Information Sciences 178 (2008) 2870–2879

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A general framework for statistical performance comparison of evolutionary computation algorithms

David Shilane [a], Jarno Martikainen [b,*], Sandrine Dudoit [a], Seppo J. Ovaska [b]

[a] Division of Biostatistics, School of Public Health, University of California, Berkeley, CA, USA
[b] Faculty of Electronics, Communications, and Automation, Helsinki University of Technology, Otakaari 5 A, P.O. Box 3000, 02015 Espoo, Finland

ARTICLE INFO

Article history:
Received 18 January 2006
Received in revised form 18 February 2008
Accepted 5 March 2008

Keywords:
Evolutionary computation
Genetic algorithms
Performance comparison
Statistics
Twofold sampling
Bootstrap
Multiple hypothesis testing

ABSTRACT

This paper proposes a statistical methodology for comparing the performance of evolutionary computation algorithms. A twofold sampling scheme for collecting performance data is introduced, and these data are analyzed using bootstrap-based multiple hypothesis testing procedures. The proposed method is sufficiently flexible to allow the researcher to choose how performance is measured, does not rely upon distributional assumptions, and can be extended to analyze many other randomized numeric optimization routines. As a result, this approach offers a convenient, flexible, and reliable technique for comparing algorithms in a wide variety of applications.

1. Introduction

Evolutionary algorithms (EAs) [1,9] are used to estimate the solution to difficult optimization problems. EAs are often hand-crafted to meet the requirements of a particular problem because no single optimization algorithm can solve all problems competitively [19]. When alternative algorithms are proposed, their relative efficacies should be assessed. Because EAs follow a stochastic process, statistical analysis is appropriate for algorithm comparison. This paper seeks to provide a general methodology for comparing the performance of EAs based on statistical sampling and hypothesis testing.

Prior research in the statistical design and analysis of EAs has considered a variety of approaches. Based upon a large number of experimental trials, Penev and Littlefair [12] demonstrate that the Free Search algorithm improves upon previous results from a variety of stochastic competitors on several optimization problems. This comparison consists of defining a number of performance metrics and computing average values for each algorithm. However, like many other evolutionary computation studies, these results are not statistically analyzed and substantiated. Because of the large sample size and clear observed differences in their results, we have no reason to doubt the specific findings of the Free Search study. Indeed, a statistical analysis of these data would likely add weight to the conclusions. In proposing a general framework for statistical performance comparison of EAs and similar randomized optimization algorithms, we seek to provide an experimental framework in which the results of similar studies may be assessed according to appropriate statistical tests.

Christensen and Wineberg [3] explain the use of appropriate statistics in artificial intelligence and propose non-parametric tests to verify the distribution of an EA's estimate of a function's optimal value. Flexer [8] proposes general guidelines for

* Corresponding author.
E-mail address: martikainen@iki.fi (J. Martikainen).

0020-0255/$ - see front matter © 2008 Published by Elsevier Inc.
doi:10.1016/j.ins.2008.03.007

statistical evaluation of neural networks that can also be applied to EAs. Although a variety of non-parametric tests are available, these procedures are often limited to specific parameters of interest. For instance, the Mann–Whitney test (also called Wilcoxon's *rank sum test* [16]) may be used to assess the equality of two populations' medians without requiring any information about the data's distribution. However, such a test is not easily adapted to other parameters, such as the mean difference between the two populations, the simultaneous comparison of more than two populations at once, or a simultaneous test of both the median and another parameter of interest. Czarn [4] discuss the use of the analysis of variance (ANOVA) in comparing the performance of EAs. Similarly, Castillo-Valdivieso et al. [2] and Rojas et al. [17] employ ANOVA methods to optimize the parameter values in the design of improved EAs for specific optimizations, whereas François and Lavergne [10] rely upon a generalized linear model. However, these procedures all require distributional assumptions that are not necessarily valid and also limit the class of performance metrics that can be used. Because EAs produce results according to complex stochastic processes, often very little is known about the distribution of results across algorithmic trials. We seek to address this problem by relying solely on empirical data generated from repeated trials of competing EAs. The proposed methodology employs a bootstrap-based multiple hypothesis testing framework [6,5,15,13] that may be applied to any parameter of interest, number of simultaneous hypotheses, and data distribution. The resulting procedure establishes an experimental framework in which EAs may be compared based upon empirical data.

An EA's initial population (Section 2) consists of a set of starting values for the evolution process. Most previous EA performance comparisons have only considered results for a single initial population or even provided different inputs for each algorithm studied. Supplying different single inputs to each EA may result in a *founder effect*, in which a population's initial advantage is continually propagated to successive generations. Furthermore, relying upon a single choice of initial population can at best determine the plausibility of preferring one candidate EA to another given suitable initial conditions. We can alleviate these issues by assessing relative performance over each of a representative sample of initial populations.

For each particular initial population sampled, two EAs may be compared by testing the null hypothesis of equal performance according to a specified performance metric. Student's t-statistics [11] are commonly used to test the equality of two population means. However, the parametric t-test assumes that the data are normally distributed. If this assumption is not valid, the resulting inference may not be meaningful. Therefore, we require a more general and objective framework for statistical performance comparison of EAs.

Because we are proposing a scientific method for performance comparison, it is important to design an effective experiment that specifies how data are collected and analyzed. To collect data, we propose a twofold sampling scheme to perform repeated EA trials at each of a representative sample of possible inputs. The candidate EAs' efficacies are then assessed in a multiple hypothesis testing framework that relies upon bootstrap resampling [6,5,15,13] to estimate the joint distribution of the test statistics. This methodology establishes a procedure for EA comparison that can be considered general in the following aspects: First, the results do not rely heavily on a single advantageous input. Second, the bootstrap-based testing procedure is applicable to any distribution and requires no *a priori* model assumptions. Finally, this methodology can be applied to essentially any function of the data collected, so the researcher is free to choose how performance should be evaluated. The result is a general framework for performance comparison that may be used to compare EAs or other stochastic optimization algorithms based upon empirical data.

The paper is organized as follows: Section 2 provides a brief introduction to EAs and presents a twofold sampling scheme for data collection. Section 3 places performance comparison in a multiple hypothesis testing framework. Section 4 shows how to use the bootstrap to estimate the test statistics' underlying distribution. Section 5 introduces a variety of multiple testing procedures. Section 6 provides an example comparing the performance of two EAs seeking to minimize Ackley's function. Section 7 discusses further applications of statistics in EA performance comparison and concludes the paper.

2. Evolutionary algorithms and data collection

An EA's *cost* (or *objective*) *function* is a map $f : \mathbb{R}^D \to \mathbb{R}$ to be optimized. Any candidate solution is specified by an *individual* with a vector of *genes* (or *traits*, used interchangeably) $\mathbf{y} = (y_1, \ldots, y_D)$. Each individual has a corresponding cost given by $f(\mathbf{y})$. Given a *population* of individuals, an EA uses *evolutionary mechanisms* to successively create *offspring*, or new individuals. The evolutionary mechanisms often consist of some combination of selection, reproduction, and mutation operators. The *selection* mechanism ranks individuals by cost, determines which individuals shall produce offspring, and assigns individuals to *mating groups*. Given a mating group, *reproduction* combines the genes of individuals within the mating group one or more times to produce offspring. Finally, the *mutation* mechanism randomly alters the genetic profile of offspring immediately following conception.

An EA's *initial population* (or *input*, used interchangeably) is a set of individuals that serve as starting values for the algorithm, and its *result* is given by the minimum observed cost among all individuals produced in G generations. Once the evolutionary mechanisms are specified, one ordered iteration of these processes in sequence is considered one *generation*, and the *evolution* process proceeds for a user-specified number of generations $G \in \mathbb{Z}^+$.

An EA's result is determined by a stochastic process with two sources of variation: the initial cost and the algorithm's improvements to this cost produced by G generations of the random evolution process. Because an EA's result depends both on its initial cost and its efficacy given this initial population, a sample of result data should be collected in a *twofold sampling* scheme: we first generate a representative sample of initial populations, and then, for each of these inputs, we perform a

number of trials of each candidate EA. If we specify the number of generations G, the data are collected via the following algorithm:

1. Generate M initial populations of H individuals. Each individual is described by a D-dimensional vector of genes. The value of the dth gene of the hth individual of the mth population is labeled y_{mhd}. When referring to an overall population y_m or single individual y_{mh} within that population, the unnecessary indices will be dropped. Populations of individuals are constructed from genes randomly generated from an associated D-dimensional distribution function P. The resulting sample of individuals, and hence the population samples, are independent and identically distributed (i.i.d.).
2. Because an EA with a particular input follows a stochastic process, we sample results for each of the inputs generated in Step 1. For each initial population y_m, perform $n_a, a \in \{1, 2\}$, trials of algorithm a, allowing each trial to iterate for G generations. Save the ith trial's result as the $[m,i]$th entry of an $M \times n_a$ data matrix X_a. The number of generations G will be dropped from much of the subsequent notation, but it should be emphasized that the data collected are realizations of an experiment conducted by running an EA for G generations, and therefore, this analysis is only valid for the specified value of G.

The values n_a specify the sample size, and M represents the number of hypotheses, each of which correspond to an initial population. In general, one should collect as much data as possible given the computational constraints of the problem. In designing a performance comparison, the sample size should be selected by considering the variability of each EA's performance data. Choosing the sample size in terms of a pre-specified margin of error is a standard procedure in tests of a single hypothesis but is currently an open problem in multiple hypothesis testing applications.

3. Multiple hypothesis testing framework

For any comparison, we must first specify the theoretical *parameter of interest* $\mu_a(y_m)$, which in this setting is an EA's measure of performance given initial population y_m and the number of generations G. A typical choice for $\mu_a(y_m)$ is the EA's expected result after G generations. This parameter is estimated by a statistic $\hat{\mu}_a(y_m)$, which is just a function of the observed data X_a. When the expected result is the parameter of interest, the corresponding estimator is the sample mean:

$$\hat{\mu}_a(y_m) = \frac{1}{n_a}\sum_{i=1}^{n_a} X_a[m,i]; \quad m = 1,\ldots,M; \;\; a = 1,2 \tag{1}$$

and the estimated variance of the result is

$$\hat{\sigma}_a^2(y_m) = \frac{1}{n_a}\sum_{i=1}^{n_a}(X_a[m,i] - \hat{\mu}_a(y_m))^2; \quad m = 1,\ldots,M; \;\; a = 1,2. \tag{2}$$

It should be noted that the numerator of (2) may also be divided by $n_a - 1$ if so desired, but the convention of bootstrap variance estimates typically divides by n_a [7].

A multiple hypothesis testing framework is needed to compare algorithmic performance based on the data collected in Section 2. The null hypothesis can take many forms depending on the researcher's priorities. For example, one may wish to show that a new algorithm's expected optimal cost after G generations is greater than that of an established standard or that its performance falls in a particular range. Typically we wish to demonstrate that the candidate EAs differ significantly in performance given an initial population, so a skeptical null hypothesis would assume for each input that no difference in performance exists between the two algorithms. This corresponds to the multiple null hypotheses

$$H_m : \mu_1(y_m) - \mu_2(y_m) = 0; \quad m = 1,\ldots,M. \tag{3}$$

We then test (3) at multiple significance levels α (e.g. FWER 0.05 – Section 5). To do so, we must construct test statistics and corresponding decision rules that reject the null hypotheses when the test statistics exceed to-be-determined cut-offs. We test each component null hypothesis using a two-sample t-statistic:

$$t_m = \frac{\hat{\mu}_1(y_m) - \hat{\mu}_2(y_m)}{\sqrt{\frac{\hat{\sigma}_1^2(y_m)}{n_1} + \frac{\hat{\sigma}_2^2(y_m)}{n_2}}}; \quad m = 1,\ldots,M. \tag{4}$$

In order to specify cut-offs that probabilistically control a suitably defined *Type I error rate* (Section 5), we must estimate the underlying joint distribution of (4). When the data are assumed to follow a normal distribution, Student's t-distribution is appropriate for the marginal distributions of (4). However, if this assumption is not valid, the test statistics may not follow any mathematically simple distribution. Under either of these circumstances, the joint distribution of (4) can be estimated using the bootstrap.

4. Using the bootstrap in hypothesis testing

The bootstrap is a simulation-based resampling method that uses the data collected to estimate a statistic's distribution in a mathematically simple but computationally intensive way. This estimate is consistent, asymptotically efficient, and does

not rely upon parametric assumptions, so it is widely applicable to many problems in statistical inference [7]. In the setting of hypothesis testing, we can estimate the underlying joint distribution of (4) via the following algorithm [6,5,15,13]:

1. Specify a number $B \in \mathbb{Z}^+$ (typically at least 10,000 for multiple hypothesis testing) of bootstrap iterations.
2. Let $n = n_1 + n_2$. Concatenate the columns of X_1 and X_2 to form an $M \times n$ data matrix X. For each $b \in \{1, \ldots, B\}$, sample n columns at random with replacement from X and store this resampling in an $M \times n$ matrix $X^{\#b}$.
3. For $b = 1, \ldots, B$, compute bootstrap test statistics on each resampled matrix $X^{\#b}$. To do so, treat the first n_1 columns of $X^{\#b}$ as if it were the data set X_1 and the last n_2 columns as X_2. Then apply (4) to these subsets of $X^{\#b}$ to compute a vector of M test statistics from the resampled data. Because this procedure is repeated for each of the B bootstrap iterations, these test statistics may be stored in an $M \times B$ matrix T. The reader may refer to [6,5,13,15] for further details.
4. Obtain an $M \times B$ matrix Z by shifting T about its row means and scaling by its row standard deviations for $m = 1, \ldots, M; b = 1, \ldots, B$:

$$Z[m,b] = \sqrt{\min\left[1, \frac{1}{\frac{1}{B}\sum_{b=1}^{B}\left(T[m,b] - \frac{1}{B}\sum_{b=1}^{B} T[m,b]\right)^2}\right]}\left(T[m,b] - \frac{1}{B}\sum_{b=1}^{B} T[m,b]\right). \tag{5}$$

The estimate of the test statistic (4)'s joint distribution is given by the empirical distribution of the columns of Z in (5). In the next section, we will use this joint distribution to compute a measure of the observed data's extremity under the null hypothesis (3). For hypothesis testing applications, the bootstrap is implemented in the *MTP* function of the *R* statistical programming environment's *multtest* package [14].

5. Multiple testing procedures

The significance level α, the observed test statistics t_m (4), and the matrix of bootstrap test statistics Z (5) constitute the input to a multiple testing procedure (MTP). In this setting, a variety of methods that reflect a diversity of attitudes toward risk are available. Statistical tests can generate two types of errors: a *Type I error* (or *false positive*) occurs when a true null hypothesis is incorrectly rejected, and a *Type II error* (or *false negative*) occurs when a false null is not rejected. When testing M hypotheses simultaneously, as in (3), we define the following random variables: The number of Type I errors V, which is not observed, and the number of rejected hypotheses R, which is observed. Classical MTPs seek to control the Family-Wise Error Rate (FWER). More recent research has been developed to control the generalized Family-Wise Error Rate (gFWER), False Discovery Rate (FDR), and the Tail Probability for the Proportion of False Positives (TPPFP), which are defined in Table 1.

As described in [6,5,15,13,14,18], Table 2 lists a selection of available MTPs for each Type I error rate. The results of a multiple hypothesis test can be summarized in terms of several measures. A *rejection region* provides a set of values for which each hypothesis H_m of (3) is rejected while controlling the desired Type I error rate at level α. A $1 - \alpha$ *confidence region* estimates a plausible range of values for the parameter of interest based upon the estimator's inherent variability. If the

Table 1
Type I error rates

Type I error rate	Parameter	Parameter controlled
FWER	–	$\Pr(V > 0)$
gFWER	$k \in \mathbb{Z}^+$	$\Pr(V > k)$
FDR	–	$E[V/R]$
TPPFP	$q \in [0, 1]$	$\Pr(V/R > q)$

The Family-Wise Error Rate (FWER) is the probability of obtaining at least 1 false positive in the simultaneous test of M hypotheses. The *generalized* Family-Wise Error Rate (gFWER) is the probability of obtaining at least $k + 1$ false positives. The False Discovery Rate (FDR) is the expected proportion of false positives among all rejected hypotheses. Finally, the Tail Probability for the Proportion of False Positives (TPPFP) is the probability that the proportion of false positives among all rejected hypotheses exceeds a given threshold.

Table 2
MTPs by Type I error rate

Type I error rate	Multiple testing procedures
FWER	Single step (SS) max T, SS minP, step down (SD) maxT, SD minP, Bonferroni, Holm, Hochberg, SS Šidák, SD Šidák
gFWER	Augmentation procedure (AMTP), SS common cut-off, SS common quantile, empirical Bayes
FDR	Conservative augmentation, restrictive augmentation, Benjamini–Yekutieli (BY), Benjamini–Hochberg (BH)
TPPFP	AMTP, empirical Bayes

For details about these procedures, please refer to Dudoit and van der Laan [6].

experiment were repeated a large number of times, a proportion of approximately $1-\alpha$ of the resulting confidence regions would contain the true parameter. Finally, *adjusted p-values* [6] provide a measure of the data's extremity under the null hypothesis. The adjusted p-value for null hypothesis H_m is defined as the minimum value of Type I error level α for which H_m is rejected. Smaller adjusted p-values correspond to stronger evidence against the validity of the null hypothesis. Adjusted p-values from different MTPs controlling the same Type I error rate may be directly compared, with smaller values reflecting a less conservative test [6]. The MTPs of Table 2 are implemented in the *MTP* function of the *R multtest* package [14]. The user need only supply the data, the value of α, the form of the null hypothesis, the test statistic, Type I error rate to control, and select the MTP.

6. Example: Ackley's function minimization

6.1. Defining Ackley's function

We seek to compare two candidate EAs that approximate the minimum of a $D = 10$-dimensional Ackley function [1]. With $y_{mh} = (y_{mh1}, \ldots, y_{mhD})$ as in Section 2, Ackley's multi-modal function, which achieves a known minimum at the origin, is defined as

$$f(y_{mh}) = -c_1 \exp\left(-c_2\sqrt{\frac{1}{D}\sum_{d=1}^{D} y_{mhd}^2}\right) - \exp\left(\frac{1}{D}\sum_{d=1}^{D}\cos(c_3 y_{mhd})\right) + c_1 + \exp(1) \tag{6}$$

with the following parameters supplied for this example:

$$c_1 = 20, \quad c_2 = 0.2, \quad c_3 = 2\pi, \quad D = 10, \quad y_{mhd} \in (-20, 30).$$

6.2. Candidate EAs $Ackley_1$ and $Ackley_2$

The algorithms $Ackley_1$ and $Ackley_2$ were devised to estimate the minimum of (6). Each EA takes an input population y_m as described in Section 2. Each individual y_{mh} of this population has associated cost $f(y_{mh})$ given by (6). At each generation, both algorithms include a selection, reproduction, and mutation phase. $Ackley_1$ and $Ackley_2$ differ only in the choice of the mutation rate and are otherwise identical algorithms. The evolutionary mechanisms of these EAs are as follows:

Selection: For simplicity, the population size H is assumed to be a multiple of 4, though this method can be generalized with floor and ceiling operators for other values of H. Sort and re-label the H individuals in order of increasing $f(y_{mh}), h = 1, \ldots, H$. The $H/2$ best-fit individuals – those with the smallest values of (6) – are selected for reproduction, while the other members will not breed. For $h = 1, \ldots, H/4$, pair individuals $y_{\lfloor m(2h-1)\rfloor}$ and $y_{\lfloor m(2h)\rfloor}$ for mating. Although selection is the last phase of a generation, it is presented first because the initialization process that creates the 0th generation requires selection before the first generation of the evolution process may commence.

Reproduction: Selection in the previous generation pairs individuals $y_{\lfloor m(2h-1)\rfloor}$ and $y_{\lfloor m(2h)\rfloor}, h = 1, \ldots, H/4$, for mating. Each pair produces two offspring to replace individuals not selected. For the first child ($c = 1$), a uniform random variable λ_1 is generated on $(0, 1)$, and the second child ($c = 2$) receives $\lambda_2 = 1 - \lambda_1$. Genes are inherited (vector-wise) by the weighted average

$$y_{\lfloor m(H/2+2(h-1)+c)\rfloor} = \lambda_c y_{\lfloor m(2h-1)\rfloor} + (1-\lambda_c) y_{\lfloor m(2h)\rfloor}. \tag{7}$$

Mutation: Each offspring $y_{\lfloor H/2+1\rfloor}, \ldots, y_H$ may randomly mutate in a single gene at birth with probability θ_a. When mutation occurs, the gene is selected from a uniform random variable on $\{1, \ldots, D\}$, and this trait is assigned a uniform random variable on $(-20, 30)$. In this example, mutation probabilities for $Ackley_1$ and $Ackley_2$ are $\theta_1 = 0.1$ and $\theta_2 = 0.8$, respectively. Because only one of an individual's $D = 10$ genes may mutate, the expected proportions of mutating genes in $Ackley_1$ and $Ackley_2$ are 0.01 and 0.08.

Except for the mutation probability, $Ackley_1$ and $Ackley_2$ are identical EAs. The initial population is considered the completion of the reproduction and mutation phases for the 0th generation, and the first generation begins after selection of the initial population. The process of reproduction, mutation, and selection repeats a total of G generations, and the EAs' results are given by

$$\text{Result} = \min_{h\in\{1,\ldots,H\}} f(y_{mh}). \tag{8}$$

The value of (8) observed for EA a after G generations on the ith trial given initial population y_m is stored as the $[m, i]$th entry of the data matrix X_a. Because the reproduction and mutation phases have random components at each generation, the value $X_a[m, i]$ is a random variable.

It should be noted that $Ackley_1$ and $Ackley_2$ were designed solely to provide an example of our comparison methodology. Different population sizes, reproduction schemes, or mutation rates may lead to improved estimates of (6)'s minimum.

6.3. Study design and results

Using the twofold sampling scheme of Section 2, we generated $M = 100$ initial populations $y_1, \ldots, y_M$, each consisting of $H = 100$ individuals with $D = 10$ genes apiece. Each individual's traits were initialized using pseudo-random number generation from a uniform distribution on the interval $(-20, 20)$. It should be noted that subsequent mutations allowed genes to take any value in $(-20, 30)$, so only mutant genes and their offspring can reach the interval $[20, 30)$. Function (6) was used to assess each individual's cost. Then, for each initial population $m = 1, \ldots, M$, we collected result data on $n_1 = n_2 = 50$ trials of the EAs. On each trial, both $Ackley_1$ and $Ackley_2$ were allowed to evolve for $G = 10,000$ generations.

The data for the $Ackley_1$ and $Ackley_2$ trials are displayed in Fig. 1 as a function of initial population index. Fig. 2 shows the average performance of the EAs for each initial population. Though $Ackley_2$ produces a better (i.e. smaller) mean value of (8) than $Ackley_1$ at each initial population, Fig. 1 shows that $Ackley_1$ is capable of producing competitive results for some trials across all inputs. Furthermore, $Ackley_1$ appears to exhibit greater variance than $Ackley_2$ in its estimates. Therefore, it is not immediately clear that $Ackley_2$ does indeed perform better than $Ackley_1$.

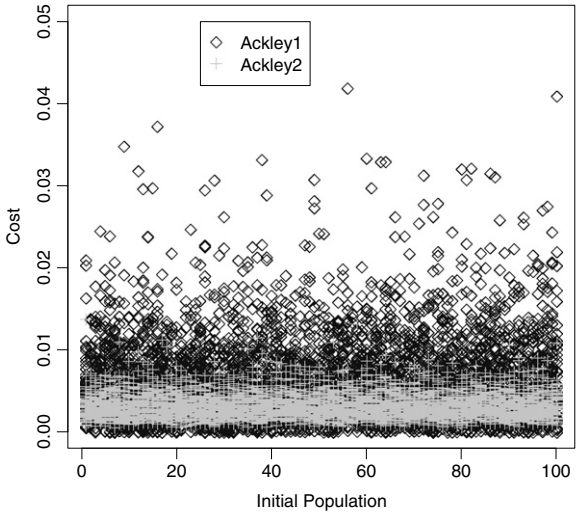

Fig. 1. Cost data for $Ackley_1$ and $Ackley_2$ trials by initial population.

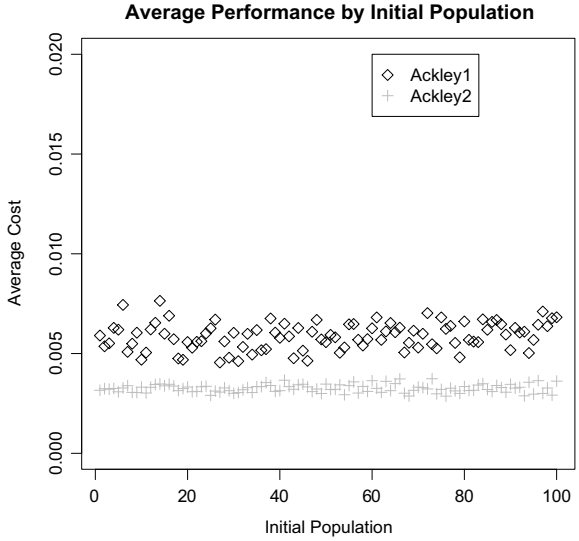

Fig. 2. Average cost of $Ackley_1$ and $Ackley_2$ by initial population.

We conducted two-sided tests of the multiple null hypotheses (3) corresponding to no difference in mean performance between $Ackley_1$ and $Ackley_2$ at each given input versus the alternative of unequal mean performance. Note that one could also perform one-sided tests that designate one candidate EA as superior to the other in the null hypothesis.

Hypotheses (3) were tested using the *multtest* package [14] of *R* based on the data collected and the test statistic (4). We first employed the FWER-controlling SS maxT MTP at nominal level $\alpha = 0.05$. Figs. 3–6 provide summary plots of the SS maxT results. Fig. 3 shows how the number of rejected hypotheses *R* grows as a function of α. The second plot depicts the SS maxT adjusted *p*-values in sorted order, which relates the growth of α to the number of rejected hypotheses. This curve indicates that 91 hypotheses are rejected at level $\alpha = 0.05$. Fig. 5 shows how the SS maxT adjusted *p*-values decrease with the absolute value of the test statistics. Here the adjusted *p*-values approach 0.05 as the test statistics increase toward -2.75, indicating that a given hypothesis is rejected for all test statistics smaller than this value. Fig. 6 displays the unordered SS maxT adjusted *p*-values, which allow one to identify the initial populations that result in significant (<0.05) performance differences.

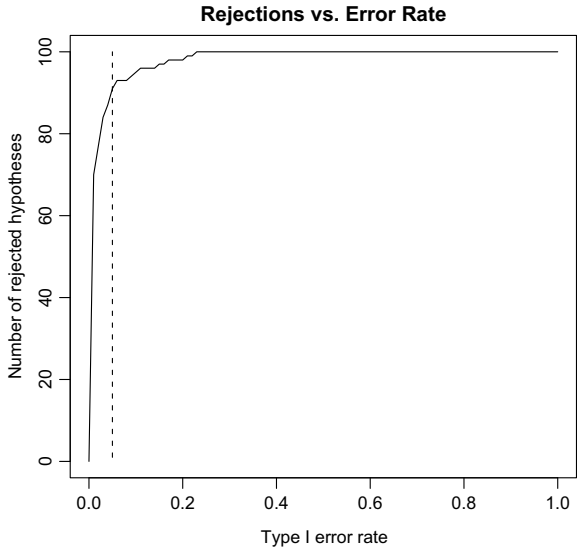

Fig. 3. The number of rejected hypotheses as a function of Type I error rate in SS maxT testing. Each hypothesis is rejected if its adjusted *p*-value is smaller than Type I error rate α.

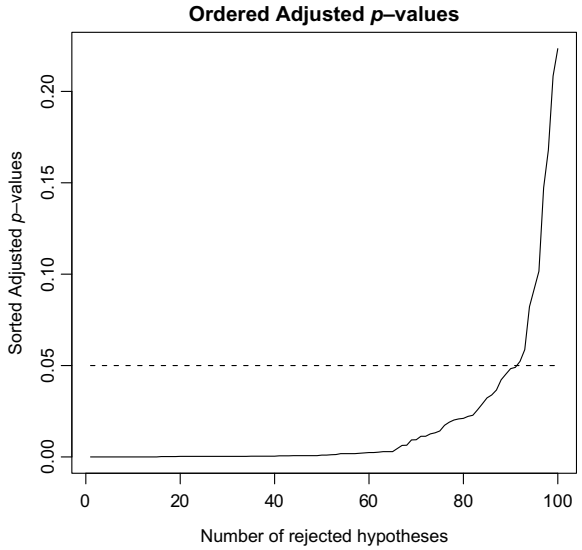

Fig. 4. Adjusted SS maxT *p*-values in sorted order.

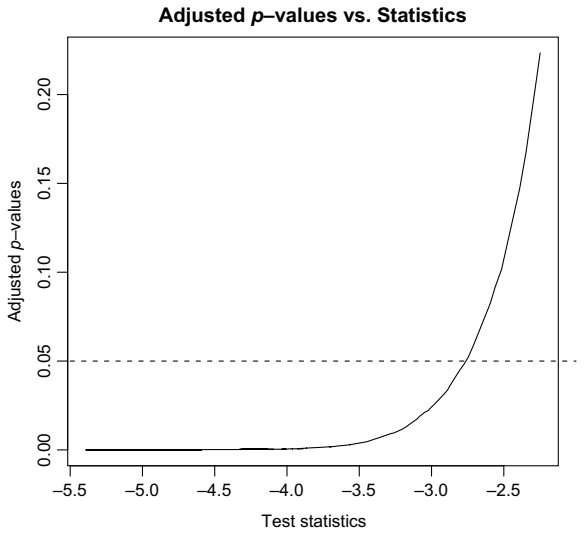

Fig. 5. Adjusted p-values as a function of test statistic value in the SS maxT test. From this result we can approximate the rejection and non-rejection regions for the test. It appears that the test rejected a component null hypothesis when the test statistic did not exceed a value of approximately -2.75.

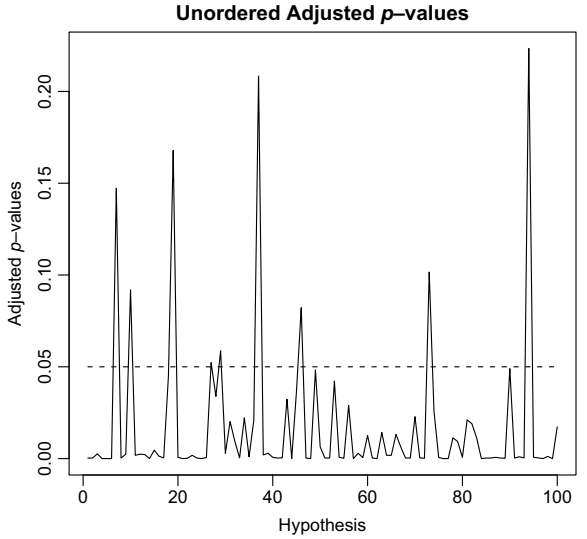

Fig. 6. Unordered adjusted p-values for the SS maxT test. With a significance level of $\alpha = 0.05$ (dashed line), we can identify the nine sampled initial populations that led to insignificant differences between the two Ackley EAs.

We then implemented a selection of the MTPs listed in Table 2 to test (3) under different Type I error rates. Table 3 displays the number of hypotheses rejected by each MTP at varying Type I error levels α. The following procedures reject all 100 hypotheses at level $\alpha = 0.05$: Holm, Hochberg, SD Šidák, and the TPPFP Augmentation Procedure with values $q = 0.07$ and $q = 0.10$.

For the gFWER and TPPFP-controlling augmentation procedures, the question remains whether the allowed number k or rate q of false positives is tolerable in testing EA performance differences. This question is epistemological in nature and must be decided by subject matter specialists. In practice, a maximum value for these parameters should be established before comparison takes place. Although the particular benchmark is somewhat arbitrary (much like the choice of $\alpha = 0.05$ in hypothesis testing), establishing a uniform standard is necessary for future studies.

The results of the test of (3) suggest a performance difference between $Ackley_1$ and $Ackley_2$. On each of the $M = 100$ sample input populations, $Ackley_2$ achieved a smaller average observed minimum. All MTPs rejected at least 84 of the $M = 100$ hypotheses at level $\alpha = 0.05$, and a number of procedures rejected all hypotheses even at level $\alpha = 0.01$. Therefore, based upon the data collected, we conclude that $Ackley_2$ significantly outperforms $Ackley_1$ in estimating the minimum of (6) when

Table 3
The number of rejected hypotheses R as a function of α for a selection of MTPs

Rate	MTP	0.01	0.03	0.05	0.07	0.1
FWER	SS maxT	70	84	91	93	95
	Bonferroni	85	89	91	94	95
	Holm	100	100	100	100	100
	Hochberg	100	100	100	100	100
	SS Šidák	85	89	91	94	95
	SD Šidák	100	100	100	100	100
gFWER	AMTP $k = 5$	70	84	91	93	95
	AMTP $k = 10$	75	89	96	98	100
FDR	Conservative AMTP	70	84	91	93	95
	Restricted AMTP	75	90	98	100	100
	BY	66	76	84	89	95
	BH	66	76	84	89	95
TPPFP	AMTP $q = 0.07$	99	100	100	100	100
	AMTP $q = 0.10$	100	100	100	100	100

the expected result obtained after G generations of evolution is the parameter of interest. Because the two algorithms only differed in their mutation probabilities, it appears that the increased mutation rate of $Ackley_2$ is beneficial in this application. Future comparisons may consider further increases in mutation to search for the frequency that best tunes the EA to the Ackley cost function (6).

7. Discussion

This paper's methodology provides a general approach to EA performance comparison. The proposed framework allows the researcher to choose the parameter of interest in an EA comparison. When parameters other than the expected optimum cost are used (such as the trimmed mean, the median, the 75th percentile, or other quantiles), our methodology is applicable provided that the necessary data are collected and appropriate estimators (1), null hypotheses (3), and test statistics (4) are chosen. In crafting an EA for a particular optimization problem, this paper's approach can be used iteratively to select the best among a set of candidate parameter values for quantities such as the mutation rate, population size, and selection proportion. When three or more EAs are simultaneously compared, null hypotheses of equality in means may be tested using F-statistics.

For illustration purposes, we considered an example in Section 6 involving a simple objective function (6), measure of performance $\mu_a(y_m)$, and sampling scheme based upon i.i.d. inputs. However, this methodology is applicable for general choices of the objective function, parameters of interest, sampling scheme, null hypotheses, test statistics, and number of algorithms to compare. Furthermore, although this paper studies performance comparison within the field of evolutionary computation, the general framework can be applied to essentially any stochastic optimization routine.

The reader should be cautioned that issues of sample size cannot be neglected. Determining an adequate sample size in multiple hypothesis testing settings is currently an open problem in the statistics literature. In general, the bootstrap approximation of (4)'s joint distribution grows more accurate as the values B and n_a increase. In practice, researchers may choose to collect as much data as a pre-specified time limit will allow. Data-adaptive study designs may also be implemented to halt data collection once a pre-specified level of *statistical power* is achieved. In hypothesis testing, the power of a test is defined as the probability of correctly rejecting a set of false null hypotheses given the true parameter values.

If competing algorithms draw from different input sets, then the test of a single hypothesis ($M = 1$) concerning average results from representative input samples may be considered. When the input sets are identical, an alternative to the approach of this paper may choose to average all trials in a single hypothesis test provided that all inputs are i.i.d. The choice of which approach to use is philosophical: this paper assumes that EAs should be compared using the same input sample. In this setting, the parameter of interest is the expected result obtained in G generations given the initial population. This allows the algorithm to be assessed solely on the merits of its evolutionary mechanisms without any possibility of a founder effect. However, if one views the input generation and resulting evolution as inextricably linked in the same algorithm, then a single hypothesis testing framework may be more appropriate, and this paper's methodology is otherwise applicable. In this scenario, the parameter of interest shifts to the unconditional expectation of performance. Though a single test may simplify the interpretation of performance differences, this approach lacks the appeal of direct performance comparison on the same trial inputs.

The researcher may also wish to compare EAs as a function of time by collecting data at regular generational intervals. Displaying performance curves and confidence regions graphically may allow one to quickly determine decision criteria

and search for clues about an algorithm's rate of convergence and asymptotic result. Finally, an EA's efficacy should be considered in terms of both performance and computational complexity. Researchers may consider performing a comparison in which each candidate algorithm is allowed to run for the same amount of time instead of the same number of generations to satisfy both objectives simultaneously.

Acknowledgements

The authors would like to thank the anonymous referees for their insightful comments and suggestions aimed at improving this paper. In addition, the authors gratefully acknowledge Frances Tong for her helpful comments and suggestions during the process of preparing the initial manuscript.

References

[1] T. Bäck, Evolutionary Algorithms in Theory and Practice, Oxford University Press, 1996.
[2] P.A. Castillo-Valdivieso, J.J. Merelo, A. Prieto, I. Rojas, G. Romero, Statistical analysis of the parameters of a neuro-genetic algorithm, IEEE Transactions on Neural Networks 13 (6) (2002) 1374–1394.
[3] S. Christensen, M. Wineberg, Using appropriate statistics – statistics for artificial intelligence, in: Tutorial Program of the Genetic and Evolutionary Computation Conference, Seattle, WA, 2004, pp. 544–564.
[4] A. Czarn, C. MacNish, K. Vijayan, B. Turlach, R. Gupta, Statistical exploratory analysis of genetic algorithms, IEEE Transactions on Evolutionary Computation 8 (4) (2004) 405–421.
[5] S. Dudoit, M.J. van der Laan, K.S. Pollard, Multiple testing. Part I. Single-step procedures for control of general type I error rates, Statistical Applications in Genetics and Molecular Biology 3 (1) (2004). Article 13.
[6] S. Dudoit, M.J. van der Laan, Multiple Testing Procedures and Applications to Genomics, Springer, New York, NY, 2008.
[7] B. Efron, R. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, Boca Raton, FL, 1994.
[8] A. Flexer, Statistical evaluation of neural network experiments: minimum requirements and current practice, in: R. Trappel (Ed.), Proceedings of the 13th European Meeting on Cybernetics and Systems Research, Cybernetics and Systems '96, Vienna, Austria, vol. 2, 1996, pp. 1005–1008.
[9] D.B. Fogel, Evolutionary Computation: Toward a New Philosophy of Machine Intelligence, Wiley-IEEE Press, Hoboken, NJ, 2000.
[10] O. François, C. Lavergne, Design of evolutionary algorithms – a statistical perspective, IEEE Transactions on Evolutionary Computation 5 (2) (2001) 129–148.
[11] J.S. Milton, J.C. Arnold, Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences, McGraw-Hill, New York, NY, 1990.
[12] K. Penev, G. Littlefair, Free search – a comparative analysis, Information Sciences 172 (2005) 173–193.
[13] K.S. Pollard, M.D. Birkner, M.J. van der Laan, S. Dudoit, Test statistics null distributions in multiple testing: simulation studies and applications to genomics, Journal de la Société Française de Statistique 146 (1–2) (2005) 77–115.
[14] K.S. Pollard, S. Dudoit, M.J. van der Laan, Multiple testing procedures: the multtest package and applications to genomics, Bioinformatics and Computational Biology Solutions using *R* and Bioconductor, Statistics for Biology and Health, Springer-Verlag, New York, 2005. pp. 249–271 (Chapter 15).
[15] K.S. Pollard, M.J. van der Laan, Choice of a null distribution in resampling-based multiple testing, Journal of Statistical Planning and Inference 125 (1–2) (2004) 85–100.
[16] J.A. Rice, Mathematical Statistics and Data Analysis, Duxbury Press, Pacific Grove, CA, 2006.
[17] I. Rojas, J. González, H. Pomares, J.J. Merelo, P.A. Castillo, G. Romero, Statistical analysis of the main parameters involved in the design of a genetic algorithm, IEEE Transactions on Systems, Man, and Cybernetics–Part C 32 (1) (2002) 31–37.
[18] M.J. van der Laan, S. Dudoit, K.S. Pollard, Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives, Statistical Applications in Genetics and Molecular Biology 3 (1) (2004). Article 15.
[19] D.H. Wolpert, W.G. MacReady, No free lunch theorems for optimization, IEEE Transactions on Evolutionary Computation 1 (1) (1997) 67–82.