ESs II Theoretical Aspects

Algorithms won their first reputation in practice;
i.e. they did their job
in finding good solutions
for difficult tasks

What is the aim of theory, then?

Insight: why and when do algorithms work

effectivity, robustness

efficiency, convergence velocity (order)

Insight II: how to improve algorithms

Important questions:

n-dependency of necessary effort,
speedup for parallel implementations

In the following: restriction to real-valued
parameter optimization
$(1+1)$, $(1,\lambda)$, and (μ,λ) ESs

results from Rechenberg, Rappl, Bäck,
Rudolph, Beyer, ...

7.1 Some Def. from Prob. Th. & Statistics

7.1.1 Random Variables, Distribution & Density Functions

Let $(\Omega, \mathcal{A}, p)$ a probability space

$X : \Omega \to \mathbb{R}$ a continuous random variable

Ω a set of elementary events

$\mathcal{A}$ an algebra of the events

p a probability measure,

then $F_X : \mathbb{R} \to [0,1]$ distribution function

$$x \mapsto F_X(x) = p(X \leq x)$$

$f_X : \mathbb{R} \to \mathbb{R}$ density function

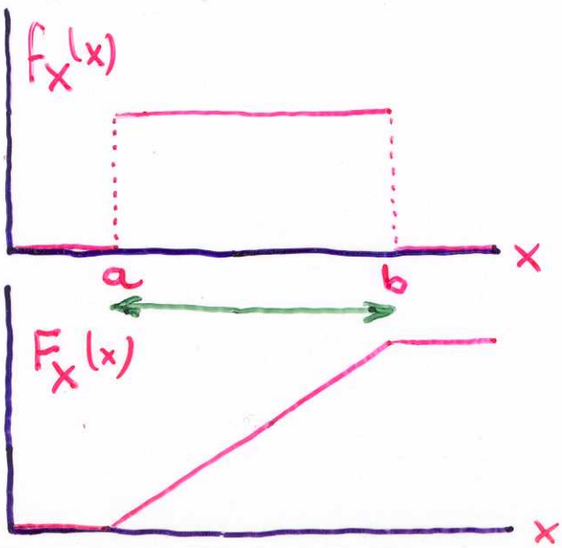
$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx \quad \forall x \in \mathbb{R}$$

(implicit definition)

or $f_X(x) = \frac{\partial}{\partial x} F_X(x)$

The set $\{x \in \mathbb{R} \mid f_X(x) > 0\}$ is called support (X)

e.g. uniform random number



$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

$a < x < b$ support (X)

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

7.1.2 Characteristic Values of p.d.s

One dimensional case

$$\xi = E[X] := \int_{-\infty}^{+\infty} x f_X(x)\,dx$$

expectation
measure of additive mean

$$\sigma^2 = D^2[X] := \int_{-\infty}^{+\infty} (x-\xi)^2 f_X(x)\,dx$$

variance
measures of dispersion

$$\sigma = D[X]$$

standard deviation

Let $h: \mathbb{R} \to \mathbb{R}$ a monotonous function of a continuous random variable with density f_X, then

$$E[h(X)] = \int_{-\infty}^{+\infty} h(x) f_X(x)\,dx$$

e.g. $h(x) = e^x$ used later on

if X normally distributed

then $Y := e^X$ lognormally distributed

additive mean in this case irrelevant

7.1.2 continued

Multidimensional case

$$X = (X_1, X_2, \dots, X_n)^T : \Omega \to \mathbb{R}^n$$

n-dimensional random vector

$$\xi = E[X] := (E[X_1], E[X_2], \dots, E[X_n])^T$$ expectation

Dispersion of two random variables X_i and X_j which may be independent from each other, or not

$$\Sigma_X = \mathrm{Cov}[X_i, X_j] := \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & & & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

covariance matrix

variances

$$\sigma_{ij} = \sigma_{ji} = E[(X_i - \xi_i)(X_j - \xi_j)]$$ covariances

$\sigma_{ii} = \sigma_i^2$ diagonal elements

If $\sigma_{ij} = 0 \quad \forall\, i \neq j$

then X_i and X_j are uncorrelated

Stochastically independent random variables are always uncorrelated

(the inverse, however, does not hold, generally)

7.1.3 Special Distributions

The Gaussian or normal distribution $N(\xi, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\xi)^2}{2\sigma^2}}$$ density function (bell shaped)

$$F_X(x) = \Phi(x) = \int_{-\infty}^{x} f_X(x)\,dx$$ distribution function

with expectation ξ and standard deviation σ

standard normal distribution $N(0, 1)$ (ξ, σ^2)

If $X \sim N(\xi, \sigma^2)$ then $Y := \frac{X-\xi}{\sigma} \sim N(0,1)$

$Y \sim N(0,1)$

$X := \xi + \sigma Y \sim N(\xi, \sigma^2)$

addition theorem

for stochastically independent $X_1, X_2, \ldots, X_n$

$X_i \sim N(\xi_i, \sigma_i^2)$ $\forall i$

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \xi_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

the sum is also normally distributed!

especially: if $\xi_i = \xi$ and $\sigma_i = \sigma$ $\forall i$

then $\xi_{sum} = n\xi$, $\sigma_{sum} = \sqrt{n}\,\sigma$

$\sigma^2_{sum} = n\sigma^2$

7.1.3 continued

The n-dimensional normal distribution

$$X = (X_1, X_2, \dots, X_n)^T$$

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma_X|}} e^{-\frac{1}{2}(x-\xi)^T \Sigma_X^{-1}(x-\xi)}$$

ξ : vector of expectations

Σ_X : covariance matrix ; $|\Sigma_X|$ its determinant

in this case: stochastic independence
equivalent to uncorrelatedness
off-diagonal elements $\sigma_{ij} = 0 \quad \forall i \neq j$
Σ_X diagonal matrix

The multidimensional normal distribution belongs to the class of elliptical distributions

$Z = RX$ is elliptically distributed if
- X is uniformly distributed over an n-dim. ellipsoid and
- R is a non-negative random variable that is stochastically independent of X

special case: spherical distribution
if X is uniformly distributed over a hypersphere
i.e. $\sigma_i = \sigma \quad \forall i$ and $\sigma_{ij} = 0 \quad \forall i \neq j$
within an n-dimensional normal distribution

Then, R^2 is χ^2 distributed

7.1.3 continued II

The χ^2 distribution

Let $X_i \sim N(0,1) \quad \forall\, i = 1, 2, \dots, n$

then $Y := \sum_{i=1}^{n} X_i^2$ has a χ^2_n distribution with n degrees of freedom

$$E[Y] = n$$

$$D^2[Y] = 2n$$

If $n \gg 1$ the χ^2 distribution can be approximated by an $N(n, 2n)$ distribution

i.e. $\chi^2_n \to n + \sqrt{2n}\, N$ for $n \gg 1$

$R := +\sqrt{Y}$ may be handled as if $R \sim N\left(\sqrt{n}, \sqrt{\frac{1}{2}}\right)$

(follows from rules for $E[h(x)]$ $D^2[h(x)]$ above)

7.1.3 continued III

Order Statistics

Let $X_1, X_2, \dots, X_n$ be stochastically independent random variables with F_{X_i} and f_{X_i}

then $Y_1 = \max\{X_1, X_2, \dots, X_n\}$

the distribution function F_{Y_1} of which is

$$F_{Y_1}(x) = \prod_{i=1}^{n} F_{X_i}(x)$$

$$= p(X_1 \leq x) \cdot p(X_2 \leq x) \cdot \dots \cdot p(X_n \leq x)$$

If all X_i are identically distributed

$$F_{Y_1}(x) = [F_X(x)]^n$$

$$f_{Y_1} = \frac{\partial}{\partial x} F_{Y_1} = n f_X(x) [F_X(x)]^{n-1}$$

More generally:

Let Y_m be the m-th largest X_i

then

$$f_{Y_m}(x) = \underbrace{n\binom{n-1}{m-1}}_{c} f_X(x) \underbrace{[F_X(x)]^{n-m}}_{b} \underbrace{[1-F_X(x)]^{m-1}}_{a}$$

a: $m-1$ larger

b: $n-m$ smaller

elements X_i in vector X

c: combinatorial manifold of situations

7.2 Convergence behavior of ES's

1971 $(1+1)$ two membered
$(\mu+1)$ first multimembered
| sequential Rechenberg

1974 $(1+\lambda)$, $(1,\lambda)$
$(\mu+\lambda)$, (μ,λ)
| parallel Schwefel

plus-versions: elitist
comma-versions: non-elitist $\kappa = 1$ $\kappa = \infty$

1995 (μ, κ, λ) ES κ = max. life span

generally: theory can assure something only on the basis on certain assumptions (simplifications)

efficiency results need more rigorous ass. than effectiveness results

nevertheless: except by chance

no method can be better in the more general case than has been proven for simplified situations

e.g. Newton-Raphson method
just one step necessary if objective quadratic
but may be divergent in slightly modified situations

EAs aim at robustness
however: no theory for general efficiency

7.2.1 Convergence reliability

here: only continuous case $x \in \mathbb{R}^n$
[see Rudolph, 1994 for $x \in \mathbb{N}^n$]

here: only $(1+1) \in S$
[see Rudolph, 1995 for comma-versions]

We start at $x^{(0)}$ and iterate

$$x^{(t+1)} = x^{(t)} + z \quad \text{with } z \sim N(0, \sigma^2 I_n)$$

We assume a regular optimization problem, i.e.

$$f^* = f(x^*) = \min\{ f(x) \mid x \in M \subseteq \mathbb{R}^n \}$$

with

a) $f^* > -\infty$

b) $x^* \in \text{int}(M)$ — interior

c) $\mu(\{x \in M \mid f(x) \in U_\varepsilon(f^*)\}) > 0 \quad \forall \varepsilon > 0$

Lebesgue measure

[f^* global minimum ; x^* solution or global minimizer]

a) necessary, otherwise there is no finite minimizer
b) only simplifies analysis
c) excludes singular optima (needle in a haystack) that cannot be found with probability > 0

7.2.1 continued

Theorem A

Let $\varepsilon > 0$ and $p_t := p\left(x^{(t)} \in \{x \in M \mid f(x) \in U_\varepsilon(f^*)\}\right)$

the prob. that a $(1+1)$ ES has reached $x^{(t)}$ at iteration t the obj. function value of which is closer to f^* than ε ;

then assuming that

$$\sum_{t=0}^{\infty} p_t = \infty \qquad A_1$$

implies

$$p\left(\lim_{t \to \infty} (f(x^{(t)}) - f^*) = 0\right) = 1$$

for any starting point $x^{(0)} \in M$

more practically (Lemma A'):

If $M \subseteq \text{support}(f_z)$ with f_z probab. dens. of vector z of the mutation operator, and M is bounded, then A_1 is valid

more loosely:

you can reach x^* at any iteration t immediately as long as $\sigma^{(t)} > 0$

This result is of academic interest only !

More interesting : Time to achieve a certain approximation

7.2.2 Convergence velocity (or rate)

Let $\delta_t := E[f(x^{(t)}) - f^*]$ expected error at step t

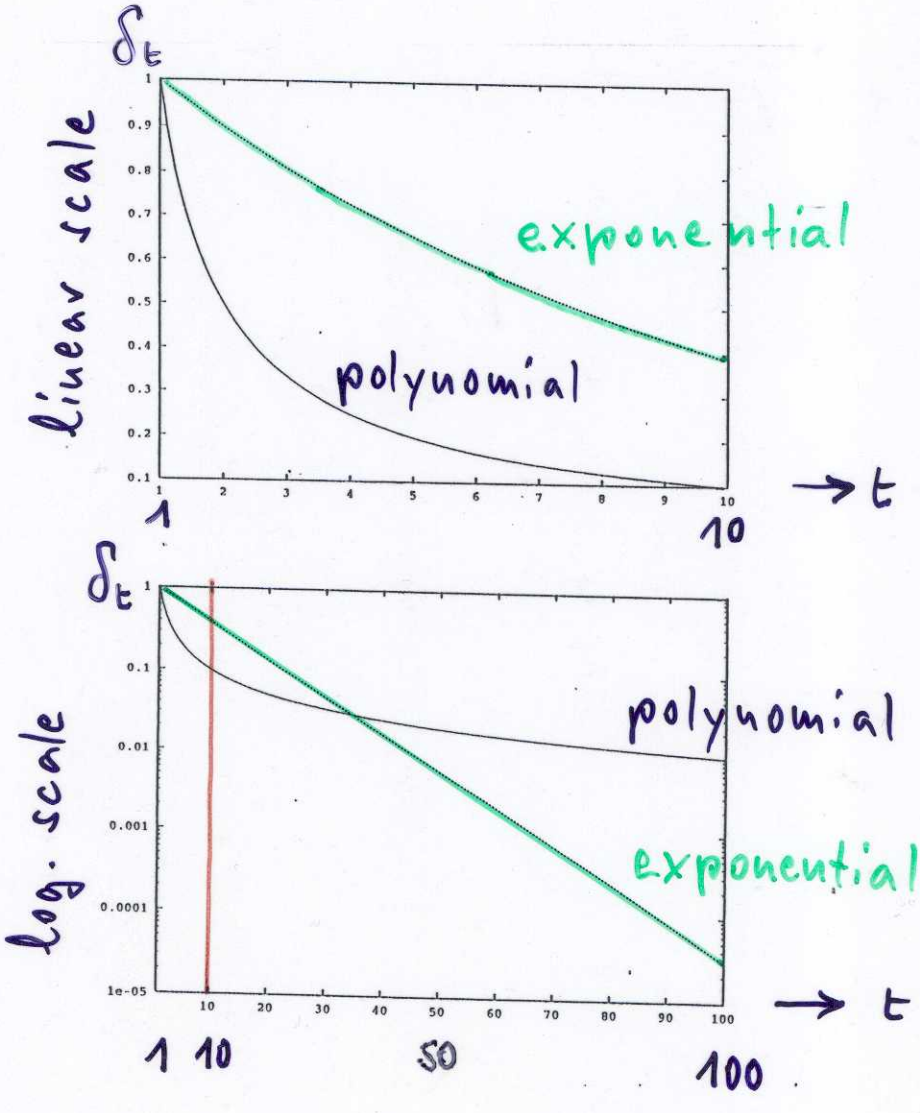
There are different kinds of convergence order

$\delta_t = \mathcal{O}(t^{-\alpha})$, $\alpha > 0$ polynomial

$\delta_t = \mathcal{O}(\beta^t)$, $\beta \in (0,1)$ exponential

Convergence order

→ two graphs



7.2.2 continued

special assumptions necessary
here: strictly convex problems
(implies continuous differentiability, unimodality)

This is not the domain of EAs, but

any method that is good for difficult problems should behave even better in simple situations.

Thus, we may gain best-case results, now.
(stronger ones in the future, hopefully)

Theorem B

Let $f: M \to \mathbb{R}$ strictly convex objective function
mut. step size of a $(1+1)$ ES spherically distrib.
with $z = RU$, $\text{support}(R) = (0, a) \neq \varnothing$

Then $\delta_t = \begin{cases} O(t^{-2/n}) & \text{for a constant} \\ O(\beta^t) & \text{for an adaptive} \end{cases}$ mut. step size

at step t for any start position $x^{(0)} \in M$

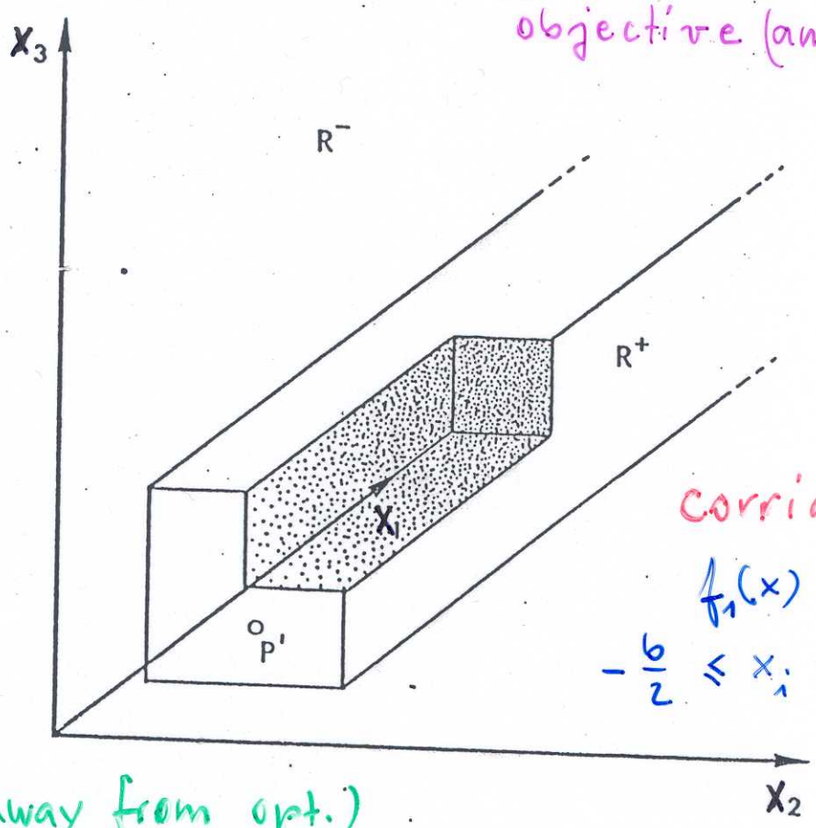
with $\beta \in (0,1)$ and step size adaptation
according to $R^{(t+1)} = \|\nabla f(x^{(t)})\| R^{(t)}$.

or 1/5 success rule (Rechenberg, Rappl)

How large is β?

two (n-dimensional) 'fitness' functions

objective (and constraints)

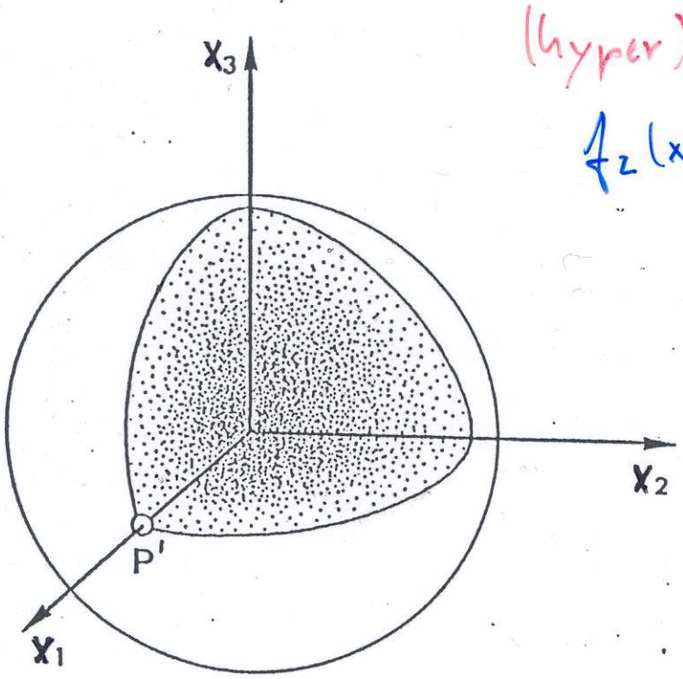


corridor (ridge)

$f_1(x) = -a\,x_1$

$-\frac{b}{2} \leq x_i \leq \frac{b}{2} \quad \forall i = 2,3,\ldots,n$

(far away from opt.)



(hyper) sphere

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

$= r^2$

(close to opt.)

i.e. two different 'typical' situations

7.3 The Sphere Model — n-dimensional

$$f(x) := \|x - x^*\|^2 = r^2$$

Results are valid for many more models with concentric hypersurfaces $f(x) = \text{const.}$

What we want to know: n-dependency and σ- "

of $\varphi := E[r^{(t)} - r^{(t+1)}]$

mean convergence velocity $\varphi(n, \sigma, r)$

for ES's with just one common σ

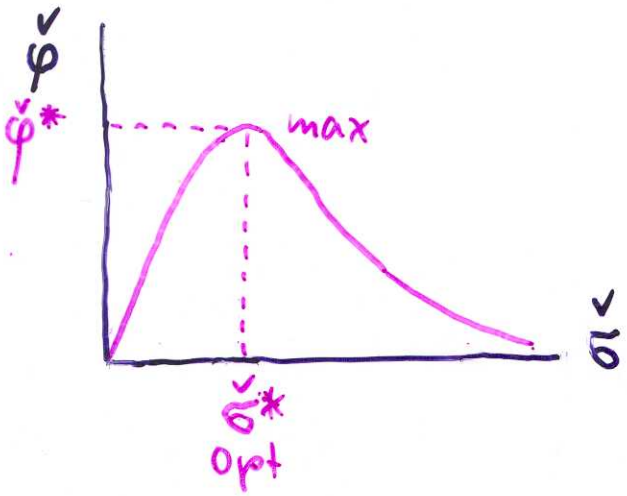
most interesting: $\varphi_{max}(n, \sigma_{opt}, r)$

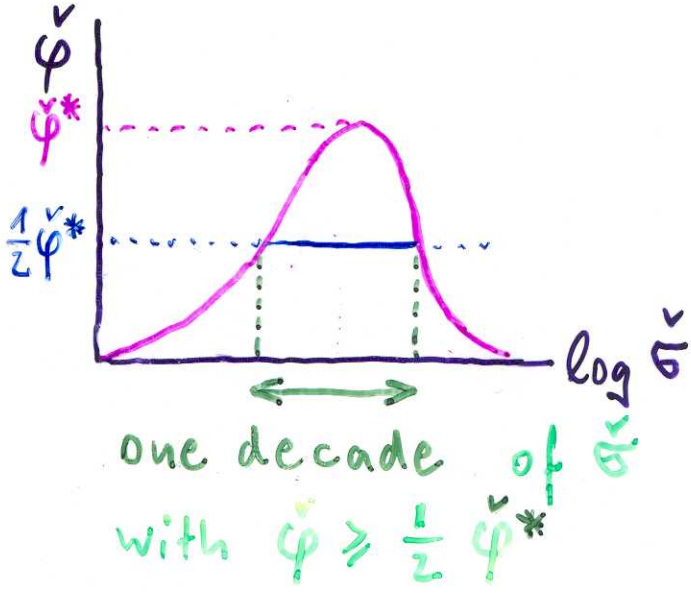
7.3.1 The (1+1) ES — / two membered

$$\varphi_{1+1} \approx \frac{r}{n}\left\{\frac{\sigma n}{\sqrt{2\pi}\,r} e^{-\left(\frac{\sigma n}{\sqrt{8}\,r}\right)^2} - \left(\frac{\sigma n}{\sqrt{2}\,r}\right)^2\left[1 - \Phi\left(\frac{\sigma n}{2r}\right)\right]\right\}$$

result — for $n \gg 1$

with 'dimensionless' $\check{\varphi} = \varphi\frac{n}{r}$, $\check{\sigma} = \sigma\frac{n}{r}$





in between:

$$r^{(t+1)} = \| x^{(t)} - x^* + z \| = \sqrt{\sum_{i=1}^{n} \left((x_i^{(t)} - x_i^*) + \sigma N_i \right)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i^{(t)} - x_i^*)^2 + 2\sigma \sum_{i=1}^{n} (x_i^{(t)} - x_i^*) N_i + \sigma^2 \underbrace{\sum_{i=n}^{n} N_i^2}_{\chi^2}}$$

$$\sim \sqrt{r^2 + 2\sigma r N + \sigma^2 (n + \sqrt{2n}\, N')} \quad \leftarrow \text{addition th.}$$

$$\sim \sqrt{r^2 + n\sigma^2 + \sqrt{2n\sigma^4 + 4r^2\sigma^2}\, N''}$$

$$p := \frac{r^{(t+1)} - r}{r} \qquad \text{relative progress}$$

$$= 1 - \sqrt{1 + \frac{n\sigma^2}{r^2} + \sqrt{\frac{2n\sigma^4}{r^4} + 4\frac{\sigma^2}{r^2}}\, N''}$$

with $\sigma \frac{n}{r} = \check{\sigma}$

$$p = 1 - \sqrt{1 + \frac{\check{\sigma}^2}{n} + \frac{\check{\sigma}^2}{n}\sqrt{\frac{4}{\check{\sigma}^2} + \frac{2}{n}}\, N''}$$

small if n large

$$\approx 1 - \sqrt{1 + \frac{\check{\sigma}^2}{n} + \frac{2\check{\sigma}}{n} N''} \qquad \text{with } \sqrt{1+\varepsilon} \approx 1 + \frac{\varepsilon}{2}$$

$$\approx 1 - \left(1 + \frac{\check{\sigma}^2}{2n} + \frac{\check{\sigma}}{n} N''\right) = -\frac{\check{\sigma}^2}{2n} - \frac{\check{\sigma}}{n} N''$$

$$\sim N\left(-\frac{\check{\sigma}^2}{2n}, \frac{\check{\sigma}^2}{n^2}\right)$$

$$E[p_{l+1}] = \int_0^\infty x\, f_p(x)\, dx := \frac{\varphi_{l+1}}{r} = \frac{\check{\varphi}_{l+1}}{n}$$

only positive moves accepted

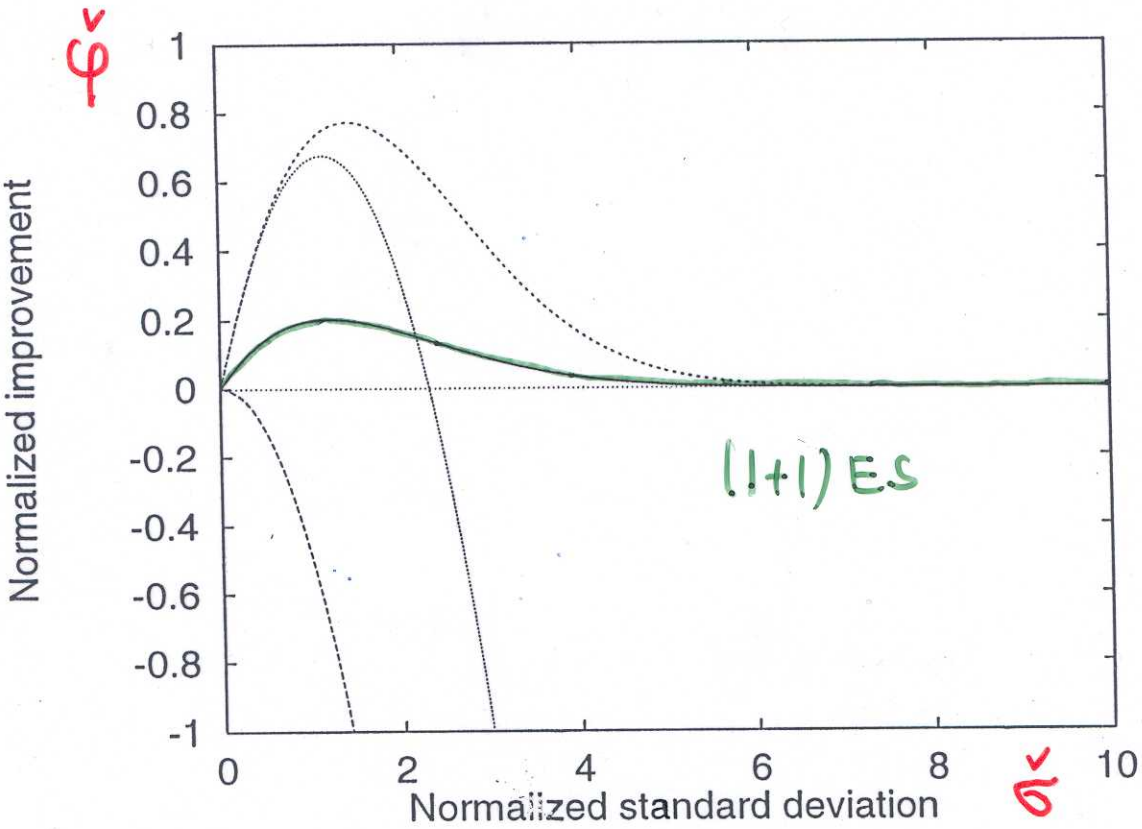
within $(l+1) \in S$

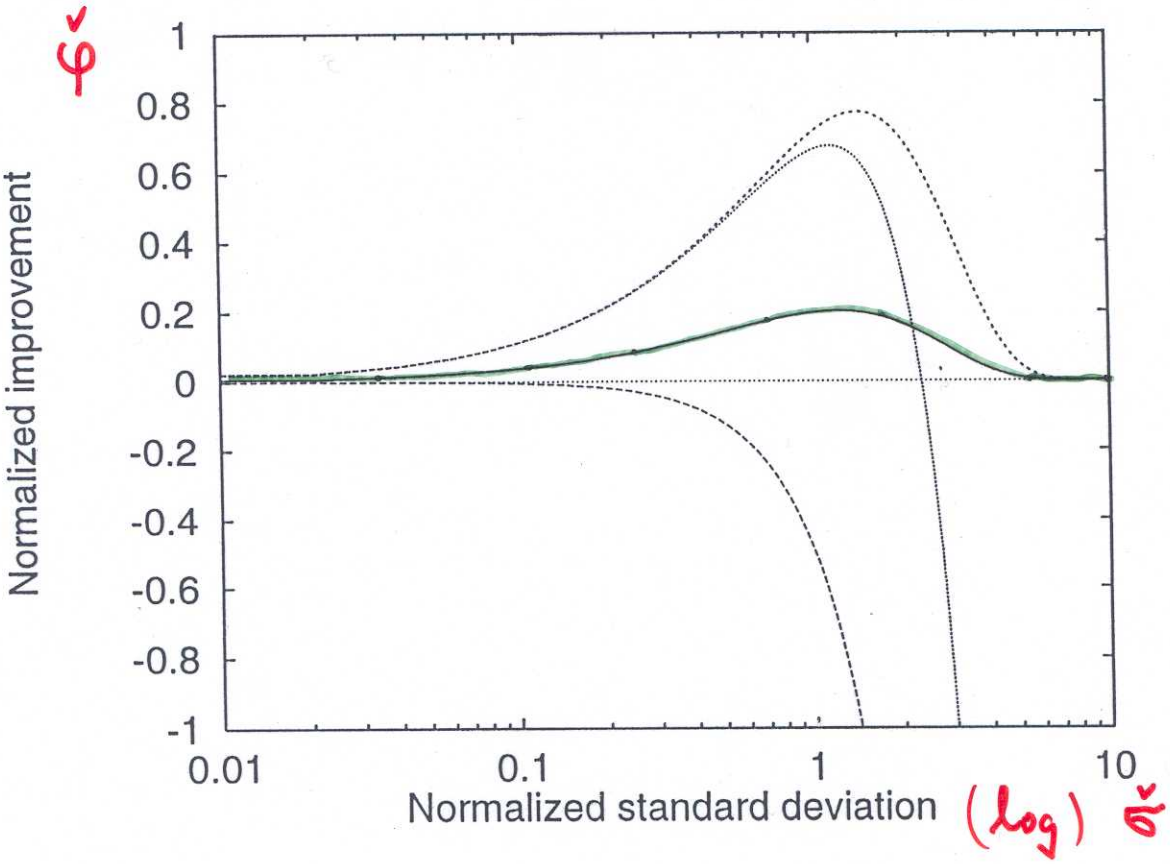
$$f_p(x) = \frac{1}{\sqrt{2\pi}\,\frac{\check{\sigma}}{n}} e^{-\frac{1}{2}\frac{\left(x+\frac{\check{\sigma}^2}{2n}\right)^2}{2\frac{\check{\sigma}^2}{n^2}}}$$

$$= \frac{n}{\sqrt{2\pi}\,\check{\sigma}} e^{-\frac{1}{2}\left(\frac{nx}{\check{\sigma}}+\frac{\check{\sigma}}{2}\right)^2}$$

$$E[P_{t+1}] = \int_0^\infty \frac{nx}{\sqrt{2\pi}\,\check{\sigma}} e^{-\frac{1}{2}\left(\frac{nx}{\check{\sigma}}+\frac{\check{\sigma}}{2}\right)^2} dx$$

$$= \frac{1}{n}\left\{\frac{\check{\sigma}}{\sqrt{2\pi}} e^{-\frac{\check{\sigma}^2}{8}} - \frac{\check{\sigma}^2}{2}\left[1-\Phi\left(\frac{\check{\sigma}}{2}\right)\right]\right\}$$


Normalized improvement
Normalized standard deviation
(1+1) ES


Normalized improvement
Normalized standard deviation
(log)

7.3.1 continued

$$\varphi^*_{l+1} = 0.2025\,\frac{r}{n} \quad \text{at} \quad \sigma^* = 1.224\,\frac{r}{n} = \frac{0.612}{n}\,\|\nabla f(x)\|$$

$\sim \frac{1}{n}$ $\sim \frac{1}{n}$ inversely proportional to the # of variables

Success probability

$$p^*_s \approx \Phi\left(-\frac{\sigma^*}{2}\right) \approx 0.270$$

Similar results available for an n-dimensional 'corridor' model of width b

$$\varphi^*_{l+1} \sim \frac{b}{n}\,, \quad \sigma^* \sim \frac{b}{n}\,, \quad p^*_s = \frac{1}{2e} \approx 0.184$$

⇒ formulation of 1/5 success rule

7.3.2 The $(\mu+1)$ ES

extinction of the worst

Similar to EVOP strategy of G.E.P. Box

polyhedron strategies

Simplex	Nelder and Mead
Complex	M.J. Box

no longer in use
since self-adaptation of σ does not work within the $(\mu+1)$ scheme

7.3.3 The $(1+\lambda)$ ES

Already calculated: relative change per generation of each successor

$$P \approx -\frac{\check{\sigma}^2}{2n} + \frac{\check{\sigma}}{n} N''$$

The corresponding distribution function is

$$F_P(x) = \Phi\left((x-\theta)/\eta\right)$$

with expectation $\theta = -\frac{\check{\sigma}^2}{2n}$

and standard deviation $\eta = \frac{\check{\sigma}}{n}$

From order statistics:
The best out of λ successors

$$F_{P_\lambda}(x) = \Phi^\lambda\left(\frac{x-\theta}{\eta}\right)$$

with $$f_{P_\lambda}(\hat{x}) = \frac{d}{dx}\Phi^\lambda(\hat{x}) = \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{\hat{x}^2}{2}} \Phi^{\lambda-1}(\hat{x})$$

from above

$$E[P_\lambda] = \frac{1}{\eta}\int_{-\infty}^{\infty} x f_{P_\lambda}\left(\frac{x-\theta}{\eta}\right) dx$$

$$E[P_{1+\lambda}] = E[\max\{0, P_\lambda\}]$$

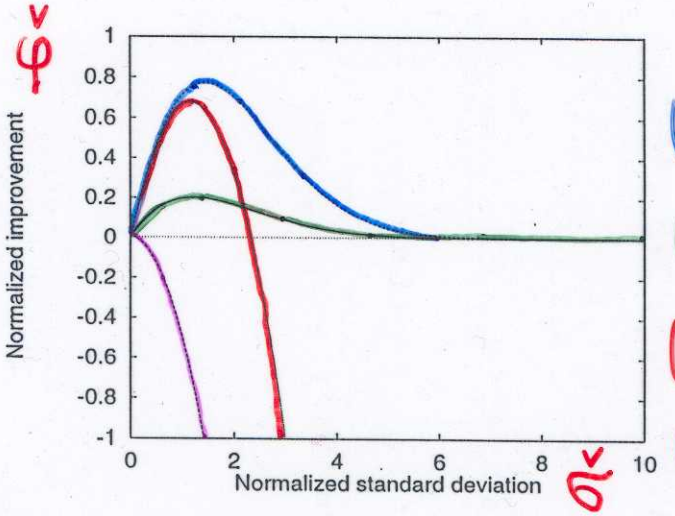
$$= \frac{1}{\eta}\int_{0}^{\infty} x f_{P_\lambda}\left(\frac{x-\theta}{\eta}\right) dx$$

substituting $z := (x-\theta)/\eta$ yields

$$E[P_{1+\lambda}] = \int_{-\theta/\eta}^{\infty} (z\eta+\theta) f_{P_\lambda}(z)\,dz$$

$$= \frac{\check{\sigma}}{n} \int_{\check{\sigma}/2}^{\infty} z f_{P_\lambda}(z)\,dz - \frac{\check{\sigma}^2}{2n}\left[1-\Phi^\lambda\left(\frac{\check{\sigma}}{2}\right)\right]$$

integral must be evaluated numerically



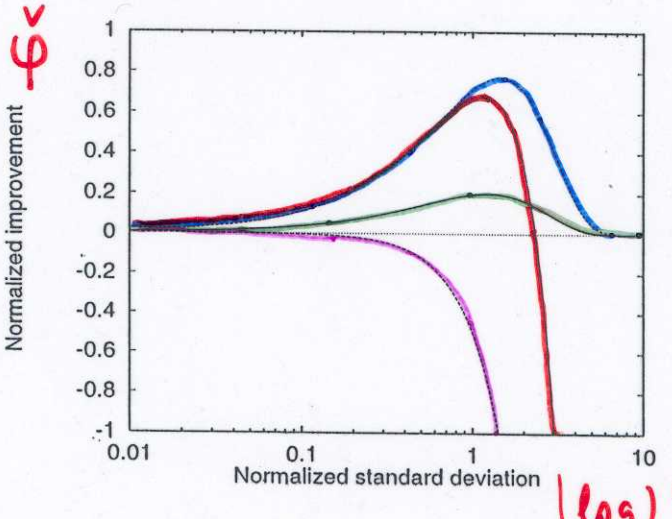
$\check{\varphi}$ $\check{\sigma}$

(1+5) ES

(1+1) ES

(1,5) ES

(1,1) random walk



$\check{\varphi}$ (log) $\check{\sigma}$

7.3.4 The $(1,\lambda)$ ES

non-elitist ! $\Rightarrow$ it may diverge if σ too large

Rudolph (1994) has shown : under certain conditions the $(1,\lambda)$ ES converges on average almost sure.

For mean convergence rate we just have to change the lower bound of the integral only

$$E[P_{1,\lambda}] = \int_{-\infty}^{\infty} (z\eta + \theta)\, f_{P_\lambda}(z)\, dz$$

$$= \frac{\check{\sigma}}{n} \underbrace{\int_{-\infty}^{\infty} z\, f_{P_\lambda}(z)\, dz}_{c_{1,\lambda}} - \frac{\check{\sigma}^2}{2n}$$

$c_{1,\lambda}$ new constant

$$= (2\check{\sigma}\, c_{1,\lambda} - \check{\sigma}^2)/2n$$

Looking for a maximum at some optimal σ^*

$$\frac{\partial}{\partial \check{\sigma}} E[P_{1,\lambda}] = \frac{1}{2n}(2c_{1,\lambda} - 2\check{\sigma})$$

becomes zero if $\check{\sigma} = c_{1,\lambda}$

$$\frac{\partial^2}{\partial \check{\sigma}^2} E[P_{1,\lambda}] = -\frac{1}{n} < 0$$

Thus, the optimal mutation step size is

$$\sigma^* = c_{1,\lambda}\frac{r}{n} = \frac{c_{1,\lambda}}{2n}\|\nabla f(x)\|$$

$$E^*[P_{1,\lambda}] = \frac{1}{2n}\, c_{1,\lambda}^2$$

7.3.4 continued

in other terms

$$\check{\varphi}^*_{1,\lambda} = \frac{1}{2} c^2_{1,\lambda} = \frac{1}{2} \check{\sigma}^{*2}$$

$c_{1,\lambda}$: tabulated

For large n and large λ it has been shown that

$$c_{1,\lambda} \sim \sqrt{2 \ln \lambda}$$

$$\check{\varphi}^*_{1,\lambda} \sim \ln \lambda$$

The same holds for the $(1+\lambda)$ ES!

$$\check{\varphi}^*_{1,+\lambda} \sim \ln \lambda$$

$\Rightarrow$ logarithmic speedup
if λ successors are evaluated in parallel

7.3.5 The (μ, λ) ES

Beyer (1994) has shown in a rather lengthly paper:

$$\check{\varphi}^*_{\mu,\lambda} \sim \ln \frac{\lambda}{\mu}$$

$\Rightarrow$ It's the selection pressure that speeds up

Is $\mu = 1$ optimal ? No !

One may even achieve linear speedup by increasing μ at constant λ/μ

with recombination !

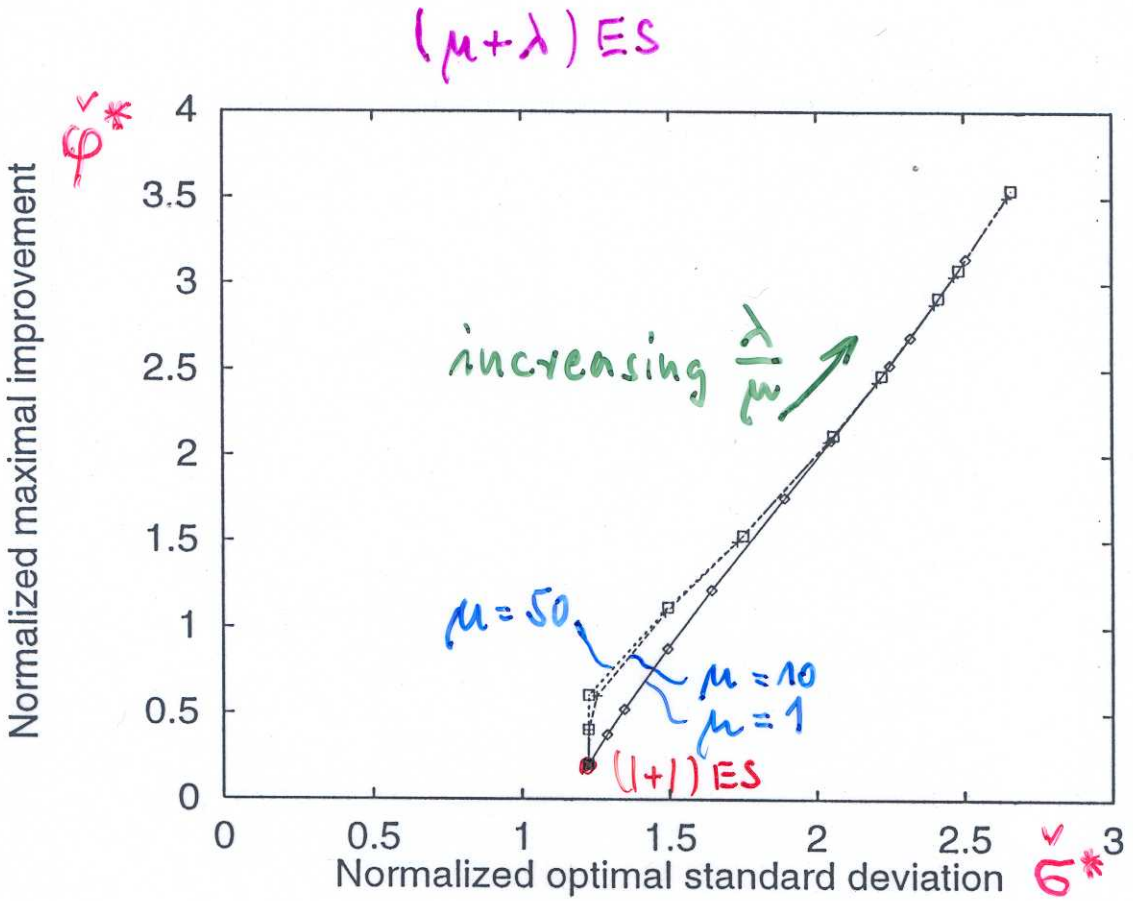
Beyer (1994) investigated panmictic recomb.

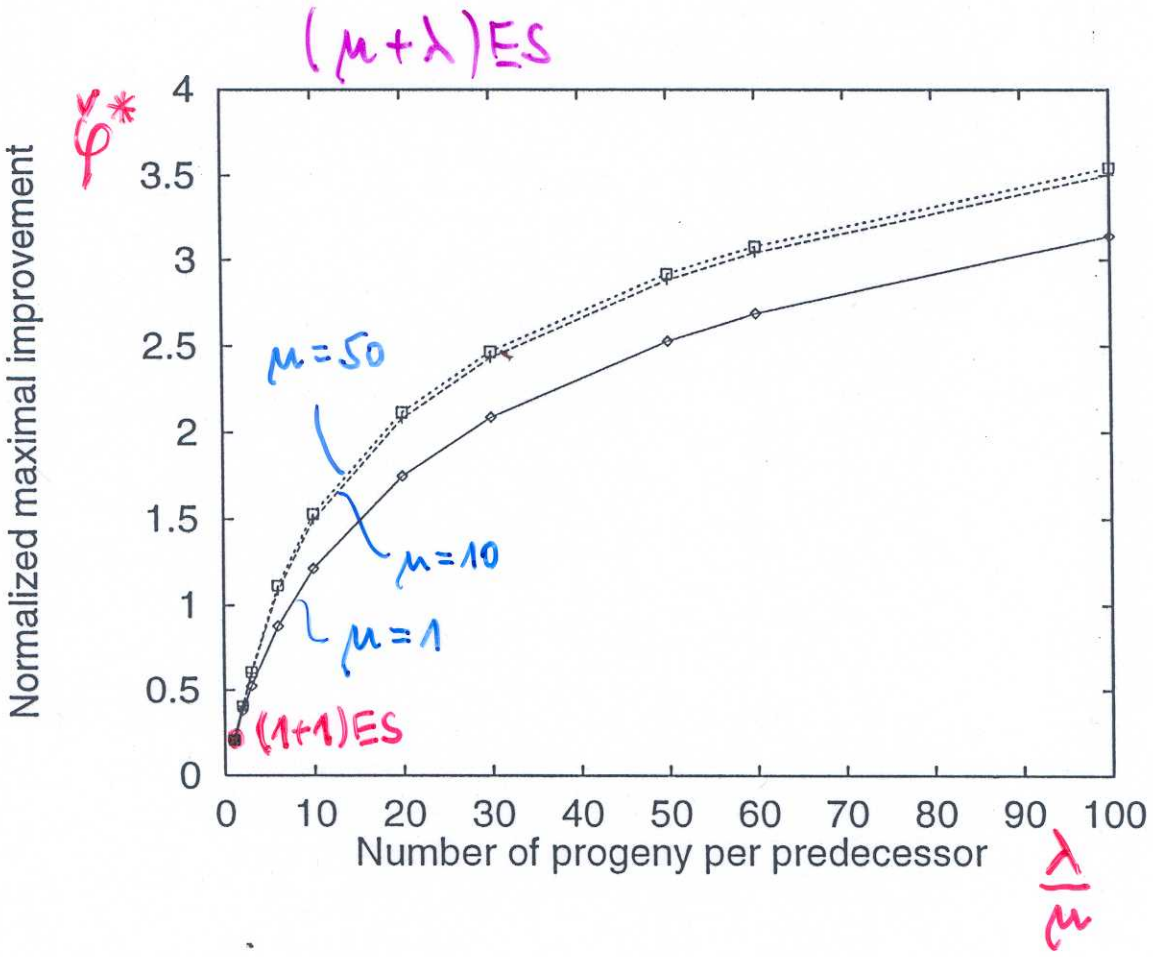
- intermediary (globally)
- global discrete

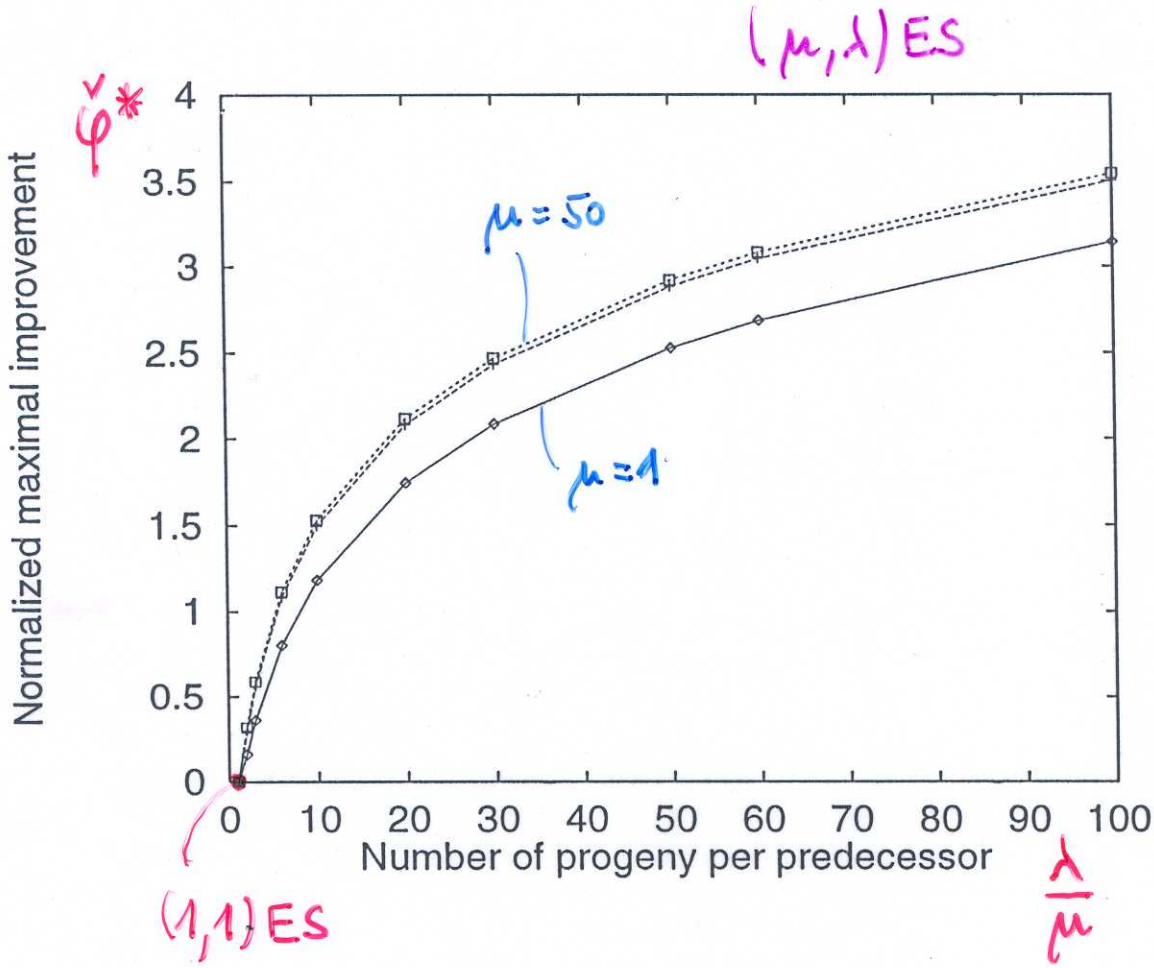
I add a bit speculation (supported by graphs):

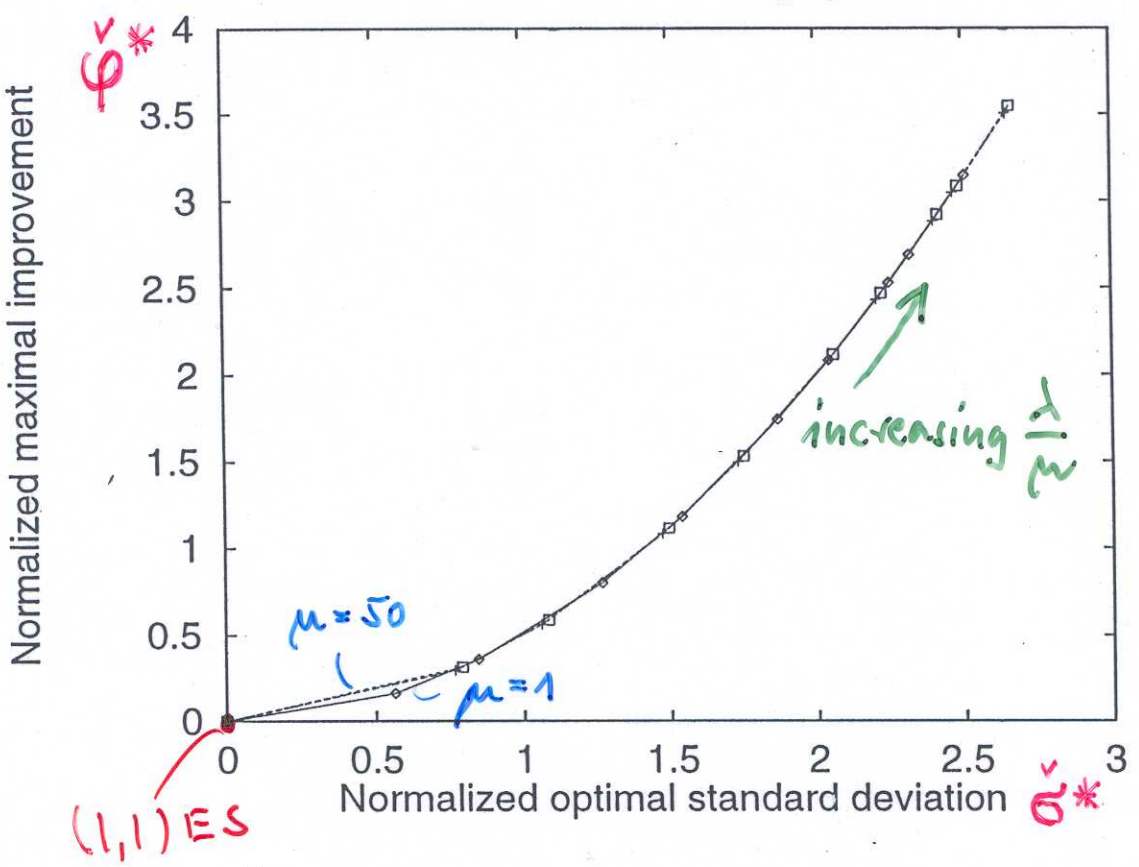
$$\check{\varphi}^*_{\mu, \varkappa=1, \lambda, \varsigma=\mu} \sim \mu \ln \frac{\lambda}{\mu}$$

$$\sigma^* = \begin{cases} \sqrt{2\varphi^*} & \text{for uniform crossover} \\ \sqrt{2\mu\varphi^*} & \text{for global interm. reco.} \end{cases}$$


(μ+λ) ES
φ*
Normalized maximal improvement
4
3.5
3
2.5
2
1.5
1
0.5
0
increasing λ/μ
μ = 50
μ = 10
μ = 1
(1+1) ES
0
0.5
1
1.5
2
2.5
3
Normalized optimal standard deviation
σ*


(μ+λ)ES
φ*
Normalized maximal improvement
4
3.5
3
2.5
2
1.5
1
0.5
0
μ=50
μ=10
μ=1
(1+1)ES
0
10
20
30
40
50
60
70
80
90
100
Number of progeny per predecessor
λ/μ


(μ,λ) ES
φ*
Normalized maximal improvement
4
3.5
3
2.5
2
1.5
1
0.5
0
μ = 50
μ = 1
0
10
20
30
40
50
60
70
80
90
100
Number of progeny per predecessor
λ/μ
(1,1) ES



conclusion:

efficiency mainly depends on selection pressure
(if μ and λ not too small)

without recombination:
population size not thus important
(if μ and λ not too small)

if μ and λ not too small and $\frac{\lambda}{\mu} > 6$:
no difference between
comma-ES and
plus-ES (elitist version)