In Proceedings of the 2013 IEEE Congress on Evolutionary Computation (CEC-2013)
2013. IEEE Press.

Measure-Theoretic Analysis of Performance in Evolutionary Algorithms

Alan J Lockett
IDSIA
Galleria 2
6928 Manno-Lugano
Switzerland
alan.lockett@gmail.com

***Abstract*—The performance of evolutionary algorithms has been studied extensively, but it has been difficult to answer many basic theoretical questions using the existing theoretical frameworks and approaches. In this paper, the performance of evolutionary algorithms is studied from a measure-theoretic point of view, and a framework is offered that can address some difficult theoretical questions in an abstract and general setting. It is proven that the performance of continuous optimizers is in general nonlinear and continuous for finitely determined performance criteria. Since most common optimizers are continuous, it follows that in general there is substantial reason to expect that mixtures of optimization algorithms can outperform pure algorithms on many if not most problems. The methodology demonstrated in this paper rigorously connects performance analysis of evolutionary algorithms and other optimization methods to functional analysis, which is expected to enable new and important theoretical results by leveraging prior work in these fields.**

I. Introduction and Motivation

The performance of evolutionary algorithms has been studied extensively, but it has been difficult to answer many basic theoretical questions using the existing theoretical frameworks and approaches. This study extends the measure-theoretic framework of Lockett and Miikkulainen [6] to account for the analysis of optimizer performance. This treatment provides a bridge between the mathematical analysis of evolutionary algorithms and the well-studied theory of random variables and stochastic processes. Using this framework, the tools and techniques of stochastic analysis can be used directly on evolutionary algorithms to draw broad conclusions in a general setting. Previously, such techniques have been used to analyze the asymptotic convergence of genetic algorithms using Markov chains [8]. The No Free Lunch theorems for optimization also proceed along similar lines [12, 2]. Since such seminal results in the theory of evolutionary algorithms have been achieved using similar principles, it is hoped that a rigorous application of stochastic analysis to the performance of evolutionary algorithms will also yield substantial new insights.

This paper will introduce the *optimization process*, a stochastic process representing the path of an evolutionary algorithm through the search domain. A performance criterion will be defined as the expectation of a random variable depending on the optimization process and the fitness function. It will be shown how some common performance metrics such as the final error, success probability, and runtime are realized in this framework. Then a theorem is presented stating that if the performance criterion depends only on finitely many steps of the algorithm, then the average performance of the algorithm will vary continuously as either the algorithm or the fitness is varied.

This final result has direct practical consequences. As Lockett and Miikkulainen showed, many sets of evolutionary algorithms are convex subsets of normed vector spaces, and hence there is a line or spectrum of evolutionary algorithms between any two fixed algorithms. This paper shows that performance changes continuously and nonlinearly along this line, leaving open the possibility that interior points on this line in algorithm space might represent better-performing evolutionary algorithms than the endpoints. In fact, For example, Lehre and Özcan [5] have already exhibited instances in which this conjecture is borne out. They constructed pairs of algorithms such that one of their convex combinations (a *mixed strategy* in their terminology) has better average runtime than either of the original algorithms, a possibility predicted by this theory. It is not yet known in general whether interior points in the convex span of state-of-the-art evolutionary algorithms perform better than the original algorithms. The mathematical analysis of performance explored in this paper suggests that there should be many problems for which these interior algorithms are best.

II. Review of Optimizer Analysis

This analysis of performance is built on the analysis of iterative stochastic optimizers put forward by Lockett and Miikkulainen [6]. In this section, the basic ideas are reviewed, and the concept of the optimization process is introduced to represent the infinite trajectory through the search space followed by an evolutionary algorithm.

A. Formal Setting

The search domain is assumed to be a Hausdorff topological space (X, τ), and an optimization algorithm samples each point from a probability distribution on the Borel-measurable space $(X, \mathcal{B}_\tau)$, where $\mathcal{B}_\tau$ is the Borel σ-algebra, a technical tool specifying which subsets of the search domain X may be assigned a probability, including at a minimum the open and closed sets. In general, binary spaces, Euclidean spaces, and neural network spaces can be represented within this framework, and the more general expression makes it possible to draw conclusions about a wider range of algorithms.

A formal optimizer may be defined as follows. Suppose the fitness function is a static real function $f \in \mathbb{R}^X$. Let $\mathcal{P} = \mathcal{P}[X, \mathcal{B}_\tau]$ be the set of probability measures on $(X, \mathcal{B}_\tau)$. Let $\mathcal{T} = \mathcal{T}[X]$ be the set of finite sequences in X of arbitrary length, i.e. $\mathcal{T} = \bigcup_{n=0}^{\infty} X^n$. Then $\mathbb{P} \in \mathcal{P}$ represents the probability distribution used to generate new populations (that is, the randomized genetic operators), and $\mathcal{T}$ represents the sequence of solutions (organisms) that can be proposed by an algorithm. At each time step, an optimizer takes a history of previous solutions and a fitness function and returns a probability distribution to generate the next population. A *one-step optimizer* is defined as a functional $\mathcal{G} : \mathcal{T} \times \mathbb{R}^X \to \mathcal{P}$, that is, given a prior history $t \in \mathcal{T}$ and a fitness $f \in \mathbb{R}^X$, the optimizer samples $\mathcal{G}[t, f]$ to create the next proposed solution. The set of all one-step optimizers is written as $\mathcal{PF}$ (for probability-valued functionals).

A one-step optimizer is allowed to depend on the fitness function arbitrarily. The definition can be restricted by requiring an optimizer to depend only on the trajectory of fitness evaluations over the population history. Thus if $\mathcal{G}[t, f] = \mathcal{G}[t, g]$ whenever $f(x) = g(x)$ for all $x \in t$, $\mathcal{G}$ is said to be *trajectory-restricted*. The trajectory-restricted optimizers are the most commonly studied optimizers in evolutionary computation; for example, most No Free Lunch results pertain to this class [12, 2, 7]. If $\mathcal{G}$ is further allowed to depend only on the last population and its fitness values, then $\mathcal{G}$ is *population-Markov*; this set of objects coincides with Vose's work on Randomized Search Heuristics [10, 9]. Note that important optimization techniques such as Newton methods can be described by one-step optimizers, but are not necessarily trajectory-restricted, since they can depend on features such as the gradient of the fitness function.

One-step optimizers form a closed convex subset of a normed vector space [6], which means that the set of one-step optimizers lies inside a continuous space and thus it is possible to smoothly blend any two one-step optimizers. In particular, for any $\mathcal{G}_1, \mathcal{G}_2$ and any $\alpha \in [0, 1]$, one may define a line of optimizers by $\mathcal{G}_\alpha = \alpha\mathcal{G}_1 + (1 - \alpha)\mathcal{G}_2$, where addition and multiplication are taken pointwise. In practical terms, one may sample from $\mathcal{G}_\alpha$ by flipping a coin with bias α to choose either $\mathcal{G}_1$ or $\mathcal{G}_2$, and then sampling from the selected optimizer. A general convex combination is given by $\mathcal{G}_\alpha = \sum_{i=1}^{N} \alpha_i \mathcal{G}_i$, where α is a vector whose components sum to one, i.e., $\sum_{i=1}^{N} \alpha_i = 1$. In this case, the probability distributions produced by $\mathcal{G}$ are mixture distributions over the set of probability distributions produced by $\{\mathcal{G}_i\}_{i=1}^{N}$.

In order to study performance, one must examine the points an optimizer proposes over many steps. The next section examines the long-running properties of optimizers.

B. The Optimization Process

When a stochastic optimizer $\mathcal{G}$ is run on a particular objective f, it is initialized with the empty trajectory $\emptyset$, and $\mathcal{G}[\emptyset, f]$ is sampled to obtain a random evaluation point Z_1. This point is added to the trajectory, and $\mathcal{G}[(Z_1), f]$ is sampled to get Z_2. The process continues iteratively, so that $Z_{n+1} \sim \mathcal{G}[(Z_m)_{m=1}^{n}, f]$ for each n. In this way, an infinite random process $Z = (Z_n)_{n=1}^{\infty}$ can be generated. The finite-dimensional distributions of this process are generated by iterative applications of $\mathcal{G}$, and so the Kolmogorov Extension Theorem [4] implies that the process is well defined. The process generated in this way is termed the *optimization process* of an optimizer $\mathcal{G}$ on an objective f.

The optimization process is generated by a probability measure over infinite sequences in X; this measure depends on the fitness function f. The space of infinite sequences is $X^{\mathbb{N}}$; denote by $\mathcal{B}[X^{\mathbb{N}}]$ the σ-algebra extending $\mathcal{B}_\tau$ generated by the Kolmogorov Extension Theorem. Thus a *long-running optimizer* is a functional $\mathcal{G}_f : \mathbb{R}^X \to \mathcal{P}[X^{\mathbb{N}}, \mathcal{B}[X^{\mathbb{N}}]]$, where $\mathcal{P}[X^{\mathbb{N}}, \mathcal{B}[X^{\mathbb{N}}]]$ is the set of probability measures on $\left(X^{\mathbb{N}}, \mathcal{B}[X^{\mathbb{N}}]\right)$. The notation $\mathcal{G}_f$ indicates that the long-running optimizer is generated by a one-step optimizer $\mathcal{G}$ on a fitness function f, although the relationship is not one-to-one, since different one-step optimizers that differ on population histories of $\mathcal{G}_f$-measure zero generate the same long-running optimizer.

In the notation of [6], the marginal distribution of Z_n at any particular point in time can be stated succinctly as $Z_n \sim (\bigstar_{i=1}^{n} \mathcal{G})[\emptyset, f]$. Conditional on $(Z_m)_{m=1}^{n-1}$, it holds that $Z_n \mid Z_1, \ldots, Z_{n-1} \sim \mathcal{G}\left[(Z_m)_{m=1}^{n-1}, f\right]$.

Performance will be defined as the expected value of a functional of the optimization process Z. If g is a functional on $X^{\mathbb{N}}$, then $\mathbb{E}_{\mathcal{G}f}[g(Z)]$ is the expected value of the functional $g(Z)$ with respect to the probability measure $\mathcal{G}_f$.

In general, a property holds $\mathcal{G}_f$-almost surely ($\mathcal{G}_f$-$a.s.$) if there is some subset A of $X^{\mathbb{N}}$ such that $\mathcal{G}_f(A) = 1$ and the property holds on A. In what follows, $\mathcal{G}_f$ will sometimes be treated as though it were a measure over trajectories in $\mathcal{T}[X]$. Thus a set of trajectories $T \subseteq \mathcal{T}[X]$ corresponds to the set of sequences in $X^{\mathbb{N}}$ that infinitely expand any trajectory in T. The set $T \subseteq \mathcal{T}[X]$ is described as having $\mathcal{G}_f$-measure zero if the set of all sequences that infinitely expand it has $\mathcal{G}_f$-measure zero. Also, if a property holds for all trajectories except on a set of $\mathcal{G}_f$-measure zero, then this property is said to hold $\mathcal{G}_f$-$a.s.$

C. Continuity of One-Step Optimizers

Lockett and Miikkulainen [6] also studied the continuity of one-step optimizers. A one-step optimizer is a function of two arguments and can be continuous in either argument. Continuity answers two questions: (1) if an optimizer is altered slightly, does it still produce similar populations? and (2) if the fitness function is altered slightly, does an optimizer still produce similar populations? The specific meaning of the phrase "altered slightly" is determined by the particular choice of topology used for trajectories, fitness functions, and probability measures.

In this paper, the topology over trajectories is constructed from the given topology τ; the expression $t_n \to t$ will be used to indicate convergence for this topology, but the exact meaning of this formulation is left open. At a minimum, the expression $t_n \to t$ should imply that there is an N such that for $n > N$, $|t_n| = |t|$, and that for all $1 \leq i \leq |t|$, $t_n^i \to t^i$ in τ. Here $|t|$ indicates the length of a trajectory. In the case where τ is a metric topology with metric ρ, then a metric topology

for $\mathcal{T}[X]$ can be generated from the metric

$$d_\rho(t_1, t_2) = |\,|t_1| - |t_2|\,| + \sum_{i=1}^{|t_1| \wedge |t_2|} \rho\left(t_1^i, t_2^i\right). \quad (1)$$

The topology over fitness functions used here is the topology of pointwise convergence: $f_n \to f \equiv \forall x f_n(x) \to f(x)$. For probability measures, this paper uses the topology of the total variation norm, $||\mathbb{P} - \mathbb{Q}||_{\mathcal{M}} = \sup_{A \in \mathcal{B}_\tau} |\mathbb{P}(A) - \mathbb{Q}(A)|$. Other choices of topology are possible but not explored here.

A function between two topological spaces is continuous if the inverse image of every open set is an open set. For the sake of simplicity, the following definitions are used in this paper.

Definition 1: A one-step optimizer $\mathcal{G} \in \mathcal{MF}[X]$ is *continuous in objectives* at f if for any sequence of fitness functions $\{f_n\}$, $f_n \to f$ implies $||\mathcal{G}[t, f] - \mathcal{G}[t, f_n]||_{\mathcal{M}} \to 0$.

Definition 2: A one-step optimizer $\mathcal{G} \in \mathcal{MF}$ is *continuous in trajectories* at t if for any sequence of trajectories $\{t_n\} \subseteq \mathcal{T}[X]$, $t_n \to t$ implies $||\mathcal{G}[t_n, f] - \mathcal{G}[t, f]||_{\mathcal{M}} \to 0$.

A one-step optimizer is *continuous* at a pair t, f if it is continuous both in objectives and trajectories at t, f.

If a one-step optimizer $\mathcal{G}$ is continuous for every choice of t, f, it is *continuous everywhere*. If $\mathcal{G}$ is continuous in trajectories for a given function f on a set of trajectories $T \subseteq \mathcal{T}$ that has $\mathcal{G}_f$-probability one, then $\mathcal{G}$ is *$\mathcal{G}_f$-a.s continuous in trajectories at f*, or, more tersely, $\mathcal{G}$ is *continuous $\mathcal{G}_f$-a.s.*. Almost sure continuity is the building block for studying the continuity of performance, since it ignores discontinuities that caused by population histories that do not occur practically.

Most popular evolutionary algorithms are continuous in trajectories and objectives for most fitness functions and *trajectories of unambiguous fitness*. In general, continuity applies for fitness functions that are the limit of continuous functions. Trajectories of ambiguous fitness are trajectories in which the last population contains two or more distinct points that have the same fitness value. Trajectories of ambiguous fitness will often have $\mathcal{G}_f$-probability zero unless the fitness function has plateaus. For further discussion of when a one-step optimizer is continuous, see [6].

The basic concepts of continuity are used to assess the continuity of long-running optimizers and ultimately the continuity of performance criteria, which are introduced next.

III. Performance Criteria

This section introduces *performance criteria* that formalize common notions of what it means for an optimizer to perform well on an objective. These formalizations will make it possible to study performance as a general abstract concept.

A. Definition of Performance

A *performance criterion* is defined as the expected value of a positive function of the optimization process. Conceptually, performance is determined by a scoring function $h(z, f)$ that scores the value of a particular infinite trajectory z through the search domain on a particular fitness function f. The performance of an optimizer $\mathcal{G}$ is determined by averaging the score of each path z weighted according to the probability that $\mathcal{G}_f$ will follow that path.

Definition 3: Let $\mathcal{G} \in \mathcal{PF}$ and $f \in X^{\mathbb{R}}$, and let $Z = (Z_n)_{n \in \mathbb{N}}$ be the optimization process induced by $\mathcal{G}$ on f. Then a function $\phi : \mathcal{PF} \times \mathbb{R}^X \to [0, \infty)$ is a performance criterion if there exists a measurable and nonnegative function h with $h : X^{\mathbb{N}} \times \mathbb{R}^X \to [0, \infty)$ such that

$$\phi(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}[h(Z, f)] = \int_{X^{\mathbb{N}}} h(z, f)\, \mathcal{G}_f(dz) \quad (2)$$

whenever the integrals exist; h is called the kernel of ϕ.

Performance criteria can be used to compare optimizers to each other, and to analyze how the performance of an optimizer varies as the objective changes.

The next subsection defines classes of possible performance criteria that correspond broadly to the kinds of results reported in the experimental literature on optimizers. These examples are given in four groups: (1) evaluation by average error, (2) hitting times for an error bound, (3) probability of attaining an error bound, and (4) error at a stopping time.

B. Evaluation by Average Error

A first approach to evaluating optimizers is to average the magnitude of the errors the optimizer makes at each time step. This metric combines the total accuracy along with the speed of convergence, at the risk of disproportionately penalizing optimizers for early errors due to exploration of the objective. Such a metric is not traditionally reported, but could prove useful, since it contains information about the convergence speed of the optimizer.

Let $f \in \mathbb{R}^X$, $\mathcal{G} \in \mathcal{PF}$, and let $Z = (Z_n)$ be the optimization process generated by $\mathcal{G}$ on f. Define a performance criterion by

$$\phi_w(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}\left[\sum_{n=1}^{\infty} w_n\, |f(Z_n^*) - f^*|\right], \quad (3)$$

where $w = (w_n)$ is a sequence of weights that can be used to discount later values and $f^* = \inf_{x \in X} f(x)$ is the minimum of f. Three basic choices for w_n are (1) $w_n = 1$, which treats all errors equally but only results in ϕ_w finite when $\mathcal{G}$ converges on f at a fast enough rate, (2) $w_n = 2^{-n}$, which places more weight on earlier errors but is finite whenever the objective is almost surely finite on $\mathcal{G}[\emptyset, f]$, and (3) $w_n = 1$ for $n \leq N$ for some fixed $N < \infty$ and zero otherwise, which considers only a finite number of steps. Another scheme might ignore initial errors up to a finite time, allowing optimizers to explore more broadly in earlier stages without penalty.

C. Evaluation by Hitting Time

In existing literature, when evaluating a proposed optimizer, the optimizer is often run on a benchmark set of problems for which the optima are known (see e.g. [1, 3]). A common performance criterion for ranking optimizers is to count the number of points that must be generated before obtaining a solution whose fitness is within a fixed error from

the globally optimal fitness. The theoretical study of this metric is referred to as *runtime analysis*.

For a fixed error $\epsilon \geq 0$, define the hitting time for ϵ as the first time when an evaluation point has global error less than ϵ, i.e. $\tau_\epsilon \equiv \min\{n : |f(Z_n) - f^*| \leq \epsilon\}$. Then define a performance criterion by

$$\psi_\epsilon(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}[\tau_\epsilon], \tag{4}$$

which is the average hitting time for ϵ over all runs of the algorithm $\mathcal{G}$ on the objective f.

This formula has a flaw for non-convergent optimizers. If $\mathcal{G}$ has a positive probability of failing to attain error less than ϵ, then $\psi_\epsilon = \infty$. Even if $\mathcal{G}$ succeeds quickly in 99.999% of trials, it will still hold that $\psi_\epsilon = \infty$. Additionally, from the standpoint of approximation, only finite computational time is available, and thus cases in which τ_ϵ is large cannot be distinguished computationally from cases in which it is infinite.

One alternative is to place a finite limit on the stopping time; that is, for $N < \infty$,

$$\psi_\epsilon^N(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}[\tau_\epsilon \wedge N], \tag{5}$$

where the notation $\tau_\epsilon \wedge N = \min\{\tau_\epsilon, N\}$ as usual. The criterion $\psi_\epsilon^N(\mathcal{G}, f)$ can be estimated reasonably by running $\mathcal{G}$ on f several times for at most N evaluations. This performance criterion also reflects a natural criterion for comparing optimizers; it measures the average number of steps the optimizer must be run before it produces a solution correct within error ϵ.

D. Evaluation by Success Probability

The hitting time tests how long it takes on average to attain an error threshold ϵ. However, it does not test how often the threshold is attained. Define the sets $T_\epsilon = \{\tau_\epsilon < \infty\}$ and $T_\epsilon^N = \{\tau_\epsilon < N\}$ to represent respectively the sequences that asymptotically attain a given error bound and those that attain it within a fixed number of evaluations. Then the *success probability* is the probability of attaining a bound asymptotically, and the *finite success probability* is the probability of attaining the bound within a finite time window [11]. Each of these are performance criteria given by

$$\sigma_\epsilon(\mathcal{G}, f) = \mathcal{G}_f(T_\epsilon), \quad \sigma_\epsilon^N(\mathcal{G}, f) = \mathcal{G}_f(T_\epsilon^N). \tag{6}$$

To see that σ_ϵ and σ_ϵ^N are performance criteria, recall that $\mathcal{G}_f(T_\epsilon) = \mathbb{E}_{\mathcal{G}f}[1_A(Z)]$ where 1_{T_ϵ} is the indicator set of T_ϵ, i.e. $1_{T_\epsilon}(z) = 1$ if $z \in T_\epsilon$ and is zero otherwise. The finite success probability is the preferred criterion of these two, since σ_ϵ^N can be estimated experimentally, whereas σ_ϵ cannot. Notice that σ_ϵ does not conform to the convention that lower performance values should be better and zero should be optimal. The convention is ignored here because the success probability has an intuitive meaning in its own right. In situations where the convention is important, the performance criterion $1 - \sigma_\epsilon$ can be used instead.

Given the finite success probability, it is of interest to know the average hitting time for sequences that attain the error bound. The average hitting time on successful trajectories is a performance criterion, given by

$$\hat{\psi}_\epsilon^N(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}[(\tau_\epsilon \wedge N) 1_{T_\epsilon}(Z)]. \tag{7}$$

On its own, this quantity is not useful, since it may be zero when the optimizer fails, i.e. when $\mathcal{G}_f(T_\epsilon^N) = 0$. However, the pair $\left(\hat{\psi}_\epsilon^N, \sigma_\epsilon^N\right)$ disambiguates this situation, and these two values can be reported together for completeness [1].

E. Evaluation by Error at a Stopping Time

Optimizers are often tested by running the algorithm for a fixed number of evaluations and then reporting the final error. As a generalization of this type of evaluation, suppose that an optimizer is run until some criterion is satisfied, not necessarily connected to the number of evaluations. As one example of why this generalization may be useful, suppose that rather than stopping after a fixed number of evaluations, one wishes to stop an optimizer after it uses up a fixed amount of resources, such as CPU cycles or calendar time. Such a criterion can be modeled as a stopping time, and the error magnitude at this stopping time is a performance criterion.

Let T be a stopping time equal to the generation in which this resource limit is first expended, and define a performance criterion by

$$\zeta_T(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}|f(Z_T^*) - f^*|, \tag{8}$$

so that ζ_T is the smallest error attained within the allocated resources, where Z_n^* is the running minimum on Z_n as above.

A substantial number of performance criteria have now been introduced. The next two sections discuss the mathematical properties of performance criteria, such as nonlinearity, decomposability, and continuity.

IV. Performance Properties

It is clear that a wide variety of performance criteria exists. These criteria can be analyzed in general according to their mathematical properties. This section examines two such properties that a performance criterion may possess: nonlinearity and progressive decomposability. The question of continuity in performance criteria is a larger topic and will be addressed separately in the next section.

A. Nonlinearity

All non-trivial performance criteria are nonlinear in both arguments. For a given objective function, the location and nature of the optima are nonlinear qualities. The location of the global optimum for $f+g$ bears no general relationship to the location of the optimum for g or f. Thus for any useful performance criterion ϕ, one expects that $\phi(\mathcal{G}, f+g) \neq \phi(\mathcal{G}, f) + \phi(\mathcal{G}, g)$ in general, ignoring trivial choices of ϕ, $\mathcal{G}$, f, and g.

Performance criteria are also nonlinear in optimizers as well. For an n-dimensional cylinder set A restricting the first n coordinates of Z, the probability that A contains Z for an optimizer $\mathcal{G} + \mathcal{H}$ is given by

$$(\mathcal{G}+\mathcal{H})_f(Z \in A) = \int_{A_1} \cdots \int_{A_n} \prod_{i=1}^{n} \mathcal{G}+\mathcal{H}\left[(Z_m)_{m=1}^{i-1}, f\right](dx_i). \tag{9}$$

It is thus clear that $(\mathcal{G}+\mathcal{H})_f \neq \mathcal{G}_f + \mathcal{H}_f$ except under special circumstances, due to the cross terms under the product. In general, $\phi(\mathcal{G}+\mathcal{H}, f) \neq \phi(\mathcal{G}, f) + \phi(\mathcal{H}, f)$.

The nonlinearity of most performance criteria has an important consequence: It opens the possibility that a convex combination over a bank of one-step optimizers may outperform any of the given optimizers. This topic is discussed below in Section V-C, where an example is also shown in Figure 1.

B. Progressive Decomposability

A progressively decomposable performance criterion can be broken down into an infinite sum of finitely determined random variables. A random variable $h(Z)$ of the optimization process is *finitely determined* if it depends only on a finite prefix of Z, i.e., $h(Z) = h\left((Z_n)_{n=1}^m\right)$. It is typically easier to reason about finitely determined random variables.

Definition 4: A performance criterion ϕ is progressively decomposable if there exists a sequence of functions $h_m : X^m \times \mathbb{R}^X \to \mathbb{R}$ such that

$$\phi\left(\mathcal{G}, f\right) = \sum_{m=1}^{\infty} \mathbb{E}_{\mathcal{G}f}\left[h_m\left((Z_n)_{n=1}^m, f\right)\right], \tag{10}$$

where $(Z_n)_{n=1}^m$ is the vector in $\mathbb{R}^m$ formed by taking the first m elements of the optimization process.

Perhaps surprisingly, it is simple to prove that every performance criterion is progressively decomposable by conditioning on the natural filtration of the optimization process.

Theorem 4.1: Every performance criterion as defined in Definition 3 is progressively decomposable.

Proof: Let $(\mathcal{Z}_n)$ be the natural filtration of the optimization process. Given $h(z, f)$, let $h_1(Z_1, f) = \mathbb{E}_{\mathcal{G}f}[h(Z, f) \mid \mathcal{Z}_1]$ and recursively define

$$h_m\left((Z_n)_{n=1}^m, f\right) = \mathbb{E}_{\mathcal{G}f}[h(Z, f) \mid \mathcal{Z}_m] - h_{m-1}\left((Z_n)_{n=1}^{m-1}, f\right) \tag{11}$$

Observe that $\mathbb{E}_{\mathcal{G}f}[h_1(Z_1, f)] = \phi(\mathcal{G}, f)$, and for $m > 1$, $\mathbb{E}_{\mathcal{G}f}[h_m((Z_n)_{n=1}^m, f)] = 0$. As a result,

$$\phi(\mathcal{G}, f) = \mathbb{E}_{\mathcal{G}f}[h(Z, f)] = \sum_{m=1}^{\infty} \mathbb{E}_{\mathcal{G}f}[h_m\left((Z_n)_{n=1}^m, f\right)], \tag{12}$$

which concludes the proof. ∎

The properties of nonlinearity and progressive decomposability are useful for reasoning in general terms about multiple performance metrics. But perhaps the most interesting property of performance for this paper is continuity, which is studied next.

V. Continuity of Performance

Continuous performance criteria are of interest for a variety of reasons. A continuous performance criterion must score an optimizer similarly on similar objective functions. If performance is continuous, then several other opportunities exist, some of which include: 1) the convex span of two or more optimizers may contain better performing optimizers than the endpoints; 2) optimizers that are difficult to compute may be approximated by simpler optimizers with similar performance; and 3) for a fitness function that is expensive to calculate, parameter settings for an algorithm may be reliably evaluated based on their performance on approximations of the expensive function. In this section, it is shown that finitely determined performance criteria are continuous on continuous optimizers.

A. Continuity of the Optimization Process

When does continuity of an optimizer imply that the optimization process generated by that optimizer is continuous? Specifically, suppose $f_n \to f$, and let $\mathcal{G} \in \mathcal{PF}$ be continuous in objectives. Does $\mathcal{G}_{f_n} \to \mathcal{G}_f$ in the total variation norm? Because the optimization process is infinite, it may be possible for $\mathcal{G}_{f_n}$ to diverge from $\mathcal{G}_f$ even if $\mathcal{G}$ is continuous everywhere. Thus it is not possible to extend continuity from the one-step optimizer to arbitrary performance criteria

It is possible to prove that the continuity of a one-step optimizer on sufficiently many trajectories implies that the long-running optimizer yields similar average values for finitely-determined random variables of the optimization process: It will be shown that for any optimizer $\mathcal{G}$ that is continuous $\mathcal{G}_f$-$a.s.$, and for any finitely determined random variable Y,

$$\mathbb{E}_{\mathcal{G}f_n}[Y(Z)] \to \mathbb{E}_{\mathcal{G}f}[Y(Z)]. \tag{13}$$

The condition of finiteness is needed because the infinitesimal differences between $\mathcal{G}_{f_n}$ and $\mathcal{G}_f$ can cause divergence of the integral after infinitely many time steps.

The space of random variables on $\left(X^{\mathbb{N}}, \mathcal{B}[X^{\mathbb{N}}]\right)$ is the set of functionals on $Y : X^{\mathbb{N}} \to \mathbb{R}$ whose backward projections are $\mathcal{B}[X^{\mathbb{N}}]$-measurable, that is, $Y^{-1}(A) \in \mathcal{B}[X^{\mathbb{N}}]$ for every A in the Borel σ-algebra on $\mathbb{R}$. These random variables may be written either in lower case as $y(Z)$ or in upper case as $Y(Z)$. If written in upper case, the argument may be omitted, e.g. $Y = Y(Z)$.

If $Y(Z)$ is a random variable of this sort, then $\mathbb{E}_{\mathcal{G}f}\left[Y(Z)\right]$ integrates over X countably many times. But if Y is finitely determined, it depends on only finitely many components in $X^{\mathbb{N}}$. The remaining (infinitely many) steps can be integrated out. If x_1^m is the trajectory formed by taking the first m components of $x \in X^{\mathbb{N}}$ and $Y(x) = \hat{Y}(x_1, \ldots, x_m)$, m integrals are required, since

$$\begin{aligned}\mathbb{E}_{\mathcal{G}f}\left[Y\left(Z\right)\right] &= \int_{X^{\mathbb{N}}} \hat{Y}(z_1, \ldots, z_m) \prod_{k=1}^{\infty} \mathcal{G}\left[z_1^{k-1}, f\right](dz_k) \\ &= \int_{X^m} \hat{Y}(z_1, \ldots, z_m) \prod_{k=1}^{m} \mathcal{G}\left[z_1^{k-1}, f\right](dz_k). \end{aligned} \tag{14}$$

Along any particular trajectory t, the optimization processes of $\mathcal{G}_f$ and $\mathcal{G}_{f_n}$ cannot move far apart when $\mathcal{G}$ is continuous in objectives on the trajectory t. If $\mathbb{E}_{\mathcal{G}f}\left[Y(Z)\right]$ depends on finitely many optimization steps, then for large n, $\mathbb{E}_{\mathcal{G}f}\left[Y(Z)\right]$ must be close to $\mathbb{E}_{\mathcal{G}f_n}\left[Y(Z)\right]$ as well when $\mathcal{G}$ is continuous in objectives at f with $\mathcal{G}_f$-probability one. This claim is made rigorous with the following theorem.

Theorem 5.1: Let $\mathcal{G} \in \mathcal{PF}$ be continuous $\mathcal{G}_f$-$a.s.$ at an objective f, and let $f_n \to f$ pointwise. Let $g\left(x_1, \ldots, x_m\right)$ be a real function on X^m with $m < \infty$ fixed, and suppose that $\mathbb{E}_{\mathcal{G}f}\left|g\left(Z_1, \ldots, Z_m\right)\right| < \infty$ and $\mathbb{E}_{\mathcal{G}f_n}\left|g\left(Z_1, \ldots, Z_m\right)\right| < \infty$. Then $\mathbb{E}_{\mathcal{G}f_n}\left[g\left(Z_1, \ldots, Z_m\right)\right] \to \mathbb{E}_{\mathcal{G}f}\left[g\left(Z_1, \ldots, Z_m\right)\right]$.

Proof: Fix $\epsilon > 0$. Assume $||\mathcal{G}[t, f]||_{\mathcal{M}} \leq M < \infty$. Suppose J and L are two index sets of positive integers less than or equal to m. J and L will be termed *complementary* if $J \cap L = \emptyset$ and $J \cup L = \{1, \ldots, m\}$. Let $\mathcal{K}$ be the set of all complementary pairs of index sets. There are exactly 2^m such

pairs. These complementary sets can be used to state the joint distribution of $Z_1, \ldots, Z_m$ as a sum.

Let T be a set of trajectories such that $\mathcal{G}_f(T) = 1$ and $\mathcal{G}$ is continuous at t, f for all $t \in T$.

Let $t \in T$ be of length at least $m < \infty$. Let t_1^m be the trajectory formed by the first m components of t. Then

$$\begin{aligned}\mathbb{P}\left((Z_n)_{n=1}^m \in dt_1^m\right) &= \prod_{k=1}^m \mathcal{G}\left[t_1^{k-1}, f\right]\left(dt^k\right) \\ &= \prod_{k=1}^m \left[\left(\mathcal{G}[t_1^{k-1}, f](dt^k) - \mathcal{G}[t_1^{k-1}, f_n](dt^k)\right)\right. \\ &\qquad \left. + \mathcal{G}[t_1^{k-1}, f_n](dt^k)\right] \qquad (15) \\ &= \sum_{J,L\in\mathcal{K}} \left[\prod_{j\in J}\left(\mathcal{G}[t_1^{j-1}, f](dt^j) - \mathcal{G}[t_1^{j-1}, f_n](dt^j)\right)\right. \\ &\qquad \left. \times \prod_{\ell\in L} \mathcal{G}[t_1^{\ell-1}, f_n](dt^\ell)\right]. \qquad (16)\end{aligned}$$

Equation 16 expands the product in Equation 15 by cross multiplying the difference with the joint distribution over f_n. This sum contains 2^m terms, one for each pair of complementary index sets. With the exception of the complementary sets given by $J_0 = \emptyset, L_0 = \{1, \ldots, m\}$, every pair of complementary index sets in $\mathcal{K}$ yields a product in Equations 16 with at least one factor of the form

$$\mathcal{G}\left[t_1^{j-1}, f\right]\left(dt^j\right) - \mathcal{G}\left[t_1^{j-1}, f_n\right]\left(dt^j\right).$$

Because m is finite and $\mathcal{G}$ is continuous in objectives, n can be chosen so that $\left|\mathcal{G}[t_1^{j-1}, f] - \mathcal{G}[t_1^{j-1}, f_n]\right| < \frac{\epsilon}{2^{2m}}$ for each j. Thus each term in the sum except for the one at J_0, L_0 is less than $\frac{\epsilon}{2^m}$, since $|\mathcal{G}[t_1^{j-1}, f] - \mathcal{G}[t_1^{j-1}, f_n]| < 2$. Further, the term in the sum at J_0, L_0 reduces to

$$\prod_{k=1}^m \mathcal{G}\left[t_1^{k-1}, f_n\right]\left(dt^k\right),$$

and therefore for $A \in \mathcal{B}_{\tau^m}$,

$$\begin{aligned}&\int_A \left|\prod_{k=1}^m \mathcal{G}\left[t_1^{k-1}, f\right]\left(dt^k\right) - \prod_{k=1}^m \mathcal{G}\left[t_1^{k-1}, f_n\right]\left(dt^k\right)\right| \\ &\leq \sum_{J,L\in\mathcal{K}\setminus\{J_0,L_0\}} \int_A \prod_{j\in J}\left|\mathcal{G}\left[t_1^{j-1}, f\right]\left(dt^j\right) - \mathcal{G}\left[t_1^{j-1}, f_n\right]\left(dt^j\right)\right| \\ &< 2^m \frac{\epsilon}{2^{2m}} 2^m = \epsilon.\end{aligned}$$

Because of the integrability assumptions on g, it follows that

$$\left|\mathbb{E}_{\mathcal{G}f_n}\left[g\left(Z_1, \ldots, Z_m\right)\right] - \mathbb{E}_{\mathcal{G}f}\left[g\left(Z_1, \ldots, Z_m\right)\right]\right| \to 0, \quad (18)$$

which concludes the proof ∎

Corollary 5.2: Under the same general assumptions as Theorem 5.1, let A be a set in $\mathcal{B}[X^{\mathbb{N}}]$ such that for fixed $m < \infty$, A is independent of Z_n for $n > m$ under $\mathcal{G}_f$ and $\mathcal{G}_{f_n}$. Then $\mathcal{G}_{f_n}(A) \to \mathcal{G}_f(A)$.

Proof: Note that $\mathcal{G}_f(A) = \mathbb{E}_{\mathcal{G}f}[1_A]$. Define $g(Z_1, \ldots, Z_m) = \mathbb{E}_{\mathcal{G}f}[1_A \mid Z_1, \ldots, Z_m]$. Because A is independent of Z_n for $n > m$, $g(Z_1, \ldots, Z_m) = 1_A(Z)$ by the definition of conditional expectations. The result follows directly from Theorem 5.1. ∎

If the fitness function is held constant, but the optimizer is altered slightly, a similar theorem holds without continuity assumptions. Integrals over finitely determined random variables change continuously with the optimizer, regardless of whether the optimizer is continuous. The next theorem shows that the average value of a functional under $\mathcal{G}_n f$ converges to its average value under $\mathcal{G}f$, again if the functional depends on finitely many steps of the optimization process. This result will be used to demonstrate that performance criteria are continuous over optimizers.

Theorem 5.3: Let $\mathcal{G} \in \mathcal{PF}$, and let $f \in \mathbb{R}^X$. Let $\mathcal{G}_n \to \mathcal{G}$. Let $g(x_1, \ldots, x_m)$ be a real function with $m < \infty$ fixed, and suppose that both $\mathbb{E}_{\mathcal{G}f}|g(Z_1, \ldots, Z_m)| < \infty$ and also $\mathbb{E}_{\mathcal{G}_n f}|g(Z_1, \ldots, Z_m)| < \infty$. Then it follows from these facts that $\mathbb{E}_{\mathcal{G}_n f}[g(Z_1, \ldots, Z_m)] \to \mathbb{E}_{\mathcal{G}f}[g(Z_1, \ldots, Z_m)]$.

Proof: Repeat the proof of Theorem 5.1, but replacing $\mathcal{G}[t_1^{k-1}, f_n]$ by $\mathcal{G}_n[t_1^{k-1}, f]$; note continuity is not needed. ∎

Theorem 5.1 and 5.3 are sufficient to prove the continuity of performance criteria on continuous optimizers.

B. Performance Continuity in Objectives

Continuity in objectives is a strong requirement, and it will not be possible to achieve it for all cases. In this section, something slightly weaker will be proven. Given any sequence f_n such that $f_n \to f$ uniformly,[1] it will be shown that $\phi(\mathcal{G}, f_n) \to \phi(\mathcal{G}, f)$ if $\mathcal{G}$ is continuous $\mathcal{G}_f$-$a.s.$ The following general theorem proves that $\phi(\mathcal{G}, f_n) \to \phi(\mathcal{G}, f)$ when the kernel of ϕ is finitely determined and converges in expectation under $\mathcal{G}_f$ as $f_n \to f$. Additionally, it is required that $\phi < \infty$ on f_n and f. It will then be shown that this type of convergence follows for the performance criteria in Section III when $f_n \to f$ uniformly.

Theorem 5.4: Suppose ϕ is a performance criterion and $\mathcal{G} \in \mathcal{PF}$ is continuous $\mathcal{G}_f$-$a.s.$ in objectives. Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of functions converging pointwise to f. Suppose additionally that the kernel h of ϕ is finitely determined by the first m steps of the optimization process and has the property that $\mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f_n)] \to \mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f)]$. Then if $\phi(\mathcal{G}, f) < \infty$ and $\phi(\mathcal{G}, f_n) < \infty$, $\phi(\mathcal{G}, f_n) \to \phi(\mathcal{G}, f)$.

Proof: First suppose $\phi(\mathcal{G}, f) < \infty$ and $\phi(\mathcal{G}, f_n) < \infty$ for all n. Fix $\epsilon > 0$. Let $f_n \to f$. Suppose without loss of generality that $f_n^* = f^* = 0$, since otherwise $f - f^*$ and $f_n - f_n^*$ will satisfy this equality. The conditions for Theorem 5.1 are met, and so

$$\begin{aligned}&|\phi(\mathcal{G}, f) - \phi(\mathcal{G}, f_n)| \\ &\quad = |\mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f)] - \mathbb{E}_{\mathcal{G}f_n}[h((Z_k)_{k=1}^m, f_n)]| \\ &\quad \leq |\mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f_n)] - \mathbb{E}_{\mathcal{G}f_n}[h((Z_k)_{k=1}^m, f_n)]| \\ &\qquad + \mathbb{E}_{\mathcal{G}f}|h((Z_k)_{k=1}^m, f) - h((Z_k)_{k=1}^m, f_n)| \\ &\quad < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \qquad (19)\end{aligned}$$

[1] That is, for any $\epsilon > 0$ there is an N such that $|f_n(x) - f(x)| < \epsilon$ for $n > N$, and N does not depend on x.

In these equations, the left term is less that $\frac{\epsilon}{2}$ by an application of Theorem 5.1, and the right term by the assumptions on h. ■

The following corollaries apply this theorem to the classes of performance criteria defined in Section III, beginning with the average error ϕ_w.

Corollary 5.5: If $f_n \to f$ uniformly and $\mathcal{G}$ is continuous $\mathcal{G}_f$-$a.s.$ in objectives, then $\phi_w(\mathcal{G}, f_n) \to \phi_w(\mathcal{G}, f)$ whenever $\phi_w(\mathcal{G}, f_n) < \infty$, $\phi_w(\mathcal{G}, f) < \infty$, and there exists m such that $w_n = 0$ for all $n > m$.

Proof: Suppose without loss of generality that $f^* = 0$ and $f_n^* = 0$. The result will hold if $\mathbb{E}_{\mathcal{G}f}[h((Z_k)_{n=1}^m, f_n)] \to \mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f)]$. For ϕ_w, $h(z, f) = \sum_{k=1}^m w_k f(z_k^*)$ under the assumptions. Because $f_n \to f$ uniformly, it follows that $h(z, f_n) \to h(z, f)$ uniformly, which proves that $\mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f_n)] \to \mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f)]$. The desired result follows from Theorem 5.4. ■

The functional ζ_T is continuous under the same conditions, provided that the stopping time T is almost surely finite and does not introduce discontinuities.

Corollary 5.6: Suppose $f_n \to f$ uniformly and $\mathcal{G}$ is continuous $\mathcal{G}_f$-$a.s.$ in objectives. Let $T = T_f(z)$ be a stopping time s.t. for some $m < \infty$, with $\mathcal{G}_f$- and $\mathcal{G}_{f_n}$-probability one, it holds that $T_{f_n} \leq m$, $T_f \leq m$, and $T_{f_n} \to T_f$ uniformly. Then $\zeta_T(\mathcal{G}, f_n) < \infty$ and $\zeta_T(\mathcal{G}, f) < \infty$ imply $\zeta_T(\mathcal{G}, f_n) \to \zeta_T(\mathcal{G}, f)$.

Proof: For ζ_T, $h(z, f) = \sum_{k=1}^m f(z_k^*) 1_{\{t:T_f(t)=k\}}(z)$. Because the stopping times are discrete, there is an N independent of z such that $T_{f_n}(z) = T_f(z)$ $\mathcal{G}_f$-$a.s.$ for all $n > N$. Because $f_n \to f$ uniformly, $h(z, f_n) \to h(z, f)$ uniformly, and therefore $\mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f_n)] \to \mathbb{E}_{\mathcal{G}f}[h((Z_k)_{k=1}^m, f)]$. The result follows by applying Theorem 5.4. ■

Corollary 5.6 begs the question of when T varies uniformly with the fitness f. One simple answer is that any stopping time that is independent of the fitness function will have this property.

The performance criteria ψ_ϵ, ψ_ϵ^N, σ_ϵ, and σ_ϵ^N require more stringent criteria in order to prove convergence, because there exist sequences of objectives $f_n \to f$ such that $f_n - f_n^* > \epsilon$ while $f - f^* = \epsilon$. As a simple example of discontinuity, let $f_n(x) = f(x) = 0$ on $(0, 1)$, and let $f_n(x) = \epsilon + n^{-1}$ and $f(x) = \epsilon$ on $[1, 2)$. Let $\mathcal{G}$ be uniform over $(0, 2)$. Then $f_n \to f$ uniformly, but $\psi_\epsilon(\mathcal{G}, f) = 1$ and $\psi_\epsilon(\mathcal{G}, f_n) = \sum_{n=1}^\infty n 2^{-n} = 2$. The discontinuity is caused by objectives with plateaus located at a distance of precisely ϵ away from the optimum. This problem does not arise if the trajectories with error ϵ have $\mathcal{G}_f$ measure zero. The next corollary applies to ψ_ϵ^N and σ_ϵ^N in general, but only apply to ψ_ϵ and σ_ϵ when they are finitely determined.

Corollary 5.7: Suppose $f_n \to f$ uniformly, and let $\mathcal{G}$ be an optimizer that is continuous $\mathcal{G}_f$-$a.s.$ Suppose the set

$$Z_\epsilon = \{z \in X^{\mathbb{N}} : |f(z_m) - f^*| = \epsilon \text{ for some m}\}$$

has $\mathcal{G}_f$-measure zero. Then $\phi(\mathcal{G}, f_n) < \infty$ and $\phi(\mathcal{G}, f) < \infty$ imply $\phi(\mathcal{G}, f_n) \to \phi(\mathcal{G}, f)$ when ϕ is one of ψ_ϵ^N or σ_ϵ^N.

Proof: On the set $X^{\mathbb{N}} \setminus Z_\epsilon$, it is not possible to have $f(z_m^*) - f^* = \epsilon$. Thus $f_n(z_m^*) - f_n^*$ must eventually be on the same side of ϵ as $f(z_m^*) - f^*$. The kernel of ψ_ϵ^N is $h(z, f) = \sum_{k=1}^N 1_{(\epsilon,\infty)}(f(z_k^*) - f^*)$. On $X^{\mathbb{N}} \setminus Z_\epsilon$, $h(z, f_n) = h(z, f)$ for all $n > N$ with N independent of z. The kernel of σ_ϵ^N is $h(z, f) = \sum_{k=1}^N 1_{B_\epsilon^{f,k}}(z)$ with

$$B_\epsilon^{f,k} = \{x \in \mathbb{R}^m : |f(x_k) - f^*| \leq \epsilon \text{ and } |f(x_i) - f^*| > \epsilon \ \forall i < k\}.$$

Once again, $h_m(z, f_n) = h(z, f)$ for all $n > N$ on $X^{\mathbb{N}} \setminus Z_\epsilon$. Thus in either case, $\mathbb{E}_{\mathcal{G}f}[h_m(Z_1^m, f_n)] \to \mathbb{E}_{\mathcal{G}f}[h_m(Z_1^m, f)]$ because $\mathcal{G}_f(Z_\epsilon) = 0$, and the result follows from Theorem 5.4. ■

Thus we have conditions to determine when $\phi(\mathcal{G}, f_n) \to \phi(\mathcal{G}, f)$ in many cases for the specific performance criteria introduced in Section III.

C. Continuity in Optimizers

As a final result, performance criteria are continuous in optimizers whenever they are finite and finitely determined, without the complications that arose analyzing continuity in objectives. The following theorem is analogous to Theorem 5.4 but with much weaker assumptions.

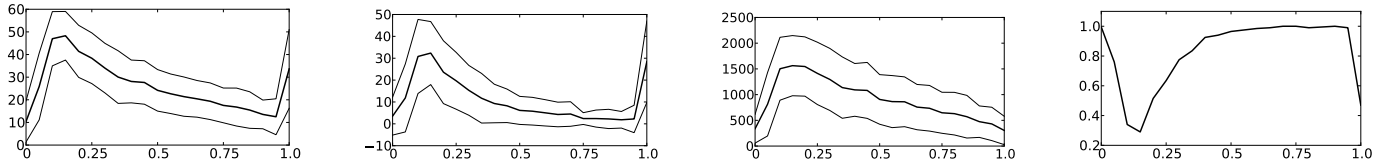
Theorem 5.8: Suppose $\mathcal{G}_n \to \mathcal{G}$ in the total variation norm. Then for any f and for any ϕ with a finitely determined kernel, $\phi(\mathcal{G}, f) < \infty$ and $\phi(\mathcal{G}_n, f) < \infty$ imply $\phi(\mathcal{G}_n, f) \to \phi(\mathcal{G}, f)$. That is, every finitely determined performance criterion ϕ is continuous over optimizers wherever ϕ is finite.

Proof: The result follows directly from Theorem 5.3. ■

Theorem 5.8 proves that for finite and finitely determined performance criteria, performance always changes smoothly as one moves from one optimizer to another along a line through $\mathcal{PF}$. Similar optimizers perform similarly on the same objective.

As discussed above, performance criteria are in general nonlinear. If one defines a line in optimizer space by $\mathcal{G}_\alpha = \alpha\mathcal{G}_1 + (1-\alpha)\mathcal{G}_2$ for two one-step optimizers $\mathcal{G}_1$ and $\mathcal{G}_2$, then the image of this line in performance space for fitness f is given by $\hat{\phi}(\alpha) = \phi(\mathcal{G}_\alpha, f)$. Due to nonlinearity, it is entirely possible that there is an α_0 such that $\hat{\phi}(\alpha_0) < \phi(\mathcal{G}_1, f)$ and $\hat{\phi}(\alpha_0) < \phi(\mathcal{G}_2, f)$. That is, the internal points of the line may outperform the endpoints. In fact, Lehre and Özcan have already exhibited choices of $\mathcal{G}_1$ and $\mathcal{G}_2$ for which this conjecture can be proven theoretically in the case of runtime analysis (ψ_ϵ). Thus the predictions of the theory are supported in some concrete cases.

To demonstrate the shape of $\hat{\phi}(\alpha)$ experimentally, the one-step optimizers for Differential Evolution (DE) and Particle Swarm Optimization (PSO) were convexly combined to generate a line in optimizer space as above. Figure 1 shows the performance of $\mathcal{G}_\alpha$ on Schwefel's function for various values of α. Once again, the change in performance is smooth but non-linear, as predicted by the theory. It can be seen that at $\alpha = 0.95$, $\mathcal{G}_\alpha$ outperforms both PSO and DE, although the result is statistically insignificant and is therefore merely suggestive. This observation suggests that the best performance on a particular fitness function may lie strictly inside the convex span of commonly used optimizers. Thus the theory suggests a new class of optimization methods that might be classified as *convex control*, for which improved performance may be possible.



(a) $\phi_1(\mathcal{G}_\alpha, f)$, $\alpha \in (0,1)$ (b) $\zeta_T(\mathcal{G}_\alpha, f)$, $\alpha \in (0,1)$ (c) $\hat{\psi}_\epsilon^N(\mathcal{G}_\alpha, f)$, $\alpha \in (0,1)$ (d) $\sigma_\epsilon^N(\mathcal{G}_\alpha, f)$, $\alpha \in (0,1)$

Fig. 1. Change in performance as the optimizer changes smoothly from PSO with $\omega = -.5, \phi_g = \phi_p = 2$ ($\alpha = 0$) to DE with CR=.2, F=.2 ($\alpha = 1$). The x-axis ranges over values of α, the y-axis over performance values. The first standard deviation is plotted where possible. The panels show the performance criteria ϕ_1, ζ_T, $\hat{\psi}_\epsilon^N$, and σ_ϵ^N, respectively, with $\epsilon = 25$ for Schwefel's function. As predicted by the theory, performance on these optimizers changes smoothly and nonlinearly as a function of the optimizer. Interestingly, at $\alpha = .95$, $\mathcal{G}_\alpha$ outperforms PSO and DE on ζ_T, although the result is not statistically significant.

VI. Discussion and Future Work

This paper has presented a measure-theoretic analysis of performance for iterative stochastic optimizers, including evolutionary algorithms. A formal model of performance was introduced, and the idea of the optimization process was developed to support analysis of this model. It was shown that for a finite number of steps, continuity of one-step optimizers implies continuity of the optimization process. Further, if the performance criterion is finite and only depends on a finite number of steps, then performance varies continuously as either the optimizer or the fitness function is altered, although the fitness function must be altered uniformly. Several commonly used performance metrics were defined within the formal framework of this paper, and the continuity results were applied specifically to these metrics.

This research demonstrates that the principles of functional analysis can be applied rigorously to study evolutionary algorithms. Further, such an analysis raises interesting practical possibilities. One opportunity suggested by this analysis is the idea of convex control, where one searches for a convex combination of a bank of optimizers that performs best on a particular problem. The theoretical observations in this paper are also relevant to the problem of parameter selection for a parameterized family of optimization algorithms. In many cases, small changes of parameters cause correspondingly small changes in the optimizer; these results suggest that the change in performance due to the change will also be small.

In addition to varying the optimizer, one may also vary the fitness function. One useful idea is that an easily computed approximation to an expensive fitness function might be sufficient for parameter selection, since the performance change from the expensive fitness function is small due to continuity.

The applications of this research go beyond examination of properties such as continuity. This type of performance analysis can also be used to determine the exact nature of priors over fitness functions that can induce No Free Lunch theorems. Furthermore, averaging these performance criteria over a random fitness function results in a duality between optimizers and random fitness functions. Thus the results presented in this paper are just the beginning of what is possible using a measure-theoretic approach to performance analysis.

VII. Conclusion

This paper has presented a measure-theoretic framework for analyzing performance criteria for optimizers. Specific categories of performance criteria were presented corresponding to the experimental quantities that are commonly reported in the literature. These performance criteria were shown to be continuous under certain conditions, suggesting the idea of convex control as a practical way to improve performance on a problem. This way of thinking about performance offers a novel means for achieving theoretical results about performance as well as practical ideas for developing new optimizers. Such methods may be used in the future in order to discover exciting new possibilities in optimization.

References

[1] D. Ashlock. Taxonomic clustering of genetic algorithms using unique performance signatures. private communication, 2011.

[2] A. Auger and O. Teytaud. Continuous lunches are free! In *Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO-2007)*, New York, 2007. ACM Press.

[3] K. M. Bryden, D. A. Ashlock, S. Corns, and S. Willson. Graph-based evolutionary algorithms. *IEEE Transaction on Evolutionary Computation*, 10(5), 2006.

[4] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. 1933.

[5] P. K. Lehre and E. Özcan. A runtime analysis of simple hyper heuristics: To mix or not to mix operators. In *Proceedings of the 12th Workshop on Foundations of Genetic Algorithms (FOGA-2013)*. ACM, 2013.

[6] A. Lockett and R. Miikkulainen. A measure-theoretic analysis of stochastic optimization. In *Proceedings of the 12th Workshop on Foundations of Genetic Algorithms (FOGA-2013)*. ACM, 2013.

[7] J. E. Rowe, M. D. Vose, and A. H. Wright. Reinterpreting no free lunch. *Evolutionary Computation*, 17(1), 2009.

[8] G. Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE Transactions on Neural Networks*, 5(1), 1994.

[9] M. Vose. *The Simple Genetic Algorithm*. MIT Press, Cambridge, Massachusetts, 1999.

[10] M. D. Vose. Random heuristic search. *Theoretical Computer Science*, 229:103–142, 1999.

[11] I. Wegener. On the expected runtime and the success probability of evolutionary algorithms, 2000.

[12] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 1997.