Selected Research Topics in Stochastic Blackbox Optimization

Dimo Brockhoff

KanGAL, IIT Kanpur, December 8, 2010
Dimo Brockhoff

2000-2005
study of CS (Dipl. inform.) in Dortmund, Germany

2005-2009
Dr. sc. ETH at ETH Zurich, Switzerland

2009-2010
postdoc at INRIA Saclay---Ile-de-France

since November 2010
postdoc at Ecole Polytechnique
Main Interests

- theory of EAs
- algorithm design
- blackbox optimization
- Multiobjective single-objective
- stochastic algorithms

Theory

Algorithms ↔ Applications
Blackbox Optimization: The Big Picture

Optimization problems occur frequently in practice...

\[
\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))
\]

where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

objective function

\[ f(X) \]

single-objective: total order
Optimization problems occur frequently in practice...

$$\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))$$

where $x \in X \mapsto f(x) \in \mathbb{R}^k$
Optimization problems occur frequently in practice...

\[
\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))
\]

where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

multiple objectives:

dominance relation (partial order)
Optimization problems occur frequently in practice…

\[
\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))
\]

where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

Pareto Front

multiple objectives:

dominance relation (partial order)
Optimization problems occur frequently in practice...

\[
\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))
\]

where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

Issues:
- non-linear
- noisy
- uncertain
- non-differentiable
- expensive
- (integrated simulations)
- huge search spaces
- many objectives
- many constraints

Pareto Front

X

cost

power

consumption
Blackbox Optimization: The Big Picture

Optimization problems occur frequently in practice...

\[
\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))
\]

where \( x \in X \leftrightarrow f(x) \in \mathbb{R}^k \)

---

Black box optimization

\[x \in X \rightarrow f \rightarrow (f_1(x), \ldots, f_k(x))\]

\(\rightarrow\) no assumptions

---

- many objectives
- huge search spaces
- many constraints

---

© Dimo Brockhoff, LIX, Ecole Polytechnique

Sel. Research Topics in Stochastic BB Optimization, IITK, December 3, 2010
Blackbox Optimization: The Big Picture

Optimization problems occur frequently in practice...

\[
\min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x))
\]

where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

Black box optimization

\( x \in X \rightarrow f \rightarrow (f_1(x), \ldots, f_k(x)) \)

\( \rightarrow \) no assumptions

Evolutionary Algorithms (EAs)

Evolutionary algorithms = randomized search algorithms optimizing on solution sets

© Dimo Brockhoff, LIX, Ecole Polytechnique
Optimization problems occur frequently in practice...

\[ \min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x)) \]

where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

Black box optimization

\( x \in X \rightarrow f \rightarrow (f_1(x), \ldots, f_k(x)) \)

→ no assumptions

Evolutionary Algorithms (EAs)

Evolutionary algorithms = randomized search algorithms

optimizing on solution sets

- robust
- multimodal problems
- “sampling” the (Pareto-) optimal solutions to inform DM
General Principle of Evolutionary Algorithms

How these algorithms work...

- variation
- evaluation
- update
- survival

create $\lambda$ offspring

black box!

select $\mu$ to keep

based on total order on solutions
Objective Reduction

Hypervolume-based search

Mirroring and Sequential Selection
Overview

Objective Reduction

Hypervolume-based search

Mirroring and Sequential Selection

multiobjective

single-objective
Objective Reduction: Motivation

- Which objectives are the most important ones?
- What is the relationship between the objectives?
- Are all objectives necessary or can objectives be omitted?
- Are additional objectives always bad?
- Can efficient methods be developed?
- How can user preferences be incorporated into the search?

→ Learning about the problem
Recall: Multiobjective Optimization

- **w.l.o.g.** \( \min_{x \in X} f(x) = (f_1(x), \ldots, f_k(x)) \)
  
  where \( x \in X \mapsto f(x) \in \mathbb{R}^k \)

- weak Pareto dominance relation wrt set \( \mathcal{F} = \{f_1, \ldots, f_k\} \) of objectives: \( x \preceq_{\mathcal{F}} y \iff \forall f_i \in \mathcal{F} : f_i(x) \leq f_i(y) \)

- incomparable/comparable/indifferent

- \( x^* \in X \) Pareto-optimal: \( \nexists x \in X : x \preceq_{\mathcal{F}} x^* \land x^* \not\preceq_{\mathcal{F}} x \)

**Goal** (without decision maker):

- find or approximate set of Pareto-optimal solutions (Pareto set)
- in practice: as close as possible & as diverse as possible
Set Problem View and Refinements

**set problem:** generalize Pareto dominance on sets

\[ A \preceq \mathcal{F} B \iff \forall b \in B : \exists a \in A : a \preceq \mathcal{F} b \]

**Sought:** total refinement

i.e. a total order on sets that is compliant with dominance

\textit{Definition 2.4:} Given a set \( \Psi \). Then the preference relation \( \preceq_{\text{ref}} \) refines \( \preceq \) if for all \( A, B \in \Psi \) we have

\[ (A \preceq B) \land (B \not\preceq A) \Rightarrow (A \preceq_{\text{ref}} B) \land (B \not\preceq_{\text{ref}} A). \]

from [ztb2010a in IEEETEC’10]
Many-Objective Optimization

Main Problem

weak Pareto dominance gives no search direction with many objectives

Needed:

“more total” order

One Idea:

Reduce the number of objectives automatically

→ omitting objectives results in a refinement!
Automated Objective Reduction

Related Work:

- MCDM approaches: [gl1977a, agre1997a, mali2006a, mt2007a, mt2008a, mali2008a]
  - for linear objectives only
- PCA-based: Deb and Saxena [ds2006a, sd2007a, sd2008b]
  - no control over dominance relation (“what happens?”)

**MOSS: The Minimum Objective Subset Problem**

Given a set $A$ of solutions with relations $\preceq_f \subseteq A \times A$,

Find minimum objective set $\mathcal{F}' \subseteq \mathcal{F}$ preserving the relation ($\preceq_{\mathcal{F}'} = \preceq_{\mathcal{F}}$)

**MOSS is NP-complete**

- Reduction from SETCOVER
- As a result, consideration of objective sets of fixed size is not sufficient
An Example

values

omit cost 2

still the same relations

values

© Dimo Brockhoff, LIX, Ecole Polytechnique
Sel. Research Topics in Stochastic BB Optimization, IITK, December 3, 2010
Algorithms for the MOSS Problem [bz2007d]

Exact algorithm
- Correctness proof
- Runtime: $O(|A|^2 \cdot k \cdot 2^k)$
- Worst case: $\Omega(|A|^2 \cdot 2^{k/3})$

Simple greedy heuristic
- Correctness proof
- Runtime: $O(k \cdot |A|^2)$
- Best possible approximation ratio
  of $\Theta(\log |A|)$ [feigl998a]

$S := \emptyset$

for each pair $x, y \in A$ of solutions do
  \[S_x := \{i \mid i \in \{1, \ldots, k\} \land x \preceq_i y \land y \not\preceq_i x\}\]
  \[S_y := \{i \mid i \in \{1, \ldots, k\} \land y \preceq_i x \land x \not\preceq_i y\}\]
  \[S_{xy} := S_x \cup S_y \text{ where}\]
  \[S_1 \cup S_2 := \{s_1 \cup s_2 \mid s_1 \in S_1 \land s_2 \in S_2\}\]
  \[\land (\exists p_1 \in S_1, p_2 \in S_2 : p_1 \cup p_2 \subseteq s_1 \cup s_2)\}\]
  if $S_{xy} = \emptyset$ then $S_{xy} := \{1, \ldots, k\}$
  \[S := S \cup S_{xy}\]
end for
Output a smallest set $s_{\text{min}}$ in $S$

$E := \preceq^C \cap E$ where $\preceq^C := (A \times A) \setminus \preceq$
$I := \emptyset$
while $E \neq \emptyset$ do
  choose an $i \in \{1, \ldots, k\} \setminus I$
  such that $|\preceq^C \cap E \setminus \preceq_i|$ is maximal
  \[E := E \setminus \preceq^C \cup \{i\}\]
  \[I := I \cup \{i\}\]
end while
Generalizations

1. Exactly conserving dominance structure sometimes too strict
   → $\delta$-error versions (based on $\varepsilon$-dominance) [bz2006d]

\[ \delta = 0.5 \]

2. Omitting Objectives during search might yield bad objective values
   → Aggregation of objectives [bz2010a]

   → Interestingly: also a refinement for weighted sum
Objective Reduction Results (Excerpts)

- Comparison of greedy vs. exact methods for random objectives.
- Graphs showing the number of objectives needed for different problem sizes.
- Bar charts illustrating the number of objectives in minimal sets for various problem instances.

- Observations:
  - For $|A| = 200$ and $k = 25$, there are 25 objectives in the minimal set.
  - For $|A| = 200$ and $k = 25$, similar trends are observed.

- Conclusions:
  - The greedy method performs well compared to exact methods.
  - The number of objectives needed decreases as the number of objectives increases.
Objective Reduction: Open Problems

Online objective reduction
- fast algorithms
- objective aggregation
- objective decomposition
- one idea: using multi-armed bandits
- ...

Decision Making
- what are the most important objectives?
- what can we learn more about the objective relations?
- incorporation of decision space
- ...

Sel. Research Topics in Stochastic BB Optimization, IITK, December 3, 2010

© Dimo Brockhoff, LIX, Ecole Polytechnique
Overview

Objective Reduction

Hypervolume-based search

Mirroring and Sequential Selection

- multiobjective
- single-objective
A “Classical” Algorithm: NSGA-II

Selection:

- 1st criterion: Pareto dominance
- 2nd criterion: crowding distance
- Optimizing crowding distance introduces cycles!

NSGA-II [dapm2002a]

**Dominance depth**

\[ d(i) = \sum_{m=1}^{\sigma} |f_m(i-1) - f_m(i+1)| \]

**Crowding distance**

\[ d_2(i) \]

\[ d_1(i) \]
**A “Classical” Algorithm: NSGA-II**

**Selection:**
- 1\textsuperscript{st} criterion: Pareto dominance
- 2\textsuperscript{nd} criterion: crowding distance
- Optimizing crowding distance introd

![Diagram of NSGA-II selection process](image-url)

**NSGA-II**

```
domiance depth
```

![Graphs comparing NSGA-II and SPEA](image-url)
State-of-the-art algorithms (SMS-EMOA, MO-CMA-ES, HypE, …) use hypervolume indicator as 2nd selection criterion: no cycles! refinement!

Main idea
Delete solutions with the smallest hypervolume loss
\[ d(s) = l_H(P,r) - l_H(P / \{s\},r) \]
iteratively
Optimal $\mu$-Distributions

When the goal is to maximize the hypervolume...

- this yields sets with only Pareto-optimal solutions \cite{flei2003a @ EMO’03}
- those sets, if unrestricted in size, cover the entire Pareto front
- many hypervolume-based EMO algorithms have a population size $\mu$!

Optimal $\mu$-Distribution:
A set of $\mu$ solutions that maximizes the hypervolume indicator among all sets of $\mu$ solutions is called optimal $\mu$-distribution.
Questions:

- how are optimal $\mu$-distributions characterized?
  - understand the bias of the indicator (influence on DM)
  - how can it be changed?
- what is their indicator value?
  - helpful for performance assessment (target values)
- what is the influence of the indicator's parameters on optimal $\mu$-distributions?
  - guidelines for practical usage
- do algorithms converge to optimal $\mu$-distributions?
Optimal $\mu$-Distributions

Questions:

- how are optimal $\mu$-distributions characterized?
  - understand the bias of the indicator (influence on DM)
  - how can it be changed?
- what is their indicator value?
  - helpful for performance assessment (target values)
- what is the influence of the indicator's parameters on optimal $\mu$-distributions?
  - guidelines for practical usage
- do algorithms converge to optimal $\mu$-distributions?
Results for 2 objectives only… (except [abb2010a])

\[
f : x \in D \mapsto f(x)
\]

hypervolume indicator:

\[
I_H((x_1, \ldots, x_\mu)) := \sum_{i=1}^{\mu} (x_{i+1} - x_i)(f(x_0) - f(x_i))
\]
A Necessary Condition [abbz2009a]

**Proposition 1. (Necessary condition for optimal \( \mu \)-distributions)** If \( f \) is continuous, differentiable and \( (x_1^\mu, \ldots, x_{\mu}^\mu) \) denote the \( x \)-coordinates of a set of \( \mu \) points maximizing the hypervolume indicator, then for all \( x_{\min} < x_i^\mu < x_{\max} \)

\[
f'(x_i^\mu)(x_{i+1}^\mu - x_i^\mu) = f(x_i^\mu) - f(x_{i-1}^\mu), \quad i = 1 \ldots \mu \tag{3}
\]

where \( f' \) denotes the derivative of \( f \), \( f(x_0^\mu) = r_2 \) and \( x_{\mu+1}^\mu = r_1 \).

**Proof idea:**

\( I_H \) \( \text{max} \) \( \Rightarrow \) derivative is 0 at each \( x_i^\mu \) or \( x_i^\mu \) is at the boundary of the domain.
Example: equal distances (only) on linear fronts

\[ f : x \in [x_{\text{min}}, x_{\text{max}}] \mapsto \alpha x + \beta \]

\[ \alpha \left( x_{i+1}^{\mu} - x_i^{\mu} \right) = f(x_i^{\mu}) - f(x_{i-1}^{\mu}) = \alpha(x_i^{\mu} - x_{i-1}^{\mu}) \]

generalization of results in [ebn2005a,bne2007a]
**Previous Belief About the Hypervolume**

*“Belief” about Bias:*

- **“biased towards the boundary solutions”** [dmm2005a]
- Focuses on knee points; points less dense on extremes [bne2007a]
- **“convex regions might be preferred to concave regions”** [zt1998b]
A Density Result: When $\mu$ Goes to Infinity

**Observation:**

general front shapes too difficult to investigate for finite $\mu$

**Question:**

can we characterize optimal $\mu$-distributions with respect to a density

\[
\delta(x) = \lim_{\mu \to \infty, h \to 0} \left( \frac{1}{\mu h} \sum_{i=1}^{\mu} 1_{[x, x+h]}(x_i^\mu) \right)
\]

[abz2009a]
Result and Interpretation

The resulting density is

$$\delta(x) = \frac{\sqrt{-f'(x)}}{\int_0^{x_{\text{max}}} \sqrt{-f'(x)} \, dx}$$

How can we interpret this?

- bias only depends on slope of $f$ in contrast to [dmm2005a, zt1998b]
- density highest where slope $= 45^\circ$ compliant to [bne2007a]
- experimental results for finite and small $\mu$ support the result

Conclusion:

only theoretical results make it possible to understand the bias
### How to Use the Result in Performance Assess.

<table>
<thead>
<tr>
<th>Problem</th>
<th>front description</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td>bi-objective sphere</td>
<td>[ f(x) = \left( (b - a) - x^{1/\alpha} \right)^{\alpha} ]</td>
<td>[ \delta(x) = C \cdot \sqrt{(b - a - x^{1/\alpha})^{\alpha - 1} \cdot x^{\frac{\alpha - \alpha}{\alpha}}} ]</td>
</tr>
<tr>
<td>ZDT1, ZDT4 [24]</td>
<td>[ f(x) = 1 - \sqrt{x} ]</td>
<td>[ \delta(x) = \frac{3}{4x^{1/4}} ]</td>
</tr>
<tr>
<td>ZDT2 [24]</td>
<td>[ f(x) = 1 - x^2 \text{ for } x \in [0, 1] ]</td>
<td>[ \delta(x) = \frac{3}{2} \sqrt{x} ]</td>
</tr>
<tr>
<td>ZDT3* [24]</td>
<td>[ f(x) = 1 - \sqrt{x} - x \cdot \sin(10\pi x) ]</td>
<td>[ \delta(x) = 1.5609 \cdot \sqrt{x} \cdot \sin(10\pi x) + 10\pi x \cos(10\pi x) ]</td>
</tr>
<tr>
<td></td>
<td>for all ( x \in F ) where ( F = [0, 0.0830015349] \cup [0.1822287280, 0.2577623634] \cup [0.4093136748, 0.4538821041] \cup [0.6183967944, 0.6525117038] \cup [0.8233317983, 0.8518328654] ]</td>
<td></td>
</tr>
<tr>
<td>ZDT6 [24]</td>
<td>[ f(x) = 1 - x^2 ] for ( x \in \left[ \frac{\arctan(9\pi)}{6\pi}, 1 \right] \approx [0.08146, 1] ]</td>
<td>with ( C = \frac{3}{2} \left( 1 - \frac{\arctan(9\pi)}{6\pi} \right)^{3/2} \approx 1.53570 )</td>
</tr>
<tr>
<td>DTLZ1 [8]</td>
<td>[ f(x) = \frac{1}{2} - x ]</td>
<td>[ \delta(x) = 1 ]</td>
</tr>
<tr>
<td>DTLZ2, DTLZ3, DTLZ4 [8]</td>
<td>[ f(x) = \sqrt{1 - x^2} ]</td>
<td>[ \delta(x) = 1.1803 \cdot \frac{x}{\sqrt{1 - x^2}} ]</td>
</tr>
<tr>
<td>DTLZ7* [8]</td>
<td>[ f(x) = 4 - x(1 + \sin(3\pi x)) ]</td>
<td>[ \delta(x) = 0.6566 \cdot \sqrt{1 + \sin(3\pi x) + 3\pi x \cos(3\pi x)} ]</td>
</tr>
<tr>
<td></td>
<td>for all ( x \in F ) where ( F = [0, 0.2514118361] \cup [0.6314265307, 0.8594008566] \cup [1.3596178368, 1.5148392681] \cup [2.0518383519, 2.1164268079] )</td>
<td></td>
</tr>
</tbody>
</table>
### How to Use the Result in Performance Assess.

<table>
<thead>
<tr>
<th>Problem</th>
<th>front description</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td>bi-objective</td>
<td><img src="image1" alt="ZDT1 and ZDT4" /></td>
<td><img src="image2" alt="ZDT2" /></td>
</tr>
<tr>
<td>ZDT1, ZDT2, ZDT3* [24]</td>
<td><img src="image1" alt="ZDT1 and ZDT4" /></td>
<td><img src="image2" alt="ZDT2" /></td>
</tr>
<tr>
<td>ZDT6 [24]</td>
<td><img src="image1" alt="ZDT1 and ZDT4" /></td>
<td><img src="image2" alt="ZDT2" /></td>
</tr>
<tr>
<td>DTLZ1 [8]</td>
<td><img src="image1" alt="ZDT1 and ZDT4" /></td>
<td><img src="image2" alt="ZDT2" /></td>
</tr>
<tr>
<td>DTLZ2, DTLZ7* [8]</td>
<td><img src="image1" alt="ZDT1 and ZDT4" /></td>
<td><img src="image2" alt="ZDT2" /></td>
</tr>
<tr>
<td>DTLZ7* [8]</td>
<td><img src="image1" alt="ZDT1 and ZDT4" /></td>
<td><img src="image2" alt="ZDT2" /></td>
</tr>
</tbody>
</table>

Numerical optimization of $I_H(A,r,\mu)$ possible!

Then: plot difference to optimal hypervolume

$\rightarrow$ should decrease log-linear!
How to Change the Bias?

**Goal:** Incorporate user preferences into search (interactive optimization)
- (p)reference points, stressing extremes
- simulate classical scalarizing function approaches
- while keeping the refinement property

**Idea:** [zbt2007a]
Examples of Weight Functions

preference point

\[ w(\vec{z}) = \frac{1}{(2\pi)^k/2 |\Sigma|^{1/2} \cdot \frac{1}{2}} \]

stressing one objective

\[ w(z_1, \ldots, z_k) = \begin{cases} 
\left( \prod_{i \neq s} (b_i^u - b_i^l) \right)^{-1} \lambda e^{-\lambda(z_s - b_s^l)} & \vec{z} \in B \\
0 & \vec{z} \notin B 
\end{cases} \]
Results in 2D

stress $f_1$

preference point

Tchebycheff

desirability function

uniform

$\epsilon$-constraint

SPEA2

NSGA-II
Results in 3D

- stress $f_1$
- preference point
- Tchebycheff
- desirability function
- uniform
- $\varepsilon$-constraint
Results in 7D

- **stress $f_3$**
  - Minimize

- **preference point**

- **Tchebycheff**
  - Ideal point

- **desirability function**
**Question:**
How do optimal $\mu$-distributions for the weighted hypervolume indicator look like?

[abbz2009c, abbz2011a]
A New Idea of How to Articulate Preferences

Idea: [abbz2009c]

- compute theoretical result for weighted case
  \[ \delta(x) = \frac{\sqrt{-f'(x)w(x, f(x))}}{\int_{0}^{x_{\max}} \sqrt{-f'(x)w(x, f(x))} \, dx} \]

- use „inverse“:
  - define a desired density
  - compute the corresponding weight
  - optimize with hypervolume-based algorithm

Problems:

- theoretical result for weight on front only
- front in practice not known
- efficient calculation of the hypervolume
A New Idea of How to Articulate Preferences

Idea:

- compute theoretical result for weighted case
  \[ \delta(x) = \frac{\sqrt{-f'(x)w(x, f(x))}}{\int_0^{x_{\text{max}}} \sqrt{-f'(x)w(x, f(x))}dx} \]
- use „inverse“:
  - define a desired density
  - compute the corresponding weight
  - optimize with hypervolume-based algorithm

Problems:

- theoretical result for weight on front only (extend with const. \( w \))
- front in practice not known (assume expected front)
- efficient calculation of the hypervolume (dynamic programming)
- define density as function of angle \( \phi \) instead of \( x \)
Results 1
Results II
Hypervolume: Open Questions

Optimal $\mu$-distributions
- uniqueness
- more objectives
- other indicators
- exact results
- faster algorithms to compute them
- convergence (greedy approach, HypE)
- linear convergence

Articulating User Preferences
- changing preferences over time
- simulating other classical approaches (from AI?)
- interactive
Overview

Objective Reduction

Hypervolume-based search

**Mirroring and Sequential Selection**

- multiobjective
- single-objective
The CMA-ES \cite{ho1996a, ho2001a}

The “best” single-objective blackbox algorithm:
- Covariance Matrix Adaptation Evolution Strategy and variances
- continuous optimization

\[ x_i = m + \sigma N(0, C) \]

create $\lambda$ offspring

variation

evaluation

update

survival

take best $\mu$
The CMA-ES: Equations

**Input:** $m \in \mathbb{R}^n$, $\sigma \in \mathbb{R}_+$, $\lambda$

**Initialize:** $C = I$, and $p_c = 0$, $p_\sigma = 0$,

**Set:** $c_c \approx 4/n$, $c_\sigma \approx 4/n$, $c_1 \approx 2/n^2$, $c_\mu \approx \mu_w/n^2$, $c_1 + c_\mu \leq 1$, $d_\sigma \approx 1 + \sqrt{E_w/n}$, and $w_{i=1,...,\lambda}$ such that $\mu_w = \frac{1}{\sum_{i=1}^{\mu} w_i^2} \approx 0.3 \lambda$

**While not terminate**

\[
x_i = m + \sigma y_i, \quad y_i \sim \mathcal{N}(0, C), \quad \text{for } i = 1, \ldots, \lambda
\]

sampling

\[
m \leftarrow \sum_{i=1}^{\mu} w_i x_i; \lambda = m + \sigma y_w \quad \text{where } y_w = \sum_{i=1}^{\mu} w_i y_i; \lambda
\]

discretization

\[
p_c \leftarrow (1 - c_c) p_c + \mathbb{1}_{\{\|p_\sigma\| < 1.5 \sqrt{n}\}} \sqrt{1 - (1 - c_c)^2} \sqrt{\mu_w} y_w
\]

cumulation for $C$

\[
p_\sigma \leftarrow (1 - c_\sigma) p_\sigma + \sqrt{1 - (1 - c_\sigma)^2} \sqrt{\mu_w} C^{-\frac{1}{2}} y_w
\]

cumulation for $\sigma$

\[
C \leftarrow (1 - c_1 - c_\mu) C + c_1 p_c p_c^T + c_\mu \sum_{i=1}^{\mu} w_i y_i; \lambda y_i; \lambda^T
\]

update $C$

\[
\sigma \leftarrow \sigma \times \exp \left( \frac{c_\sigma}{d_\sigma} \left( \frac{\|p_\sigma\|}{\mathbb{E}[\|N(0, I)\|]} - 1 \right) \right)
\]

update of $\sigma$

**Not covered** on this slide: termination, restarts, useful output, boundaries and encoding

© N. Hansen & A. Auger
The CMA-ES: Ideas

\[ x_i = m + \sigma y_i, \quad y_i \sim \mathcal{N}(0, C) \]

\[ C_{\mu} \]

\[ C \leftarrow \frac{1}{\mu} \sum_{i} y_i:\lambda y_i^T:\lambda \left(1 - \frac{1}{\mu} \right) + 1 \times C_{\mu} \]

\[ m_{\text{new}} \leftarrow m + \frac{1}{\mu} \sum_{i} y_i:\lambda \]

© N. Hansen & A. Auger
Two Independent Ideas to Make Local (CMA-)ES Faster

- local: only a few children ($\lambda$ small)
- derandomized mutations
- stopping generations whenever better than parent

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

in $[-0.2, 0.8]^n$

for $n = 10$

© N. Hansen & A. Auger
**Mirrored Mutations**

**Idea**

use one random vector to generate two offspring

mutation vector independent
mirrored vector (dependent)
Mirrored Mutations

Idea

use one random vector
to generate two
offspring

Reasoning

often "good" and "bad" in opposite
directions

mutation vector
independent
mirrored vector
(dependent)
Sequential Selection

Idea

stop generation of new offspring as soon as a solution, better than the parent, is found

if all worse: selection as usual

○ parent

better offspring

worse offspring

evaluation saved
**Sequential Selection**

**Idea**
stop generation of new offspring as soon as a solution, better than the parent, is found

**Reasoning**
if sublevel sets convex one better is enough in particular with mirroring

if all worse: selection as usual

- parent
- better offspring
- worse offspring
- evaluation saved
Theoretical Results on Convergence Rates

for several variants of scale-invariant ($\sigma_t = \sigma |X_t|$) Evolution Strategies

**Theorem 4.** For a $(1, 2^s)$-ES with scale-invariant step-size ($\sigma_k = \sigma ||X_k|| > 0$) on the sphere function $g(||x||)$, for $g \in \mathcal{M}$, linear convergence holds and

$$\frac{1}{T_k} \ln \frac{\|X_k\|}{\|X_0\|} \xrightarrow{k \to \infty} \frac{1}{2} \frac{1}{2 - p_s(\sigma)} \times E \left[ \ln \left( 1 - 2\sigma \|\mathcal{N}\|_1 + \sigma^2 \|\mathcal{N}\|^2 \right) \right] \text{ a.s.}$$

where $T_k$ is the random variable for the number of function evaluations until iteration $k$, $\mathcal{N}$ is a random vector following a multivariate normal distribution, and $p_s(\sigma) = \Pr(2\|\mathcal{N}\|_1 + \sigma \|\mathcal{N}\|^2 < 0)$ is the probability that the first offspring is successful.

- can be estimated via Monte Carlo Sampling
Results: Estimated Convergence Rates

scale-invariant \((1,2^5_m)\)-ES

dimension 20

- \(c(\sigma) \cdot \text{dimension}\)
- \(\sigma \cdot \text{dimension}\)

- Extract minima
Implementaiton in CMA-ES

results on sphere

\[ \lambda = 2 \]

\[
\begin{align*}
(1,\lambda)\text{-CMA} \\
(1,\lambda^2)\text{-CMA} \\
(6/6_w,12)\text{-CMA} \\
(1,\lambda_m)\text{-CMA} \\
(1+1)\text{-CMA} \\
(1,\lambda_m^s)\text{-CMA}
\end{align*}
\]

\[ \lambda = 4 \]

results on BBOB'2010

- \((1,4_m^s)\text{-CMA-ES}\) turned out to be fastest local non-elitist strategy tested
- 3rd best of BBOB'2009/10 on Gallagher with 101 peaks (3x faster than \((1+1)\text{-CMA-ES})
- even more competitive on noisy functions
Mirroring and Sequ. Selection: Open Questions

- how to implement in $\mu/\mu_w$-CMA-ES without bias in step-size?
- further mirroring (more dependencies)
- does it make sense in multiobjective CMA-ES?
- ...

© Dimo Brockhoff, LIX, Ecole Polytechnique
Sel. Research Topics in Stochastic BB Optimization, IITK, December 3, 2010
Conclusions

1 Objective Reduction
   - idea, algorithms
   - omission and aggregation of objectives

2 Hypervolume-based Search
   - weighted hypervolume indicator
   - optimal $\mu$-distributions
   - a new way to articulate user preferences in 2D

3 Mirroring and Sequential Selection
   - idea, results
   - log-linear convergence


Announcement

EMO session @ MCDM’2011 in Jyväskylä, Finland
organizers: Dimo Brockhoff and Kalyanmoy Deb
tentative deadline: Jan. 31, 2011 (full papers & abstracts)
http://emoatmcdm.gforge.inria.fr
References I


References II


