcal{X})$ with $f \circ \pi = g$. Thus, it follows:

Corollary 2. *If the number of local minima of every function f in a non-empty subset $F \subset \mathcal{F}$ is constrained to be smaller than the maximal possible $\max_{f \in F} l^{\max}(f)$, 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 theorems 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 measure where algorithms have different performance when averaged over all the considered objective functions.

Acknowledgments

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.

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 $\{x | f(x) = y\}$ of y under f. Further, we call two functions f, g *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*

$$\binom{|\mathcal{X}| + |\mathcal{Y}| - 1}{|\mathcal{X}|}$$

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 π 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 $\binom{|\mathcal{X}|+|\mathcal{Y}|-1}{|\mathcal{X}|}$, i.e., the number of *distinguishable distributions* (e.g., [3], p. 38). Two basis classes B_{h_1} and B_{h_2}, $h_1 \neq h_2$, 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 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 π_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 $y \in \mathcal{Y}$ and $\pi \in \Pi(\mathcal{X})$ it holds $\left|(f \circ \pi)^{-1}(y)\right| = \sum_{x \in \mathcal{X}} \delta(y, f(\pi(x))) = \sum_{x \in \mathcal{X}} \delta(y, f(x)) = \left|f^{-1}(y)\right|$, because π 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 $O_f = \bigcup_{\pi \in \Pi(\mathcal{X})} \{f \circ \pi\}$ be the orbit of f under permutations π. 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 π 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 $F = \bigcup_{h:\, h \in \mathcal{H} \wedge F_h \neq \emptyset} B_h$ is unique.

□

Proof of theorem 2. By lemma 1(a), the number of different basis classes is given by $\binom{|\mathcal{X}|+|\mathcal{Y}|-1}{|\mathcal{X}|}$. The number of different, non-empty 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 $2^{\binom{|\mathcal{X}|+|\mathcal{Y}|-1}{|\mathcal{X}|}} - 1$. By lemma 1(d), this is the number of non-empty subsets of $\mathcal{F}$ that are c.u.p. □

References

[1] S. Droste, T. Jansen, and I. Wegener. Perhaps not a Free Lunch but at least a free appetizer. In W. Banzhaf, J. Daida, A. Eiben, M. H. Garzon, V. Honovar, M. Jakiela, and R. E. Smith, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '99)*, pages 833–839, Orlando, FL, USA, 1999. Morgan Kaufmann.

[2] S. Droste, T. Jansen, and I. Wegener. Optimization with randomized search heuristics – The (A)NFL theorem, realistic scenarios, and difficult functions. *Theoretical Computer Science*, 287(1):131–144, 2002.

[3] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. John Wiley & Sons, New York, 3. edition, 1971.

[4] C. Schumacher, M. D. Vose, and L. D. Whitley. The No Free Lunch and description length. In L. Spector, E. Goodman, A. Wu, W. Langdon, H.-M. Voigt, M. Gen, S. Sen, M. Dorigo, S. Pezeshk, M. Garzon, and E. Burke, editors, *Genetic and Evolutionary Computation Conference (GECCO 2001)*, pages 565–570, San Francisco, CA, USA, 2001. Morgan Kaufmann.

[5] D. Whitley. A Free Lunch proof for gray versus binary encodings. In W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, and R. E. Smith, editors, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'99)*, volume 1, pages 726–733, Orlando, FL, USA, 1999. Morgan Kaufmann.

[6] D. H. Wolpert and W. G. Macready. No Free Lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.