

Step Length on Linear Fitness Functions

Self-Adaptation and Beyond

Nikolaus Hansen

`Nikolaus.Hansen@inf.ethz.ch`

Outline

- Self-Adaptation (SA)
- $(1, 2)$ - σ -SA-ES on f_{linear}
- (μ, λ) - σ -SA-ES on f_{linear}
- Beyond σ -self-adaptation: population diversity on f_{linear}

Preliminaries

(Affine) Linear Fitness Function

- Example: $f_1 : \mathbb{R}^n \rightarrow \mathbb{R}$, $f_1(\mathbf{x}) = x_1$ to be maximized
- **General** (any affine linear transformation of f_1):

$$f_{\text{linear}} : \mathbb{R}^n \rightarrow \mathbb{R}$$
$$\mathbf{x} \mapsto f_0 + \langle \mathbf{v}, \mathbf{x} \rangle = f_0 + \sum_{i=1}^n v_i x_i$$

where the constants $f_0 \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$, and $\mathbf{v} \neq \mathbf{0}$.

- Even more general: $g \circ f_{\text{linear}} : \mathbf{x} \mapsto g(f_{\text{linear}}(\mathbf{x}))$, where $g : \mathbb{R} \rightarrow \mathbb{R}$ is strictly monotonic.

Optimal step length on f_{linear} :

the larger the better, in particular in gradient direction

The Principle of Self-Adaptation

Given: strategy parameter $\theta^{(0)}$ and search point $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

New individuals at generation g obey for $k = 1, \dots, \lambda$

$$\begin{aligned}\theta_k^{(g+1)} &= \theta^{(g)} + Y_k(\tau) \\ \mathbf{x}_k^{(g+1)} &= \mathbf{x}^{(g)} + \gamma\left(\theta_k^{(g+1)}\right) \mathcal{N}_k(\mathbf{0}, \mathbf{I}) \quad ,\end{aligned}$$

where $E\{Y_k(\tau)\} = 0$ (i.e. θ is *unbiased*), $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gauss-distributed random vector, γ is some function.

Truncation selection yields (in case of $(1, \lambda)$ -selection)

$$\begin{aligned}\theta^{(g+1)} &= \theta_{1;\lambda}^{(g)} \\ \mathbf{x}^{(g+1)} &= \mathbf{x}_{1;\lambda}^{(g)}\end{aligned}$$

σ -Self-Adaptation

Given: strategy parameter $\theta^{(0)}$ and search point $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

New individuals at generation g obey for $k = 1, \dots, \lambda$

$$\begin{aligned}\theta_k^{(g+1)} &= \theta^{(g)} + Y_k(\tau) \\ \mathbf{x}_k^{(g+1)} &= \mathbf{x}^{(g)} + \gamma\left(\theta_k^{(g+1)}\right) \mathcal{N}_k(\mathbf{0}, \mathbf{I}) \quad ,\end{aligned}$$

where $\mathbb{E}\{Y_k(\tau)\} = 0$ (i.e. θ is *unbiased*), $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gauss-distributed random vector, γ is some function.

Using $\gamma \equiv \exp$, where $\sigma \equiv \gamma(\theta) \in \mathbb{R}$, yields σ -self-adaptation.

σ -Self-Adaptation

Given: step-size $\sigma^{(0)} \in \mathbb{R}_+$ and search point $\mathbf{x}^{(0)} \in \mathbb{R}^n$.

New individuals at generation g obey for $k = 1, \dots, \lambda$

$$\begin{aligned}\ln \sigma_k^{(g+1)} &= \ln \sigma^{(g)} + Y_k(\tau) \\ \mathbf{x}_k^{(g+1)} &= \mathbf{x}^{(g)} + \sigma_k^{(g+1)} \mathcal{N}_k(\mathbf{0}, \mathbf{I}) \quad ,\end{aligned}$$

where $\mathbb{E}\{Y_k(\tau)\} = 0$, and often $Y_k \sim \mathcal{N}(0, \tau^2)$.

Unbiased We call σ *unbiased*, if there exists a function γ such that $\mathbb{E}\{\gamma(\sigma^{(g+1)}) | \sigma^{(g)}\} = \gamma(\sigma^{(g)})$.

Scale-invariance This formulation leads to a scale-invariant algorithm (not just a coincidence).

The analysis becomes in particular independent of $\sigma^{(0)}$.

Invariance

- results are invariant under certain transformations
- well-defined notion of **generalization** and robustness
- a conceptual view onto **strategy parameter adaptation** is to introduce new invariance properties (while preserving existing invariances)

(1, 2)- σ -Self-Adaptation on f_{linear}

Theorem 1 (σ -distribution) *On a linear fitness function the σ -distribution of the (1, 2)- σ -SA-ES is identical before and after selection, that is*

$$\sigma^{(g+1)} = \sigma_{1;2}^{(g+1)} \sim \sigma_{2;2}^{(g+1)} \sim \sigma_1^{(g+1)} \sim \sigma_2^{(g+1)}$$

Remark 1 *The same holds under random selection.*

Proof We will consider three cases.

1. Case ($p = 0.5$)

$$f(\mathbf{x}_1^{(g+1)}) > f(\mathbf{x}^{(g)}) > f(\mathbf{x}_1^{(g+1)}) \text{ or}$$

$$f(\mathbf{x}_2^{(g+1)}) > f(\mathbf{x}^{(g)}) > f(\mathbf{x}_2^{(g+1)})$$

does not influence the σ -distribution

2. Case ($p = 0.25$)

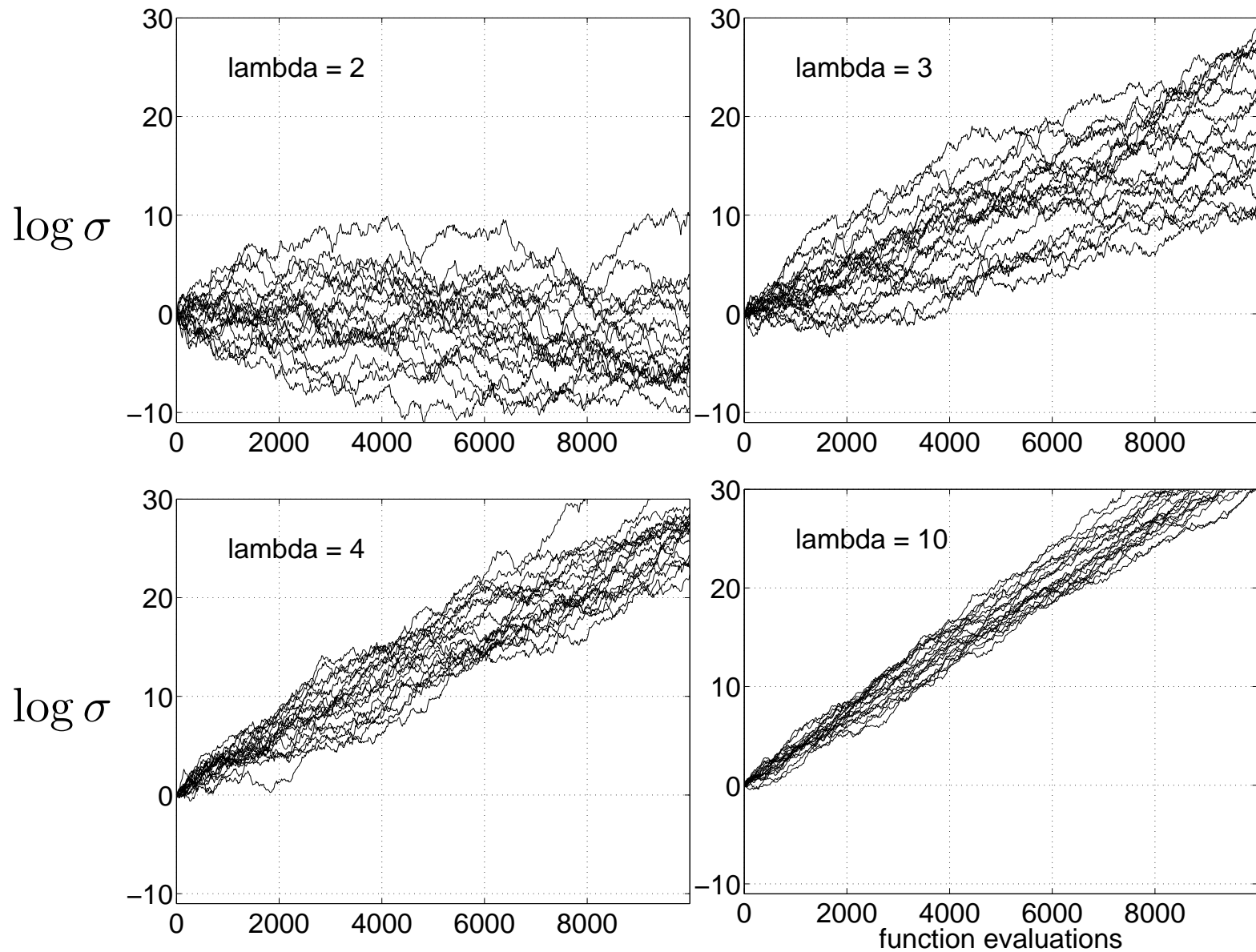
$$f(\mathbf{x}_1^{(g+1)}) > f(\mathbf{x}^{(g)}) \text{ and } f(\mathbf{x}_2^{(g+1)}) > f(\mathbf{x}^{(g)})$$

3. Case ($p = 0.25$)

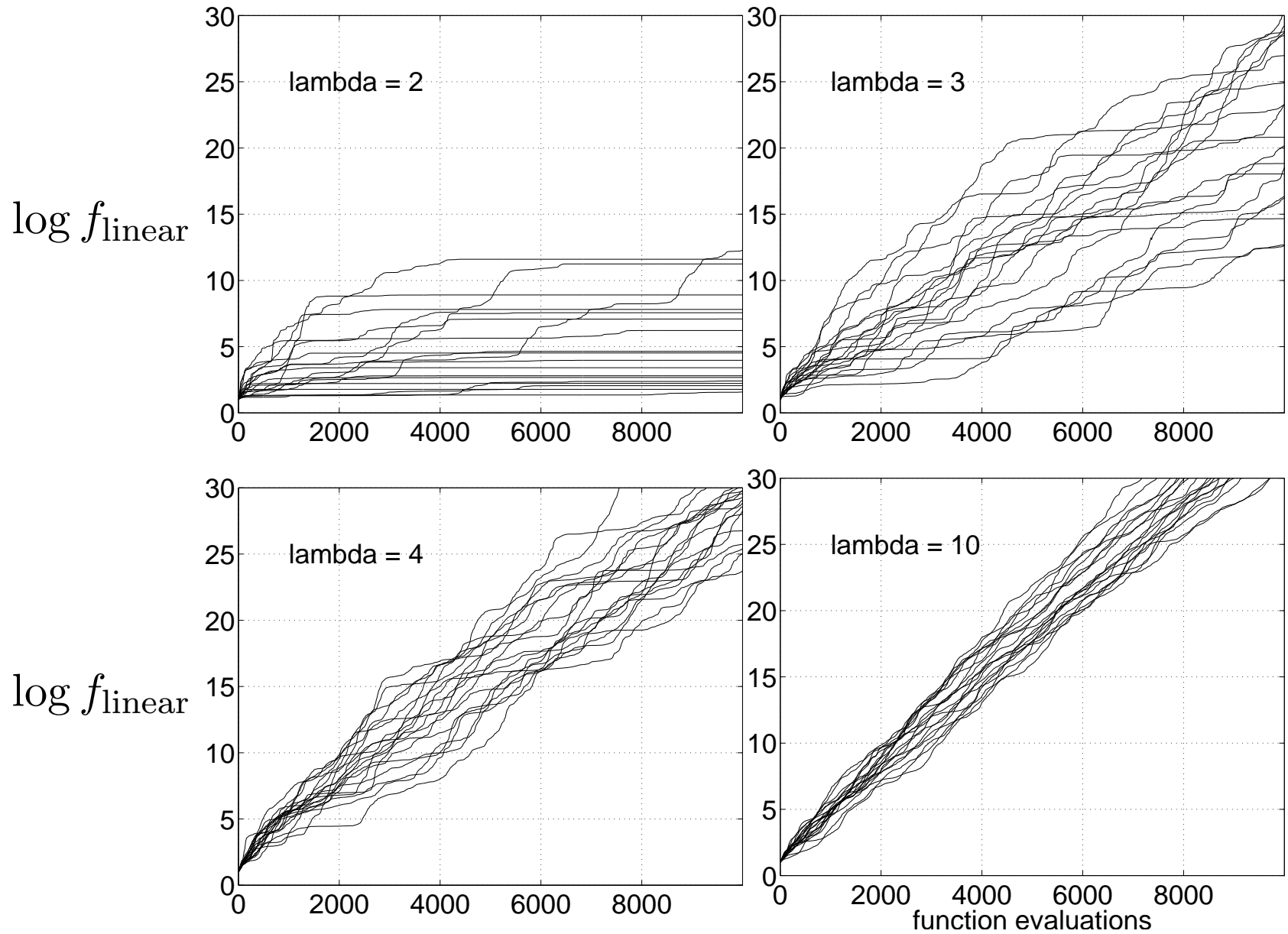
$$f(\mathbf{x}_1^{(g+1)}) < f(\mathbf{x}^{(g)}) \text{ and } f(\mathbf{x}_2^{(g+1)}) < f(\mathbf{x}^{(g)})$$

Case 2 and 3 are complementary events with respect to σ . □

Evolution of σ in the $(1, \lambda)$ -ES on f_{linear}



Evolution of the Fitness on f_{linear}



(1, 2)- σ -Self-Adaptation on f_{linear}

Corollary 1 (σ -stationarity) For the (1, 2)- σ -SA-ES *step-size σ is unbiased* on a linear fitness function, that is

$$\mathbb{E} \left\{ \ln \sigma^{(g+1)} \mid \sigma^{(g)} \right\} = \ln \sigma^{(g)}$$

Proof The step-size obeys

$$\ln \sigma_k^{(g+1)} = \ln \sigma^{(g)} + Y_k(\tau)$$

and we have $\sigma^{(g+1)} \sim \sigma_k^{(g+1)}$, and $\mathbb{E}\{Y_k\} = 0$. □

(1, 2)- σ -Self-Adaptation on f_{linear}

Corollary 2 (exponential increase of σ in mean) *For unbiased σ , where $\mathbb{E}\{\ln \sigma^{(g+1)} | \sigma^{(g)}\} = \ln \sigma^{(g)}$, holds $\mathbb{E}\{(\sigma^{(g+1)})^\alpha | \sigma^{(g)}\} > (\sigma^{(g)})^\alpha$ for all $\alpha > 0$, that is, the population variance increases exponentially in mean.*

Proof With Jensen's inequality we have

$$\begin{aligned}\mathbb{E}\left\{\left(\sigma^{(g+1)}\right)^\alpha\right\} &= \exp \ln \mathbb{E}\left\{\left(\sigma^{(g+1)}\right)^\alpha\right\} \\ &> \exp \mathbb{E}\left\{\ln \left(\sigma^{(g+1)}\right)^\alpha\right\} \\ &= \exp \alpha \ln \sigma^{(g)} \\ &= \left(\sigma^{(g)}\right)^\alpha\end{aligned}$$

Remark 2 *The corollaries hold under random selection as well.* □

Exponential increase of the population variance (or step length, in expectation) is not a sufficient demand on f_{linear} .

Postulate 1 *On f_{linear} , an evolutionary algorithm should increase the expected logarithm of step length in gradient direction linearly in time.*

Emerging Questions

- only valid for (1, 2)-selection?
- only valid for σ -self-adaptation?

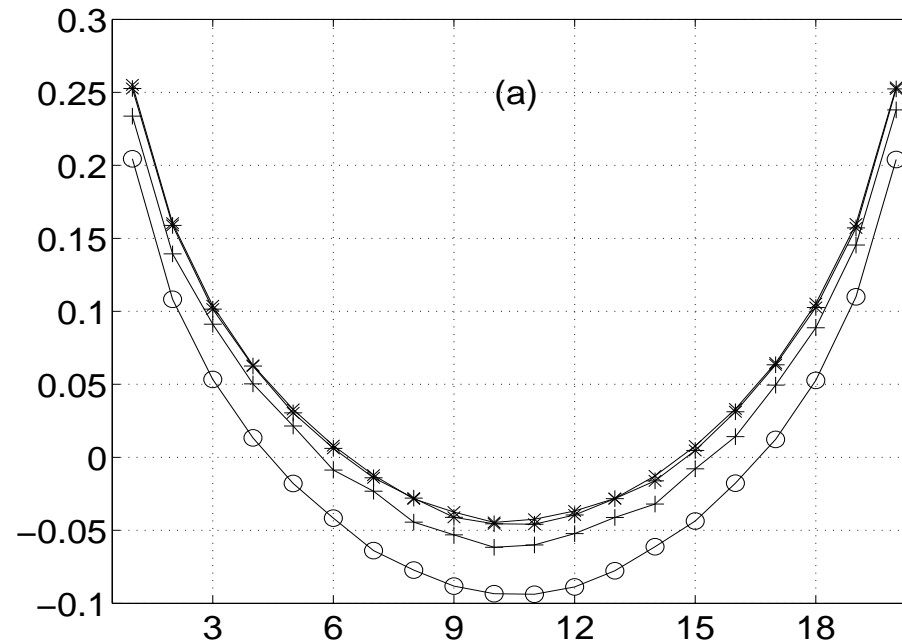
Beyond (1, 2)-Selection

Again, we consider σ -self-adaptation in an evolutionary algorithm.

Theorem 2 *Given a point symmetrical distribution of the offspring population, the i -th best and i -th worst individual have the same σ -distribution on f_{linear} .*

Proof Selection on f_{linear} yields the same σ -distribution as selection on $-f_{\text{linear}}$, because the offspring population is symmetrical. The i -th best individual on f_{linear} corresponds to the i -th worst individual on $-f_{\text{linear}}$. □

Simulation of $E\{\log \sigma\}$



Expected **logarithmic step-size change**, i.e., $E\{\ln \sigma^{(g+1)} | \sigma^{(g)}\} - \text{mean} \ln \sigma^{(g)}$
 $\text{mean}_{1e3 \leq g < 5e4} (\log(\sigma_{i:20}^{(g+1)}) - \text{mean}_{j \in \{I_{\text{sel}}\}} (\log_{10}(\sigma_j^{(g)})))$
versus fitness rank of descendants from left (best) to right (worst) on f_{linear} .

(a): $(\mu/\mu_I, 20)$ -ES with arithmetic recombination for σ .

Results given for $\mu = 1; 5; 10; 15$ (○; +; *; ×), where problem dimension $n = 2$.

The data are in accordance with Theorem 2

Implications

- Increase of σ between two generations can be related to a **bias of the recombination operator** on σ and can explain divergent behavior as observed e.g. by Kursawe (1995), Arnold and Beyer (2000), Beyer and Sendhoff (2005).

Theorem 3 (($\mu, 2\mu$)-ES) *Given a **point symmetrically distributed offspring population** and **unbiased** mutation and recombination of the step-sizes σ , then the ($\mu, 2\mu$)- σ -SA-ES leaves the step-sizes unbiased.*

Presumably this means for $\mu > \lambda/2$ that the step size decrease.

... **but** ...

violation of one of the assumptions lead to reasonable
strategy behavior on f_{linear}

... and Beyond σ -Self-Adaptation

We consider selection on f_{linear} regardless of how the individuals were generated.

Theorem 4 *Given a point symmetrical distribution of the offspring population, the i -th best and the i -th worst individual have the same **step length** distribution on f_{linear} .*

Step length refers to the distance to the symmetry point.

Remarks and Consequences

- (Distribution of steps) only selecting **less than half** of a point symmetrically distributed offspring population **increases** the step lengths compared to the original population.
- Selection on f_{linear} *always* **decreases the within-population variance** in gradient direction.

“Practical” implication:

Postulate 2 *The step lengths in gradient direction on f_{linear} should **increase faster** than step lengths under random selection.*

Summary

On f_{linear}

- The $(1, 2)$ - σ -self-adaptation-ES yields an unbiased random walk of $\ln \sigma^{(g)}$.
- Given a symmetric population distribution and unbiased operators, ^a the $(\mu, 2\mu)$ - σ -self-adaptation-ES yields $\mathbb{E}\{\ln \sigma^{(g+1)} | \sigma^{(g)}\} = \ln \sigma^{(g)}$.
- In general, given a symmetrical population distribution, the step-length distribution of i -th best and i -th worst individual are identical.
- We gave **two postulates** on the step lengths (or population variance) that can easily be verified (empirically) for any algorithm.

^awhich is typically the case for the mutation operator

Thank you