A priori comparison of binary crossover operators: No universal statistical measure, but a set of hints.

Leila Kallel and **Marc Schoenauer**
CMAP – URA CNRS 756
Ecole Polytechnique
Palaiseau 91128, France
FirstName.Name@polytechnique.fr

Abstract. The choice of an operator in evolutionary computation is generally based on comparative runs of the algorithms. However, some statistical a priori measures, based on samples of (parents-offspring), have been proposed to avoid such brute comparisons. This paper surveys some of these works in the context of binary crossover operators. We first extend these measures to overcome some of their limitations. Unfortunately, experimental results on well-known binary problems suggest that any of the measures used here can give false indications. Being all defined as averages, they can miss the important parts: none of the tested measures have been found to be a good indicator alone. However, considering together the mean improvement to a target value and the Fitness Operator Correlation gives the best predictive results. Moreover, detailed insights on the samples, based on some graphical layouts of the best offspring fitnesses, seem to allow more accurate predictions on the respective performances of binary crossover operators.

1 Introduction

The most simple and common way to compare some parameter settings (including the choice of operators) of evolutionary algorithms (EAs) is to run numerous experiments for each one of the settings, and to compare averages of on-line or off-line fitness. This trial-and-error experimental method is unfortunately time consuming. Nevertheless, the lack of efficient and robust measures of difficulty and/or adequation of settings for a given problem has made this experimental comparative method very popular since the early conferences of EAs (see e.g. [14]).

Some studies have proposed heuristical approaches to try to estimate the efficiency of different operators a priori. These methods are usually based on statistical measures computed on random samples of points of the search space. Two kinds of such methods can be distinguished :

The first set of methods do not take into account the specific characteristics of operators, but rather implicitly rely on conceptual notions that are common to all operators. For instance, the strong causality of mutations is assumed in [10], where a distance in the genotype space is used to assess correlation length, and in [8], where the correlation between fitness and distance to optimum is described as a good predictor of EA performances. In the same line, the respectfulness of operators toward schemata [6, 12] or epistasis [4] is used to characterize the suitability of the representation and its associated operators.

The second kind of methods directly address the issue of operator performance relative to the fitness landscape. In the early days of evolution strategies, Rechenberg [13] derived the famous 1/5 rule by theoretical calculation of the expected progress rate toward the optimum for a zero mean Gaussian mutation, on the sphere and corridor models in $\mathbb{R}^n$, in relation to the probability of a successful mutation.

More recently, Weinberger [15] introduced the notion of fitness operator correlation, for asexual operators, as the correlation between the fitnesses of the parents and their iterated offspring. Manderick [10] adapted this correlation measure to crossover operators. He devoted attention to the correlation between mean parental fitness and mean offspring fitness, and indicated the possible usefulness of this measure to assess the role of an operator on the NK-landscapes and the TSP problems.

However, as noted by Altenberg [1], such fitness operator correlation suffers from theorical limitations: a maximal correlation is possible even in the absence of any improvement of fitness from parents to offspring. Altenberg further suggested that the performance of a genetic algorithm is related to the probability distribution of offspring fitness indexed by their parents' fitness that result from application of the genetic operator during the GA. Unfortunately, Altenberg noted that empirical estimation of the fitness probability distribution could be problematic if the searched surface has multiple domains of attraction.

Grefenstette [7] used the probability distribution of offspring fitnesses indexed by the mean parents fitnesses to yield a first-order linear approximation of the suitability of a genetic operator. Another approach was suggested by Fogel and Ghozeil [5]: to compare different selection schemes and operators, they use the mean improvement of offspring, and concludes that the uniform intermediate recombinations yields better results (higher mean improvement) than the one-point crossover, for all three selection methods they tested, on some two-dimensional functions in the real-valued framework.

This paper is concerned with statistical a priori measures that can be used to assess crossover efficiency: All the above measures (fitness operator correlation, probability distributions, and mean improvement) have been tested separately. As far as we know, no comparison has been made of their efficiency on the same problem. The goal here is to achieve such a comparison on different binary problems from the GA literature. The different selection schemes will not be discussed here, and for all experiments linear ranked-based selection will be used.

The paper is organized as follows. Section 2 presents in detail and discusses the statistical measures briefly introduced above. A modified version of the fitness operator correlation is introduced to address some of its limitations. Mean improvement method is modified to only account for strictly positive improvements. Section 3 experimentally compares the estimates of those different measures on various test cases in the light of actual results of GA runs, demonstrating the inadequacy of the fitness operator correlation for most problems, and the potentialities of the statistical measures based on best offspring fitnesses. These results yield a better understanding of the domain of validity and the limitations of each statistical measure. Moreover, a detailed look at some samples appeals for on-line measures: Further research directions are sketched in the last section, in light of the obtained results.

2 Comparing Crossover operators

This section details some of statistical measures that have been proposed in previous work, and discusses some of their limitations. This leads to proposing alternative measures.

Consider the problem of optimizing a real-valued fitness function $\mathcal{F}$ defined on a metric search space $\mathcal{E}$ and suppose that there exists one global optimum of $\mathcal{F}$. The common basis of all statistical measures described below is the use of a sample of random individuals in $\mathcal{E}$. The operators to be assessed are then applied on all (or part of) the sample.

2.1 Mean Fitness Operator Correlation

Manderick & al. [10] define the Mean Fitness Operator Correlation (M-FOC) as follows:

From a sample of n couples of individuals (x_i,y_i) , let (f_{pi}) be the mean fitness of couple i, and (f_{ci}) the mean fitness of their two offspring with crossover OP. The *M-FOC coefficient* of the operator OP is defined by:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(f_{pi}-\bar{f}_p)(f_{ci}-\bar{f}_c)}{\sigma_{F_p}\sigma_{F_d}}, \qquad (1)$$

where $\bar{f}_p, \bar{f}_c, \sigma_{F_p}$ and σ_{F_d} are respectively the means and standard deviations of parents' and offspring mean fitness.

This coefficient measures the extent to which the mean fitness of a couple of parents can give some information about the mean fitness of their offspring. A very high correlation between parents and offspring (M-FOC coefficient close to 1) indicates that offspring mean fitness would be almost proportional to parental mean fitness. No correlation (M-FOC coefficient close to 0) shows that no information about offspring mean fitness can be gained from the parents mean fitness. GAs using high-correlated operators are expected to achieve best performances as demonstrated in [10] on the NK-landscapes and the TSP problems.

Unfortunately, suppose there is a quasi-linear relationship between the mean fitnesses of both sets of parents and offspring. The M-FOC correlation will be maximum (1), whatever the slope of that linear relation. In particular, if that slope is less than 1, no offspring fitness can be better (in mean) than its parents': the correlation can be maximal when no improvement is to be expected! However unrealistic this particular case can be, it nevertheless shows the limits of M-FOC as a measure of operator's efficiency.

In that line, as noted by Fogel and Ghozeil [5], M-FOC cannot be used on problems that yield zero mean difference between parents and offspring fitnesses, as in the case of linear fitness functions with real-valued representations. The correlation is always maximal, and cannot reflect the high dependency of their convergence rate on the the standard deviation of the Gaussian mutation [13].

2.2 Best Fitness Operator Correlation

To address the above-mentioned limitations of M-FOC, we propose to study the correlation between the best (rather than mean) parent and offspring fitnesses. The definition of the B-FOC coefficient is the same as in equation 1, except that $\bar{f}_p, \bar{f}_c, \sigma_{F_p}$ and σ_{F_d} are replaced by the means and standard deviations of the best fitnesses of parents and offspring. This coefficient should avoid some of the ambiguities pointed out in the preceding subsection.

2.3 Mean Improvement

Fogel and Ghozeil [5] use 1000 couples of parents to compare different selection schemes and operators. To generate each couple, 100 random individuals are drawn. The couple is then chosen among them either randomly, or as the two best individuals, or as one random and the best. The mean improvement of the best offspring (of each couple) over the best individual of the initial 100 individuals is then computed. Such mean improvements are finally averaged over the 1000 couples. As noted in [5], one limitation of this approach is its computational cost (102.000 fitness evaluations). The following definitions are a tentative to achieve the same goals at lower cost: Consider a sample of N random individuals as $N/100$ disjoint subsets of individuals. Among each of these 100 size subsets, 100 couples of parents are selected randomly (selection scheme is presented in 3.1). Two cases are considered:

Mean Improvement over the best parent
In the same line as Fogel and Ghozeil's second procedure (computing the improvement over the best parent), we propose to compute the improvement of the best offspring of each couple over its best parent. The mean improvement of N couples of parents is computed as follows:

$$\text{MI-P} = \frac{1}{N}\sum_{i=1}^{N} \max(0, \mathcal{F}(\text{best offspring}) - \mathcal{F}(\text{best parent}))$$

Mean Improvement to target
Fogel and Ghozeil's first procedure amounts to computing improvement of the offspring over a target set to the best fitness of the current set of 100 individuals. Rather, we define a target $\mathcal{T}$ as the median of the set $\{f_i\}_{i=1,\frac{N}{100}}$ where f_i is the fitness of the best individual among those of the ith -100 size- subset generated. The mean improvement to target is then computed as follows:

MI-T=

$$\frac{1}{N}\sum_{i=1}^{N}\begin{cases}\max(0,\mathcal{F}(\text{best offspring})-\mathcal{T}) \text{ if } \mathcal{F}(\text{best offspring}) > \mathcal{F}(\text{best parent})\\ 0 \text{ otherwise}\end{cases}$$

2.4 Probability of Improvement

The above definitions of MI-P and MI-T do not consider the number of offspring that actually are source of improvement. In order to separate the mean improvement from the probability of improvement, the latter is defined as the fraction of offspring actually having a better fitness than their parents (PI-P) or the target (PI-T). Two new measures are then defined, termed MI-P+ and MI-T+, by counting in the definitions of MI-P and MI-T only those offspring.

3 Experimental Validation

3.1 Experimental conditions

From now on, the search space will be the binary space $\{0,1\}^n$ ($n = 900$ for all problems, except the Royal Road), and the underlying EA some standard binary GA: population size 100, linear ranked-based selection, crossover rate of 0.6, mutation rate of $1/n$ per bit, elitist generational replacement.

Two operators are compared: The uniform crossover (UC) and the one-point crossover (1C). For each of them, the statistical measures are computed either on samples of 10000 individuals, with one crossover per parent, or on samples of 200 individuals, each couple generating 100 offspring (by repeating 50 crossovers operations). The parents are selected with a linear ranking selection applied to each disjoint subset of 100 individuals. In both cases, the computational cost is 20000 fitness evaluations per operator.

The goal of these experiments is to compare operators under the same conditions of initialization and selection. Therefore neither the selection scheme (as in [5]) nor the initialization procedure (as in [9]) will be modified.

The stability of all the following statistical measures have been checked carefully. The relative difference between measures on UC and 1C, has been found stable over 10 trials in most cases. However, all unstability cases are marked with a "*" in the tables.

3.2 Test cases

Onemax problems
In the well-known *onemax* problem, the fitness function is simply the number of 1's in the bitstring. Hence, the fitness is equal to n minus the distance to the global optimum. In order to avoid structural bias, modified versions of the onemax problem were designed: The fitness function is defined as the Hamming distance to a fixed bitstring. That bitstring is chosen to have a certain number

O of 1's, randomly placed in the bitstring. Such problem is termed the *(O,n-O)-onemax* problem. Different values for O were experimented: $(900,0)$, $(800,100)$ and $(450,450)$.

Gray-coded Baluja functions F1 and F3
Consider the functions of k variables $(x_1, \ldots x_k)$, with $x_i \in [-2.56, 2.56]$ [2]:

$$F_1(\mathbf{x}) = \frac{100}{10^{-5} + \sum_{i=0}^{k-1} |y_i|}, y_0 = x_0 \text{ and } y_i = x_i + y_{i-1} \text{ for } i = 1, \ldots, k-1$$

$$F_3(\mathbf{x}) = \frac{100}{10^{-5} + \Sigma_i |.024 * (i+1) - x_i|}$$

They reach their maximum value of 10^7 at point $(0, \ldots 0)$. The Gray-encoded versions of F_i, with 100 variables encoded on 9 bits each are considered.

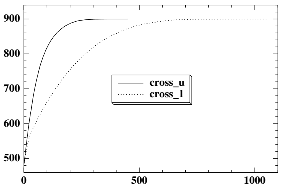
The 4 peaks problem
In the FourPeaks problem [3], the fitness is the maximum of the length of the sequence of 0's starting at the first bit position and the length of the sequence of 1's ending at the last position, plus a reward of n if both sequences are larger than a given threshold T. There are two global optima, made of a block of 0's followed by a block of 1's, of lengths T and $n-T$ or $n-T$ and T. However, there are two local optima, the all 1's and all 0's bitstrings, from which it is difficult for a GA to escape. The most difficult instance of that problem (i.e. $T = n/4$) is used here.
The Ugly problem
The Ugly problem [16] is defined from an elementary deceptive 3-bit problem (F(x) = 3 if x = 111; F(x) = 2 for x in $0\star\star$, and F(x) = 0 otherwise). The – deceptive – full problem is composed of 300 concatenated elementary deceptive problems. The maximum is the all 1's bitstring.

The Royal Road problem
The 64-bits Royal Road problem used here was conceived to study into details the combination of features most adapted to GA search (laying a *Royal Road*). A precise definition as well as an analysis of the unexpected difficulties of this problem can be found in [11].

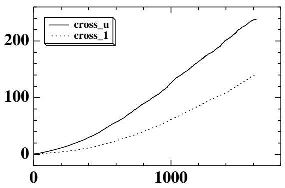
3.3 Results

For all problems above, the statistical measures defined in section 2 have been be computed, and their predictions compared to actual GA runs.

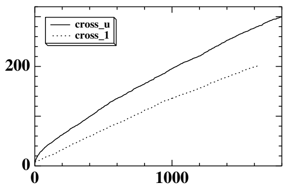
Average on-line results of the runs (over 21 runs) are presented in Figure 1, and the statistical measures for both uniform (UC) and one-point (1C) crossovers are presented in tables 1 and 2. Each table presents FOC and MI-P coefficients with one crossover per parent, as well as MI-T coefficients for both 1 and 50 crossovers per parent. Only one of the three onemax problems is presented since the two others gave almost the same results.



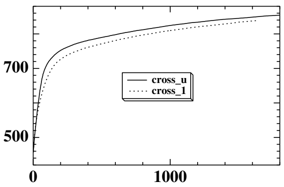
(a) onemax problems



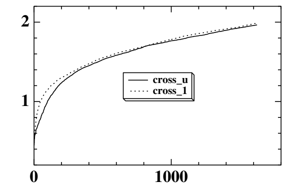
(b) F3



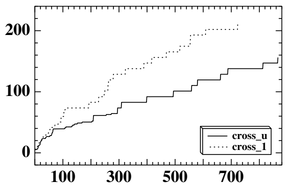
(c) 4peaks



(d) ugly



(e) F1g



(f) 4peaks

Figure 1: *Average on-line results of 21 GA runs on different problems (fitness $\times$generation): influence of the crossover operators.*

Large unstability has been found for MI-T with the Royal Road problem. This is due to the highly discontinuous fitness values it yields (observed values are limited to 1, 5.2, 5.4, 21.6). Hence, when computing MI-T on few individuals, this can amount to contradictory results for different samples. More generally, MI-T has been found slightly unstable. This is not surprising when considering that very low probabilities are involved (lines PI-T in the tables). However, when considering the case of 50 crossovers per parent, the unstability does no longer alter the significance of the differences between MI-T coefficients of operators. A last general remark is that all MI coefficients should be considered relatively to the actual fitness range in the sample.

FOC Coefficients The only case where higher M-FOC coefficient and better convergence coincide is obtained with the F1g problem! On the 4peaks, Ugly and F3g problems the reverse situation is observed: Better on-line performances (Figure 1) are obtained for the operator with lower M-FOC values (Tables 1 and 2). On the opposite, runs for the Royal Road problem with either UC or 1C are similar (the 99% T-test fails) until generation 500, though, their respective M-FOC coefficients are quite different (Table 2).

On the same problems, B-FOC give similar indications than M-FOC, and hence suffer the same defects. On the Onemax problems, where the M-FOC always equals 1, the B-FOC coefficient reaches higher values for the 1C, while the runs (Figure 1) using UC are far better than those using 1C.

Hence neither M-FOC not B-FOC seems an accurate predictor of relative performance of operators, at least under linear ranking selection.

	onemax problems		F3g		4peaks	
	UC	1C	UC	1C	UC	1C
M-FOC	1	1	0.88	0.99	0.7	0.93
B-FOC	0.7	0.8	0.6	0.8	0.6	0.99
MI-P	3.4(0.03)	2.7(0.03)	0.01	0.008	0.4	0.002
MI-P+	7(0.08)	6(0.06)	0.02	0.016	2	2
PI-P	0.23	0.23	0.26	0.25	0.1	0.003
MI-T	0.07(0.012)	0.06(0.016)	0.00026	0.00021	0.1	0.002
MI-T+	5.6(0.3)	5.6(0.5)	0.02	0.02	2.2	1.8
PI-T	0.006	0.005	0.006	0.005	0.005	6e-5
MI-T (50)	**3.2(1)**	**1.2(0.7)**	**0.01**	**0.004**	**0.6**	**0.02**
MI-T+ (50)	**8(1)**	**6(1)**	**0.028**	**0.025**	**2.7**	**1.7**
PI-T (50)	**0.02**	**0.018**	**0.018**	**0.016**	**0.01**	**0.0001**

Table 1: *Mean(standard deviation) of statistical a priori measures for different problems, on a 10000 individuals sample, 2 offspring for each parent with the uniform and one-point crossover operators. Bold fonts indicates measures on a 200 individuals sample, and 100 offspring for each parent. The + concerns strictly positive improvements only.*

	ugly		F1g		royal	
	UC	1C	UC	1C	UC	1C
M-FOC	0.74	0.99	0.2	0.7	0.18	0.92
B-FOC	0.5	0.8	0.1(0.01)	0.6(0.005)	0.18	0.91
MI-P	5.9(0.07)	3.6(0.05)	0.05(7e-4)	0.03(4e-4)	0.19	0.019
MI-P+	12.3(0.1)	7.7(0.1)	0.09(9e-4)	0.05(8e-3)	4.2	4
PI-P	0.27	0.23	0.3	0.28	0.02	0.002
MI-T	0.08(0.01)	0.07(0.008)	7e-4(1e-4)	5e-4(1e-4)	e-3(8e-4)	2e-4(4e-4)
MI-T+	6.6(0.5)	6.8(0.5)	0.05(2e-3)	0.05(5e-3)	1.5(1.2)	0.9(2.3)
PI-T	0.006	0.005	0.007	0.006	0.0004	0.0001
MI-T(50)	**5(2)**	**1.3(0.8)**	**0.05(0.02)**	**0.02(7e-3)**	**0.07(0.07)**	**0.01(0.02)**
MI-T+(50)	**10(2)**	**8(1.5)**	**0.07(0.01)**	**0.06(0.01)**	**2.5(2.8)**	**0.6(4)**
PI-T(50)	**0.018**	**0.016**	**0.016**	**0.015**	**4e-4(1e-4)**	**4e-4(8e-4)**

Table 2: *Mean(standard deviation) of statistical a priori measures for different problems, on a 10000 individuals sample, 2 offspring for each parent with the uniform and one-point crossover operators. Bold fonts indicates measures on a 200 individuals sample, and 100 offspring for each parent.*

Mean Improvement to Parents If we consider the F1g problem, the results of the runs (figure 1), show that the one-point crossover average run is better than the uniform one until generation 200 (confirmed with a 99 % confidence T-test). But mean improvement to parents values (Table 2), suggest the opposite. A close look at the plot of best parents $\times$ offspring (Figure 2), shows that UC yields better improvement than OP only at low fitness values of (best) parents. The situation is reversed for higher fitnesses. This situation could have been predicted by the Best FOC results: a 0.7 correlation for the 1C means that the best offspring's fitness is almost proportional to its best parent's fitness, while the 0.2 value for the UC means on the otherhand that well-fitted offspring can be obtained whatever the (best) parent's fitness. The misleading results of improvement to parents are so no longer surprising. It seems that MI-P values can be compared reliably when the B-FOC coefficients are similar, as it is the case of the F3g and the three onemax problems.

For the Ugly and Royal Road problems, the situation is similar to that of F1g problem. Average improvements to parents are very different in favor to UC, whereas the runs show identical performances: the T-test with 99 % confidence finds no difference between average runs until generation 30 (Ugly) and 500 (Royal Road). Again, at the opposite of 1C, the UC yields good fitness offspring from low fitness parents, as illustrated by the B-FOC coefficients.

Note however that the reverse is not true: different B-FOC values do not necessarily imply false MI-P previsions, as illustrated by the 4peaks problem.

Considering the probabilities of improvement (PI-P and MI-P+) for the F1g and Ugly problems does not improve the results: the same bias as for MI-P is observed, for the same reasons (different B-FOC coefficients).

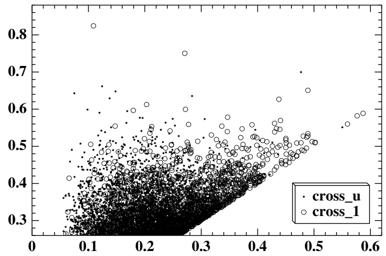


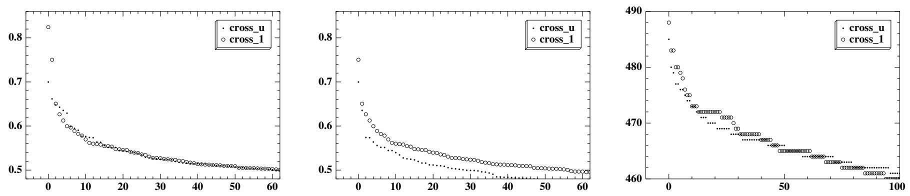
Figure 2 :*Fitness of best offspring × fitness of best parent, for the offspring that outperformed their parents, on the F1g problem.*

Mean Improvement to Target The case of 1 crossover per parent is found to be rather unstable and often involves very low improvement probabilities, which does not help comparison. Note for example that for the onemax problems (table 1), the observed differences of MI-T, with one crossover per parent, are very small (around 0.01) relatively to the range of fitness values of the problem ([400,500]). Hence only the case of 50 crossovers per parents (bold fonts in tables) will be considered in the following.

When considering the improvement to target, the dependency on the parents' fitness vanishes. So no bias is induced by the dissymmetric repartitions of improvement to parents. Nevertheless, for F1g and Ugly problems, MI-T values (Table 2) are still indicating some superiority of UC, whereas the opposite actually happens during the GA runs for F1g, and Ugly runs are similar until generation 30 (Figure 1). The main reason for that is probably that averaging the improvements hides both the evolution of their differences and the best fitted offspring obtained: For the F1g problem, when getting to higher parents' fitness, the 1C gets better improvement to target than the UC as can be seen for instance on the plot (best parent $\times$ best offspring) of Figure 2.

Considered together with the probability of improvement, some additional information about the sample repartition can be obtained. In fact, IP-T does take into account the whole set of generated offspring, while MI-T does only take into account the set of best offspring of each parent. For the problematic cases of F1g and Ugly, MI-T+ values are still favoring the wrong operator, but with much smaller differences than MI-T, and IP-T values are close for both operators. MI-T+ still suffers from the bias of averaging, and does nevertheless hide the most important information about offspring, which is their maximal fitnesses, and how they evolve when higher parents are considered.

Further directions Other measures, not presented here, have been computed for both UC and 1C operators; such as mean fitness values of all offspring, or of all offspring that outperformed their parents. None of these have been found of any help to find about operators suitability. Rather, the most important factor is found to be the 'maximal potential' of an operator, i.e. how fit are the best offspring it can give birth to.



(a) F1g :whole sample (b) F1g: Half sample (c) Ugly: whole sample

Figure 3: *Fitnesses of best offspring $\times$ their rank, with one crossover per parent, for uniform and one-point crossovers: The relative position of curves changes when only the top half of the sample is considered for offspring generation.*

To illustrate this, consider the plots of fitness repartition of offspring as a function of their rank according to their fitness. The case of F1g problem presented in Figure 3-a shows the best offspring of the whole sample. One cannot say which curve is better: 1C yields higher fitness offspring than UC, but UC gives a higher number of well-fitted offspring. (Remember that the all the considered measures suggest the superiority of UC, and that the runs are ordered the other way round.) But consider now Figure 3-b, which plots the same repartition of best offspring, but when considering only the best half of the parents in the sample. The curve for 1C gets largely better than UC. Figure 3-c presents the case of the ugly problem: curves of 1C and UC are similar (and so are the runs), while all the previous measures favor the UC.

This plot was found efficient in predicting the runs order for all the considered problems. The plot is almost unchanged when considered with the top half of the sample.
A crucial factor for operator comparisons seems to be their ability to generate high values, even rarely, or even for the best of parents. And this can be quite different from any mean value on the whole sample. The mean as a measure does not give any information about the most likely direction of evolution. This calls for alternative measures, as well as for on-line dynamic measurement.

4 Summary and conclusion

Some statistical measures to allow assessment of a priori suitability of a genetic operator have been presented and experimented to compare 1-point crossover and uniform crossover in the binary framework.

These experiments have first confirmed the inadequacy of fitness operator correlation, which was either unable to detect differences, or even predicting the inverse of actual results of GA. FOC only indicates whether well-fitted offspring are obtained from well-fitted parents or not. However, such information can be useful in addition to other measures.

Average improvement to parents (MI-P) can be very misleading for two reasons:
1- The dependency on the parents fitnesses can induce big differences (in MI-P) while the fitness distribution of offspring -independently of their parents- is the same for different operators.
2- It suffers from the loss of information induced by any measure based on averaging.
Those remarks address both cases of 1 and 50 crossovers per parents, which generally gave the same tendancies, the case of 50 crossovers, showing much higher differences.

When considering average improvement to target (MI-T), only the case of 50 crossovers per parents gave stable and significant differences relatively to the fitness range of each problem. MI-T does no longer suffers from the first bias above but still meets the second:
MI-T, even with 50 crossovers per parents, fails on both F1g and Ugly problems to predict the best crossover. In our opinion, this is due to the second bias of average measure. In fact, on the F1g problem, the relative position of offspring fitness distribution for UC and 1C operators changes when considering only fit parents. Moroever, the averaging does not capture any information about the fittest offspring of each operator. This information, is found to be crucial to compare crossovers.

So none of the proposed measures was found satisfactory alone, though MI-T appears the most useful – especially when B-FOC values are similar. To design more efficient measures, we think averaging should be avoided, and best fitness offspring should be emphasized. Whatever the measure, the most important factor seems to be the way it evolves when getting to higher fitness parents: Further work will have to design online measures.

References

1. L. Altenberg. The schema theorem and Price's theorem. In L. D. Whitley and M. D. Vose, editors, *Foundations of Genetic Algorithms 3*, pages 23–49, San Mateo, CA, 1995. Morgan Kaufmann.
2. S. Baluja. An empirical comparizon of seven iterative and evolutionary function optimization heuristics. Technical Report CMU-CS-95-193, Carnegie Mellon University, 1995.
3. S. Baluja and R. Caruana. Removing the genetics from the standard genetic algorithms. In A. Prieditis and S. Russel, editors, *Proceedings of ICML95*, pages 38–46. Morgan Kaufmann, 1995.
4. Y. Davidor. Epistasis variance: A viewpoint on representations, GA hardness, and deception. *Complex systems*, 4:369–383, 1990.
5. D. B. Fogel and A. Ghozeil. Using fitness distributions to design more efficient evolutionary computations. In T. Fukuda, editor, *Proceedings of the Third IEEE International Conference on Evolutionary Computation*, pages 11–19. IEEE, 1996.
6. D. E. Goldberg and M. Rudnick. Genetic algorithms and the variance of fitness. *Complex Systems*, 5:266–278, 1991.
7. J. J. Grefenstette. Predictive models using fitness distributions of genetic operators. In L. D. Whitley and M. D. Vose, editors, *Foundations of Genetic Algorithms 3*, pages 139–161. Morgan Kaufmann, 1995.
8. T. Jones and S. Forrest. Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In L. J. Eshelman, editor, *Proceedings of the* 6^{th} *International Conference on Genetic Algorithms*, pages 184–192. Morgan Kaufmann, 1995.
9. L. Kallel and M. Schoenauer. Alternative random initialization in genetic algorithms. In Th. Bäeck, editor, *Proceedings of the* 7^{th} *International Conference on Genetic Algorithms*. Morgan Kaufmann, 1997. To appear.
10. B. Manderick, M. de Weger, and P. Spiessens. The genetic algorithm and the structure of the fitness landscape. In R. K. Belew and L. B. Booker, editors, *Proceedings of the* 4^{th} *International Conference on Genetic Algorithms*, pages 143–150. Morgan Kaufmann, 1991.
11. M. Mitchell and J.H. Holland. When will a genetic algorithm outperform hill-climbing ? In S. Forrest, editor, *Proceedings of the* 5^{th} *International Conference on Genetic Algorithms*, page 647, 1993.
12. N. J. Radcliffe and P. D. Surry. Fitness variance of formae and performance prediction. In L. D. Whitley and M. D. Vose, editors, *Foundations of Genetic Algorithms 3*, pages 51–72. Morgan Kaufmann, 1995.
13. I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution.* Fromman-Holzboog Verlag, Stuttgart, 1973.
14. J. D. Schaffer, R. A. Caruana, L. Eshelman, and R. Das. A study of control parameters affecting on-line performance of genetic algorithms for function optimization. In J. D. Schaffer, editor, *Proceedings of the* 3^{rd} *International Conference on Genetic Algorithms*, pages 51–60. Morgan Kaufmann, 1989.

15. E.D. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics*, 63:325–336, 1990.
16. D. Whitley. Fundamental principles of deception in genetic search. In G. J. E. Rawlins, editor, *Foundations of Genetic Algorithms.* Morgan Kaufmann, 1991.