Analysis of Information Geometric Optimization with Isotropic Gaussian Distribution Under Finite Samples

Kento Uchida
Yokohama National University
Yokohama, Japan
uchida-kento-nc@ynu.jp

Shinichi Shirakawa
Yokohama National University
Yokohama, Japan
shirakawa-shinichi-bg@ynu.ac.jp

Youhei Akimoto
University of Tsukuba
Tsukuba, Japan
akimoto@cs.tsukuba.ac.jp

ABSTRACT

In this article, we theoretically investigate the convergence properties of the information geometric optimization (IGO) algorithm given the family of isotropic Gaussian distributions on the sphere function. Differently from previous studies, where the exact natural gradient is taken, i.e., the infinite samples are assumed, we consider the case that the natural gradient is estimated from finite samples. We derive the rates of the expected decrease of the squared distance to the optimum and the variance parameter as functions of the learning rates, dimension, and sample size. From the rates of decrease deduces that the rates of decreases of the squared distance to the optimum and the variance parameter must agree for geometric convergence of the algorithm. In other words, the ratio between the squared distance to the optimum and the variance must be stable, which is observed empirically but is not derived in the previous theoretical studies. We further derive the condition on the learning rates that the rates of decreases agree and derive the stable value of the ratio. We confirm in simulation that the derived rates of decreases and the stable value of the ratio well approximate the behavior of the IGO algorithm.

CCS CONCEPTS

• **Theory of computation → Stochastic control and optimization**; **Theory of randomized search heuristics**;

KEYWORDS

Natural Gradient, Finite Samples, Convergence Property, Information Geometric Optimization, Theory

ACM Reference Format:
Kento Uchida, Shinichi Shirakawa, and Youhei Akimoto. 2018. Analysis of Information Geometric Optimization with Isotropic Gaussian Distribution Under Finite Samples. In *GECCO '18: Genetic and Evolutionary Computation Conference, July 15–19, 2018, Kyoto, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3205455.3205487

1 INTRODUCTION

Information geometric optimization (IGO) [14] is a unified framework of stochastic search algorithms for optimization problems on an arbitrary domain. The IGO algorithm takes a parametrized family of probability distributions as sampling distributions. Starting from a given initial distribution, the IGO algorithm repeats sampling candidate solutions, evaluating them on an objective, and updating the distribution parameters until it finds a satisfactory solution. The parameter update follows the steepest ascent step in the parameter space of the probability distributions equipped with the Fisher metric. The steepest ascent direction in such a space is known to be computed by the inverse Fisher information matrix times the vanilla gradient, called the natural gradient [5]. The natural gradient is estimated by Monte Carlo using the candidate solutions. The IGO framework recovers several known evolutionary algorithms such as the pure rank-μ update covariance matrix adaptation evolution strategy (CMA-ES) [12] and the population-based incremental learning (PBIL) [8].

The objective of this paper is to investigate the effect of the parameters of the IGO algorithm on its behavior. When instantiating an IGO algorithm, we have two important parameters, the sample size (aka population size, number of candidate solutions at each iteration) and the learning rates for the distribution parameter updates. However, the design principle of the IGO algorithm does not include the guide for these parameter settings. Since these parameters are critical in the performance of the algorithm, it is important to understand the influence of these parameters on the distribution parameter update.

Some variants of the IGO algorithm with a family of multivariate Gaussian distribution in continuous domain has been analyzed in the literature [1, 2, 9, 11]. Since the analysis of stochastic algorithms in continuous domain in general is usually mathematically intricate, different levels of approximations have been applied to model the dynamics of the real algorithms. One way is to consider the deterministic ordinary differential equation associated with the infinite sample size and the infinitesimal learning rate of the distribution parameter update [2, 9, 11]. Another way is to consider the deterministic difference equation associated with the infinite sample size but a finite learning rate [1]. In the latter case, the geometric convergence of the mean vector and the covariance matrix of the sampling distribution is derived. However, these theoretical results do not well describe the behavior of the real IGO algorithm. More precisely, it is observed that the convergence rate of the standard deviation of the sampling distribution is the same as that of the distance between the optimal solution and the mean vector, while the theoretical results under the infinite samples tell that these convergence speeds can be different. It indicates that we need to incorporate the stochastic property of the algorithm in some way into the analysis.

GECCO '18, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5618-3/18/07…$15.00
https://doi.org/10.1145/3205455.3205487

Different analysis methodologies show that the rates of geometric convergence of the standard deviation and the distance to the optimum must be the same. Markov chain analysis [7], for example, shows that the ratio between the distance to the optimum and the standard deviation admits a stationary distribution, meaning that the distance to the optimum and the standard deviation converges at the same rate. However, it is very complicated and difficult to derive the influence of the strategy parameters such as the sample size and the learning rates on the rate of convergence and the stationary distribution of the ratio explicitly. Quality gain or progress rate analysis, which considers a single step expected behavior, provides the optimal step-size given a mean vector in an explicit form and shows the optimal recombination weights in the evolution strategy in the limit of search space dimension to infinity [3, 6]. The analysis methodology most related to this paper is probably the so-called dynamical systems approach [13]. Roughly speaking, it describes the algorithm dynamics by a lower number of parameters and studies the steady-state behavior. Nonetheless, these analysis has not been applied to the IGO algorithms.

In this paper, we analyze the IGO algorithm given the family of isotropic Gaussian distributions solving the sphere function. Differently from the previous theoretical studies of the IGO algorithms where the infinite samples are assumed, we consider the expected dynamics of the IGO algorithm with finite samples. Taking into account the variance and covariance of the estimated natural gradient under finite samples, we show that the rate of the expected decrease of the squared distance to the optimum depends on the ratio of the squared distance from the distribution mean to the optimum and the distribution variance. This implies that the ratio needs to be stable for the algorithm to converge geometrically, which is empirically observed but is not derived in the previous analysis assuming the infinite samples. Then, we derive the stable value of the ratio and the conditions on the learning rates such that the ratio admits a stable value. The simulation results show that the theoretical results derived in this paper well describe the behavior of the real algorithm.

The rest of this paper is organized as follows. In Section 2, we describe the IGO framework. In Section 3, we investigate the convergence properties under infinite samples as well as under finite samples on the sphere function. In Section 4, we study the condition on the learning rates that the expected decrease of the squared distance from the distribution mean to the optimum and the expected decrease of the distribution variance agree. In Section 5, the derived results are compared to the behavior of the real IGO algorithm in simulation, and we confirm that they match precisely. Finally, we discuss the simulation results and the future works in Section 6 and conclude this paper in Section 7.

2 INFORMATION GEOMETRIC OPTIMIZATION

Information geometric optimization (IGO) is a unified framework of stochastic algorithms for black-box optimization on an arbitrary search space X. We assume an objective function $f : X \to \mathbb{R}$ to be minimized on the search space X. Given a family of probability distributions $\{P_\theta\}$ on X, IGO transforms the original minimization problem into the maximization of the expected value of a utility function u over the parameter space of $\{P_\theta\}$ as

$$\arg\max_\theta \mathbb{E}_{P_\theta}[u(x)] \ . \tag{1}$$

The utility function u is a strictly decreasing transformation of f, that means $u(x) > u(x')$ for all $f(x) < f(x')$.

IGO updates the parameter of the probability distribution along the natural gradient direction which is the steepest direction with respect to the Fisher metric [5]. The Fisher metric is a Riemannian metric defined by the Fisher information matrix. Let p_θ be the probability density function induced by P_θ, the Fisher information matrix is defined as

$$F(\theta) = \int \left(\frac{\partial \ln p_\theta(x)}{\partial \theta}\right)\left(\frac{\partial \ln p_\theta(x)}{\partial \theta}\right)^{\mathrm{T}} p_\theta(x)\mathrm{d}x \ . \tag{2}$$

The natural gradient is given by the product of the inverse of the Fisher information matrix F and the vanilla gradient, derived as

$$\tilde{\nabla}\mathbb{E}_{P_\theta}[u(x)] = F^{-1}(\theta)\nabla\mathbb{E}_{P_\theta}[u(x)] \tag{3}$$

$$= \int u(x)(\tilde{\nabla}\ln p_\theta(x))p_\theta(x)\mathrm{d}x \tag{4}$$

where $\tilde{\nabla}\ln p_\theta(x) = F^{-1}(\theta)\nabla\ln p_\theta(x)$. However, the above integral cannot be performed analytically since f is black-box in our setting. Instead, we take the Monte Carlo estimate using n i.i.d. samples drawn from probability distribution P_θ, namely,

$$\tilde{\nabla}\mathbb{E}_{P_\theta}[u(x)] \approx \frac{1}{n}\sum_{i=1}^{n} u(x_i)\tilde{\nabla}\ln p_\theta(x_i) \ . \tag{5}$$

Introducing the learning rate η^t at iteration t, the update rule of IGO reads

$$\theta^{t+1} = \theta^t + \frac{\eta^t}{n}\sum_{i=1}^{n} u(x_i)\tilde{\nabla}\ln p_\theta(x_i) \ . \tag{6}$$

In this paper, we analyze the IGO algorithm given the family of isotropic Gaussian distributions on $X = \mathbb{R}^d$. The isotropic Gaussian distribution, $\mathcal{N}(m, \beta)$, is a multivariate Gaussian distribution where the covariance matrix is the identity times a scalar. We parametrize it by its mean vector $m \in \mathbb{R}^d$ and the variance parameter $\beta \in \mathbb{R}_{>0}$, i.e., $\theta = (m^{\mathrm{T}}, \beta)^{\mathrm{T}}$. The probability density function is written as

$$p_\theta(x) = \frac{1}{(\sqrt{2\pi\beta})^d}\exp\left(-\frac{\|x-m\|^2}{2\beta}\right) \ . \tag{7}$$

In the analysis we set the utility function as the negative of the objective function, $u = -f$, which is also used in previous theoretical studies [1, 9]. The estimated negative natural gradient of each parameter at iteration t is written as

$$\tilde{\nabla}_m\mathbb{E}_{P_\theta}[f(x)] \approx \widehat{\delta m^t} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)(x_i - m^t) \tag{8}$$

$$\tilde{\nabla}_\beta\mathbb{E}_{P_\theta}[f(x)] \approx \widehat{\delta\beta^t} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)\left(\frac{\|x_i - m^t\|^2}{d} - \beta^t\right) \tag{9}$$

and the update rules read

$$m^{t+1} = m^t - \eta_m^t\widehat{\delta m^t} \tag{10}$$

$$\beta^{t+1} = \beta^t - \eta_\beta^t\widehat{\delta\beta^t}, \tag{11}$$

where $\eta_m^t, \eta_\beta^t \in \mathbb{R}_{>0}$ are the learning rates.

3 CONVERGENCE ANALYSIS

We investigate the convergence properties of the IGO algorithm (10) and (11) on a convex quadratic function $f(x) = (x-x^*)^{\mathrm{T}} A(x-x^*)$, where A is a positive definite symmetric matrix and x^* is global optimum solution, with a special focus on the sphere function $f(x) = \|x-x^*\|^2$. Under the isotropic Gaussian distribution, the expected objective function value is

$$\mathbb{E}_{P_\theta}[f(x)] = \beta \mathrm{Tr}(A) + (m-x^*)^{\mathrm{T}} A (m-x^*) \,. \tag{12}$$

Therefore, the optimal parameter is $\theta^* = ({x^*}^{\mathrm{T}}, 0)^{\mathrm{T}}$ and we are going to investigate how θ^t approaches the optimal parameter. Since the following result is invariant under any translation, we assume without loss of generality (w.l.g.) that the optimum of the objective function f is located at $x^* = \mathbf{0}$.

3.1 Analysis of Infinite Sample Model

The Monte Carlo estimates of the component-wise natural gradients, (8) and (9), are both unbiased and consistent. Therefore, in the limit of the sample size n to infinity leads the estimates (8) and (9) to converge to

$$\tilde{\nabla}_m \mathbb{E}_{P_\theta}[f(x)] = 2\beta^t A m^t \tag{13}$$

$$\tilde{\nabla}_\beta \mathbb{E}_{P_\theta}[f(x)] = 2(\beta^t)^2 \frac{\mathrm{Tr}(A)}{d} \,, \tag{14}$$

respectively. In this section, we set the learning rates as

$$\eta_m^t = \frac{c_m}{2\lambda_1(A)\beta^t} \,, \tag{15}$$

$$\eta_\beta^t = \frac{c_\beta}{2\lambda_1(A)\beta^t} \,, \tag{16}$$

where c_m and c_β are constant values and $\lambda_k(A)$ is the k-th largest eigenvalue of A. This setting serves as the normalization of the objective function values by $2\lambda_1(A)\beta^t$. In other words, our utility is $u(x) = -f(x)/(2\lambda_1(A)\beta^t)$ instead of $u(x) = -f(x)$ and the learning rates are c_m and c_β instead of η_m^t and η_β^t, respectively. Similar settings are found in [1, 9].

The following proposition proves the geometric convergence of each parameter with the rate of convergence depending on the learning rate.

Proposition 3.1. *Consider the deterministic sequence of m^t and β^t defined by* (10) *and* (11) *with $\widehat{\delta m^t}$ and $\widehat{\delta\beta^t}$ replaced by* (13) *and* (14)*, respectively. Then, for $0 < c_m < 2$ and $0 < c_\beta < 1$, both $\|m^t\|^2$ and β^t converge towards zero with the rates of convergence*

$$\min_k \left(1 - c_m \frac{\lambda_k(A)}{\lambda_1(A)}\right)^2 \le \frac{\|m^{t+1}\|^2}{\|m^t\|^2} \le \max\left\{(1-c_m)^2, (1-c_m \mathrm{Cond}^{-1}(A))^2\right\} \tag{17}$$

and

$$\frac{\beta^{t+1}}{\beta^t} = 1 - c_\beta \frac{\mathrm{Tr}(A)/d}{\lambda_1(A)} \,, \tag{18}$$

where $\mathrm{Cond}(A) = \lambda_1(A)/\lambda_d(A)$ is the condition number of A.

Proof. The convergence rate of the variance parameter β^t is as follows:

$$\frac{\beta^{t+1}}{\beta^t} = \frac{\beta^t - \eta_\beta^t \tilde{\nabla}_\beta \mathbb{E}_{P_\theta}[f(x)]}{\beta^t} = 1 - c_\beta \frac{\mathrm{Tr}(A)/d}{\lambda_1(A)} \,. \tag{19}$$

Since $\lambda_1(A) \ge \mathrm{Tr}(A)/d$, the variance parameter always converges when $0 < c_\beta < 1$ is satisfied.

The square norm of the updated mean vector is

$$\|m^{t+1}\|^2 = \left\|m^t - 2\eta_m^t \beta^t A m^t\right\|^2 \tag{20}$$

$$= (m^t)^{\mathrm{T}} \left(\mathrm{I}_d - \frac{c_m}{\lambda_1(A)} A\right)^2 m^t \,. \tag{21}$$

The eigenvalues of the matrix inside the parentheses are upper bounded by the larger of $1 - c_m$ and $1 - c_m/\mathrm{Cond}(A)$, which leads to

$$\frac{\|m^{t+1}\|^2}{\|m^t\|^2} \le \max\left\{(1-c_m)^2, (1-c_m \mathrm{Cond}^{-1}(A))^2\right\} \,. \tag{22}$$

Note that the right-hand side is strictly smaller than 1, implying geometric convergence. Moreover, the lower bound is

$$\frac{\|m^{t+1}\|^2}{\|m^t\|^2} \ge \lambda_d \left(\left(\mathrm{I}_d - \frac{c_m}{\lambda_1(A)} A\right)^2\right) = \min_k \left(1 - c_m \frac{\lambda_k(A)}{\lambda_1(A)}\right)^2 \,. \tag{23}$$

This ends the proof. □

Proposition 3.1 implies that both c_m and c_β must be strictly greater than zero for geometric convergence. In other words, η_m^t and η_β^t must be in $\Theta(1/\beta^t)$. Therefore, the scaling of the learning rates introduced in (15) and (16) are necessary for geometric convergence of the IGO algorithm. This is explained as follows. The function value $f(x)$ is proportional to β^t (given that $\|m^t\|^2 \propto \beta^t$) and the length of the natural gradient scales down as $f(x)$ does if $u = -f$. The $\Theta(1/\beta^t)$ learning rates are necessary to prevent the natural gradient from scaling down depending on the function value scale. Utility transformations such as truncation selection keep the utility values constant and itself serves as the normalization of the natural gradient [9]. Therefore, we observe geometric convergence of IGO algorithms with truncation selection utility and constant η_m^t and η_β^t in practice.

If we consider the case $A = \mathrm{I}_d$, i.e., the sphere function $f(x) = \|x\|^2$, then the rate of convergence of each parameter reads

$$\frac{\|m^{t+1}\|^2}{\|m^t\|^2} = (1-c_m)^2 \tag{24}$$

$$\frac{\beta^{t+1}}{\beta^t} = 1 - c_\beta \,. \tag{25}$$

This result corresponds what is derived in [1], where an arbitrary symmetric positive definite covariance matrix is considered instead of β. It says that we can control the convergence rates of m^t and β^t independently by setting different values for c_m and c_β. However, this is not true for a real algorithm with a finite sample size, as we will observe in simulation in Section 5. More precisely, $\|m\|^2/\beta$ must be stable, which is not only observed in empirical studies but also implied by theoretical results such as Markov chain analysis [7] and progress rate and quality gain analysis [3, 4, 6]. This is not achieved in the analysis of the infinite sample size model of the IGO algorithm.

3.2 Estimation Variance of Monte Carlo

In the infinite sample model, the estimation variance of the natural gradient ((8) and (9)) due to a finite sample size, n, is not taken into account. This is considered one of the reasons that we fail to obtain a result that describes the behavior of a real algorithm. The following lemma shows the covariance matrix and the variance of the components of the natural gradient estimates, which are proportional to $1/n$.

Lemma 3.2. *The covariance and the variance of the components of estimated negative natural gradient,* (8) *and* (9)*, on a convex quadratic function $f(x) = x^{\mathrm{T}}Ax$ are*

$$\mathrm{Cov}[\widehat{\delta m^t}] = \frac{1}{n} V_m(m^t, \beta^t, A) \tag{26}$$

$$\mathrm{Var}[\widehat{\delta \beta^t}] = \frac{1}{nd^2} V_\beta(m^t, \beta^t, A) \ , \tag{27}$$

where V_m and V_β are defined as

$$\begin{aligned} V_m(m, \beta, A) =& 4\beta^3 \mathrm{Tr}(A)A + 8\beta^3 A^2 \\ &+ \beta^3 \left(2\mathrm{Tr}(A^2) + \mathrm{Tr}^2(A)\right) \mathrm{I}_d \\ &+ 2\beta^2 (\mathrm{Tr}(A)\,\mathrm{I}_d + 2A) m^{\mathrm{T}} A m \\ &+ 4\beta^2 \left(\|Am\|^2 \,\mathrm{I}_d + (Am)(Am)^{\mathrm{T}}\right) \\ &+ \beta^t \,\mathrm{I}_d (m^{\mathrm{T}} A m)^2 \end{aligned} \tag{28}$$

$$\begin{aligned} V_\beta(m, \beta, A) =& 2\beta^4 (d+12) \left(\mathrm{Tr}^2(A) + 2\mathrm{Tr}(A^2)\right) \\ &+ 4\beta^3 (d+4) \left(\mathrm{Tr}(A) m^{\mathrm{T}} A m + 2\|Am\|^2\right) \\ &+ 2\beta^2 d (m^{\mathrm{T}} A m)^2 \\ &- 4\beta^4 \mathrm{Tr}^2(A) \end{aligned} \tag{29}$$

3.3 Analysis of Finite Sample Algorithm

In the following, we focus on the case $A = \mathrm{I}_d$. This means the objective function is the sphere function $f(x) = \|x\|^2$. Moreover, we define $c_m, c_\beta \in \mathbb{R}_{>0}$ similarly to Section 3.1,

$$\eta_m^t = \frac{c_m}{2\beta^t} \tag{30}$$

$$\eta_\beta^t = \frac{c_\beta}{2\beta^t} \ . \tag{31}$$

Then, we consider the expected progress of $\|m^{t+1}\|^2$ and β^{t+1} given m^t and β^t.

First, we consider the expected dynamics of the variance parameter β^t. The expected value of β^{t+1} is simply

$$\mathbb{E}[\beta^{t+1}] = \beta^t (1 - c_\beta) \ , \tag{32}$$

and the rate of decrease of the variance parameter is given in the next theorem.

Theorem 3.3. *Consider the single update of β in* (11) *given m^t and β^t on the sphere function. The expected decrease of β is given by*

$$\frac{\mathbb{E}[\beta^{t+1}]}{\beta^t} = 1 - c_\beta \ . \tag{33}$$

Next, we consider the expected dynamics of the mean vector $\|m^t\|^2$. Note that the expected value of $\|m^t\|$ decreases when the expected value of $\|m^t\|^2$ decreases. When updating the mean vector along (10), the expected value of $\|m^{t+1}\|^2$ is written as

$$\begin{aligned} &\mathbb{E}[\|m^{t+1}\|^2] = \\ &\quad \|m^t\|^2 - 2\eta_m^t (m^t)^{\mathrm{T}} \mathbb{E}\left[\widehat{\delta m^t}\right] + (\eta_m^t)^2 \mathbb{E}\left[\|\widehat{\delta m^t}\|^2\right] \ . \end{aligned} \tag{34}$$

The following theorem shows the rate of decrease of the expected squared distance between the mean vector to the optimum.

Theorem 3.4. *Consider the single update of m in* (10) *given m^t and β^t on the sphere function. The expected decrease of $\|m\|^2$ is given by*

$$\frac{\mathbb{E}[\|m^{t+1}\|^2]}{\|m^t\|^2} = (1 - c_m)^2 + \frac{c_m^2}{4n} K \ , \tag{35}$$

where

$$K = (d+4)(d+2)d \frac{\beta^t}{\|m^t\|^2} + 2(d^2 + 4d + 2) + d \frac{\|m^t\|^2}{\beta^t} \ . \tag{36}$$

Note the decreasing rate depends on the ratio $\|m^t\|^2/\beta^t$.

The right-hand side of (35) *is less than* 1 *if and only if the following two conditions hold:*

- *Condition on ratio $\|m^t\|^2/\beta^t$*

$$\begin{aligned} &\frac{1}{d}\left(M(d,n) - \sqrt{M^2(d,n) - d^2(d+2)(d+4)}\right) \\ &\quad < \frac{\|m^t\|^2}{\beta^t} < \frac{1}{d}\left(M(d,n) + \sqrt{M^2(d,n) - d^2(d+2)(d+4)}\right) \ , \end{aligned} \tag{37}$$

  where

$$M(d,n) = 2n\left(\frac{2}{c_m} - 1\right) - (d^2 + 4d + 2) \ . \tag{38}$$

- *Condition on learning rate coefficient c_m (necessary and sufficient condition for* (37) *to have a nonzero interval)*

$$0 < c_m < \frac{4n}{s(d,n)} \ , \tag{39}$$

  where

$$s(d,n) = d^2 + 4d + 2 + d\sqrt{(d+2)(d+4)} + 2n \ . \tag{40}$$

Proof. On the sphere function, V_m in Lemma 3.2 is

$$\begin{aligned} V_m(m^t, \beta^t, \mathrm{I}_d) =& (\beta^t)^3 (d+4)(d+2)\,\mathrm{I}_d \\ &+ 2(\beta^t)^2 \left(\|m^t\|^2 (d+4)\,\mathrm{I}_d + 2m^t (m^t)^{\mathrm{T}}\right) \\ &+ \beta^t \|m^t\|^4 \,\mathrm{I}_d \end{aligned} \tag{41}$$

and the last term in (34) can be derived from (26) as

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\delta m^t}\|^2\right] &= \mathrm{Tr}\left(\mathrm{Cov}[\widehat{\delta m^t}] + \mathbb{E}[\widehat{\delta m^t}]\mathbb{E}[\widehat{\delta m^t}]^{\mathrm{T}}\right) \\ &= \mathrm{Tr}\left(\mathrm{Cov}[\widehat{\delta m^t}]\right) + \left\|\mathbb{E}[\widehat{\delta m^t}]\right\|^2 \\ &= \frac{K}{n}(\beta^t)^2 \|m^t\|^2 + 4(\beta^t)^2 \|m^t\|^2 \ . \end{aligned} \tag{42}$$

The right-hand side of (35) is derived from (34) and the above equations.

Considering the inequality that the right-hand side of (35) is less than 1 as a second order inequality of the ratio $\|m^t\|^2/\beta^t$, we can

verify that (37) is the necessary and sufficient condition for the inequality to be satisfied.

Finally we consider the condition for the interval (37) to be nonempty. It is clear that the interval is nonempty if

$$M^2(d,n) - d^2(d+4)(d+2) \leq 0 \ . \tag{43}$$

It is satisfied if one of the following holds:

$$M(d,n) \leq -d\sqrt{(d+2)(d+4)} \tag{44}$$

$$M(d,n) \geq d\sqrt{(d+2)(d+4)} \ . \tag{45}$$

Since the ratio $\|m^t\|^2/\beta^t$ is nonnegative, so is $M(d,n)$. Therefore, only (45) can be the necessary and sufficient condition for (37) to be nonempty, and we derive the upper bound of (39) from (45). This ends the proof. □

Theorems 3.3 and 3.4 imply that the rates of geometric convergence of $\|m^t\|^2$ and β^t towards zero must be the same, which is what we observe in practice and what we cannot derive from the analysis of the infinite sample model of the IGO algorithm. To see this, we assume the rates of geometric convergence of $\|m^t\|^2$ and β^t are different. Then, it is easy to see that ratio $\|m^t\|^2/\beta^t$ tends to either ∞ or 0, both of which lead to divergence of K. Theorem 3.4 then tells that the expected change rate of the squared distance of the mean vector will be greater than 1, meaning that it fails to converge. To prevent the divergence, one needs to decrease c_m towards zero as K increases, such that

$$0 < c_m < \frac{8n}{4n+K} \ , \tag{46}$$

which is necessary for (35) to be smaller than 1. However, it leads to the rate of convergence to 1, i.e., it fails to converge geometrically. By taking into account a finite sample size, we manage to derive such an important criterion.

4 STABLE VALUE OF THE RATIO

In this section, we study the condition on the learning rate coefficients c_m and c_β for the rates of expected progress of m^t and β^t to be the same on the sphere function. Moreover, we derive the stable value of $\|m^t\|^2/\beta^t$.

Theorem 4.1. *Consider the single updates of m and β in* (10) *and* (11). *We have*

$$\frac{\mathbb{E}[\|m^{t+1}\|^2]}{\|m^t\|^2} = \frac{\mathbb{E}[\beta^{t+1}]}{\beta^t} = 1 - c_\beta < 1 \tag{47}$$

if and only if the following three conditions hold

$$0 < c_\beta \leq \frac{2n}{s(d,n)} \tag{48}$$

$$\frac{2n - \sqrt{4n^2 - 2nc_\beta s(d,n)}}{s(d,n)} \leq c_m \leq \frac{2n + \sqrt{4n^2 - 2nc_\beta s(d,n)}}{s(d,n)} \tag{49}$$

$$\frac{\|m^t\|^2}{\beta^t} = \frac{1}{d}\left(M(d,n) - 2n\frac{c_\beta}{c_m^2} \pm \sqrt{\left(M(d,n) - 2n\frac{c_\beta}{c_m^2}\right)^2 - d^2(d+2)(d+4)}\right) \ , \tag{50}$$

where s and M are defined in Theorem 3.4.

Proof. Since the ratio $\|m^t\|^2/\beta^t$ is nonnegative, we can transform (47) into quadratic equation

$$d\left(\frac{\|m^t\|^2}{\beta^t}\right)^2 - 2\left(M(d,n) - 2n\frac{c_\beta}{c_m^2}\right)\frac{\|m^t\|^2}{\beta^t} + d(d+2)(d+4) = 0 \, , \tag{51}$$

and the discriminant of the equation is

$$D = \left(M(d,n) - 2n\frac{c_\beta}{c_m^2}\right)^2 - d^2(d+2)(d+4) \, . \tag{52}$$

Condition $D \geq 0$ is satisfied if and only if one of the followings holds

$$M(d,n) - 2n\frac{c_\beta}{c_m^2} \leq -d\sqrt{(d+2)(d+4)} \tag{53}$$

$$M(d,n) - 2n\frac{c_\beta}{c_m^2} \geq d\sqrt{(d+2)(d+4)} \ . \tag{54}$$

When either (53) or (54) is satisfied, the solutions to the above quadratic equation are given by (50). Note that both solutions in (50) become negative if the left-hand side of (53) or (54) is negative. Since the ratio $\|m^t\|^2/\beta^t$ must be nonnegative but (53) does not accept a positive solution, (54) is the necessary and sufficient condition for (50). Conditions (48) and (49) derive from (54). This ends the proof. □

In (50) we have two candidate values of a stationary point of the ratio $\|m^t\|^2/\beta^t$. Let R_{small} and R_{large} denote the smaller and greater values in (50), respectively. We expect that only R_{small} can be the stable stationary point of $\|m^t\|^2/\beta^t$. The reason is described as follows. If $\|m^t\|^2/\beta^t \in (R_{\text{small}}, R_{\text{large}})$, we have

$$\frac{\mathbb{E}[\|m^{t+1}\|^2]}{\|m^t\|^2} < \frac{\mathbb{E}[\beta^{t+1}]}{\beta^t} \ , \tag{55}$$

hence the ratio will become smaller. If not, the ratio will become greater. This implies that R_{small} is attractive for $\|m\|^2/\beta < R_{\text{large}}$, and $\|m\|^2/\beta$ tends to diverge if $\|m\|^2/\beta > R_{\text{large}}$.

If we consider the case when c_m, c_β violate the condition (48) and (49), we always have

$$\frac{\mathbb{E}[\beta^{t+1}]}{\beta^t} = 1 - c_\beta < (1-c_m)^2 + \frac{c_m^2}{4n}K_{\min} \leq \frac{\mathbb{E}[\|m^{t+1}\|^2]}{\|m^t\|^2} \ , \tag{56}$$

where $K_{\min}$ is the minimal possible K in (36) given by

$$K_{\min} = \min_{\|m\|^2/\beta} K = 2\left((d^2+4d+2) + d\sqrt{(d+2)(d+4)}\right) \ . \tag{57}$$

This means the ratio is expected to diverge when c_m, c_β violate (48) and (49).

When the decreasing rates of $\|m^t\|^2$ and β^t satisfy (47), the decreasing rate of the expected objective function value is given by

$$\frac{\mathbb{E}[f(x)|m = m^{t+1}, \beta = \beta^{t+1}]}{\mathbb{E}[f(x)|m = m^t, \beta = \beta^t]} = 1 - c_\beta \ . \tag{58}$$

It is intuitive that it is the same as those of $\|m\|^2$ and β, however, it cannot be derived by the analysis with the infinite samples.

Note that the interval of c_m in (49) depends on c_β. In particular, the interval becomes shorter around $\frac{2n}{s(d,n)}$ as c_β is set greater. Both intervals of c_m and c_β depend heavily on the sample size n

and the dimension d of the search space. If we take the limit for the sample size to infinity, we obtain

$$\lim_{n\to\infty} \frac{2n - \sqrt{4n^2 - 2nc_\beta s(d,n)}}{s(d,n)} = 1 - \sqrt{1-c_\beta} \tag{59}$$

$$\lim_{n\to\infty} \frac{2n + \sqrt{4n^2 - 2nc_\beta s(d,n)}}{s(d,n)} = 1 + \sqrt{1-c_\beta} \ . \tag{60}$$

Therefore, (48) and (49) read

$$0 < c_\beta \leq 1 \tag{61}$$

$$1 - \sqrt{1-c_\beta} \leq c_m \leq 1 + \sqrt{1-c_\beta} \ . \tag{62}$$

These are different from the conditions given in Proposition 3.1, where the infinite sample size is assumed from the beginning of the derivation. Moreover, given c_m and c_β,

$$\lim_{n\to\infty} \frac{R_{\text{large}}}{n} = \frac{4(2c_m - c_m^2 - c_\beta)}{c_m^2 d} \tag{63}$$

$$\lim_{n\to\infty} nR_{\text{small}} = \frac{c_m^2 d(d+2)(d+4)}{4(2c_m - c_m^2 - c_\beta)} \ . \tag{64}$$

These limit values imply that, with Landau's O-notation, the stable ratio R_{small} of $\|m\|^2/\beta$ becomes smaller in $O(1/n)$ and its region of attraction, $[0, R_{\text{large}}]$, becomes wider as $R_{\text{large}} \in O(n)$.

When we consider the case $d \gg 1$, we can easily see that the intervals (48) and (49) tend to zero. In other words, the learning rates satisfying these conditions need to be infinitesimally small if $d \gg 1$ while n is fixed. Now we introduce α_m, α_β as

$$c_m = \alpha_m \frac{2n}{s(d,n)} \ , \qquad c_\beta = \alpha_\beta \frac{2n}{s(d,n)} \ . \tag{65}$$

The variables α_m, α_β represent the proportion of c_m, c_β to the value $2n/s(d,n)$, which is the maximum value in (48) and the center value in (49). Note that when the dimension d is sufficiently large, $2n/s(d,n)$ is approximated by n/d^2. The ranges of c_m and c_β satisfying (48) and (49) in Theorem 4.1 reads

$$0 < \alpha_\beta \leq 1 \tag{66}$$

$$\left(1 - \sqrt{1-\alpha_\beta}\right) \leq \alpha_m \leq \left(1 + \sqrt{1-\alpha_\beta}\right) \ , \tag{67}$$

and the limit values of R_{large} and R_{small} divided by d are derived as

$$\lim_{d\to\infty} \frac{R_{\text{large}}}{d} = \left(\frac{4}{\alpha_m} - \frac{2\alpha_\beta}{\alpha_m^2} - 1\right) + \sqrt{\left(\frac{4}{\alpha_m} - \frac{2\alpha_\beta}{\alpha_m^2} - 1\right)^2 - 1} \ , \tag{68}$$

$$\lim_{d\to\infty} \frac{R_{\text{small}}}{d} = \left(\frac{4}{\alpha_m} - \frac{2\alpha_\beta}{\alpha_m^2} - 1\right) - \sqrt{\left(\frac{4}{\alpha_m} - \frac{2\alpha_\beta}{\alpha_m^2} - 1\right)^2 - 1} \ . \tag{69}$$

It implies that the stable value of the ratio $\|m\|^2/\beta$ scales as $O(d)$. This agrees with our empirical knowledge and the theoretical fact that the optimal step-size on the sphere function is proportional to the distance to the optimum times d [3, 6].

5 NUMERICAL SIMULATION

The analysis in the previous sections studies the expected single step behavior and it does not guarantee the long-term behavior itself. In this section, we perform the IGO algorithm given the isotropic Gaussian distribution on the sphere function, and see how precisely the results derived in the previous sections describe the behavior of the long-term behavior of the real algorithm. Particularly, we observe that the rate of convergence is approximated by (47) and the ratio $\|m^t\|^2/\beta^t$ tends to the value derived in Theorem 4.1 under the conditions on the learning rate coefficients derived there. We also check what happens if the conditions on the learning rate coefficients are not satisfied.

5.1 Experiment Setting

We run the IGO algorithm described by (10) and (11) with the natural gradients (8) and (9) estimated from the finite samples on the sphere function with dimension $d = 10$. The sample size is $n = 10$ in this simulation. The learning rates are set as (30) and (31) with different c_m and c_β values. The initial variance parameter is fixed to $\beta^0 = 1$ and the distance of the mean vector to the origin (optimum) is set in different values as $\|m^0\|^2 = 10^{-1}, 10^0, 10^1, 10^2$.

We remark that the domain of β^t is a positive real space $\mathbb{R}_{>0}$, but the update (11) with (9) sometimes leads to a negative value, especially when c_β is relatively large. To prevent β^t from being negative, we replace (11) by

$$\beta^{t+1} = \left|\beta^t - \eta_\beta^t \widehat{\delta\beta^t}\right| \ . \tag{70}$$

The artificial effect of this modification is discussed later.

5.2 Simulation Results

First, we set $c_m = 0.1$ and $c_\beta = 0.01$, which satisfy the conditions (48) and (49). Figure 1 shows the median values and 25th and 75th percentile ranges of the transitions of the ratio $\|m^t\|^2/\beta^t$, the minimum function value at each iteration, $\|m^t\|^2$ and β^t in 1000 trials for each setting. We plot them on first 100 iterations, which is sufficient for the ratio to reach the stable state. In Figure 1 we observe that the rates of geometric convergence of $\|m^t\|^2$ and β^t and the decreasing rate of minimum evaluation value were well approximated by $1 - c_\beta$ if the initial ratio is smaller than R_{large}, and observe that the ratio, minimum evaluation value, $\|m^t\|^2$ and β^t all diverged if not. Moreover, the ratio $\|m^t\|^2/\beta^t$ tended to R_{small} in (50) if the initial ratio is smaller than R_{large}. The simulation results matched the theoretical results presented in the previous section and the discussion.

Next, we investigate the setting $c_m = c_\beta = 0.1$, which do not satisfy (49). The result is shown in Figure 2. In this setting, $\|m\|^2$ did not increase when the ratio $\|m\|^2/\beta$ was satisfying (37) and it started diverging once (37) was violated. This behavior is explained by Theorem 3.4, however, the variance parameter β also diverged unlike the result of Theorem 3.3. Moreover, the ratio $\|m\|^2/\beta$ fluctuates around some value. The fluctuation is due to unstable update of β. This may be because of the modification (70): too large update of β in a negative direction sometimes results in increasing β. This is discussed in the next section.

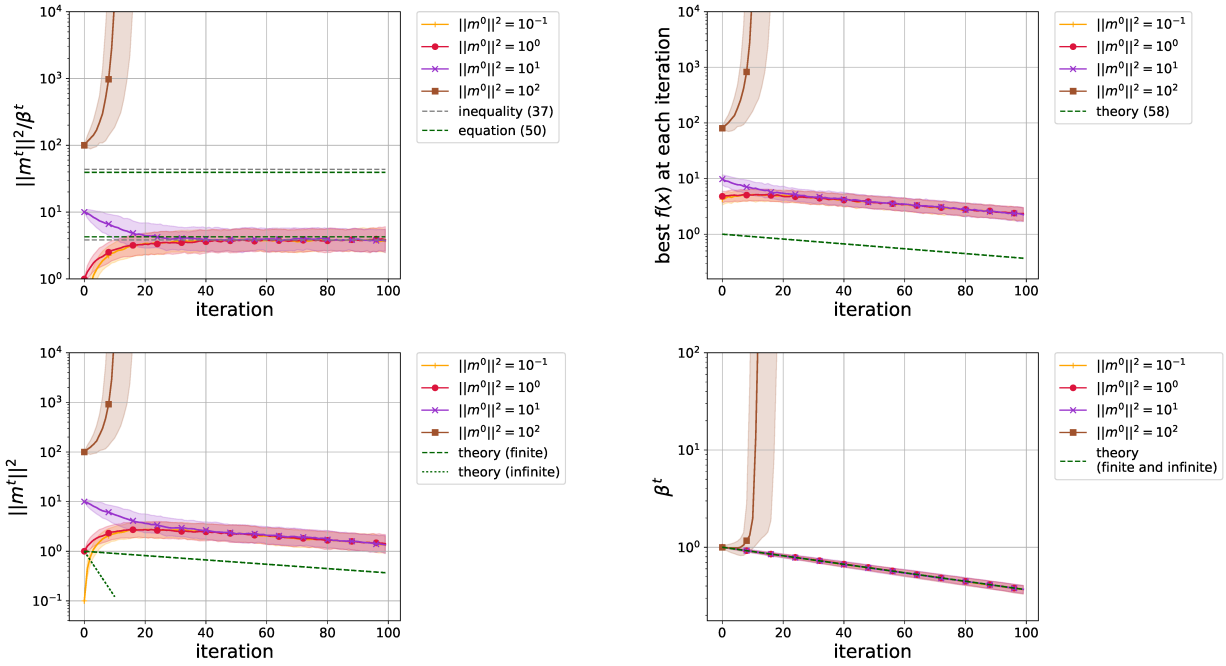

Figure 1: The transitions of each value when c_m and c_β satisfy the conditions (48), (49). These figures show the transitions of the ratio $\|m^t\|^2/\beta^t$, the minimum evaluation, the mean vector $\|m^t\|^2$ and the variance parameter β^t. We plot the median values and 25th and 75th percentile ranges of 1000 trials in each setting.

6 FURTHER DISCUSSION

First, we discuss the effect of the modified update rule (70). As we mentioned, the ratio $\|m\|^2/\beta$ theoretically continues to be greater the whole time if it does not satisfy the stability condition (49). However, we did not observe such a behavior. The reason may be described as follows. The variance of the β-update on the sphere function is

$$\mathrm{Var}\left[\frac{\beta^{t+1}}{\beta^t}\right] = \frac{c_\beta^2}{2nd^2}\left(d(d^2+12d+24) + 2(d+4)(d+2)\frac{\|m^t\|^2}{\beta^t} + d\left(\frac{\|m^t\|^2}{\beta^t}\right)^2\right), \quad (71)$$

which increases to the plus infinity as the ratio $\|m\|^2/\beta$ increases. Once the ratio becomes large, there is high chance that the variance parameter is updated in the negative direction and the modified update rule (70) results in increasing the variance parameter. Therefore, the variance parameter β increases and the ratio $\|m\|^2/\beta$ fluctuates in the case of Figure 2.

Remark that we employed the $-f$-proportional utility in our analysis as it has been employed in the previous works. In practice, however, we employ a ranking-based utility. Under a proper utility function and a proper learning rate, the variance parameter is guaranteed to be positive. Moreover, in the CMA-ES, the covariance matrix is guaranteed to be positive definite. When using a ranking-based utility such as the truncation selection [10], we observed that the ratio tends to be a stable value when the learning rates are reasonable, and that it tends to diverge if the learning rates are improper. These behaviors are similar to the simulation results described in Section 5, and the analysis with a ranking-based utility makes our understanding even better. It is an important direction of a future work.

Next, we discuss the optimal sample size n^*. The optimal sample size is defined as

$$n^* = \arg\min_{n\geq 1} \sqrt{n\frac{\mathbb{E}_n[\|m^{t+1}\|^2]}{\|m^t\|^2}}, \quad (72)$$

where $\mathbb{E}_n[\|m^{t+1}\|^2]$ is the expected value of $\|m^{t+1}\|^2$ under the sample size n. Note that the optimal value depends on c_m and c_β. If we consider the case where the learning rates are set to the maximal values that admits a stable value of the ratio $\|m\|^2/\beta$, derived in Theorem 4.1, we have $n^* = 1$. It is natural from the stochastic approximation perspective: smaller samples with a smaller learning rate tend to be faster than larger samples with a greater learning rate because of a small but frequent update of the parameter. Note, however, that it is not necessarily optimal on other functions. Moreover, it is not possible to set $n = 1$ when a ranking-based utility u is employed.

7 CONCLUSION

We have studied the convergence properties of the IGO algorithm given the family of isotropic Gaussian distributions under finite samples on the sphere function. We have derived the variance and covariance of estimated natural gradients and derived the rate of

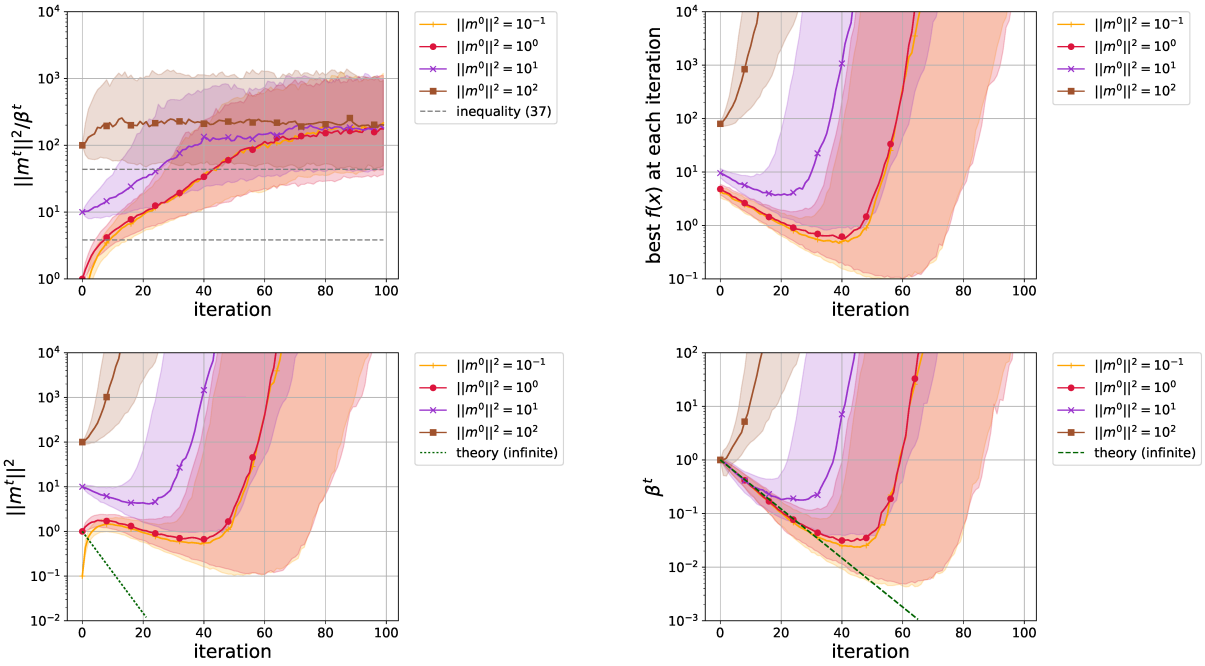

Figure 2: The transitions of each value when c_m and c_β do not satisfy the condition (49). These figures show the transitions of the ratio $\|m^t\|^2/\beta^t$, the minimum evaluation, the mean vector $\|m^t\|^2$ and the variance parameter β^t.

decrease of the expected squared distance to the optimum and the expectation of the variance parameter. We have shown that the rate of decrease of the expected squared distance to the optimum depends on the ratio $\|m^t\|^2/\beta^t$, which is empirically observed but is not derived in the previous analysis assuming the infinite samples. We have derived the stable value of the ratio $\|m^t\|^2/\beta^t$ and the conditions on the learning rates that admit a stable value. The simulation results show that the ratio $\|m^t\|^2/\beta^t$ tends to the theoretically derived stable value, and the rate of decrease of the expected squared distance to the optimum and the expectation of the variance parameter is well approximated by the theoretically derived value.

We consider our analysis can be expanded to more general algorithms such as the CMA-ES. A generalization of our analysis to the IGO with a family of Gaussian distribution with arbitrary positive definite covariance matrix is an ongoing future work. Moreover, we would like to investigate the convergence of the covariance matrix to the inverse Hessian of the objective function times a scalar factor, which is derived with its rate of convergence under the infinite samples [1]. This part is also left for the future work.

REFERENCES

[1] Youhei Akimoto. 2012. Analysis of a Natural Gradient Algorithm on Monotonic Convex-Quadratic-Composite Functions. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO) 2012*. 1293–1300.

[2] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2012. Convergence of the Continuous Time Trajectories of Isotropic Evolution Strategies on Monotonic C^2-composite Functions. In *Parallel Problem Solving from Nature - PPSN XII, Part I (Lecture Notes in Computer Science)*, Vol. 7491. Springer, 42–51.

[3] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2017. Quality Gain Analysis of the Weighted Recombination Evolution Strategy on General Convex Quadratic Functions. In *Proceedings of the 14th Conference on Foundations of Genetic Algorithms (FOGA) 2017*. 111–126.

[4] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. 2017. Quality Gain Analysis of the Weighted Recombination Evolution Strategy on General Convex Quadratic Functions. *arXiv:1604.00772* (2017). arXiv:1608.04813 https://arxiv.org/pdf/1608.04813.pdf

[5] Shun-ichi Amari. 1998. Natural Gradient Works Efficiently in Learning. *Neural Computation* 10 (1998), 251–276.

[6] Dirk V. Arnold. 2005. Optimal Weighted Recombination. In *Proceedings of the 8th Conference on Foundations of Genetic Algorithms (FOGA) 2005*. Springer, 215–237.

[7] Anne Auger. 2005. Convergence results for the (1,λ)-SA-ES using the theory of φ-irreducible Markov chains. *Theoretical Computer Science* 334, 1–3 (2005), 35–69.

[8] Shummet Baluja. 1994. *Population-Based Incremental Learning: A Method for Integrating Genetic Search Based Function Optimization and Competitive Learning*. Technical Report. Carnegie Mellon University Pittsburgh.

[9] Hans Georg Beyer. 2014. Convergence Analysis of Evolutionary Algorithms That Are Based on the Paradigm of Information Geometry. *Evolutionary Computation* 22, 4 (2014), 679–709.

[10] James F. Crow and Motoo Kimura. 1979. Efficiency of Truncation Selection. In *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 76. 396–399.

[11] Tobias Glasmachers. 2012. Convergence of the IGO-Flow of Isotropic Gaussian Distributions on Convex Quadratic Problems. In *Parallel Problem Solving from Nature - PPSN XII, Part I (Lecture Notes in Computer Science)*, Vol. 7491. Springer, 1–10.

[12] Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. 2003. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation* 11, 1 (2003), 1–18.

[13] Silja Meyer-Nieberg and Hans-Georg Beyer. 2012. *The Dynamical Systems Approach — Progress Measures and Convergence Properties*. Springer-Verlag Berlin Heidelberg, 741–814.

[14] Yann Ollivier, Ludovic Arnold, Anne Auger, and Nikolaus Hansen. 2017. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *Journal of Machine Learning Research* 18, 1 (2017), 564–628.