On Risky Methods for Local Selection under Noise

Günter Rudolph

Universität Dortmund, Fachbereich Informatik, D–44221 Dortmund / Germany

Abstract. The choice of the selection method used in an evolutionary algorithm may have considerable impacts on the behavior of the entire algorithm. Therefore, earlier work was devoted to the characterization of selection methods by means of certain distinguishing measures that may guide the design of an evolutionary algorithm for a specific task. Here, a complementary characterization of selection methods is proposed, which is useful in the presence of noise. This characterization is derived from the interpretation of iterated selection procedures as sequential non-parametric statistical tests. From this point of view, a selection method is risky if there exists a parameterization of the noise distributions, such that the population is more often directed into the wrong than into the correct direction, i.e., if the error probability is larger than $1/2$. It is shown that this characterization actually partitions the set of selection methods into two non-empty sets by presenting an element of each set.

1   Introduction

Selection methods may occur on two occasions during the population's life cycle of an evolutionary algorithm. They are used for choosing mating partners if recombination operators are used, and if the parents produce a surplus of offspring then they are used for keeping the population at a constant size. In both cases, these selection methods are responsible for moving the population towards regions with better fitness values. Since this happens differently fast or reliable for the variety of selection methods commonly in use, it has been tried to characterize these selection methods by quantities like takeover time, takeover probability, selection intensity, and related metrics [1–5]. Recently, it was also examined to which extent these measures are affected by noisy fitness functions [6–8]. Here, a complementary characterization of selection methods in the presence of noise is proposed. The key idea rests on the observation that the repeated application of some selection method to a population of random elements may be interpreted as a sequential non-parametric statistical test [9]. From this point of view, there are many measures that may serve to characterize the statistical power of a selection method. For example, a first simple distinguishing feature of a selection method might be based on its ability of keeping the error probability below $1/2$. Remarkably, this type of characterization actually partitions the set of selection methods into two non-empty sets. This is shown in sections 3 and 4 by presenting a member of each class. The theoretical foundation of these sections is introduced next.

2 Theoretical Framework

Suppose that the determination of the fitness value (to be maximized) is stochastically perturbed by additive noise which continuous distribution (with support $\mathbb{R}$) is symmetrical with respect to zero. More specifically, let μ be the true, unperturbed fitness value of some individual. Then the perturbed fitness value is given by $\mu + \sigma Z$ where $\sigma > 0$ and the median of random variable Z is zero. If $\mathsf{E}[\,|Z|\,] < \infty$ then also $\mathsf{E}[\,Z\,] = 0$, otherwise the expectation of Z does not exist. An assumption regarding the scale of Z is not yet necessary.

Let the initial population consist of n individuals (n even) where $n/2$ individuals are of type x and the remaining half of type y. An individual is said to be of type x (resp. y) if its random fitness value X (resp. Y) possesses the distribution function

$$F_X(z) = F_Z\left(\frac{z - \mu_x}{\sigma_x}\right) \quad \text{resp.} \quad F_Y(z) = F_Z\left(\frac{z - \mu_y}{\sigma_y}\right). \tag{1}$$

Without loss of generality it is assumed that $\mu_x > \mu_y$. In this case, $\mu_x > \mu_y$ if and only if $\mathsf{P}\{\,X < Y\,\} < 1/2$. Repeated application of some selection method will lead to a uniform population with probability one, i.e., each individual is either of type x or of type y. One might expect that a proper selection method leads more often to a uniform population of type x individuals than to type y individuals. As it is shown in the subsequent sections this property is not valid for all selection methods commonly used in evolutionary algorithms.

In general, this scenario can be modeled by homogeneous finite Markov chains which state space and transition probabilities depend on the selection method under consideration and on the spatial structure, if any, of the population. In any case, the resulting Markov chain has only two absorbing states, namely, the states representing uniform populations.

Definition
Let the initial population consists of $n/2$ individuals of type x and $n/2$ individuals of type y with distribution functions as specified in equation (1). A selection method is called *risky* if there exists a parameterization of the distributions, i.e., parameter values $(\mu_x, \mu_y, \sigma_x, \sigma_y)$ with $\mu_x > \mu_y$, such that the absorption probability to a uniform population with lower median μ_y is larger than the absorption probability to a uniform population with higher median μ_x. $\qquad\square$

For the sake of brevity, only three local selection methods on a certain spatially structured population will be investigated. The imposition of this limitation has the compensating advantage that the associated Markov chain models reduce to less complex random walk models. Suppose that the individuals are arranged in a one-dimensional array of size n. Initially, the first $n/2$ cells of the array are filled with individuals of type x and the remaining cells with individuals of type y. Prior to selection the random fitness values of the individuals are calculated. The type of each cell after selection only depends on the fitness values before selection of the cell itself and its nearest neighbors. It is clear that the type of a cell is unaltered if its type is identical to the type of both

neighboring cells. Therefore, it is sufficient to restrict the attention to the section of the array where type x individuals meet type y individuals. For this purpose consider the 4-tuple (X_1, X_2, Y_1, Y_2) of independent random variables. Notice that the leftmost and rightmost cell will not alter their type since there are further type x cells to the left and type y cells to the right. Only the two cells in between can change their type. Thus, there are four possible arrangements after selection: $(xyyy)$, $(xxxy)$, $(xxyy)$, and $(xyxy)$. Assume that the probability of the last outcome is zero whereas

$$
\left.\begin{array}{l}
\mathsf{P}\{\,(xxyy) \rightarrow (xyyy)\,\} = \alpha > 0 \\
\mathsf{P}\{\,(xxyy) \rightarrow (xxxy)\,\} = \beta > 0 \\
\mathsf{P}\{\,(xxyy) \rightarrow (xxyy)\,\} = 1 - (\alpha + \beta) .
\end{array}\right\} \tag{2}
$$

Let N_k be the random number of type x cells at step $k \geq 0$. Then N_k performs a random walk on the state space $\{0, 1, \ldots, n\}$ so that the transition probabilities given in equation (2) are now expressible by

$$
\begin{array}{ll}
\mathsf{P}\{\,N_{k+1} = i - 1 \mid N_k = i\,\} = \alpha \\
\mathsf{P}\{\,N_{k+1} = i + 1 \mid N_k = i\,\} = \beta \\
\mathsf{P}\{\,N_{k+1} = i \mid N_k = i\,\} \quad = 1 - (\alpha + \beta)
\end{array}
$$

for $i = 2, \ldots, n - 2$. If $i \in \{1, n - 1\}$ then the transition probabilities will generally be different from α and β, but these differences may be neglected if the population size is large enough. A formal proof of this claim will be published in a subsequent paper. Under the assumption that the population size is sufficiently large the probability of absorption a_n from initial state $n/2$ to state n is

$$
a_n = \frac{1}{1 + (\alpha/\beta)^{n/2}}
$$

whereas the probability of absorption to state zero is $a_0 = 1 - a_n$ [10]. Thus, $a_0 > a_n$ if $\alpha > \beta$ or, equivalently, if the *replacement error* $\rho = \alpha/(\alpha + \beta) > 1/2$. If this case may occur for some local selection method then it will be classified *risky*.

3 A Risky Local Selection Method

3.1 Characteristics of Local Best Offspring Selection

The local best offspring selection method works as follows: Each cell adopts the type of that cell with largest fitness value among the cell itself and its nearest neighbors. To determine the transition probabilities α and β consider the random tuple (X_1, X_2, Y_1, Y_2). The second cell changes its type if and only if $\max\{X_1, X_2\} < Y_1$ whereas the third cell changes its type if and only if $\max\{Y_1, Y_2\} < X_2$. Notice that these events are mutual exclusive. As a consequence, one obtains

$$
\alpha = \mathsf{P}\{\,X_{2:2} < Y\,\} \qquad \text{and} \qquad \beta = \mathsf{P}\{\,Y_{2:2} < X\,\}
$$

where $X_{2:2} = \max\{X_1, X_2\}$ and $Y_{2:2} = \max\{Y_1, Y_2\}$. These probabilities can be calculated via

$$\alpha = \mathsf{P}\{\, X_{2:2} < Y\,\} = F_{X_{2:2}}(Y) = const. = \mathsf{E}[\,F_{X_{2:2}}(Y)\,]$$

$$= \int_{-\infty}^{\infty} F_{X_{2:2}}(y)\, f_Y(y)\, dy = \int_{-\infty}^{\infty} F_X^2(y)\, f_Y(y)\, dy \tag{3}$$

where $f_Y(y) = \frac{d}{dy} F_Y(y)$, and analogously for β. In general, the inequality $\alpha < \beta$ is nonlinear and can be solved only in exceptional cases. For example, the integrals can be used to consider the case $\sigma_x = \sigma_y = \eta > 0$. Owing to equations (1) and (3) one easily obtains

$$\alpha = \int_{-\infty}^{\infty} F_Z^2(z - \xi)\, f_Z(z)\, dz < \int_{-\infty}^{\infty} F_Z^2(z + \xi)\, f_Z(z)\, dz = \beta$$

where $\xi = (\mu_x - \mu_y)/\eta > 0$. The situation changes if σ_y is sufficiently larger than σ_x.

3.2 Determination of Critical Parameter Ranges

Unless the distribution of the noise is specified it is hardly possible to determine the parameter ranges for which $\alpha > \beta$. A parameterization with this property will be termed critical.

To consider the most usual case let $G_i \sim N(\mu, \sigma^2)$ and $Z_i \sim N(0, 1)$ with $i = 1, 2$ be independent normal random variables. As usual, the symbol "$\sim$" means that the random variable on its left hand side possesses the distribution specified on its right hand side. Similarly, the symbol "$\sim_a$" indicates that the distributional relationship is approximately valid.

Since $G_i \overset{d}{=} \mu + \sigma Z_i$ it follows that $G_{2:2} \overset{d}{=} \mu + \sigma Z_{2:2}$ and hence

$$\mathsf{E}[\,G_{2:2}\,] = \mu + \sigma\, \mathsf{E}[\,Z_{2:2}\,] = \mu + \frac{\sigma}{\sqrt{\pi}} \quad \text{and} \quad \mathsf{V}[\,G_{2:2}\,] = \sigma^2\, \mathsf{V}[\,Z_{2:2}\,] = \sigma^2\, \frac{\pi - 1}{\pi}$$

where the operator $\overset{d}{=}$ indicates that the random variables on its left and right hand side possess the same distribution. The approximation of the replacement error rests on the observation that the distribution of $G_{2:2}$ is well approximated by a normal distribution with expectation $\mathsf{E}[\,G_{2:2}\,]$ and variance $\mathsf{V}[\,G_{2:2}\,]$. As a consequence, if $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_i \sim N(\mu_y, \sigma_y^2)$ are normally distributed random variables with $\mu_x > \mu_y$ then

$$X_{2:2} - Y \sim_a N\left(\mu_x - \mu_y + \frac{\sigma_x}{\sqrt{\pi}}, \sigma_x^2\, \frac{\pi - 1}{\pi} + \sigma_y^2\right)$$

$$Y_{2:2} - X \sim_a N\left(\mu_y - \mu_x + \frac{\sigma_y}{\sqrt{\pi}}, \sigma_x^2 + \sigma_y^2\, \frac{\pi - 1}{\pi}\right)$$

and hence

$$\alpha = \mathsf{P}\{\, X_{2:2} - Y < 0\,\} \approx 1 - \Phi\left(\frac{\delta\, \pi^{1/2} + \eta}{\eta\,(\pi - 1 + c^2\, \pi)^{1/2}}\right)$$

$$\beta = \mathsf{P}\{\, Y_{2:2} - X < 0\,\} \approx \Phi\left(\frac{\delta\, \pi^{1/2} - c\, \eta}{\eta\,(\pi + c^2\,(\pi - 1))^{1/2}}\right)$$

where $\delta = \mu_x - \mu_y > 0$, $\eta = \sigma_x$, and $\sigma_y = c\,\sigma_x$ with $c > 0$. Assume that δ and η are fixed. Since

$$\alpha \to 1 - \Phi(0) \qquad\qquad = \frac{1}{2} \qquad \text{strictly monotonically increasing whereas}$$

$$\beta \to 1 - \Phi\big((\pi - 1)^{-1/2}\big) < \frac{1}{4} \qquad \text{strictly monotonically decreasing}$$

as $c \to \infty$ there must exist a value $c_0 > 0$ such that $\alpha > \beta$ and therefore $\rho > 1/2$ for all $c > c_0$. It remains to ensure that this property is not an artifact of the approximation via the normal distribution above. Owing to equation (3) the values for α and β can be reliably calculated via numerical integration. For this purpose let $\mu_x = 1$, $\mu_y = 0$, and $\eta = 1$. Figure 1 reveals that $c_0 \approx 4.2$ for this particular choice of parameters. Thus, the approximation via the normal distribution already offers considerable insight into the situation. One may conclude that for every choice of the triple (μ_x, μ_y, σ_x) with $\mu_x > \mu_y$ there exists a critical value $c_0 > 0$ such that the replacement error ρ is larger than $1/2$ for every $\sigma_y = c\,\sigma_x$ with $c > c_0$. As a consequence, this type of local selection may lead more often into the wrong than into the correct direction—at least for the specific initial population considered here.
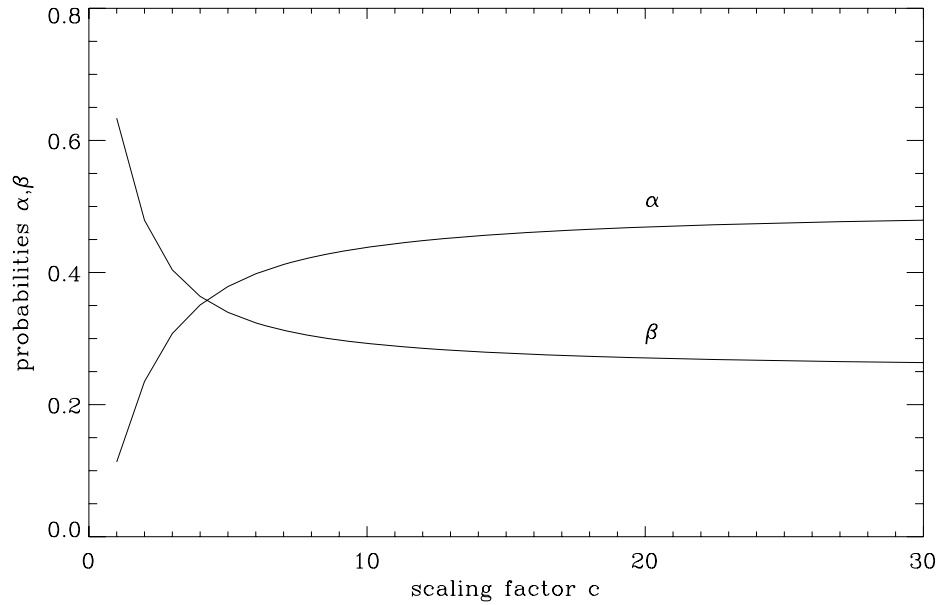

Fig. 1. Probabilities α and β for $(\mu_x, \mu_y, \sigma_x, \sigma_y) = (1, 0, 1, c)$ and varying scaling factor c.

3.3 Numerical Validation for Random Initial Populations

The analysis presented so far presupposes a very special initial population: The first $n/2$ cells are of type x whereas the last $n/2$ cells are of type y. It is by no means obvious that the results remain valid if the initial population is a random permutation of the initial population considered previously. In this case there are

$$\binom{n}{n/2} \sim 2^{n+1/2}/\sqrt{n\,\pi}$$

equally likely initial populations with $n/2$ cells of each type x and y. The existence of the critical scale parameter c_0 in this more general situation may be validated by numerical experiments with random initial populations. More specifically, for each population size $n \in \{50, 100\}$ and scale parameter $c = 2(0.05)6$ the process was run 1000 times with random initial populations. The relative frequency of the event "absorption at uniform population of type y cells" is an estimator of the absorption probability a_0. Figure 2 reveals that there actually exists a value c_0 for which $a_0 > 1/2$ if $c > c_0$ and vice versa. Moreover, the value of c_0 is apparently between 4.20 and 4.25, which is in agreement with the value found in the previous subsection. This observation provides evidence that the random walk model is an appropriate approximation of the more general situation.
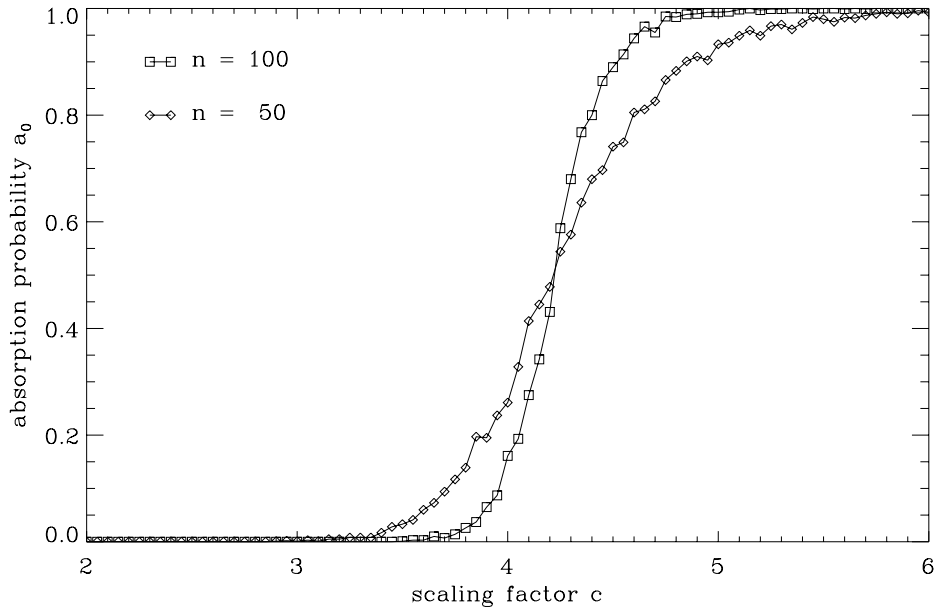

Fig. 2. Empirical absorption probability a_0 based on 1000 experiments per population size $n \in \{50, 100\}$ with parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y) = (1, 0, 1, c)$ and random initial population.

To sum up, we can say that this selection method may become risky if the scale of the noise is a strictly monotonous increasing function of the distance between the individual's fitness value μ and the maximum fitness value μ^*, i.e., $\sigma = h(\mu^* - \mu)$ for some strictly monotonous increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$ with $h(x) = 0$ if and only if $x = 0$. In this case it is ensured that individuals with low true fitness will encounter larger additive noise than individuals with high true fitness. Notice that this is exactly the situation which may lead to predominantly wrong selection decisions.

4 Are there Non-Risky Local Selection Methods?

4.1 Characteristics of Local Random Neighbor Tournament Selection

Local random neighbor tournament selection [9] works as follows: For each cell the individual chooses either its left or right neighbor with the same probability and adopts the chosen neighbor's type if this neighbor's fitness value is larger than the fitness value of the cell itself. Again, consider the random tuple (X_1, X_2, Y_1, Y_2) of fitness values. Only the second and third cell may change their type. The second cell changes its type if and only if it competes with its right neighbor and $X_2 < Y_1$ whereas the third cell changes its type if and only if it competes with its left neighbor and $X_2 > Y_1$. Notice that these event are mutual exclusive. As a consequence, one obtains

$$\alpha = \frac{\gamma}{2} \qquad \text{and} \qquad \beta = \frac{1 - \gamma}{2}$$

where $\gamma = \mathsf{P}\{ X < Y \} < 1/2$ (because of $\mu_x > \mu_y$; see section 2). Since $\alpha < \beta$ (or $\rho < 1/2$) this leads to $a_n > a_0$ regardless of the scaling parameters σ_x and σ_y, i.e., despite potentially arbitrarily scaled noise the local random neighbor tournament selection method leads more often into the correct than into the wrong direction.

If the initial population is drawn at random then the situation is considerably more complicated since the probability distribution of the number of type x cells may now depend on the ordering of all n in lieu of only four random variables. A formal analysis of this situation is beyond the scope of this paper. Needless to say, the non-existence of a critical parameterization for random initial populations cannot be proven by numerical experiments. But as shown next, there is a similar local selection method that is certainly non-risky.

4.2 Characteristics of Alternating Local Binary Tournament Selection

Let the population be arranged on a ring instead of a linear array and let $c_i \in \{x, y\}$ denote the type of cell $i \in \{0, 1, \ldots, n - 1\}$. At iteration $k \geq 0$ the population is grouped into pairs (c_i, c_{i+1}) such that $i \in \mathbb{Z}_n$ is odd if k is even and vice versa. Each pair performs a binary tournament and the types of the pair are set to that of the winner. Thus, there are $n/2$ *independent* binary tournaments. Let the initial population be drawn at random and let $d > 0$ be the number[1] of pairs with type (x, y) or (y, x)

---

[1] If $d = 0$ then the frequencies of type x and y cells is not changed. Notice that this event does not affect the absorption probabilities. But if this event has been occurred then $d > 0$ for the next iteration—or the population is uniform.

of the current population. The probability that such a pair transitions to a pair of type (y, y) is $\gamma = \mathsf{P}\{ X < Y \} < 1/2$ so that the probability distribution of the number D of (y, y)-pairs after selection is binomially distributed with parameters (d, γ). Since $\mathsf{P}\{ D = i \} > \mathsf{P}\{ D = d - i \}$ for $0 \leq i < d/2$ if and only if $\gamma < 1/2$ it follows that a decrease of type y cells is uniformly more likely than an increase, regardless of the current state of the population. Since the initial population has the same number of type x and y cells the property above ensures that $a_n > a_0$. As a consequence, this selection method leads more often into the correct than into the wrong direction.

5 Conclusions

The distinction between risky and non-risky methods for selection under noise leads to a clear recommendation which selection methods should be avoided in the presence of additive noise. A quantitative determination of the absorption probabilities, however, may become a very complex task. Therefore, it should be aimed at developing simpler yet sufficient conditions permitting a distinction between risky and non-risky methods.

The observation that the local best offspring selection rule is risky only for state-dependent noise might lead to the conjecture that all selection methods commonly used in evolutionary algorithms are non-risky under constant additive noise. Its verification would be a pleasant result.

The interpretation of repeated selection as a sequential statistical test offers the opportunity of transferring typical measures known from statistical test theory to selection methods under noise. This may open the door to more detailed guidelines for the design of evolutionary algorithms that operate in the presence of noise.

Acknowledgments

References

1. D. E. Goldberg and K. Deb. A comparative analysis of selection schemes used in genetic algorithms. In G. J. E. Rawlins, editor, *Foundations of Genetic Algorithms*, pages 69–93. Morgan Kaufmann, San Mateo (CA), 1991.
2. M. de la Maza and B. Tidor. An analysis of selection procedures with particular attention paid to proportional and Boltzman selection. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 124–131. Morgan Kaufmann, San Mateo (CA), 1993.
3. T. Bäck. Selective pressure in evolutionary algorithms: A characterization of selection mechanisms. In *Proceedings of the First IEEE Conference on Evolutionary Computation, Vol. 1*, pages 57–62. IEEE Press, Piscataway (NJ), 1994.
4. T. Blickle and L. Thiele. A comparison of selection schemes used in evolutionary algorithms. *Evolutionary Computation*, 4(4):361–394, 1996.

5. U. Chakraborty, K. Deb, and M. Chakraborty. Analysis of selection algorithms: A Markov chain approach. *Evolutionary Computation*, 4(2):133–167, 1996.
6. B. L. Miller and D. E. Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 4(2):113–131, 1996.
7. Y. Sakamoto and D. E. Goldberg. Takeover time in a noisy environment. In T. Bäck, editor, *Proceedings of the 7th International Conference on Genetic Algorithms*, pages 160–165. Morgan Kaufmann, San Francisco (CA), 1997.
8. D.B. Fogel and A. Ghozeil. The schema theorem and the misallocation of trials in the presence of stochastic effects. In *Proceedings of the 7th Annual Conference on Evolutionary Programming*. Springer, Berlin, 1998.
9. G. Rudolph. Reflections on bandit problems and selection methods in uncertain environments. In T. Bäck, editor, *Proceedings of the 7th International Conference on Genetic Algorithms*, pages 166–173. Morgan Kaufmann, San Fransisco (CA), 1997.
10. M. Iosifescu. *Finite Markov Processes and Their Applications*. Wiley, Chichester, 1980.