Evolution Strategies are NOT Gradient Followers

Hans-Georg Beyer
Hans-Georg.Beyer@fhv.at
https://homepages.fhv.at/hgb/

Center for Process and Product Engineering
Vorarlberg University of Applied Sciences

FHV

On the Search Behavior of ES in $\mathbb{R}^N$

How does the ES explore the search space?

- often used picture: Population traces the gradient path
- this is based on the following observations
  1. ES exhibits linear convergence order just like classical gradient strategies
  2. Claims in publications:
     - ⋆ "Evolution strategies (ES) can be best described as a gradient descent method which uses gradients estimated from stochastic perturbations around the current parameter value."[1]
     - ⋆ ". . . instead of computing the exact gradient, ES computes an approximation from all the sample points (called pseudo-offspring) generated from parent"[2]

NB: This is due to a misleading statement in a paper by Salimans et al. (2017): Evolution Strategies as a Scalable Alternative to Reinforcement Learning.[3]

[1] https://www.inference.vc/evolutionary-strategies-embarrassingly-parallelizable-optimization/

[2] X. Zhang, J. Clune, and K.O. Stanley: On the Relationship Between the OpenAI EvolutionStrategy and Stochastic Gradient Descent. ArXiv e-prints, abs/1712.06564

[3] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. ArXiv e-prints, abs/1703.03864

Recall: Gradient Strategies

If one wants to minimize a function $f(\mathbf{y}), \mathbf{y} \in \mathbb{R}^N$
Iterative scheme:

$$\mathbf{y}^{(g+1)} = \mathbf{y}^{(g)} - \eta^{(g)} \nabla f(\mathbf{y}^{(g)}) \tag{1}$$

or more general

$$\mathbf{y}^{(g+1)} = \mathbf{y}^{(g)} - \eta^{(g)} \mathbf{C}^{(g)} \nabla f(\mathbf{y}^{(g)}) \tag{2}$$

as long as $\mathbf{C}^{(g)} \in \mathbb{R}^{N \times N}$ is *positive definite*, or even more general

$$\mathbf{y}^{(g+1)} = \mathbf{y}^{(g)} - \tilde{\mathbf{c}}[\nabla f(\mathbf{y}^{(g)}), g] \tag{3}$$

SALIMANS ET AL. used normally distributed mutations $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ and called

$$\mathbf{y}^{(g+1)} = \mathbf{y}^{(g)} - \alpha \sum_{i=1}^{\lambda} f(\mathbf{y}^{(g)} + \mathbf{z}_i)\mathbf{z}_i \tag{4}$$

this update scheme Evolution Strategy (with reference to RECHENBERG)

What is the meaning of $\alpha \sum_{i=1}^{\lambda} f(\mathbf{y} + \mathbf{z}_i)\mathbf{z}_i$?
Since in high-dimensional spaces $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ the length of $\mathbf{z}$ is

$$\mathrm{E}[||\mathbf{z}||] \simeq \sigma\sqrt{N} \tag{5}$$

thus, we have a Monte Carlo estimator of a *surface integral* in $\mathbb{R}^N$

$$\alpha \sum_{i=1}^{\lambda} f(\mathbf{y} + \mathbf{z}_i)\mathbf{z}_i \simeq \oiint_{\partial V} f(\mathbf{y} + \mathbf{x})\, \mathrm{d}\mathbf{A}(\mathbf{x}) \tag{6}$$

Applying Gauss' Theorem: $\oiint_{\partial V} f(\mathbf{x})\, \mathrm{d}\mathbf{A} = \iiint_V \nabla f \, \mathrm{d}V$ and divide by the volume V of the ball and taking the limit $V \to 0$, i.e. $\sigma \to 0$

$$\lim_{V\to 0} \frac{\alpha}{V} \sum_{i=1}^{\lambda} f(\mathbf{y}+\mathbf{z}_i)\mathbf{z}_i \simeq \lim_{V\to 0} \frac{1}{V} \oiint_{\partial V} f(\mathbf{y}+\mathbf{x})\, \mathrm{d}\mathbf{A}(\mathbf{x}) = \lim_{V\to 0} \frac{1}{V} \iiint_V \nabla f \, \mathrm{d}V = \nabla f \tag{7}$$

one recovers the *coordinate-free* definition of the gradient!

☞ SALIMANS ET AL. "Evolution Strategy" is a vanilla gradient strategy!

☞ ... and this is not SALIMANS' ET AL. invention, but was already proposed by R. SALOMON in the late 1990s "Evolutionary Gradient Search" [1]

❸ if one projects N-dimensional individual $\mathbf{y} := (y_1, \ldots, y_N)^{\mathrm{T}}$ into (x_1, x_2)-plane using (RECHENBERG)

$$x_1 := \sqrt{y_1^2 + \cdots + y_{(N/2)}^2}\,, \qquad x_2 := \sqrt{y_{(N/2)+1}^2 + \cdots + y_N^2}, \tag{8}$$

one observes indeed some kind of "gradient diffusion"

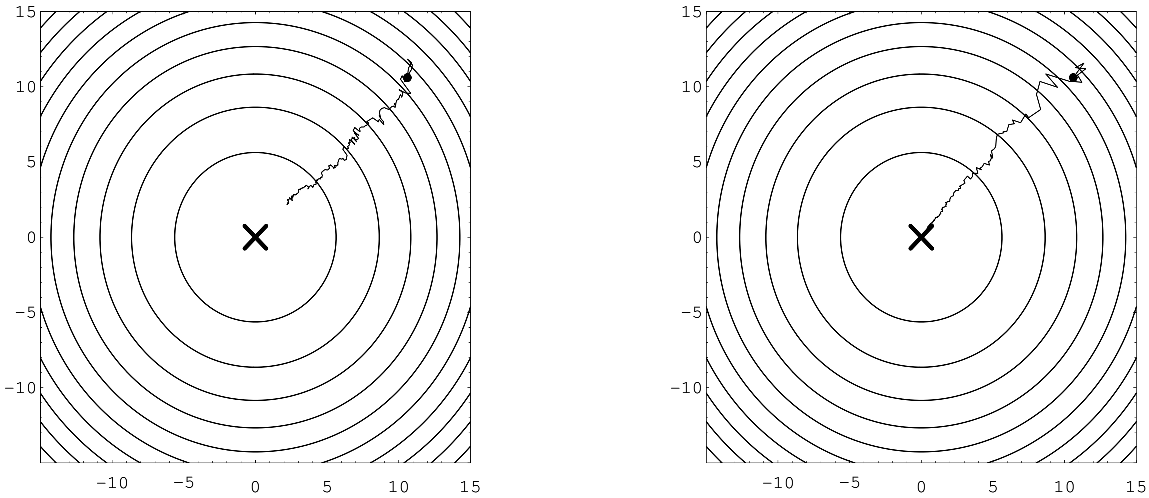

Figure 1: Path of the best individual in a $(4, 20)$-ES (left) and a $(4/4_I, 20)$-ES (right) on the $N = 100$-dimensional sphere model after Projection (8) into 2D over 200 generations. "•": start, "×": optimizer.

- Fig. 1 presents a strong support for the gradient diffusion picture, however

⇒ What would be the use of ES at all?

⇒ probability of leaving local attractors would be very small

⇒ one should better use multi-start gradient strategies

Is this the real picture of the search behavior of ES?

- No, Projection (8) is misleading:
- lumping together $N/2$ components ⇒ central limit theorem of statistics dampens the variance of the random components by a factor of $2/N$
- behavior of single components of the $\mathbf{y}$ vector is not correctly reflected

☞ single components of $\mathbf{y}$ must be considered

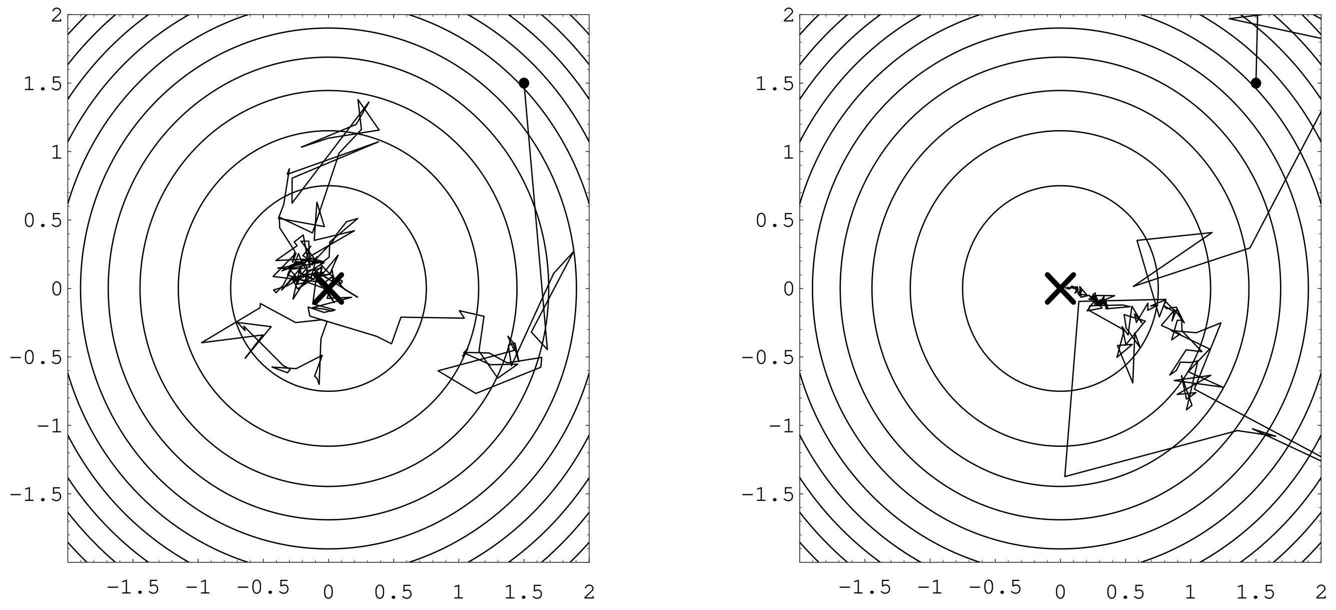

Figure 2: The $x_1 := y_1$ and $x_2 := y_2$ components (x_1 horizontal axis, x_2 vertical axis) of the evolution path of the best individual of the ES runs of Fig. 1, Slide 5 are displayed. Left: $(4, 20)$-ES, right: $(4/4, 20)$-ES.

- actually realized evolution path is much more random as can be seen on Slide 7
- however, this random walk is restricted by selection
- approach to the optimizer $\Leftrightarrow$ EXPLOITATIVE POWER of the EA
- can be described by the Evolutionary Progress Principle (EPP)
- note, concrete form of EPP depends on the definition of "progress"
- however, it is always related to a decomposition of the mutation vector $\mathbf{z}$ or the vector describing the change of the parental centroid from g to $g+1$
- **general observation:**

$$\left\{ \begin{array}{lllll} \text{gain part} & \Leftrightarrow & x\text{-component} & \Leftrightarrow & \text{EXPLOITATION} \\ \text{loss part} & \Leftrightarrow & \mathbf{h}\text{-vector} & \Leftrightarrow & \text{EXPLORATION} \end{array} \right. \tag{9}$$

Q: How to quantify Exploitation/Exploration?

Different options to define the exploitation/exploration ratio

1. decomposition of the expected value of the parental centroid change $\langle \mathbf{y} \rangle^{(g)} - \langle \mathbf{y} \rangle^{(g+1)}$ according to

$$\frac{\text{Exploitation}}{\text{Exploration}} := \frac{\mathrm{E}[R - \tilde{R}]}{\mathrm{E}[\|\mathbf{h}\|]} = \frac{\varphi}{\mathrm{E}[\|\mathbf{h}\|]} \tag{10}$$

2. relating the fictive length of the expected change in local gradient direction to the perpendicular part (perpendicular w.r.t. the local gradient) of the parental centroid change
   - fictive length is also referred to as *normal progress* φ_R

$$\varphi_R = \frac{\overline{Q}}{\|\nabla F(\mathbf{y}_\mathrm{p})\|}, \quad \overline{Q} - \text{QUALITY GAIN}, \quad \mathbf{y}_\mathrm{p} = \langle \mathbf{y} \rangle^{(g)} \tag{11}$$

   - where quality gain is defined by

$$\overline{Q} = \mathrm{E}\left[F\left(\langle \mathbf{y} \rangle^{(g+1)}\right) - F(\mathbf{y}_\mathrm{p})\right] \tag{12}$$

   - and the exploitation/exploration ratio reads

$$\frac{\text{Exploitation}}{\text{Exploration}} := \frac{\varphi_R}{\mathrm{E}[\|\mathbf{h}\|]} \tag{13}$$

$$\begin{cases} \text{progress in local gradient direction} & \Leftrightarrow \text{EXPLOITATION} \\ \text{perpendicular part} & \Leftrightarrow \text{EXPLORATION} \end{cases} \tag{14}$$

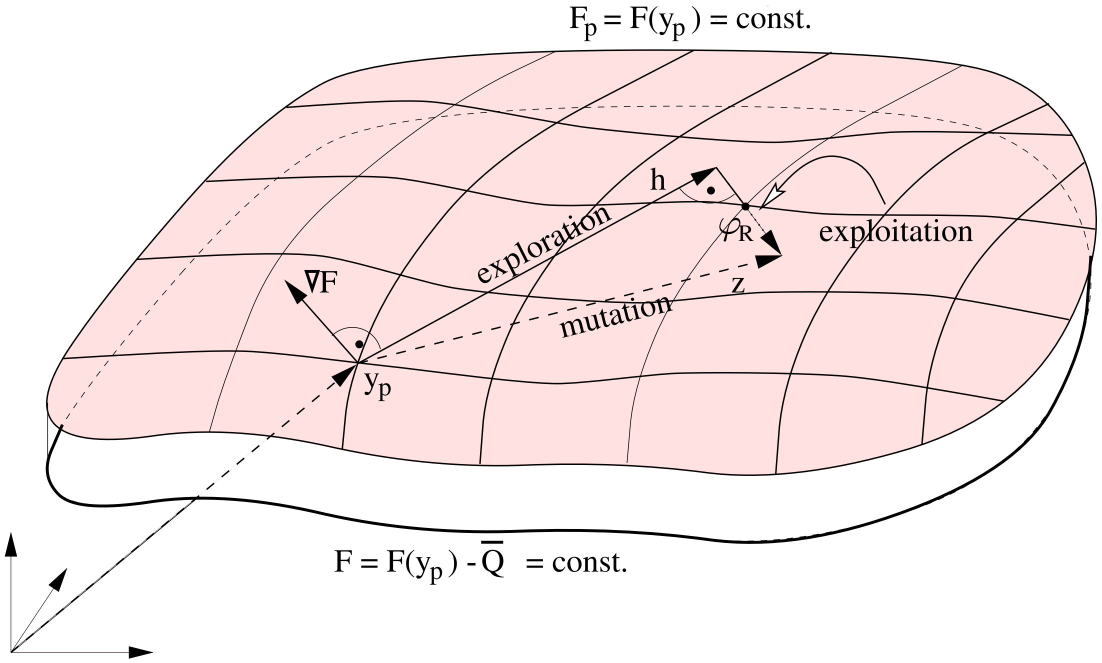

Figure 3: Visualization of exploration vs. exploitation based on normal progress. The surface displayed represents equal function values (i.e., $\mathbf{y} \in \mathbb{R}^3$).

Asymptotic $N \to \infty$ exploration-exploitation behavior (sphere model)

- isotropic Gaussian mutations: $E[\|\mathbf{h}\|] \simeq \sigma\sqrt{N}$
- as for $(\mu/\mu_I, \lambda)$-ES on sphere model, Definition (9) yields

$$\max[\varphi_{\mu/\mu,\lambda}] \simeq \frac{R}{N}\mu\frac{c_{\mu/\mu,\lambda}^2}{2} \quad \Leftrightarrow \quad \sigma = \mu c_{\mu/\mu,\lambda}\frac{R}{N}$$

and

$$\mathrm{E}[\|\mathbf{h}_{\mu/\mu,\lambda}\|] \simeq \frac{R}{N}\mu c_{\mu/\mu,\lambda}\sqrt{N}$$

- thus

$$\boxed{\frac{\text{Exploitation}}{\text{Exploration}} \simeq \left(\frac{1}{\sqrt{N}}\right)} \qquad (15)$$

- this also holds for each single mutation

On the Search Behavior of ES in $\mathbb{R}^N$

First Summary

1. **Exploitation:** ability of an EA to evolve into a desirable progress direction
   ☞ acts locally in one dimension
2. **Exploration:** process that drives the offspring away from the local progress direction
   ☞ *random walk* on an $(N-1)$-dimensional manifold, locally perpendicular to local progress direction
3. actual "path" of the population in search space does *not* follow the local gradient
4. Are ESs path-oriented search methods?
   ☞ Yes, Brownian random path
5. actual "path" of population in search space is reminiscent of *serpentines* in mountainous regions

Mean Value Dynamics of Self-Adaptive ESs

Goals of a theoretical analysis:

- getting a general understanding how Evolution Strategies (ES) do work
- given a objective function model $f(\mathbf{y})$ to be optimized, how fast does the ES approach the optimizer?
- how is the influence of the model parameters (e.g. condition number) on the ES performance?
- not only interested in convergence order, but also in the computational resources needed to get a predefined improvement
- ideally, we want to calculate the dynamics describing the approach towards the optimizer
- getting information how strategy specific parameters (e.g. population size, truncation ratio) influence the performance

Goal Function:

$$f(\mathbf{y}) = \sum_{i=1}^{N} a_i y_i^2, \qquad a_i > 0 \tag{16}$$


```
σ^(0) ← σ_init
y^(0) ← y_init
g ← 0
do
  for l = 1, ..., λ begin
    σ̃_l ← σ^(g) e^{τ N_l(0,1)}
    z_l ← N_l(0, I)
    x_l ← σ̃_l z_l
    ỹ_l ← y^(g) + x_l
    F̃_l ← F(ỹ_l)
  end
  F̃_sort ← sort(F̃_{1...λ})
  σ^(g+1) ← (1/μ) Σ_{m=1}^{μ} σ̃_{m;λ}
  y^(g+1) ← (1/μ) Σ_{m=1}^{μ} ỹ_{m;λ}
  g ← g + 1
until termination
```

Figure 4: The $(\mu/\mu_I, \lambda)$-σSA-ES

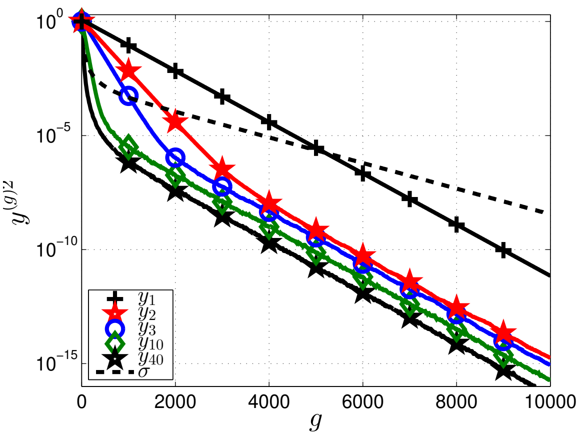

Figure 5: Dynamics of the $(3/3_I, 10)$-ES on a fitness function (16) with $a_i = i$ and $N = 40$. The quadratic deviation of y_i from the optimizer is displayed for the components $i = 1, 2, 3, 10, 40$. Additionally, the mutation strength σ has been plotted. ES learning parameter: $\tau = 1/\sqrt{N}$. Note, the graphs are averages over 1000 independent runs.

- mean value dynamics are described by a system of $N+1$ difference equations:

$$\left(y_i^{(g+1)}\right)^2 = \left(y_i^{(g)}\right)^2 \left(1 - \frac{2c_{\mu/\mu,\lambda}\sigma^{(g)} a_i}{\sqrt{\sum_{j=1}^{N} a_j^2 \left(y_j^{(g)}\right)^2}}\right) + \frac{\left(\sigma^{(g)}\right)^2}{\mu} \tag{17}$$

$$\sigma^{(g+1)} = \sigma^{(g)} \left[1 + \tau^2 \left(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - \frac{c_{\mu/\mu,\lambda}\sigma^{(g)} \sum_{j=1}^{N} a_j}{\sqrt{\sum_{j=1}^{N} a_j^2 \left(y_j^{(g)}\right)^2}}\right)\right] \tag{18}$$

- note this system is *non-linear* and a closed-form solution is excluded
- however, one can derive an *asymptotically* exact solution for $g \to \infty$
- this is also referred to as *steady state* solution:

- the steady state solution reads:

$$\left(y_i^{(g)}\right)^2 = b_i \mathrm{e}^{-\nu g}, \qquad b_i > 0,\ \nu > 0 \tag{19}$$

$$\sigma^{(g)} = \sigma_0 \mathrm{e}^{-\frac{\nu}{2} g}, \qquad \sigma_0 > 0 \tag{20}$$

note, this already implies *linear convergence order*.

- here, $\nu > 0$ is the smallest eigenvalue of the eigenvalue problem (21)

$$\nu b_i = 2\sigma^*_{\mathrm{ss}} c_{\mu/\mu,\lambda} \frac{a_i}{\sum_{j=1}^{N} a_j} b_i - \frac{\left(\sigma^*_{\mathrm{ss}}\right)^2 \sum_{j=1}^{N} a_j^2 b_j}{\mu \left(\sum_{j=1}^{N} a_j\right)^2}, \tag{21}$$

$$\nu = \tau^2 \left(2\sigma^*_{\mathrm{ss}} c_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1\right), \tag{22}$$

and ν, b_i, and $\sigma^*_{\mathrm{ss}} = \sigma_0 \sum_{j=1}^{N} a_j / \sqrt{\sum_{j=1}^{N} a_j^2 b_j}$ are unknowns

- getting a closed form solution for ν is a challenge, however, for $N \to \infty$ one can asymptotically assume $\nu \to 0$

Important Results

- considering the general model case $f(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\mathbf{Q}\mathbf{y}$ and the eigenvalues a_i of $\mathbf{Q}$, one finds

$$\boxed{\nu \simeq 2\sigma^*_{\mathrm{ss}} c_{\mu/\mu,\lambda} \min(a_i)/\mathrm{Tr}[\mathbf{Q}]} \quad (23)$$

- expected running time: How many generations are needed to reduce $f(\mathbf{y})$ by a factor of $2^{-\beta}$?

$$\boxed{G \simeq \frac{\beta \ln(2)}{2\sigma^*_{\mathrm{ss}} c_{\mu/\mu,\lambda}} \frac{\mathrm{Tr}[\mathbf{Q}]}{\min(a_i)}.} \quad (24)$$

- that is, the resources (number of function evaluations) the ES needs is basically determined by the trace of $\mathbf{Q}$ divided by the *smallest* eigenvalue
- steady state σ^*_{ss}:

$$\boxed{\sigma^*_{\mathrm{ss}} \simeq \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \cdot \frac{1}{1 - \min(a_i)/\left(\tau^2 \mathrm{Tr}[\mathbf{Q}]\right)}.} \quad (25)$$

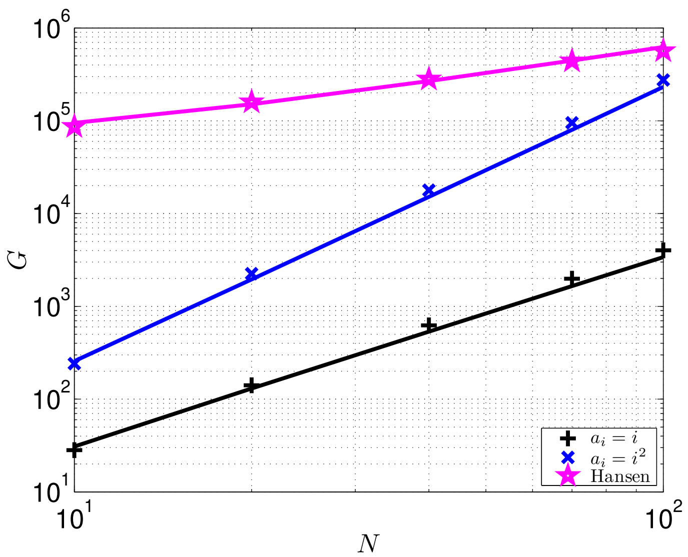

Figure 6: Expected runtime experiments for the $(3/3_I, 10)$-σSA-ES with $\tau = 1/\sqrt{N}$ on the ellipsoid models $a_i = i$, i^2, and Hansen's with $\alpha = 5$. The predictions of (24) for $\beta = 2$ are displayed by curves.

- interestingly, Hansen's f-model $f(\mathbf{y}) := \sum_{i=1}^{N} 10^{\alpha \frac{i-1}{N-1}} y_i^2$ is asymptotically *not harder than the sphere model*, i.e. $G = \mathcal{O}(N)$

ES mean value dynamics *does not* follow the gradient of $f(\mathbf{y})$

- coming back to the claim that ES follows the gradient path (on average)
- this would mean that it mimics a classical gradient strategy
- however, look at (19), this is not the case:

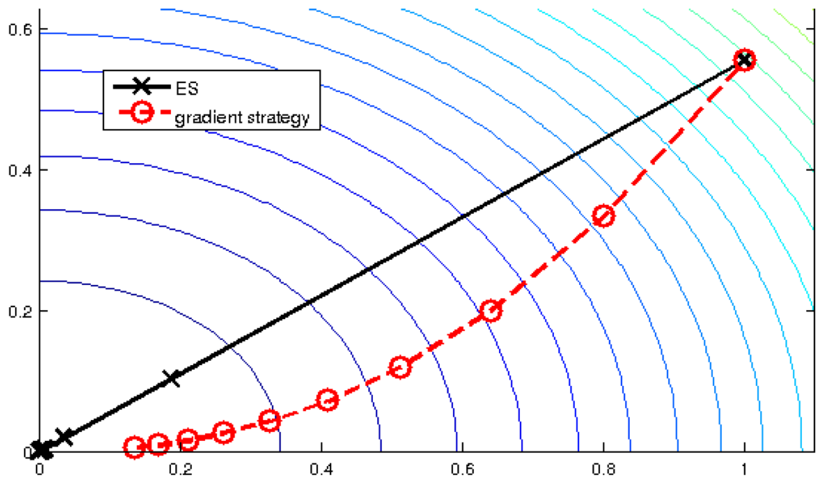

Figure 7: In the steady state, the ES follows in expectation a straight line towards the optimizer when applied to quadratic objective functions.

Summary

- not all ESs labeled as ES are ESs
- using an inappropriate visualization may lead to wrong conclusions
- regarding the search behavior of ES, one has to look at the actual search paths
- these search paths are more like restricted random walks than gradient descents/ascents
- one may consider this locally as an exploration process in $N-1$ dimensions and an exploitation in one dimension
- the search path of ES resembles serpentine paths in mountain regions
- even if one considers the mean value dynamics, the ES does not approximate the gradient path, except for the sphere
- in the steady state, the ES approximates on average the Newton-direction even though only isotropic mutations are used
- **not considered:** When does a gradient strategy behave like an ES?

The End

?

R. Salomon.
Evolutionary Search and Gradient Search: Similarities and Differences.
IEEE Transactions on Evolutionary Computation, 2(2):45–55, 1998.

H.-G. Beyer.
On the "Explorative Power" of ES/EP-like Algorithms.
In V.W. Porto, N. Saravanan, D. Waagen, and A.E. Eiben, editors, *Evolutionary Programming VII: Proceedings of the Seventh Annual Conference on Evolutionary Programming*, pages 323–334, Heidelberg, 1998. Springer-Verlag.
DOI: 10.1007/BFB0040785.

H.-G. Beyer and A. Melkozerov.
The Dynamics of Self-Adaptive Multi-Recombinant Evolution Strategies on the General Ellipsoid Model.
IEEE Transactions on Evolutionary Computation, 18(5):764–778, 2014.
DOI: 10.1109/TEVC.2013.2283968.