Foundations of Real-Valued Evolutionary Algorithms

Dirk V. Arnold

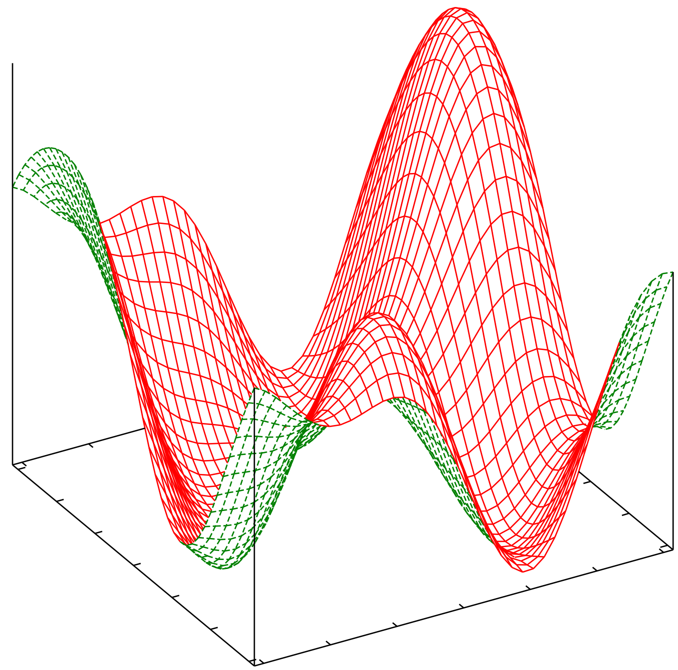
Faculty of Computer Science
Dalhousie University
Halifax, Nova Scotia
Canada

`dirk@cs.dal.ca`

DALHOUSIE UNIVERSITY
Inspiring Minds

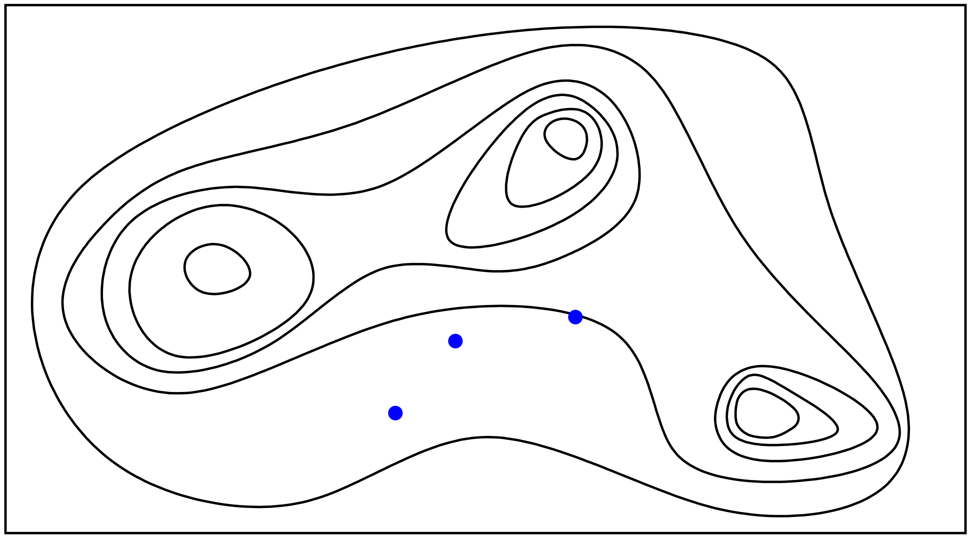
Optimisation Problems

- optimisation problems are abundant in science and engineering
- mathematically, we want to minimise (or maximise) some function $f : D \to R$



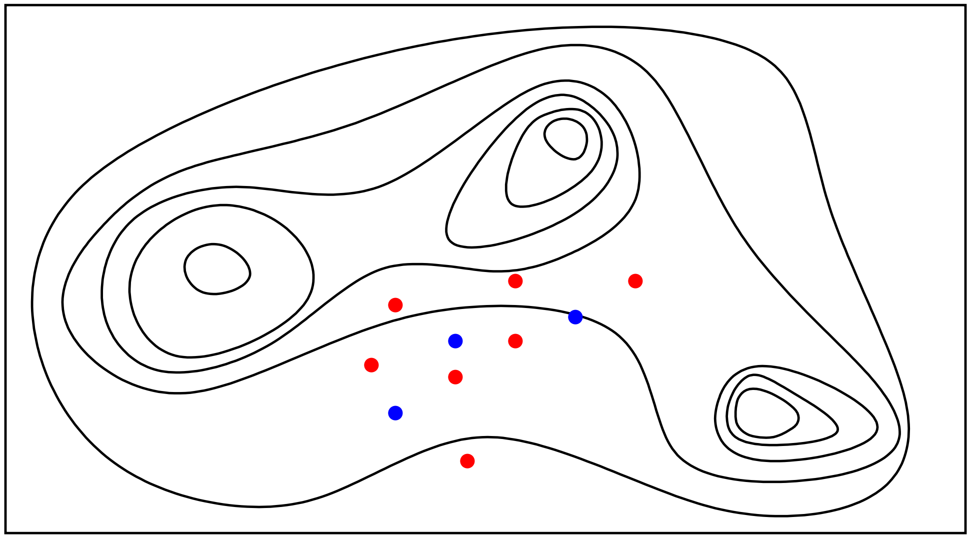
Evolutionary Algorithms

- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function



Evolutionary Algorithms

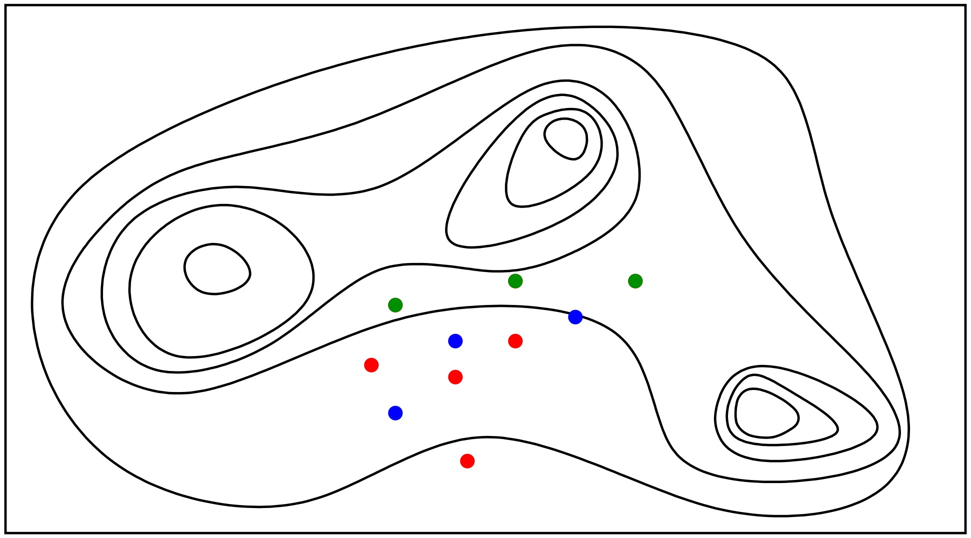
- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function



DALHOUSIE UNIVERSITY Inspiring Minds

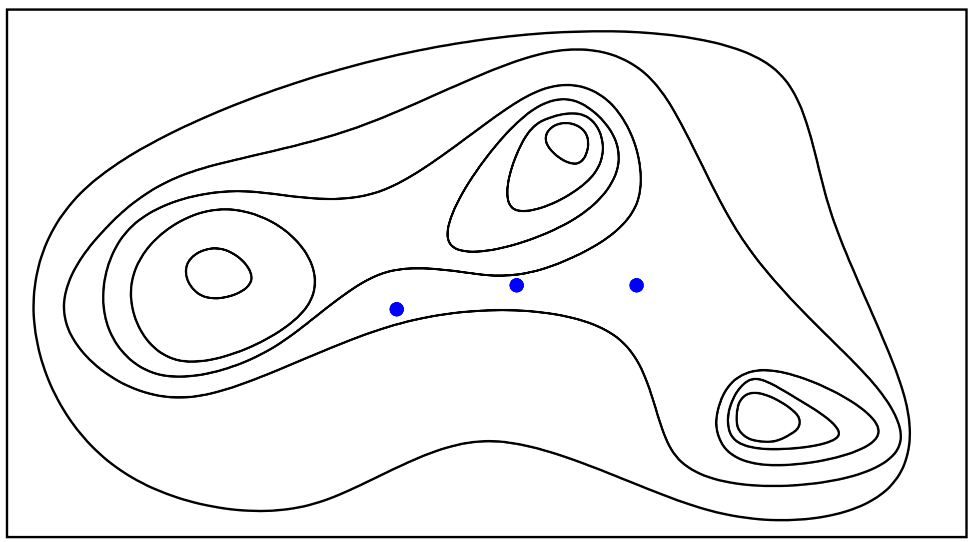
Evolutionary Algorithms

- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function



Evolutionary Algorithms

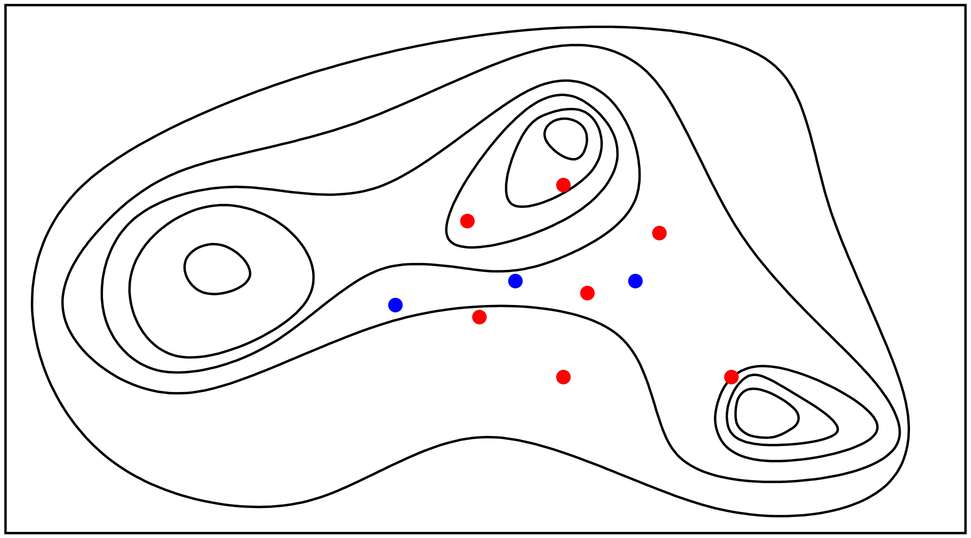
- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function



DALHOUSIE UNIVERSITY
Inspiring Minds

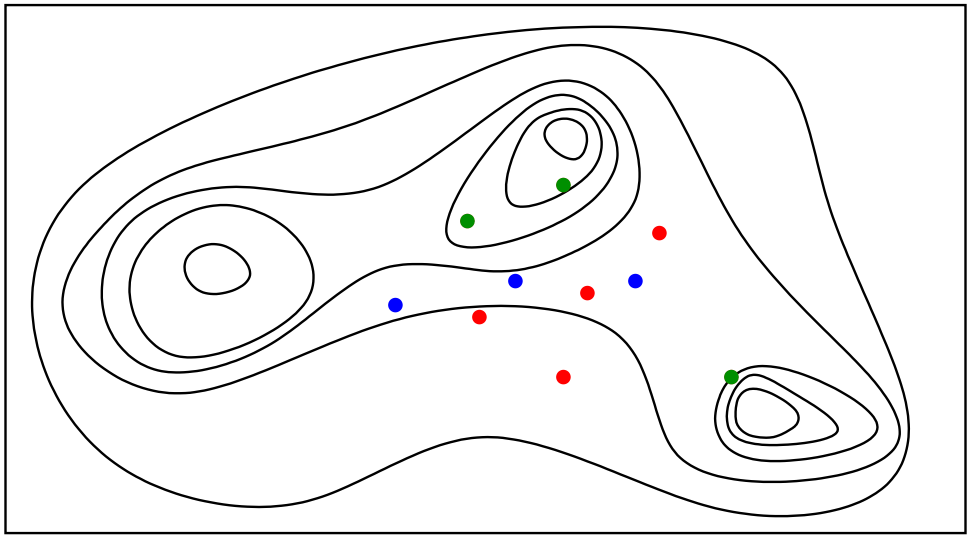
Evolutionary Algorithms

- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function



Evolutionary Algorithms

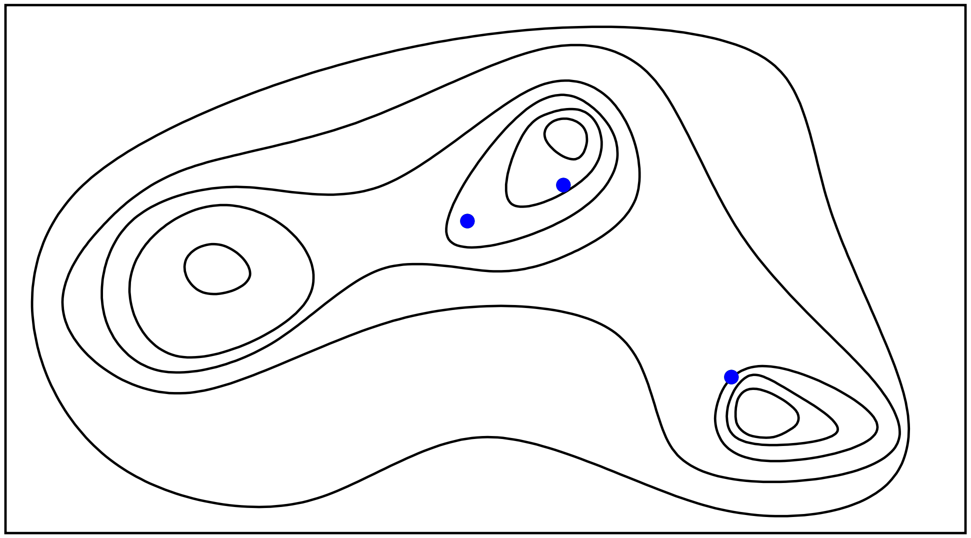
- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function



DALHOUSIE UNIVERSITY Inspiring Minds

Evolutionary Algorithms

- evolutionary algorithms (EAs) are optimisation strategies which derive inspiration from Darwinian evolution
- they model the interplay of variation and selection in a population of individuals
- fitness is determined by the objective function

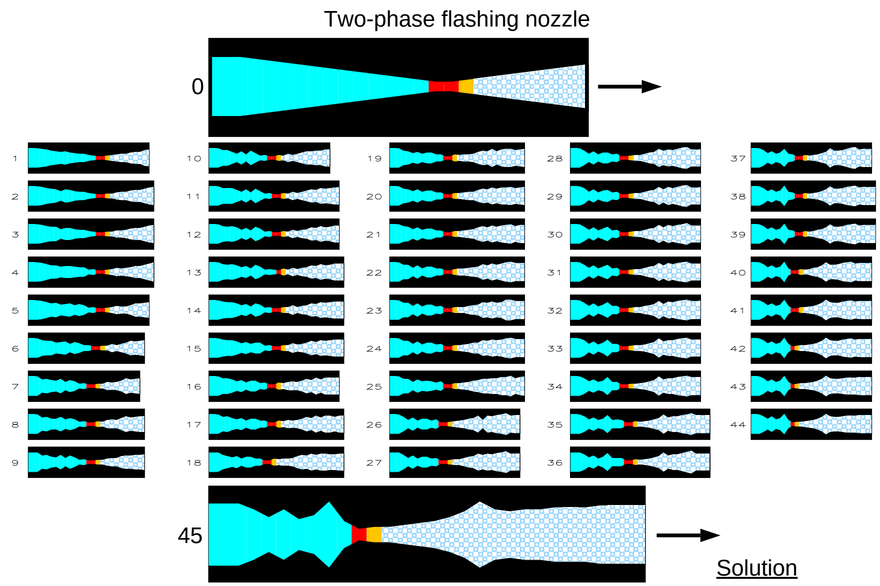


DALHOUSIE UNIVERSITY
Inspiring Minds

Two-phase Jet Nozzle

- optimisation of a single-component two-phase nozzle
- objective: maximisation of efficiency
- 330 compatible segments

  $\Rightarrow$ 10^{60} different configurations
- the efficiency increased from 55% to 80%



J. Klockgether and H.-P. Schwefel, 1970. “Two-phase nozzle and and hollow core jet experiments”, *Proc. 11th Symp. Engineering Aspects of Magnetohydrodynamics*, pp. 141-148.

Optimisation of Coffee Blends (1)

- most coffees are blends of up to ten different kinds of single-origin coffee
- the quality and availability of the single-origin coffees varies from year to year
- brand name coffees have distinct “target tastes” that need to be matched
- experts judge based on criteria such as aroma, brightness, acidity, and body, and achieve the target taste using their experience

DALHOUSIE UNIVERSITY
Inspiring Minds

Optimisation of Coffee Blends (2)

- interactive evolution of coffee blends:
  - replace the experts' heuristics with random steps
  - experts pick the best among a population of five blends
- after 11 generations, the taste of the blend was indistinguishable from the target
- the composition of the blend was very different from what the experts would have chosen
- different blends can have identical flavour; the expert solution is not necessarily the cheapest one

M. Herdy, 1997. "Evolutionary optimisation based on subjective selection – evolving blends of coffee", *Proc. 5th European Congress on Intelligent Techniques and Soft Computing*, pp. 640-644.

Real-Valued Optimisation

- in this tutorial, we consider optimisation problems of the form $f : \mathbb{R}^N \longrightarrow \mathbb{R}$
- unconstrained numerical optimisation algorithms:
  - quasi-Newton and conjugate gradient methods
  - implicit filtering
  - pattern search
  - stochastic approximation
  - response surface methods
  - evolutionary algorithms
  - $\ldots$

Real-Valued Evolutionary Algorithms

- evolutionary algorithms
  - are easy to understand and implement
  - do not rely on derivative information and make no assumptions with regard to the function being optimised
  - typically employ populations
  - involve some element of randomness
  - proceed based on incomplete information
  - strive to be “adaptive”
- types of real-valued EAs include
  - evolution strategies
  - evolutionary programming
  - differential evolution

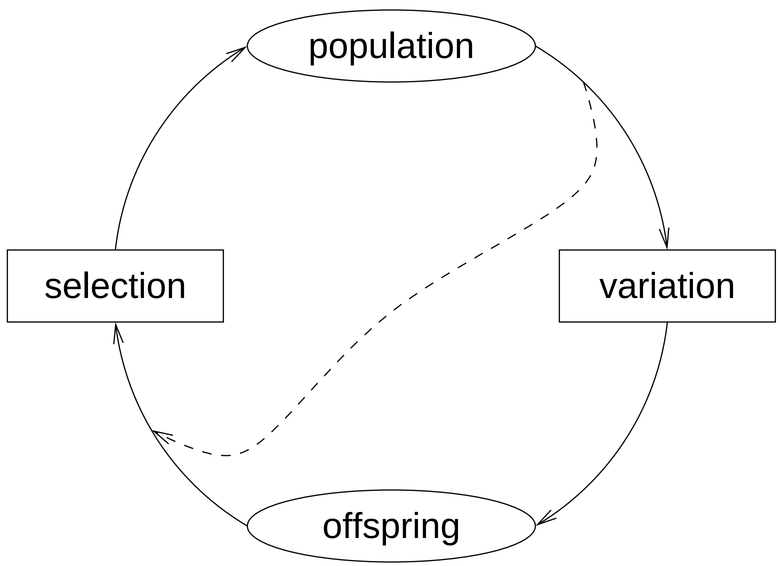
Outline

- evolution strategies
  - the $(\mu/\rho \overset{+}{,} \lambda)$-ES
  - performance for the line model
  - performance for the sphere model
  - performance for other problems
- covariance matrix adaptation
  - CMA-ES
  - other approaches

Benefits of “Keeping It Simple”

- derive scaling laws
- compare with optimal behaviour
- highlight differences between strategy variants; reveal strengths and weaknesses
- recommend parameter settings
- develop intuition with regard to working principles of operators
- develop and improve adaptation strategies

Evolution Strategies: The $(\mu/\rho \overset{+}{,} \lambda)$-ES (1)



- population size: $|\mathcal{P}| = \mu$
- number of offspring: $|\mathcal{Q}| = \lambda$

selection: the μ best candidate solutions in

- $\mathcal{P} \cup \mathcal{Q}$ for plus-selection
- $\mathcal{Q}$ for comma-selection

survive ("truncation selection")

variation: for every offspring to be generated, randomly choose ρ parents; recombine and mutate

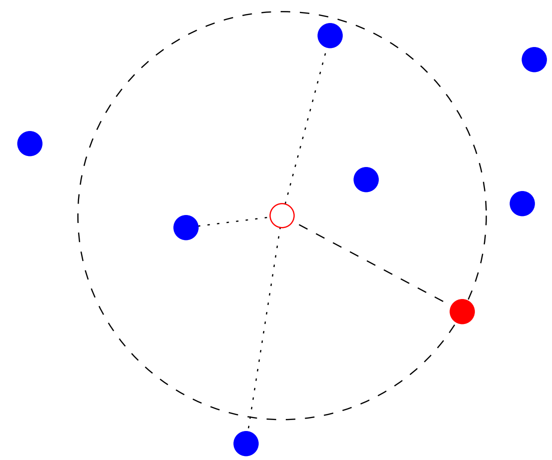
Evolution Strategies: The $(\mu/\rho \overset{+}{,} \lambda)$-ES (2)

recombination: with $i_1, i_2, \ldots, i_\rho$ the indices of the (randomly chosen) parents of a candidate solution to be generated, let

$$\mathbf{x} = \frac{1}{\rho} \sum_{j=1}^{\rho} \mathbf{x}_{i_j}$$

mutation: add a normally distributed random vector with mean zero

$$\mathbf{y} = \mathbf{x} + \sigma \mathbf{z}$$



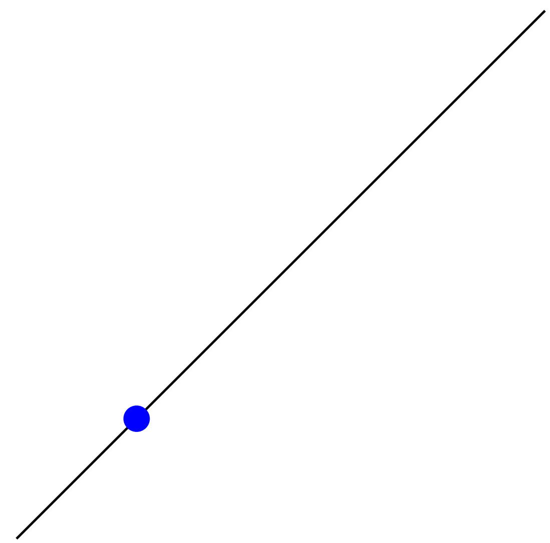
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

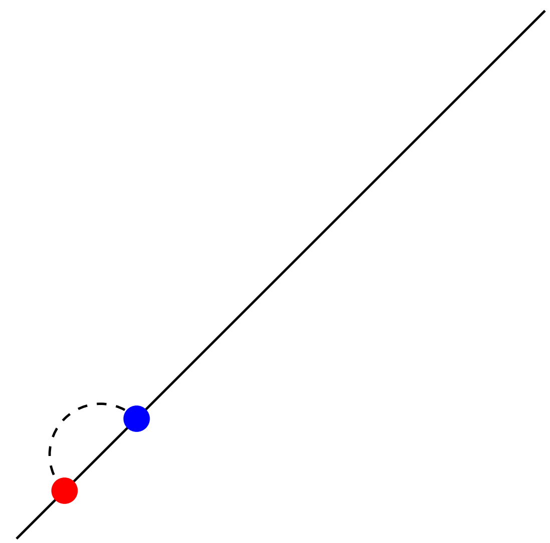
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The $(1+1)$-ES

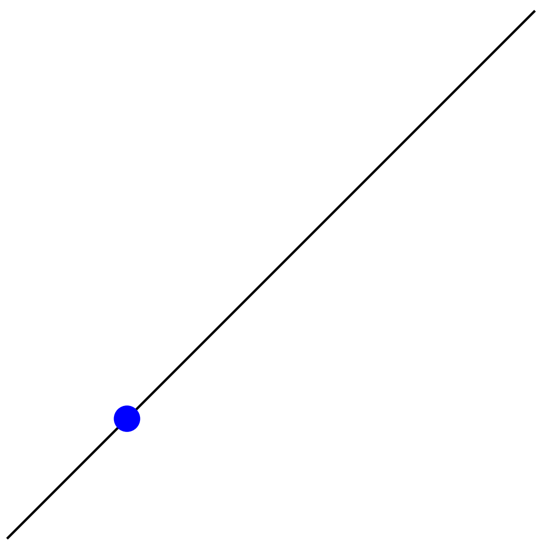
- consider maximisation of objective function $f(x) = x$

- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

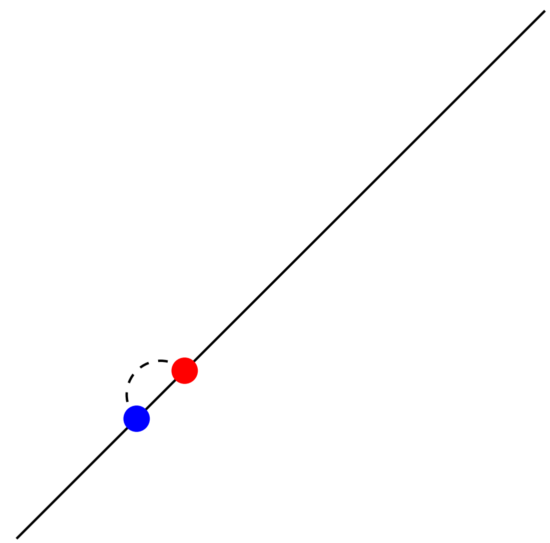
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY
Inspiring Minds

Line Model: The $(1+1)$-ES

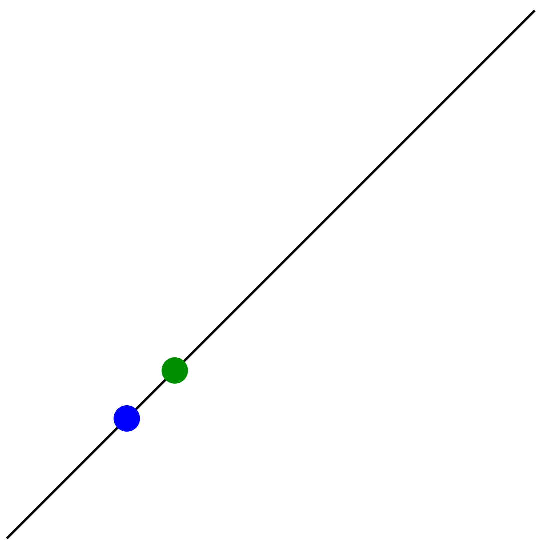
- consider maximisation of objective function $f(x) = x$

- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The $(1+1)$-ES

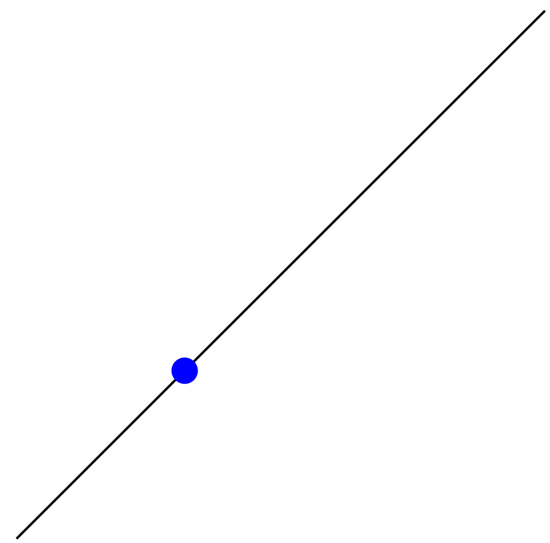
- consider maximisation of objective function $f(x) = x$

- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



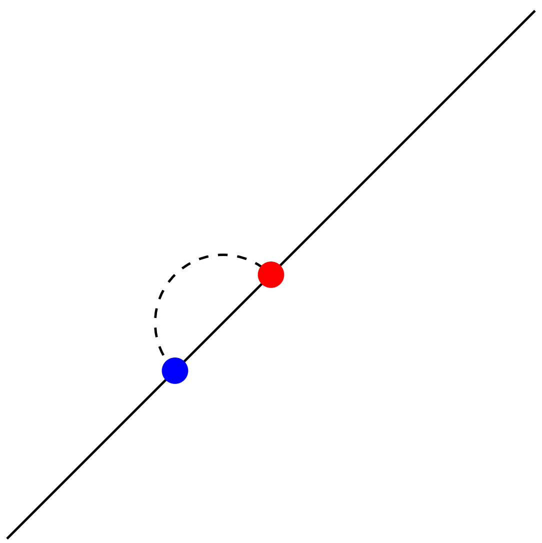
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY
Inspiring Minds

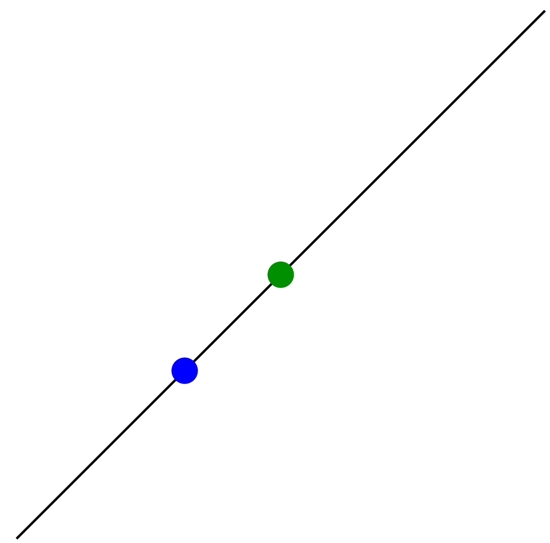
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY Inspiring Minds

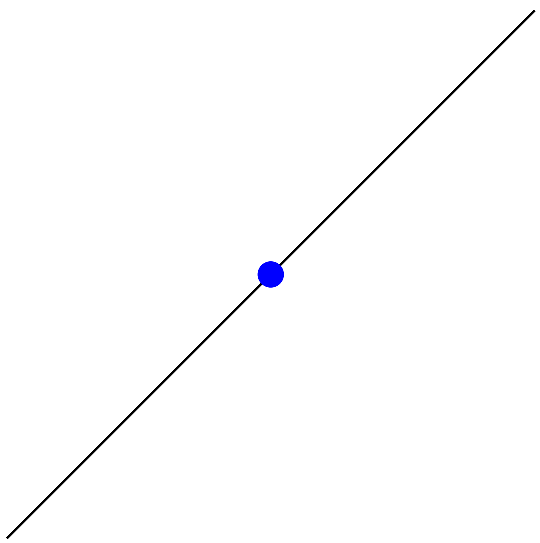
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



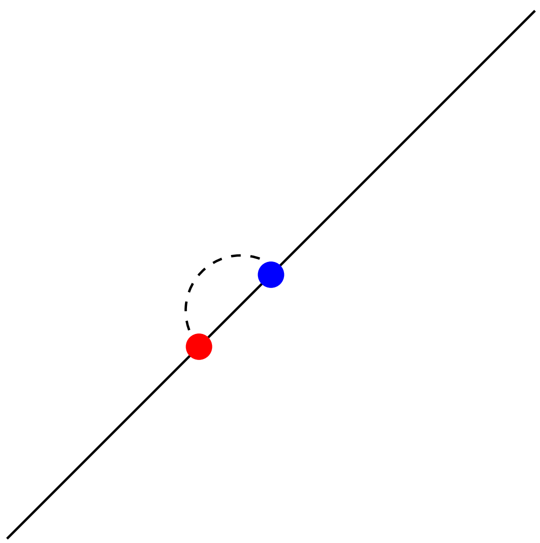
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



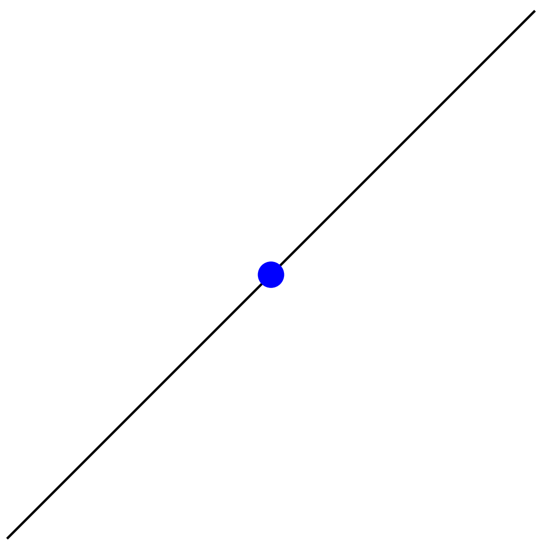
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^{\infty} z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



Line Model: The $(1+1)$-ES

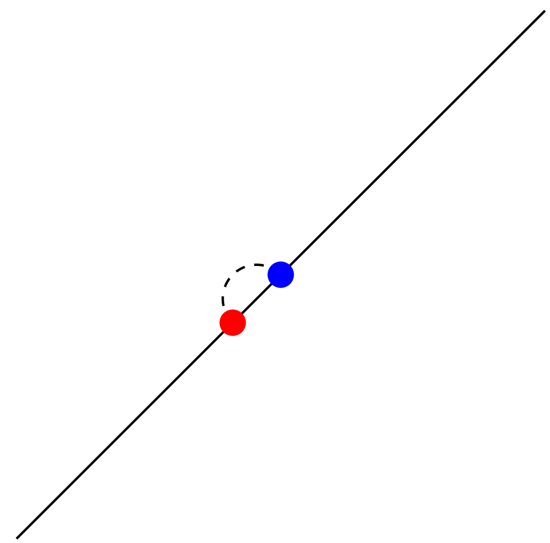
- consider maximisation of objective function $f(x) = x$

- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^{\infty} z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY
Inspiring Minds

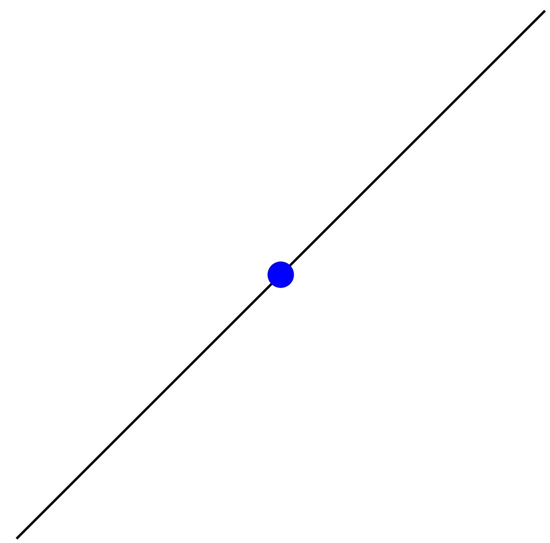
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY
Inspiring Minds

Line Model: The $(1+1)$-ES

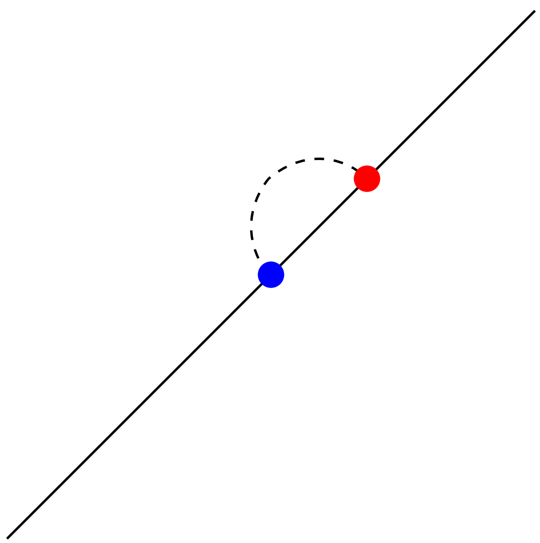
- consider maximisation of objective function $f(x) = x$

- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

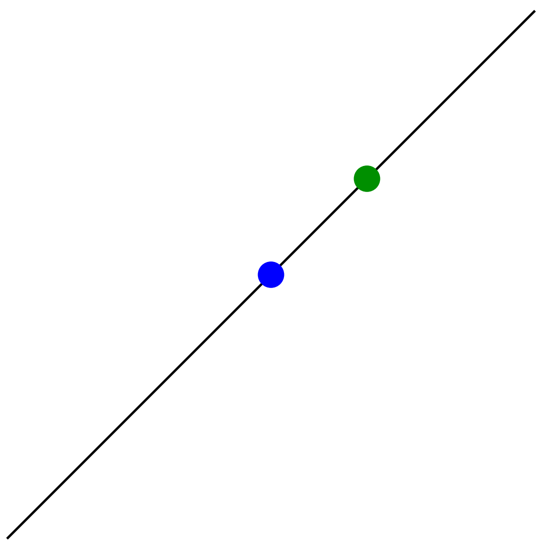
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\,\mathrm{e}^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$



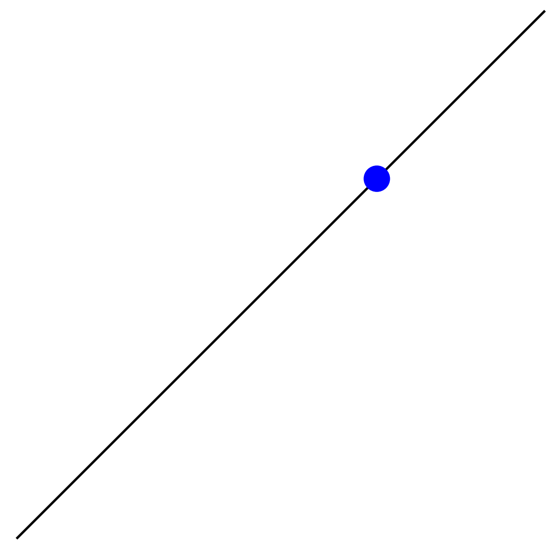
Line Model: The $(1+1)$-ES

- consider maximisation of objective function $f(x) = x$
- $(1+1)$-ES:

$$x^{(t+1)} = \begin{cases} x^{(t)} + \sigma z & \text{if } f(x^{(t)} + \sigma z) > f(x^{(t)}) \\ x^{(t)} & \text{otherwise} \end{cases}$$

- progress rate:

$$\begin{aligned} \varphi_{1+1} &= \mathrm{E}\left[x^{(t+1)} - x^{(t)}\right] \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_0^\infty z\, \mathrm{e}^{-\frac{1}{2}z^2}\, \mathrm{d}z \\ &= \frac{\sigma}{\sqrt{2\pi}} \end{aligned}$$

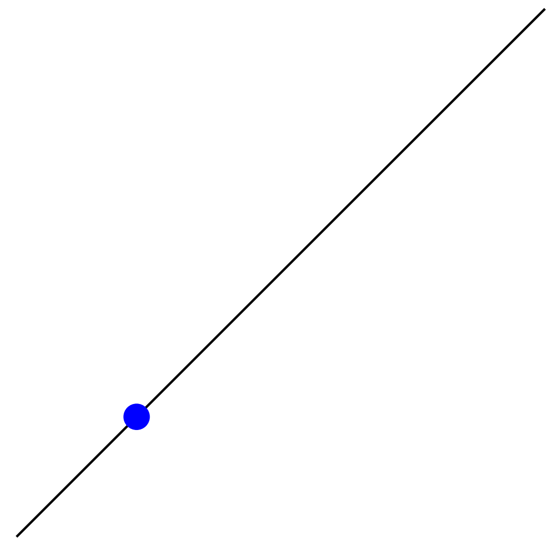


Line Model: The $(1, \lambda)$-ES

- $(1, \lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

  where $k;\lambda$ denotes the index of the kth best offspring



- progress rate:

$$\begin{aligned}\varphi_{1,\lambda} &= \sigma \mathrm{E}[z_{1;\lambda}] \\ &\equiv \sigma c_{1,\lambda}\end{aligned}$$

- $c_{1,\lambda}$ is referred to as the $(1, \lambda)$-progress coefficient

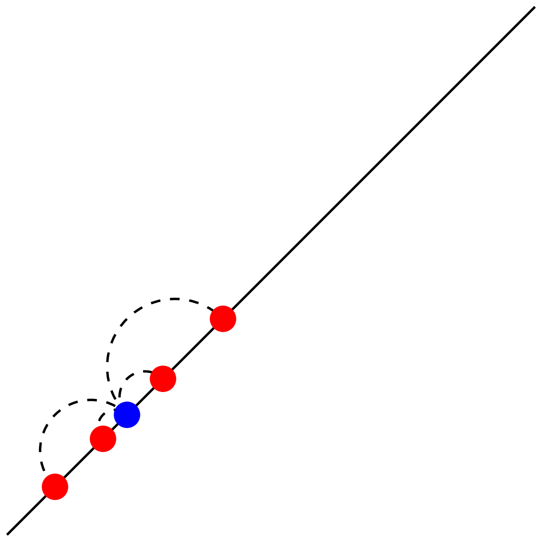
DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The $(1, \lambda)$-ES

- $(1, \lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

where $k; \lambda$ denotes the index of the kth best offspring



- progress rate:

$$\varphi_{1,\lambda} = \sigma \mathrm{E}[z_{1;\lambda}] \equiv \sigma c_{1,\lambda}$$

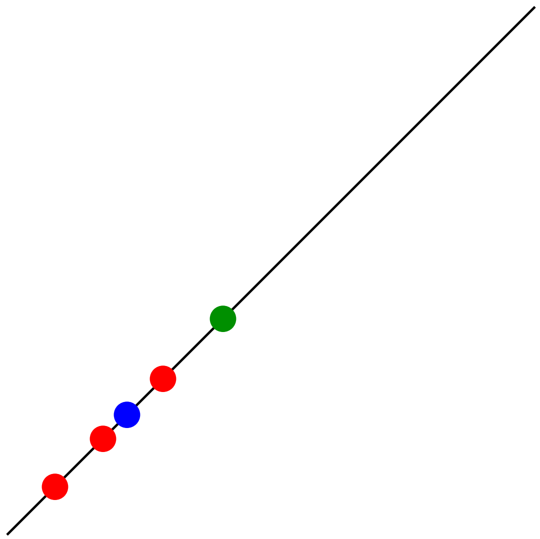
- $c_{1,\lambda}$ is referred to as the $(1, \lambda)$-progress coefficient

Line Model: The $(1, \lambda)$-ES

- $(1, \lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

where $k;\lambda$ denotes the index of the kth best offspring



- progress rate:

$$\begin{aligned}\varphi_{1,\lambda} &= \sigma \mathrm{E}[z_{1;\lambda}] \\ &\equiv \sigma c_{1,\lambda}\end{aligned}$$

- $c_{1,\lambda}$ is referred to as the $(1, \lambda)$-progress coefficient

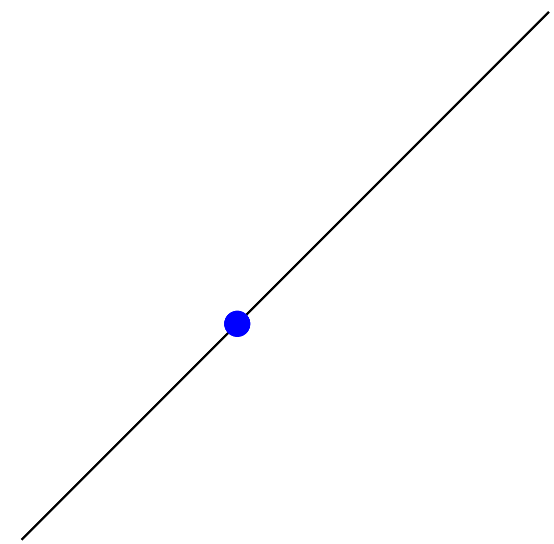
DALHOUSIE UNIVERSITY Inspiring Minds

Line Model: The $(1, \lambda)$-ES

- $(1, \lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

  where $k;\lambda$ denotes the index of the kth best offspring



- progress rate:

$$\begin{aligned}\varphi_{1,\lambda} &= \sigma \mathrm{E}[z_{1;\lambda}] \\ &\equiv \sigma c_{1,\lambda}\end{aligned}$$

- $c_{1,\lambda}$ is referred to as the $(1, \lambda)$-progress coefficient

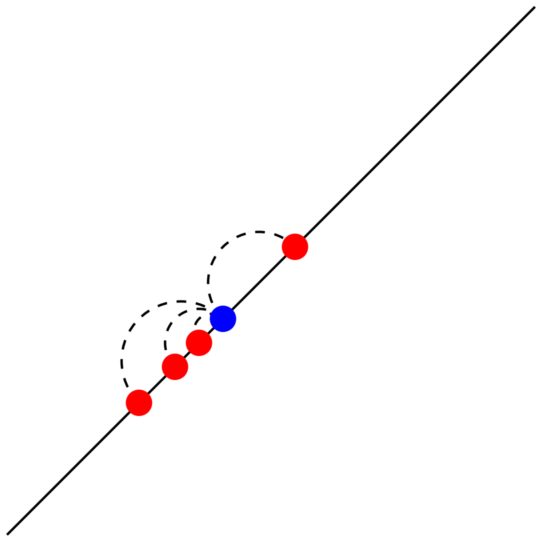
DALHOUSIE UNIVERSITY
Inspiring Minds

Line Model: The $(1, \lambda)$-ES

- $(1, \lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

where $k;\lambda$ denotes the index of the kth best offspring



- progress rate:

$$\begin{aligned} \varphi_{1,\lambda} &= \sigma \mathrm{E}[z_{1;\lambda}] \\ &\equiv \sigma c_{1,\lambda} \end{aligned}$$

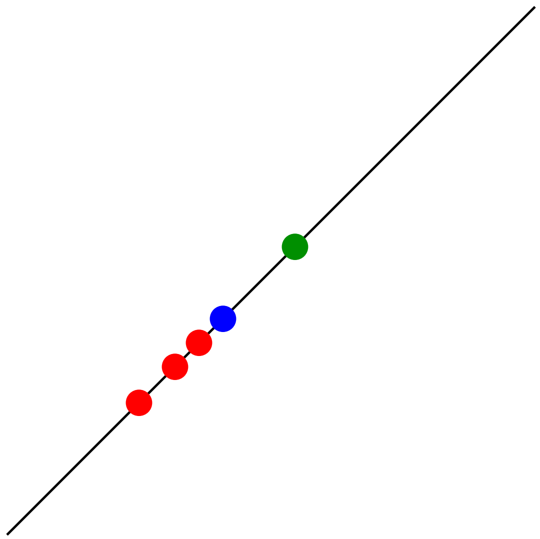
- $c_{1,\lambda}$ is referred to as the $(1, \lambda)$-progress coefficient

Line Model: The $(1,\lambda)$-ES

- $(1,\lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

where $k;\lambda$ denotes the index of the kth best offspring



- progress rate:

$$\varphi_{1,\lambda} = \sigma \mathrm{E}[z_{1;\lambda}]$$
$$\equiv \sigma c_{1,\lambda}$$

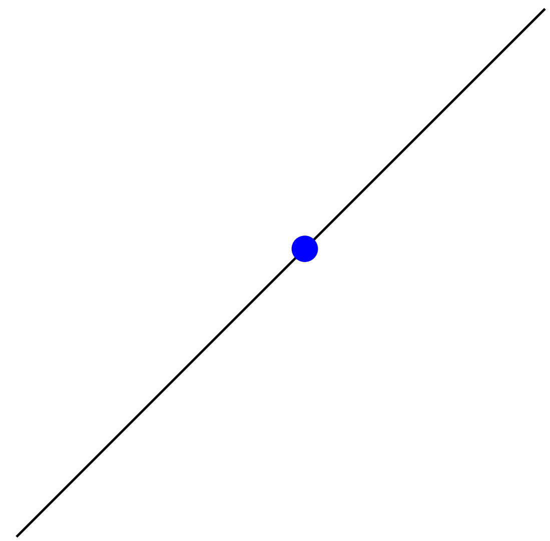
- $c_{1,\lambda}$ is referred to as the $(1,\lambda)$-progress coefficient

Line Model: The $(1, \lambda)$-ES

- $(1, \lambda)$-ES:

$$x^{(t+1)} = x^{(t)} + \sigma z_{1;\lambda}$$

where $k;\lambda$ denotes the index of the kth best offspring



- progress rate:

$$\begin{aligned}\varphi_{1,\lambda} &= \sigma \mathrm{E}[z_{1;\lambda}] \\ &\equiv \sigma c_{1,\lambda}\end{aligned}$$

- $c_{1,\lambda}$ is referred to as the $(1, \lambda)$-progress coefficient

Line Model: The $(1, \lambda)$-ES

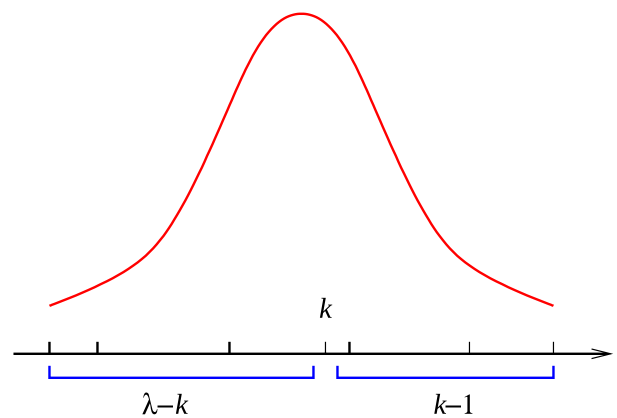
- $z_{k;\lambda}$ is the $(\lambda + 1 - k)$th order statistic of a sample of λ independent, standard normally distributed random variables

$$\mathrm{E}[z_{k;\lambda}] = \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, \mathrm{e}^{-\frac{1}{2}z^2} [\Phi(z)]^{\lambda-k} [1 - \Phi(z)]^{k-1}\, \mathrm{d}z$$

- for large λ,

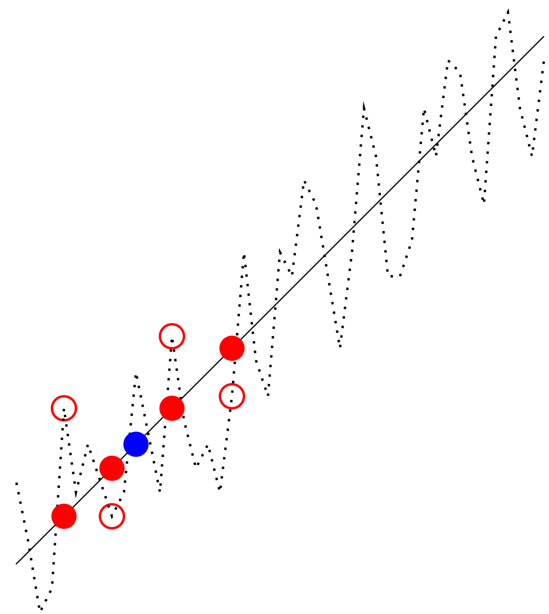
$$c_{1,\lambda} \propto \sqrt{2 \log \lambda}$$

$\Rightarrow$ the growth of the progress rate with λ is very slow



H.-G. Beyer, 2001. *The Theory of Evolution Strategies*, Springer.

Noisy Line Model: The $(1, \lambda)$-ES



- sources of noise in real-world problems:
  - inaccurate measurements
  - Monte Carlo methods
  - subjective selection
  - …
- noisy fitness: $f_\epsilon(\mathbf{x}) = f(\mathbf{x}) + \sigma_\epsilon z_\epsilon$
- progress rate:

$$\varphi_{1,\lambda} = \frac{\sigma c_{1,\lambda}}{\sqrt{1 + \vartheta^2}}$$

$\vartheta = \sigma_\epsilon / \sigma$: noise-to-signal ratio

$\Rightarrow$ larger steps reduce the noise-to-signal ratio by amplifying the signal

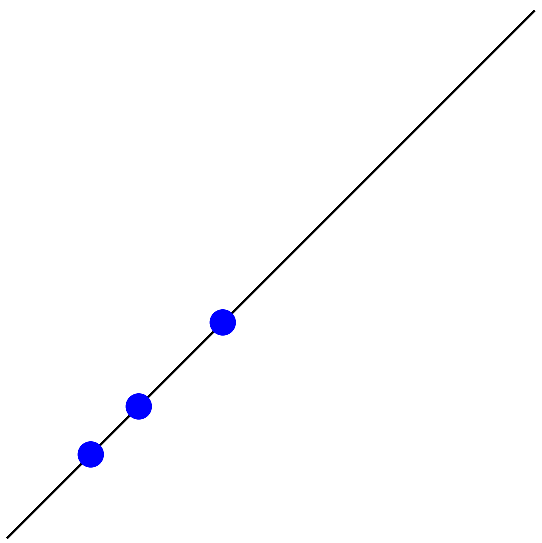
Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination
- problem: the distribution of the population in search space needs to be modelled
- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



H.-G. Beyer, 1995. "Toward a theory of evolution strategies: The (μ, λ)-theory", *Evolutionary Computation*, 2(4):381-407.

DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination

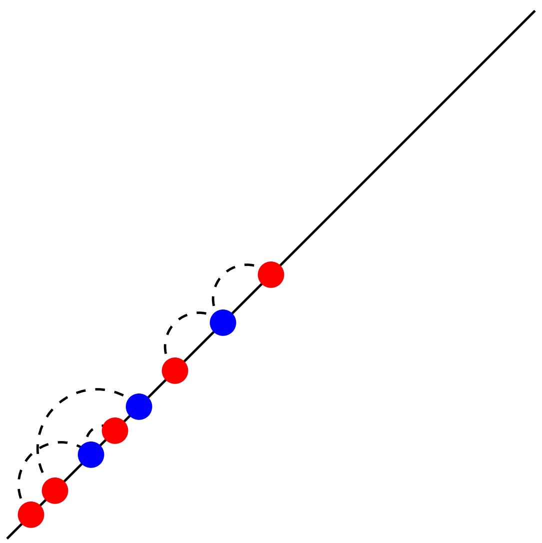
- problem: the distribution of the population in search space needs to be modelled

- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



H.-G. Beyer, 1995. "Toward a theory of evolution strategies: The (μ, λ)-theory", *Evolutionary Computation*, 2(4):381-407.

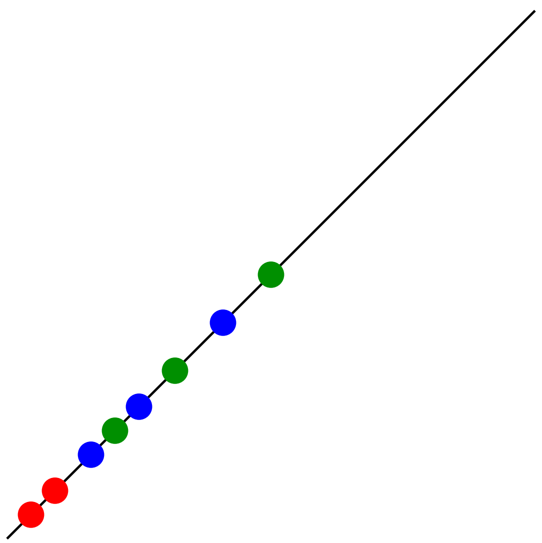
Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination
- problem: the distribution of the population in search space needs to be modelled
- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



H.-G. Beyer, 1995. "Toward a theory of evolution strategies: The (μ, λ)-theory", *Evolutionary Computation*, 2(4):381-407.

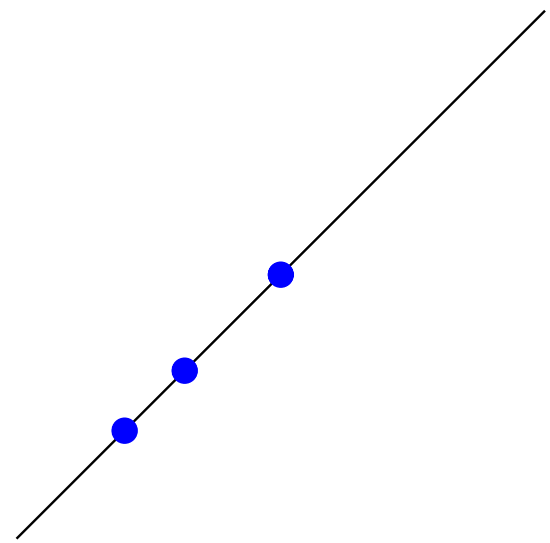
Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination
- problem: the distribution of the population in search space needs to be modelled
- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



H.-G. Beyer, 1995. "Toward a theory of evolution strategies: The (μ, λ)-theory", *Evolutionary Computation*, 2(4):381-407.

DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination

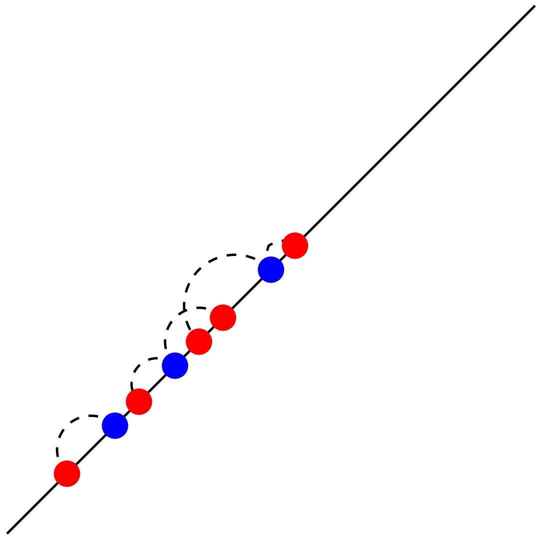
- problem: the distribution of the population in search space needs to be modelled

- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



H.-G. Beyer, 1995. "Toward a theory of evolution strategies: The (μ, λ)-theory", *Evolutionary Computation*, 2(4):381-407.

DALHOUSIE UNIVERSITY Inspiring Minds

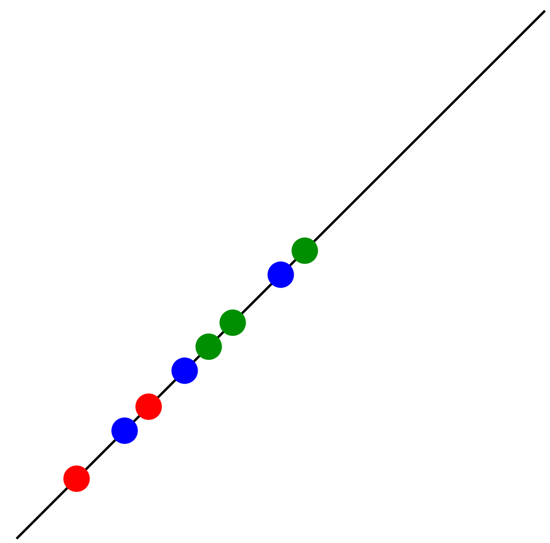
Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination
- problem: the distribution of the population in search space needs to be modelled
- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



H.-G. Beyer, 1995. “Toward a theory of evolution strategies: The (μ, λ)-theory”, *Evolutionary Computation*, 2(4):381-407.

Line Model: The (μ, λ)-ES

- (μ, λ)-ES: short for $(\mu/1, \lambda)$-ES; no recombination

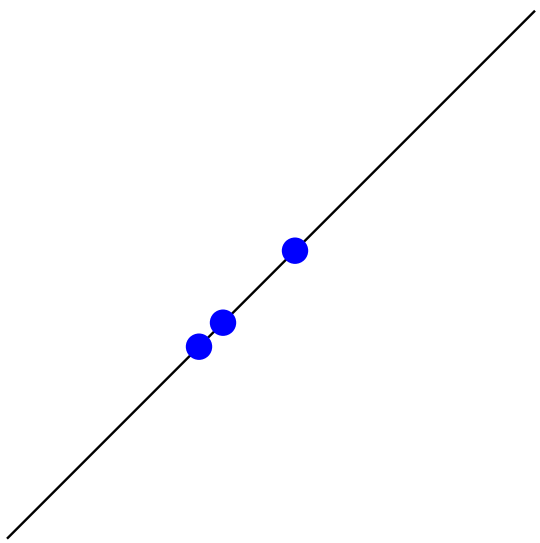
- problem: the distribution of the population in search space needs to be modelled

- progress rate:

$$\varphi_{\mu,\lambda} = \sigma c_{\mu,\lambda}$$

- it can be shown that $c_{\mu,\lambda} \leq c_{1,\lambda}$ for any μ

  $\Rightarrow$ keeping any but the best offspring deteriorates performance



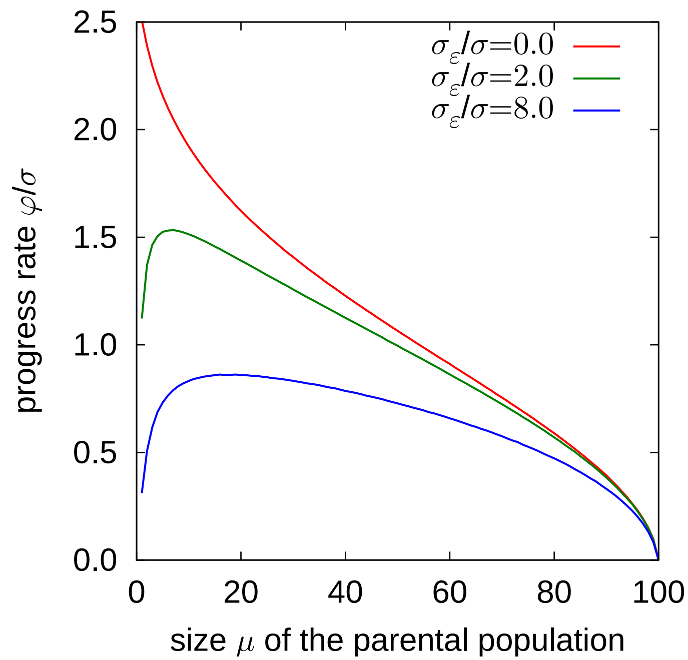
H.-G. Beyer, 1995. "Toward a theory of evolution strategies: The (μ, λ)-theory", *Evolutionary Computation*, 2(4):381-407.

Noisy Line Model: The (μ, λ)-ES

- the signal strength is $\sqrt{\sigma^2 + D^2}$, where D^2 is the variance of the population

- D is proportional to σ and increases with increasing μ

  $\Rightarrow$ increasing the size of the population decreases the noise-to-signal ratio



D. V. Arnold and H.-G. Beyer, 2001. "Investigation of the (μ, λ)-ES in the presence of noise", *2001 IEEE Congress on Evolutionary Computation*, pp. 332-339.

D. V. Arnold and H.-G. Beyer, 2003. "On the benefits of populations for noisy optimization", *Evolutionary Computation*, 11(2):111-127.

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

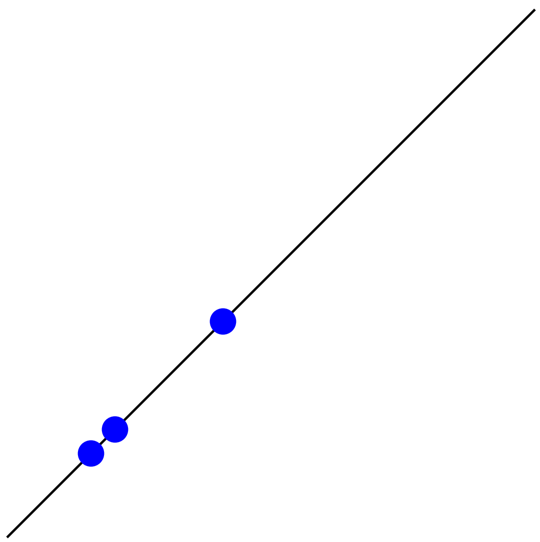
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



DALHOUSIE UNIVERSITY Inspiring Minds

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

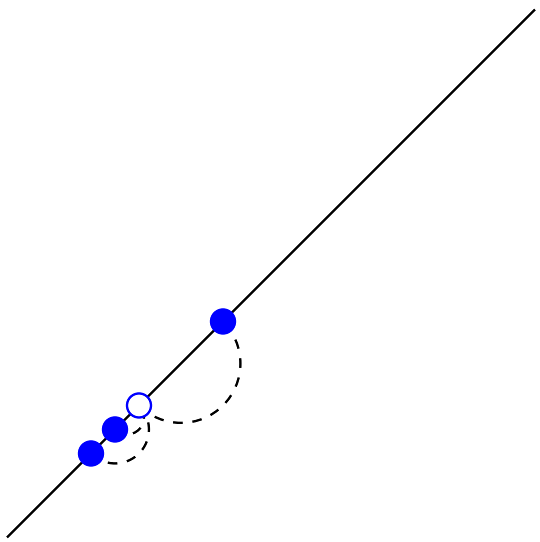
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

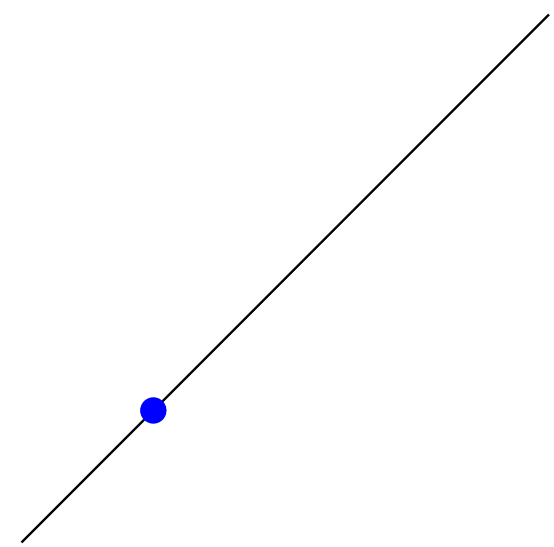
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1 + \vartheta^2}}$$



Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

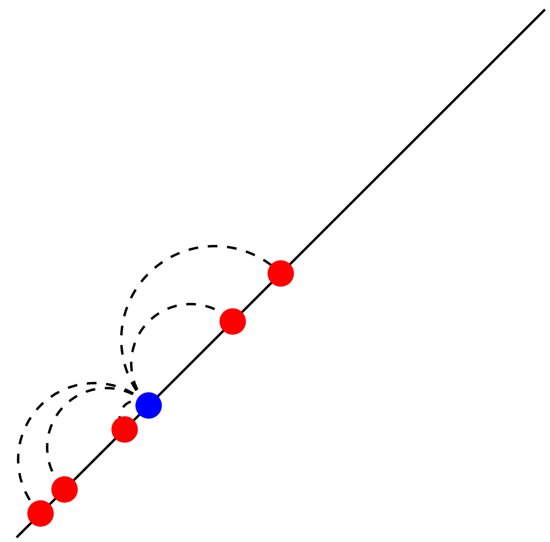
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

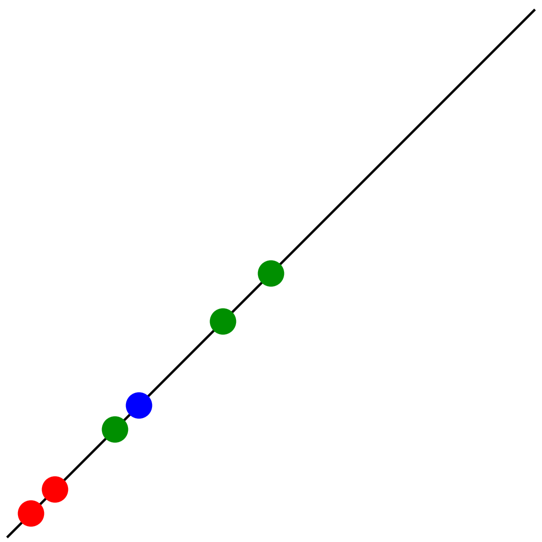
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu}\sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



DALHOUSIE UNIVERSITY Inspiring Minds

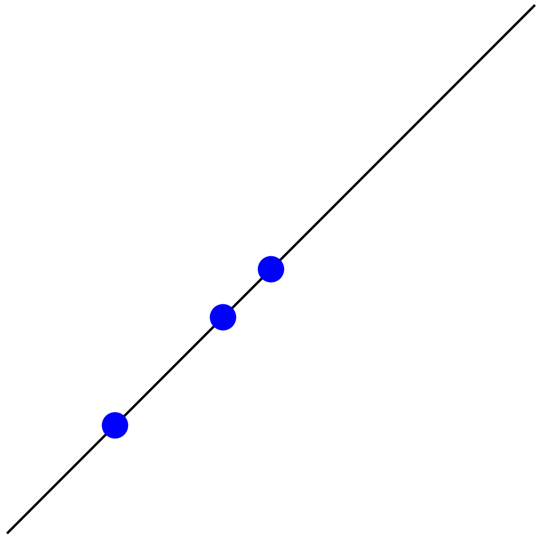
Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu}\sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$



- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$

DALHOUSIE UNIVERSITY *Inspiring Minds*

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

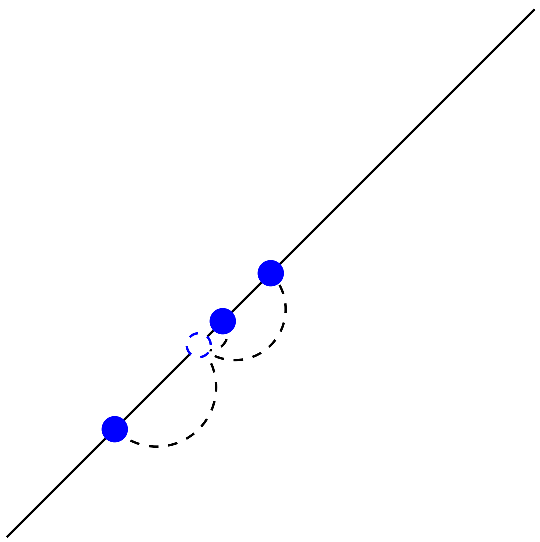
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



DALHOUSIE UNIVERSITY
Inspiring Minds

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

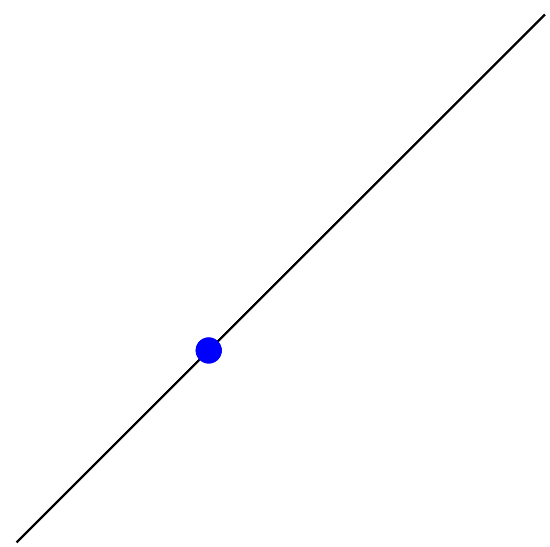
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



Line Model: The $(\mu/\mu,\lambda)$-ES

- $(\mu/\mu,\lambda)$-ES: recombination of $\rho=\mu$ parents contracts the population to a point

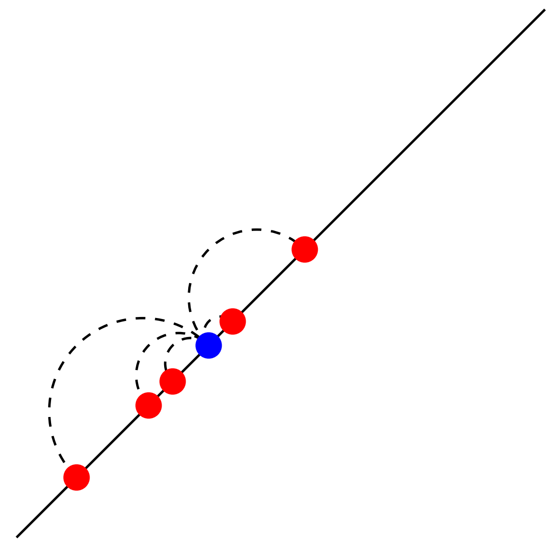
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu}\sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$



Dalhousie University Inspiring Minds

Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

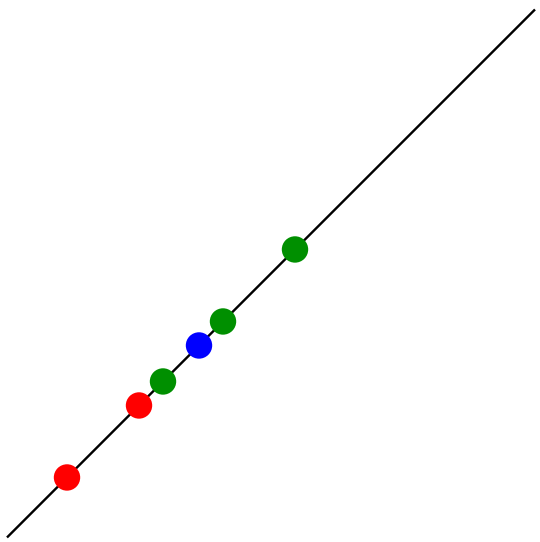
$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$

- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1 + \vartheta^2}}$$



DALHOUSIE UNIVERSITY
Inspiring Minds

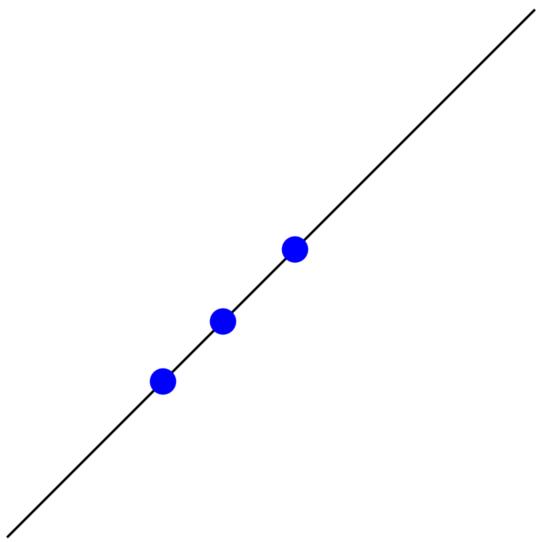
Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-ES: recombination of $\rho = \mu$ parents contracts the population to a point

$$x^{(t+1)} = x^{(t)} + \frac{\sigma}{\mu} \sum_{k=1}^{\mu} z_{k;\lambda}$$

- progress rate:

$$\varphi_{\mu/\mu,\lambda} = \sigma c_{\mu/\mu,\lambda}$$



- in the presence of noise:

$$\varphi_{\mu/\mu,\lambda} = \frac{\sigma c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}}$$

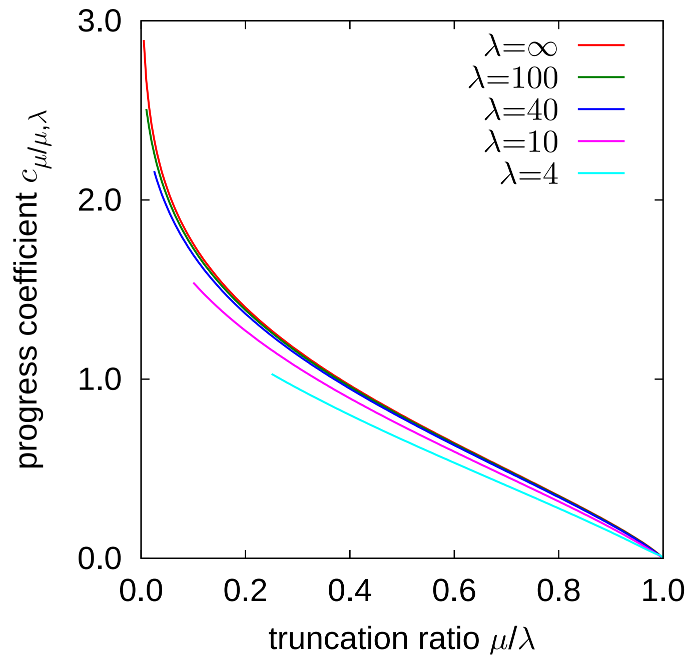
Line Model: The $(\mu/\mu, \lambda)$-ES

- $(\mu/\mu, \lambda)$-progress coefficient:

$$c_{\mu/\mu,\lambda} = \frac{\lambda - \mu}{2\pi} \binom{\lambda}{\mu} \int_{-\infty}^{\infty} z\, \mathrm{e}^{-z^2} [\Phi(z)]^{\lambda-\mu-1} [1 - \Phi(z)]^{\mu-1}\, \mathrm{d}z$$

- in general, $c_{\mu/\mu,\lambda} \lessapprox c_{\mu,\lambda} \leq c_{1,\lambda}$

H.-G. Beyer, 1995. “Toward a theory of evolution strategies: On the benefit of sex – the $(\mu/\mu, \lambda)$-theory”, *Evolutionary Computation*, 3(1):81-111.



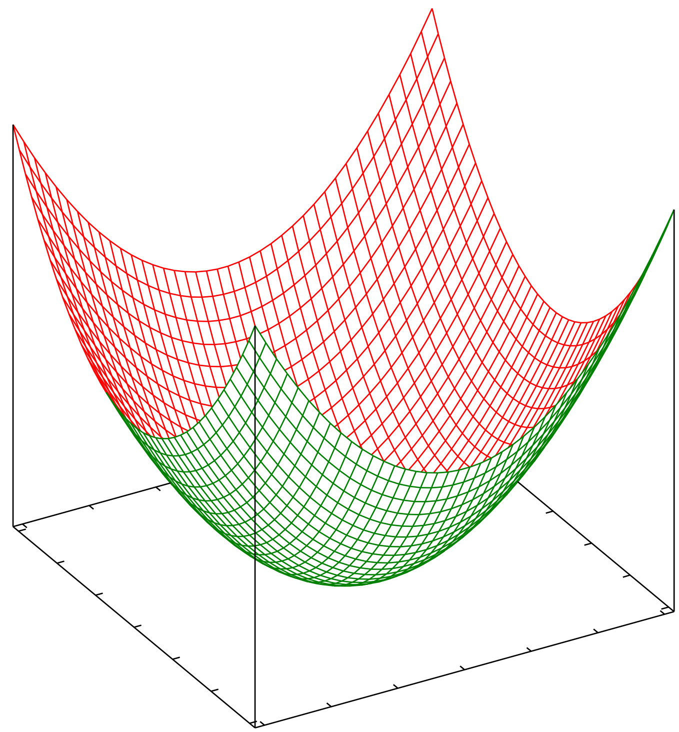
DALHOUSIE UNIVERSITY Inspiring Minds

The Sphere Model (1)

- sphere model: minimise

$$f(\mathbf{x}) = \sum_{i=1}^{N} x_i^2$$

- assume that N is large



I. Rechenberg, 1994. *Evolutionsstrategie '94*, Frommann-Holzboog.

H.-G. Beyer, 2001. *The Theory of Evolution Strategies*, Springer.

D. V. Arnold, 2002. *Noisy Optimization with Evolution Strategies*, Kluwer.

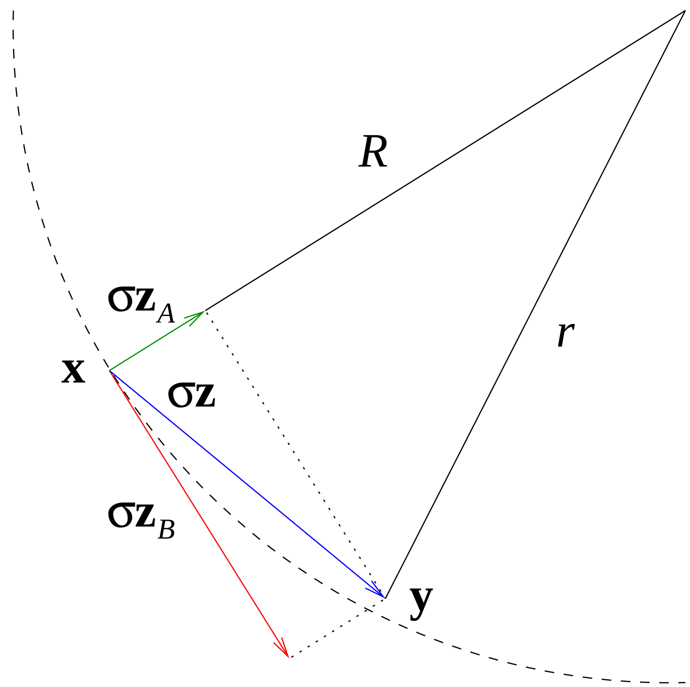
The Sphere Model (2)

- consider candidate solution

$$\mathbf{y} = \mathbf{x} + \sigma \mathbf{z}$$

- fitness:

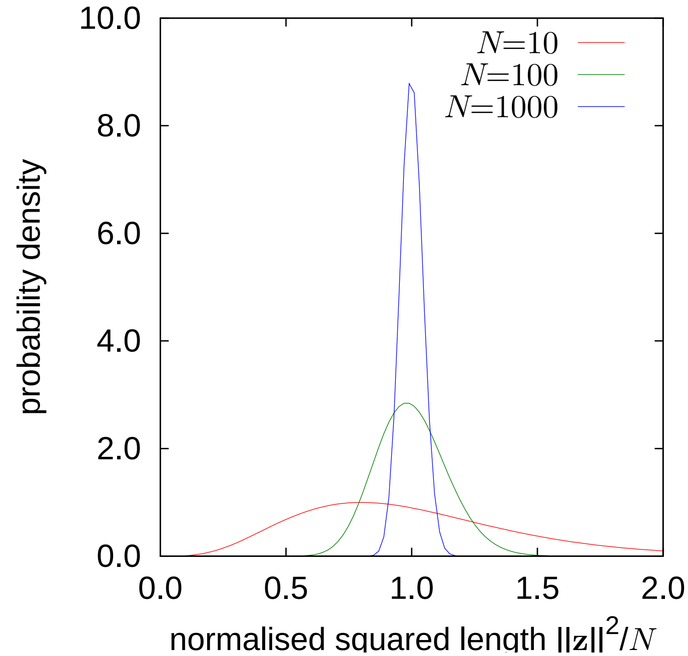
$$\begin{aligned} f(\mathbf{y}) &= r^2 \\ &= (R - \sigma z_A)^2 + \sigma^2 \|\mathbf{z}_B\|^2 \\ &= R^2 - 2R\sigma z_A + \sigma^2 \|\mathbf{z}\|^2 \\ &= f(\mathbf{x}) - 2R\sigma z_A + \sigma^2 \|\mathbf{z}\|^2 \end{aligned}$$



- $\sigma^2 \|\mathbf{z}\|^2$ deteriorates fitness, limiting useful mutation strengths

The Sphere Model (3)

- if $\mathbf{z}$ is a mutation vector, then
  1. z_A is standard normally distributed
  2. $\|\mathbf{z}\|^2$ is χ^2_N-distributed
- the χ^2_N-distribution has mean N and standard deviation $\sqrt{2N}$



- the coefficient of variation of the χ^2_N-distribution tends to zero as N increases

  $\Rightarrow$ for the range of mutation strengths of interest and large enough N, $\|\mathbf{z}\|^2$ can be replaced with N

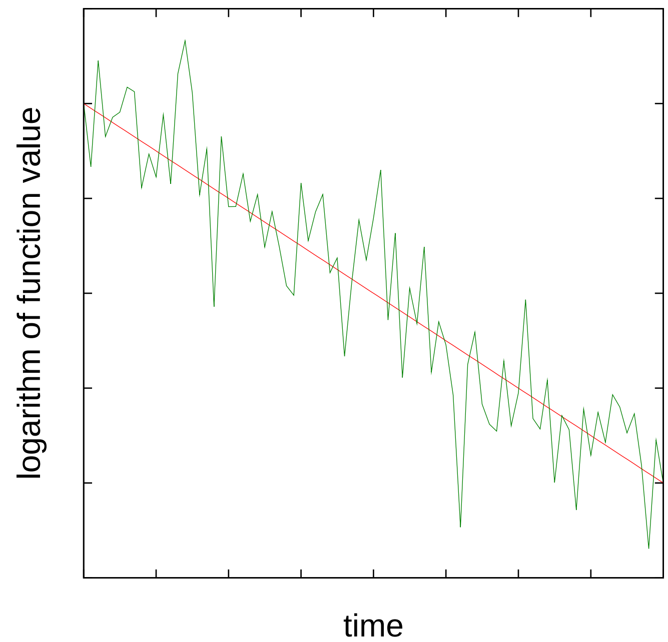
The Sphere Model (4)

- offspring fitness:

$$f(\mathbf{y}) = f(\mathbf{x}) - 2R\sigma z_A + N\sigma^2$$

- with normalised mutation strength $\sigma^* = \sigma N/R$:

$$f(\mathbf{y}) = f(\mathbf{x}) \left[1 - \frac{2}{N}\left(\sigma^* z_A - \frac{\sigma^{*2}}{2}\right)\right]$$



- evolution strategies converge linearly on the sphere model provided that the mutation strength is adapted properly

- the rate of convergence is inversely proportional to N

DALHOUSIE UNIVERSITY
Inspiring Minds

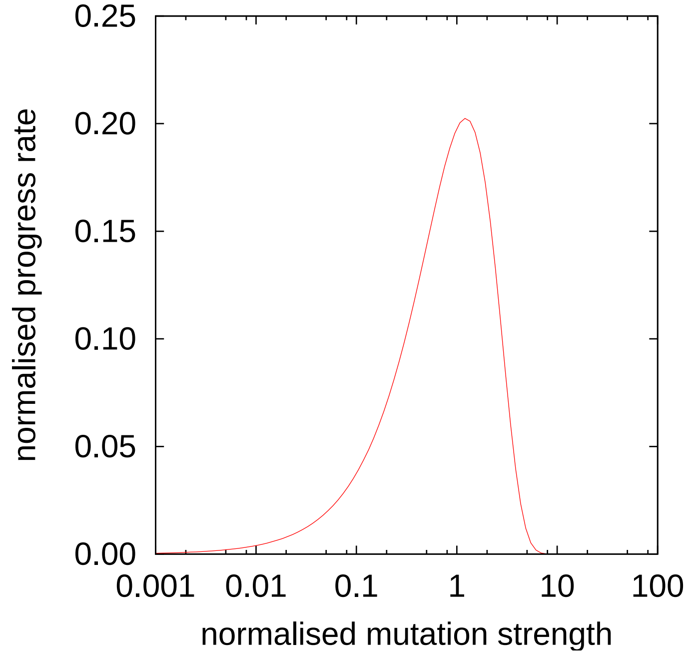
Sphere Model: The $(1+1)$-ES

- progress rate (Rechenberg, 1973):

$$\varphi^*_{1+1} = \frac{\sigma^*}{\sqrt{2\pi}} e^{-\frac{1}{8}\sigma^{*2}} - \frac{\sigma^{*2}}{2}\left[1 - \Phi\left(\frac{\sigma^*}{2}\right)\right]$$

- success probability:

$$P_{succ} = 1 - \Phi\left(\frac{\sigma^*}{2}\right)$$



- maximal progress rate:

$$\varphi^*_{1+1} = 0.202$$

at mutation strength $\sigma^* = 1.224$

DALHOUSIE UNIVERSITY Inspiring Minds

Sphere Model: The $(1+1)$-ES

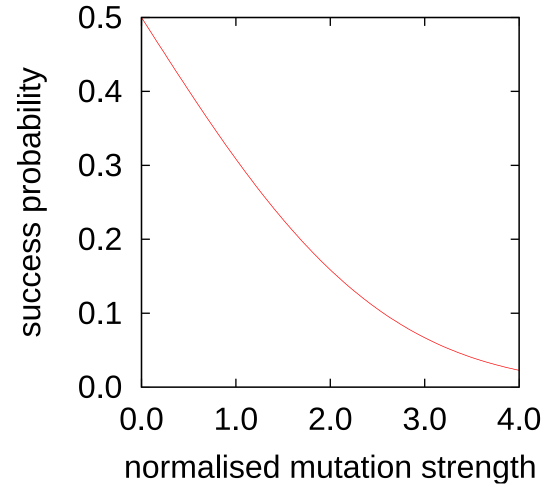
- 1/5th success rule (Rechenberg, 1973):

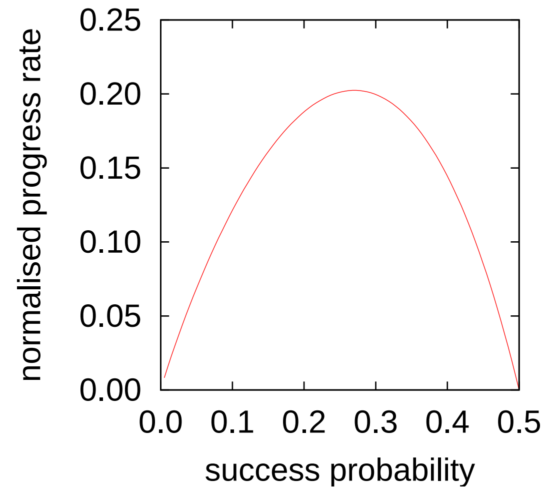
  Decrease the mutation strength if the percentage of successful mutations is below one fifth; increase it if it is above.

- simple implementation:

$$\sigma^{(t+1)} = \sigma^{(t)} \cdot \begin{cases} 2^{1/N} & \text{on success} \\ 2^{-0.25/N} & \text{otherwise} \end{cases}$$

S. Kern et al., 2004. "Learning probability distributions in continuous evolutionary algorithms – a comparative review", *Natural Computing*, 3:77-112.





DALHOUSIE UNIVERSITY
Inspiring Minds

Sphere Model: The $(1,\lambda)$-ES

- progress rate (Rechenberg, 1984):

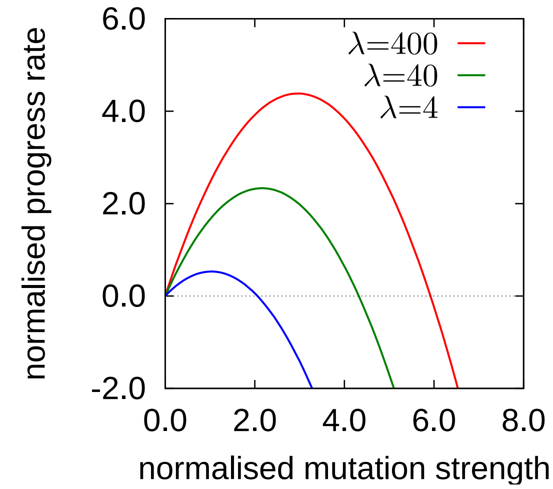
$$\varphi^*_{1,\lambda} = \sigma^* c_{1,\lambda} - \frac{{\sigma^*}^2}{2}$$

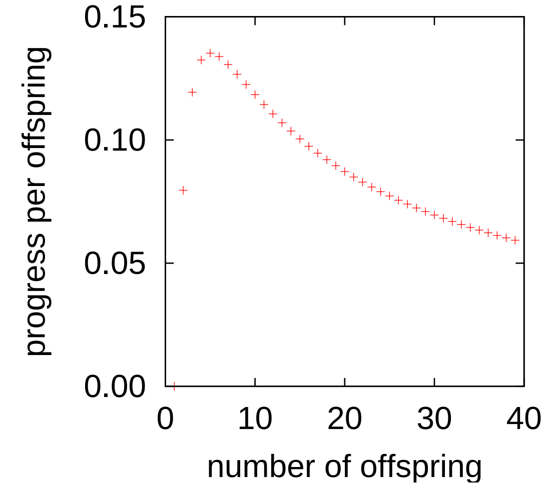
- optimal progress rate:

$$\varphi^*_{1,\lambda} = \frac{c^2_{1,\lambda}}{2} \propto \log \lambda$$

  at mutation strength $\sigma^* = c_{1,\lambda}$

- the $(1,\lambda)$-ES is less efficient than the simple $(1+1)$-ES unless offspring can be evaluated in parallel





DALHOUSIE UNIVERSITY *Inspiring Minds*

Noisy Sphere Model: The $(1, \lambda)$-ES

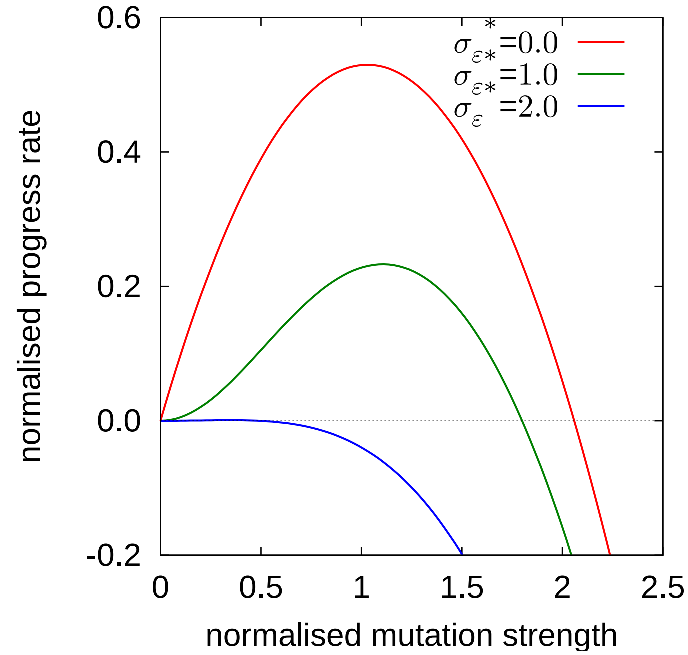
- assume noise of strength $\sigma_\epsilon(\mathbf{x})$ proportional to R^2

- progress rate (Beyer, 1993):

$$\varphi^*_{1,\lambda} = \frac{\sigma^* c_{1,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{{\sigma^*}^2}{2}$$

  where $\vartheta = \sigma^*_\epsilon / \sigma^*$ is the noise-to-signal ratio

- averaging over multiple fitness evaluations reduces the noise strength, but is expensive



Noisy Sphere Model: Rescaled Mutations

- $(1, \lambda)$-ES with rescaled mutations:
  - generate and evaluate offspring

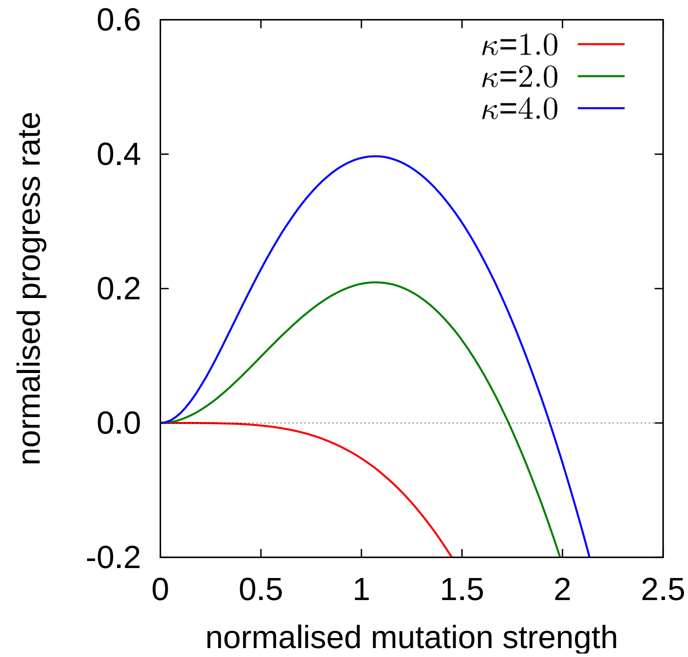
$$\mathbf{y}_i = \mathbf{x}^{(t)} + \kappa\sigma\mathbf{z}_i$$

  - then let

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \sigma\mathbf{z}_{1;\lambda}$$

- progress rate:

$$\varphi^* = \frac{\sigma^* c_{1,\lambda}}{\sqrt{1 + (\sigma^*_\epsilon/(\kappa\sigma^*))^2}} - \frac{{\sigma^*}^2}{2}$$



I. Rechenberg, 1994. *Evolutionsstrategie '94*, Frommann-Holzboog.

H.-G. Beyer, 1998. “Mutate large, but inherit small! On the analysis of rescaled mutations in $(\tilde{1}, \tilde{\lambda})$-ES with noisy fitness data”, *Parallel Problem Solving from Nature*, 5, pp. 109-118, Springer.

DALHOUSIE UNIVERSITY Inspiring Minds

Mutative Self-Adaptation

- every candidate solution carries its own set of strategy parameters
  - mutation:

$$\sigma_i = \sigma^{(t)} \exp(\tau N(0,1))$$
$$\mathbf{y}_i = \mathbf{x}^{(t)} + \sigma_i \mathbf{z}_i$$

  - selection:

$$\sigma^{(t+1)} = \sigma_{1;\lambda}$$
$$\mathbf{x}^{(t+1)} = \mathbf{y}_{1;\lambda}$$

- have a competition of strategic ideas; a “good” set of strategy parameters increases the chance of generating a good set of object parameters and is thus likely to prevail under selection

S. Meyer-Nieberg and H.-G. Beyer, 2007. “Self-adaptation in evolutionary algorithms”, in F. Lobo et al. (eds.), *Parameter Setting in Evolutionary Algorithms*, Springer.

Hierarchically Organised Evolution Strategies

- problems with mutative self-adaptation:
  - selection of strategy parameters is indirect and noisy
  - rewarding short-term success may be shortsighted

- hierarchically organised ES “try out” strategy parameter settings for longer periods of time

  $\Rightarrow$ evolve several populations in isolation from each other; compare their relative success after a number of time steps

M. Herdy, 1992. “Reproductive isolation as strategy parameter in hierarchically organized evolution strategies”, *Parallel Problem Solving from Nature*, 2, pp. 207-217, Elsevier.

D. V. Arnold and A. MacLeod, 2006. “Hierarchically organised evolution strategies on the parabolic ridge”, *Genetic and Evolutionary Computation Conference — GECCO 2006*, pp. 437-444.

Sphere Model: The $(\mu/\mu, \lambda)$-ES

- progress rate (Rechenberg, 1994):

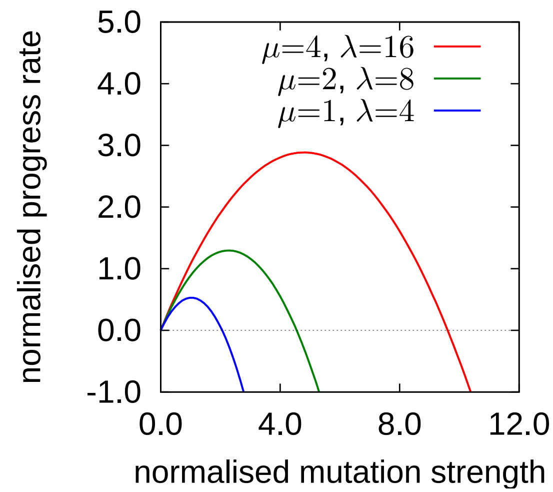
$$\varphi^*_{\mu/\mu,\lambda} = \sigma^* c_{\mu/\mu,\lambda} - \frac{{\sigma^*}^2}{2\mu}$$

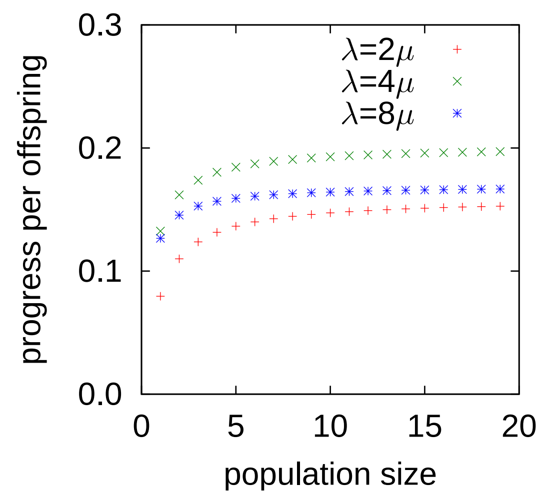
- maximal progress rate:

$$\varphi^*_{\mu/\mu,\lambda} = \frac{\mu c^2_{\mu/\mu,\lambda}}{2} \propto \mu$$

at mutation strength $\sigma^* = \mu c_{\mu/\mu,\lambda}$

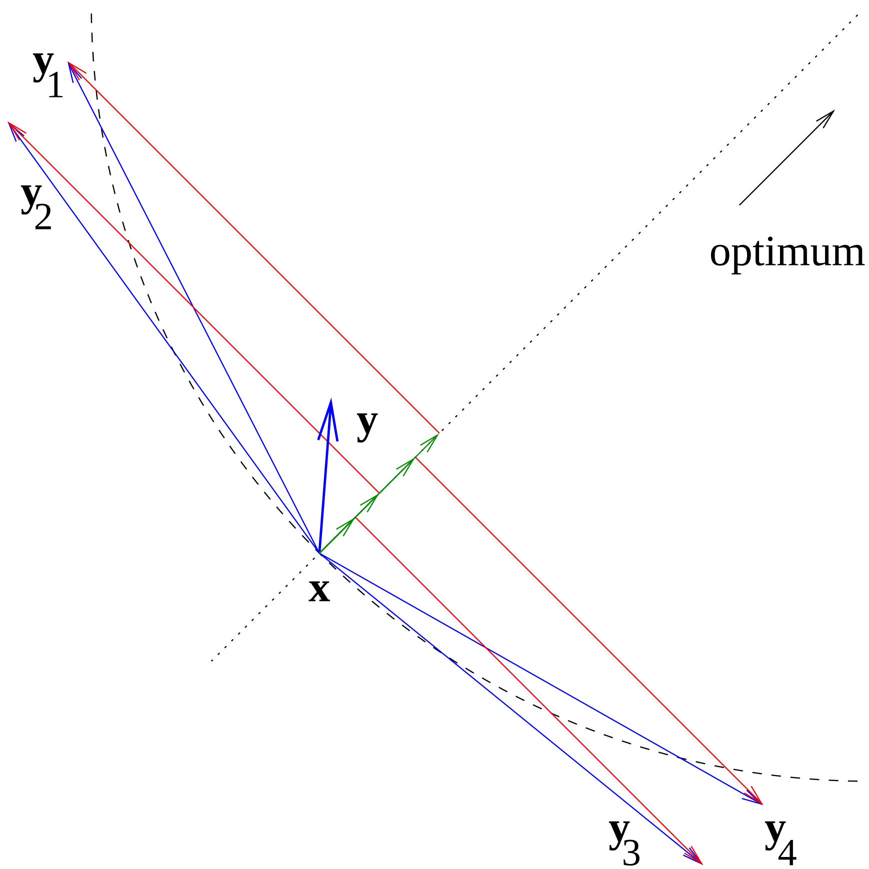
- the maximal (serial!) efficiency is asymptotically equal to that of the $(1+1)$-ES





Sphere Model: The $(\mu/\mu, \lambda)$-ES

- the components toward the optimum of the selected mutation vectors are correlated, the other components are not



- consider vector

$$\mathbf{z}^{(\text{avg})} = \frac{1}{\mu} \sum_{k=1}^{\mu} \mathbf{z}_{k;\lambda}$$

- the length of the “harmful” components of the mutation vectors is reduced

- the purpose of recombination is similarity extraction;

  $\Rightarrow$ “genetic repair principle” (Beyer, 1995)

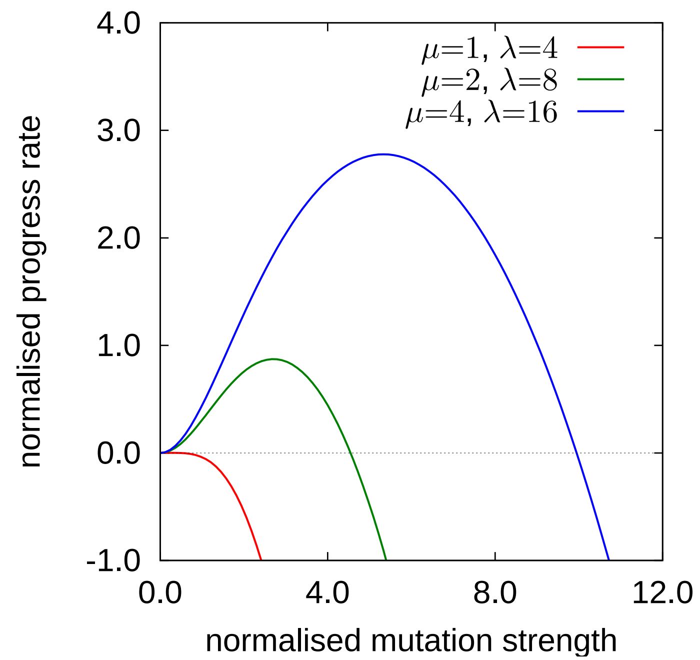
DALHOUSIE UNIVERSITY Inspiring Minds

Noisy Sphere Model: The $(\mu/\mu, \lambda)$-ES

- progress rate in the presence of noise:

$$\varphi^*_{\mu/\mu,\lambda} = \frac{\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{1+\vartheta^2}} - \frac{{\sigma^*}^2}{2\mu}$$

  where $\vartheta = \sigma^*_\epsilon/\sigma^*$ is the noise-to-signal ratio



- the larger mutation strengths (compared to the $(1, \lambda)$-ES) reduce ϑ

- the $(\mu/\mu, \lambda)$-ES implicitly rescales mutation vectors

D. V. Arnold and H.-G. Beyer, 2000. "Local performance of the $(\mu/\mu, \lambda)$-ES in a noisy environment", *Foundations of Genetic Algorithms*, 6, pp. 127-141.

DALHOUSIE UNIVERSITY Inspiring Minds

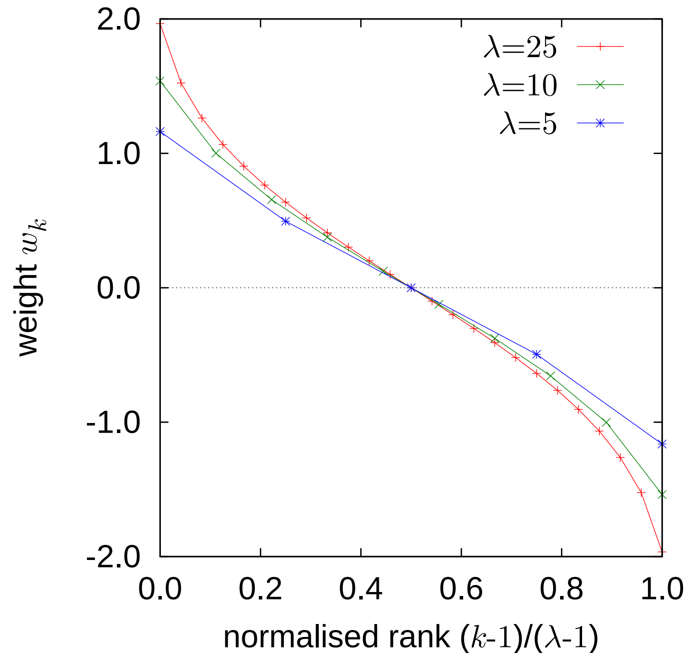
Sphere Model: Optimally Weighted Recombination

- weighted recombination: replace

$$\mathbf{x} = \frac{1}{\mu}\sum_{k=1}^{\mu} \mathbf{x}_{k;\lambda}$$

with

$$\mathbf{x} = \sum_{k=1}^{\lambda} w_k \mathbf{x}_{k;\lambda}$$

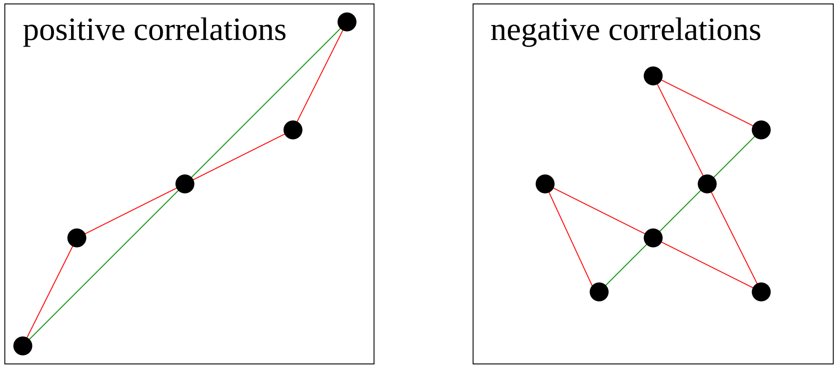


- for maximal progress, choose $w_k \propto \mathrm{E}[z_{k;\lambda}]$

- the proportionality constant determines the amount of implicit re-scaling; the speed-up is 2.5-fold

D. V. Arnold, 2006. “Weighted multirecombination evolution strategies”, *Theoretical Computer Science*, 361(1):18-37.

Cumulative Step Length Adaptation (1)



- postulate: consecutive steps of the strategy should be uncorrelated

- if consecutive steps are positively correlated, then the step length should be increased

- if consecutive steps are negatively correlated, then the step length should be decreased

A. Ostermeier, A. Gawelczyk, and N. Hansen, 1994. “Step-size adaptation based on non-local use of selection information”, *Parallel Problem Solving from Nature*, 3, pp. 189-198, Springer.

DALHOUSIE UNIVERSITY Inspiring Minds

Cumulative Step Length Adaptation (2)

- in order to detect correlations, information from a number of steps needs to be accumulated

  $\Rightarrow$ (for the $(\mu/\mu, \lambda)$-ES) define the search path

$$\mathbf{s}^{(t+1)} = (1-c)\mathbf{s}^{(t)} + \sqrt{\mu c(2-c)}\,\mathbf{z}^{(\text{avg})}$$

- under random selection, the expected squared length of the search path is N

- the step length is updated according to

$$\sigma^{(t+1)} = \sigma^{(t)} \exp\left(\frac{\|\mathbf{s}^{(t+1)}\|^2 - N}{2DN}\right)$$

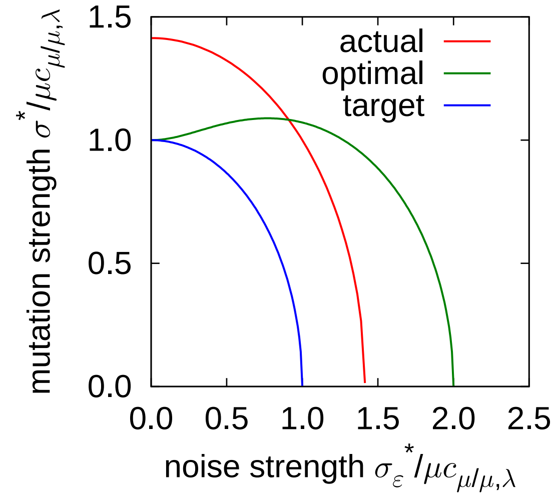
Cumulative Step Length Adaptation (3)

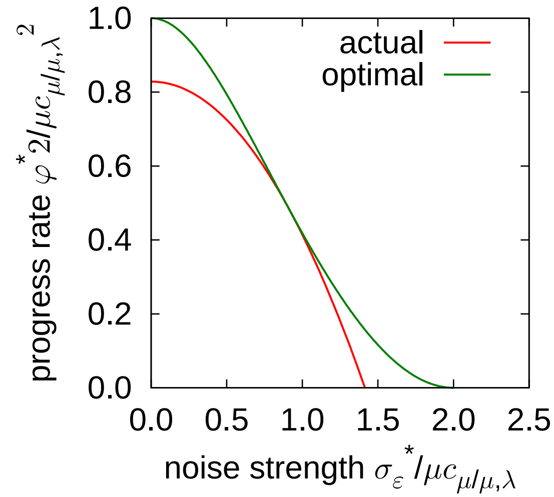
- on the noisy sphere, cumulative step length adaptation generates

$$\sigma^* = \mu c_{\mu/\mu,\lambda}\sqrt{2 - \left(\frac{\sigma^*_\epsilon}{\mu c_{\mu/\mu,\lambda}}\right)^2}$$

and achieves progress rate

$$\varphi^* = \frac{\sqrt{2}-1}{2}\mu c^2_{\mu/\mu,\lambda}\left(2 - \left(\frac{\sigma^*_\epsilon}{\mu c_{\mu/\mu,\lambda}}\right)^2\right)$$





D. V. Arnold and H.-G. Beyer, 2004. "Performance analysis of evolutionary optimization with cumulative step length adaptation", *IEEE Transactions on Automatic Control*, 49(4):617-622.

Comparison of Strategies (1)

Direct Pattern Search: (Hooke and Jeeves, 1961) precursor of many direct search strategies

Multi-Directional Search: (Torczon, 1989) successor of Nelder and Mead's simplex method, the most popular strategy for noisy optimisation

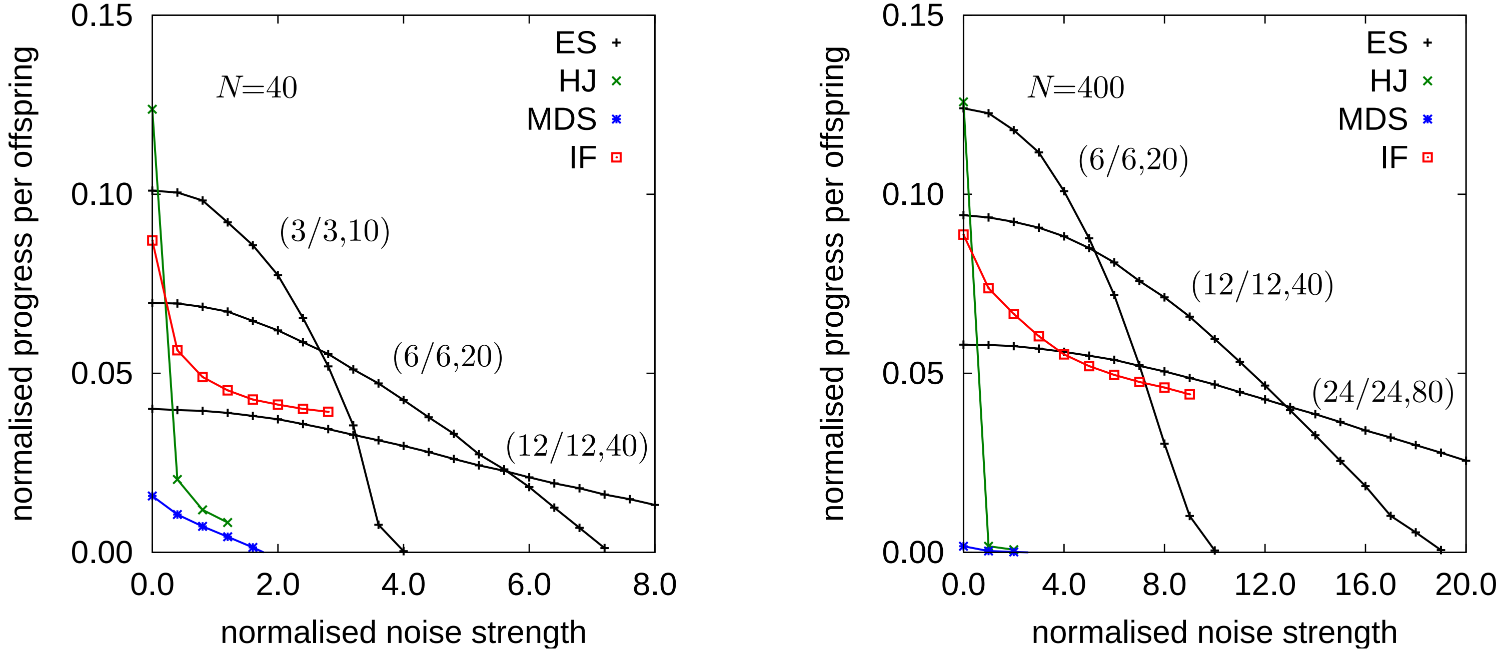
Implicit Filtering: (Gilmore and Kelley, 1995) gradient strategy using finite differencing and Armijo line searches; designed for noisy optimisation

Evolution Strategy: $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation and various population sizes

D. V. Arnold and H.-G. Beyer, 2003. "A comparison of evolution strategies with other direct search methods in the presence of noise", *Computational Optimization and Applications*, 24(1):135-159.

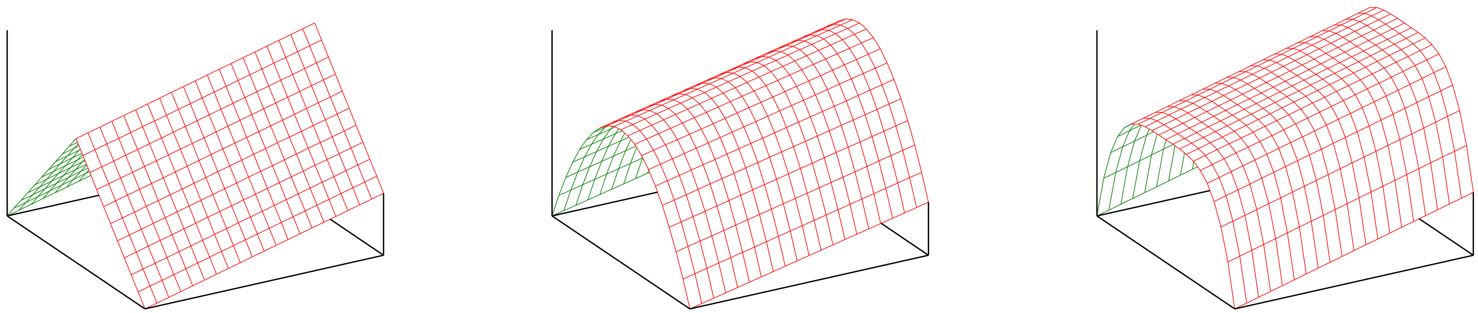
DALHOUSIE UNIVERSITY *Inspiring Minds*

Comparison of Strategies (2)



- incomplete graphs result from failure to achieve linear convergence
- larger populations buy robustness at the price of efficiency
- strengths of ES: genetic repair and relatively robust step length adaptation

Ridge Model



- ridge model:

$$f(\mathbf{x}) = x_1 - d\left(\sum_{i=2}^{N} x_i^2\right)^{\alpha/2}$$

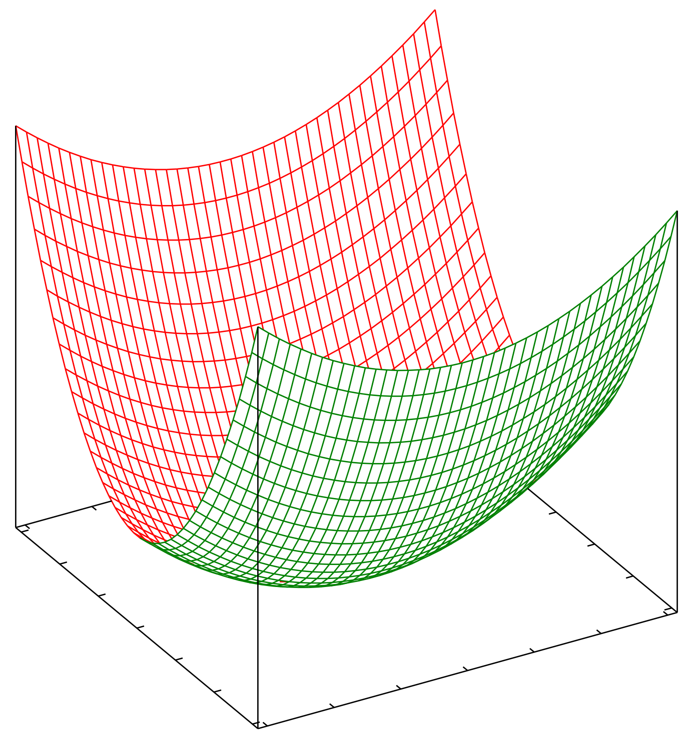
A. I. Oyman, H.-G. Beyer, and H.-P. Schwefel, 1998. “Where elitists start limping: Evolution strategies at ridge functions”, *Parallel Problem Solving from Nature*, 5, pp. 109-118, Springer.

D. V. Arnold and A. MacLeod, 2006. “Step length adaptation on ridge functions”, Technical Report CS-2006-08, Faculty of Computer Science, Dalhousie University.

Convex Quadratic Functions (1)

- convex quadratic functions:

$$f(\mathbf{x}) = \sum_{i=1}^{N} (a_i x_i)^2$$



- Hessian matrix:

$$\mathbf{H} = \begin{pmatrix} 2a_1^2 & 0 & \cdots & 0 \\ 0 & 2a_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2a_N^2 \end{pmatrix}$$

- condition number: $a_{\max}/a_{\min}$

- for strategies that are not rotationally invariant, the coordinate system should be rotated

DALHOUSIE UNIVERSITY *Inspiring Minds*

Convex Quadratic Functions (2)

- examples of convex quadratic functions:

$$f_{\text{cigar}}(\mathbf{x}) = x_1^2 + \sum_{i=2}^{N} (1000 x_i)^2$$

$$f_{\text{discus}}(\mathbf{x}) = (1000 x_1)^2 + \sum_{i=2}^{N} x_i^2$$

$$f_{\text{ellipsoid}}(\mathbf{x}) = \sum_{i=1}^{N} \left(1000^{(i-1)/(N-1)} x_i \right)^2$$

$$f_{\text{twoaxes}}(\mathbf{x}) = \sum_{i=1}^{\lfloor N/2 \rfloor} x_i^2 + \sum_{\lfloor N/2 \rfloor + 1}^{N} (1000 x_i)^2$$

Convex Quadratic Functions (3)

- performance of the $(1+1)$-ES using isotropically distributed mutations (Jägersküpper, 2006):
  - if the condition number is $\mathcal{O}(1)$, the number of steps needed to reduce the approximation error to a 2^{-b}-fraction is $\Theta(bN)$
  - for f_{twoaxes} with condition number ξ polynomially bounded in N such that $1/\xi \to 0$ as $N \to \infty$, the number of steps is $\Theta(b\xi N)$
- similar results hold for the $(\mu/\mu, \lambda)$-ES with cumulative step length adaptation (Arnold, 2007)

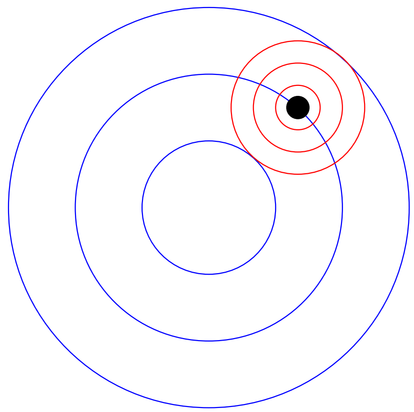
J. Jägersküpper, 2006. "How the $(1+1)$-ES using isotropic mutations minimizes positive definite quadratic forms", *Theoretical Computer Science*, 361(1):38-56.

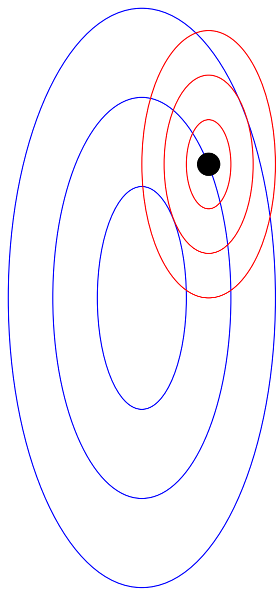
D. V. Arnold, 2007. "On the use of evolution strategies for optimising certain positive definite quadratic forms", *Genetic and Evolutionary Computation Conference — GECCO 2007*.

Nonisotropically Distributed Mutations

- nonisotropic mutation distributions can be vastly more efficient than isotropic ones

- ideally, the mutation covariance matrix should be proportional to the inverse of the local Hessian of the objective





H.-P. Schwefel, 1981. *Numerical Optimization of Computer Models*, Wiley.

G. Rudolph, 1992. "On correlated mutations in evolution strategies", *Parallel Problem Solving from Nature*, 2, pp. 105-114, Elsevier.

Covariance Matrix Adaptation (1)

- the CMA-ES adapts the mutation covariance matrix based on information gathered in past steps (Hansen and Ostermeier, 2001)
  - variances in directions that have previously been successful are increased
  - other variances decay over time
- the strategy is rotationally invariant

N. Hansen and A. Ostermeier, 2001. “Completely derandomized self-adaptation in evolution strategies”, *Evolutionary Computation*, 9(2):159-195.

N. Hansen, 2005. “The CMA evolution strategy: A tutorial”, `http://www.bionik.tu-berlin.de/user/niko/cmatutorial.pdf`.

Covariance Matrix Adaptation Evolution Strategy (1)

- state variables of the $(\mu/\mu, \lambda)$-CMA-ES: $\mathbf{x}$, σ, $\mathbf{C}$, $\mathbf{s}_\sigma$, $\mathbf{s}_\mathbf{C}$
- in every iteration:

1. Compute an eigen decomposition $\mathbf{C} = \mathbf{BD}(\mathbf{BD})^\mathrm{T}$.
2. Generate λ offspring $\mathbf{y}_i = \mathbf{x} + \sigma\mathbf{BD}\mathbf{z}_i$.
3. Compute the mean of the μ best offspring

$$\mathbf{z}^{(\text{avg})} = \frac{1}{\mu}\sum_{k=1}^{\mu}\mathbf{z}_{k;\lambda}.$$

4. Update the search point

$$\mathbf{x} \leftarrow \mathbf{x} + \sigma\mathbf{BD}\mathbf{z}^{(\text{avg})}.$$

5. Update $\mathbf{s}_\sigma$ and $\mathbf{s}_\mathbf{C}$.
6. Update $\mathbf{C}$ and σ.

Covariance Matrix Adaptation Evolution Strategy (2)

- update of the search paths:

$$\mathbf{s}_\sigma \leftarrow (1 - c_\sigma)\mathbf{s}_\sigma + \sqrt{\mu c_\sigma (2 - c_\sigma)}\mathbf{B}\mathbf{z}^{(\text{avg})}$$

$$\mathbf{s}_\mathbf{C} \leftarrow (1 - c_\mathbf{C})\mathbf{s}_\mathbf{C} + \sqrt{\mu c_\mathbf{C} (2 - c_\mathbf{C})}\mathbf{B}\mathbf{D}\mathbf{z}^{(\text{avg})}$$

- update of the step length:

$$\sigma \leftarrow \sigma \exp\left(\frac{\|\mathbf{s}_\sigma\|^2 - N}{2DN}\right)$$

- update of the covariance matrix:

$$\mathbf{C} \leftarrow (1 - c_{\text{cov}})\mathbf{C} + c_{\text{cov}}\mathbf{s}_\mathbf{C}\mathbf{s}_\mathbf{C}^\mathrm{T}$$

- c_σ, $c_\mathbf{C}$, and c_{cov} determine how quickly old information fades

DALHOUSIE UNIVERSITY
Inspiring Minds

Covariance Matrix Adaptation Evolution Strategy (3)

- better use can be made of large populations by using covariance matrix update

$$\mathbf{C} \leftarrow (1 - c_{\text{cov}})\mathbf{C} + c_{\text{cov}} \left(\alpha_{\text{cov}}\mathbf{s}_{\mathbf{C}}\mathbf{s}_{\mathbf{C}}^{\mathrm{T}} + (1 - \alpha_{\text{cov}})\mathbf{Z}\right)$$

where

$$\mathbf{Z} = \mathbf{BD}\left(\frac{1}{\mu}\sum_{k=1}^{\mu}\mathbf{z}_{k;\lambda}\mathbf{z}_{k;\lambda}^{\mathrm{T}}\right)(\mathbf{BD})^{\mathrm{T}}$$

- c_{cov} can be chosen larger (roughly by a factor of μ) than for the previous update rule

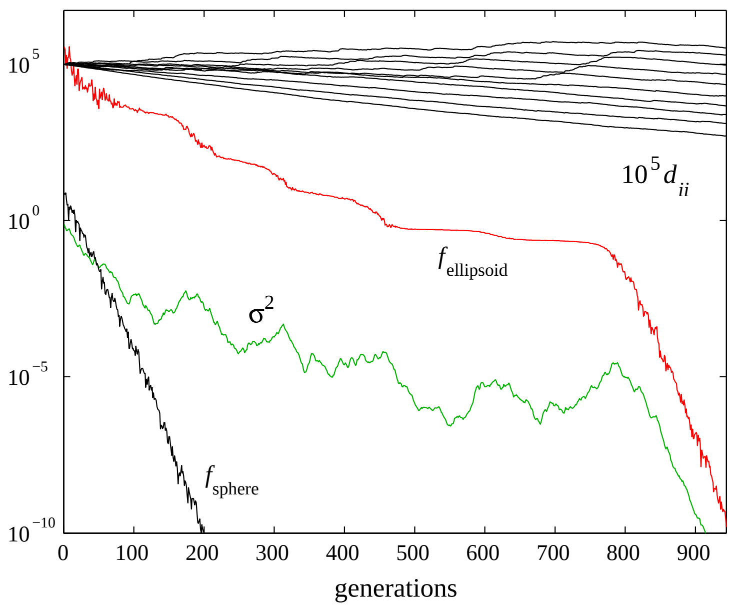
- α_{cov} weights the path-based and population-based contributions

N. Hansen, S. D. Müller, and P. Koumoutsakos, 2003. “Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)”, *Evolutionary Computation*, 11(1):1-18.

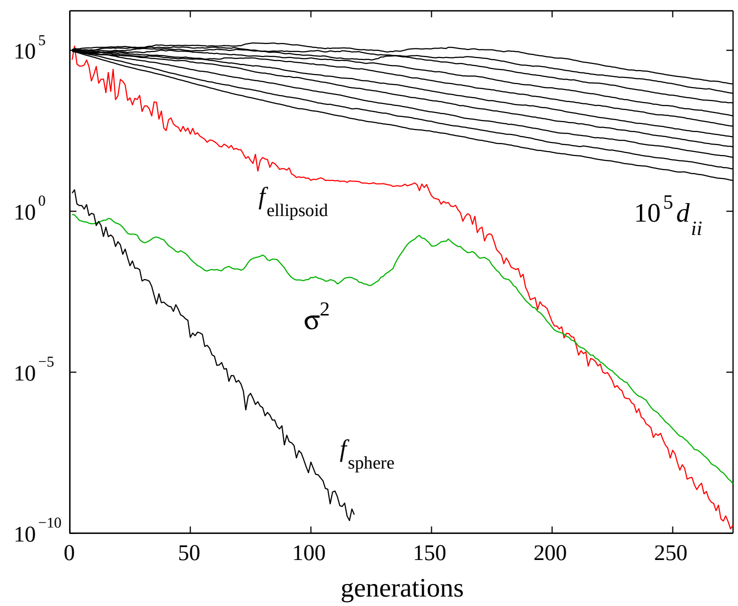
Covariance Matrix Adaptation Evolution Strategy (4)

- performance on $f_{\text{ellipsoid}}$ with $N = 10$

small population ($\mu = 2$, $\lambda = 8$):



large population ($\mu = 10$, $\lambda = 40$):



DALHOUSIE UNIVERSITY Inspiring Minds

Active Covariance Matrix Adaptation (1)

- idea: use information not only from successful, but also from unsuccessful offspring

  $\Rightarrow$ actively decrease variances in directions that repeatedly yield bad offspring

- new update rule:

$$\mathbf{C} \leftarrow (1 - c_{\mathsf{cov}})\mathbf{C} + c_{\mathsf{cov}}\mathbf{s}_{\mathbf{C}}\mathbf{s}_{\mathbf{C}}^{\mathrm{T}} + \beta\mathbf{Z}$$

  where

$$\mathbf{Z} = \mathbf{BD}\left(\frac{1}{\mu}\sum_{k=1}^{\mu}\mathbf{z}_{k;\lambda}\mathbf{z}_{k;\lambda}^{\mathrm{T}} - \frac{1}{\mu}\sum_{k=\lambda-\mu+1}^{\lambda}\mathbf{z}_{k;\lambda}\mathbf{z}_{k;\lambda}^{\mathrm{T}}\right)(\mathbf{BD})^{\mathrm{T}}$$

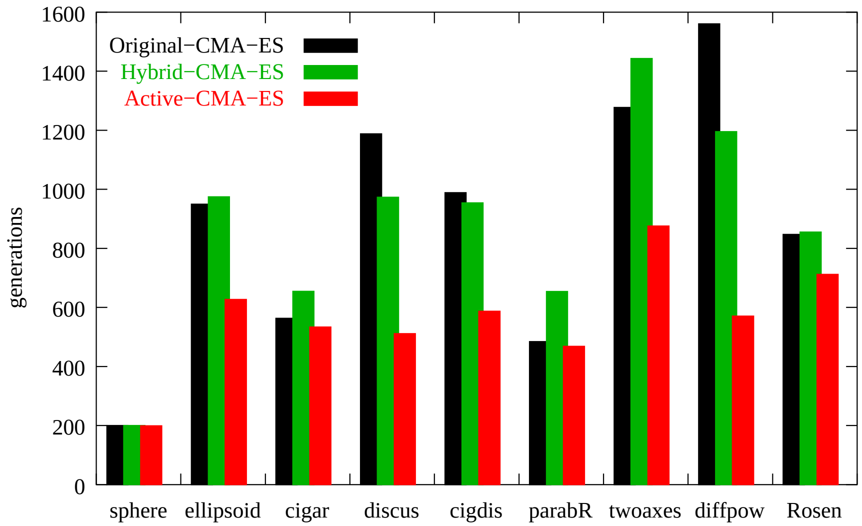
G. A. Jastrebski and D. V. Arnold, 2006. “Improving evolution strategies through active covariance matrix adaptation”, *Proc. IEEE Congress on Evolutionary Computation*, pp. 9719-9726.

DALHOUSIE UNIVERSITY *Inspiring Minds*

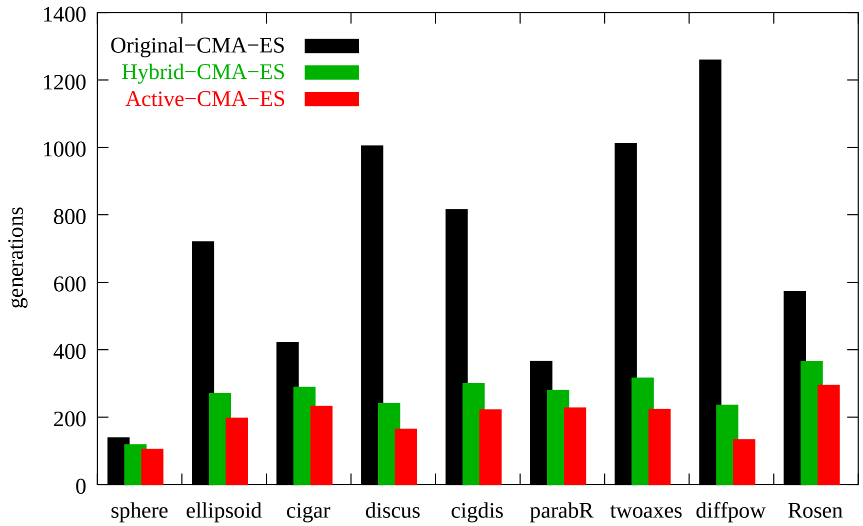
Active Covariance Matrix Adaptation (2)

- performance across a set of test functions ($N = 10$)

small population ($\mu = 2$, $\lambda = 8$):



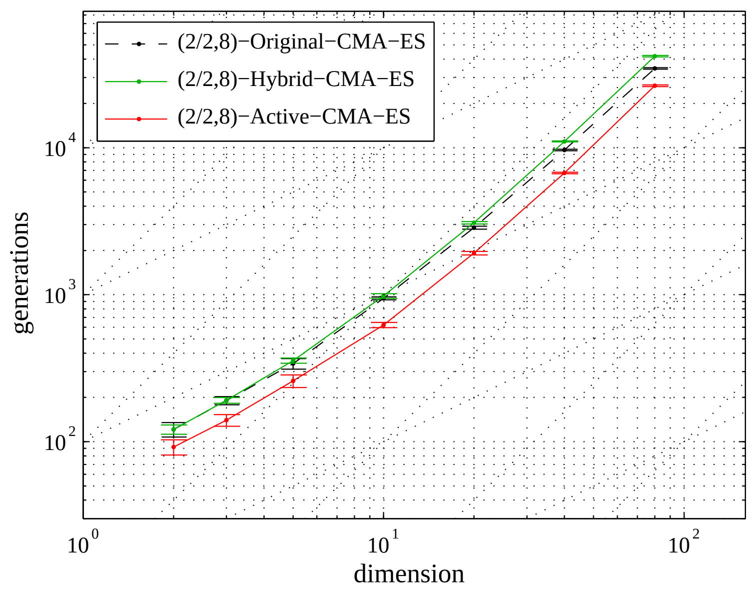
large population ($\mu = 10$, $\lambda = 40$):



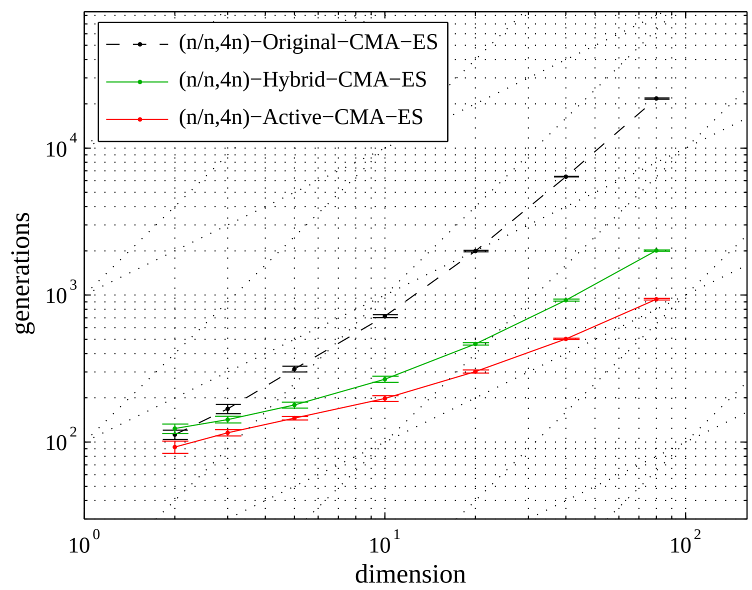
Active Covariance Matrix Adaptation (3)

- dependence on N
- test function: $f_{\text{ellipsoid}}$

small population ($\mu = 2$, $\lambda = 8$):



large population ($\mu = N$, $\lambda = 4N$):



DALHOUSIE UNIVERSITY Inspiring Minds

$(1+1)$-CMA-ES

- the eigen decomposition is computationally expensive for large N
- for the $(1+1)$-CMA-ES,
  - the mutation strength can be controlled using the $1/5$th rule
  - matrix $\mathbf{A} = \mathbf{BD}$ can be updated directly, with no need to decompose the covariance matrix
- for multimodal (and presumably for noisy) functions, the longer steps of the $(\mu/\mu, \lambda)$-CMA-ES are advantageous

C. Igel, T. Suttorp, and N. Hansen, 2006. "A computational efficient covariance matrix update and a (1+1)-CMA for evolution strategies", *Genetic and Evolutionary Computation Conference — GECCO 2006*, pp. 453-460.

Differential Evolution

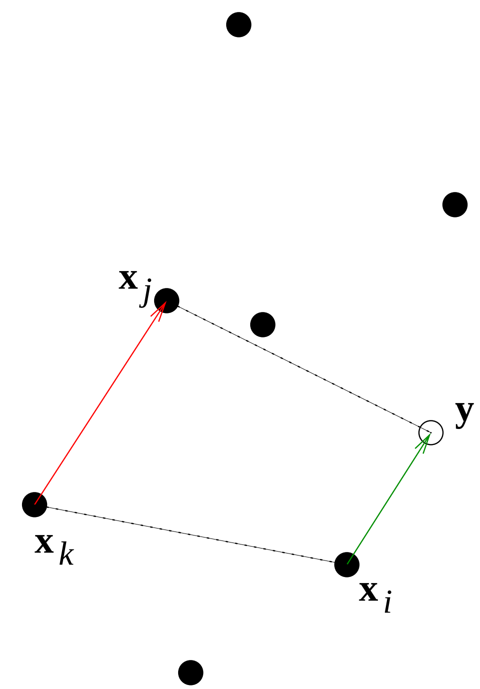
- mutation (DE/rand/1): randomly pick $i \neq j \neq k$ and let

$$\mathbf{y} = \mathbf{x}_i + k(\mathbf{x}_j - \mathbf{x}_k)$$

  $\Rightarrow$ the step length is determined by the diversity of the population

- the rate at which diversity decreases is influenced by the population size, the replacement mechanism, and the factor k

- adaptive variants exist

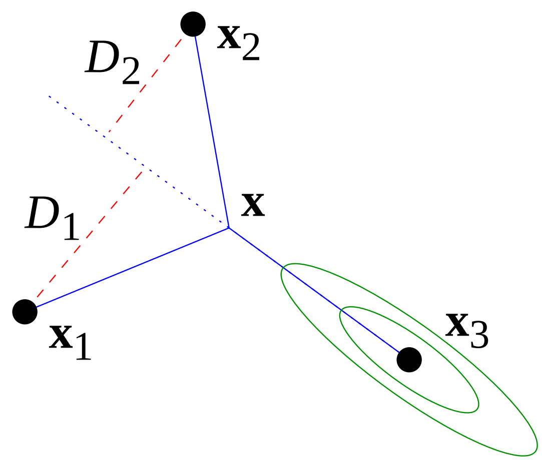


K. V. Price, R. Storn, and J. Lampinen, 2005. *Differential Evolution — A Practical Approach to Global Optimization*, Springer.

PCX — Parent Centric Recombination

- assume wlog that $\mathbf{x}_1$ has been picked as parent; compute

$$\mathbf{y} = \mathbf{x}_1 + \sigma_1 z_1 \mathbf{d}_1 + \sigma_2 D \sum_{i=2}^{\mu} z_i \mathbf{e}_i$$



where $\mathbf{x}$ is the population centroid, $\mathbf{d}_i = \mathbf{x}_i - \mathbf{x}$, $\mathbf{e}_i$ is the normalised vector consisting of those components of $\mathbf{d}_i$ that are perpendicular to $\mathbf{x}_1$, and D is the average distance of the $\mathbf{x}_i$ from the line through $\mathbf{x}$ and $\mathbf{x}_1$

K. Deb, A. Anand, and D. Joshi, 2002. “A computationally efficient evolutionary algorithm for real-parameter optimization”, *Evolutionary Computation*, 10(4):371-395.

Summary and Further Topics

topics covered:

- evolution strategies:
  - benefits of populations and recombination
  - coping with noise
- step length adaptation: success rate based; mutative; cumulative
- covariance matrix adaptation: passive and active; alternative approaches

further topics:

- multimodal problems: restart strategies; niching methods; spatially organised populations
- constraint handling techniques