Threshold Selection, Hypothesis Tests, and DOE Methods

Thomas Beielstein
Department of Computer Science XI, University of Dortmund, D-44221 Dortmund, Germany.
tom@ls11.cs.uni-dortmund.de
Sandor Markon
FUJITEC Cm.Ltd. World Headquarters, 28-10, Shoh 1-chome, Ibaraki, Osaka, Japan.
markon@rd.fujitec.co.jp

Abstract- Threshold selection – a selection mechanism for noisy evolutionary algorithms – is put into the broader context of hypothesis testing. Theoretical results are presented and applied to a simple model of stochastic search and to a simplified elevator simulator. Design of experiments methods are used to validate the significance of the results.

1 Introduction

Many real world optimization problems have to deal with noise. Noise arises from different sources, such as measurements errors in experiments, the stochastic nature of the simulation process, or the limited amount of samples gathered from a large search space. Evolutionary algorithms (EA) can cope with a wide spectrum of optimization problems [16]. Common means used by evolutionary algorithms to cope with noise are resampling, and adaptation of the population size. Newer approaches use efficient averaging techniques, based on statistical tests, or local regression methods for fitness estimation [3, 1, 17, 6, 15].

In the present paper we concentrate our investigations on the selection process. From our point of view the following case is fundamental for the selection procedure in noisy environments: *Reject or accept a new candidate, while the available information is uncertain. Thus, two errors may occur: An α error as the probability of accepting a worse candidate due to noise and a β error, the error probability of rejecting a better candidate.*

A well established technique to investigate these error probabilities is hypothesis testing. We state that threshold selection (TS) can be seen as a special case of hypothesis testing. TS is a fundamental technique, that is used also used in other contexts and not only in the framework of evolutionary algorithms. The TS-algorithm reads: *Determine the (noisy) fitness values of the parent and the offspring. Accept the offspring if its noisy fitness exceeds that of the parent by at least a margin of τ; otherwise retain the parent.*
The theoretical analysis in [13], where TS was introduced for EAs with noisy fitness function values, were based on the progress rate theory on the sphere model and were shown for the $(1+1)$-evolution strategy (ES). These results were subsequently transfered to the S-ring, a simplified elevator model. Positive effects of TS could be observed. In the current paper we will base our analysis on mathematical statistics.

This paper is organized as follows: In the next section we give an introduction into the problems that arise when selection in uncertain (e.g. noisy) environments takes place. The basic idea of TS is presented in the following section. To show the interconnections between the threshold value and the critical value, statistical hypothesis testing is discussed. Before we give a summary, we show the applicability of TS to optimization problems: A stochastic search model – similar to the model that was used by Goldberg in his investigation of the mathematical foundations of Genetic Algorithms – and the S-ring – a simplified elevator simulator – are investigated [7, 13].

2 Selection in Uncertain Environments

Without loss of generality we will restrict our analysis in the first part of this paper to maximization problems. A candidate is 'better' ('worse'), if its fitness function value is 'higher' ('lower') than the fitness function value of its competitor. Suppose that the determination of the fitness value is stochastically perturbed by zero mean Gaussian noise. Let $\tilde{f}$ denote the perturbed fitness function value, while $\overline{f}$ denotes the average fitness function value. Obviously four situations may arise in the selection process: A $\{$ better $|$ worse$\}$ candidate can be $\{$ accepted $|$ rejected $\}$. This situation is shown in Fig. 1. The chance of accepting a good (respectively of rejecting a worse) candidate plays a central role in our investigations. In the next section, we shall discuss the details of the TS process.

3 Threshold Selection

3.1 Definitions

Threshold selection is a selection method, that can reduce the error probability of selecting a worse or rejecting a good candidate. Its general idea is relatively simple and already known in other contexts. Nagylaki states that a similar principle is very important in plant and animal breeding: *Accept a new candidate if its (noisy) fitness value is* significantly *better than*

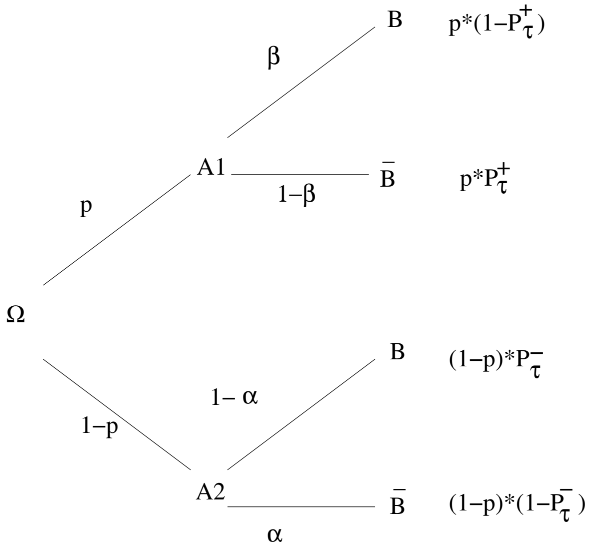

Figure 1: Decision tree visualizing the general situation of a selection process in uncertain environments. The events are labeled as follows: A_1: $f(Y) > f(X)$, A_2: $f(Y) \leq f(X)$, B: $\overline{f}(Y) < \overline{f}(X)$, and $\overline{B}$: $\overline{f}(Y) \geq \overline{f}(X)$.

that of the parent [14].

DEFINITION 1 (THRESHOLD ACCEPTANCE PROBABILITY) *Let $\overline{f}(X) := \sum_{i=1}^{n} \tilde{f}(X_i)/n$ be the sample average of the perturbed values, and f denote the unperturbed fitness function value. The conditional probability, that the fitness value of a better candidate Y is higher than the fitness value of the parent X by at least a threshold τ,*

$$P_\tau^+ := P\{\overline{f}(Y) > \overline{f}(X) + \tau \mid f(Y) > f(X)\}, \quad (1)$$

is called a threshold acceptance probability.

DEFINITION 2 (THRESHOLD REJECTION PROBABILITY) *The conditional probability, that a worse candidate Y has a lower noisy fitness value than the fitness value of parent X by at least a threshold τ,*

$$P_\tau^- := P\{\overline{f}(Y) \leq \overline{f}(X) + \tau \mid f(Y) \leq f(X)\}. \quad (2)$$

is called a threshold rejection probability.

The investigation of the requirements for the determination of an optimal threshold value reveals similarities between TS and hypothesis tests.

4 Hypothesis Tests

4.1 Hypothesis and Test Statistics

The determination of a threshold value can be interpreted in the context of hypothesis testing as the determination of a critical value. To formulate a statistical test, the question of interest is simplified into two competing hypotheses between which we have a choice: the null hypothesis, denoted H_0, is tested against the alternative hypothesis, denoted H_1. The

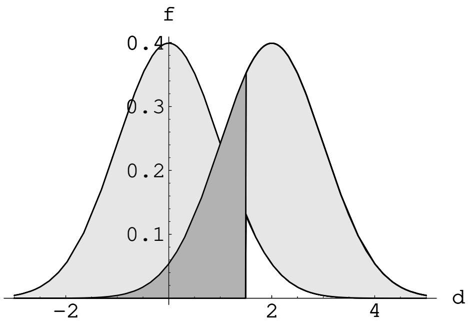

Figure 2: Error of the first (light region) and of the second kind (darkest region): *P.d.f. of two normal-distributed r.v. $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(2,1)$.*

decision is based on a quantity T calculated from a sample of data using a test function or test statistic. In the following we will use the r. v.

$$\overline{Z}_{m,n} := \frac{1}{n}\sum_{i=1}^{n} \tilde{f}(Y_{t,i}) - \frac{1}{m}\sum_{i=1}^{m} \tilde{f}(X_{t,i}) \quad (3)$$

as a test function. m and n define the number of samples taken from the parentX_t respectively offspring Y_t at time step t.

4.2 Critical Value and Error Probabilities

The *critical value* $c_{1-\alpha}$ for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected. We are seeking a value $c_{1-\alpha}$, such that

$$P\{T > c_{1-\alpha} \quad | H_0 \text{ true }\} \leq \alpha. \quad (4)$$

Making a decision under this circumstances may lead to two errors: an error of the first kind occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected with an error probability α. If the null hypothesis H_0 is not rejected when it is in fact false, an error of the second kind happens. β denotes the corresponding error probability.

5 Hypothesis Testing and Threshold Selection

5.1 The Relationship between α, β, and $P_\tau^\pm$

Let us consider the hypothesis H_0, that the fitness of the offspring is not better than the parental fitness. Furthermore we will use the test function defined in Eq. 3. Regarding selection in uncertain environments from the point of view of hypothesis tests, we obtain:

THEOREM 5.1
Suppose that $T = \overline{Z}_{m,n}$, $c_{1-\alpha} = \tau$, and $H_0 : f(Y_t) \leq f(X_t)$. Then we get: The conditional rejection probability

P_τ^- *and the error of the first kind are 'complementary' probabilities.*

$$P_\tau^- = P\{\overline{Z}_{m,n} > \tau \mid f(Y_t) \le f(X_t)\} = 1 - \alpha. \quad (5)$$

Proof This can be seen directly by combining Eq. 2 and Eq.4.

COROLLARY 5.2 (TO THEOREM 5.1)
The conditional acceptance probability P_τ^+ and the error of the second kind are 'complementary' probabilities:

$$P_\tau^+ = 1 - \beta. \quad (6)$$

5.2 Normal Distributed Noise

In the following, $\Phi(x)$ denotes the normal d.f., whereas z_α defines the (α)-quantile of the $\mathcal{N}(0,1)$-distribution: $\Phi(z_\alpha) = \alpha$, and t_α defines the (α)-quantile of the t-distribution. We are able to analyze TS with the means of hypothesis tests: Assuming stochastically independent samples $X_{t,i}$ and $Y_{t,i}$ of $\mathcal{N}(\mu_X, \sigma_X^2)$ respectively $\mathcal{N}(\mu_Y, \sigma_Y^2)$ distributed variables, we can determine the corresponding threshold τ for given error of the first kind α. The equation $P\{\overline{Z}_{m,n} \le \tau | H_0\} = 1 - \alpha$ leads to

$$\tau = z_{1-\alpha} \cdot \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}. \quad (7)$$

5.3 Unknown, but Equal Variances σ_X^2 and σ_Y^2

In many real world optimization problems, the variances are unknown, so that the test is based on empirical variances:
Let the r.v.

$$S_X^2 := \frac{1}{m-1}\sum_{i=1}^{m}(\tilde{f}(X_{t,i}) - \overline{f}(X_t))^2 \quad (8)$$

be the empirical variance of the sample. In this case, we obtain:

$$\tau = (t_{m+n-2;1-\alpha}) \cdot \sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{n+m-2}}\sqrt{\frac{m+n}{m \cdot n}}. \quad (9)$$

If the observations are paired (two corresponding programs are run with equal sample sizes ($m = n$) and with the same random numbers) and the variance is unknown, Eq. 9 reads:

$$\tau = \frac{t_{n-1;1-\alpha} \cdot s_d}{\sqrt{n}}, \quad (10)$$

with

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \overline{d})^2, \quad (11)$$

an estimate of the variance σ_d^2, and
$\overline{d} = \sum_{i=1}^{m}\left\{\tilde{f}(Y_{t,i}) - \tilde{f}(X_{t,i})\right\}/m$.

These results provide the basis for a detailed investigation of the TS mechanism. They can additionally be transferred to real-world optimization problems as shown in the following sections.

6 Applications

6.1 Example 1: A Simple Model of Stochastic Search in Uncertain Environments

In our first example, we analyze the influence of TS on the selection process in a simple stochastic search model. This model possesses many crucial features of real-world optimization problems, i. e. a small probability of generating a better offspring in an uncertain environment.

6.1.1 Model and Algorithm

DEFINITION 3 (SIMPLE STOCHASTIC SEARCH)
Suppose that the system to be optimized is at time t in one of the consecutive discrete states $X_t = i$, $i \in \mathbb{Z}$. In state i, we can probe the system to obtain a fitness value $\tilde{f}(X_t) = i \cdot \delta + U$. $\delta \in \mathbb{R}^+$ represents the distance between the expectation of the fitness values of two adjacent states. The random variable (r.v.) U possesses normal $\mathcal{N}(0, \sigma_\epsilon^2)$ distribution. The goal is to take the system to a final state $X_t = i$ with i as high as possible (maximization problem) in a given number of steps.

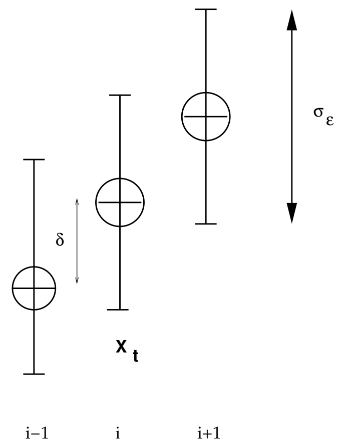

Figure 3: Simple stochastic search. Adjacent states.

Let us consider the following

ALGORITHM 1 (SIMPLE SEARCH WITH TS)
1. Initialize: *Initial state $X_{t=0} = 0$.*

2. Generate offspring: *At the t-th step, with current state $X_t = i$, flip a biased coin: Set the candidate of the new state Y_t to $i+1$ with probability p and to $i-1$ with probability $(1-p)$.*

3. Evaluate: *Draw samples (fitness values) from the current and the candidate states:*

$$\tilde{f}(X_{t,j}) \text{ and } \tilde{f}(Y_{t,k}), \quad (12)$$

with the measured fitness value $\tilde{f}(X) := f(X) + w$. w is the realization of a r.v., representing normal distributed noise, $W \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

4. Select: *Determine a threshold value τ. If $\overline{f}(Y_t) + \tau > \overline{f}(X_t)$, accept Y_t as the next state: $X_{t+1} := Y_t$; otherwise , keep the current state: $X_{t+1} := X_t$.*

5. Terminate: *If $t < t_{\max}$, increment t and go to step 2.*

REMARK 1
In this model, p is given; it is interpreted as the probability of generating a better candidate. In general, the experimenter has no control over p, which would be some small value for non-trivial optimization tasks.

THEOREM 6.1
ALGORITHM *1 can be represented by a Markov chain $\{X_t\}$ with the following properties:*

1. $X_0 = 0$.
2. $P\{X_{t+1} = i+1 | X_t = i\} = p \cdot P_\tau^+$
3. $P\{X_{t+1} = i-1 | X_t = i\} = (1-p) \cdot (1 - P_\tau^-)$
4. $P\{X_{t+1} = i | X_t = i\} = p \cdot (1 - P_\tau^+) + (1-p) \cdot P_\tau^-$, *with*

$$P_\tau^\pm := \Phi\left(\frac{\delta \mp \tau}{\sqrt{\frac{m+n}{mn}}\sigma_\epsilon}\right). \tag{13}$$

6.1.2 Search Rate and Optimal τ

The measurement of the local behavior of an EA can be based on the expected distance change in the object parameter space. This leads to the following definition:

DEFINITION 4 (SEARCH RATE)
Let R be the number of advance in the state number t in one step:

$$R := X_{t+1} - X_t. \tag{14}$$

The **search rate** *is defined as the expectation*

$$E[R(\delta, \sigma_\epsilon, p, t)], \tag{15}$$

to be abbreviated $E[R]$.

THEOREM 6.2
Let $E[R_\tau]$ be the search rate as defined in Eq. 15. Then Eq. 13 leads to

$$E[R_\tau] = p \cdot P_\tau^+ - (1-p) \cdot (1 - P_\tau^-). \tag{16}$$

COROLLARY 6.3 (TO THEOREM 6.2)
In this example (simple stochastic search model) it is possible to determine the optimal τ_{opt} value with regard to the search rate, if the fitness function is disturbed with normal-distributed noise:

$$\tau_{opt} = \frac{\sigma_\epsilon^2}{\delta} \log \frac{1-p}{p}. \tag{17}$$

p	τ_{opt}	$E[R_{\tau=0}]$	$E[R_{\tau_{opt}}]$
0.1	4.394	-0.262	0.00005
0.2	2.773	-0.162	0.003
0.3	1.695	-0.062 (D)	0.018 (C)
0.4	0.811	0.038 (B)	0.059 (A)
0.5	0.0	0.138	0.138

Table 1: Simple stochastic search. The noise level σ_ϵ equals 1.0, the distance δ is 0.5. Labels (A) to (D) refer to the results of the corresponding simulations shown in Fig. 4.

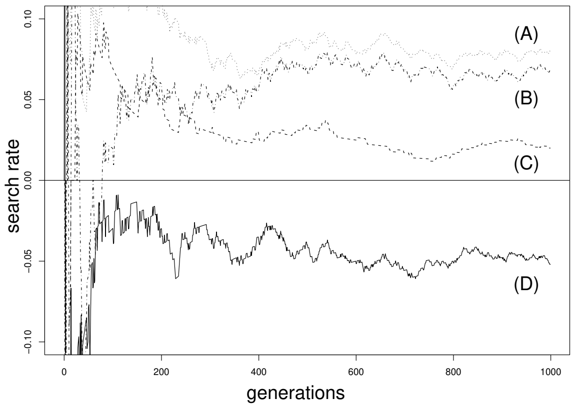

Figure 4: Simple stochastic search. Simulations performed to analyze the influence of TS on the search rate.

Assume there is a very small success probability p. Then the search can be misled, although the algorithm selects only 'better' candidates. We can conclude from Eq. 16, that a decreasing success probability ($p \searrow 0$) leads to a negative search rate. Based on Eq. 17, we calculated the optimal threshold value for 5 different success probabilities to illustrate the influence of TS on the search rate, cp. Tab 1. Corresponding values of the search rate are shown in the third column. TS can enhance the search rate and even avoid that the search rate becomes negative. This can be seen from the values in the last column.

Fig. 4 reveals that simulations lead to the same results. For two different p-values, the influence of TS on the search rate is shown. The search rate becomes *negative*, if p is set to 0.3 and no TS is used (D). The situation can be improved, if we introduce TS: The search rate becomes positive (C). A comparison of (A), where a zero threshold was used, and (B), where the optimal threshold value was used, shows that TS can improve an already positive search rate. These results are in correspondence with the theoretical results in Tab. 1.

6.2 Example 2: Application to the S-ring Model

6.2.1 The S-Ring as a Simplified Elevator Model

In the following we will analyze a '*S-ring model*', that is a simplified version of the elevator group control problem [12, 13]. The S-ring has only a few parameters: the number of elevator cars s_m, the number of customers s_n, and the passenger arrival rate s_λ. Therefore, the rules of operation are very simple, so that this model is easily reproducible and suitable for benchmark testing. However, there are important similarities with real elevator systems. The S-ring and real elevator systems are discrete-state stochastic dynamical systems, with high-dimensional state space and a very complex behavior. Both are found to show suboptimal performance when driven with simple 'greedy' policies. They exhibit a characteristic instability (commonly called 'bunching' in case of elevators). The policy π, that maps system states to decisions, was represented by a linear discriminator (perceptron) [13]. An EA was used to optimize the policy π.

6.2.2 DOE-Methodology

The analysis of many real-world optimization problems requires a different methodology than the analysis of the optimization of a fitness function f, because f remains unknown or can only be determined approximately. We use an approach that is similar to the concept discussed in [8]: From the complex real-world situation we proceed to a simulation model. In a second step we model the relationship between the inputs and outputs of this model through a regression model (meta-model). The analysis of the meta-model is based on DOE methods. Let the term *factor* denote a parameter

Factor	low value	medium value	high value
(A)Selection:	comma-strategy	plus-strategy	TS-strategy
(B)Selective Pressure:	4.0	6.0	9.0
(C)Population Size:	2.0	4	7.0

Table 2: EA–parameter and factorial designs

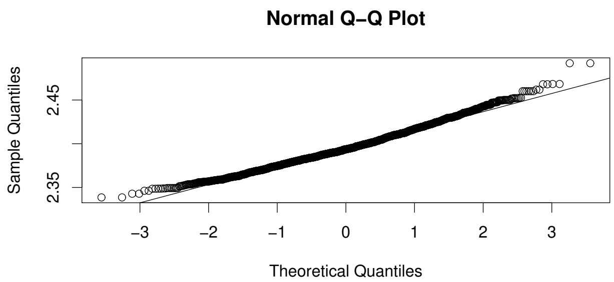

Figure 5: Diagnostic plot.

or input variable of our model. DOE methods can be defined as *selecting the combinations of factor levels that will be actually simulated when experimenting with the simulation model* [5, 4, 8, 9, 11, 10]. It seems reasonable to use DOE methods on account of the exponential growth in the number of factor levels as the number of factors grows. Based on these methods, we investigate the S-ring model[1]. The principal aim is to minimize the number of waiting customers, so we consider a minimization problem. A prototype S-ring

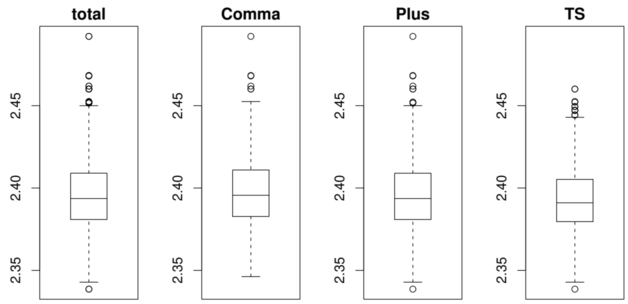

Figure 6: Box plot. Different selection schemes. Comma-selection, plus-selection, and TS, cp. Tab. 2.

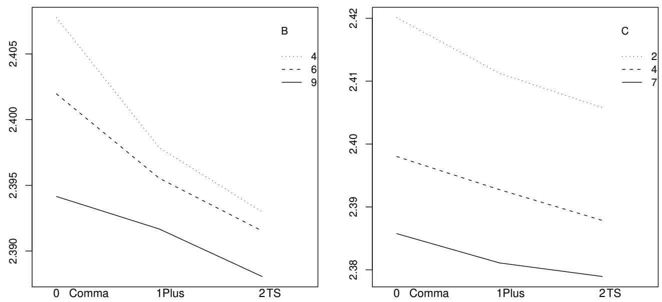

Figure 7: Plot of the means of the responses. The labels on the x-axis represent different selection mechanism. 0: Comma, 1: Plus, and 2: TS. B and C represent the selective strength $(4, 6, 9)$, resp. the population size $(2, 4, 7)$, cp. Tab. 2.

with the following parameter settings was used as a test case: customers $s_n = 6$, servers $s_m = 2$, and arrival rate $s_\lambda = 0.3$. The number of fitness function evaluations was set to 10^5, and every candidate was reevaluated 5 times. Eq. 10 was used to determine the threshold. The TS-scheme was compared to a comma-strategy and a plus-strategy. Global intermediate recombination was used in every simulation run. 50 experiments were performed for every ES-parameter setting. The population size and the selective pressure (defined as the ratio λ/μ) were varied. The corresponding settings are shown in Tab. 2.

6.2.3 Validation and Results

Before we are able to present the results of our simulations, the underlying simulation model has to be validated. The nor-

[1] The applicability of DOE methods to EAs is discussed in detail in [2].

mal Q–Q plot in Fig. 5 shows that the values are approximately standard normal. This is an important assumption for the applicability of the F-test, that was used in the regression analysis to determine the significance of the effects and of the interactions. Further statistical analysis reveals that the effects of the main factors are highly significant.

The results are visualized in two different ways. Box plots, shown in Fig. 6, give an excellent impression how the change of a factor influences the results. Comparing the comma-selection plot and the plus-selection plot to the TS selection plot, we can conclude that TS improves the result. In addition to the box plots, it may be also important to check for interaction effects (Fig. 7): Obviously TS performs better than the other selection methods.

7 Summary and Outlook

The connection between TS and hypothesis tests was shown. A formulae for the determination of the optimal τ value in a simple search model and a formulae for the determination of the threshold value for the error of the first kind α and the (estimated) variance s_d^2 were derived. Theoretical results were applied to a simplified elevator group control task problem. TS performs significantly better than other selection methods.

This work will be extended in the following way: To reduce the number of fitness function evaluations it might be sufficient to determine the noise level only at the beginning and after a certain number of time steps, instead of in every generation. Furthermore, we will investigate the situation shown in Fig. 1 from the viewpoint of Bayesian statistics.

Acknowledgments

T. Beielstein's research was supported by the DFG as a part of the collaborative research center 'Computational Intelligence' (531).

Bibliography

[1] D. V. Arnold. Evolution strategies in noisy environments — A survey of existing work. In L. Kallel, B. Naudts, and A. Rogers, editors, *Theoretical Aspects of Evolutionary Computing*, Natural Computing, pages 239–249. Springer, Berlin, 2001.

[2] T. Beielstein. Design of experiments and evolutionary algorithms. Part I: The classical model. Technical report, Universität Dortmund, Fachbereich Informatik, 2001. (to appear).

[3] H.-G. Beyer. Evolutionary algorithms in noisy environments: Theoretical issues and guidelines for practice. *CMAME (Computer methods in applied mechanics and engineering)*, 186:239–267, 2000.

[4] G. E. P. Box and N. R. Draper. *Experimental Model Building and Response Surfaces*. Wiley, 1987.

[5] G. E. P. Box, G. Hunter, William, and J. S. Hunter. *Statistics for experimenters*. Wiley series in probabilty and mathematical statistics: Applied probability and statistics. Wiley, 1978.

[6] J. Branke, C. Schmidt, and H. Schmeck. Efficient fitness estimation in noisy environments. In L. S. et al., editor, *Genetic and Evolutionary Computation Conference (GECCO'01)*. Morgan Kaufmann, 2001.

[7] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[8] J. Kleijnen. *Statistical Tools for Simulation Practitioners*. Marcel Dekker, New York, 1987.

[9] J. Kleijnen. Validation of models: Statistical techniques and data availability. In P. Farrington, H. Nembhard, D. Sturrock, and G. Evans, editors, *Proceedings of the 1999 Winter Simulation Conference*, pages 647–654, 1999.

[10] J. Kleijnen. Experimental designs for sensitivity analysis of simulation models. In A. H. et al, editor, *Proceedings of EUROSIM 2001*, 2001.

[11] A. M. Law and W. D. Kelton. *Simulation Modelling and Analysis*. McGraw-Hill Series in Industrial Egineering and Management Science. McGraw-Hill, New York, 3 edition, 2000.

[12] S. Markon. *Studies on Applications of Neural Networks in the Elevator System*. PhD thesis, Kyoto University, 1995.

[13] S. Markon, D. V. Arnold, T. Bäck, T. Beielstein, and H.-G. Beyer. Thresholding – a selection operator for noisy es. In J.-H. Kim, B.-T. Zhang, G. Fogel, and I. Kuscu, editors, *Proc. 2001 Congress on Evolutionary Computation (CEC'01)*, pages 465–472, Seoul, Korea, May 27–30, 2001. IEEE Press, Piscataway NJ.

[14] T. Nagylaki. *Introduction to Theoretical Population Genetics*. Springer, Berlin, Heidelberg, 1992.

[15] Y. Sano and H. Kita. Optimization of Noisy Fitness Functions by Means of Genetic Algorithms using History of Search. In M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. Merelo, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature (PPSN VI)*, volume 1917 of *LNCS*, pages 571–580. Springer, 2000.

[16] H.-P. Schwefel. *Evolution and Optimum Seeking*. Sixth-Generation Computer Technology. Wiley Interscience, New York, 1995.

[17] P. Stagge. Averaging efficiently in the presence of noise. In A.Eiben, editor, *Parallel Problem Solving from Nature, PPSN V*, pages 188–197, Berlin, 1998. Springer-Verlag.