

Computational Intelligence

Winter Term 2013/14

Prof. Dr. Günter Rudolph

Lehrstuhl für Algorithm Engineering (LS 11)

Fakultät für Informatik

TU Dortmund

mutation: $Y = X + Z$

$Z \sim N(0, C)$ multinormal distribution



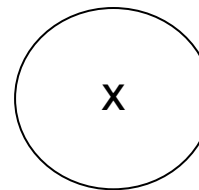
maximum entropy distribution for support \mathbb{R}^n , given expectation vector and covariance matrix

how should we choose covariance matrix C ?

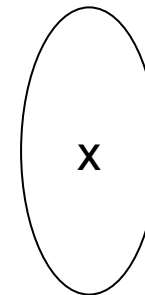
unless we have not learned something about the problem during search

⇒ don't prefer any direction!

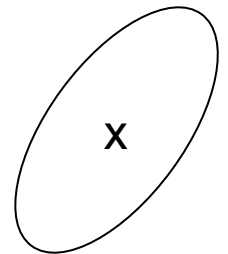
⇒ covariance matrix $C = I_n$ (unit matrix)



$C = I_n$

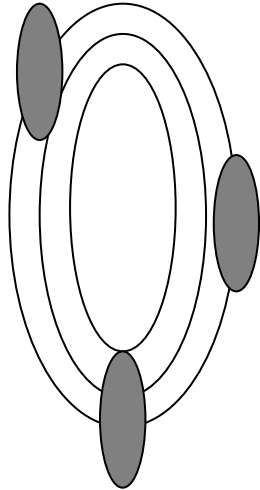


$C = \text{diag}(s_1, \dots, s_n)$



C orthogonal

claim: mutations should be aligned to isolines of problem (Schwefel 1981)



if true then covariance matrix should be inverse of Hessian matrix!

\Rightarrow assume $f(x) \approx \frac{1}{2} x'Ax + b'x + c \quad \Rightarrow H = A$

$Z \sim N(0, C)$ with density

$$f_Z(x) = \frac{1}{(2\pi)^{n/2} |C|^{1/2}} \exp\left(-\frac{1}{2} x' C^{-1} x\right)$$

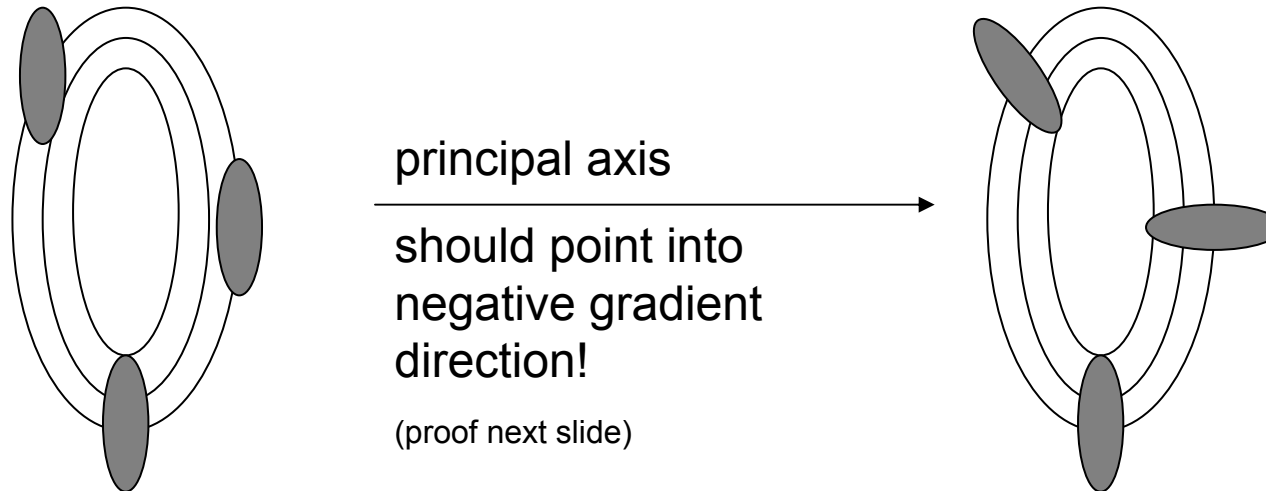
since then many proposals how to adapt the covariance matrix

\Rightarrow extreme case: use $n+1$ pairs $(x, f(x))$,

apply multiple linear regression to obtain estimators for A , b , c

invert estimated matrix A ! OK, **but:** $O(n^6)$! (Rudolph 1992)

doubts: are equi-aligned isolines really optimal?



most (effective) algorithms behave like this:

run roughly into negative gradient direction,

sooner or later we approach longest main principal axis of Hessian,

now negative gradient direction coincidences with direction to optimum,

which is parallel to longest main principal axis of Hessian,

which is parallel to the longest main principal axis of the inverse covariance matrix

(Schwefel OK in this situation)

$$Z = rQu, A = B'B, B = Q^{-1}$$

$$\begin{aligned}
 f(x + rQu) &= \frac{1}{2} (x + rQu)' A (x + rQu) + b'(x + rQu) + c \\
 &= \frac{1}{2} (x' Ax + 2rx' AQu + r^2 u' Q' AQu) + b'x + rb'Qu + c \\
 &= f(x) + rx' AQu + rb'Qu + \frac{1}{2} r^2 u' Q' AQu \\
 &= f(x) + r(Ax + b + \frac{r}{2} AQu)' Qu \\
 &= f(x) + r(\nabla f(x) + \frac{r}{2} AQu)' Qu \\
 &= f(x) + r \nabla f(x)' Qu + \frac{r^2}{2} u' Q' AQu \\
 &= f(x) + r \nabla f(x)' Qu + \frac{r^2}{2}
 \end{aligned}$$

if Qu were deterministic ...

\Rightarrow set $Qu = -\nabla f(x)$ (direction of steepest descent)

Apart from (inefficient) regression, how can we get matrix elements of Q ?

⇒ iteratively: $C^{(k+1)} = \text{update}(C^{(k)}, \text{Population}^{(k)})$

basic constraint: $C^{(k)}$ must be positive definite (p.d.) and symmetric for all $k \geq 0$,
otherwise Cholesky decomposition impossible: $C = Q'Q$

Lemma

Let A and B be quadratic matrices and $\alpha, \beta > 0$.

a) A, B symmetric $\Rightarrow \alpha A + \beta B$ symmetric.

b) A positive definite and B positive semidefinite $\Rightarrow \alpha A + \beta B$ positive definite

Proof:

ad a) $C = \alpha A + \beta B$ symmetric, since $c_{ij} = \alpha a_{ij} + \beta b_{ij} = \alpha a_{ji} + \beta b_{ji} = c_{ji}$

ad b) $\forall x \in \mathbb{R}^n \setminus \{0\}$: $x'(\alpha A + \beta B)x = \underbrace{\alpha x'Ax}_{> 0} + \underbrace{\beta x'Bx}_{\geq 0} > 0$ ■

Theorem

A quadratic matrix $C^{(k)}$ is symmetric and positive definite for all $k \geq 0$,

if it is built via the iterative formula $C^{(k+1)} = \alpha_k C^{(k)} + \beta_k v_k v_k'$

where $C^{(0)} = I_n$, $v_k \neq 0$, $\alpha_k > 0$ and $\liminf \beta_k > 0$.

Proof:

If $v \neq 0$, then matrix $V = vv'$ is symmetric and positive semidefinite, since

- as per definition of the dyadic product $v_{ij} = v_i \cdot v_j = v_j \cdot v_i = v_{ji}$ for all i, j and
- for all $x \in \mathbb{R}^n$: $x' (vv') x = (x'v) \cdot (v'x) = (x'v)^2 \geq 0$.

Thus, the sequence of matrices $v_k v_k'$ is symmetric and p.s.d. for $k \geq 0$.

Owing to the previous lemma matrix $C^{(k+1)}$ is symmetric and p.d., if

$C^{(k)}$ is symmetric as well as p.d. and matrix $v_k v_k'$ is symmetric and p.s.d.

Since $C^{(0)} = I_n$ symmetric and p.d. it follows that $C^{(1)}$ is symmetric and p.d.

Repetition of these arguments leads to the statement of the theorem. ■

Idea: Don't estimate matrix C in each iteration! Instead, approximate iteratively!

(Hansen, Ostermeier et al. 1996ff.)

→ **C**ovariance **M**atrix **A**daptation **E**volutionary **A**lgorithm (CMA-EA)

Set initial covariance matrix to $C^{(0)} = I_n$

$$C^{(t+1)} = (1-\eta) C^{(t)} + \eta \sum_{i=1}^{\mu} w_i d_i d_i'$$

η : "learning rate" $\in (0,1)$

$$m = \frac{1}{\mu} \sum_{i=1}^{\mu} x_{i:\lambda}$$

mean of all selected parents

complexity:
 $\mathcal{O}(\mu n^2 + n^3)$

$$d_i = (x_{i:\lambda} - m) / \sigma$$

sorting: $f(x_{1:\lambda}) \leq f(x_{2:\lambda}) \leq \dots \leq f(x_{\lambda:\lambda})$

dyadic product: $dd' =$

$$\begin{pmatrix} d_1 d_1 & d_1 d_2 & \dots & d_1 d_\mu \\ d_2 d_1 & d_2 d_2 & \dots & d_2 d_\mu \\ \vdots & \vdots & \ddots & \vdots \\ d_\mu d_1 & d_\mu d_2 & \dots & d_\mu d_\mu \end{pmatrix}$$

is positive semidefinite
dispersion matrix

variant:

$$m = \frac{1}{\mu} \sum_{i=1}^{\mu} x_{i:\lambda} \quad \text{mean of all selected parents}$$

$$p^{(t+1)} = (1 - \chi) p^{(t)} + (\chi (2 - \chi) \mu_{\text{eff}})^{1/2} (m^{(t)} - m^{(t-1)}) / \sigma^{(t)} \quad \text{“Evolution path“}$$

$$p^{(0)} = 0 \quad \chi \in (0, 1)$$

$$C^{(0)} = I_n$$

$$C^{(t+1)} = (1 - \eta) C^{(t)} + \eta p^{(t)} (p^{(t)})'$$

complexity: $\mathcal{O}(n^2)$

→ Cholesky decomposition: $\mathcal{O}(n^3)$ für $C^{(t)}$

State-of-the-art: CMA-EA (currently many variants)

→ successful applications in practice

available in WWW:

- http://www.lri.fr/~hansen/cmaes_inmatlab.html →
- <http://shark-project.sourceforge.net/> (EALib, C++)
- ...

**C, C++, Java
Fortran, Python,
Matlab, R, Scilab**

main arguments against EA in \mathbb{R}^n :

1. **Evolutionary Algorithms have been developed heuristically.**
2. **No proofs of convergence have been derived for them.**
3. **Sometimes the rate of convergence can be very slow.**

what can be done? \Rightarrow **disable arguments!**

ad 1) not really an argument against EAs ...

EAs use principles of biological evolution as pool of inspiration purposely:

- to overcome traditional lines of thought
- to get new classes of optimization algorithms

\Rightarrow the new ideas may be bad or good ...

\Rightarrow necessity to analyze them!

stochastic convergence \neq “empirical convergence“

frequent observation:

N runs on some test problem / averaging / comparison

\Rightarrow **this proves nothing!**

- no guarantee that behavior stable in the limit!
- N lucky runs possible
- etc.

formal approach necessary:

$D_k = |f(X_k) - f^*| \geq 0$ is a random variable

we shall consider the stochastic sequence D_0, D_1, D_2, \dots

Does the stochastic sequence $(D_k)_{k \geq 0}$ converge to 0?

If so, then evidently „convergence to optimum“!

But: there are many modes of **stochastic convergence**!

→ therefore here only the most frequently used ...

notation: $\mathcal{P}^{(t)}$ = population at time step $t \geq 0$, $f_b(\mathcal{P}^{(t)}) = \min\{f(x) : x \in \mathcal{P}^{(t)}\}$

Definition

Let $D_t = |f_b(\mathcal{P}(t)) - f^*| \geq 0$. We say: The EA

(a) **converges completely** to the optimum, if $\forall \varepsilon > 0$

$$\lim_{t \rightarrow \infty} \sum_{k=1}^t P\{D_k > \varepsilon\} < \infty ;$$

(b) **converges almost surely** or **with probability 1 (w.p. 1)** to the optimum, if

$$P\left\{\lim_{t \rightarrow \infty} D_k = 0\right\} = 1 ;$$

(c) **converges in probability** to the optimum, if $\forall \varepsilon > 0$

$$\lim_{t \rightarrow \infty} P\{D_t > \varepsilon\} = 0 ;$$

(a) **converges in mean** to the optimum, if $\forall \varepsilon > 0$

$$\lim_{t \rightarrow \infty} E\{D_t\} = 0 .$$



Lemma

- $(a) \Rightarrow (b) \Rightarrow (c)$.
- $(d) \Rightarrow (c)$.
- If $\exists K < \infty : \forall t \geq 0 : D_t \leq K$, then $(d) \Leftrightarrow (c)$.
- If $(D_t)_{t \geq 0}$ stochastically independent sequence, then $(a) \Leftrightarrow (b)$. ■

Typical *modus operandi*:

1. Show convergence in probability (c). Easy! (in most cases)
2. Show that convergence fast enough (a). This also implies (b).
3. Sequence bounded from above? This implies (d).

Let $(X_k)_{k \geq 1}$ be sequence of independent random variables.

$$\text{distribution: } P\{X_k = 0\} = 1 - \frac{1}{k} \quad P\{X_k = 1\} = \frac{1}{k}$$

$$1. \quad P\{X_k > \varepsilon\} = P\{X_k = 1\} = \frac{1}{k} \rightarrow 0 \text{ for } k \rightarrow \infty$$

\Rightarrow convergence in probability (c)

$$2. \quad \sum_{k=1}^{\infty} P\{X_k > \varepsilon\} = \sum_{k=1}^{\infty} P\{X_k = 1\} = \sum_{k=1}^{\infty} \frac{1}{k} = \infty$$

\Rightarrow convergence too slow! Consequently, no complete convergence!

$$3. \quad \text{Note: } \forall k \geq 0 : 0 \leq X_k \leq 1. \quad \text{Hence: sequence bounded with } K = 1.$$

since convergence in prob. (c) and bounded \Rightarrow convergence in mean (d)

let $(X_k)_{k \geq 1}$ be sequence of independent random variables.

distribution:

	(a)	(c)	(d)
$P\{X_k = 0\} = 1 - \frac{1}{k}$ $P\{X_k = 1\} = \frac{1}{k}$	(-)	(+)	(+)
$P\{X_k = 0\} = 1 - \frac{1}{k^2}$ $P\{X_k = 1\} = \frac{1}{k^2}$	(+)	(+)	(+)
$P\{X_k = 0\} = 1 - \frac{1}{k}$ $P\{X_k = k\} = \frac{1}{k}$	(-)	(+)	(-)
$P\{X_k = 0\} = 1 - \frac{1}{k^2}$ $P\{X_k = k\} = \frac{1}{k^2}$	(+)	(+)	(+)
$P\{X_k = 0\} = 1 - \frac{1}{k}$ $P\{X_k = k^2\} = \frac{1}{k}$	(-)	(+)	(-)
$P\{X_k = 0\} = 1 - \frac{1}{k^2}$ $P\{X_k = k^2\} = \frac{1}{k^2}$	(+)	(+)	(-)

timeline of theoretical work on convergence

1971 – 1975	Rechenberg / Schwefel	convergence rates for simple problems
1976 – 1980	Born	convergence proof for EA with genetic load
1981 – 1985	Rappl	convergence proof for (1+1)-EA in \mathbb{R}^n
1986 – 1989	Beyer	convergence rates for simple problems

all publications in German and for EAs in \mathbb{R}^n

⇒ results only known to German-speaking EA nerds!

timeline of theoretical work on convergence

1989	Eiben	a.s. convergence for elitist GA
1992	Nix/Vose	Markov chain model of simple GA
1993	Fogel	a.s. convergence of EP (Markov chain based)
1994	Rudolph	a.s. convergence of elitist GA non-convergence of simple GA (MC based)
1994	Rudolph	a.s. convergence of non-elitist ES (based on supermartingales)
1996	Rudolph	conditions for convergence

⇒ convergence proofs are no issue any longer!

Theorem:

Let $D_k = |f(x_k) - f^*|$ with $k \geq 0$ be generated by (1+1)-EA,

$S^* = \{x^* \in S : f(x^*) = f^*\}$ is set of optimal solutions and

$P_m(x, S^*)$ is probability to get from $x \in S$ to S^* by a single mutation operation.

If for each $x \in S \setminus S^*$ holds $P_m(x, S^*) \geq \delta > 0$, then $D_k \rightarrow 0$ completely and in mean.

Remark:

The proofs become simpler and simpler.

Born's proof (1978) took about 10 pages.

Eiben's proof (1989) took about 2 pages.

Rudolph's proof (1996) takes about 1 slide ...

Proof:

For the (1+1)-EA holds: $P(x, S^*) = 1$ for $x \in S^*$ due to elitist selection.

Thus, it is sufficient to show that the EA reaches S^* with probability 1:

Success in 1st iteration: $P_m(x, S^*) \geq \delta$.

No success in 1st iteration: $\leq 1 - \delta$.

No success in kth iteration: $\leq (1 - \delta)^k$.

\Rightarrow at least one success in k iterations: $\geq 1 - (1 - \delta)^k \rightarrow 1$ as $k \rightarrow \infty$.

Since $P\{D_k > \varepsilon\} \leq (1 - \delta)^k \rightarrow 0$ we have convergence in probability and

since $\sum_{k=0}^{\infty} (1 - \delta)^k < \infty$ we actually have complete convergence.

Moreover: $\forall k \geq 0: 0 \leq D_k \leq D_0 < \infty$, implies convergence in mean. ■

Observation:

Sometimes EAs have been very slow ...

Questions:

Why is this the case?

Can we do something against this?

⇒ no speculations, instead: **formal analysis!**

first hint in Schwefel's masters thesis (1965):

observed that step size adaptation in \mathbb{R}^2 useful!

convergence speed without „step size adaptation“ (pure random search)

$$f(x) = \|x\|^2 = x'x \rightarrow \min! \text{ where } x \in S_n(r) = \{x \in \mathbb{R}^n : \|x\| \leq r\}$$

Z_k is uniformly distributed in $S_n(r)$

$$X_{k+1} = Z_k \text{ if } f(Z_k) < f(X_k), \text{ else } X_{k+1} = X_k$$

$\Rightarrow V_k = \min \{ f(Z_1), f(Z_2), \dots, f(Z_k) \}$ best objective function value until iteration k

$$P\{\|Z\| \leq z\} = P\{Z \in S_n(z)\} = \text{Vol}(S_n(z)) / \text{Vol}(S_n(r)) = (z/r)^n, \quad 0 \leq z \leq r$$

$$P\{\|Z\|^2 \leq z\} = P\{\|Z\| \leq z^{1/2}\} = z^{n/2} / r^n, \quad 0 \leq z \leq r^2$$

$$P\{V_k \leq v\} = 1 - (1 - P\{\|Z\|^2 \leq v\})^k = 1 - (1 - v^{n/2} / r^n)^k$$

$$E[V_k] \rightarrow r^2 \Gamma(1 + 2/n) k^{-2/n} \text{ for large } k$$

no adaptation:

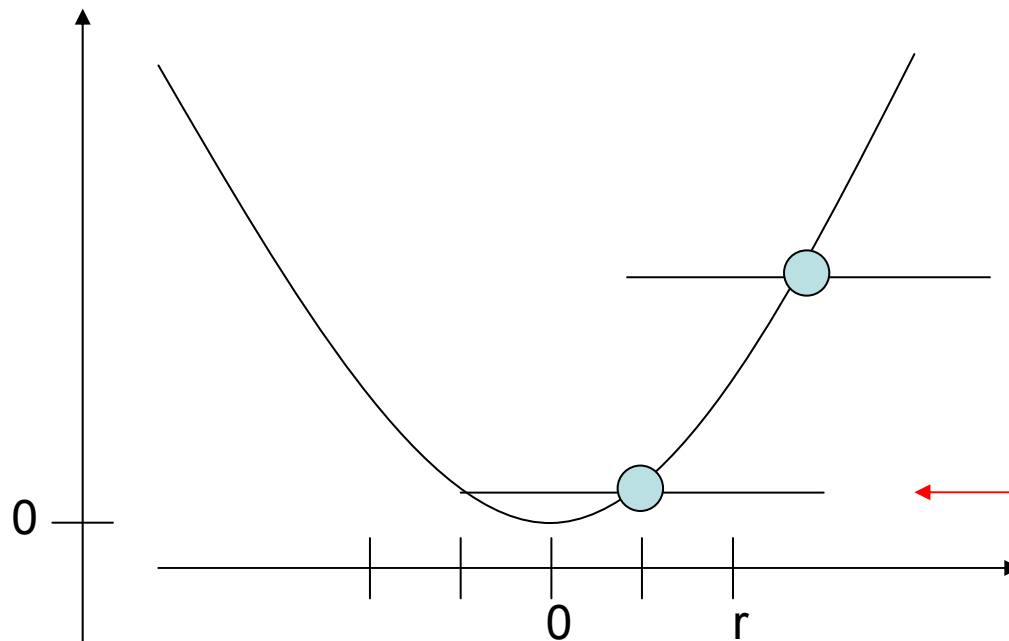
$$D_k = \Theta(k^{-2/n})$$

convergence speed without „step size adaptation“ (local uniformly distr.)

$$f(x) = \|x\|^2 = x'x \rightarrow \min! \text{ where } x \in S_n(r) = \{x \in \mathbb{R}^n : \|x\| \leq r\}$$

Z_k uniformly distributed in $[-r, r]$, $n = 1$

$$X_{k+1} = X_k + Z_k \text{ if } f(X_k + Z_k) < f(X_k), \text{ else } X_{k+1} = X_k$$



no adaptation:
 $D_k = \Theta(k^{-2/n})$

from now on:
 resembling pure
 random search!

convergence speed with „step size adaptation“ (uniform distribution on $S_n(1)$)

$(1, \lambda)$ -EA mit $f(x) = \|x\|^2$

$$\begin{aligned}
 \|Y_k\|^2 &= \|X_k + r_k U_k\|^2 = (X_k + r_k U_k)' (X_k + r_k U_k) \\
 &= X_k' X_k + 2r_k X_k' U_k + r_k^2 U_k' U_k \\
 &= \|X_k\|^2 + 2r_k X_k' U_k + r_k^2 \underbrace{\|U_k\|^2}_{=1} = \|X_k\|^2 + 2X_k' U_k + r_k^2
 \end{aligned}$$

note: random scalar product $x'U$ has same distribution like $\|x\| B$,

where r.v. B beta-distributed with parameters $(n-1)/2$ on $[-1, 1]$. It follows, that

$$\|Y_k\|^2 = \|X_k\|^2 + 2r_k \|X_k\| B + r_k^2.$$

Since $(1, \lambda)$ -EA selects best value out of λ trials in total, we obtain

$$\|X_{k+1}\|^2 = \|X_k\|^2 + 2r_k \|X_k\| B_{1:\lambda} + r_k^2$$

$$\|X_{k+1}\|^2 = \|X_k\|^2 + 2r_k \|X_k\| B_{1:\lambda} + r_k^2$$

↓ conditional expectation on both sides

$$E\|X_{k+1}\|^2 = \|X_k\|^2 + 2r_k \|X_k\| E[B_{1:\lambda}] + r_k^2$$

↓ assume: $r_k = \gamma \|X_k\|$

$$E\|X_{k+1}\|^2 = \|X_k\|^2 + 2\gamma \|X_k\|^2 E[B_{1:\lambda}] + \gamma^2 \|X_k\|^2$$

↓ symmetry of B implies $E[B_{1:\lambda}] = -E[B_{\lambda:\lambda}] < 0$

$$\begin{aligned} E\|X_{k+1}\|^2 &= \|X_k\|^2 - 2\gamma \|X_k\|^2 E[B_{\lambda:\lambda}] + \gamma^2 \|X_k\|^2 \\ &= \|X_k\|^2 (1 - 2\gamma E[B_{\lambda:\lambda}] + \gamma^2) \end{aligned}$$

with adaptation:

$$D_k \sim \mathcal{O}(c^k), c \in (0,1)$$

minimum at $\gamma^* = E[B_{\lambda:\lambda}]$, thus $E\|X_{k+1}\|^2 = \|X_k\|^2 (1 - E[B_{\lambda:\lambda}])^2$

problem in practice:

how do we get $\|X_k\|$ for $r_k = \|X_k\| \cdot E[B_{\lambda:\lambda}]$?

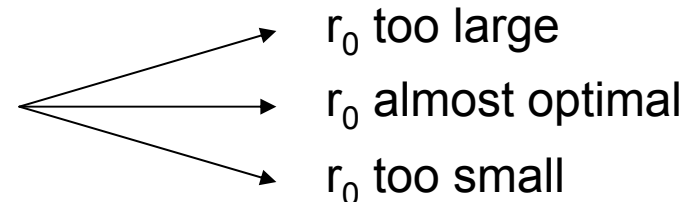
We know from analysis: $E\|X_{k+1}\|^2 = \|X_k\|^2 (1 - E[B_{\lambda:\lambda}]^2)$

assume: r_k was optimally adjusted

$$\Rightarrow r_{k+1} = \|X_{k+1}\| E[B_{\lambda:\lambda}] \approx \|X_k\| \underbrace{(1 - E[B_{\lambda:\lambda}]^2)^{1/2} E[B_{\lambda:\lambda}]}_{\text{constant!}}$$

\Rightarrow multiply r_k with constant: $r_{k+1} = c \cdot r_k$

but: how do we get r_0 or $\|X_0\|$?

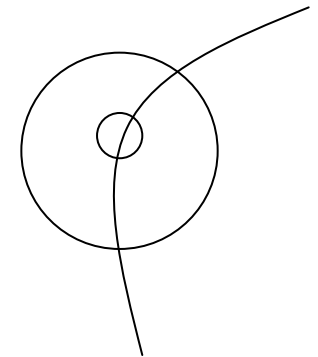


(1+1)-EA with step-size adaptation (1/5 success rule, Rechenberg 1973)

Idea:

- If many successful mutation, then step size too small.
- If few successful mutations, then step size too large.

for infinitesimal
small radius:
success rate = $1/2$



approach:

- count successful mutations in certain time interval
- if fraction larger than some threshold (z. B. $1/5$),
then increase step size by factor > 1 ,
else decrease step size by factor < 1 .

empirically known since 1973:

step size adaptation increases convergence speed dramatically!

about 1993 EP adopted multiplicative step size adaptation
(was additive)

no proof of convergence!

1999	Rudolph	no a.s. convergence for all continuous functions
2003	Jägersküppers	shows a.s. convergence for convex problems and linear convergence speed

⇒ same order of local convergence speed like gradient method!