Convergence of the IGO-Flow of Isotropic Gaussian Distributions on Convex Quadratic Problems

Tobias Glasmachers

Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany
tobias.glasmachers@ini.rub.de

Abstract. The information geometric optimization (IGO) flow has been introduced recently by Arnold et al. This distinguished mathematical flow on the parameter manifold of a family of search distributions constitutes a novel approach to the analysis of several randomized search heuristics, including modern evolution strategies. Besides its appealing theoretical properties, it offers the unique opportunity to approach the convergence analysis of evolution strategies in two independent steps. The first step is the analysis of the flow itself, or more precisely, the convergence of its trajectories to Dirac peaks over the optimum. In a second step it remains to study the deviation of actual algorithm trajectories from the continuous flow. The present study approaches the first problem. The IGO flow of isotropic Gaussian search distributions is analyzed on convex, quadratic fitness functions. Convergence of all trajectories to the Dirac peak over the optimum is established.

1 Introduction

Our theoretical understanding of evolution strategies (ESs) lags behind their practical successes. ESs are powerful optimization techniques that can work under adverse conditions, like non-smooth, discontinuous, or even noisy fitness functions. However, useful convergence guarantees exist only for the simplest algorithms on restricted problem classes [5,2]. Narrowing this gap between practically relevant problems and theoretical guarantees is a long-standing goal of evolutionary algorithms research.

In this context we view the recently introduced information geometric optimization (IGO) flow [1] as a promising tool towards a unified analysis of randomized search algorithms. Various invariance properties make this flow on the parameter manifold of a family of search distributions a canonical means for optimization. It can be interpreted as a continuous time version of various iterative, randomized search techniques. In particular, it resembles the behavior of existing evolution strategies [4,3] in the limit of large populations and small search strategy updates.

This hints at a two-step analysis: Convergence of the flow trajectories on an as large as possible class of problems should be separated from bounding

C.A. Coello Coello et al. (Eds.): PPSN 2012, Part I, LNCS 7491, pp. 1–10, 2012.

the deviation of discrete trajectories of actual algorithms from the continuous trajectories of the flow. With the present work we progress towards the first goal. We provide a complete convergence analysis of the IGO flow of isotropic Gaussian distributions on convex, quadratic fitness functions to the Dirac peak over the optimum. This problem class is of prime interest, since it approximates local optima of twice differentiable fitness functions. The result is non-trivial, since there exist counter examples where the flow converges prematurely.

2 The Information Geometric Optimization Flow

The IGO flow is defined in the context of randomized search for the minimum of a fitness function $f : X \to \mathbb{R}$ in the black box model. Iterative, randomized search algorithms like evolutionary algorithms can be interpreted as defining a sequence of search distributions. The IGO flow resembles this process with a continuous time flow on the parameter manifold Θ of a family P_θ of search distributions (with densities p_θ). In the limit of large populations and small learning rates popular ESs such as CMA-ES [4] and NES [6,3] closely follow this flow [1]. The IGO framework lifts optimization from the search space X to the parameter manifold Θ. For example, for isotropic Gaussians the parameter space $\Theta = \mathbb{R}^d \times \mathbb{R}^+$ is composed of the mean vector and the standard deviation.

In a first step the fitness function is normalized w.r.t. the current search distribution, which also makes it invariant under monotonic transformations. We need the following notation. Let $B(x_0, r) = \{x \in \mathbb{R}^d \,|\, \|x - x_0\| < r\}$ denote the open ball of radius r around x_0. Let $u^f(y) = \{x \in \mathbb{R}^d \,|\, f(x) < y\}$ denote the sub-level sets of the fitness function, and let $q^f_\theta(y) = P_\theta(u^f(y))$ denote the (lower) quantile function, measuring the probability to sample a solution x with fitness $f(x)$ better than y under the search distribution encoded by θ. This function is assumed to be continuous.[1] Combining these definitions we write $u^f_\theta(q) = u^f(y)$ if $q = q^f_\theta(y)$. The composition $q^f_\theta \circ f$ assigns to each point the probability to sample a better point under P_θ. Importantly, this is a monotone, rank preserving transformation of the fitness function, which is itself (by construction) invariant under rank-preserving transformations of the fitness values.

In a second step a non-increasing weight function $w : [0, 1] \to \mathbb{R}$ is introduced that puts user-defined emphasis on different quantiles. A simple choice is the indicator $w = \mathbf{1}_{[0,q]}$ for some quantile q. The function $W^f_\theta = w \circ q^f_\theta \circ f$ is a monotonically decreasing transformation of f. Thus, for fixed θ, maximization of W^f_θ is equivalent to minimization of f. This is an objective function on the search space X, which can be transferred to the parameter manifold Θ in the form $J(\theta, \theta') = \mathbb{E}_{\theta'}[W^f_\theta(x)]$. For fixed θ this is an objective function in θ'.

The parameter manifold Θ is naturally equipped with the Fisher metric. Maximization in the resulting statistical manifold can be achieved locally by gradient

[1] Otherwise the subsequent technical analysis is unnecessarily complicated by the need to distinguish upper and lower quantiles, see e.g. [1], equation (3). The assumption is always fulfilled for the distributions and fitness functions considered in this paper.

ascent. The gradient in the inner geometry of distributions pulled back to the parameter manifold Θ is the natural gradient, denoted by the symbol $\widetilde{\nabla}$. Steepest ascent is thus realized by following the vector field $V(\theta) = \widetilde{\nabla}_{\theta'}|_{\theta'=\theta} J(\theta, \theta')$. The formula

$$V(\theta) = \int W_\theta^f(x)\, \widetilde{\nabla}_{\theta'} \log\left(p_{\theta'}(x)\right) dP_\theta(x) \tag{1}$$

(equation (10) in [1]) connects the vector field V to the natural gradient of the logarithmic density in equation (2). It also ensures the continuity of V, provided that p_θ is non-zero and continuous.

The IGO flow is the solution of the differential equation $\dot{\phi}^t(\theta) = V(\phi^t(\theta))$. Here $\phi^t(\theta)$ denotes a trajectory with initial condition $\phi^0(\theta) = \theta$. The upper index t denotes time. This flow is invariant under coordinate changes of θ and under rank-preserving (strictly monotone) transformations of fitness values [1].

The IGO vector field is defined by means of a natural gradient operator. However, its definition is *not* of the form $V(\theta) = \widetilde{\nabla}_\theta J'(\theta)$ for some potential function J'. The existence of such a potential function would greatly simplify the analysis of the IGO flow, but there are counter-examples where it does not exist. It remains unclear whether such a function exists for the case of isotropic Gaussians and convex, quadratic fitness function.

In this context it is worth mentioning that the family of NES algorithms [6,3] is commonly derived for the potential function $J'(\theta) = \mathbb{E}_\theta\left[f(x)\right]$ of expected fitness. It has been argued in [1] that practical NES algorithms follow the IGO flow instead. This is because NES algorithms are rendered invariant under rank-preserving transformations of the fitness function by a technique called fitness shaping. Expected fitness has the desirable property to be a potential function of the corresponding flow. However, its drawbacks are that depending on the fitness function (over which there is no control in a black box setting) the expectation may not always exist, the resulting flow is not invariant under monotone fitness transformations, and existing algorithms do not approximate the corresponding flow. Consequently we focus on the IGO flow in this study, albeit expected fitness has its merits, e.g., on finite search spaces (where the expectation always exists).

Isotropic Gaussian search distributions on $X = \mathbb{R}^d$ with densities

$$p_{\mu,\sigma}(x) = \frac{1}{(\sqrt{2\pi}\cdot\sigma)^d} \cdot \exp\left(-\frac{\|x-\mu\|^2}{2\sigma^2}\right)$$

are characterized by a mean vector $\mu \in \mathbb{R}^d$ and a variance $\sigma^2 \in \mathbb{R}^+$. In this paper we use the parameterization $\theta = (\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}^+ = \Theta$. For isotropic Gaussians the flow is invariant under translation, scaling, and rotation of the search space (provided that the initial conditions are transformed accordingly).

The natural gradient of the logarithmic density (see equation (1)) can be computed as

$$G(\theta, x) = \widetilde{\nabla}_\theta \log\left(p_\theta(x)\right) = \begin{pmatrix} x-\mu \\ \frac{\sigma}{4d}\left[\left(\frac{\|x-\mu\|}{\sigma}\right)^2 - d\right] \end{pmatrix} . \tag{2}$$

Analogously, the vector field is decomposed into components $V = (V_\mu, V_\sigma)$ describing the evolution of the mean and the standard deviation under the flow.

The IGO flow of Gaussian distributions is of particular interest for its connection to evolution strategies [1,4,3].

3 Analysis of the IGO Flow

We start with the core technical lemma. This auxiliary result decomposes the IGO vector field into additive components, with each component corresponding to a tractable, geometric problem.

Lemma 1. *The IGO flow vector field $V(\theta)$ can be written in the form*

$$V(\theta) = \int_{[0,1]} \left[\int_{u_\theta^f(q)} G(\theta, x) \; dP_\theta(x) \right] dg(q)$$

$$= \int_{[0,1]\times[0,\infty)} \left[\int_{u_\theta^f(q)\cap B(\mu,r)} G(\theta, x) \; dx \right] dh_\theta(q, r)$$

w.r.t. non-negative measures $g(q)$ and $h_\theta(q, r)$.

Proof. We rewrite $W_\theta^f(x) = \int_0^1 \mathbf{1}_{u_\theta^f(q)}(x)\, dg(q)$ as an integral of constant functions on sub-level sets of f (which are super-level sets of W_θ^f). Analogously, we rewrite $P_{\mu,\sigma} = \int_0^\infty U_{B(\mu,r)}\, d\beta_\sigma(r)$ as a superposition of uniform distributions $U_{B(\mu,r)}$ over balls around the center μ. By construction the measures g and β_σ are non-negative. Plugging both decompositions into equation (1) and choosing h_θ as the product of g and β_σ completes the proof. ∎

The above lemma allows us to analyze the IGO flow based on the natural gradient of the logarithmic density given by equation (2), restricted to the intersection of a ball with a sub-level set. For convex functions the integration area $u_\theta^f(q)\cap B(\mu, r)$ is convex (possibly empty). This lemma will be applied multiple times in the following.

3.1 Linear Objective Functions

The goal of minimization of a linear fitness function $f(x) = v^T x$ is to move the center of the distribution into the direction $-v$ as quickly as possible, and to drive the step size σ to infinity. The invariance properties of the IGO flow allow us to assume $v = (1, 0, \ldots, 0) \in \mathbb{R}^d$, $\mu = 0$, and $\sigma = 1$.

Using Lemma 1 we write $V_\mu(0, 1)$ as an integral over terms of the form $\int_{u_\theta^f(q)\cap B(0,r)} x\, dx$. The half-space $u_\theta^f(q)$ is given by the inequality $x_1 < y$, with $y = 0$ for the median ($q = 1/2$). The inner product of the above term with

v yields the same expression with integrand x_1 (first component of x) instead of x. There are three cases: The integral is zero if the ball is fully contained in or disjoint to the half-space. Otherwise it is negative. It follows $V_\mu(0,1) = (-c, 0, \ldots, 0) \in \mathbb{R}^d$ for some $c > 0$, and for symmetry and invariance reasons it holds $V_\mu(\mu, \sigma) = -\sigma \cdot c \cdot v$. Thus, the flow moves the center μ in direction $-v$. However, it may converge prematurely if σ decays too quickly.

Lemma 1 allows us to write the component $V_\sigma(0,1)$ as an integral over terms of the form $\int_{u_\theta^f(q)} (\|x\|^2 - d)\, dP_{(0,1)}(x)$. The expectation of the integrand over the whole space vanishes, and so it does (for symmetry reasons) restricted to the half-space $x_1 < 0$ ($q = 1/2$). However, for $x_1 < y$ with $y < 0$ ($q < 1/2$) the integral is positive, since compared to the half-space $x_1 < 0$ probability mass is missing particularly for shorter-than-average vectors x. It follows with an analog argument that the integral is negative for $y > 0$ ($q > 1/2$). Thus, depending on the choice of the weight function w, it is possible that $V_\sigma(0,1)$ is negative. In this case the step size σ decays exponentially, resulting in (premature) convergence of the IGO flow trajectories to Dirac delta peaks. For example, for the so-called "selection quantile" weight function $w(t) = \mathbf{1}_{[0,q]}(t)$ trajectories convergence prematurely for $q > 1/2$, and σ grows exponentially for $q < 1/2$. The ability to handle a (close to) linear objective function is a must for any reasonable optimization scheme. Care should be taken to impose sufficient selection pressure by the choice of the weight function. This assumption is formalized as follows:

Assumption. Let L be defined as $V_\sigma(0,1)$ for a linear objective function $f(x) = v^T x$ with slope $\|v\| \neq 0$. Using translation and scale invariance this is equivalent to $V_\sigma(\mu, \sigma) = \sigma \cdot L$. We assume in the following that w is chosen such that $L > 0$.

3.2 Convex Quadratic Objective Functions

The core contribution of the present work is the analysis of the IGO flow on objective functions of the form $f(x) = x^T Q x$, where $Q \in \mathbb{R}^{d \times d}$ is symmetric and positive definite. This situation is analyzed in the following lemmas.

Lemma 2. *V is scale invariant: it holds $V(\lambda \cdot \theta) = \lambda \cdot V(\theta)$ for all $\lambda > 0$.*

Proof. The lemma follows directly from equations (1) and (2) and the scale invariance of the level sets of $f(x) = x^T Q x$. ■

As a consequence, the vector field V is fully described by its values on a section of co-dimension one through the equivalence classes $[\theta] = \mathbb{R}^+ \cdot \theta$ in Θ. The set $S = \left\{ \theta \in \Theta \,\middle|\, \|\theta\| = 1 \right\}$ is such a section, with $\|\cdot\|$ denoting the Euclidean two-norm on $\Theta \subset \mathbb{R}^{d+1}$.

Lemma 3. *For $\mu \neq 0$ the inner product $\langle V_\mu(\mu, \sigma), \mu \rangle$ is negative.*

Note that the above inner product is the time derivative of $\frac{1}{2}\|\mu\|^2$ under the flow (since by definition V_μ is the time derivative of μ). Thus, the center component μ moves towards the optimum, although not necessarily straight.

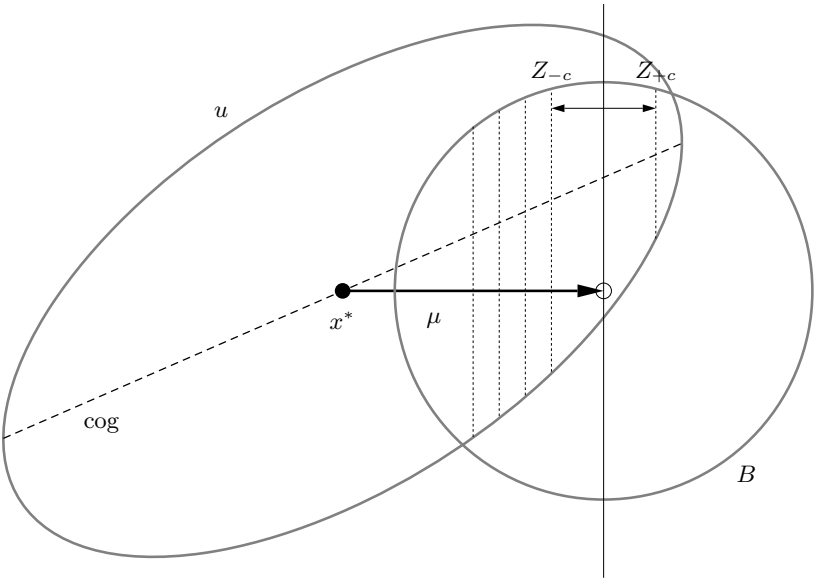

Fig. 1. The figure depicts the sets B (circular outline), u (elliptic outline), the optimum x^* in the origin, the mean vector μ (arrow), the hyperplane H_0 (vertical line), the parameterized line $\mathrm{cog}(c)$ of centers of gravity of $Y_c = H_c \cap u$ (dashed line), as well as a number of sets Z_c (vertical, dotted lines). Refer to the proof of Lemma 3 for further details.

Proof. This proof amounts to a non-trivial application of Lemma 1. The proof is based on an involved construction, see Figure 1.

Fix $q \in [0,1]$ and $r > 0$, and the corresponding sets $u = u_\theta^f(q)$ and $B = B(\mu, r)$. We define the hyperplanes $H_c = \{x \in \mathbb{R}^d \mid \langle x, \mu\rangle = \|\mu\|^2 + c\}$ orthogonal to μ, as well as their subsets $Y_c = H_c \cap u$ and $Z_c = Y_c \cap B$. Let $\mathcal{M}$ denote the (d-1)-dimensional Lebesgue measure on the hyperplanes H_c. Then the center of gravity of Y_c is defined as $\mathrm{cog}(c) = 1/\mathcal{M}(Y_c) \cdot \int_{Y_c} x\, dx$. For a convex, quadratic objective function the set u is the interior of an ellipsoid, and analogously, each set Y_c is the interior on an ellipsoid in $d-1$ dimensions. The centers of gravity $\mathrm{cog}(c)$ as a function of c form a parameterized line.

Recall that the μ-component of the natural gradient of the logarithmic density is $x - \mu$. The relevant expression for the application of Lemma 1 is the inner product of $x - \mu$ with μ. The sets Z_c form sections of $u \cap B$ such that $\langle x - \mu, \mu\rangle$ takes the constant value c. Now fix a positive constant $c > 0$ and consider the pair of sections Z_{+c} and Z_{-c}, as well as the translation $\psi_c : H_{+c} \to H_{-c}$ along the line $\mathrm{cog}(c)$.

By construction it holds $\psi_c(Y_{c+}) \subset Y_{-c}$, and again by construction it holds $\psi_c(Z_{+c}) \subset Z_{-c}$ (see Figure 1), and since the translation ψ_c is measure preserving

it follows $\mathcal{M}(Z_{+c}) \leq \mathcal{M}(Z_{-c})$. In those cases where Z_c is bounded by the ellipsoid and not only by the ball (these cases exist if $u \cap B \neq \emptyset$ and $u \cap B \neq B$) the inequality is strict, because the ellipsoid Y_{+c} is strictly smaller than for Y_{-c} (see also Figure 1).

The inner term of Lemma 1 projected onto the direction μ becomes

$$\begin{aligned}\left\langle \int_{u \cap B} (x - \mu)\, dx, \mu \right\rangle &= \int_{-\infty}^{\infty} c \cdot \mathcal{M}(Z_c)\, dc \\ &= \int_{0}^{\infty} c \cdot \big(\mathcal{M}(Z_{+c}) - \mathcal{M}(Z_{-c})\big)\, dc \;<\; 0 \;\; .\end{aligned}$$

Finally, the application of Lemma 1 yields $\langle V_\mu(\mu, \sigma), \mu \rangle < 0$. ■

The next three lemmas analyze the evolution of the step size. Their proofs rely on the following types of topological arguments: Continuous functions map compact sets in the preimage onto compact sets in the image, and preimages of open sets are open. This implies two handy properties: First, a continuous function attains infimum and supremum on a compact set, which means that minimum and maximum exist. Second, if a continuous function is positive in one point, then it is positive in a (small) open neighborhood of this point.

We define the set $M = (\mathbb{R}^d \times \mathbb{R}_0^+) \setminus \{(0, 0)\}$ and the continuous[2] function $n : M \to [0, \infty]$, $n(\mu, \sigma) = \|\mu\|/\sigma$, measuring normalized distance of the search distribution to the optimum. Because of $n(\theta) = n(\lambda \cdot \theta)$ for all $\theta \in M$ and $\lambda > 0$ the function is uniquely described by its values on the compact half-sphere $\overline{S} = \big\{\theta \in M \,\big|\, \|\theta\| = 1\big\}$, which is the topological closure of the open half-sphere $S \subset \Theta$.

Lemma 4. *It holds $V(0, \sigma) = (0, -c \cdot \sigma)$ for some $c > 0$.*

Proof. We apply Lemma 1 to compute $V_\mu(0, \sigma)$. The sub-level set $u_\theta^f(q)$ as well as the ball $B(\mu, r) = B(0, r)$ are symmetric around the origin, and so is their intersection. The inner term in the integration is x, such that the integral over $u_\theta^f(q) \cap B(0, r)$ vanishes.

The form $V_\sigma(0, \sigma) = -c \cdot \sigma$ follows from Lemma 2. It remains to show that V_σ is negative. We apply Lemma 1 again and consider the inner term

$$\int\limits_{u_\theta^f(q)} \frac{\sigma}{4d} \left[\left(\frac{\|x\|}{\sigma} \right)^2 - d \right] dP_\theta(x) \;\; .$$

The integration, when spanning the whole search space, amounts to zero. However, the set $u_\theta^f(q)$ is convex and symmetric around the origin and thus puts more probability mass on smaller-than-average vectors. As a result the above expression is negative, and we obtain $V_\sigma(0, \sigma) < 0$ from Lemma 1. ■

Lemma 5. *There exists $c_1 < \infty$ such that $n(\mu, \sigma) > c_1$ implies $V_\sigma(\mu, \sigma) > 0$.*

[2] The set $[0, \infty]$ is equipped with the standard one-point-compactification topology.

Proof. The objective function $f(x) = x^T Q x$ is differentiable and can thus, locally, be approximated arbitrarily well by its first order Taylor expansion. Thus, for fixed $\mu \neq 0$ the fitness approaches an affine linear function with non-zero slope in the limit $\sigma \to 0$. The limit $\lim_{\sigma \to 0} V_\sigma(\mu, \sigma)/\sigma = L > 0$ exists for all $\mu \neq 0$, and V_σ/σ is continuous. This allows us to extend the domain of V_σ/σ as a continuous function from Θ to M, or analogously from S to $\overline{S}$. Let $S_\mu = \left\{\mu \in \mathbb{R}^d \,\middle|\, \|\mu\| = 1\right\}$ denote the unit sphere in $\mathbb{R}^d$. We use $\overline{n} = n|_{\overline{S}}$ as a shorthand notation for the function n restricted to $\overline{S}$. Then the pre-image of infinity under $\overline{n}$ takes the form $\overline{n}^{-1}(\infty) = S_\mu \times \{0\} \subset M$, and the function V_σ/σ has the constant value L on this set.

The continuity of V_σ/σ implies that there exists an open neighborhood $N \subset \overline{S}$ of $S_\mu \times \{0\}$ with $V_\sigma(\mu, \sigma)/\sigma > 0$ for all $(\mu, \sigma) \in N$. The set $\overline{S} \setminus N$ is compact, and therefore also its image $\overline{n}(\overline{S} \setminus N)$. By construction this set does not contain infinity. Thus, the choice $c_1 = \max\left(\overline{n}(\overline{S} \setminus N)\right)$ concludes the proof. ∎

Lemma 6. *There exists $c_2 > 0$ such that $n(\mu, \sigma) < c_2$ implies $V_\sigma(\mu, \sigma) < 0$.*

Proof. The proof is analogous to the previous one. Consider the point $(\mu, \sigma) = (0, 1) \in \overline{S}$. Lemma 4 implies $V_\sigma(0, 1) < 0$, and it holds $\overline{n}^{-1}(\{0\}) = \{(0, 1)\}$. From the continuity of V_σ we conclude the existence of an open neighborhood $N' \subset \overline{S}$ of $(0, 1)$ with $V_\sigma(\mu, \sigma) < 0$ for all $(\mu, \sigma) \in N'$. The set $\overline{n}(\overline{S} \setminus N') \subset [0, \infty]$ is closed and does not contain zero, which allows for the choice $c_2 = \min\left(\overline{n}(\overline{S} \setminus N')\right)$. ∎

Theorem 1. *For all $\theta_0 \in \Theta$ the IGO flow trajectory $\phi^t(\theta_0)$ converges to a Dirac peak over the optimum: It holds $\lim_{t \to \infty} \phi^t(\theta_0) = (0, 0)$.*

Proof. For $b \geq 0$ we define the open neighborhood $B \subset \overline{\Theta} = \mathbb{R}^d \times \mathbb{R}_0^+$ of $\theta^* = (0, 0) \in \overline{\Theta}$ as $B = \{(\mu, \sigma) \in \overline{\Theta} \,|\, \sigma < b, \|\mu\| < c_2 \cdot b\}$. Since b is arbitrary, showing that the trajectory $\phi^t(\theta_0)$ enters B in finite time and stays there will prove the statement. Based on lemmas 5 and 6 we split the parameter space into three dynamic regimes

$$
\begin{aligned}
R_1 &= \{\theta_1 \in \Theta \,|\, c_1 \leq n(\theta_1)\} \\
R_2 &= \{\theta_2 \in \Theta \,|\, c_2 \leq n(\theta_2) \leq c_1\} \\
R_3 &= \{\theta_3 \in \Theta \,|\, n(\theta_3) \leq c_2\}
\end{aligned}
$$

of qualitatively different behavior. The constraints imposed by the various lemmas on the vector field are illustrated in Figure 2. In particular, Lemma 3 implies that the flow can only shrink μ, which corresponds to the vector field pointing to the "left" in Figure 2. In addition, the vertical component is by Lemma 5 restricted to point "upwards" ($V_\sigma > 0$) in R_1, and according to Lemma 6 "downwards" ($V_\sigma < 0$) in R_3.

For initial conditions $\theta_1 \in R_1$, $\theta_2 \in R_2$, or $\theta_3 \in R_3$ we define compact sets C_1, C_2, and C_3 in which the trajectory $\phi^t(\theta_i)$ is restricted to stay for $t > 0$ according to the above conditions until it enters the set B. Figure 2 (right) illustrates these sets, which will be considered w.l.o.g. as closed (otherwise consider the closure).

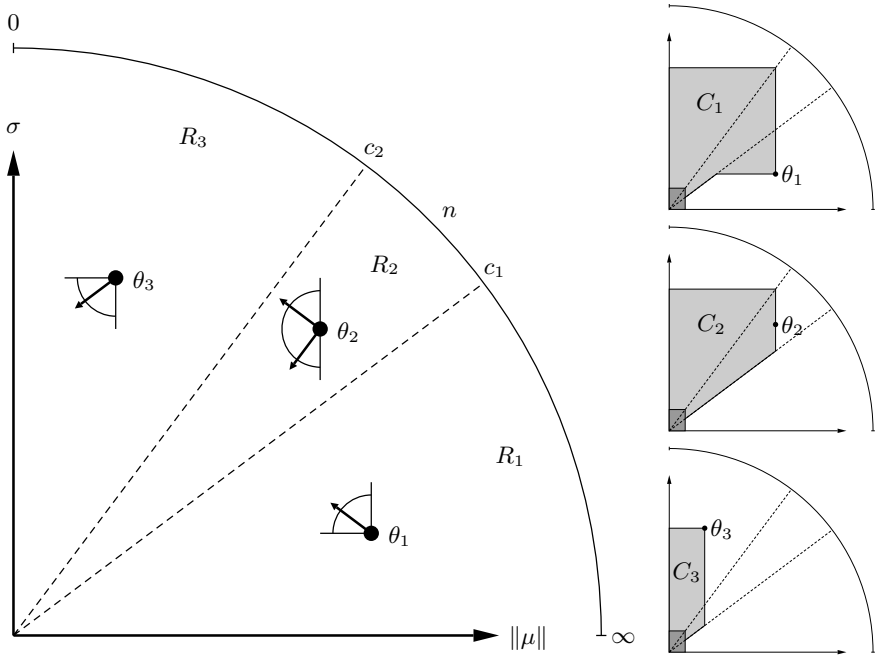

Fig. 2. Left: Illustration of the different dynamic regimes R_1, R_2, and R_3. The quarter-circles and the half-circle attached to the prototypical points $\theta_i \in R_i$, $i \in \{1, 2, 3\}$, illustrate how the vector field $V(\theta)$ is constrained by the various lemmas. Right: Illustration of the compact regions C_i (gray areas), in downscaled versions of the same figure. The second and third of the small figures also depict the open neighborhood B of $(\mu, \sigma) = (0, 0)$ (dark gray area).

They are compact, since they are also bounded away from infinity and from the boundary of Θ. These sets are split into

$$C_i' = \left\{ (\mu, \sigma) \in C_i \;\middle|\; \|\mu\| \geq \frac{c_2 \cdot b}{2} \right\} \quad \text{and} \quad C_i'' = \left\{ (\mu, \sigma) \in C_i \;\middle|\; \|\mu\| \leq \frac{c_2 \cdot b}{2} \right\} .$$

Lemma 3 together with the condition $\mu \geq c_2 \cdot b/2$ implies that restricted to the sets C_i' it holds $\langle V_\mu, \mu \rangle < 0$. Each of these sets is compact, and thus the maximum of this function exists, which is a negative value. This value provides a non-zero lower bound on the velocity of the movement of the trajectory towards smaller $\|\mu\|$ ("to the left" in Figure 2). Thus, the flow leaves the set C_i' in finite time. Assume the flow did not reach B, then it must enter the corresponding set C_i''. By construction, these compact sets are fully contained in regime R_3. There the function V_σ is negative, and with the same argument the maximum exists and is negative, which provides a lower bound on the velocity of the flow moving towards smaller σ ("downwards"). Thus, the flow enters B in finite time. The shape of B is constructed so that the flow stays inside (see Lemmas 3 and 6). ■

As a comment and without proof we want to add that the same compactness arguments give rise to the existence of a linear convergence rate.

4 Discussion

It has been proven that all trajectories of the IGO flow on isotropic Gaussian distributions converge to the Dirac peak over the optimum. Due to invariance properties this result holds for all convex quadratic functions and rank-preserving transformations thereof, given that the quantile weights are chosen so that the flow does not get stuck on a linear slope. The importance of this result is that it describes the dynamics of the flow in the proximity of local optima of twice differentiable fitness functions.

This is a promising result, although we view it rather as a first step. The author has good faith that most of the statements brought forward in the various lemmas can be generalized. This is because the proof techniques are kept as general as possible. In particular, geometric and topological arguments have been preferred over an algebraic treatment of the (linear or quadratic) objective function. Thus, large parts of the analysis should be generalizable, which holds in particular for the proof of the theorem.

This leaves us with a considerable body of future work. The analysis can be extended into different directions. First, the class of search distributions can be broadened. Gaussian distributions with fully adaptive covariance matrix are of primary interest, since the corresponding flow is resembled by state-of-the-art evolution strategies [1,4,3]. Second, the class of fitness functions can be extended. An ambitious goal is to cover the full class of all smooth, uni-modal problems. Third, the present understanding of how closely actual evolutionary algorithms follow the IGO flow is limited. The idea of transferring results from the IGO flow to evolution strategies drives the present ongoing investigation and is therefore a primary research goal.

References

1. Arnold, L., Auger, A., Hansen, N., Ollivier, Y.: Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. Technical Report arXiv:1106.3708v1, arxiv.org (2011)
2. Auger, A.: Convergence results for the $(1, \lambda)$-SA-ES using the theory of φ-irreducible Markov chains. Theoretical Computer Science 334(1-3), 35–69 (2005)
3. Glasmachers, T., Schaul, T., Sun, Y., Wierstra, D., Schmidhuber, J.: Exponential Natural Evolution Strategies. In: Genetic and Evolutionary Computation Conference, GECCO (2010)
4. Hansen, N., Ostermeier, A.: Completely Derandomized Self-Adaptation in Evolution Strategies. Evolutionary Computation 9(2), 159–195 (2001)
5. Jägersküpper, J.: Analysis of a Simple Evolutionary Algorithm for Minimization in Euclidean Spaces. In: Baeten, J.C.M., Lenstra, J.K., Parrow, J., Woeginger, G.J. (eds.) ICALP 2003. LNCS, vol. 2719, pp. 1068–1079. Springer, Heidelberg (2003)
6. Wierstra, D., Schaul, T., Peters, J., Schmidhuber, J.: Natural Evolution Strategies. In: Congress on Evolutionary Computation (CEC 2008), Hongkong. IEEE Press (2008)