Convergence of the Continuous Time Trajectories of Isotropic Evolution Strategies on Monotonic $\mathcal{C}^2$-composite Functions

Youhei Akimoto, Anne Auger, and Nikolaus Hansen

TAO Team, INRIA Saclay-Ile-de-France, LRI, Paris Sud University, France
{Youhei.Akimoto,Anne.Auger,Nikolaus.Hansen}@lri.fr

Abstract. The *Information-Geometric Optimization (IGO)* has been introduced as a unified framework for stochastic search algorithms. Given a parametrized family of probability distributions on the search space, the IGO turns an arbitrary optimization problem on the search space into an optimization problem on the parameter space of the probability distribution family and defines a natural gradient ascent on this space. From the natural gradients defined over the entire parameter space we obtain continuous time trajectories which are the solutions of an ordinary differential equation (ODE). Via discretization, the IGO naturally defines an iterated gradient ascent algorithm. Depending on the chosen distribution family, the IGO recovers several known algorithms such as the pure rank-μ update CMA-ES. Consequently, the continuous time IGO-trajectory can be viewed as an idealization of the original algorithm.

In this paper we study the continuous time trajectories of the IGO given the family of isotropic Gaussian distributions. These trajectories are a deterministic continuous time model of the underlying evolution strategy in the limit for population size to infinity and change rates to zero. On functions that are the composite of a monotone and a convex-quadratic function, we prove the global convergence of the solution of the ODE towards the global optimum. We extend this result to composites of monotone and twice continuously differentiable functions and prove local convergence towards local optima.

1 Introduction

Evolution Strategies (ESs) are stochastic search algorithms for numerical optimization. In ESs, candidate solutions are sampled using a Gaussian distribution parametrized by a mean vector and a covariance matrix. In state-of-the art ESs, those parameters are iteratively adapted using the ranking of the candidate solutions w.r.t. the objective function. Consequently, ESs are invariant to applying a monotonic transformation to the objective function. Adaptive ES algorithms are successfully applied in practice and there is ample empirical evidence that they converge linearly towards a local optimum of the objective function on a wide class of functions. However, their theoretical analysis even on simple functions is difficult as the state of the algorithm is given by both the mean vector and the covariance matrix that have a stochastic dynamic that needs to be simultaneously controlled. Their linear convergence to local optima is so far only proven for functions

C.A. Coello Coello et al. (Eds.): PPSN 2012, Part I, LNCS 7491, pp. 42–51, 2012.

that are composite of a monotonic transformation with a convex quadratic function—hence function with a single optimum—for rather simple search algorithms compared to the covariance matrix adaptation evolution strategy (CMA-ES) that is considered as the state-of-the-art ES [1–4]. In this paper, instead of analyzing the exact stochastic dynamic of the algorithms, we consider the deterministic time continuous model underlying adaptive ESs that follows from the Information-Geometric Optimization (IGO) setting recently introduced [5].

The Information-Geometric Optimization is a unified framework for randomized search algorithms. Given a family of probability distributions parametrized by $\theta \in \Theta$, the original objective function, f, is transformed to a fitness function J_θ defined on Θ. The IGO algorithm defined on Θ performs a natural gradient ascent aiming at maximizing J_θ. For the family of Gaussian distributions, the IGO algorithm recovers the pure rank-μ update CMA-ES [6], for the family of Bernoulli distributions, PBIL [7] is recovered. When the step-size for the gradient ascent algorithm (that corresponds to a learning rate in CMA-ES and PBIL) goes to zero, we obtain an ordinary differential equation (ODE) in θ. The set of solutions of this ODE, the IGO-flow, consists of continuous time models of the recovered algorithms in the limit of the population size going to infinity and the step-size (learning rate for ES or PBIL) to zero.

In this paper we analyze the convergence of the IGO-flow for isotropic ESs where the family of distributions is Gaussian with covariance matrix equal to an overall variance times the identity. The underlying algorithms are step-size adaptive ESs that resemble ESs with derandomized adaptation [8] and encompass xNES [9] and the pure rank-μ update CMA-ES with only one variance parameter [6]. Previous works have proposed and analyzed continuous models of ESs that are solutions of ODEs [10, 11] using the machinery of stochastic approximation [12, 16]. The ODE variable in these studies encodes solely the mean vector of the search distribution and the overall variance is taken to be proportional to $H(\nabla f)$ where H is a smooth function with $H(0) = 0$. Consequently the model analyzed looses invariance to monotonic transformation of the objective function and scale-invariance, both being fundamental properties of virtually all ESs. The technique relies on the Lyapunov function approach and assumes the stability of critical points of the ODE [10, 11]. In this paper, our approach also relies on the stability of the critical points of the ODE that we analyze by means of Lyapunov functions. However one difficulty stems from the fact that when convergence occurs, the variance typically converges to zero which is at the boundary of the definition domain Θ. To circumvent this difficulty we extend the standard Lyapunov method to be able to study stability of boundary points.

Applying the extended Lyapunov's method to the IGO-flow in the manifold of isotropic Gaussian distributions, we derive a sufficient condition on the so-called weight function w—parameter of the algorithm and usually chosen by the algorithm designer—so that the IGO-flow converges to the global minimum independently of the starting point on objective functions that are composite of a monotonic function with a convex quadratic function. We will call those functions *monotonic convex-quadratic-composite* in the sequel. We then extend this result to functions that are the composition of a monotonic transformation and a twice continuously differentiable function, called monotonic C^2-composite in the rest of the paper. We prove local convergence to a local optimum of the function in the sense that starting close enough from a local optimum, with a small enough variance, the IGO-flow converges to this local optimum.

The rest of the paper is organized as follows. In Section 2 we introduce the IGO-flow for the family of isotropic Gaussian distributions, which we call *ES-IGO*-flow. In Section 3 we extend the standard Lyapunov's method for proving stability. In Section 4 we apply the extended method to the ES-IGO-flow and provide convergence results of the ES-IGO-flow on monotonic convex-quadratic-composite functions and on monotonic $\mathcal{C}^2$-composite functions.

Notation. For $A \subset X$, where X is a topological space, we let A^c denote the complement of A in X, A^o the interior of A, $\overline{A}$ the closure of A, $\partial A = \overline{A} \setminus A^o$ the boundary of A. Let $\mathbb{R}$ and $\mathbb{R}^d$ be the sets of real numbers and d-dimensional real vectors, $\mathbb{R}_{\geqslant 0}$ and $\mathbb{R}_+$ denote the sets of non-negative and positive real numbers, respectively. Let $\|x\|$ represent the Euclidean norm of $x \in \mathbb{R}^d$. The open and closed balls in $\mathbb{R}^d$ centered at θ with radius $r > 0$ are denoted by $B(\theta, r)$ and $\overline{B}(\theta, r)$.

Let μ_{Leb} denote the Lebesgue measure on either $\mathbb{R}$ or $\mathbb{R}^d$. Let P_1 and P_d be the probability measures induced by the one-variate and d-variate standard normal distributions, p_1 and p_d the probability density function induced by P_1 and P_d w.r.t. μ_{Leb}. Let p_θ and P_θ represent the probability density function w.r.t. μ_{Leb} and the probability measure induced by the Gaussian distribution $\mathcal{N}(m(\theta), C(\theta))$ parameterized by $\theta \in \Theta$, where the mean vector $m(\theta)$ is in $\mathbb{R}^d$ and the covariance matrix $C(\theta)$ is a positive definite symmetric matrix of dimension d. We sometimes abbreviate $m(\theta(t))$ and $C(\theta(t))$ to $m(t)$ and $C(t)$. Let $\mathrm{vec} : \mathbb{R}^{d \times d} \to \mathbb{R}^{d^2}$ denote the vectorization operator such that $\mathrm{vec} : C \mapsto [C_{1,1}, C_{1,2}, \dots, C_{1,d}, C_{2,1}, \dots, C_{d,d}]^{\mathrm{T}}$, where $C_{i,j}$ is the i,j-th element of C. We use both notations: $\theta = [m^{\mathrm{T}}, \mathrm{vec}(C)^{\mathrm{T}}]^{\mathrm{T}}$ and $\theta = (m, C)$.

2 The ES-IGO-Flow

The IGO framework for continuous optimization with the family of Gaussian distributions is as follows. The original objective is to minimize an objective function $f : \mathbb{R}^d \to \mathbb{R}$. This objective function is mapped into a function on Θ. Hereunder, we suppose that f is μ_{Leb}-measurable. Let $w : [0, 1] \to \mathbb{R}$ be a bounded, non-increasing weight function. We define the weighted quantile function [5] as

$$W_\theta^f(x) = w\big(P_\theta[y : f(y) \leqslant f(x)]\big) \ . \tag{1}$$

The function $W_\theta^f(x)$ is a preference weight for x according to the P_θ-quantile. The fitness value of θ' given θ is defined as the expectation of the preference W_θ^f over $P_{\theta'}$, $J_\theta(\theta') = \mathbb{E}_{x \sim P_\theta}\left[W_\theta^f(x)\right]$. Note that since $W_\theta^f(x)$ depends on θ so does $J_\theta(\theta')$. The function J_θ is defined on a statistical manifold $(\Theta, \mathcal{I})$ equipped with the Fisher metric $\mathcal{I}$ as a Riemannian metric. The Fisher metric is the natural metric. It is compatible with relative entropy and with KL-divergence and is the only metric that does not depend on the chosen parametrization. Using log-likelihood trick and exchanging the order of differentiation and integration, the "vanilla" gradient of J_θ at $\theta' = \theta$ can be expressed as $\nabla_{\theta'} J_\theta(\theta')|_{\theta'=\theta} = \mathbb{E}_{x \sim P_\theta}\left[W_\theta^f(x) \nabla_\theta \ln(p_\theta(x))\right]$. The natural gradient, that is, the gradient taken w.r.t. the Fisher metric, is given by the product of the inverse of the Fisher information matrix $\mathcal{I}_\theta$ at θ and the vanilla gradient, namely $\mathcal{I}_\theta^{-1}\nabla_{\theta'}J_\theta(\theta')|_{\theta'=\theta}$. The IGO ordinary differential equation is defined as

$$\frac{\mathrm{d}\theta}{\mathrm{dt}} = \mathcal{I}_\theta^{-1}\nabla_{\theta'}J_\theta(\theta')\big|_{\theta'=\theta} \ . \tag{2}$$

Since the right-hand side (RHS) of the above ODE is independent of t the IGO ODE is autonomous. The IGO-flow is the set of solution trajectories of the above ODE (2).

When the parameter θ encodes the mean vector and the covariance matrix of the gaussian distribution in the following way $\theta = [m^{\mathrm{T}}, \mathrm{vec}(C)^{\mathrm{T}}]^{\mathrm{T}}$, the product of the inverse of the Fisher information matrix $\mathcal{I}_\theta^{-1}$ and the gradient of the log-likelihood $\nabla_\theta \ln(p_\theta(x))$ can be written in an explicit form [14] and (2) reduces to

$$\frac{\mathrm{d}\theta}{\mathrm{dt}} = \int W_\theta^f(x)\begin{bmatrix} x-m \\ \mathrm{vec}\big((x-m)(x-m)^{\mathrm{T}} - C\big)\end{bmatrix} P_\theta(\mathrm{d}x) \ . \tag{3}$$

The pure rank-μ update CMA-ES [6] can be considered as an Euler scheme for solving (3) with a Monte-Carlo approximation of the integral. Let $x_1, \dots, x_n$ be samples independently generated from P_θ. Then, the quantile $P_\theta[y : f(y) \leqslant f(x_i)]$ in (1) is approximated by the number of solutions better than x_i divided by n, i.e., $\big|\{x_j, j = 1, \dots, n : f(x_j) \leqslant f(x_i)\}\big|/n =: R_i/n$. Then $W_\theta^f(x_i)$ is approximated by $w\big((R_i - 1/2)/n\big)$, where w is the given weight function. The Euler scheme for approximating the solutions of (3) where the integral is approximated by Monte-Carlo leads to

$$\theta^{t+1} = \theta^t + \eta\sum_{i=1}^{n}\frac{w\big((R_i - 1/2)/n\big)}{n}\begin{bmatrix} x_i - m^t \\ \mathrm{vec}\big((x_i - m^t)(x_i - m^t)^{\mathrm{T}} - C^t\big)\end{bmatrix} \ , \tag{4}$$

where η is the time discretization step-size. This equation is equivalent to the pure rank-μ update CMA-ES when the learning rates η_m and η_C, for the update of m^t and C^t respectively, are set to the same value η, while they have different values in practice ($\eta_m = 1$ and $\eta_C \leqslant 1$). The summation on the RHS in (4) converges to the RHS of (3) with probability one as $\lambda \to \infty$ (Theorem 4 in [5]).

In the following, we study the simplified IGO-flow where the covariance matrix is parameterized by only a single variance parameter v as $C = vI_d$. Under the parameterization $\theta = [m^{\mathrm{T}}, v]^{\mathrm{T}}$, (2) reduces to $\frac{\mathrm{d}\theta}{\mathrm{dt}} = \int W_\theta^f(x)\big[\begin{smallmatrix} x-m \\ \|x-m\|^2/d - v\end{smallmatrix}\big]P_\theta(\mathrm{d}x)$. Using the change of variable $z = (x-m)/\sqrt{v}$, the above ODE reads

$$\frac{\mathrm{d}\theta}{\mathrm{dt}} = F_\theta(\theta) \ , \quad F_\theta(\theta) = \int W_\theta^f(m + \sqrt{v}z)\begin{bmatrix} \sqrt{v}z \\ v(\|z\|^2/d - 1)\end{bmatrix} P_d(\mathrm{d}z) \tag{5}$$

and we rewrite it by part

$$\tfrac{\mathrm{d}m}{\mathrm{dt}} = F_m(\theta) \ , \quad F_m(\theta) = \sqrt{v}\textstyle\int W_\theta^f(m + \sqrt{v}z) z P_d(dz) \tag{6}$$

$$\tfrac{\mathrm{d}v}{\mathrm{dt}} = F_v(\theta) \ , \quad F_v(\theta) = v\textstyle\int W_\theta^f(m + \sqrt{v}z)(\|z\|^2/d - 1)P_d(dz) \ . \tag{7}$$

The domain of this ODE is $\Theta = \{\theta = (m, v) \in \mathbb{R}^d \times \mathbb{R}_+\}$. We call (5) the *ES-IGO* ordinary differential equation. The following proposition shows that for a Lipschitz continuous weight function w, solutions of the ODE (5) exist for any initial condition $\theta(0) \in \Theta$ and are unique.

Proposition 1 (Existence and Uniqueness). *Suppose w is Lipschitz continuous. Then the initial value problem: $\frac{\mathrm{d}\theta}{\mathrm{dt}} = F_\theta(\theta)$, $\theta(0) = \theta_0$, has a unique solution on $[0, \infty)$ for each $\theta_0 \in \Theta$, i.e. there is only one solution $\theta : \mathbb{R}_{\geqslant 0} \to \Theta$ to the initial value problem.*

Proof. We can obtain a lower bound $a(t) > 0$ and an upper bound $b(t) < \infty$ for $v(t)$ for each $t \geqslant 0$ under a bounded w. Similarly, we can have an upper bound $c(t) < \infty$ for $\|m(t)\|$. Then we have that $(m(t), v(t)) \in E(t) = \{x \in \mathbb{R}^d : \|x\| \leqslant c(t)\} \times \{x \in \mathbb{R}_+ : a(t) \leqslant x \leqslant b(t)\}$ and $E(t)$ is compact for each $t \geqslant 0$. Meanwhile, F_θ is locally Lipschitz continuous for a Lipschitz continuous w. Since $E(t)$ is compact, the restriction of F_θ into $E(t)$ is Lipschitz continuous. Applying Theorem 3.2 in [15] that is an extension of the theorem known as Picard-Lindelöf theorem or Cauchy-Lipschitz theorem, we have the existence and uniqueness of the solution on each bounded interval $[0, t]$. Since t is arbitrary, we have the proposition. □

Now that we know that solutions of the ES-IGO ODE exist and are unique, we define the ES-IGO-flow as the mapping $\varphi : \mathbb{R}_{\geqslant 0} \times \Theta \to \Theta$, which maps (t, θ_0) to the solution $\theta(t)$ of (5) with initial condition $\theta(0) = \theta_0$. Note that we can extend the domain of F_θ from $\Theta = \mathbb{R}^d \times \mathbb{R}_+$ to $\overline{\Theta} = \mathbb{R}^d \times \mathbb{R}_{\geqslant 0}$. It is easy to see from (5) that the value of $F_\theta(\theta)$ at $\theta = (m, 0)$ is 0 for any $m \in \mathbb{R}^d$. However, we exclude the boundary $\partial\Theta$ from the domain for reasons that will become clear in the next section. Because the initial variance must be positive and the variance starting from positive region never reach the boundary in finite time, solutions $\varphi(t, \cdot)$ will stay in the domain Θ. However, as we will see, they can converge asymptotically towards points of the boundary.

Since J_θ is *adaptive*, i.e. $J_{\theta_1}(\theta) \neq J_{\theta_2}(\theta)$ for $\theta_1 \neq \theta_2$ in general, it is not trivial to determine whether the solutions to (2) converge to points where $F_\theta(\theta) = 0$[1]. Even knowing that they converge to zeros of $F_\theta(\theta)$ is not helpful at all, because we have $F_\theta(\theta) = 0$ for any θ with variance zero and we are actually interested in convergence to the point $(\mathrm{x}^*, 0)$ where x^* is a local optimum of f.

Remark 1. Because of the invariance property of the natural gradient, the mean vector $m(\theta)$ and the variance $v(\theta)$ obey (6) and (7) under re-parameterization of the Gaussian distributions. Therefore, the trajectories of m and v are also independent of the parameterization. For instance, we obtain the same trajectories $v(\theta)$ for any of the following parameterizations: $\theta_{d+1} = v$, $\theta_{d+1} = \sqrt{v}$, and $\theta_{d+1} = \frac{1}{2}\ln v$, although the trajectories of the parameters θ_{d+1} are of course different. Consequently, the same convergence results for $m(\theta)$ and $v(\theta)$ (see Section 4) will hold under any parameterization. Parameterizations $\theta = (m, v)$ and $\theta = (m, \frac{1}{2}\ln v)$ correspond to the pure rank-μ update CMA-ES and the xNES with only one variance parameter. Thus, the continuous model to be analyzed encompasses both algorithms.

Remark 2. Theory of stochastic approximation says that a stochastic algorithm $\theta^{t+1} = \theta^t + \eta h^t$ follows the solution trajectories of the ODE $\frac{\mathrm{d}\theta}{\mathrm{dt}} = \mathbb{E}[h^t \mid \theta^t = \theta]$ in the limit

[1] If J_θ is not adaptive and defined to be the expectation of the objective function $f(x)$ over P_θ, convergence to the zeros of the RHS of (2) is easily obtained. For example, see Theorem 12 and its proof in [13], where the solution to the system of a similar ODE whose RHS is the vanilla gradient of the expected objective function is derived and the convergence of the solution trajectory to the critical point of the expected function is proven.

for η to zero under several conditions. In our setting, θ encodes m and v and the noisy observation $h^t = \sum_{i=1}^{\lambda} w_{R_i} \mathcal{I}_\theta^{-1} \nabla_\theta \ln p_{\theta^t}(x_i)$, where w_i, $i = 1, \ldots, \lambda$, are predefined weights and R_i is the ranking of x_i. If we define $w(p) = \sum_{i=1}^{\lambda} w_i \binom{\lambda-1}{i-1} p^{i-1}(1-p)^{\lambda-i}$ in (1), then $F_\theta(\theta) = \mathbb{E}[h^t \mid \theta^t = \theta]$ and the ODE agrees with (5). Therefore, (5) can be viewed as the limit behavior of adaptive-ES algorithms not only in the case $\eta \to 0$ and $\lambda \to \infty$ but also in the case $\eta \to 0$ and finite λ. Indeed, it is possible to bound the difference between $\{\theta^t, t \geqslant 0\}$ and the solution $\theta(\cdot)$ of the ODE (5) by extending Lemma 1 in Chapter 9 of [16].The details are omitted due to the space limitation.[2]

3 Extension of Lyapunov Stability Theorem

When convergence occurs, the variance typically converges to zero. Hence the study of the convergence of the solutions of the ODE will be carried out by analyzing the stability of the points $\theta^* = (\mathrm{x}^*, 0)$. However, because points with variance zero are excluded from the domain Θ, we need to extend classical definitions of stability to be able to handle points located on the boundary of Θ.

Definition 1 (Stability). *Consider the following system of differential equation*

$$\dot{\theta} = F(\theta), \quad \theta(0) = \theta_0 \in D, \tag{8}$$

where $F : D \mapsto \mathbb{R}^{d_\theta}$ is a continuous map and $D \subset \mathbb{R}^{d_\theta}$ is open. Then $\theta^* \in \overline{D}$ is called

- *stable in the sense of Lyapunov*[3] *if for any $\varepsilon > 0$ there is $\delta > 0$ such that $\theta_0 \in D \cap \overline{B}(\theta^*, \delta) \implies \theta(t) \in D \cap \overline{B}(\theta^*, \varepsilon)$ for all $t \geqslant 0$, where $t \mapsto \theta(t)$ is any solution of* (8)*;*
- *locally attractive if there is $\delta > 0$ such that $\theta_0 \in D \cap \overline{B}(\theta^*, \delta) \implies \lim_{t\to\infty} \|\theta(t) - \theta^*\| = 0$ for any solution $t \mapsto \theta(t)$ of* (8)*;*
- *globally attractive if $\lim_{t\to\infty} \|\theta(t) - \theta^*\| = 0$ for any $\theta_0 \in D$ and any solution $t \mapsto \theta(t)$ of* (8)*;*
- *locally asymptotically stable if it is stable and locally attractive;*
- *globally asymptotically stable if it is stable and globally attractive.*

We can now understand why we need to exclude points with variance zero from the domain Θ. Indeed, points with variance zero are points from where solutions of the ODE will never move because $F_\theta(\theta) = 0$. Consequently, if we include points $(x, 0)$

[2] When $H(\theta)$ is a (natural) gradient of a function, the stochastic algorithm is called a stochastic gradient method. The theory of stochastic gradient method (e.g., [17]) relates the convergence of the stochastic algorithm with the zeros of $H(\theta)$. However, it is not applicable to our algorithm due to the reason mentioned above Remark 1.

[3] Usually, stability is defined for stationary points. However, it is not the only case that a point is stable in our definition. Let $\theta^* \in \overline{D}$ be a stable point. If $\theta^* \in D$ or F can be prolonged by continuity at θ^* as $\lim_{\theta\to\theta^*} F(\theta) = F(\theta^*)$, then $F(\theta^*) = 0$. That is, θ^* is a stationary point. However, $\lim_{\theta\to\theta^*} F(\theta)$ does not always exist for a stable boundary point $\theta^* \in \partial D$. For example, consider the ODE: $\mathrm{d}\theta_1/\mathrm{d}t = -\theta_1/\sqrt{\theta_1^2 + \theta_2^2}$, $\mathrm{d}\theta_2/\mathrm{d}t = -\theta_2$. The domain is $\mathbb{R} \times \mathbb{R}_+$. Then, $|\theta_1|$ and θ_2 are monotonically decreasing to zero. Hence, $(0, 0)$ is globally asymptotically stable. However, $\lim_{\theta\to(0,0)} F(\theta)$ does not exist.

in Θ, none of these points can be attractive as in a neighborhood we always find $\theta_0 = (x_0, 0)$ such that a solution starting in θ_0 stays there and cannot thus converge to any other point.

A standard technique to prove stability is Lyapunov's method that consists in finding a scalar function $V : \mathbb{R}^{d_\theta} \to \mathbb{R}_{\geqslant 0}$ that is positive except for a candidate stable point θ^* with $V(\theta^*) = 0$, and that is monotonically decreasing along any trajectory of the ODE. Such a function is called *Lyapunov function* (and is analogous to a potential function in dynamical systems). Lyapunov's method does not require the analysis of the solutions of the ODE. The standard Lyapunov's stability theorem gives practical conditions to verify that a function is indeed a Lyapunov function. However, because our candidate stable points are located on $\partial\Theta$, we need to extend this standard theorem.

Lemma 1 (Extended Lyapunov Stability Method). *Consider the autonomous system* (8)*, where $F : D \to \mathbb{R}^{d_\theta}$ is a map and $D \subset \mathbb{R}^{d_\theta}$ is the open domain of θ. Let $\theta^* \in \overline{D}$ be a candidate stable point. Suppose that there is an $R > 0$ such that*
(A1)*: $F(\theta)$ is continuous on $D \cap B(\theta^*, R)$;*
(A2)*: there is a continuously differentiable $V : \mathbb{R}^{d_\theta} \to \mathbb{R}$ such that for some strictly increasing continuous function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying* $\lim_{p\to\infty} \alpha(p) = \infty$*,*

$$V(\theta^*) = 0, \quad V(\theta) \geqslant \alpha(\|\theta - \theta^*\|) \quad \forall \theta \in D \cap B(\theta^*, R) \setminus \{\theta^*\} \tag{9}$$

$$\text{and} \qquad \nabla V(\theta)^{\mathrm{T}} F(\theta) < 0 \quad \forall \theta \in D \cap B(\theta^*, R) \setminus \{\theta^*\}; \tag{10}$$

(A3)*: for any r_1 and r_2 such that $0 < r_1 \leqslant r_2 < R$, if a solution $\theta(\cdot)$ to* (8) *starting from $D_{r_1,r_2} = \{\theta \in D : r_1 \leqslant \|\theta - \theta^*\| \leqslant r_2\}$ stays in D_{r_1,r_2} for $t \in [0, \infty)$, then there is a $T \geqslant 0$ and a* compact *set $E \subset D_{r_1,r_2}$ such that $\theta(t) \in E$ for $t \in [T, \infty)$.*

Then, θ^* is locally asymptotically stable. If (A1) *and* (A2) *hold with D replacing $D \cap B(\theta^*, R)$ and* (A3) *holds with $R = \infty$, then θ^* is globally asymptotically stable.*

Proof. We follow the proof of Theorem 4.1 in [15]. We have from assumptions (A1) and (A2) that there is $\delta < R$ such that θ^* is stable and $V(\theta(t)) \to \tilde{V} \geqslant 0$ for each $\theta_0 \in D \cap B(\theta^*, \delta)$. Moreover, under (A1) and (A2) with D replacing $D \cap B(\theta^*, R)$ we have that $V(\theta(t)) \to \tilde{V} \geqslant 0$ for each $\theta_0 \in D$. Since $\lim_{t\to\infty} V(\theta(t)) \to 0$ implies $\lim_{t\to\infty} \|\theta - \theta^*\| = 0$ by (9), it is enough to show $\tilde{V} = 0$. We show $\tilde{V} = 0$ by contradiction argument. Assume that $\tilde{V} > 0$. Then, we have that for each $\theta_0 \in D$ (or $\in D \cap B(\theta^*, \delta)$ for the case of local asymptotic stability) there are r_1 and r_2 such that $0 < r_1 \leqslant r_2$ ($\leqslant \delta$) and $\theta(t)$ lies in D_{r_1,r_2} for $t \geqslant 0$. Note that D_{r_1,r_2} is not necessarily a compact set. This is different from Theorem 4.1 in [15]. By assumption (A3) we have that there is a compact set E and $T \geqslant 0$ such that $\theta(t) \in E$ for $t \geqslant T$. Since V is continuously differentiable and F is continuous, $\nabla V(\theta)^{\mathrm{T}} F(\theta)$ is continuous. Then, the function $\theta \mapsto V(\theta)^{\mathrm{T}} F(\theta)$ has its maximum $-\beta$ on the compact E and $-\beta < 0$ by (10). This leads to $V(\theta(t)) \leqslant V(\theta(T)) - \beta(t - T) \downarrow -\infty$ as $t \to \infty$. This contradicts the hypothesis that $V > 0$. Hence, $\tilde{V} = 0$ for any $\theta_0 \in D$ (or $\in D \cap B(\theta^*, \delta)$). □

4 Convergence of the ES-IGO-Flow

In this section we study the convergence properties of the ES-IGO-flow $\varphi : (t, \theta_0) \mapsto \theta(t)$, where $\theta(\cdot)$ represents the solution to the ES-IGO ODE (5) with initial value

$\theta(0) = \theta_0$, i.e., $\frac{\mathrm{d}\varphi(t,\theta_0)}{\mathrm{d}t} = F_\theta(\varphi(t,\theta_0))$ and $\varphi(0,\theta_0) = \theta_0$. By the definition of asymptotic stability, the global asymptotic stability of $\theta^* \in \overline{\Theta}$ implies the global convergence, that is, $\lim_{t\to\infty} \varphi(t,\theta_0) = \theta^*$ for all $\theta_0 \in \Theta$. Moreover, the local asymptotic stability of $\theta^* \in \overline{\Theta}$ implies the local convergence, that is, $\exists \delta > 0$ such that $\lim_{t\to\infty} \varphi(t,\theta_0) = \theta^*$ for all $\theta_0 \in \Theta \cap B(\theta^*, \delta)$. We will prove convergence properties of the ES-IGO-flow by applying Lemma 1. In order to prove our result we need to make the following assumption on w:
(B1): w is non-increasing and Lipschitz continuous with $w(0) > w(1)$;
(B2): $\int w(P_1[y : y \leqslant z])(z^2/d - 1/d)P_1(\mathrm{d}z) = \alpha > 0$.

Assumption (B1) is not restrictive. Indeed, the non-increasing and non-constant property of $w(\cdot)$ is a natural requirement and any weight setting in (4) can be expressed, for any given population size n, as a discretization of some Lipschitz continuous weight function. Assumption (B2) is satisfied if and only if the variance v diverges exponentially on a linear function. In fact, $F_v(\theta)$ defined in (7) reduces to $v \int w(P_1[y : y \leqslant z])(z^2/d - 1/d)P_1(\mathrm{d}z)$ when $f(x) = a^{\mathrm{T}}x$ for $\forall a \in \mathbb{R}^d \setminus \{0\}$ and we have that $\dot{v} = \alpha v$ and the solution is $v(t) = v_0 \exp(\alpha t)$. Then, $v(t) \to \infty$ as $t \to \infty$. Assumption (B2) holds, for example, if w is convex and not linear.

Let $\mathcal{G}$ be the set of strictly increasing functions $g : \mathbb{R} \to \mathbb{R}$ that are μ_{Leb}-measurable and $\mathcal{C}^2$ be the set of twice continuously differentiable functions $h : \mathbb{R}^d \to \mathbb{R}$ that are μ_{Leb}-measurable. Under (B1) and (B2), we have the following main theorems.

Theorem 1. *Suppose that the objective function f is a monotonic convex-quadratic-composite function $g \circ h$, where $g \in \mathcal{G}$ and h is a convex quadratic function $x \mapsto (x - \mathrm{x}^*)^{\mathrm{T}} A(x - \mathrm{x}^*)/2$ where A is positive definite and symmetric. Assume that* (B1) *and* (B2) *hold. Then, $\theta^* = (\mathrm{x}^*, 0) \in \overline{\Theta}$ is the globally asymptotically stable point of the ES-IGO. Hence, we have the global convergence of $\varphi(t,\theta_0)$ to θ^*.*

Proof. Since the ES-IGO does not explicitly utilize the function values but uses the quantile $P_\theta[y : f(y) \leqslant f(x)]$ which is equivalent to $P_\theta[y : g^{-1} \circ f(y) \leqslant g^{-1} \circ f(x)]$, without loss of generality we assume $f = h$.
According to Lemma 1, it is enough to show that (A1) and (A2) hold with $D(= \Theta)$ replacing $D \cap B(\theta^*, R)$ and (A3) holds with $R = \infty$. As is mentioned in the proof of Proposition 1, F_θ is locally Lipschitz continuous for a Lipschitz continuous w. Thus, (A1) is satisfied under (B1).

We can choose as a Lyapunov candidate function $V(\theta) = \sum_{i=1}^{d} (m_i - \mathrm{x}^*{}_i)^2 + d \cdot v = \|m - \mathrm{x}^*\|^2 + \mathrm{Tr}(vI_d)$. All the conditions on V described in (A2) are obvious except for the negativeness of $\nabla V(\theta)^{\mathrm{T}} F_\theta(\theta)$. To show the negativeness, rewrite $F_\theta(\theta)$ as $\int W_\theta^f(m + \sqrt{v}z)F_\theta(\theta, z)P_d(\mathrm{d}z)$. The idea is to show the (strictly) negative correlation between $W_\theta^f(m + \sqrt{v}z)$ and $\nabla V(\theta)^{\mathrm{T}} F_\theta(\theta, z)$ by using an extension of the result in [18, Chapter 1] and apply the inequality $\int W_\theta^f(m + \sqrt{v}z)\nabla V(\theta)^{\mathrm{T}} F_\theta(\theta, z)P_d(\mathrm{d}z) < \int W_\theta^f(m + \sqrt{v}z)P_d(\mathrm{d}z) \int \nabla V(\theta)^{\mathrm{T}} F_\theta(\theta, z)P_d(\mathrm{d}z) = 0$. We use the non-increasing property of w with $w(0) > w(1)$ in (B1) to show the negative correlation.

To prove (A3), we require (B2). Since a continuously differentiable function can be approximated by a linear function at any non-critical point $\bar{\mathrm{x}}$, the natural gradient F_θ is approximated by that on a linear function in a small neighborhood of $(\bar{\mathrm{x}}, 0)$. We use the property $\mu_{\mathrm{Leb}}[x : f(x) = \bar{f}] = 0$ to approximate F_θ. As is mentioned above, (B2) implies F_v on a linear function is positive. By using the approximation and this property, we can show that $E = D_{r_1,r_2} \cap \{\theta : v \geqslant \bar{v}\}$ satisfies (A3) for some $\bar{v} > 0$. □

We have that for any initial condition $\theta(0) = (m_0, v_0)$, the search distribution P_θ weakly converges to the Dirac measure δ_{x^*} concentrated at the global minimum point x^*. This result is generalized to monotonic $\mathcal{C}^2$-composite functions using a quadratic Taylor approximation. However, global convergence becomes local convergence.

Theorem 2. *Suppose that the objective function f is a monotonic $\mathcal{C}^2$-composite function $g \circ h$, where $g \in \mathcal{G}$ and $h \in \mathcal{C}^2$ has the property that $\mu_{\mathrm{Leb}}[x : h(x) = s] = 0$ for any $s \in \mathbb{R}$. Assume that* (B1) *and* (B2) *hold. Let* x^* *be a critical point of h, i.e.* $\nabla h(\mathrm{x}^*) = 0$, *with a positive definite Hessian matrix A. Then, $\theta^* = (\mathrm{x}^*, 0) \in \overline{\Theta}$ is a locally asymptotically stable point of the ES-IGO. Hence, we have the local convergence of $\varphi(t, \theta_0)$ to θ^*. Moreover, if $\bar{\mathrm{x}}$ is not a critical point of $h(\cdot)$, for any $\theta_0 \in \Theta$, $\varphi(t, \theta_0)$ will never converge to $\bar{\theta} = (\bar{\mathrm{x}}, 0)$.*

Proof. As in the proof of Theorem 1, we assume $f = h$ without loss of generality. The proofs of (A1) and (A3) carry over from Theorem 1 because we only used the property $\mu_{\mathrm{Leb}}[x : f(x) = \bar{f}] = 0$. To show (A2), we use the Taylor approximation of the objective function f. Since f is approximated by a quadratic function in a neighborhood of a critical point x^*, we approximate the natural gradient by the corresponding natural gradient on the quadratic function. Then, employing the same Lyapunov candidate function as in the previous theorem we can show (A2). Because of the approximation, we only have local asymptotic stability. The last statement of Theorem 2 is an immediate consequence of the approximation of the natural gradient and (B2). □

We have that starting from a point close enough to a local minimum point x^* with a sufficiently small initial variance, the search distribution weakly converges to δ_{x^*}. It is not guaranteed for the parameter to converge somewhere when the initial mean is not close enough to the local optimum or the initial variance is not small enough. Theorem 2 also states that the convergence $(m(t), v(t)) \to (\bar{\mathrm{x}}, 0)$ does not happen for $\bar{\mathrm{x}}$ such that $\nabla h(\bar{\mathrm{x}}) \neq 0$. That is, the continuous time ES-IGO does not prematurely converge on a slope of the landscape of f.

5 Conclusion

In this paper we have proven the local convergence of the continuous time model associated to step-size adaptive ESs towards local minima on monotonic $\mathcal{C}^2$-composite functions. In the case of monotonic convex-quadratic-composite functions we have proven the global convergence, i.e. convergence independently of the initial condition (provided the initial step-size is strictly positive) towards the unique minimum. Our analysis relies on investigating the stability of critical points associated to the underlying ODE that follows from the Information Geometric Optimization setting. We use a classical method for the analysis of stability of critical points, based on Lyapunov functions. We have however extended the method to be able to handle convergence towards solutions at the boundary of the ODE definition domain. We believe that our approach is general enough to handle more difficult cases like the CMA-ES with a more general covariance matrix. We want to emphasize that the model we have analyzed is the correct model for step-size *adaptive* ESs as the ODE encodes both the mean vector *and* step-size and preserves fundamental invariance properties of the algorithm.

Acknowledgments. This work was partially supported by the ANR-2010-COSI-002 grant (SIMINOLE) of the French National Research Agency and the ANR COSINUS project ANR-08-COSI-007-12.

References

1. Auger, A.: Convergence results for the (1, λ)-SA-ES using the theory of φ-irreducible Markov chains. Theoretical Computer Science 334(1-3), 35–69 (2005)
2. Jägersküpper, J.: Probabilistic runtime analysis of (1+, λ), ES using isotropic mutations. In: Proceedings of the 2006 Genetic and Evolutionary Computation Conference, GECCO 2006, pp. 461–468. ACM (2006)
3. Jägersküpper, J.: How the (1 + 1) ES using isotropic mutations minimizes positive definite quadratic forms. Theoretical Computer Science 361(1), 38–56 (2006)
4. Jägersküpper, J.: Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. Theoretical Computer Science 379(3), 329–347 (2007)
5. Arnold, L., Auger, A., Hansen, N., Ollivier, Y.: Information-geometric optimization algorithms: a unifying picture via invariance principles. arXiv:1106.3708v1 (2011)
6. Hansen, N., Muller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evolutionary Computation 11(1), 1–18 (2003)
7. Baluja, S., Caruana, R.: Removing the genetics from the standard genetic algorithm. In: Proceedings of the 12th International Conference on Machine Learning (1995)
8. Ostermeier, A., Gawelczyk, A., Hansen, N.: A derandomized approach to self-adaptation of evolution strategies. Evolutionary Computation 2(4), 369–380 (1994)
9. Glasmachers, T., Schaul, T., Yi, S., Wierstra, D., Schmidhuber, J.: Exponential natural evolution strategies. In: Proceedings of Genetic and Evolutionary Computation Conference, pp. 393–400. ACM (2010)
10. Yin, G.G., Rudolph, G., Schwefel, H.P.: Establishing connections between evolutionary algorithms and stochastic approximation. Informatica 1, 93–116 (1995)
11. Yin, G.G., Rudolph, G., Schwefel, H.P.: Analyzing the (1, λ) evolution strategy via stochastic approximation methods. Evolutionary Computation 3(4), 473–489 (1996)
12. Kushner, H.J., Yin, G.G.: Stochastic approximation and recursive algorithms and applications, 2nd edn. Springer (2003)
13. Malagò, L., Matteucci, M., Pistone, G.: Towards the geometry of estimation of distribution algorithms based on the exponential family. In: Proceedings of Foundations of Genetic Algorithms (FOGA 2011), pp. 230–242. ACM (2011)
14. Akimoto, Y., Nagata, Y., Ono, I., Kobayashi, S.: Theoretical foundation for CMA-ES from information geometry perspective. Algorithmica, Online First (2011)
15. Khalil, H.K.: Nonlinear systems. Prentice-Hall, Inc. (2002)
16. Borkar, V.S.: Stochastic approximation: a dynamical systems viewpoint. Cambridge University Press (2008)
17. Bonnabel, S.: Stochastic gradient descent on Riemannian manifolds. arXiv:1111.5280v2 (2011)
18. Thorisson, H.: Coupling, stationarity, and regeneration. Springer (2000)