Natural Evolution Strategies Converge on Sphere Functions

Tom Schaul
Courant Institute of Mathematical Sciences
New York University
715, Broadway, New York, NY 10003
schaul@cims.nyu.edu

ABSTRACT

This theoretical investigation gives the first proof of convergence for (radial) natural evolution strategies, on d-dimensional sphere functions, and establishes the conditions on hyper-parameters, as a function of d. For the limit case of large population sizes we show asymptotic linear convergence, and in the limit of small learning rates we give a full analytic characterization of the algorithm dynamics, decomposed into transient and asymptotic phases. Finally, we show why omitting the natural gradient is catastrophic.

Categories and Subject Descriptors

F.1.2 [**Analysis of Algorithms and Problem Complexity**]: Numerical Algorithms and Problems

General Terms

Algorithms, Theory

Keywords

black-box optimization, evolution strategies, stability theory, sphere function, convergence proof, natural gradient

1. INTRODUCTION

One of the most widely used algorithm classes for continuous black-box optimization are evolution strategies (ES). Their modern variants adapt multi-variate search distributions and include covariance matrix adaptation (CMA-ES [9]) and natural evolution strategies (NES [16]). While these methods were widely adopted for practical applications, and successfully so, theoretical work has lagged behind their developments, because it remained mostly focused on (1+1) ES (and its self-adaptive variant). For a recent review of the state of the field, see [4], and references therein. A noteworthy (negative) result for the broader category of purely comparison-based algorithms is that convergence can at best be linear [15].

GECCO'12, July 7-11, 2012, Philadelphia, Pennsylvania, USA.

The present paper aims to establish first positive convergence results for a popular class of evolutionary optimization algorithms, by studying the convergence properties of a canonical version of NES on d-dimensional sphere functions, a problem class which serves as a minimal condition for convergence to arbitrary precision.

Our approach consists in studying the dynamics of the algorithm as a stochastic process using tools from stability theory (resembling the dynamic systems approach taken in [6], and to a lesser degree related to the phi-irreducible Markov chains used in [3]). We make use of a novel type of utility function (section 2.2) that delivers manageable analytic expressions. Beyond establishing convergence itself, we obtain an explicit characterization of the dynamics, and exact convergence rates, in the limit cases of large populations sizes or small learning rates. Apart from the numerous technical results, in section 3.3 we also provide the (more casual) reader with visualizations of the dynamics, and intuitions for why the algorithm dynamics are composed of three distinct regimes.

Our investigation also includes a number of practical ramifications, giving new guidelines for hyper-parameter settings (and validating existing heuristics) in section 6.

It is a notable strength of the NES framework that the user may choose among many types of search distributions, and weighting functions for the samples, properties that were prerequisites for our explicit analysis. Nevertheless, we expect the techniques from this paper to be applicable to studying other, related algorithms.

2. NATURAL EVOLUTION STRATEGIES

In real-valued optimization, we call $f : \mathcal{S} \subset \mathbb{R}^d \mapsto \mathbb{R}$ the *fitness* function on some search space $\mathcal{S}$, which admits an optimum $f^* = f(\mathbf{z}^*)$.

Natural evolution strategies (NES) [16] are a class of evolutionary algorithms for this type of problems, which maintain a search distribution π and adapt the distribution parameters θ towards higher expected fitness J, that is, maximizing

$$J(\theta) = \mathbb{E}_\theta[f(\mathbf{z})] = \int f(\mathbf{z})\ \pi(\mathbf{z}\,|\,\theta)\ d\mathbf{z} \qquad (1)$$

Each iteration the algorithm produces n samples $\mathbf{z}_i \sim \pi(\mathbf{z}|\theta)$, $i \in \{1, \ldots, n\}$, i.i.d. from its search distribution, which is parameterized by θ. The gradient w.r.t. the parameters θ can be rewritten (see [16]) as

$$\nabla_\theta J(\theta) \quad = \nabla_\theta \int f(\mathbf{z})\ \pi(\mathbf{z}\,|\,\theta)\ d\mathbf{z} \quad = \mathbb{E}_\theta\left[f(\mathbf{z})\ \nabla_\theta \log \pi(\mathbf{z}\,|\,\theta)\right]$$

from which we obtain a Monte Carlo estimate

$$\nabla_\theta J(\theta) \approx \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{z}_i)\ \nabla_\theta \log \pi(\mathbf{z}_i \,|\, \theta)$$

of the search gradient. The key step then consists in replacing this gradient by the *natural gradient* [1], defined as $\mathbf{F}^{-1}\nabla_\theta J(\theta)$ where $\mathbf{F} = \mathbb{E}\left[\nabla_\theta \log \pi\left(\mathbf{z}|\theta\right) \nabla_\theta \log \pi\left(\mathbf{z}|\theta\right)^\top\right]$ is the Fisher information matrix. The search distribution is iteratively updated using natural gradient ascent

$$\theta \leftarrow \theta + \eta \mathbf{F}^{-1}\nabla_\theta J(\theta) \tag{2}$$

with learning rate parameter η.

This general formulation is applicable to arbitrary parameterizable search distributions [16, 12], including multivariate Gaussians [8] and Cauchy distributions [13].

2.1 Radial NES

To facilitate our study, we consider the simple but useful case of NES with radial Gaussian distributions (as in [5]), with search distribution $\pi_t = \mathcal{N}(\boldsymbol{\mu}_t, \sigma_t^2\mathbb{I})$, where $\mathbb{I}$ is the d-dimensional identity matrix. The distribution parameters are a mean vector $\boldsymbol{\mu}_t \in \mathbb{R}^d$ and a scale coefficient $\sigma_t \in \mathbb{R}^+$. The update equations at time-step t follow from equation 2 and read:

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t + \eta_{\boldsymbol{\mu}}\sigma_t \sum_{i=1}^{n} u_i \mathbf{s}_i \\ \sigma_{t+1} &= \sigma_t \cdot \exp\left(\frac{\eta_\sigma}{2}\sum_{i=1}^{n} u_i(\|\mathbf{s}_i\|^2 - d)\right)\end{aligned} \tag{3}$$

where a utility weight u_i has substituted the fitness $f(\mathbf{z}_i)$, and where the $\mathbf{s}_i$ are standard multi-normal sample points in the *natural coordinate system*, that is $\mathbf{z}_i = \boldsymbol{\mu}_t + \sigma_t\mathbf{s}_i$. Operating on natural coordinates is the reason for the multiplicative update on σ_t (see also [8] for a more extensive discussion of natural coordinate systems). Besides the initial distribution $\langle \boldsymbol{\mu}_0, \sigma_0\rangle$, the algorithm has three effective parameters:

- the population size n,
- the two learning rates $\eta_{\boldsymbol{\mu}} > 0$ and $\eta_\sigma > 0$, and
- the function u_i, which assigns 'utility' weights to each sample $\mathbf{z}_i$, based exclusively on the rank of its fitness. It satisfies $\sum^n |u_i| = 1$, which helps disentangle the effects of the learning rate from those of redistributing sample utility. It is also rank-preserving: $f(\mathbf{z}_i) > f(\mathbf{z}_j) \Rightarrow u_i > u_j$.

Using a rank-based utility function, instead of the raw fitness values f, increases robustness [16, 12] and makes the algorithm invariant under monotone transformations of the fitness function.

2.2 Gaussian utility functions

One way to simplify the subsequent analysis, is to choose a utility function that has the property of becoming a Gaussian density in the limit of infinitely many samples.

Under the search distribution π_t, the values of the sphere function $f_i = f(\mathbf{z}_i) \propto \|\mathbf{z}_i/\sigma_t\|^2$ follow a noncentral χ^2-distribution with d degrees of freedom and noncentrality parameter $\lambda = \Delta_t^2/\sigma_t^2$, where $\Delta_t = \|\boldsymbol{\mu}_t\|$. Thus, the (normalized) rank function $\mathrm{rank}(f_i)$ coincides with the corresponding cumulative distribution function (CDF; with value 0 for the best point, 1 for the worst):

$$\mathrm{rank}(f_i) = \mathrm{CDF}\left[\chi^2_{\nu=d,\lambda=\Delta_t^2/\sigma_t^2}\right](f_i)$$

In order to numerically[1] compute $\mathrm{rank}^{-1}(x)$, we need to estimate Δ_t somehow. One option is to take the distance to the currently best sample as a reasonable (but underestimating) proxy: $\Delta_t \approx \|\boldsymbol{\mu}_t - \mathbf{z}_b\|$, where $\mathbf{z}_b$ is the best sample of the iteration: $b = \arg\max_{1\le i\le n}(f_i)$. This requires using a utility function that changes in time, and is similar to an existing (successful) speedup technique, called 'distance-weighted' NES (DX-NES [7]), which adapts learning rates online (which has a similar effect to adjusting the utility function), to one of three discrete settings, based on an estimated value of Δ_t^2/σ_t^2.

In section 3.3, we will see that the quantity Δ_t^2/σ_t^2 reaches an equilibrium point near d, and we thus take that point as default value, if a static utility function is desired.

Using this inverse rank function, we constrain $u_i(\mathrm{rank}(f_i))$ to be approximately Half-Gaussian:

$$\begin{aligned}u_i(\mathrm{rank}(f_i)) &= 2\phi_{\langle 0,1\rangle}\left(\sqrt{f_i}/\sigma_t\right) \\ \Leftrightarrow u_i(x) &= 2\phi_{\langle 0,1\rangle}\left(\sqrt{\mathrm{rank}^{-1}(x)}/\sigma_t\right)\end{aligned}$$

where $\phi_{\langle\mu,\sigma^2\rangle}$ denotes the Gaussian density function. Now, in the limit of many samples, $u_i(\mathrm{rank}(f_i)) \cdot \pi_t(\mathbf{z}_i)$ is again

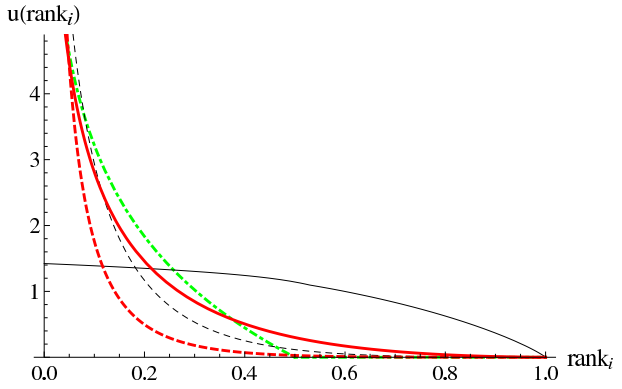

Figure 1: Different utility functions. With $\Delta_t^2/\sigma_t^2 = r_t \gg 1$, the utilities resemble the classical exponential decay from best to worst (dashed lines; thin black is for $d = 1$, thick red is for $d = 10$). At the other extreme, with $r_t \ll 1$, the utilities decay exponentially for large dimensions (solid thick red, $d = 10$), but are very flat for $d = 1$ (solid thin black). For comparison, we also show the most common utility function used in previous work, which is proportional to $\log(\mathrm{rank}_i)$, but uses a cutoff (dashed-dotted, green): the approximation by the Gaussian utilities is close if r_t is not too small, which is the case as $r_t \to r_* \approx d$. Also note that in all cases the plots are normalized to an area under the curve of 1; in practice this normalization constant depends on the number of samples n.

[1]There is no closed form for the inverse, but numerical approximations can be based on [11], for example.

normally distributed. In the natural coordinates we obtain

$$u(\mathbf{z}) \cdot \pi_t(\mathbf{s}) \approx \phi_{\left\langle \frac{-\boldsymbol{\mu}_t}{\sigma_t}, \mathbb{I} \right\rangle}(\mathbf{s}) \cdot \phi_{\langle \mathbf{0}, \mathbb{I} \rangle}(\mathbf{s}) = \phi_{\langle \boldsymbol{\mu}', \sigma'^2 \mathbb{I} \rangle}(\mathbf{s})$$

where the resulting distribution has parameters

$$\boldsymbol{\mu}_t' = \frac{1 \cdot \frac{-\boldsymbol{\mu}_t}{\sigma_t} + \mathbf{0}}{1+1} = \frac{-\boldsymbol{\mu}_t}{2\sigma_t} \quad , \quad \sigma'^2 = \frac{1}{2} \tag{4}$$

Figure 1 shows how the function u_i varies with rank, for different values of the ratio $r_t = \Delta_t^2/\sigma_t^2$ and different problem dimensions d. Note that this new utility function closely resembles the most popularly used (and heuristically chosen) utility function [8] (shown as dashed-dotted green line on Figure 1); we expect this approximation to be sufficient because empirical results indicate that the precise utility function does not affect performance substantially [12].

3. CONVERGENCE ON SPHERE

The goal of any stochastic search algorithm is to find a solution that is arbitrarily close in value to the optimum $f(\mathbf{z}^*)$. Given the stochastic nature of algorithms with search distributions, an appropriate success criterion can be formulated as follows: For any given $\epsilon > 0$, at least half of the samples $\mathbf{z} \sim \pi_t$, drawn from the search distribution (at iteration $t = T$) have a fitness value not worse than ϵ from the optimal one. Or more concisely:

Criterion 1 (Stochastic convergence).

$$\exists T \in \mathbb{N}, \forall t > T, \mathbf{z} \sim \pi_T \Rightarrow P(|f(\mathbf{z}^*) - f(\mathbf{z})| < \epsilon) > \frac{1}{2}$$

If we further assume that the fitness function has bounded curvature around the optimum:

$$|f(\mathbf{z}^*) - f(\mathbf{z})| < \epsilon \Rightarrow \frac{|f(\mathbf{z}^*) - f(\mathbf{z})|}{\|\mathbf{z}^* - \mathbf{z}\|} < K$$

for some constant $K < \infty$, then criterion 1 is also implied by convergence based on distance:

$$\forall \mathbf{z} \in \mathcal{S}, \|\mathbf{z}^* - \mathbf{z}\| < \epsilon' \Rightarrow |f(\mathbf{z}^*) - f(\mathbf{z})| < \epsilon \tag{5}$$

for any $\epsilon' < \frac{\epsilon}{K}$. That is, if the distribution is close enough to the optimum, then the fitness will also be good enough.

3.1 Sphere functions

Let f be the simple but common *sphere function*, in d dimensions:

$$\forall \mathbf{z} \in \mathbb{R}^d, f(\mathbf{z}) = \mathbf{z}^\top \mathbf{z},$$

where the objective is to minimize it. The sphere function is commonly used as a test function for optimization algorithms, because any smooth function is locally quadratic near its optimum, and thus convergence on the sphere function is a necessary condition for convergence on any smooth function.

Translation invariance: For convenience of notation, we use the sphere function centered at zero, but all our results can trivially be generalized to the translated variant $f(\mathbf{z}) = (\mathbf{z} - \mathbf{z}^*)^\top (\mathbf{z} - \mathbf{z}^*)$.

Intrinsic one-dimensionality: The sphere function is rotation-symmetric (around $\mathbf{z}^* = \mathbf{0}$), and the search distribution is also rotation symmetric (around $\boldsymbol{\mu}_t$). Without loss of generality, we can therefore rotate the coordinate system at each iteration, such that $\boldsymbol{\mu}_t = (\Delta_t, 0, \cdots, 0)$, where $\Delta_t = \|\boldsymbol{\mu}_t - \mathbf{0}\|$, the distance of the current distribution's center to the optimum.

We can now reformulate criterion 1 for the case of the sphere function:

$$\begin{aligned}
\frac{1}{2} < & \int_{\|\mathbf{z}\| < \epsilon'} \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{1}{2\sigma_t^2}\|\mathbf{z} - \boldsymbol{\mu}_t\|^2\right) d\mathbf{z} \\
= & \int_{\|\mathbf{z}\|^2 < \epsilon'^2} \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{1}{2\sigma_t^2}\|\mathbf{z} - \boldsymbol{\mu}_t\|^2\right) d\mathbf{z} \\
= & \int_{\|\mathbf{z}'\|^2 < \epsilon'^2/\sigma_t^2} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|\mathbf{z}' - \boldsymbol{\mu}_t/\sigma_t\|^2\right) d\mathbf{z}' \\
= & \int_0^{\frac{\epsilon'^2}{\sigma_t^2}} \chi^2_{\nu=d, \lambda=\Delta_t^2/\sigma_t^2}(x)\, dx
\end{aligned} \tag{6}$$

where $\chi^2_{\nu,\lambda}$ is the density of the noncentral χ^2-distribution, with $\nu = d$ degrees of freedom and noncentrality parameter $\lambda = \Delta_t^2/\sigma_t^2$. In other words, the inequality holds if and only if ϵ'^2/σ_t^2 is larger than the distribution's median $m_{\nu,\lambda}$. From [14] we know that

$$\nu - 2 + \lambda < m_{\nu,\lambda} \leq \nu + \lambda,$$

therefore the inequality 6 is verified if

$$d + \frac{\Delta_t^2}{\sigma_t^2} < \frac{\epsilon'^2}{\sigma_t^2} \Leftrightarrow \Delta_t^2 + \sigma_t^2 d < \epsilon'^2 \tag{7}$$

Thus, we find that criterion 1 together with equations 5 and 7 lead to a simple new success criterion for the sphere function:

Criterion 2 (Convergence on sphere).

$$\exists T \in \mathbb{N}, \forall t > T, \Delta_T^2 + \sigma_T^2 d < \epsilon'^2$$

In other words, both Δ_t and σ_t must converge to zero.

3.2 Update dynamics

This section contains a number of auxiliary derivations of the expectations of parameter updates, required for the analysis in the subsequent sections.

Using the Gaussian utilities from section 2.2, we can rewrite the update equations 3 as

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \eta_{\boldsymbol{\mu}} \sigma_t \sum_{i=1}^n u_i \mathbf{s}_i = \boldsymbol{\mu}_t + \eta_{\boldsymbol{\mu}} \sigma_t \frac{1}{n} \sum_{i=1}^n \mathbf{s}_i'$$

where the samples $\mathbf{s}_i'$ are now drawn from $\mathcal{N}(\boldsymbol{\mu}', \sigma'^2\mathbb{I})$, with parameters from equation 4. We can decompose the sample vectors $\mathbf{s}_i$ into their components $\mathbf{s}_{i,j}$ with $j \in \{1, \ldots, d\}$:

$$\begin{aligned}
\Delta_{t+1}^2 &= \left\|\boldsymbol{\mu}_{t+1}\right\|^2 = \left\|\boldsymbol{\mu}_t + \eta_{\boldsymbol{\mu}} \sigma_t \frac{1}{n} \sum_{i=1}^n \mathbf{s}_i'\right\|^2 \\
&= \left(\Delta_t + \eta_{\boldsymbol{\mu}} \sigma_t \frac{1}{n} \sum_{i=1}^n \mathbf{s}_{i,1}'\right)^2 + \eta_{\boldsymbol{\mu}}^2 \sigma_t^2 \sum_{j=2}^d \left(\frac{1}{n} \sum_{i=1}^n \mathbf{s}_{i,j}'\right)^2
\end{aligned}$$

So, denoting by ξ_k independent standard normal (1-dimensional) samples we can explicitly separate the stochastic parts from the deterministic ones: The intrinsic one-dimensionality gives

us $\mathbf{s}_{i,1} = \frac{-\Delta_t}{2\sigma_t} + \frac{1}{\sqrt{2}}\xi_1$ and $\mathbf{s}_{i,j} = \frac{1}{\sqrt{2}}\xi_j$ for $j > 1$; therefore

$$\begin{aligned}\Delta_{t+1}^2 = & \left(\eta_{\boldsymbol{\mu}}\sigma_t \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\Delta_t}{\eta_{\boldsymbol{\mu}}\sigma_t} + \frac{-\Delta_t}{2\sigma_t} + \frac{1}{\sqrt{2}}\xi_i\right)\right)^2 \\ & + \eta_{\boldsymbol{\mu}}^2\sigma_t^2 \sum_{j=2}^{d}\left(\frac{1}{n\sqrt{2}}\sum_{i=1}^{n}\xi_{jd+i}\right)^2 \\ = & \frac{(2-\eta_{\boldsymbol{\mu}})^2}{4}\Delta_t^2 + \frac{\eta_{\boldsymbol{\mu}}(2-\eta_{\boldsymbol{\mu}})\sigma_t\Delta_t}{n\sqrt{2}}\sum_{i=1}^{n}\xi_i + \frac{\eta_{\boldsymbol{\mu}}^2\sigma_t^2}{2n^2}\sum_{j=1}^{nd}\xi_j^2\end{aligned}$$

giving, in expectation

$$\mathbb{E}[\Delta_{t+1}^2] = \frac{(2-\eta_{\boldsymbol{\mu}})^2}{4}\Delta_t^2 + \frac{\eta_{\boldsymbol{\mu}}^2 d}{2n}\sigma_t^2 \quad (8)$$

$$\mathbb{V}\mathrm{ar}[\Delta_{t+1}^2] = \frac{\eta_{\boldsymbol{\mu}}^2(2-\eta_{\boldsymbol{\mu}})^2}{2n}\sigma_t^2\Delta_t^2 + \frac{\eta_{\boldsymbol{\mu}}^4 d}{2n^2}\sigma_t^4 \quad (9)$$

Similarly, we rewrite the update of σ_t (from equation 3):

$$\begin{aligned}\sigma_{t+1}^2 &= \sigma_t^2 \exp\left(\eta_\sigma \sum_{i=1}^{n} u_i(\|\mathbf{s}_i\|^2 - d)\right) \\ &= \sigma_t^2 \exp\left(-\eta_\sigma d + \frac{\eta_\sigma}{n}\sum_{i=1}^{n}\|\mathbf{s}_i'\|^2\right) \\ &= \sigma_t^2 \exp\left(-\eta_\sigma d + \frac{\eta_\sigma}{2n}\sum_{i=1}^{n}\left[\left(-\frac{\Delta_t}{\sigma_t} + \xi_i\right)^2 + \sum_{j=2}^{d}\xi_j^2\right]\right) \\ &= \sigma_t^2 \exp\left(-\eta_\sigma d + \frac{\eta_\sigma\Delta_t^2}{2\sigma_t^2} - \frac{\eta_\sigma\Delta_t}{n\sigma_t}\sum_{i=1}^{n}\xi_i + \frac{\eta_\sigma}{2n}\sum_{j=1}^{nd}\xi_j^2\right)\end{aligned}$$

To compute the expectation, we can decompose this into a product, because all the ξ_i are independent of each other. Using the auxiliary results

$$\mathbb{E}\left[\exp\left(\tfrac{\eta_\sigma}{2n}\xi^2\right)\right] = \left(1 - \tfrac{\eta_\sigma}{n}\right)^{-1/2}$$

$$\mathbb{E}\left[\exp\left(-\tfrac{\eta_\sigma\Delta_t}{n\sigma_t}\xi + \tfrac{\eta_\sigma}{2n}\xi^2\right)\right] = \left(1 - \tfrac{\eta_\sigma}{n}\right)^{-1/2}\exp\left(\tfrac{\eta_\sigma^2\Delta_t^2}{2n\sigma_t^2(n-\eta_\sigma)}\right)$$

we obtain

$$\begin{aligned}\mathbb{E}[\sigma_{t+1}^2] = & \sigma_t^2 \exp\left(-\eta_\sigma d + \tfrac{\eta_\sigma\Delta_t^2}{2\sigma_t^2}\right) \\ & \cdot\mathbb{E}\left(\exp\left[-\tfrac{\eta_\sigma\Delta_t}{n\sigma_t}\xi_i + \tfrac{\eta_\sigma}{2n}\xi_i^2\right]\right)^n \mathbb{E}\left(\exp\left[\tfrac{\eta_\sigma}{2n}\xi_j^2\right]\right)^{n(d-1)} \\ = & \sigma_t^2\left(1 - \tfrac{\eta_\sigma}{n}\right)^{-nd/2}\exp\left(-\eta_\sigma d + \tfrac{\eta_\sigma\Delta_t^2}{2\sigma_t^2} + \tfrac{\eta_\sigma^2\Delta_t^2}{2\sigma_t^2(n-\eta_\sigma)}\right) \\ \approx & \sigma_t^2 \exp\left(-\tfrac{\eta_\sigma}{2}d + \tfrac{n\eta_\sigma}{2(n-\eta_\sigma)}\tfrac{\Delta_t^2}{\sigma_t^2}\right) \quad (10)\end{aligned}$$

where we used $\log(1 - \frac{\eta_\sigma}{n}) \approx -\frac{\eta_\sigma}{n}$, with the assumption $\eta_\sigma \ll n$. Similarly, the variance of the update becomes:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\sigma_{t+1}^2] = & -\mathbb{E}[\sigma_{t+1}^2]^2 + \sigma_t^4 \exp\left(-2\eta_\sigma d + \tfrac{\eta_\sigma\Delta_t^2}{\sigma_t^2}\right) \\ & \cdot\mathbb{E}\left(\exp\left[-\tfrac{2\eta_\sigma\Delta_t}{n\sigma_t}\xi_i + \tfrac{\eta_\sigma}{n}\xi_i^2\right]\right)^n \mathbb{E}\left(\exp\left[\tfrac{\eta_\sigma}{n}\xi_j^2\right]\right)^{n(d-1)} \\ = & \sigma_t^4 \exp\left(-2\eta_\sigma d + \tfrac{n\eta_\sigma}{n-\eta_\sigma}\tfrac{\Delta_t^2}{\sigma_t^2}\right) \\ & \cdot\left[\left(1 - \tfrac{2\eta_\sigma}{n}\right)^{-nd/2} - \left(1 - \tfrac{\eta_\sigma}{n}\right)^{-nd}\right] \\ \approx & \sigma_t^4 \exp\left(-\eta_\sigma d + \tfrac{n\eta_\sigma}{n-\eta_\sigma}\tfrac{\Delta_t^2}{\sigma_t^2}\right)\left[\exp\left(\tfrac{\eta_\sigma^2 d}{2n-4\eta_\sigma}\right) - 1\right] \quad (11)\end{aligned}$$

3.3 Phase plane analysis

The fixed points of the dynamics are those values for which the expected update is the identity function. As a prerequisite for our stability analysis, we thus determine the fixed point for each variable. We have a fixed point for Δ_t^2 if the equality holds in

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}^2] &\leq \Delta_t^2 \\ \Leftrightarrow \frac{\eta_{\boldsymbol{\mu}}^2 d}{2n}\sigma_t^2 + \frac{\eta_{\boldsymbol{\mu}}^2 - 4\eta_{\boldsymbol{\mu}}}{4}\Delta_t^2 &\leq 0 \\ \Leftrightarrow \frac{\Delta_t^2}{\sigma_t^2} &\geq \frac{2\eta_{\boldsymbol{\mu}}}{n(4-\eta_{\boldsymbol{\mu}})}d \quad (12)\end{aligned}$$

Note that we only have a fixed point only if $\eta_{\boldsymbol{\mu}} < 4$, so for larger values, we have divergence. Similarly, we have a fixed point of σ_t^2 if the equality holds in

$$\begin{aligned}\mathbb{E}[\sigma_{t+1}^2] &\leq \sigma_t^2 \\ \Leftrightarrow \exp\left(-\frac{\eta_\sigma}{2}d + \frac{n\eta_\sigma}{2(n-\eta_\sigma)}\frac{\Delta_t^2}{\sigma_t^2}\right) &\leq 1 \\ \Leftrightarrow \frac{\Delta_t^2}{\sigma_t^2} &\leq \frac{n-\eta_\sigma}{n}d \quad (13)\end{aligned}$$

Given that the left term is positive, we only have a fixed point for σ_t if $\eta_\sigma < n$.

Thus, it is clear that under these conditions, the *single* fixed point for the joint updates is where these two lines intersect, namely at $\langle\Delta_t^2 = 0, \sigma_t^2 = 0\rangle$.

Based on the inequalities 12 and 13, we find that the dynamics can be divided into three distinct regimes:

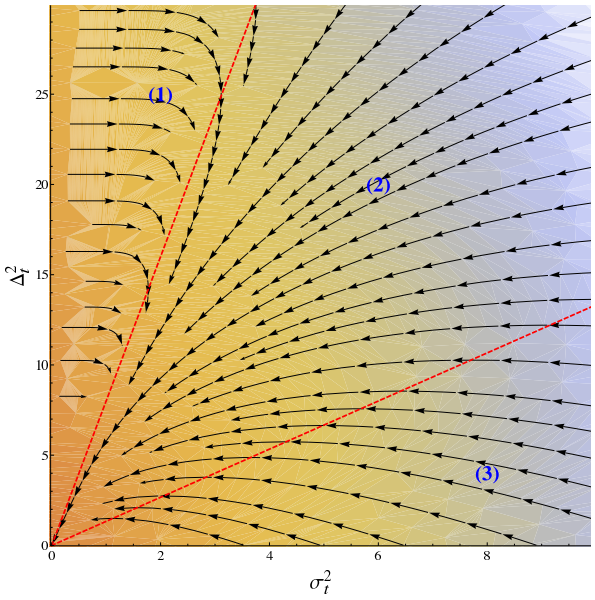

Figure 2: NES phase plane. Black arrows give the directions of the expected combined update of Δ_t^2 and σ_t^2 (magnitudes are not to scale). The settings used are $d = 10$, $n = 5$, $\eta_{\boldsymbol{\mu}} = 1$ and $\eta_\sigma = 1$. The dashed red lines indicate the null clines (fixed points) for the two variables, and at the same time delimit the three dynamic regimes. The underlaid color gradient is corresponds to the objective from criterion 2. Visibly, regimes (1) and (3) are transient, and all updates in (2) lead toward the optimum (at the origin). Note also the very drastic increase of σ_t^2 whenever it is much smaller than Δ_t^2/d, in regime (1).

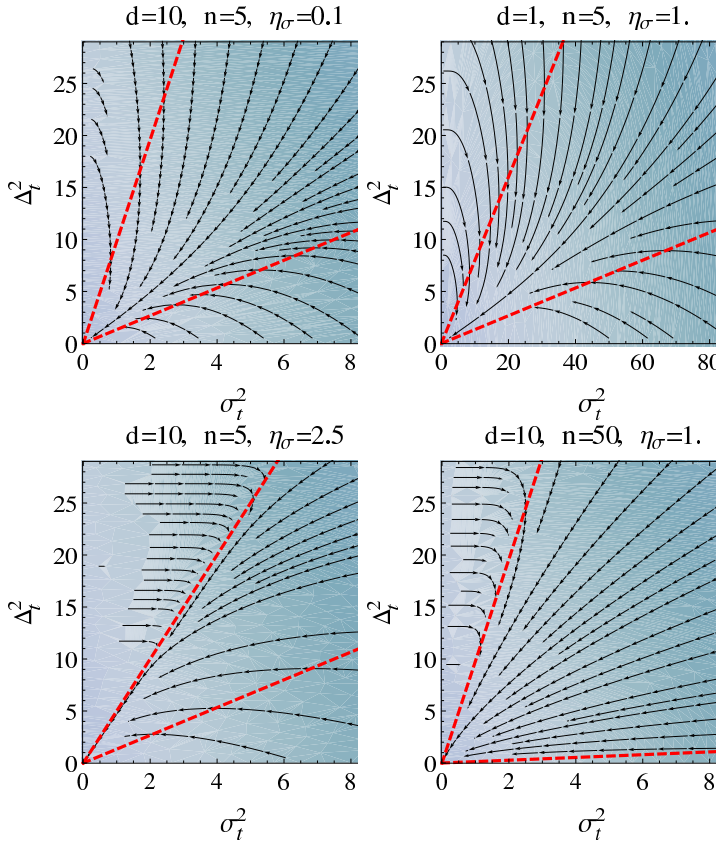

Figure 3: Additional NES phase planes. Compared to figure 2, we varied d, n and η_σ. In the top left, we see that reducing η_σ by a factor 10 smoothens the transition between regimes (1) and (2). On the top right, we illustrate how similar the dynamics remain (as compared to the corresponding settings top left), even when reducing the dimension by a factor 10. The bottom left plot shows how increasing η_σ substantially changes both the borders of the regimes, narrowing regime (2), and the dynamics within. Finally, the bottom right plot reduces the noise by increasing n by a factor 10; as expected, regime (3) almost vanishes.

(1) If $\frac{\Delta_t^2}{\sigma_t^2} \geq \frac{n-\eta_\sigma}{n} d$, then σ_t^2 increases, Δ_t^2 decreases, so $\frac{\Delta_t^2}{\sigma_t^2}$ decreases and we approach regime (2).

(2) If $\frac{n-\eta_\sigma}{n} d > \frac{\Delta_t^2}{\sigma_t^2} \geq \frac{2\eta_\mu}{n(4-\eta_\mu)} d$, then σ_t^2 decreases, Δ_t^2 decreases, and we approach the goal of $\Delta_t^2 + \sigma_t^2 d < \epsilon'$, and thus converge according to criterion 2.

(3) If $\frac{2\eta_\mu}{n(4-\eta_\mu)} d > \frac{\Delta_t^2}{\sigma_t^2} > 0$, then σ_t^2 decreases, Δ_t^2 increases so $\frac{\Delta_t^2}{\sigma_t^2}$ increases and we approach regime (2).

From this, we can deduce that regimes (1) and (3) are transient[2], and that the asymptotic regime is (2). Therefore, the algorithm will always converge according to criterion 2 in expectation.

The intuitive interpretations of these are that when Δ_t is much larger than σ_t (regime 1), the algorithm is far from the solution, and consequently increases the scale of the search distribution; which then allows it to reduce Δ_t with increasingly larger steps. On the other hand, if Δ_t is much smaller than σ_t (regime 3), it is so close to the optimum that the noise effects tend to temporarily lead it further away. Otherwise (regime 2), both Δ_t and σ_t decrease to obtain convergence.

The phase planes in figures 2 and 3 visualize the update directions in the different regimes, for a number of different dimensions and parameter settings.

3.4 Noise-free discrete dynamics

In our first limit case we study the dynamics in the absence of noise, which we can obtain by taking $n \to \infty$. We start by looking at the ratio $r_t = \frac{\Delta_t^2}{\sigma_t^2}$, which is omnipresent because it corresponds to the scale-invariant part of the state variables. We have:

$$\begin{aligned}\mathbb{E}\left[\frac{\Delta_{t+1}^2}{\sigma_{t+1}^2}\right] = \mathbb{E}\Big[\exp\left(\eta_\sigma d - \tfrac{\eta_\sigma \Delta_t^2}{2\sigma_t^2}\right)\Big[\tfrac{(2-\eta_\mu)^2}{4}\Delta_t^2 \\ + \tfrac{\eta_\mu(2-\eta_\mu)\sigma_t \Delta_t}{n\sqrt{2}}\textstyle\sum_{i=1}^n \xi_i + \tfrac{\eta_\mu^2 \sigma_t^2}{2n^2}\textstyle\sum_{j=1}^{nd}\xi_j^2\Big] \\ \cdot\ \sigma_t^{-2}\exp\left(\tfrac{\eta_\sigma \Delta_t}{n\sigma_t}\textstyle\sum_{i=1}^n \xi_i - \tfrac{\eta_\sigma}{2n}\textstyle\sum_{j=1}^{nd}\xi_j^2\right)\Big]\end{aligned}$$

which, following a derivation analogous to the one for 10, and using $n \to \infty$ gives

$$\mathbb{E}[r_{t+1}] = \mathbb{E}\left[\frac{\Delta_{t+1}^2}{\sigma_{t+1}^2}\right] = \frac{(2-\eta_\mu)^2}{4}\exp\left(\frac{1}{2}\eta_\sigma(d - r_t)\right) r_t$$

and thus the following expression for the fixed point r_*:

$$\begin{aligned}\mathbb{E}[r_{t+1}] &= r_t \\ \Leftrightarrow r_* &= d - \frac{4\log(2) - 4\log(2-\eta_\mu)}{\eta_\sigma}\end{aligned} \tag{14}$$

under the constraint that

$$\begin{aligned} r_* &\geq 0 \\ \Leftrightarrow \eta_\sigma d &\geq 4\log(2) - 4\log(2-\eta_\mu) \\ \Leftrightarrow (2-\eta_\mu)^2 &\geq 4\exp(-1/2\eta_\sigma d) \\ \Leftarrow \eta_\mu &\leq 2(1-\exp(-1/4\eta_\sigma d))\end{aligned} \tag{15}$$

Replacing in 10, this gives the asymptotic expected update

$$\begin{aligned}\mathbb{E}[\sigma_{t+1}^2] &= \sigma_t^2 \exp\left(\frac{\eta_\sigma^2 d}{2(n-\eta_\sigma)} - \frac{2n\log(2) - 2n\log(2-\eta_\mu)}{n-\eta_\sigma}\right) \\ &\approx \frac{(2-\eta_\mu)^2}{4}\sigma_t^2\end{aligned}$$

Thus there exists a time τ, for which

$$\sigma_{\tau+t}^2 = \left(\frac{(2-\eta_\mu)^2}{4}\right)^t \sigma_\tau^2 \geq \exp\left(-\frac{\eta_\sigma d}{2}t\right)\sigma_\tau^2$$

(from condition in equation 15), decay is analogously for $\Delta_{\tau+t}^2$. In other words, in the absence of noise, we have asymptotic linear convergence with a convergence rate $\exp\left(-\frac{\eta_\sigma d}{4}\right)$, which could in principle be boosted to arbitrary values by increasing η_σ. However, we will see below that for finite n the quantity $\eta_\sigma d$ should be bounded.

[2] A technical condition is that the updates are small enough to avoid oscillation between regimes (1) to (3), while regime (2) is 'over-jumped'. It can be shown that $\eta_\sigma < \frac{2}{d}\log\frac{4n}{\eta_\mu^2(n+2)}$ is a sufficient condition to avoid such oscillations (assuming $\eta_\mu \leq 1$).

4. CONTINUOUS-TIME DYNAMICS

Taking the learning rates to smaller values $\eta_{\boldsymbol{\mu}} \to dt \cdot k_{\boldsymbol{\mu}}$ and $\eta_\sigma \to dt \cdot k_\sigma$ while doing more iterations smoothens the updates of the algorithm. In the limit of infinitesimal steps, we obtain a stochastic continuous-time flow in $\mathbb{R}^2$ (as in [2]). Using our calculations in section 3.2, we can express it as a system of stochastic differential equations:

$$d\Delta_t^2 = -k_{\boldsymbol{\mu}}\Delta_t^2 dt + k_{\boldsymbol{\mu}}\sqrt{\frac{2}{n}}\,\sigma_t\Delta_t\; dW \quad (16)$$

$$d\sigma_t^2 = k_\sigma\left(\frac{1}{2}\Delta_t^2 - \frac{d}{2}\sigma_t^2\right)dt + k_\sigma\sqrt{\frac{d}{2n}}\sigma_t^2 dW \quad (17)$$

where dW denotes standard Brownian motion. Unlike in the discrete-time case, we can now explicitly compute the stability matrix from the Itō drift terms:

$$\begin{pmatrix} -k_{\boldsymbol{\mu}} & 0 \\ \frac{k_\sigma}{2} & -\frac{k_\sigma d}{2} \end{pmatrix}$$

which has two negative eigenvalues[3], which implies that the dynamics correspond to an asymptotically stable (i.e., absorbing) node at the origin.

The joint density $p(x, y, t)$ at time t of the coupled stochastic process (where $x = \Delta_t^2$ and $y = \sigma_t^2$) is described by the Fokker-Planck equation [10]:

$$\begin{aligned}\frac{\partial p}{\partial t} &= k_{\boldsymbol{\mu}}\frac{\partial}{\partial x}[xp] + \frac{k_\sigma}{2}\frac{\partial}{\partial y}[(yd - x)p] \\ &\quad + \frac{1}{n}\frac{\partial^2}{\partial x^2}[xyp] + \frac{k_\sigma^2 d}{2n}\frac{\partial^2}{\partial y^2}[y^2 p] \\ &= \left[k_{\boldsymbol{\mu}} + \frac{n + 2k_\sigma}{2n}k_\sigma d\right]f + \left[k_{\boldsymbol{\mu}}x + \frac{2}{n}\right]\frac{\partial f}{\partial x} + \frac{1}{n}xy\frac{\partial^2 f}{\partial x^2} \\ &\quad + \frac{k_\sigma}{2}\left[-x + d\frac{n + 4k_\sigma}{n}y\right]\frac{\partial f}{\partial y} + \frac{k_\sigma^2 d}{2n}y^2\frac{\partial^2 f}{\partial y^2} \quad (18)\end{aligned}$$

These full dynamics appear to not be analytically tractable[4], but the differential equation can be numerically simulated, which shows qualitatively that there is a transient phase where x may grow, and after the ratio $r_t = x/y$ stabilizes the densities converge exponentially to a δ-function above the origin.

In the next subsections, we attempt to formally characterize these dynamical properties. First, we study the exponential convergence after the transient adaptation of r_t has taken place (section 4.1). Second, we look at infinite population sizes (i.e., noise-free dynamics as in section 3.4, but now for the continuous case), which is a proxy for characterizing the dynamics of the mean of the distribution, in section 4.2.

4.1 Geometric Brownian motion

In the discrete (noisy) case, there is no analytical solution for r_* (the fixed point of r_t)[5], but here, with continuous dynamics, we can identify it:

$$\begin{aligned}\frac{\Delta^2}{\sigma_t^2} &= \frac{\Delta^2 + d\Delta^2}{\sigma_t^2 + d\sigma_t^2} \\ \Leftrightarrow \frac{\Delta^2}{\sigma_t^2} &= \frac{\Delta^2 - k_{\boldsymbol{\mu}}\Delta^2 dt}{\sigma_t^2 + \frac{k_\sigma}{2}\left(-\sigma_t^2 d + \Delta^2\right)dt} \\ \Leftrightarrow \frac{k_\sigma}{2}\left(-\sigma_t^2 d + \Delta^2\right)dt &= -k_{\boldsymbol{\mu}}\sigma_t^2 dt \\ \Leftrightarrow \frac{\Delta^2}{\sigma_t^2} &= d - \frac{2k_{\boldsymbol{\mu}}}{k_\sigma} \equiv r_* \quad (19)\end{aligned}$$

which is stable, but exists only if $k_\sigma > \frac{2}{d}k_{\boldsymbol{\mu}}$. Under this condition, we call the *transient* phase the early part of the run until r_t has (approximately) converged to this fixed point[6].

If the adaptation of r_t is transient, then there is a time τ such that $\forall t \geq \tau,\ r_t \approx r_*$. We can study the asymptotic dynamics by just considering the updates of σ_t^2, because $\Delta_t^2 \approx r_*\sigma_t^2$, simply. The system of stochastic differential equations 17 becomes a single equation

$$d\sigma_t^2 \approx -k_\sigma\frac{d - r_*}{2}\sigma_t^2 dt + k_\sigma\sqrt{\frac{d}{2n}}\sigma_t^2 dW \quad (20)$$

This has the well-known form of a *geometric Brownian motion*, which is known to converge linearly to zero in expectation, if the drift term is negative (which is always true, because $r_* < d$ is guaranteed by equation 19). Thus, following [10] and replacing the value of r_*, we have

$$\sigma_{\tau+t}^2 = \sigma_\tau^2\exp\left[-\left(k_{\boldsymbol{\mu}} + \frac{k_\sigma^2 d}{4n}\right)t + k_\sigma\sqrt{\frac{d}{2n}}dW\right]$$

and

$$\begin{aligned}\mathbb{E}(\sigma_{t+\tau}^2) &= \sigma_\tau^2\exp\left(-k_{\boldsymbol{\mu}}t\right) \\ \mathbb{V}\mathrm{ar}(\sigma_{t+\tau}^2) &= \sigma_\tau^4\exp\left(-k_{\boldsymbol{\mu}}t\right)\left[\exp\left(\frac{k_\sigma^2 d}{2n}t\right) - 1\right] \\ &< \sigma_\tau^4\exp\left[\left(-k_{\boldsymbol{\mu}} + \frac{k_\sigma^2 d}{2n}\right)t\right]\end{aligned}$$

Thus, we also have an exponential decrease in variance if

$$\begin{aligned}-k_{\boldsymbol{\mu}} + \frac{k_\sigma^2 d}{2n} &< 0 \\ \Leftrightarrow k_\sigma &< \sqrt{\frac{2nk_{\boldsymbol{\mu}}}{d}} \quad (21)\end{aligned}$$

which is in turn a guarantee that the density $p(x, y, t)$ will converge to a δ-peak above the origin.

4.2 Noise-free continuous dynamics

An alternative avenue to characterize convergence explicitly is taking $n \to \infty$, then the noise effects vanish, and from equation 16 we immediately obtain the expression of the exponential decay for Δ_t^2

$$\Delta_t^2 = \Delta_0^2\exp\left(-k_{\boldsymbol{\mu}} \cdot t\right) \quad (22)$$

which, when replacing in 17, gives the following inhomogeneous differential equation for σ_t^2:

$$\frac{d\sigma_t^2}{dt} = \frac{-k_\sigma d}{2}\sigma_t^2 + \frac{k_\sigma\Delta_0^2}{2}\exp\left(-k_{\boldsymbol{\mu}} \cdot t\right)$$

[3]The eigenvalues can be identical, but the node remains asymptotically stable.

[4]A good candidate guess for p would be the two-dimensional Wishart distribution where the three scale parameters and the degrees-of-freedom parameter are a function of time. But our attempts at identifying the explicit form failed due to the last (y^2) term in equation 18.

[5]It always exists, because for $r_t \ll 1$, we have $\mathbb{E}[r_{t+1}] \gg 0$ while for $r_t \gg 1$, we have $\mathbb{E}[r_{t+1}] \ll 1$, so they must intersect.

[6]A possible heuristic for bypassing this transient phase, is to initialize the algorithm with $\sigma_0^2 \approx \hat{\Delta}_0^2/d$, if the user can provide a guess $\hat{\Delta}_0$ of the initial distance to the optimum.

Using the method of underdetermined coefficients, and assuming $\frac{k_\sigma d}{2} \neq k_{\boldsymbol{\mu}}$, we guess at the form

$$\sigma_t^2 = C_1 \exp(-\frac{k_\sigma d}{2} t) + C_2 \exp(-k_{\boldsymbol{\mu}} t)$$

Differentiating and identifying the coefficients, we find:

$$C_1 = \sigma_0^2 - C_2 \quad , \quad C_2 = \frac{k_\sigma \Delta_0^2}{k_\sigma d - 2k_{\boldsymbol{\mu}}}$$

Otherwise, if $\frac{k_\sigma d}{2} = k_{\boldsymbol{\mu}}$, we guess at the form

$$\sigma_t^2 = (C_3 + C_4 t) \exp(-k_{\boldsymbol{\mu}} t)$$

and identify

$$C_3 = \sigma_0^2 \quad , \quad C_4 = \frac{k_\sigma \Delta_0^2}{2}$$

Clearly, Δ_t^2 converges to zero in $\mathcal{O}(e^{-k_{\boldsymbol{\mu}} t})$. For the convergence of σ_t^2 we distinguish three cases: if $\frac{k_\sigma d}{2} > k_{\boldsymbol{\mu}}$ then the convergence is in $\mathcal{O}(e^{-k_{\boldsymbol{\mu}} t})$, if we have equality it is in $\mathcal{O}(t \cdot e^{-k_{\boldsymbol{\mu}} t})$, otherwise it is in $\mathcal{O}(e^{-\frac{k_\sigma d}{2} t})$. The asymptotically best of these cases is when $\frac{k_\sigma d}{2} > k_{\boldsymbol{\mu}}$, and for which we obtain the total convergence rate on the relevant quantity $\sigma_t^2 d + \Delta_t^2$ of $\mathcal{O}(e^{-k_{\boldsymbol{\mu}} t})$. Explicitly:

$$\begin{aligned} \sigma_t^2 &= \sigma_0^2 \exp\left(-\frac{k_\sigma d}{2} t\right) \\ &\quad - \frac{k_\sigma \Delta_0^2}{k_\sigma d - 2k_{\boldsymbol{\mu}}} \left(\exp\left(-\frac{k_\sigma d}{2} t\right) - \exp\left(-k_{\boldsymbol{\mu}} t\right)\right) \quad (23) \end{aligned}$$

Note that this holds for all $t > 0$, unlike the strictly asymptotic characterization in section 3.4. We may see transient growth in σ_t^2, depending on the initial ratio $r_0 = \frac{\Delta_0^2}{\sigma_0^2}$, before the dominant term takes over, and the algorithm converges exponentially.

5. OMITTING THE NATURAL GRADIENT

It is not obvious at first sight that NES would perform qualitatively differently, were one to remove the natural gradient, as that would still optimize the right objective (equation 1). It is known [12] that this 'vanilla' version is not scale-invariant. With the tools introduced here, we can now characterize its contrasting convergence properties. The vanilla update equations are

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &= \boldsymbol{\mu}_t + \frac{\eta_{\boldsymbol{\mu}}}{\sigma_t} \sum_{i=1}^{n} u_i \mathbf{s}_i = \boldsymbol{\mu}_t + \frac{\eta_{\boldsymbol{\mu}}}{n\sigma_t} \sum_{i=1}^{n} \mathbf{s}_i' \\ \sigma_{t+1} &= \sigma_t + \frac{\eta_\sigma}{2\sigma_t} \sum_{i=1}^{n} u_i (\|\mathbf{s}_i\|^2 - d) \\ &= \sigma_t - \frac{\eta_\sigma d}{2\sigma_t} + \frac{\eta_\sigma}{2n\sigma_t} \sum_{i=1}^{n} \|\mathbf{s}_i'\|^2 \end{aligned}$$

leading to the expectations

$$\begin{aligned} \mathbb{E}[\Delta_{t+1}^2] &= \frac{(2\sigma_t^2 - \eta_{\boldsymbol{\mu}})^2}{4\sigma_t^4} \Delta_t^2 + \frac{\eta_{\boldsymbol{\mu}}^2 d}{2n\sigma_t^2} \quad (24) \\ \mathbb{E}[\sigma_{t+1}] &= \sigma_t - \frac{\eta_\sigma d}{4\sigma_t} + \frac{\eta_\sigma \Delta_t^2}{4\sigma_t^3} \end{aligned}$$

The fixed points are

$$\Delta_t^2 = \frac{2\eta_{\boldsymbol{\mu}} d}{n(\eta_{\boldsymbol{\mu}} + 4\sigma_t^2)} \sigma_t^2 \quad , \quad \sigma_t = \frac{\Delta_t}{\sqrt{d}}$$

which hold simultaneously if

$$\Delta_t^2 = \sigma_t^2 d \quad = \quad \frac{\eta_{\boldsymbol{\mu}} d(n-2)}{4n}$$

As this equilibrium is not at the origin (as it is for NES), this implies that we need to choose

$$\eta_{\boldsymbol{\mu}} < \frac{2n}{d\,(n-2)} \epsilon'^2$$

in order for the convergence condition from criterion 2 to hold eventually (at the equilibrium). This $\eta_{\boldsymbol{\mu}}$ is an infinitesimal quantity for any ambitious ϵ', so equation 24 becomes

$$\mathbb{E}[\Delta_{t+1}^2] \quad \approx \quad \left(1 - \frac{\eta_{\boldsymbol{\mu}}}{\sigma_t^2}\right) \Delta_t^2$$

which clearly shows that even if the algorithm converges (i.e., when noise effects are sufficiently attenuated with a large value of n), convergence must be very slow.

6. PARAMETER GUIDELINES

From the above results we can deduce guidelines for hyperparameter choices. Note that all of these, while useful, are to be taken with a grain of salt, because they result from the analysis of idealized scenarios, and not the exact algorithm.

- Section 5 justifies always using the natural gradient.
- Section 4.1 recommends the initialization $\sigma_0^2 \approx \frac{\hat{\Delta}_0^2}{d}$.
- Section 3.3 recommends setting $n > \frac{2d+6}{3d}$, where the larger n is, the closer the dynamics will be to those describes in section 3.4.
- Sections 4.1 and 4.2 recommend choosing $\frac{2}{d} < \eta_\sigma < \sqrt{\frac{2n}{d}}$, and section 3.4 showed that (at least if n is large) η_σ should be as large as possible[7].
- Section 3.4 recommends $\eta_{\boldsymbol{\mu}}$ to be as large as possible with $\eta_{\boldsymbol{\mu}} \leq 2 - 2\exp(-\frac{1}{4}\eta_\sigma d)$.

For example, given the often desirable choice of a minimal n, easy to remember learning rates (that satisfy the above) are $\eta_\sigma = \frac{2}{\sqrt{d}}$ and $\eta_{\boldsymbol{\mu}} = 1$.

A common choice in previous work on NES [12] has been $n = 4 + \lfloor 3\log(d) \rfloor$, $\eta_{\boldsymbol{\mu}} = 1$ and $\eta_\sigma = \frac{3+\log(d)}{5\sqrt{d}}$. These settings satisfy the constraint in equation 21 only for $d < 10^{63}$, but it is safe to say that includes all realistic cases.

7. DISCUSSION AND CONCLUSIONS

This paper has established a number of novel convergence results for natural evolution strategies. Using stability theory, we determined the parameter settings for which radial NES is guaranteed to converge on the sphere function.

For the most general case (section 3.3), we established a qualitative decomposition into three distinct regimes. For the limit case of large population sizes (section 3.4), we proved asymptotic linear convergence; for the orthogonal

[7]Note that in practice, choices of $\eta_\sigma < \frac{2}{d}$ often give robust empirical performance (especially on problems more difficult than sphere functions, see [8]), despite the absence of a distinct transient phase.

limit case of small learning rates (section 4.1), we determined the conditions for which both the mean and the variance of the diffusion converge linearly. Finally, combining both limit cases (section 4.2), we gave a full characterization of the dynamics, for both the transient and the asymptotic phases.

7.1 Generalizations

The performance on the sphere function generalizes to any function that is *strictly* monotonous in the distance to the optimum $f(\mathbf{z}) = g(\|\mathbf{z} - \mathbf{z}^*\|)$ (i.e., that has circular contour lines). The next milestone will be to determine whether convergence can be proven for general quadratic functions, or possibly even for all smooth convex functions.

All our results for the convergence or radial NES on sphere functions also carry over to *separable* multivariate Gaussian search distributions (SNES, [13]). This is because SNES can be seen as a collection of d independent radial NES algorithms that optimize parameters in each dimension separately; for each dimension, our convergence results hold separately, so they must hold for the composite case.

7.2 Future work

The presented may lend themselves to an extension for NES variants that adapt the full covariance matrix (e.g., xNES [8]), which must converge if none of the eigenvalues of the normalized covariance matrix vanish, (which holds trivially if $\eta_{\mathbf{B}} = 0$). Other interesting cases are non-Gaussian search distributions; here the full analysis will probably be limited to members of the exponential family (which admit conjugate priors).

A different direction is to develop improved algorithms using well-founded schemes for adapting the learning rates η_σ depending on the current regime (as in [7]), or time-dependent the population sizes n_t, based on the presented analysis.

Acknowledgments

I would like to thank Tobias Glasmachers for providing the motivation for this work, and Claudia Clopath for many helpful comments. I am also grateful to the anonymous reviewers for the careful proofreading.

This work was funded in part through AFR postdoc grant number 2915104, of the National Research Fund Luxembourg.

References

[1] S. I. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10:251–276, 1998.

[2] L. Arnold, A. Auger, N. Hansen, and Y. Ollivier. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. Technical report, June 2011.

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

[4] A. Auger and N. Hansen. Theory of Evolution Strategies: a New Perspective. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, chapter 10, pages 289–325. World Scientific Publishing, 2011.

[5] G. Cuccu, F. Gomez, and T. Glasmachers. Novelty-based restarts for evolution strategies. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 158–163, 2011.

[6] S. Finck and H.-G. Beyer. Performance analysis of the simultaneous perturbation stochastic approximation algorithm on the noisy sphere model. *Theoretical computer science*, 419(C):50–72, Feb. 2012.

[7] N. Fukushima, Y. Nagata, S. Kobayashi, and I. Ono. Proposal of distance-weighted exponential natural evolution strategies. In *2011 IEEE Congress of Evolutionary Computation*, pages 164–171. IEEE, 2011.

[8] T. Glasmachers, T. Schaul, Y. Sun, D. Wierstra, and J. Schmidhuber. Exponential Natural Evolution Strategies. In *Genetic and Evolutionary Computation Conference (GECCO)*, Portland, OR, 2010.

[9] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *IEEE Transactions on Evolutionary Computation*, 9:159–195, 2001.

[10] N. Ikeda and S. Watanabe. Stochastic differential equations and diffusion processes. 1989.

[11] M. Sankaran. Approximations to the non-central chi-square distribution. *Biometrika*, 1963.

[12] T. Schaul. *Studies in Continuous Black-box Optimization*. Ph.D. thesis, Technische Universität München, 2011.

[13] T. Schaul, T. Glasmachers, and J. Schmidhuber. High Dimensions and Heavy Tails for Natural Evolution Strategies. In *Genetic and Evolutionary Computation Conference (GECCO)*, Dublin, Ireland, 2011.

[14] P. K. Sen. The Mean-Median-Mode Inequality and Noncentral Chi Square Distributions. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 51 (1):pp. 106–114, 1989.

[15] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. *Parallel Problem Solving from Nature-PPSN IX*, 2006.

[16] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. Natural Evolution Strategies. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, Hong Kong, China, 2008.