Introduction to Evolution Strategies

Thomas Bäck
Leiden Institute of Advanced Computer Science (LIACS)
Leiden University, Niels Bohrweg 1, 2333 CA Leiden
The Netherlands
t.h.w.baeck@liacs.leidenuniv.nl

http://www.sigevo.org/gecco-2014/

GECCO'14, July 12–16, 2014, Vancouver, BC, Canada.
ACM 978-1-4503-2662-9/14/07.

1

Instructor

Thomas Bäck is full Professor of Computer Science at the Leiden Institute of Advanced Computer Science (LIACS) at Leiden University.

Thomas Bäck has authored or co-authored more than 250 publications on evolutionary computation, as well as a book on evolutionary algorithms, entitled *Evolutionary Algorithms: Theory and Practice.*

He is editorial board member and associate editor of a number of journals on evolutionary and natural computation, and has served as program chair for the major conferences in evolutionary computation. He received the best dissertation award from the Gesellschaft für Informatik (GI) in 1995 and is an elected fellow of the International Society for Genetic and Evolutionary Computation for his contributions to the field.

From 2000 - 2009, Thomas was CEO of NuTech Solutions GmbH and CTO of NuTech Solutions, Inc., until November 2009. Thomas has ample experience in working with Fortune 1000 customers such as Air Liquide, BMW Group, Beiersdorf, Daimler, Corning, Inc., Ford of Europe, Honda, Johnson & Johnson, P&G, Symrise, Siemens, Unilever, and others.

He is co-editor of the Handbook of Evolutionary Computation and the Handbook of Natural Computing (Springer, 2012). In 2013, he published the book *Contemporary Evolution Strategies* (with C. Foussette and P. Krause).

Home page: www.natcomp.liacs.nl

2

Abstract

This tutorial gives a basic introduction to **evolution strategies**, a class of evolutionary algorithms. Key features such as mutation, recombination and selection operators are explained, and specifically the concept of **self-adaptation** of strategy parameters is introduced.

All algorithmic concepts are explained to a level of detail such that an implementation of basic evolution strategies is possible.

In addition, the tutorial also presents a brief taxonomy of contemporary evolution strategy variants, including e.g. the CMA-ES and variations thereof, and compares their performance for a small number of function evalutions – which represents many of today's practical application cases.

Some guidelines for utilization as well as some application examples are also given.

3

Agenda

- Introduction: Optimization and EAs
- Evolution Strategies
- Examples

4

A True Story …

During my PhD

- Ran artificial test problems
- n=30 maximum dimensionality
- Evaluation took „no“ time
- No constraints
- Thought these were difficult

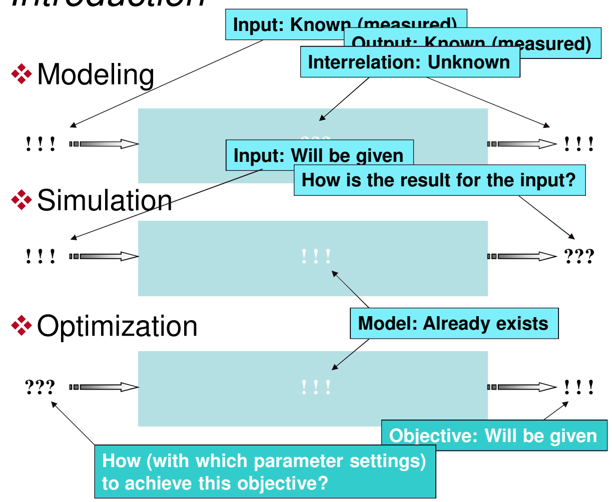
Now

- Real-world problems
- n=150, n=10,000
- Evaluation can take 20 hours
- 50 nonlinear constraints
- Tip of the iceberg

5

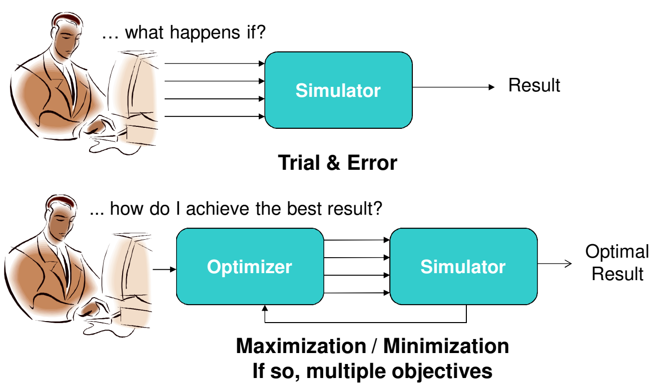
Introduction

- Modeling
- Simulation
- Optimization



6

Simulation vs. Optimization



7

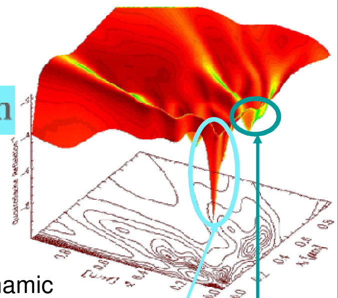
Introduction:

Optimization Evolutionary Algorithms

8

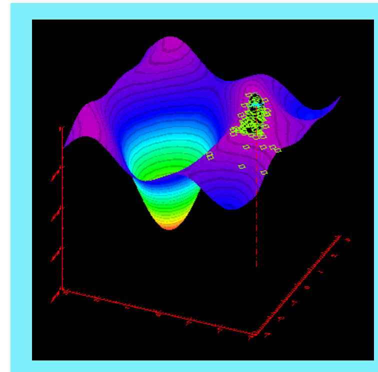
Optimization

$$f : M \to \Re, f(\vec{x}) \to \min$$



- ❖ f : objective function
  - High-dimensional
  - Non-linear, multimodal
  - Discontinuous, noisy, dynamic
- ❖ $M \subseteq M_1 \times M_2 \times ... \times M_n$ heterogeneous
- ❖ Restrictions possible over $M, f(\vec{x})$
- ❖ Good local, robust optimum desired
- ❖ Realistic landscapes are like that!

Dynamic Optimization



- ❖ Dynamic Function
- ❖ 30-dimensional
- ❖ 3D-Projection

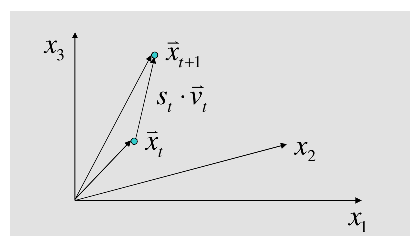
Classification of Optimization Algorithms

- ❖ Direct optimization algorithm: Evolutionary Algorithms — $f(\vec{x})$
- ❖ First order optimization algorithm: e.g., gradient method — $f(\vec{x}), \nabla f(\vec{x})$
- ❖ Second order optimization algorithm: e.g., Newton method — $f(\vec{x}), \nabla f(\vec{x}), \nabla^2 f(\vec{x})$

Iterative Optimization Methods

- ❖ General description: $\vec{x}_{t+1} = \vec{x}_t + s_t \cdot \vec{v}_t$

New Point
Actual Point
Directional vector
Step size (scalar)



- At every Iteration:
  - Choose direction
  - Determine step size
- Direction:
  - Gradient
  - Random
- Step size:
  - 1-dim. optimization
  - Random
  - Self-adaptive

The Fundamental Challenge

- Global convergence with probability one:

$$\lim_{t\to\infty} \Pr(\vec{x}^* \in P(t)) = 1$$

- General, but for practical purposes useless
- Convergence velocity:

$$\varphi = E(f_{\max}(P(t+1)) - f_{\max}(P(t)))$$

- Local analysis only, specific (convex) functions

13

Theoretical Statements

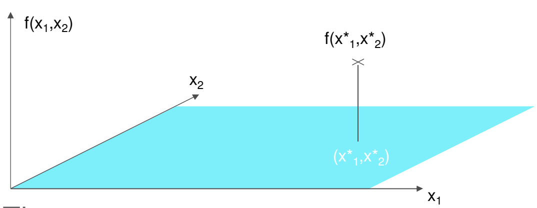
- Global convergence (with probability 1):

$$\lim_{t\to\infty} \Pr(\vec{x}^* \in P(t)) = 1$$

- General statement (holds for all functions)
- Useless for practical situations:
  - Time plays a major role in practice
  - Not all objective functions are relevant in practice

14

An Infinite Number of Pathological Cases !



- NFL-Theorem:
  - All optimization algorithms perform equally well iff performance is averaged over all possible optimization problems.
- Fortunately: We are not Interested in „all possible problems“

15

Theoretical Statements

- Convergence velocity:

$$\varphi = E(f_{\max}(P(t+1)) - f_{\max}(P(t)))$$

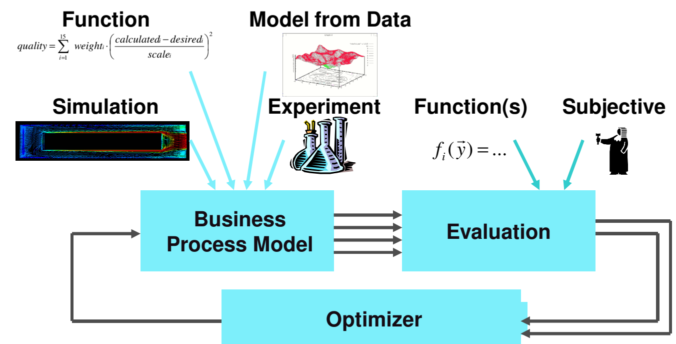
- Very specific statements
  - Convex objective functions
  - Describes convergence in local optima
  - Very extensive analysis for Evolution Strategies

16

Evolution Strategies

17

Model-Optimization-Action



18

Evolutionary Algorithms Taxonomy

Evolutionary Algorithms

Evolution Strategies

- Mixed-integer capabilities
- Emphasis on mutation
- Self-adaptation
- Small population sizes
- Deterministic selection
- Developed in Germany
- Theory focused on convergence velocity

Genetic Algorithms

- Discrete representations
- Emphasis on crossover
- Constant parameters
- Larger population sizes
- Probabilistic selection
- Developed in USA
- Theory focused on schema processing

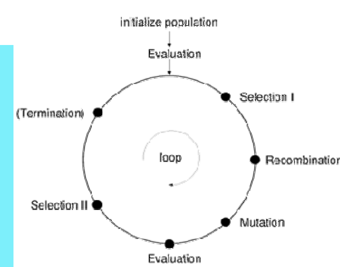
Other

- Evolutionary Prog
- Differential Evol.
- GP
- PSO
- EDA
- Real-coded Gas
- …

19

Generalized Evolutionary Algorithm

t := 0;
***initialize*(P(t));**
***evaluate*(P(t));**
while not *terminate* do
  P'(t) := *mating_selection*(P(t));
  P''(t) := *variation*(P'(t));
  ***evaluate*(P''(t));**
  P(t+1) := *environmental_selection*(P''(t) ∪ Q);
  t := t+1;
od



20

Optimization Creating Innovation

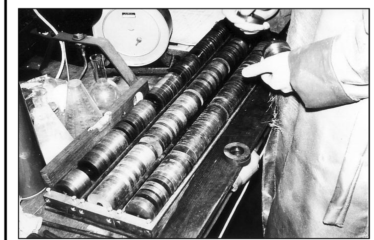
- Illustrative Example: Optimize Efficiency
  - Initial:

  - Evolution:

- 32% Improvement in Efficiency !

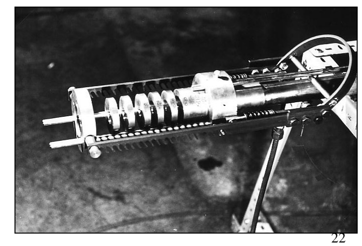
21

Nozzle Experiment (I)



collection of conical nozzle parts

device for clamping nozzle parts



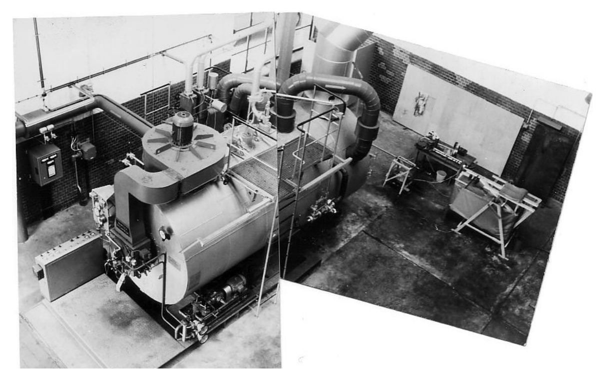
22

Nozzle Experiment (II)

Hans-Paul Schwefel
while changing nozzle parts

23

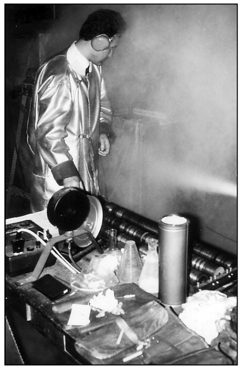
Nozzle Experiment (III)

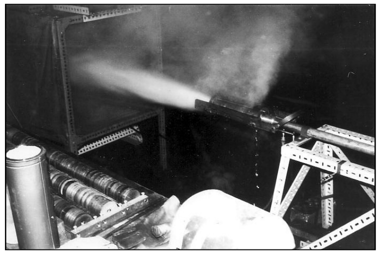


steam plant / experimental setup

24

Nozzle Experiment (IV)





the nozzle in operation …

… while measuring degree of efficiency

25

The Simple (1+1)-ES

$t := 0$;
initialize $P(0) := \{\vec{x}(0)\} \in I,\ I = I\!R^n,\ \vec{x} = (x_1, \dots, x_n)$;
evaluate $P(0): \{f(\vec{x}(0))\}$
while not *terminate*$(P(t))$ **do**
  mutate: $\vec{x}'(t) := m(\vec{x}(t))$
    where $x'_i := x_i + \sigma(t) \cdot N_i(0,1)\ \forall i \in \{1, \dots, n\}$
  evaluate: $P'(t) := \{\vec{x}'(t)\} : \{f(\vec{x}'(t))\}$
  select: $P(t+1) := s_{(1+1)}(P(t) \cup P'(t))$;
  $t := t + 1$;
  if $(t \bmod n = 0)$ **then**

$$\sigma(t) := \begin{cases} \sigma(t-n)/c & \text{, if } p_s > 1/5 \\ \sigma(t-n) \cdot c & \text{, if } p_s < 1/5 \\ \sigma(t-n) & \text{, if } p_s = 1/5 \end{cases}$$

    where p_s *is the relative frequency of successful mutations, measured over intervals of, say,* $10 \cdot n$ *trials;*
    and $0.817 \le c \le 1$;
  else
    $\sigma(t) := \sigma(t-1)$;
  fi
26 **od**

26

Evolution Strategy – Basics

- Mostly real-valued search space IR^n
  - also mixed-integer, discrete spaces
- Emphasis on mutation
  - n-dimensional normal distribution
  - expectation zero
- Different recombination operators
- Deterministic selection
  - (μ, λ)-selection: Deterioration possible
  - (μ+λ)-selection: Only accepts improvements
- λ >> μ, i.e.: Creation of offspring surplus
- Self-adaptation of strategy parameters.

27

Representation of search points

- Simple ES with 1/5 success rule:
  - Exogenous adaptation of step size σ
  - Mutation: N(0, σ)

$$\vec{a} = (x_1, \dots, x_n)$$

- Self-adaptive ES with single step size:
  - One σ controls mutation for all x_i
  - Mutation: N(0, σ)

$$\vec{a} = ((x_1, \dots, x_n), \sigma)$$

28

Representation of search points

- Self-adaptive ES with individual step sizes:
  - One individual σ_i per x_i
  - Mutation: $N_i(0, \sigma_i)$

$$\vec{a} = ((x_1,\ldots,x_n),(\sigma_1,\ldots,\sigma_n))$$

- Self-adaptive ES with correlated mutation:
  - Individual step sizes
  - One correlation angle per coordinate pair
  - Mutation according to covariance matrix: N(**0**, **C**)

$$\vec{a} = ((x_1,\ldots,x_n),(\sigma_1,\ldots,\sigma_n),(\alpha_1,\ldots,\alpha_{n(n-1)/2}))$$

Evolution Strategy:

Algorithms
Mutation

Operators: Mutation – one s

- Self-adaptive ES with one step size:
  - One σ controls mutation for all x_i
  - Mutation: $N(0, \sigma)$

$$\vec{a} = ((x_1,\ldots,x_n),\sigma)$$ Individual before mutation

$$\vec{a}' = ((x_1',\ldots,x_n'),\sigma')$$ Individual after mutation

$$\sigma' = \sigma \cdot \exp(\tau_0 \cdot N(0,1))$$ 1.: Mutation of step sizes

$$x_i' = x_i + \sigma' \cdot N_i(0,1)$$ 2.: Mutation of objective variables

Here the new σ' is used!

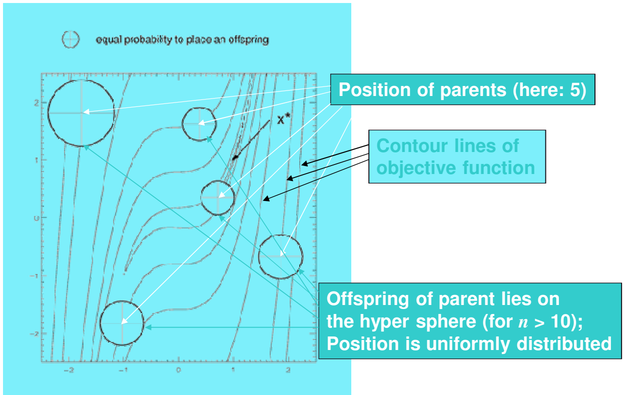
Operators: Mutation – one σ

- Thereby τ_0 is the so-called learning rate
  - Affects the speed of the σ-Adaptation
  - τ_0 bigger: faster but more imprecise
  - τ_0 smaller: slower but more precise
  - How to choose τ_0?
  - According to recommendation of Schwefel*:

$$\tau_0 = \frac{1}{\sqrt{n}}$$

*H.-P. Schwefel: Evolution and Optimum Seeking, Wiley, NY, 1995.

Operators: Mutation – one σ



33

Pros and Cons: One σ

- Advantages:
  - Simple adaptation mechanism
  - Self-adaptation usually fast and precise
- Disadvantages:
  - Bad adaptation in case of complicated contour lines
  - Bad adaptation in case of very differently scaled object variables
    - -100 < x_i < 100 and e.g. -1 < x_j < 1

34

Operators: Mutation – individual σ_i

- Self-adaptive ES with individual step sizes:
  - One σ_i per x_i
  - Mutation: $N_i(0, \sigma_i)$

$$\vec{a} = ((x_1,\ldots,x_n),(\sigma_1,\ldots,\sigma_n))$$
$$\vec{a}' = ((x_1',\ldots,x_n'),(\sigma_1',\ldots,\sigma_n'))$$
$$\sigma_i' = \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1))$$
$$x_i' = x_i + \sigma_i' \cdot N_i(0,1)$$

Individual before Mutation

Individual after Mutation

1.: Mutation of individual step sizes

2.: Mutation of object variables

The new individual σ_i' are used here!

35

Operators: Mutation – individual σ_i

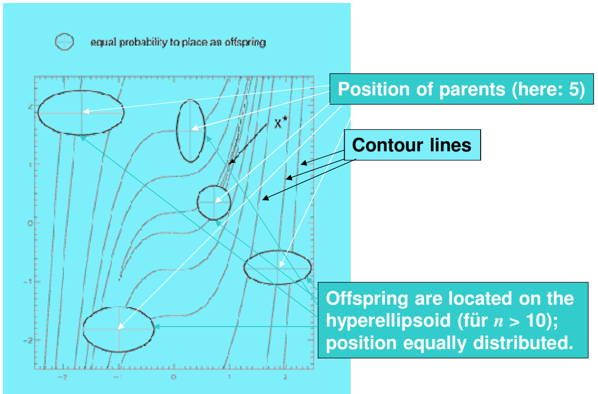
- τ, τ' are learning rates, again
  - τ': Global learning rate
  - N(0,1): Only one realisation
  - τ: local learning rate
  - $N_i(0,1)$: n realisations
  - Suggested by Schwefel*:

$$\tau' = \frac{1}{\sqrt{2n}} \qquad \tau = \frac{1}{\sqrt{2\sqrt{n}}}$$

*H.-P. Schwefel: Evolution and Optimum Seeking, Wiley, NY, 1995.

36

Operators: Mutation – individual σ_i



Pros and Cons: Individual σ_i

- Advantages:
  - Individual scaling of object variables
  - Increased global convergence reliability
- Disadvantages:
  - Slower convergence due to increased learning effort
  - No rotation of coordinate system possible
    - Required for badly conditioned objective function

Operators: Correlated Mutations

- Self-adaptive ES with correlated mutations:
  - Individual step sizes
  - One rotation angle for each pair of coordinates
  - Mutation according to covariance matrix: N(**0**, **C**)

$$\vec{a} = ((x_1,\dots,x_n),(\sigma_1,\dots,\sigma_n),(\alpha_1,\dots,\alpha_{n(n-1)/2}))$$ Individual before

$$\vec{a}' = ((x_1',\dots,x_n'),(\sigma_1',\dots,\sigma_n'),(\alpha_1',\dots,\alpha_{n(n-1)/2}'))$$ Individual after mutation

$$\sigma_i' = \sigma_i \cdot \exp(\tau' \cdot N(0,1) + \tau \cdot N_i(0,1))$$ 1.: Mutation of Individual step sizes

$$\alpha_j' = \alpha_j + \beta \cdot N_j(0,1)$$ 2.: Mutation of rotation angles

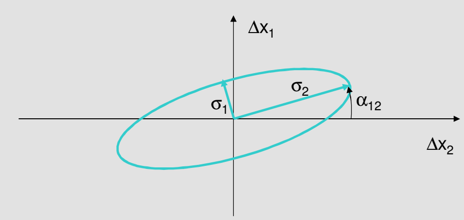
3.: Mutation of object variables

$$x_i' = x_i + \vec{N}(\vec{0}, C')$$ New convariance matrix C' used here!

Operators: Correlated Mutations

- Interpretation of rotation angles α_{ij}
- Mapping onto convariances according to

$$c_{ij(i\neq j)} = \frac{1}{2}(\sigma_i^2 - \sigma_j^2)\tan(2\alpha_{ij})$$



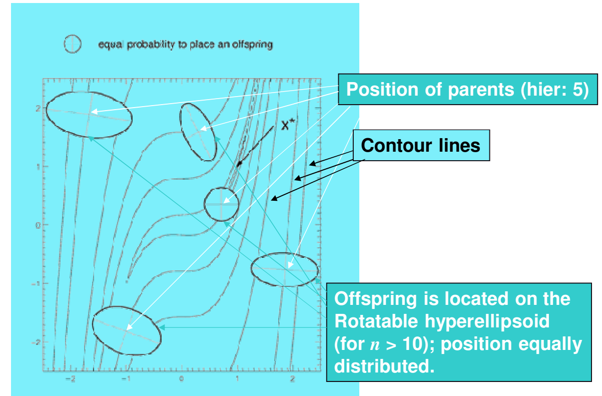
Operators: Correlated Mutation

- τ, τ‘, β are again learning rates
  - τ, τ‘ as before
  - β = 0,0873 (corresponds to 5 degree)
  - Out of boundary correction:

$$|\alpha_j'| > \pi \Rightarrow \alpha_j' \leftarrow \alpha_j' - 2\pi \cdot sign(\alpha_j')$$

41

Correlated Mutations for ES



42

Operators: Correlated Mutations

- How to create $\vec{N}(\vec{0}, C')$?
  - Multiplication of uncorrelated mutation vector with $n(n-1)/2$ rotational matrices

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

  - Generates only feasible (positiv definite) correlation matrices

43

Operators: Correlated Mutations

- Structur of rotation matrix

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

44

Operators: Correlated Mutations

- Implementation of correlated mutations

```
nq := n(n-1)/2;
for i:=1 to n do
   σu[i] := σ[i] * N_i(0,1);
for k:=1 to n-1 do
   n1 := n-k;
   n2 := n;
   for i:=1 to k do
        d1     := σu[n1]; d2:= σu[n2];
        σu[n2] := d1*sin(α[nq])+ d2*cos(α[nq]);
        σu[n1] := d1*cos(α[nq])- d2*sin(α[nq]);
        n2     := n2-1;
        nq     := nq-1;
   od
od
```

Generation of the uncorrelated mutation vector

Rotations

45

Pros and Cons: Correlated Mutations

- Advantages:
  - Individual scaling of object variables
  - Rotation of coordinate system possible
  - Increased global convergence reliability
- Disadvantages:
  - Much slower convergence
  - Effort for mutations scales quadratically
  - Self-adaptation very inefficient

46

Operators: Mutation – Addendum

- Generating N(0,1)-distributed rnd numbers?

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

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

47

Evolution Strategy:

Algorithms
Recombination

48

Operators: Recombination

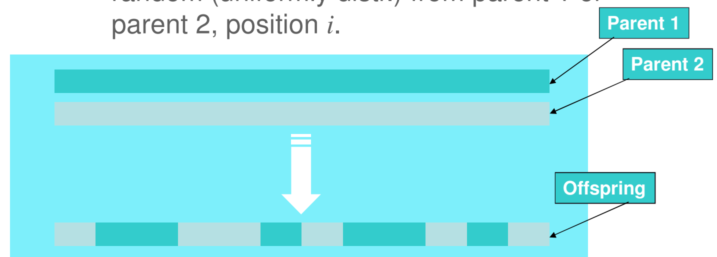
- Only for $\mu > 1$
- Directly after Selection
- Iteratively generates λ offspring:

```
for i:=1 to λ do
   choose recombinant r1 uniformly at random
        from parent_population;
   choose recombinant r2 <> r1 uniformly at random
        from parent population;
   offspring := recombine(r1,r2);
   add offspring to offspring_population;
od
```

49

Operators: Recombination

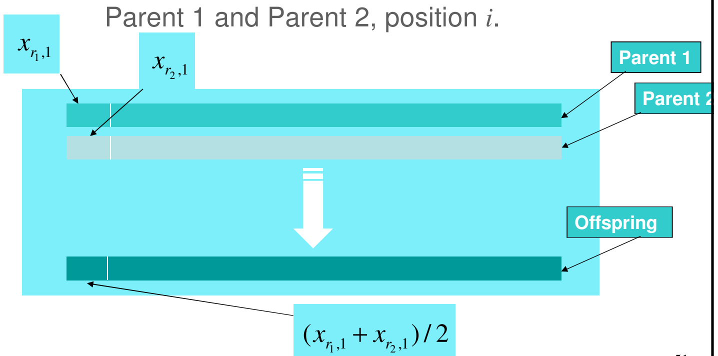
- How does recombination work?
- Discrete recombination:
  - Variable at position i will be copied at random (uniformly distr.) from parent 1 or parent 2, position i.



50

Operators: Recombination

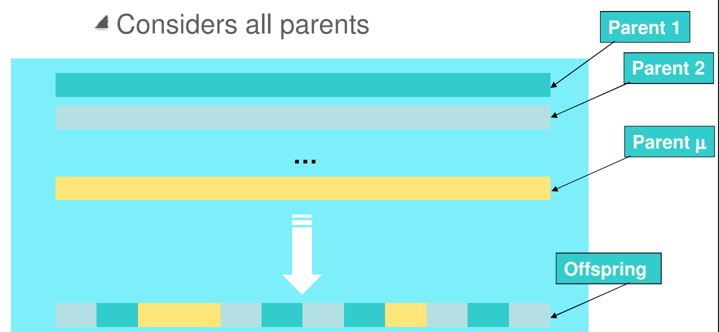
- Intermediate recombination:
  - Variable at position i is arithmetic mean of Parent 1 and Parent 2, position i.



51

Operators: Recombination

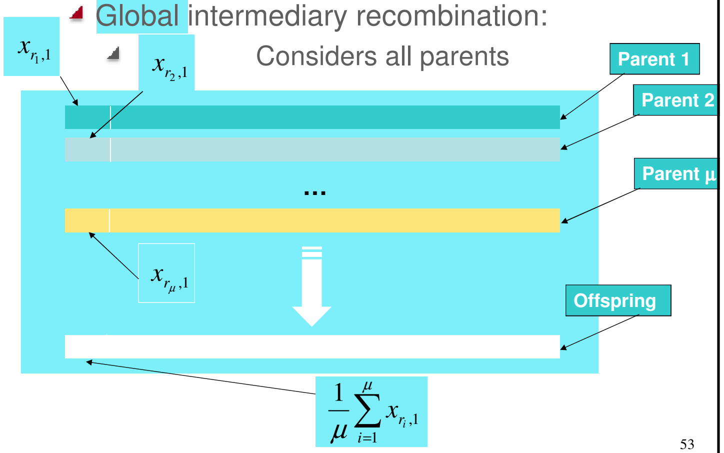
- Global discrete recombination:
  - Considers all parents



52

Operators: Recombination

- Global intermediary recombination:
  - Considers all parents



53

Evolution Strategy

Algorithms
Selection

54

Operators: (μ+λ)-Selection

= New solution candidate

- (μ+λ)-Selection means:
  - μ parents produce λ offspring by
    - (Recombination +)
    - Mutation
  - These μ+λ individuals will be considered together
  - The μ best out of μ+λ will be selected („survive“)
    - Deterministic selection
  - This method guarantees monotony
    - Deteriorations will never be accepted

Recombination may be left out
Mutation always exists!

55

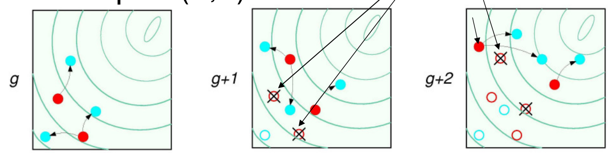
Operators: (μ,λ)-Selection

- (μ,λ)-Selection means:
  - μ parents produce λ >> μ offspring by
    - (Recombination +)
    - Mutation
  - λ offspring will be considered alone
  - The μ best out of λ offspring will be selected
    - Deterministic selection
  - The method doesn‘t guarantee monotony
    - Deteriorations are possible
    - The best objective function value in generation $t+1$ may be worse than the best in generation t.

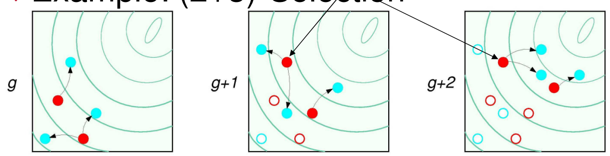
56

Operators: Selection

- Example: (2,3)-Selection



- Example: (2+3)-Selection



57

Operators: Selection

Exception!

- Possible occurrences of selection
  - (1+1)-ES: One parent, one offspring, 1/5-Rule
  - (1,λ)-ES: One Parent, λ offspring
    - Example: (1,10)-Strategy
    - One step size / n self-adaptive step sizes
    - Mutative step size control
    - Derandomized strategy
  - (μ,λ)-ES: $\mu > 1$ parents, $\lambda > \mu$ offspring
    - Example: (2,15)-Strategy
    - Includes recombination
    - Can overcome local optima
  - (μ+λ)-strategies: elitist strategies

58

Evolution Strategy:

Self adaptation of step sizes

59

Self-adaptation

- No deterministic step size control!
- Rather: Evolution of step sizes
  - Biology: Repair enzymes, mutator-genes
- Why should this work at all?
  - Indirect coupling: step sizes – progress
  - Good step sizes improve individuals
  - Bad ones make them worse
  - This yields an indirect step size selection

60

Self-adaptation: Example

- How can we test this at all?
- Need to know optimal step size …
  - Only for very simple, convex objective functions
  - Here: Sphere model

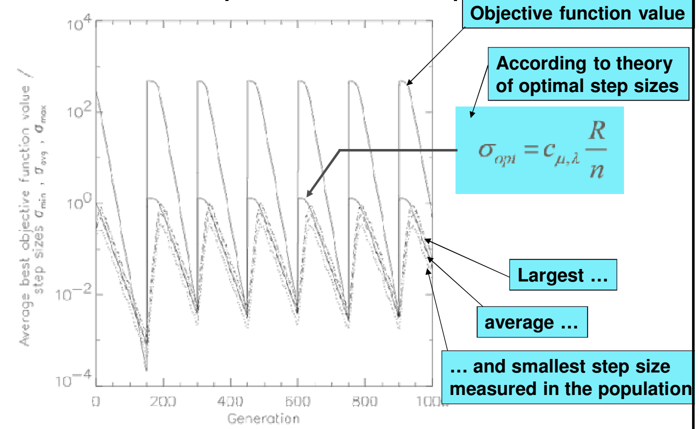
$$f(\vec{x}) = \sum_{i=1}^{n} (x_i - x_i^*)^2$$

$\vec{x}^*$: Optimum

  - Dynamic sphere model
    - Optimum locations changes occasionally

61

Self-adaptation: Example



62

Self-adaptation

- Self-adaptation of one step size
  - Perfect adaptation
  - Learning time for back adaptation proportional n
  - Proofs only for convex functions
- Individual step sizes
  - Experiments by Schwefel
- Correlated mutations
  - Adaptation much slower

63

Evolution Strategy:

Derandomization

64

Derandomization

- Goals:
  - Fast convergence speed
  - Fast step size adaptation
  - Precise step size adaptation
  - Compromise convergence velocity – convergence reliability
- Idea: Realizations of $N(0, \sigma)$ are important!
  - Step sizes and realizations can be much different from each other
  - Accumulates information over time

65

Derandomized (1,λ)-ES

- Current parent: $\vec{x}^g$ in generation g
- Mutation ($k=1,\ldots,\lambda$):

$$\vec{x}^g_{N_k} = \vec{x}^g + \delta^g \cdot \vec{\delta}^g_{scal} \cdot \vec{Z}_k$$

Offspring k

Global step size in generation g

$\vec{Z} = (z_1, \ldots, z_n) \qquad z_i \sim N(0,1)$

Individual step sizes in generation g

- Selection: Choice of best offspring

$$\vec{x}^{g+1} = \vec{x}^g_{N_{sel}}$$

Best of λ offspring in generation g

66

Derandomized (1,λ)-ES

- Accumulation of selected mutations:

$$\vec{Z}^g_A = (1-c) \cdot \vec{Z}^{g-1}_A + c \cdot \vec{Z}_{sel}$$

The particular mutation vector, which created the parent!

- Also: weighted history of good mutation vectors!
- Initialization:

$$\vec{Z}^0_A = \vec{0}$$

- Weight factor: $c = \dfrac{1}{\sqrt{n}}$

67

Derandomized (1,λ)-ES

- Step size adaptation:

Norm of vector

$$\delta^{g+1} = \delta^g \cdot \left(\exp\left(\frac{\left|\vec{Z}^g_A\right|}{\sqrt{n} \cdot \sqrt{\frac{c}{2-c}}} \right) - 1 + \frac{1}{5n} \right)^{\beta}$$

Vector of absolute values

$$\vec{\delta}^{g+1}_{scal} = \vec{\delta}^g_{scal} \cdot \left(\frac{\left|\vec{Z}^g_A\right|}{\sqrt{\frac{c}{2-c}}} + 0.35 \right)^{\beta_{scal}}$$

Regulates adaptation speed and precision

68

Derandomized (1,λ)-ES

- Explanations:
  - Normalization of average variations in case of missing selection (no bias):

$$\sqrt{\frac{c}{2-c}}$$

  - Correction for small n: $1/(5n)$
  - Learning rates:

$$\beta = \sqrt{1/n}$$
$$\beta_{scal} = 1/n$$

69

Evolution Strategy:

Rules of thumb

70

Some Theory Highlights

- Convergence velocity:

$$\varphi \sim 1/n$$

Problem dimensionality

- For (1,λ)-strategies:

$$\varphi \sim \ln \lambda$$

Speedup by λ is just logarithmic – more processors are only to a limited extend useful to increase φ.

- For (μ,λ)-strategis (discrete and intermediary recombination)

$$\varphi \sim \mu \ln \frac{\lambda}{\mu}$$

Genetic Repair Effect of recombination!

71

...

- For strategies with global intermediary recombination:

$$\lambda = 4 + \lfloor 3 \log n \rfloor$$
$$\mu = \lfloor \lambda / 2 \rfloor$$

n	λ	μ
10	10.91	5.45
20	12.99	6.49
30	14.20	7.10
40	15.07	7.53
50	15.74	7.87
60	16.28	8.14
70	16.75	8.37
80	17.15	8.57
90	17.50	8.75
100	17.82	8.91
110	18.10	9.05
120	18.36	9.18
130	18.60	9.30
140	18.82	9.41
150	19.03	9.52

- Good heuristic for (1,λ):

$$\lambda = 10$$

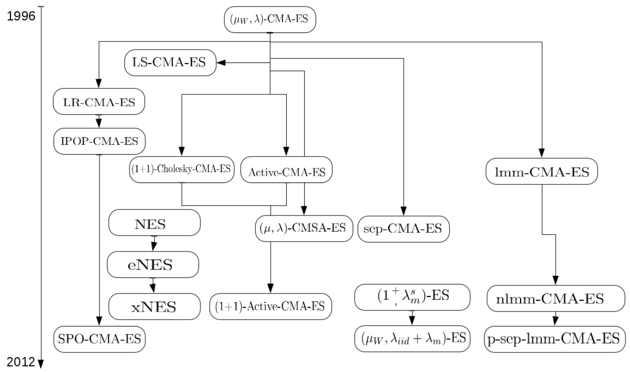
- General:

$$\lambda \approx 7\mu$$

72

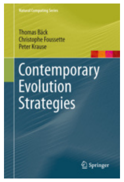
And beyond CMA-ES …

❖ Many strategy variations since 1996



For an Overview:

❖ Bäck, Foussette, Krause: *Contemporary Evolution Strategies*. Natural Computing Series, Springer, 2013.



Empirical Investigation

❖ ES algorithms investigated

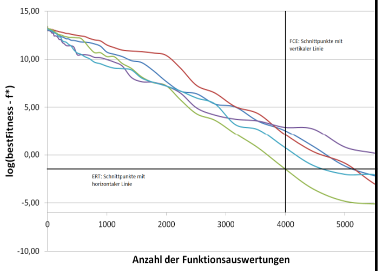
- 9 modern algorithms:
  CMA-ES (with weighted recombination), Active-CMA-ES, (1+1)-Cholesky-CMA-ES, (1+1)-Active-CMA-ES, xNES, sep-CMA-ES, (μ,λ)-CMSA-ES, LS-CMA-ES, $(1,4_m)$-ES
- 5 „old“ algorithms
  - (1+1)-ES
  - (1, λ)-MSC-ES
  - DR1 bzw. (1, λ)-ES with derandomized mutative step-size
  - DR2 bzw. (1, λ)-ES with derandomized mutative step-size control using accumulated information
  - DR3 bzw. (1, λ)-GSA-ES (generating set adaptation)
- Matlab/Octave author‘s implementations for CMA-ES, Active-CMA-ES, $(1,4_m)$-ES and xNES, own implementations for the other ones

Empirical Investigation

❖ Efficiency measures

❖ Expected Running Time (ERT)
- Horizontal viewpoint
- Recommended in BBOB
- Advantage due to quantitative comparison
- For runs with very small number of function evaluations, choice of fixed error Δf* is problematic

❖ Fixed Cost Error (FCE)
- Vertical point of view
- Allows only for qualitative comparison
- Directly applicable for runs with very small number of function evaluations



Empirical Investigation

❖ BBOB Framework

- Framework with 24 test functions
  - 3 unimodal, separable (e.g., Sphere)
  - 7 unimodal, not separable (e.g., Bent Cigar)
  - 2 multimodal, separable (e.g., Rastrigin)
  - 12 multimodal, not separable (e.g., Schwefel)
- Saves convergence behavior as tuples (number of function evaluations, Δf*)
- Dimensionality n: 2, 5, 10, 20, 40 and 100
- 100 runs per setting, algorithm and test function, with 100n function evaluations

Empirical Investigation

❖ Performance assessment

- Using non-parametric tests (Studen-t) significant differences ($p < 0.05$) of FCE-values of two competing algorithms are counted as a „win“ for the better algorithm
- Difference peak performance and average performance
  - peak performance: For each k=5 runs, only the best is used for the assessment
  - average performance: Considers all FCE
- Accumulated rank orders
  - For each test function class
  - For all test functions

Empirical Investigation

❖ Results: Ranks for all test functions

ES	$n=2$		$n=5$		$n=10$		$n=20$		$n=40$		$n=100$	
	p	a	p	a	p	a	p	a	p	a	p	a
(1+1)-ES	8	9	9	9	7	8	6	7	8	9	9	7
(μ,λ)-MSC-ES	10	7	13	11	13	13	13	13	11	13	11	11
DR1	5	6	6	4	6	5	9	6	7	5	10	9
DR2	6	10	5	6	4	4	4	4	3	3	3	3
DR3	12	14	12	13	12	12	11	12	12	12	12	12
(μ_W,λ)-CMA-ES	3	4	2	2	2	3	2	2	2	2	2	2
LS-CMA-ES	11	12	11	11	10	11	10	10	10	10	5	5
(1+1)-Cholesky-CMA-ES	7	10	7	10	9	9	7	7	9	8	7	6
Active-CMA-ES	1	2	1	1	1	1	1	1	1	1	1	1
(μ,λ)-CMSA-ES	13	13	7	8	8	6	5	5	4	3	4	4
sep-CMA-ES	14	5	14	14	14	14	14	14	14	14	14	14
(1+1)-Active-CMA-ES	2	8	4	7	5	7	7	9	5	7	8	7
$(1,4^s_m)$-CMA-ES	4	3	2	3	3	2	3	3	6	6	6	10
xNES	9	1	10	5	11	9	12	11	13	11	13	13

Tabelle 2: Aggregierte Rangfolgen über alle 24 Testfunktionen mit Zielkosten $100 \cdot n$ für die Dimensionen $n \in \{2, 5, 10, 20, 40, 100\}$. Die Spalten p zeigen die Rangfolge für die *peak performance* und die Spalten a stellen die Rangfolge für die *average performance* dar. Rang 1 ist das beste, Rang 14 das schlechteste ES-Verfahren.

Empirical Investigation

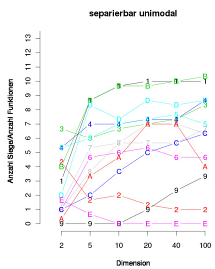
❖ Results: For 100n function evaluations

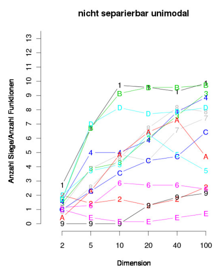
- Active-CMA-ES wins, followed by (μ,λ)-CMA-ES
  - „aktive“ covariance matrix adaptation very successful
- DR2 is a useful alternative to using (μ,λ)-CMA-ES for large n ($O(n^2)$ instead of $O(n^3)$)
- sep-CMA-ES (linear complexity) performs worst

Empirical Investigations

❖ Results: For 25n and 50n function evaluations

- Especially for unimodal functions, various (1+1) methods outperform CMA-ES





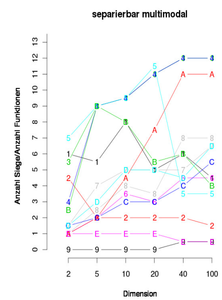
1: (1+1)-ES
2: (μ,λ)-MSC-ES
3: (μ,λ)-CMA-ES
4: Active-CMA-ES
5: $(1,4_m)$-CMA-ES
6: xNES
7: DR1
8: DR2
9: DR3
A: (μ,λ)-CMSA-ES
B: (1+1)-Cholesky-CMA-ES
C: LS-CMA-ES
D: (1+1)-Active-CMA-ES
E: sep-CMA-ES

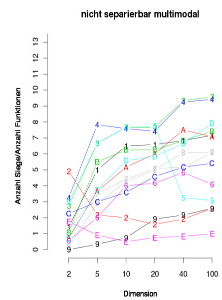
81

Empirical Investigations

❖ Results: For 25n function evaluations

- Für multimodal functions, Active-CMA-ES are (μ,λ)-CMA-ES are clear winners





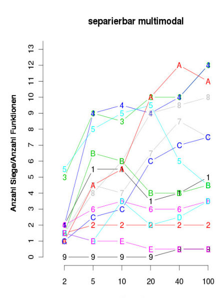
1: (1+1)-ES
2: (μ,λ)-MSC-ES
3: (μ,λ)-CMA-ES
4: Active-CMA-ES
5: $(1,4_m)$-CMA-ES
6: xNES
7: DR1
8: DR2
9: DR3
A: (μ,λ)-CMSA-ES
B: (1+1)-Cholesky-CMA-ES
C: LS-CMA-ES
D: (1+1)-Active-CMA-ES
E: sep-CMA-ES

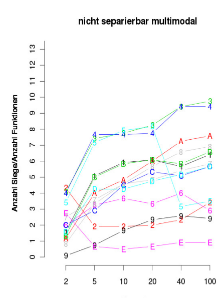
82

Empirical Investigations

❖ Results: For 50n function evaluations

- Für multimodal functions, Active-CMA-ES are (μ,λ)-CMA-ES are clear winners





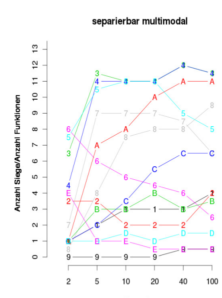
1: (1+1)-ES
2: (μ,λ)-MSC-ES
3: (μ,λ)-CMA-ES
4: Active-CMA-ES
5: $(1,4_m)$-CMA-ES
6: xNES
7: DR1
8: DR2
9: DR3
A: (μ,λ)-CMSA-ES
B: (1+1)-Cholesky-CMA-ES
C: LS-CMA-ES
D: (1+1)-Active-CMA-ES
E: sep-CMA-ES

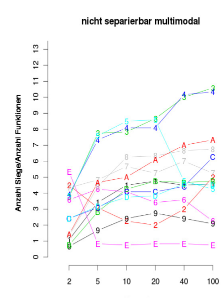
83

Empirical Investigations

❖ Results: For 100n function evaluations

- Für multimodal functions, Active-CMA-ES are (μ,λ)-CMA-ES are clear winners





1: (1+1)-ES
2: (μ,λ)-MSC-ES
3: (μ,λ)-CMA-ES
4: Active-CMA-ES
5: $(1,4_m)$-CMA-ES
6: xNES
7: DR1
8: DR2
9: DR3
A: (μ,λ)-CMSA-ES
B: (1+1)-Cholesky-CMA-ES
C: LS-CMA-ES
D: (1+1)-Active-CMA-ES
E: sep-CMA-ES

84

Summary

- Empirical investigation for very small number of function evaluations
  - (μ_w,λ)-CMA-ES with active covariance matrix adaptation is clear winner
  - For 25n and 50n function evaluations, (1+1) are useful as well

Mixed-Integer Evolution Strategies

Mixed-Integer Evolution Strategy

- Generalized optimization problem:

$$f(r_1,\ldots,r_{n_r},z_1,\ldots,z_{n_z},d_1,\ldots,d_{n_d}) \to min$$

subject to:

$$r_i \in [r_i^{min}, r_i^{max}] \subset \mathbb{R},\ \ i=1,\ldots,n_r$$

$$z_i \in [z_i^{min}, z_i^{max}] \subset \mathbb{Z},\ \ i=1,\ldots,n_z$$

$$d_i \in D_i = \{d_{i,1},\ldots,d_{i,|D_i|}\}, i=1,\ldots,n_d$$

Mixed-Integer ES: Mutation

for $i = 1,\ldots,n_r$ **do**
  $s_i' \leftarrow s_i \exp(\tau_g N_g + \tau_l N(0,1))$
  $r_i' = r_i + N(0, s_i')$
end for
for $i = 1,\ldots,n_z$ **do**
  $q_i' \leftarrow q_i \exp(\tau_g N_g + \tau_l N(0,1))$
  $z_i' \leftarrow z_i + G(0, q_i')$
end for
$p_i' := 1/[1 + \frac{1-p_i}{p_i} * \exp(-\tau_l * N(0,1))]$
for $i \in \{1,\ldots,n_d\}$ **do**
  if $U(0,1) < p_i'$ **then**
    $d_i' \leftarrow$ uniformly randomly value from D_i
  end if
end for

Learning rates (global)

Learning rates (global)

Geometrical distribution

Mutation Probabilities

Some Application Examples

Mostly Engineering Problems

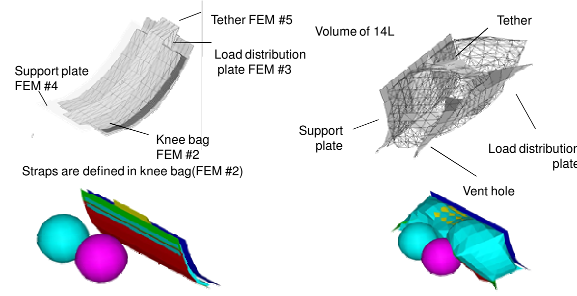
89

Examples I: Inflatable Knee Bolster Optimization

Low Cost ES: **0.677**
GA (Ford): 0.72
Hooke Jeeves DoE: 0.88

mazda — Initial position of knee bag model

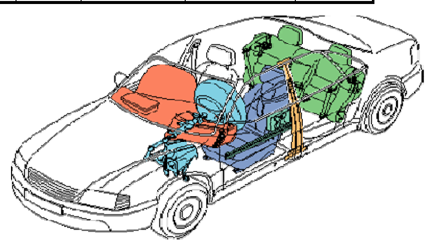
Ford — deployed knee bag (unit only)



90

IKB: Previous Designs

# Variables	Characteristics	HIC	CG	Left foot load	Right foot load	$P_{Combined}$
4	Unconstrained	576,324	44,880	4935	3504	12,393
5	Unconstrained	384,389	41,460	4707	4704	8,758
9	Unconstrained	292,354	38,298	5573	5498	6,951
10	Constrained	305,900	39,042	6815	6850	7,289



91

IKB: Problem Statement

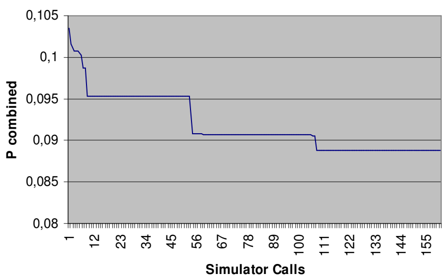
- Objective: Min Ptotal

Subject to: Left Femur load <= 7000
Right Femur load <= 7000

Design Variable	Description	Base Design 1	Base Design 2	GA (Yan Fu)
dx	IKB center offset x	0	0	0,01
dz	IKB center offset y	0	0	-0,01
rcdex	KB venting area ratio	1	1	2
massrat	KB mass inflow ratio	1	1	1,5
rcdexd	DB venting area ratio	1	1	2,5
Dmassratf	DB high output mass inflow ratio	1	1	1,1
Dmassratl	DB low output mass inflow ratio	1	1	1
dbfire	DB firing time	0	0	-0,003
dstraprat	DB strap length ratio	1	1	1,5
emr	Load of load limiter (N)	3000	3000	2000
Performance Response	**Description**			
NCAP_HIC_50	HIC	590	555.711	305,9
NCAP_CG_50	CG	47	47.133	39,04
NCAP_FMLL_50	Left foot load	760	6079	6815
NCAP_FMRL_50	Right foot load	900	5766	6850
P combined (Quality)		13.693	13.276	7,289

92

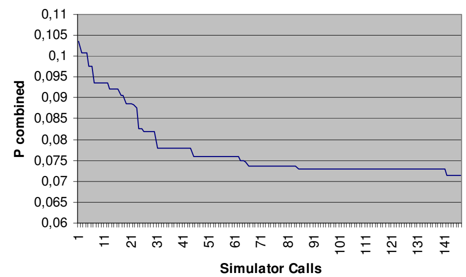
IKB Results I: Hooke-Jeeves



Quality: 8.888 Simulations: 160

93

IKB Results II: (1+1)-ES



Quality: 7.142 Simulations: 122

94

Engineering Optimization

95

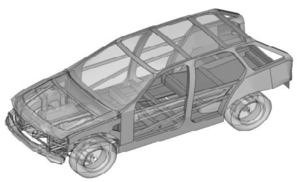
Safety Optimization – Pilot Study

- Aim: Identification of most appropriate Optimization Algorithm for realistic example!
- Optimizations for 3 test cases and 14 algorithms were performed (28 x 10 = 280 shots)
  - Body MDO Crash / Statics / Dynamics
  - MCO B-Pillar
  - MCO Shape of Engine Mount
- NuTech’s ES performed significantly better than Monte-Carlo-scheme, GA, and Simulated Annealing
- Results confirmed by statistical hypothesis testing

96

MDO Crash / Statics / Dynamics

- Minimization of body mass
- Finite element mesh
  - Crash ~ 130.000 elements
  - NVH ~ 90.000 elements
- Independent parameters: Thickness of each unit: 10
- Constraints: 18

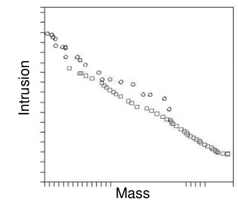


Algorithm	Avg. reduction (kg)	Max. reduction (kg)	Min. reduction (kg)
Best so far	-6.6	-8.3	-3.3
Our ES	-9.0	-13.4	-6.3

97

MCO B-Pillar – Side Crash

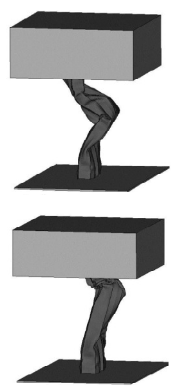
- Minimization of mass & displacement
- Finite element mesh
  - ~ 300.000 elements
- Independent parameters: Thickness of 10 units
- Constraints: 0
- ES successfully developed Pareto front



98

MCO Shape of Engine Mount

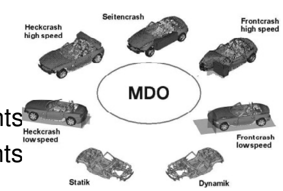
- Mass minimal shape with axial load > 90 kN
- Finite element mesh
  - ~ 5000 elements
- Independent parameters: 9 geometry variables
- Dependent parameters: 7
- Constraints: 3
- ES optimized mount
  - less weight than mount optimized with best so far method
  - geometrically better deformation



99

Safety Optimization – Example

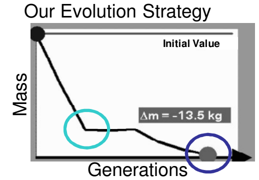
- Production Run !
- Minimization of body mass
- Finite element mesh
  - Crash ~ 1.000.000 elements
  - NVH ~ 300.000 elements
- Independent parameters:
  - Thickness of each unit: 136
- Constraints: 47, resulting from various loading cases
- 180 (10 x 18) shots ~ 12 days
- No statistical evaluation due to problem complexity



100

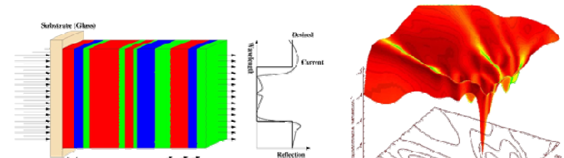
Safety Optimization – Example of use

Our Evolution Strategy



- 13,5 kg weight reduction by ES
- Beats best so far method significantly
- Typically faster convergence velocity of ES
  ~ 45% less time (~ 3 days saving) for comparable quality needed
- Still potential of improvements after 180 shots.
- Reduction of development time from 5 to 2 weeks allows for process integration

Optical Coatings: Design Optimization



- Nonlinear mixed-integer problem, variable dimensionality.
- Minimize deviation from desired reflection behaviour.
- Excellent synthesis method; robust and reliable results.

Dielectric Filter Design Problem

Client:

Corning, Inc.,
Corning, NY

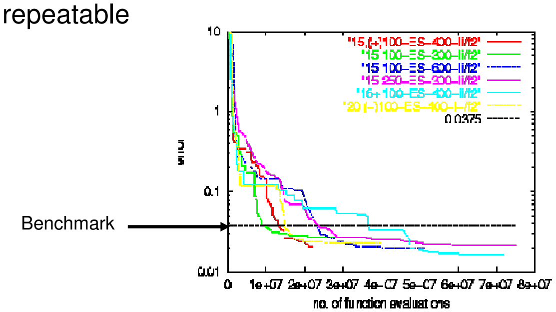
- Dielectric filter design.
- n=40 layers assumed.
- Layer thicknesses xi in [0.01, 10.0].
- Quality function: Sum of quadratic penalty terms.

$$quality = \sum_{i=1}^{15} weight \cdot \left(\frac{calculated - desired}{scale}\right)^2 \rightarrow \min$$

- Penalty terms = 0 iff constraints satisfied.

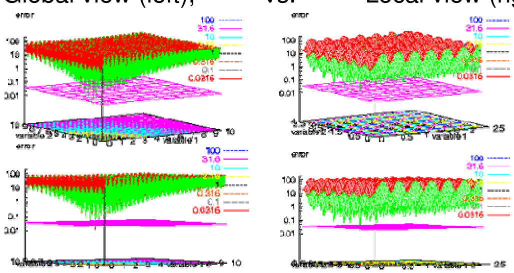
Results: Overview of Runs

- Factor 2 in quality.
- Factor 10 in effort.
- Reliable, repeatable results.



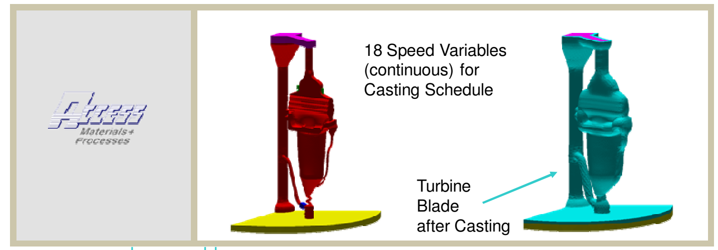
Problem Topology Analysis: An Attempt

- Grid evaluation for 2 variables.
- Close to the optimum (from vector of quality 0.0199).
- Global view (left), vs. Local view (right).



105

Bridgman Casting Process

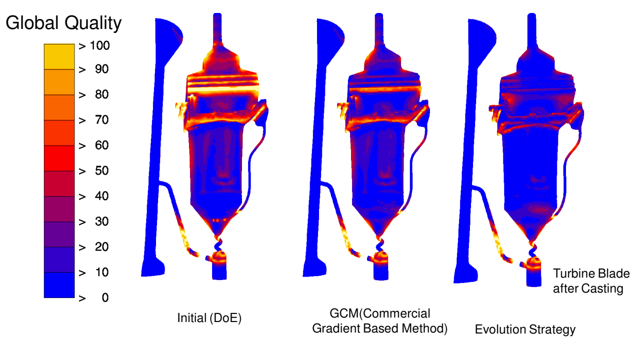


large problem:

- run time varies: 16 h 30 min to 32 h (SGI, Origin, R12000, 400 MHz)
- at each run: 38,3 GB of view factors (49 positions) are treated!

106

Bridgman Casting Process



Quality Comparison of the Initial and Optimized Configurations

107

Traffic Light Control

Ministerie van Verkeer en Waterstaat

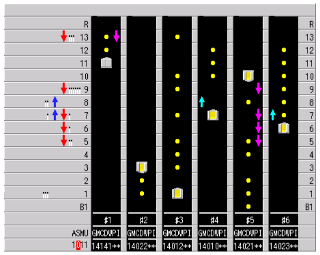
- Generates green times for next switching schedule.
- Minimization of total delay / number of stops.
- Better results (3 – 5%) / higher flexibility than with traditional controllers.
- Dynamic optimization, depending on actual traffic (measured by control loops).

- **Client:**
  Dutch Ministry of Traffic
  Rotterdam, NL

108

Elevator Control

FUJITEC

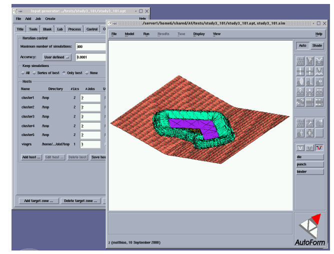


- Minimization of passenger waiting times.
- Better results (3 – 5%) / higher flexibility than with traditional controllers.
- Dynamic optimization, depending on actual traffic.

- **Client:**
  Fujitec Co. Ltd., Osaka, Japan

109

Metal Stamping Process

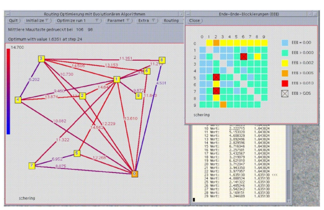


- Minimization of defects in the produced parts.
- Optimization on geometric parameters and forces.
- Fast algorithm; finds very good results.

- **Client:**
  AutoForm Engineering GmbH, Dortmund

110

Network Routing

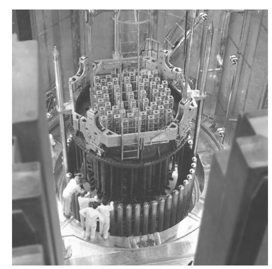


- Minimization of end-to-end-blockings under service constraints.
- Optimization of routing tables for existing, hard-wired networks.
- 10%-1000% improvement.

- **Client:**
  SIEMENS AG, München

111

Nuclear Reactor Refueling



- Minimization of total costs.
- Creates new fuel assembly reload patterns.
- Clear improvements (1%-5%) of existing expert solutions.
- Huge cost saving.

- **Client:**
  SIEMENS AG, München

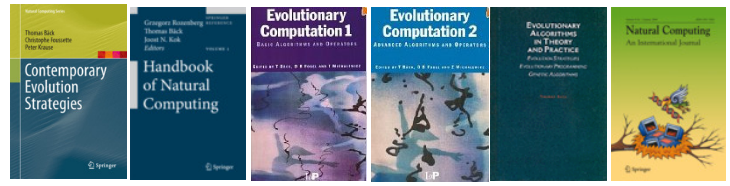
112

Business Issues

- Supply Chain Optimization
- Scheduling & Timetabling
- Product Development, R&D
- Management Decision Making, e.g., project portfolio optimization
- Optimization of Marketing Strategies; Channel allocation
- Multicriteria Optimization (cost / quality)
- ... And many others

113

Literature …



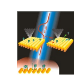
114

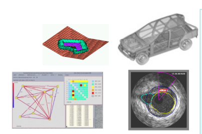
Leiden Institute of Advanced Computer Science (LIACS)

- See www.liacs.nl and http://natcomp.liacs.nl
- Masters in
  - Comp. Science
  - ICT in Business
  - Media Technology
- Elected „Best Comp. Sci. Study“ by students.
- Excellent job opportunities for our students.
- Research education with an eye on business.

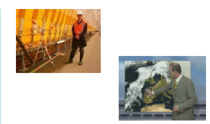
115

LIACS Research





Algorithms
Prof. J.N. Kok, Prof. T. Bäck
- Novel Algorithms
- Data Mining
- Natural Computing
- Applications
  - Drug Design
  - Medicine
  - Engineering
  - Logistics
  - Physics



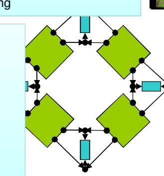
Technology and Innovation Management
Prof. B. Katzy
- Coevolution of Technology and Social Structures
- Entrepreneurship
- Innovation Management

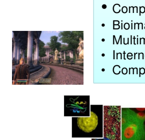
Synergies & Collaboration

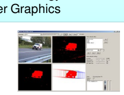
Core Computer Technologies
Prof. H. Wijshoff, Prof. E. Deprettere
- Embedded Systems
- Parallel / Distributed Computing
- Compiler Technology
- Data Mining

Imagery and Media
Dr. M. Lew, Dr. F. Verbeek
- Computer Vision and Audio/Video
- Bioimaging
- Multimedia Search
- Internet Technology
- Computer Graphics

Foundations of Software Technology
Prof. F. Arbab, Prof. J.N. Kok
- Software Systems
- Embedded Systems
- Service Composition
- Multicore Systems
- Formal Methods
- Coordination / Concurrency







116

Leiden Institute of Advanced Computer Science 116

Literature

- H.-P. Schwefel: *Evolution and Optimum Seeking*, Wiley, NY, 1995.
- I. Rechenberg: *Evolutionsstrategie 94*, frommann-holzboog, Stuttgart, 1994.
- Th. Bäck: *Evolutionary Algorithms in Theory and Practice*, Oxford University Press, NY, 1996.
- Th. Bäck, D.B. Fogel, Z. Michalewicz (Hrsg.): *Handbook of Evolutionary Computation, Vols. 1,2*, Institute of Physics Publishing, 2000.
- Th. Bäck, C. Foussette, P. Krause: *Contemporary Evolution Strategies*, Springer, Berlin, 2013.

117