Experimental Research in Evolutionary Computation

Thomas Bartz-Beielstein
Mike Preuss

Algorithm Engineering
Universität Dortmund

July, 9th 2006
Introduction
   Goals
   History

Comparison
   Observed Significance Plots
   Beyond the NFL

Sequential Parameter Optimization (SPO)
   Similar Approaches
   Rosenberg Study: Problem Definition and Scientific Hypothesis
   Basics

Overview
   Models
   Heuristic

SPO Toolbox (SPOT)
   Demo
   Community Discussion

Parametrized Performance
   Parametrized Algorithms
   Parameter Tuning
   Performance Measuring

Reporting and Visualization
   Reporting Experiments
   Visualization
Scientific Goals?

- Why is astronomy considered scientific—and astrology not?
- And what about experimental research in EC?

Figure: Nostradamus
Goals in Evolutionary Computation

(RG-1) *Investigation.* Specifying optimization problems, analyzing algorithms. Important parameters; what should be optimized?

(RG-2) *Comparison.* Comparing the performance of heuristics


(RG-4) *Quality.* Robustness (includes insensitivity to exogenous factors, minimization of the variability) [Mon01]
Goals in Evolutionary Computation

- Given: Hard real world optimization problems, e.g., chemical engineering, airfoil optimization, bioinformatics
- Many theoretical results are too abstract, do not match with reality
- Real programs, not algorithms
- Develop problem specific algorithms, experimentation is necessary
- Experimentation requires statistics
A Totally Subjective History of Experimentation in Evolutionary Computation

- Palaeolithic
- Yesterday
- Today
- Tomorrow
Stone Age: Experimentation Based on Mean Values

- First phase (foundation and development, before 1980)
- Comparison based on mean values, no statistics
- Development of standard benchmark sets (sphere function etc.)
- Today: Everybody knows that mean values are not sufficient
Stone Age Example: Comparison Based on Mean Values

Example (PSO swarm size)

- Experimental setup:
  - 4 test functions: Sphere, Rosenbrock, Rastrigin, Griewangk
  - Initialization: asymmetrically
  - Termination: maximum number of generations
  - PSO parameter: default

- Results: Table form, e.g.,

  Table: Mean fitness values for the Rosenbrock function

<table>
<thead>
<tr>
<th>Population</th>
<th>Dimension</th>
<th>Generation</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10</td>
<td>1000</td>
<td>96,1725</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>1500</td>
<td>214,6764</td>
</tr>
</tbody>
</table>

- Conclusion: “Under all the testing cases, the PSO always converges very quickly”
Yesterday: Mean Values and Simple Statistics

- Second phase (move to mainstream, 1980-2000)
- Statistical methods introduced, mean values, standard deviations, tutorials
- $t$ test, $p$ value, \ldots
- Comparisons mainly on standard benchmark sets
- Questionable assumptions (NFL)
Yesterday: Mean Values and Simple Statistics

Example (GAs are better than other algorithms (on average))

Theorem (NFL)

There is no algorithm that is better than another over all possible instances of optimization problems

Figure: [Gol89]
Today: Based on Correct Statistics

- Third phase (Correct statistics, since 2000)
  - Statistical tools for EC
  - Conferences, tutorials, workshops, e.g., Workshop On Empirical Methods for the Analysis of Algorithms (EMAA) ([http://www.imada.sdu.dk/~marco/EMAA](http://www.imada.sdu.dk/~marco/EMAA))
  - New disciplines such as algorithm engineering
- But: There are three kinds of lies: lies, damned lies, and statistics (Mark Twain or Benjamin Disraeli), why should we care?
- Because it is the only tool we can rely on (at the moment, i.e., 2006)
Today: Based on Correct Statistics

Example (Good practice)

Table 3: Results of the algorithms with population of 20

<table>
<thead>
<tr>
<th>Test functions</th>
<th>SGA mean best (std. dev.)</th>
<th>OGA mean best (std. dev.)</th>
<th>MGA mean best (std. dev.)</th>
<th>IGAL mean best (std. dev.)</th>
<th>t-value between SGA to the best FDGA</th>
<th>Best algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>g_1</td>
<td>8.060e-000</td>
<td>8.568e-000</td>
<td>8.654e-000</td>
<td>8.272e-000</td>
<td>-6.76 °</td>
<td>SGA</td>
</tr>
<tr>
<td>g_2</td>
<td>7.803e-000, 4.583e+000</td>
<td>4.247e-000, 1.321e+000</td>
<td>3.544e+000, 2.087e+000</td>
<td>3.509e+000, 1.497e+000</td>
<td>-4.00 °</td>
<td>SGA</td>
</tr>
<tr>
<td>g_3</td>
<td>6.495e+000</td>
<td>9.272e+000</td>
<td>8.66e+000</td>
<td>8.637e+000</td>
<td>-5.65 °</td>
<td>SGA</td>
</tr>
<tr>
<td>g_4</td>
<td>1.350e+002</td>
<td>9.200e+002</td>
<td>8.207e+002</td>
<td>8.227e+002</td>
<td>-11.69 °</td>
<td>SGA</td>
</tr>
<tr>
<td>g_5</td>
<td>3.338e+002</td>
<td>2.807e+002</td>
<td>2.599e+002</td>
<td>2.485e+002</td>
<td>-3.87 °</td>
<td>SGA</td>
</tr>
<tr>
<td>g_6</td>
<td>2.747e+002</td>
<td>6.823e+002</td>
<td>8.205e+002</td>
<td>6.247e+002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_7</td>
<td>3.062e+002</td>
<td>5.477e+002</td>
<td>5.204e+002</td>
<td>5.599e+002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_8</td>
<td>2.079e-003</td>
<td>2.705e-005</td>
<td>2.591e-005</td>
<td>2.538e-005</td>
<td>15.81 °</td>
<td>FDGA</td>
</tr>
<tr>
<td>g_9</td>
<td>9.184e-004</td>
<td>3.527e-006</td>
<td>3.329e-006</td>
<td>2.737e-006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_10</td>
<td>2.079e-003</td>
<td>4.333e+011</td>
<td>4.619e+011</td>
<td>4.006e+011</td>
<td>1.91 °</td>
<td>FDGA</td>
</tr>
<tr>
<td>g_11</td>
<td>9.184e-004</td>
<td>7.490e+012</td>
<td>8.049e+012</td>
<td>8.329e+012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_12</td>
<td>7.121e+001</td>
<td>5.015e+001</td>
<td>5.177e+001</td>
<td>4.064e+001</td>
<td>3.13 °</td>
<td>FDGA</td>
</tr>
<tr>
<td>g_13</td>
<td>7.121e+001</td>
<td>4.187e+010</td>
<td>3.757e+010</td>
<td>4.106e+010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_14</td>
<td>1.485e+001</td>
<td>5.128e+001</td>
<td>4.618e+001</td>
<td>4.656e+001</td>
<td>11.33 °</td>
<td>FDGA</td>
</tr>
<tr>
<td>g_15</td>
<td>6.257e-002</td>
<td>4.193e-003</td>
<td>1.6727e-002</td>
<td>1.285e-002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_16</td>
<td>9.212e-002</td>
<td>7.252e-002</td>
<td>6.498e-002</td>
<td>6.486e-002</td>
<td>2.94 °</td>
<td>FDGA</td>
</tr>
<tr>
<td>g_17</td>
<td>6.105e-002</td>
<td>2.138e+002</td>
<td>2.180e+002</td>
<td>2.402e+002</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The value of \( t \) with 49 degrees of freedom is significant at \( \alpha = 0.05 \) by a one tailed test.

Figure: [CAF04]
Today: Based on Correct Statistics

Example (Good practice?)

- Authors used
  - Pre-defined number of evaluations set to 200,000
  - 50 runs for each algorithm
  - Population sizes 20 and 200
  - Crossover rate 0.1 in algorithm A, but 1.0 in B
  - A outperforms B significantly in $f_6$ to $f_{10}$

- We need tools to
  - Determine adequate number of function evaluations to avoid floor or ceiling effects
  - Determine the correct number of repeats
  - Determine suitable parameter settings for comparison
  - Determine suitable parameter settings to get working algorithms
  - Draw meaningful conclusions
Today: Based on Correct Statistics

Example (Good practice?)

- Authors used
  - Pre-defined number of evaluations set to 200,000
  - 50 runs for each algorithm
  - Population sizes 20 and 200
  - Crossover rate 0.1 in algorithm $A$, but 1.0 in $B$
  - $A$ outperforms $B$ significantly in $f_6$ to $f_{10}$

- We need tools to
  - Determine adequate number of function evaluations to avoid floor or ceiling effects
  - Determine the correct number of repeats
  - Determine suitable parameter settings for comparison
  - Determine suitable parameter settings to get working algorithms
  - Draw meaningful conclusions
Today: Based on Correct Statistics

- We claim: Fundamental ideas from statistics are misunderstood!
- For example: What is the \( p \) value?

**Definition (\( p \) value)**

The \( p \) value is the probability that the null hypothesis is true.
Today: Based on Correct Statistics

- We claim: Fundamental ideas from statistics are misunderstood!
- For example: What is the $p$ value?

Definition ($p$ value)
The $p$ value is the probability that the null hypothesis is true. **No!**
Today: Based on Correct Statistics

- We claim: Fundamental ideas from statistics are misunderstood!
- For example: What is the $p$ value?

**Definition ($p$ value)**
The $p$ value is $p = P\{ \text{result from test statistic, or greater} \mid \text{null model is true} \}$

- $\Rightarrow$ The $p$ value is not related to any probability whether the null hypothesis is true or false
Tomorrow: Correct Statistics and Correct Conclusions

- Problems of today:
  Adequate statistical methods, but wrong scientific conclusions

- Tomorrow:
  - Consider scientific meaning
  - Severe testing as a basic concept
  - First Symposium on Philosophy, History, and Methodology of Error, June 2006

Scientific inquiry or problem

How to generate and analyze empirical observations and to evaluate scientific claims

Statistical inquiry: Testing hypotheses

Model of hypotheses
Model of experimental test
Model of data

(1) (2) (3)
Tomorrow: Correct Statistics and Correct Conclusions

- Generally: Statistical tools to decide whether \( a \) is better than \( b \) are necessary
- Today: Sequential parameter optimization (SPO)
  - Heuristic, but implementable approach
  - Extension of classical approaches from statistical design of experiments (DOE)
  - Other (better) approaches possible
  - SPO uses plots of the observed significance
Tests and Significance

- Plots of the observed significance level based on [May83]
- Rejection of the null hypothesis $H : \theta = \theta_0$ by a test $T^+$ based on an observed average $\bar{x}$
- Alternative hypothesis $J : \theta > \theta_0$

**Definition (Observed significance level)**

The observed significance level is defined as

$$\alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x}|\theta)$$  \hspace{1cm} (1)
Plots of the Observed Significance

- Observed significance level
  \[ \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta) \]
- Observed average \( \bar{x} = 51.73 \)

- Rejection of the null hypothesis
  \[ H : \theta = \theta_0 = 0 \]
  by a test \( T^+ \) in favor of an alternative
  \[ J : \theta > \theta_0 \]
  Then \( \hat{\alpha}(\theta) = 0.0530 \)

- Interpretation: Frequency of erroneously rejecting \( H \) ("there is a difference in means as large as \( \theta_0 \) or larger") with such an \( \bar{x} \)
Plots of the Observed Significance

- Observed significance level

\[ \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta) \]

- Observed average \( \bar{x} = 51.73 \)

- Rejection of the null hypothesis

\[ H : \theta = \theta_0 = 10 \]

by a test \( T^+ \) in favor of an alternative

\[ J : \theta > \theta_0 \]

Then \( \hat{\alpha}(\theta) = 0.0961 \)

- Interpretation: Frequency of erroneously rejecting \( H \) ("there is a difference in means as large as \( \theta_0 \) or larger") with such an \( \bar{x} \)
Plots of the Observed Significance

- Observed significance level

\[ \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta) \]

- Observed average \( \bar{x} = 51.73 \)

- Rejection of the null hypothesis

\[ H : \theta = \theta_0 = 20 \]

by a test \( T^+ \) in favor of an alternative

\[ J : \theta > \theta_0 \]

Then \( \hat{\alpha}(\theta) = 0.1607 \)

- Interpretation: Frequency of erroneously rejecting \( H \) ("there is a difference in means as large as \( \theta_0 \) or larger") with such an \( \bar{x} \)
Plots of the Observed Significance

• Observed significance level

\[ \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x}|\theta) \]

• Observed average \( \bar{x} = 51.73 \)

• Rejection of the null hypothesis

\[ H : \theta = \theta_0 = 30 \]

by a test \( T^+ \) in favor of an alternative

\[ J : \theta > \theta_0 \]

Then \( \hat{\alpha}(\theta) = 0.2485 \)

• Interpretation: Frequency of erroneously rejecting \( H \) ("there is a difference in means as large as \( \theta_0 \) or larger") with such an \( \bar{x} \)
Plots of the Observed Significance

- Observed significance level

\[ \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta) \]

- Observed average \( \bar{x} = 51.73 \)

- Rejection of the null hypothesis

\[ H : \theta = \theta_0 = 40 \]
by a test \( T^+ \) in favor of an alternative

\[ J : \theta > \theta_0 \]

Then \( \hat{\alpha}(\theta) = 0.3570 \)

- Interpretation: Frequency of erroneously rejecting \( H \) ("there is a difference in means as large as \( \theta_0 \) or larger") with such an \( \bar{x} \)
Plots of the Observed Significance

- Observed significance level
  \[ \alpha(\bar{x}, \theta) = \hat{\alpha}(\theta) = P(\bar{X} \geq \bar{x} | \theta) \]

- Observed average \( \bar{x} = 51.73 \)

- Rejection of the null hypothesis
  \[ H : \theta = \theta_0 = 50 \]
  by a test \( T^+ \) in favor of an alternative
  \[ J : \theta > \theta_0 \]
  Then \( \hat{\alpha}(\theta) = 0.4784 \)

- Interpretation: Frequency of erroneously rejecting \( H \) ("there is a difference in means as large as \( \theta_0 \) or larger") with such an \( \bar{x} \)
Small $\alpha$ Values

- Rejecting $H$ with a $T^+$ test with a small size $\alpha$ indicates that $J : \theta > \theta_0$
- If any and all positive discrepancies from $\theta_0$ are scientifically important $\Rightarrow$ small size $\alpha$ ensures that construing such a rejection as indicating a scientifically important $\theta$ would rarely be erroneous
- **Problems** if some $\theta$ values in excess of $\theta_0$ are not considered scientifically important
- Small size $\alpha$ does not prevent a $T^+$ rejection of $H$ from often being misconstrued when relating it to the scientific claim
- $\Rightarrow$ Small $\alpha$ values alone are not sufficient
Largest Scientifically Unimportant Values

- [May83] defines $\theta_{un}$ the largest scientifically unimportant $\theta$ value in excess of $\theta_0$
- But what if we do not know $\theta_{un}$?
- Discriminate between legitimate and illegitimate construals of statistical results by considering the values of $\hat{\alpha}(\theta')$ for several $\theta'$ values
**Figure:** Plots of the observed difference. *Left:* This is similar to Fig. 4.3 in [May83]. Based on $n = 50$ experiments, a difference $\bar{x} = 51.3$ has been observed, $\hat{\alpha}(\theta)$ is the area to the right of the observed difference $\bar{x}$. *Right:* The $\hat{\alpha}(\theta)$ value is plotted for different $n$ values.
OSL Plots

- Bootstrap procedure $\Rightarrow$ no assumptions on the underlying distribution necessary
- Summary:
  - $p$ value is not sufficient
  - OSL plots one tool to derive meta-statistical rules
  - Other tools needed

Figure: Same situation as above, bootstrap approach
The Art of Comparison

Orientation

The NFL\(^1\) told us things we already suspected:

- We cannot hope for the one-beats-all algorithm (solving the general nonlinear programming problem)
- Efficiency of an algorithm heavily depends on the problem(s) to solve and the exogenous conditions (termination etc.)

In consequence, this means:

- The posed question is of extreme importance for the relevance of obtained results
- The focus of comparisons has to change from:
  \textit{Which algorithm is better?}
  
to
  \textit{What exactly is the algorithm good for?}

\(^1\)no free lunch theorem
Most existing experimental studies focus on the efficiency of optimization algorithms, but:

- Adaptability to a problem is not measured, although
- It is known as one of the important advantages of EAs

Interesting, previously neglected aspects:

- Interplay between adaptability and efficiency?
- How much effort does adaptation to a problem take for different algorithms?
- What is the problem spectrum an algorithm performs well on?
- Systematic investigation may reveal inner logic of algorithm parts (operators, parameters, etc.)
Similarities and Differences to Existing Approaches

- Agriculture, industry: Design of Experiments (DoE)
- Evolutionary algorithms: Meta-algorithms
- Algorithm engineering: Rosenberg Study (ANOVA)
- Statistics: Design and Analysis of Computer Experiments (DACE)
Empirical Analysis: Algorithms for Scheduling Problems

- Problem:
  - Jobs build binary tree
  - Parallel computer with ring topology
- 2 algorithms:
  - Keep One, Send One (KOSO) to my right neighbor
  - Balanced strategy KOSO*: Send to neighbor with lower load only
- Is KOSO* better than KOSO?
Empirical Analysis: Algorithms for Scheduling Problems

- **Problem:**
  - Jobs build binary tree
    - Parallel computer with ring topology
  - 2 algorithms:
    - *Keep One, Send One (KOSO)* to my right neighbor
    - Balanced strategy KOSO*: Send to neighbor with lower load only
  - Is KOSO* better than KOSO?
Empirical Analysis: Algorithms for Scheduling Problems

- Problem:
  - Jobs build binary tree
  - Parallel computer with ring topology

- 2 algorithms:
  - Keep One, Send One (KOSO) to my right neighbor
  - Balanced strategy KOSO*: Send to neighbor with lower load only

- Is KOSO* better than KOSO?
Empirical Analysis: Algorithms for Scheduling Problems

- Problem:
  - Jobs build binary tree
  - Parallel computer with ring topology
- 2 algorithms:
  - Keep One, Send One (KOSO) to my right neighbor
  - Balanced strategy KOSO*: Send to neighbor with lower load only
- Is KOSO* better than KOSO?
Empirical Analysis: Algorithms for Scheduling Problems

- Problem:
  - Jobs build binary tree
  - Parallel computer with ring topology

- 2 algorithms:
  - Keep One, Send One (KOSO) to my right neighbor
  - Balanced strategy KOSO*: Send to neighbor with lower load only

- Is KOSO* better than KOSO?
Empirical Analysis: Algorithms for Scheduling Problems

- Hypothesis: Algorithms influence running time
- But: Analysis reveals
  - # Processors and # Jobs explain 74% of the variance of the running time
  - Algorithms explain nearly nothing
- Why?
  - Load balancing has no effect, as long as no processor starves.
  - But: Experimental setup produces many situations in which processors do not starve
- Furthermore: Comparison based on the optimal running time (not the average) makes differences between KOSO und KOSO*.
- Summary: Problem definitions and performance measures (specified as algorithm and problem design) have significant impact on the result of experimental studies
Designs

- Sequential Parameter Optimization based on
  - Design of Experiments (DOE)
  - Design and Analysis of Computer Experiments (DACE)
- Optimization run = experiment
- Parameters = design variables or factors
- Endogenous factors: modified during the algorithm run
- Exogenous factors: kept constant during the algorithm run
  - Problem specific
  - Algorithm specific
Example (Algorithm design)

Particle swarm optimization. Set of exogenous strategy parameters

- Swarm size $s$
- Cognitive parameter $c_1$
- Social parameter $c_2$
- Starting value of the inertia weight $w_{\text{max}}$
- Final value of the inertia weight $w_{\text{scale}}$
- Percentage of iterations for which $w_{\text{max}}$ is reduced
- Maximum value of the step size $v_{\text{max}}$
Example (Problem design)

Sphere function $\sum_{i=1}^{d} x_i^2$ and a set of $d$-dimensional starting points, performance measure, termination criterion

- Tuning (efficiency):
  - Given one problem instance $\Rightarrow$ determine improved algorithm parameters

- Robustness (effectivity):
  - Given one algorithm $\Rightarrow$ test several problem instances
SPO Overview

1. Pre-experimental planning
2. Scientific thesis
3. Statistical hypothesis
4. Experimental design: Problem, constraints, start-/termination criteria, performance measure, algorithm parameters
5. Experiments
7. Solution good enough?
   - Yes: Goto step 8
   - No: Improve the design (optimization). Goto step 5
8. Acceptance/rejection of the statistical hypothesis
9. Objective interpretation of the results from the previous step
• Response $Y$: Regression model and random process
• Model:
\[ Y(x) = \sum_{h} \beta_h f_h(x) + Z(x) \]
• $Z(\cdot)$ correlated random variable
• Stochastic process.
• DACE stochastic process model
• Until now: DACE for deterministic functions, e.g. [SWN03]
• New: DACE for stochastic functions
Expected Model Improvement

Design and Analysis of Computer Experiments (DACE)

Figure: Axis labels left: function value, right: expected improvement. Source: [JSW98]

(a) Expected improvement: 5 sample points
(b) Another sample point $x = 2.8$ was added
Heuristic for Stochastically Disturbed Function Values

- Latin hypercube sampling (LHS) design: Maximum spread of starting points, small number of evaluations
- Sequential enhancement, guided by DACE model
- Expected improvement: Compromise between optimization ($\min Y$) and model exactness ($\min \text{MSE}$)
- Budget-concept: Best search points are re-evaluated
- Fairness: Evaluate new candidates as often as the best one

Table: SPO. Algorithm design of the best search points

<table>
<thead>
<tr>
<th>$Y$</th>
<th>$s$</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$w_{\text{max}}$</th>
<th>$w_{\text{scale}}$</th>
<th>$w_{\text{iter}}$</th>
<th>$\nu_{\text{max}}$</th>
<th>Conf.</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.055</td>
<td>32</td>
<td>1.8</td>
<td>2.1</td>
<td>0.8</td>
<td>0.4</td>
<td>0.5</td>
<td>9.6</td>
<td>41</td>
<td>2</td>
</tr>
<tr>
<td>0.063</td>
<td>24</td>
<td>1.4</td>
<td>2.5</td>
<td>0.9</td>
<td>0.4</td>
<td>0.7</td>
<td>481.9</td>
<td>67</td>
<td>4</td>
</tr>
<tr>
<td>0.061</td>
<td>32</td>
<td>1.8</td>
<td>2.1</td>
<td>0.8</td>
<td>0.4</td>
<td>0.5</td>
<td>9.6</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>0.058</td>
<td>32</td>
<td>1.8</td>
<td>2.1</td>
<td>0.8</td>
<td>0.4</td>
<td>0.5</td>
<td>9.6</td>
<td>41</td>
<td>8</td>
</tr>
</tbody>
</table>
Data Flow and User Interaction

- User provides parameter ranges and tested algorithm
- Results from an LHS design are used to build model
- Model is improved incrementally with new search points
- User decides if parameter/model quality is sufficient to stop
SPO in Action

- Sequential Parameter Optimization Toolbox (SPOT)
- Introduced in [BB06]

- Software can be downloaded from http://ls11-www.cs.uni-dortmund.de/people/tom/ExperimentalResearchPrograms.html
SPO Installation

- Create a new directory, e.g., g:\my_spot
- Unzip SPO toolbox: http://ls11-www.cs.uni-dortmund.de/people/tom/spot03.zip
- Unzip MATLAB DACE toolbox: http://www2.imm.dtu.dk/~hbn/dace/
- Unzip ES package: http://ls11-www.cs.uni-dortmund.de/people/tom/esmatlab03.zip
- Start MATLAB
- Add g:\my_spot to MATLAB path
- Run demoSpotMatlab.m
Region of interest (ROI) files specify the region, over which the algorithm parameters are tuned.

```
name low high isint pretty
NPARENTS 1 10 TRUE 'NPARENTS'
NU 1 5 FALSE 'NU'
TAU1 1 3 FALSE 'TAU1'
```

Figure: demo4.roi
SPO Configuration file

- *Configuration* files (CONF) specify SPO specific parameters, such as the regression model

```plaintext
new=0
defaulttheta=1
loval=1E-3
upval=100
spotrmodel='regpoly2'
spotcmodel='corrgauss'
isotropic=0
repeats=3
...
```

*Figure: demo4.m*
SPO Output file

- *Design* files (DES) specify algorithm designs
- Generated by SPO
- Read by optimization algorithms

```
TAU1 NPARENTS NU TAU0 REPEATS CONFIG SEED STEP
0.210507 4.19275 1.65448 1.81056 3 1 0 1
0.416435 7.61259 2.91134 1.60112 3 2 0 1
0.130897 9.01273 3.62871 2.69631 3 3 0 1
1.65084 2.99562 3.52128 1.67204 3 4 0 1
0.621441 5.18102 2.69873 1.01597 3 5 0 1
1.42469 4.83822 1.72017 2.17814 3 6 0 1
1.87235 6.78741 1.17863 1.90036 3 7 0 1
0.372586 3.08746 3.12703 1.76648 3 8 0 1
2.82925 5.85851 2.29289 2.28194 3 9 0 1
...
```

**Figure:** demo4.des
Algorithm: Result File

- Algorithm run with settings from design file
- Algorithm writes *result file* (RES)
- RES files provide basis for many statistical evaluations/visualizations
- RES files read by SPO to generate stochastic process models

```
Y NPARENTS FNAME ITER NU TAU0 TAU1 KAPPA NSIGMA RHO DIM CONFIG SEED
3809.15 1 Sphere 500 1.19954 0 1.29436 Inf 1 2 2 1 1
0.00121541 1 Sphere 500 1.19954 0 1.29436 Inf 1 2 2 1 2
842.939 1 Sphere 500 1.19954 0 1.29436 Inf 1 2 2 1 3
2.0174e-005 4 Sphere 500 4.98664 0 1.75367 Inf 1 2 2 2 1
0.000234033 4 Sphere 500 4.98664 0 1.75367 Inf 1 2 2 2 2
1.20205e-007 4 Sphere 500 4.98664 0 1.75367 Inf 1 2 2 2 3
...
```

**Figure:** demo4.res
Summary: SPO Interfaces

- SPO requires CONF and ROI files
- SPO generates DES file
- Algorithm run with settings from DES
- Algorithm writes *result file* (RES)
- RES files read by SPO to generate stochastic process models
- RES files provide basis for many statistical evaluations/visualizations (EDA)

**Figure: SPO Interfaces**
SPO live

- Tuning of an ES on the sphere (demo4)
- Compare best from initial LHD and tuned design (demo5)
- Include recommendations from literature (demo6)
- How do the results change if the dimension is increased? (demo8)
SPO and EDA

- Interaction plots
- Main effect plots
- Regression trees
- Scatter plots

- Box plots
- Trellis plots
- Design plots
- ...

Plotted against TAU1
How to Perform an Experimental Analysis

- **Scientific claim:** “ES with small populations perform better than ES with larger ones on the sphere.”

- **Statistical hypotheses:**
  - ES with, say \( \mu = 2 \), performs better than ES with \( mu > 2 \) if compared on problem design \((1)\)
  - ES with, say \( \mu = 2 \), performs better than ES with \( mu > 2 \) if compared on problem design \((2)\)
  - . . .
  - ES with, say \( \mu = 2 \), performs better than ES with \( mu > 2 \) if compared on problem design \((n)\)
SPO Open Questions

- Models?
  - (Linear) Regression models
  - Stochastic process models
- Designs?
  - Space filling
  - Factorial
- Statistical tools
- Significance
- Standards
SPOT Community

- Provide SPOT interfaces for important optimization algorithms
- Simple and open specification
- Currently available (April 2006) for the following products:

<table>
<thead>
<tr>
<th>Program</th>
<th>Language</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolution Strategy</td>
<td>JAVA, MATLAB</td>
<td><a href="http://www.springer.com/3-540-32026-1">http://www.springer.com/3-540-32026-1</a></td>
</tr>
<tr>
<td>Genetic Algorithm and Direct Search Toolbox</td>
<td>MATLAB</td>
<td><a href="http://www.mathworks.com/products/gads">http://www.mathworks.com/products/gads</a></td>
</tr>
<tr>
<td>Particle Swarm Optimization Toolbox</td>
<td>MATLAB</td>
<td><a href="http://psotoolbox.sourceforge.net">http://psotoolbox.sourceforge.net</a></td>
</tr>
</tbody>
</table>
Discussing SPO

- SPO is not the final solution—it is one possible (but not necessarily the best) solution
- Goal: continue a discussion in EC, transfer results from statistics and the philosophy of science to computer science
What is the Meaning of Parameters?

Are Parameters “Bad”?  

Cons:

- Multitude of parameters dismays potential users
- It is often not trivial to understand parameter-problem or parameter-parameter interactions
  - Parameters complicate evaluating algorithm performances

But:

- Parameters are simple handles to modify (adapt) algorithms
- Many of the most successful EAs have lots of parameters
- New theoretical approaches: Parametrized algorithms / parametrized complexity, (“two-dimensional” complexity theory)
Possible Alternatives?

Parameterless EAs:

- Easy to apply, but what about performance and robustness?
- Where did the parameters go?

Usually a mix of:

- Default values, sacrificing top performance for good robustness
- Heuristic rules, applicable to *many* but not *all* situations; probably not working well for completely new applications
- (Self-)Adaptation techniques, these cannot learn too many parameter values at once, and not necessarily reduce the number of parameters

⇒ We can reduce number of parameters, but usually at the cost of either performance or robustness
Parameter Control or Parameter Tuning?

The time factor:
- Parameter control: during algorithm run
- Parameter tuning: before an algorithm is run

But: Recurring tasks, restarts, or adaptation (to a problem) blur this distinction

And: How to find meta-parameter values for parameter control?
⇒ Parameter control *and* parameter tuning
Tuning and Comparison

What do Tuning Methods (e.g. SPO) Deliver?

- A best configuration from \( \{ \text{perf}(\text{alg}(\text{arg}_t^{\text{exo}})) | 1 \leq t \leq T \} \) for \( T \) tested configurations
- A spectrum of configurations, each containing a set of single run results
- A progression of current best tuning results
How do Tuning Results Help?
...or Hint to new Questions

What we get:
- A near optimal configuration, permitting top performance comparison
- An estimation of how good any (manually) found configuration is
- A (rough) idea how hard it is to get even better

*No excuse:* A first impression may be attained by simply doing an LHS

Yet unsolved problems:
- How much amount to put into tuning (fixed budget, until stagnation)?
- Where shall we be on the spectrum when we compare?
- Can we compare spectra (⇒ adaptability)?
“Traditional” Measuring in EC

*Simple Measures*

- MBF: mean best fitness
- AES: average evaluations to solution
- SR: success rates, $SR(t) \Rightarrow$ run-length distributions (RLD)
- best-of-n: best fitness of $n$ runs

But, even with all measures given: Which algorithm is better?

(figures provided by Gusz Eiben)
Success Performances:

- **SP1 [HK04]** for equal expected lengths of successful and unsuccessful runs $\mathbb{E}(T^s) = \mathbb{E}(T^{us})$:
  \[
  SP1 = \frac{\mathbb{E}(T^s)}{\rho_s} 
  \]  
  (2)

- **SP2 [AH05]** for different expected lengths, unsuccessful runs are stopped at $FE_{max}$:
  \[
  SP2 = \frac{1 - \rho_s}{\rho_s} FE_{max} + \mathbb{E}(T^s_A) 
  \]  
  (3)

Probably still more aggregated measures needed (parameter tuning depends on the applied measure)
Choose the Appropriate Measure

- Design problem: Only best-of-n fitness values are of interest
- Recurring problem or problem class: Mean values hint to quality on a number of instances
- Cheap (scientific) evaluation functions: exploring limit behavior is tempting, but is not always related to real-world situations

In real-world optimization, $10^4$ evaluations is a lot, sometimes only $10^3$ or less is possible:
- We are relieved from choosing termination criteria
- Substitute models may help (Algorithm based validation)
- We encourage more research on short runs

Selecting a performance measure is a very important step
Current “State of the Art”

Around 40 years of empirical tradition in EC, but:

- No standard scheme for reporting experiments
- Instead: one (“Experiments”) or two (“Experimental Setup” and “Results”) sections in papers, providing a bunch of largely unordered information
- Affects readability and impairs reproducibility

Other sciences have more structured ways to report experiments, although usually not presented in full in papers. Why?

- Natural sciences: Long tradition, setup often relatively fast, experiment itself takes time
- Computer science: Short tradition, setup (implementation) takes time, experiment itself relatively fast

⇒ We suggest a 7-part reporting scheme
Suggested Report Structure

ER-1: **Focus/Title** the matter dealt with
ER-2: **Pre-experimental planning** first—possibly explorative—program runs, leading to task and setup
ER-3: **Task** main question and scientific and derived statistical hypotheses to test
ER-4: **Setup** problem and algorithm designs, sufficient to replicate an experiment
ER-5: **Experimentation/Visualization** raw or produced (filtered) data and basic visualizations
ER-6: **Observations** exceptions from the expected, or unusual patterns noticed, plus additional visualizations, no subjective assessment
ER-7: **Discussion** test results and necessarily subjective interpretations for data and especially observations

This scheme is well suited to report 12-step SPO experiments
Objective Interpretation of the Results

Comparison. Run-length distribution

Empirical CDF

Number of function evaluations

F(X)

PSO

PSO*

PSOC

PSOC*

Bartz-Beielstein/Preuss (Universität Dortmund)
(Single) Effect Plots

Useful, but not Perfect

- Large variances originate from averaging
- The $\tau_0$ and especially $\tau_1$ plots show different behavior on extreme values (see error bars), probably distinct (averaged) effects/interactions
One-Parameter Effect Investigation

Effect Split Plots: Effect Strengths

- Sample set partitioned into 3 subsets (here of equal size)
- Enables detecting more important parameters visually
- Nonlinear progression 1–2–3 hints to interactions or multimodality
Two-Parameter Effect Investigation

*Interaction Split Plots: Detect Leveled Effects*
Updates

- Please check [http://ls11-www.cs.uni-dortmund.de/people/tom/ExperimentalResearchSlides.html](http://ls11-www.cs.uni-dortmund.de/people/tom/ExperimentalResearchSlides.html) for updates, software, etc.
Discussion

- Standards for good experimental research
- Review process
- Research grants
- Meetings
- Building a community
- Teaching
- ...

Bartz-Beielstein/Preuss (Universität Dortmund)
Experimental Research
July, 9th 2006


Nikolaus Hansen and Stefan Kern.


