Evaluation of Evolutionary and Genetic Optimizers: No Free Lunch*

Thomas M. English
Bounded Theoretics
http://BoundedTheoretics.com

Abstract

The recent "no free lunch" theorems of Wolpert and Macready indicate the need to reassess empirical methods for evaluation of evolutionary and genetic optimizers. Their main theorem states, loosely, that the average performance of all optimizers is identical if the distribution of functions is average. [*An "optimizer" selects a sample of the values of the objective function. Its "performance" is a statistic of the sample.*] The present work generalizes the result to an uncountable set of distributions. The focus is upon the ~~conservation of information as an optimizer evaluates points~~ [*statistical independence of the selection process and the selected values*]. ~~It is shown that the information an optimizer gains about unobserved values is ultimately due to its prior information of value distributions.~~ [*The paper mistakes selection bias for prior information of the objective function.*] ~~Inasmuch as information about one distribution is misinformation about another, there is no generally superior function optimizer~~. Empirical studies are best regarded as attempts to ~~infer the prior information optimizers have about distributions~~ [*match selection biases to constrained problems*] — i.e., to determine which tools are good for which tasks.

0 Sampling Bias Is Not Information

This preface (Sect. 0) addresses expository errors in the original paper, but stands as a self-contained report. Specific corrections and amplifications appear in the remainder of the text (Sects. 1–6).

*Published in *Evolutionary Programming V: Proceedings of the Fifth Annual Conference on Evolutionary Programming*, L. J. Fogel, P. J. Angeline, and T. Bäck, eds., pp. 163–169. Cambridge, Mass: MIT Press, 1996.

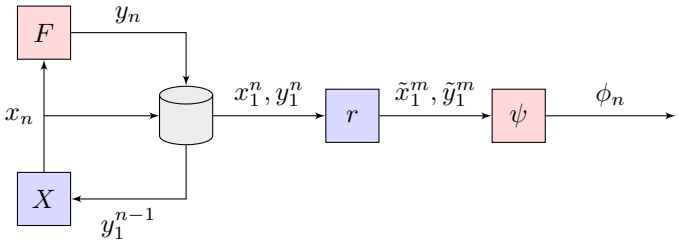

Figure 1: Objective function F and quality measure ψ comprise the optimization problem. An optimizer is a sampler X, along with a reduction r of the selection x_1^n and the sample y_1^n to outputs $\tilde{x}_1^m = x_{j_1}, \ldots, x_{j_m}$ and $\tilde{y}_1^m = y_{j_1}, \ldots, y_{j_m}$, with j_1^m strictly increasing. NFL analyses make four assumptions: the selection is non-repeating, the reduction is identical for all optimizers under consideration, the output $\tilde{y}_1^m$ depends only on the sample, and quality depends only on $\tilde{y}_1^m$. Accordingly, write $\tilde{y}_1^m = r(y_1^n)$ and $\phi_n = \psi(\tilde{y}_1^m)$. Then the quality $\psi(r(y_1^n))$ of the optimizer's output is a statistic $\phi = \psi \circ r$ of the sample it generates.

Black-box optimization may be decomposed into generation of a sample of the values of the objective function, and use of the sample to produce a result. "No free lunch" (NFL) analyses assume explicitly that sampling is without repetition, and define quality of the optimization result to be a statistic of the sample (Wolpert and Macready, 1995, 1997). Fig. 1 unpacks the tacit assumptions, the most important of which is that the optimizers under consideration differ only in selection of samples. Fig. 2 shows how postprocessing of the sample, assumed to be identical for all optimizers, is embedded in the statistic. Although the statistic combines part of the optimizer with part of the optimization problem, the NFL literature refers to it as the performance measure, and to samplers as optimization (or search) algorithms.

The better known of NFL theorems, including that of Sect. 3.2 (which appears also in Streeter, 2003, and Igel and Toussaint, 2005), are reinventions of basic results in probability (Häggström, 2007). They address sampling and statistics. This claim is justified by showing that the selection process

$$X_i \equiv X(F(X_1), \ldots, F(X_{i-1})),$$

$1 \leq i \leq n$, is statistically independent of the selected values

$$F_1^n(X) \equiv F(X_1), \ldots, F(X_n),$$

despite its data processing. Sect. 3.1 supplies proof for the special case of deterministic selection, with the domain and codomain of the random objective function restricted to finite sets. Here the result is generalized to random selection, with repetition allowed, and with the domain and codomain possibly infinite, though countable. Furthermore, the selection process is equipped to terminate. Although this would seem to complicate matters, the proof is greatly simplified.

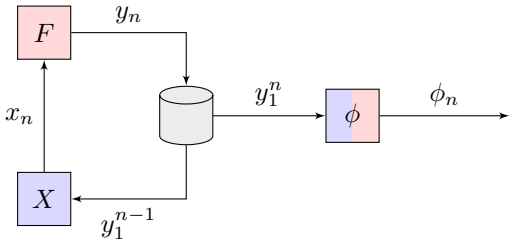

Figure 2: Random sampler X is statistically independent of random objective function F, and the selection $X(\lambda) = x_1, \ldots, X(y_1, \ldots, y_{n-1}) = x_n$ is statistically independent of the sample $F(x_1) = y_1, \ldots, F(x_n) = y_n$. The statistic $\phi = \psi \circ r$ is the composition of the quality measure ψ on optimizer outputs, given in the optimization problem, and the reduction r of samples to outputs, assumed to be identical for all optimizers. The selection and the sample are abbreviated $X_1^n = x_1^n$ and $F_1^n(X) = y_1^n$, respectively.

The selection process cannot be regarded as having or gaining information of unselected values of the objective function when it is statistically independent of the selected values. Data processing is a source of *bias* in the selection. The paper fails, in exposition, to recognize statistical independence for what it is, and refers instead to "conservation of information." (The formal results are correct.) It furthermore equates the propensity of an optimizer to perform better for some distributions of the random objective function and worse for others with prior information that somehow inheres in the optimizer. The notion that one entity simply has, rather than acquires, varying degrees of prior information and misinformation of others is incoherent.

The following subsection formalizes the system sketched in Fig. 2, proves that the selection is indeed statistically independent of the selected values, and goes on to demonstrate that a general NFL theorem, so called, follows almost immediately. This exercise leaves very little room in which to contend that the NFL theorems address something other than sampling. The preface concludes with a subsection that briefly deconstructs the paper's erroneous claims about information.

0.1 Formal Results on Sampling

0.1.1 Definitions

Sampling processes do not necessarily terminate, and it is convenient to treat them all as infinite. The distinguished symbol $\diamond$ is interpreted as the *sample terminator*. Let countable sets $\mathcal{X}_\diamond = \mathcal{X} \cup \{\diamond\}$ and $\mathcal{Y}_\diamond = \mathcal{Y} \cup \{\diamond\}$, where sets $\mathcal{X}$ and $\mathcal{Y}$ exclude $\diamond$. The random *objective function* F maps $\mathcal{X}_\diamond$ to $\mathcal{Y}_\diamond$, with $F(x) = \diamond$ if and only if $x = \diamond$.

Finite sequences of the forms $\alpha_1, \ldots, \alpha_n$ and $\gamma(\alpha_1), \ldots, \gamma(\alpha_n)$ are abbreviated α_1^n and $\gamma_1^n(\alpha)$, respectively. A *sampler* is a random function, statistically

independent of F, from the set of all finite sequences on $\mathcal{Y}_\diamond$ to $\mathcal{X}_\diamond$. A *selection* is a random vector X_1^n with

$$X_i \equiv X(F(X_1), \ldots, F(X_{i-1}))$$

$1 \leq i \leq n$, where X is a sampler. The selection is *non-repeating* if $X_1^n \in \pi(\mathcal{X})$ surely, where $\pi(\mathcal{X})$ is the set of all non-repeating sequences on $\mathcal{X}$.

A *statistic* is a function with the set of all finite sequences on $\mathcal{Y}_\diamond$ as its domain. Assume, without loss of generality, that the codomain of a statistic is a countable superset of its range, which is countable like the domain.

0.1.2 Statistical Independence of Selection and Sample

Probability measure is unproblematic when considering the selection X_1^n and the corresponding sequence $F_1^n(X)$, because both take values in countable sets. The following lemma establishes that the selection is statistically independent of the selected values, i.e., that it is correct to refer to $F_1^n(X)$ as a *sample* of the values $\{F(x) \mid x \in \mathcal{X}_\diamond\}$ of the objective function. The data processing in extensions of the selection, highlighted in the proof, is a potential source of selection bias, not information about the objective function.

Lemma. *Selection $X_1, \ldots, X_n$ is statistically independent of $F(X_1), \ldots, F(X_n)$.*

Proof. Let x_1^n and y_1^n be nonempty sequences on $\mathcal{X}_\diamond$ and $\mathcal{Y}_\diamond$, respectively, such that $P(F_1^n(x) = y_1^n) > 0$. Then

$$\begin{aligned} P(X_1^n = x_1^n, F_1^n(X) = y_1^n) &= P(X_1^n = x_1^n, F_1^n(x) = y_1^n) \\ &= P(X_1^n = x_1^n \mid F_1^n(x) = y_1^n) \cdot P(F_1^n(x) = y_1^n), \end{aligned}$$

and

$$\begin{aligned} &P(X_1^n = x_1^n \mid F_1^n(x) = y_1^n) \\ &\quad = P(X(F_1^0(x)) = x_1, \ldots, X(F_1^{n-1}(x)) = x_n \mid F_1^n(x) = y_1^n) \\ &\quad = P(X(y_1^0) = x_1, \ldots, X(y_1^{n-1}) = x_n \mid F_1^n(x) = y_1^n) \\ &\quad = P(X(y_1^0) = x_1, \ldots, X(y_1^{n-1}) = x_n) \end{aligned}$$

because sampler X is statistically independent of objective function F. □

Corollary 1. *For all statistics ϕ, selection $X_1, \ldots, X_n$ is statistically independent of $\phi(F(X_1), \ldots, F(X_n))$.*

0.1.3 Simple Derivation of a So-Called NFL Theorem

The following theorem uses the symbol $\overset{\mathcal{D}}{=}$ to denote equality in probability distribution. It says, loosely, that the distribution of a statistic depends on the choice of *sampler* if and only if the distribution of the statistic depends on the choice of *sample*.

Theorem. *Let $x_1, \ldots, x_n$ be a nonempty, non-repeating sequence on $\mathcal{X}$, and let ϕ be a statistic. Then*

$$\phi(F(X_1), \ldots, F(X_n)) \stackrel{\mathcal{D}}{=} \phi(F(x_1), \ldots, F(x_n)) \tag{1}$$

for all non-repeating selections $X_1, \ldots, X_n$ if and only if

$$\phi(F(w_1), \ldots, F(w_n)) \stackrel{\mathcal{D}}{=} \phi(F(x_1), \ldots, F(x_n)) \tag{2}$$

for all non-repeating sequences $w_1, \ldots, w_n$ on $\mathcal{X}$.

Proof.

($\Rightarrow$) Suppose that (1) holds for all non-repeating selections X_1^n, and let w_1^n be a non-repeating sequence on $\mathcal{X}$. There exists sampler X with constant selection $X_1^n = w_1^n$, and thus (2) follows.

($\Leftarrow$) Suppose that (2) holds for all non-repeating sequences w_1^n on $\mathcal{X}$, and let X_1^n be a non-repeating selection. A condition stronger than (1) follows. For each and every realization $X_1^n = w_1^n$, non-repeating on $\mathcal{X}$ by definition,

$$\phi(F_1^n(X)) \stackrel{\mathcal{D}}{=} \phi(F_1^n(w)) \stackrel{\mathcal{D}}{=} \phi(F_1^n(x)).$$

The first equality holds by Corollary 1, and the second by assumption. □

Corollary 2. *Let $x_1, \ldots, x_n$ be a nonempty, non-repeating sequence on $\mathcal{X}$. Then*

$$F(X_1), \ldots, F(X_n) \stackrel{\mathcal{D}}{=} F(x_1), \ldots, F(x_n) \tag{3}$$

for all non-repeating selections $X_1, \ldots, X_n$ if and only if

$$F(w_1), \ldots, F(w_n) \stackrel{\mathcal{D}}{=} F(x_1), \ldots, F(x_n) \tag{4}$$

for all non-repeating sequences $w_1, \ldots, w_n$ on $\mathcal{X}$.

Proof. Set the statistic ϕ in the theorem to the identity function. □

The corollary is essentially the NFL theorem of Sect. 3.2, and also (Streeter, 2003; Igel and Toussaint, 2005). When equality (4) holds for all n, the random values $\{F(x) \mid x \in \mathcal{X}\}$ of the objective function are said to be *exchangeable* (Häggström, 2007).

0.2 Understanding the Misunderstanding of Information

The root error is commitment to the belief that information is the cause of performance in black-box optimization (search). The NFL theorems arrived at a time when researchers commonly claimed that evolutionary optimizers gained information about the fitness landscape, and adapted themselves dynamically to improve performance. Wolpert and Macready (1995) observe that superior performance on a subset of functions is offset precisely by inferior performance on the complementary subset. In online discussion of their paper, Bill Spears

referred to this as *conservation of performance*. My paper suggests that conservation of information accounts for conservation of performance.

The lemma of Sect. 3.1, "Conservation of Information," expresses the absolute uninformedness of the sample selection process. The performance of a black-box optimizer has nothing whatsoever to do with its information of the objective function. But the paper recognizes only that information *gain* is impossible, and claims incoherently that prior information resides in the optimizer itself. Conservation of this chimeral information supposedly accounts for conservation of performance in optimization. Here are the salient points of illogic:

1. Information causes performance.
2. The optimizer gains no exploitable information by observation, so it must be prior information that causes performance.
3. There is no input by which the optimizer might gain prior information, so it must be that prior information inheres in the optimizer.
4. Prior information of one objective function is prior misinformation of another. Conservation of performance is due to conservation of information.

It should have been obvious that prior information is possessed only after it is acquired. The error is due in part to a mangled, literalistic half-reading of (Wolpert and Macready, 1995, p. 8, emphasis added):

> The NFL theorem illustrates that even if we know something about [the objective function] ... but don't *incorporate that knowledge into* [*the sampler*] then we have no assurances that [the sampler] will be effective; we are simply relying on a fortuitous matching between [the objective function] and [the sampler].

The present work (Sect. 6) concludes that:

> The tool literally carries information about the task. Furthermore, optimizers are literally tools — an algorithm implemented by a computing device is a physical entity. In empirical study of optimizers, the objective is to determine the task from the information in the tool.

This reveals confusion of one type of information with another. When a toolmaker imparts form to matter, the resulting tool is in-formed to suit a task. But such form is not prior information. Having been formed to perform is different from having registered a signal relevant to the task. An optimization practitioner may gain information of a problem by observation, and then form a sampler to serve as a proxy in solving it. Although the sampler is informed to act as the practitioner would, it is itself uninformed of the problem to which it is applied, and thus cannot justify its own actions. The inherent form that accounts for its performance is sampling bias.

Unjustified application of a biased sampler to an optimization problem is merely biased sampling by proxy. The NFL theorems do not speak to this fundamental point. They specify conditions in which all of the samplers under

consideration are equivalent in overall performance, or are nearly so. Ascertaining that none of these theorems applies to a real-world circumstance does not justify a bias, but instead suggests that justification may be possible. There is never a "free lunch" for the justifier.

1 Introduction

In "No Free Lunch Theorems for Search," Wolpert and Macready (1995) have established that there exists no generally superior function optimizer. There is no "free lunch" in the sense that an optimizer "pays" for superior performance on some functions with inferior performance on others. Their paper shows that if the distribution of functions is uniform, then gains and losses balance precisely, and all optimizers have identical average performance.

The news of "no free lunch" spread rapidly through the evolutionary computation (EC) community. Empirical comparison of genetic and evolutionary optimizers has long been the cornerstone of research in the field (Bäck and Schwefel, 1993; Fogel, 1995). Furthermore, many workers regard natural evolution as an optimized optimizer that produces outstanding results under all circumstances. Thus it is not surprising that the significance of "no free lunch" has been debated vigorously in the community. It is surprising, however, that few understand the fundamental reasons for the result.

It is even more surprising that the fundamental reasons have changed several times during the preparation of this paper. The primary objective was, and is, to provide EC practitioners with an accessible explanation. It is purely fortuitous that simplification of the formal presentation has led to new results on "no free lunch" distributions — that is, function distributions for which random walks are optimal. The formal demonstration depends primarily upon a theorem that describes how information is ~~conserved~~ [*not gained*] in optimization. This Conservation Lemma states that when an optimizer evaluates points, the posterior joint distribution of values for those points is exactly the prior joint distribution. Put simply, observing the values of a randomly selected function does not change the distribution.

To see the usefulness of the lemma, suppose that function values are independent and identically distributed (iid). The Conservation Lemma indicates that the values observed by an optimizer will also be iid. In essence, the points are identical roulette wheels, and all ways of visiting n distinct points correspond to identically distributed value sequences. Thus there is no distinction between optimizers' value-sequence distributions. Any distinction between the value-sequence distributions on a subset of functions is "canceled" by a distinction on the complementary subset. This cancellation is most intuitive when values are iid uniform — i.e., when the distribution of functions is uniform (Wolpert and Macready, 1995).

The following section presents several definitions and concepts required in Section 3, where the formal results are derived. Section 4 discusses the significance of the results. Section 5 makes suggestions for future research in genetic

and evolutionary optimization. Section 6 states several conclusions, and briefly argues that no conclusions about natural evolution are justified.

2 Definitions, Concepts, and Notation

The section starts with a brief review of functions. Then the notion of a random distribution of functions is presented in a simple, but somewhat unusual, way. After that, the notions of mutual independence and mutual information are explained in terms relevant to the present work. Finally, the term *walk* (as in "random walk") is defined as an abstraction of an optimizer's decisions to evaluate points in a particular order. The theorems of Section 3 will refer to walks, and not to optimizers.

2.1 Functions

Let S and T be sets. A function f from domain S to codomain T, denoted $f : S \to T$, is a subset of $S \times T$ such that for each $x \in S$ there is exactly one $y \in T$ for which $(x, y) \in f$. The expression $y = f(x)$ is equivalent to $(x, y) \in f$. The range of f is $f(S) = \{f(x) : x \in S\}$.

To paraphrase loosely, a function is a set of value assignments. Every domain element has exactly one value in the codomain. The codomain elements that are actually "used" as values comprise the range. For instance, if $f = \{(b, 2), (a, 2), (c, 1)\}$ then the domain is $\{a, b, c\}$ and the range is $\{1, 2\}$. Any superset of the range may be regarded as the codomain. Note that f is equivalent to the union of three disjoint and non-empty functions; i.e., $f = \{(a, 2)\} \cup \{(b, 2)\} \cup \{(c, 1)\}$.

In the present work, as in (Wolpert and Macready, 1995), the domain and codomain are assumed to be finite. The significance of this restriction is mathematical, rather than practical, because digital computers represent only finite sets. Extension of the formal results to infinite sets is straightforward, but interpretation becomes considerably more complicated.

2.2 Distributions of Functions

It is no more odd to say that a random variable is distributed on a set of functions than to say that it is distributed on a set of Presidential candidates. The random variable models uncertainty about an event, and there are no restrictions upon the set of possible outcomes. For instance, let random variable F be distributed on the set of all functions from S to T. The expression $P(F = f)$ denotes the probability that the outcome [*realization*] of F is a particular function $f : S \to T$. If F is uniformly distributed, then $P(F = f) = |T|^{-|S|}$ for each of the $|T|^{|S|}$ functions f from S to T.

Recalling the definition of a function, F can be written

$$F = \{(x_1, F(x_1)), (x_2, F(x_2)), \ldots (x_n, F(x_n))\},$$

where $|S| = n$. The random variables $F(x)$, $x \in S$, are sometimes referred to as the values of F. Note that x is an index, not an argument, in the expression $F(x)$. This notation permits the straightforward statement,

$$P(F = f) = P(F(x_1) = f(x_1), \ldots, F(x_n) = f(x_n))$$

for all functions f. That is, the distribution of a random function corresponds to a joint distribution of its values.

As an example of the relationship between the distribution of F and the distributions of its values, note that F is uniform on the set of all functions from S to T if and only if the distributions of its values are mutually independent and uniform on T.

It is sometimes convenient to refer to an observed outcome of a random variable, X, as a *realization* of X. In the context of function optimization, some additional terminology is helpful. The optimizer operates upon a realization of F. *Evaluating* or *visiting* point x is equivalent to observing the realization of the value $F(x)$. It is said that the value of x has been *observed*. The values of unvisited points are said to be *unobserved*.

2.3 Mutual Independence

The random variables $F(x)$, $x \in S$, are mutually independent if and only if

$$P(F = f) = \prod_x P(F(x) = f(x))$$

for all functions f from S to T. That is, for every joint outcome the probability is the product of the probabilities of the individual outcomes. Equivalently, the values are mutually independent if and only if all conditional distributions are identical to their unconditional distributions. That is,

$$P(G = g \mid H = h) = P(G = g),$$

for all outcomes g of G and h of H, where G and H are disjoint subsets of F. Put simply, the distributions of unobserved values do not change when some values have been observed.

2.4 Entropy and Mutual Information

This subsection provides a sufficient set of concepts and definitions to understand the information theoretic analysis of function optimization in later sections.

2.4.1 Entropy

The entropy of a distribution [*probability distribution p of random variable X*] is the average uncertainty of the outcome or, equivalently, the average information

gained when the outcome is observed. For each possible outcome x, $-\log_2 p(x)$ [*bits*] is the information gained when x is observed. The average information is

$$H(X) = -\sum_x p(x) \log_2 p(x),$$

where x ranges over the possible outcomes of random variable X.

This measure of information has profound physical significance. Consider a scenario in which an observer uses a binary code and some method of transmission to tell a non-observer about the outcomes. A compelling measure of information in the outcomes is the minimum bit rate (bits per outcome) that suffices to keep the non-observer fully informed. Determining the minimum bit rate over all possible codes is a daunting prospect, however. In each code, outcomes have binary names, and the bit rate is $\sum_x p(x)n(x)$, where $n(x)$ is the length of the name for outcome x. Fortunately, a fundamental theorem of information theory states that the bit rate is minimized when each outcome is assigned a name of $-\log_2 p(x)$ bits. [*This assumes a prefix code.*] These ideal lengths are not necessarily integers, but there is an efficient algorithm that always succeeds in generating a real code with bit rate [*approaching the bound*]

$$\sum_x p(x)n(x) = -\sum_x p(x) \log_2 p(x).$$

[*It may be necessary to encode outcomes in large blocks, rather than individually.*] Thus the entropy of a distribution is the minimum bit rate that allows a non-observer to be fully informed of outcomes. Other forms of entropy are $H(X, Y)$, the joint entropy, and $H(X|Y)$, the conditional entropy, where X and Y are random variables. As the names and notations suggest, $H(X, Y)$ is the average information in joint outcomes of X and Y, and $H(X \mid Y)$ is the average information in the outcome of X when the outcome of Y is known. The identity $p(x, y) = p(x \mid y)p(y)$ has as its analog $H(X, Y) = H(X \mid Y) + H(Y)$.

2.4.2 Mutual Information

Another measure of information is defined in terms of entropy. The mutual information of the distributions of X and Y is

$$I(X; Y) = H(X) - H(X \mid Y).$$

The difference $H(X) - H(X \mid Y)$ is the reduction in average uncertainty when the outcome of Y is known. The reduction is non-negative, although some outcomes of Y may supply negative information (i.e., increase the uncertainty of the outcome of X).

In the present work, the mutual information of the distribution of a function value and a joint distribution of function values is of particular interest. For a random variable F distributed on the set of all functions from S to T, the mutual information of the distributions of $F(x)$ and $G \subset F$, $x \in S$, $F(x) \notin G$, is

$$I(F(x); G) = H(F(x)) - H(F(x) \mid G).$$

This may be interpreted as potential reduction in an ~~optimizer's~~ [*practitioner's*] average uncertainty about the value of x after visiting a certain set of points. Actual reduction requires knowledge of the joint distribution of G and $F(x)$. ~~That is, when the realization of G is g, $p(y \mid g) = p(y, g)/p(g)$ for all possible realizations y of $F(x)$. The optimizer gains information about $F(x)$ to the degree that it obtains $p(y|g)$. Thus the information comes from the optimizer's prior information of $p(y, g)$, not g itself.~~ [*Knowledge may be a matter of degree, described perhaps in terms of the relative entropy of the assumed and actual distributions.*]

By definition, the values of F are mutually independent if and only if $H(F(x) \mid G) = H(F(x))$ for all $F(x) \in F$ and $G \subset F$, $F(x) \notin G$. But $H(F(x) \mid G) = H(F(x))$ is equivalent to $I(F(x); G) = 0$. Thus mutual independence is equivalent to zero mutual information of all values and subsets of values. In this context, it is instructive to note that the total amount of information gained about unobserved values from observed values is $\sum_x H(F(x)) - H(F)$ bits, on average. The difference is zero if and only if the values of F are mutually independent.

2.5 Walks of Functions

[*A so-called walk is a* non-repeating *selection.*]

The notion of 'honest' selection of a sequence of points is formalized with the definition of walk. In essence, it is dishonest to evaluate points and omit them from the sequence. It is also dishonest to conceal the order in which points are evaluated.

Formally, let $\mathbf{x}$ denote any finite sequence of points in the domain of function f. Sequence $\mathbf{x}$ is a walk of f if and only if $\mathbf{x}$ is empty or $\mathbf{x} = \mathbf{x}'x$ such that

1. $\mathbf{x}'$ is a walk of f,
2. x does not occur in $\mathbf{x}'$, and
3. x is selected without reference to values of points other than those in $\mathbf{x}'$.

Note that the definition does not preclude stochastic selection of $m > 1$ points at a time. Any permutation of simultaneously selected points may be added to the end of the walk, subject to the constraint that no points in the walk are duplicated. Thus the parallel exploration that characterizes evolutionary and genetic algorithms is not excluded.

When functions are randomly distributed, the value sequence of an arbitrary walk $\mathbf{x} = x_1 \ldots x_n$, is randomly distributed as well. The expression $p_{\mathbf{x}}(y_1, \ldots, y_n)$ denotes the probability that value sequence $y_1 \ldots y_n$ corresponds to $\mathbf{x}$.

3 Fundamental Theorems

This section explores the relationship between properties of the distribution of functions and properties of the distribution of value sequences for a walk. Of

particular interest are the conditions under which all walks $\mathbf{w}$ and $\mathbf{x}$ of identical length have identical value-sequence distributions $p_{\mathbf{w}}$ and $p_{\mathbf{x}}$. When the value-sequence distribution depends only upon the length of the walk, it clearly does not depend upon the procedure by which the walk is generated.

3.1 Conservation of Information

[*In the notation of the preface, the selection X_1^n and the sample $F_1^n(X)$ are statistically independent. This means that samplers operate with no information whatsoever of unrealized values of the objective function. Nonexistent information is not conserved.*]

An important property of walks is that they provide no information about the values of visited points. The ubiquitous claim that optimizers gain and exploit information about functions is ~~somewhat misleading~~ [*wrong*]. The following proof shows that optimizers ~~gain information about the values of unvisited points without gaining information about the function~~ [*are nothing but biased samplers*]. ~~This apparent paradox is resolved by noting that optimizers exploit the redundancy (mutual information) of value distributions.~~

Lemma (Conservation). Let F be distributed on the set of all functions from finite set S to finite set T. If $\mathbf{x} = x_1 \dots x_n$ is a walk of F and $(y_1, \dots, y_n)$ is in T^n then

$$p_{\mathbf{x}}(y_1, \dots, y_n) = P(F(x_1) = y_1, \dots, F(x_n) = y_n).$$

The proof proceeds by induction on n. For $n = 1$, $p_{\mathbf{x}}(y_1) = P(F(x_1) = y_1)$ because x_1 is by definition selected without reference to the value of any point. Now suppose that the equality holds for $n = k$, $1 \le k < |S|$. For arbitrary walk $\mathbf{x} = x_1 \dots x_{k+1} = \mathbf{w}x_{k+1}$, the prefix $\mathbf{w}$ is by definition a walk, and

$$p_{\mathbf{x}}(y_1, \dots, y_{k+1}) = p_{\mathbf{x}}(y_{k+1} \mid y_1, \dots, y_k)\ p_{\mathbf{w}}(y_1, \dots, y_k).$$

The induction step is completed by showing that each factor in the right-hand side can be rewritten as the corresponding factor in

$$\begin{aligned} &P(F(x_1) = y_1, \dots, F(x_{k+1}) = y_{k+1}) \\ &= P(F(x_{k+1}) = y_{k+1} \mid F(x_1) = y_1, \dots, F(x_k) = y_k) \\ &\qquad \times P(F(x_1) = y_1, \dots, F(x_k) = y_k). \end{aligned}$$

By definition, x_{k+1} is selected without reference to values of points other than $x_1, \dots, x_k$, and therefore

$$p_{\mathbf{x}}(y_{k+1} \mid y_1, \dots, y_k) = P(F(x_{k+1}) = y_{k+1} \mid F(x_1) = y_1, \dots, F(x_k) = y_k).$$

By hypothesis,

$$p_{\mathbf{w}}(y_1, \dots, y_k) = P(F(x_1) = y_1, \dots, F(x_k) = y_k).$$

This establishes that the equality stated in the lemma holds for $n = 1, \dots, |S|$.

3.2 Distributions Independent of Walk Selection

[*This is a poor argument that the distribution of the sample is independent of the sampler if and only if the random values* $\{F(x) \mid x \in \mathcal{X}\}$ *are* exchangeable *(Häggström, 2007).*]

It follows from the Conservation Lemma that $p_{\mathbf{x}} = p_{\mathbf{w}}$ for all walks $\mathbf{x} = x_1 \dots x_n$ and $\mathbf{w} = w_1 \dots w_n$ if and only if the distributions of all sets $\{F(x_1), \dots, F(x_n)\}$ and $\{F(w_1), \dots, F(w_n)\}$ of n values are identical. If all sets of n values are identically distributed for $n = 1, \dots, |S|$, then the distribution of value sequences is identical for all ways of selecting walks. Furthermore, all conditional distributions $p_{\mathbf{x}}(y_{k+1} \mid y_1, \dots, y_k)$ depend only upon k. This indicates that all sets of k observed values supply the same information about each of the unobserved values.

3.3 Mutually Independent Value Distributions

[*If the random values in* $\{F(x) \mid x \in \mathcal{X}\}$ *are mutually independent, then so are those in* $F_1^n(X)$*, by independence of the non-repeating selection* X_1^n *and the sample. In this circumstance the practitioner cannot use realized values to predict unrealized values. There are no decisions to be made while sampling, and the practitioner can do no better than to select* $X_1^n = x_1^n$ *in advance.*]

If function values are mutually independent, each of the posterior distributions $p_{\mathbf{x}}(y_{k+1} \mid y_1, \dots, y_k)$ is identical to the corresponding prior, $P(F(x_{k+1}) = y_{k+1})$. Thus an optimizer [*no, the practitioner*] can exploit only prior information [*of individual random values* $F(x)$] — there is no mutual information. To be more explicit, no strategy is better than one that selects a fixed walk on the basis of the prior distribution of values, irrespective of the realization of F.

Theorem (Independent Values). Let F be distributed on the set of all functions from finite set S to finite set T. Also let $\mathbf{x} = x_1 \dots x_n$ be a walk of F and let $(y_1, \dots, y_n)$ be an element of T^n. If the values $F(x)$, $x \in S$, are mutually independent then

$$p_{\mathbf{x}}(y_1, \dots, y_n) = P(F(x_1) = y_1) \times \cdots \times P(F(x_n) = y_n).$$

This is demonstrated by writing

$$\begin{aligned} p_{\mathbf{x}}(y_1, \dots, y_n) &= P(F(x_1) = y_1, \dots, F(x_n) = y_n) \\ &= P(F(x_1) = y_1) \times \cdots \times P(F(x_n) = y_n). \end{aligned}$$

The first step is justified by the Conservation Lemma, and the second by the mutual independence of $F(x_1), \dots, F(x_n)$.

3.4 Independent and Identical Value Distributions

[*If the random values in* $\{F(x)\}$ *are i.i.d., then they are exchangeable. When* F *is uniform on finite* Y^X*, the values in* $\{F(x)\}$ *are i.i.d. uniform.*]

If values are not only mutually independent, but identically distributed, then every ordering of points corresponds to an identical sequence of value distributions. That is, sequences of n iid values are iid n-sequences of values. As is evident in the following corollary, the distribution $p_{\mathbf{x}}$ depends only upon the length of x.

Corollary (IID Values). If, in addition to the hypotheses of the Independent Values Theorem, the values $F(x)$, $x \in S$, are identically distributed as random variable Y then

$$p_{\mathbf{x}}(y_1, \ldots, y_n) = \prod_i P(Y = y_i).$$

To verify, substitute $P(Y = y_i)$ for $P(F(x_i) = y_i)$, $i = 1, \ldots, n$, in the equality of the theorem.

3.5 "Needle in a Haystack" Functions

[*Exchangeable values $\{F(x)\}$ are mutually independent only if identically distributed. To see this, suppose that exchangeable $F(x_1)$ and $F(x_2)$ are independently, but not identically, distributed. Then the sequence $F(x_1), F(x_2)$ is not distributed as $F(x_2), F(x_1)$, a contradiction. In short, the practitioner cannot necessarily exploit the mutual information of values $\{F(x)\}$.*]

The IID Values Corollary is much stronger than a statement that all walks of a given length have identically distributed value sequences. It says that all points in all walks have identically distributed values. As one might guess, mutual independence is not a necessary condition for the walk selection procedure to be irrelevant to the distribution of value sequences. There are function distributions in which the mutual information of value distributions cannot be exploited by any optimizer. It cannot be exploited because every set of k observed values provides the same information about each of the unobserved values (Section 3.2).

This is illustrated by constructing a distribution of "needle in a haystack" functions. Let the domain and codomain be $S = \{a, b, c, d\}$ and $T = \{0, 1\}$, respectively. Let random variable F be uniformly distributed on the four functions from S to T that assign 1 to exactly one element of the domain. It is easily verified that all value subsets of equal size are identically distributed. Thus all procedures for generating walks yield identical value-sequence distributions.

It remains to be shown that the value distributions are mutually informative. For each $x \in S$, $P(F(x) = 0) = 3/4$ and $P(F(x) = 1) = 1/4$. This gives entropy of $H(F(x)) \approx 0.81$ bits for each value. With four equiprobable realizations of F, $H(F) = 2$, and the total mutual information of the value distributions is $\sum_x H(F(x))H(F) \approx 3.24 - 2 = 1.24$ bits.

4 Discussion

4.1 The Source of Information

The Conservation Lemma indicates that a walk-generating procedure [*begins with no information, and*] gains no information about [*unrealized values of*] the function. As mentioned in Section 2.4.2, mutual information is not gained from observed values. ~~An optimizer exploits mutual information only to the degree that it is informed of the prior distribution of functions.~~ [*The practitioner possibly exploits mutual information.*] Mutual information is a measure of the information that can be gained ~~from~~ [*about unrealized values from realized values by knowing*] the prior distribution. [*But this information is not necessarily useful. Exchangeable* $\{F(x)\}$ *generally are not independent.*]

Note that not all parts of the joint distribution are equally relevant to locating optima. In some cases there is strong regularity in the structure of the distribution which can be captured in a simple procedure — e.g., consider optimization of quadratic functions. ~~Thus there are subtle questions as to what prior information is embedded in the algorithm, and as to how it is encoded.~~ [*A practitioner with prior information of an optimization problem may form a sampler to serve as a proxy in solving the problem. Although the sampler is informed to decide as the practitioner would, it cannot obtain prior information of the problem as the practitioner does. See Sect. 6.*]

4.2 No Free Lunch

The present work does not contradict earlier results (Wolpert and Macready, 1995), because a distribution of functions is uniform if and only if the values are iid uniform. The uniform distribution is a very special case, because it is the average of all distributions. The notion that an optimizer has to "pay" for its superiority on one subset of functions with inferiority on the complementary subset is easiest to understand in the case of the uniform. The issue of whether the distribution of problems in the world is uniform is irrelevant. The point is to gain insight into the economy of information and optimization performance.

4.3 Optimizing Uniformly Distributed Functions

[*This subsection remains of value because it explodes a false intuition. However, almost all objective functions are implemented by no program much shorter than than one that stores all of the values in a lookup table. Unless the solution space is small, physical existence of such a program is implausible.*]

The obvious interpretation of "no free lunch" is that no optimizer is faster, in general, than any other. This misses some very important aspects of the result, however. One might conclude that all of the optimizers are slow, because none is faster than enumeration. And one might also conclude that the unavoidable slowness derives from the perverse difficulty of the uniform distribution of test functions. Both of these conclusions would be wrong.

	Probability [p]		
Fraction [q]	**0.01**	**0.001**	**0.0001**
0.99	458	687	916
0.999	4603	6904	9206
0.9999	46049	69074	92099
0.99999	460515	690772	921029

Table 1: Number of trials required to obtain a particular quality at a particular probability. [*Number of trials for which the probability is p that all values are in the lower fraction q of the range. A* trial *is an evaluation of a domain element drawn uniformly at random.*]

If the distribution of functions is uniform, the optimizer's best-so-far value is the maximum of n realizations of a uniform random variable. The probability that all n values are in the lower q fraction of the codomain is $p = q^n$. Exploring $n = \log_2 p$ [$n = \log(p)/\log(q)$] points makes the probability p that all values are in the lower q fraction. Table 1 shows n for several values of q and p.

It is astonishing that in 99.99% of trials a value better than 99.999% of those in the codomain is obtained with fewer than one million evaluations. This is an average over all functions, of course. It bears mention that one of them has only the worst codomain value in its range, and another has only the best codomain value in its range.

Breeden (1994) has given an analogous distribution-free result for finite functions. Suppose that all points have been ranked according to value, with ties broken arbitrarily. Further, let it be the rank function, rather than the given test function, that is optimized. If a point is drawn randomly from the domain, the value is uniform on the set of ranks. It follows that randomly drawing n points, with replacement, is equivalent to sampling a uniform random variable n times. This is precisely the condition underlying the computations in the table above. Thus the table also describes the relationship between rank, probability, and number of evaluations in random optimization of any finite function. In this case, however, the numbers do not represent an average over functions. They apply to each rank function individually.

How can test functions from a distribution with absolutely no structure be so easy, on average, to optimize? When function values are drawn independently from a uniform distribution, high values are as likely as low values. High and low values, both, tend to be spread throughout the domain. Every point is a good one to try, and the order in which points are tried is irrelevant. Intuitively, when there is no structure to help the optimizer find good points, there is also no structure to hide good points. As the next subsection shows, the number of good points is also very important.

4.4 The Hardest Distributions Are the Easiest

The "needle in a haystack" distribution (Section 3.5) is the hardest distribution for function maximizers. Knowing which element of the domain is assigned the good value is equivalent to knowing the identity of the function. Thus the entropy of the distribution of good points is equal to the entropy of the distribution of functions. The location of good points cannot be more uncertain.

The most difficult distribution for maximizers is the least difficult for minimizers. When the sense of optimality is reversed, the problem is to find the "hay." The entropy of the location of good points cannot be smaller without being zero. Changing the optimality criterion does not change the distribution of values, and thus there is no mutual information to be exploited by minimizers. That is, there is no strategy for avoiding the one point with a bad value.

The reason that the mutual information cannot be exploited is interesting. The location of the maximum is uniformly distributed on the domain, and all non-maxima have identical values. Observing 0's at visited points yields no information as to which of the unvisited points has the value of 1. The mutual information corresponds to reduction in uncertainty as to whether one of the unvisited points is the optimum. Observing a 1 removes all uncertainty about the values of unvisited points.

Recall that the domain of the four functions in the distribution is $\{a, b, c, d\}$, and that $H(F(x)) = 0.811$ bits for all points x. When the values of three points have been observed there is no uncertainty in the value of the remaining point. This indicates that the joint entropy of any three value distributions is 2 bits. It is easy to determine that the joint entropy of any two distributions is 1.5 bits. Thus, on average, the first value observation supplies 0.811 bits, the second $1.5 - 0.811 = 0.689$ bits, the third $2 - 1.5 = 0.5$ bits, and the fourth $2 - 2 = 0$ bits of information. The corresponding mutual information values are 0, 0.122, 0.311, and 0.811 bits. As indicated in Section 3.5, the total mutual information of the value distributions is 1.244 bits.

5 Ramifications for Future Studies

From the discussion of preceding sections, there emerges a clear picture of what empirical studies can and should do. ~~Perhaps the most important observation is that each optimizer has knowledge of some distribution of functions.~~ [*A sampler does not have knowledge, i.e., justified true belief. It may operate with* bias, *i.e., totally unjustified belief about the problem.*] Thus empirical performance assessment in the absence of a distribution of problems is meaningless. Of course, the class of multimodal functions is often identified in the EC literature. But if one takes multimodal to mean "not unimodal," then virtually all functions are multimodal. Thus the implicit uniform distribution on the class makes the performance of all optimizers nearly identical. Specifying a class that is too broad is not much better than specifying no class at all.

The literature is dominated, however, by continuous functions selected for

response surface shape. This amounts to a strong Euclidean bias. The bias is particularly clear in many descriptions of how an evolutionary algorithm "finds its way" to an optimum, perhaps adapting itself to the orientation of "ravines" in the response surface. (Similar, but not Euclidean, biases are evident in genetic algorithms. The number of variations of genetic algorithms makes general statements difficult.) An appropriate way for research to proceed would be for the properties of interest — they do exist — to be explicitly identified, and for the distribution of functions with those properties to be sampled. Means for random sampling do not necessarily exist, but that does not reduce the importance of obtaining a representative sample by some reasonable means. Note, also, that difficulties in random sampling generally diminish as the class is restricted.

5.1 "Promising" Algorithms

Anyone slightly familiar with the EC literature recognizes the paper template "Algorithm X was treated with modification Y to obtain the best known results for problems P_1 and P_2." Anyone who has tried to find subsequent reports on "promising" algorithms knows that they are extremely rare. Why should this be the case?

A claim that an algorithm is the very best for two functions is a claim that it is the very worst, on average, for all but two functions. Indeed, the only way to make an algorithm faster on a given set of functions is to make it slower on others. It is possible, and clearly undesirable, for speed-ups on particular functions to be attended by slow-downs on other functions in the class the algorithm is intended to handle well.

In studies of algorithm X-with-Y, there is rarely explicit indication of the class of functions the algorithm should optimize rapidly. It is easy to discern (particularly after conversations with authors) that X-with-Y is "promising" because the author believes that some additional modification will give excellent performance on at least one benchmark in $\{P_3, ..., P_N\}$ without detracting from the performance on P_1 and P_2. Obviously the class of interest is the set of all popular benchmark problems.

It is due to the diversity of the benchmark set that the "promise" is rarely realized. Boosting performance for one subset of the problems usually detracts from performance for the complement. In any case, the notion that the best algorithm is one that works well for a wide range of problems is highly dubious, in light of "no free lunch." The standard X-with-Y studies have always been subject to criticism, if only because of their exclusion of bad results. There is now a strong basis for saying that they are totally illegitimate.

5.2 Extensive Benchmark Studies

There is a class of studies of an entirely different caliber, which compares algorithm variants or different algorithms on more than a few benchmark problems. A major justification for such studies is that they assess algorithms with a diverse collection of problems, and that if one algorithm does better than another

on a wide range of problems, it is genuinely a better algorithm. Algorithms that do well on only two or three problems are generally considered to be tuned to those problems.

It is not so much that the rationale for extensive benchmark studies is wrong as it is that the objective of finding a generally better algorithm does not appear to be well founded. It is much as though the community is insisting that tools be Swiss army knives instead of hammers and screwdrivers.

Revision of objectives is strongly indicated by link of performance to ~~information of the prior distribution~~ [*sampling bias*]. "How good is the optimizer?" is more appropriately "What does the optimizer do?" The preoccupation with the best optimizer should shift to an interest in finding the right optimizer for the job. The benchmark problems would profitably be replaced by a collection of diagnostic distributions. That is, the distributions would be designed to provide information as to how an optimizer works.

Researchers who select distributions, sample them, and give a full characterization of the results of trials provide information that can be used in many ways.

5.3 Applications

Studies of applications have the advantage that distributions are given. The main concern is to obtain a random and representative sample, as it always has been. There are no apparent ramifications of "no free lunch" for these studies.

It is extremely interesting, however, that in most applications the basic algorithm is tuned to fit the problem domain. For some applications, the algorithm fails miserably prior to modification. This is not a rare event, and it is just what one would expect on the basis of the "no free lunch" arguments.

6 Conclusion

[*The following discussion of the information of tools is correct, but only because it shifts to a different sense of* information. *When a toolmaker imparts form to matter, the resulting tool is literally in-formed to suit a task. This kind of information is not prior information. Having form is different from observing beforehand. An optimization practitioner may gain information of a problem by observation, and then form a sampler to serve as a proxy in solving it. Although the sampler is informed to decide as the practitioner would, it cannot gain information of the problem as the practitioner does, and thus cannot justify decisions. Its sampling bias is form imparted by the practitioner.*]

Hammers contain information about the distribution of nail-driving problems. Screwdrivers contain information about the distribution of screw-driving problems. Swiss army knives contain information about a broad distribution of survival problems. Hammers and screwdrivers do their own jobs very well, but they do each others jobs very poorly. Swiss army knives do many jobs, but none par-

ticularly well. When the many jobs must be done under primitive conditions, however, Swiss army knives are ideal.

The tool literally carries information about the task. Furthermore, optimizers are literally tools — an algorithm implemented by a computing device is a physical entity. In empirical study of optimizers, the objective is to determine the task from the information in the tool. The problem of the EC researcher is similar to that of an anthropologist trying to explain excavated artifacts. EC researchers make and bury the tools before digging them up and trying to explain them, however. This anomaly derives from the fact that the algorithms are biologically inspired, but poorly understood.

Do the arguments of this paper contradict the evidence of remarkable adaptive mechanisms in biota? The question is meaningful only if one regards evolutionary adaptation as function optimization. Unfortunately, that model has not been validated. It is well known that biota are components of complex, dynamical ecosystems. Adaptive forces can change rapidly and nonlinearly, due in part to the fact that evolutionary adaptation is itself ecological change. In terms of function optimization, evaluation of points changes the fitness function. The Conservation Lemma clearly does not apply to such a process.

Acknowledgments

The novel results of this paper derive from questions posed by a reviewer. D. Wolpert disabused the author of the notion that the uniform distribution is perversely difficult.

References

T. Bäck and H.-P. Schwefel. An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1(1):1–23, 1993.

J. L. Breeden. An EP/GA synthesis for optimal state space representations. In A. V. Sebald and L. J. Fogel, editors, *Proceedings of the Third Annual Conference on Evolutionary Programming*, pages 216–223. World Scientific, River Edge, NJ, 1994.

D. B. Fogel. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. IEEE Press, Piscataway, NJ, 1995.

O. Häggström. Intelligent design and the NFL theorems. *Biology & Philosophy*, 22(2):217–230, 2007.

C. Igel and M. Toussaint. A no-free-lunch theorem for non-uniform distributions of target functions. *Journal of Mathematical Modelling and Algorithms*, 3(4): 313–322, 2005.

M. J. Streeter. Two broad classes of functions for which a no free lunch result does not hold. In *Genetic and Evolutionary Computation – GECCO 2003*, pages 1418–1430. Springer, 2003.

D. H. Wolpert and W. G. Macready. No free lunch theorems for search. Technical report, SFI-TR-95-02-010, Santa Fe Institute, 1995.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.