Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Towards a steady-state analysis of an evolution strategy on a robust optimization problem with noise-induced multi-modality

Hans-Georg Beyer, Bernhard Sendhoff

2017

Preprint:

This is an accepted article published in IEEE Transactions on Evolutionary Computation. The final authenticated version is available online at: https://doi.org/10.1109/TEVC.2017.2668068

Toward a Steady-State Analysis of an Evolution Strategy on a Robust Optimization Problem With Noise-Induced Multimodality

Hans-Georg Beyer and Bernhard Sendhoff, *Senior Member, IEEE*

***Abstract*—A steady state analysis of the optimization quality of a classical self-adaptive evolution strategy (ES) on a class of robust optimization problems is presented. A novel technique for calculating progress rates for nonquadratic noisy fitness landscapes is presented. This technique yields asymptotically exact results in the infinite population size limit. This technique is applied to a class of functions with noise-induced multimodality. The resulting progress rate formulas are compared with high-precision experiments. The influence of fitness resampling is considered and the steady state behavior of the ES is derived and compared with simulations. The questions whether one should sample and average fitness values and how to choose the truncation ratio are discussed giving rise to further research perspectives.**

***Index Terms*—Evolution strategies (ESs), functions with noise-induced multimodality (FNIM), progress rate analysis, robust optimization.**

I. Introduction

A PRIME application domain of evolutionary algorithms (EAs) is the search for robust designs, a special type of optimization task where one seeks to find solutions $\hat{\mathbf{y}} \in \mathcal{Y}$ of high quality *and* which are also insensitive (with respect to quality) to environmental changes and/or fluctuations $\delta\mathbf{y}$ around the nominal design $\hat{\mathbf{y}}$ (see [1], [2] for comprehensive introductions and overviews). Assuming the environmental and design changes to be of stochastic nature, the optimization task becomes noisy. That is, the function $f(\mathbf{y})$ to be optimized becomes a conditional random variate $\tilde{f}|_{\mathbf{y}}$. For example, in the case of *actuator noise* $\delta\mathbf{y} = \mathbf{z}$ (a special case of systematic noise) we have $\tilde{f}|_{\mathbf{y}} = f(\mathbf{y} + \mathbf{z})$. However, optimization of $\tilde{f}|_{\mathbf{y}}$ (with respect to $\mathbf{y}$) is not well posed, one has to define the meaning of robustness. This is usually done by introducing some kind of utility functional measuring the quality of the noisy function $\tilde{f}$ (parameterized by $\mathbf{y}$). One might interpret such a procedure as a measurement of design robustness also referred to as *robustness measure*.

In context of evolutionary optimization expected fitness robustness has been considered by various authors [3]–[6]. Further robustness measures have been introduced in [7] in order to evaluate the performance of evolution strategies (ESs) on noisy optimization problems: the *threshold measures* and the *expected utility measure* (a generalization of the expected fitness approach). The latter poses a robust optimization problem into a maximization (or minimization) problem of the expected value of a utility function $U(\tilde{f}_{|\mathbf{y}})$ to be defined by the user. Most widely used is $U(\tilde{f}) = \tilde{f}$ yielding the *expected fitness measure*, thus the robust maximizer is defined as

$$\hat{\mathbf{y}} := \arg\max_{\mathbf{y}} \Big[\mathrm{E}\Big[\tilde{f}|\mathbf{y}\Big]\Big]. \tag{1}$$

While the introduction and use of special robustness measures depends on the specific design optimization problem, the question arises whether the optimization strategies used are able to *approximate* the robust maximizer $\hat{\mathbf{y}}$ arbitrarily well. This question is of special interest when using *direct* optimization strategies (so-called zeroth-order strategies) which directly use the *noisy* quality information to guide the search in $\mathcal{Y}$ [2], [12].

In the field of evolutionary computation, this question is usually tackled by empirical investigations using a set of test functions. While in ordinary optimization[1] generally accepted and well discussed test function suites exist, e.g., COCO platform (http://coco.gforge.inria.fr/), in robust optimization several authors have suggested test functions as well (see [4], [5], [8], [9]). However, it remains unclear to which degree these tests functions cover mathematically interesting as well as practically relevant robust optimization problems. Besides the use of special and general quadratic models with different kinds of noise, the class of *functions with noise-induced multimodality* (FNIM), gleaned from model considerations in optimal air-wing design, first proposed in [10], appears to be a challenging and promising test function class. Unlike other test functions proposed in [4], the FNIMs provides a class of problems:

1) where the robust optimizers $\hat{\mathbf{y}}$ as well as the modality of the optima depend on the noise strength ε,

Manuscript received October 11, 2016; revised January 17, 2017; accepted February 3, 2017. Date of publication February 13, 2017; date of current version July 27, 2017.

H.-G. Beyer is with the Research Center Process and Product Engineering, Vorarlberg University of Applied Sciences, A-6850 Dornbirn, Austria (e-mail: hans-georg.beyer@fhv.at).

B. Sendhoff is with Honda Research Institute Europe GmbH, 63073 Offenbach, Germany (e-mail: bernhard.sendhoff@honda-ri.de).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEVC.2017.2668068

[1] "Ordinary optimization" refers to the task of finding the optimizer $\mathbf{y}^*$ of a deterministic function $f(\mathbf{y})$, e.g., $\mathbf{y}^* = \arg\max_{\mathbf{y}} f(\mathbf{y})$. Note that in our nomenclature the term "optimum" refers to the function value $f(\mathbf{y}^*)$ at the optimizer $\mathbf{y}^*$.

1089-778X © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

exhibiting some kind of solution bifurcation controlled by ε;
2) which are *scalable* with respect to the problem dimension N (search space dimensionality);
3) which were derived as a model from a real-world problem describing the behavior observed in air-wing design, thus, being probably more related to practice then other test functions.

First theoretical as well as systematic empirical investigations of the performance of ESs on robust optimization problems presented in [9] and [10], respectively, suggested that the final solution quality of the strategies is a monotonous function of the offspring population size λ (assuming a constant truncation ratio $\vartheta = \mu/\lambda$, μ – parental population size). That is, there were theoretical as well as empirical evidences that by increasing λ, the approximation quality increases and in the asymptotic limit case the strategy approaches the robust optimizer $\hat{\mathbf{y}}$ defined by (1). This gave rise to attempts to design ES algorithms that are able to approximate the robust optimizer arbitrarily exact by successively increasing the population size.

A first ES algorithm designed for that task has been proposed in [11] and is displayed in Fig. 1. It is a standard $(\mu/\mu_I, \lambda)$-ES with σ-self-adaptation (σSA) enhanced with a population size control rule. After initialization, λ offspring are generated in lines 6–10. This is done by first mutating the parental mutation strength $\langle\sigma\rangle$ using a log-normal random multiplication. The resulting mutation strength $\tilde{\sigma}_l$ is then used to mutate the parental (centroid) individual $\langle\mathbf{y}\rangle$ by addition of an isotropic normally distributed random vector with component-wise standard deviation $\tilde{\sigma}_l$. The fitness of the resulting offspring $\tilde{\mathbf{y}}_l$ is determined in line 9. In line 12, the new parental mutation strength is calculated by averaging the σ values of the μ best offspring individuals. Intermediate multirecombination of the $\tilde{\mathbf{y}}$-vectors of the μ best offspring individuals (centroid calculation) is performed in line 13. In line 14, the average fitness $\langle F\rangle$ of the μ best offspring is calculated and used in line 15 to update $\overline{\Delta F}$ by exponential smoothing. $\overline{\Delta F}$ quantifies the average f-change between two consecutive generations. In the case of minimization, this quantity should be on average ≤ 0. If after an isolation period of $G = N$ (N – search space dimensionality) generations this minimization tendency is not observed (line 16), it is assumed that the ES has reached a steady state, i.e., the average fitness $\langle F\rangle$ does not improve further and the population size must be increased. This is done by multiplicatively increasing the parental population size μ in line 17 and the offspring population size λ is increased in line 18, respectively. The whole process is repeated until a termination criterion is fulfilled (e.g., number of function evaluations). The final $\langle\mathbf{y}\rangle$ serves as an approximation of $\hat{\mathbf{y}}$.

Given a fixed number of function evaluations, one can empirically evaluate the approximation quality of the ES in Fig. 1. On the other hand, by increasing the maximal number of function evaluations, one should expect to be able to approach the robust optimizer arbitrarily close. Investigations presented in [12] have shown, however, that this is not always the case. While most of the simpler test functions, such as the

$(\mu/\mu_I, \lambda)$-σSA-ES (Maximization)

1 $c_\mu = 4; \quad c_f = 1/N; \quad \tau = 1/\sqrt{N}; \quad G = N$
2 $\langle\mathbf{y}\rangle := \mathbf{y}^{(\text{init})}; \quad \langle\sigma\rangle := \sigma^{(\text{init})};$
3 $\mu := \mu^{(\text{init})}; \quad \lambda := \lceil\mu/\vartheta\rceil; \quad g := 0$
4 $\langle F\rangle^{(g)} := \frac{1}{\mu}\sum_{m=1}^{\mu} f_2\left(\mathbf{y}^{(\text{init})}\right); \quad \overline{\Delta F} := 0$
5 **Repeat**
6 **For** $l := 1$ **To** λ
7 $\tilde{\sigma}_l := \langle\sigma\rangle \exp[\tau\mathcal{N}_l(0,1)]$
8 $\tilde{\mathbf{y}}_l := \langle\mathbf{y}\rangle + \tilde{\sigma}_l\,\boldsymbol{\mathcal{N}}_l(\mathbf{0},\mathbf{I})$
9 $\tilde{f}_l := f_2(\tilde{\mathbf{y}}_l)$
10 **End For**
11 $g := g + 1$
12 $\langle\sigma\rangle := \frac{1}{\mu}\sum_{m=1}^{\mu}\tilde{\sigma}_{m;\lambda}$
13 $\langle\mathbf{y}\rangle := \frac{1}{\mu}\sum_{m=1}^{\mu}\tilde{\mathbf{y}}_{m;\lambda}$
14 $\langle F\rangle^{(g)} := \frac{1}{\mu}\sum_{m=1}^{\mu}\tilde{f}_{m;\lambda}$
15 $\overline{\Delta F} := (1-c_f)\overline{\Delta F} + c_f\left(\langle F\rangle^{(g)} - \langle F\rangle^{(g-1)}\right)$
16 **If** $(\text{mod}(g,G) = 0) \wedge (\overline{\Delta F} \leq 0)$ **Then**
17 $\mu := \lceil\mu c_\mu\rceil$
18 $\lambda := \lceil\mu/\vartheta\rceil$
19 **End If**
20 **Until** Termination_Condition

Fig. 1. Pseudocode of the $(\mu/\mu_I, \lambda)$-σSA-ES with population size control.

sphere model and the general quadratic model as well as a special class of FNIMs (denoted as "f_4" in [12]), can be optimized by the ES arbitrarily precise, there is a class of FNIMs where this is only possible for a specific truncation ratio. Since the truncation ratio is an exogenous strategy parameter, one cannot expect the ES to find the robust optimizer. A function that provides such a difficulty for the ES is ($\mathbf{y} \in R^N$)

$$f_2(\mathbf{y}) := -\frac{(y_{N-1}+\delta)^2 + \sum_{i=1}^{N-2} y_i^2}{y_N^2 + b} - y_N^2,\ \delta \sim \varepsilon\mathcal{N}(0,1) \quad (2)$$

where $b > 0$. Using probability theory and calculus, one can easily calculate the robust maximizer of f_2. One finds [8]

$$\left.\begin{array}{ll} \hat{\mathbf{y}} = \mathbf{0} & \text{for} \quad \varepsilon \leq b \\ \hat{\mathbf{y}} = \left(0,\ldots,0,\pm\sqrt{\varepsilon - b}\right)^T & \text{for} \quad \varepsilon > b \end{array}\right\}. \quad (3)$$

If one runs the self-adaptive $(\mu/\mu_I, \lambda)$-ES with population size control (as shown in Fig. 1) on this problem one observes the dynamics presented in Fig. 2. As one can see, the parental subspace distance r

$$r := \sqrt{\sum_{i=1}^{N-1}\langle y\rangle_i^2} \quad (4)$$

converges on average steadily toward the robust maximizer $\forall i = 1,\ldots,N-1 : \hat{y}_i = 0$. However, the Nth component y_N does *not* converge to $\hat{y}_N$ no matter how the population size λ is increased. There remains a steady state deviation of the Nth $\mathbf{y}$-component from the global maximizer value $\hat{y}_N = \sqrt{\varepsilon - b}$. An extensive parameter study in [12], reproduced in Fig. 3, showed that the parental steady state $\langle y_N\rangle$ value depends on

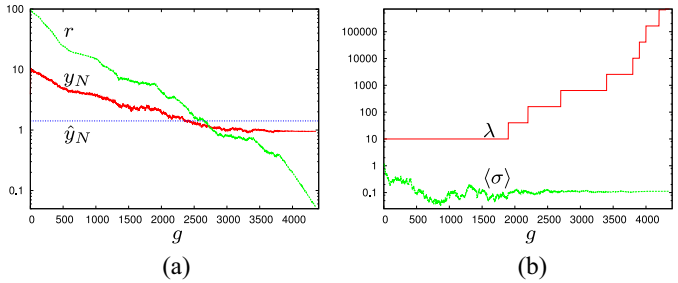


Fig. 2. Dynamics of a σSA-ES with population size control on f_2 (2). The function specific parameters are $b = 1$, $\varepsilon = 3$, and $N = 100$. Truncation ratio used: $\vartheta = 0.4$. (a) Horizontal dashed line represents the global maximizer state $\hat{y}_N$ of the object parameter y_N given by (3). The steady state y_N of the ES does not approximate the robust maximizer state $\hat{y}_N = \sqrt{\varepsilon - b}$. (b) Parental σ and offspring population size λ versus generations.

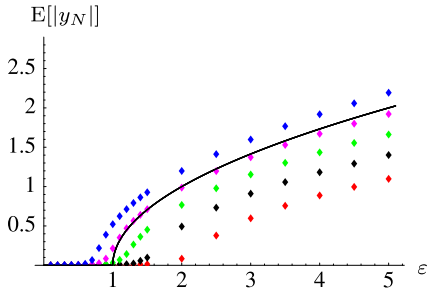


Fig. 3. Mean steady state y_N values of ES runs on test function f_2 with $N = 40$ and $b = 1.0$ depending on noise strength ε for ES with *different* truncation ratios ϑ. The data points presented are averages obtained from 20 independent runs each. The truncation ratios displayed are (from bottom to top) $\vartheta = 0.3$, $\vartheta = 0.4$, $\vartheta = 0.5$, $\vartheta = 0.6$, and $\vartheta = 0.7$ (data points are partially overlapping). The curve presents the y_N values of the global maximizer $\hat{y}_N$ as given by (3).

the truncation ratio $\vartheta := \mu/\lambda$ (μ – parental population size). That is, one can approximate the robust maximizer component $\hat{y}_N$ provided that one knows the right ϑ. For the f_2 parameters given in Fig. 3 and the noise strength $\varepsilon = 3$ the ES approximates the robust optimizer well for a truncation ratio of about $\vartheta = 0.61$.

The important message of those investigations is that given a noisy objective function, one cannot be sure whether the EA can locate the robust optimizer correctly. A stability analysis of the objective function f_2 at the nominal robust optimizer $\hat{\mathbf{y}}$ presented in [12] showed that in those cases $\hat{\mathbf{y}}$ is not a stationary point of f_2, i.e., the first derivatives of f_2 do not identically vanish at $\hat{\mathbf{y}}$. Furthermore, the resulting induced noisy fitness distribution gets skewed. As a result of these effects, initializing the ES at $\hat{\mathbf{y}}$, there is a tendency to depart from that state. On the other hand, choosing the right truncation ratio can counteract this tendency. Alternatively, one may resort to a numerical estimation of the expected value $E[f_2|\mathbf{y}]$ and apply a standard ES to this estimate. However, estimating $E[f_2|\mathbf{y}]$ numerically by averaging the noisy f-values for a fixed $\mathbf{y}$ candidate solution κ times

$$\langle f\rangle_\kappa(\mathbf{y}) := \frac{1}{\kappa}\sum_{k=1}^{\kappa} f_2(\mathbf{y}) \tag{5}$$

requires usually a very high number κ of samples to get sufficiently precise estimates of $E[f_2|\mathbf{y}]$.[2] This is a waste of computational resources. The question arises whether one could resort to small κ values in combination with the population sizes increase mechanism. That is, it should not be the aim of the averaging procedure (5) to get a nearly noise-free $E[f_2|\mathbf{y}]$ estimate, but to de-skew the resulting average resulting in a (nearly) symmetric $\langle f\rangle_\kappa$ distribution that can be more effectively treated by the σSA-ES.

Function evaluations for many applications, e.g., in shape or topology optimization, are prohibitively expensive. Therefore, from a practitioners point of view, the estimation of a minimal viable κ would be highly relevant. Even though the limit $\lambda \to \infty$ is practically equally infeasible, both from the theory as well as from many experiments, we have an understanding how the residual error scales with λ and therefore, we can qualitatively estimate the reliability of the achieved optimization results. For the FNIM functions discussed in this paper, up to now we are pretty much left in the dark besides some experimental analysis presented in [12]. The situation is worsened by the fact that the FNIMs have been abstracted from a concrete application problem as we mentioned before.

Besides estimating the impact of κ, it is equally important to find out whether the optimization problem we have to deal with in any application requires sampling or not, i.e., if convergence can be guaranteed for $\lambda \to \infty$ and whether the estimation of the residual error holds. We will come back to this point in the conclusion.

It is the aim of this paper to analyze the above given qualitative arguments on a more quantitative level. Ideally, the dynamical processes, depicted in Fig. 2, should be derived using the dynamical systems approach [13]. To this end, progress rates and the self-adaptation response function for f_2 (2) must be determined and the system of evolution equations must be solved. This turns out to be a rather ambitious program that is still in statu nascendi. In this paper we present first steps toward such an analysis and calculate the expected value of the steady state y_N for large population sizes λ. For this purpose we first have to derive an integral representation for the progress rate φ_{y_N} of the y_N component of the parental centroid and its asymptotic expansion for $\lambda \to \infty$. Using this expansion, an asymptotically exact approximation for φ_{y_N} will be derived and the predictive quality is compared with experiments. Using the approximate φ_{y_N}, the steady state y_N^{ss} is determined by the zero of φ_{y_N}. The predictions will be compared with experiments. In a next step, the averaging of f-values is included in the progress rate analysis and comparisons with experiments are presented concluding with a discussion.

II. Asymptotic Progress Rate Theory

A. General Considerations

As defined in [14], the progress rate is the expected change of parental quantities in the search space from generation g to $g+1$

$$\varphi_w := w^{(g)} - \mathrm{E}\left[w^{(g+1)}\middle|\mathbf{y}^{(g)}, \sigma^{(g)}\right] \tag{6}$$

[2]The standard deviation of the average decreases only with the factor $1/\sqrt{\kappa}$.

where w is an (in general) aggregated quantity of the parental centroid $\langle \mathbf{y} \rangle$ appearing in line 13 of Fig. 1. That is, the progress rate describes the expected change of the ES from generation g to $g+1$. The state of the ES at g is given by $\sigma := \langle \sigma \rangle^{(g)}$ and $\mathbf{y} := \langle \mathbf{y} \rangle^{(g)}$. Since we are interested basically in the steady state of the Nth parental centroid component $\langle y_N \rangle$, it suffices to consider φ_{y_N}

$$\varphi_{y_N} := y_N - \mathrm{E}\big[\langle \tilde{y}_N \rangle | \mathbf{y}, \sigma\big]. \tag{7}$$

The steady state is characterized by the zero of the progress rate

$$\varphi_{y_N}(\mathbf{y}, \sigma) \stackrel{!}{=} 0. \tag{8}$$

According to line 13 of Fig. 1, the new parental state is the centroid of the μ best offspring individuals $\tilde{\mathbf{y}}_l$. For $N \to \infty$ it holds $\tau \to 0$ (line 1 of Fig. 1) and one can approximate the offspring generation

$$\tilde{\mathbf{y}}_l = \mathbf{y} + \sigma \mathbf{z} \quad \text{where} \quad \mathbf{z} \sim \boldsymbol{\mathcal{N}}_l(\mathbf{0}, \mathbf{1}). \tag{9}$$

An offspring's Nth y-component is therefore generated as $\tilde{y}_N = y_N + \sigma z$, where $z \sim \mathcal{N}(0,1)$. If inserted into (7) one gets

$$\varphi_{y_N} = y_N - \mathrm{E}\big[\langle y_N + \sigma z \rangle\big] = -\sigma \mathrm{E}[\langle z \rangle] \tag{10}$$

where $E[\langle z \rangle]$ is the expected value[3] of the sum of μ standard normal variates those induced order statistics produced the μ largest f_2 values (maximization considered). In order to determine $E[\langle z \rangle]$ we will make use of the following theorem the proof of which is given in Appendix A.

Theorem 1 (Integral Representation of the Expectation of the Mean of Induced Order Statistics): Let z be a random variate with density $p_z(z)$ and support $z \in [z_l, z_u]$. Consider a second random variate $q \in [q_l, q_u]$ the density of which depends conditionally on z, i.e., $p_q(q|z)$ and the conditional distribution function $P_q(q|z) = \Pr(Q \le q|z)$. Produce λ samples Z_l from $p_z(z)$ and the corresponding λ samples Q_l from $p_q(q|Z_l)$. Define $z_{m;\lambda}$ as the random variable denoting the z-value associated with the mth greatest realization of Q. Consider the average $\langle z \rangle$ of these μ $z_{m;\lambda}$-variates

$$\langle z \rangle := \frac{1}{\mu} \sum_{m=1}^{\mu} z_{m;\lambda}. \tag{11}$$

Then the expected value of the mean is given by the integral

$$\mathrm{E}[\langle z \rangle] = \frac{\lambda!}{(\lambda - \mu - 1)!\mu!} \int_{z=z_l}^{z=z_u} z p_z(z) \times \int_{x=0}^{x=1} \Big[1 - P_q\Big(P_q^{-1}(x)|z\Big)\Big] x^{\lambda-\mu-1}(1-x)^{\mu-1} \,\mathrm{d}x \,\mathrm{d}z \tag{12}$$

[3]Note, for the sake of brevity the conditional state has been dropped in the expected value expressions.

where $P_q^{-1}(x)$ is the quantile function of the random variate q and the cumulative distribution function (cdf) of which is given by

$$P_q(q) = \int_{z=z_l}^{z=z_u} P_q(q|z) p_z(z) \,\mathrm{d}z. \tag{13}$$

The actual calculation of the integral (12) appears intractable. However, since reaching the steady state in robust optimization problems requires large population sizes λ [see Fig. 2(b)], one can consider the asymptotic case $\lambda \to \infty$. This will simplify the calculations. It holds (for the proof, see Appendix B)

Theorem 2 (Infinite Population Integral Representation): Consider the limit case $\lambda \to \infty$ of infinite population sizes such that the truncation ratio

$$\vartheta = \frac{\mu}{\lambda} = \text{const.} \quad \text{and} \quad 0 < \vartheta < 1. \tag{14}$$

Then an asymptotically exact expression for the expected value of the mean (11) of the induced order statistics is given by

$$\mathrm{E}[\langle z \rangle] \simeq \frac{1}{\vartheta} \int_{z=z_l}^{z=z_u} z p_z(z) \Big[1 - P_q\Big(P_q^{-1}(1-\vartheta)|z\Big)\Big] \,\mathrm{d}z. \tag{15}$$

The most demanding and next step in calculating φ_{y_N} concerns the approximation of (15). To this end, one needs to approximate $P_q(q|z)$ in such a manner that the integral in (15) can be solved in closed form. The density $p_z(z)$ is given by the probability density function (pdf) $\phi(z)$ of the $\mathcal{N}(0,1)$ standard normal variate

$$p_z(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}. \tag{16}$$

$P_q(q)$ is the cdf of the *offspring's* f_2-value

$$P_q(q) = \Pr\big[f_2(\mathbf{y} + \sigma \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{I})) \le q\big]. \tag{17}$$

$P_q(q|z)$ is the *conditional* cdf of the f_2 values keeping the Nth component of the mutation vector constant at σz, that is

$$P_q(q|z) = \Pr\Big[f_2\Big(\mathbf{y}^{(g)} + \sigma\big(\mathcal{N}(0,1), \ldots, \mathcal{N}(0,1), z\big)^T\Big) \le q\Big]. \tag{18}$$

Note, it holds

$$P_q(q) = \int_{-\infty}^{\infty} P_q(q|z) \phi(z) \,\mathrm{d}z. \tag{19}$$

Both $P_q(q)$ and $P_q(q|t)$ cannot be expressed in terms of closed analytical expressions and the same holds for (15). Therefore, asymptotically exact approximations will be derived for $N \to \infty$ and small mutation strengths $\sigma \to 0$. The approximations must be of such a kind that:

1) P_q can be inverted analytically in order to obtain the quantile function P_q^{-1};
2) $P_q(q|t)$ must provide an integrand $z\phi(z)P_q(q|z)$ that allows for a closed-form integration.

There seems to be only one nontrivial cdf that fulfills requirement 1). It is the cdf $\Phi(t)$ of the standard normal variate

$$\Phi(t) = \int_{-\infty}^{z=t} \phi(z) \,\mathrm{d}z. \tag{20}$$

As have been shown in [14], integrands containing functions of type $\Phi(\alpha+\beta t)$ can be integrated in closed form. Therefore, we will develop expansions that allow to take advantage of the identities [14, p. 330ff]

$$\int_{-\infty}^{\infty} \phi(t)\Phi(\alpha+\beta t)\,\mathrm{d}t = \Phi\left(\frac{\alpha}{\sqrt{1+\beta^2}}\right) \tag{21}$$

and

$$\int_{-\infty}^{\infty} t\phi(t)\Phi(\alpha+\beta t)\,\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\frac{\beta}{\sqrt{1+\beta^2}}\exp\left(-\frac{1}{2}\frac{\alpha^2}{1+\beta^2}\right). \tag{22}$$

B. Approximating the Probabilities

In a first step we will derive an asymptotically exact expression for the conditional cdf (18) needed in (15). To this end, we start with the conditional probability $\Pr[f_2(\mathbf{y}) \le q|\mathbf{y}]$ already determined in [9, p. 511, eq. (23)]

$$\begin{aligned}\Pr\big[f_2(\mathbf{y}) \le q|\mathbf{y}\big] &= 1 - \Pr\big[f_2(\mathbf{y}) > q|\mathbf{y}\big] \\ &= 1 - \Phi\left(\frac{-y_{N-1}+\sqrt{\xi}}{\varepsilon}\right) \\ &\quad + \Phi\left(\frac{-y_{N-1}-\sqrt{\xi}}{\varepsilon}\right)\end{aligned} \tag{23}$$

where

$$\xi \coloneqq -\Big(y_N^2+q\Big)\Big(y_N^2+b\Big) - u^2 \quad \text{and} \quad u^2 \coloneqq \sum_{i=1}^{N-2} y_i^2. \tag{24}$$

In order to get the fitness distribution (18) of the mutated state (offspring) $\tilde{\mathbf{y}}$ one has to consider (9), $\tilde{\mathbf{y}} = \mathbf{y} + \sigma\mathbf{z}$, keeping the Nth component constant, and integrate (23) using the respective densities of the random variates

$$\begin{aligned}P_q(q|z) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Pr\Big[f_2 < q|(y_1+\sigma z_1,\ldots, \\ &\qquad y_{N-1}+\sigma z_{N-1}, y_N+\sigma z)^T\Big]\phi(z_1)\ldots\phi(z_{N-1}) \\ &\qquad \mathrm{d}z_1\ldots\mathrm{d}z_{N-1} \\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} P_q(q|z_1,\ldots,z_{N-2},z) \\ &\qquad \times\ \phi(z_1)\ldots\phi(z_{N-2})\,\mathrm{d}z_1\ldots\mathrm{d}z_{N-2}\end{aligned} \tag{25}$$

where

$$\begin{aligned}&P_q(q|z_1,\ldots,z_{N-2},z) \\ &\quad\coloneqq \int_{-\infty}^{\infty} \Pr\big[f_2 < q|\big(y_1+\sigma z_1,\ldots, \\ &\qquad\qquad y_{N-1}+\sigma z_{N-1}, y_N+\sigma z)^T\big]\phi(z_{N-1})\,\mathrm{d}z_{N-1}.\end{aligned} \tag{26}$$

The integration with respect to z_{N-1} using (23) and (21) and taking $\Phi(-x) = 1-\Phi(x)$ into account yields immediately

$$\begin{aligned}&P_q(q|z_1,\ldots,z_{N-2},z) \\ &\quad= \int_{-\infty}^{\infty}\phi(z_{N-1})\left[1-\Phi\left(\frac{-y_{N-1}-\sigma z_{N-1}+\sqrt{\xi}}{\varepsilon}\right)\right. \\ &\qquad\left. +\ \Phi\left(\frac{-y_{N-1}-\sigma z_{N-1}-\sqrt{\xi}}{\varepsilon}\right)\right]\mathrm{d}z_{N-1} \\ &\quad= 2-\Phi\left(\frac{\sqrt{\xi}-y_{N-1}}{\sqrt{\varepsilon^2+\sigma^2}}\right)-\Phi\left(\frac{\sqrt{\xi}+y_{N-1}}{\sqrt{\varepsilon^2+\sigma^2}}\right).\end{aligned} \tag{27}$$

Equation (27) can be simplified provided that steady state conditions can be assumed. Considering the r-dynamics in Fig. 2(a) it becomes clear that with increasing generation numbers $r = \sqrt{\sum_{i=1}^{N-1} y_i^2} \to 0$ and therefore $|y_{N-1}| \to 0$ (in probability).[4] Under this assumption, the steady state fluctuations of y_{N-1} will be much smaller than the actuator noise δ. As one can easily show using (2) and (24) $|y_{N-1}+\delta| = \sqrt{\xi}$. Therefore

$$|y_{N-1}| \ll \sqrt{\xi} \tag{28}$$

holds almost surely. Assumption (28) is not immediately needed for the next calculations to come. However, in order to keep formulas as simple as possible, it will be applied to (27) resulting in

$$P_q(q|z_1,\ldots,z_{N-2},z) \simeq 2-2\Phi\left(\frac{\sqrt{\xi}}{\sqrt{\varepsilon^2+\sigma^2}}\right). \tag{29}$$

Now, we consider the integration over the first $N-2$ random normal variates $z_1,\ldots,z_{N-2}$ in (25). Having a closer look at ξ, defined in (24), one sees that the offspring's ξ in (25) is given by

$$\begin{aligned}\tilde{\xi} = &-\underbrace{\Big((y_N+\sigma z)^2+q\Big)\Big((y_N+\sigma z)^2+b\Big)}_{=:A_\xi} \\ &-\underbrace{\sum_{i=1}^{N-2}(y_i+\sigma z_i)^2}_{=:B_\xi}.\end{aligned} \tag{30}$$

That is, with respect to the first $N-2$ y-components, $\tilde{\xi}$ comprises a noncentral χ^2-distribution with $N-2$ degrees of freedom and a constant A_ξ. Therefore, the integral over the $N-2$ standard normal variates in (25) can be transformed into an 1-D integration using the χ^2_{N-2} density. While even this integration cannot be carried out analytically, the χ^2_{N-2} distribution approaches asymptotically normality. This opens the way for an asymptotically exact treatment of (25). Results derived in [7] yield immediately

$$\tilde{\xi} \simeq -A_\xi - \overline{B}_\xi + \sqrt{\mathrm{Var}\big[B_\xi\big]}\,\mathcal{N}(0,1) \tag{31}$$

[4]This is the typical behavior known from the sphere model. Keeping y_N constant, f_2 can be regarded as a sphere model with radius r and actuator noise $\delta \sim \varepsilon\mathcal{N}(0,1)$. Increasing λ in such a manner that the ES operates in the vicinity of the steady state, the ES will continuously reduce r.

where

$$\overline{B}_\xi = u^2 + (N-2)\sigma^2 \qquad \text{and} \tag{32}$$

$$\mathrm{Var}\big[B_\xi\big] = 4u^2\sigma^2 + 2(N-2)\sigma^4 =: D_\xi^2. \tag{33}$$

Thus, one gets *asymptotically*

$$\tilde{\xi} \simeq \underbrace{-A_\xi - u^2 - N\sigma^2}_{=:a_\xi} + \underbrace{2\sigma\sqrt{u^2 + \frac{N\sigma^2}{2}}}_{b_\xi} t,\ t \sim \mathcal{N}(0,1). \tag{34}$$

Furthermore, considering the variation coefficient

$$\frac{D_\xi}{\overline{B}_q} = \frac{2\sigma\sqrt{u^2 + (N-2)\sigma^2/2}}{u^2 + (N-2)\sigma^2} \le \frac{2\sigma\sqrt{u^2 + (N-2)\sigma^2}}{u^2 + (N-2)\sigma^2} = \frac{2\sigma}{\sqrt{u^2 + (N-2)\sigma^2}} \tag{35}$$

one sees that the fluctuations in (34) can be asymptotically neglected compared to the expected value of B_ξ as $N \to \infty$ or $\sigma \to 0$. We will take advantage of this result later on. For the time being, we use the property (35) to derive a first order approximation of $\sqrt{\tilde{\xi}}$ such that the resulting expression depends linearly on z. This is carried out with the intention to obtain an asymptotically exact integral instead of (25) which in turn can be integrated using (21). Since

$$\sqrt{a_\xi + b_\xi t} = \sqrt{a_\xi} + \frac{1}{2}\frac{b_\xi}{\sqrt{a_\xi}}t - \frac{1}{2}\frac{b_\xi^2}{a_\xi\sqrt{a_\xi}}t^2 + \dots \tag{36}$$

we get with $t \sim \mathcal{N}(0,1)$

$$\sqrt{\tilde{\xi}} \simeq \sqrt{-A_\xi - u^2 - N\sigma^2} + \frac{\sigma\sqrt{u^2 + N\sigma^2/2}}{\sqrt{-A_\xi - u^2 - N\sigma^2}}t \tag{37}$$

for the square root of (34). The newly obtained (asymptotically exact) approximation can now be used in (29) to determine (25). One gets

$$P_q(q|z) \simeq \int_{-\infty}^{\infty} 2\left[1 - \Phi\left(\frac{\sqrt{-A_\xi - u^2 - N\sigma^2}}{\sqrt{\varepsilon^2 + \sigma^2}} + \frac{\sigma\sqrt{u^2 + N\sigma^2/2}}{\sqrt{\varepsilon^2 + \sigma^2}\sqrt{-A_\xi - u^2 - N\sigma^2}}t\right)\right]\phi(t)\,\mathrm{d}t. \tag{38}$$

Using (21), we immediately obtain

$$P_q(q|z) \simeq 2 - 2\Phi\left(\frac{\sqrt{-A_\xi(q,z) - u^2 - N\sigma^2}}{\sqrt{\varepsilon^2 + \sigma^2 - \sigma^2\frac{u^2 + N\sigma^2/2}{A_\xi(q,z) + u^2 + N\sigma^2}}}\right) =: 2 - 2\Phi(G(q,z)). \tag{39}$$

Equation (39) could be used to derive $P_q(q)$ using (19) [the inverse of which is needed in (15)]. However, the resulting integral does not allow for a closed form solution. Therefore, the standard approach would be to perform Taylor expansion of the argument of Φ in (15) at $z = 0$. Breaking off after the linear term, one ends up with an integral of the form (21). However, the resulting $P_q(q)$ expression is not invertible with respect to q, thus, rendering such a calculation useless. The solution to this problem is to consider the small σ-case only. That is, the $\sigma\mathcal{N}(0,1)$ fluctuations about the parental y_N are assumed to be neglectable in A_ξ. This is especially correct if $\varepsilon \gg \sigma$, a behavior which is indeed observed in real ES runs [see Fig. 2(b)] in the vicinity of the steady state. Applying this assumption to (39), one ends up with a remarkably simple $P_q(q)$ expression

$$P_q(q) = 2 - 2\Phi\left(\frac{1}{\varepsilon}\sqrt{-\big(y_N^2 + q\big)\big(y_n^2 + b\big) - u^2 - N\sigma^2}\right). \tag{40}$$

For the calculation of the expected value integral in (15) the equation $1 - \vartheta = P_q(q)$ must be solved in order to get $q = P_q^{-1}(1-\vartheta)$. This can be easily done using (40)[5]

$$P_q^{-1}(1-\vartheta) = q \simeq -\frac{\varepsilon^2\Big[\Phi^{-1}\Big(\frac{1}{2} + \frac{\vartheta}{2}\Big)\Big]^2 + u^2 + N\sigma^2}{y_N^2 + b} - y_n^2 \tag{41}$$

where Φ^{-1} denotes the quantile function of the standard normal distribution.

C. Derivation of the φ_{y_N}-Approximation

In order to determine the integral in (15) the conditional cdf $P_q(q|z)$, given by (39), must be approximated such that the integration formula (22) can be applied. To this end, the argument function

$$G(q,z) = \frac{-A_\xi(q,z) - u^2 - N\sigma^2}{\sqrt{\sigma^2\big(u^2 + N\sigma^2/2\big) - \big(\varepsilon^2 + \sigma^2\big)\big(A_\xi(q,z) + u^2 + N\sigma^2\big)}} \tag{42}$$

being the argument in (39), must be approximated by a linear function of z

$$G(q,z) \simeq G_0(q) + G_1(q)z. \tag{43}$$

G_0 and G_1 are obtained by Taylor expansion of G at $z = 0$. Using (30) and (42) one obtains

$$G_0(q) = G(q,z)|_{z=0} = \frac{-\big(y_N^2 + q\big)\big(y_N^2 + b\big) - u^2 - N\sigma^2}{\sqrt{\sigma^2\Big(u^2 + \frac{N\sigma^2}{2}\Big) - \big(\varepsilon^2 + \sigma^2\big)\Big(\big(y_N^2 + q\big)\big(y_N^2 + b\big) + u^2 + N\sigma^2\Big)}} \tag{44}$$

and

$$G_1(q) = \left.\frac{\partial G(q,z)}{\partial z}\right|_{z=0} = \frac{1}{2}\left.\frac{\partial A_\xi(q,z)}{\partial z}\right|_{z=0} \times \frac{\big(\varepsilon^2 + \sigma^2\big)\big(A_\xi(q,0) + u^2 + N\sigma^2\big) - 2\sigma^2\Big(u^2 + \frac{N\sigma^2}{2}\Big)}{\Big[\sigma^2\Big(u^2 + \frac{N\sigma^2}{2}\Big) - \big(\varepsilon^2 + \sigma^2\big)\big(A_\xi(q,0) + u^2 + N\sigma^2\big)\Big]^{3/2}} \tag{45}$$

[5]The main reason for requiring (28) was to get rid of the unsymmetry in (27) that would not allow for an analytically closed solution of $1 - \vartheta = P_q(q)$.

where

$$A_\xi(q,0) = \left(y_N^2 + q\right)\left(y_N^2 + b\right) \tag{46}$$

and

$$\left.\frac{\partial A_\xi(q,z)}{\partial z}\right|_{z=0} = 2\sigma y_N\left(2y_N^2 + b + q\right). \tag{47}$$

Considering (30) and (15) with (16), one sees that a linear z approximation corresponds to the small σ limit. That is, the final φ_{y_N} formula will describe the $\sigma \to 0$ behavior asymptotically exact. Due to the symmetry of $\phi(z)$, the Taylor expansion of G has been developed at the maximum point of $\phi(z)$, i.e., at $z = 0$.

Using (39) with (43) and plugging this together with G_0 and G_1 into (15) one gets for the progress rate integral (10)

$$\varphi_{y_N} \simeq 2\frac{\sigma}{\vartheta}\int_{z=-\infty}^{z=\infty} z\,\phi(z)\left\{1 - \Phi\left[G_0\left(P_q^{-1}(1-\vartheta)\right) + G_1\left(P_q^{-1}(1-\vartheta)\right)z\right]\right\} \mathrm{d}z. \tag{48}$$

Applying the integral formula (22) one finally obtains

$$\varphi_{y_N} \simeq -\frac{\sigma}{\vartheta}\sqrt{\frac{2}{\pi}}\frac{G_1\left(P_q^{-1}(1-\vartheta)\right)}{\sqrt{1+\left[G_1\left(P_q^{-1}(1-\vartheta)\right)\right]^2}} \times \exp\left(-\frac{1}{2}\frac{\left[G_0\left(P_q^{-1}(1-\vartheta)\right)\right]^2}{1+\left[G_1\left(P_q^{-1}(1-\vartheta)\right)\right]^2}\right) \tag{49}$$

where

$$G_0\left(P_q^{-1}(1-\vartheta)\right) = \frac{\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2}{\sqrt{\sigma^2\left(u^2+\frac{N\sigma^2}{2}\right)+\left(\varepsilon^2+\sigma^2\right)\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2}} \tag{50}$$

and

$$G_1\left(P_q^{-1}(1-\vartheta)\right) = \frac{\sigma y_N}{y_N^2+b}\left[\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2 - \left(y_N^2+b\right)^2\right] \times \frac{2\sigma^2\left(u^2+\frac{N\sigma^2}{2}\right)+\left(\varepsilon^2+\sigma^2\right)\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2}{\left[\sigma^2\left(u^2+\frac{N\sigma^2}{2}\right)+\left(\varepsilon^2+\sigma^2\right)\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2\right]^{3/2}}. \tag{51}$$

D. Comparison With Simulations and Discussion

Equation (49) together with (50) and (51) is a rather complex progress rate expression derived under the small mutation strength assumption. That is, when using (49) as *approximation* for φ_{y_N}, one can expect a high predictive quality for small σ only. Similarly, this holds for the influence of the parameter space dimensionality N ($N \to \infty$) and the population size λ ($\lambda \to \infty$, $0 < \vartheta < 1$).

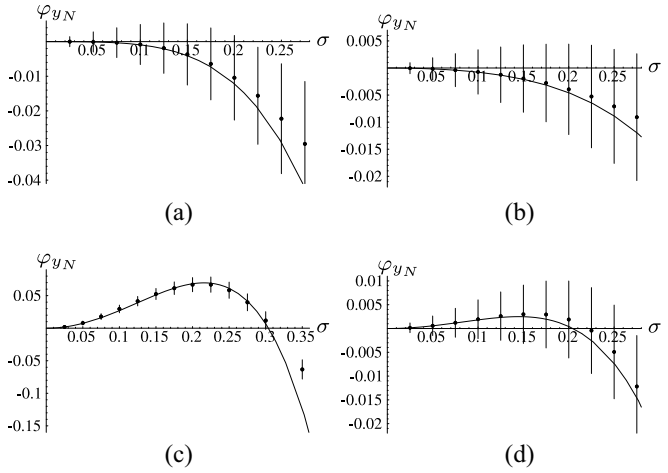


Fig. 4. Progress rate φ_{y_N} simulations versus asymptotic expression (49). Each data point (displayed as dot) has been obtained as average over 50,000 one-generation experiments simulating the progress rate definition (7) with initial $r = 0$. The vertical lines indicate ± 1 standard deviation of the outcome of the one-generation experiments. (a) $\vartheta = 0.3$, $y_N = 0.1$. (b) $\vartheta = 0.6$, $y_N = 0.1$. (c) $\vartheta = 0.3$, $y_N = 1.5$. (d) $\vartheta = 0.6$, $y_N = 1.5$.

The simulations presented in Fig. 4 (and in Appendix C-A of the supplementary material) are intended to show that (49) can yet be used to derive conclusions for $\sigma > 0$, $N < \infty$, and $\lambda < \infty$. In order to ensure the applicability of the integral representation (15) one has to consider large populations. However, this is not a serious restriction since robust optimization has to work with large population sizes λ in order to keep the optimizer approximation error small. The simulations have been done using $\lambda = 1000$ offspring and typical truncation ratios $\vartheta = 0.3$ and $\vartheta = 0.6$. The initial parental centroid $\mathbf{y}$ has been chosen in such a manner that its Nth component y_N is fixed in the experiments at $y_N = 0.1$ and 1.5, respectively, in Fig. 4 (and $y_N = 0.5, 1.0, 2.0, 2.5$ additionally in Appendix C-A of the supplementary material) in order to simulate different distances to the robust maximizer $\hat{y}_N$ (3). The remaining first $(N-1)$ y-components are randomly chosen on an $(N-1)$-dimensional hyper-sphere of radius r defined by (4), however, kept constant for the simulation of a (single) φ_{y_N} value. Further simulation parameters are: $b = 1$, $\varepsilon = 3$, $N = 100$, and $r = 0$ (for $r = 2$, see Appendix C-A of the supplementary material).

As one can see, the asymptotic φ_{y_N} formula (49) exhibits good predictive quality as long as the mutation strength σ is sufficiently small. Considering the definition (7) of the progress rate, it holds that in expectation the new parental state is given by $y_N^{(g+1)} = y_N^{(g)} - \varphi_{y_N}(\sigma^{(g)}, y_N^{(g)})$. Considering Fig. 4(a) and (b) one sees that y_N will be increased whereas Fig. 4(c) and (d) predicts a decrease of y_N. This makes perfect sense since in this example the robust maximizer (3) is at $\hat{y}_N = \sqrt{3-1} \approx 1.414$. In Fig. 4, one also notices a certain tendency that the *relative* strength of the fluctuations about the expected value φ_{y_N} gets smaller when the absolute value of the parental y_N is increased. Actually, it is maximal for $y_N = 0$ (no simulations shown for this setting). In that case φ_{y_N} is identically zero. That is, the ES has no preferred evolution direction for y_N. As a result, the evolution can in principle proceed in the positive or negative y_N-direction.

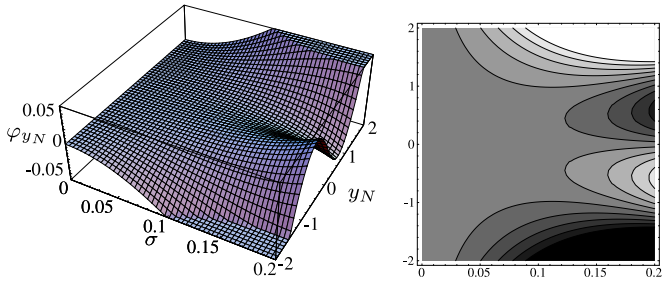


Fig. 5. Progress rate landscape depending on mutation strength σ and the parental y_N state (the other $N-1$ parental components are chosen to be zero, i.e., $r=0$) for $\varepsilon=3$, $b=1$, $\vartheta=0.3$, and $N=100$. Note, $\varphi_{y_N}(\sigma, y_N)$ values have been truncated for $\varphi_{y_N} > 0.05$ and $\varphi_{y_N} < -0.05$ in the 3-D-plot.

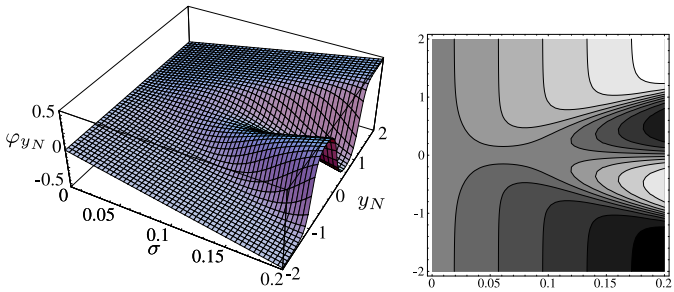


Fig. 6. Progress rate landscape depending on mutation strength σ and the parental y_N state (the other $N-1$ parental components are chosen to be zero, i.e., $r=0$) for $\varepsilon=0.2$, $b=1$, $\vartheta=0.3$, and $N=100$.

Whether a small fluctuation derives the ES away from the $y_N = 0$ line or not can be determined by closer inspection of the $\varphi_{y_N}(\sigma, y_N)$ landscape given by (49). Fig. 5 shows the behavior for $\varepsilon = 3$ given $b = 1$ and the parental state $r = 0$. According to (3), the Nth component of the global maximizer is $\pm\sqrt{\varepsilon - b} = \pm\sqrt{2} \neq 0$. Consider, e.g., $\sigma = 0.2$. While $\varphi|_{y_N=0} = 0$, we see that small deviations δ of the parental state $y_N = 0$ result in $\varphi|_{y_N=\delta} \neq 0$. Since by definition (7) positive φ_{y_N} causes a $y_N^{(g+1)}$ decrease and negative φ_{y_N} results in a $y_N^{(g+1)}$ increase, one clearly sees that the parental $y_N = 0$ state is *instable*. That is, the ES will leave this state during the evolution.

Magnifying the φ_{y_N}-landscape in Fig. 5 in the vicinity of $\sigma = 0$ one can easily verify that this behavior also holds for arbitrarily small mutation strengths $\sigma > 0$. As a result, y_N changes over the generations finally reaching a stationary point y_N^{ss} where $\varphi(y_N^{ss}) = 0$. As one can infer from the 3-D-plot, this new state is *stable*, because negative deviations δ from y_N^{ss} cause negative $\varphi(y_N^{ss} + \delta)$ values, thus, resulting in a y_N increase counteracting the δ fluctuations. Conversely, positive deviations δ from y_N^{ss} cause positive $\varphi(y_N^{ss} + \delta)$, thus, resulting in a y_N decrease.

Unlike the case $\varepsilon > b$, considered in Fig. 5, case $\varepsilon = 0.2$, i.e., $\varepsilon < b$, displayed in Fig. 6, represents a somewhat more complex situation. For $0 < \sigma \leq \sigma_1$ the progress rate $\varphi_{y_N}(y_N)$ has only one zero. For $\sigma > \sigma_1$, one again observes three zeros as in the case of Fig. 5 (where this holds for all $\sigma > 0$). This has consequences for the steady state behavior of the ES. Recall that for $\varepsilon < b$ the robust maximizer $\hat{y}_N$ is at $y_N = 0$ [see (3)]. Consider the case $0 < \sigma \leq \sigma_1$ first. If an ES is in a parental state $y_N < 0$ then the progress rate is negative and the expected parental state at the next generation is increased. Conversely, if $y_N > 0$ then the progress rate is positive and the expected parental state is decreased. That is, the global robust maximizer state $\hat{y}_N = 0$ is a stable attractor of the mean value dynamics of the ES. This does not longer hold if $\sigma > \sigma_1$. In that case we have three zeros of $\varphi(y_N)$ similar to the situation $\varepsilon > b$ displayed in Fig. 5. That is, the ES dynamics has two stable attractors $y_N^{ss} \neq 0$. Put it another way, since we are already considering the infinite population limit, we can conclude that for mutation strengths $\sigma > \sigma_1$ the ES is not able to approximate the robust maximizer $\hat{\mathbf{y}}$ arbitrarily close for test function f_2.

The critical σ_1 for which the ES evolves to a wrong steady state value can be calculated analytically. To this end, we consider the behavior of φ_{y_N} in Fig. 6 again. Provided that $\sigma < \sigma_1$, the function $\varphi_{y_N}(y_N)$ is a strictly monotonous increasing function of y_N, i.e., $\forall y_N : (\partial/\partial y_N)\varphi_{y_N}(y_N) > 0$. However, if $\sigma > \sigma_1$ the derivative of $\varphi_{y_N}(y_N)$ at $y_N = 0$ must necessarily change the sign because $\varphi_{y_N}(y_N)$ has a local maximum and a local minimum resulting in a negative derivative at $y_N = 0$. Therefore, for the case $\sigma = \sigma_1$ the derivative must vanish at $y_N = 0$

$$\left.\frac{\partial \varphi_{y_N}(\sigma, y_N)}{\partial y_N}\right|_{y_N=0} = 0 \quad \Longleftrightarrow \quad \sigma = \sigma_1. \tag{52}$$

After a long but straightforward calculation, one finds for the partial derivative of (49) with respect to y_N at $y_N = 0$

$$\begin{aligned}
&\left.\frac{\partial \varphi_{y_N}(\sigma, y_N)}{\partial y_N}\right|_{y_N=0} \\
&= \frac{\sigma^2}{\vartheta b}\sqrt{\frac{2}{\pi}}\left[b^2 - u^2 - N\sigma^2 - \varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2\right] \\
&\quad \times \frac{2\sigma^2\left(u^2 + \frac{N\sigma^2}{2}\right) + \left(\varepsilon^2+\sigma^2\right)\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2}{\left[\sigma^2\left(u^2 + \frac{N\sigma^2}{2}\right) + \left(\varepsilon^2+\sigma^2\right)\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2\right]^{3/2}} \\
&\quad \times \exp\left[-\frac{1}{2}\frac{\varepsilon^4\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^4}{\sigma^2\left(u^2 + \frac{N\sigma^2}{2}\right) + \left(\varepsilon^2+\sigma^2\right)\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2}\right].
\end{aligned} \tag{53}$$

Applying (52) to (49), using (50) and (51), one sees that a nontrivial zero of (52) requires that the outer bracket in the first line of (53) vanishes. Thus, we immediately obtain

$$\sigma_1 = \frac{1}{\sqrt{N}}\sqrt{b^2 - u^2 - \varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2}. \tag{54}$$

Formula (54) allows for a discussion of the qualitative behavior of the ES in the $\varepsilon < b$ regime (i.e., where the global maximizer is at $\mathbf{0}$). From the experimental observation we know that u, defined in (24), can in principle be reduced arbitrarily.[6]

[6]In order to show this theoretically, φ_u as defined by (6) must be calculated and the respective steady state equations must be solved. This remains as a task for future research.

Therefore, it suffices to consider the influence of the remaining terms in (54). The parameter space dimension N reduces σ_1. Assuming that the ES approaches a constant (self-adapted) σ [see Fig. 2(b)] this might cause problems for larger N. It quests for the design of (self-) adaptive ES that reduce the mutation strength σ when approaching the global optimizer. On the other hand one can play around with the truncation ratio ϑ. Since $\Phi^{-1}(1/2) = 0$, one can reduce the influence of ε by choosing a low truncation ratio. However, there still remains the upper bound $b/\sqrt{N}$ for σ_1 that cannot be increased further.

E. Steady State Behavior

In the progress rate plots, Fig. 5, of Section II-D we have seen that for $\varepsilon > b$ the ES approaches a steady state $y_N^{ss} \neq 0$. From the experiments we know that this steady state usually deviates from the global maximizer $\hat{y}_N = \sqrt{\varepsilon - b}$ (3). The question arises whether one can calculate y_N^{ss}. This can indeed be done using (49)–(51). To this end, the zeros of $\varphi_{y_N}(y_N)$ with respect to y_N must be determined. Since the exponential function in (49) is greater than zero and the same holds for the square root in the denominator of (49), finding the zeros reduces to finding the zeros of $G_1(P_q^{-1}(1-\vartheta))$, defined by (51). This immediately yields the trivial solution

$$y_{N_0}^{ss} = 0 \tag{55}$$

and additionally the equation

$$\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2 - \left(y_N^2 + b\right)^2 \stackrel{!}{=} 0 \tag{56}$$

to be solved for y_N. One gets

$$y_{N_{1,2}}^{ss} = \pm\sqrt{\sqrt{\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2} - b}. \tag{57}$$

The question arises which of the solutions y_N^{ss} are actually realized by the ES. Having a closer look at (57), one sees that the outer square root does not provide a real-valued solution if

$$\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2 < b^2. \tag{58}$$

The limit case of equality corresponds to $y_N^{ss} = 0$. This is consonant with condition (54) for the critical σ value σ_1 corresponding to vanishing y_N-derivative (53). That is, if (58) is fulfilled, the y_N-derivative (53) is positive, and $y_N = 0$ is the stable steady state attractor. Otherwise, the positive or negative root in (57) is realized as the stable steady state attractor of the ES, thus, we finally get

$$y_N^{ss} = \begin{cases} 0, & \text{for} \quad b^2 > \varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2 \\ \pm\sqrt{\sqrt{\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2} - b}, & \text{else.} \end{cases} \tag{59}$$

Whether the positive or negative root is realized in an actual ES runs is clearly a random decision that is strongly conditioned by the initial parents chosen.

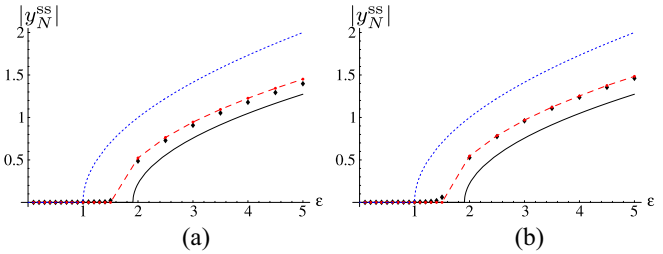
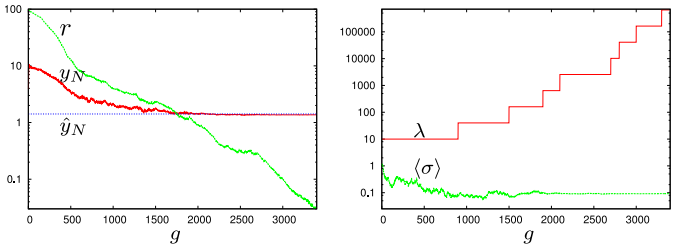


Fig. 7. Steady state y_N-values (displayed as black diamonds) of $(\lfloor\vartheta\lambda\rfloor/\lfloor\vartheta\lambda\rfloor_I, \lambda)$-ES ($\lambda = 10,000$, $\vartheta = 0.4$) on f_2 with (a) $b = 1, N = 40$ and (b) $b = 1, N = 100$ depending on the noise strength ε. The dots connected by the dashed lines are the data points predicted by (59). The bottom right curve is obtained from (59) using $\sigma = 0$ and $u = 0$. The dotted top curve displays the global maximizer $\hat{y}_N$ as given by (3). Each data point was obtained as average over (a) eight independent runs with 20 000 000 and (b) four independent runs with 30 000 000 function evaluations per run. Within a single run, the last 500 generations have been used to calculate the expected steady state value of the specific run. The observed standard deviations of the averaged data points are smaller than the size of the black diamonds.

Note, (59) contains the two dynamical quantities u^2 and σ^2, which cannot be predicted by the theory developed so far. However, in order to evaluate its approximation quality, one can plug the measured average u and average σ into (59). Fig. 7 shows $|y_N^{ss}|$ values derived from results of real ES runs using $(\lfloor\vartheta\lambda\rfloor/\lfloor\vartheta\lambda\rfloor_I, \lambda)$-ES with a final $\lambda = 10\,000$. As one can see, the asymptotic formula (59) predicts the real ES behavior well (for further simulations, see Appendix C-B of the supplementary material). Due to the asymptotic approach, the approximation quality increases with the parameter space dimension N as expected. The deviation of the prediction (data points connected by dashed curve) from the measured steady state values (black diamonds) increases slightly with increasing ε. Also, it might be seen that there is a deviation in the vicinity of the bifurcation point (the ε value where $|y_N^{ss}|$ starts to get greater than zero) in the plot in Fig. 7(b). The latter deviation is due to a simulation time chosen too small. Reaching the steady state regime in the vicinity of the bifurcation point appears to take a considerably long time compared to the other ε regimes. This has consequences for real ES runs in so far as one cannot stop the search for the robust optimizer with a simple stopping rule. Instead one has to observe the ES dynamics thoroughly.

Apart from the comparison of the observed steady state values (black diamonds) with the robust maximizer function (3), displayed as dotted curve, Fig. 7 shows also the asymptotic steady state curve when $u \to 0$ and $N\sigma^2 \to 0$, which represents the ideal case. As one can see, there are considerable deviations to the dashed curve predicted by theory. These deviations are basically due to the nonvanishing mutation strength σ of the real self-adaptive ES.

From the ES dynamics (see Fig. 2) we know that $u \to 0$ if the population size increases steadily, whereas the mutation strength σ approaches a constant steady state value. However, even if σ were zero, comparing (59) and (3) we see that in general $y_N^{ss} \neq \hat{y}_N$. This has been the puzzling observation showing that the ES is not able to approximate the robust maximizer $\hat{y}_N$ arbitrarily exact as the population size λ goes to infinity. We can now discuss this behavior in detail using our predictive



(a) (b)

Fig. 8. Dynamics of the evolution of the σSA-ES on f_2 with resampling of size $\kappa = 10$ (total number of function evaluations 1 200 000 000). (a) Horizontal dashed line represents the global maximizer state $\hat{y}_N$ of the object parameter y_N given by (3). (b) Parental σ and offspring population size λ versus generations. Unlike the same optimization problem presented in Fig. 2, the resampling version gets close to the robust maximizer component $\hat{y}_N \approx 1.414$.

formula (59). As we have also seen in Fig. 3, considering a *fixed* noise strength ε, one can make the ES approximate the robust optimizer by choosing the *correct* truncation ratio $\hat{\vartheta}$. This can also be derived from (59) using (3)

$$\hat{y}_N = y_N^{ss} \Longleftrightarrow \varepsilon \stackrel{!}{=} \sqrt{\varepsilon^2\left[\Phi^{-1}\left(\frac{1}{2}+\frac{\vartheta}{2}\right)\right]^2 + u^2 + N\sigma^2}$$

$$\Longleftrightarrow \vartheta \stackrel{!}{=} 2\Phi\left(1-\frac{u^2+N\sigma^2}{\varepsilon^2}\right) - 1 =: \hat{\vartheta}. \quad (60)$$

That is, depending on u and σ one can calculate the truncation ratio $\hat{\vartheta}$ which guarantees (within the limits of the approximation) that the (expected) steady state y_N approximates the robust maximizer $\hat{y}_N$ well. In the ideal case of asymptotically vanishing u and $\sigma \to 0$, one even obtains a fixed truncation ratio

$$\hat{\vartheta} = 2\Phi(1) - 1 \approx 0.6827 \quad (61)$$

valid for arbitrary f_2 parameters. That is, provided that the conditions used to derive the progress rate expression (49) are fulfilled, in order to get close to the robust maximizer, roughly 2/3 of the offspring should survive as parents for the next generation. In order to get (61) from (60), we had also to assume that the mutation strength σ is sufficiently small. This is a somewhat conflicting goal when considering ES as a global searcher in multimodal fitness landscapes. Also, up to now we neither have a theory to predict the actually observed σ values nor do we have a practically applicable rule for controlling the mutation strength if need be. Both issues should be addressed in future research.

III. Resampling

A. Motivation

In the previous section, it has been shown that the expected steady state y_N can approximate the robust optimizer if one is able to choose the right truncation ratio ϑ. On the other hand, it has been shown empirically in [12] that an ES with moderate resampling (5) can approach the robust optimizer even for small sample sizes κ. Fig. 8 shows such an experiment using $\kappa = 10$. Since using an unnecessarily large resampling size is a waste of computational resources, κ should be as small as possible. To approach this topic on a quantitative level, the progress rate theory developed so far will be extended to include the resampling.

B. Progress Rate Theory for y_N With Resampling

In this section, we will sketch the derivation of the progress rate formula. The derivations do only slightly deviate from those presented in Sections II-A–II-C. Therefore, we present those derivation details that are different. Taking the symmetry of $p_z(z) = \phi(z)$ into account, the asymptotic progress rate integral is given by (10) and (15)

$$\varphi_{y_{N_\kappa}} \simeq \frac{\sigma}{\vartheta}\int_{t=-\infty}^{t=\infty} t\,\phi(t)\,P_{q_\kappa}\left(P_{q_\kappa}^{-1}(1-\vartheta)|t\right)\mathrm{d}t. \quad (62)$$

Here P_q is to be replaced by the respective cdf P_{q_κ} that takes the resampling into account. This concerns the expressions (23), (39), and (40). The details of the calculations are presented in Appendix B of the supplementary material. One obtains the progress rate formula

$$\varphi_{y_{N_\kappa}} \simeq -\frac{\sigma}{\vartheta}\frac{1}{\sqrt{2\pi}}\frac{G_{\kappa 1}\left(P_{q_\kappa}^{-1}(1-\vartheta)\right)}{\sqrt{1+\left[G_{\kappa 1}\left(P_{q_\kappa}^{-1}(1-\vartheta)\right)\right]^2}} \times \exp\left(-\frac{1}{2}\frac{\left[G_{\kappa 0}\left(P_{q_\kappa}^{-1}(1-\vartheta)\right)\right]^2}{1+\left[G_{\kappa 1}\left(P_{q_\kappa}^{-1}(1-\vartheta)\right)\right]^2}\right) \quad (63)$$

where

$$G_{\kappa 0}\left(P_{q_\kappa}^{-1}(1-\vartheta)\right) \simeq \frac{\Phi^{-1}(\vartheta)\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]}{\sqrt{\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]^2+\left(\frac{2\kappa\sigma}{\varepsilon^2}\right)^2\left(u^2+\frac{N\sigma^2}{2}\right)}} \quad (64)$$

the derivation of which can be found in Appendix B of the supplementary material. There one can also find the derivation of $G_{\kappa 1}$ [being (106) and (107) in Appendix B of the supplementary material]

$$G_{\kappa 1}\left(P_{q_\kappa}^{-1}(1-\vartheta)\right) \simeq \frac{2\kappa\sigma y_N}{\varepsilon^2}\left[y_N^2+b-\frac{\frac{\varepsilon^2}{2\kappa}\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]^2+u^2+N\sigma^2}{y_N^2+b}\right] \times \left[\frac{\Phi^{-1}(\vartheta)\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]}{\left(\sqrt{\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]^2+\left(\frac{2\kappa\sigma}{\varepsilon^2}\right)^2\left(u^2+\frac{N\sigma^2}{2}\right)}\right)^3} - \frac{\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]^{-1}\left[2\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]}{\sqrt{\left[\Phi^{-1}(\vartheta)+\sqrt{2\kappa-1}\right]^2+\left(\frac{2\kappa\sigma}{\varepsilon^2}\right)^2\left(u^2+\frac{N\sigma^2}{2}\right)}}\right]. \quad (65)$$

In order to check the approximation quality, extensive one-generation experiments have been conducted that have shown that the $\varphi_{y_{N_\kappa}}$ (63) works well at least for sufficiently small σ-values. Four examples are displayed in Fig. 9.

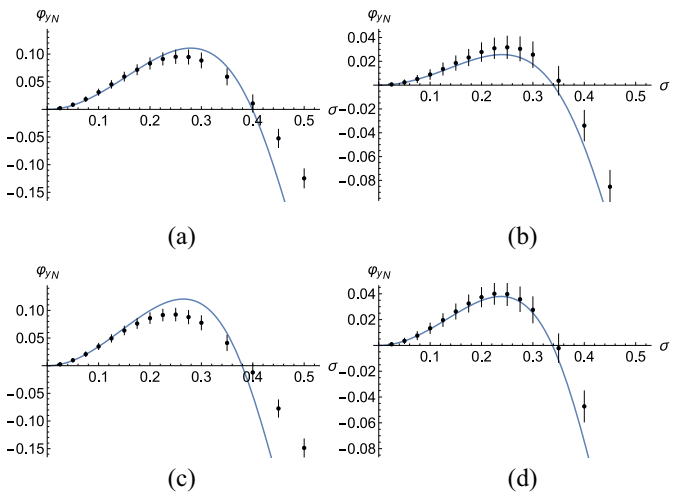


Fig. 9. Resampling progress rate $\varphi_{y_{N_\kappa}}$ simulations versus asymptotic expression (63). Each data point (displayed as dot) has been obtained as average over 50 000 one-generation experiments simulating the progress rate definition (7) with initial $u = 2.0$ and $y_N = 2.0$. The vertical lines indicate ± 1 standard deviation of the outcome of the one-generation experiments. (a) $\vartheta = 0.3, \kappa = 4$. (b) $\vartheta = 0.6, \kappa = 4$. (c) $\vartheta = 0.3, \kappa = 10$. (d) $\vartheta = 0.6, \kappa = 10$.

C. Steady State Behavior

While the progress rate (63) is a rather complex expression, it can be easily used to investigate the steady state behavior of the ES with resampling. Since the steady state $y^{ss}_{N_\kappa}$ is characterized by (expected) zero progress, the condition

$$\varphi_{y_{N_\kappa}}\left(y^{ss}_{N_\kappa}\right) = 0 \tag{66}$$

defines the steady state y_N to be obtained by resolving (66) for $y^{ss}_{N_\kappa}$. Considering (63), one sees that the zero of $\varphi_{y_{N_\kappa}}$ is determined by $G_{\kappa 1}(P^{-1}_{q_\kappa}(1-\vartheta))$. Since $G_{\kappa 1}(P^{-1}_{q_\kappa}(1-\vartheta))$ is given by (65) one sees that the zero (with respect to y_N) depends on the vanishing of the first line in (65) only [because the expressions in the second and third line of (65) do not depend on y_N]. Thus one gets

$$y_N\left[y_N^2 + b - \frac{\frac{\varepsilon^2}{2\kappa}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]^2 + u^2 + N\sigma^2}{y_N^2 + b}\right] \stackrel{!}{=} 0. \tag{67}$$

This leads to the trivial solution $y^{ss}_{N_\kappa} = 0$ and the nontrivial condition

$$y_N^2 + b - \frac{\frac{\varepsilon^2}{2\kappa}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]^2 + u^2 + N\sigma^2}{y_N^2 + b} \stackrel{!}{=} 0 \tag{68}$$

the solution of which yields immediately

$$y^{ss}_{N_\kappa\,1,2} = \pm\sqrt{\sqrt{\frac{\varepsilon^2}{2\kappa}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]^2 + u^2 + N\sigma^2} - b}. \tag{69}$$

Summarizing, one gets similarly to (59) the steady state solutions

$$y^{ss}_{N_\kappa} = \begin{cases} 0, & \text{for} \quad b^2 > \frac{\varepsilon^2}{2\kappa}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]^2 + u^2 + N\sigma^2 \\ \pm\sqrt{\sqrt{\frac{\varepsilon^2}{2\kappa}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]^2 + u^2 + N\sigma^2} - b}, & \text{else.} \end{cases} \tag{70}$$

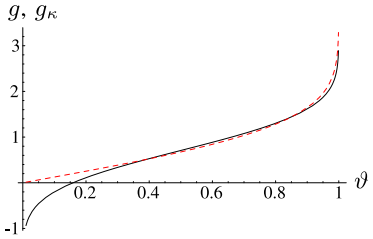


Fig. 10. Comparison of the approximation quality of $g_\kappa(\vartheta, 1)$ displayed as continuous (black) curve. The function $g(\vartheta)$ is displayed as the dashed curve.

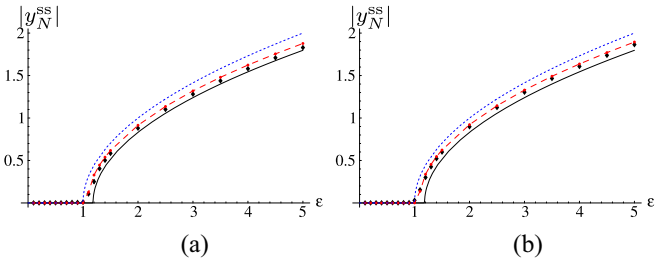


Fig. 11. Steady state y_N values (displayed as black diamonds) of $(\lfloor\vartheta\lambda\rfloor/\lfloor\vartheta\lambda\rfloor_I, \lambda)$-ES ($\lambda = 10\,000$, $\vartheta = 0.4$) on f_2 with (a) $b = 1, N = 40$ and (b) $b = 1, N = 100$ depending on the noise strength ε. The dots connected by the dashed lines are the data points predicted by (70). The bottom right (continuous) curve is obtained from (70) using $\sigma = 0$ and $u = 0$. The dotted blue curve displays the robust maximizer $\hat{y}_N$ as given by (3). Each data point was obtained as average over (a) four independent runs with 40 000 000 and (b) two independent runs with 100 000 000 function evaluations per run. Within a single run, the last 500 generations have been used to calculate the expected steady state value of the specific run. The observed standard deviations are smaller than the size of the black diamonds. Note, even though the resampling size is only $\kappa = 4$, the resulting expected steady state values get close to the robust maximizer (top dotted curve). For further results, see Appendix C-B of the supplementary material.

A comparison with (59) shows that the only difference between both progress rate formulas is in the expressions associated with the square brackets

$$\text{without resampling:} \quad g(\vartheta) := \Phi^{-1}\left(\frac{1+\vartheta}{2}\right) \tag{71}$$

$$\text{with resampling:} \quad g_\kappa(\vartheta, \kappa) := \frac{1}{\sqrt{2\kappa}}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]. \tag{72}$$

While for $\kappa = 1$ – in the ideal case – both expressions should be equal to each other, i.e., $\forall\vartheta : g(\vartheta) = g_\kappa(\vartheta, 1)$, one sees that the approximation $g_\kappa(\vartheta, 1)$ (which is based on Fisher's χ^2-cdf approximation [see (95) in Appendix B of the supplementary material], which is asymptotically exact for $\kappa \to \infty$) necessarily deviates from $g(\vartheta)$. However, as one can see in Fig. 10, the approximation quality is reasonably good for moderate truncation ratios commonly used in practice.

Results of y^{ss}_N steady state simulations similar to those presented in Fig. 7 are shown in Fig. 11 for $\kappa = 4$.

Further experimental results can be found in Appendix C-B of the supplementary material. Comparing the predictions of (70), displayed as dashed curve, with the observed data points (black diamonds) confirms that the predictive quality is surprisingly high even for this small κ. Therefore, (70) cannot only be used to discuss the asymptotic scaling behavior with respect to the sampling size κ, but also for reasonably small sampling sizes.

Let us first consider the $\kappa \to \infty$ case. From the discussion of the averaging formula (5) in Section I, it is known that the ES will approximate the robust maximizer $\hat{\mathbf{y}}$ arbitrarily well. This should also hold for the steady state y_N derived from the progress rate theory. That is, (70) must approach the robust maximizer state $\hat{y}_N = \sqrt{\varepsilon - b}$ given by (3) provided that $u \to 0$ and $\sigma \to 0$. Considering the κ-dependent part g_κ (72), in (70), one sees that

$$g_\kappa(\vartheta,\kappa)^2 = \frac{1}{2\kappa}\left[\Phi^{-1}(\vartheta) + \sqrt{2\kappa - 1}\right]^2 = 1 + \frac{\sqrt{2}\Phi^{-1}(\vartheta)}{\sqrt{\kappa}} + \mathcal{O}\left(\kappa^{-1}\right). \tag{73}$$

Therefore, $g_\kappa^2 \to 1$ and provided that $u \to 0$ and $\sigma \to 0$ one immediately sees that

$$\left|y_{N_\kappa}^{ss}\right| \xrightarrow{\kappa\to\infty} \left|\hat{y}_N\right| \tag{74}$$

theoretically holds. In practice, however, $\hat{y}_N$ is not exactly reached: There remains a deviation due to the $N\sigma^2$ term in (70) which does not vanish in (mutative) σ-self-adaptive ES.[7] These deviations, however, are usually small. Yet, this may be regarded as a flaw of current self-adaptive ES. As an alternative, one might consider *cumulative step size adaptation* (CSA-ES) [15]. Unfortunately, under strong noise conditions CSA-ES exhibits premature convergence for small population sizes and instable behavior when working with large populations [7].

Equation (70) can also be used to assess the influence of the resampling size κ on the approximation quality. Comparing the ratio $|y_{N_\kappa}^{ss}|/|\hat{y}_N|$ that should be one in the ideal case, one gets with (70) and (73), assuming $u = 0$ and $\sigma = 0$, and taking $\sqrt{1+x} = 1 + x/2 + \mathcal{O}(x^2)$ into account

$$\frac{\left|y_{N_\kappa}^{ss}\right|}{\left|\hat{y}_N\right|} \simeq \sqrt{1 + \frac{\varepsilon\left[\frac{\Phi^{-1}(\vartheta)}{\sqrt{2\kappa}} + \sqrt{1-\frac{1}{2\kappa}} - 1\right]}{\varepsilon - b}} \simeq \sqrt{1 + \frac{\varepsilon}{\varepsilon - b}\frac{\Phi^{-1}(\vartheta)}{\sqrt{2\kappa}}} \simeq 1 + \frac{\varepsilon}{\varepsilon - b}\frac{\Phi^{-1}(\vartheta)}{\sqrt{8}}\frac{1}{\sqrt{\kappa}}.$$

Thus, the relative error becomes

$$\frac{\left|y_{N_\kappa}^{ss}\right| - \left|\hat{y}_N\right|}{\left|\hat{y}_N\right|} \simeq \frac{\varepsilon}{\varepsilon - b}\frac{\Phi^{-1}(\vartheta)}{\sqrt{8}}\frac{1}{\sqrt{\kappa}}. \tag{75}$$

That is, the (relative) approximation error vanishes asymptotically with $\kappa^{-1/2}$. Therefore, increasing the sample size κ does only slowly improve the approximation quality. As we have seen, usually a small sample size, such as $\kappa = 4, \ldots, 10$, yields often satisfactory results. However, not always, as one can infer from (75). The relative robust optimizer approximation error increases monotonically with the truncation ratio ϑ. Since $\Phi^{-1}(\vartheta)$ increases super-linearly, choosing ϑ too large is not desirable. This is also confirmed in Fig. 12, where one can see that in the example considered the sample size of $\kappa = 4$,

[7]Note, one can reach $\hat{y}_N$ exactly by tuning ϑ in (70). However, in practice the correct ϑ value is unknown.

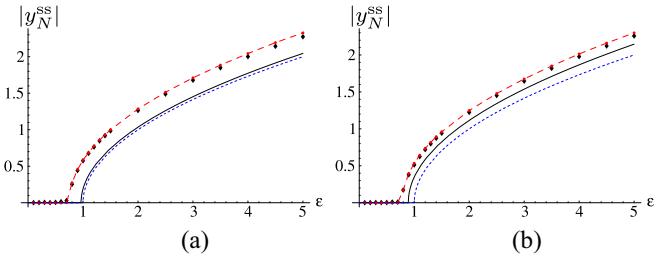


Fig. 12. Comparison of ES's robust optimizer approximation quality (a) without and (b) with resampling ($\kappa = 4$) using a truncation ratio $\vartheta = 0.7$ (further information can be taken from Figs. 7 and 11, respectively). As one can see, there is only a small effect of resampling when using a large truncation ratio.

Fig. 12(b) does only slightly improve the ES's $\hat{y}_N$ approximation quality. From viewpoint of (75) choosing the truncation ratio of $\vartheta = 1/2$ seems to be the best choice. However, one has to keep in mind that this choice is only optimal asymptotically under the condition of vanishing σ and u.

IV. Conclusion

EA are usually regarded as good at approximating solutions to noisy optimization problems. Since in robust optimization noise is injected into the function to be optimized, the resulting function is noisy and, thus, it belongs to the application domain where EAs are regarded to excel. However, as we have seen, ES do not always hold that claim. There are relatively simple functions the robust optimizer of which cannot be well approximated by the $(\mu/\mu_I, \lambda)$-ES. Only when the correct truncation ratio $\vartheta = \mu/\lambda$ is chosen, the ES is able to approximate the robust optimizer $\hat{\mathbf{y}}$ arbitrarily exact. It is an open question whether this observation can also be transferred to other EAs such as differential evolution (DE), particle swarm optimization (PSO), or genetic algorithms (GAs).

From viewpoint of ES theory and in order to evaluate the efficiency of the ES, it is important to predict the steady state behavior of the ES on the respective test functions. Up until now, all nontrivial analyses done concerned quadratic test functions. This paper is the *first* that extends the progress rate analysis to a nonquadratic noisy function. To this end, a novel asymptotic analysis technique for large population sizes has been developed and applied to a class of FNIM. Using a newly derived asymptotically exact progress rate integral, we were able to calculate the progress rate of the critical objective variable y_N for both ES without and with resampling.

From an application point of view the theoretical analysis that the relative error can only be reduced by the square root of κ is both highly relevant and frustrating at the same time. For the first time, we know that choosing a small κ is basically a waste of resources for the problem class of FNIMs, because it will not substantially increase convergence of the optimization. At the same time, as we mentioned in the introduction, using a large κ (> 100) is impossible for many applications. Therefore, from a practical point of view, there are two options how to deal with the dilemma. First, we can try to better understand the role of the optimal truncation ratio on the relative error, in particular with regard to the range of truncation values that would enable convergence even with a small κ (or even

with $\kappa = 1$). Second, we need to find out whether the optimization problem we face has the convergence problem as shown in this paper for this type of FNIM.[8] If we could derive a quantity during the optimization process, that would indicate that the problem is a "difficult" FNIM, we could at least determine whether the optimization result is reliable or not. Again from a practical point of view this is not entirely satisfactory, however, it is better than being left in the dark again about the quality of the optimization result.

Let us come back to the progress rate approach presented in this paper. The importance of the progress rate theory lies in the fact that the condition $\varphi(\mathbf{y}) = 0$ defines the manifold of those states $\mathbf{y}$ for which the ES cannot improve the solution any further. That is, $\varphi(\mathbf{y}) = 0$ defines the steady state of the ES. It appears as a real surprise that the steady state solutions can be obtained in terms of *closed* expressions. This allows for the direct assessment of the scaling behavior of the ES approximation quality on f_2. Moreover, even the predictive quality of the formulas obtained is surprisingly high. Therefore, they can be used to explain the deviations of the real ES behavior from the desired optimizer behavior. As one can clearly infer from (59) and (70), respectively, the observed deviations from the robust optimizer $\hat{\mathbf{y}}$ (3) are mainly due to the nonvanishing mutation strength σ. This uncovers a problem of classical self-adaptive ES in environments with nonvanishing noise. The mutation strength stabilizes at a certain expected value greater than zero (see also issues 2 and 3 below). This observation cannot be predicted theoretically up to now leading us to the future research perspectives.

1) While this paper lays the ground for an asymptotically exact progress rate theory, the analysis is still incomplete. The focus was on the behavior of the y_N component. From experiments we know that the aggregated state variable u, defined in (24), decreases monotonously with increasing population size λ. However, the decrease rate remains still to be derived. While [8] offered a first estimate of the steady state u (the R_∞ in [8, eq. (25)]), the assumptions made are only rough approximations. What is really needed is a progress rate theory for u and y_{N-1}.
2) Calculating the steady state mutation strength $\sigma > 0$ (for $g \to \infty$) is a challenge. Up until now, only the σSA-ES on the noisy sphere and the ridge model have been analyzed in the Ph.D. thesis [16]. There it has been shown that on the sphere model the *normalized* steady state mutation strength σ^* is proportional to $\sqrt{\mu}$ (μ parental population size). Together with the residual distance to the optimizer R_∞-formula this implies also a constant steady state mutation strength σ.
3) The nonvanishing steady state mutation strength σ is a problem in the final phase of the evolution runs. It would be desirable to have a strategy that copes with the problem of nonvanishing σ in self-adaptive ES under noise. The CSA-ES by [15] is *not* the solution so far. It exhibits premature stagnation indicating that the mutation strength goes down too fast in the strong-noise regime. It is still an open question whether one can design a stable working CSA-ES without the flaw of premature stagnation.
4) The focus in this paper was on the prediction of the behavior of ESs. The question arises how other EAs, such as DE, GAs, PSO, and other direct search algorithms will cope with FNIMs. Clearly, if a direct search strategy removes the noise by excessive resampling, any optimization strategy should approach the robust optimizer. However, in these cases the efficiency issue becomes dominant. Up to now, there are neither theoretical nor empirical investigations that evaluate search strategies in such scenarios.

[8]We have shown in a previous paper [9] that "simple" FNIMs exist without convergence problems.

APPENDIX A
PROOFS OF THEOREMS

Integral Representation of the Expectation of the Mean of Induced Order Statistics (Theorem 1)

Proof: Using the definition of the expected value one gets

$$\mathrm{E}[\langle z\rangle] = \frac{1}{\mu}\mathrm{E}\Big[\sum_{m=1}^{\mu} z_{m;\lambda}\Big] = \frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\big[z_{m;\lambda}\big] \tag{76}$$

with

$$\mathrm{E}\big[z_{m;\lambda}\big] = \int_{z=z_l}^{z=z_u} z p_{m;\lambda}(z)\,\mathrm{d}z. \tag{77}$$

That is, the pdf $p_{m;\lambda}(z)$ of the $z_{m;\lambda}$ variates must be derived. To this end, consider a sample $Z = z$ of the random variate obeying the density $p_z(z)$. In order to be accepted as the mth best individual the associated $Q = q$ value, generated from the density $p(q|z)$, must be the mth *greatest* Q-value from the set of the λ Q-values. That is, $m-1$ of the Q-values must be greater than q and $\lambda - m$ must be less than (or equal) to q. The probability of this specific event is therefore

$$\begin{aligned}\Pr\big[Q > q\big]^{m-1}\Pr\big[Q \le q\big]^{\lambda-m} &= \big[1-\Pr\big[Q \le q\big]\big]^{m-1}\Pr\big[Q \le q\big]^{\lambda-m}\\ &= \big[1-P_q(q)\big]^{m-1}\big[P_q(q)\big]^{\lambda-m}\end{aligned}$$

where $P_q(q)$ is the cdf of the q-variate. Since the ordering within the $m-1$ larger Q-samples and $\lambda - m$ smaller Q-samples is without relevance, the number of different orderings belonging to the specific "$m;\lambda$"-individual becomes $\lambda!/((m-1)!(\lambda-m)!)$. Since the generation of a specific q value is conditioned on z with the density $p_q(q|z)$, the probability density of $z_{m;\lambda}$ is finally obtained by integrating over all q states, yielding

$$p_{m;\lambda}(z) = p_z(z)\frac{\lambda!}{(m-1)!(\lambda-m)!} \times \int_{q=q_l}^{q=q_u} p_q(q|z)\big[1-P_q(q)\big]^{m-1}\big[P_q(q)\big]^{\lambda-m}\,\mathrm{d}q. \tag{78}$$

If we back-substitute (78) into (77) and then in (76), we get

$$\mathrm{E}[\langle z\rangle] = \frac{\lambda!}{\mu}\int_{z=z_l}^{z=z_u} z p_z(z) \int_{q=q_l}^{q=q_u} p_q(q|z) \times \sum_{m=1}^{\mu} \frac{\big[1-P_q(q)\big]^{m-1}\big[P_q(q)\big]^{\lambda-m}}{(m-1)!(\lambda-m)!}\,\mathrm{d}q\,\mathrm{d}z. \quad (79)$$

By means of the regularized incomplete beta function [17, p. 83], the sum in (79) can be transformed into an integral

$$\sum_{m=1}^{\mu}\frac{[1-P]^{m-1}P^{\lambda-m}}{(m-1)!(\lambda-m)!} = \frac{1}{(\lambda-\mu-1)!(\mu-1)!}\int_{x=0}^{x=P} x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x. \quad (80)$$

One gets

$$\mathrm{E}[\langle z\rangle] = \frac{\lambda!}{(\lambda-\mu-1)!\mu!}\int_{z=z_l}^{z=z_u} z p_z(z) \int_{q=q_l}^{q=q_u} p_q(q|z) \times \int_{x=0}^{x=P_q(q)} x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x\,\mathrm{d}q\,\mathrm{d}z. \quad (81)$$

Exchanging the order of the two inner integrations taking the quantile function (i.e., the inverse) of the cdf of $P_q(q)$ into account, one obtains

$$\mathrm{E}[\langle z\rangle] = \frac{\lambda!}{(\lambda-\mu-1)!\mu!}\int_{z=z_l}^{z=z_u} z p_z(z) \int_{x=0}^{x=1} \int_{q'=P^{-1}(x)}^{q'=q_u} p_q\big(q'|z\big)\, x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}q'\,\mathrm{d}x\,\mathrm{d}z. \quad (82)$$

The integration over q' can be immediately carried out yielding

$$\int_{q'=P_q^{-1}(x)}^{q'=q_u} p_q\big(q'|z\big)\,\mathrm{d}q' = 1 - P_q\Big(P_q^{-1}(x)|z\Big) \quad (83)$$

where $P_q(q|z)$ is the conditional cdf of the q-distribution given a (fixed) z. Plugging this into (82) one finally obtains (12). ■

Appendix B

Infinite Population Integral Representation, Theorem 2

Proof: Using Theorem 1, the expected value of $\langle z\rangle$ can be expressed by the integral

$$\mathrm{E}[\langle z\rangle] = \frac{\lambda}{\mu}\int_{z=z_l}^{z=z_u} z p_z(z) I(z)\,\mathrm{d}z \quad (84)$$

where I, is of the from

$$I(z) = \frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!}\int_{x=0}^{x=1} f(x,z)\; x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x \quad (85)$$

with

$$f(x,z) \coloneqq 1 - P_q\Big(P_q^{-1}(x)|z\Big). \quad (86)$$

Considering $x^{\lambda-\mu-1}(1-x)^{\mu-1} =: g(x)$ in the integrand of (85), one sees that this function has a maximum at $\hat{x} \in (0,1)$ which gets sharper with increasing λ (keeping ϑ constant). Therefore, expanding the function $f(x)$ about $\hat{x}$ yields a Taylor series that can be integrated term-by-term. The location $\hat{x}$ of the maximum can be easily obtained

$$\begin{aligned}
\max\Big[x^{\lambda-\mu-1}(1-x)^{\mu-1}\Big] &= \hat{x}^{\lambda-\mu-1}\big(1-\hat{x}\big)^{\mu-1}\\
\iff \quad \frac{\mathrm{d}}{\mathrm{d}x}x^{\lambda-\mu-1}(1-x)^{\mu-1} &\stackrel{!}{=} 0\\
(\lambda-\mu-1)x^{\lambda-\mu-2} - (\mu-1)(1-x)^{\mu-2} &\stackrel{!}{=} 0\\
\Rightarrow \quad (\lambda-\mu-1)\big(1-\hat{x}\big) &= (\mu-1)\hat{x}\\
\hat{x} \;=\; 1 - \frac{\mu-1}{\lambda-2}.
\end{aligned} \quad (87)$$

According to [17, p. 79], the identity

$$\frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!}\int_{x=0}^{x=1} x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x = 1 \quad (88)$$

holds and can be used to calculate the series terms arising from the expansion of $f(x,z)$ in (85)

$$I(z) = \sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k f(x,z)}{\partial x^k}\bigg|_{x=\hat{x}}\frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!} \times \int_{x=0}^{x=1}\big(x-\hat{x}\big)^k x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x. \quad (89)$$

Adopting the integral identity 15.3.1 of [17, p. 215]

$$\frac{(\lambda-1)!}{(\lambda-\mu-1)!(\mu-1)!}\int_{x=0}^{x=1}\big(x-\hat{x}\big)^k x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x = \big(-\hat{x}\big)^k\,{}_2F_1\big(-k,\lambda-\mu;\lambda;1/\hat{x}\big) \quad (90)$$

where ${}_2F_1(a,b;c;z)$ is the hyper-geometric function (see [17, p. 213] for its definition), (89) becomes

$$I(z) = \sum_{k=0}^{\infty}\frac{1}{k!}\frac{\partial^k f(x,z)}{\partial x^k}\bigg|_{x=\hat{x}}\big(-\hat{x}\big)^k\,{}_2F_1\big(-k,\lambda-\mu;\lambda;1/\hat{x}\big). \quad (91)$$

For $k = 0,1,2,3$ one finds (by simple integration) with (14), (87), and (88)

$$\begin{aligned}
{}_2F_1\big(0,\lambda-\mu;\lambda;1/\hat{x}\big) &= 1\\
{}_2F_1\big(-1,\lambda-\mu;\lambda;1/\hat{x}\big) &= \frac{1}{\lambda}\,\frac{1-2\vartheta}{(1-\vartheta-1/\lambda)}
\end{aligned}$$

$$
\begin{aligned}
{}_2F_1\big(-2, \lambda-\mu; \lambda; 1/\hat{x}\big) &= \frac{1}{\lambda+1} \\
&\quad \times \frac{(1-\vartheta)\vartheta + \big(1-8\vartheta+8\vartheta^2\big)/\lambda + 1/\lambda^2}{(1-\vartheta-1/\lambda)^2} \\
{}_2F_1\big(-3, \lambda-\mu; \lambda; 1/\hat{x}\big) &= \frac{(1-2\vartheta)}{(\lambda+1)(\lambda+2)} \\
&\quad \times \frac{(1-\vartheta)\vartheta(5-22/\lambda) + (1+3/\lambda)/\lambda + 2/\lambda^3}{(1-\vartheta-1/\lambda)^3}.
\end{aligned}
$$

As one can see, for $k \geq 1$ the values of ${}_2F_1(-k, \lambda-\mu; \lambda; 1/\hat{x})$ have an upper bound $\mathcal{O}(1/\lambda)$. Therefore, provided that the derivatives of (86) are bounded, one gets for (90) asymptotically (i.e., $\lambda \to \infty$) zero (for $k \neq 0$) and (91) becomes simply

$$I(z) \simeq f(1-\vartheta, z). \tag{92}$$

Plugging this into (84) we gets

$$\mathrm{E}[\langle z \rangle] \simeq \frac{\lambda}{\mu} \int_{z=z_l}^{z=z_u} z p_z(z)\, f(1-\vartheta, z)\, \mathrm{d}z. \tag{93}$$

Inserting (86) into (93), one finally obtains the asymptotically exact integral representation (15). ■

REFERENCES

[1] G.-J. Park, T.-H. Lee, K. H. Lee, and K.-H. Hwang, "Robust design: An overview," *AIAA J.*, vol. 44, no. 1, pp. 181–191, 2006.

[2] H.-G. Beyer and B. Sendhoff, "Robust optimization—A comprehensive survey," *Comput. Methods Appl. Mech. Eng.*, vol. 196, nos. 33–34, pp. 3190–3218, 2007.

[3] S. Tsutsui, A. Ghosh, and Y. Fujimoto, "A robust solution searching scheme in genetic search," in *Parallel Problem Solving from Nature—IV*, H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, Eds. Heidelberg, Germany: Springer, 1996, pp. 543–552.

[4] S. Tsutsui and A. Ghosh, "Genetic algorithms with a robust solution searching scheme," *IEEE Trans. Evol. Comput.*, vol. 1, no. 3, pp. 201–208, Sep. 1997.

[5] J. Branke, "Creating robust solutions by means of evolutionary algorithms," in *Parallel Problem Solving from Nature—V*, A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, Eds. Heidelberg, Germany: Springer, 1998, pp. 119–128.

[6] I. Paenke, J. Branke, and Y. Jin, "Efficient search for robust solutions by means of evolutionary algorithms and fitness approximation," *IEEE Trans. Evol. Comput.*, vol. 10, no. 4, pp. 405–420, Aug. 2006.

[7] H.-G. Beyer, M. Olhofer, and B. Sendhoff, "On the behavior of $(\mu/\mu_I, \lambda)$-ES optimizing functions disturbed by generalized noise," in *Foundations of Genetic Algorithms, 7*, K. A. De Jong, R. Poli, and J. Rowe, Eds. San Francisco, CA, USA: Morgan Kaufmann, 2003, pp. 307–328.

[8] B. Sendhoff, H.-G. Beyer, and M. Olhofer, "The influence of stochastic quality functions on evolutionary search," in *Recent Advances in Simulated Evolution and Learning* (Advances in Natural Computation), K. Tan, M. Lim, X. Yao, and L. Wang, Eds. New York, NY, USA: World Sci., 2004, pp. 152–172.

[9] H.-G. Beyer and B. Sendhoff, "Functions with noise-induced multimodality: A test for evolutionary robust optimization–properties and performance analysis," *IEEE Trans. Evol. Comput.*, vol. 10, no. 5, pp. 507–526, Oct. 2006.

[10] B. Sendhoff, H.-G. Beyer, and M. Olhofer, "On noise induced multi-modality in evolutionary algorithms," in *Proc. 4th Asia–Pac. Conf. Simulat. Evol. Learn.–SEAL*, vol. 1. Singapore, 2002, pp. 219–224.

[11] H.-G. Beyer and B. Sendhoff, "Evolution strategies for robust optimization," in *Proc. WCCI Conf.*, Piscataway Township, NJ, USA, 2006, pp. 4489–4496.

[12] H.-G. Beyer and B. Sendhoff, "Evolutionary algorithms in the presence of noise: To sample or not to sample," in *Proc. 1st IEEE Symp. Found. Comput. Intell. (FOCI)*, Honolulu, HI, USA, 2007, pp. 17–24.

[13] S. Meyer-Nieberg and H.-G. Beyer, "The dynamical systems approach—Progress Measures and convergence properties," in *Handbook of Natural Computing*, G. Rozenberg, T. Bäck, and J. Kok, Eds. Berlin, Germany: Springer, 2012, pp. 741–814.

[14] H.-G. Beyer, *The Theory of Evolution Strategies* (Natural Computing Series). Heidelberg, Germany: Springer, 2001, doi: 10.1007/978-3-662-04378-3.

[15] N. Hansen and A. Ostermeier, "Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation," in *Proc. IEEE Int. Conf. Evol. Comput. (ICEC)*, Nagoya, Japan, 1996, pp. 312–317.

[16] S. Meyer-Nieberg, "Self-adaptation in evolution strategies," Ph.D. dissertation, Comput. Sci. Dept., Univ. of Dortmund, Dortmund, Germany, 2007.

[17] M. Abramowitz and I. A. Stegun, *Pocketbook of Mathematical Functions*. Thun, Switzerland: Verlag Harri Deutsch, 1984.

[18] N. L. Johnson and S. Kotz, *Continuous Univariate Distributions-1*. Boston, MA, USA: Houghton Mifflin, 1970.



Hans-Georg Beyer received the Diploma degree in theoretical electrical engineering from Ilmenau Technical University, Ilmenau, Germany, in 1982, and the Ph.D. degree in physics from the Bauhaus-University Weimar, Weimar, Germany, in 1989, and the Habilitation degree from the University of Dortmund, Dortmund, Germany, and qualified for a professorship in Computer Science (CS) in 1997.

From 1982 to 1984, he was a Research and Development Engineer with the Reliability Physics Department, VEB Gleichrichterwerk, Stahnsdorf, Germany. From 1984 to 1989, he was a Research and Teaching Assistant and a Post-Doctoral Fellow with the Physics Department and the CS Department, Bauhaus-University Weimar. From 1990 to 1992, he was a Senior Researcher with the Electromagnetic Fields Theory Group, Darmstadt University of Technology, Darmstadt, Germany. From 1993 to 2004, he was with the CS Department of the University of Dortmund. In 1997, he became a DFG Heisenberg Fellow with German Research Foundation, Bonn, Germany, where he was a Leader of a working group and from 2003 to 2004 a Professor of CS. Since 2004, he has been a Professor with the Vorarlberg University of Applied Sciences, Dornbirn, Austria. He has authored the book entitled *The Theory of Evolution Strategies* (Heidelberg: Springer-Verlag, 2001) and has authored/co-authored over 100 papers.

Dr. Beyer was the Editor-in-Chief of the MIT Press journal *Evolutionary Computation* from 2010 to 2016. He has been serving as an Associate Editor for the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION since 1997. He is an Advisory Board Member of the Elsevier journal *Swarm and Evolutionary Computation*.



Bernhard Sendhoff (M'99–SM'05) received the Ph.D. degree in applied physics from the Ruhr-Universität Bochum, Bochum, Germany, in 1998.

From 1999 to 2002, he was with Honda Research and Development Europe (Deutschland) GmbH, Offenbach, Germany, last in the position of the Deputy Division Manager and the Chief Scientist. In 2003, he joined the Honda Research Institute Europe GmbH, Offenbach, as a Chief Technology Officer. Since 2007, he has been the Honorary Professor with the School of Computer Science, University of Birmingham, Birmingham, U.K., and since 2008, he has been the Honorary Professor with the Technical University of Darmstadt, Darmstadt, Germany. Since 2011, he has been the President of the Honda Research Institute Europe GmbH. He has authored or co-authored over 180 scientific publications.

Dr. Sendhoff is a Senior Member of the ACM and a member of the SAE.