An Overview of Evolution Strategies

Prof. Dr. Thomas Bäck

NuTech Solutions GmbH · Martin-Schmeisser-Weg 15
D-44227 Dortmund

www.nutechsolutions.de
baeck@nutechsolutions.de

Tel: +49 (0) 231 72 54 63 10
Fax: +49 (0) 231 72 54 63 29

Overview

- Basic introduction:
  Representation, strategy parameters, mutation, recombination, selection.
- Self-adaptation of strategy parameters.
- The (1+1)-evolution strategy.
- Convergence velocity theory.
- Application examples.

Main Characteristics (at a glance)

- Often continuous search spaces, $\mathbb{R}^n$.
- Emphasis on mutation: n-dimensionally normally distributed, expectation zero.
- Various recombination operators.
- Deterministic (μ,λ)-selection.
- *Self-adaptation* of strategy parameters: First self-adaptive EA.
- Generation of an offspring surplus $\lambda \gg \mu$.

Representation (1)

Spaces:

- Phenotype space:

$$\mathbb{R}^n$$

- Strategy parameter space (standard deviations and rotation angles of mutation):

$$\mathcal{S} = \mathbb{R}_+^{n_\sigma} \times [-\pi, \pi]^{n_\alpha}$$

- Individual space (genotype):

$$I = \mathbb{R}^n \times \mathcal{S}$$

One individual:

$$\vec{a} = (\underbrace{(x_1, \ldots, x_n)}_{\vec{x}}, \underbrace{(\sigma_1, \ldots, \sigma_{n_\sigma})}_{\vec{\sigma}}, \underbrace{(\alpha_1, \ldots, \alpha_{n_\alpha})}_{\vec{\alpha}}) \in I$$

Representation (2)

The three parts of an individual:

$\vec{x}$	:	Object variables	$\Rightarrow$	Fitness $f(\vec{x})$
$\vec{\sigma}$	:	Standard deviations	$\Rightarrow$	Variances
$\vec{\alpha}$	:	Rotation angles	$\Rightarrow$	Covariances

A strategy parameter set ($s = (\vec{\sigma}, \vec{\alpha}) \in \mathcal{S}$):

- Is part of an individual.
- Represents the probability density function (p.d.f.) for its mutation.

n_σ	n_α	Remark
1	0	standard mutation
n	0	standard mutations
n	$n \cdot (n-1)/2$	correlated mutations
$1 \le n_\sigma \le n$	$(n - \frac{n_\sigma}{2})(n_\sigma - 1)$	general case (correlated mutations)

Possible settings of n_σ and n_α.

Genetic operators: mutation (1)

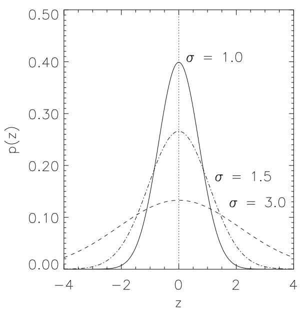
Simple mutation I

- Simple mutation makes use of normally distributed variations, $N(\xi, \sigma)$.

$$p(\Delta x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\Delta x_i - \xi)^2}{2\sigma^2}\right)$$

- Expectation (ξ) is assumed to equal 0.
- Standard deviation (σ) must be adapted.

The one-dimensional case.



Genetic operators: mutation (2)

Simple mutation I (continued)

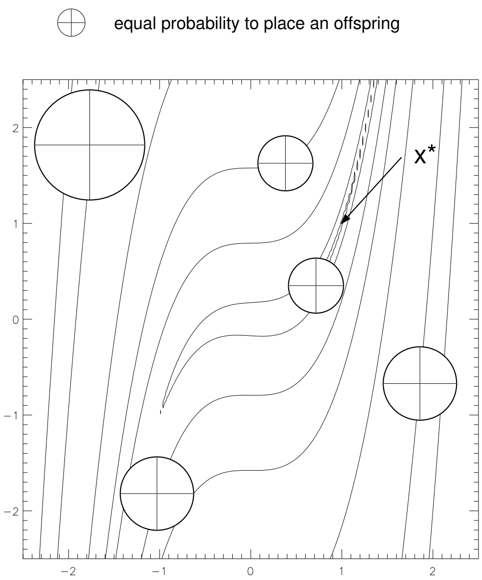
- x_i is mutated by adding some Δx_i from a normal probability distribution.
- σ is mutated by multiplying by e^Γ, with Γ from a normal probability distribution.

$$\begin{aligned} I &= \mathbb{R}^n \times \mathbb{R}_+ \\ m'_{\{\tau_0\}}(\vec{x}, \sigma) &= (\vec{x}', \sigma') \\ \tau_0 &\sim 1/\sqrt{n} \end{aligned}$$

$$\begin{aligned} \sigma' &= \sigma \cdot \exp(\tau_0 \cdot N(0,1)) \\ x'_i &= x_i + \sigma' \cdot N_i(0,1) \end{aligned}$$

Genetic operators: mutation (3)

Simple mutation I (continued)



Simple mutations, $n = 2$, $n_\sigma = 1$, ($\Rightarrow n_\alpha = 0$).

Genetic operators: mutation (4)

Simple mutation II

- x_i is mutated by adding some Δx_i from a normal probability distribution.
- σ_j is mutated by multiplying by e^{Γ_j} with Γ_j from a normal probability distribution.

$$\begin{array}{lcl} I & = & \mathrm{IR}^n \times \mathrm{IR}^n_+ \\ m'_{\{\tau,\tau'\}}(\vec{x},\vec{\sigma}) & = & (\vec{x}',\vec{\sigma}') \\ \tau & \sim & 1/\sqrt{2\sqrt{n}} \\ \tau' & \sim & 1/\sqrt{2n} \end{array}$$

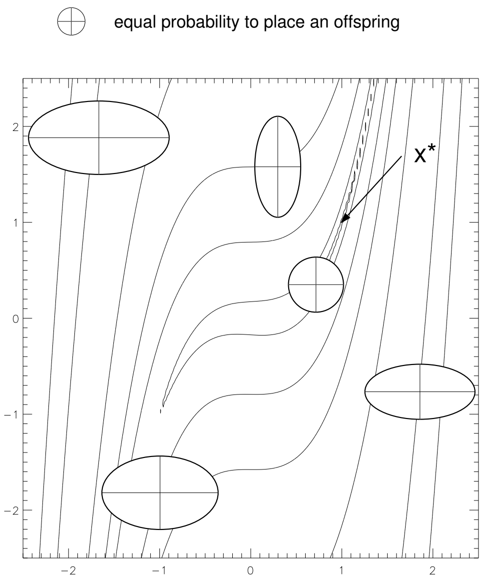
$$\begin{array}{lcl} \sigma'_i & = & \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1)) \\ x'_i & = & x_i + \sigma'_i \cdot N_i(0,1) \end{array}$$

Boundary rule for preserving standard deviations larger than zero:

$$\sigma'_i < \varepsilon_\sigma \quad \Rightarrow \quad \sigma'_i := \varepsilon_\sigma$$

Genetic operators: mutation (5)

Simple mutation II (continued)



Simple mutations, $n = 2$, $n_\sigma = 2$, $(n_\alpha = 0)$.

Genetic operators: mutation (6)

Correlated mutation

- Correlated mutation uses the following probability distribution function for $\overline{\Delta x}$:

$$p(\overline{\Delta x}) = \sqrt{\frac{\det C}{(2\pi)^n}} \cdot \exp\left(-\frac{1}{2}\overline{\Delta x}^T \cdot C\overline{\Delta x}\right)$$

- Where C^{-1} is the covariance matrix:

$$\begin{array}{lcl} c_{ii} & = & \sigma_i^2 \\ c_{ij,(i\neq j)} & = & \left\{ \begin{array}{ll} 0 & \text{no correlations} \\ \frac{1}{2}(\sigma_i^2 - \sigma_j^2)\tan(2\alpha_{ij}) & \text{correlations} \end{array} \right. \end{array}$$

Genetic operators: mutation (7)

Correlated mutation (continued)

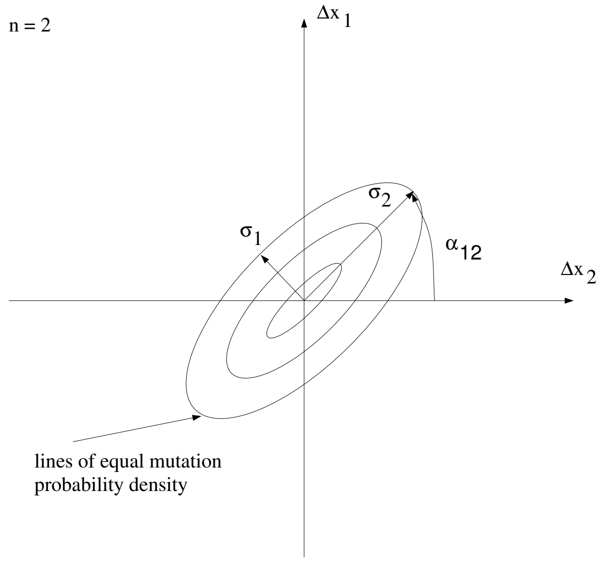


Illustration of the mutation ellipsoid for the case $n = 2$, $n_\sigma = 2$, $n_\alpha = 1$.

Genetic operators: mutation (8)

Correlated mutation (continued)

- $\overline{x}$ is mutated by adding some $\overline{\Delta_x}$ from an n-dimensional normal distribution.
- σ_i is mutated by multiplying by e^{Γ_i} with Γ_i from a normal probability distribution.
- α_j is mutated by adding some Δa_j from a normal probability distribution.

Genetic operators: mutation (9)

Correlated mutation (continued)

$$
\begin{array}{lcl}
n_\alpha & = & n \cdot (n-1)/2 \\
I & = & \mathrm{IR}^n \times \mathrm{IR}_+^n \times [-\pi, \pi]^{n_\alpha} \\
m'_{\{\tau, \tau' \beta\}}(\vec{x}, \vec{\sigma}, \vec{\alpha}) & = & (\vec{x}', \vec{\sigma}', \vec{\alpha}') \\
\tau & \sim & 1/\sqrt{2\sqrt{n}} \\
\tau' & \sim & 1/\sqrt{2n} \\
\beta & \approx & 5^\circ
\end{array}
$$

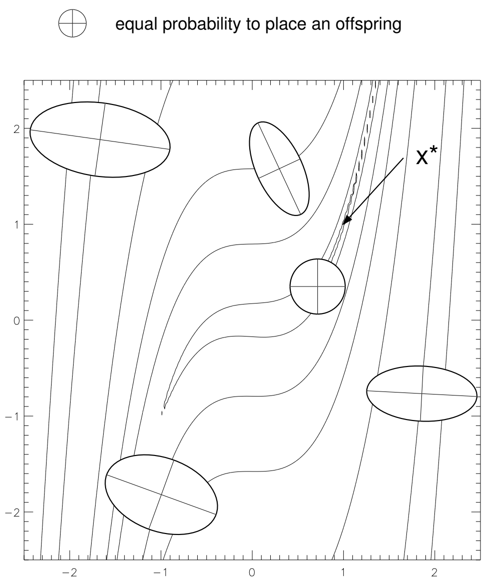
$$
\begin{array}{lcl}
\sigma'_i & = & \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1)) \\
\alpha'_j & = & \alpha_j + \beta \cdot N_j(0,1) \\
\vec{x}' & = & \vec{x} + \vec{N}(\vec{0}, C')
\end{array}
$$

Boundary rule for keeping rotation angles feasible:

$$|\alpha'_j| > \pi \quad \Rightarrow \quad \alpha'_j := \alpha'_j - 2\pi \cdot \mathrm{sign}(\alpha'_j)$$

Genetic operators: mutation (10)

Correlated mutation (continued)



Correlated mutations, $n = 2$, $n_\sigma = 2$, $n_\alpha = 1$.

Genetic operators: mutation (11)

Some remarks:

- Biological model:
  Repair enzymes, mutator genes.
- No deterministic control:
  strategy parameters evolve.
- Indirect link between fitness and useful strategy parameter settings.
- $\vec{\sigma}$, $\vec{\alpha}$ are conceivable as an *internal model* of the local topology.

Genetic operators: mutation (12)

Some remarks (continued):

- Standard strategy: $n_\sigma = n$, $n_\alpha = 0$.
- For correlated mutations:
  - $\vec{\sigma}_c \sim \vec{N}(\vec{0}, C)$ is generated by a multiplication of the uncorrelated random vector $\vec{\sigma}_u$ by n_α rotation matrices (Schwefel 1981, Rudolph 1992).

$$\vec{\sigma}_c = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} R(\alpha_{ij}) \cdot \vec{\sigma}_u \quad .$$

  - Exactly the feasible (positive definite) correlation matrices C can be created this way (Rudolph 1992).

Genetic operators: mutation (13)

The structure of a single rotation matrix:

$$R(\alpha_{ij}) = \begin{pmatrix} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\alpha_{ij} & & & -\sin\alpha_{ij} & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & 1 & & & \\ & & & \sin\alpha_{ij} & & & \cos\alpha_{ij} & & \\ & & & & & & & 1 & \\ 0 & & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix}$$

Genetic operators: mutation (14)

The correlated mutation procedure (Algorithm):

$n_q := n \cdot (n-1)/2;$
for $i := 1$ **to** n **do**
    $\sigma_{u,i} := \sigma_i \cdot N_i(0,1);$
od
for $k := 1$ **to** $n-1$ **do**
    $n_1 := n - k;$
    $n_2 := n;$
        for $i := 1$ **to** k **do**
            $d_1 := \sigma_{u,n_1};$
            $d_2 := \sigma_{u,n_2};$
            $\sigma_{u,n_2} := d_1 \cdot \sin(\alpha_{n_q}) + d_2 \cdot \cos(\alpha_{n_q});$
            $\sigma_{u,n_1} := d_1 \cdot \cos(\alpha_{n_q}) - d_2 \cdot \sin(\alpha_{n_q});$
            $n_2 := n_2 - 1;$
            $n_q := n_q - 1;$
        od
od

Genetic operators: mutation (15)

Generating normally distributed random numbers:

$$\begin{aligned} u &= 2 \cdot U[0,1) - 1 \\ v &= 2 \cdot U[0,1) - 1 \\ w &= u^2 + v^2 \\ x_1 &= u \cdot \sqrt{\frac{-2\log(w)}{w}} \quad , \text{ if } w > 1 \\ x_2 &= v \cdot \sqrt{\frac{-2\log(w)}{w}} \quad , \text{ if } w > 1 \end{aligned}$$

$U[0,1)$ denotes a uniform random number.

Then: $x_1, x_2 \sim N(0,1)$

Genetic operators: recombination (1)

Basic ideas:

- $I^{\mu} \rightarrow I$, μ parents yield 1 offspring.
- Is applied λ times, typically $\lambda \gg \mu$.
- Is applied to object variables as well as strategy parameters.
- Per offspring gene two corresponding parent genes are involved.
- Two ways to recombine two parent alleles:
  - Discrete recombination: choose one randomly.
  - Intermediate recombination: average the values.
- Might involve two or μ parents (global recombination).

Genetic operators: recombination (2)

The operator:

1. For each object variable:
   (a) Choose two parents.
   (b) Apply discrete recombination on the corresponding variables.
2. For each strategy parameter:
   (a) Choose two parents.
   (b) Apply intermediate recombination on the corresponding parameters.

Genetic operators: recombination (3)

Recombination illustrated

1.2	-2.4	0.56	8.7	0.3	0.01	0.4	2.4	Parent 1
-8.2	0.2	-6.7	2.3	0.8	1.8	2.9	20	Parent 2
1.2	0.2	-6.7	2.3	0.55	0.905	1.65	11.2	Offspring
discrete				intermediate				

Discrete recombination on x_i, intermediate on σ_i.

Selection (1)

- Strictly deterministic, rank-based.
- The μ best ranks are handled equally.
- The μ best offspring ($P''(t)$) survive.
  - Important for self-adaptation.
  - Applicable also for noisy objective functions, moving optima.
- N.B. μ selected from λ; notation: (μ, λ).
- Selective pressure: very high.

Selection (2)

Takeover time τ^*:

Definition:
number of generations until repeated application of selection completely fills the population with copies of the best individual.

Remarks:

- Goldberg and Deb 1991:

$$\boxed{\tau^* = \frac{\ln \lambda}{\ln(\lambda/\mu)}}$$

- $\tau^* \approx 2$ generations for a (15,100)-ES
(15 and 100 are typical values for the standard ES).

- Proportional selection in GAs:
$\tau^* \approx \lambda \ln \lambda = 460$ generations!

Other components

- Initialization:
  - x_i, α_i: randomly
  - σ_i: $\delta x_i/\sqrt{n}$, with δx_i a very rough measure for the distance to the optimum.
- Termination:
  - Termination after a number of generations.
  - Or iff $\max\{f(\vec{x}_i(t))\} - \min\{f(\vec{x}_i(t))\} \leq c(P(t))$.
    * $c(P(t))$ absolute ($= \varepsilon_1 > 0$), or
    * $c(P(t))$ relative ($= \varepsilon_2 \cdot |\bar{f}|$)).
- Constraints:
  - Handled by repeating creation and evaluation of individuals.

Reproduction Cycle

Generational ES model:

```
t := 0;
initialize P(t);
evaluate P(t);
while not terminate do
      P'(t) := recombine (P(t));
      P''(t) := mutate(P'(t));
      evaluate(P''(t));
      P(t + 1) := select(P''(t) ∪ P(t));
//    P(t + 1) := select(P''(t));
      t := t + 1;
od
```

- *recombine*: Recombination applied to all individuals.

- $P'(t)$ has size $\lambda > \mu$, $P(t)$ size μ.

- *mutate:* Normally distributed variations, all individuals.

- *select:* (μ+λ) or (μ,λ).

Self-adaptation principles

- Biological model: Repair enzymes, mutator genes.
- No deterministic control: strategy parameters *evolve*.
- *Indirect* link between fitness and useful strategy parameter settings.
- Strategy parameters are conceivable as an *internal model* of the local topology.
- Individual space:

$$I = M \times \mathcal{S}$$

  - M: Search space.
  - $\mathcal{S}$: Strategy parameter space.

The crucial claim (Schwefel 1987, 1992):

Self-adaptation of strategy parameters works

- Without exogenous control.
- By recombining/mutating the strategy parameters.
- By exploiting the implicit link between fitness and useful internal model.

Necessary conditions (found by experiments):

- Generation of a surplus, $\lambda > \mu$
- (μ, λ)-selection (to guarantee extinction of misadapted individuals.
- A not too strong selective pressure e.g., (15,100) where $\lambda/\mu \approx 7$, but clearly $\mu > 1$ is neccesary.
- Recombination also on strategy parameters (especially: intermediate recombination).

Empirical Test Design

- With simple functions (with predictable optimal σ_i values), check whether it works.
- Investigate impact of selection.
- Compare with optimal behavior (if known).

Test functions for experiments

- One common step size ($n_\sigma = 1$): Sphere model.

$$f_1(\vec{x}) = \sum_{i=1}^{n} x_i^2$$

- Appropriate scaling of variables ($n_\sigma = n$):

$$f_2(\vec{x}) = \sum_{i=1}^{n} i \cdot x_i^2$$

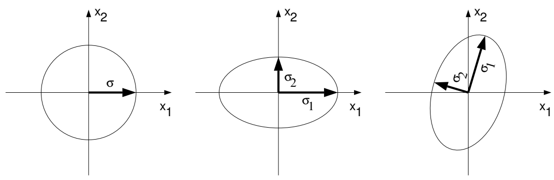
- A metric ($n_\sigma = n$, $n_\alpha = n \cdot (n-1)/2$):

$$f_3(\vec{x}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} x_j \right)^2$$

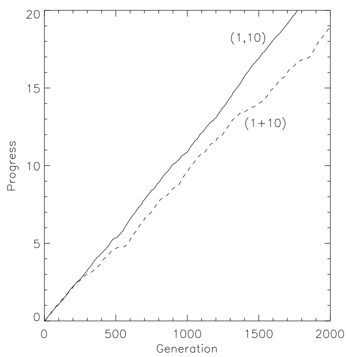
Experiments

Sketch of the lines of equal probability density

- Left: Standard mutations, $n_\sigma = 1$.
- Middle: Standard mutations, $n_\sigma = 2$.
- Right: Correlated mutations, $n_\sigma = 2$, $n_\alpha = 1$.



Experimental Results on Sphere Model (1)



Convergence velocity of a $(1,10)$-ES vs. that of a $(1+10)$-ES (sphere model f_1 with $n = 30$ and $n_\sigma = 1$).

Experimental Results on Sphere Model (2)

Progress measure:

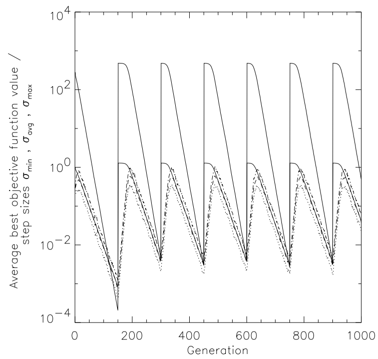
$$P_g = \log \sqrt{\frac{f_{\text{min}}(0)}{f_{\text{min}}(g)}}$$

- Counterintuitive: Elitist strategy is a bad choice.
- Misadapted σ might survive in an elitist strategy.
- Forgetting is necessary to prevent stagnation periods.

Time-Varying Sphere Model (1)

- Sphere model, $f(\vec{x}) = \|\vec{x} - \vec{x}^*\|^2 = R^2$.
- Optimum location $\vec{x}^*$ is shifted every 150 generations.
- (15,100)-ES, $n_\sigma = 1$, $n = 30$, no recombination.
- Simple model of a dynamic environment (with "catastrophes").

Time-varying Sphere Model (2)



Best objective function value and minimum, average, maximum and optimal standard deviation.

Time-varying Sphere Model (3)

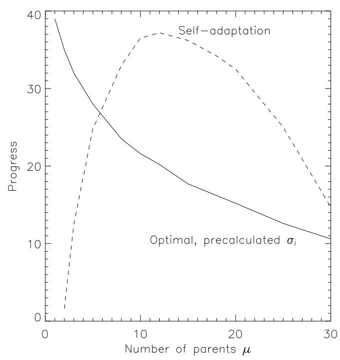
- Standard deviation σ adapts to the optimum value

$$\sigma_{opt} = c_{\mu,\lambda} \frac{R}{n} = c_{\mu,\lambda} \frac{\sqrt{f(\overline{x})}}{n}$$

- Transition time is $g \propto n$ (Beyer 1995).

⇒ The principle *learns* the optimal setting of the mutation rate ("internal strategy") without exogenous control.

Self-Adaptation is Collective Learning (1)

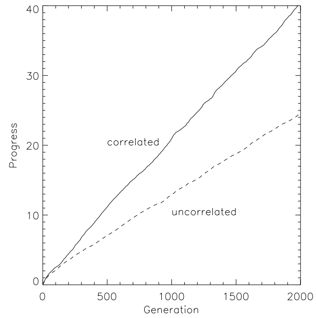


Average convergence velocity on f_2

Self-Adaptation is Collective Learning (2)

- $(\mu, 100)$-ES with $\mu \in \{1, \ldots, 30\}$
- $n_\sigma = n = 30$, and the optimum $\sigma_i \propto 1/\sqrt{i}$ is known.
- Optimum setting of σ_i: $\mu = 1$ performs best.
- Self-adaptation: $\mu = 12$ imperfect, diverse parents are as good as the optimal strategy.
- Individuals exchange information about their "internal models" by recombination.

Self-Adaptation of Covariances (1)



Convergence velocity of ES with correlated mutations vs. one with self-adaptation of standard deviations only, on f_3.

Self-Adaptation of Covariances (2)

- $(15, 100)$-ES, $n = n_\sigma = 10$, $n_\alpha = 45$.
- Recombination:
  - Intermediary on x_i.
  - Global intermediary on σ_i.
  - None on α_j (covariances).

Covariances increase effectiveness in case of rotated coordinate systems.

Other Variants for Continuous Search Spaces

- Original EP:
  $$\sigma' = \sigma \cdot (1 + \alpha \cdot N(0,1))$$
  Equivalent to log-normal with $n_\sigma = 1$, $\tau_0 = \alpha$ (Beyer 1995).
- Two-point distribution:
  $$\sigma' = \begin{cases} \sigma \cdot \alpha & , \text{ if } u \sim U(0,1) \leq 1/2 \\ \sigma / \alpha & , \text{ if } u \sim U(0,1) > 1/2 \end{cases}$$
  (Mutational step size control after Rechenberg, $\alpha = 1.3$).
- Substitution of $N(0,1)$ by other distributions (e.g., one-dimensional Cauchy, Yao and Liu 1996).

Self-Adaptation: Conclusions

- Powerful & robust parameter control scheme.
- Optimal conditions concerning selection, population size, etc.?
- Perfect adaptation vs. useful diversity — or a mixture ?
- Optimal speed of self-adaptation (i.e., learning rate settings) ?
- Few theoretical results.

Self-Adaptation: Individuals as Agents

- Individuals are *autonomous*; internal control of their behavior (mutation).
- Individuals *communicate* by exchanging partial information (recombination).
- Individuals are *reactive* to their environment (objective function).
- Further possibilities:
  - Spatial communication structure (graph).
  - Parallel implementation.
  - More complex internal strategies; including symbolic representation.

Basic Theory: The (1+1)-Strategy

- Properties of the mutation vector.
- The (1+1)-evolution strategy.
- Convergence velocity: Sphere model, corridor model.
- 1/5-success rule.
- Evolution window.
- ES vs. gradient strategy.

The mutation vector (1)

$$\Delta\vec{x} = \vec{z} = (z_1, \ldots, z_n)$$

$Z_1, \ldots, Z_n$: $(0, \sigma)$-normally distributed random variables.

$\Rightarrow S^2 = \sum_{i=1}^{n} Z_i^2$ is χ^2-distributed.

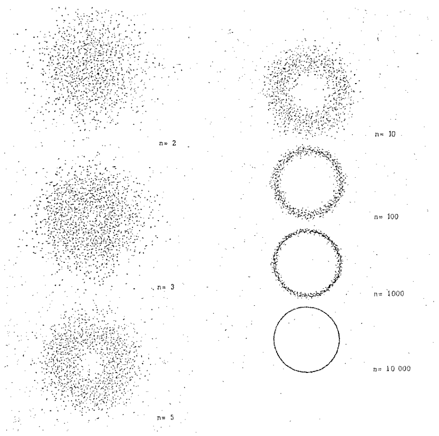
Random variable $S = \sqrt{S^2}$:
Length of the mutation vector $\vec{z}$.

After some math:

$$E(S) \approx \sigma\sqrt{n} \quad , \quad V(S) = \frac{1}{2}\sigma$$

The mutation vector (2)

- Variance $V(S)$ is independent of n.
- For large n: Offspring located on hypersphere of radius $E(S) \approx \sigma\sqrt{n}$.



The simple (1+1)-ES (1)

Algorithm:

$t := 0$;
initialize $P(0) := \{\vec{x}(0)\} \in I$, $I = I\!R^n$, $\vec{x} = (x_1, \ldots, x_n)$;
evaluate $P(0)$: $\{f(\vec{x}(0))\}$
while not *terminate*$(P(t))$ **do**
  mutate: $\vec{x}'(t) := m(\vec{x}(t))$
    where $x_i' := x_i + \sigma(t) \cdot N_i(0,1)\ \forall i \in \{1, \ldots, n\}$
  evaluate: $P'(t) := \{\vec{x}'(t)\} : \{f(\vec{x}'(t))\}$
  select: $P(t+1) := s_{(1+1)}(P(t) \cup P'(t))$;
  $t := t + 1$;
  if $(t \bmod n = 0)$ **then**
    $$\sigma(t) := \begin{cases} \sigma(t-n)/c & \text{, if } p_s > 1/5 \\ \sigma(t-n) \cdot c & \text{, if } p_s < 1/5 \\ \sigma(t-n) & \text{, if } p_s = 1/5 \end{cases}$$
    where p_s *is the relative frequency of successful mutations, measured over intervals of, say,* $10 \cdot n$ *trials;*
    and $0.817 \leq c \leq 1$;
  else
    $\sigma(t) := \sigma(t-1)$;
  fi
od

(1+1)-ES: Convergence velocity (1)

Convergence velocity: Expectation of the distance towards the optimum covered per generation.

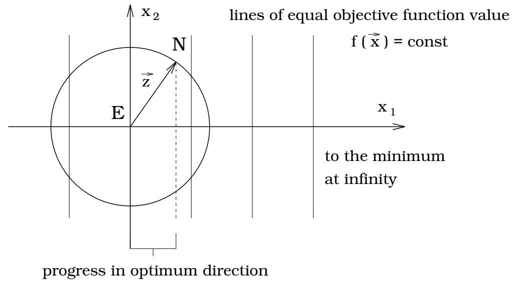
$$\varphi = E(\|\vec{x}^* - \vec{x}_t\| - \|\vec{x}^* - \vec{x}_{t+1}\|)$$

Alternatively:

$$\tilde{\varphi} = E(|f(\vec{x}^*) - f(\vec{x}_t)| - |f(\vec{x}^*) - f(\vec{x}_{t+1})|)$$

(1+1)-ES: Convergence velocity (2)

Example: The linear model.



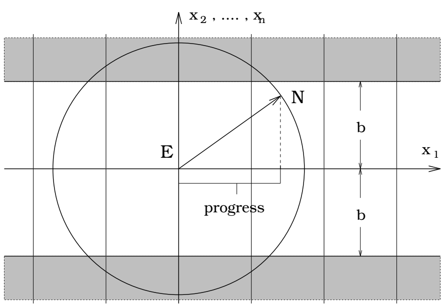
$$\begin{aligned}\varphi &= E(Z_1)\\ &= \sigma \cdot E(Z_1')\\ &= \sigma \int_0^\infty z_1 \phi(z_1) dz_1\\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z_1 \cdot \exp\left(-\frac{z_1^2}{2}\right) dz_1\\ &= \frac{\sigma}{\sqrt{2\pi}}\end{aligned}$$

Notice: $Z_1 \sim N(0, \sigma)$, $Z_1' \sim N(0,1)$

(1+1)-ES: Convergence velocity (3)

Example: The corridor model.

$$f(\vec{x}) = c \cdot x_1 \qquad -b \le x_2, \ldots, x_n \le b$$



lines of equal objective function value

(1+1)-ES: Convergence velocity (4)

Result of analysis:

$$\varphi = \frac{\sigma}{\sqrt{2\pi}} \left(1 - \frac{\sigma}{\sqrt{2\pi} b}\right)^{n-1}$$

After normalization of variables $\varphi' = \frac{\varphi n}{b}$, $\sigma' = \frac{\sigma n}{b}$:

$$\varphi' \approx \frac{\sigma'}{\sqrt{2\pi}} \exp\left(-\frac{\sigma'}{\sqrt{2\pi}}\right) \qquad \text{for } n \gg 1$$

Success probability: $w_e = P(f(\vec{x}') \le f(\vec{x}))$

$$w_e \approx \frac{1}{2} \exp\left(-\frac{\sigma'}{\sqrt{2\pi}}\right) \qquad \text{for } n \gg 1$$

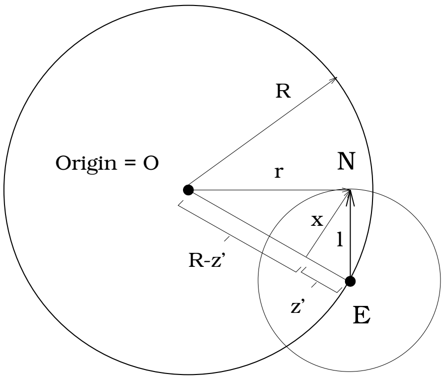
From these results:

- Optimal standard deviation: $\sigma'_{opt} = \sqrt{2\pi}$
- Maximum convergence velocity: $\varphi'_{max} = \frac{1}{e}$
- Optimal success probability: $w_{e_{opt}} = \frac{1}{2e}$

(1+1)-ES: Convergence velocity (5)

Example: The sphere model.

$$f(\vec{x}) = \sum_{i=1}^{n} x_i^2 = r^2$$



Geometry: $z'^2 + x^2 = l^2$, $x^2 + (R - z')^2 = r^2$, i.e.,

$$r^2 = l^2 + R^2 - 2Rz'$$

(1+1)-ES: Convergence velocity (6)

Derivation:

$$\begin{aligned} \tilde{\varphi} &= E(R^2 - r^2) = E(2RZ' - l^2) \\ &= E(2R\sigma Z - \sigma^2 n) \\ &= \ldots \\ &= 2R\sigma \int_{z_{\min}}^{\infty} z\phi(z)dz - \sigma^2 n \int_{z_{\min}}^{\infty} \phi(z)dz \\ &= \frac{2R\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\sigma^2 n^2}{8R^2}\right) - \sigma^2 n \left(1 - \Phi\left(\frac{\sigma n}{2R}\right)\right) \end{aligned}$$

(1+1)-ES: Convergence velocity (7)

Normalization of variables:

$$\varphi' = \frac{\sigma'}{\sqrt{2\pi}} \exp\left(-\frac{\sigma'^2}{8}\right) - \frac{\sigma'^2}{2}\left(1 - \Phi\left(\frac{\sigma'}{2}\right)\right)$$

Success probability:

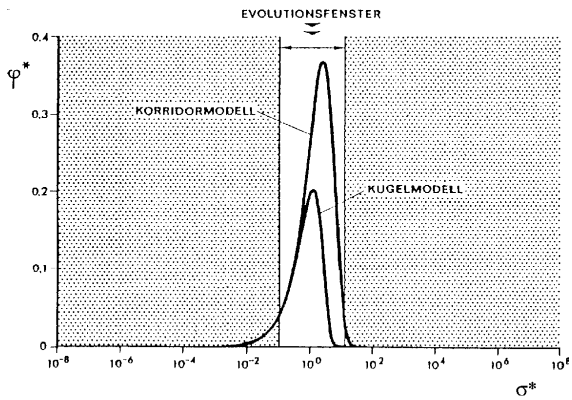
$$w_e = 1 - \Phi(\sigma'/2)$$

From these results:

- Optimal standard deviation: $\sigma'_{opt} \approx 1.224$
- Maximum convergence velocity: $\varphi'_{max} \approx 0.2025$
- Optimal success probability: $w_{e_{opt}} \approx 0.270$

(1+1)-ES: 1/5-success rule (1)

Progress window:



(1+1)-ES: 1/5-success rule (2)

Formulation of the 1/5-success rule:

> $w_{e_{opt}}$ should be about 1/5. If w_e — measured during execution of the (1+1)-ES — is larger than 0.2, increase σ. If it is smaller than 0.2, decrease σ.

Algorithmically:

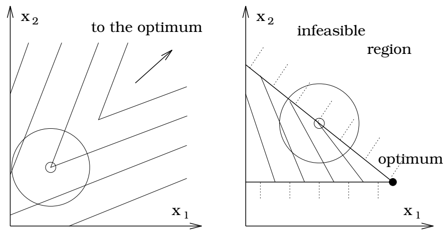
$$\sigma(t+n) = \begin{cases} \sigma(t) & , \quad w_e = 0.2 \\ \sigma(t) \cdot k & , \quad w_e < 0.2 \\ \sigma(t)/k & , \quad w_e > 0.2 \end{cases}$$

Choice of k: $k \approx 0.82$ (theoretical arguments, Schwefel 1977).

(1+1)-ES: 1/5-success rule (2)

Disadvantages of the (1+1)-ES:

- Certainly a more local search method.
- 1/5 success rule may fail.



(1+1)-ES vs. gradient method (1)

Gradient method:

- n sampling steps (partial derivative determinations).
- 1 working step, length $s = \sqrt{\sum_{i=1}^{n} z_i^2}$.

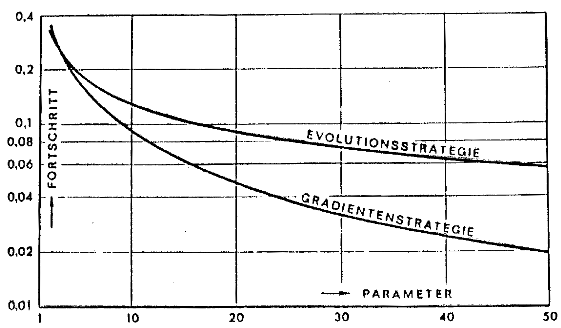
$$\varphi_{gradient} = \frac{s}{n+1} \sim \frac{1}{n}$$

(1+1)-ES:

$$\varphi_{(1+1)} = \frac{\sigma}{\sqrt{2\pi}} = \frac{E(S)}{\sqrt{2\pi n}} \sim \frac{1}{\sqrt{n}}$$

(since $E(S) = \sigma\sqrt{n}$).

(1+1)-ES vs. gradient method (2)



Progress rate as a function of n, ES vs. gradient.

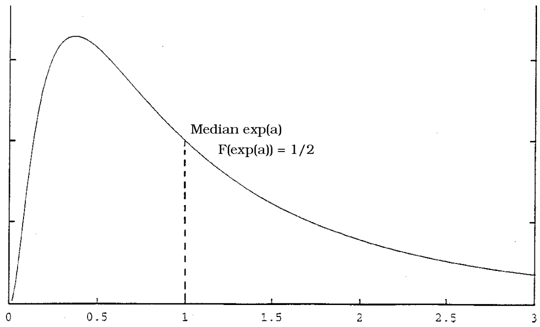
Extended Theory

- The lognormal distribution (self-adaptation).
- (1+λ) and (1,λ)-strategies.
- (μ,λ)-strategies with recombination:
  Discrete recombination, intermediary recombination.

Log-normal distribution (1)

- Probability density function:

$$f_X(x) = \frac{1}{\sigma x \sqrt{2\pi}} \cdot \exp\left(-\frac{(\ln\ x - \mu)^2}{2\sigma^2}\right)$$



log-normal distribution, $\sigma = 1, \mu = 0$.

Log-normal distribution (2)

- Expectation:

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

- Median:

$$\exp(\mu) \qquad \left(F_X(\exp(\mu)) = \frac{1}{2}\right)$$

Characteristics:

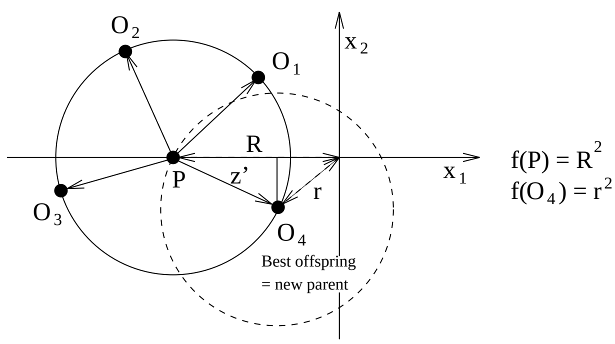
- Identical probability to sample x and $\frac{1}{x}$.
- Small changes more likely than large ones.
- σ_i are guaranteed to remain positive.

Convergence velocity of multi-membered ESs (1)

Simplifications:

- No self-adaption.
- One step-size.
- No recombination.
- $\mu = 1$

$\Rightarrow$ $(1 + \lambda)$-strategies, $(1, \lambda)$-strategies.



(1,4)-strategy, sphere model

Convergence velocity of multi-membered ESs (2)

Definition:
$Z_1, Z_2, \ldots, Z_\lambda$ i.i.d. random variables with p.d.f. $p(z)$.

$$Z_{1:\lambda} \le Z_{2:\lambda} \le \ldots \le Z_{\lambda:\lambda}$$

is called order statistics of the Z_i. $p_{v:\lambda}(z)$ denotes the p.d.f. of $Z_{v:\lambda}$.

Idea:
Best of offspring individual has

- smallest value of $r \Rightarrow r_{1:\lambda}$
- largest value of $z' \Rightarrow Z'_{\lambda:\lambda}$

Z': projection into direction of origin.

$$Z'_{v:\lambda} \sim N(0, \sigma)$$
$$Z_{v:\lambda} \sim N(0, 1)$$

Convergence velocity of multi-membered ESs (3)

$$\begin{aligned}\tilde{\varphi}_{(1\overset{+}{,}\lambda)} &= E(R^2 - r^2_{1:\lambda}) \\ r^2_{v:\lambda} &= l^2 + R^2 - 2R \cdot Z'_{\lambda-v+1:\lambda}\end{aligned}$$

Some math:

$$\begin{aligned}\tilde{\varphi}_{(1\overset{+}{,}\lambda)} &= E(2R \cdot Z'_{\lambda:\lambda} - \sigma^2 n) \\ &= E(2R\sigma \cdot Z_{\lambda:\lambda} - \sigma^2 n) \\ &= \int_{z_{min}}^{\infty} (2R\sigma \cdot z - \sigma^2 n) \cdot p_{\lambda:\lambda}(z)\, dz \\ &= 2R\sigma \int_{z_{min}}^{\infty} z \cdot p_{\lambda:\lambda}(z)\, dz - \\ &\quad \sigma^2 n \int_{z_{min}}^{\infty} p_{\lambda:\lambda}(z)\, dz\end{aligned}$$

Convergence velocity of multi-membered ESs (4)

With:

$$p_{\lambda:\lambda}(z) = \lambda\, \phi(z)\, (\Phi(z))^{\lambda-1} = \frac{d}{dz}\, (\Phi(z))^\lambda$$

It follows that:

$$\begin{aligned}\tilde{\varphi}_{(1\overset{+}{,}\lambda)} &= 2R\sigma \int_{z_{min}}^{\infty} z \cdot \frac{d}{dz}\, (\Phi(z))^\lambda\, dz - \\ &\quad \sigma^2 n \int_{z_{min}}^{\infty} \frac{d}{dz}\, (\Phi(z))^\lambda\, dz\end{aligned}$$

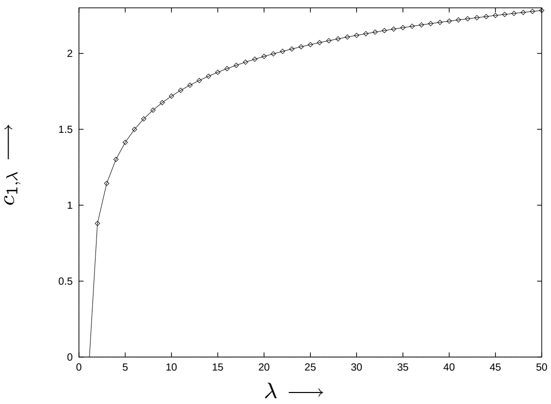
Convergence velocity of $(1, \lambda)$-ESs (1)

When accepting everything (non-elitist), $z_{min} = -\infty$.

$$\boxed{\tilde{\varphi}_{(1,\lambda)} = 2R\sigma \cdot c_{1,\lambda} - \sigma^2 n}$$

$$c_{1,\lambda} := E(Z_{\lambda:\lambda}) \begin{cases} \text{progress coefficent (Rechenberg)} \\ \text{selection intensity (Mühlenbein)} \end{cases}$$

Convergence velocity of $(1, \lambda)$-ESs (2)



The progress coefficient $c_{1,\lambda}$.

- Asymptotic behaviour: $c_{1,\lambda} \approx \sqrt{2 \ln \lambda}$.

Convergence velocity of $(1, \lambda)$-ESs (3)

- Normalisation of $\tilde{\varphi}$, with

$$\varphi \approx \frac{\tilde{\varphi}}{2R} \quad , \varphi' = \frac{\varphi n}{R} \quad , \sigma' = \frac{\sigma n}{R}$$

$$\boxed{\varphi'_{1,\lambda} = c_{1,\lambda}\sigma' - \tfrac{1}{2}\sigma'^2}$$

  – Optimal standard deviation:

$$\sigma'_{opt} = c_{1,\lambda}$$

  – Maximum convergence velocity:

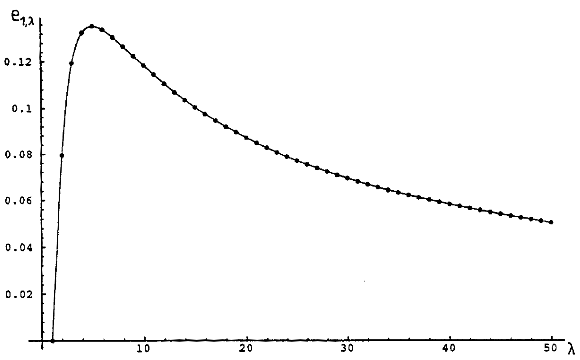
$$\varphi'_{max} = \frac{1}{2} c_{1,\lambda}^2 \approx \ln \lambda$$

- Evolution condition: $\sigma' < 2c_{1,\lambda}$
(Guarantees $\varphi' > 0$).

Evolution efficiency

Maximum progress per individual

$$e_{1,\lambda} = \frac{\varphi'_{max}}{\lambda}$$



Convergence velocity of $(1 + \lambda)$-ESs

- From $r \leq R$ it follows that

$$z_{min} = \frac{\sigma n}{2R}.$$

- Thus:

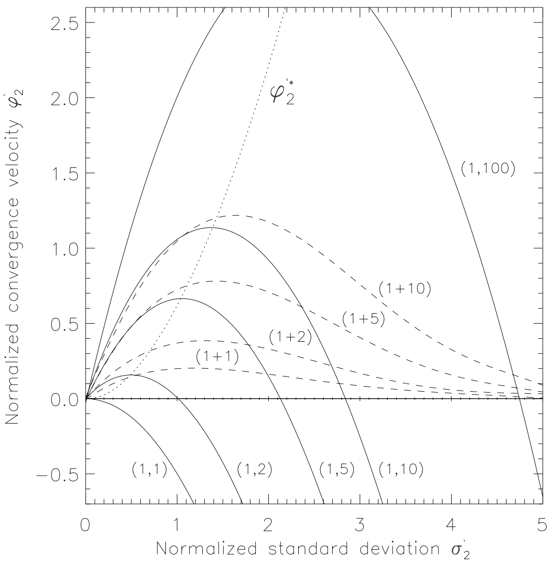
$$\varphi'_{(1+\lambda)} = \sigma' c_{1+\lambda}(\sigma') - \frac{\sigma'^2}{2}(1 - \Phi^{\lambda}(\frac{\sigma'}{z}))$$

- Where

$$c_{1+\lambda}(x) = \int_{\frac{x}{2}}^{\infty} z \frac{z}{dz} \Phi^{\lambda}(z)\, dz$$

No further analytical simplifications are possible.

Convergence velocity: illustration



Normalized convergence velocity φ' as a function of normalized standard deviation σ' for $(1, \lambda)$- and $(1 + \lambda)$-evolution strategies.

Convergence velocity of (μ, λ)-ESs (1)

Simplifications:

- No self-adaptation.
- One step-size.
- Recombination:
  - center of mass recombination μ/μ_I (intermediary), or
  - global discete recombination μ/μ_D.

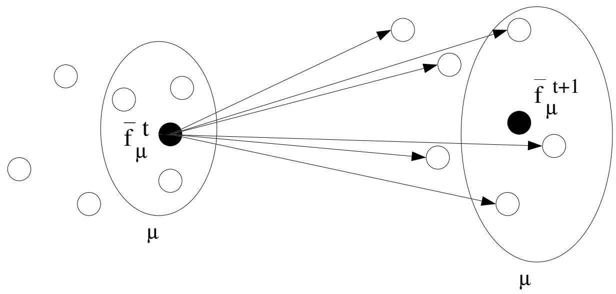


Illustration of center of mass recombination

.

Convergence velocity of (μ, λ)-ESs (2)

$$\begin{aligned}\varphi_{\mu,\lambda} &= \langle R \rangle - E(\langle \tilde{R} \rangle_{\mu,\lambda}) \\ &= \frac{1}{\mu}\sum_{v=1}^{\mu} R_v - \frac{1}{\mu}\sum_{v=1}^{\mu} r_{v:\lambda}\end{aligned}$$

Where:

- $\langle R \rangle$: Average distance to the optimum of parents.
- $\langle \tilde{R} \rangle_{\mu,\lambda}$: Average distance to the optimum of the μ best ofspring.

Convergence velocity of $(\mu/\mu_I, \lambda)$-ESs (1)

Without derivation (Rechenberg '94, Beyer '96):

$$\varphi'_{\mu/\mu_I,\lambda} = c_{\mu,\lambda} \cdot \sigma' - \frac{\sigma'^2}{2\mu}$$

(For $\sigma' \ll n, \mu^2 \ll n$)

- Optimal standard deviation:

$$\sigma'_{opt} = \mu \cdot c_{\mu,\lambda}$$

- Maximum convergence velocity:

$$\varphi'_{max} = \frac{1}{2}\mu \cdot c^2_{\mu,\lambda}$$

Convergence velocity of $(\mu/\mu_I, \lambda)$-ESs (2)

Progress coefficient ($Z_{v:\lambda} \sim N(0,1)$):

$$c_{\mu,\lambda} = \frac{1}{\mu} \sum_{v=\lambda-\mu+1}^{\lambda} E(Z_{v:\lambda})$$
$$\approx \frac{\lambda}{\mu} \cdot \phi(\Phi^{-1}(1 - \frac{\mu}{\lambda}))$$
$$\approx O\left(\sqrt{\ln \frac{\lambda}{\mu}}\right)$$

Conjecture:

$$\varphi'_{max} \approx \mu \cdot \ln \frac{\lambda}{\mu}$$

Convergence velocity of $(\mu/\mu_D, \lambda)$-ESs (1)

Without derivation (Rechenberg '94, Beyer '96):

$$\varphi'_{(\mu/\mu_D,\lambda)} = \sqrt{\mu} \cdot c_{\mu,\lambda} \sigma' - \frac{\sigma'^2}{2}$$

(For $\sigma' \ll n, \mu^2 \ll n$)

- Optimal standard deviation:
$$\sigma'_{opt} = \sqrt{\mu} \cdot c_{\mu,\lambda}$$
- Maximum convergence velocity:
$$\varphi'_{max} = \frac{1}{2}\mu \cdot c^2_{\mu,\lambda}$$

Again:

$$\varphi'_{max} \approx \mu \cdot \ln \frac{\lambda}{\mu}$$

Interpretation of results

- Genetic repair (Beyer '96):
μ/μ_I-recombination decreases the harmful part of mutation.
- Incest taboo:
μ/μ_I-recombination is only useful, if parents are different from each other.
- Implicit genetic repair:
μ/μ_D-recombination estimates the center of mass corresponding to a species centered around the wild-type.