Convergence Analysis of Evolutionary Algorithms That are Based on the Paradigm of Information Geometry*

Hans-Georg Beyer Hans-Georg.Beyer@fhv.at
Department of Computer Science, Vorarlberg University of Applied Sciences,
Hochschulstr. 1, A-6850 Dornbirn, Austria.

Abstract

The convergence behaviors of so-called natural evolution strategies (NES) and of the information-geometric optimization (IGO) approach are considered. After a review of the NES/IGO ideas, which are based on information geometry, the implications of this philosophy w.r.t. optimization dynamics are investigated considering the optimization performance on the class of positive quadratic objective functions (the ellipsoid model). Exact differential equations describing the approach to the optimizer are derived and solved. It is rigorously shown that the original NES philosophy optimizing the expected value of the objective functions leads to very slow (i.e. sublinear) convergence towards the optimizer. This is the real reason why state-of-the-art implementations of IGO algorithms optimize the expected value of transformed objective functions, e.g. by utility functions based on ranking. It is shown that these utility functions are localized fitness functions that change during the IGO flow. The governing differential equations describing this flow are derived. In the case of convergence, the solutions to these equations exhibit an exponentially fast approach to the optimizer (i.e. linear convergence order). Furthermore, it is proven that the IGO philosophy leads to an adaptation of the covariance matrix that equals in the asymptotic limit – up to a scalar factor – the inverse of the Hessian of the objective function considered.

Keywords

Convergence Analysis, Evolution Strategies, Information Gain, Information Geometry, Relative Entropy Loss

1 Introduction

For decades, engineering of Evolution Strategies (ES) as well as of other Evolutionary Algorithm (EA) related designs (including Genetic Algorithms, Particle Swarm Optimization and Differential Evolution to name a few) was mainly based on biomimicry (also referred to as bionics). That is, principles gleaned from biology were translated into optimization algorithms. While there were attempts to put the design philosophy of ESs on some kind of "axiomatic" base by formulating design principles, see Beyer and Deb (2001); Beyer (2001); Hansen (2006); Beyer (2007), a new development by Wierstra et al. (2008); Sun et al. (2009) called "Natural Evolution Strategies" (NES) promised a more principled approach to the design of EAs. This new approach yielded algorithm implementations that were similar to the well-known covariance matrix adaptation (CMA) ES of Hansen et al. (2003) with rank-μ update, see Glasmachers et al. (2010). This striking similarities gave rise to attempts grounding CMA-ES on information geometry (Amari and Nagaoka (2000)) resulting in a work of Akimoto et al. (2012b).

An alternative view on NES is to consider it as an algorithmic implementation of an information-geometric optimization (IGO) flow in the parameter space of the respective probability distribution family. That is, an IGO differential equation describing the time evolution of

*Revised 2020 version. The author is grateful to Li Zhenhua for pointing out mistakes to be corrected here.

the distribution parameters is set up and solved numerically by discretizing the time (thus getting a generational algorithm, i.e. an EA) and estimating the expected values by Monte-Carlo sampling (the mutation and recombination process in ES), see Akimoto et al. (2012a).

From a bird's-eye view, most of the research done in this field follows directly or indirectly this line of approach with the aim to derive real EA implementations. On the other hand, the IGO flow differential equation can be used to derive theoretical assertions w.r.t. the convergence behavior and generally the time evolution of the IGO flow system. This line of research has been opened by the work of Glasmachers (2012) and Akimoto et al. (2012a). Considering isotropic mutations they proved convergence to the optimizer of convex quadratic functions using different techniques avoiding the direct solution of the IGO equation system. The most advanced results have been provided by Akimoto (2012) who proved linear convergence order on functions with ellipsoidal level sets.

It should be clear that proving convergence of the IGO flow without calculating the real time evolution is but a first step. However, the direct solution of the IGO flow system yields the maximum information which allows for a deeper understanding of the evolutionary dynamics of IGO. This is the primary goal of this paper. It is devoted to the calculation of the exact time evolution of NES algorithms optimizing convex quadratic objective functions (the so-called ellipsoid model). The NES to be analyzed uses normally distributed mutations with a fully developed covariance matrix $\mathbf{C}$. Both cases are considered, the ordinary expected fitness value optimization and the case of rank-based utility optimization. It will be shown that the original NES philosophy, i.e. the "natural gradient" ascent in the expected fitness landscape leads to sublinear convergence. That is, the original NES idea results in convergent but slowly performing algorithms. Introducing a rank-based evaluation instead of the direct fitness changes the behavior drastically. This rank-based evaluation localizes the search in the sense that the globally defined objective function is replaced by a locally (in time) acting utility function. Such a utility can be obtained in different ways. For example, considering the optimal ES weighting of Arnold (2006), one gets a locally standardized fitness. Alternatively, (μ, λ)-ES truncation selection has a similar effect. In both cases asymptotical exponentially fast convergence to the optimizer will be proven. These results provide the theoretical explanation why recent NES implementations always rely on rank-based utility functions. Additionally, the analysis technique presented also allows for the calculation of the dynamics of the covariance matrix evolution. As will be shown, $\mathbf{C}(t)$ approaches up to a scalar factor the inverse of the Hessian of the objective function.

The rest of the paper is organized as follows. The philosophy of NES and IGO will be introduced in a self-contained manner. However, this introduction will not consider implementation aspects since these are not needed in this paper. In Sect. 2 it will be shown – also, but not only for didactical reasons – that ordinary gradient ascent in expected fitness landscapes may cause problems in that different parameterizations may yield qualitatively different (and sometimes undesirable) behaviors. Thereafter, the idea of information distance constrained gradient ascent is introduced building the philosophical basis of NES. The resulting IGO flow will be analyzed in Sect. 3 by solving the differential equation system yielding a slowly converging evolution dynamics. In Sect. 4 the idea of localized fitness evaluation is added to the NES framework and analyzed for optimal weighting and truncation selection assuming normally distributed fitness values. Finally, in Sect. 5 conclusions are drawn.

2 NES/IGO in a Nutshell

Given a function $f(\boldsymbol{x})$ to be optimized, one has numerous options to perform a gradual approach to the optimizer $\hat{\boldsymbol{x}}$. One way – as pursued in evolutionary algorithms – consists in randomly sampling $\boldsymbol{x}$ values from a family of probability distributions $P(\boldsymbol{\theta})$ the density of which may be

given by $p(\boldsymbol{x}|\boldsymbol{\theta})$.[1] Here, the set of distribution parameters $\boldsymbol{\theta}$ evolves over time t (the generation counter) with the goal to change the distribution in such a manner that with increasing t the distribution is more and more concentrated about the optimizer $\hat{\boldsymbol{x}}$. The manner in which the $\boldsymbol{\theta}$ parameter set is changed over the generations t characterizes the different evolutionary algorithm (EA) class such as estimation of distribution algorithms (EDAs) and Evolution Strategies (ESs). Alternatively, $\boldsymbol{\theta}$ can even be implicitly presented by the population of candidate solutions as is usually done in genetic algorithms (GAs). However, in this work, we will focus on EAs where the distribution parameters $\boldsymbol{\theta}$ are explicitly represented. In the next section, we will start with a naive approach by minimizing the expected value of f. It will be shown that this approach does not always lead to a stable performing EA. Therefore, in an attempt to get a theoretically principled approach based on expected value maximization, the class of so-called *natural evolution strategies* (NES) will be considered (Wierstra et al. (2008)). The underlying ideas will be presented and investigated in subsequent sections.

2.1 Maximizing the Expected Value of f

While each real implementation of EAs has to produce a population of random samples $\boldsymbol{x}_l$ and has to evaluate these samples w.r.t. fitness, i.e. $f_l = f(\boldsymbol{x}_l)$, the information obtained from these samples can be used in different manner. From the EA modeling perspective, one possible objective is to consider the *expected value*

$$E_f(\boldsymbol{\theta}) := \mathrm{E}[f|\boldsymbol{\theta}] = \int f(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\theta}) \, \mathrm{d}^N \boldsymbol{x} \tag{1}$$

as the target quantity to be optimized. That is, in the case of f-maximization this leads to a *transformed* optimization problem

$$\hat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}). \tag{2}$$

This approach is the starting point of the so-called *natural evolution strategies* (NES) by Wierstra et al. (2008). In order to have a meaningful optimization problem (2) one has to ensure that the parameter set $\boldsymbol{\theta}$ allows to express the estimate of the optimizer of $f(\boldsymbol{x})$ with arbitrary precision. This can be achieved if the expected value vector $\overline{\boldsymbol{x}} := \mathrm{E}[\boldsymbol{x}|\boldsymbol{\theta}]$ is itself part of the parameter set $\boldsymbol{\theta}$. For example, this is attained by using multivariate Gaussian normal distributions $\boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C})$ with mean $\overline{\boldsymbol{x}}$ and the symmetric covariance matrix $\mathbf{C} = \mathbf{C}^{\mathrm{T}}$, i.e.

$$\boldsymbol{\theta} = (\overline{\boldsymbol{x}}, \mathbf{C}). \tag{3}$$

The idea behind such a parameterization is that $\overline{\boldsymbol{x}}$ evolves towards the optimizer $\hat{\boldsymbol{x}}$ and the covariance matrix $\mathbf{C}$ shrinks in such a manner that the distribution of the samples gets more and more concentrated about the optimizer $\hat{\boldsymbol{x}}$.

An implementation of the idea of maximizing the expected value function according to (2) starts by realizing a gradient ascent in the parameter space of $\boldsymbol{\theta}$. To this end, the gradient is needed, symbolized by the nabla operator $\nabla_{\boldsymbol{\theta}}$. In real-world cases, however, the expected value (1) cannot be calculated analytically. Therefore, a Monte-Carlo sampling technique is applied in order to *estimate* the gradient of (1), see Sun et al. (2009)

$$\nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}) \approx \frac{1}{\lambda} \sum_{l=1}^{\lambda} f(\boldsymbol{x}_l) \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{x}_l|\boldsymbol{\theta}), \quad \text{where} \quad \boldsymbol{x}_l \sim \boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C}) \tag{4}$$

[1] In this work, only real-valued optimization is considered, i.e., $\boldsymbol{x} \in \mathbb{R}^N$ and therefore, probability distributions will be described by their probability density functions (pdfs) $p(\boldsymbol{x}|\boldsymbol{\theta})$ parameterized by sets of distribution parameters.

and λ is the number of samples taken. λ is also referred to as the offspring population size in Evolution Strategies (ESs). The gradient estimate (4) is then used to perform a hill climbing step in the $\boldsymbol{\theta}$-space, thus forming an iterative update formula

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta \, \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \tag{5}$$

where η is a step-size factor. While in practice one is interested in an η that provides a fast approach to the optimizer, one can also consider the $\eta \to 0$ limit case. Subtracting $\boldsymbol{\theta}^{(t)}$ on both sides of (5), dividing by η, and taking the limit, the t-discrete $\boldsymbol{\theta}$ values become a continuous time t function and one obtains

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}. \tag{6}$$

This is an ordinary differential equation (ODE) system that can serve as a model description of ESs. A modified version of (6) has been used already in the so-called information-geometric optimization (IGO) framework in order to investigate the convergence behavior of infinite population size models of ESs, see Akimoto et al. (2012a); Glasmachers (2012), and the remainder of this paper.

Before considering more advanced versions of (1) and (6), being the basis for mature version of NES and IGO algorithms, a motivation for those will be given here. Equation (6) describes the continuous time evolution of the $\boldsymbol{\theta}$ trajectories. Provided that f is given in simple form, one can calculate this time evolution. To this end, the maximization of the *general ellipsoid model*

$$f_Q(\boldsymbol{x}) := \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x} \tag{7}$$

is considered with the symmetric positive definite matrix $\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}$ and an arbitrary (constant) vector $\boldsymbol{a}$. Note that the fitness model (7) can be regarded as a local approximation of more complex objective functions $f(\boldsymbol{x})$. In that case, $\mathbf{Q}$ is simply proportional to the Hessian of f, i.e. $\mathbf{Q} = -\frac{1}{2}\mathbf{H}_f(\boldsymbol{x}) = -\frac{1}{2}\nabla\nabla^{\mathrm{T}} f(\boldsymbol{x})$.

Using Gaussian random vectors $\boldsymbol{x} \sim \boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C})$, the expected value (1) can be easily calculated using results from (Beyer, 2001, p. 122)

$$E_f(\boldsymbol{\theta}) = E_f(\overline{\boldsymbol{x}}, \mathbf{C}) = \boldsymbol{a}^{\mathrm{T}}\overline{\boldsymbol{x}} - \overline{\boldsymbol{x}}^{\mathrm{T}}\mathbf{Q}\overline{\boldsymbol{x}} - \mathrm{Tr}[\mathbf{Q}\mathbf{C}]. \tag{8}$$

Here, $\mathrm{Tr}[\mathbf{Q}\mathbf{C}]$ represents the trace of the matrix product $\mathbf{Q}\mathbf{C}$. Calculating the gradient of (8) yields

$$\nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\overline{\boldsymbol{x}}} E_f \\ \nabla_{\mathbf{C}} E_f \end{pmatrix} = \begin{pmatrix} \boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}} \\ -\mathbf{Q} \end{pmatrix}. \tag{9}$$

Inserting this into (6), one obtains

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}(t)}{\mathrm{d}t} = \boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}(t), \tag{10a}$$

$$\frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} = -\mathbf{Q}. \tag{10b}$$

Equation (10a) does not depend on $\mathbf{C}$. A particular solution to the inhomogeneous equation is given by the stationary state condition $\mathrm{d}\overline{\boldsymbol{x}}/\,\mathrm{d}t = \mathbf{0}$ leading to $\boldsymbol{a} = 2\mathbf{Q}\hat{\overline{\boldsymbol{x}}}$. That is, the particular solution is just the optimizer $\hat{\boldsymbol{x}}$ of (7)

$$\hat{\boldsymbol{x}} = \frac{1}{2}\mathbf{Q}^{-1}\boldsymbol{a}. \tag{11}$$

Consider the deviation $\boldsymbol{z}$ measuring the distance of $\overline{\boldsymbol{x}}(t)$ to the optimizer $\hat{\boldsymbol{x}}$

$$\boldsymbol{z} := \overline{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}, \tag{12}$$

Eq. (10a) can be rewritten using (11)

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = -2\mathbf{Q}\boldsymbol{z}. \tag{13}$$

As one can easily check by insertion, its solution is given by

$$\boldsymbol{z}(t) = \exp(-2\mathbf{Q}t)\boldsymbol{z}(0). \tag{14}$$

Thus, one obtains

$$\overline{\boldsymbol{x}}(t) = \hat{\boldsymbol{x}} + \mathrm{e}^{-2\mathbf{Q}t}(\overline{\boldsymbol{x}}(0) - \hat{\boldsymbol{x}}), \tag{15}$$

i.e., the model equation approaches the optimizer $\hat{\boldsymbol{x}}$ exponentially fast (linear convergence order).

The solution of the second equation in (10) is very simple, as one can check by insertion

$$\mathbf{C}(t) = \mathbf{C}(0) - \mathbf{Q}t. \tag{16}$$

However, this result indicates a problem since there is a t_0 above which the covariance matrix $\mathbf{C}(t)$ loses positive definiteness. That is, implementing an ES on the basis of the $\boldsymbol{\theta} = (\overline{\boldsymbol{x}}, \mathbf{C})$ parameterization might cause convergence problems due to inappropriate covariance matrix evolution.

This covariance matrix adaptation problem can be avoided by using another ad hoc parameterization. Using a decomposition $\mathbf{C} = \mathbf{A}\mathbf{A}$ with a symmetric $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$ matrix resolves the problem. That is, one uses $\boldsymbol{\theta} = (\overline{\boldsymbol{x}}, \mathbf{A})$ as parameterization of the Gaussian distribution family.[2] This changes $\mathrm{Tr}[\mathbf{Q}\mathbf{C}]$ in (8) to $\mathrm{Tr}[\mathbf{Q}\mathbf{A}\mathbf{A}] = \mathrm{Tr}[\mathbf{A}\mathbf{Q}\mathbf{A}]$ and the derivatives w.r.t. $\mathbf{A}$ become $\nabla_{\mathbf{A}}\mathrm{Tr}[\mathbf{Q}\mathbf{A}^2] = \mathbf{Q}\mathbf{A} + \mathbf{A}\mathbf{Q}$. As a result, (10b) changes to

$$\frac{\mathrm{d}\mathbf{A}(t)}{\mathrm{d}t} = -(\mathbf{Q}\mathbf{A}(t) + \mathbf{A}(t)\mathbf{Q}). \tag{17}$$

This equation can be solved using the *Ansatz* $\mathbf{A}(t) = \exp(-\mathbf{L}t)\mathbf{A}(0)\exp(-\mathbf{L}t)$ yielding

$$\mathbf{A}(t) = \mathrm{e}^{-\mathbf{Q}t}\mathbf{A}(0)\mathrm{e}^{-\mathbf{Q}t}, \tag{18}$$

as can be easily checked by inserting (18) into (17). As one can see, $\mathbf{A}$ and therefore $\mathbf{C}$ remains positive definite for $t < \infty$ because $\mathbf{Q}$ is positive definite. The result (18) is in contrast to (16). Using the $\mathbf{A}$ parameterization avoids the problem of evolving a non-positive definite covariance matrix. Let us have a closer look at the $\mathbf{C}$-dynamics imposed by (18). To this end, $\mathbf{A}(t)$ is projected into the eigen system of $\mathbf{Q}$. Consider the eigenvalue problem

$$\mathbf{Q}\boldsymbol{u}_k = q_k\boldsymbol{u}_k, \quad \text{where} \quad \boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_k = \delta_{ik} \quad \text{and} \quad 0 < q_i \leq q_k \text{ for } i \leq k. \tag{19}$$

Note, a non-decreasing ordering of the eigenvalues has been chosen in (19), i.e. q_1 is the smallest eigenvalue. Using q_k ($k = 1, \ldots, N$) as basis, Eq. (18) becomes

$$(\mathbf{A}(t))_{ik} = \boldsymbol{u}_i^{\mathrm{T}}\mathbf{A}(t)\boldsymbol{u}_k = \mathrm{e}^{-q_i t}\boldsymbol{u}_i^{\mathrm{T}}\mathbf{A}(0)\boldsymbol{u}_k\mathrm{e}^{-q_k t} = \mathrm{e}^{-(q_i+q_k)t}(\mathbf{A}(0))_{ik}. \tag{20}$$

[2] This has the additional advantage that generating the $\boldsymbol{x} \sim \mathcal{N}(\overline{\boldsymbol{x}}, \mathbf{C})$ samples can be done directly by the transformation of isotropic standard normally distributed random vector components $\boldsymbol{x} \sim \overline{\boldsymbol{x}} + \mathbf{A}\mathcal{N}(\mathbf{0}, \mathbf{I})$. That is, no matrix square root operations are needed.

That is, $\mathbf{A}(t)$ and therefore $\mathbf{C}(t) = \mathbf{A}(t)^2$ shrinks exponentially fast. However, this shrinking appears at different time constants $1/q_k$. In the asymptotic limit $t \to \infty$ the dynamics are dominated by the slowest decaying $\mathrm{e}^{-(q_i+q_k)}$ mode, i.e. the $i = k = 1$ mode. Therefore

$$(\mathbf{A}(t))_{ik} \overset{t\to\infty}{\longrightarrow} \begin{cases} \mathrm{e}^{-2q_1 t}(\mathbf{A}(0))_{11}, & \text{for } i = k = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

Transforming back, the covariance matrix $\mathbf{C} = \mathbf{A}^2$ becomes

$$\mathbf{C}(t) \overset{t\to\infty}{\longrightarrow} \mathrm{e}^{-4q_1 t}(\mathbf{A}(0))_{11}^2 \, \boldsymbol{u}_1 \boldsymbol{u}_1^{\mathrm{T}}. \tag{22}$$

This means that $\boldsymbol{x}$ samples are *predominantly* produced in the direction of the largest principal axis while the other directions related to the $\boldsymbol{u}_k$ with $k > 1$ are over-proportionally dampened. In other words, given an ellipsoid success domain (defined by Eq. (7)), search is predominantly conducted in the major axis direction. That is, for large t, the search degenerates in an one-dimensional subspace of the $\mathbb{R}^N$. Similarly, if the ground state is degenerated by a multiplicity of m, $q_1 = q_2 = \ldots = q_m$, then there are m orthogonal directions $\boldsymbol{u}_k$ resulting in a degenerated search in the corresponding m-dimensional subspace of $\mathbb{R}^N$.

2.2 How to Get Unbiased

Degeneration of the search distribution is undesirable. Instead of (22), it would be desirable to get an asymptotic behavior like this

$$\mathbf{C}(t) \overset{t\to\infty}{\longrightarrow} c(t)\mathbf{Q}^{-1}. \tag{23}$$

This would ensure that the sample points are distributed according to the shape of the ellipsoid described by $\mathbf{Q}$ and the covariance matrix could shrink in even manner. How can such a behavior be ensured by first principles?

It is quite clear that such a behavior can only be obtained by constraining the evolution of the search distribution $p(\boldsymbol{x}|\boldsymbol{\theta}(t))$. It is the goal to change $p(\boldsymbol{x}|\boldsymbol{\theta}(t))$ from t to $t + \delta t$ only minimally while being maximally *unbiased*. That is, the information gained in this step should be rather small. A quantity measuring this change is reminiscent of a relative entropy loss or information gain I considering the ratio $h := p(\boldsymbol{x}|\boldsymbol{\theta} + \delta\boldsymbol{\theta})/p(\boldsymbol{x}|\boldsymbol{\theta})$ (see Rényi (1961))

$$I(P(\boldsymbol{\theta} + \delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) := \int h(\boldsymbol{x}) \ln h(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{\theta}) \, \mathrm{d}^N \boldsymbol{x}. \tag{24}$$

The logarithm with base e has been used instead of the usual base 2 resulting in an additional factor of $\ln 2$ which is, however, without any relevance for the considerations here. Equation (24) evaluates to

$$I(P(\boldsymbol{\theta} + \delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) = \int \ln\left(\frac{p(\boldsymbol{x}|\boldsymbol{\theta} + \delta\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta})}\right) p(\boldsymbol{x}|\boldsymbol{\theta} + \delta\boldsymbol{\theta}) \, \mathrm{d}^N \boldsymbol{x} =: \mathrm{KL}(P(\boldsymbol{\theta} + \delta\boldsymbol{\theta})||P(\boldsymbol{\theta})), \tag{25}$$

where the resulting integral can be also interpreted as the Kullback-Leibler divergence KL. While it is well-known that the information gain defined in (25) is not a distance measure with the meaning of a metric (see, e.g. Eguchi and Copas (2006)), its infinitesimal version is. Interpreting $\delta\boldsymbol{\theta}$ as small quantities, the Taylor expansion of I up to the second order in the $\delta\theta_k$ components yields after a simple calculation

$$I(P(\boldsymbol{\theta} + \delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) = \frac{1}{2}\sum_{i,j} I_{ij}(\boldsymbol{\theta})\delta\theta_i\delta\theta_j + \mathcal{O}((\delta\theta)^3), \tag{26}$$

where

$$I_{ij}(\boldsymbol{\theta}) = \int \frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j} p(\boldsymbol{x}|\boldsymbol{\theta}) \, \mathrm{d}^N \boldsymbol{x} = - \int \frac{\partial^2 \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} p(\boldsymbol{x}|\boldsymbol{\theta}) \, \mathrm{d}^N \boldsymbol{x} \tag{27}$$

are the components of the Fisher information matrix $\boldsymbol{\mathcal{I}}$, see Lehmann and Casella (1998).[3]

While for non-infinitesimal deviations $\Delta\boldsymbol{\theta}$ it generally holds that $I(P(\boldsymbol{\theta} + \Delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) \neq I(P(\boldsymbol{\theta})|P(\boldsymbol{\theta}+\Delta\boldsymbol{\theta}))$, one can easily show that $I(P(\boldsymbol{\theta}+\delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) = I(P(\boldsymbol{\theta})|P(\boldsymbol{\theta}+\delta\boldsymbol{\theta}))$ holds for infinitesimal deviations. That is, the rhs of (26) can be considered as a differential length element on the statistical manifold induced by $P(\boldsymbol{\theta})$. This is the starting point of *information geometry* by noting that I_{ij} is just the metric tensor of the statistical manifold, see Amari and Nagaoka (2000). However, for the considerations in this paper the apparatus of Riemannian geometry is not really needed. All what is needed is the distance formula (26).

As we have seen at the end of Section 2.1, a gradient ascent in the $\boldsymbol{\theta}$-parameter space on the expected value landscape $E_f(\boldsymbol{\theta})$ can result in a degeneration of the search distribution $P(\boldsymbol{\theta})$ with the result that the search gets finally restricted in subspaces of the R^N. Such a collapse might be avoided by *constraining* the search step $\delta\boldsymbol{\theta}$ in the $\boldsymbol{\theta}$ parameter space such that the information gain I is kept at a (given) small level ε, i.e. $I(P(\boldsymbol{\theta} + \delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) = \varepsilon$, while searching for the $\delta\boldsymbol{\theta}$ that provides the largest increase of $E_f(\boldsymbol{\theta})$. Using Taylor expansion for $E_f(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$ this yields the *constrained* maximization problem

$$E_f(\boldsymbol{\theta} + \delta\boldsymbol{\theta}) - E_f(\boldsymbol{\theta}) = \sum_i \frac{\partial E_f(\boldsymbol{\theta})}{\partial \theta_i} \delta\theta_i + \ldots \quad \rightarrow \quad \text{Max!} \tag{28a}$$

$$\text{s.t.} \quad I(P(\boldsymbol{\theta} + \delta\boldsymbol{\theta})|P(\boldsymbol{\theta})) = \varepsilon \tag{28b}$$

that can be solved by using Lagrange's method. Using (26) and (28), neglecting higher order $\delta\theta$ terms, one obtains the Lagrange function

$$L(\delta\boldsymbol{\theta}, \kappa) = \sum_i \frac{\partial E_f(\boldsymbol{\theta})}{\partial \theta_i} \delta\theta_i + \kappa \left(\varepsilon - \frac{1}{2} \sum_{i,j} I_{ij}(\boldsymbol{\theta}) \delta\theta_i \delta\theta_j \right). \tag{29}$$

Taking the derivatives w.r.t. $\delta\theta_k$ and κ, one gets

$$\frac{\partial L}{\partial \delta\theta_k} = \frac{\partial E_f(\boldsymbol{\theta})}{\partial \theta_k} - \kappa \sum_j I_{kj}(\boldsymbol{\theta}) \delta\theta_j, \tag{30a}$$

$$\frac{\partial L}{\partial \kappa} = \varepsilon - \frac{1}{2} \sum_{i,j} I_{ij}(\boldsymbol{\theta}) \delta\theta_i \delta\theta_j. \tag{30b}$$

Equating (30a) to zero in order to find the stationary point $\delta\hat{\boldsymbol{\theta}}$, one obtains in matrix notation $\nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}) - \kappa \boldsymbol{\mathcal{I}} \delta\hat{\boldsymbol{\theta}} = \mathbf{0}$. Solving for $\delta\hat{\boldsymbol{\theta}}$ yields

$$\delta\hat{\boldsymbol{\theta}} = \frac{1}{\kappa} \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \, \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}), \tag{31}$$

where $\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})$ is the inverse of the Fisher matrix (27). Equating (30b) to zero and inserting (31), one gets in matrix notation (note, $\boldsymbol{\mathcal{I}} = \boldsymbol{\mathcal{I}}^{\mathrm{T}} \;\Rightarrow\; \boldsymbol{\mathcal{I}}^{-1} = (\boldsymbol{\mathcal{I}}^{-1})^{\mathrm{T}}$)

$$\varepsilon = \frac{1}{2\kappa^2} \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} E_f(\boldsymbol{\theta}) \, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \, \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}). \tag{32}$$

[3]Here, calligraphic $\boldsymbol{\mathcal{I}}$ has been used in order to distinguish this matrix from the unity matrix $\mathbf{I}$.

Solving for κ and reinserting the result into (31) yields

$$\delta\hat{\boldsymbol{\theta}} = \sqrt{\frac{2\varepsilon}{\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} E_f(\boldsymbol{\theta})\, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\, \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta})}}\, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\, \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta}). \tag{33}$$

Noting that the square root in (33) is an infinitesimal quantity as $\varepsilon \to 0$, one can interpret this as an infinitesimal time change from t to $t+\delta t$ that causes a change from $\boldsymbol{\theta}(t)$ to $\boldsymbol{\theta}(t)+\delta\hat{\boldsymbol{\theta}}$ leading to $\boldsymbol{\theta}(t+\delta t) - \boldsymbol{\theta}(t) = \delta\hat{\boldsymbol{\theta}}$. Therefore, dividing by δt, the rhs of (33) can be interpreted as the time derivative of $\boldsymbol{\theta}$. Thus, one gets the ordinary differential equation (ODE) system

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\, \nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}. \tag{34}$$

This ODE system is different from Eq. (6) in that the gradient direction given by $\nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta})$ is transformed by the inverse of the Fisher information matrix. According to Amari (1998), the rhs of (34) is called "natural gradient"[4] in contrast to the ordinary gradient $\nabla_{\boldsymbol{\theta}} E_f(\boldsymbol{\theta})$.

The idea of "natural gradient" descent in the expected value landscape $E_f(\boldsymbol{\theta})$ is the key idea of the so-called "natural evolution strategies" (NES)[5] (see e.g. Wierstra et al. (2008)). The differential Eq. (34) is also referred to as "information-geometric optimization" (IGO) differential equation in Akimoto et al. (2012a). The information flow generated by this differential equation will be subject of investigation in the next section. A very astonishing result will be derived for the ellipsoid model (7) showing that the IGO flow (34) results in a *sublinear* convergence order.

3 On the Dynamics of IGO without Utility Functions

According to (Amari, 1998, p. 251),

> "the ordinary gradient does not give the steepest direction of a target function; rather, the steepest direction is given by the natural (or contravariant) gradient."

This assertion was made under the premise of a "Riemannian metric structure" (Amari, 1998, p. 251). From the analysis presented so far in this paper, this assertion seems hard to be justified under the premise of expected value maximization of (1). While the family of Gaussian distributions gives rise to a non-flat metric, it remains an open question how fast (34) approaches the steady state and the optimizer $\hat{\boldsymbol{x}}$. To shed some light on this matter, we will investigate the dynamics of IGO flow on the general quadratic function f given by (7) using Gaussian samples parameterized by $\boldsymbol{\theta} = (\overline{\boldsymbol{x}}, \mathbf{C})$, i.e. $\boldsymbol{x} \sim \boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C})$. To this end, the inverse Fisher information matrix $\boldsymbol{\mathcal{I}}^{-1}$ of the Gaussian distribution $\boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C})$ must be calculated in order to get the concrete form of (34).

3.1 Derivation of the IGO Differential Equation

As a first step, $\boldsymbol{\mathcal{I}}$ must be calculated. This calculation starting from (27) is simple, but somewhat lengthy. Therefore, we abstain from presenting it here and use the result of (Kay, 1993, p. 47)

$$I_{\alpha,\beta} = \frac{\partial\overline{\boldsymbol{x}}^{\mathrm{T}}}{\partial\theta_\alpha}\mathbf{C}^{-1}\frac{\partial\overline{\boldsymbol{x}}}{\partial\theta_\beta} + \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\theta_\alpha}\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial\theta_\beta}\right]. \tag{35}$$

Here, α and β are (multi-) indices corresponding to the components of the vector $\overline{\boldsymbol{x}}$ and the covariance matrix $\mathbf{C}$, respectively.[6] Since $\overline{\boldsymbol{x}}$ does not depend on $\mathbf{C}$ and vice versa, the resulting

[4] The term "natural gradient" is put into quotes throughout this paper because in the author's opinion there is nothing "natural" here. As have been shown, the "natural gradient" yields just a special ascent direction.

[5] NES inherits its "natural" from the "natural gradient" ascent, thus, being a rather special, i.e, synthetic choice. A better characterizing term would be "synthetic ES" (SES) instead of NES.

[6] The part of $\boldsymbol{\mathcal{I}}$ that corresponds to $\overline{\boldsymbol{x}}$ is indexed by a single index whereas the $\mathbf{C}$-matrix related part needs two indices, e.g. $\alpha = (\alpha_1, \alpha_2)$.

Fisher information (35) subdivides into a part acting on $\overline{\boldsymbol{x}}$ and another on $\mathbf{C}$. That is, the cross terms $I_{\alpha,(\beta_1\beta_2)}$ and $I_{(\alpha_1\alpha_2),\beta}$ vanish. As for the components w.r.t. $\overline{\boldsymbol{x}}$ one immediately gets from (35)

$$\overline{\boldsymbol{x}}: \quad I_{\alpha,\beta} = C^{-1}_{\alpha\beta}. \tag{36}$$

Treating the $\mathbf{C}$-related part in (35) using $\frac{\partial C_{ab}}{\partial C_{cd}} = \frac{1}{2}\left(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}\right)$ (here, symmetry of $\mathbf{C}$ has been taken into account) yields

$$\begin{aligned} I_{(\alpha_1\alpha_2),(\beta_1\beta_2)} &= \frac{1}{2}\sum_{k,l,m,n} C^{-1}_{kl}\frac{\partial C_{lm}}{\partial C_{\alpha_1\alpha_2}}C^{-1}_{mn}\frac{\partial C_{nk}}{\partial C_{\beta_1\beta_2}} \\ &= \frac{1}{8}\sum_{k,l,m,n} C^{-1}_{kl}(\delta_{l\alpha_1}\delta_{m\alpha_2} + \delta_{l\alpha_2}\delta_{m\alpha_1})C^{-1}_{mn}(\delta_{n\beta_1}\delta_{k\beta_2} + \delta_{n\beta_2}\delta_{k\beta_1}). \end{aligned} \tag{37}$$

Thus, one gets for the $\mathbf{C}$-related part of $\boldsymbol{\theta}$ (taking the symmetry of $\mathbf{C}^{-1}$ into account; for an alternative derivation, see Appendix A in the supplement material)

$$\mathbf{C}: \quad I_{(\alpha_1\alpha_2),(\beta_1\beta_2)} = \frac{1}{4}\left(C^{-1}_{\alpha_1\beta_1}C^{-1}_{\alpha_2\beta_2} + C^{-1}_{\alpha_1\beta_2}C^{-1}_{\alpha_2\beta_1}\right). \tag{38}$$

Due to the block structure of $\boldsymbol{\mathcal{I}}$, calculating the inverse of $\boldsymbol{\mathcal{I}}$ reduces to the calculation of the inverse of (36) and (38) separately. Considering (36), one immediately concludes that

$$\overline{\boldsymbol{x}}: \quad I^{-1}_{\alpha,\beta} = C_{\alpha\beta}. \tag{39}$$

The correctness of

$$\mathbf{C}: \quad I^{-1}_{(\alpha_1\alpha_2),(\beta_1\beta_2)} = 2C_{\alpha_1\beta_2}C_{\beta_1\alpha_2} \tag{40}$$

is proven directly by checking

$$\sum_{\beta_1,\beta_2} I^{-1}_{(\alpha_1\alpha_2),(\beta_1\beta_2)}\, I_{(\beta_1\beta_2),(\gamma_1\gamma_2)} = \frac{1}{2}\left(\delta_{\alpha_1\gamma_2}\delta_{\alpha_2\gamma_1} + \delta_{\alpha_1\gamma_1}\delta_{\alpha_2\gamma_2}\right).$$

Now, the IGO differential equation can be directly obtained from (34) using (39) and (40). It reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\overline{\boldsymbol{x}}(t)\\ \mathbf{C}(t)\end{pmatrix} = \begin{pmatrix}\mathbf{C}(t)\,\nabla_{\overline{\boldsymbol{x}}}E_f\big(\boldsymbol{\theta}(t)\big)\\ 2\mathbf{C}(t)\,\nabla_{\mathbf{C}}E_f\big(\boldsymbol{\theta}(t)\big)\,\mathbf{C}(t)\end{pmatrix}. \tag{41}$$

While the $\overline{\boldsymbol{x}}$-dynamics were directly obtained as a matrix vector product, the $\mathbf{C}$-dynamics needed an intermediate step:

$$\sum_{\beta_1,\beta_2} I^{-1}_{(\alpha_1\alpha_2),(\beta_1\beta_2)}\frac{\partial E_f(\boldsymbol{\theta})}{\partial C_{\beta_1\beta_2}} = 2\sum_{\beta_1,\beta_2} C_{\alpha_1\beta_2}C_{\beta_1\alpha_2}\frac{\partial E_f(\boldsymbol{\theta})}{\partial C_{\beta_1\beta_2}} = \big(2\mathbf{C}\,\nabla_{\mathbf{C}}E_f(\boldsymbol{\theta})\,\mathbf{C}\big)_{\alpha_1\alpha_2}.$$

Using the gradients already calculated in (9) one finally obtains the non-linear IGO differential equation system

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}(t)}{\mathrm{d}t} = \mathbf{C}(t)\big(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}(t)\big), \tag{42a}$$

$$\frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} = -2\mathbf{C}(t)\mathbf{Q}\mathbf{C}(t). \tag{42b}$$

In the next section, a solution to this system (42) will be derived.

3.2 On the Time Evolution of the IGO System

A closer look at (42) reveals that (42b) is independent of (42a). That is, in a first step (42b) must be solved.

Theorem 1 (IGO C-dynamics). *The non-linear ordinary differential equation (ODE) system* $\frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} = -2\mathbf{C}(t)\mathbf{Q}\mathbf{C}(t)$ *with the symmetric matrices* $\mathbf{Q}$ *and* $\mathbf{C}(t)$*, both invertible, and the initial condition* $\mathbf{C}(0) = \mathbf{C}_0$ *has the solution*

$$\mathbf{C}(t) = \left(\mathbf{C}_0^{-1} + 2t\mathbf{Q}\right)^{-1}. \tag{43}$$

Proof. Let $\mathbf{G}(t)$ be an invertible matrix. Let $\mathbf{C}(t) = \mathbf{Q}^{-1}\mathbf{G}(t)$ and substitute this expression in (42b), then multiply from the left with $\mathbf{Q}$, this yields the ODE

$$\frac{\mathrm{d}\mathbf{G}(t)}{\mathrm{d}t} = -2\mathbf{G}^2(t). \tag{44}$$

Now, differentiate the identity $\frac{\mathrm{d}\mathbf{I}(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{G}\mathbf{G}^{-1}) = \frac{\mathrm{d}\mathbf{G}}{\mathrm{d}t}\mathbf{G}^{-1} + \mathbf{G}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{G}^{-1} = \mathbf{0}$ and multiply from the right with $\mathbf{G}$. This yields after rearrangement

$$\frac{\mathrm{d}\mathbf{G}}{\mathrm{d}t} = -\mathbf{G}\frac{\mathrm{d}\mathbf{G}^{-1}}{\mathrm{d}t}\mathbf{G}. \tag{45}$$

Plugging this result into (44) and multiplication from the left and the right with $\mathbf{G}^{-1}$ yields

$$\frac{\mathrm{d}\mathbf{G}^{-1}}{\mathrm{d}t} = 2\mathbf{I}. \tag{46}$$

This ODE system can be easily solved yielding $\mathbf{G}^{-1} = \mathbf{K} + 2t\mathbf{I}$, where $\mathbf{K}$ is a constant (time independent) matrix. Thus, one gets $\mathbf{G}(t) = (\mathbf{K} + 2t\mathbf{I})^{-1}$. Recalling that $\mathbf{C}(t) = \mathbf{Q}^{-1}\mathbf{G}(t)$ was originally substituted, one gets

$$\mathbf{C}(t) = \mathbf{Q}^{-1}(\mathbf{K} + 2t\mathbf{I})^{-1} = \left[(\mathbf{K} + 2t\mathbf{I})\mathbf{Q}\right]^{-1}. \tag{47}$$

The constant matrix $\mathbf{K}$ is obtained by considering the initial condition $\mathbf{C}(0) = \mathbf{C}_0$. Using (47) one gets $\mathbf{C}(0) = (\mathbf{K}\mathbf{Q})^{-1} \stackrel{!}{=} \mathbf{C}_0$ and therefore $\mathbf{K} = \mathbf{C}_0^{-1}\mathbf{Q}^{-1}$. Introduced in the rhs of (47) finally yields (43). □

Rewriting the $\mathbf{C}$ dynamic (43) for $t > 0$, one gets for the asymptotic $t \to \infty$ behavior

$$\mathbf{C}(t) = \frac{1}{2t}\left(\frac{1}{2t}\mathbf{C}_0^{-1} + \mathbf{Q}\right)^{-1} \quad \Longrightarrow \quad \boxed{\mathbf{C}(t) \simeq \frac{1}{2t}\mathbf{Q}^{-1}.} \tag{48}$$

That is, for sufficiently large time, IGO forgets the initial covariance matrix and approaches a matrix that is proportional to the inverse of the $\mathbf{Q}$ matrix as postulated by (24). However, this approach is rather slow since it obeys an $1/t$ law. The question arises whether this slow approach transfers also to the $\overline{\boldsymbol{x}}$-dynamics.

In order to derive a solution for $\overline{\boldsymbol{x}}$ it should be first noted that (42a) has a fixed point (attractor), characterized by $\frac{\mathrm{d}\overline{\boldsymbol{x}}(t)}{\mathrm{d}t} = \mathbf{0}$, which is determined by the solution of $\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}(t) = \mathbf{0}$ already calculated in Eq. (11). That is, $\overline{\boldsymbol{x}}$ approaches the maximizer $\hat{\boldsymbol{x}}$ of $f(\boldsymbol{x})$ for $t \to \infty$. Let us investigate the approach to the maximizer by considering the evolution of the deviation $\boldsymbol{z} = \overline{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}$. Using (11), the ODE (42a) becomes

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = \mathbf{C}(t)\left(\mathbf{Q}\mathbf{Q}^{-1}\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}(t)\right) = \mathbf{C}(t)\mathbf{Q}\left(2\hat{\boldsymbol{x}} - 2\overline{\boldsymbol{x}}(t)\right) = -2\mathbf{C}(t)\mathbf{Q}\boldsymbol{z}(t). \tag{49}$$

Inserting (47) with $\mathbf{K} = \mathbf{C}_0^{-1}\mathbf{Q}^{-1}$ into (49) and applying matrix algebra, one gets step-by-step

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} &= -2\left[\left(\mathbf{C}_0^{-1}\mathbf{Q}^{-1} + 2t\mathbf{I}\right)\mathbf{Q}\right]^{-1}\mathbf{Q}\boldsymbol{z}(t) \\ &= -2\left[\mathbf{Q}^{-1}\left(\mathbf{C}_0^{-1}\mathbf{Q}^{-1} + 2t\mathbf{I}\right)\mathbf{Q}\right]^{-1}\boldsymbol{z}(t) \\ &= -2\left[\mathbf{Q}^{-1}\left(\mathbf{C}_0^{-1} + 2t\mathbf{Q}\right)\right]^{-1}\boldsymbol{z}(t) \\ &= \underbrace{-2\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\right)^{-1}}_{:=\mathbf{D}(t)}\boldsymbol{z}(t). \end{aligned} \tag{50}$$

This is a linear ODE system with a time dependent coefficient matrix $\mathbf{D}(t)$ that can be solved using the *Ansatz*

$$\boldsymbol{z}(t) = \mathrm{e}^{\mathbf{B}(t)}\boldsymbol{z}_0. \tag{51}$$

Calculating the time derivative of (51), one gets

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = \frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}\mathrm{e}^{\mathbf{B}(t)}\boldsymbol{z}_0 = \frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}\boldsymbol{z}. \tag{52}$$

Here it was implicitly assumed that $\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t}$ and $\mathbf{B}$ commute.[7] Comparing (52) with (50) one obtains the ODE

$$\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}t} = -2\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\right)^{-1}. \tag{53}$$

This ODE can be formally integrated using the matrix logarithm yielding $\mathbf{B}(t) = \ln\mathbf{B}_0 - \ln\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\right)$ where $\mathbf{B}_0$ is a constant matrix to be determined below. Inserting this result in (51), one gets

$$\boldsymbol{z}(t) = \mathbf{B}_0\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\right)^{-1}\boldsymbol{z}_0. \tag{54}$$

Taking the initial condition $\boldsymbol{z}(0) = \boldsymbol{z}_0$ into account, one gets using (54)

$$\boldsymbol{z}(0) = \mathbf{B}_0\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\right)^{-1}\boldsymbol{z}_0 \quad\Longrightarrow\quad \mathbf{B}_0\mathbf{C}_0\mathbf{Q} = \mathbf{I} \quad\Longrightarrow\quad \mathbf{B}_0 = \mathbf{Q}^{-1}\mathbf{C}_0^{-1}. \tag{55}$$

Plugging this into (54) it follows the

Theorem 2 (IGO residual distance to optimizer dynamics). *Consider the IGO system (42) where samples are generated using Gaussian normal vectors $\boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C})$ and the fitness is given by the ellipsoid model (7). Then the residual distance $\boldsymbol{z} = \overline{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}$ to the optimizer $\hat{\boldsymbol{x}}$ of the ellipsoid model (7) obeys the ODE $\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = -2\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\right)^{-1}\boldsymbol{z}(t)$ and its solution is given by*

$$\boldsymbol{z}(t) = \mathbf{Q}^{-1}\mathbf{C}_0^{-1}\left(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\right)^{-1}\boldsymbol{z}_0. \tag{56}$$

Proof. Since the ODE regarding $\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t}$ has been already derived in Eq. (50), it remains to insert the solution (56) into both sides of (50) and to show their equality. To this end, (56) is expressed in terms of the $\mathbf{D}$ matrix defined in (50)

$$\boldsymbol{z}(t) = -\frac{1}{2}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\mathbf{D}\boldsymbol{z}_0. \tag{57}$$

Let us first calculate the lhs of (50) by making use of (45) (replacing $\mathbf{B}$ by $\mathbf{D}$)

$$\mathrm{lhs} = \frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} = -\frac{1}{2}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\frac{\mathrm{d}\mathbf{D}}{\mathrm{d}t}\boldsymbol{z}_0 = \frac{1}{2}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\mathbf{D}\frac{\mathrm{d}\mathbf{D}^{-1}}{\mathrm{d}t}\mathbf{D}\boldsymbol{z}_0. \tag{58}$$

[7] This assumption and also the formal integration step presented below will get their justification by finally proving that the resulting solution fulfills (50).

Since

$$\frac{\mathrm{d}\mathbf{D}^{-1}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(-\frac{1}{2}\big(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\big)\right) = -\mathbf{I},$$

one gets for (58)

$$\mathrm{lhs} = -\frac{1}{2}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\mathbf{D}\mathbf{D}\boldsymbol{z}_0. \tag{59}$$

The rhs of (50) yields

$$\mathrm{rhs} = \mathbf{D}\boldsymbol{z} = -\frac{1}{2}\mathbf{D}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\mathbf{D}\boldsymbol{z}_0. \tag{60}$$

Equality of (59) and (60) is proven if $\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\mathbf{D}$ is equal to $\mathbf{D}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}$. This transfers to their inverses, i.e. $\mathbf{D}^{-1}\mathbf{C}_0\mathbf{Q}$ and $\mathbf{C}_0\mathbf{Q}\mathbf{D}^{-1}$. Direct calculation yields

$$\mathbf{D}^{-1}\mathbf{C}_0\mathbf{Q} = -\frac{1}{2}\big(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\big)\mathbf{C}_0\mathbf{Q} = -\frac{1}{2}\big(\mathbf{I} + 2t\mathbf{C}_0\mathbf{Q}\big), \tag{61a}$$

$$\mathbf{C}_0\mathbf{Q}\mathbf{D}^{-1} = -\frac{1}{2}\mathbf{C}_0\mathbf{Q}\big(\mathbf{Q}^{-1}\mathbf{C}_0^{-1} + 2t\mathbf{I}\big) = -\frac{1}{2}\big(\mathbf{I} + 2t\mathbf{C}_0\mathbf{Q}\big). \tag{61b}$$

Since (61a) and (61b) agree, Eqs. (59) and (60) agree as well and thus, $\mathrm{lhs} = \mathrm{rhs}$ completes the proof. □

3.3 Discussion

Let us have a closer look at Eq. (56) and investigate the asymptotic $t \to \infty$ behavior

$$\boldsymbol{z}(t) = \frac{1}{2t}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\left(\frac{\mathbf{Q}^{-1}\mathbf{C}_0^{-1}}{2t} + \mathbf{I}\right)^{-1}\boldsymbol{z}_0 \quad \Longrightarrow \quad \boxed{\boldsymbol{z}(t) \simeq \frac{1}{2t}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0.} \tag{62}$$

This is a very astonishing result for the "natural gradient" ascent: The optimizer $\hat{\boldsymbol{x}}$ is approached very slowly obeying an $1/t$-law being in contrast to the exponential decay of the ordinary gradient ascent given by Eq. (14). There is also another difference concerning the manner in which the optimizer is approached: According to (62) the IGO flow follows a straight line the direction of which is given by $\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0$. That is, each component of the deviation vector $\boldsymbol{z}$ changes in proportional manner. This can also be seen using (49) in conjunction with the asymptotic Eq. (48) yielding

$$\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} \simeq -\frac{1}{t}\boldsymbol{z}(t). \tag{63}$$

This is in contrast to the ordinary gradient ascent dynamics (13) that locally transforms the ascent direction by the $\mathbf{Q}$ matrix. The latter yields $\boldsymbol{z}(t)$ trajectories (14) being bended in the $\overline{\boldsymbol{x}}$ space while the "natural gradient" case produces a straight line. In that sense one could argue that in the asymptotic limit the "natural gradient" ascent proceeds along a "geodesic" in a flat $\overline{\boldsymbol{x}}$-space. Interestingly, the search behavior of $(\mu/\mu_I, \lambda)$-ES with σ-self-adaptation does also exhibit such a search behavior when considering the expected value dynamics in the steady state, see Beyer and Melkozerov (2013). However, following a geodesic does *not* necessarily guarantee a fast approach to the optimizer $\hat{\boldsymbol{x}}$. While the $(\mu/\mu_I, \lambda)$-ES approaches $\hat{\boldsymbol{x}}$ exponentially fast (thus, exhibiting linear convergence order in expectation), the IGO dynamics (42a) yields according to (62) only a disappointingly slow $1/t$ behavior, i.e. sublinear convergence order. While this result seems astonishing at first glance, it does not really come as a surprise when considering the premises under which the "natural gradient" has been derived: The "natural gradient" direction is a result of the constrained optimization problem (28) that "penalizes" the steepest ascent of $E_f(\boldsymbol{\theta})$ such that the information gain (28b) is fixed at a small value ε. That is, constraining the

information gain (24) results necessarily in a slowly changing search distribution (48). The real surprise is, however, that using the "natural gradient" does change the convergence behavior in such a drastic manner. Since state-of-the-art implementations of NESs do not exhibit such a sublinear convergence behavior, another "ingredient" of these strategies comes into focus: the application of "utility" functions.

4 NES/IGO Localized – On the Use of "Utility" Transforms

4.1 Assessing Utility by Individual Ranking

While the NES approach has a certain "scientific appeal," taking it too literally, one ends up with slowly converging algorithms on quadratic functions. Actually, even one of the first publications on NES, Wierstra et al. (2008), did not literally implement the NES update (5) to perform the maximization of the expected value (1). Instead, *transformed* fitness functions have been introduced replacing $f(\boldsymbol{x})$ in (1). This technique has been called *fitness shaping* in Wierstra et al. (2008). In latter publications it was emphasized that the original NES algorithm "converges slowly or even prematurely" (Glasmachers et al., 2010, p. 393), thus, supporting the new theoretical findings of Sect. 3 from the empirical perspective. In recent versions of NES, as e.g. exponential NES, fitness shaping is realized by assigning *utility* values to the individuals sampled, i.e. weights, reflecting the *ranking* of those individuals. Thus, NES became increasingly similar to classical Evolution Strategies, see Beyer and Schwefel (2002); Hansen et al. (2003); Beyer (2007). It is *very important* to stress once more: The use of fitness shaping techniques *cannot be deduced* from the "natural gradient" ascent philosophy. It must be introduced in an *ad hoc* manner. In literature, its use is mainly justified by rendering "the algorithm invariant under monotonically growing (i.e. rank preserving) transformations of the fitness function" (Glasmachers et al., 2010, p. 394). The same argument is used to explain the application of weighted (μ, λ)-selection in CMA-ES, see Hansen (2006). Therefore, it does not come as a surprise that in the fully-developed IGO framework, see Akimoto et al. (2012a), a rank-based utility function $W_f(\boldsymbol{x})$ replaces the $f(\boldsymbol{x})$ in (1). Thus, (1) changes to

$$E_W(\boldsymbol{\theta}) := \mathrm{E}[W|\boldsymbol{\theta}] = \int W_f(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}) \, \mathrm{d}^N\boldsymbol{x}. \tag{64}$$

Here, it is important to notice that $W_f(\boldsymbol{x}|\boldsymbol{\theta})$ itself depends on the distribution parameters $\boldsymbol{\theta}$. That is, the *globally* defined fitness function $f(\boldsymbol{x})$ is replaced by a *locally* acting function $W_f(\boldsymbol{x}|\boldsymbol{\theta})$. The local character of this weighting function becomes immediately clear when looking at standard ES with (μ, λ)-truncation selection. There, λ offspring $\boldsymbol{x}_l$ $(l = 1, \ldots, \lambda)$ are generated according to the offspring distribution $\boldsymbol{x}_l \sim \boldsymbol{\mathcal{N}}(\overline{\boldsymbol{x}}, \mathbf{C})$. The best μ individuals (i.e. those with the largest f-values in the case of maximization) are taken with weights $1/\mu$ and the rest gets weights equal to zero[8, 9]

$$W_f(\boldsymbol{x}_{[m:\lambda]}|\boldsymbol{\theta}) = \begin{cases} \frac{\lambda}{\mu}, & \text{if } \lambda - \mu + 1 \leq m \leq \lambda, \\ 0, & \text{if } 1 \leq m \leq \lambda - \mu. \end{cases} \tag{65}$$

This ranking depends clearly on the choice of the distribution parameters $\boldsymbol{\theta}$. That is, $W_f(\boldsymbol{x}|\boldsymbol{\theta})$ changes with every iteration step. Thus, in the continuous limit it becomes a function of time. As a replacement for $f(\boldsymbol{x})$, it can be regarded as some kind of *local* fitness. Climbing up

[8] Here, the concomitants notation $[m : \lambda]$ of order statistics has been used where "m" indicates that individual with the f-value being in the mth position of the ascendingly ordered λ f-values, see e.g. Arnold et al. (1992).

[9] Note, in the first line of (65) λ/μ has been written instead of $1/\mu$. This guarantees that the corresponding Monte-Carlo approximation (4) yields a update formula for the $\boldsymbol{x}$-vectors that agrees with the intermediate multi-recombination used in standard $(\mu/\mu_I, \lambda)$-ES.

the expected value landscape (1) is therefore governed by the $\boldsymbol{\theta}$ gradient acting on the density function $p(\boldsymbol{x}|\boldsymbol{\theta})$ exclusively. That is, instead of the IGO differential equation (34), one now has to consider

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\,\nabla_{\boldsymbol{\theta}} E_W(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)} \tag{66}$$

(Ollivier et al., 2011) and the gradient must be evaluated according to

$$\nabla_{\boldsymbol{\theta}} E_W(\boldsymbol{\theta}) = \int W_f(\boldsymbol{x}|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})\,\mathrm{d}^N\boldsymbol{x}. \tag{67}$$

That is, the gradient operator *must not* act on $W_f(\boldsymbol{x}|\boldsymbol{\theta})$.

Besides (μ,λ)-selection (65), rank-based weights putting non-linear emphasis on the best individuals are state-of-the-art in CMA-ES. For example, Hansen (2006) proposed the heuristic formula

$$W_f(\boldsymbol{x}_{[m:\lambda]}) = \begin{cases} \lambda \frac{\ln(\mu+1)-\ln(\lambda-m+1)}{\sum_{m=1}^{\mu}\left(\ln(\mu+1)-\ln(m)\right)}, & \text{if } \lambda-\mu+1 \le m \le \lambda, \\ 0, & \text{if } 1 \le m \le \lambda-\mu. \end{cases} \tag{68}$$

that puts different weights on the μ best individuals (i.e., those with the largest f-values). Moreover, ES theory even allows for the determination of *optimal* weights based on asymptotical sphere model assumptions. Arnold (2006) has shown for ESs with isotropic mutations that choosing

$$W_f(\boldsymbol{x}_{[m:\lambda]}) = \mathrm{E}[z_{m:\lambda}], \quad z \sim \mathcal{N}(0,1), \tag{69}$$

guarantees maximal progress towards the optimizer provided that the mutation strength is controlled correctly. Here the lhs in (69) is the expected value of the mth order statistics of the standard normal variate given by the integral

$$\mathrm{E}[z_{m:\lambda}] = \frac{\lambda!}{(m-1)!(\lambda-m)!}\int_{-\infty}^{\infty} z\phi(z)[\Phi(z)]^{m-1}[1-\Phi(z)]^{\lambda-m}\,\mathrm{d}z, \tag{70}$$

where $\phi(z)$ and $\Phi(z)$ are the pdf and cdf of the standard normal variate, respectively. For large λ (i.e. large sample sizes), (70) can be expressed asymptotically (Arnold et al., 1992, p. 128, Eq. (5.5.2))

$$\mathrm{E}[z_{m:\lambda}] \simeq \Phi^{-1}\left(\frac{m}{\lambda+1}\right), \tag{71}$$

where Φ^{-1} is the quantile function of the standard normal variate. These optimal weights are different to the heuristic weights (65) in that each of the λ individuals gets a weight and not only the best μ individuals. Unlike (68) and (65) the weight equation (69) yields also negative weights for individuals those f-values are below the median of f. The weights are anti-symmetric about the median, i.e. it holds $W_f(\boldsymbol{x}_{[k:\lambda]}) = -W_f(\boldsymbol{x}_{[\lambda+1-k:\lambda]})$.

In the remaining part of this section, the effect of the local weight models on the dynamics of ES will be investigated considering linear and quadratic fitness models (7). It turns out that using optimal weights results in surprisingly simple expressions that allow for closed solutions of the IGO differential equation.

4.2 Asymptotic Fitness Distributions and Search Gradients

In order to investigate the dynamics of IGO with local weighting, one first has to calculate the search gradient (67) – a rather difficult task due to the N-dimensional integral. Taking into

account, however, that the local weight function depends only on the f-values, (67) can be transformed to an one-dimensional integral by the mapping $\boldsymbol{x} \mapsto f$. Thus, (67) changes to

$$\nabla_{\boldsymbol{\theta}} E_W(\boldsymbol{\theta}) = \int W_f(f|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} p(f|\boldsymbol{\theta}) \, \mathrm{d}f. \tag{72}$$

This integral is tractable provided that the pdf $p(f|\boldsymbol{\theta})$ can be expressed by simple expressions, e.g. in terms of a Gaussian distribution

$$p(f|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_f(\boldsymbol{\theta})} \exp\left[-\frac{1}{2} \left(\frac{f - \overline{f}(\boldsymbol{\theta})}{\sigma_f(\boldsymbol{\theta})} \right)^2 \right]. \tag{73}$$

This holds exactly for linear fitness functions $f(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$. Provided that the eigenvalue spectrum of $\mathbf{Q}$ of the quadratic fitness function (7) does not contain singularly dominating eigenvalues, (73) holds also for the quadratic case if $N \to \infty$, due to the central limit theorem of statistics. In order to proceed under these assumptions, the expected value $\overline{f}$ and the variance $\mathrm{Var}[f] = \sigma_f^2$ must be determined. To this end, the local fitness function is considered that follows from (7) by substituting $\boldsymbol{x} = \overline{\boldsymbol{x}} + \boldsymbol{z}$ and $\boldsymbol{z} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{C})$

$$\begin{aligned} f(\boldsymbol{z} + \overline{\boldsymbol{x}}) &= \boldsymbol{a}^{\mathrm{T}}(\overline{\boldsymbol{x}} + \boldsymbol{z}) - (\overline{\boldsymbol{x}} + \boldsymbol{z})^{\mathrm{T}}\mathbf{Q}(\overline{\boldsymbol{x}} + \boldsymbol{z}) \\ &= \boldsymbol{a}^{\mathrm{T}}\overline{\boldsymbol{x}} - \overline{\boldsymbol{x}}^{\mathrm{T}}\mathbf{Q}\overline{\boldsymbol{x}} + \big(\boldsymbol{a}^{\mathrm{T}} - 2\overline{\boldsymbol{x}}^{\mathrm{T}}\mathbf{Q}\big)\boldsymbol{z} - \boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\boldsymbol{z}. \end{aligned} \tag{74}$$

Since $\mathrm{E}[\boldsymbol{z}] = \mathbf{0}$, one only has to consider $-\mathrm{E}[\boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\boldsymbol{z}]$ that yields $-\mathrm{Tr}[\mathbf{QC}]$ (Beyer, 2001, p. 122). Thus, one gets

$$\overline{f}(\boldsymbol{\theta}) = \boldsymbol{a}^{\mathrm{T}}\overline{\boldsymbol{x}} - \overline{\boldsymbol{x}}^{\mathrm{T}}\mathbf{Q}\overline{\boldsymbol{x}} - \mathrm{Tr}[\mathbf{QC}]. \tag{75}$$

As for the variance of (74) it is noted that constant terms do not contribute to the variance of a random variate. Therefore

$$\mathrm{Var}[f] = \mathrm{Var}\big[\underbrace{\big(\boldsymbol{a}^{\mathrm{T}} - 2\overline{\boldsymbol{x}}^{\mathrm{T}}\mathbf{Q}\big)}_{:=\tilde{\boldsymbol{a}}^{\mathrm{T}}} \boldsymbol{z} - \boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\boldsymbol{z} \big] \tag{76}$$

Making use of the standard deviation formula (4.59) in (Beyer, 2001, p. 123), one finds *mutatis mutandis*

$$\sigma_f(\boldsymbol{\theta}) = \sqrt{\big(\boldsymbol{a}^{\mathrm{T}} - 2\overline{\boldsymbol{x}}^{\mathrm{T}}\mathbf{Q}\big)\mathbf{C}\big(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}\big) + 2\mathrm{Tr}[(\mathbf{QC})^2]}. \tag{77}$$

Note, Eq. (77) contains also the case of linear fitness as special case by setting $\mathbf{Q} = \mathbf{0}$ yielding $\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}\boldsymbol{a}}$.

The next step concerns the calculation of the gradient in (72). To this end, it is noted that $\nabla_{\boldsymbol{\theta}} p(f|\boldsymbol{\theta}) = p(f|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ln p(f|\boldsymbol{\theta})$. Thus (72) changes to

$$\nabla_{\boldsymbol{\theta}} E_W(\boldsymbol{\theta}) = \int W_f(f|\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \ln p(f|\boldsymbol{\theta}) \; p(f|\boldsymbol{\theta}) \, \mathrm{d}f. \tag{78}$$

Taking the logarithm of (73), one gets

$$\ln p(f|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi - \ln \sigma_f(\boldsymbol{\theta}) - \frac{1}{2}\left(\frac{f - \overline{f}(\boldsymbol{\theta})}{\sigma_f(\boldsymbol{\theta})} \right)^2 \tag{79}$$

and applying the $\boldsymbol{\theta}$ gradient calculation yields

$$\nabla_{\boldsymbol{\theta}} \ln p(f|\boldsymbol{\theta}) = \frac{1}{\sigma_f}\left[-\nabla_{\boldsymbol{\theta}}\sigma_f + \left(\frac{f - \overline{f}}{\sigma_f} \right) \nabla_{\boldsymbol{\theta}}\overline{f} + \left(\frac{f - \overline{f}}{\sigma_f} \right)^2 \nabla_{\boldsymbol{\theta}}\sigma_f \right]. \tag{80}$$

In order to complete (80), the gradients of $\overline{f}$ and σ_f are to be calculated. Comparing (75) with (8) taking (9) into account, one immediately gets

$$\nabla_{\overline{\boldsymbol{x}}}\overline{f} = \boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}} \qquad \text{and} \qquad \nabla_{\mathbf{C}}\overline{f} = -\mathbf{Q}. \tag{81}$$

Straightforward calculation using (77) yields

$$\nabla_{\overline{\boldsymbol{x}}}\sigma_f = -\frac{2}{\sigma_f}\mathbf{Q}\mathbf{C}\big(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}\big). \tag{82}$$

Concerning the $\mathbf{C}$ gradient, a detailed component-wise calculation is presented:

$$\begin{aligned}
\frac{\partial\sigma_f}{\partial C_{mn}} &= \frac{1}{2}\frac{1}{\sigma_f}\frac{\partial}{\partial C_{mn}}\left(\sum_{i,j,k,l}(a_i - 2\overline{x}_k Q_{ki})C_{ij}(a_j - 2\overline{x}_l Q_{lj}) + 2Q_{ij}C_{jk}Q_{kl}C_{li}\right)\\
&= \frac{1}{2}\frac{1}{\sigma_f}\sum_{i,j,k,l}\Bigg((a_i - 2\overline{x}_k Q_{ki})\frac{1}{2}(\delta_{im}\delta_{jn} + \delta_{in}\delta_{jm})(a_j - 2\overline{x}_l Q_{lj})\\
&\qquad + Q_{ij}(\delta_{jm}\delta_{kn} + \delta_{jn}\delta_{km})Q_{kl}C_{li} + Q_{ij}C_{jk}Q_{kl}(\delta_{lm}\delta_{in} + \delta_{ln}\delta_{im})\Bigg)\\
&= \frac{1}{2}\frac{1}{\sigma_f}\sum_{k,l}\big((a_m - 2\overline{x}_k Q_{km})(a_n - 2\overline{x}_l Q_{ln}) + 4Q_{mk}C_{kl}Q_{ln}\big)
\end{aligned} \tag{83}$$

Rewriting in matrix vector notation yields

$$\nabla_{\mathbf{C}}\sigma_f = \frac{1}{2}\frac{1}{\sigma_f}\big((\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})^{\mathrm{T}} + 4\mathbf{Q}\mathbf{C}\mathbf{Q}\big). \tag{84}$$

The results obtained are to be inserted into (80) and then into (72). Since the final substitution step depends on the weighting function W_f, it will be performed in separate sections focussing on optimal weights (71) and truncation selection weights (65), respectively.

4.3 Dynamics of IGO with Arnold's Optimal Weights

In order to analyze the IGO flow (66) it shall be recalled that this differential equation describes an infinite population model, i.e., $\lambda \to \infty$. That is, in expectation the rank m of an individual can be directly inferred from its f-value, i.e., $m_f = \lambda P(f|\boldsymbol{\theta})$ where $P(f|\boldsymbol{\theta})$ is the (local) cumulative distribution function of the fitness samples. Assuming normality given by (73), this transfers to

$$m_f = \lambda\,\Phi\left(\frac{f - \overline{f}(\boldsymbol{\theta})}{\sigma_f(\boldsymbol{\theta})}\right). \tag{85}$$

Plugging this into Arnold's optimal (sphere) weight formula (69) using the asymptotically exact version (71), one ends up with the surprisingly simple expression (for $\lambda \to \infty$)

$$W_f(f|\boldsymbol{\theta}) = \frac{f - \overline{f}(\boldsymbol{\theta})}{\sigma_f(\boldsymbol{\theta})}. \tag{86}$$

That is, Arnold's optimal weights transform to a utility function being simply the *local standardization* of $f(\boldsymbol{x})$ in the infinite population size model. This result shares also a similarity with the so-called *fitness baseline* method found in older versions of NES, see Wierstra et al. (2008). However, the standardization by $\sigma_f(\boldsymbol{\theta})$ is missing there. As we will see below, this is a difference that causes considerable differences in the convergence behavior of the ES.

The calculation of the gradient (78) can be completed now. First, using (80) and (73) one gets

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{1}{\sqrt{2\pi\sigma_f^2}} \int W_f(f|\boldsymbol{\theta}) \left[-\nabla_{\boldsymbol{\theta}}\sigma_f + \left(\frac{f-\overline{f}}{\sigma_f}\right) \nabla_{\boldsymbol{\theta}}\overline{f} + \left(\frac{f-\overline{f}}{\sigma_f}\right)^2 \nabla_{\boldsymbol{\theta}}\sigma_f \right] \times \exp\left[-\frac{1}{2}\left(\frac{f-\overline{f}}{\sigma_f}\right)^2 \right] \mathrm{d}f. \tag{87}$$

Inserting the result (86), one obtains

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{1}{\sqrt{2\pi\sigma_f^2}} \int \frac{f-\overline{f}}{\sigma_f} \left[-\nabla_{\boldsymbol{\theta}}\sigma_f + \left(\frac{f-\overline{f}}{\sigma_f}\right) \nabla_{\boldsymbol{\theta}}\overline{f} + \left(\frac{f-\overline{f}}{\sigma_f}\right)^2 \nabla_{\boldsymbol{\theta}}\sigma_f \right] \times \exp\left[-\frac{1}{2}\left(\frac{f-\overline{f}}{\sigma_f}\right)^2 \right] \mathrm{d}f. \tag{88}$$

Performing a variable transformation $t := (f-\overline{f})/\sigma_f$, (88) changes to

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{1}{\sigma_f}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(-\nabla_{\boldsymbol{\theta}}\sigma_f \; t + \nabla_{\boldsymbol{\theta}}\overline{f} \; t^2 + \nabla_{\boldsymbol{\theta}}\sigma_f \; t^3\right) \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t. \tag{89}$$

Since the integrals over odd t powers vanish, one obtains

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{1}{\sigma_f(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}}\overline{f}(\boldsymbol{\theta}). \tag{90}$$

Using (81) one gets

$$\nabla_{\boldsymbol{\theta}} E_W(\boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\overline{\boldsymbol{x}}} E_W \\ \nabla_{\mathbf{C}} E_W \end{pmatrix} = \frac{1}{\sigma_f(\boldsymbol{\theta})} \begin{pmatrix} \boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}} \\ -\mathbf{Q} \end{pmatrix}. \tag{91}$$

Now, the IGO differential equations (66) can be set up using (41) (replacing E_f by E_W)

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}(t)}{\mathrm{d}t} = \frac{\mathbf{C}(t)\big(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}(t)\big)}{\sigma_f\big(\overline{\boldsymbol{x}}(t), \mathbf{C}(t)\big)}, \tag{92a}$$

$$\frac{\mathrm{d}\mathbf{C}(t)}{\mathrm{d}t} = -2\frac{\mathbf{C}(t)\mathbf{Q}\mathbf{C}(t)}{\sigma_f\big(\overline{\boldsymbol{x}}(t), \mathbf{C}(t)\big)}, \tag{92b}$$

where σ_f is given by (77). Finding a closed solution to the non-linear ODE system (92) seems generally excluded. However, an asymptotic solution for large $t \to \infty$ can be derived.

4.3.1 Solving the Quadratic Fitness Case

In a first step, it is again noticed that (92a) has a fixed point where $\frac{\mathrm{d}\overline{\boldsymbol{x}}(t)}{\mathrm{d}t} = \mathbf{0}$. This fixed point is the optimizer $\hat{\boldsymbol{x}}$ of the quadratic fitness function $f(\boldsymbol{x})$. This is fully analogous to (10a) and (42a). Therefore, Eq. (92a) can be changed to an ODE system describing the evolution of the distance vector $\boldsymbol{z}$ to the optimizer defined by (12). Using $\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}} = -2\mathbf{Q}\boldsymbol{z}$ it follows (see also Eq. (49))

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = -\frac{2\mathbf{C}\mathbf{Q}\boldsymbol{z}}{\sigma_f}, \tag{93a}$$

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = -\frac{2\mathbf{C}\mathbf{Q}\mathbf{C}}{\sigma_f}, \tag{93b}$$

where σ_f, Eq. (77), changes to

$$\sigma_f = \sqrt{4\boldsymbol{z}^{\mathrm{T}}\mathbf{QCQ}\boldsymbol{z} + 2\mathrm{Tr}[(\mathbf{QC})^2]}. \tag{94}$$

As the next step, a functional connection between $\boldsymbol{z}(t)$ and $\mathbf{C}(t)$ will be derived. To this end, Eq. (93b) is multiplied by $\mathbf{C}^{-1}\boldsymbol{z}$ from the right yielding

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t}\mathbf{C}^{-1}\boldsymbol{z} = -\frac{2\mathbf{CQ}\boldsymbol{z}}{\sigma_f} \overset{(93a)}{=} \frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t}. \tag{95}$$

Multiplying (95) with $\mathbf{C}^{-1}$ from the left yields

$$\mathbf{C}^{-1}\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t}\mathbf{C}^{-1}\boldsymbol{z} = \mathbf{C}^{-1}\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t}. \tag{96}$$

Taking the identity

$$\mathbf{C}^{-1}\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t}\mathbf{C}^{-1} = -\frac{\mathrm{d}\mathbf{C}^{-1}}{\mathrm{d}t} \tag{97}$$

into account (compare the inverse case, Eq. (45)), one gets

$$-\frac{\mathrm{d}\mathbf{C}^{-1}}{\mathrm{d}t}\boldsymbol{z} = \mathbf{C}^{-1}\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} \qquad \Longrightarrow \qquad \frac{\mathrm{d}\mathbf{C}^{-1}}{\mathrm{d}t}\boldsymbol{z} + \mathbf{C}^{-1}\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \mathbf{0}. \tag{98}$$

This leads immediately to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\mathbf{C}^{-1}\boldsymbol{z}\right) = \mathbf{0} \qquad \Longrightarrow \qquad \mathbf{C}^{-1}\boldsymbol{z} = \boldsymbol{b}, \tag{99}$$

where $\boldsymbol{b}$ is a constant vector to be determined below. Solving (99) for $\boldsymbol{z}$, one gets

$$\boldsymbol{z}(t) = \mathbf{C}(t)\boldsymbol{b}. \tag{100}$$

This is a very simple dependence showing that the evolution of the distance vector to the optimizer is governed by the evolution of the covariance matrix. The constant vector $\boldsymbol{b}$ is obtained by applying the initial conditions $\boldsymbol{z}_0 = \boldsymbol{z}(0)$ and $\mathbf{C}_0 = \mathbf{C}(0)$ to (100) yielding $\boldsymbol{z}_0 = \mathbf{C}_0\boldsymbol{b}$. Solving for $\boldsymbol{b}$ yields

$$\boldsymbol{b} = \mathbf{C}_0^{-1}\boldsymbol{z}_0 \tag{101}$$

and reintroduced into (100), one obtains the

Lemma 1 (Optimal weight IGO z-dynamics are governed by C-evolution). *Consider the IGO dynamics with locally fitness standardized weights (86) and $\mathcal{N}(\overline{\boldsymbol{x}}(t), \mathbf{C}(t))$ offspring sampling acting on a general quadratic fitness model (7) that generates normally distributed fitness values. Let $\boldsymbol{z}$ be the vector representing the distance of $\overline{\boldsymbol{x}}(t)$ to the optimizer $\hat{\boldsymbol{x}}$ according to (12). Let $\mathbf{C}_0 = \mathbf{C}(0)$ and $\boldsymbol{z}_0 = \overline{\boldsymbol{x}}(0) - \hat{\boldsymbol{x}}$ be the initial values. Then the $\boldsymbol{z}$-dynamics are given by*

$$\boldsymbol{z}(t) = \mathbf{C}(t)\mathbf{C}_0^{-1}\boldsymbol{z}_0. \tag{102}$$

Proof. See the derivation given above. □

The simple result of Lemma 1 obtained from the ODE system (93) is remarkable if one takes into account that the ODE system is non-linear. Unfortunately, the next step calculating the $\mathbf{C}$ dynamics seems intractable in general. Therefore, one has to settle for an asymptotic solution of (93b) that describes the evolution of $\mathbf{C}$ for $t \to \infty$. To this end, (93b) is multiplied

from both the left and the right with $\mathbf{C}^{-1}$ taking (97) into account. Furthermore, $\mathbf{z}$ is substituted in σ_f, Eq. (94), by (100) yielding

$$\frac{\mathrm{d}\mathbf{C}^{-1}}{\mathrm{d}t} = \frac{2\mathbf{Q}}{\sqrt{4\boldsymbol{b}^{\mathrm{T}}\mathbf{C}\mathbf{Q}\mathbf{C}\mathbf{Q}\mathbf{C}\boldsymbol{b} + 2\mathrm{Tr}[(\mathbf{Q}\mathbf{C})^2]}}. \tag{103}$$

Multiplying from the right with $\mathbf{Q}^{-1}$, one obtains using the abbreviation

$$\mathbf{D} := \mathbf{C}^{-1}\mathbf{Q}^{-1} = (\mathbf{Q}\mathbf{C})^{-1} \tag{104}$$

$$\frac{\mathrm{d}\mathbf{D}}{\mathrm{d}t} = \frac{\mathbf{I}}{\sqrt{\boldsymbol{b}^{\mathrm{T}}\mathbf{Q}^{-1}\mathbf{D}^{-3}\boldsymbol{b} + \frac{1}{2}\mathrm{Tr}[\mathbf{D}^{-2}]}} \tag{105}$$

That is, all off-diagonal elements of $\mathbf{D}$ must be constants since according to (105) $\frac{\mathrm{d}D_{ij}}{\mathrm{d}t} = 0$. Therefore, $\forall i \neq j : D_{ij}(t) = D_{ij}(0)$. Furthermore, all diagonal elements of $\frac{\mathrm{d}D_{ij}}{\mathrm{d}t}$ have the *same* scalar function on the rhs in (105). As a result $\forall i : D_{ii}(t) = D_{ii}(0) + q(t)$, where $q(t)$ is a function to be determined. To summarize, the general solution to (105) reads

$$\mathbf{D}(t) = \mathbf{D}(0) + q(t)\mathbf{I}. \tag{106}$$

Inserting this into (105), one obtains

$$\frac{\mathrm{d}q}{\mathrm{d}t} = q\frac{1}{\sqrt{\frac{1}{q}\boldsymbol{b}^{\mathrm{T}}\mathbf{Q}^{-1}\left(\frac{\mathbf{D}(0)}{q} + \mathbf{I}\right)^{-3}\boldsymbol{b} + \frac{1}{2}\mathrm{Tr}\left[\left(\frac{\mathbf{D}(0)}{q} + \mathbf{I}\right)^{-2}\right]}}. \tag{107}$$

Since the denominator in (105) is $\sigma_f(t)/2$ and $\sigma_f(0) > 0$, it holds $\forall i : \frac{\mathrm{d}D_{ii}}{\mathrm{d}t} > 0$. By virtue of (106) it follows that $\frac{\mathrm{d}q}{\mathrm{d}t} \geq 0$ and $q(0) = 0$. Thus, $q(t) \geq 0$ is a monotonously increasing unbounded function and for $t \to \infty : q(t) \to \infty$. Therefore, an asymptotically exact solution to (107) can be constructed by letting $q(t) \to \infty$ in the denominator of (107) leading to the simple differential equation

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{1}{\sqrt{\frac{1}{2}\mathrm{Tr}[\mathbf{I}]}}q = \sqrt{\frac{2}{N}}q. \tag{108}$$

The solution of which is

$$q(t) = \mathrm{e}^{\gamma t} - 1 \simeq \mathrm{e}^{\gamma t}, \tag{109}$$

with the inverse time constant

$$\gamma = \sqrt{\frac{2}{N}}. \tag{110}$$

Equating (106) with (104) and resolving for $\mathbf{C}$ yields

$$\mathbf{C}(t) = \mathbf{Q}^{-1}\left(\mathbf{D}(0) + q(t)\mathbf{I}\right)^{-1} = \frac{1}{q(t)}\mathbf{Q}^{-1}\left(\frac{1}{q(t)}\mathbf{D}(0) + \mathbf{I}\right)^{-1}. \tag{111}$$

Now inserting the asymptotic solution (109) in (111), one obtains

$$\mathbf{C}(t) \simeq \mathbf{Q}^{-1}\mathrm{e}^{-\gamma t}. \tag{112}$$

This result gives rise to

Theorem 3 (Optimal weight IGO asymptotic dynamics). *Consider the IGO dynamics governed by (93) for $t \to \infty$ with locally fitness standardized weights (86) and $\mathcal{N}(\overline{\boldsymbol{x}}(t), \mathbf{C}(t))$ offspring sampling acting on a general quadratic fitness model (7) that generates normally distributed fitness values. Let $\boldsymbol{z}$ be the vector representing the distance of $\overline{\boldsymbol{x}}(t)$ to the optimizer $\hat{\boldsymbol{x}}$ according to (12). Let $\mathbf{C}_0 = \mathbf{C}(0)$ and $\boldsymbol{z}_0 = \overline{\boldsymbol{x}}(0) - \hat{\boldsymbol{x}}$ be the initial values. Then the asymptotic $\mathbf{C}$-dynamics are given by*

$$\boxed{\mathbf{C}(t) \simeq \mathbf{Q}^{-1} \mathrm{e}^{-\sqrt{\frac{2}{N}}t}} \tag{113}$$

and the asymptotic $\boldsymbol{z}$-dynamics obey

$$\boxed{\boldsymbol{z}(t) \simeq \mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0 \mathrm{e}^{-\sqrt{\frac{2}{N}}t}.} \tag{114}$$

Proof. As for Eq. (113), the proof is obtained by following the derivation presented above yielding (112) and finally inserting (110). In order to prove Eq. (114) one has to make use of Lemma 1, Eq. (102). Inserting Eq. (113) into (102) already yields (114). □

The results obtained are remarkable. At first, (114) shows that IGO approaches the optimizer exponentially fast, thus providing linear convergence order. The rate of approach is determined by the time constant

$$\tau_{\mathrm{IGO}} = \sqrt{N/2} \tag{115}$$

that depends on the search space dimension N only with an unexpected square root law and no dependency on $\mathbf{Q}$. While the latter might seem as a surprise, one should recall that we have derived the asymptotical $t \to \infty$ solution to (93). According to (113), the covariance matrix gets proportional to the inverse of $\mathbf{Q}$. That is, the IGO ES "sees" effectively a sphere model. Therefore, asymptotically, the IGO ES has transformed the ellipsoidal level sets of f into spherical ones and the dependencies on $\mathbf{Q}$ vanish. Note, according to (111) also the influence of the initially chosen covariance matrix $\mathbf{C}(0) = \mathbf{C}_0$ vanishes exponentially fast.

The $\sqrt{N}$-law in (115) is somewhat peculiar. Considering the performance of isotropic $(\mu/\mu_I, \lambda)$-ES with σ mutation strength control using self-adaptation or cumulative step-size adaptation on the sphere model, Arnold and Beyer (2004) and Meyer-Nieberg and Beyer (2005), respectively, found time constants $\propto N$. That is, the expected running time for a fixed relative improvement in f-values is proportional to N whereas for the IGO ODE it grows only with the square root of the search space dimensionality. It is currently an open question whether this square root law indicates a general lower N bound for algorithms derived from the IGO ODE taking into account that the IGO ODE is an infinite population size model.

Considering (113), one sees that the covariance matrix adapts to the desired behavior (23) exponentially fast. That is, we have *proven* that this type of IGO approximates the inverse of $\mathbf{Q}$ up to a scalar factor. Such a behavior has been observed in real ES using covariance matrix adaptation, as e.g. the evolutionary gradient search ES (Arnold and Salomon, 2007, p. 492, Footnote 3). A similar result has been derived in Akimoto (2012) using difference equations. Where it was shown that $\mathbf{C}$ evolves up to a scalar factor to the inverse of the Hessian of $f(\boldsymbol{z}) = \boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\boldsymbol{z}$. In that work, a specially tailored IGO NES has been considered that deviates from the model (93). It relies on an *ad hoc* constructed time-dependent step-size scaling factor that uses eigenvalue information taken from the actual $\mathbf{C}(t)$ matrix.

Comparing the results of Theorem 3 with those of Theorems 1 and 2 regarding NES without local utility transformation, one sees that the rank-based weighting yields qualitatively better performance on quadratic fitness models. Having a closer look at the governing ODEs (42) and (92) one notices the σ_f terms in the denominator of the latter. Getting closer to the optimizer, σ_f gets smaller, thus, counteracting the decrease of the numerators in the rhs of (92a) and (92b)

with the result of larger derivatives. Tracing back the additional σ_f terms, one finds that these are due to the $W_f(f|\boldsymbol{\theta})$ term (86) in the integral (88). That is, the linear convergence order of this IGO version is clearly a result of that (sphere) optimal rank-based weighting (69) leading to the locally standardized utility function (86).

4.3.2 The Linear Fitness Case

The investigation of the behavior of IGO with weights (86) and linear fitness $f(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ is a special case of (7) assuming $\mathbf{Q} = \mathbf{0}$. If introduced in (92) and (77) one obtains the IGO system

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}}{\mathrm{d}t} = \frac{\mathbf{C}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}\boldsymbol{a}}}, \tag{116a}$$

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = \mathbf{0}. \tag{116b}$$

This IGO ODE system is exact in that it does not require the condition of fitness normality. Given $\boldsymbol{x} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, \mathbf{C})$, the linear function $f(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ yields always normally distributed fitness values. Solving (116) is trivial, one immediately gets for (116b)

$$\boxed{\mathbf{C}(t) = \mathbf{C}_0.} \tag{117}$$

If inserted in (116a), this yields

$$\boxed{\overline{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}}_0 + \frac{\mathbf{C}_0\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}_0\boldsymbol{a}}}\, t.} \tag{118}$$

That is, $\overline{\boldsymbol{x}}$ increases linearly in time. If the initial covariance matrix is chosen isotropically, i.e. $\mathbf{C}_0 = \mathbf{I}$, then the evolution is in direction of the gradient $\boldsymbol{a}$.

4.4 Dynamics of IGO with (μ, λ)-Truncation Selection

Truncation selection, aka (μ, λ)-selection, in Evolution Strategies is the standard selection that takes (in the case of maximization) those μ individuals $\boldsymbol{x}$ out of the sample of λ offspring individuals that produce the μ greatest fitness values $f(\boldsymbol{x})$. In the infinite population model that means that only individuals above the $(1-\vartheta)$ f-quantile $f_{1-\vartheta}$ are used in the calculation of the $\boldsymbol{\theta}$ gradient (67). Let ϑ be the truncation ratio

$$\vartheta := \frac{\mu}{\lambda}, \tag{119}$$

then the local weighting function (65) returns $1/\vartheta$ for f values $f \geq f_{1-\vartheta}$ and zero otherwise (for $\lambda \to \infty$)

$$W_f(x|\boldsymbol{\theta}) = \begin{cases} \frac{1}{\vartheta}, & \text{if } f(\boldsymbol{x}) \geq f_{1-\vartheta}(\boldsymbol{\theta}) \\ 0, & \text{otherwise.} \end{cases} \tag{120}$$

Assuming normally distributed fitness values f, the pdf is given by (73). Therefore, the $f_{1-\vartheta}$-quantile obeys the equation

$$\Phi\left(\frac{f_{1-\vartheta} - \overline{f}}{\sigma_f}\right) = 1 - \vartheta. \tag{121}$$

That is, the first line in (120) is fulfilled for $f(\boldsymbol{x})$ values that fulfill

$$\frac{f - \overline{f}(\boldsymbol{\theta})}{\sigma_f(\boldsymbol{\theta})} \geq \Phi^{-1}(1-\vartheta), \tag{122}$$

where Φ^{-1} is the inverse of the cdf (i.e. the quantile function) of the standard normal variate. Now, plugging (120) into (87) yields

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{1}{\sqrt{2\pi\sigma_f^2}} \int_{f=f_{1-\vartheta}}^{\infty} \frac{1}{\vartheta} \left[-\nabla_{\boldsymbol{\theta}}\sigma_f + \left(\frac{f-\overline{f}}{\sigma_f}\right) \nabla_{\boldsymbol{\theta}}\overline{f} + \left(\frac{f-\overline{f}}{\sigma_f}\right)^2 \nabla_{\boldsymbol{\theta}}\sigma_f \right] \times \exp\left[-\frac{1}{2}\left(\frac{f-\overline{f}}{\sigma_f}\right)^2 \right] \mathrm{d}f. \tag{123}$$

Change of the integration variable f to $t := (f-\overline{f})/\sigma_f$ and noting that (122) transforms to $t \geq \Phi^{-1}(1-\vartheta)$ yields

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{1}{\vartheta}\frac{1}{\sigma_f}\frac{1}{\sqrt{2\pi}} \int_{t=\Phi^{-1}(1-\vartheta)}^{\infty} \left((t^2-1)\nabla_{\boldsymbol{\theta}}\sigma_f + t\,\nabla_{\boldsymbol{\theta}}\overline{f}\right) \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t. \tag{124}$$

The t integration can be performed using integral formula (A.16) and (A.17) in (Beyer, 2001, p. 331). One gets

$$h_1(\vartheta) := \frac{1}{\sqrt{2\pi}} \int_{t=\Phi^{-1}(1-\vartheta)}^{\infty} t \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\Phi^{-1}(1-\vartheta)\right)^2 \right] \tag{125}$$

and

$$\begin{aligned} &\frac{1}{\sqrt{2\pi}} \int_{t=\Phi^{-1}(1-\vartheta)}^{\infty} (t^2-1) \exp\left(-\frac{1}{2}t^2\right) \mathrm{d}t \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\Phi^{-1}(1-\vartheta)\right)^2 \right] \Phi^{-1}(1-\vartheta) = h_1(\vartheta)h_2(\vartheta), \end{aligned} \tag{126}$$

where

$$h_2(\vartheta) := \Phi^{-1}(1-\vartheta). \tag{127}$$

Thus, one obtains for (124)

$$\nabla_{\boldsymbol{\theta}} E_W = \frac{h_1(\vartheta)}{\vartheta}\frac{1}{\sigma_f}\left(\nabla_{\boldsymbol{\theta}}\overline{f} + h_2(\vartheta)\nabla_{\boldsymbol{\theta}}\sigma_f\right). \tag{128}$$

Inserting the gradients w.r.t. $\overline{\boldsymbol{x}}$ and $\mathbf{C}$ given by (81), (82), and (84) into (128) yields

$$\nabla_{\overline{\boldsymbol{x}}} E_W = \frac{h_1(\vartheta)}{\vartheta}\frac{1}{\sigma_f}\left[(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}) - h_2(\vartheta)\frac{2}{\sigma_f}\mathbf{Q}\mathbf{C}\big(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}\big) \right] \tag{129}$$

and

$$\nabla_{\mathbf{C}} E_W = \frac{h_1(\vartheta)}{\vartheta}\frac{1}{\sigma_f}\left\{ -\mathbf{Q} + \frac{h_2(\vartheta)}{2}\frac{1}{\sigma_f}\big[(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})^{\mathrm{T}} + 4\mathbf{Q}\mathbf{C}\mathbf{Q}\big] \right\}. \tag{130}$$

The IGO ODE system (41) becomes (replacing E_f by E_W)

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}}{\mathrm{d}t} = \frac{h_1(\vartheta)}{\vartheta}\frac{1}{\sigma_f}\left[\mathbf{C}(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}) - h_2(\vartheta)\frac{2}{\sigma_f}\mathbf{C}\mathbf{Q}\mathbf{C}\big(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}}\big) \right] \tag{131}$$

and

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = \frac{h_1(\vartheta)}{\vartheta}\frac{1}{\sigma_f}\left\{ -2\mathbf{C}\mathbf{Q}\mathbf{C} + \frac{h_2(\vartheta)}{\sigma_f}\big[\mathbf{C}(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})^{\mathrm{T}}\mathbf{C} + 4\mathbf{C}\mathbf{Q}\mathbf{C}\mathbf{Q}\mathbf{C}\big] \right\}. \tag{132}$$

Its asymptotic solution will be investigated in the next section.

4.4.1 Dynamics of the Quadratic Fitness Case

Comparing (92a) with (131) and (93a), one sees that $(\boldsymbol{a} - 2\mathbf{Q}\overline{\boldsymbol{x}})$ can be substituted again by $-2\mathbf{Q}\boldsymbol{z}$. Thus, the variable transformation $\boldsymbol{z} = \overline{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}$ is introduced in (131) and (132). This leads to

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = -\frac{h_1(\vartheta)}{\vartheta}\frac{2}{\sigma_f}\left[\mathbf{C}\mathbf{Q}\boldsymbol{z} - h_2(\vartheta)\frac{2}{\sigma_f}\mathbf{C}\mathbf{Q}\mathbf{C}\mathbf{Q}\boldsymbol{z}\right] \tag{133}$$

and

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = -\frac{h_1(\vartheta)}{\vartheta}\frac{2}{\sigma_f}\left[\mathbf{C}\mathbf{Q}\mathbf{C} - 2\frac{h_2(\vartheta)}{\sigma_f}\left(\mathbf{C}\mathbf{Q}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\mathbf{C} + \mathbf{C}\mathbf{Q}\mathbf{C}\mathbf{Q}\mathbf{C}\right)\right], \tag{134}$$

where σ_f is again given by (94). Pulling $\mathbf{C}\mathbf{Q}\boldsymbol{z}$ out of (133) and $\mathbf{C}\mathbf{Q}\mathbf{C}$ out of (134) one obtains the ODE system

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = \frac{h_1(\vartheta)}{\vartheta}\left[\mathbf{I} - 2h_2(\vartheta)\frac{\mathbf{C}\mathbf{Q}}{\sigma_f}\right]\left(-\frac{2\mathbf{C}\mathbf{Q}\boldsymbol{z}}{\sigma_f}\right), \tag{135a}$$

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = \frac{h_1(\vartheta)}{\vartheta}\left[\mathbf{I} - 2h_2(\vartheta)\frac{\mathbf{C}\mathbf{Q}}{\sigma_f}\left(\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{C}^{-1} + \mathbf{I}\right)\right]\left(-\frac{2\mathbf{C}\mathbf{Q}\mathbf{C}}{\sigma_f}\right). \tag{135b}$$

This is a non-linear ODE system where closed-form solutions seem difficult to be obtained. Yet, it shares similarities with (93). A direct asymptotic solution can be given for the special case of truncation ratio $\vartheta = 1/2$. Due to (127), the second term in the brackets of (135) vanish and the resulting ODE system gets proportional to the ODE system (93) with the factor $h_1(\vartheta)/\vartheta$. That is, time t undergoes a linear transformation $t \mapsto t\, h_1(\vartheta)/\vartheta$. This is equivalent to a change of the inverse time constant γ, Eq. (110), to $\gamma' = \gamma\, h_1(\vartheta)/\vartheta$. Thus, the asymptotic dynamics can be taken from Theorem 3, Eq. (113) and (114), yielding

$$\vartheta = \frac{1}{2}: \quad \mathbf{C}(t) \simeq \mathbf{Q}^{-1}\exp\left(-\frac{2}{\sqrt{\pi N}}t\right), \tag{136a}$$

$$\boldsymbol{z}(t) \simeq \mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0\exp\left(-\frac{2}{\sqrt{\pi N}}t\right), \tag{136b}$$

where $h_1(1/2)$ was calculated using (125). As for this special case, we have shown that the IGO ODE exhibits linear convergence order. Considering $\vartheta \neq 1/2$ is much more involved and not completely solved up until now. Therefore, we will first discuss the qualitative behavior of the ODE system when changing $\vartheta \neq 1/2$ and afterwards, we will derive a solution that holds in the vicinity of $\vartheta = 1/2$.

Considering the different matrix terms in (135b), one notices that all these single terms are positive definite, i.e. it holds for $t < \infty$ that $\mathbf{C}\mathbf{Q}\mathbf{C} > 0$, $\mathbf{C}\mathbf{Q}\mathbf{C}\mathbf{Q}\mathbf{C} > 0$, and $\mathbf{C}\mathbf{Q}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\mathbf{C} > 0$. As for the latter case this becomes clear by substitution $\boldsymbol{y} := \mathbf{C}\mathbf{Q}\boldsymbol{z}$ and noting that $\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}$ is positive definite.

First, consider the $\vartheta = 1/2$ case in (135b). The term with the h_2 factor vanishes because $h_2(1/2) = 0$. The remaining expression in the bracket is positive definite, i.e. $\mathbf{C}\mathbf{Q}\mathbf{C} > 0$. Due to the negative sign in front of the bracket, the remaining ODE describes the *contraction* of the covariance matrix $\mathbf{C}$. The dynamics of which is given by (113).

Increasing the truncation ratio, i.e. $\vartheta > 1/2$, it holds $h_2 < 0$ and $\mathbf{C}\mathbf{Q}\mathbf{C}\mathbf{Q}\mathbf{C}$ and $\mathbf{C}\mathbf{Q}\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{Q}\mathbf{C}$ increase the positiveness of the square bracket resulting in a faster contraction of the covariance matrix $\mathbf{C}$. If, however, the contraction rate is too fast, the contraction of the $\boldsymbol{z}$ vector cannot keep pace with that evolution. In such a case, the ODE system describes permature convergence of the IGO algorithm.

Conversely, decreasing the truncation ratio, i.e. $\vartheta < 1/2$, it holds $h_2 > 0$. In that case the positiveness of the square bracket is decreased compared to $\mathbf{C}\mathbf{Q}\mathbf{C}$. Actually, the expression in

the square bracked can become negative definite. As a result, the covariance matrix expands. This behavior is desirable to a certain extend, especially in cases where the initial covariance matrix $\mathbf{C}(0) = \mathbf{C}_0$ was chosen too small. However, it can also result in an uncontrolled growth.

While a quantitative analysis of the contraction/expansion behavior depending on ϑ remains still to be done, the behavior of the ODE system in the vicinity of $\vartheta = 1/2$ can be derived by some kind of continuation. To this end, the solution (136) for $\vartheta = 1/2$ is used as an *Ansatz* with an unknown inverse time constant $\tilde{\gamma}$

$$\mathbf{C}(t) = \mathbf{Q}^{-1}\mathrm{e}^{-\tilde{\gamma}t}, \tag{137a}$$

$$\boldsymbol{z}(t) = \mathbf{Q}^{-1}\mathbf{C}_0\boldsymbol{z}_0\mathrm{e}^{-\tilde{\gamma}t}, \quad \tilde{\gamma} > 0. \tag{137b}$$

This Ansatz is inserted into the ODE system (135) in order to determine $\tilde{\gamma}$. At first, the $\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{C}^{-1} + \mathbf{I}$ term in (135b) is considered. This term becomes $\mathbf{I}$ for $t \to \infty$ as one can easily check by inserting (137)

$$\boldsymbol{z}\boldsymbol{z}^{\mathrm{T}}\mathbf{C}^{-1} + \mathbf{I} = \mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0\boldsymbol{z}_0^{\mathrm{T}}\mathbf{C}_0^{-1}\mathrm{e}^{-\tilde{\gamma}t} + \mathbf{I} \simeq \mathbf{I}. \tag{138}$$

As the next step, the expression $\mathbf{CQ}/\sigma_f$ in (137) is investigated. Using (94), one obtains

$$\begin{aligned}\frac{\mathbf{CQ}}{\sigma_f} &= \frac{\mathbf{Q}^{-1}\mathbf{Q}\mathrm{e}^{-\tilde{\gamma}t}}{\sqrt{4\boldsymbol{z}_0^{\mathrm{T}}\mathbf{C}_0^{-1}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0\mathrm{e}^{-3\tilde{\gamma}t} + 2\mathrm{Tr}[(\mathbf{Q}\mathbf{Q}^{-1})^2]\mathrm{e}^{-2\tilde{\gamma}t}}} \\ &= \frac{\mathbf{I}}{\sqrt{4\mathrm{e}^{-\tilde{\gamma}t}\boldsymbol{z}_0^{\mathrm{T}}\mathbf{C}_0^{-1}\mathbf{Q}^{-1}\mathbf{C}_0^{-1}\boldsymbol{z}_0 + 2N}} \simeq \frac{1}{\sqrt{2N}}\mathbf{I}\end{aligned} \tag{139}$$

Plugging (138) and (139) into (135), one gets a simplified ODE system that holds for large t

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}t} = g(\vartheta)\left(-\frac{2\mathbf{CQ}\boldsymbol{z}}{\sigma_f}\right), \tag{140a}$$

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = g(\vartheta)\left(-\frac{2\mathbf{CQC}}{\sigma_f}\right) \tag{140b}$$

with

$$g(\vartheta) := \frac{h_1(\vartheta)}{\vartheta}\left(1 - h_2(\vartheta)\sqrt{\frac{2}{N}}\right). \tag{141}$$

Now, the linear time transformation

$$\tilde{t} := g(\vartheta)t \tag{142}$$

can be applied to (140a) and (140b) yielding the ODE system

$$\frac{\mathrm{d}\boldsymbol{z}}{\mathrm{d}\tilde{t}} = -\frac{2\mathbf{CQ}\boldsymbol{z}}{\sigma_f}, \tag{143a}$$

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}\tilde{t}} = -\frac{2\mathbf{CQC}}{\sigma_f}. \tag{143b}$$

This ODE system is similar to the system (93) except the time parameter. Therefore, Theorem 3 can be applied using $\tilde{t}$, Eq. (142), instead of t in Eqs. (113) and (114). Thus, the exponent in (113) and (114) becomes $-\sqrt{\frac{2}{N}}\tilde{t} = -\sqrt{\frac{2}{N}}g(\vartheta)t$. That is, writing $\tilde{\gamma} = \sqrt{\frac{2}{N}}g(\vartheta)$, one gets for the inverse time constant using (141), (125), and (127)

$$\tilde{\gamma}(\vartheta, N) = \frac{1}{\sqrt{\pi N}}\frac{1}{\vartheta}\exp\left[-\frac{1}{2}\left(\Phi^{-1}(1-\vartheta)\right)^2\right]\left(1 - \sqrt{\frac{2}{N}}\,\Phi^{-1}(1-\vartheta)\right) \tag{144}$$

and the asymptotic dynamics become

$$\mathbf{C}(t) \simeq \mathbf{Q}^{-1}\mathrm{e}^{-\tilde{\gamma}t}, \tag{145a}$$

$$\boldsymbol{z}(t) \simeq \mathbf{Q}^{-1}\mathbf{C}_0\boldsymbol{z}_0\mathrm{e}^{-\tilde{\gamma}t}. \tag{145b}$$

The special case (136) is contained in (145) as can be checked by inserting $\vartheta = 1/2$ in (144) and (145) and comparing with (136). Concerning $\vartheta \neq 1/2$ values, it should be pointed out that the validity range of ϑ in (145) cannot be determined by the calculations presented. Further research is needed to determine the ϑ range for (145) that guarantees linear convergence order.

An alternative convergence proof for IGO using the special case of isotropic mutations has been presented by Glasmachers (2012). However, in that work "the existence of a linear convergence rate" was only claimed without proof. Another approach proving convergence to the optimizer using Lyapunov's stability analysis was proposed by Akimoto et al. (2012a). However, that analysis did not provide any assertions concerning the convergence order. Therefore, it does not contribute to the question why and how a local weighting function W_f is needed for efficiently working ESs.

Considering (145a), one sees that the covariance matrix gets asymptotically similar to the inverse of the Hessian of $f(\boldsymbol{x})$, i.e., the $\mathbf{Q}$-matrix of the quadratic model (7). This is a proof of the long-conjectured property of CMA-ES like algorithms using truncation selection (for the case $\vartheta \approx 1/2$) that is empirically observed when running such algorithms on goal functions that can be locally approximated by quadratic functions (see Arnold and Salomon (2007)).

4.4.2 The Linear Fitness Case

The IGO ODEs can be directly obtained from the quadratic case (131), (132), and (77) for $\mathbf{Q} = \mathbf{0}$. This yields

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}}{\mathrm{d}t} = \frac{h_1(\vartheta)}{\vartheta}\frac{\mathbf{C}\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}\boldsymbol{a}}}, \tag{146a}$$

$$\frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = \underbrace{\frac{h_1(\vartheta)h_2(\vartheta)}{\vartheta}}_{\varkappa(\vartheta)}\frac{\mathbf{C}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\mathbf{C}}{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}\boldsymbol{a}}. \tag{146b}$$

This system hold for all linear fitness functions. In order to derive the solution, Eq. (146b) is multiplied by $\boldsymbol{a}^{\mathrm{T}} \neq \mathbf{0}^{\mathrm{T}}$ from the left. This yields

$$\frac{\mathrm{d}\boldsymbol{a}^{\mathrm{T}}\mathbf{C}}{\mathrm{d}t} = \varkappa(\vartheta)\frac{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}\mathbf{C}}{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}\boldsymbol{a}} = \varkappa(\vartheta)\boldsymbol{a}^{\mathrm{T}}\mathbf{C} \quad \Longrightarrow \quad \frac{\mathrm{d}\mathbf{C}}{\mathrm{d}t} = \varkappa(\vartheta)\mathbf{C}. \tag{147}$$

Therefore, one immediately obtains

$$\boxed{\mathbf{C}(t) = \mathbf{C}_0\mathrm{e}^{\varkappa(\vartheta)t}.} \tag{148}$$

Plugging this result into (146a), one gets

$$\frac{\mathrm{d}\overline{\boldsymbol{x}}}{\mathrm{d}t} = \frac{h_1(\vartheta)}{\vartheta}\frac{\mathbf{C}_0\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}_0\boldsymbol{a}}}\exp\left(\frac{\varkappa(\vartheta)}{2}t\right). \tag{149}$$

The solution to (149) is easily obtained, it reads

$$\overline{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}}_0 + \frac{2h_1(\vartheta)}{\vartheta\varkappa(\vartheta)}\frac{\mathbf{C}_0\boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{C}_0\boldsymbol{a}}}\left[\exp\left(\frac{\varkappa(\vartheta)}{2}t\right) - 1\right], \tag{150}$$

as can be easily checked by inserting (150) into (149). Taking the definition of $\varkappa(\vartheta)$ in (146b) into account and (127), one finally obtains

$$\boxed{\overline{\boldsymbol{x}}(t) = \overline{\boldsymbol{x}}_0 + \frac{2}{\Phi^{-1}(1-\vartheta)} \frac{\mathbf{C}_0 \boldsymbol{a}}{\sqrt{\boldsymbol{a}^{\mathrm{T}} \mathbf{C}_0 \boldsymbol{a}}} \left[\exp\left(\frac{\varkappa(\vartheta)}{2} t \right) - 1 \right],} \tag{151}$$

where

$$\varkappa(\vartheta) = \frac{1}{\vartheta} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\Phi^{-1}(1-\vartheta) \right)^2 \right] \Phi^{-1}(1-\vartheta). \tag{152}$$

As one can see, the behavior of IGO with truncation selection on linear functions is qualitatively influenced by the truncation ratio ϑ. For $\vartheta > 1/2$ it follows that $\varkappa < 0$ and according to (148) the covariance matrix $\mathbf{C}$ shrinks exponentially fast. Thus, one observes premature convergence. In the opposite case $\vartheta < 1/2$ one gets an exponential growth of $\mathbf{C}$ and $\overline{\boldsymbol{x}}$ increases exponentially fast. Since $\boldsymbol{a}^{\mathrm{T}} \mathbf{C}_0 \boldsymbol{a} > 0$, the $f(t) = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{x}$ dynamics increase exponentially as well. It should be mentioned that the influence of the truncation ratio on the convergence behavior has also been found by Glasmachers (2012) considering IGO with isotropic mutations, i.e. for the special case $\mathbf{C} = \sigma^2 \mathbf{I}$.

5 Conclusions

Deriving Evolutionary Algorithms (EAs) from first principles is a tempting approach to put EA engineering on a sound scientific base. In an attempt to provide such an approach, the so-called "natural evolution strategies" have been proposed at the end of the last decade. A closer look at the premises of that approach reveals a paradigm that might be condensed into an *information gain constrained gradient ascent in the expected value fitness landscape*. According to Amari (1998), the resulting ascend direction is referred to as the "natural gradient" being the "steepest direction". However, as have been shown in this paper, considering quadratic objective functions, the "natural gradient" ascent when applied to expected fitness landscapes results in a slowly (but) converging information geometric flow towards the optimizer. The approach to the optimizer obeys an $1/t$-law, Eq. (48). While Amari's claim concerning the steepest descent direction is not wrong, because this is the best direction one can get given the constraints imposed on the information gain, "natural gradient" ascent alone does not necessarily yield efficient EAs.

Satisfactorily, however, is the desired result concerning the covariance matrix $\mathbf{C}$ evolution. In all cases investigated that rely on "natural gradient" ascent it has been proven that in the case of convergence $\mathbf{C}$ gets asymptotically proportional to the inverse of the Hessian of f. This behavior is desirable since it counteracts degeneration tendencies of the search distribution in subspaces.

The slow convergence behavior of the original NES is due to the combination of two properties of the original NES approach: On the one hand, the optimization of the objective function f is transformed into a globally defined expected value landscape. On the other hand, fast changes of the search distribution are suppressed by the "natural gradient". Since returning to ordinary gradient descent causes problems for some distribution parameterizations (as have been shown in Sect. 2.1), the remaining option is to localize the evaluation. That is, the utility of candidate solutions must be evaluated in a time-local manner.

As might have become clear in the above given discussion, the NES/IGO ODE design contains two decisions, which are rather independent:

(a) the choice of an appropriate statistical manifold and an ascent or descent principle, respectively,

(b) the choice of a utility function that transforms the original objective function.

The IGO ODE design choice (a) seems to be obviously fixed by the steepest ascent/descent in accordance with the metric determined by the distribution family chosen. Apart from the fact that the choice of the distribution family cannot be deduced from the IGO principle, real ES implementations, e.g. xNES, see Glasmachers et al. (2010), also depart from the "natural gradient" direction by introducing different learning rates for $\overline{\boldsymbol{x}}$ and $\mathbf{C}$. This is clearly for the sake of real algorithm performance and marks limitations of the infinite population size assumption inherent in the IGO ODE approach. Such implementational tweaks are hard to be deduced from the IGO philosophy.

The same holds for the design choice (b) regarding the utility used. In order to get linear convergence order in the IGO ODE framework, one has to localize the evaluation of the f-values generated. There are different options to localize the evaluation. Arnold's optimal weight function derived for the $(\mu/\mu, \lambda)$-ES on the sphere model yields in the asymptotic population limit a utility function that is just the *locally standardized* fitness. That is, f-values below $\overline{f}(\boldsymbol{\theta})$ get a negative evaluation. While a similar fitness baseline can already be found in early NES versions, the local standard deviation σ_f in the denominator of the standardization formula (86) makes the decisive difference. It ensures large utility "signals" when getting closer to the optimizer. It would be interesting to implement this theoretical finding into a new NES version and evaluate its performance on standard test beds. This new NES version would work without ranking, similar to the evolutionary gradient search of Arnold and Salomon (2007).

As an alternative option, truncation selection has been considered. In the case of f-maximization it accepts f-values above the *local* $(1-\vartheta)$ quantile $f_{1-\vartheta}(\boldsymbol{\theta})$. That is, instead of using the expected value of utility (86), one has to consider the expected value

$$E_W(\boldsymbol{\theta}) = \frac{1}{\vartheta} \int_{f=f_{1-\vartheta}}^{\infty} p(f|\boldsymbol{\theta}) \, \mathrm{d}f = \frac{1}{\vartheta} \Pr[f \geq f_{1-\vartheta}]. \tag{153}$$

This quantity can be given the simple interpretation of being proportional to the probability of f realizations that are greater than or equal to the local $f_{1-\vartheta}$ quantile. Thus, gradient ascent is aiming here at the increase of the probability of generating above $f_{1-\vartheta}(\boldsymbol{\theta})$ values. $f_{1-\vartheta}(\boldsymbol{\theta}(t))$ may be regarded as a threshold that changes with time. It gradually increases (in the case of f-maximization) during the IGO flow. This again ensures – similar to the σ_f in the denominator of (86) – that there is a large utility signal.

As we have seen, the choice of the utility function has a strong impact on the performance of the IGO system. This performance also depends on the class of fitness functions considered. For example, Arnold's weight scheme yields an exponentially fast approach to the optimizer of the ellipsoid model. However, as Eq. (118) shows, it performs rather slow on linear functions. Truncation selection as an alternative option yields an exponential $\boldsymbol{x}$ growth (151), provided that the truncation ratio is less than $1/2$. Therefore, it is better suited for these linear functions. It should be clear that it is impossible to draw general conclusions regarding the usefulness of specific utility functions without fixing the objective function class to be optimized.

While all utility functions considered are *f-compliant*, i.e. they emphasize the selection of *locally* better f-values, one yet can call this into question. As a matter of fact, f-compliance can imply an emphasis on local search. Even the seemingly well-posed arguments regarding the advantages of truncation selection ensuring the invariance under monotone f-transformations can be challenged when considering demands of robust optimization, see Beyer and Sendhoff (2007).

The theoretical investigations done in this paper concerned the performance of IGO on linear and quadratic models. From this analysis one cannot draw reliable conclusions regarding

optimization on multimodal objective functions. Considering global optimization, the question arises whether utility functions that allow for *non-f-compliance* to a certain degree could be a means to improve global search. To this end, IGO ODEs of simple multimodal test functions should be considered in a future research program.

As has been shown in this paper, the use of localized function evaluations yields strong enough utility signals to counteract the information conserving property of the "natural gradient". As a result, one can obtain exponentially fast convergence to the optimizer in the case of quadratic objective functions. Considering (114) and (145), one sees that the exponential approach takes place with a time constant proportional to $\sqrt{N}$. The time constant does *not* depend on $\mathbf{Q}$. This is in contrast to the naive expected value maximization the dynamics of which are given by Eq. (14). Obviously, using the "natural gradient" approach makes the dynamics asymptotically independent of the shape of the ellipsoid model defined by $\mathbf{Q}$. That is, using the Fisher metric, the IGO framework asymptotically transforms (for $t \to \infty$) the ellipsoidal problem into a spherical one. While we have considered the choice of the metric and the utility function as independent design choices, it yet seems remarkable that this transformation result is independent of the three different utilities chosen. It raises the question, how sensitive this result is w.r.t. other utility functions (e.g. non-f-compliant versions).

Let us consider the principal limitations of the IGO ODE theory. Due to the infinite population size it is hard – if not impossible – to get meaningful assertions w.r.t. the real behavior of real NES implementations derived from IGO. Obviously, this concerns especially properties that are related to the population size, as e.g. learning rates. However, it also concerns the parameterization used. While the IGO ODE theory is invariant w.r.t. $\boldsymbol{\theta}$ parameter transformations due to the differential geometry imposed by the Fisher metric, in real NES implementations the choice of an appropriate parameterization of the distribution family seems to have considerable influence on the performance of the NES, see e.g. Glasmachers et al. (2010). Furthermore, in advanced NES versions, the direction of the IGO flow of the $\boldsymbol{\theta}$ parameters departs from the "natural gradient" in that different step-size factors (learning parameters) are assigned to the $\overline{\boldsymbol{x}}$ and the $\mathbf{C}$ gradients. That is, the flow vector does no longer point into the "natural gradient" direction. However, such deviations could be incorporated into the IGO ODE framework. In the simplest case this would lead to different factors in (93) and (135). Finding closed form solutions to those ODEs might be a challenge for future research.

While the IGO ODE theory presented provides useful insights into the dynamic behavior of such systems, one should be cautious when transferring these infinite population size results to real NES implementations. As already pointed out, real NES implementation do usually deviate considerably from the IGO ODE such that the original ODE does not correctly describe the real ES. This concerns Monte-Carlo gradient estimations, explicit methods of step-size control and evolution path cumulation (as used in CMA-ES, see Hansen et al. (2003)), and different learning parameters. The current development of IGO theory cannot account for sampling aspects, path cumulations, etc. Analyzing the convergence behavior of real NES poses the same problems as in the case of classical ES theory and needs similar approaches and techniques as have been developed for the ES in Beyer (2001). With regard to NES, a first treatise has been provided by Schaul (2012).

Probably the most important benefit of the IGO philosophy lies in its "inspiring power" for deriving new EA variants on a scientifically grounded base. Even though final implementations do (most often) considerably deviate from the pure theory, the IGO philosophy can provide new starting points for research. One such starting point (proposed by one of the reviewers of this paper) could concern the combination of evolutionary methods, system dynamics, and time-series analysis that tracks the evolution of $\boldsymbol{\theta}$ to make predictions for $\boldsymbol{\theta}(t+1)$. Future will show whether this idea leads to improved ES algorithms.

Acknowledgements

This work was partially supported by the Austrian Science Fund FWF under grant P22649-N23. The author would like to thank the anonymous reviewers and Michael Hellwig for helpful discussions. Special thanks goes to Zhenhua Li who discovered an incorrectness in the original derivation of Eq. (37) resulting in a wrong Eq. (38), however, without any consequences for the final Eq. (40) (the same holds for Eq. (83) and Eq. (84)).

References

Akimoto, Y. (2012). Analysis of the Natural Gradient Algorithm on Monotonic Convex-Quadratic-Composite Functions. In Soule et al., T., editor, *GECCO-2012: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 11293–1300, New York. ACM.

Akimoto, Y., Auger, A., and Hansen, N. (2012a). Convergence of the Continuous Time Trajectories of Isotropic Evolution Strategies on Monotonic C^2-composite Functions. In Coello Coello, C., Cutello, V., Deb, K., Forrest, S., Nicosia, G., and Pavone, M., editors, *Parallel Problem Solving from Nature–PPSN XII*, pages 42–51, Berlin. Springer. Lecture Notes in Computer Science Vol. 7491.

Akimoto, Y., Nagata, Y., Ono, I., and Kobayashi, S. (2012b). Theoretical Foundation for CMA-ES from Information Geometry Perspective. *Algorithmica*, 64(4):698–716.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276.

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*. American Mathematical Society and Oxford University Press.

Arnold, B., Balakrishnan, N., and Nagaraja, H. (1992). *A First Course in Order Statistics*. Wiley, New York.

Arnold, D. (2006). Weighted multirecombination evolution strategies. *Theoretical Computer Science*, 361:18–37.

Arnold, D. and Beyer, H.-G. (2004). Performance Analysis of Evolutionary Optimization With Cumulative Step Length Adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622.

Arnold, D. and Salomon, R. (2007). Evolutionary Gradient Search Revisited. *IEEE Transactions on Evolutionary Computation*, 11(4):480–495.

Beyer, H.-G. (2001). *The Theory of Evolution Strategies*. Natural Computing Series. Springer, Heidelberg.

Beyer, H.-G. (2007). Evolution Strategies. *Scholarpedia*, 2(8):1965.

Beyer, H.-G. and Deb, K. (2001). On Self-Adaptive Features in Real-Parameter Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, 5(3):250–270.

Beyer, H.-G. and Melkozerov, A. (2013). The Dynamics of Self-Adaptive Multi-Recombinant Evolution Strategies on the General Ellipsoid Model. *IEEE Transactions on Evolutionary Computation*. accepted, DOI 10.1109/TEVC.2013.2283968.

Beyer, H.-G. and Schwefel, H.-P. (2002). Evolution Strategies: A Comprehensive Introduction. *Natural Computing*, 1(1):3–52.

Beyer, H.-G. and Sendhoff, B. (2007). Robust Optimization - A Comprehensive Survey. *Computer Methods in Applied Mechanics and Engineering*, 196(33–34):3190–3218.

Eguchi, S. and Copas, J. (2006). Interpreting Kullback-Leibler Divergence with the Neyman-Pearson Lemma. *Journal of Multivariate Analysis*, 97(9):2034–2040.

Glasmachers, T. (2012). Convergence of the IGO-Flow of Isotropic Gaussian Distributions on Convex Quadratic Problems. In Coello Coello, C., Cutello, V., Deb, K., Forrest, S., Nicosia, G., and Pavone, M., editors, *Parallel Problem Solving from Nature–PPSN XII*, pages 1–10, Berlin. Springer. Lecture Notes in Computer Science Vol. 7491.

Glasmachers, T., Schaul, T., Sun, Y., Wierstra, D., and Schmidhuber, J. (2010). Exponential Natural Evolution Strategies. In Branke et al., J., editor, *GECCO-2010: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 393–400, New York. ACM.

Hansen, N. (2006). The CMA Evolution Strategy: A Comparing Review. In Lozano, J., Larrañaga, P., Inza, I., and Bengoetxea, E., editors, *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*, pages 75–102. Springer.

Hansen, N., Müller, S., and Koumoutsakos, P. (2003). Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18.

Kay, S. (1993). *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, Englewood Cliffs, NJ.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

Meyer-Nieberg, S. and Beyer, H.-G. (2005). On the Analysis of Self-Adaptive Recombination Strategies: First Results. In *Proceedings of the CEC'05 Conference*, pages 2341–2348, Piscataway, NJ. IEEE.

Ollivier, Y., Arnold, L., Auger, A., and Hansen, N. (2011). Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. Technical Report arXiv:1106.3708v1.

Rényi, A. (1961). On measures of information and entropy. In *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561.

Schaul, T. (2012). Natural Evolution Strategies Converge on Sphere Functions. In Soule et al., T., editor, *GECCO-2012: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 329–336, New York. ACM.

Sun, Y., Wierstra, D., Schaul, T., and Schmidhuber, J. (2009). Stochastic Search using the Natural Gradient. In Schaffer, J. D., editor, *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1161–1168, New York. ACM.

Wierstra, D., Schaul, T., Peters, J., and Schmidhuber, J. (2008). Natural Evolution Strategies. In *CEC 2008, IEEE World Congress on Computational Intelligence, 2008*, pages 3381–3387, Piscataway, NJ. IEEE.

A An alternative derivation of the C-related part of Fisher information, Eq. (38)

According to Eq. (27) the $\mathbf{C}$-related part of the Fisher information is given as the expected value

$$I_{(\alpha_1\alpha_2),(\beta_1\beta_2)} = \mathrm{E}\left[\left(\nabla_{\mathbf{C}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C})\right)_{\alpha_1\alpha_2}\left(\nabla_{\mathbf{C}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C})\right)_{\beta_1\beta_2}\right] \tag{154}$$

where $\mathbf{x} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{C})$ and $\left(\nabla_{\mathbf{C}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C})\right)_{ab} = \frac{\partial}{\partial C_{ab}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C})$. Starting from

$$\ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\det\mathbf{C} - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\overline{\mathbf{x}}) \tag{155}$$

for the N-dimensional Gaussian, one obtains (for a derivation, see Appendix B)

$$\nabla_{\mathbf{C}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C}) = -\frac{1}{2}\mathbf{C}^{-1} + \frac{1}{2}\mathbf{C}^{-1}(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}\mathbf{C}^{-1}. \tag{156}$$

Thus, one reads

$$\left(\nabla_{\mathbf{C}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C})\right)_{\alpha_1\alpha_2} = -\frac{1}{2}C^{-1}_{\alpha_1\alpha_2} + \frac{1}{2}\sum_{i,k} C^{-1}_{\alpha_1 i}(x_i-\overline{x}_i)(x_k-\overline{x}_k)C^{-1}_{k\alpha_2} \tag{157}$$

$$\left(\nabla_{\mathbf{C}} \ln p(\mathbf{x}|\overline{\mathbf{x}}, \mathbf{C})\right)_{\beta_1\beta_2} = -\frac{1}{2}C^{-1}_{\beta_1\beta_2} + \frac{1}{2}\sum_{l,m} C^{-1}_{\beta_1 l}(x_l-\overline{x}_l)(x_m-\overline{x}_m)C^{-1}_{m\beta_2} \tag{158}$$

Inserting this in (154), one gets

$$\begin{aligned}
\mathrm{E}\left[\left(\nabla_{\mathbf{C}} \ln p\right)_{\alpha_1\alpha_2}\left(\nabla_{\mathbf{C}} \ln p\right)_{\beta_1\beta_2}\right] \;=\; & \frac{1}{4}\mathrm{E}\left[C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2}\right] \\
& -\frac{1}{4}C^{-1}_{\alpha_1\alpha_2}\sum_{l,m} C^{-1}_{\beta_1 l}\mathrm{E}\left[(x_l-\overline{x}_l)(x_m-\overline{x}_m)\right]C^{-1}_{m\beta_2} \\
& -\frac{1}{4}C^{-1}_{\beta_1\beta_2}\sum_{i,k} C^{-1}_{\alpha_1 i}\mathrm{E}\left[(x_i-\overline{x}_i)(x_k-\overline{x}_k)\right]C^{-1}_{k\alpha_2} \\
& +\frac{1}{4}\sum_{i,k,l,m} C^{-1}_{\alpha_1 i}C^{-1}_{\beta_1 l}\mathrm{E}\left[(x_i-\overline{x}_i)(x_k-\overline{x}_k)(x_l-\overline{x}_l)(x_m-\overline{x}_m)\right]C^{-1}_{k\alpha_2}C^{-1}_{m\beta_2}.
\end{aligned} \tag{159}$$

Noting that $\mathrm{E}\left[(x_a-\overline{x}_a)(x_b-\overline{x}_b)\right] = C_{ab}$ and $\mathrm{E}\left[(x_i-\overline{x}_i)(x_k-\overline{x}_k)(x_l-\overline{x}_l)(x_m-\overline{x}_m)\right] = C_{il}C_{mk} + C_{lm}C_{ik} + C_{kl}C_{im}$, see (Beyer, 2001, p. 357), one obtains

$$\begin{aligned}
I_{(\alpha_1\alpha_2),(\beta_1\beta_2)} \;=\; & \frac{1}{4}C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} \\
& -\frac{1}{4}C^{-1}_{\alpha_1\alpha_2}\sum_{l,m} C^{-1}_{\beta_1 l}C_{lm}C^{-1}_{m\beta_2} - \frac{1}{4}C^{-1}_{\beta_1\beta_2}\sum_{i,k} C^{-1}_{\alpha_1 i}C_{ik}C^{-1}_{k\alpha_2} \\
& +\frac{1}{4}\sum_{i,k,l,m} C^{-1}_{\alpha_1 i}C^{-1}_{\beta_1 l}\left(C_{il}C_{mk} + C_{lm}C_{ik} + C_{kl}C_{im}\right)C^{-1}_{k\alpha_2}C^{-1}_{m\beta_2}
\end{aligned} \tag{160}$$

and further taking the symmetry of $\mathbf{C}^{-1}$ into account, it follows

$$\begin{aligned}
I_{(\alpha_1\alpha_2),(\beta_1\beta_2)} &= \frac{1}{4}C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} - \frac{1}{4}C^{-1}_{\alpha_1\alpha_2}\sum_l C^{-1}_{\beta_1 l}\delta_{l\beta_2} - \frac{1}{4}C^{-1}_{\beta_1\beta_2}\sum_i C^{-1}_{\alpha_1 i}\delta_{i\alpha_2} \\
&\quad + \frac{1}{4}\sum_{i,k,l} C^{-1}_{\alpha_1 i}C^{-1}_{\beta_1 l}\big(C_{il}\delta_{k\beta_2} + \delta_{\beta_2 l}C_{ik} + C_{kl}\delta_{i\beta_2}\big)C^{-1}_{k\alpha_2} \\
&= \frac{1}{4}C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} - \frac{1}{4}C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} - \frac{1}{4}C^{-1}_{\beta_1\beta_2}C^{-1}_{\alpha_1\alpha_2} \\
&\quad + \frac{1}{4}\sum_{i,k} C^{-1}_{\alpha_1 i}\big(\delta_{\beta_1 i}\delta_{k\beta_2} + C^{-1}_{\beta_1\beta_2}C_{ik} + \delta_{\beta_1 k}\delta_{i\beta_2}\big)C^{-1}_{k\alpha_2} \\
&= -\frac{1}{4}C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} + \frac{1}{4}\sum_i C^{-1}_{\alpha_1 i}\big(\delta_{\beta_1 i}C^{-1}_{\beta_2\alpha_2} + C^{-1}_{\beta_1\beta_2}\delta_{i\alpha_2} + \delta_{i\beta_2}C^{-1}_{\beta_1\alpha_2}\big) \\
&= -\frac{1}{4}C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} + \frac{1}{4}\big(C^{-1}_{\alpha_1\beta_1}C^{-1}_{\beta_2\alpha_2} + C^{-1}_{\alpha_1\alpha_2}C^{-1}_{\beta_1\beta_2} + C^{-1}_{\alpha_1\beta_2}C^{-1}_{\beta_1\alpha_2}\big) \\
&= \frac{1}{4}C^{-1}_{\alpha_1\beta_1}C^{-1}_{\alpha_2\beta_2} + \frac{1}{4}C^{-1}_{\alpha_1\beta_2}C^{-1}_{\alpha_2\beta_1}. \qquad (161)
\end{aligned}$$

B Derivation of Eq. (156)

A calculation of the gradient of (155) w.r.t. the symmetric matrix $\mathbf{C}$ will be sketched without explicitly relying on partial derivatives. The idea (see Boyd and Vandenberghe (2010)) is based on calculating the differential $\mathrm{d}f$ of a function f depending on matrix $\mathbf{X}$ and identifying the derivative in the inner product of the linear part of the corresponding Taylor expansion

$$f(\mathbf{X} + \mathrm{d}\mathbf{X}) = f(\mathbf{X}) + \sum_{ik}\frac{\partial f}{\partial X_{ik}}\,\mathrm{d}X_{ik}, \qquad (162)$$

thus,

$$\begin{aligned}
\mathrm{d}f(\mathbf{X}) &= f(\mathbf{X} + \mathrm{d}\mathbf{X}) - f(\mathbf{X}) = \sum_{ik}\frac{\partial f}{\partial X_{ik}}\,\mathrm{d}X_{ik} \\
\mathrm{d}f(\mathbf{X}) &= \sum_{ik}\big(\nabla_{\mathbf{X}} f\big)_{ik}\,\mathrm{d}X_{ik} \qquad (163)
\end{aligned}$$

Recalling the definition of the trace of a matrix applied to a matrix product $\mathbf{A}^{\mathrm{T}}\mathbf{B}$

$$\mathrm{Tr}[\mathbf{A}^{\mathrm{T}}\mathbf{B}] = \sum_k (\mathbf{A}^{\mathrm{T}}\mathbf{B})_{kk} = \sum_k\left(\sum_i (\mathbf{A}^{\mathrm{T}})_{ki}(\mathbf{B})_{ik}\right) = \sum_{ki} A_{ik}B_{ik} \qquad (164)$$

Eq. (163) can be expressed as

$$\mathrm{d}f(\mathbf{X}) = \mathrm{Tr}[(\nabla_{\mathbf{X}} f)^{\mathrm{T}}\,\mathrm{d}\mathbf{X}]. \qquad (165)$$

It is now the goal to calculate the differential of (155) by transforming the expressions in such a manner that one obtains trace expressions that allow for identification of the matrix gradient. To this end, $\ln\det\mathbf{C}$ is considered first

$$\begin{aligned}
\mathrm{d}\ln\det\mathbf{C} &= \ln\det(\mathbf{C} + \mathrm{d}\mathbf{C}) - \ln\det\mathbf{C} = \ln\big(\det(\mathbf{C} + \mathrm{d}\mathbf{C})(\det\mathbf{C})^{-1}\big) \\
&= \ln\big(\det(\mathbf{C} + \mathrm{d}\mathbf{C})\det\big(\mathbf{C}^{-1}\big)\big) = \ln\det\big((\mathbf{C} + \mathrm{d}\mathbf{C})\mathbf{C}^{-1}\big) \\
&= \ln\det\big(\mathbf{I} + \mathrm{d}\mathbf{C}\mathbf{C}^{-1}\big) \\
&\simeq \ln\big(1 + \mathrm{Tr}[\,\mathrm{d}\mathbf{C}\mathbf{C}^{-1}]\big) \\
&\simeq \mathrm{Tr}[\,\mathrm{d}\mathbf{C}\mathbf{C}^{-1}] = \mathrm{Tr}[\mathbf{C}^{-1}\,\mathrm{d}\mathbf{C}]. \qquad (166)
\end{aligned}$$

Here, the fourth line has been obtained by recalling the definition of a determinant specified for $\det(\mathbf{I} + \mathrm{d}\mathbf{Y})$ as the sum of the signed products of all permutations of mutually excluding row and column indexed matrix entries. As for $(\mathbf{I} + \mathrm{d}\mathbf{Y})$, except for the product of the diagonal elements, all products are at least of 2nd order differential products. Considering the product of the diagonal elements one has $(1 + \mathrm{d}Y_{11})(1 + \mathrm{d}Y_{22}) \cdots (1 + \mathrm{d}Y_{NN}) = 1 + \mathrm{d}Y_{11} + \mathrm{d}Y_{22} + \ldots \mathrm{d}Y_{NN} +$higher order terms $= 1 + \mathrm{Tr}[\,\mathrm{d}\mathbf{Y}] +$higher order terms. Using the first order Taylor approximation of $\ln(1 + x) \simeq x$ one obtains (166) and comparing with (165) one finally gets

$$\nabla_{\mathbf{C}} \ln \det \mathbf{C} = (\mathbf{C}^{-1})^{\mathrm{T}} = \mathbf{C}^{-1}. \tag{167}$$

In order to determine $\nabla_{\mathbf{C}}$ of the third term in (155) one has

$$\mathrm{d}(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) = (\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \big(\mathbf{C} + \mathrm{d}\mathbf{C}\big)^{-1} (\mathbf{x} - \overline{\mathbf{x}}) - (\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x} - \overline{\mathbf{x}}). \tag{168}$$

Considering

$$\begin{aligned} (\mathbf{C} + \mathrm{d}\mathbf{C})^{-1} &= \big(\mathbf{C}(\mathbf{I} + \mathbf{C}^{-1}\,\mathrm{d}\mathbf{C})\big)^{-1} \\ &= (\mathbf{I} + \mathbf{C}^{-1}\,\mathrm{d}\mathbf{C})^{-1} \mathbf{C}^{-1} \\ &\simeq (\mathbf{I} - \mathbf{C}^{-1}\,\mathrm{d}\mathbf{C}) \mathbf{C}^{-1} \\ &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\,\mathrm{d}\mathbf{C}\mathbf{C}^{-1}. \end{aligned} \tag{169}$$

Here, the first order Taylor expansion has been used in order to get the third line. Inserting this result in (168) yields

$$\mathrm{d}(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) = -(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1}\,\mathrm{d}\mathbf{C}\mathbf{C}^{-1} (\mathbf{x} - \overline{\mathbf{x}}). \tag{170}$$

A quadratic form $\mathbf{y}^{\mathrm{T}}\mathbf{B}\mathbf{y}$ ($\mathbf{y}$ - vector, $\mathbf{B}$ - matrix) can be expressed by a trace operation, it holds $\mathbf{y}^{\mathrm{T}}\mathbf{B}\mathbf{y} = \mathrm{Tr}[(\mathbf{y}\mathbf{y}^{\mathrm{T}})\mathbf{B}]$. Identifying $\mathbf{y} := \mathbf{C}^{-1}(\mathbf{x} - \overline{\mathbf{x}})$ and $\mathbf{B} = \mathrm{d}\mathbf{C}$ one gets

$$\begin{aligned} \mathrm{d}(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) &= -\mathrm{Tr}\left[\mathbf{C}^{-1}(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1}\,\mathrm{d}\mathbf{C}\right] \\ &= -\mathrm{Tr}\left[\left(\mathbf{C}^{-1}(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1}\right)^{\mathrm{T}}\,\mathrm{d}\mathbf{C}\right] \end{aligned} \tag{171}$$

and comparing with (165) one finally obtains

$$\nabla_{\mathbf{C}} (\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x} - \overline{\mathbf{x}}) = -\mathbf{C}^{-1}(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^{\mathrm{T}} \mathbf{C}^{-1}. \tag{172}$$

Using this result and (167), the matrix gradient of (155) is obtained as (156).

References

Boyd, S. and Vandenberghe, L. (2010). *Convex Optimization*. Cambridge University Press, UK.